Inférence en génétique spatiale des populations vue par le prisme de la coalescence

Raphaël Leblois

UNIVERSITÉ DE MONTPELLIER
ÉCOLE DOCTORALE GAIA

DOCUMENT DE SYNTHÈSE

en vue d'une candidature à une

Habilitation à diriger des recherches

Inférence en génétique spatiale des populations vue par le prisme de la coalescence

par

Raphaël LEBLOIS

Soutenue le XX octobre 2023 devant le jury composé de

Lounès CHIKHI	Directeur de Recherche, CNRS Toulouse	Examinateur
Flora JAY	Chargée de Recherche, Univ. Paris Saclay	Examinatrice
Amaury LAMBERT	Professeur, Collège de France	Rapporteur
Anna-Sapfo MALSPINAS	Professeure, Univ. de Lausanne	Rapportrice
Denis ROZE	Directeur de Recherche, CNRS Roscoff	Rapporteur
Laure SEGUREL	Chargée de Recherche, CNRS Lyon	Examinatrice
Céline SCORNAVACCA	Directrice de Recherche, CNRS Montpellier	Examinatrice

Membre invité au jury :

John NOVEMBRE	Professeur, Université de Chicago

Table des matières

Avant-propos

Structure de ce document

Ce document résume l'ensemble des travaux de recherche que j'ai réalisé durant toute ma carrière de 1999 à aujourd'hui, (i) lors de mon DEA (Diplôme d'études Approfondies en écologie et évolution) et ma thèse sous la direction d'Arnaud Estoup et François Rousset de 1999 à 2004, (ii) de mon post-doctorat à l'Université de Californie, Berkeley, avec Monty Slakin en 2005, (iii) de mon séjour au Muséum National d'Histoire Naturelle en tant que maître de conférence de 2006 à 2010, puis (iv) en tant que chargé de recherche INRAE au CBGP de 2010 jusqu'à aujourd'hui. Ces travaux ont été initiés par des longues discussions d'introduction à la génétique des populations, avec Arnaud Estoup lors de mon stage de maîtrise réalisé à l'Université du Queensland en Australie dans le laboratoire de Zoologie dirigé par Craig Moritz, sur l'estimation de la dispersion chez le crapaud de la canne à sucre (Figure 1), espèce de lutte biologique devenue envahissante.

Je me permets de repartir d'aussi loin car depuis cette première expérience de recherche, ma carrière a été extrêmement linéaire et cohérente, toujours axée sur le développement méthodologique, le test et l'application de différentes approches d'inférences démographiques et historiques à partir de données génétiques, avec un fort accent sur la génétique spatiale des populations et l'estimation des paramètres locaux et actuels de dispersion, de tailles et de densité des populations.

Du point de vue de sa forme, ce document de synthèse comprend trois parties principales et deux annexes :

- Une première partie, courte, résumant l'ensemble de mes travaux dans l'esprit de l'HDR de faire un bilan de la trajectoire des recherches passées.
- La seconde partie, la plus scientifique, développe en 5 chapitres la partie de mes recherches portant sur l'inférence de paramètres démographiques actuels et locaux sous des modèles spatialisés de génétique des population dit d'"isolement par la distance". Ça a été le fil directeur de toute ma carrière, et ce le sera probablement jusqu'à ma retraite. Après avoir introduit les principales notions de génétique des populations nécessaires pour la suite, ce second chapitre reprend les principaux arguments théoriques et expérimentaux qui ont servi de base à mes travaux et les replace dans un contexte global et historique de la ("ma" ?) génétique spatiale des populations.
- La troisième partie est la continuation directe des deux précédentes, et présente mon projet de recherche à court et moyen terme.

Les partie 1 et 3 ne sont pas indispensable à la compréhension du document scientifique, et peuvent donc être lu indépendamment. J'ai tenté de rédiger la partie 2 comme une introduction à la génétique spatiale des populations, et plus spécifiquement l'inférence de la dispersion et des densités de population à partir de données génétiques, à destination par exemple des étudiants de master ou en thèse.

Une première annexe comporte la "fiche de synthèse" ainsi qu'un résumé de mes collaborations, productions, financements obtenus et étudiants encadrés, la seconde présente un recueil d'articles de génétique spatiale des populations auxquels j'ai participé, essentiellement en lien avec la partie 2. Ces deux annexes ainsi que les parties 1 et 2 représentent donc la partie administrative demandée pour l'HDR. La partie 2, la plus importante pour moi, s'inspire de certaines partie de ma thèse, des mes publications, de la thèse de Thimothée Virgoulay (Virgoulay, 2022) ainsi que du livre de Rousset (2004).

FIGURE 1 – Un beau spécimen de crapaud de la cane a sucre (“cane toad”, *Rhinella marina*). © Queensland Department Of Environment and Science/Reuters.

Vocabulaire

Précisons dés à présent quelques termes et notations utilisées dans ce document :

On appellera *gène* la copie d'une information génétique. Cette définition a un sens immédiat lorsque l'on s'intéresse à la partie du patrimoine génétique qui contribue à l'expression du phénotype des individus. La notion de variation (ou de *polymorphisme*) génétique implique que les différentes copies d'une même information (ou gènes) ne sont pas nécessairement identiques. Pour certains marqueurs génétiques dits évolutivement *neutres* (qui, par définition, ne codent pas d'information génétique), on ne retiendra de cette définition du gène que la notion de copie de matériel génétique. Un individu diploïde possède deux copies de la même information génétique. Chez une espèce où l'hérédité est biparentale, l'une des deux copies provient du père, tandis que l'autre copie provient de la mère. On appellera *locus* la classe d'homologie d'un gène, en ce sens que seuls deux gènes homologues peuvent "*ségréger*". Enfin, un *allèle* (ou *état allélique*) représentera une classe de gènes tous équivalents. Selon ces définitions, deux gènes sont donc dans le même état allélique, si l'information qu'ils portent est codée par la même séquence d'ADN, ou s'ils sont la copie exacte d'un ancêtre commun.

Dans ce qui suit, nous considérerons un *échantillon* de gènes constitué de plusieurs *sous-échantillons*. L'échantillon est pris dans ce qu'on appellera une *population*, potentiellement structurée en *sous-populations*, ou *dèmes*, dans lesquelles sont pris les sous-échantillons. Cette structuration peut être la conséquence d'une dispersion localisée (c.a.d. limitée dans l'espace) ou de barrières à la dispersion. On considérera que la reproduction est *panmictique* (c.a.d. union aléatoire des gamètes de la population) au sein de chaque sous-population et non dans la population entière. Dans ce document, on supposera toujours que les gènes auxquels on s'intéresse sont homologues (situés sur un même locus), et que les approches se généralisent à un échantillon multilocus en considérant des locus indépendants, sauf si précisé différemment.

Un *modèle*, servant à étudier un *processus*, est défini par des *paramètres*, dit *canoniques*, tels que les tailles de populations (N), des taux de mutations ($/mu$) et des taux des migrations (m). Nous considérerons que toutes quantités qui sont uniquement fonction des paramètres qui définissent le modèle (e.g. les paramètres *composites* comme $\theta = 2N\mu$, les probabilités d'identité ou les F-statistiques), sont aussi des paramètres. Les valeurs prises par ces paramètres dans une population "vraie" seront des *statistiques* ou des *estimations*, que l'on pourra mesurer dans cette population à l'aide d'*estimateurs*. L'estimateur et les statistiques sont alors traités comme des *variables aléatoires*. Les *espérances* de ces variables aléatoires sont fonctions des valeurs des paramètres du modèle. Si un estimateur est sans biais, son espérance correspond alors exactement aux valeurs des paramètres du modèle.

Cycle de vie

Tous les organismes passent par une séquence de changements se répétant de manière cyclique à travers les générations : on parle de cycle de vie (Bell & Koufopanou, 1991). Dans le cas des organismes sexués ce cycle est dit "haplo-diploïde" puisqu'il consiste en l'alternance d'une phase haploïde à n chromosomes et d'une phase diploïde à $2n$ chromosomes (Figure 2, à l'exception notable de certaines algues rouges tel que *Antithamnionella* sp. pour lesquels il existe une phase supplémentaire). Ce cycle de vie peut être, en termes de temps passé dans chaque phase, majoritairement haploïde (comme chez de nombreuses algues, e.g *Spirogyra* sp.), majoritairement diploïde (comme chez la plupart des vertébrés e.g *Homo sapiens*) ou un équilibre entre les deux (e.g *Laminaria digitata*, de nombreux insectes sociaux).

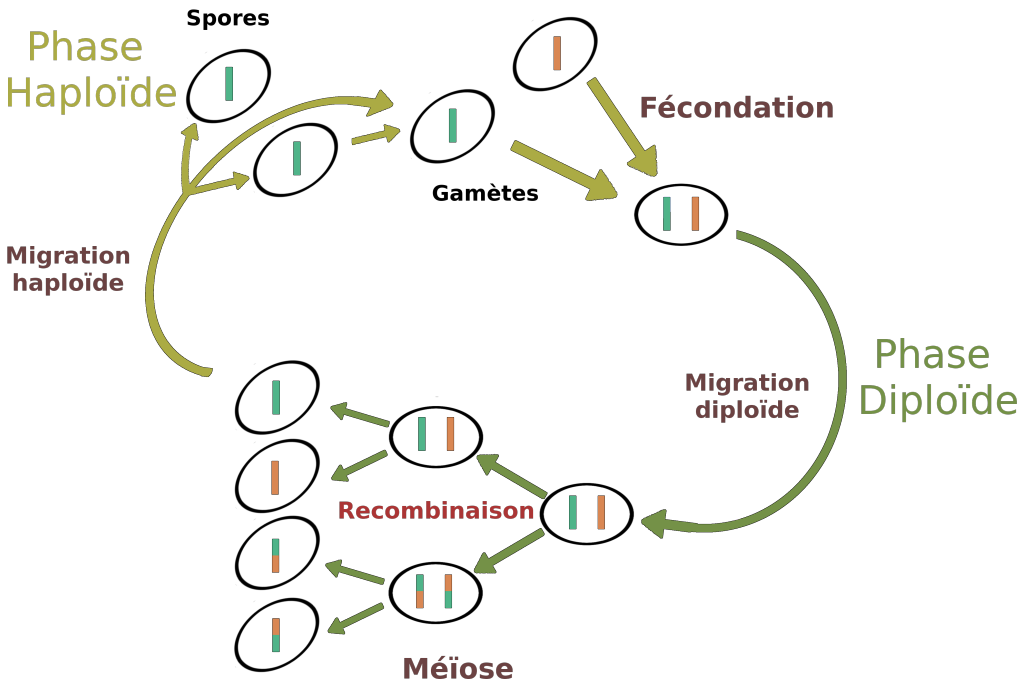


FIGURE 2 – Cycle de vie "haplo-diploïde" représentant le passage des chromosomes au cours des différentes phases possibles d'un organisme. Ici un chromosome de la paire de chromosomes maternels (en vert) et de la pair de chromosome paternels (en orange) échangent du matériel génétique durant la méiose. Figure issue de la thèse de Thimothée Virgoulay, 2022.

Le cycle de vie considéré dans tout ce document est donc composé de 5 étapes définissant une génération : (i) à chaque événement de reproduction, chaque adulte donne naissance à une infinité de gamètes et meurt ; (ii) les gamètes subissent l'effet de la mutation ; (iii) les gamètes dispersent (facultatif) ; (iv) dans chaque dème, des juvéniles diploïdes sont formés à partir du pool de gamètes ; (v) les juvéniles dispersent (facultatif) et (v) la compétition ramène le nombre d'adultes à N.

Dans certains cas, nous considérerons que la dispersion n'est que gamétique, dans d'autres que juvéniles, mais ne nous considérerons pas ici de cycles avec les deux types de dispersions (par ex. la dispersion du pollen et des graines chez les plantes).

Déclaration de valeurs et principes

" Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche."

Courte synthèse de mon parcours scientifique

Mon travail a toujours été partagé entre, d'une part, la mise au point, le test et la mise à disposition de la communauté de nouvelles méthodes d'analyses de données de génétique des populations, et d'autre part, l'application de ces méthodes (et d'autres) à des jeux de données réelles, le plus souvent en collaboration étroite avec les chercheurs ayant produit ces données. A cela s'ajoute une participation relativement constante à différentes formations, notamment à travers l'organisation de différents modules d'enseignement, et l'encadrement régulier d'un à quatre étudiants par an.

Je me suis toujours placé à l'interface entre les théoriciens (généticiens des populations théoriques, statisticiens, mathématiciens) et les empiristes, place que je trouve spécialement importante et intéressante. Ceci a été possible grâce à cette répartition de mon temps de travail entre les aspects méthodologiques et les applications empiriques, toujours à travers des collaborations, que ce soit avec des théoriciens à la pointe de ce qui se fait en statistiques/génétique des populations ou lors d'échanges poussés avec des spécialistes des modèles biologiques, de la mise au point des projets à l'analyses finale des données, en passant par l'échantillonnage sur le terrain.

Le fil directeur de mes recherches a toujours été l'estimation de paramètres démographiques à partir de données génétiques. Mes principales "ré-orientations" ont été naturellement dictées par l'environnement scientifique dans lequel j'ai évolué. Ainsi, lors de mon arrivée au Muséum, je me suis plus intéressé à des approches touchant à l'inter-specifique ("Barcode ADN" et modèles de divergence avec flux de gènes) et à des applications touchant principalement à la description de la biodiversité. Cependant, j'étais assez isolé du point de vue méthodologique, et j'ai donc choisi d'intégrer le CBGP, dont le cadre est plus favorable aux développements théoriques. Cela a été une belle opportunité, j'évolue maintenant dans un cadre optimal pour les développements méthodologiques. Petit à petit, je me suis donc aussi tourné vers des collaborations plus locales pour des applications concernant des organismes d'intérêt agronomique, des espèces envahissantes et/ou importantes pour la santé humaine, étudiées au CBGP.

Dans le même temps, la génération de gros jeux de données NGS est devenue possible sur des organismes non-modèles et je me suis donc naturellement mis à l'analyse de données NGS. Plus récemment, depuis maintenant 5 ans, j'ai repris des travaux de développements méthodologiques en génétique des populations spatialisées, un des aspects de la génétique des populations que je préfère mais que j'avais mis un peu de côté. Pour cela, j'ai décidé de me focaliser sur des approches d'inférences par simulation basées sur le développement d'un nouveau simulateur de données génomiques en populations spatialisées le plus flexible possible. C'est maintenant mon principal axe de travail et cela le sera probablement pendant de

nombreuses années puisque la génétique des population spatialisées est en plein essor au CBGP et s'ancre bien dans une thématique qui m'est chère : l'utilisation de la génétique des populations pour mieux comprendre la fonctionnement actuel et récent des populations ; et ainsi participer à la mise au point de nouveaux outils de gestion durable des populations dans des contextes agro-écologique, de santé ou de biologie de la conservation.

Motivations

Mes recherches concernent la génétique des populations et je m'intéresse donc aux mécanismes qui sous-tendent l'évolution des gènes dans les populations, non seulement pour mieux comprendre l'histoire évolutive des populations et des espèces, mais également pour participer à la mise en place de mesures efficaces pour la gestion d'un grand nombre d'espèces d'intérêt socio-économique.

Le développement des techniques de biologie moléculaires durant les 20 dernières années a permis d'augmenter considérablement la masse des données disponibles dans le cadre d'études du polymorphisme génétique au niveau des populations. La nécessité d'un traitement efficace de ces données est plus que jamais d'actualité. Parallèlement à cette explosion de données génomiques, les outils d'analyse en génétique des populations ont connu ces dernières années une évolution importante, du fait notamment de l'utilisation croissante de techniques statistiques avancées, reposant sur la vraisemblance, des méthodes d'inférences basées sur la simulation et/ou sur des méthodes d'intelligence artificielle (IA). Ces méthodes « modernes », qui nécessitent une puissance de calcul importante, permettent (1) soit de prendre en compte de l'ensemble de l'information contenue dans les données de polymorphisme (vraisemblance), (2) soit de résumer un maximum de cette information dans un large ensemble de statistiques résumantes (ABC-like, IA) et sont donc beaucoup plus puissantes que d'autres méthodes fondées par exemple sur l'utilisation d'un petit nombre de statistiques résumées (par ex. FST). L'objectif global de mes recherches est de contribuer au développement de ces méthodes, qui ouvrent de nouveaux horizons pour l'inférence démographique et historique à partir de données génétiques.

Enfin, la démocratisation des technologies de production de données génomiques à haut débit autorise désormais l'application de ces méthodes à un nombre croissant d'organismes qu'ils soient « modèles » ou « non-modèles ». Ceci ouvre des perspectives immenses car un très grand nombre de domaines d'application possibles concerne des questions d'ordre économique, sociétal et environnemental, notamment :

- Pour les espèces invasives ou en déclin, l'inférence de l'histoire démographique passée est primordiale pour proposer des mesures de gestion efficaces. Dans ce contexte, comment identifier et dater les changements démographiques passés, pour mieux analyser les dynamiques de populations en déséquilibre démographique, caractéristiques des espèces invasives ou menacées ?
- Des analyses précises des dynamiques d'invasion d'espèces bio-invasives ou d'agents pathogènes passent par une bonne compréhension des capacités de dispersion des organismes. Dans ce contexte, comment mieux caractériser l'organisation spatiale de la diversité génétique ? Comment mesurer de façon précise, à partir de cette organisation spatiale, les capacités de dispersion des individus ?
- L'application d'approches agro-écologiques à un agrosystème nécessitent souvent de mieux connaître la biologie des organismes d'intérêt agronomiques,

notamment le fonctionnement démographique local de leurs populations afin de mieux comprendre les dynamiques populationnelles à différentes échelles, de la parcelle au bassin de production. Peut-on alors, grâce aux outils de la génétique des populations, avoir accès à de l'information précise sur les tailles de populations, leurs densités ainsi que les capacités de dispersion de ces organismes, notamment en lien avec les éléments du paysage au voisinage des cultures ?

Axes de recherche

Historiquement, la génétique des populations s'est développée le long de deux axes principaux : (1) la génétique des populations « théorique » s'est intéressée, par le biais de modèles mathématiques, aux « forces évolutives » qui déterminent l'évolution des gènes dans les populations ; (2) la génétique des populations « empirique » s'est focalisée quant à elle sur la mesure, toujours plus précise, du niveau et de la distribution du polymorphisme génétique dans les populations naturelles ou artificielles. Le lien entre ces deux axes théoriques et empiriques n'a pas toujours été aussi fort qu'il pourrait être, dans la mesure où les méthodes utilisées pour analyser les données réelles reposent généralement sur des modèles démo-génétiques simples et des hypothèses assez peu réalistes (équilibre démographique, populations clairement délimitées dans l'espace, paramètres de dispersion constants dans le temps et dans l'espace, etc.). De plus, l'organisation spatiale des individus et des populations, bien qu'elle soit fondamentale pour la dynamique de la reproduction des individus dans de très nombreuses espèces, n'est en général pas prise en compte dans les modèles. Enfin, d'un point de vue plus pratique, les méthodes statistiques d'analyses de données développées par les théoriciens ne sont pas toujours faciles à appréhender car souvent très complexes mathématiquement/statistiquement, ni faciles à utiliser par les généticiens des populations empiriques car les outils mis à disposition de la communauté sont parfois mal « carrossés ».

Pourtant, il est désormais possible (1) de développer des modèles qui permettent l'analyse de scénarios démographiques complexes ; (2) d'utiliser ces modèles pour l'inférence statistique de paramètres démographiques à partir de données de polymorphisme génétique. Ces avancées méthodologiques ont notamment été possibles grâce à la combinaison de deux types de développements théoriques majeurs : (1) la théorie de la coalescence, qui offre un modèle probabiliste des généalogies de gènes dans les populations et qui permet ainsi de simuler des généalogies sous un grand nombre de modèles démographiques ; (2) Des méthodes d'inférence statistique performantes, utilisant au maximum l'information présente dans les données, telles que (a) les méthodes de Monte Carlo, qui rassemblent un grand nombre d'algorithmes comme l' « échantillonnage pondéré » (Importance Sampling, IS), les algorithmes de Monte Carlo par chaînes de Markov (Monte Carlo Markov Chains, MCMC-coa), les algorithmes de Markov cachés (Hidden Markov Models, HMM) ou (b) les méthodes d'inférence par simulation (par ex. Approximate Bayesian Computations, ABC) incorporant de plus en plus souvent des algorithmes d'intelligence artificielle (par ex. Random-Forest, Résaux de neurones). Enfin, le développement de logiciels d'analyse bien documentés et accompagnés d'interfaces graphiques permet une utilisation plus large et plus aisée par des non-spécialistes.

Après avoir testé une méthode des moments, basé sur les F-statistiques, pour l'inférence de la dispersion à partir de l'augmentation de la différenciation avec la distance, j'ai commencé à utiliser ces méthodes dès mon travail de thèse sur l'estima-

tion de la dispersion à partir de données microsatellites. J'ai continué dans la même direction pendant mes 4 ans au MNHN, en essayant d'ajouter à mes recherches une dimension inter-spécifique adaptée à la thématique « systématique » du MNHN. Mes recherches actuelles au CBGP ont toujours pour ambition de développer des méthodes d'inférence statistique innovantes et d'adapter celles déjà existantes, dans le but d'augmenter le réalisme et la pertinence des modèles démo-génétiques sous-jacents, mais aussi pour étendre leur champ d'application. Mes recherches visent notamment à (1) contribuer au développement de méthodes d'inférence de la démographie actuelle et historique des populations ; (2) évaluer leurs performances pour inférer la la dispersion et les densités à une échelle spatiale fine, potentiellement en relation avec le paysage, à l'aide de données simulées et leur pertinence lors d'applications à des données réelles ; tout cela en (3) mettant à disposition de la communauté scientifique des logiciels d'analyse de données robustes et faciles à utiliser.

Inférences démographiques intra-spécifique par vraisemblance

Les méthodes de maximum de vraisemblance (Maximum Likelihood) ont un rôle central en statistiques, car elles permettent de capturer toute l'information contenue dans les données. Leur développement en génétique des populations a été freiné par l'absence d'expression mathématique simple pour la vraisemblance dans la plupart des modèles. Dans les années 90, des algorithmes de simulation ont été développés pour approximer la vraisemblance des données en utilisant la simulation de généalogies des gènes échantillonnés basées sur la théorie de la coalescence. Ces algorithmes appartiennent à une classe de méthodes de Monte Carlo par chaînes de Markov (MCMC-coa), développés à l'origine par Felsenstein et ses collègues (Kuhner et al. 1995, Felsenstein et al. 1999), ainsi qu'à une classe de techniques d'échantillonnage pondéré (IS-coa) des histoires généalogiques, introduites par Griffiths et ses collègues (Griffiths et Tavaré 1994 ; De Iorio et Griffiths 2004a, b). Les algorithmes MCMC-coa, plus simples à implémenter, ont été plus largement développés (voir, par exemple Kuhner et al. 1998, Beerli et Felsenstein 2001, Hey et Nielsen 2007), bien qu'ils soient en général plus lents et que leur convergence soit souvent difficile à évaluer. Contrairement aux MCMC-coa, les algorithmes IS-coa sont parallisables, ne nécessitent pas de convergence vers un état stationnaire, mais sont nettement plus compliqué à adapter à différents modèles démographiques et mutationels (Rousset *et al.*, 2018).

J'ai commencé à utiliser ces algorithmes IS-coa pendant ma thèse avec F. Rousset en collaboration avec R. Griffiths sur un modèle à deux populations échangeant des migrants. Nous avons ensuite étendu la méthode pour considérer un modèle d'isolement par la distance (IBD) en une dimension en 2007, puis en deux dimensions en 2012. Ces travaux nous ont conforté dans notre choix de concentrer nos efforts sur ces approches IS-coa. En effet, nos résultats montrent une très bonne efficacité des méthodes IS-coa par rapport aux MCMC-coa, en termes de précision des estimations, de couverture des intervalles de confiance, et des temps de calculs tout à fait compétitifs. De plus ces approches ont l'avantage de pouvoir être complètement parallélisables contrairement au MCMC-coa. Dans le contexte de l'augmentation constante du nombre de cœurs des ordinateurs, c'est un avantage crucial qui nous permet d'analyser des jeux de données en un temps raisonnable.

De 2010 à 2018, j'ai principalement travaillé sur l'extension de ces méthodes à des modèles en déséquilibre démographiques, en commençant par un modèle simple

d'une population ayant subi un changement passé de taille de population, c.a.d. contraction ou expansion (Leblois et al. 2014). Cette méthode a l'avantage, par rapport aux méthodes existantes, de permettre non seulement une meilleure détection des changements passés mais aussi d'être plus robuste aux processus mutationnels complexes des marqueurs microsatellites grâce à l'implémentation d'un modèle de mutation par pas généralisé (GSM, implémenté en collaboration avec P. Pudlo, statisticien à l'université de Montpellier qui a passé un an en délégation au CBGP). Nous avons aussi implémenté : (1) avec Coralie Merle (Stagiaire de M2 en 2013, puis en thèse de 2013 à 2016) de nouveaux algorithmes de ré-échantillonnage plus efficaces (Merle et al. 2017) ; (2) avec Champak Reddy Beeravolu (Post doctorant 2011-2014), différents modèles mutationnels adaptés aux NGS (Séquences ADN et SNPs) alors que nous avions jusqu'alors uniquement considéré l'analyse de données microsatellites ; et (3) un modèle de type « Fondation-Expansion » à deux changements démographiques dans lequel une nouvelle population est fondée à partir d'un petit nombre d'individus issus d'une population ancestrale inconnue (effet de fondation), et croit ensuite de manière exponentielle (phase d'expansion). Ce modèle paraît très pertinent pour de nombreuses études, notamment dans le cadre des bioinvasions, sujet d'étude majeur au CBGP. Tout ces développements ont été décrits dans un article résumant nos travaux sur l'IS (Rousset et al. 2018) et appliqué, en collaboration, à de nombreux jeux de données (par ex. Macedo et al. 2019, Juhel et al. 2019 ; Tournayre et al. 2019, Ledoux et al. 2018, Pestopoulos 2018, Wereszczuk et al. 2017, Lalis et al. 2016a, b, Berthier et al. 2016, Zenboudji et al. 2016, Vignaud et al. 2014a, b). Enfin, tous ces développements sont implémentés dans notre logiciel Migraine écrit en C++ et en R (voir section 3.2.2). Cette partie de mes recherche sur l'inférence par vraisemblance à partir de données génétiques est détaillé dans le chapitre 3.

Il existe cependant un grand nombre de situations où la vraisemblance ne peut pas être calculée, ni même estimée par le biais de techniques IS-coa ou MCMC-coa. Pour de telles situations, des méthodes reposant sur un calcul approché de la vraisemblance ont été développées (voir Tavaré et al. 1997, Pritchard et al. 1999, Beaumont et al. 2002). Certaines de ces méthodes sont connues sous l'acronyme ABC (pour Approximate Bayesian Computations) et le CBGP est un des leaders mondiaux dans ce domaine. Ainsi, j'ai commencé en 2018 à travailler sur des méthodes d'inférence par simulation, en collaboration avec François Rousset (CNRS, ISEM), Arnaud Estoup (INRA, CBGP) et Jean-Michel Marin (Université de Montpellier, Statisticien), dans le cadre des modèles spatialisés d'IBD sur lesquels j'avais déjà longuement travaillé.

Inférences démographiques populationnelles spatialisées

Depuis ma thèse, j'ai montré un intérêt certain pour l'utilisation de modèles spatialisés de génétique des populations afin de caractériser finement les densités et les caractéristiques de dispersion chez des espèces animales ou végétales. J'ai considéré pour ces analyses spatialisées les modèles d'isolement par la distance (IBD) qui prennent en compte le fait que, chez une majorité d'espèces, les individus résident près de leurs lieux de naissance du fait d'une dispersion limitée dans l'espace. Une telle dispersion localisée est un facteur majeur de la structuration spatiale des populations, notamment à petite et moyenne échelles géographiques, mais a toujours été difficile à quantifier. Des méthodes basées sur les F-statistiques ont été développées pour estimer certains paramètres de dispersion, à partir de l'augmentation de la

différenciation génétique avec la distance géographique (méthode de la régression, voir Rousset 1997, 2000, Watts et al. 2007). Des comparaisons entre estimations génétiques (par ces méthodes) et démographiques (par capture-marquage-recapture) indépendantes sur une dizaine de jeux de données réels, ont montré une excellente concordance entre les estimations génétiques et démographiques, validant ainsi la méthode d'estimation (précision et surtout robustesse) et le modèle d'IBD (réalisme) (Leblois 2004, Guillot et al. 2009). C'est la première fois qu'une telle concordance entre estimations directes (démographiques) et indirectes (génétiques) est observée, et ceci sur tous les jeux de données que nous avons trouvés comportant des données génétiques et démographiques pertinentes pour de telles estimations, validant ainsi le modèle d'isolement par la distance et la méthode de la régression. Ces travaux sont présentés en détails dans le chapitre 2 du document scientique.

Cette méthode d'inférence basée sur les F_{ST} présente toutefois d'importantes limitations : (1) elle ne considère pas de manière efficace toute l'information présente dans les données génétiques ; (2) elle ne permet pas l'estimation de la forme de la distribution de dispersion ; enfin (3) elle fait l'hypothèse que la dispersion et la densité sont homogènes sur tout l'habitat échantillonné. Pour essayer de pallier aux deux premières limitations, nous avons implémenté avec F. Rousset des algorithmes IS-coa sous un modèle IBD dans Migraine (Rousset & Leblois 2007, 2012, Rousset et al. 2018) et nos résultats montrent de bonnes performances par rapport aux méthodes existantes. Cette méthode a été appliquée en collaboration sur différents jeux de données (Lippens et al. 2017, Gauffre et al. 2020), et d'autres applications sont en cours. Les principales limitations de cette nouvelle méthode IS-coa d'inférence de la dispersion sont (1) l'utilisation des approximations classiques du n-coalescent (grandes tailles de populations, petits taux de migration) qui ne permettent pas de considérer des population en IBD dans lesquelles les individus sont répartis de façon homogène sur un habitat continu, sans sous-structuration en dèmes ; et (2) les temps de calculs qui ne permettent pas de considérer un très grand nombre de marqueurs, ni l'ajout d'hétérogénéités spatiales ou temporelles des paramètres démographiques.

Et maintenant...inférence par simulation de la dispersion et des densités

Pour aller plus sur l'inférence des paramètres de dispersion et de densité à petite échelle spatiale, je souhaitais donc élargir les champs d'application des méthodes d'inférences spatialisées en développant de nouvelles approches, pour pouvoir prendre en compte des modèles IBD individuels en habitat continu, tout en utilisant au maximum l'information des données génétiques spatialisées. Le but à plus long terme étant aussi de pouvoir considérer des modèles spatialisés plus complexes intégrant par exemple des barrières à la dispersion (ex. barrière paysagère telles que des rivières ou des montagne,... ; barrière écologique telles que différentes plantes hôtes, différents écotypes,...) et/ou des changements temporels des paramètres démographiques (ex. une expansion spatiale). Enfin, tout cela nécessitait d'évaluer, en profondeur pour chaque type de modèle, quels types de paramètres démographiques fins (par exemples, la forme de la distribution de dispersion) peuvent être estimés à partir de données génétiques/génomiques. Tout cela m'a poussé à me tourner vers d'autres méthodes non basées sur le calcul de la vraisemblance mais utilisant des approches statistiques d'inférences par simulation, adaptées à tout modèle complexe sous lequel on peut simuler des données assez rapidement.

Cela nécessite donc la simulation de données sous des modèles spatialisés de gé-

nétique des populations. Depuis 20 ans maintenant, j'ai développé, étendu et mis à disposition de la communauté scientifique un logiciel de simulation de données génétiques sous des modèles spatialisés d'IBD : IBDSim (Leblois et al. 2009). Ce logiciel ne permet pas de prendre en compte la recombinaison, et simule donc uniquement des locus indépendants non-recombinant. De plus, j'ai commencé à écrire ce programme de simulation en master, sans aucune formation en programmation, le code n'est donc pas très propre. De ce fait, j'ai décidé en 2018 de développer avec Thimothée Virgoulay (en doctorat) un nouveau simulateur de données génomiques spatialisées basé sur les principes de l'IBD (Virgoulay et al. 2021). Parallèlement, j'ai commencé à développer un pipeline d'inférence par simulation sous R pour estimer la densité, la dispersion et la taille d'une population sous IBD. Ce pipeline permet de coupler les simulateurs GSpace et IBDSim à des librairies d'inférence par simulation telles que abcrf développée par Jean-Michel Marin et Arnaud Estoup, et Infusion développée par François Rousset. Nous avons tout récemment obtenue nos premiers résultats d'inférences et ils sont extrêmement intéressant. Il semblerait en effet, que l'on puisse inférer avec une très bonne précision la densité et la dispersion de façon indépendante, et ceci même avec peu de marqueurs. C'est le principal axe de recherche sur lequel je travaille en ce moment, et pour quelques années encore... Ces travaux sont présentés en détails dans le chapitre 4 du document scientifique.

Applications et transfert méthodologique

Depuis le début de ma carrière, je porte donc une attention particulière à l'application et au transfert des méthodes développées aux modèles biologiques étudiés dans le laboratoire auquel j'appartiens mais également, en fonction des opportunités de collaborations, aux modèles biologiques étudiés dans d'autres laboratoires. Il me semble en effet essentiel d'appliquer, de tester et de valider les nouvelles méthodes statistiques que je développe sur des données réelles : il existe non seulement un premier intérêt, évident, qui concerne une communauté scientifique d'« empiristes » en attente de solutions méthodologiques nouvelles ; mais aussi un second intérêt, pas moins important, qui concerne les théoriciens eux-mêmes dans la mesure ou les études de situations et jeux de données concrets sont très souvent à l'origine d'idées nouvelles. L'expérience montre en effet que les collaborations étroites entre « théoriciens » et « empiristes » ont une influence positive très forte sur les développements méthodologiques.

Pour faciliter ces transferts méthodologiques, nous nous efforçons au CBGP de produire des interfaces logicielles claires, simples d'utilisation, et bien documentées qui fournissent aux empiristes des outils leur permettant de réaliser sur leurs propres modèles biologiques des analyses pertinentes et robustes de leurs données. Enfin ces transferts méthodologiques sont également favorisés via l'élaboration de modules d'enseignement de génétique des populations dans diverses formations académiques ou bien d'organisation et de participation à des rencontres avec d'autres chercheurs (écoles chercheurs, colloques/workshops de petite taille). Tous ces transferts sont détaillés en Annexe.

Chapitre 1

Quelques notions de génétique des populations

La génétique des populations étudie la répartition et l'évolution de la variation (ou *polymorphisme*) génétique des populations dans le temps et dans l'espace. Cette discipline est née de la synthèse des théories de Mendel, de Darwin et des biométriciens du début du XX ème siècle, notamment Ronald A. Fisher, Sevall Wright et John B. S. Haldane, qui en sont les fondateurs. Dans les années 30 à 60, ces derniers ont chacun à leur manière posé les bases conceptuelles et une formalisation mathématique de l'évolution de la variation génétique dans les populations. Les premiers travaux s'attachent essentiellement à décrire l'effet des forces évolutives sur le polymorphisme génétique à travers divers modèles mathématiques (approche "prospective"), et immédiatement émerge l'idée que la structure spatiale des populations est un facteur primordial dans les processus d'adaptation, notamment l'idée que la dispersion limite la différenciation des populations. La dispersion pourrait donc, dans une certaine mesure, limiter l'adaptation locale des population à leur environnement, voir les premiers stades de la spéciation allopatrique. Plus largement, la dispersion, les densités et tailles de populations, entre autres facteurs démographiques, apparaissent rapidement comme des acteurs majeurs de l'histoire évolutive et de l'adaptation des populations à leur environnement.

Ces paramètres démographiques ne sont pas faciles estimer avec des méthodes démographique (par ex. par capture-marquage-recapture) et les estimer à partir des patrons de structuration génétique, et donc de données génétiques (génotypes multilocus caractérisés à l'aide de marqueurs génétiques) issues d'échantillons de populations structurées, a été rapidement un des buts de la génétique des populations. De telles inférences de paramètres démographiques, dites "indirectes" car à partir de données génétiques et non démographiques, ont été initiées dés les années 40, notamment par Dobzhansky & Wright (1941), étude dans laquelle ils estiment le nombre de migrant Nm dans un modèle en îles à partir d'échantillons de populations naturelles de drosophiles. L'inférence de paramètres démographiques à partir de données génétiques a ensuite littéralement explosé grâce aux progrès de l'informatique et à trois progrès majeurs de la biologie moléculaire ayant fortement affecté la génétique des populations : (i) l'électrophorèse enzymatique dans les années 60 ; (ii) le génotypage en routine de marqueurs microsatellites dans les années 80 ; et surtout (iii) le développement des nouvelles techniques de séquençage haut débit (NGS, pour "Next Generation Sequencing") dans les années 2000. Ainsi, la génétique des population s'est largement tournée vers des approches d'inférences "retrospectives" pour comprendre, à partir de données actuelles de polymorphisme génétique, le fonction-

nement évolutif actuel des populations, mais aussi leurs histoires démographiques et adaptatives passées.

Une caractéristique majeure de l'analyse du polymorphisme en populations naturelles, et plus généralement en biologie évolutive, est que l'on travaille sur des données "expérimentales" sans réplicats et pour lesquelles les conditions initiales de l'"expérience" ne sont pas connues. Ceci a des implications extrêmement importantes sur l'analyse des données. Ainsi, lorsqu'on étudie un échantillon d'individus d'une population génotypés à plusieurs locus, les états alléliques des différents locus peuvent être statistiquement dépendants si les locus sont proches sur le même chromosome (liaison génétique). De plus, pour chaque locus, les états alléliques des différents individus sont statistiquement dépendants du fait de l'histoire généalogique qu'ils partagent. Ces dépendances statistiques sont le résultat de l'histoire commune des événements de mutation, de recombinaison et de coalescence des lignées génétiques ancestrales de l'échantillon. Ces facteurs doivent être intégrés dans l'analyse statistique des données. Une solution consiste à modéliser le passé (l'histoire évolutive menant à l'échantillon observé) à l'aide d'un modèle stochastique approprié, dont un exemple que nous détaillerons ci-dessous est le coalescent. Comme nous allons le voir dans cette synthèse, (i) les histoires évolutives passées possibles pour un échantillon sont souvent très complexes (grands espaces de paramètres) et impliquent de nombreux processus stochastiques de forte variance même pour des modèles relativement simples, et (ii) les échantillons sont (tout au moins étaient...) de relativement petite taille La génétique des population s'est donc focalisée sur des modèles simples associés à des méthodes d'inférence statistique performantes.

Au début des années 2000, de nombreuses études remettent en question la robustesse des inférences démographiques "indirectes" du fait de l'inadéquation entre les modèles utilisés et une réalité beaucoup plus complexe (combinaison de processus démographiques actuels, passés, de processus mutationnels et des processus sélectifs, peu ou mal pris en compte). Il est notamment considéré que les approches directes permettent d'estimer les paramètres actuels tandis que les approches indirectes ne permettraient d'estimer que les valeurs passées des paramètres (Boileau *et al.*, 1992; Koenig *et al.*, 1996). Ces études soulèvent très justement des questions pertinentes sur la robustesse des estimations de tailles de populations et de la dispersion. Nous en discuterons en détails dans les sections suivantes, mais ce sentiment semble s'estomper dans les années 2000 avec la généralisation des données NGS qui permettent des estimations plus précises sous des modèles plus complexes et donc potentiellement plus réalistes (Luikart *et al.*, 2003; Marchi *et al.*, 2021). Si certaines de ces méthodes ont pu donner des résultats raisonnables et parfois même impressionnants comme l'inférence détaillée des variations passées de la taille d'une population à partir d'un seul génome diploïde (Li & Durbin, 2011), elles ont encore de nombreuses limitations. De plus, même si ces gros jeux de données permettent une compléxification accrue des modèles, ils restent une description extrêmement simplifiée de la réalité biologique des populations naturelles. La robustesse des inférence est donc toujours une question centrale en génétique des populations.

Dans ce document, je présenterai une introduction à l'inférence de paramètres démographique à partir de données génétiques, à travers la synthèse et la contextualisation des principaux résultats issus de mes recherches collaboratives sur l'inférence de la dispersion dans des modèles spatialisés. Je commencerai par introduire quelques notions de génétiques des populations nécessaires à la compréhension des travaux présentés par la suite. Ainsi, dans un premier chapitre, j'introduirai la théorie de la coalescence, la notion d'identité génétique et de F-statistiques, puis je décrirai

quelques modèles de population structurées permettant de modéliser la dispersion. Cette introduction à divers outils classiques et centraux de la génétique des populations sera complétée par une présentation rapide de quelques principes d'inférence statistique en se focalisant sur les méthodes qui seront ensuite explorées : l'inférence par la méthode des moments, par vraisemblance et l'inférence par simulation. Dans le second chapitre, je décrirai une méthode d'inférence de la densité et de la dispersion par moment basée sur les F-statistiques et présenterai les détails des tests de performances ayant permis de valider la robustesse de cette méthode pour l'inférence des paramètres démographiques locaux et actuels des populations. Nous verrons aussi comment ces différents tests permettent une certaine validation des modèles démo-génétiques d'isolement par la distance, qui apparaissent donc comme des modèles adéquates pour étudier les processus de dispersion limitée en populations naturelles. Nous nous intéresserons ensuite à d'autres approches d'inférences pour essayer de palier certaines limites de l'inférence par F-statistiques mise en évidence précédemment. Le chapitre trois présentera deux méthodes permettant l'inférence par vraisemblance à partir de données génétiques, ainsi que des tests de performances nous permettant encore une fois de mieux comprendre le comportement de ces méthodes, leurs intérêts et leurs limites. Je finiraidans le chapitre quatre en présentant mes travaux actuels sur l'inférence par simulation de la dispersion, des différents outils que nous avons récemment développés dans ce but, et quelques résultats malheureusement trop préliminaires mais ayant tout de même un impact très fort sur tout ce qui a été vu et discuté précédemment.

Le but de ce document est de présenter la démarche scientifique sous-jacente à l'ensemble de mon parcours de recherche en me focalisant donc sur l'inférence de paramètres de dispersion et de densité dans des modèles spatialisés de génétique des populations. Ce n'est donc pas une revue bibliographique de la génétique spatiale des populations, mais plutôt une vision biaisée illustrant une démarche que je pense pertinente et importante à transmettre (je me la pète...). Ainsi, j'espère que ce document pourra aussi servir d'introduction à la génétique (spatiale) des populations et à l'estimation de paramètres démographiques à partir de données génétiques, pour des étudiants commençant un travail dans ce domaine.

Ce document est donc une présentation biaisée d'un champ de la "génétique spatiale des populations", notamment par le fait que l'on se soit focalisé sur (i) les approches par coalescence, et (ii) des données issues de marqueurs génétiques 'ìndépendants" supposés neutres, et (iii) sur l'inférence de paramètres démographique. Je n'évoquerai pas du tout la sélection et les processus adaptatifs, en considérant qu'ils ont un effet négligeable à petite échelle géographiques et temporelles sur les inférences décrites ici. Je ne parlerai pas non plus des méthodes que j'appellerai descriptives qui relient une description statistique des données (statistique résumante, patron génétique, corrélations, inférence sous un modèle purement statistique, etc) ne se basant pas clairement sur un modèle biologique (Beaumont *et al.*, 2010). Je ne présenterai pas non plus la multitude d'approches alternatives développées, notamment ces dernières années, pour l'inférence de paramètres démographiques et historiques en génétiques des populations. Le lecteur pourra se référer, entre autre, aux livres de Gillespie (2004); Hein *et al.* (2005); Nielsen & Sltakin (2013) pour des introductions plus générales à la génétique des populations, ainsi Hartl & Clark (2007); Balding *et al.* (2007) pour une approche plus approfondie. Quoiqu'il en soit, nous essayerons tout au long de ce document de se baser sur exemples précis mais d'appliquer des raisonnements généraux et de tirer des conclusions ayant une portée la plus générale possible, et potentiellement utile dans tous ces autres champs

disciplinaires.

1.1 La théorie de la coalescence

Considérons un locus particulier dans le génome d'une espèce. Quel que soit l'échantillon que l'on considère, toutes les copies à ce locus (les gènes homologues) sont reliées entre elles et à un ancêtre commun ("most recent common ancestor", MRCA) par leur histoire généalogique, que l'on peut représenter sous la forme d'un arbre généalogique, dit arbre de coalescence (Figure 1.1 et 1.2). Un intérêt majeur de la coalescence est que la simulation de la généalogie peut s'effectuer sans prendre en considération les gènes des autres individus de la population car les processus démographiques (et reproductifs, par ex. l'auto-fécondation) ne dépendent pas de la configuration génétique de la population entière, contrairement aux processus sélectifs (Figure 1.1). En effet, par définition, le nombre de "descendants" d'un gène et leur probabilité de disperser ne dépend pas de son état allélique. On peut ainsi dissocier, sous l'hypothèse de neutralité des marqueurs génétiques utilisés, le processus de mutation du processus généalogique. C'est le principe de la théorie de la coalescence : on retrace la généalogie des gènes de l'échantillon suivant les processus démographiques et on ajoute ensuite sur cet arbre de coalescence les effets des événements de mutation selon les processus mutationnel, ce qui détermine in fine les états alléliques des gènes considérés.

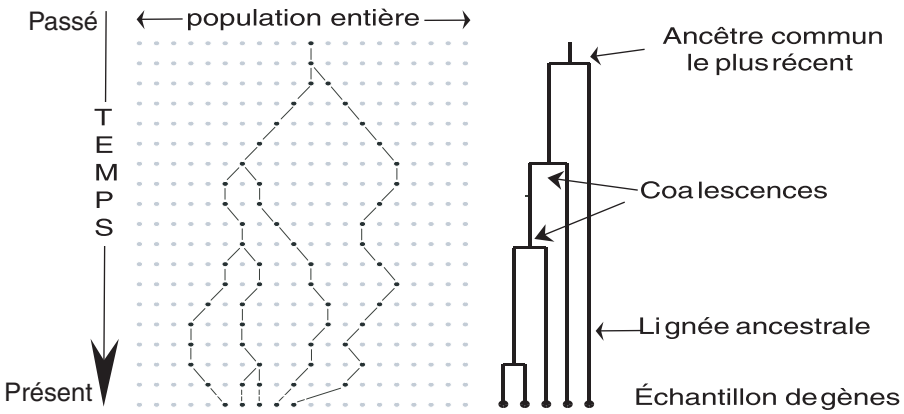


FIGURE 1.1 – Représentation du principe de la théorie de la coalescence. A gauche est représenté le "trajet" des lignée ancestrales d'un échantillon de 5 gènes au sein d'une population de Wright-Fisher de taille $N = 20$ gènes. A droite est représenté l'arbre de coalescence de ce même échantillon.

Sur la Figure 1.2, le patron de polymorphisme de l'échantillon au locus considéré reflète donc l'histoire des coalescences des lignées ancestrales, représentée dans l'arbre, et l'histoire des mutations. Tout comme les mutations sont distribuées de façon aléatoire dans les processus évolutifs, la généalogie de gènes l'est aussi et on doit donc considérer ces deux sources de variations dans les modèles. On a donc besoin de modèles qui nous permettent de décrire des généalogies aléatoires de gènes. La suite de cette section est une présentation succincte de quelques exemples classiques de tels modèles.

La théorie de la coalescence est un cadre conceptuel naturel pour analyser les modèles "démo-génétiques" de génétique des populations, et se dérive en plusieurs catégories de modèles mathématiques et/ou simulatoires que l'on précisera par la suite. Bien que l'idée originelle de la coalescence, pour une paire de gène, remonte à

Gustave Malécot [1] et ses travaux sur “l’identité par descendance” (Malécot, 1972), la théorie de la coalescence pour un échantillon de plus de deux gènes, a été formalisée de façon indépendante à travers le n-coalescent, à la fin des années 80, par Kingman (1982a,b) et Tajima (1983). Cette théorie repose sur l’unique principe énoncé ci-dessus : le polymorphisme génétique observé à un locus neutre sur un échantillon de gènes d’une même espèce est le résultat de l’histoire généalogique et mutationnelle des lignées ancestrales de cet échantillon, qui peut être générée (i) en remontant dans le passé jusqu’à l’ancêtre commun le plus récent des gènes de l’échantillon (MRCA) à l’aide d’un modèle démographique pour créer l’arbre de coalescence ; puis (ii) en y ajoutant les mutations en redescendant l’arbre. Les modèles qui en dérivent, tels que le n-coalescent ou le coalescent structuré, reposent parfois sur des hypothèses supplémentaires que nous discuterons dans les trois sections suivantes. Les intérêts de cette théorie et des modèles qui en découlent sont multiples : (i) la structure des données génétiques reflète, en grande partie, la généalogie sous-jacente aux données. De ce fait, l’étude de la généalogie permet une compréhension qualitative des patrons de variation des données génétiques (voir Nordborg & Tavare, 2002; Nielsen & Sltakin, 2013), souvent plus intuitive que dans d’autres cadres conceptuels de la génétique des populations tels que, par exemple, la théorie de la diffusion ; (ii) les analyses quantitatives sont généralement plus faciles avec des méthodes généalogiques qu’avec les approches traditionnelles qui retracent la composition de la population entière en avançant dans le temps, au moins pour l’analyse du polymorphisme neutre ; (iii) l’utilisation de la théorie de la coalescence donne des méthodes de simulation extrêmement efficaces ; et (iv) la coalescence permet de faire des inférences utilisant toute l’information des données génétiques (sous réserve d’utiliser des méthodes d’inférence par vraisemblance).

L’arbre généalogique d’un échantillon de n gènes pris dans une population panmictique de taille constante N au cours du temps est modélisé par un processus stochastique connu sous le nom de n-coalescent introduit par Kingman (1982a) comme une approximation de la généalogie de gènes évoluant suivant le modèle neutre de "Wright-Fisher". Ce modèle n-coalescent a une formalisation mathématique très simple que nous développerons dans la section suivante. Nous verrons ensuite comment adapter ce modèle pour des populations structurée en développant quelques aspects du coalescent structuré. De nombreux autres modèles ont été développés pour s’ajuster à des situations biologiques plus complexes considérant de la recombinaison, de l’autofécondation ainsi que des variations de tailles de populations au cours du temps. Certains auteurs ont aussi développés des modèles non neutres prenant en compte la sélection au locus considéré. Le lecteur pourra notamment se référer à Hudson (1990), au livre de Wakeley (2008), et la revue de Nordborg (2001), ainsi que Neuhauser & Krone (1997) et Wakeley (2010) pour la sélection.

Dans la littérature, le n-coalescent est souvent appelé le “coalescent”, et il n’y pas de distinction claire entre les terme “coalescence” et“coalescent” pour caractériser d’une par le principe de la théorie de la coalescence et d’autre part le modèle du n-coalescent. Pour ce document, nous considérerons que le terme “coalescence” se rapporte à un évènement de coalescence ou au principe de la coalescence, et que le terme "coalescent", se rapporte au modèle du n-coalescent et ses approximations.

1. “Le lien de filiation qui unit l’un [des] individus à l’un des ancêtres [...] est une chaîne d’ascendance. Deux chaînes d’ascendance aboutissant à un même ancêtre forment une chaîne de parenté entre les deux individus pris chacun sur l’une des chaînes d’ascendance exclusivement.” Malécot 1966

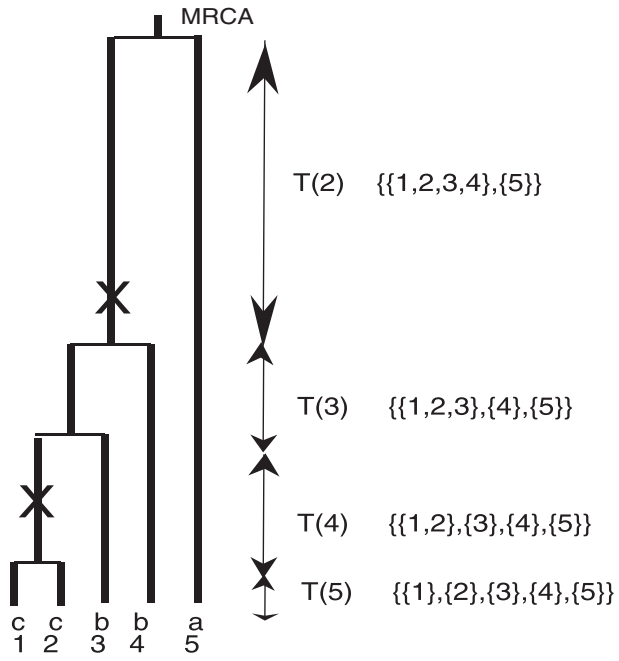


FIGURE 1.2 – La généalogie, ou arbre de coalescence, d'un échantillon de gènes peut être décrite en terme de topologie et de longueur de branche. La topologie peut être représentée comme des classes d'équivalence des lignées ancestrales. La longueur des branches correspond aux temps $T(k)$ entre deux événements de coalescence. Le polymorphisme de l'échantillon, représenté par les feuilles, numérotées de 1 à 5, ayant les état alléliques a, b et c, est du aux mutations (les X) ayant eu lieu sur les différentes branches.

1.1.1 Le n-coalescent

Le n-coalescent est le premier modèle et toujours le plus standard pour décrire la généalogie d'un échantillon de n gènes dans une population haploïde de Wright-Fisher (WF) avec le cycle de vie présenté p. 4, c'est à dire dans une population panmictique, isolée, de taille constante, avec des générations non chevauchantes et dont tous les individus ont la même variance de succès reproducteur. Dans une population de N gènes, la probabilité que deux lignées (c.a.d. gènes) aient un ancêtre commun à la génération précédente et donc coalescent est $1/N$ (on choisi le premier parent au hasard parmi les N parents possibles, puis la probabilité de choisir le même parent pour la seconde lignée est $1/N$). La probabilité qu'elles restent distinctes est donc $1-1/N$. Puisque les générations sont indépendantes, la probabilité qu'elles restent distinctes plus de t générations dans le passé est donc $(1-1/N)^t$. La probabilité qu'une coalescence ait lieu à t générations dans le passé suit alors une loi géométrique de paramètre $(1-1/N)$

$$\Pr(C_t|n=2) = (1-\frac{1}{N})^{t-1}\frac{1}{N}, \tag{1.1}$$

De cette formule, on peut facilement comprendre l'approximation standard en temps continue du n-coalescent valable pour des grandes tailles de populations ($N \to \infty$). Considérons un changement d'échelle du temps tel que la nouvelle unité de temps τ corresponde à N générations. La probabilité que deux lignées coalescent à τ unités de temps dans le passé est alors

$$\begin{aligned}\Pr(C_t|n=2) &= (1-\frac{1}{N})^{t-1}\frac{1}{N} \approx_{N\to\infty} \frac{1}{N}\left(e^{-\frac{1}{N}}\right)^{t-1} = \frac{1}{N}e^{-\frac{t-1}{N}} \\ \Pr(C_\tau|n=2) &= (1-\frac{1}{N})^{\lceil N\tau\rceil} \approx_{N\to\infty} e^{-\tau},\end{aligned} \tag{1.2}$$

où $\lceil N\tau \rceil$ est le plus grand entier plus petit que $N\tau$. Le temps de coalescence (exprimé en N générations) d'une paire de gène suit donc une loi exponentielle, $e^{-\tau}$, de moyenne 1 beaucoup plus "pratique" mathématiquement que la loi géométrique.

Considérons maintenant un échantillon de k lignées (dans lequel il y a donc $k(k-1)/2$ paires de lignées pouvant coalescer), la probabilité qu'aucune de ces lignées ne coalesce à la génération précédente est

$$\Pr(C_t > 1 | n = k) = \prod_{i=1}^{k-1}(1 - \frac{i}{N}) = 1 - \frac{k(k-1)}{2N} + O(\frac{1}{N^2}), \tag{1.3}$$

ou $O(\frac{1}{N^2})$ correspond à la probabilité des coalescences simultanées de plus de deux gènes à une génération donnée (coalescence multiple de 3, 4,... lignées), que l'on peut négliger quand N est grand. Définissons T_k, le temps de la première coalescence de deux lignées dans un échantillon de k lignées (voir Figure 1.2). On a alors, selon le même raisonnement que précédemment, l'approximation exponentielle des temps de coalescence pour de grandes tailles de populations :

$$\Pr(T_k = \tau) \approx^{N\to\infty} \frac{k(k-1)}{2} e^{-\tau \frac{2}{k(k-1)}} \tag{1.4}$$

Sous l'approximation en temps continu, le nombre de lignées ancestrale d'un échantillon décroît donc pas à pas en fonction des $T(k)$ pour $k = n, .., 2$, les temps nécessaire pour passer de k à $k-1$ lignées (Figure 1.2). En résumé, le modèle du n-coalescent décrit la généalogie d'un échantillon de n gènes comme un arbre avec des bifurcations aléatoires, où les $n-1$ temps de coalescence $\{T(n), T(n-1), .., T(3), T(2)\}$ sont des variables aléatoires mutuellement indépendantes suivant des lois exponentielles. Les mutations sont ensuite surimposées sur ces branches, en redescendant le temps (du MRCA jusqu'au présent), selon une loi binomiale avec comme paramètres μ le taux de mutation par unité de temps et t ou τ la longueur de la branche à la bonne échelle. Cette loi binomiale souvent approchée par une loi de poisson de paramètre μt sous l'hypothèse $\mu \to 0$.

On voit bien ici que la simulation d'un échantillon sous ce modèle du n-coalescent est extrêmement efficace. Il suffit de simuler $n-1$ variables aléatoires selon une loi exponentielle, correspondant aux temps de coalescence, et de construire de façon indépendante une topologie aléatoire des bifurcations, les coalescences, en choisissant au hasard les paires de lignées qui coalescent. On ajoute ensuite sur l'arbre les mutations.

Quelques propriétés importantes de génétique des populations se comprennent facilement à l'aide de la coalescence et sont présentée dans de nombreux ouvrages (voir par ex. Nordborg (2001); Wakeley (2008); Nielsen & Sltakin (2013)). Nous évoquerons ici uniquement deux propriétés de l'échantillon génétique. Saunders *et al.* (1984) ont montré que la probabilité que le MRCA d'un échantillon de taille n soit le MRCA de la population entière est $(n-1)/(n+1)$. On voit donc bien qu'il n'est pas nécessaire de prendre un grand échantillon pour remonter le plus loin possible dans le passé. Ceci implique également que les inférences sur des processus démographiques anciens seront limitées du fait que le MRCA d'un échantillon peut être récent, notamment lorsque les tailles de populations sont faibles. Enfin, il est intéressant de noter que, dans la mesure où le nombre de mutations est proportionnel à la longueur des branches et que les copies d'un gène sont étroitement liées du fait de leur généalogie commune, une augmentation de la taille de l'échantillon (au dessus d'une valeur relativement faible d'environ 20-30 individus) n'accroîtra que peu la

puissance des analyses de génétique des populations sauf pour des processus très récents et spatialisé qui repose plus sur la répartition spatiale du polymorphisme (répartition spatiale des branches terminales) que sur les temps d'apparition des mutations et leurs fréquences.

Enfin, le n-coalescent s'adapte facilement, grâce à des changements d'échelle de temps judicieux, à de nombreuses extension du modèle de WF permettent de prendre en compte des processus biologiques tels que des générations chevauchantes, des sexes séparés avec biais de sexe ratio, et d'autres systèmes de reproduction dans l'approche du n-coalescent. Ceci est dû au fait que ces processus biologiques ne changent pas la topologie de l'arbre mais seulement les longueurs de branche. Par exemple, il est possible de considérer que la variance du nombre de descendants n'est pas 1 mais α avec une unité de temps de N/α au lieu de N. La considération d'une échelle non linéaire avec le temps peut aussi permettre de prendre en compte certaines variations simples (exponentielle) des tailles de population dans le temps (de nombreux exemples de ces changements d'échelle sont revus dans Nordborg, 2001; Wakeley, 2008). Enfin, des formulation alternatives peuvent facilement être développée pour de nombreux autres modèles de population, dont le coalescent structuré décrit ci-dessous.

1.1.2 Le coalescent structuré

Considérons une population d'individus haploïdes de taille N_{tot} subdivisée en n_d sous-populations de tailles N_i telles que $\sum_i N_i = N_{tot}$ avec toujours le même cycle de vie présenté en p. 4. A chaque génération, chaque adulte produit une infinité de gamètes qui migrent de leur dème d'origine i vers le dème j avec la probabilité m_{ij}. Enfin, la compétition entre juvéniles ramène le nombre d'individus à N_i adultes dans chaque sous-population. On définit les quantités $c_i \equiv \frac{N_i}{N_{tot}}$ et $B_{ij} \equiv N_{tot} b_{ij}$, ou les b_{ij} sont les probabilités de migration "arrière" (en remontant le temps). En d'autres termes b_{ij} est la probabilité qu'un gène du dème i ait son parent dans le dème j. Ces quantités sont définies en fonction des taux de migration "avant" m_{ji} par

$$b_{ij} = \frac{N_j m_{ji}}{\sum_k N_k m_{ki}}. \tag{1.5}$$

Comme pour le n-coalescent, on suppose que ces quantités sont constantes quand $N_{tot} \to \infty$ et que le temps est mesuré en N_{tot} générations. C'est à dire que l'on considère que les probabilités que plusieurs événements de migration et/ou de coalescence aient lieu en une seule génération sont des $O(1/N_{tot}^2)$ et donc négligeables. Dans la limite où $N_{tot} \to \infty$, les seuls événements possibles sont donc une coalescence au sein d'un dème ou une migration entre deux dèmes. C'est ce qu'on appellera les approximations du coalescent, aussi souvent dénommées approximation de diffusion. Le temps d'occurrence d'un de ces événements (c.a.d. le temps attendu avant qu'un événement se produisent, comme on a vu précédemment pour les évènements de coalescence uniquement) suit alors une loi exponentielle de moyenne, r, le taux d'événement global, somme des taux de chaque événement possible,

$$r(\mathbf{n}) = \sum_i \Big(\frac{k_i(k_i-1)}{2c_i} + \sum_{j \neq i} k_i B_{ij} \Big), \tag{1.6}$$

où k_i est le nombre de lignées présentes dans le dème i au temps considéré et $\mathbf{n} = \{k_i\}$, pour $i \in [1, .., n_\mathrm{d}]$, est la configuration (répartition) globale des lignées dans

les différents dèmes. Si un événement a lieu, c'est une coalescence dans le dème i avec la probabilité

$$\frac{k_i(k_i-1)/2c_i}{r(\mathbf{n})}, \tag{1.7}$$

ou c'est une migration ("arrière") du dème i vers le dème j avec la probabilité

$$\frac{k_i B_{ij}}{r(\mathbf{n})}. \tag{1.8}$$

Avec ces modèles de structuration, il n'y a pas de changement d'échelle possible pour se rapprocher du n-coalescent car la structuration des populations ne change pas uniquement la longueur des branches mais aussi la topologie de l'arbre. Ainsi, si la migration est faible, les lignées échantillonnées dans un même dème vont coalescer rapidement entre elles et le temps nécessaire pour que deux lignées échantillonnées dans deux dèmes différents coalescent va être beaucoup plus long que dans le modèle panmictique. Intuitivement, on comprend que, si la migration est faible, ces lignées mettront un certain temps avant de se retrouver dans le même dème pour pouvoir coalescer. Cette caractéristique est illustrée par la figure 1.3. Une limite de ce modèle réside dans le fait que l'on peut considérer uniquement des grandes tailles de populations afin de ne pas avoir à considérer d'événements multiples pour avoir une expression simple de temps d'attente entre deux événements. Pour une lecture approfondie sur le coalescent structuré, le lecteur pourra se référer à Takahata (1991) et Wakeley (2010).

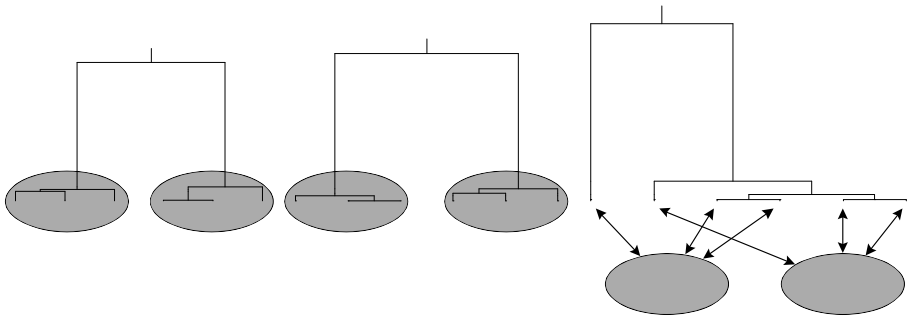


FIGURE 1.3 – Trois réalisations du coalescent structuré sous un modèle symétrique de migration à deux dèmes avec $N_i = 3$. Les lignées ont tendance à coalescer au sein d'un dème mais pas toujours comme le montre l'arbre de droite. (figure issue de Nordborg, 2001)

Il est intuitif de dire qu'une faible migration aura un effet important sur la structure des généalogies. A l'inverse, une migration forte va, d'une certaine manière, nous rapprocher des généalogies obtenues sous un modèle panmictique. Une migration forte implique que les événements de migration sont beaucoup plus fréquents que les événements de coalescence et $\lim_{N_{tot}} Bij = \lim_{N_{tot}\to\infty} N_{tot} b_{ij} = \infty$. Dans le cas limite où $N_{tot} \to \infty$, il y aura donc une infinité d'événements de migration entre deux événements de coalescence. C'est ce que l'on appelle une séparation des échelles de temps : les événements de migration ont lieu sur une échelle de temps beaucoup plus rapide que les coalescences. Les événements de coalescence ayant lieu entre deux lignées présentes dans le même dème, ils sont alors fonction de la distribution stationnaire $\Pi = \{\pi_i, i \in [1, .., n_\mathrm{d}]\}$ des lignées entre les différents dèmes. Selon ces notations, la coalescence d'une paire de lignées a lieu dans le dème i avec la probabilité π_i^2/c_i, puisque la probabilité que deux lignées se retrouvent dans le dème i est π_i^2. En prenant comme échelle de temps N_{tot}/α avec $\alpha \equiv \sum_i \pi_i^2/c_i$, le taux de coalescence total de l'échantillon, on se ramène alors au modèle du n-coalescent

(Nagylaki, 1980; Notohara, 1993). Cette séparation des échelles de temps simplifie donc considérablement l'analyse mathématique du coalescent structuré mais ignore certains effets spécifiques de la migration et, dans ce cas, l'utilisation du n-coalescent avec la mise à l'échelle N_{tot}/α ne permet pas d'inférences sur les processus de migration. Cette notion de séparation des échelles de temps est discutée en détail dans Fu (1997); Nordborg (1997); Nordborg & Krone (2002) et Rousset (2006).

Une notion similaire de séparation des échelles de temps sera retrouvée dans le cadre de l'analyses des probabilités d'identité dans une population panmictique et dans le modèle en îles en section 1.2, pour lequel on s'intéressera à la séparation des événements de coalescence intra- et inter-classe (par ex. individu, dèmes). Nous verrons que cette séparation des événements de coalescence intra- et inter-classe permet une meilleure interprétation des propriétés de certains paramètres du modèle. Ainsi lorsqu'une séparation des échelles de temps est possible, certains paramètres s'inscrivent dans l'échelle de temps longs alors que d'autres sont dans l'échelle de temps courts et seront de ce fait moins influencés par des processus passés.

1.1.3 Simulation par coalescence

Précédemment, nous avons vu que la théorie de la coalescence semble être un moyen très efficace de simuler le polymorphisme d'un échantillon génétique sous différents modèles démographiques et mutationnels car elle ne se base pas sur la simulation de la population entière mais uniquement des lignée ancestrales de l'échantillon. Il existe trois principes de simulation permettant de générer des échantillons de gènes ayant évolué sous des modèles démo-génétiques plus ou moins variés, avec plus ou moins d'approximations.

La première méthode, développée par Donnelly (1999) et proche de la formule d'échantillonage de Ewens ("Ewens sampling formula", Ewens, 2004), ne s'applique que dans le cadre du n-coalescent. C'est une méthode qui s'apparente à un tirage de boules de couleur dans une urne avec les couleurs correspondant aux états alléliques des gènes de l'échantillon, et où les probabilités de tirage d'une boule de même couleur est donnée par la probabilité de coalescence, et celle de tirer une boule d'une autre couleur par le processus de mutation. Cette méthode de simulation considère donc en même temps la mutation et la coalescence, et n'a pas de notion d'échelle de temps ni de lignée ancestrale. L'avantage de cette méthode est son extrême rapidité, puisqu'elle est plus rapide que l'approche directe du n-coalescent mentionnée dans la section 1.1.1. Son inconvénient principal est qu'elle ne s'applique que dans des modèles panmictiques de taille constante.

La seconde méthode, développée par Hudson (1990, 1993, 1998), s'applique dans un cadre plus général pouvant comporter une structuration géographique et/ou des variations des paramètres démographiques (par ex. Cornuet & Luikart, 1996; Estoup *et al.*, 2001) tant que l'approximation en temps continu est valide. Elle est par exemple utilisable dans le cadre du coalescent structuré. Le principe est de remonter le temps événement par événement, sans considérer les mutations (donc uniquement les coalescences et migrations), selon les probabilités données par les équations (1.7) et (1.8). Pour chaque événement de coalescence ou de migration, on tire le moment dans le passé auquel cet événement s'est produit dans la loi exponentielle de moyenne égale à l'inverse des probabilités mentionnées ci-dessus. Cela permet de décrire, en même temps, la topologie et la longueur des branches de l'arbre de coalescence. Lorsqu'on arrive au MRCA de l'échantillon, l'arbre de coalescence de l'échantillon est complètement décrit et il suffit d'ajouter les mutations sur les différentes branches

en redescendant dans le temps en tirant dans une loi binomiale ou de Poisson comme vu précédemment.

La troisième méthode de simulation d'échantillons de gènes par coalescence correspond aux algorithmes dits génération par génération, que nous utiliserons beaucoup par la suite. Le principe est très simple et assez proche de la simulation en approximation continue de Hudson. La différence est que l'on ne remonte pas le temps événement par événement mais génération par génération. En d'autres termes, on envisage tous les événements possibles de coalescence ou de migration de tous les gènes de l'échantillon à chaque génération pour créer l'arbre de coalescence. Les mutations sont ensuite ajoutées sur l'arbre comme pour la méthode de Hudson. Un exemple très détaillé d'un tel algorithme sera développé dans le chapitre suivant.

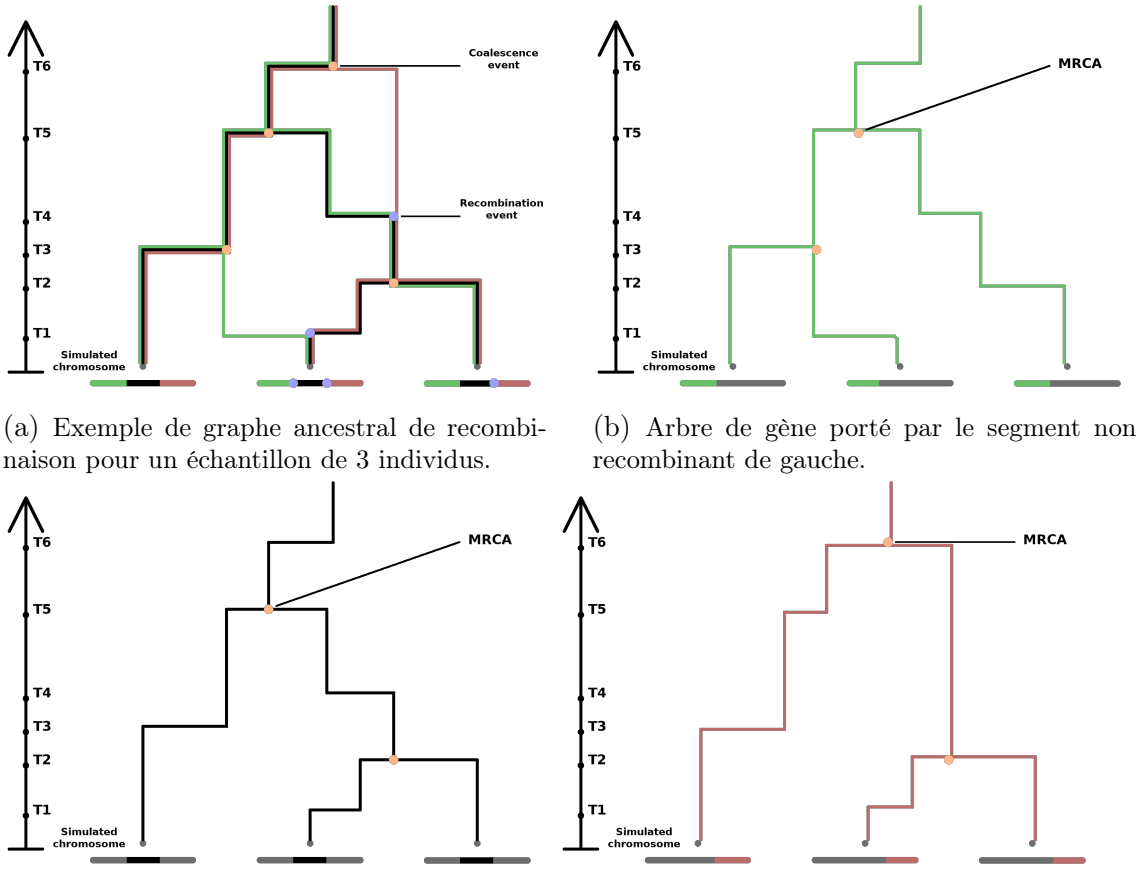
Ces trois méthodes différent essentiellement par leur rapidité et par la complexité des modèles démographiques qu'elles peuvent prendre en compte. La plus rapide, le modèle en Urne, ne s'applique que dans des situations simples s'apparentant au n-coalescent. La méthode la plus lente, l'algorithme génération par génération, peut s'appliquer sous n'importe quel type de modèle démographique de type WF sans aucune approximation. C'est une description exacte des processus de Wright-Fisher sous-jacents aux modèles démographiques considérés. Enfin, l'algorithme en temps continu de Hudson peut s'appliquer à de nombreux modèles pour peu que l'on dispose d'une loi donnant l'expression des temps d'attente entre deux événements. Ceci n'est pas trivial dans de nombreux cas, par exemple lorsque l'on considère des petites populations, des taux de migration forts ou encore des démographies très variables dans le temps. Dans ces cas, seul l'algorithme génération par génération permettra de simuler des patrons génétiques de manière satisfaisante.

De nombreux simulateurs de données génétiques implémentent ces algorithmes de coalescence dans le cadre de différents modèles démo-génétiques plus ou moins flexibles. Les plus classiquement utilisés sont : `ms` qui implémente l'algorithme de Hudson en temps continue avec prise en compte de la recombinaison (Hudson, 2002) ; `SIMCOAL2` basé sur un algorithme génération par génération avec recombinaison (Laval & Excoffier, 2004) ; et `fastsimcoal2` (Excoffier & Foll, 2011) qui considère un algorithme en temps continu prenant aussi en compte la recombinaison mais selon l'approximation dite SMC de McVean & Cardin (2005) ; et plus récemment, `MSprime` version améliorée de `ms`, qui implémente un algorithme en temps continu plus efficace pour la simulation de gros jeux de données génomiques (Kelleher *et al.*, 2016) et a largement pris le dessus sur tous les autres. De notre coté, nous avons implémenté `IBDsim` (Leblois *et al.*, 2009) et `GSpace` (Virgoulay *et al.*, 2021) car nous trouvons qu'il est toujours plus confortable (voir nécessaire pour faire du développement comme on le souhaite) d'utiliser ses propres outils. Ce sont deux simulateurs générations par générations spécifiquement développés pour faire des simulations spatiales, et dont nous détaillerons les caractéristiques dans les prochains chapitres.

coalescence et recombinaison

Considérer la recombinaison dans un processus de coalescence n'est pas trivial, complique fortement les calculs et ralentit les simulations. En effet, en remontant dans le temps, la coalescence va "rassembler" les lignées ancestrales qui coalescent et donc réduire le nombre de lignées ancestrales suivies. Au contraire, la recombinaison va les "séparer" en créant deux arbres ancestraux de coalescence pour chacune des

lignées ainsi séparées. Cette approche mène à reconstruire non plus un arbre mais un ensemble d'arbres pour un échantillon donné, appelé graphe ancestral de recombinaison ("ancestral recombination graph" ARG, voir figure 1.4a ; Hudson 1983). Chacun des arbres du graphe ancestral de recombinaison correspond à l'arbre de coalescence d'une partie du génome n'ayant jamais recombiné entre le présent et les MRCA de l'échantillon (appelé "segment non recombinant", voir figure 1.4). J'évoquerai rapidement la coalescence avec recombinaison dans le chapitre 4 mais le lecteur pourra se référer au livre de Hein *et al.* (2005) qui illustre très bien la coalescence avec recombinaison le long des génomes.



(a) Exemple de graphe ancestral de recombinaison pour un échantillon de 3 individus.

(b) Arbre de gène porté par le segment non recombinant de gauche.

(c) Arbre de gène porté par le segment non recombinant central.

(d) Arbre de gène porté par le segment non recombinant de droite.

FIGURE 1.4 – Exemple de graphe ancestral de recombinaison et de l'ensemble des arbres de gène le composant. On peut remarquer que les topologies des arbres de gène portés par deux segments non recombinants voisins ne sont pas indépendantes. Elles ne diffèrent qu'à partir du moment (en remontant dans le temps) où un évènement de recombinaison va séparer les lignées ancestrales portées par les deux segments. Les points mauves représentent les évènements de recombinaison et les points jaunes les évènements de coalescence. Figure issue de Virgoulay (2022).

Nous avons vu dans cette section le principe de la théorie de la coalescence, comment elle permet de dériver des modèles stochastiques simples pour une population panmictique ou une population structurée, et les algorithmes de simulation en découlant. Voir la génération d'un échantillon en remontant dans le temps est un peu contre-intuitif, mais une fois que le raisonnement est rodé, c'est souvent la façon la plus simple de comprendre l'histoire des populations ayant généré le polymorphisme observé. Quelques exemples de raisonnements simples sur le polymorphisme attendu dans un échantillon à partir de la théorie de la coalescence peuvent être trouvés dans

Nielsen & Sltakin (2013), notamment p.56 pour les exemples classiques de population ayant subi des contractions ou expansions passées. On peut noter ici que le coalescent (avec approximation donc) est insensible (invariant) à des changement de valeurs de certains paramètres canoniques du modèle de Wright-Fisher à produit constant. Ainsi, le coalescent pour $N' = 3N$ et $\mu' = \mu/3$ sera strictement identique (par homothétie) au coalescent pour N et μ. Cela indique que toute méthode d'inférence basée sur le coalescent ne pourra estimer que le produit $N\mu$ et non N et μ séparément. Ainsi, on peut définir l'inférence d'une méthode en termes de paramètres canoniques du modèle d'inférence, et non des paramètres canoniques du modèle démo-génétique sous-jacent mais non-approché. Ceci est valable pour toutes les méthodes d'inférences mais parfois oublié lors de l'application de ces inférences sur données réelles.

Nous verrons par la suite comment la coalescence peut aider à l'interprétation des performance de certaines méthodes d'inférence et notamment de la robustesse de certaines estimations. Nous verrons aussi comment la coalescence peut être utilisée pour estimer la vraisemblance d'un échantillon de gène sous certains modèles. Cependant, cette estimation de la vraisemblance par coalescence est très complexe et nous voyons tout de suite en quoi le calcul des probabilités d'identité d'une seule paire de gènes, et non de l'échantillon complet, était et reste très utile en génétique des populations.

1.2 Probabilités d'identité, F-statistiques, et temps de coalescence

Dans cette partie, je m'attacherai à définir et discuter des concepts d'identité de paires de gènes et des F-statistiques qui sont à la source de nombreuses analyses de génétique des population et abondamment utilisés dans la suite de ce document. Nous n'envisagerons ici que l'identité génétique à un locus, pour une extension à plusieurs locus le lecteur peut lire par exemple Vitalis & Couvet (2001b,a). Une brève description des probabilités d'identité entre paires de locus est aussi décrite en section 4.2.2. Décrire les différents états alléliques et déterminer l'identité exacte, au sens généalogique, des gènes d'un échantillon n'est pas tout à fait équivalent. En effet, seule la description des états alléliques est envisageable en pratique, tandis que la détermination de la vrai généalogie des gènes parait difficilement accessible. Il s'agit là de la différence entre l'identité par descendance ("identity by descent", Crow, 1954) et l'identité par état ("identity in state", Kempthorne, 1954), qui ne sera que très brièvement discutée ci-dessous. La notation IBD, largement utilisée dans la littérature pour désigner l'identité par descendance, sera ici strictement réservée dans ce document aux modèles d'isolement par la distance (IBD = "isolation by distance") que nous décrirons dans les sections suivantes. Pour la suite, on s'intéressera à la probabilité d'identité par descendance et, sauf précision contraire, on omettra d'expliciter, mais non de discuter, l'effet des modèles mutationnels. Les différentes approches de l'analyse des modèles en terme de probabilités d'identité et F-statistiques que j'ai choisies de présenter dans ce document proviennent en grande partie de Rousset (2004) et de ses travaux antérieurs.

1.2.1 Probabilités d'identité

Comme on le comprend intuitivement maintenant, toute paire de gène homologues échantillonnée dans une même espèce est issue d'un évènement de coalescence. On cherche ici à calculer la probabilité que les deux gènes soient identiques par descendance, et donc qu'il n'y ait pas eu de mutation depuis la génération t de leur ancêtre commun. Cette notion de probabilité d'identité par descendance entre paires de gènes, que l'on notera Q, a été utilisée pour la première fois par Gustave Malécot (voir Nagylaki, 1989), puis repris par de nombreux généticiens des populations à partir des années 60 jusqu'à maintenant. Nous n'en donnerons ici qu'une présentation minimaliste permettant de bien comprendre la suite de ce document.

Les probabilités d'identité par descendance peuvent donc directement être exprimées en fonction de la probabilité C_t que la paire de gène considérée coalesce t générations dans le passé et de la probabilité qu'il n'y ait pas eu de mutation, par l'équation $Q = \sum_{t=0}^{\infty} C_t(1-\mu)^{2t}$, ou μ est la probabilité ("taux") de mutation par génération à ce locus (Malécot, 1975; Slatkin, 1991). En fonction du "type" de la paire de gène que l'on considère (par ex., au sein d'un individu, entre deux individus, au sein d'un même dème ou entre dèmes différents), ces probabilités d'identité peuvent être facilement calculées en fonction des paramètres pour de nombreux modèles. Et comme on peut en calculer des estimateurs sur des données réelles, elles peuvent aussi servir à extraire de l'information des données génétiques et faire des inférences à l'aide de la méthode des moments ou d'approches basées sur la simulation. C'est un point que nous allons détailler dans divers sections de ce document.

Comme nous l'avons vu précédemment, dans une population panmictique constituée de N individus diploïdes et suivant le cycle de vie présenté p. 4, la probabilité que deux individus partagent un même parent est de $\frac{1}{N}$, et la probabilité qu'ils aient reçu une copie du même gène parental est alors $\frac{1}{2N}$ puisque le parent est diploïde. Par conséquent, on peut écrire la probabilité d'identité par descendance Q au temps $t+1$ (chez les descendants) de deux gènes pris au hasard dans la population, en fonction de cette même probabilité Q au temps t (chez les parents), comme

$$Q(t+1) = (1-\mu)^2\Big[\frac{1}{2N} + \Big(1 - \frac{1}{2N}\Big)Q(t)\Big]. \tag{1.9}$$

En notant $\gamma \equiv (1-\mu)^2$ la probabilité qu'aucun des deux gènes n'ait muté entre t et $t+1$, on obtient à l'équilibre (c.a.d. en posant $Q(t+1) = Q(t)$)

$$Q = \frac{\gamma}{2N(1-\gamma)+\gamma} \approx^{\mu\to 0} \frac{1}{1+4N\mu}. \tag{1.10}$$

La diversité génétique (ou hétérozygotie "attendue") correspondant à $H_e = 1 - Q = \theta/(1+\theta)$, on a donc très facilement obtenu sa définition en fonction de $\theta = 4N\mu$, le produit de la taille de la population N et du taux de mutation μ, les deux paramètres canonique du modèle panmictique.

Plus généralement, il est utile de définir des probabilités d'identité de paires des gènes à différents niveaux hiérarchiques pour décrire différents modèles de populations, subdivisées ou non. Il est toujours possible de définir des classes de gènes de telle sorte que le modèle est entièrement décrit par les probabilités d'identité de paires de gènes intra et inter-classes, ceci pour n'importe quel type de structure (voir par ex. Rousset, 1999a,b). Dans l'exemple précédent de la population panmictique isolée, on définira Q_0 la probabilité d'identité de paires de gènes pris au sein d'un individu diploïde et Q_1 la probabilité d'identité de paires de gènes pris dans

deux individus différents de la population. Dans un modèle en population subdivisée, on ajoutera par exemple Q_2 la probabilité d'identité de paires de gènes pris dans deux sous-populations différentes. Ou encore, dans un modèle d'isolement par la distance, on considéreraQ_r la probabilité d'identité de paires de gènes séparées par une distance géographique r. Nous reviendrons plus en détails sur ces modèles démographiques de populations structurées dans la section 1.3.

Ainsi, dans une population structurée en n_d dèmes de taille identiques N individus diploïdes échangeant des migrants à un même taux $m/(n_d-1)$ pour toutes les paires de dèmes (le modèle en îles... que nous verrons un peu plus tard), on s'intéressera tout d'abord à l'effet de la migration. On appellera m le taux d'émigration, représentant donc la probabilité qu'un individu disperse à chaque génération. Ainsi dans le modèle en île, pour tout i et j, $m_{ij} = m/(n_d-1)$. On peut tout d'abord définir $a = (1-m)^2 + m^2/(n_d-1)$ la probabilité que deux gènes pris dans un même dème soient (des copies de gènes) issus d'un même dème à la génération précédente. De la même manière, pour une paire de gènes pris dans deux dèmes différents, $b = 2m(1-m)/(n_d-1)$ correspond à la probabilité qu'un des deux gènes n'ait pas migré, et que l'autre provienne du même dème, et $c = (n-2)\big(m/(n_d-1)\big)^2$ à la la probabilité que les deux gènes aient migré à partir du même dème. Ainsi, la probabilité que deux gènes pris dans deux dèmes différents soient issus du même dème est donc $d = b + c = 2m(1-m)/(n_d-1) + (n-2)\big(m/(n_d-1)\big)^2 = (1-a)/(n-1)$. Avec le même raisonnement que pour l'équation 1.10, on peut donc écrire

$$Q_1(t+1) = \gamma\left[a\Big(\frac{1}{2N} + \Big(1-\frac{1}{2N}\Big)Q_1(t)\Big) + (1-a)Q_2(t)\right] \tag{1.11}$$

et

$$Q_2(t+1) = \gamma\left[d\Big(\frac{1}{2N} + \Big(1-\frac{1}{2N}\Big)Q_1(t)\Big) + (1-d)Q_2(t)\right]. \tag{1.12}$$

Nous utiliserons ces récurrences dans la section suivante pour exprimer le F_{ST} en fonction des paramètres N, m et n_d du modèle en îles.

Dans le cas de données observées à un marqueur génétique par une certaine technique de biologie moléculaire, on dira que deux gènes sont *identiques par état* s'ils appartiennent à la même classe allélique. Par exemple pour un locus microsatellite (voir Estoup & Angers, 1998; Estoup *et al.*, 2002), on dira que deux gènes sont identiques si les fragments d'ADN amplifiés sont de longueur identique (et donc migrent à la même distance sur un gel d'électrophorèse). En revanche, les séquences de ces deux gènes peuvent être différentes. Cet exemple illustre le concept d'*homoplasie*, qui se défini plus généralement comme le fait que deux gènes sont identiques par état mais pas par descendance, ce qui arrive aussi si une mutation qui survient après une autre rétablit l'état initial (mutation reverse). L'homoplasie provient donc en partie de notre perception des états alléliques des marqueurs génétiques mais aussi de la nature des mutations. Pour des marqueurs SNP, l'homoplasie est le plus souvent négligée du fait de la faible probabilité de mutation reverse sur un seul site à des échelles évolutives pas trop grandes (c.a.d. populationnelles).

Plusieurs modèles ont été développés pour traiter les processus de mutation. Le modèle à nombre d'allèles infini ("infinite allele model", IAM, Kimura & Crow, 1964), dans lequel chaque mutation créé un allèle différent de tous les allèles déjà présents dans la population, décrit donc l'identité par descendance. Deux autres modèles ont été développés pour traiter le polymorphisme enzymatique puis utilisés pour d'autres marqueurs tels que les microsatellites. Ce sont le modèle à K allèles ("K

allele model", KAM, Crow & Kimura, 1970) et le modèle par pas ("stepwise mutation model" SMM, Ohta & Kimura, 1973). Sous le modèle KAM, la mutation engendre un allèle parmi K allèles possibles de façon équiprobable. Pour le modèle par pas, les possibilités de mutation sont beaucoup plus restreintes que pour les modèles précédents : une mutation diminue ou augmente, en proportions égales, le nombre de répétitions d'une unité. Un quatrième modèle mutationnel, spécifiquement adapté au cas des marqueurs microsatellites (mais le SMM est aussi bien adapté), est le GSM ("generalised stepwise mutation"). Sous ce modèle, une mutation augmente ou diminue le nombre de répétitions d'un certain nombre d'unités, tiré dans une loi géométrique. Le principe de l'intégration des modèles mutationnels dans les calculs d'identité par état à partir de l'identité par descendance est décrit dans Rousset (1996, 2004), mais n'est pas nécessaire pour la compréhension de la suite.

1.2.2 F-statistiques et F_{ST}

Afin de quantifier la structure des populations, Dobzhansky & Wright (1941) ont utilisé le rapport des variances des fréquences alléliques entre sous-populations sur la variance intra-populations, $\mathrm{Var}(p)/[\bar{p}(1-\bar{p})]$ ou $\mathrm{Var}(p)$ est la variance et $\bar{p}$ la moyenne des fréquences alléliques entre sous-populations, mieux connu sous le nom de F_{ST} (Wright, 1951, 1969) et aussi appelés *coefficients de consanguinité* inter populations ("inbreeding coefficients"). Bien que cette définition du F_{ST} en terme de variance des fréquences alléliques soit la plus connue et la plus utilisée en génétique des population, la relation entre la variance et la moyenne des fréquences alléliques avec l'apparentement des individus au sein et entre les sous-population n'est pas évidente, et l'interprétation des fréquences alléliques comme variables aléatoires ou comme paramètres du modèle peut mener à certaines confusions (voir Rousset, 2006). C'est pourquoi il peut paraître plus utile de considérer une définition alternative du F_{ST}, et plus généralement des F-statistiques, fondée sur les probabilités d'identité, dont l'interprétation en terme de parenté, et le lien direct avec la théorie de la coalescence, sont plus directs comme montrés ci-dessus. Depuis Wright, la définition et l'estimation des F-statistiques ont fait l'objet d'un vaste débat (Chakraborty & Danker-Hopfe, 1991; Excoffier, 2001; Rousset, 2001; Weir & Cockerham, 1984). Notons que l'appellation F-statistiques est en complète contradiction avec notre terminologie puisque nous les considérons comme des paramètres, et seul leurs estimateurs sont des statistiques. Nous retiendrons dans ce document uniquement l'approche développée par Cockerham (1969, 1973). Le point important de ces développements est que la décomposition de la variance totale du modèle considéré par Cockerham conduit naturellement à l'expression des F-statistiques en terme de rapports de probabilités d'identité par état (Rousset, 2001). Ainsi, on peut définir les paramètres F_{IS} F_{ST} et F_{IT} comme

$$F_{\mathrm{IS}} \equiv \frac{Q_0 - Q_1}{1 - Q_1};\ F_{\mathrm{ST}} \equiv \frac{Q_1 - Q_2}{1 - Q_2};\ F_{\mathrm{IT}} \equiv \frac{Q_0 - Q_2}{1 - Q_2} \tag{1.13}$$

(voir Cockerham & Weir, 1987; Rousset, 1996). Ces expression mesurent la divergence entre gènes inter-classes relativement à la divergence intra-classe. Les F-statistiques définies par les équations (1.13) sont donc bien des paramètres et non des statistiques. Cette écriture permet également de proposer des estimateurs de la forme

$$\hat{F}_{\mathrm{IS}} \equiv \frac{\hat{Q}_0 - \hat{Q}_1}{1 - \hat{Q}_1};\ \hat{F}_{\mathrm{ST}} \equiv \frac{\hat{Q}_1 - \hat{Q}_2}{1 - \hat{Q}_2};\ \hat{F}_{\mathrm{IT}} \equiv \frac{\hat{Q}_0 - \hat{Q}_2}{1 - \hat{Q}_2}. \tag{1.14}$$

Rousset (2001) montre que ces estimateurs sont exactement identiques à ceux de Weir & Cockerham (1984).

Les F-statistiques peuvent donc être définies en terme de probabilités d'identité de paires de gènes intra- et inter-classes, qui elle mêmes peuvent être exprimées en fonction des paramètres du modèle considéré, comme nous l'avons vu dans le section précédente pour la diversité génétique H_e. Il est donc facile d'obtenir les expression des F-statistiques en fonction des paramètres du modèle.

Ainsi, pour le F_{IS} dans une population panmictique diploïde avec un taux d'auto-fécondation s, l'équation de récurrence 1.9 est adaptée pour donner $Q_1(t+1) = (1-\mu)^2\Big[\frac{1}{2N}\frac{Q_0(t)+1}{2} + \Big(1-\frac{1}{2N}\Big)Q_1(t)\Big]$, et on peut aisément poser $Q_0(t+1) = (1-\mu)^2\Big[s\frac{Q_0(t)+1}{2} + \Big(1-s\Big)Q_1(t)\Big]$. On voit déjà que si $s = 0$ alors les deux équations sont équivalentes et F_{IS} sera nul. Pour $s > 0$, on obtient à l'équilibre $F_{IS} = \gamma(Ns-1)/(2N-\gamma(Ns-1))$, et pour des taux de mutation faibles $F_{IS} = (s-1/N)/(2-(s-1/N))$.

De manière équivalente, on peut calculer les attendus à l'équilibre des probabilités d'identité Q_1 et Q_2 dans un modèle en îles. La résolution des équation de récurrences 1.11 et 1.12 par une analyse matricielle des probabilités d'identité permet notamment d'exprimer les ratios de probabilités d'identité suivant, en fonction des paramètres du modèle

$$\frac{Q_1}{1-Q_1} = \frac{1}{2Nn_d}\Big(\frac{\gamma}{1-\gamma} + (n_\mathrm{d}-1)\frac{\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}{1-\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}\Big), \tag{1.15}$$

et

$$\frac{Q_2}{1-Q_1} = \frac{1}{2Nn_d}\Big(\frac{\gamma}{1-\gamma} - \frac{\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}{1-\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}\Big). \tag{1.16}$$

La différence entre ces deux expressions amène à l'expression suivante, d'une forme plus simple,

$$\frac{Q_1-Q_2}{1-Q_1} = \frac{F_\mathrm{ST}}{1-F_\mathrm{ST}} = \frac{1}{2N}\frac{\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}{1-\gamma(1-m\frac{n_\mathrm{d}}{n_\mathrm{d}-1})^2}. \tag{1.17}$$

Contrairement aux expressions (1.15) et (1.16), cette expression a une limite finie quand $\mu \to 0$. On peut noter que toutes ces expressions sont fonctions du ratio $n_\mathrm{d}/(1-n_\mathrm{d})$ et ne dépendent donc que peu du nombre d'îles du modèle sauf si ce nombre est faible. Dans le cas limite où $n_\mathrm{d} \to \infty$, considéré par Wright dans son modèle à nombre d'îles infini, on a

$$\frac{F_\mathrm{ST}}{1-F_\mathrm{ST}} = \frac{1}{2N}\frac{\gamma(1-m)^2}{1-\gamma(1-m)^2} \approx \frac{1}{2N}\frac{\gamma(1-2m)}{1-\gamma(1-2m)}. \tag{1.18}$$

De cette expression, on obtient facilement le fameux résultat de Wright

$$F_\mathrm{ST} = \frac{1}{1+2N(2\mu+2m)} \approx_{\mu\to 0} \frac{1}{1+4Nm}. \tag{1.19}$$

On peut aussi retrouver l'expression correspondante pour un nombre d'îles fini, donnée par Li (1976),

$$F_\mathrm{ST} = \frac{1}{1+2N(2\mu+2\frac{n_\mathrm{d}}{n_\mathrm{d}-1}m)} \approx_{\mu\to 0} \frac{1}{1+4N\frac{n_\mathrm{d}}{n_\mathrm{d}-1}m}. \tag{1.20}$$

Les F-statistiques ont fait et font encore l'objet d'une littérature très abondante (voir par ex. les discussions de Jakobsson *et al.*, 2013; Rousset, 2013). La première

raison est qu'elles peuvent être utilisées pour estimer certains paramètres populationnels à partir d'un échantillon génétique avec les formules vues ci-dessus. La formule de Wright de l'estimateur du *nombre de migrants par génération*, dans le modèle à nombre d'îles infini, $Nm = (1/F_{\mathrm{ST}} - 1)/4$ a été largement utilisée pour décrire la structuration des populations naturelles, bien que les conditions d'application de cette formule, liées aux hypothèses peu réalistes du modèle à nombre d'îles infini, ne soient généralement pas remplies (Whitlock & McCauley, 1999), voir section 1.5. Les expressions (1.19) et (1.20) ont aussi largement été utilisées pour caractériser des taux de migration entre paires de sous-populations en calculant un $\hat{F}_{\mathrm{ST}}$ entre deux sous-populations et en exprimant le résultat en terme de nombre de migrants, $\hat{Nm} \approx (1/\hat{F} - 1)/4$ (inférence par la méthode des moments, voir section 1.4). Ce type de raisonnement n'est en aucun cas correct puisqu'un F_{ST} entre deux sous-populations n'est pas, dans le modèle en île, fonction du nombre de migrants entre ces deux sous-populations, mais du nombre de migrants moyen entre une sous-population et toutes les autres sous-populations du modèle. De plus, il est attendu que deux sous-populations, n'échangeant aucun migrant entre-elles mais échangeant des migrants avec d'autres sous-populations, aient entre-elles un F_{ST} non nul et donc en apparence un nombre de migrants, Nm, non nul. C'est un bon exemple de mauvaise compréhension des modèles en génétique des populations et du manque d'exploration de la robustesse d'un tel raisonnement. La formule de Wright $F_{\mathrm{ST}} = 1/(1 + 4Nm)$ a aussi largement contribué à imposer l'idée que la différenciation entre sous-population est (uniquement) fonction du nombre de migrants échangés à chaque génération. C'est vrai pour ce modèle en îles, mais nous verrons par la suite que la distribution des distances de dispersion entre parents et descendants, est aussi un aspect crucial mais souvent négligé de la différenciation des populations naturelles. Enfin, la seconde raison, expliquant l'utilisation massive des F-statistiques, est qu'elles apparaissent également dans les modèles d'adaptation en populations subdivisées (Gandon & Rousset, 1999; Roze & Rousset, 2003; Whitlock, 2003).

1.2.3 Liens entre probabilités d'identités, F-statistiques et temps de coalescence

Comme nous l'avons vu ci-dessus, il existe une relation étroite entre les probabilités d'identité par état de paires de gènes, les F-statistiques et les temps de coalescence (Malécot, 1975; Slatkin, 1991), notamment via la relation $Q_I = \sum_{t=1}^{\inf} \gamma^t C_I(t)$ ou I prenant la valeur 0 si les gènes sont pris au sein d'un même individu, 1 s'ils sont pris dans deux individus distincts de la même population, 2 s'ils sont pris dans deux populations différentes et r s'ils sont séparés par une distance géographique r. D'après cette définition des probabilités d'identité, et de la définition des F-statistiques en terme de probabilités d'identité (éq.1.13), on peut exprimer les F-statistiques en fonction des temps moyens de coalescence pour différentes paires de gènes. Ainsi, pour F_{ST} on a

$$\lim_{\mu \to 0}(F_{\mathrm{ST}}) = \frac{T_2 - T_1}{T_2} \tag{1.21}$$

où $T_I = \sum_{t=1}^{\infty} t C_I(t)$ est le temps moyen de coalescence de paires de gènes dans la classe I. On peut alors mieux comprendre et interpréter les propriétés des F-statistiques à travers l'étude des probabilités de coalescence des paires de gènes pris dans différentes classes (Figure 1.5).

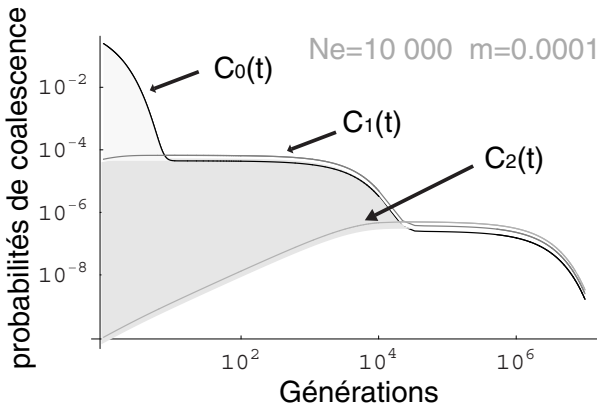


FIGURE 1.5 – Probabilités de coalescence en fonction du temps dans un modèle en îles. Les probabilités de $C_I(t)$ que deux gènes coalescent au temps t sont représentées pour différentes paires de gènes : 0 pour des gènes intra-individu, 1 pour des gènes intra-dème mais inter-individu et 2 pour des gènes inter-dème. N=10000, m=0.0001 et n=100. Pour cette figure, nous avons considéré un taux d'autofécondation de $s = 0.5$ afin d'avoir une différence marquée entre $C_0(t)$ et $C_1(t)$. L'aire gris clair correspond au F_{IS} et celle en gris foncée à F_{ST}. L'échelle des deux axes est logarithmique. D'après Rousset (2004).

La figure 1.5 montre que, pour des temps anciens, la distribution des probabilités de coalescence dans une classe de gènes, par exemple $C_0(t)$, est proportionnelle à la distribution des probabilités de coalescence dans une classe de gènes moins apparentés, par exemple $C_1(t)$. Au contraire, dans une période de temps assez récente, les distributions différent. On peut donc décomposer la surface couverte par $C_0(t)$ en la somme de la surface couverte par $C_1(t)$ et une surface représentée par la région gris clair sur la figure 1.5. Cette surface gris claire représente une masse de probabilité équivalente à F_{IS} (Rousset, 2001). De même, F_{ST} (région gris foncé sur la figure 1.5) est approximativement équivalent à la masse de probabilité correspondant à la différence des distributions $C_1(t)$ et $C_2(t)$. D'après cette figure, on comprend aisément que F_{IS} ne dépend que des événements récents de coalescence et qu'il est donc très peu influencé par la mutation (Rousset, 1996). Par contre, dans le modèle en îles considéré ici, F_{ST} dépend des événements de coalescence plus anciens. Il sera donc plus sensible à la mutation (c.a.d. taux de mutation et modèle mutationnel) que le F_{IS}. Nous verrons, dans la suite de ce document, comment différents facteurs démographiques (tailles de population et taux de migration) influencent ces courbes, et comment elles nous permettent de mieux appréhender les effets des processus mutationnels et démographiques sur les F-statistiques.

Cette interprétation des F-statistiques par la coalescence et une notion similaire à celle de séparation des échelles de temps discutée précédemment. On retrouve, par exemple, qu'une augmentation du nombre de dèmes diminuera la probabilité de coalescence inter-dème tout en gardant une probabilité de coalescence intra-dème constante. Dans le cas limite où $n_{\text{d}} \to \infty$, la probabilité de coalescence inter-dème $C_2(t)$ va tendre vers zéro pour tout t. La figure 1.6 montre bien que lorsque l'on augmente le nombre de dèmes , $C_2(t)$ diminue pour tout t.

1.3 Vers une dispersion réaliste

Comme nous l'avons vu en introduction, la distribution du polymorphisme peut nous renseigner sur les paramètres démographiques et génétiques influant sur la structuration des populations tels que les tailles (et/ou densités) des populations

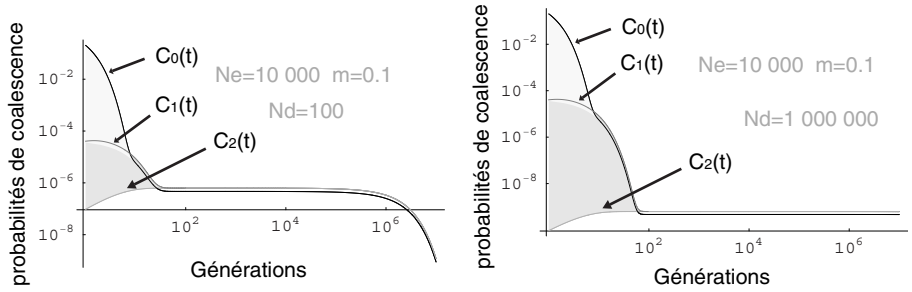


FIGURE 1.6 – Probabilité de coalescence dans un modèle en îles pour deux nombres de dèmes. (a) 100 dèmes (b) 1 000 000 dèmes. Les probabilités de $C_I(t)$ que deux gènes coalescent au temps t sont représentées pour différentes paires de gènes $0, 1, 2$. Pour cette figure, nous avons considéré un taux d'autofécondation de $s = 0.5$ afin d'avoir une différence marquée entre $C_0(t)$ et $C_1(t)$. L'échelle est une double échelle logarithmique.

ainsi que sur les flux de gènes potentiels entre sous populations et/ou leurs caractéristiques de dispersion. Ces patrons spatiaux de polymorphisme sont complexes et le développement de modèles en permet une meilleure analyse. Le premier modèle de populations structurées, le plus simple, est le modèle en îles de Wright. Un des points faibles de ce modèle est la modélisation de la dispersion. En effet, l'hypothèse est faite que la dispersion se fait de façon équiprobable entre toutes les sous-populations, ce qui intuitivement semble en désaccord avec la réalité dans de nombreuses situations biologiques. Malgré cette faiblesse, le modèle en îles a été, et est toujours, largement utilisé pour comprendre les conséquences évolutives de la dispersion, voir dans des procédures d'inférence.

Dans de nombreuses espèces, la dispersion est restreinte dans l'espace, et la différenciation génétique est plus faible à petite distance qu'à grande distance. Ceci est la pierre angulaire des modèles d'isolement par la distance introduits par Wright (1943, 1946). Dans cette section, je tacherai d'introduire ces différents modèles démographiques, en donnant les grandes lignes des analyses que l'on peut en faire et les principaux résultats qui en découlent, ceci sans entrer dans les détails mathématiques que le lecteur pourra trouver, par exemple, dans Rousset (2004).

Gardons en mémoire que nous essayons le plus souvent d'analyser des échantillons provenant de grands ensembles de sous-populations. Non seulement (i) en raison du problème des populations existantes mais non échantillonnées, mais aussi (ii) parce que de nombreuses populations naturelles peuvent être décrites comme un vaste réseau de petits dèmes panmictiques reliées par une forte dispersion. Souvent même, aucune sous-population n'existe vraiment et la population est décrite par un réseau d'individus ou des couples, et non de dèmes, répartis sur un habitat continu avec une dispersion forte mais localisée, c.a.d. à des distances beaucoup plus petites que la taille de l'habitat (Sumner *et al.*, 2001; Fenster *et al.*, 2003; Watts *et al.*, 2006). Les valeurs typiques des scénarios biologiques nous intéressant serait donc des tailles de dème de 1 à 100 individus, et des probabilités d'émigration de 0.2 à 0.5 sur des distances assez faibles. Un exemple de répartition des individus dans l'espace en populations naturelles est illustrée sur la Figure 1.7.

1.3.1 Modèle en îles et stepping-stone

Le premier modèle défini pour prendre en compte des populations structurées est donc le modèle en île (Wright, 1931), représenté schématiquement en Figure 1.8 et déjà évoqué précédemment. Ce modèle considère une population constituée de

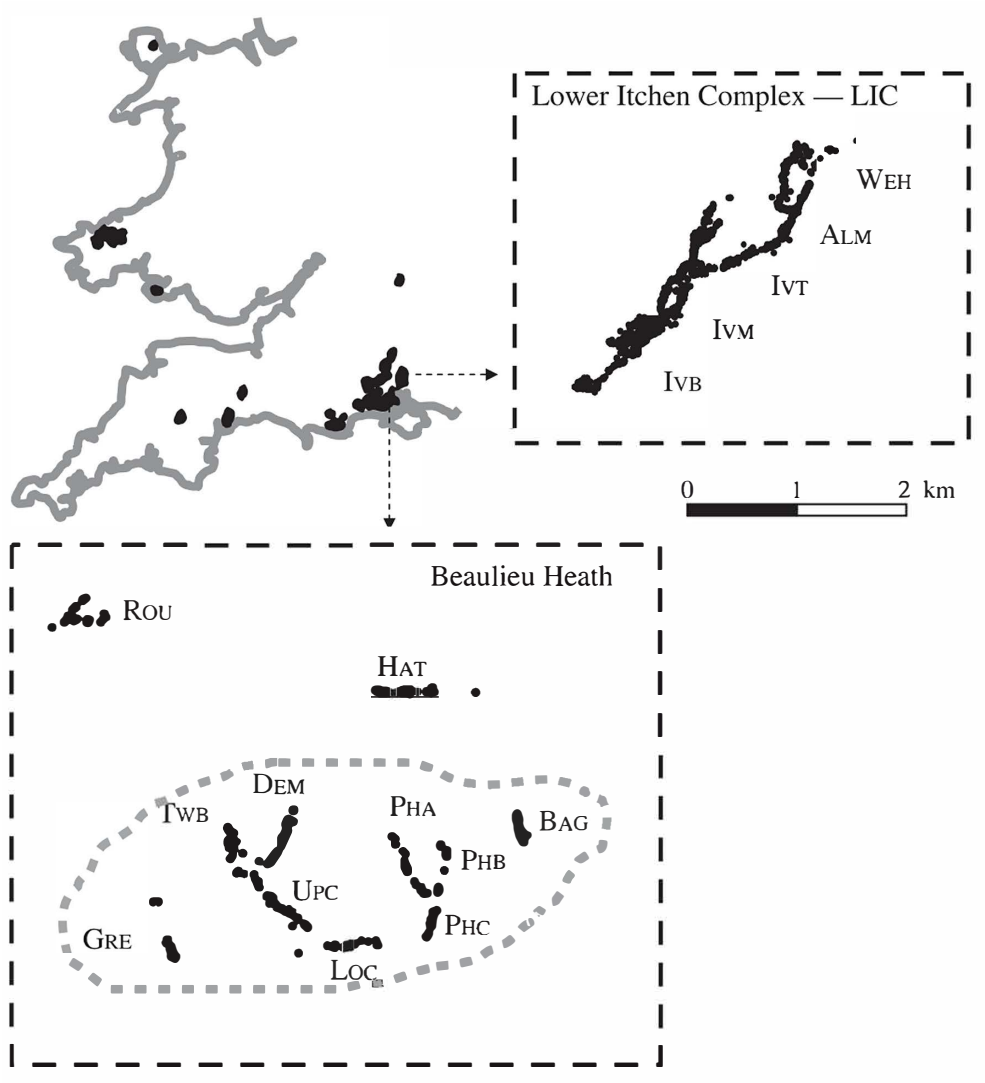


FIGURE 1.7 – Distribution approximative des libellules demoiselles *Coenagrion mercuriale* à travers le Royaume-Uni et localisation des sites d'étude de Watts *et al.* (2006). Les zones agrandies montrent une distribution plus précise de *C. mercuriale* dans chaque site. Figure issue de Watts *et al.* (2006)

n_d sous-populations panmictiques (ou dèmes), chacune de taille constante égale à $2N$ gènes (ou N individus adultes diploïdes). Le cycle de vie est le même que celui considéré précédemment (voir p.4). On considère qu'une proportion des gamètes migrent de leur dème d'origine vers un des $n_d - 1$ autres dèmes avec la probabilité $m/(n_d - 1)$. On peut noter ici que le nombre de migrants n'est pas strictement égal à Nm, mais est une variable aléatoire suivant une loi binomiale de moyenne Nm. Enfin, la compétition entre juvéniles ramène le nombre d'individus à N adultes. Ce modèle n'est pas spatial car les places des dèmes peuvent être échangées sans aucune conséquence sur le modèle puisque les dèmes sont tous équivalents du fait de la migration isotrope.

Ce modèle est toujours largement utilisé pour étudier théoriquement l'effet de la migration (Mazet *et al.*, 2016; Arredondo *et al.*, 2021), et dans de nombreux modèles d'inférence pour prendre en compte une structuration simple des populations mais dans lesquels l'inférence précise de la migration n'est pas le but principal (voir par ex. Vitalis *et al.*, 2014). Comme nous l'avons vu précédemment, il a été largement utilisé, à plus ou moins bon escient, pour estimer un nombre de migrants entre dèmes

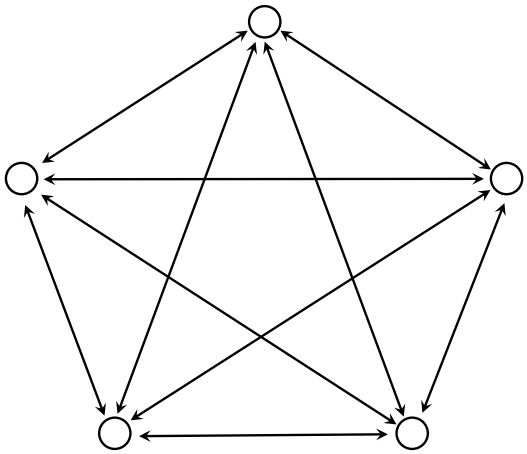


FIGURE 1.8 – **Représentation schématique du modèle en îles avec 5 îles.** Les cercles représentent les dèmes (sous-populations panmictiques) de taille N et les flèches la migration de taux $m' = m/(n_d - 1)$.

à partir d'un échantillon génétique grâce aux expressions (1.19) et (1.20) dans les années 80 à 2000. Pour ces formules, et plus largement dans divers méthodes d'inférence, les approximations du coalescent (grandes tailles de populations et faibles taux d'événements) ont souvent été appliquée lors de la considération du modèle en île. Ainsi, dans la plupart des cas, le modèle en îles est défini par les deux paramètres mis à l'échelle $\theta = 4n_dN\mu$ et $M = 4Nm$, avec $N \to \infty$ et μ et $m \to 0$.

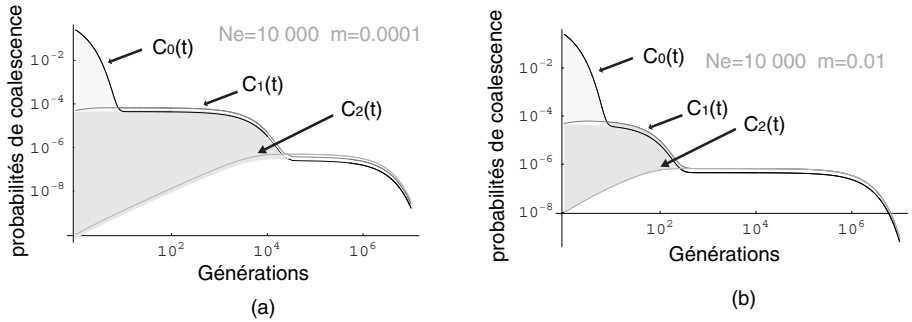


FIGURE 1.9 – **Probabilités de coalescence dans un modèle en îles pour deux taux de migration : (a) migration faible $m = 0.0001$ (b) migration forte $m = 0.01$.** Les probabilités, $C_I(t)$, que deux gènes coalescent au temps t sont représentées pour différentes paires de gènes 0,1,2. Pour cette figure, $n_\mathrm{d} = 100$, $Ne = 10000$ et nous avons considéré un taux d'autofécondation de $s = 0.5$ afin d'avoir une différence marquée entre $C_0(t)$ et $C_1(t)$. L'échelle des deux axes est logarithmique.

Si l'on reprend les raisonnements basés sur les relations entre F-statistiques et les temps de coalescence, on comprends bien que si les taux de migrations sont faibles et les tailles de populations grandes, alors F_ST et le estimations qui en découlent, seront fortement influencées par les processus du passés lointain. La figure 1.9 illustre bien que le paramètre F_ST, correspondant à l'aire gris foncé, dépendra fortement de la mutation, des changements démographiques et adaptatifs passés et ce d'autant plus que la migration est faible. Au contraire, si les taux de migration sont forts et/ou si les tailles de populations sont petites, il est attendu que l'estimation de Nm par les F_ST corresponde plus à la valeur actuelle du paramètre qu'à une moyenne des valeurs passées et elle sera moins influencée par les processus de mutation des marqueurs génétiques (Rousset, 2004).

L'application du modèle en îles avec les approximations du coalescent, ainsi que la modélisation trop simpliste de la migration (c.a.d. isotrope), sont sans doute les principales raisons pour lesquelles (i) Whitlock & McCauley (1999) concluent que le F_{ST} ne peut pas être utilisé pour estimer correctement la migration dans une population structurée ; (ii) Slatkin (1994) suggère qu'il y a une cohérence qualitative des estimations de la dispersion à partir de données génétiques mais que les résultats diffèrent quantitativement à cause de l'influence des processus passés ; et plus largement, (iii) de nombreuses études des années 2000 remettent clairement en question la pertinence des inférences démographiques "indirectes" du fait de l'inadéquation entre les modèles utilisés et une réalité beaucoup plus complexe (combinaison de processus démographiques actuels, passés, de processus mutationnels et des processus sélectifs, peu ou mal pris en compte, Boileau *et al.*, 1992; Koenig *et al.*, 1996).

Le modèle en stepping stone (Kimura & Weiss 1964) a été développé pour palier la dispersion non localisée du modèle en îles. C'est un modèle où les dèmes sont placés sur une grille régulière et n'échangent des migrants qu'avec leurs voisins immédiats. Il prend donc bien en compte la dispersion limitée dans l'espace et donc l'aspect spatial. Cependant, bien que la dispersion soit limitée, il ne permet pas de simuler d'événements de dispersion à longue distance qui, bien que rares, sont des évènements impactant fortement l'organisation de la diversité au sein des populations (Endler 1977, Excoffier & Ray 2008).

Il est aussi possible de considérer un modèle en île non-homogène, et pouvant donc considérer des situations plus complexes avec des hétérogénéités spatiales des paramètres démographiques. Un tel modèle peut être défini par des tailles de dèmes N_i et des taux de migration entre paires de dèmes m_{ij} tous différents. Cependant, le nombre de paramètres explose rapidement avec le nombre de dèmes (par ex. 5+20 = 25 paramètres pour 5 dèmes) et il est statistiquement peu réaliste de vouloir estimer correctement chaque paramètre à partir d'un petit jeu de données. Comme expliqué régulièrement dans ce document, nous préférerons chercher des "caractéristiques communes" biologiques de dispersion et de densité implémenté dans des modèles simple, donc souvent homogènes dans l'espace, en testant leur robustesse par rapport à cette homogénéité. L'hétérogénéité spatiale des paramètres démographiques peut ensuite être ajouté à ces modèles, pour petit à petit prendre en compte l'influence d'un habitat hétérogène, par exemple en lien avec des caractéristiques paysagères. Nous expliciterons cela plus en détail à différentes reprises dans ce document.

1.3.2 Dispersion en population naturelles

Une caractéristique majeure du modèle en îles est donc que les immigrants peuvent provenir, de façon équiprobable, de n'importe laquelle des sous-populations. Or, dans la réalité, la dispersion est le plus souvent localisée dans l'espace et se fait donc préférentiellement entre deux sous-populations géographiquement proches. Puisque l'on considère ici la dispersion des gènes et non les mouvements nets d'individus, cela revient à dire qu'il y a une plus forte probabilité pour que des individus se reproduisent avec d'autres individus nés à proximité qu'avec des individus nés à plus grande distance.

Les jeux de données sur les distances de dispersion sont relativement rares, sans doute parce que la dispersion est un facteur difficile à estimer de manière rigoureuse. Endler (1977) a fait une revue extensive de la littérature de cette époque et a montré que généralement la dispersion est très limitée dans l'espace. Portnoy & Willson

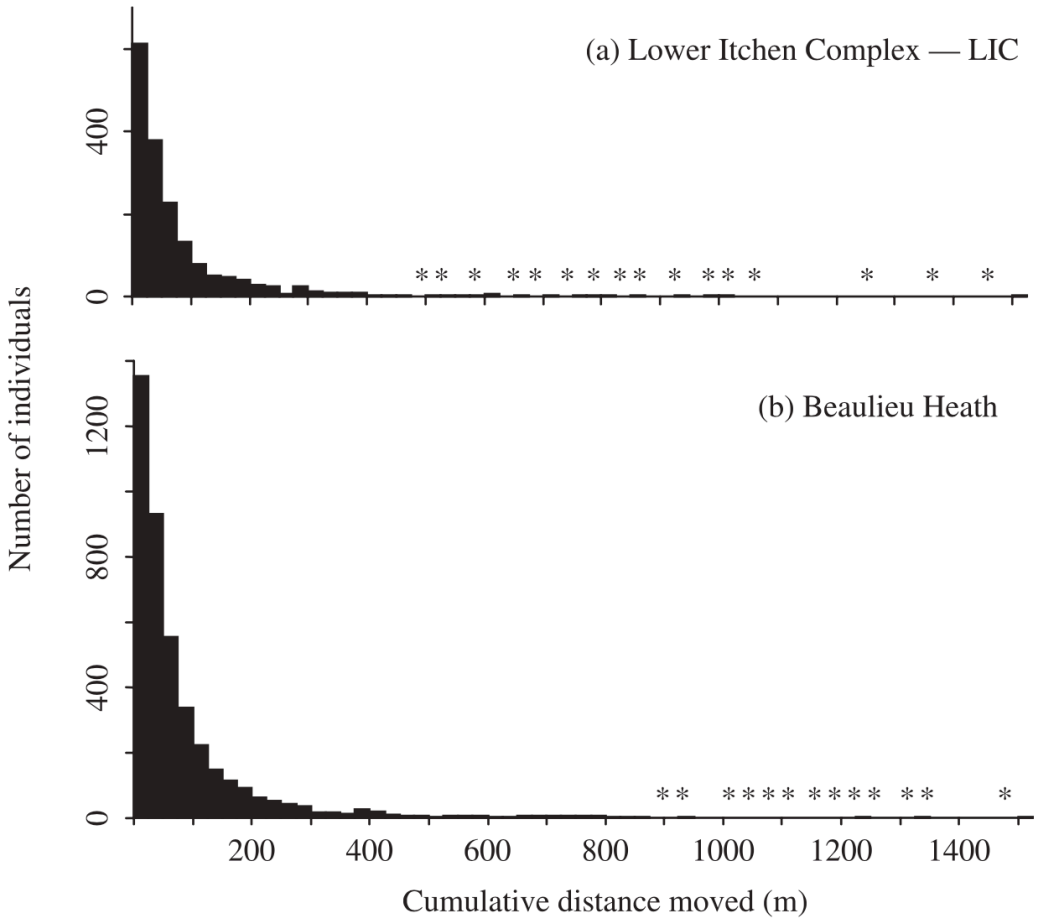


FIGURE 1.10 – Fréquences des déplacements cumulés au cours de la vie des adultes de demoiselles *Coenagrion mercuriale*. Regroupées par intervalles de distance de 25 m pour deux sites étudiés en capture-marquage-recapture (a) le Lower Itchen Complex (LIC) et (b) Beaulieu Heath, à la même échelle. En ordonnées, n est nombre d'individus recapturés ; * souligne qu'il y a eu des individus recapturés à cet intervalle distance mais que l'échelle ne permet pas de bien les distinguer, et illustre donc des déplacements peu fréquents (n = 1 ou 2). Figure issue de Watts *et al.* (2006).

(1993) et Bullock *et al.* (2017) se sont plus récemment intéressé à la forme de le dispersion des graines dans de nombreux jeux de données, quelques autres études ont aussi démontré des distances de dispersion restreintes chez les plantes (Fenster *et al.*, 2003; Crawford, 1984; Vekemans & Hardy, 2004) et chez les animaux (Rousset, 1997, 2000; Spong & Creel, 2001; Sumner *et al.*, 2001).

Les distributions de dispersion peuvent être caractérisées par leurs différents moments. Un moment $\mathcal{M}_k$ non centré d'ordre k est défini par $\mathcal{M}_k \equiv \mathrm{E}\left[X^k\right] = \sum_x x^k \Pr(X = x)$, un moment centré est défini par $\mathcal{M}'_k \equiv \mathrm{E}\left[X - \mathrm{E}(X)\right]^k = \sum_x (x - \bar{x})^k \Pr(X = x)$. Parmi les moments communément utilisés, la moyenne est le moment non centré d'ordre 1 et la variance est le moment centré d'ordre 2, $V(X) = \mathrm{E}\left[X - \mathrm{E}(X)\right]^2 = \mathcal{M}_2 - \mathcal{M}_1^2$. La kurtosis, définie par $\mathcal{M}'_4/\mathcal{M}'_3 - 3$, donne l'importance de la dispersion à courte et longue distance par rapport aux dispersions intermédiaires (le -3 est une convention pour que la kurtosis d'une loi normale soit nulle). Un autre moment qui apparaît souvent dans les modèles génétiques (par exemple lors de l'étude des clines, Barton & Gale, 1993) est σ^2, le moment d'ordre 2 [2] de la distance axiale de dispersion, ou encore la moyenne des carrés des distances [3] de dispersion parents-descendants C'est un moment de la distribution de dispersion qui nous intéressera spécialement par la suite.

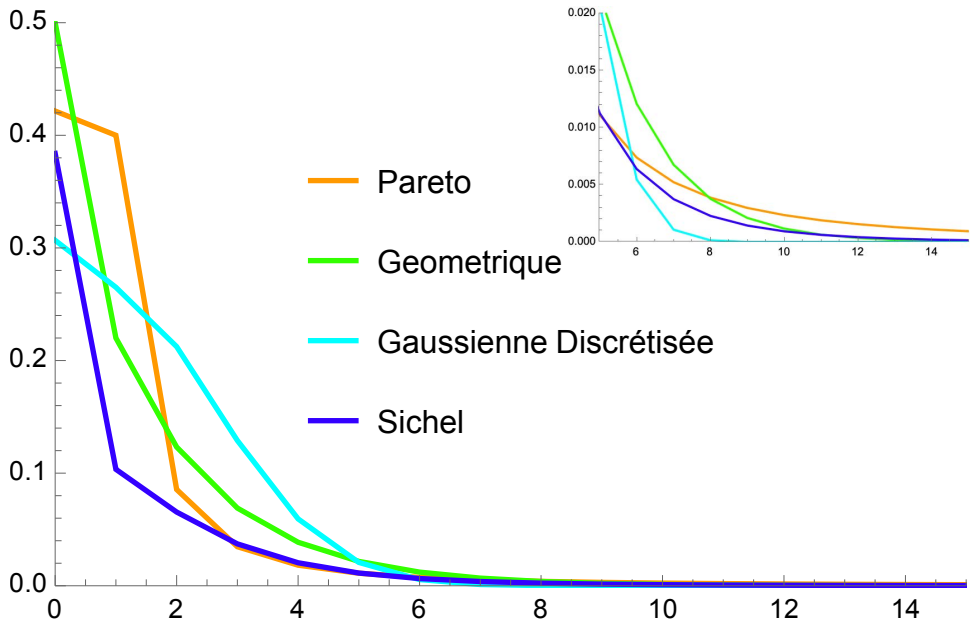


FIGURE 1.11 – Exemples de distributions de Pareto, Géométrique, Gaussienne discrétisée et Sichel avec un taux de dispersion assez fort et un $\sigma^2 = 4$. Le cadre en haut à droite représente un agrandissement des ces distributions sur les plus grandes distances.

Une des distributions les plus utilisées est la loi normale du fait de la convexité de la courbe à petites distances et qu'elle est "pratique" à considérer du point de vue mathématique. Mais les distributions de dispersion en population naturelles sont

2. Il s'agit du moment non centré d'ordre deux si la dispersion est mesurée en valeur absolue ; cependant puisque nous nous intéressons a la distance axiale (voir note suivante) la moyenne de la distribution de dispersion est nulle et les moments centrés sont alors identiques au moment non centrés.

3. nous considérerons ici et dans le reste du document que les distances correspondent à de distances vectorielles (les coordonnées (x, y) d'une entité par rapport à une autre) et non les distances euclidiennes ($r \equiv \sqrt{x^2 + y^2}$), plus classiquement utilisées. x et y sont appelées distances axiales et peuvent être négatives, contrairement à la distance euclidienne.

souvent leptokurtiques, c'est à dire qu'elles ont un excès de dispersion à faible et longue distance par rapport à la dispersion à des distances intermédiaires (Bateman, 1950; Clark *et al.*, 1999; Nathan *et al.*, 2012, revue dans Endler, 1977; Portnoy & Willson, 1993; Kot *et al.*, 1996; Bullock *et al.*, 2017). Un bel exemple de distribution empirique leptokurtique issu de l'étude de Watts *et al.*, 2006 est présenté en Figure 1.10. On dit aussi qu'elles ont une longue queue, ou une forte kurtosis. Le problème de la loi normale est qu'elle ne prend pas en compte cette caractéristique majeure des distributions de dispersion. Une autre caractéristique des distributions de dispersion est qu'elles doivent avoir une forte kurtosis tout en ayant un taux d'émigration global assez fort (peu d'individus se reproduisent exactement où leurs parents se sont reproduits mais beaucoup juste à coté...). Cette caractéristique fait que beaucoup de distributions communément utilisées, telles que les distributions exponentielles ou géométriques, ne sont pas non plus complètement pertinentes pour la modélisation des processus de dispersion (nous utiliserons tout de même la distribution géométrique pour certains développements). Certaines familles de distributions permettent de combiner des taux de migration forts et une forte kurtosis, parmi celles-ci on peut citer les distributions discrètes de la forme $f_k = f_{-k} = M/k^n$, où f_k est la probabilité de migrer à une distance k. Pour ces distributions, M contrôle approximativement le taux de migration global et n la kurtosis. Elles correspondent à des distributions discrètes de Pareto (ou Zeta) tronquées (voir, par ex. Patil & Joshi, 1968). Ce sont ces distributions que l'on utilisera pour la simulation dans la section 2.1.

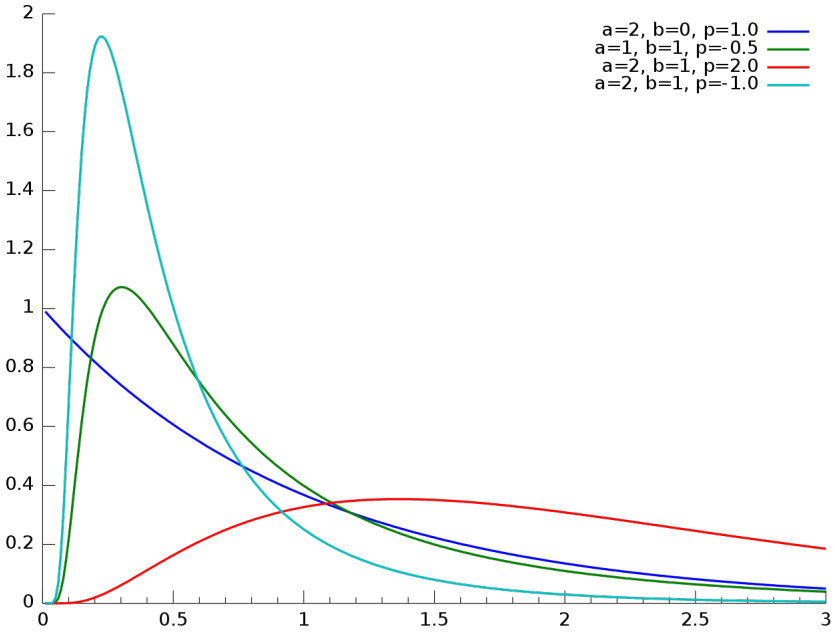


FIGURE 1.12 – Exemples de formes que peut prendre la distribution Sichel en fonction des valeurs de ses 3 paramètres a, b, p (réduits à 2 dans le texte, je ne sais plus comment...).

En pratique, les distributions de dispersion sont extrêmement diverses et il parait donc préférable de ne pas se focaliser sur une famille de distributions, telle que la loi normale, mais de considérer des modèles généraux applicables à n'importe quels types de distributions, ou des distributions très souples avec plusieurs paramètres permettant de jouer "indépendamment" sur les différents moments. Un exemple est la distribution Gamma réciproque de Poisson (dite distribution de Sichel, Chesson & Lee, 2005). Cette distribution est d'apparence gaussienne à courte distance, mais a une grande queue de distribution (en puissance), et peut donc présenter une kurtosis

élevée avec un taux de dispersion fort, typique de la dispersion chez de nombreuses espèces. Ses deux paramètres $\gamma < 0$ et κ déterminent la puissance $\gamma - 1$ de la queue et le second moment $\sigma^2 = -\kappa/[2(1+\gamma)]$, et permettent ainsi de modéliser une grande diversité de distributions de dispersion (voir figure 1.11). Nous l'utiliserons en section 3.2.2.

1.3.3 Isolement par la distance sur un réseau

Le modèle en isolement par la distance ("isolation by distance" IBD, Wright 1943) est un modèle permettant de décrire la dispersion des populations naturelles de manière plus pertinente que les précédents. L'idée est de modéliser la dispersion non plus uniquement par un taux d'émigration mais aussi par une distribution donnant la probabilité qu'un descendant disperse (et se reproduise) dans un dème en fonction de la distance entre ce dème et le dème de son(ses) parent(s). Ceci implique que les modèles d'isolement par la distance "contiennent" les modèles précédents. En effet, si la dispersion est uniforme sur l'ensemble de l'habitat, l'IBD est équivalent à un modèle en île. Si cette dispersion est limitée aux dèmes adjacents, l'IBD équivaut alors à un modèle en stepping-stone. Bien sûr, la distribution de dispersion peut avoir d'autres formes permettant de modéliser à la fois de nombreux évènements de dispersion à courte distance et de rares évènements de dispersion longue distance.

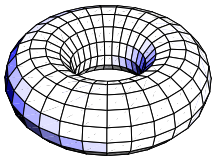


FIGURE 1.13 – Représentation graphique d'un tore

Les modèles considérés pour les analyses de l'isolement par la distance sont des modèles en réseaux, dans lesquels un ensemble de dèmes sont placées sur une grille régulière. Pour certaines analyses, notamment l'analyse sous IBD en terme de probabilités d'identité et F-statistiques, le réseau forme un cercle, en une dimension, ou un tore, en deux dimensions, pour éviter les effets de bord et avoir ainsi une surface parfaitement homogène (un tore est représenté sur la Figure 1.13). Sinon, la grille forme généralement une ligne ou un rectangle avec des effets de bords à spécifier, et tout autre forme d'habitat est envisageable.

N individus adultes composent chacune des sous-populations, dont la position sur le réseau est donnée par deux coordonnées, $r \equiv (x, y)$. L'unité de longueur est la distance inter-dèmes, c'est à dire la distance entre deux dèmes adjacentes. Pour des données réelles, on pourra exprimer cette distance inter-dème en fonction des distances géographiques mesurées entre les individus échantillonnés. Il suffira d'exprimer la densité dans la même unité de distance. Comme pour les autres modèles, au cours du cycle de vie (4), chaque adulte produit une infinité de gamètes qui subissent les effets de la mutation avec une probabilité μ. Chaque gamète migre ou non de façon indépendante vers un autre dème et la compétition ramène le nombre d'adulte dans chaque dème à N. Cette hypothèse de migration des gamètes sera relâchée dans certaines analyses, notamment celle utilisant le simulateur `GSpace` qui peut considérer que la dispersion se fait par les individus (juvéniles) et non les gamètes. Un modèle de dispersion individuelle est plus pertinent pour les espèces

animales car il prend en compte la corrélation plus forte entre histoires généalogiques à différents locus du fait de cette migration jointe des deux haplotypes des juvéniles diploïdes dispersant. Au contraire, les approches sans recombinaison, simulant souvent des histoires généalogiques différentes pour chaque locus indépendant, comme dans `IBDsim` ne considère pas du tout cette dépendance due à la migration d'haplotypes. Nous verrons en section 4.2.4 que cela peut avoir des conséquences fortes sur la précision des estimations, jusqu'ici peu prises en compte en génétique des populations. Pour les végétaux, il faudrait considérer une dispersion gamétique par le pollen, et une dispersion "individuelle" par les graines. Ce point n'est pas évoqué dans ce document et le lecteur pourra se référer au chapitre 8 de Rousset (2004) pour plus de détails sur la modélisation de la dispersion chez les plantes.

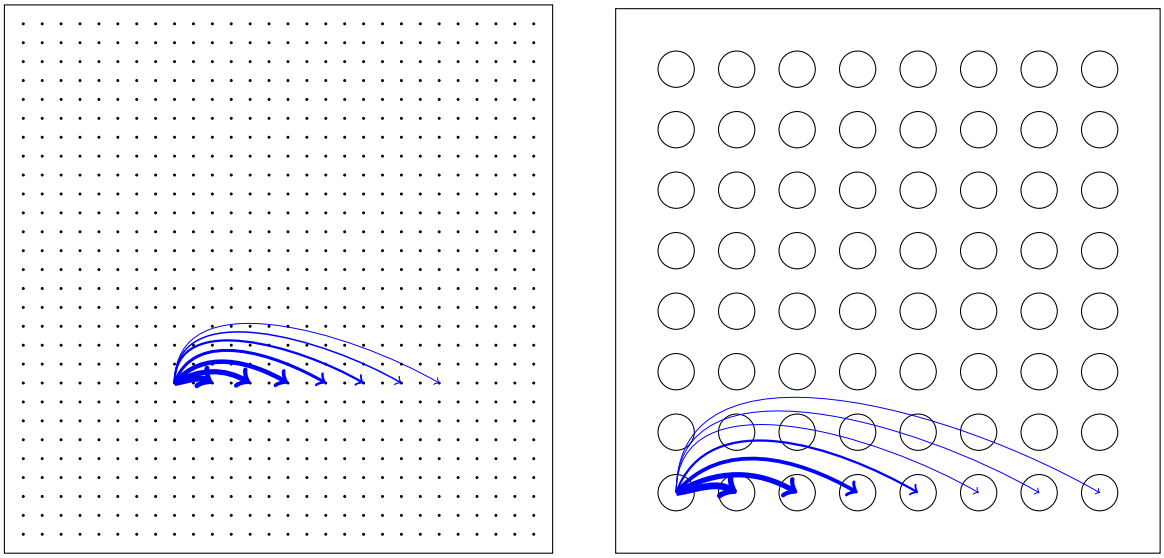


FIGURE 1.14 – Représentation graphique de deux modèle d'IBD individuels en habitat continu (a gauche) et en dèmes (a droite), sur un réseau de 28x28 et 8x8, respectivement. La dispersion est représentée à partir d'un unique point (ou dème) de l'habitat par les flèches bleues dont l'épaisseur indique l'intensité (taux) de la dispersion à cette distance. Les cercles représentent les dèmes, et les points les individus. Le cadre représente l'habitat.

Dans ce modèle, chaque dème se comporte comme une population panmictique dans lesquelles tous les individus sont équivalents. Un tel modèle est bien adapté à l'étude d'espèces vivant dans des habitats en petits patches pour lesquelles la dispersion est "illimitée" au sein d'un patche (et serait donc "panmictique") mais limitée par la distance entre les patches. On peut par exemple penser à des amphibiens se reproduisant dans des mares, ou des animaux commensaux (par ex. la souris domestique) vivant en petits groupes dans des maison ou des petits villages. Pour des espèces vivant dans des habitats continus et pour lesquelles la dispersion est très localisée par rapport à la taille de ces habitats, un modèle plus réaliste ne considérerait pas forcément de structure en dèmes et les individus pourraient se placer n'importe où sur une surface continue définissant l'habitat. La position des individus de la population varierait entre générations, entraînant une variation des densités locales entre générations. De tels modèles de populations continues ont été développés (Wright, 1943, 1946; Malécot, 1967; Sawyer, 1977) mais ne suivent pas un ensemble d'hypothèses biologiques et sont en ce sens non réalistes et difficilement utilisables (Maruyama, 1972; Felsenstein, 1975). Le fait qu'il n'y ait aucune régulation explicite des densités locales conduit, par exemple, à une certaine forme d'agrégation des individus dans l'espace. Quelques tentatives ont été faites pour s'affranchir de

ce problème (Sawyer & Felsenstein, 1981; Wilkins & Wakeley, 2002; Wilkins, 2004) mais elles sont basées sur des hypothèses peu réalistes de régulation de la densité et/ou entrainent des difficultés mathématiques (Nagylaki, 1986; Barton *et al.*, 2010).

Dans ce contexte, Robledo-Arnuncio & Rousset (2010) ont formulé des récurrences exactes pour les probabilités d'identité dans des populations continues en IBD, avec divers niveaux de fluctuations démographiques, permettant de définir une dispersion efficace (σ_e^2) et une densité efficace (D_e). Ces travaux permettent de généraliser les résultats des modèles en réseau que nous verrons dans le chapitre 2. Ainsi, la méthode d'inférence de la régression que nous détaillerons reste valide en population continue mais les estimations de densité et dispersion, bien que robustes dans la majorité des cas, pourraient être fortement biaisées dans des situations extrêmes où les individus auraient une forte tendance à se regrouper spatialement.

Plus récemment, Nick Barton, Amandine Véber et leurs collègues on défini un nouveau modèle, appelé le processus spatial Λ-Fleming-Viot, pour modéliser de l'isolement par la distance dans un habitat continu (Barton *et al.*, 2010, 2013; Véber & Wakolbinger, 2015). Les auteurs décrivent la dynamique ("avant") d'une population évoluant dans un habitat continu non borné (c.a.d sans effets de bords, donc sur un cercle ou un torre), mais aussi le processus de "coalescence" (arrière) correspondant qui caractérise la généalogie des individus échantillonnés à partir duquel Guindon *et al.* (2016) et Ringbauer *et al.* (2017) ont développé des méthodes d'inférences de la dispersion. Du fait de la position variable et non prédictible des parents à la génération précédente, la validité de leur approche par coalescence ne nous semblait pas claire. Quelques comparaison d'estimations sur données réelles avec la méthode de Ringbauer *et al.* (2017) et la méthode de la régression que nous décrirons ci-dessous suggèrent que ce modèle estime bien les mêmes paramètres de densité et de dispersion (Smith *et al.*, 2022). Ces résultats étant très récents, nous n'avons pas considérer ce modèle Λ-Fleming-Viot dans ce document mais il mérite sûrement d'être plus largement exploré.

A part peut-être le modèle Λ-Fleming-Viot, et vu les conclusions de Robledo-Arnuncio & Rousset (2010), le meilleur modèle de population continue que nous avons à présent est donc le modèle en réseau avec un individu ou un couple par nœud. Ce modèle peut être considéré comme une approximation de la population continue avec une régulation locale de la densité (par ex. lorsque la compétition locale est forte, Malécot, 1975; Rousset, 2000). Nous détaillerons l'analyse mathématique de ces modèles d'IBD (en réseau) dans la section suivante. On peut noter dès à présent que ces modèles d'IBD (en réseau) semblent plus réalistes, notamment du point de vue de la dispersion, et que leurs caractéristiques, telles que de faibles tailles de dème et une dispersion relativement forte mais limitée dans l'espace, semblent favorable à l'inférence robuste des paramètres de densité et de dispersion locaux et actuels.

1.4 Inférence par moment vs. vraisemblance vs. par simulation

On peut distinguer deux approches principales pour l'inférence statistique (Cox & Hinkley, 1974) : les méthodes des moments et les méthodes fondées sur la vraisemblance. Les méthodes des moments, dont nous avons évoqué des exemples à travers l'estimation de Nm à partir du F_{ST} dans un modèle en îles, consistent à estimer certains paramètres d'un modèle en égalisant certains moments théo-

riques (qui dépendent de ces paramètres) avec leurs contreparties empiriques. Ce qui revient à "approcher" une espérance mathématique par un estimateur empirique (comme on l'a vu avec l'estimateur du nombre de migrant dans le modèle en îles, $\hat{Nm} = (1/\hat{F_{\text{ST}}} - 1)/4$, défini en terme d'estimateur du F_{ST} calculé sur l'échentillon). En génétique des populations, l'approche des moment la plus courante est celle basée sur les F-statistiques, et donc sur l'analyse des corrélations des fréquences alléliques entre différents niveau hiérarchiques définis au sein de la population étudiée. Elle suppose que l'on sait exprimer ces F-statistiques en fonction des paramètres du modèle considère et que l'on puisse en calculer un estimateur sur un jeu de données génétiques. Elles n'utilisent ainsi qu'une partie des données résumées dans une seule statistique pour l'inférence de chaque paramètre d'intérêt.

Les méthodes par vraisemblance nécessitent un calcul explicite de la probabilité d'un échantillon génétique en fonction des paramètres du modèle ou des algorithmes de Monte Carlo permettant d'estimer cette vraisemblance par simulation si le calcul n'est pas possible. Ce sont des méthodes d'inférence optimales du point de vue statistique car elles utilisent toute l'information du jeu de données et bénéficie d'une cadre statistique rodé que ce soit en maximum de vraisemblance ou en inférence Bayesienne. La démarche consiste simplement à considérer la vraisemblance $\mathcal{L}(\mathcal{P}; D) \equiv \Pr(D; \mathcal{P})$ d'un modèle défini par les valeurs prises sur l'ensemble de ses paramètres $\mathcal{P}$, sachant les données $\mathcal{D}$[4]. Il faut ensuite trouver le jeu de paramètres $\mathcal{P}_{\text{MLE}}$ pour lequel la vraisemblance est maximum. $\mathcal{P}_{\text{MLE}}$ est alors l'estimateur par maximum de vraisemblance.

De nombreuses méthodes ont été développées pour calculer ou estimer la vraisemblance de différents types de données génétiques selon différents modèles de structuration de populations. Il est important de noter qu'il existe de nombreuses méthodes d'estimation de paramètres démographiques et historiques par vraisemblance fondées sur les fréquences alléliques d'un échantillon n'utilisant pas la coalescence. La plupart sont basées sur des approximations de la distribution des fréquences alléliques dans un modèle de Wright-Fisher par des distributions statistiques instrumentales, c.a.d. des lois normales ou beta approchant les distributions sous le vrai processus de WF, et donc pas directement reliées aux paramètres du modèle démo-génétique (tout au moins dans le sens explicité dans ce document). De plus, ces approximations, souvent dérivées dans le cadre de la théorie de la diffusion, permettent principalement de considérer des modèles de pure dérive, et sont moins adaptés à la modélisation de la migration et aux populations structurées. Nous nous focaliserons donc sur l'estimation par vraisemblance des paramètres d'un modèle démo-génétique à partir d'un échantillon de gènes par des méthodes fondées sur la théorie de la coalescence. Le lecteur trouvera une synthèse détaillée de ces approches alternatives et des références vers les méthodes d'inférence les implémentant dans Tataru *et al.* (2016).

D'un point de vue purement statistique, un des aspects les plus intéressants de la théorie de la coalescence est de permettre une analyse par vraisemblance du polymorphisme de marqueurs neutres et donc d'utiliser l'information complète des données génétiques. Mis à part quelques cas spécifiques comme pour une population panmictique avec un modèle mutationnel à nombre infini d'allèles ("Ewens sampling

4. Par extension (ou abus de langage), on parlera aussi bien de la vraisemblance d'un modèle sachant les données que de la vraisemblance des données sous un modèle, bien que la dernière formulation corresponde moins à l'esprit du maximum de vraisemblance. En effet, puisque l'on cherche les valeurs des paramètres qui maximisent la vraisemblance, ce sont les paramètres qui varient et non les données.

formula", Ewens, 2004), il n'existe pas de formules explicites de la vraisemblance pour les modèles démo-génétiques de populations structurées.

Ces méthodes ne considèrent donc pas de formules explicites pour la vraisemblance mais utilisent la simulation d'arbres de coalescence pour obtenir une estimation de la vraisemblance d'un échantillon. Dans le contexte de la coalescence, on peut écrire la vraisemblance des paramètres d'un modèle en fonction des données comme

$$\mathcal{L}(\mathcal{P}; D) = \int_G \Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P}), \tag{1.22}$$

où l'intégrale représente une somme sur toutes les généalogies compatibles avec l'échantillon. Un estimateur non biaisé de la vraisemblance est alors

$$\mathcal{L}(\mathcal{P}; D) = E\left[\Pr(D|G; \mathcal{P})\right], \tag{1.23}$$

où E correspond à l'espérance d'une chaîne de Markov de distribution stationnaire $\Pr(G; \mathcal{P})$. En d'autre mots, l'estimation de $\mathcal{L}(\mathcal{P}; D)$ par l'équation (1.23) se fera en calculant la moyenne de $\Pr(D|G; \mathcal{P})$ sur un grand nombre de généalogie simulées selon la distribution stationnaire $\Pr(G; \mathcal{P})$. Cependant, simuler directement selon la distribution $\Pr(G; \mathcal{P})$ peut s'avérer difficile techniquement ou peu efficace (c.a.d. beaucoup de généalogies échantillonnées auront une probabilité $\Pr(D|G; \mathcal{P})$ très faible).

L'équation (1.22) peut alors être mise sous la forme

$$\mathcal{L}(\mathcal{P}; D) = \int_G \frac{\Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P})}{f(G)} f(G), \tag{1.24}$$

où $f(G)$ est ce que l'on appellera la fonction d'échantillonnage pondéré, ou d'importance ("Importance Sampling", IS). Cette nouvelle distribution f impliquera que l'on simulera les généalogies selon une distribution stationnaire différente de $\Pr(G; \mathcal{P})$. Dans ce cas, un estimateur non biaisé de la vraisemblance est

$$\mathcal{L}(\mathcal{P}; D) = E\left[\frac{\Pr(D|G; \mathcal{P}) \Pr(G; \mathcal{P})}{f(G)}\right], \tag{1.25}$$

où E correspond à l'espérance d'une chaîne de Markov de distribution stationnaire $f(G)$. L'utilisation de fonctions d'échantillonnage d'importance permet d'explorer l'espace des possibles (ici les différentes généalogies) par la simulation selon f de façon plus efficace que directement selon la distribution stationnaire $\Pr(G; \mathcal{P})$ (c.a.d. l'exploration de zones de forte probabilité sera favorisée au dépens des zones de faible probabilité). Si l'on explore un grand nombre de généalogies $\{G_1, G_2, .., G_n\}$ selon la distribution f, on a alors

$$\hat{\mathcal{L}}(\mathcal{P}; D) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\Pr(D|G_i; \mathcal{P}) \Pr(G_i; \mathcal{P})}{f(G_i)}. \tag{1.26}$$

Il suffit alors pour estimer $\hat{\mathcal{L}}(\mathcal{P}; D)$ de déterminer la probabilité $\Pr(D|G_i; \mathcal{P}) \Pr(G_i; \mathcal{P})$ d'un échantillon de gène D pour chaque généalogie G_i explorée.

Selon les implémentations, (i) l'espace des paramètres peut être exploré en même temps que celui des généalogies en suivant un algorithme de Monte Carlo par Chaîne de Markov (MCMC) sur l'espace joint des paramètres et des généalogies, généralement implémenté dans une approche d'inférence Bayesienne, sinon (ii) les généalogies sont explorées dans un premier temps par simulation pour estimer la vraisemblance

en un certains nombre de points de l'espace des paramètres ; et l'espace des paramètres est ensuite exploré en interpolant une surface de vraisemblance à partir de ces points (ou en utilisant des algorithmes de type "expectation-maximisation" si l'on ne veut pas la surface de vraisemblance mais juste le maximum) pour trouver l'estimateur par maximum de vraisemblance $\mathcal{P}_{\text{MLE}}$ et construire des intervalles de confiance.

L'inférence basée sur la vraisemblance est donc souvent limitée par la difficulté de calculer ou d'estimer la vraisemblance sous les modèles d'intérêt. Ceci a motivé, notamment en génétique des populations, le développement de nouvelles méthodes d'inférence dont le but est d'approcher la vraisemblance, en utilisant des simulations et non une formulation mathématique du modèle (Marjoram & Tavaré, 2006). C'est ce que l'on appellera l'inférence par simulation, qui peut se décrire intuitivement sous sa forme originelle simple (mais statistiquement peu robuste) de l'algorithme de rejet appliqué en génétique des populations par Tavaré *et al.* (1997) et Pritchard *et al.* (1999).

L'algorithme de rejet consiste tout simplement à (i) simuler un grand nombre de jeux de données, avec les mêmes caractéristiques que le jeu de données réelles ("observé") que l'on veut analyser (c.a.d le même nombre d'individus et de locus, et le même type de marqueurs) dans le cadre d'un modèle donnée. Les paramètres des jeux de données simulés sont échantillonnés à partir d'une distribution a priori définie par l'utilisateur ; (ii) les données simulées sont ensuite réduites à un ensemble de statistiques, dites "statistiques résumantes", et (iii) les paramètres de chaque simulation sont acceptés ou rejetés sur la base d'une distance maximale entre les statistiques résumantes simulées et observées ; enfin (iv) le sous-échantillon des valeurs acceptées de paramètres permet de construire la distribution marginale à posteriori pour chaque paramètre. Le principe de l'algorithme de rejet est donc de générer un grand nombre de jeux de données simulés, les réduire à un ensemble de statistiques résumantes calculées sur chaque simulation, et puis de faire l'inférence à partir des valeurs de paramètres des simulations retenues comme étant les plus proches du jeu de données observé, par comparaison entre statistiques résumantes simulées et observées.

Ce principe s'applique assez naturellement dans un cadre d'inférence Bayesienne, et de nombreuses versions améliorées de cet algorithme de rejet ont été développées sous le nom d' "Aproximate Bayesian Computations" (ABC, Beaumont *et al.*, 2002) que nous détaillerons en section 4.1. Pour nos travaux, nous avons choisi d'utiliser une approche alternative d'inférence par simulation, la vraisemblance résumée ("summary likelihood" SL, Rousset *et al.*, 2017), que nous présenterons et testerons dans le chapitre 4. Par ailleurs, la coalescence, de part son efficacité et sa flexibilité à simuler une grande variété de modèles démo-génétiques, a été naturellement utilisée pour générer les jeux de données simulés dans une majorité des développements ABC de génétique des populations.

Enfin, on peut noter que l'inférence par simulation repose donc sur (1) l'analyse incomplète des données, car résumées dans un (petit) ensemble de statistiques résumantes décrivant les données, (ii) des algorithmes d'apprentissage automatique supervisés permettant de trouver les paramètres correspondant au mieux aux données observées (réduites), sur la base de données d'apprentissage simulées. Ce lien avec les méthodes d'apprentissage automatique et l'explosion récente des méthodes dites d'intelligence artificielle (IA), ont naturellement conduit au développement récents d'approche d'inférence par simulations basées sur ce type de méthodes, plus ou

moins intégré dans un cadre ABC (voir par ex. Sanchez *et al.*, 2021). Ces nouvelles approches ont notamment pour but d'améliorer l'apprentissage (et donc de réduire les temps de simulation), et d'essayer de se passer des statistiques résumantes ou de pouvoir en considérer un très grand nombre. En effet, l'apprentissage automatique, notamment profond, semble particulièrement adapté à l'extraction d'informations pertinentes à partir de données génétiques. Ainsi, Sanchez *et al.* (2021) montrent que la combinaison de l'apprentissage profond par réseau de neurone et de l'ABC peut améliorer les performances des inférences en tirant parti des deux approches. Ce domaine d'inférence par simulation, notamment basé sur l'apprentissage automatique, est en pleine explosion, et une des idées serait de se passer des statistiques résumantes, et donc pourvoir utiliser toute l'information des données.

1.5 Robustesse des inférences

Une des spécificités de la biologie évolutive lorsque l'on cherche à étudier les processus évolutifs à partir de données issues de populations naturelles est que l'expérience menant aux données observées a été faite une fois, dans des conditions initiales et de déroulement de l'expérience inconnues, qu'il est impossible de la refaire, et qu'il n'existe pas de réplicat. Il existe donc une incertitude quasi-totale sur le déroulement du processus évolutif du passé infiniment lointain jusqu'au moment de l'observation des données, que l'on considérera ici comme le présent. Dans de nombreuses situations, cette incertitude sur l'histoire évolutive doit être prise en compte lors de l'analyse des ces données.

En génétique des populations, de même que dans d'autres domaines de la biologie évolutive, cette incertitude peut être prise en compte à l'aide de modèle stochastique de l'évolution, dont les approches par coalescence et les modèles de Wright-Fisher étendus que nous avons vu précédemment sont des exemples. Malgré une demande potentiellement forte des empiristes pour des modèles complexes, quelques principes de base, sur lesquels nous reviendrons régulièrement, suggèrent plutôt de se focaliser sur des modèles simples mais robustes quand ils s'agit de faire des inférences à partir de données issues de populations naturelles sur les processus évolutifs les ayant générées.

Une première raison purement statistique est que plus un modèle à de paramètres moins ces paramètres peuvent être estimés avec précision à partir d'un jeu de donnée de taille finie. Un corollaire est que plus le jeu de données est petit (moins il "contiendra d'information"), plus il faut considérer des modèles simples.

De plus, les données issues de population naturelles, mais aussi de populations très anthropisées de type agricole ou encore de populations Humaines, ont été générées dans des contextes beaucoup plus complexes que n'importe quel modèle même complexe utilisé pour l'inférence. Il existe donc de nombreux facteurs non-contrôlés pouvant influencer les estimations, en les biaisant, en augmentant leur variance ou encore en compliquant ou invalidant l'interprétation biologique des paramètres estimés.

Enfin, comme nous l'avons vu pour le n-coalescent, de nombreuses méthodes reposent sur les approximations de grandes tailles de populations et de petits taux d'évènements, et il n'est alors pas toujours possible d'estimer séparément tous les paramètres canoniques du modèle démo-génétique non-approché. Par exemple, nous avons vu que le coalescent était invariant à des transformations de N et μ à produit

$N\mu$ constant, il est donc uniquement possible d'inférer le produit $N\mu$[5], et pas N et μ séparément, sous ces hypothèses. Ce raisonnement peut être étendu à tous les paramètres biologiques apparaissant dans le modèle d'inférence uniquement sous forme de combinaison des paramètres canoniques du modèle démo-génétique non-approché. Il est alors important de bien distinguer les paramètres estimables ou non à partir des données selon les méthodes utilisées. Cette différence n'est pas forcément toujours claire pour les utilisateurs des méthodes d'inférence, il paraît crucial de paramétrer les modèles d'inférences en terme de paramètres estimables (c.a.d les paramètres canoniques du modèle approché) et non des paramètres canoniques du modèle démo-génétique non-approché. Cela sera abordé régulièrement dans les trois chapitres suivants.

Pendant longtemps, les jeu de données génétiques pour les espèces non modèles ont été de tailles très réduites, non seulement en terme de nombre d'individus (et c'est encore le cas), mais surtout parce que l'on avait accès à un très petit nombre de marqueurs génétiques. Les inférences de l'histoire adaptative des populations étaient de fait limité à des modèles simples (et sous lesquelles les estimations obtenues étaient en plus souvent très imprécises). Ceci n'a pas empêché l'émergence de nombreuses "croyances" basées sur le manque de prise en considération de la robustesse des ces inférences. Ainsi, comme évoqué précédemment, il a souvent été considéré que le F_{ST} estimé sur un échantillons issus de 2 "dèmes" échantillonnés dans une population comportant de nombreux autres dèmes renseigne sur la migration entre cette paire de pop. Sauf si les 2 dèmes sont complètement isolés du reste de la population, Ceci est complètement faux puisque c'est un estimateur de l'ensemble des immigrants arrivants dans ces populations à partir de toutes les autres populations. C'est même plus exactement un estimateur du nombre d'immigrants moyen arrivant dans chacun de ces dèmes sous l'hypothèse que ce taux est équivalent pour toutes les paires de dèmes, comme nous l'avons vu en section 1.3.1. C'est un exemple assez caricatural mais non moins commun, et il en existe bien d'autres tout aussi problématiques.

Un point important à prendre en compte est donc de développer et utiliser des méthodes robustes vis à vis des écarts les plus probables au modèle. Nous verrons dans les deux chapitres suivants de nombreux exemples de tests de robustesse de l'inférence de la dispersion et de la densité en isolement par la distance. Quelques exemples classiques de facteurs importants à tester sont l'influence des processus anciens vs. récents, des processus locaux vs. avoisinants ou lointains, des processus mutationnels des marqueurs utilisés, ou encore des processus d'adaptation. Un autre point important de robustesse, illustré dans l'exemple du F_{ST} ci-dessus et sur lequel nous reviendrons souvent, est le nombre de dèmes échantillonnés vs. nombre de dèmes du système biologique, et plus largement la définition de (sous-)populations dans les modèles de populations structurées.

Ainsi, malgré l'arrivée des NGS permettant maintenant des estimations extrêmement précises sous des modèles avec de nombreux paramètres, ces modèles restant une extrême simplification de la réalité biologique, il parait toujours important de privilégier la robustesse à la précision pour éviter d'aboutir à des conclusions erronées. C'est ce type d'approche que je tenterai de décrire dans la suite de ce document,

5. $N\mu$ est souvent (abusivement) appelé paramètre de taille de population mis à l'échelle de la mutation, ce qui est en effet plus parlant quand on s'intéresse à l'inférence démographique et c'est effectivement comme cela qu'il est généralement interprété, c.a.d on divise $N\mu$ par μ pour avoir accès à N. Cependant, c'est un taux de mutation mis à l'échelle de la taille de la populations, comme tous les autres paramètres du coalescent.

toujours partir de modèles simples et de méthodes statistiques adaptées et performantes, que l'on teste profondément afin de bien comprendre leur comportement, notamment la robustesse des estimations (ce qui peut durer très longtemps...). Puis les complexifier petits à petits pour répondre à des questions biologiques différentes ou de manière plus précise. J'espère que le lecteur comprendra bien tout au long de sa lecture, qu'in-fine, ce sera toujours la robustesse qui déterminera ce que l'on pourra inférer.

Enfin, par principe, toute inférence statistique devrait partir de la construction d'un ensemble de modèles paraissant a priori convenir pour les données et le modèle biologique considéré, choisir le ou les meilleurs modèles par une procédure de choix de modèle, puis faire sous ce(s) modèle(s) l'inférence des paramètres d'intérêt et finir par un test de qualité de l'ajustement de ce(s) modèle(s) aux données ("goodness of fit" GOF). Par manque de temps et de place, mais aussi parce que ce sont des questions que je n'ai pas directement traitées dans mes travaux, je n'aborderai pas dans ce document ces aspects pourtant primordiaux. Je laisserai le lecteur trouver la lecture statistique la plus adaptée à son niveau, dans l'immense littérature scientifique statistique consacrée à ces questions (bien que Cox, 2006 semble être une référence de choix :).

Chapitre 2

Inférences par F-statistiques sous IBD

Nous avons vu dans le chapitre d'introduction comment l'ajustement de modèles démo-génétiques peut permettre, au moins dans certains cas, l'estimation de paramètres démographiques à partir de données génétiques par la méthode des moments (cf. éq.1.19 par exemple) et que le modèle d'IBD permet une meilleur prise en compte de la dispersion et de l'organisation spatiale des population naturelles. Dans ce chapitre, nous nous focaliserons sur la "méthode de la régression" de Rousset (1997, 2000), une méthode d'inférence de la dispersion et de la densité sous IBD en réseau fondées sur l'augmentation de la différenciation génétique entre individus ou entre dèmes en fonction de la distance géographique. C'est une méthode des moments reposant sur le calcul des probabilités d'identité et d'une F-statistique sous isolement par la distance, introduits ci-dessous.

2.1 Structuration génétique en isolement par la distance et méthode de la régression

Si les premières analyses des modèles d'isolement par la distance ont été faite par Wright (1943, 1946) et Malécot (1948), toutes les analyses rigoureuses découlent du modèle en réseau formulé par Malécot (1950). Dans ces modèles, toutes les distributions de dispersion peuvent être considérées, à condition d'avoir des moments finis jusqu'à l'ordre 3, impliquant une dispersion locale. On suppose également que la densité (ou les tailles de dèmes) et la migration sont homogènes dans l'espace, c'est à dire que la densité et la distribution de dispersion sont identiques en tout point du réseau. Un point intéressant est que les modèles IBD en dèmes et en habitat continu (dans lesquels il n'y a pas de dèmes mais juste des individus ou des couples à chaque nœud du réseau) peuvent être traité de manière similaire. Nous considérerons donc par souci de simplification que l'on parlera abusivement parfois de dème pouvant avoir une taille $N = 1$ ou deux quand on ne souhaitera pas différencier les modèles en dèmes ou individuel en habitat continu. Généralement la notation D sera utilisée pour la densité dans le modèle individuel et N pour les tailles de dèmes dans l'autre modèle mais si l'unité de mesure considéré est la distance inter-individuelle ou la distance inter-dème, ce qui sera la cas dans tous ce document, les deux mesures peuvent être vues comme équivalentes puisque $D = \frac{N}{\text{unité}^2} = \frac{N}{1^2} = N$.

Comme pour les analyses en terme de probabilité d'identité des modèles de population panmictique ou du modèle en îles présentée en section 1.2, l'analyse ma-

thématique du modèle d'isolement par la distance permet l'expression des équations de récurrences des différentes probabilités d'identité par descendance Q_{w} au sein d'un dème (Q_{wd}) ou au sein d'un individu (Q_{wi}) et Q_{r} entre dèmes séparés par une distance géographique r en fonction des paramètres du modèle et ce système d'équation est résolu à l'équilibre en utilisant les transformées de Fourrier. L'outil principal dans l'analyse des modèles en réseau, notamment pour résoudre le système d'équation de récurrence des Q est l'analyse de Fourier, dont je ne donnerai pas les détails car ils sont assez complexes. Pour l'application de l'analyse de Fourier aux modèles d'isolement par la distance dont nous ne donnerons donc ici que les principaux résultats, le lecteur pourra se référer à Sawyer (1977) et Rousset (1997, 2004). Tous les résultats que l'on présentera dans cette partie sont valables pour des réseaux de taille infinie ou sans effet de bords, donc sur un cercle ou un tore. On a alors, tout d'abord pour un modèle d'IBD en une dimension,

$$\frac{Q_{\mathrm{r}}}{1-Q_{\mathrm{w}}} \approx \frac{e^{-\sqrt{2\mu}r/\sigma}}{4N\sigma\sqrt{2\mu}}, \tag{2.1}$$

où σ est la racine carré du moment d'ordre deux de la distribution de distance de dispersion parent-descendant, comme défini en section 1.3.2. L'équation (2.1) est une approximation pour des grandes distances géographiques (c.a.d. $r \to \infty$, mais en pratique, $r \gg sigma$). Pour $r = 0$, on a alors

$$\frac{Q_{\mathrm{w}}}{1-Q_{\mathrm{w}}} \approx \frac{1}{4N\sigma\sqrt{2\mu}} + \frac{A_1}{4N\sigma}, \tag{2.2}$$

où A_1 est une constante, relativement compliquée, déterminée uniquement par la distribution de dispersion (mais pas uniquement fonction de σ^2). Sawyer (1977) en donne la définition suivante

$$A_1 \equiv 2\sigma\left(\frac{1}{\pi}\int_0^{\pi}\frac{\psi^2(x)}{1-\psi^2(x)} - \frac{1}{\sigma^2x^2}dx - \frac{1}{\pi^2\sigma^2}\right), \tag{2.3}$$

où ψ est la fonction caractéristique des probabilités de dispersion m_r, $\psi(z) \equiv \sum_r m_r e^{\imath rz}$ (où $\imath$ désigne la partie imaginaire d'un nombre complexe). De telles fonctions caractéristiques sont couramment utilisées dans le cadre des analyses de Fourier (voir Rousset, 2004).

En deux dimensions, pour deux gènes à la distance euclidienne $r \equiv \sqrt{x^2+y^2}$, on a

$$\frac{Q_{\mathrm{r}}}{1-Q_{\mathrm{w}}} \approx \frac{K_0(\sqrt{2\mu}r/\sigma)}{4N\pi\sigma^2}, \tag{2.4}$$

où K_0 est la fonction modifiée de Bessel de second type et d'ordre zéro (voir par ex. Abramovitz & Stegun, 1972). L'équation (2.4) est aussi une approximation pour des r grands et une autre expression doit être considérée pour $r = 0$:

$$\frac{Q_{\mathrm{w}}}{1-Q_{\mathrm{w}}} \approx \frac{-\ln(\sqrt{2\mu}+2\pi A_2)}{4N\pi\sigma^2}, \tag{2.5}$$

où A_2 est de la même nature que A_1. Le lecteur pourra en trouver une expression dans Sawyer (1977, éq.3.4). Pour le modèle en dèmes, Rousset (1997) montre que la F-statistiques la mieux adapté pour décrire la différenciation en fonction de la distance et des paramètres du modèle est $F_{\mathrm{ST}}/(1-F_{\mathrm{ST}}) = \frac{Q_{\mathrm{wd}}-Q_{\mathrm{r}}}{1-Q_{\mathrm{wd}}}$. On comprendra tout de suite pourquoi ci-dessous. Pour le modèle individuel en habitat continu,

Rousset (2000) définit le paramètre $a_r = \frac{Q_{wi}-Q_r}{1-Q_{wi}}$, analogue à $F_{ST}/(1-F_{ST})$ mais entre paires d'individus et non entre paires de dèmes. Si l'on pose F_r correspondant à $F_{ST}/(1-F_{ST})$ ou a_r selon le modèle, on a alors en réarrangeant les équations ci-dessus

$$\begin{aligned} F_r \equiv \frac{Q_w - Q_r}{1-Q_w} &\approx \frac{1-e^{-\sqrt{2\mu}r/\sigma}}{4N\sigma\sqrt{2\mu}} + \frac{A_1}{4N\sigma} \\ &\approx^{r \text{ et } \mu \text{ petit}} \frac{r}{4N\sigma^2} + \frac{A_1}{4N\sigma} \approx^{N\to D} \frac{r}{4D\sigma^2} + \frac{A_1'}{4D\sigma} \end{aligned} \tag{2.6}$$

en une dimension, et en deux dimensions on a

$$\begin{aligned} F_r &\approx \frac{-\ln(\sqrt{2\mu}) - K_0(\sqrt{2\mu}r\sigma) + 2\pi A_2}{4N\pi\sigma^2} \\ &\approx^{r \text{ et } \mu \text{ petit}} \frac{\ln(r\sigma) - 0.116 + 2\pi A_2}{4N\pi\sigma^2} \approx^{N\to D} \frac{\ln(r)}{4\pi D\sigma^2} + \frac{\ln(\sigma) - 0.116 + 2\pi A_2'}{4\pi D\sigma^2} \end{aligned} \tag{2.7}$$

Les dernières expressions correspondent au passage des tailles de populations N à la densité d'individus sur le réseau D. En deux dimensions, peu importe l'unité de longueur utilisée, il suffit que la densité soit exprimée avec la même unité de longueur que σ, et $D\sigma^2$ est exprimé en nombre d'individus. En une dimension, l'unité de $D\sigma^2$ et un nombre d'individu fois l'inverse d'une distance (individus$\times m^{-1}$ par exemple). Une unité simple, que l'on utilisera ensuite, est la maille du réseau (la distance entre deux dèmes adjacents). C'est l'unité qui est utilisée pour σ lorsque l'on considère des tailles de populations N (modèle avec structure démique) et non les densités. C'est à ce niveau que se fait le lien entre les modèles à structuration démique dans lesquelles les individus sont regroupés en dèmes et les modèles en populations continues dans lesquels les individus sont répartis de façon homogène sur toute la surface définissant l'habitat. On notera aussi que $D\sigma^2 = Nm$ dans un modèle stepping-stone et prend une valeur infinie dans un modèle en îles. Les modèles d'isolement par distance comprennent donc le modèle en îles et le modèle stepping-stone comme sous-modèles "extrême" du point de vus de la dispersion.

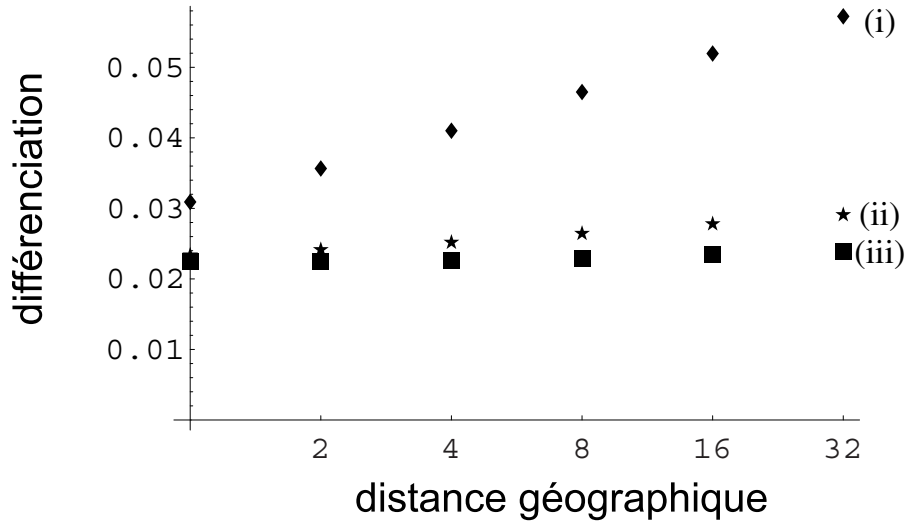


FIGURE 2.1 – Différenciation en fonction du logarithme de la distance en isolement par la distance en deux dimensions avec : (i) une structuration forte ; (ii) avec une structuration moins forte $\sigma_{ii}^2 > \sigma_i^2$; et (iii) dans un modèle en îles avec le même taux total de migration 4/9. Noter l'échelle logarithmique de la distance géographique. Figure adaptée de Rousset, 2004.

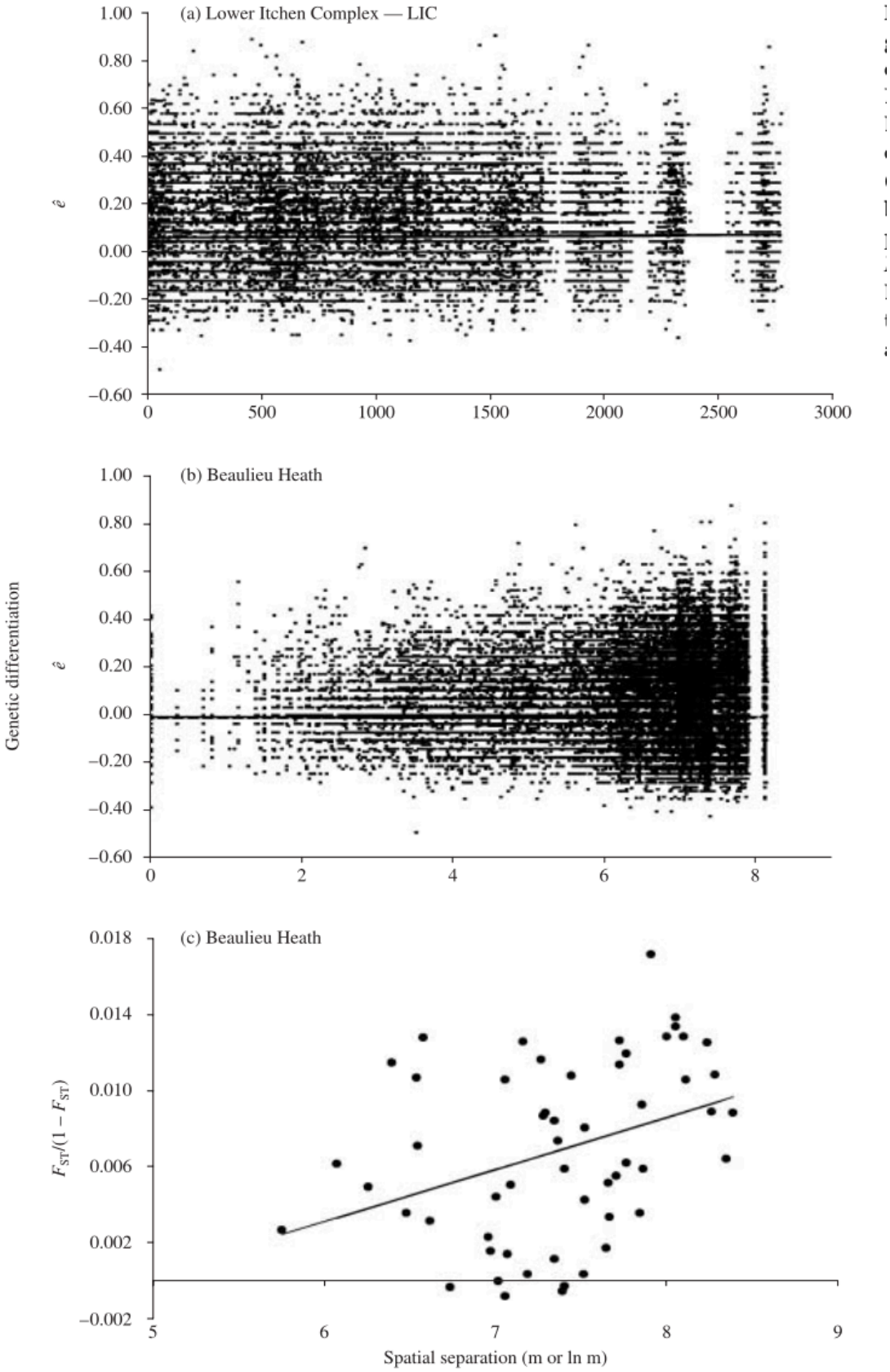


FIGURE 2.2 – Régressions linéaires entre la distance géographique et la différenciation génétique par paire d'individus ou de sous-populations de *Coenagrion mercuriale* dans trois sites géographiques. Ces graphiques sont typiques des regressions obtenues dans le cadre de l'analyse IBD par la méthode de la régression de Rousset (1997, 2000). Figure issue de Watts *et al.* (2006).

Le résultat principal de ces analyses, illustrant la raison du choix de $F_r = \frac{Q_w - Q_r}{1 - Q_w}$, est que l'on a donc une relation linéaire entre F_r et la distance géographique en une dimension, et entre F_r et le logarithme de la distance géographique en deux dimensions (Figure 2.1). On remarquera que les approximations sont faites en considérant tout d'abord que les distances sont grandes (éq.2.1 et éq.2.4), et petites ensuite (éq.2.6 et éq.2.7). La relation linéaire sera donc valide pour des distance intermédiaires. Que ce soit en une ou deux dimensions, la pente de cette relation linéaire est fonction de $D\sigma^2$. Ces approximations permettent une description relativement simple de la différenciation attendue sous ce modèle et permettent d'envisager une estimation du produit $D\sigma^2$ par la pente de la relation entre la différenciation, observée sur des marqueurs génétiques et mesurée par une estimateur de F_r, et la distance géographique (ou le logarithme de la distance pour le modèle en deux dimensions, voir Figure 2.2). C'est la méthode d'inférence de $D\sigma^2$ dite "méthode de la régression" de Rousset (1997, 2000). Nous verrons en détail les caractéristiques d'une telle estimation et testerons ses performances dans la section suivante 2.2.

On peut noter aussi que la différenciation sous ces modèles n'est pas uniquement fonction de $D\sigma^2$ mais également d'autres caractéristiques de la dispersion "contenues" dans les constantes A. Aucune méthode à ce jour ne permet d'extraire cette information sur la dispersion contenue dans l'intercepte de la régression. De plus, pour des taux de migration faibles, la différenciation entre deux sous-populations adjacentes est proche de la différenciation attendue sous un modèle en îles avec le même nombre d'émigrants (Figure 2.1). Ceci confirme que $D\sigma^2$ n'est pas la seule caractéristique de dispersion jouant sur la différenciation génétique, mais que Nm joue aussi un rôle important, et peut sans doute être co-estimé avec $D\sigma^2$ à partir d'un jeu de données réel avec des méthodes autres que la régression (voir sections 1.4).

Il existe de nombreuses raisons de considérer d'autres statistiques de différenciation que a_r pour les inférences sous IBD. La relation entre la statistique utilisée et les paramètres démographiques du modèle doit être bien définie ; et cette dernière relation doit être, dans une certaine mesure, robuste aux processus de mutation et à la conception de l'échantillonnage. Le lecteur trouvera des exemples de conception et de tests statistiques différents dans Hardy & Vekemans (1999) et Hardy (2003) pour les plantes, ainsi que dans Watts *et al.* (2006) pour une amélioration du a_r pour les populations avec un faible IBD (c.ad. avec $D\sigma^2$ grand). On peut également choisir d'utiliser des statistiques biaisées avec une faible variance pour tester l'IBD avec plus de puissance, puis utiliser des statistiques non biaisées pour faire des inférences démographiques (ce qui peut être fait avec e_r puis a_r, voir Rousset, 2008).

Ainsi, Slatkin (1993) avait mis au point la première méthode d'inférence par régression du produit Nm dans un modèle stepping-stone, basée sur la régression (linéaire) des estimateurs de $M \equiv (1/FST - 1)/2$ avec la distance géographique sur une double échelle logarithmique (Nm est donnée par l'ordonnée à l'origine). L'utilisation de M présente deux intérêts principaux : (i) M est à peu près indépendant des taux de mutation et du plan d'échantillonnage ; (ii) la relation entre M, Nm et la distance géographique r est simple dans le modèle stepping-stone ($M = Nm/r$). Cette méthode permet des inférences quantitatives de la dispersion dans le cadre de ce modèles, mais pas dans le cadre de modèles IBD plus généraux, car la relation simple entre M et le nombre de migrants n'est plus valable.

Sur la figure 2.3, l'aire gris foncé représente la masse de probabilité correspondant à F_r. Sous isolement par la distance, la relation entre les F-statistiques et

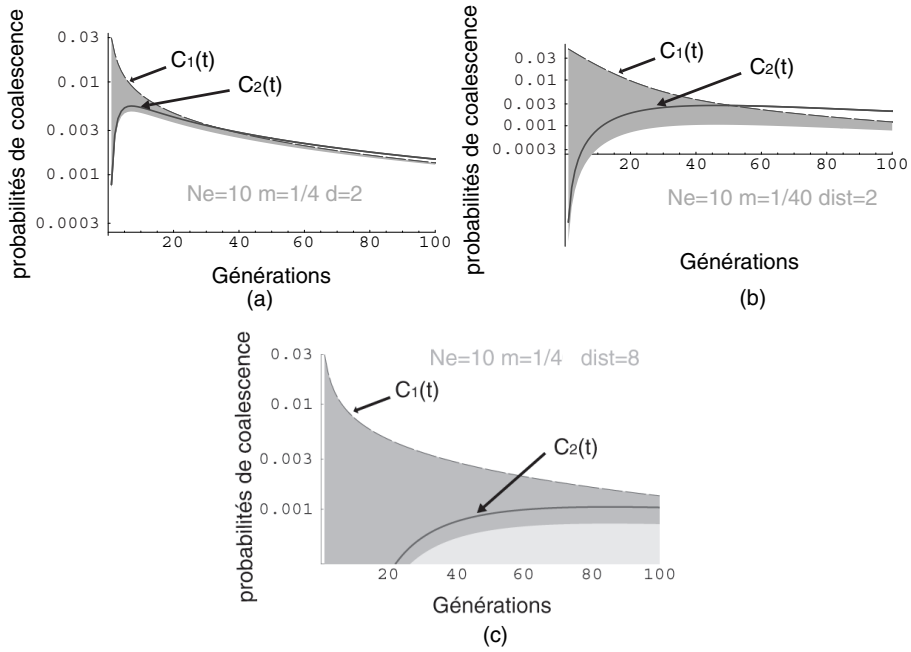


FIGURE 2.3 – Distribution des probabilité de coalescence sous isolement par la distance en fonction du taux de migration et de l'éloignement des deux gènes considérés. $C_1(t)$ correspond aux temps de coalescence de gènes intra-dèmes. $C_2(t)$ correspond aux temps de coalescence de gènes inter-dèmes situé à 2 unités du réseau pour les figures (a) et (b) et à 8 pas de distance pour (c). La migration se fait uniquement entre dèmes adjacents (stepping-stone) avec un taux de migration de 1/4 pour (a) et (c) ; et 1/40 pour (b). Pour toutes ces figures, le réseau est en une dimension, le nombre de dèmes est 100 et chaque dème est constitué de 10 adultes haploïdes. L'aire gris foncée correspond à F_r. L'échelle de l'axe des ordonnées est logarithmique. Figure adaptée de Rousset, 2004

les temps de coalescence nous indique les même tendances que pour le modèle en îles. L'influence des mutations et des fluctuations démographiques passées sera donc d'autant plus faible que les taux de migration sont forts (Figure 2.3a vs. b) et, dans une moindre mesure que les tailles des dèmes (ou les densités) sont petites. On peut aussi souligner que cette influence sera d'autant plus faible que les dèmes (ou les individus) comparées seront proches géographiquement (Figure 2.3a vs. c).

2.2 Robustesse de la méthode de la régression

Les performances de la méthode de la régression ont été testées par simulation par Leblois *et al.* (2003) et Leblois *et al.* (2004). Comme vu ci-dessus, cette méthode repose sur une régression des estimateurs du paramètres F_r et le logarithme de la distance géographique, dont la pente est utilisée pour estimer le paramètre $D\sigma^2$, ou D est la densité d'individus adultes et σ^2 le carré moyen de la distance de dispersion parent-descendant. Cette méthode est en principe plus performante que les méthodes précédentes traitant des modèles d'isolement par la distance pour au moins deux raisons : (i) le modèle démographique sous-jacent, introduit en section 2.1, fait peu d'hypothèses quant aux distributions de dispersion et il est particulièrement robuste pour des distributions leptokurtiques, une caractéristique communément observée dans les populations naturelles (voir la section 1.3.2 relative aux distributions de dispersion) ; (ii) les variations des F-statistiques avec la distance géographique donnent des résultats plus facilement interprétables que les F-statistiques en elles-mêmes. Ce dernier point peut être illustré par les équations (2.6) et (2.7)

de la section 2.1 dans lesquelles on voit bien que les constantes "A", entrant dans la définition du paramètres a_{r}, sont des fonctions complexes des distributions de dispersion alors que la variation de a_{r} avec la distance géographique n'est fonction que de $D\sigma^2$.

Un autre intérêt majeur de cette méthode est que l'analyse de la différenciation est faite à une petite échelle géographique (échelle géographique locale), ou plutôt, devrait être faite à échelle intermédiaire pour respecte les hypothèse sous-jacentes aux approximations considérée dans la méthode d'inférence. Cette notion d'échelle géographique et de validité de la méthode ne sont pas très claires à ce stade de la réflexion car basées sur des approximations mathématiques (en plus contradictoire, $r \to \infty$ des éq.2.1 et éq.2.4), et $r \to 0$ ensuite (éq.2.6 et éq.2.7). Nous verrons par la suite que cette échelle géographique intermédiaire peut être caractérisée relativement à la valeur σ, permettant une interprétation plus aisée de ces contraintes spatiales, et correspondra in fine à de très petites échelles géographiques.

Or les hypothèses de stabilité démographique sont moins critiques quand on considère des petites surfaces (Slatkin, 1993). Plus précisément, les études à des échelles géographiques locales donnent de meilleures estimations car l'hétérogénéité des paramètres démographiques (tels que la densité ou la dispersion) est réduite et leur influence sur l'hétérogénéité des processus génétiques tels que la dérive le sont aussi (Rousset, 2001). Mais ceci ne fait qu'atténuer le problème de l'hétérogénéité potentielle des paramètres démographiques. Dans le cas de fortes variations dans un passé pas si lointain, il est légitime de se poser la question de la signification exacte des paramètres de dispersion et de densité estimés. Plus précisément, obtient-on un estimateur du paramètres actuel ou bien l'estimation subit-elle largement l'influence des variations antérieures ? D'autre part même si l'homogénéité spatiale est plus probable à des échelles démographiques locales, il est possible que la présence de zones présentant des densité plus fortes que d'autres ait une influence majeure sur les estimations du produit $D\sigma^2$, en particulier si l'échantillon a été récolté en partie ou totalement dans de telles zones. Enfin, le processus mutationnel des marqueurs génétiques sont souvent complexes et mal connus, et ne correspondent pas aux hypothèse d'identité par descendance sur laquelle est basé la méthode de la régression.

Il nous a paru intéressant de quantifier par simulation l'effet de tels écarts par rapport aux hypothèses de base du modèle sur l'estimation de $D\sigma^2$ par cette méthode des moments. Dans ce but, nous avons développé un algorithme de simulation génération par génération fondé sur la théorie de la coalescence (voir section 1.1) pour tester l'influence sur l'estimation du produit $D\sigma^2$: (i) de l'échelle d'échantillonnage des individus, (ii) des processus mutationnels (taux et modalités de mutation) des marqueurs utilisés, avec une référence spéciale aux microsatellites, et (iii) d'hétérogénéités spatiales et temporelles des paramètres démographiques.

Pour les modèles IBD, il n'existe pas de traitement analytique des probabilités et temps de coalescence pour plus de deux gènes. Les modèles classiques du n-coalescent et du coalescent structuré ne s'appliquent pas non plus puisque la migration est forte et les dèmes de très petite taille (un individu). Nous avons donc utilisé un algorithme de simulation de coalescence génération par génération que nous avons implémenté dans `IBDsim` (Leblois *et al.*, 2009). `IBDsim` permet de simuler de manière exacte le processus d'isolement par la distance sur un réseau dans la cadre général des processus de Wright-Fisher décrit en introduction, en considérant à chaque génération tous les évènements possibles de dispersion et de coalescence, jusqu'au MRCA de l'échantillon.

Notons que lorsque que l'habitat est hétérogène (réseau avec des tailles de sous-populations différentes), `IBDsim` calcule les distributions de dispersion "arrière" à partir des distributions "avant" choisies par l'utilisateur. Chaque point du réseau a alors une distribution de dispersion "arrière" propre, dépendant de la densité en chaque dème susceptible d'être le dème d'origine du gène considéré (à savoir l'ensemble des dèmes situés à une distance du dème d'origine inférieure ou égale à la distance maximale de dispersion). Soit $N_{x,y,G}$ le nombre d'individus au dème (x, y) à la génération parentale G. La probabilité de dispersion "arrière" de dx pas sur la première dimension et dy pas sur la deuxième dimension au dème (x_i, y_i) est alors de la forme

$$b_{dx,dy} = \frac{N_{x_i+dx,y_i+dy,G} \cdot f_{dx,dy}}{\sum_{dx,dy \leq d_{max}} N_{x_i+dx,y_i+dy,G}, f_{dx,dy}}. \tag{2.8}$$

où f est la distribution de dispersion avant, et d_{max} représente la distance maximale de dispersion.

La validation des outils que nous développons est un point essentiel, et `IBDsim` a été testé par comparaisons des probabilités d'identité de deux gènes obtenues par simulation avec celles obtenues analytiquement comme expliquée ci-dessus (différences relatives $< 10 - 4$).

Pour chaque question ci-dessous, nous avons donc simulé un échantillon de 100 (10×10) individus caractérisés par leurs coordonnées sur un réseau de taille (100×100) avec un individu par nœud, et génotypés à 13 locus polymorphes multi-alléliques de type marqueurs microsatellites, sauf précision contraire. La distribution de dispersion est de la forme $f_k = f_{-k} = M/k^n$ avec $\sigma^2 = 4$ et dont les paramètres sont

$$M_1 = 0.06,\ M_2 = 0.03 \text{ et pour } 2 < dx < 49,\ M = 0.802,\ n = 2.518. \tag{2.9}$$

Des arbres de coalescence indépendants (donc des simulations indépendantes) ont été utilisés pour simuler des données multilocus à des locus indépendants. Chaque simulation a été répété 1 000 fois générant 1 000 échantillons multilocus pour 100 individus partageant la même histoire démographique (à l'indépendance des marqueurs prés). Pour chaque jeu de données simulé, la valeur de la pente de la droite de régression entre $\hat{a}_r$ et le logarithme de la distance géographique est calculée par le logiciel `Genepop` (Rousset, 2008).

La qualité de l'inférence est évaluée par le calcul de son biais relatif (c.a.d. moyenne de (observé - attendu)/attendu), du carré moyen des erreurs relatives ("mean square error", MSE), et la proportion d'estimations tombant dans un intervalle de facteur 2 par rapport à la valeur attendue (c.a.d. [attendu/2 ; 2× attendu]) sur la valeur de la pente, pour éviter les valeurs infinies de $D\sigma^2$ pour une pente nulle. Pour tout le reste de ce document nous appellerons valeurs attendues les valeurs des paramètres choisies pour les simulations.

Il est important de noter que les biais observés dans nos simulations peuvent provenir : (i) d'un biais, inhérent à la méthode, dû à l'effet des forts taux de mutation sur la valeur du paramètre (que nous appellerons "biais paramétrique") ; (ii) d'un biais dû à la déviation des estimateurs par rapport à la valeur du paramètre lorsque l'on considère un échantillon fini d'individus et de locus ("biais d'échantillonnage") ; et (iii) d'un biais introduit par les écarts aux hypothèses du modèle démographique (variations spatiales et temporelles des paramètres démographiques).

2.2.1 Influence de l'échelle d'échantillonnage et de la taille de l'habitat

Comme vu ci-dessus, la méthode de la régression repose sur l'hypothèse que la différenciation est observée à une échelle intermédiaire vis à vis de σ^2. De même, des simulations ("avant") antérieures (Rousset, 2000) ont suggéré que l'inférence est optimale si l'on échantillonne $(100D\sigma^2)$ individus sur une surface d'environ $(10\sigma \times 10\sigma)$. Aussi, puisque le nombre d'individus à échantillonner est nécessairement limité, la méthode considérée ici sera moins efficace si $D\sigma^2$ est grand.

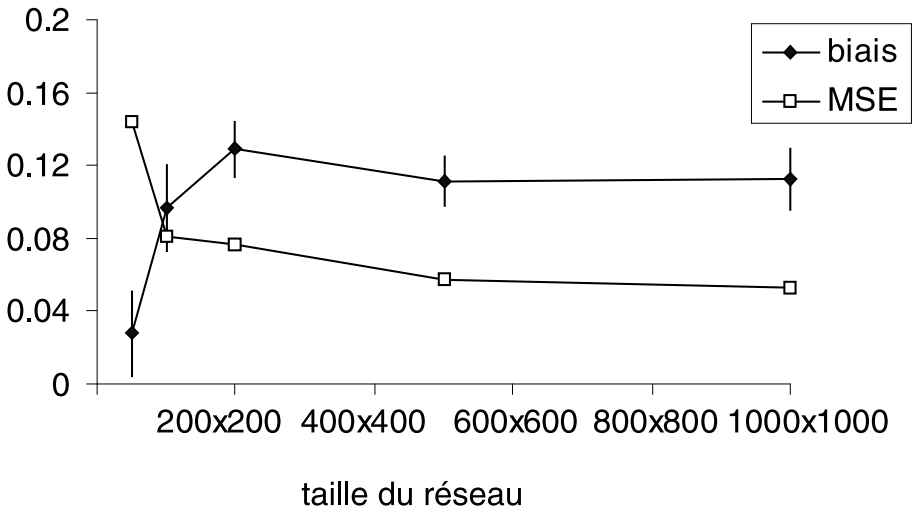


FIGURE 2.4 – Influence de la taille du réseau sur l'estimation du produit $D\sigma^2$. Les barres verticales représente l'erreur standard sur le biais. 500 répétitions par taille du réseau ont été utilisé pour faire cette figures. Les paramètres utilisés sont décrit ci-dessus et l'échantillon est pris sur une surface de (20×20).

Les premiers tests ont été fait en considérant différentes tailles de réseau, et montrent que la taille du réseau a peu d'effet sur l'estimation à condition que le réseau soit plus de 10 fois plus grand que la dispersion moyenne (Figure 2.4). En effet, à l'exception du cas où la taille du réseau est particulièrement petite (50×50), le biais et la MSE sont proches de ceux obtenus pour un très grand réseau (1000×1000).

TABLE 2.1 – Influence de l'échelle d'échantillonnage sur l'estimation de $1/(4\pi D\sigma^2)$. L'aire d'échantillonnage est exprimée en unité du réseau (voir le texte pour les détails) et la taille du réseau est de 500x500.

	échelle d'échantillonnage (surface)			
	1 (10×10)	2 (20×20)	5 (50×50)	10 (100×100)
Biais	0.219	0.130	-0.056	-0.205
(erreur standard)	(0.0077)	(0.0077)	(0.0072)	(0.0064)
MSE	0.106	0.0763	0.0554	0.082
Facteur 2	0.999	0.996	0.967	0.93

Quatre différents schémas d'échantillonnage de 100 individus ont ensuite été testés : (i) tous les nœuds sur une aire de $(5\sigma \times 5\sigma$, donc 10x10), (ii) tous les deux nœuds sur une aire de $(10\sigma \times 10\sigma$, c.a.d. 20x20), (iii) tous les cinq nœuds sur une aire de $(25\sigma \times 25\sigma$, c.a.d. 50x50), et (iv) tous les dix nœuds sur une aire

de ($50\sigma \times 50\sigma$, c.a.d. 100x100). Ces tests montrent que l'échelle d'échantillonnage semble avoir un effet limité sur la MSE de l'estimation de $D\sigma^2$ (Tableau 2.1). Quelle que soit l'échelle considérée, la MSE est petit (par ex. entre 5% et 12% dans les cas étudiés). Par contre, l'échelle d'échantillonnage a un effet plus important sur le biais. Un échantillon pris sur une aire deux fois plus petite que l'aire recommandée entraîne un fort biais positif (22%) et donc une sous-estimation de $D\sigma^2$. Le biais diminue ensuite lorsque la surface échantillonnée augmente et atteint de fortes valeurs négatives (c.a.d. sur-estimation de $D\sigma^2$) lorsque la surface recommandée est largement dépassée (-21% pour la troisième colonne du Tableau 2.1). L'estimation reste toutefois correcte même pour des échelles d'échantillonnage extrêmes puisque une large majorité des estimations tombent dans l'intervalle de facteur 2 (>93%).

Pour la suite, nous considérerons un échantillon de 100 individus pris sur une surface de ($10\sigma \times 10\sigma = 20 \times 20$) évoluant sur un tore de (100×100).

2.2.2 Influence des processus de mutation

Nous avons ensuite testé l'effet des processus mutationnels car la méthode de la régression, reposant sur une analyse en probabilité d'identité par descendance et non par état, ne prends pas en compte l'homoplasie. De plus, elle fait l'hypothèse que les taux de mutations sont faibles. Elle pourrait donc être sensibles aux forts taux de mutation et aux processus mutationnels complexes des marqueurs microsatellites, largement utilisés à l'époque ou nous avons fait ces tests. Les microsatellites n'étant plus tellement d'actualité, je ne résumerai que brièvement nos résultats et le lecteur pourra se référer à Leblois *et al.* (2003) pour plus de détails.

Cinq modèles mutationnels (voir section 1.2.1) : (i) le modèle à nombre d'allèles infini (IAM), (ii) le modèle à K allèles (KAM), avec un choix arbitraire de $K = 10$, (iii) le modèle de mutation par pas (SMM), (iv) le modèle de mutation par pas généralisé (GSM), avec une variance de la loi géométrique de 0.36 (correspondant à une loi géométrique de paramètre $p_{GSM} = 0.22$), correspondant à la valeur estimée sur le grand jeu de données de mutation microsatellitaires de Dib *et al.* (1996) chez l'homme, et (v) un GSM avec des contraintes sur les tailles des allèles (range de taille de 10 ou 20 répétitions).

TABLE 2.2 – Influence des modèles mutationnels sur l'estimation de $1/(4\pi D\sigma^2)$ à diversité génétique constante.

	modèle mutationnel à diversité génétique constante 0.68					
	IAM	KAM ($K = 10$)	SMM	GSM	GSM borné ($K = 10$)	GSM borné ($K = 20$)
Taux de mutation	0.0001	0.000218	0.000342	0.00012	0.0005	0.0002
Biais	0.111	0.104	0.118	0.121	0.0997	0.108
(erreur standard)	(0.01)	(0.01)	(0.015)	(0.012)	(0.0101)	(0.01)
MSE	0.119	0.109	0.119	0.159	0.112	0.121
Facteur 2	0.96	0.97	0.96	0.938	0.976	0.962

Pour un taux de mutation donné, la diversité génétique ($1 - Q_0$), correspondant à la proportion attendue d'individus hétérozygotes dans la population, est proportionnelle au taux de mutation mais varie en fonction du modèle mutationnel

considéré. Puisque la diversité génétique peut avoir un effet important sur l'estimation, nous avons considéré tous les modèles mutationnels précédents à diversité génétique constante en ajustant les taux de mutations selon les calculs de Rousset (1996) (Tableau 2.2). Les tests montrent que la nature du modèle mutationnel a très peu d'influence sur l'estimation de $D\sigma^2$ (Tableau 2.2). Quel que soit le modèle mutationnel considéré, le biais est positif et d'environ 10% et les différences de MSE sont mineures. Pour tous les modèles mutationnels considérés plus de 93% des estimations tombent dans l'intervalle de facteur 2. Les conclusions sont très similaires quand on considère un taux de mutation fixé à $5 \cdot 10^{-4}$ (voir Leblois *et al.* (2003)).

L'influence du taux de mutation, ou de la diversité génétique, a ensuite été étudiée en considérant un des modèles mutationnels les plus réalistes par rapport aux processus mutationnels des microsatellites, le GSM (Estoup & Cornuet, 1998, voir section 1.2.1). L'interprétation est faite en terme de diversité génétique car c'est une mesure facile à obtenir en pratique contrairement aux taux de mutation.

Nos résultats montrent que la diversité génétique a un effet important sur le biais et la MSE de l'estimation de $D\sigma^2$ (Tableau 2.3). la MSE est plus fortement influencé par la diversité génétique que le biais. Pour des diversités génétiques de l'ordre de 0.5, le biais observé est positif et inférieur à 10% mais la MSE est forte (c.a.d. supérieur à 20%). Cependant, pour cette diversité génétique plus faible que la diversité génétique souvent observée aux locus microsatellites, 84% des estimations sont dans l'intervalle de facteur 2.

TABLE 2.3 – Influence du taux de mutation sur l'estimation de $1/(4\pi D\sigma^2)$. Le modèle mutationnel est le GSM.

	Taux de mutation				
	0.00005	0.00012	0.0005	0.005	0.05
Diversité génétique	0.56	0.68	0.77	0.82	0.85
Biais	0.0972	0.121	0.104	0.00946	-0.390
(erreur standard)	(0.016)	(0.012)	(0.0086)	(0.0062)	(0.0055)
MSE	0.268	0.159	0.0852	0.0380	0.182
Facteur 2	0.84	0.94	0.99	0.99	0.761

Pour des diversités génétiques fortes (c.a.d. 0.85), le biais devient fortement négatif et la MSE augmente rapidement avec la diversité génétique. Ce résultat traduit le fait que pour des diversité génétiques fortes, le "biais paramétrique" (a priori du à la violation de l'hypothèse de faible taux de mutation), qui est négatif (Rousset, 1997), devient plus important que le "biais d'échantillonnage" et le biais global devient donc négatif.

Enfin, il est souvent considéré que les variations de taux de mutation entre locus peuvent influencer la précision des estimations en génétique des populations (Takezaki & Nei, 1996; Gonser *et al.*, 2000). Nous avons donc testé l'influence d'un taux de mutation variable entre locus, indépendamment ou en fonction de la longueur de l'allèle. Les résultats ne sont pas détaillé ici, mais ces tests montrent qu'un taux de mutation variable a peu d'effet sur l'estimation du produit $D\sigma^2$ (Leblois *et al.*, 2003). Le biais et la MSE sont inférieurs à 12% ce qui diffère peu du cas avec un taux de mutation fixe, et plus de 98% des estimations sont dans l'intervalle de facteur 2.

2.2.3 Influence d'hétérogénéités spatiales et temporelles

On peut imaginer une infinité de scénarios intégrant des hétérogénéités spatiale et temporelles, nous avons choisi de focaliser notre étude sur des scénarios démographiques souvent rencontrés en biologie de la conservation ou lors de l'étude de bio-invasions. Dans ce contexte, nous avons évalué les effets sur l'estimation de $D\sigma^2$ (i) d'un changement au cours du temps de la dispersion, (ii) d'une réduction ou d'une augmentation de la densité au cours du temps, (iii) d'une expansion spatiale à densité constante, et (iv) d'un échantillonnage dans une zone de forte densité.

Variation temporelle de la dispersion

Nous avons étudié l'effet d'une diminution des capacités de dispersion au cours du temps au travers d'une diminution du paramètre σ^2 en choisissant deux distributions de dispersion avec des σ^2 très différents : $\sigma^2 = 4$ pour le présent et passé récent et $\sigma^2 = 100$ pour le passé plus lointain, les autres paramètres de simulation étant constants.

Nos simulations montrent que le biais (calculé sur la pente $1/4\pi D\sigma^2$) dû à une réduction de la dispersion dans le temps est négatif (Tableau 2.4), et correspond donc à une sur-estimation du $D\sigma^2$ actuel, attendue si la méthode d'inférence a une certaine mémoire des caractéristiques de dispersion passées. Cependant, cette mémoire semble de courte durée puisqu'une réduction 100 générations dans le passé introduit seulement un biais très faible compensé dans les simulations par les "biais paramétrique et d'échantillonnage" (cf. première colonne du tableau 2.4). De plus, même pour une réduction récente ($G_c = 10$), le biais est inférieur à 25%, une valeur relativement faible par rapport à la forte amplitude de la variation de σ^2 modélisée (facteur 25, qui pourrait donc théoriquement introduire un biais négatif maximal de -96% et pas 2500% car il est calculé sur l'inverse de σ^2). Nos simulations montrent donc que globalement la précision de l'estimation du $D\sigma^2$ actuel est assez robuste aux changements temporels de dispersion.

TABLE 2.4 – Effet d'une diminution de la dispersion dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. Une distribution de dispersion avec $\sigma^2 = 4$ est utilisée du présent jusqu'au moment du changement dans le passé (G_c), puis de G_c à TMRCA, une distribution de dispersion avec $\sigma^2 = 100$ est considérée. Le changement de dispersion intervient à trois moments différents dans le passé suivant les simulations ($G_c = 10$, 20, 100) et à un temps infini comme simulation témoin (aucun changement).

	G_c			
	∞	100	20	10
Biais	0.444	0.0923	-0.0795	-0.234
(erreur standard)	(0.0062)	(0.0081)	(0.0076)	(0.0074)
MSE	0.228	0.0743	0.0642	0.109
Facteur 2	0.99	0.99	0.97	0.88

Variations temporelle de la densité

Une diminution et une augmentation de la densité dans le temps ont ensuite été explorées en considérant quatre modèles de réseau avec chacun un nombre d'individu

TABLE 2.5 – modèles utilisés pour évaluer l'effet de changements de densité dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu le changement de densité. TMRCA correspond au temps de l'ancêtre commun le plus récent de l'échantillon.

Changement démographique		Densité (Nombre d'individus par nœud du réseau) du présent à G_c	de G_c à TMRCA	Facteur
Goulet d'étranglement	faible	1	10	10
	fort	1/9	10	90
Explosion démographique	faible	1	1/9	9
	fort	1	1/100	100

par dèmes différents (par ex. 1, 10, 1/9, 1/100). Les cas avec moins d'un individu par dème ont été modélisés en considérant qu'une certaine proportion des dèmes sont vides (par ex. une densité de 1/9 est obtenue avec 8/9 des dèmes vides). Les différentes densités utilisées sont résumées dans le tableau 2.5. Les changements se font à trois moments dans le passé, $G_c = 10$, 20, 100 et infini comme témoin.

TABLE 2.6 – Effet d'une diminution de la densité dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu le changement de densité.

Intensité		G_c ∞	100	20	10
Faible	Biais	0.444	0.099	-0.063	-0.22
	(erreur standard)	(0.0062)	(0.0070)	(0.0064)	(0.0061)
	MSE	0.228	0.0588	0.0449	0.0868
	Facteur 2	0.99	0.99	0.99	0.92
Forte	Biais	-0.014	-0.074	-0.33	-0.53
	(erreur standard)	(0.0042)	(0.0027)	(0.0017)	(0.0012)
	MSE	0.0175	0.0128	0.115	0.278
	Facteur 2	1	1	1	0.238

Une diminution de la densité entraîne un biais négatif (tableau 2.6), correspondant à une sur-estimation de $D\sigma^2$, et reflétant la mémoire des fortes densités passées. Pour une réduction d'un facteur 10, la méthode est assez robuste quand le changement a lieu à 20 ou plus de 20 générations dans le passé : Le biais et la MSE sont faibles (c.a.d. moins de 10%) et 99% des estimations correspondent bien à la valeur du paramètres $D\sigma^2$ actuel à un facteur 2 près. Pour un changement plus récent (par ex. $G_c = 10$), le biais est nettement plus important mais la MSE reste petite et 92% des estimations tombent dans l'intervalle de facteur 2 par rapport à la valeur actuelle de $D\sigma^2$.

L'effet d'une réduction forte de densité (c.a.d. d'un facteur 90) est beaucoup plus marquée. Pour un changement récent (c.a.d. moins de 10 générations dans le passé), le biais atteint -50% et seulement 24% des estimations correspondent à la valeur actuelle de $D\sigma^2$ à un facteur 2 près. Pour un changement ayant eu lieu à 20 générations ou plus dans le passé, plus de 99% des estimations sont dans l'intervalle de facteur 2 par rapport à la valeur actuelle de $D\sigma^2$, et pour $G_c = 100$, le biais et la MSE retrouvent un niveau proche de la simulation témoin. Ainsi, même pour des goulets d'étranglement forts et relativement récents, la méthode de la régression

montre une bonne robustesse.

TABLE 2.7 – Effet d'une augmentation de la densité dans le temps sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu le changement de densité.

		G_c			
Intensité		∞	100	20	10
Faible	Biais	0.44	0.32	0.69	1.4
	(erreur standard)	(0.0062)	(0.040)	(0.043)	(0.046)
	MSE	0.228	1.72	2.33	4.07
	Facteur 2	0.99	0.45	0.38	0.24
Forte	Biais	0.43	0.65	2.2	3.9
	(erreur standard)	(0.0064)	(0.0094)	(0.015)	(0.019)
	MSE	0.228	0.508	5.27	15.8
	Facteur 2	0.99	0.89	0.003	0

Dans le cas d'une augmentation de la densité, le biais positif observé dans le tableau 2.7, correspondant à une sous-estimation du $D\sigma^2$ actuel, reflète bien la mémoire des faibles densités passées. Même pour une faible augmentation relativement ancienne de la densité (facteur 9, 100 générations), le biais et la MSE sont forts, et la proportion d'estimations correspondant au $D\sigma^2$ actuel à un facteur 2 près est faible (moins de 50%). Cet effet augmente considérablement avec l'intensité de la variation. Pour une explosion démographique d'un facteur 100 et pour $G_c = 10$, le biais atteint 390% et aucune des estimations ne tombent dans l'intervalle de facteur 2 (Tableau 2.7). Ainsi, en dépit du fait que le biais et la MSE diminuent lorsque G_c augmente, l'estimation reste incertaine dans les gammes de temps étudiées que ce soit pour une variation forte ou faible. Ces résultats contrastent fortement avec les résultats obtenus précédemment et met en avant la forte sensibilité de la méthode à des augmentations (diminution passée) des densité.

Expansion spatiale de la population à densité constante

Le quatrième type de situation étudié est une expansion spatiale de la population à densité constante (Figure2.5). La population introduite dans le nouvel habitat vide est composé d'individus ayant évolué dans une population source à l'équilibre sous certaines caractéristiques démographiques (c.a.d. densité et distribution de dispersion). La population introduite s'étend en quelques générations (en 2 générations dans nos simulations) sur tout le nouveau territoire avec les même caractéristiques démographiques que celles de la population source. L'échantillon d'individus est pris sur le nouvel habitat à une distance de 50 dèmes de la zone d'introduction.

Toutes les statistiques (Biais, MSE et proportion d'estimation dans l'intervalle de facteur 2) calculées sur ces simulations indiquent que l'estimation du $D\sigma^2$ actuel est très robuste à une expansion spatiale ayant eu lieu 20 générations ou plus dans le passé (Tableau 2.8). Pour les cas les plus récents $G_c = 10$, le biais est négatif et de 8%, ce qui correspond à une faible sur-estimation du $D\sigma^2$ actuel, indiquant que le patron d'équilibre d'IBD n'a peut être pas encore eu le temps de s'établir complètement. la MSE est faible (10%) et 98% des estimations tombent dans l'intervalle de facteur 2. Ainsi, nos simulations montrent qu'une expansion spatiale, telle qu'elle est modélisée ici, a une influence limitée sur l'estimation de $D\sigma^2$. La méthode de la

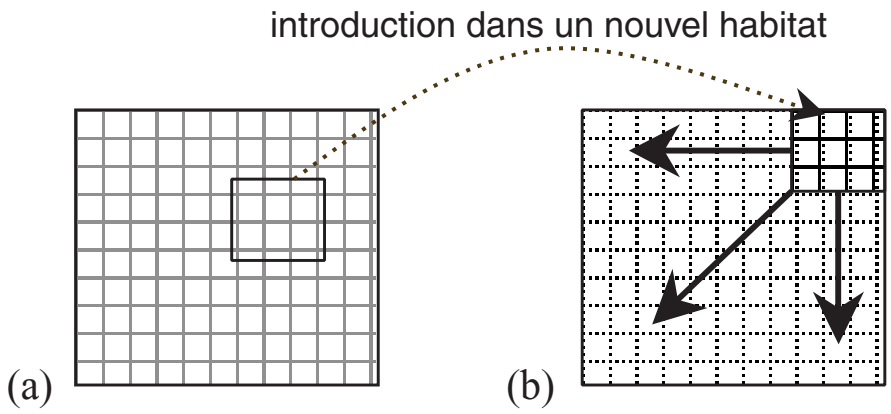


FIGURE 2.5 – Schéma d'une expansion démographique à densité constante telles que modélisées dans Leblois *et al.* (2004). La grille (a) correspond à une population source dont est issu un échantillon d'individus (petit cadre noir) introduits dans un nouvel habitat (flèche pointillée). La grille (b) correspond au nouvel habitat vide dans lequel l'échantillon introduit va s'étendre en quelques générations (flèches noires).

TABLE 2.8 – Effet d'une expansion spatiale à densité constante sur l'estimation de $1/(4\pi D\sigma^2)$. G_c indique le moment dans le passé auquel a lieu l'expansion.

	G_c			
	∞	100	20	10
Biais	0.43	0.39	0.13	-0.0824
(erreur standard)	(0.0076)	(0.013)	(0.011)	(0.010)
MSE	0.243	0.23	0.08	0.0581
Facteur 2	0.99	0.98	0.99	0.97

régression est donc précise même pour des expansions récentes.

Hétérogénéité spatiale de la densité

La dernière situation étudie dans Leblois *et al.* (2004) reflète le fait que les biologistes collectent généralement leurs échantillons sur des zones où l'espèce étudiée est plus facile à collecter, c'est à dire sur des zones de fortes densités. Pour cela nous avons considéré un modèle en réseau avec une densité homogène sauf sur une petite zone (20×20 et 40×40) sur laquelle la densité est dix fois plus forte que sur le reste du réseau (Figure 2.6, (a) sur, (b) sur et autour ou (c) en dehors de la zone d'échantillonnage. Nous avons alors évalué si l'estimation correspondait plutôt à la densité sur la zone échantillonnée ou si cette estimation était largement influencée par la densité autour de la zone échantillonnée.

Ces simulations montrent que l'estimation de $D\sigma^2$ n'est pas robuste lorsque la zone de forte densité est petite et correspond strictement à la surface échantillonnée (Tableau 2.9). Les valeurs du biais et du MSE indiquent que, dans ce cas, la zone de faible densité autour de l'échantillon influence fortement l'estimation de $D\sigma^2$, qui devient alors une mauvaise mesure aussi bien de la densité locale (c.a.d. sur la surface échantillonnée) que de la densité avoisinante (c.a.d. autour de la zone d'échantillonnage). Si la zone de forte densité plus grande que la surface échantillonnée, le biais et la MSE sont beaucoup plus faibles quand on se réfère à la densité locale plutôt

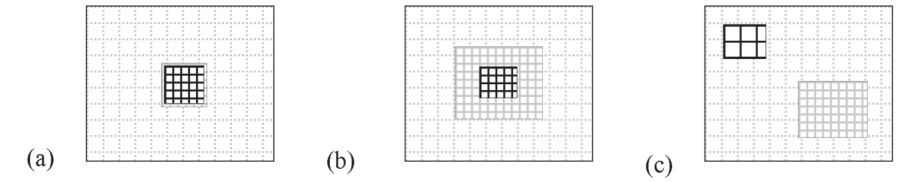


FIGURE 2.6 – Schéma d'hétérogénéités spatiales telles que modélisées dans Leblois *et al.* (2004). (a) Petite zone de forte densité (grille gris foncé) correspondant strictement à la zone échantillonnée (grille noire) sur un réseau en deux dimensions de densité plus faible (grille gris claire). (b) Grande zone de forte densité incluant la zone échantillonnée sur un réseau en deux dimensions de densité plus faible. (c) Grande zone de forte densité située hors de la zone d'échantillonnage sur un réseau en deux dimensions de densité plus faible.

TABLE 2.9 – Effet d'hétérogénéités spatiales de la densité sur l'estimation de $1/(4\pi D\sigma^2)$. Colonne 'Densité locale' : la densité attendue est la densité locale (c.a.d. sur la zone d'échantillonnage). Colonne 'Densité avoisinante' : la densité attendue est la densité avoisinante (c.a.d. autour de la zone échantillonnée). Les colonnes 'contrôle' correspondent à un réseau homogène avec une densité correspondant à la densité locale ou avoisinante pour les colonnes respectives de densité locale et avoisinante.

Hétérogénéité spatiale		Densité locale		Densité avoisinante	
		Estimation	contrôle	Estimation	contrôle
Petite zone	Biais	2.1	0.45	-0.69	0.43
de	(erreur standard)	(0.017)	(0.025)	(0.0017)	(0.0076)
forte densité	MSE	4.76	0.83	0.477	0.243
	Facteur 2	0.018	0.65	0.001	0.99
Grande zone	Biais	0.39	0.45	-0.86	0.43
de	(erreur standard)	(0.013)	(0.025)	(0.0013)	(0.0076)
forte densité	MSE	0.330	0.83	0.743	0.243
	Facteur 2	0.9	0.65	0	0.99
Grande zone de	Biais	0.45	0.43	13.5	0.45
forte densité	(erreur standard)	(0.0075)	(0.0076)	(0.075)	(0.025)
hors zone	MSE	0.256	0.243	187	0.83
d'échantillonnage	Facteur 2	0.99	0.99	0	0.65

qu'avoisinante pour la valeur attendue de $D\sigma^2$. Environ 90% des estimations correspondent, à un facteur 2 près, à la valeur locale de $D\sigma^2$, alors qu'aucune estimation ne tombe dans l'intervalle de facteur 2 par rapport à la valeur avoisinante de $D\sigma^2$. Nos simulations montrent donc que la méthode estime les paramètres démographiques locaux et qu'elle est robuste pour ces mesures quand la densité est relativement homogène autour de la zone échantillonnée (par ex. sur une surface environ égale à quatre fois la surface échantillonnée). Le troisième cas d'une zone de forte densité située hors de la zone d'échantillonnage confirme ce résultat puisqu'une zone de forte densité située à 50 nœuds de la zone échantillonnée n'a quasiment aucune influence sur l'estimation du $D\sigma^2$ local.

2.2.4 Conclusions des tests par simulations

Processus mutationnels Une première conclusion générale de ces tests par simulation est que les modèles de mutation des marqueurs génétiques ont peu d'effets sur les performances de la méthode d'estimation du produit $D\sigma^2$ par la régression mais que la diversité génétique (celle-ci étant largement influencée par le taux de mutation) a une influence importante sur l'estimation de $D\sigma^2$. Ceci est en accord avec des études antérieures démontrant que le taux de mutation est un facteur plus important que les processus mutationnels pour l'estimation de paramètres démographiques par F-statistiques (Rousset, 2001; Estoup *et al.*, 2002). De plus, les effets de taux de mutation variables entre locus et entre allèles semblent faibles. La diversité génétique typiquement observée aux locus microsatellites se situe généralement entre 0.5 et 0.8 (revue dans Estoup & Angers, 1998), ce qui correspond aux niveaux de diversité maximisant les performances de l'estimation de $D\sigma^2$. Les marqueurs microsatellites sont donc bien appropriés pour l'estimation de $D\sigma^2$ (mais un grand nombre de marqueurs SNPs devraient aussi donner de très bons résultats, sans doute même plus précis). Ces simulations ont aussi montré qu'il est préférable d'éviter des locus avec une diversité génétique trop forte (c.a.d. >0.85) car ces locus biaiseront fortement l'estimation de $D\sigma^2$.

Échelles d'échantillonnage Il est attendu que les effets des processus mutationnels et des taux de mutation élevés sur l'estimation de $D\sigma^2$ soient plus importants à large échelle géographique (comme nous l'avons précédemment suggéré par les représentations des probabilités d'identité en fonction des distributions des temps de coalescences, voir sections 1.2.3 et 2.1). En accord avec ces prédictions, nos résultats montrent qu'échantillonner à trop grande échelle entraîne une sur-estimation de $D\sigma^2$. Ainsi échantillonner sur des grandes distances diminue la probabilité de détecter un patron d'isolement par la distance. Au contraire, échantillonner sur une zone géographique trop petite entraîne une sous-estimation de $D\sigma^2$. Une explication possible est que la relation linéaire entre a_r et le logarithme de la distance est moins fiable à petites distances (c.a.d. pour $r < \sigma$, Rousset, 1997). Cependant, l'utilisation d'échantillons non appropriés au cas biologique étudié (c.a.d. plus petit ou plus grand que la surface recommandée de $(10\sigma \times 10\sigma)$) donnent des estimations relativement robustes. il est donc important de restreindre spatialement l'échantillonnage afin de rester à une échelle géographique très locale. Cependant une estimation précise demande aussi un échantillonnage plus large quand σ augmente, ce qui implique une connaissance à priori de la valeur de σ. Il paraît donc fortement conseillé de procéder en deux étapes en faisant une estimation préliminaire de σ permettant de définir une échelle d'échantillonnage adaptée qui permettra une estimation plus pré-

cise des paramètres de dispersion. En l'absence d'estimation préliminaire de σ, une estimation grossière déduite de la connaissance à priori de certaines caractéristiques de la dispersion semble utile pour définir une échelle minimale d'étude (par ex. la vitesse de colonisation du crapaud de la canne à sucre en Australie, Leblois *et al.*, 2000)

Hétérogénéité démographiques spatiales et temporelles Dans la limite des situations étudiées ici, et à l'exception d'une explosion démographique, nos résultats montrent que des fluctuations temporelles et spatiales des paramètres démographiques, si elles ne sont pas trop importantes, ni trop récentes, ont une influence limitée sur l'estimation de $D\sigma^2$ actuel et local. Il est important de noter que nous parlons ici-de changements sur une échelle de temps de l'ordre de quelques dizaines de générations. Si cette échelle peut paraître extrêmement récente en génétique des populations, elle reflète pour de nombreuses espèces (par ex. espèces menacées et bio-invasions) les changements démographiques liés à l'activité humaine. Soulignons également que les nombres de générations définissant les moments dans le passé auxquels ont eu lieu les changements doivent être considérés uniquement comme des indices approximatifs de la durée de l'effet des changements démographiques étudiés. En effet, la persistance dans le temps des effets des fluctuations démographiques dépend fortement de nombreuses caractéristiques des modèles démographiques (par ex. les valeurs de σ et de D) et des situations de déséquilibre. Les tests ayant été fait dans le modèle d'IBD individuel en habitat continu, la robustesse de la méthode de la régression devrait être moindre pour des modèles IBD en dèmes, et ceci d'autant plus que les dèmes sont grands, comme le suggèrent les raisonnement sur les probabilité d'identité en terme de distributions des temps de coalescence. Il parait donc préférable de considérer des tendances globales que des nombres précis de générations pour chaque situation.

A l'époque, la robustesse globale de la méthode de la régression à diverses fluctuations des paramètres démographiques dans l'espace et dans le temps contredisait de précédentes études sur les déséquilibres évolutifs. Dans leur commentaire, Koenig *et al.* (1996) avait conclu que l'estimation de paramètres de dispersion à partir de données génétiques donne des indications sur les flux de gènes passés plutôt qu'actuels, alors que les méthodes directes, telles que les techniques de capture-marquage-recapture, donne de meilleures estimations des paramètres de dispersion actuels. Boileau *et al.* (1992) avait montré de façon similaire que des centaines ou des milliers de générations étaient nécessaires pour effacer les effets de processus de colonisation sur des estimateurs de type F_{ST} à partir de données allozymiques dans de grandes populations. Ces auteurs concluaient que les estimations de flux de gènes à partir de données génétiques doivent être interprétées avec "précautions". Les fluctuations démographiques temporelles ont probablement un effet fort et persistant sur certains statistiques et méthodes. Néanmoins, notre étude montre que certaines méthodes indirectes, et certains marqueurs génétiques, donnent des estimations satisfaisantes de la densité et de la dispersion actuelles même si l'histoire démographique des populations étudiées inclut des fluctuations démographiques relativement récentes.

Interprétation de la robustesse générale de la méthode à l'aide des temps de coalescence Cette robustesse générale de la méthode de la régression aux facteurs mutationnels et aux hétérogénéités spatiales et temporelles des paramètres démographiques peut s'interpréter à l'aide des probabilités de coalescence, comme on l'a vu dans les sections 1.2.3 et 2.1. On a vu en effet que les F-statistiques pouvaient être déduites des différences entre les distributions de probabilités de coalescence de

différentes paires de gènes (Figure 1.5, 1.6 et 2.3). Dans le cas du F_{ST}, ou de a_{r}, ces distributions diffèrent essentiellement par un excès de probabilité de coalescence pour les gènes les plus apparentés ($C_1(t)$ sur la Figure 2.3). Pour les modèles d'isolement par la distance avec une forte migration, cet excès de probabilité est concentré sur une courte période τ dans le passé récent. Comme la sensibilité des F_{ST}, ou de a_{r}, aux processus mutationnels et démographiques dépend de la durée de cette période τ, ces processus devraient avoir d'autant moins d'influence que τ est petit. Or comme on l'a vu en section 2.1, cette période de passé récent est d'autant plus courte que les taux de migration sont forts, et dans une moindre mesure les tailles de dème petites. Il n'est donc pas étonnant que, sous les modèles IBD individuel en habitat continue pour lesquels les taux de migrations sont forts (c.a.d. de l'ordre de 50%) et les tailles de dème faibles (c.a.d. un individu par nœud du réseau), l'influence des processus mutationnels et des fluctuations démographiques passées sur l'estimation de paramètres démographiques par F-statistiques soit limitée. A l'inverse, sous le modèle classique en îles avec des grandes populations et des taux de migrations faibles, l'effet des mutations et des fluctuations démographiques doivent être plus problématiques, ce qui a été largement vérifié par ailleurs (Boileau *et al.*, 1992). De plus, puisque l'on s'intéresse à la différenciation à petite échelle géographique, ces effets seront d'autant plus faibles. En effet, comme on l'a vu sur la Figure 2.3, plus les gènes comparés sont distants géographiquement, plus la période τ s'étend dans le passé (Slatkin, 1994; Rousset, 2004).

Le même type de raisonnement peut être utilisé pour comprendre pourquoi la méthode donne des estimations de densité correspondant plutôt à la densité locale sur la zone échantillonnée qu'à la densité autour de l'échantillon. Puisque la période τ s'étend peu dans le passé, les F-statistiques, F_{ST} ou a_{r}, dépendent principalement des événements de coalescence, de migration et/ou de mutation ayant eu lieu dans le passé récent et à une échelle géographique locale, puisque la dispersion est localisée dans l'espace. Par conséquent, l'estimation de $D\sigma^2$ avec la méthode étudiée ici correspond à sa valeur locale sur la zone échantillonnée et devrait être peu influencée par des caractéristiques démographiques de zones situées géographiquement éloignées de la zone échantillonnée.

2.2.5 Tests sur données réelles

Les résultats des tests part simulation présentés ci-dessus suggèrent donc une très bonne robustesse générale de inférence de $D\sigma^2$ vis à vis des processus mutationnels et vis à vis d'hétérogénéités spatiales et temporelles des paramètres démographiques, sauf pour une augmentation récente de la densité. Cependant, les études par simulation, bien qu'ayant clairement démontré leur utilité ci-dessus, ne prennent jamais en compte tous les facteurs pouvant influencer l'inférence à partir de données réelles. Enfin, bien que suggérées par l'analyse de différentes études empiriques, les situations démographiques et les valeurs de paramètres choisies pour les simulations présenté ci-dessus ne reflète peut être pas bien la réalité biologique.

Pour tester une méthode d'inférence sur des données réelles, il faut pouvoir comparer les estimations obtenues avec des attendus indépendants des données traitées, ici les données génétiques. De nombreuses espèces ont fait l'objet d'étude démographiques directes (c.a.d. avec des outils de démographie), et les techniques de capture-marquage-recapture peuvent permettre d'estimer des densité et des distributions de dispersion (comme nous l'avons vu en section 1.3.2 sur les libellules demoiselles *Coenagrion mercuriale* dans l'étude de Watts *et al.* (2006)) à partir desquelles on peut

calculer σ^2. François Rousset a donc compilé (ou participé à) entre 1996 et 2006 un ensemble d'étude pour lesquelles il y avait, sur une même zone d'étude, (i) des données génétiques adaptées à l'inférence par la méthode de la régression et (ii) des données démographiques permettant une bonne estimation de la densité et de la distribution de dispersion.

Le plus bel exemple d'une telle comparaison entre estimations indirectes (génétiques, Figure 2.2) et directes (démographiques) est présenté dans l'étude de Watts *et al.* (2006). Le lecteur s'y référera pour y trouver tous les détails croustillants, notamment la définition d'une nouvelle F-statistique e_{r} (voir aussi section 4.2), une alternative à a_{r}, montrant un léger biais mais une moindre variance que a_{r} et plus performante dans des situations de forte dispersion (faible IBD). La conclusion principale de cette étude nous intéressant ici est que l'on observe une très bonne congruence entre les estimations de $D\sigma^2$ par la méthode de la régression et celles par capture-marquage-recapture ($D\sigma^2 = 222, 259, 753$ par la régression vs. 277, 249, 555 par la démographie, sur trois sites différents). Toutes les comparaisons sur les autres jeux de données ont aussi montré que les estimations par la méthode de la régression était cohérentes (c.a.d. équivalentes à un facteur 2 près) avec les estimations démographiques. Ainsi, des valeurs de $\hat{D\sigma^2}$ de 2.6 (indirecte) vs. 1.4 (directe) ont été estimée chez les rongeurs *Dipodomys* (Rousset, 2000; Winters & Waser, 2003), $\hat{D\sigma^2} = 21$ (indirecte) vs. 29 (directe) dans des populations humaines nomades de Papouasie - Nouvelle Guinée (Wood *et al.*, 1985; Rousset, 1997), $\hat{D\sigma^2} = 14$ (indirecte) vs. 10 (directe) pour la légumineuse *Chamaecrista fasciculata* (Fenster *et al.*, 2003), $\hat{D\sigma^2} = 3.8$ (indirecte) vs. 7.5 (directe) pour la martre d'Amérique *Martes americana* (Broquet *et al.*, 2006) et $\hat{D\sigma^2} = 5.5$ (indirecte) vs. 11.5 (directe) pour les scinques de forêt *Gnypetoscincus queenslandiae* (Sumner *et al.*, 2001).

Ces applications sur données réelles, et bien d'autres, ont tout d'abord permis d'avoir une première idée de valeurs de $D\sigma^2$ raisonnables en populations naturelles (bien qu'elles pouvaient être remises en cause à l'époque...), et notamment que ces valeurs sont généralement faibles voir très faible chez de nombreux organismes. Il est d'ailleurs assez étonnant de voir que la plupart des analyses portant sur des organismes pour lesquels on pensait a priori que les densité de population et la dispersion étaient fortes donne des estimations de $D\sigma^2$ faibles. Deux exemples assez caricaturaux sont : (i) la processionnaire du pin, avec un $\hat{D\sigma^2}$ compris entre 2.5 et 6 à petite échelle dans des populations portugaises à l'"équilibre" dans un milieu favorable de type pinèdes (Burban *et al.*, 2004) ; et (ii) la souris domestique, avec un $\hat{D\sigma^2}$ compris entre 5 et 7.4 entre villages au nord du Sénégal (Lippens *et al.*, 2017). Des estimations de $D\sigma^2$ plus grandes ont toutefois été trouvées chez les demoiselles *Coenagrion mercuriale* comme illustré précédemment ($\hat{D\sigma^2}$ autour de 200-600), des coléoptères méligèthe du colza *Brassicogethes aeneus* ($\hat{D\sigma^2}$ autour de 50-100, Juhel *et al.*, 2019) ou sur des chauve-souris grand rhinolophe *Rhinolophus ferrumequinum* ($\hat{D\sigma^2}$ autour de 20-30, Tournayre *et al.*, 2019).

Mais c'est surtout la première fois en génétiques des populations que l'on a une si bonne concordance systématique entre des estimations de paramètres démographiques à partir de données génétiques d'une part et à partir de données démographiques d'autre part et je trouve que cela à des implications très importantes...

2.3 Validation des modèles IBD et limites de la méthode de la régression

Tous ces tests de la méthode de la régression, par simulation et par comparaison avec des données démographiques, valident assez clairement le fait que cette méthode estime de façon robuste et précise les valeurs actuelles et locales du produit $D\sigma^2$ à partir de données génétiques géo-référencées. Comme nous l'avons vu tout au long de ce chapitre, cette robustesse provient de la combinaison : (i) d'un modèle (relativement) plus réaliste que ceux considérés par ailleurs, qui d'une part modélise mieux la dispersion (que le modèle en îles par exemple) grâce à la prise en compte d'une distribution de dispersion, et qui d'autre part grâce à sa version IBD individuel en habitat continu ne nécessite pas de définir des (sou-)populations qui souvent ne correspondent à aucune réalité biologique ; et (ii) d'une méthode bien pensée, précise et robuste, du fait que l'augmentation de la différenciation avec la distance est plus facile à interpréter en terme de paramètre démographique que les valeurs des F-statistiques en elles-mêmes, qu'elle fait peu d'hypothèse sur la forme de la distribution de dispersion et donc convient aux distribution fortement leptokurtiques souvent observées en population naturelles, et qu'elle "force" l'utilisateur (avisé...) à analyser un échantillon récolté à une petite échelle géographique et limite ainsi l'influence des hétérogénéités spatiales et temporelles des paramètres démographiques.

Tous ces tests valident donc non seulement la méthode d'inférence et les hypothèses sur lesquelles elle repose mais ils valident aussi plus globalement la pertinence du modèle d'IBD pour modéliser les processus évolutifs neutres de dérive et de dispersion à un niveau local en population naturelle. Les modèles IBD semblent donc être de bons modèles pour étudier l'effet de la structuration des populations à petite échelle notamment quand, comme chez de nombreuses espèces, la dispersion est limitée dans l'espace ; et la méthode de la régression une bonne méthode pour estimer $D\sigma^2$ sous ces modèles. Tout cela invalide, en partie puisque seulement pour le cas de l'IBD et de la méthode de la régression, les nombreuses critiques ayant été faites sur l'inférence de paramètres démographiques à partir de données génétiques, et nous a permis d'en comprendre les raisons grâce aux raisonnements basés sur la coalescence.

La méthode de la régression ainsi que les modèles IBD ont toutefois de nombreuses limitations. Du point de vue de l'inférence, les trois principales limites sont : (i) cette approche ne permet que l'estimation du produit $D\sigma^2$ qui est un paramètre important pour l'évolution spatiale des populations, mais qui n'a pas un sens biologique immédiat pour les biologistes ; (ii) par définition puisque c'est une méthode des moments, la méthode de la régression n'utilise que l'information comprise dans les estimateurs de $F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}})$ ou a_{r}, qui elle même est résumée par la pente de la régression de ces estimateurs avec la distance. Elle n'utilise donc pas du tout l'information complète "contenue" dans les données ; (iii) elle ne peut considérer d'un modèle d'IBD homogène (sans hétérogénéités de densité et de dispersion, sans effets de bords) et à l'équilibre (du fait de la résolution à l'équilibre du système d'équations de récurrences des des probabilités d'identité).

Il serait tout d'abord intéressant de pouvoir estimer séparément D etσ^2, mais aussi d'estimer d'autres paramètres de la distribution de dispersion tels que la distance moyenne et/ou "maximale" de dispersion, sa forme (globalement pouvoir estimer plus de moments de cette distribution), ainsi que le taux d'émigration m. Il serait aussi intéressant de pouvoir petit à petit considérer des modèles plus com-

plexes, par exemple avec des variations spatiales de la densité pour se rapprocher de la "génétique de paysage" et/ou des changements temporels de tailles d'habitat pour pouvoir inférer des changements démographiques récents et locaux des populations naturelles menacées, invasives, d'intérêt agronomique ou médical. Ce dernier point est notamment crucial car plusieurs études ont démontré l'influence très forte de la structuration des populations sur l'inférence des changements passé de tailles de populations dans le cadre d'un modèle en île (Chikhi *et al.*, 2010) mais aussi d'IBD (Leblois *et al.*, 2006, 2014). Il serait donc pertinent de développer des méthodes d'inférence combinée de la structuration et des variations passées des tailles de populations, et le modèle d'IBD semble bien adapté puisqu'il peut considérer une large gamme de dispersion, du stepping-stone au modèle en îles (dans le cas d'une dispersion géométrique par exemple, ces modèles sont obtenus avec $g = 0$ et $g = 1$, respectivement). Tous ces développements sont théoriquement possible dans le cadre de l'inférence par vraisemblance ou basée sur la simulation. C'est ce que nous verrons dans les deux chapitres suivants.

Chapitre 3

Inférences par vraisemblance sous IBD

Nous avons vu dans le chapitre précédent qu'une méthode des moments, la méthode de la régression de Rousset (1997, 2000), permet d'estimer le produit $D\sigma^2$, ou D est la densité de la population et σ^2 le carré moyen de la dispersion parent-descendant, dans un modèle d'isolement par la distance (IBD). Des tests par simulation et sur données réelles ont permis de valider le fait que cet inférence par la méthode de la régression est précise et robuste vis à vis des processus mutationnels et de processus démographiques autre que la dérive et la dispersion telles que des hétérogénéités spatiales et temporelles de ces paramètres de densité et dispersion. Ces tests ont donc montré que la méthode de la régression estime bien les paramètres locaux et actuels des populations étudiées.

Enfin, les comparaisons sur différentes jeux de données avec des estimations démographiques indépendantes, montrant systématiquement une bonne concordance entre les approches directes et indirectes, suggère fortement que les modèles IBD sont de bons modèles pour inférer les paramètres de densité et de dispersion, locaux et actuels, en populations naturelles.

Pour aller au delà des limites de cette approches d'inférence par moment tout en profitant des caractéristiques pertinentes des modèles IBD , il parait intéressant de tester ce que peuvent permettre les inférences par vraisemblance, qui théoriquement devraient être beaucoup plus puissante puisqu'elle utilisent toute l'information "contenue" dans les données, dans ce contexte d'isolement par la distance. Dans ce chapitre, nous montrerons tout d'abord comment la théorie de la coalescence et le coalescent structuré ont permis le développement de deux algorithmes de Monte Carlo pour estimer la vraisemblance d'une échantillon génétique par simulation. Nous présenterons ensuite quelques tests d'une méthode implémentant ces premiers algorithmes puis des tests plus poussé de notre implémentation des second algorithmes, principalement sous des modèles d'isolement par la distance (IBD).

3.1 Méthodes permettant le calcul de la vraisemblance basées sur la coalescence

On peut distinguer deux approches pour l'estimation de paramètres démographiques par maximum de vraisemblance basées sur la coalescence.

L'une, initialement développée par Wilson & Balding (1998), Felsenstein et collaborateurs (Felsenstein *et al.*, 1999) et Nielsen (2000), utilise le découplage des

processus démographiques et mutationnels (voir section 1.1), pour calculer la vraisemblance d'un échantillon en intégrant sur les différentes généalogies (arbres de coalescence) et histoires de mutations possibles. Les généalogies, comme les paramètres, sont alors explorées selon un algorithme de Metropolis-Hastings par chaînes de Markov en passant d'une généalogie à l'autre, ou d'une valeur de paramètre à une autre, en les modifiant légèrement. On les appelle classiquement, en génétique des populations, les approches MCMC ("Monte Carlo Markov Chain") par coalescence, que l'on notera ici MCMC-coa, basées sur l'algorithme dit de "pruning" de Felsenstein (2004). Relativement facile à adapter à différents modèles démographiques et mutationnels, cette approche a été à la source de nombreuses méthodes d'inférence telles que la suite LAMARC 2.0 (Kuhner, 2006) pour estimer des taux de croissance de population, taux de recombinaison, ou tailles de sous-populations et taux de migration (sous-programme MIGRATE, Beerli & Felsenstein (1999, 2001)), le logiciel Msvar (Beaumont, 1999; Storz & Beaumont, 2002; Beaumont, 2003), IM et ses dérivés (Nielsen & Wakeley, 2001; Hey & Nielsen, 2004, 2007; Hey, 2010), qui ont elles-même été énormément utilisées dans les années 95-2015, notamment sur données microsatellites.

L'autre approche, initiée par Griffiths et collaborateurs (Griffiths & Tavaré, 1994a,b; Bahlo & Griffiths, 2000; de Iorio & Griffiths, 2004a,b) utilise des algorithmes fondés sur l'échantillonnage pondéré, ou échantillonage d'importance, (IS, "importance sampling") pour (1) retracer par chaînes de Markov absorbantes (c.a.d avec un état final absorbant, ici le MRCA de l'arbre de coalescence) les histoires ancestrales (généalogie et mutations) possibles d'une échantillon grâce à des récurrences sur les tous états ancestraux possible de l'échantillon considéré et (2) en même temps d'en estimer la vraisemblance de l'échantillon grâce aux probabilités de transition des récurrences.

Les principes de ces deux types d'algorithmes, notamment celui de Griffiths et collaborateurs, étant très complexe à décrire, je ne ferai ici qu'une présentation générale assez succincte permettant de comprendre les tenant et aboutissant des inférences décrites ensuite, mais le lecteur pourra chercher les détails dans les synthèse de Kuhner (2009) pour les approches par MCMC-coa, et Rousset *et al.* (2018) ou, à ses risques et périls, dans les articles originaux de Griffiths et collaborateurs (Griffiths & Tavaré, 1994a; de Iorio & Griffiths, 2004a), pour l'IS-coa.

3.1.1 Approche de Felsenstein et collaborateurs

Beerli & Felsenstein (1999, 2001) ont utilisé un algorithme de Monte Carlo par chaînes de Markov pour estimer la vraisemblance $\mathcal{L}(\mathcal{P};\mathcal{D})$ dans la cas d'une population structurée, en passant donc par

$$\begin{aligned}\mathcal{L}(\mathcal{P};D) &= \int_G \frac{\Pr(D|G;\mathcal{P})\Pr(G;\mathcal{P})}{f(G)} f(G) \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \frac{\Pr(D|G_i;\mathcal{P})\Pr(G_i;\mathcal{P})}{f(G_i)}.\end{aligned} \tag{3.1}$$

ou l'on explore un grand nombre de généalogies selon la fonction d'échantillonnage d'importance f. Puisque, sous un modèle neutre, la généalogie ne dépend pas des paramètres mutationnels $\mathcal{M}$ mais uniquement des paramètres démographiques $\mathcal{D}$

(voir section 1.1), on a

$$\hat{\mathcal{L}}(\mathcal{P};D) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(D|G_i;\mathcal{M})\Pr(G_i;\mathcal{D})}{f(G_i)}, \tag{3.2}$$

avec $\mathcal{P}=(\mathcal{D},\mathcal{M})$. Beerli & Felsenstein (1999, 2001), comme la plupart des développements basés sur cette approche, ont utilisé la fonction d'échantillonnage d'importance suivante

$$f_{BF}(G) \equiv \frac{\Pr(G;\mathcal{D}_0)\Pr(D|G;\mathcal{M})}{\mathcal{L}(\mathcal{P}_0;D)} \tag{3.3}$$

pour un ensemble $\mathcal{D}_0$ de valeurs des paramètres démographiques, avec $\mathcal{P}_0=(\mathcal{D}_0,\mathcal{M})$. En remplaçant la fonction d'échantillonnage d'importance dans l'équation (3.2), on a alors

$$\frac{\mathcal{L}(\mathcal{P};D)}{\mathcal{L}(\mathcal{P}_0;D)} \simeq \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(D|G_i;\mathcal{M})\Pr(G_i;\mathcal{D})}{\Pr(G_i;\mathcal{D}_0)\Pr(D|G_i;\mathcal{M})} = \frac{1}{n}\sum_{i=1}^{n}\frac{\Pr(G_i;\mathcal{D})}{\Pr(G_i;\mathcal{D}_0)}, \tag{3.4}$$

où les généalogies G_i sont générées par une chaîne de Markov de distribution stationnaire f_{BF} avec les paramètres $\mathcal{P}_0$. L'équation (3.4) permet alors l'estimation du rapport des vraisemblances pour un ensemble de valeurs de $\mathcal{P}$ autour de $\mathcal{P}_0$ à partir de la réalisation d'une chaîne de Markov unique. Il suffit alors de trouver le jeu de paramètres $\mathcal{P}_{MLE}$ qui maximise ce ratio de vraisemblance ou, plus généralement, d'échantillonner dans la distribution stationnaire de la chaîne de Markov pour en déduire les distributions a posteriori des paramètres. Nous allons maintenant voir rapidement comment on calcule la probabilité $\Pr(G_i;\mathcal{D})$ d'une généalogie connaissant les valeurs des paramètres démographiques puis comment la chaîne de Markov d'échantillonnage des généalogies est implémentée dans leurs algorithmes.

Probabilité d'une généalogie sachant les paramètres

Sous un coalescent structuré comme défini en section 1.1.2, la généalogie d'un échantillon de gènes est alors entièrement définie par (i) la séquence d'événements de coalescence ou de migration (arrière) des lignées ancestrales, (ii) les intervalles de temps séparant ces différents événements et (iii) les lignées étant concernées par ces événements (Figure 1.2). On peut alors exprimer la probabilité d'une généalogie comme $\Pr(G;\mathcal{D}) = \Pr(\text{intervalles de temps entre événements de G};\mathcal{D}) \times \Pr(\text{séquence de coa-}$
lescences ou de migrations; $\mathcal{D}$). Comme on l'a vu dans le cadre du coalescent structuré, si l'on considère que les tailles de populations sont grandes et les taux de migration petits, les temps d'attente u_τ entre deux événements en remontant dans le temps suivent une loi exponentielle de paramètres le taux d'événements global Λ_τ, somme des taux de chaque événement possible au temps τ. On a donc

$$\Pr(\{u_1,..,u_T\};\mathcal{D}) = \prod_{\tau=1}^{T}\Lambda_\tau e^{\left(\Lambda_\tau \times u_\tau\right)} \tag{3.5}$$

avec

$$\Lambda_\tau = \sum_{i=1}^{n_d}\left[\frac{n_{\tau i}(n_{\tau i}-1)}{4N_i} + \sum_{k=1;k\neq i}^{n_d} n_{\tau i} b_{ki}\right], \tag{3.6}$$

et $n_{\tau i}$ les nombre de lignées ancestrales présentes dans le dème i pendant l'intervalle de temps τ et b_{ij} les taux de migration arrières du dème i vers le dème j. Par ailleurs les probabilités des événements de coalescence et de migration sont identiques à ceux présentés dans le cadre du coalescent structuré, on a donc

$$\Pr(\mathrm{coa}_\tau; \mathcal{D}) = \frac{\sum_{i=1}^{n_\mathrm{d}} \frac{n_{\tau i}(n_{\tau i}-1)}{4N_i}}{\sum_{i=1}^{n_\mathrm{d}} (\frac{n_{\tau i}(n_{\tau i}-1)}{4N_i} + \sum_{k=1;k\neq i}^{n_\mathrm{d}} n_{\tau i} b_{ki})}, \tag{3.7}$$

et

$$\Pr(\mathrm{mig}_\tau; \mathcal{D}) = \frac{\sum_{i=1}^{n_\mathrm{d}} \sum_{k=1;k\neq i}^{n_\mathrm{d}} n_{\tau i} b_{ki}}{\sum_{i=1}^{n_\mathrm{d}} (\frac{n_{\tau i}(n_{\tau i}-1)}{4N_i} + \sum_{k=1;k\neq i}^{n_\mathrm{d}} n_{\tau i} b_{ki})}. \tag{3.8}$$

Enfin, la probabilité qu'une lignée concernée par un événement de migration ou de coalescence appartienne à une sous-population donnée est proportionnelle au nombre de lignées présentes dans la sous-population dans l'intervalle de temps τ. La probabilité qu'une coalescence donnée se produise dans la sous-population v_τ à la fin de l'intervalle de temps τ est donc

$$\Pr(v_\tau; \mathcal{D}) = \frac{n_{\tau v_\tau}(n_{\tau v_\tau} - 1)/4N_{v_\tau}}{\sum_{i=1}^{n_\mathrm{d}} n_{\tau i}(n_{\tau i} - 1)/4N_i} \frac{2}{n_{\tau v_\tau}(n_{\tau v_\tau} - 1)}. \tag{3.9}$$

Pour un événement de migration "arrière" de la sous-population w_τ vers la sous-population v_τ, la probabilité est

$$\Pr(v_\tau, w_\tau; \mathcal{D}) = \frac{n_{\tau v_\tau} b_{w_\tau v_\tau}}{\sum_{i=1}^{n_\mathrm{d}} \sum_{k=1;k\neq i}^{n_\mathrm{d}} n_{\tau i} b_{ki}} \frac{1}{n_{\tau v_\tau}}. \tag{3.10}$$

En posant $u'_\tau \equiv \mu u_\tau$, $\theta_{v_\tau} \equiv 4N_{v_\tau}\mu$, $b'_{ij} \equiv b_{ij}/\mu$ où μ est le taux de mutation des marqueurs, la probabilité d'une généalogie G sachant les paramètres démographiques $\mathcal{D}$ est alors

$$\Pr(G; \mathcal{D}) = \prod_{\tau=1}^{T} \left[(\delta_\tau b'_{w_\tau v_\tau} + (1-\delta_\tau)\frac{2}{\theta_{v_\tau}}) \exp\left(-u'_\tau \sum_{i=1}^{n_\mathrm{d}} (\frac{n_{\tau i}(n_{\tau i}-1)}{\theta_{v_i}} + n_{\tau i} \sum_{k=1;k\neq i}^{n_\mathrm{d}} b'_{ki}) \right) \right], \tag{3.11}$$

où δ_τ est une variable indicatrice prenant la valeur 1 si on a une migration ou 0 si on a une coalescence.

Échantillonnage des généalogies par chaîne de Markov

La topologie de la première généalogie est construite à partir de l'échantillon par la méthode UPGMA ("Unweighted Pair Group Method with Arithmetic Averages", voir Swofford *et al.*, 1996) puis la méthode par parcimonie de Sankoff (1975) permet d'ajouter le nombre minimal d'événements de migration sur cette topologie. Nous considérerons pour la suite qu'une généalogie contient l'information sur les événements de migration ayant affecté le différentes lignées.

Pour explorer les différentes généalogies compatibles avec l'échantillon, Felsenstein et collaborateurs ont utilisé des *chaînes de Markov*. Le terme chaîne de Markov traduit le fait que les généalogies vont être échantillonnées par transition d'une généalogie G_i à une autre G_{i+1} selon des probabilités de transition $\Pr(G_{i+1}|G_i)$ dépendant uniquement de ces deux généalogies. Une nouvelle généalogie G_{i+1} est créée à partir de la généalogie G_i par délétion et reconstruction d'une partie de cette généalogie en fonction des paramètres démographiques $\mathcal{D}$ du modèle. L'acceptation

de la nouvelle généalogie est faite selon un algorithme de Metropolis-Hastings (Hastings, 1970), dont je ne donnerai pas les détails ici (voir Beerli & Felsenstein, 1999). En d'autres mots, l'algorithme de Metropolis-Hastings va donner les probabilités de transition entre les différentes généalogies explorées, calculées de telle manière à ce que l'espace soit exploré selon une fonction d'échantillonnage voulue, ici la fonction d'échantillonnage d'importance $f_{BF}(G)$.

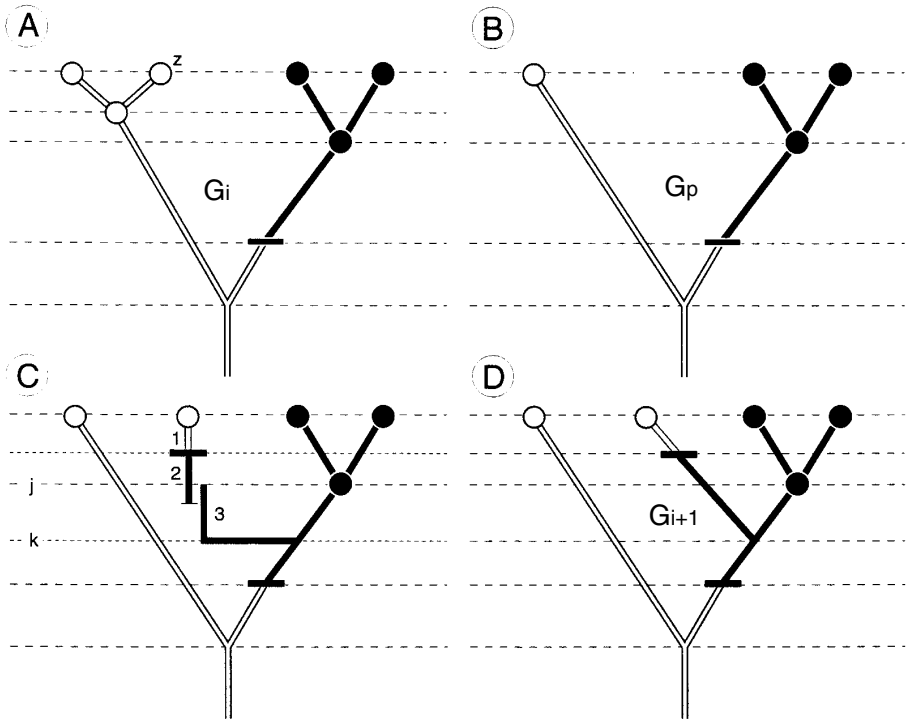


FIGURE 3.1 – Illustration du principe de délétion-construction d'une généalogie selon l'algorithme de Felsenstein et collaborateurs. Les lignées noires et blanches appartiennent à différentes sous-populations. Voir le texte pour les détails. D'après Beerli & Felsenstein, 1999.

La construction de la nouvelle généalogie G_{i+1} à partir de G_i se fait en quatre étapes : (i) un nœud de l'arbre (correspondant à une coalescence ou un gène de l'échantillon de départ) est choisi au hasard sur la généalogie (le nœud z sur la Figure 3.1) ; (ii) la lignée ancestrale de ce nœud est effacée pour obtenir une généalogie partielle G_p (étape B de la Figure 3.1) ; (iii) un nouvel intervalle de temps u est calculé, selon les probabilité des différents événements possibles considérée dans l'équation (3.6) mais conditionnellement au fait que la lignée ancestrale du nœud z est impliquée dans l'événement. Cet intervalle de temps u est donné par

$$u = -\frac{\ln(p_\tau)}{\sum_k b'_{ik} + \frac{2n'_{\tau i}}{\theta_i}}, \tag{3.12}$$

où $n'_{\tau i}$ est le nombre de lignées dans la sous-population i, qui contenait le nœud z (c.a.d les lignées blanches sur la Figure 3.1), pendant l'intervalle de temps τ de la généalogie partielle G_p (qui ne contient plus le nœud z), et p_τ est une variable aléatoire tirée dans une loi uniforme [0,1] (étape C de la Figure 3.1, "1" correspond au nouvel intervalle de temps). On choisit ensuite un nouvel événement pour cet intervalle de temps selon la probabilité relative des événements. Par exemple pour une migration arrière de i vers l, la probabilité de cet événement est

$$\Pr(\text{migration de } i \text{ vers } l\ ) = \frac{b'_{il}}{\sum_k b'_{ik} + \frac{2n'_{\tau i}}{\theta_i}}. \tag{3.13}$$

Ces probabilités correspondent aussi aux probabilités des différents événements possibles considérés dans l'équation (3.6) mais conditionnellement au fait que la lignée ancestrale du nœud z est impliquée dans l'événement. Si c'est une migration, cet exemple est illustré sur la figure 3.1C, la lignée change de population. Si c'est une nouvelle coalescence, la lignée coalesce alors avec une autre lignée de la même sous-population et une nouvelle généalogie G_{i+1} est ainsi formée. Dans le cas de la migration, des nouveaux intervalles de temps sont calculés (éq.3.12) pour finir avec une coalescence dans la bonne population (c.a.d la population de la nouvelle lignée ancestrale du nœud z, Figure 3.1D).

La nouvelle généalogie G_{i+1} est acceptée avec la probabilité de Metropolis-Hastings

$$r = \min\left(1, \frac{\Pr(D|G_{i+1})}{\Pr(D|G_i)}\right). \tag{3.14}$$

L'utilisation du critère de Metropolis-Hastings garantit le fait que la chaîne de Markov ait f_{BF} comme distribution stationnaire. L'ensemble de cette méthode a été implémentée dans le logiciel MIGRATE (Beerli & Felsenstein, 1999, 2001). Quelques tests de ce logiciel seront présentés ci-dessous dans la section 3.2.1.

3.1.2 L'approche de Griffiths et collaborateurs

La formulation de cette approche n'est pas triviale et j'essaierai ici d'en donner une explication la plus simple possible, dérivée de la reformulation de Rousset *et al.* (2018). Une description alternative de ces algorithmes, plus proche de la description originale de Griffiths et collaborateurs, pourra être trouvé dans ma thèse (Leblois, 2004), et pourra aider le lecteur dans la compréhension globale des algorithmes. Les modèles démographiques et cycles de vie sont les mêmes que dans les sections précédentes.

Formulation de l'échantillonnage d'importance séquentiel sur l'histoire ancestrale de l'échantillon

Étant donné un échantillon actuel $\mathbf{S}$, nous considérerons les états ancestraux (c'est-à-dire les nombres et types génétique/allélique) des lignées ancestrales de $\mathbf{S}$ à tout moment t, appelé "échantillon ancestral", $\mathbf{S}(t)$, du présent $t = 0$ jusqu'au moment t_τ du MRCA de l'échantillon. Nous nous intéresserons spécialement aux probabilités de transition $\hat{p}$ entre états ancestraux $\mathbf{S}(t)$ et $\mathbf{S}(t')$, ainsi qu'aux poids d'échantillonnage d'importance $\hat{w}$ définis de telle sorte que la vraisemblance d'un échantillon, notée ici $q(\mathbf{S})$ à la place de $\mathcal{L}(D)$ comme précédemment, puisse être écrite sous la forme suivante

$$q(\mathbf{S}) = \mathrm{E}_{\hat{p}}\left(\prod_{k=0}^{\tau} \hat{w}[\mathbf{S}(t_k)]\right), \tag{3.15}$$

où l'espérance est prise sur la distribution de la séquence de $(\mathbf{S}(t_k))$ des échantillons ancestraux générés par les probabilités de transition $\hat{p}$. Ces probabilités de transition définissent une chaîne de Markov sur les états ancestraux de l'échantillon actuel jusqu'à l'état absorbant MRCA atteint au moment t_τ. Chaque réalisation de cette chaîne de Markov enregistre une séquence d'événements de coalescence, de mutation et de migration jusqu'à ce que l'ancêtre commun soit atteint. L'estimation de $q(\mathbf{S})$ est alors réalisée en calculant la moyenne des $\prod_{t=0}^{\tau} \hat{w}[\mathbf{S}(t)]$ sur un grand nombre de réalisations indépendantes de cette chaîne de Markov. Le nombre de généalogies

indépendantes à explorer pour avoir une estimation correcte de la vraisemblance est un point crucial de cet approche. Nous verrons que ce nombre peut varier fortement entre les différentes versions des algorithmes IS-coa et des modèles mutationnels et démographiques considéré, typiquement de 1 à 20,000. Comme nous le verrons plus en détail par la suite, Les premières versions de Griffiths & Tavaré (1994a); Nath & Griffiths (1996); Bahlo & Griffiths (2000) était très peu efficaces et nécessitaient d'explorer près de 100,000 généalogies, alors que les nouveaux algorithmes de Stephens & Donnelly (2000) puis de de Iorio & Griffiths (2004a) sont nettement plus efficaces et nécessitent 100 à 100,000 fois moins de simulations !

de Iorio & Griffiths (2004a,b) proposent $\hat{p}$ et $\hat{w}$ sur la base d'approximations du rapport $\pi \equiv q(\mathbf{S})/q(\mathbf{S}')$ des probabilités d'échantillons différant par un événement (mutation, migration ou coalescence). Nous allons détailler comment ces approximations sont construites mais tout d'abord, nous examinerons les récurrences sur un intervalle de temps, reliant l'échantillon actuel à un échantillon ancestral $\mathbf{S}(t)$ considéré un évènement auparavant.

Ces récurrences sont obtenues par un raisonnement par coalescence. En d'autres termes, nous représentons les événements qui ont conduit à l'échantillon actuel de n gènes comme les réalisations de deux processus : un processus de coalescence déterminant la distribution marginale des généalogies ancestrales de n gènes, indépendamment des types alléliques actuels ; et, compte tenu d'une généalogie, un processus de mutation qui modifie les types alléliques le long des branches de l'arbre généalogique.

En effet, la relation entre la probabilité d'un échantillon actuel et la probabilité de l'échantillon parental ancestral peut être conçue comme la réalisation conjointe de deux processus : le processus généalogique marginal sur la dernière génération et le processus de mutation sur cette génération. Dans ce qui suit, nous considérons des échantillons provenant de populations subdivisées, où la taille de l'échantillon est définie comme un vecteur $\mathbf{n} = n_i$ de tailles d'échantillons dans des sous-populations distinctes, et les échantillons sont caractérisés par leur configuration, c'est à dire le nombre d'allèles différents dans chaque sous-population échantillonnée n_{ik} le nombre de gène de type k dans la sous-population i. Par exemple, l'échantillon $\mathbf{S} = ((0, 4, 5), (5, 4, 0))$ décrit les comptages des trois allèles parmi n = 18 individus échantillonnés dans deux sous-populations (n = (9,9)), le premier allèle ne se trouvant que dans la deuxième sous-population en 5 exemplaires, et ainsi de suite. La récurrence entre un échantillon actuel $\mathbf{S}'$ et tous les échantillons parentaux possibles $\mathbf{S}$ prend la forme suivante

$$q(\mathbf{S}') = \sum_{\mathbf{S}} \Pr(\mathbf{n}) q(\mathbf{S}) \Pr(\mathbf{S}'|\mathbf{S}), \tag{3.16}$$

où $q(\mathbf{S}') \equiv \Pr(\mathbf{S}'|\mathbf{n}')$ est la probabilité stationnaire de l'échantillon descendant $\mathbf{S}'$, étant donné la taille de l'échantillon descendant $\mathbf{n}'$; $q(\mathbf{S}) \equiv \Pr(\mathbf{S}|\mathbf{n})$ est la probabilité stationnaire de l'échantillon ancestral $\mathbf{S}$ étant donné sa taille $\mathbf{n}$; $\Pr(\mathbf{n})$=$\Pr(\mathbf{n}|\mathbf{n}')$ est la probabilité stationnaire que, sachant la taille de l'échantillon descendants $\mathbf{n}'$ (mais pas sachant $\mathbf{S}'$), les lignées parentales forment un échantillon de *ssize* gènes. Cette probabilité dépend de la probabilité stationnaire des événements de coalescence et de migration dans la dernière génération, mais l'occurrence des mutations ne modifie pas $\mathbf{n}$; et $\Pr(\mathbf{S}'|\mathbf{S})$=$\Pr(\mathbf{S}'|\mathbf{S}, \mathbf{n}')$ est la probabilité (étant donné $\mathbf{n}'$) que les événements de mutation ait conduit à l'échantillon descendant $\mathbf{S}'$ étant donné l'échantillon parental $\mathbf{S}$ et la taille de l'échantillon descendant $\mathbf{n}'$.

Cette récurrence suggère l'algorithme d'échantillonnage d'importance inefficace suivant. Nous réécrivons la récurrence en éliminant le cas où $\mathbf{S}' = \mathbf{S}$ sur la somme

de droite. L'équation résultante peut être écrite comme suit

$$q(\mathbf{S}') = \sum_{\mathbf{S}\neq\mathbf{S}'} \tilde{w}(\mathbf{S}')p(\mathbf{S}|\mathbf{S}')q(\mathbf{S}), \tag{3.17}$$

où

$$\tilde{w}(\mathbf{S}') \equiv \frac{\sum_{\mathbf{S}\neq\mathbf{S}'} \Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})}{1-\sum_{\mathbf{S}\neq\mathbf{S}'} \Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})} \tag{3.18}$$

et

$$\tilde{p}(\mathbf{S}|\mathbf{S}') \equiv \frac{\Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})}{\sum_{\mathbf{S}\neq\mathbf{S}'} \Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})}. \tag{3.19}$$

Les probabilités $\tilde{p}(\mathbf{S}|\mathbf{S}')$ définissent les probabilités de transition d'une chaîne de Markov telle que

$$q(\mathbf{S}) = \mathrm{E}_{\tilde{p}}\left(q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1}\tilde{w}[\mathbf{S}(t_k)]\right) \tag{3.20}$$

où $\mathbf{S}(0) = \mathbf{S}$ représente la configuration allélique de l'échantillon actuel, et $\mathbf{S}(\tau)$ le type allélique du MRCA. Ainsi, les $\tilde{w}$ (ou leur produit) sont des poids d'échantillonnage d'importance dans un algorithme d'échantillonnage d'importance séquentiel dont la distribution de proposition d'échantillonnage est la distribution des histoires ancestrales générées par $\tilde{p}$.

Une bonne paire (p, w) est telle que $q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1} w[\mathbf{S}(t_k)]$ a une faible variance sur les réalisations de p. La paire ci-dessus est inefficace à cet égard. Un algorithme IS optimal peut être défini comme produisant une variance nulle, et Stephens & Donnelly (2000) a caractérisé la paire optimale (p, w) en termes d'échantillons successifs et de leurs probabilités stationnaires. Pour obtenir un algorithme réalisable à partir de cette caractérisation, de Iorio & Griffiths (2004a,b) l'ont reformulée en termes de probabilités $\pi(j|d, \mathbf{S})$, pour tout j et d, qu'un gène supplémentaire prélevé dans les sous-populations d soit du type j. On peut alors définir des approximations pour l'optimum (p, w) à partir d'approximations pour les π.

p et w optimaux

Réécrivons

$$q(\mathbf{S}') = \sum_{\mathbf{S}\neq\mathbf{S}'} w(\mathbf{S}')p(\mathbf{S}|\mathbf{S}')q(\mathbf{S}) \tag{3.21}$$

comme

$$q(\mathbf{S}') = \sum_{\mathbf{S}\neq\mathbf{S}'} \hat{w}(\mathbf{S}', \mathbf{S})\hat{p}(\mathbf{S}|\mathbf{S}')q(\mathbf{S}) \tag{3.22}$$

pour certaines probabilités de transition $\hat{p}(\mathbf{S}|\mathbf{S}')$ formant une matrice de transition de Markov, et pour

$$\hat{w}(\mathbf{S}', \mathbf{S}) \equiv w(\mathbf{S}')\frac{p(\mathbf{S}|\mathbf{S}')}{\hat{p}(\mathbf{S}|\mathbf{S}')}. \tag{3.23}$$

Alors $q(\mathbf{S}) = \mathrm{E}_p\left(q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1} w[\mathbf{S}(t_k)]\right)$ devient
$q(\mathbf{S}) = \mathrm{E}_{\hat{p}}\left(q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1}\hat{w}[\mathbf{S}(t_k), \mathbf{S}(t_{k+1})]\right)$.
En considérant la chaîne de Markov définie par les probabilités de transition

$$\hat{p}(\mathbf{S}|\mathbf{S}') \equiv w(\mathbf{S}')p(\mathbf{S}|\mathbf{S}')\frac{q(\mathbf{S})}{q(\mathbf{S}')} \tag{3.24}$$

pour toute paire $(\mathbf{S}', \mathbf{S})$, alors $\hat{w}[\mathbf{S}(t_k), \mathbf{S}(t_{k+1})] = q[\mathbf{S}(t_{k+1})]/q[\mathbf{S}(t_k)]$ et toute réalisation de la chaîne de Markov sur les états ancestraux donne la vraisemblance exacte ("simulation parfaite") :

$$q(\mathbf{S}(t_\tau)) \prod_{k=0}^{\tau-1} \hat{w}[\mathbf{S}(t_k), \mathbf{S}(t_{k+1})] = q(\mathbf{S}(\tau)) \prod_{k=0}^{\tau-1} \frac{q[\mathbf{S}(t_{k+1})]}{q[\mathbf{S}(t_k)]} = q(\mathbf{S}(0)), \tag{3.25}$$

ce qui montre que $(\hat{p}, \hat{w})$ est optimal.

Formulation de p et w efficaces

Nous pouvons réécrire l'algorithme optimal d'échantillonnage d'importance en termes de probabilité $\pi(j|d, \mathbf{S})$ qu'un gène supplémentaire prélevé dans le dème d soit de type j (de sorte que la somme de tous les types possibles $\sum_j \pi(j|d, \mathbf{S}) = 1$). La probabilité stationnaire $q(\mathbf{S})$ s'écrit alors comme une espérance sur la distribution conjointe des fréquences X_{di} pour tous les allèles i dans toutes les sous-populations d,

$$q(\mathbf{S}) = \mathrm{E}\Big(\prod_d \binom{n_d}{(n_{di})} \prod_i X_{di}^{n_{di}} \Big) \tag{3.26}$$

Alors, pour tout d et j, $\pi(j|d, \mathbf{S})$ est liée aux probabilités de l'échantillon stationnaire par

$$\pi(j|d, \mathbf{S}) q(\mathbf{S}) = \mathrm{E}\left(X_{dj} \prod_d \binom{n_d}{(n_{di})} \prod_i X_{di}^{n_{di}} \right) = \frac{n_{dj} + 1}{n_d + 1} q(\mathbf{S} + \mathbf{e}_{dj}) \tag{3.27}$$

où l'espérance est prise sur la densité stationnaire des fréquences conjointes des allèles $\mathbf{x}$ dans les différents dèmes considérés. Ainsi, si deux échantillons successifs diffèrent par l'ajout d'une copie de gène de type j dans le dème d, le terme correspondant $\hat{w}[\mathbf{S}(t_k), \mathbf{S}(t_{k+1})]$ dans l'équation 3.25 peut être écrit comme suit

$$\pi(j|d, \mathbf{S}(t)) \frac{n_d(t) + 1}{n_{dj}(t) + 1} = \pi(j|d, \mathbf{S}(t_{k+1})) \frac{n_d(t_{k+1})}{n_{dj}(t_{k+1})}. \tag{3.28}$$

Si deux échantillons successifs diffèrent par une mutation de i à j dans le dème d, alors

$$\hat{w}[\mathbf{S}(t), \mathbf{S}(t_{k+1})] = \frac{\pi(j|d, \mathbf{S}(t_{k+1}))}{\pi(i|d, \mathbf{S}(t_{k+1}))} \frac{n_{di}(t_{k+1}) + 1}{n_{dj}(t_{k+1})}, \tag{3.29}$$

car la mutation peut être représentée comme l'élimination d'une copie de gène et l'ajout d'une autre copie de gène d'un autre type dans le même dème. De même, une migration du dème d vers le dème d' donne

$$\hat{w}[\mathbf{S}(t), \mathbf{S}(t_{k+1})] = \frac{\pi(j|d', \mathbf{S}(t_{k+1}))}{\pi(j|d, \mathbf{S}(t-1))} \frac{n_{d'}(t_{k+1}) n_{dj}(t_{k+1}) + 1}{(n_d(t_{k+1}) + 1) n_{d'j}(t_{k+1})}. \tag{3.30}$$

Les méthodes de coalescence considèrent généralement qu'un seul événement (coalescence, mutation ou migration) distingue les échantillons successifs. Ainsi, en termes informels, les taux de mutation et de migration sont supposés faibles et les tailles des sous-populations sont supposées importantes, de sorte qu'il est peu probable que plus d'un événement de coalescence se produise au cours d'une génération. Ensuite, le produit des poids séquentiels dans l'équation 3.25 peut être écrit, pour

toute séquence d'échantillons ancestraux, comme un produit des termes donnés dans les trois dernières équations. Toute approximation des π définit alors une approximation des poids optimaux dans un algorithme d'échantillonnage d'importance.

Une explication détaillée de l'approximation définie par de Iorio & Griffiths (2004a,b) pourra être trouvée dans l'annexe de Rousset *et al.* (2018). Cette approximation récupère les vrais π et permet donc une "simulation parfaite" dans quelques cas où la distribution stationnaire des fréquences des allèles dans les populations est connue, et elle est par ailleurs très efficace pour d'autres modèles stationnaires qui ont été étudiés (de Iorio *et al.*, 2005; Rousset & Leblois, 2007, 2012). Les arguments précédents permettent également d'obtenir des algorithmes IS pour les modèles non homogènes dans le temps où les taux d'événements dépendent des variations temporelles des valeurs des paramètres (Griffiths & S.Tavaré, 1994). L'approximation $\hat{\pi}$ de de Iorio & Griffiths (2004a,b) a été utilisée pour étendre la méthode d'inférence à des modèles dont la taille de la population varie dans le temps (Leblois *et al.*, 2014) et à des modèles avec des événements de divergence de population (divergence avec migration entre deux populations, travail non publié). Cependant, les $\hat{p}$ à tout pas de temps t ne prend en compte que les taux au temps t, et non les variations de taux plus ancestrales qui affectent aussi les probabilités de l'échantillon au temps t, ce qui entraîne une perte d'efficacité de l'algorithme IS. Des méthodes de ré-échantillonnage (Liu, 2004) ont été étudiées pour remédier à cette inefficacité (Merle *et al.*, 2017).

La méthode définie par de Iorio & Griffiths (2004a,b) fournit une approximation de la probabilité qu'un gène nouvellement échantillonné soit d'un type donné. Comme indiqué plus haut, cette approximation se réduit, dans le cadre d'un modèle de population stationnaire unique avec des mutations indépendantes des parents (PIM, c'est-à-dire lorsque le taux de mutation vers l'avant du type génétique i vers j est indépendant de i), à la véritable probabilité, et conduit donc à la distribution optimale d'échantillonnage d'importance, permettant une "simulation parfaite" dans le cadre d'un tel modèle. Dans les modèles stationnaires de populations structurées, cette approximation ne permet pas une simulation d'échantillonnage d'importance parfaite mais reste très efficace. Une grande partie du temps computationnel de ces méthodes provient du calcul, fait indépendamment pour chaque point de l'espace des paramètre, des termes $\hat{\pi}$ de de Iorio & Griffiths (2004b), et qui est nécessaire pour déterminer la distribution d'importance et des poids de l'échantillonnage d'importance.

L'heuristique de PAC-vraisemblance

L'équation 3.25 est valable pour toute séquence $(\mathbf{S}(t_k))$, même si cette séquence n'est pas une séquence biologiquement cohérente d'états ancestraux. Elle est donc valable pour toute séquence $\mathbf{S}_l$ définie comme l'addition séquentielle de toutes les copies de gènes g_l $(l = 1, \ldots, n)$ constituant l'échantillon final $\mathbf{S}$, dans n'importe quel ordre. Pour une telle séquence, l'équation 3.25 prend la forme suivante

$$q(\mathbf{S}) = \prod_{l=1}^{n} \pi\big(j(g_l)|d(g_l), \mathbf{S}_{l-1}\big) \frac{n_d(l-1)}{n_{dj}(l-1)} = \binom{n}{\mathbf{n}} \prod_{l=1}^{l=n} \pi(j(g_l)|d(g_l), \mathbf{S}_{l-1}). \qquad (3.31)$$

où $j(g_l)$ et $d(g_l)$ représentent respectivement le type allélique de la copie de gène ${}_g l$ et la sous-population où elle est ajoutée.

L'utilisation d'une approximation pour les π dans cette expression donne un Produit de Approximation Conditionnelle (PAC) de la vraisemblance (Li & Stephens,

2003). Il s'agit d'une approximation heuristique, en ce sens qu'elle n'est généralement pas un estimateur cohérent de la vraisemblance. Cependant, nous pouvons utiliser les mêmes approximations des π que dans l'algorithme d'échantillonnage d'importance (Cornuet & Beaumont, 2007), et dans ce cas, l'inférence de la vraisemblance basée sur la PAC-vraisemblance s'est avérée pratiquement équivalente à celle basée sur la vraisemblance (Rousset & Leblois, 2007, 2012; Leblois *et al.*, 2014). Le principal inconvénient de cette approximation est que, puisqu'il n'y a pas de temps ancestral attaché aux échantillons ancestraux successifs $\mathbf{S}(l)$, cette approximation PAC-vraisemblance ne peut pas remplacer l'approche IS dans les modèles avec des variations dans le temps. Cependant, certains modèles avec des taux variables dans le temps comprennent une phase démographique ancestrale stable (par ex. le modèle avec une contraction ou une augmentation de la taille de la population utilisé dans Leblois *et al.*, 2014). Dans le cadre de ces modèles, la PAC-vraisemblance peut encore être utilisée pour estimer la probabilité que les états des lignées de gènes ancestraux subsistent lorsque la phase démographique stable est atteinte en remontant dans le temps, et cette approximation a permis de réduire considérablement le temps de calcul sans perte de précision (Leblois *et al.*, 2014).

Les algorithmes IS-coa (IS et PAC-vraisemblance) décrit ci-dessus ont été implémenté dans le logiciel `Migraine` que nous détaillerons et dont nous étudierons les performances dans la section 3.2.2.

3.2 Inférences vraisemblance sous IBD

Avant de détailler les performances de l'inférence par vraisemblance sous IBD que nous avons étudié avec notre implémentation des algorithmes d'IS dans `Migraine` décrite ci-dessus, regarderons deux tests "préliminaires" obtenus avec l'approche MCMC-coa implémentée dans MIGRATE sur des données en IBD dont nous pourrons tirer quelques conclusions intéressantes et assez générales.

3.2.1 Approche MCMC-coa : test de MIGRATE sur données en IBD

Le logiciel MIGRATE (Beerli & Felsenstein, 1999, 2001) est censé traiter des modèles très généraux de migration dans lesquels la migration est définie, comme pour le coalescent structuré, par une matrice de migration $\mathbf{M} = m_{ab}$ définissant les taux de migration entre chaque paires de sous-populations et une vecteur de taille de sous-populations $\mathbf{N} = N_a$ (en nombre d'individus diploïdes). Les résultats fournis par ce logiciel pour la traitement d'un échantillon de gènes pris dans dans un ensemble de n sous-populations sont les paramètres $\theta_a = 4N_a\mu$ pour chaque sous-population a échantillonnée et les taux de migration pour chaque paire de sous-population sous la forme $4N_a m_{ab}$ où m_{ab} est le taux de migration "avant" de la sous-population a vers b. Dans le chapitre 2 de ce document, nous avons vu qu'une méthode fondée sur les F-statistiques donne de bonnes estimations du produit $D\sigma^2$, où D est la densité d'individus et σ^2 le moment d'ordre deux de la distribution de dispersion parents-descendants, à partir de données génétiques sous isolement par la distance. Cependant, le fait que l'on estime uniquement σ^2 comme paramètre de dispersion n'est pas tout à fait satisfaisant. En effet, il serait plus intéressant d'avoir une estimation de la distribution de dispersion en elle-même. Le logiciel MIGRATE aurait

en théorie ce potentiel en considérant l'ensemble des estimateurs des taux de migration $4N_a m_{ab}$ répartis par classes de distances entre sous-populations. C'est pourquoi il nous est paru intéressant d'évoluer les potentialités de ce logiciel dans le cadre de modèles d'isolement par la distance sur un jeu de données réel ainsi que sur des jeux de données simulés.

Test sur jeu de données réel

Le jeu de données utilisé est constitué de données démographiques (Wood *et al.*, 1985) et des données génétiques de marqueurs allozymiques (Long *et al.*, 1986) obtenues sur des populations humaines de Papouasie-Nouvelle-Guinée. L'avantage de ce jeu de données est qu'il comporte à la fois des données démographiques et des données génétiques.

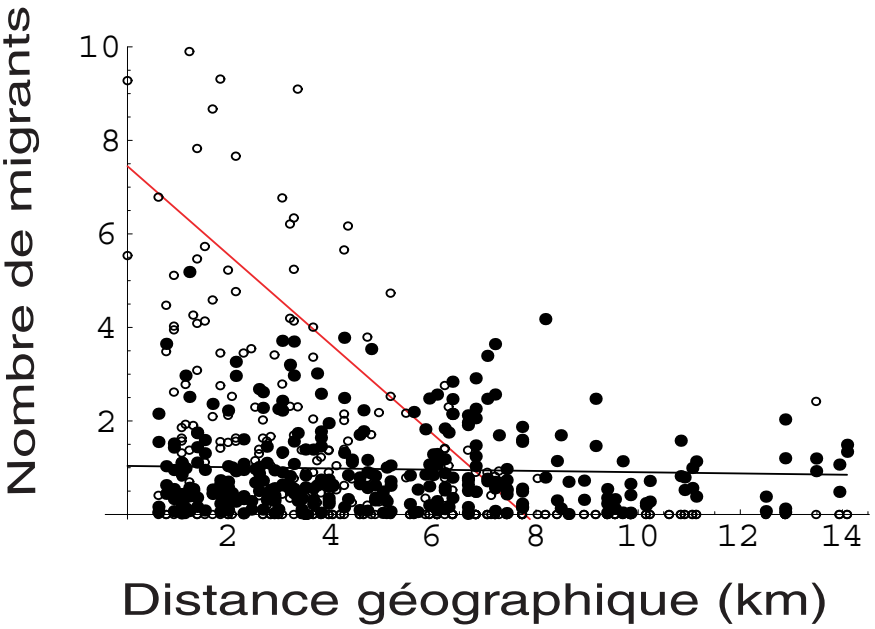


FIGURE 3.2 – Représentation du nombre de migrants en fonction de la distance géographique issu des données Humaine sur les villages de Papouasie Nouvelle-Guinée. Les cercles représentent les estimations à partir du jeu de données démographiques. Les points noirs sont les nombres de migrants estimés par MIGRATE à partir du jeu de données génétiques. Les droite correspondent aux droites de régression calculées sur chaque jeux de données. Les distances sont exprimées en nombre de pas sur le réseau.

Les données démographiques ont permis de calculer σ^2 à partir de la distribution des distances des villages des parents par rapport aux villages des descendants. Cette estimation démographique donne une estimation de σ^2=1.93 km^2/génération (Rousset, 1997). Par ailleurs, les données génétiques ont été traitées avec la méthodes de Rousset (1997), analogue à la méthode de Rousset (2000) mais considérant un modèle d'isolement par la distance avec une structure en dèmes ; cette méthode donne une estimation indirecte de $\sigma^2 = 1.4$ km^2/génération (l'estimation de σ^2 est obtenue en considérant les tailles des populations issues de données démographiques), qui est proche de l'estimation démographique.

Ce même jeu de données génétique a été traité avec le logiciel MIGRATE et le calcul de σ^2 à partir de l'ensemble des taux de migration par paires de sous-populations donne une estimation indirecte de $\sigma^2 = 16.3$ km^2/génération. Ce résultat parait largement sur-estimer la dispersion réelle (c.a.d d'un facteur 10). La

figure 3.2 représente les estimations du nombre de migrants en fonction de la distance par la méthode démographique (points gris) et par MIGRATE (points noirs). On obtient une surestimation globale par MIGRATE de la migration pour chaque classe de distances, au point de ne plus observer de patron net d'isolement par la distance (Figure 3.2).

Test sur jeux de données simulés

Dans un deuxième temps nous avons testé l'estimation des taux de migration par MIGRATE sur des jeux de données simulés avec `IBDsim`. Les simulations ont été faites en considérant un échantillon de 20 individus pris dans 11 sous-populations pour 5 locus. Les individus ont évolué sur un tore de (200×200) avec 20 individus par dème, le modèle mutationnel est le KAM à 10 allèles avec un taux de mutation de $5 \cdot 10^{-4}$. Enfin, la migration se fait uniquement entre dèmes adjacents (migration "stepping-stone") avec un taux de migration total de 1/2. Trois jeux de données simulés ont été analysés avec MIGRATE. Nous n'avons analysé qu'un petit nombre de jeux de données pour seulement 5 locus car les temps de calcul demandés par MIGRATE sont longs et nous n'avions alors pas accès à des ordinateurs puissants. Les résultats de ces simulations sont présentés sur la figure 3.3.

On voit bien sur la figure 3.3 que MIGRATE sur-estime largement les nombres de migrants entre sous-populations. En effet, puisque la migration se fait, dans notre modèle de simulation, uniquement entre dèmes adjacents, on s'attend à avoir un nombre de migrants positif pour des distances de 1 pas sur le réseau et un nombre de migrants nul pour toutes les autres distances (carrés gris Figure 3.3). On peut noter toutefois que l'estimation du nombre de migrants à une distance de 1 pas sur le réseau est bonne et correspond bien aux valeurs du modèle. Enfin, comme pour l'analyse du jeu de données réel, MIGRATE sous-estime largement l'isolement par la distance puisque le nombre de migrants estimé ne décroît que très peu avec la distance (droites de régression de la Figure 3.3), contrairement à ce qui est attendu.

Que ce soit pour le jeu de données réel ou pour les jeux de données simulés, les mauvais résultats obtenus avec le logiciel MIGRATE peuvent être dus à différents facteurs :

(i) le processus mutationnel des marqueurs (ou le modèle mutationnel simulé) ne correspond pas exactement à celui assumé dans MIGRATE (IAM dans MIGRATE, KAM dans les simulations et inconnu pour le jeu de données réel). Une analyse de robustesse de la méthode aux processus mutationnels serait nécessaire. Si la méthode n'est pas robuste vis à vis des modèles mutationnels, cela risque de limiter considérablement son utilité en pratique du fait que l'on a rarement une idée précise des processus mutationnels des marqueurs utilisés lors d'études expérimentales.

(ii) le nombre de sous-populations échantillonnées (11 dans les simulations) ne correspond pas au nombre de sous-populations total du modèle démographique (40 000 dans les simulations). Or MIGRATE considère que le nombre de sous-populations de l'échantillon est égal au nombre total de sous-populations du système étudié. Une étude de Beerli (2004) montre que, dans le contexte d'un modèle en île, le nombre de sous-populations non échantillonnées a beaucoup d'influence sur l'estimation des tailles de populations mais peu sur l'estimation des taux de migration. Cette robustesse à la présence de sous-populations non échantillonnées est rassurant. Cependant, cette étude considère que les taux de migration entre sous-populations

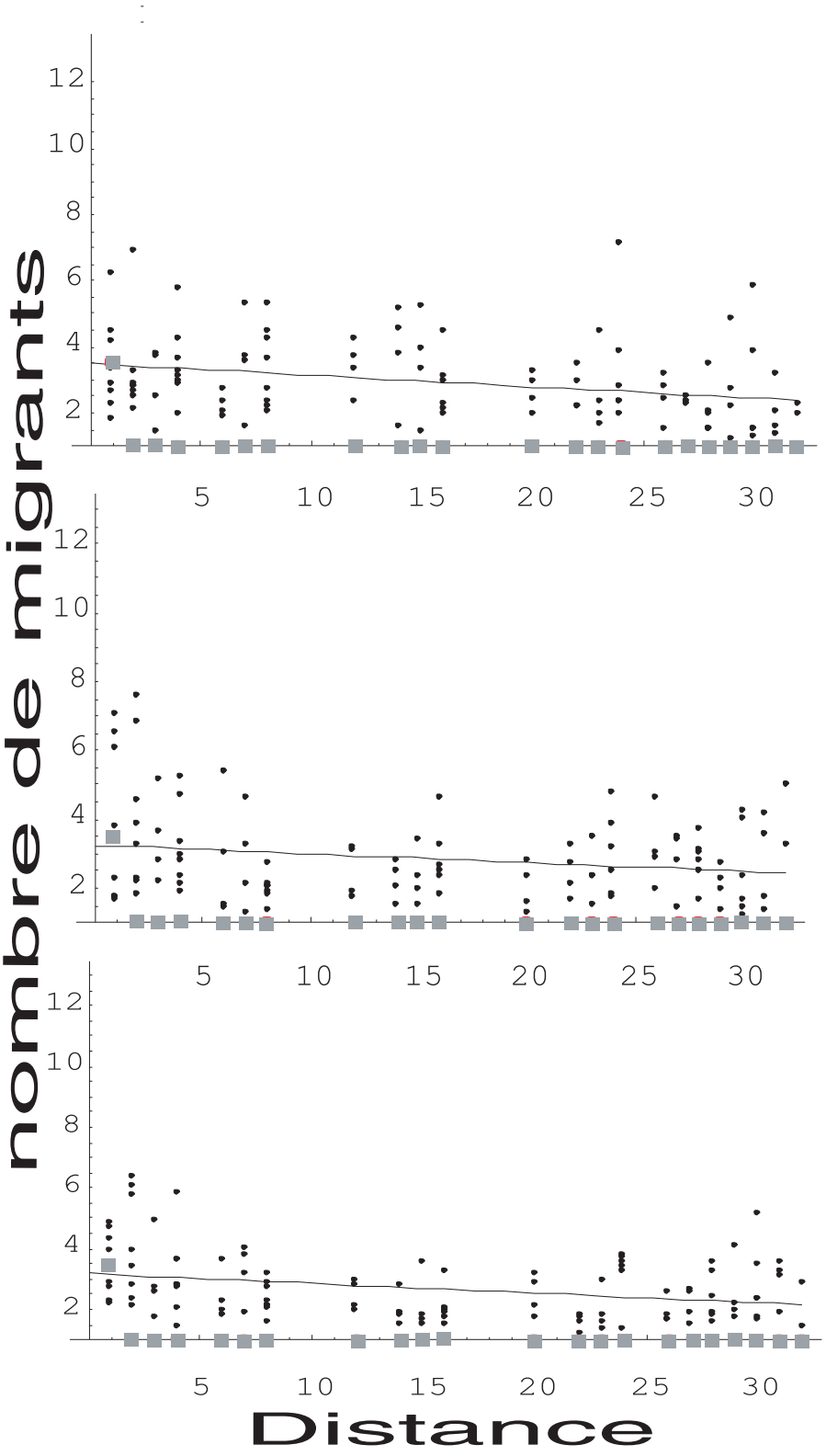


FIGURE 3.3 – Représentation du nombre de migrants en fonction de la distance géographique. Les carrés gris représentent les valeurs attendue (c.a.d les valeurs avec lesquelles ont a simulé les données). Les points noirs sont les nombres de migrants estimés par MIGRATE sur ces jeux de données simulés.

sont globalement faibles (de l'ordre de 0.0001 événements de migration par individus par génération). Slatkin (2005) prédit l'ampleur de ces effets à partir d'un simple argument de coalescence entre paires de gènes et montre, comme attendu, que ces effets augmentent avec les taux de dispersion. Des tests supplémentaires de l'influence de la présence de sous-populations non échantillonnées sur les estimations données par MIGRATE, avec des taux de migrations plus forts serait donc nécessaires pour conclure à la robustesse générale de cette méthode vis à vis de ce facteur.

(iii) la convergence est un problème récurrent dans le contexte des MCMC-coa et la configuration par défaut de MIGRATE n'est peut être pas optimale de ce point de vue. Toutefois, une estimation avec des analyses beaucoup plus longues (analyse de deux semaines par jeu de données) ont été faites sur les jeux de données simulés et aucune différence notable n'a été notée par rapport au analyses "courtes" (analyse de six jours par jeu de données) de la configuration par défaut.

Enfin (iv) il pourrait exister un biais inhérent à la méthode qui sur-estimerait les taux de migration. En effet, Beerli et Felsenstein ont observé sur des simulations un biais positif pour des paires de sous-populations n'échangeant aucun migrant (Beerli & Felsenstein, 2001). Ce biais pourrait être d'autant plus important que le nombre de paramètres estimés est élevé, réduisant ainsi la précision de l'estimation de chaque paramètre.

Tous ces facteurs nous poussent à conclure que les estimations de paramètres démographiques avec le logiciel MIGRATE doivent être interprétées avec beaucoup de précautions, notamment pour des populations en isolement par la distance. Des études approfondies de robustesse de ce logiciel vis à vis de différents facteurs mutationnels et démographiques seraient nécessaires.

3.2.2 Performances des algorithmes d'IS-coa en populations structurées

Résultats préliminaires sur les différents algorithmes d'IS-coa

Dans un premier temps, nous avons testé les performances de (i) l'algorithme de Nath & Griffiths (1996), correspondant à la fonction d'échantillonnage d'importance "peu efficace" définie par Griffiths & Tavaré (1994a) adaptée à un coalescent structuré ; puis (ii) les nouvelles fonctions d'échantillonnage d'importance de de Iorio & Griffiths (2004b) toujours dans un coalescent structuré ; et (iii) l'algorithme de de Iorio *et al.* (2005) que nous avons développé pour prendre en compte les mutations par pas dans le d'un modèle à deux populations. Ces tests ont été fait sur des jeux de données simulés par `IBDsim` dans un modèle en île homogène ou dans un modèle à deux populations avec flux de gènes asymétriques, et nous avons parfois comparé leurs performances avec celles du logiciel MIGRATE (introduit dans la section précédente). Les principaux résultats de ces études, dont le lecteur pourra trouver les détails dans Leblois (2004) et de Iorio *et al.* (2005), sont (i) l'algorithme de Griffiths & Tavaré (1994a) est extrêmement lourd à simuler car les temps de calcul nécessaires à l'estimation de la vraisemblance sont très longs (c.a.d autour de 50 000 généalogies à explorer pour avoir une estimation assez précise de la vraisemblance en différents points de l'espace des paramètres) et ne peut raisonnablement pas être utilisé pour de l'inférence en populations structurées ; (ii) que les nouvelles distributions d'échantillonnage d'importance de de Iorio & Griffiths (2004b) et de Iorio *et al.* (2005) sont beaucoup plus efficaces que les précédentes (c.a.d gain d'un facteur 20 à 1000 car il suffit de simuler 1 à 200 généalogies pour avoir une estimation

assez précise de la vraisemblance), et permettent donc d'envisager l'estimation de paramètres démographiques pour des modèles de population structurée ; et (3) les approches par IS-coa semblent trois à dix fois plus rapides, plus précises (voir tableau 3.1) et plus faciles d'utilisation que les approches MCMC-coa pour différentes raisons que nous détaillerons ci-dessous. On peut tout de suite noter qu'il n'y a pas que les algorithmes d'estimation de la vraisemblance qui rentre en compte dans ces tests de performances et comparaisons, mais aussi le cadre statistique dans lequel est fait l'inférence.

TABLE 3.1 – Précision de l'estimation de $(\theta = 2N\mu, \gamma = 2Nm)$ avec l'algorithme de de Iorio & Griffiths (2004b) et avec le Logiciel MIGRATE dans le cas d'un modèle à 2 populations avec un modèle de mutation par pas. (e.s) est l'erreur standard de l'estimation.

		Algorithme			
		De Iorio *et al.* 2004		MIGRATE	
		5 locus	20 locus	5 locus	20 locus
θ	Biais	-0.036	0.039	0.23	-0.60
	(e.s.)	(0.32)	(0.20)	(0.49)	(0.18)
	MSE	0.10	0.041	0.30	0.40
γ	Biais	0.31	0.30	1.2	0.25
	(e.s.)	(0.67)	(0.32)	(1.2)	(0.62)
	MSE	0.54	0.21	3.1	0.46

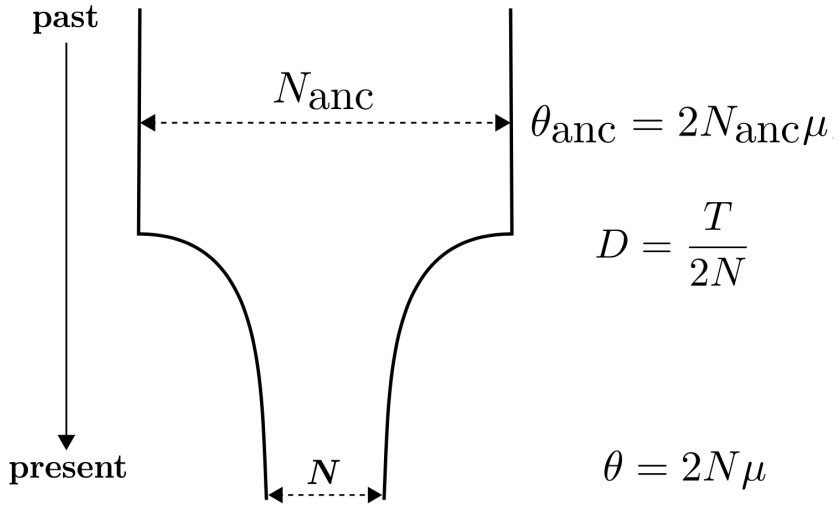


FIGURE 3.4 – Représentation du modèle démographique avec un changement passé de taille de population considéré dans Leblois *et al.* (2014) et implémenté dans `Migraine`. N est la taille actuelle de la population, N_{anc} est la taille de la population ancestrale (avant le changement démographique), T est le temps mesuré en générations depuis le présent et μ le taux de mutation des marqueurs utilisés. Ces quatre paramètres sont les paramètres canoniques du modèle de population de taille finie (c.a.d WF). θ , D et θ_{anc} sont les paramètres mis à l'échelle du coalescent. Les tailles sont exprimées en nombre de gènes, et pas en nombre d'individus diploïdes.

Pour ces différentes raisons, nous avons continué d'adapter et tester les performances des algorithmes IS-coa de de Iorio & Griffiths (2004a) pour d'autres modèles

démographiques, notamment (i) pour estimer des changements passé de tailles d'une population (modèle avec un changement passé de taille, discret ou continu, dans Leblois *et al.* (2014) pour inférer une expansion ou une contraction de population ; avec deux changements démographiques dans Rousset *et al.* (2018) pour inférer, par exemple, un évènement de fondation suivi d'une expansion) ; et (ii) pour estimer la dispersion et les tailles de population sous un modèle d'IBD en une dimension dans Rousset & Leblois (2007) et en deux dimension dans Rousset & Leblois (2012). Tout ces développements ont été implémenté dans le logiciel `Migraine`, dont certaines caractéristiques clés seront été présentées ci-dessous. Après avoir mis en avant quelques résultats pertinents obtenus lors des tests de performances en situation de changements démographiques, nous détaillerons les résultats obtenus sous IBD. Une caractéristique distinctive de tous ces travaux, par rapport à la plupart des publications sur les méthodes alternatives d'inférence, est l'accent mis sur l'évaluation de l'inférence en termes de propriétés de couverture des intervalles de confiance fondés sur la vraisemblance.

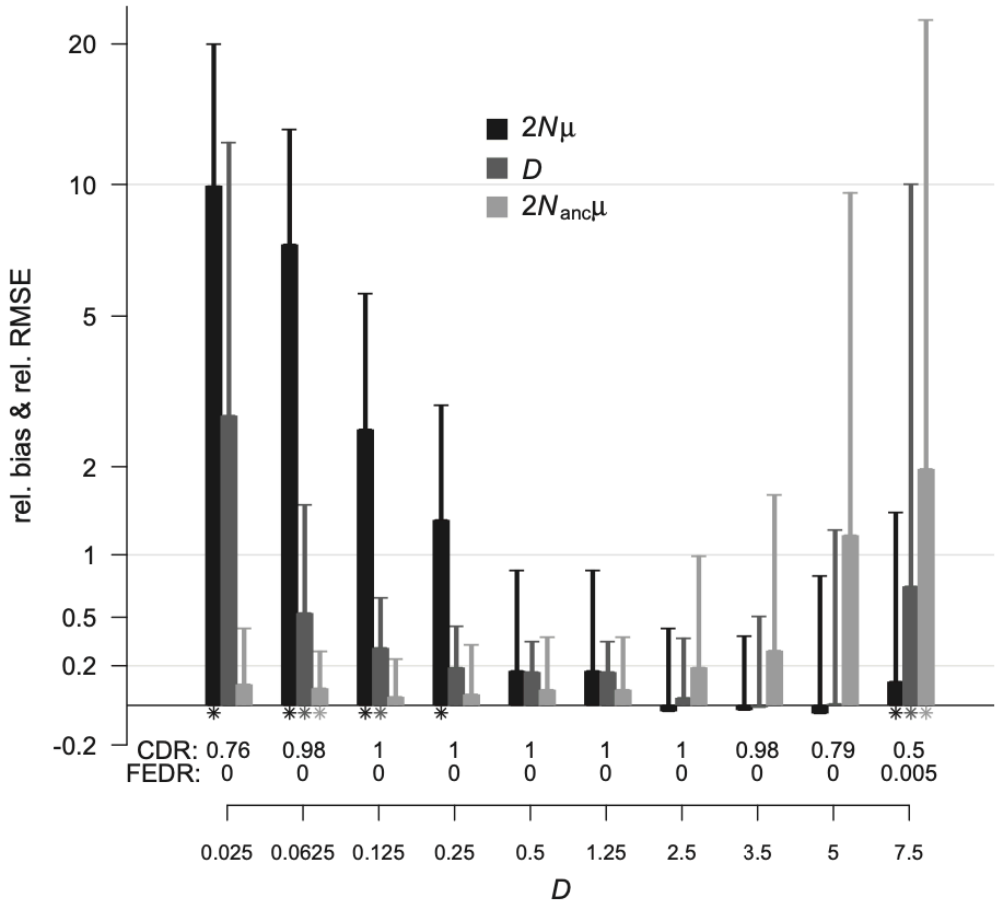


FIGURE 3.5 – **Effet du moment de la contraction de la taille de la population sur l'inférence de chaque paramètre du modèle.** Le biais relatif est indiqué par les grosses barres et la RRMSE par les lignes fines. Les étoiles indiquent les faibles P-valeurs du test de Kolmogorov-Smirnov sur la distribution des LRT P-valeurs (c.a.d <0,05). CDR : taux de détection des contractions ; FEDR : taux de détection des fausses expansions.

Leblois *et al.* (2014) illustre bien l'intérêt et les limites de ces approches pour des modèles en déséquilibre, dont la principale limite est que les fonction d'échantillonnage d'importance de Stephens & Donnelly (2000) et de Iorio & Griffiths (2004a), fondées sur l'hypothèse d'équilibre "démographique" (CF section 3.1.2 et annexe de

Rousset *et al.*, 2018), sont de moins en moins efficace pour des changements démographiques de plus en plus intense (c.a.d fort et soudains). Ce qui implique des temps de calculs plus longs sous ces modèles en déséquilibre. Notons au passage que Leblois *et al.* (2014) montre aussi une influence de la structuration des populations sur ces inférences des variations passés des tailles de populations, qui peut fortement biaiser les résultats sous ces modèles ne considérant qu'une population panmictique isolée. Cette publication illustre aussi très bien le fait que la précision des estimation de chaque paramètre démographique, notamment les tailles actuelles ou ancestrales de la population, dépend fortement de la situation démographique considérée, et comment on peut "intuiter" l'information contenue dans les données sur les différents paramètres à l'aide de la théorie de la coalescence. La figure 3.5 illustre bien cette relation entre précision des estimations et situation démographique. Ainsi, pour résumer très brièvement, la taille actuelle N ne peut être estimer avec précision que si le changements n'est pas trop récent, sinon il ne se passe pas assez de temps dans la configuration "taille actuelle" pour que le processus évolutif (les coalescences et les mutations) ne laisse de trace de cette période. Inversement, la taille ancestrale N_{anc} ne peut être estimer avec précision que si le changements n'est pas trop ancien, sinon le processus de coalescence est fini (MRCA) ou presque (juste quelques lignées ancestrales) pour que le processus évolutif (les coalescences et les mutations) ne laisse de trace de cette période. Enfin, une comparaison des performances des algorithmes IS-coa et la méthode Msvar, implémentant un modèle démographique équivalent mais avec l'approche MCMC-coa, montre encore une fois de meilleures performances, et des temps de calculs plus courts, pour l'IS-coa (voir Leblois *et al.*, 2014).

Le logiciel `Migraine`

Les algorithmes décrits ci-dessus fournissent des estimations de la vraisemblance pour des valeurs de paramètres données. Dans le contexte statistique du maximum de vraisemblance, calculer des intervalles de confiances, ainsi que des graphiques des profiles [1] de vraisemblance en une et deux dimension pour les différentes (paires de) paramètres, nécessite de déduire une surface de vraisemblance à partir des vraisemblances estimées en différents points de l'espace des paramètres. Le "Krigeage" a été utilisé classiquement pour l'inférence des surfaces de réponse (par ex. Sacks *et al.*, 1989), et nos efforts pour obtenir une bonne couverture nous ont conduits à implémenter ces méthodes dans le package R `blackbox` permettant d'explorer la surface de vraisemblance d'une manière automatique (voir Rousset & Leblois (2012); Rousset *et al.* (2018) pour plus de détails sur cette procédure de Krigeage). Une des point cruciaux est d'avoir une procédure de lissage qui prend bien en compte l'erreur d'estimation de la vraisemblance.

Les méthodes décrites ici sont toutes mises en œuvre dans un logiciel libre : le

1. le profil de vraisemblance est la fonction de vraisemblance "marginale" qui résulte de la maximisation de la fonction de vraisemblance sur tous les autres paramètres. Autrement dit, le profil de vraisemblance donne la valeur maximale de la fonction de vraisemblance pour chaque valeur possible du paramètre considéré, en fixant pour cette même valeur du paramètre considéré les autres paramètres à leur valeur de vraisemblance maximale. C'est une façon de suivre les lignes de crêtes de la fonction de vraisemblance pour les différents (ou les différentes paires de) paramètres. Ce n'est donc pas une coupe dans la vraisemblance. Cette approche est aussi assez de celle utilisée pour l'obtention des distribution marginales, notamment largement utilisées dans les approches Bayesiennes, dans laquelle on intègre sur les autres paramètre.

logiciel `Migraine` écrit en C++, met en œuvre les algorithmes d'estimation de la vraisemblance pour chaque point de paramètre donné, écrit et lance un script R qui effectue l'inférence de la surface de vraisemblance à partir des points de vraisemblance, trace diverses représentations de cette surface et d'autres diagnostics, évalue les intervalles de confiance du rapport de vraisemblance, et conçoit de nouveaux points de paramètre dont la vraisemblance doit être calculée lors d'une prochaine itération automatique de `Migraine`, et ainsi de suite jusqu'à obtenir une "bonne estimation" (selon l'utilisateur ou selon des critères de MSE des estimateurs et des bornes des IC). Une caractéristique intéressante de cette approche itérative est qu'il n'est pas très important d'avoir une estimation précise de la vraisemblance en chaque point de paramètre, car l'accumulation des estimations de la vraisemblance à proximité du maximum (ou de tout autre point cible) au cours des itérations successives fournira, grâce aux propriétés asymptotiques d'interpolation du Krigeage (Stein, 1999), une estimation précise de la vraisemblance au niveau du maximum. Par exemple, dans les modèles IBD que l'on verra ci-dessous, il suffit de simuler 20 généalogies par point de paramètre pour obtenir des propriétés de couverture presque parfaites des intervalles de confiance (Rousset & Leblois, 2012). Par contre il en faut quelques milliers dans le modèle avec variation passé de la taille de population dans lequel la fonction d'échantillonnage d'importance est moins efficace (voir section 3.1.2). Toute cette procédure itérative d'inférence est illustrée dans Rousset *et al.* (2018) à partir d'un exemple détaillé et la figure 3.6 illustre le type de graphiques que produit `Migraine`. Cette procédure itérative et le Krigeage sont deux points essentiels de `Migraine`, et expliquent sans doute en partie les meilleurs performances des algorithmes IS-coa implémenté dans `Migraine` par rapport aux approches MCMC-coa (voir discussion ci-dessous en section 3.3).

Les méthodes examinées dans le présent document ont été largement évaluées, notamment en termes de couverture des IC (comme illustré Figure 3.7, et voir Rousset & Leblois, 2012; Leblois *et al.*, 2014). L'évaluation de la couverture des IC est souvent plus utile que l'évaluation du biais et de la variance pour détecter les problèmes dans l'inférence de la surface de vraisemblance par lissage, et plus largement tout problème de code. Cette évaluation mériterait à peine d'être mentionnée, si ce n'est qu'elle n'est pas la pratique dominante dans de larges segments de la littérature liée à ce travail, que ce soit dans ses objectifs (inférence à partir de la variation génétique) ou dans ses méthodes (diverses méthodes stochastiques pour déduire les vraisemblances ou les distributions a posteriori). Par conséquent, des méthodes ou des logiciels mal évalués sont facilement disponibles et approuvés par des praticiens désireux de tirer parti de leurs données. En fait, très peu de publications testant des méthodes d'inférence génétique des populations mentionnent même les propriétés de couverture des intervalles de confiance. En outre, les quelques articles qui font état de ces informations constatent souvent que les IC sont très imprécis (par ex. Abdo *et al.*, 2004 ; Beerli, 2006 ; Hey, 2010 ; Hey *et al.*, 2015 ; annexe S3 de Peter *et al.*, 2010).

Comme pour tous les logiciels que nous développons, la procédure d'estimation de la vraisemblance a été vérifiée par rapport aux formules standard de probabilité d'identité des paires de gènes (cf section 1.2.1) et prises dans la limite $N \to \infty$ pour $N\mu$ et Nm fixés comme dans l'algorithme de coalescence. Plus globalement, le fait que les distributions des LRT P-valeurs soient uniformes (voir ci-dessous) dans la plupart des situations testées valident aussi la procédure d'inférence et la procédure de calcul de la vraisemblance, ainsi que les algorithmes de simulation dans les conditions du coalescent. C'est une autre manière, complémentaire et plus "intégrative"

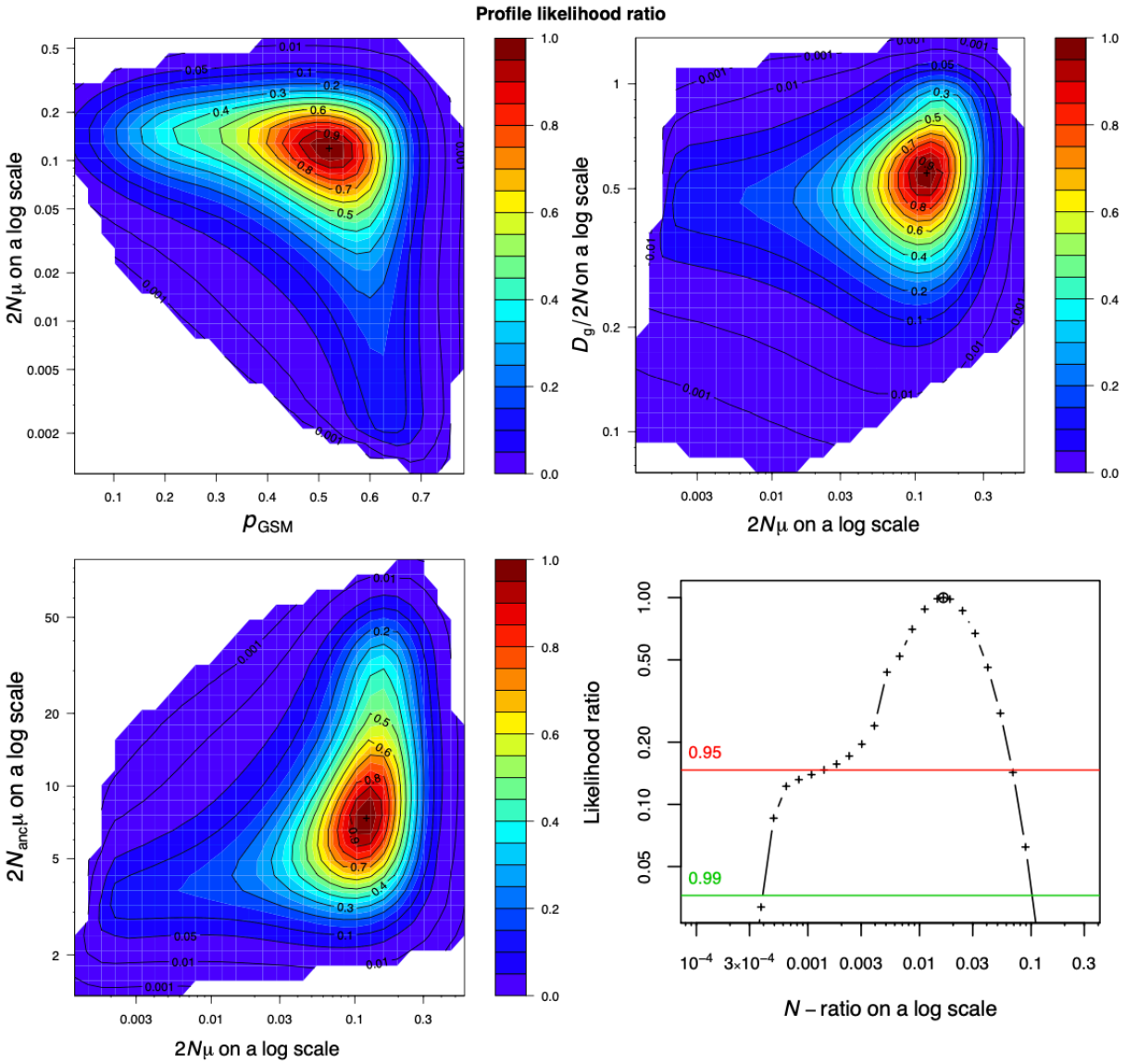


FIGURE 3.6 – Profiles 1D et 2D des LRT pour le jeu de données sur les moutons analysé dans Rousset *et al.* (2018) avec `Migraine` dans un modèle d'une seule population ayant subi un changement passé de taille de populations, schématisé en Figure 3.4. $N-$ratio est le ratio de la taille de population actuelle N sur la taille ancestrale N_{anc}.

que les probabilités d'identité, de vérifier les logiciels développés.

Performance des algorithmes IS-coa sous IBD

Les algorithmes IS-coa implémentés dans `Migraine` ont tout d'abord été testé dans le cadre d'un modèle d'IBD en une dimension (c.a.d habitat linéaire) dans Rousset & Leblois (2007), puis en deux dimensions dans Rousset & Leblois (2012). Le modèle démographique d'inférence est forcément un modèle en dème puisque les algorithmes d'IS-coa sont dérivé sous les hypothèses de grande taille de population du coalescent structuré. La dispersion axiale est modélisée par une distribution géométrique de paramètre g (paramètre de forme) donnant la probabilité de disperser à $k > 0$ pas de distance comme $\frac{m}{2}(1-g)g^{|k-1|}$, ou m est la probabilité d'émigration de chaque dème. La valeur $g = 0$ correspond à un stepping-stone, et la valeur 1 à un modèle en île (voir section 2.1). La méthode d'inférence IS-coa adaptée au modèle démographique IBD permet donc d'estimer trois paramètres canoniques, mis à l'échelle du coalescent (c.a.d sous forme de produit avec la taille de la population N qui est aussi l'unité de mesure du temps, voir section 1.1) : un taux de mutation $N\mu$, un taux de migration Nm (probabilité d'émigration mise à l'échelle) et le paramètre g décrivant la distribution géométrique des distances de dispersion. Nous avons également considéré l'estimation du paramètre composite $D\sigma^2$, ou σ^2 se calcul pour une dispersion géométrique, en fonction des paramètres m et g de la distribution selon l'équation

$$\sigma^2 = \frac{m(1+g)}{(1-g)^2}. \tag{3.32}$$

Les performances de la méthode en terme de précision, robustesse et couverture des IC, a été testée en considérant dans un premier temps que le modèle d'inférence correspond exactement au modèle ayant généré les données (notamment des grandes tailles de dèmes de 400 à 40,000 gènes, des faibles taux de migration de 0.01 à 0.001, un même nombre de dème dans la population et un même modèle mutationnel dans les simulations et dans l'inférence) avec un petit nombre de dèmes (par ex. 4 à 10 en 1D, 4x4 à 10x10 en 2D), puis avec plus de dèmes (100 en 1D, 40x40 à 80x80 en 2D), puis en s'écartant des hypothèses du modèle d'inférence notamment et en tirant parti de la méthode rapide de la PAC-vraisemblance pour pouvoir augmenter la taille de la grille, réduire la taille des dèmes, augmenter la dispersion. Sauf précision contraire, un échantillon de 4x60 gènes est pris dans 4 dèmes situés au milieu de l'habitat pour le modèle 1D, et de 8x20 gènes pris dans 4 paires de dèmes situés dans les coin de l'habitat en 2D. Le modèle mutationnel par défaut est un KAM à 4 ou 10 allèles, avec un taux de mutation entre 10^{-4} et 10^{-3} selon les situation, choisi pour obtenir une diversité suffisante. Dans chacune des situations testées, que je ne présenterai pas en détail ici, 200 jeux de données simulés ont été analysés pour en calculer le biais et l'erreur relative d'estimation de chacun des paramètres, ainsi que la couverture des intervalles de confiance grâce aux distribution de LRT P-valeurs comme détaillé ci-dessous.

Un premier résultat général est que l'estimation par IS-coa dans le cadre de modèles de moins de 20 populations est relativement facile (notamment assez rapide pour tourner sur un seul processeur en quelques heures), elle devient progressivement plus difficile à mesure que le nombre de sous-populations augmente, et l'analyse d'un jeu de données moyen de type microsatellites (par ex. une centaine d'invididus

échantillonnées à une vingtaine de marqueurs) nécessite des semaines de calcul sur un processeur lorsque plus de 40 sous-populations sont prises en compte (mais qu'une vingtaine d'heures sur un cluster du fait que les algorithmes IS-coa sont facilement parallélisable).

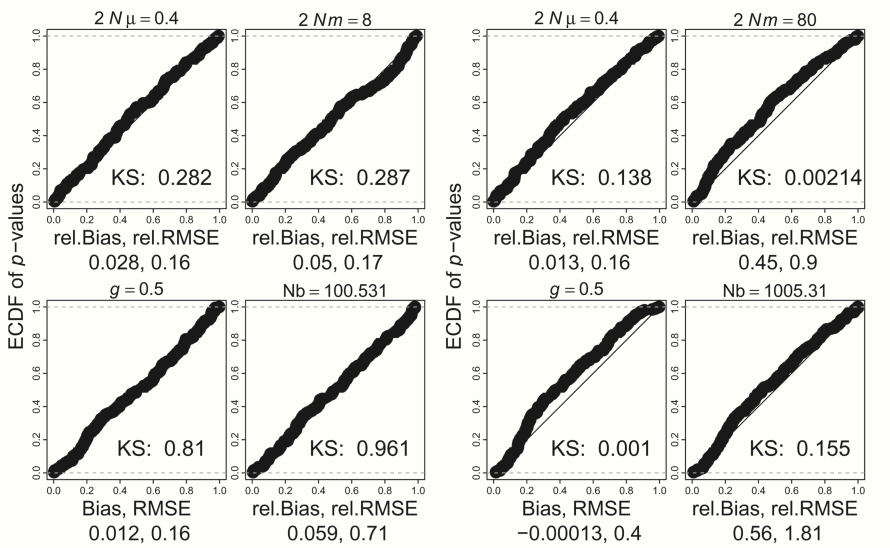


FIGURE 3.7 – Distributions des P-Valeurs des test de ratio de vraisemblance (LRT P-value) en IBD 2D, avec 4x4 dèmes, $N = 400$, $g = 0.5$, $\mu = 5.10^{-4}$ et $m = 0.001$ pour [1] (à gauche), alors que $m = 0.1$pour [2] (à droite).

Dans tous les cas idéaux, c.a.d quand les données sont simulées sous le modèle d'inférence, la précision obtenue sur les 4 paramètres $N\mu$, Nm, g, et $D\sigma^2$ est excellente pour n'importe quelle application pratique, même avec un très petit nombre de loci (par ex. 5 à 20 locus microsatellites dans cette étude). Les biais relatif sont souvent de l'ordre de 5 à 10% pour tous les paramètres, et toujours ≤ 0.5 pour $2Nmu$, ≤ 0.2 pour $2Nm$, et ≤ 0.1 pour g. Les RMSE sont généralement de l'ordre de 0.2 à 0.5 avec des valeurs parfois plus fortes pour Nm et $D\sigma^2$ (voir Rousset *et al.*, 2018 pour les détails).

Processus mutationnels Les distributions d'importance pour l'algorithme IS-coa en IBD ont été obtenues uniquement pour un modèle de mutation KAM, qui ne peut généralement pas être considéré comme une représentation exacte des processus de mutation des marqueurs utilisés. Les tests sur des données de type microsatellites simulées dans le cadre d'un modèle SMM (Ohta & Kimura, 1973, voir section 1.2.1) ont montré qu'une mauvaise spécification du modèle de mutation a peu d'impact sur la performance des estimateurs de dispersion Nm et g, mais qu'un biais de 50 à 75% sur les estimations de $N\mu$ est observé (Rousset & Leblois, 2007, 2012). Ce biais est attendu car la variation de la diversité locale dans KAM par rapport à SMM est approximativement celle qui résulte d'une variation d'un facteur 2 du taux de mutation (Rousset, 1996). L'inférence est donc très robuste aux processus mutationnels pour la dispersion, un peu moins pour la mutation/taille de population.

Notons qu'en revanche, les inférences dans les modèles démographiques dont les paramètres varient dans le temps sont beaucoup plus sensibles aux processus mutationnels. Ainsi, Leblois *et al.* (2014) montrent qu'une mauvaise spécification des

processus mutationnels peut induire une fausse détection de la contraction passée de la taille des populations à partir d'échantillons prélevés sur des populations stationnaires. Elle peut également induire des biais dans l'inférence du moment et de la force d'un changement passé de la taille de la population à partir d'échantillons prélevés dans une population qui a effectivement subi des changements démographiques dans le passé. Nous avons donc mis en œuvre des variantes des algorithmes d'échantillonnage d'importance pour différents modèles de mutation dans ces modèles d'inférence des variations passées des tailles de population. Ces travaux sont illustrés pour un modèle SMM et des modèles à une ou deux populations dans de Iorio *et al.* (2005) ; dans Leblois *et al.* (2014) pour un modèle GSM dans une seule population, et dans Rousset *et al.* (2018) pour le modèle à nombre de sites infini (ISM, Kimura, 1969), un modèle adapté aux marqueurs de type séquences d'ADN.

PAC-vraisemblance L'approximation PAC-vraisemblance a été comparée à l'analyse de vraisemblance. Dans tous les cas, leurs performances sont très similaires, si ce n'est que les estimations du taux de mutation par PAC-vraisemblance semblent non (ou moins) biaisées, alors que celles par stricte vraisemblance présentent toujours un léger biais positif (Figure 3.8). L'analyse PAC-vraisemblance étant beaucoup plus rapide (facteur 20 environs) et au moins aussi bonne que la vraisemblance stricte, nous la privilégierons pour toutes les analyses sous ces modèles IBD car elle permet notamment de considérer des modèles avec un plus grand nombre de dèmes.

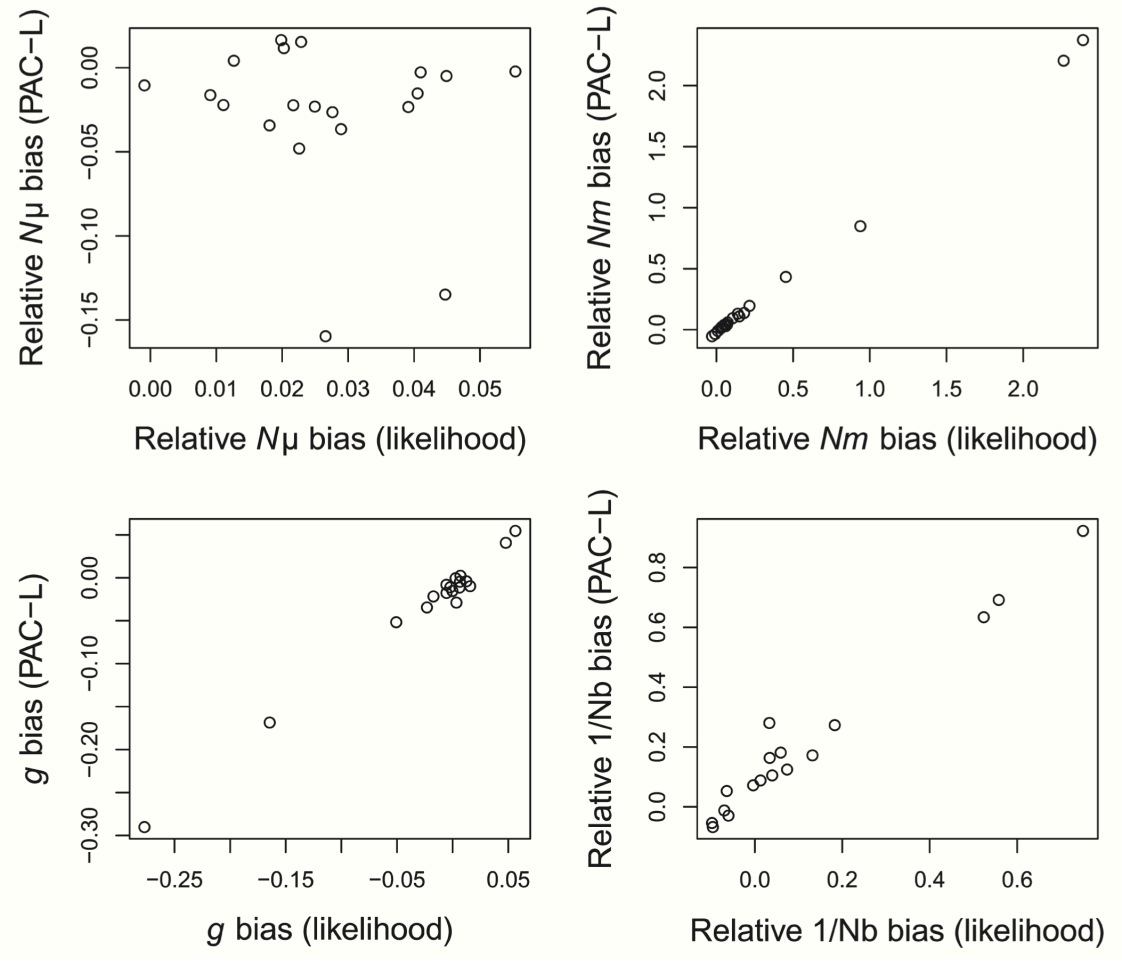


FIGURE 3.8 – Comparaison des biais obtenus par vraisemblance stricte et par PAC-vraisemblance de Rousset & Leblois (2012)

Approximations du coalescent Les propriétés de couverture des intervalles de confiance peuvent être examinées graphiquement en regardant si la distribution cumulée empirique des P-valeurs des tests de ratio de vraisemblance (LRT pour likelihood ratio test) associées à la valeur attendue (c.a.d simulée) du ou des paramètres s'aligne (ou non) sur la diagonale 1 :1. La figure 3.7-1 (à gauche) illustre un bon résultat. Les écarts par rapport à la diagonale sont testés par le test de Kolmogorov-Smirnov. Quatre sous-graphiques sont présentés, un pour chacun des paramètres canoniques et un pour $D\sigma^2$. Le biais relatif et l'erreur relative quadratique moyenne (RMSE) de chaque estimation sont également indiqués. On peut observer, comme décrit précédemment, et cela sera également vrai lorsque les intervalles de confiance ont une couverture incorrecte, que le biais et la RMSE de $N\mu$ et Nm sont faibles par rapport aux attendus classiques et pratiques. La figure 3.7-2 (à droite) présente un résultat moins satisfaisant. La seule différence avec l'exemple précédent est que m est 0.1 au lieu de 0.01. La figure 3.9 illustre aussi un bon résultat et un beaucoup moins satisfaisant, la différence entre les deux situations étudiées étant que l'une est dans les hypothèse du coalescent (à droite) avec des grandes tailles de populations et des petits taux de migration et mutation ($N = 40,000$, $g = 0.25$, $\mu = 10^{-6}$ et $m = 10^{-6}$), alors que l'autre (à gauche) ne l'est pas ($N = 40$, $g = 0.25$, $\mu = 10^{-3}$ et $m = 10^{-3}$).

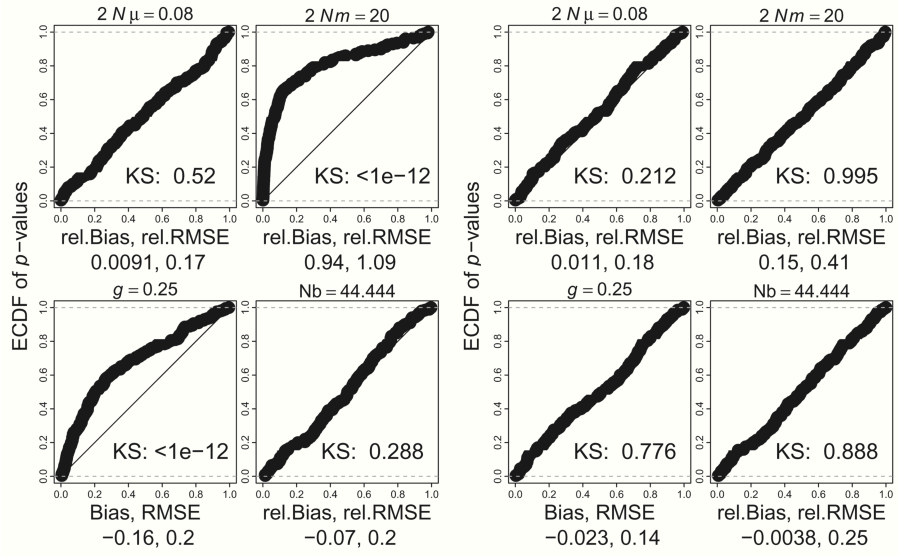


FIGURE 3.9 – Distributions des P-Valeurs des test de ratio de vraisemblance (LRT P-value)dans deux situations, en 1D, avec 16 dèmes, $N = 40$, $g = 0.25$, $\mu = 10^{-3}$ et $m = 0.001$ pour [1] (à gauche), alors que $N = 40,000$, $g = 0.25$, $\mu = 10^{-6}$ et $m = 10^{-}6$pour [2] (à droite) .

Ces résultats illustrent une forte influence des hypothèses du coalescent et donc la difficulté de considérer des petites tailles de population et des forts taux de migration à cause de forts biais dans les estimations. Les biais les plus élevés sont attendus pour des valeurs de m élevées (Figure 3.10). Le biais Nm le plus important est pour $m = 0.5$ dans un réseau linéaire de 100 dèmes, et d'autres cas avec $m = 0.1$ montrent de fortes distorsions de la distribution des LRT P-valeurs. Pour des valeurs intermédiaires de m de 0.01 à $m < 0.1$, les biais sur Nm peuvent encore être relativement importants, mais les distorsions de la distribution des LRT P-valeurs sont moins importantes, sauf dans certains cas où une mauvaise spécification des

effets de bord peut également contribuer (voir Rousset & Leblois, 2012).

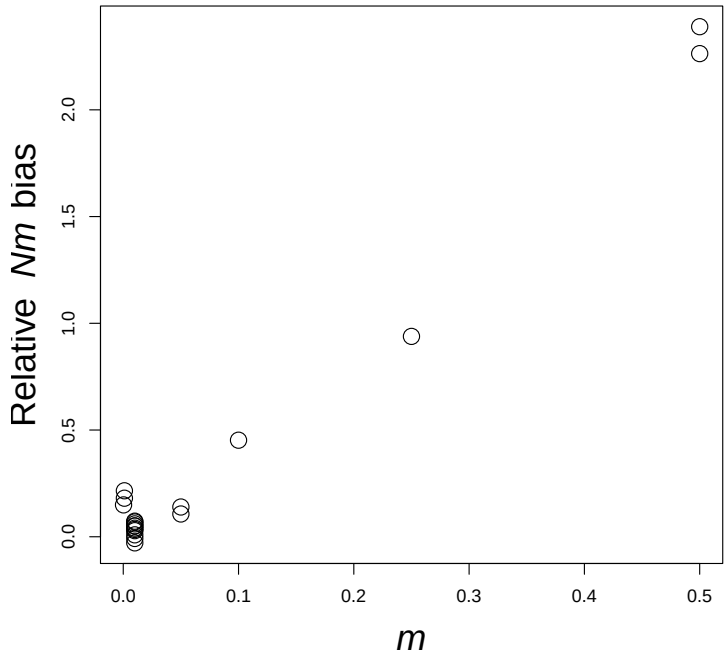


FIGURE 3.10 – Relation entre la probabilité de dispersion et le biais de l'estimation de Nm pour toutes les situations examinées dans Rousset & Leblois (2012)

Quelques difficultés du modèle en dèmes Ainsi, le nombre de dèmes du modèle d'inférence est limité par les temps de calculs et leur taille ne peut être trop petite à cause des approximation du coalescent. En populations naturelles, comme nous l'avons vu en introduction, les individus ne sont que rarement réparti dans l'espace en grand dèmes panmictiques. Pour ces raisons, il est nécessaire pour ce type d'analyse de regrouper les individus, voir les groupes d'individus échantillonnés dans un "vrai" dème, dans un ensemble de cases d'une grille régulière correspondant aux dèmes du modèle d'IBD d'inférence (comme illustré sur la Figure 3.11). Dans cette situation, il n'est pas nécessairement évident de savoir à quoi correspondent les paramètres du modèle d'inférence par rapport aux paramètre à estimer (les estimands) à partir des données groupées.

Par exemple, pour les échantillons provenant d'un réseau régulier, un estimand [2] putatif pour Nm est le nombre d'immigrants dans chaque case de la grille, c'est-à-dire, la somme du nombre d'immigrants au sein de chaque dème, diminuée du nombre d'immigrants échangés entre dèmes au sein de cette case. Pour des échantillons provenant d'une population structurée de manière peu régulière, l'estimand attendu de Nm est beaucoup moins clair. Heureusement, l'estimateur de $D\sigma^2$ devrait être invariant par rapport à la taille des cases (dans les habitats linéaires, ceci est valable à condition que la distance spatiale soit toujours mesurée dans les unités d'origine et non en nombre de largeurs de cases). Pour la mutation, on peut supposer que l'estimand est la taille de la population d'une case multipliée par la probabilité

2. Un estimand est une quantité qui doit être estimée dans une analyse statistique. Ici par ex. c'est le produit Nm du modèle d'inférence avec les regroupement par cases, différent du produit Nm du modèle avec les dèmes biologiques.

de mutation. Cependant, ces prédictions ne fonctionnent pas toujours bien (détaillé dans l'annexe de Rousset & Leblois, 2012) et les effets de ce regroupement peuvent également dépendre de la distribution des échantillons dans les cases de la grille. Ils est donc très difficile dans ces situations d'étudier les performances de la méthodes sur les inférences des paramètres autres que $D\sigma^2$.

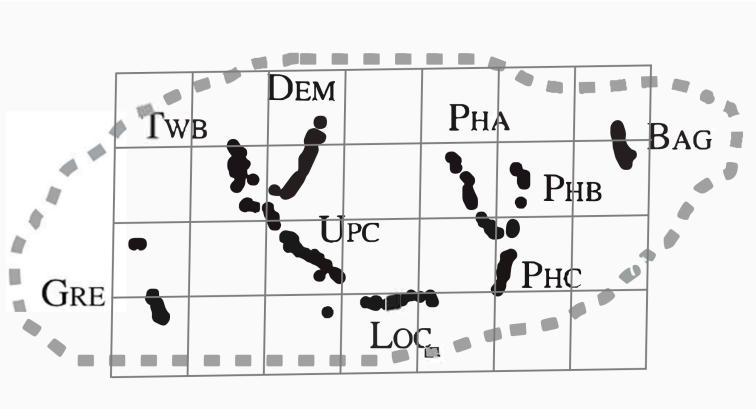


FIGURE 3.11 – Exemple de grille de 7x4 surimposée sur l'échantillonnage spatial des libellules demoiselles *Coenagrion mercuriale* sur un des sites d'étude de Watts *et al.* (2006). L'échantillonnage de ce site était déjà présenté en Figure 1.7. Une telle grille permet la modélisation de dèmes sur un échantillonage non régulier dans un habitat homogène comme utilisé dans Rousset & Leblois, 2012. Figure adaptée de Watts *et al.* (2006)

Robustesse vis à vis de la distribution de dispersion Afin d'évaluer l'effet d'une mauvaise spécification de la distribution de dispersion, une distribution de Sichel (Gamma réciproque de Poisson, Chesson & Lee, 2005, voir section 1.3.2) a été utilisée pour les simulations (voir). La valeur de σ^2 est fixée dans nos simulations en faisant varier κ pour une valeur fixe de $\gamma = -2.15$, donnant une kurtosis entre entre 20.1 et 22.5. Pour ces analyses, seuls les performances des estimations de $N\mu$ et $D\sigma^2$ sont explorées car il parait hasardeux de comparer les paramètres de forme d'une distribution géométrique avec ceux de la Sichel. Les résultats pour deux valeurs de $D\sigma^2$ de 10 et 50 ($\kappa = 0.92$ et 4.6), montrent des biais très faibles $< 5\%$, des RMSE de 0.15 à 0.3 et une bonne couverture des intervalles de confiance, même si le modèle d'inférence comporte beaucoup moins de dèmes que le modèle simulé (10x10 VS 40x40). Cependant, pour des valeurs plus forte de $D\sigma^2$ de 250 ($\kappa = 23$), il est nécessaire d'augmenter le nombre de dème à 20x20 pour le modèle d'inférence et de considérer une échelle d'échantillonage plus étendue dans le modèle de simulation (80x80 au lieu de 40x40 pour les simulations avec un σ^2 plus faible) pour avoir une bonne couverture des IC et une inférence correcte, toutefois marquée par des biais assez forts sur $D\sigma^2$, de l'ordre de 50%.

Données réelles et comparaison avec la méthode de la régression Les performance des estimations de $D\sigma^2$ par PAC-vraisemblance sous IBD a été comparée avec l'estimateur des moments de la pente de régression de $F_{ST}/(1 - F_{ST})$ ou de e_r avec le logarithme de la distance de Rousset (1997, 2000) décrit et testé en section 2.1 et 2.2. Les estimateurs de $D\sigma^2$ (ou plutot $1/4\pi D\sigma^2$ pour éviter les valeurs infinies de $D\sigma^2$) sont comparés en termes de ratio de RMSE et, comme attendu d'une méthode de vraisemblance, l'erreur de PAC-vraisemblance est plus faible que celle de la ré-

gression (ratio entre 0.25 et 0.66 selon les situations). En outre, cet écart persiste lorsque l'on considère d'autres distributions de dispersion et d'autres modèles de mutation (que ceux du modèle d'inférence) ce qui montre une robustesse certaine de l'IS-coa/PAC-vraisemblance vis à vis de ces facteurs, d'autant plus que l'on avait déjà précédemment observé une forte robustesse de la méthode de la régression à l'égard de ces facteurs en section 2.2. Enfin, le ratio d'erreur reste favorable à la PAC-vraisemblance même dans les cas ou les hypothèse du coalescent ne sont pas vérifiée (0.66 pour la situation 40x40 dèmes analysés en 20x20, $N = 50$, $m = 0.5$, $g = 0.5$, $\mu = 10^4$). En conséquence, les intervalles de confiance basés sur les moments devraient être plus larges, mais l'on observe au contraire tendance à être trop courts (Figure 3.12), comme cela a été montré précédemment (Leblois *et al.*, 2003; Watts *et al.*, 2006). L'inférence de $4\pi D\sigma^2$ par PAC-vraisemblance peut être plus robuste que par la méthode de la régression, car cette dernière ne tient pas compte des effets de bordure spatiale (c.a.d hypothèse que la population est sur un cercle ou un tore, voir section 2.1).Cette hypothèse est soutenue par l'analyse d'une situation avec de forts effets de bords (une petite grille de 10×10 dèmes avec des échantillons prélevés dans les coins), la méthode de régression a un biais d'environ 300% , alors que la méthode de PAC-vraisemblance a un biais de 30% et une couverture des IC correcte.

D'après tous ces résultats, l'analyse en PAC-vraisemblance des données originales de Watts *et al.* (2006) sur les demoiselles devrait fournir une estimation et un intervalle de confiance plus précis et plus fiables pour $4\pi D\sigma^2$ que les analyses précédentes basées sur la méthode des moments. Néanmoins, les résultats ne sont pas très différents des estimations précédentes : 1 110 [600 - 3,625] par PAC-vraisemblance contre 753 [319 - 3,162] par la régression (Watts *et al.*, 2006), mais toujours en accord avec la conclusion antérieure selon laquelle les estimations génétiques ne sont que "légèrement" plus élevées que l'estimation démographique ($4\pi D\sigma^2 = 555$, Watts *et al.* (2006)).

3.3 Conclusions sur la vraisemblance

Dans ce chapitre, nous avons vu comment la théorie de la coalescence et notamment le coalescent structuré permet d'estimer la vraisemblance d'un échantillon grâce à des algorithmes de Monte Carlo explorant l'espace des généalogies, et peut ainsi permettre l'inférence des tailles de populations et de la dispersion à partir de données génétiques échantillonnées en population naturelles.

Une première observation, sur laquelle on ne s'attardera pas, est que les temps de calculs sont un problème récurrent des ces méthodes d'inférence par vraisemblance basée sur la coalescence. Il n'y a priori pas de solution statistique ou simulatoire vraiment pertinente pour les réduire drastiquement. Les tentatives d'améliorations par échantillonnage par pont ("bridge sampling", Gelman & Meng, 1998) ne semblent pas bien adapté au calcul de la vraisemblance par coalescence (Fearnhead & Donnelly, 2001; Leblois, 2004) et le ré-échantillonnage séquentiel ("sequencial resampling", (Liu, 2004)) n'ont permit qu'une diminution très relative des temps de calculs (voir Merle *et al.*, 2017 pour une application en IS-coa). Ainsi, l'analyse d'un grand nombre de loci (par ex. quelques milliers pour les données NGS typiques d'un organisme non modèle) peut s'avérer difficile en raison des temps ,de calculs et de l'effet additif de la variance observée à chaque locus. Ces limitations expliquent la persistance et l'utilité des méthodologies alternatives en génétique des populations.

Une seconde observation, clairement attendue mais pas toujours prise en compte par les développeurs de méthodes d'analyse en génétique des populations, est que

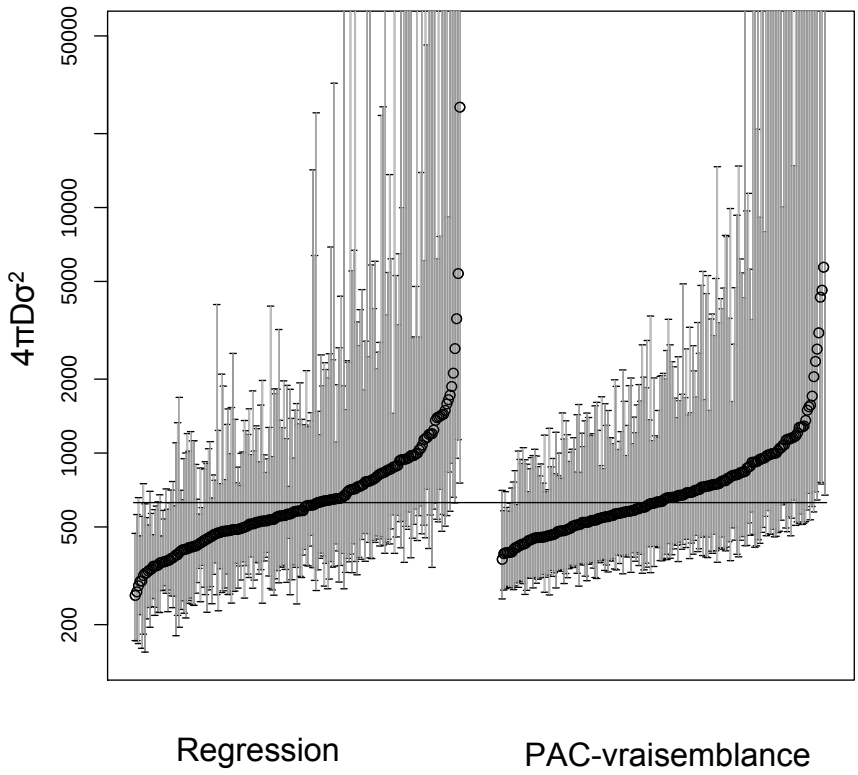


FIGURE 3.12 – Distributions des estimations et intervalles de confiance pour $4\pi D\sigma^2$, par la méthode de la régression et par PAC-vraisemblance. La situation démographique correspond à 40x40 dèmes analysés en 20x20, avec $N = 50$, $m = 0.5$, $g = 0.5$, $\mu = 10^4$. La ligne horizontale marque la valeur réelle du paramètre (Rousset & Leblois, 2012).

l'utilisation de modèle démographiques homogènes de type IBD, dans lesquels on modélise la dispersion avec un petit nombre de paramètres (Nm, et le ou les paramètres de forme de la distribution de dispersion), est plus pertinente que de considérer des modèles de structuration de populations non-homogènes (c.a.d avec des tailles de dèmes et des taux de dispersion par paires de dèmes spécifiques) et comportant donc un grand nombre de paramètres (voir section 3.2.1). En ce sens, la comparaison que l'on pourrait faire entre les performances de MIGRATE et de `Migraine` dans le cadre du modèle d'IBD en dèmes n'est pas juste puisque les mauvaises performances des algorithmes MCMC-coa de MIGRATE sont en partie dus au grand nombre de paramètres à estimer. Cette approche devrait être plus performante pour l'estimation d'un petit nombre de paramètres dans le cadre d'un modèle homogène dans l'espace. Cependant, les comparaisons plus équitables de Leblois *et al.* (2014) ont aussi montré de meilleurs performance des algorithmes IS-coa implémenté dans `Migraine` par rapport aux algorithmes MCMC-coa (implémentés dans Msvar, Beaumont (1999)). De plus, la difficulté à obtenir de bonnes estimations à partir des approches MCMC-coa ont été soulignées à divers reprises (Beerli, 2009; Kuhner, 2009). Ainsi, ces divers études ainsi que d'autres indices qui nous verrons ci-dessous nous poussent à penser que ce sera généralement le cas.

Une grande partie des observations et conclusions auxquelles nous sommes arrivées après les travaux présentés ci-dessus ne semble donc pas restreintes à des spécificités des algorithmes MCMC-coa et IS-coa considérés ici, mais avoir une portée beaucoup plus générale dans le domaine de l'inférence démographique et historique à partir de données génétiques et génomiques. La première est donc l'importance de considérer des modèles simples avec peu de paramètres, une évidence statistique pourtant souvent négligée lors de la mise au point et/ou de l'application des méthodes d'inférence en génétique des populations. La seconde est qu'il est pertinent de décrire le modèle démographique en terme de dèmes potentiellement présents dans l'habitat autour des populations étudiées plutôt qu'en restreignant le modèle aux seules populations échantillonnées. C'est à dire en essayant de considérer la configuration spatiale des populations du modèle biologique plutôt que de l'échantillon uniquement.

La troisième est que les surfaces de vraisemblance sous de nombreux modèles de "démo-génétiques" peuvent être assez complexes, c.a.d pas caractérisées par un unique pic gaussien dans les profils en 1 et 2 dimensions sur l'ensemble de l'espace des paramètres exploré. Les deux exemples représentés en figure 3.13 illustrent bien certaines de ces caractéristiques avec une situation donnant de beaux profiles bien piqués et une autre situation pourtant très proche, dans lesquels les profiles 2D montrent des surfaces plus complexes, en forme d'entonnoirs ou de croix avec un petit maximum peu marqué (voir aussi Beaumont, 1999). Ces surfaces de vraisemblances complexes sont notamment obtenues dans les modèles avec variations passées des paramètres démographiques, et sont notamment dues au fait que l'information "contenue" dans les données permettra d'estimer certaines paramètres avec plus de précision que d'autre selon la zone du paramètre (par ex. ici le temps) explorée, comme cela a été illustré en section 3.2.2 sur la Figure 3.5. Plus globalement, un échantillon génétique ne contiendra jamais de l'information sur tous les paramètres de n'importe quel modèle, mais uniquement pour certaines combinaisons de paramètres, pour certaines valeurs et sous certaines modèles, dépendant de la configuration de l'échantillon et de son histoire évolutive, qui va fortement limiter la précision de l'inférence de certains paramètres dans de nombreuses situations. Ce dernier point, assez trivial mais pas toujours bien pris en compte par les empiristes,

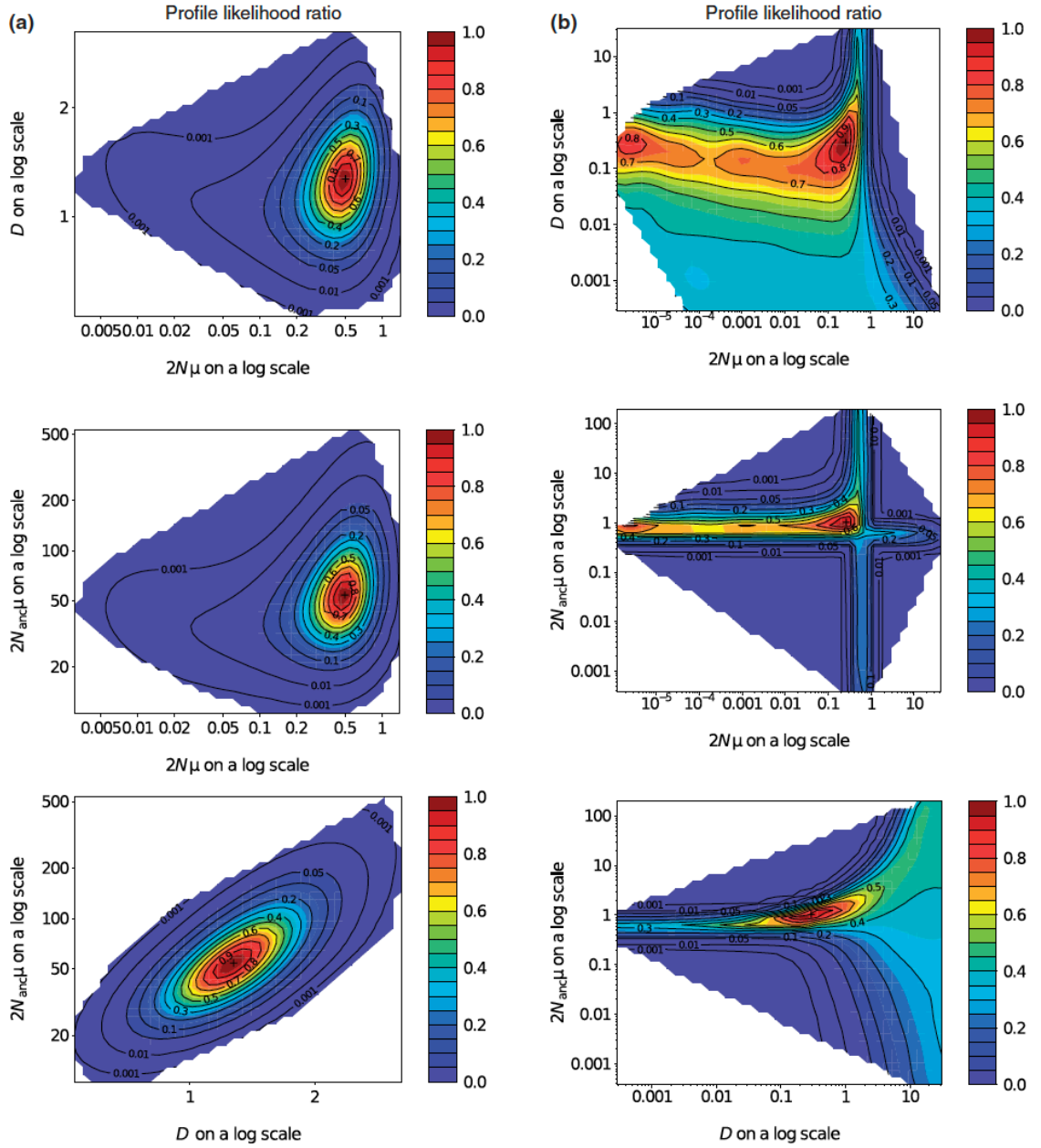


FIGURE 3.13 – Exemples typiques de profiles de rapports de vraisemblance en 2D dans le modèle d'une population avec un changement passé de taille de population pour deux situations contrastées. Les deux situations démographiques simulées correspondent à : (a) un cas de changement fort (facteur 100), ni trop récent, ni trop ancien, dans lequel les données "contiennent" de l'information sur tous les paramètres du modèle, caractérisé par $\theta = 0.4$, $D = 1.25$ et $\theta_{anc} = 40$; et (b) un cas ou le changement est de trop faible amplitude (facteur 5) dans lequel les données ne contiennent que peu d'information sur les différents paramètres, caractérisé par $\theta = 0.4$, $D = 1.25$ et $\theta_{anc} = 2$. La surface de vraisemblance est inférée à partir de de 1 240 points en deux étapes itératives (a), et de 3 720 points en trois étapes itératives (b), comme décrit dans la section 3.2.2 décrivant `Migraine`. Figure issue de Leblois *et al.* (2014)

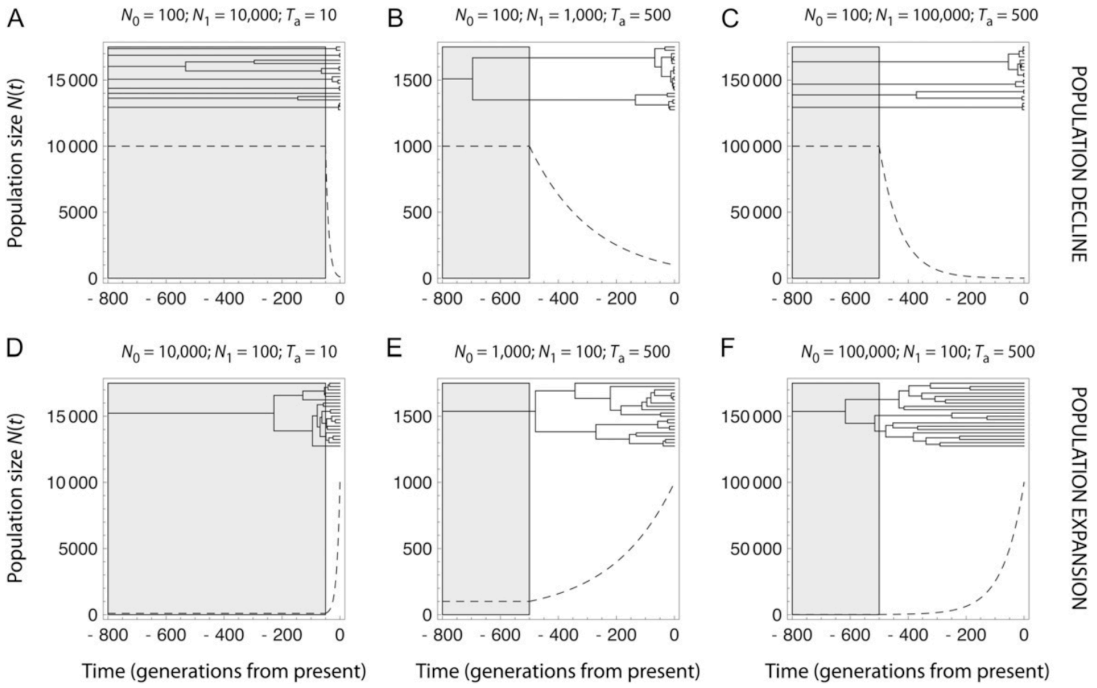


FIGURE 3.14 – **Illustration de différentes dynamique des changements de taille d'une population $N(t)$ et des généalogies attendues.** A-C correspond à une réduction de la taille de population du passé vers le présent et D-F à augmentation du passé vers le présent (courbe en pointillés). La zone ombrée de chaque graphique indique que la taille de la population ancestrale est constante. Sur chacune de ces courbe, la généalogie attendue pour 20 lignées échantillonnées est représentée. Les généalogies attendues ont été obtenues en calculant la moyenne des temps de coalescence sur 500 000 simulations de chaque scénario démographique. Notons que certaines généalogies sont incomplètes (A et C), certaines lignées n'ayant pas coalescé à 800 générations du présent. Figure issue de Girod *et al.* (2011).

constitue la quatrième observation très générale de nos études. C'est un point crucial de la génétique des populations empiriques, qui peut facilement être intuité grâce à des raisonnements simples de coalescence, rapidement présenté en section 1.1, mis en avant dans Girod *et al.* (2011) et Leblois *et al.* (2014) et régulièrement repris dans les chapitres précédents. La Figure 3.14 illustre bien ce principe : l'information contenue dans les données sur les différents paramètres d'un modèle démographique provient des évènements de coalescence et de mutation (et de recombinaison) affectant les lignées ancestrales, ainsi que de leurs "places" dans les différents "compartiments" du modèle (par ex. les dèmes dans un modèle en îles, les coordonnées spatiales dans un modèle d'IBD ou comme ici les différentes zones temporelles dans un modèle avec variations dans le temps de paramètres démographiques). Plus il y a de lignées ancestrales dans un de ces compartiments, plus on aura une estimation précise des paramètres "impliqués" dans ce compartiment, et vice-versa.

En cinquième, les différents tests et applications des méthodes MCMC-coa et IS-coa ont globalement montré qu'il est difficile d'explorer efficacement par simulation de Monte Carlo l'espace des généalogies pour estimer la vraisemblance d'un échantillon génétique avec une faible variance. Leblois *et al.* (2014) montre aussi que l'amplitude de la variance d'estimation de la vraisemblance peut fortement varier en fonction de la zone de l'espace des paramètres explorées. Ces deux facteurs, combinés à la forme parfois complexe des surfaces de vraisemblances, font qu'il est a priori

plus facile d'inférer correctement les estimateurs de maximum de vraisemblance (ou équivalent en Bayesien), des intervalles de confiance (ou de crédibilité), ou encore la surface globale de vraisemblance avec des algorithmes de lissage prenant bien en compte la variance des points observés lors de l'interpolation, plutôt qu'avec des algorithmes de type MCMC ou EM. UN PEU LEGER à consolider...

En sixième et dernier point, les approximations du coalescent de grandes tailles de populations et petits taux de dispersion, sous-jacentes à toutes ces méthodes, peuvent non seulement largement biaiser les inférences de taux de dispersion et des tailles de dèmes, mais aussi en compliquer fortement l'interprétation lorsque l'échantillon analysé provient de populations naturelles dans lesquelles la notion de dème n'est biologiquement évidente. Ces observations n'ont été faites clairement que dans le cas des tests minutieux des algorithmes IS-coa, mais toute approche basée sur les mêmes approximations de coalescence devrait montrer les mêmes limitations. Ainsi, les approches MCMC-coa, mais aussi toutes les méthodes d'inférence basées sur une distributions instrumentales des fréquences alléliques découlant des approximations de diffusion, comme la formule de diffusion de Wright (1931), devraient être affectées par le même type de biais et de problèmes d'interprétation des paramètres estimés. C'est ce qu'ont observé (Faubet *et al.*, 2007, p. 1160) en évaluant la méthode de Wilson & Rannala (2003) sur des échantillons simulé avec de petites tailles de populations et de forts taux de migration. Seule l'inférence de $D\sigma^2$ est robuste à ces approximations, point important sur lequel nous reviendrons pour conclure ce chapitre. Et un corollaire de cette estimation robuste de $D\sigma^2$ et de l'estimation non robuste du Nm est que les algorithmes basés sur ces approximations du coalescent ne sont pas les plus appropriés pour inférer la forme de la distribution de dispersion. Elles devraient en effet se limiter à des scénarios avec de faible taux de dispersion, ce qui restreindrait fortement leur utilité.

En plus de ces observations assez générales, le développement, test et application de ces méthodes par vraisemblance dans le cadre de modèles structurées de population nous a permis d'avancer concrètement sur l'inférence des paramètres locaux et actuels de taille de population, de densité et de dispersion à partir d'un échantillon génétique en population naturelles.

Ces tests intensifs et détaillés de `Migraine` illustrent à la fois les points forts et les imperfections de ces inférences : dans la plupart des cas, les estimateurs ont un faible biais et, compte tenu de la taille relativement petite des échantillons considérés, une faible MSE. Les biais de mauvaise spécification des processus mutationnels sont relativement faibles et faciles à comprendre, mais moins pour les paramètres de dispersion. Ainsi, seul le paramètre composite $D\sigma^2$ est relativement peu affecté par une mauvaise spécification (i) du nombre de dèmes du modèle biologique, (ii) de la configuration en dème de l'échantillon, (iii) de la distribution de dispersion et par (iv) les approximations du coalescent. Au contraire, toute mauvaise spécification de la distribution de dispersion ainsi que de la répartition spatiale des dèmes peuvent fausser l'estimation des paramètres de dispersion de manière complexe, et les rendre difficilement interprétable. Enfin, la comparaison avec la méthode de la régression de Rousset (1997) et Rousset (2000) montre que l'estimation de $D\sigma^2$ basée sur la vraisemblance est nettement plus précise et que ses intervalles de confiance sont plus fiables, même lorsque l'on combine les nombreux facteurs de complications discuté ci-dessus.

Une caractéristique distinctive de ce dernier travail, par rapport à la plupart des publications sur les méthodes alternatives d'inférence, est l'accent mis sur l'évalua-

tion de l'inférence en termes de propriétés de couverture des intervalles de confiance fondés sur la vraisemblance. Ceci a été permis par la procédure de Krigeage que l'on a implémenté dans `Migraine`et, comme nous l'avons vu dans ce chapitre, cette approche aide à l'interprétation des performances d'une inférence. De plus, à nos yeux, de bonnes propriétés de couverture des IC est, en combinaison avec une bonne analyse de la précision attendue et de la robustesse à différents facteurs, la meilleure façon de valider une méthode d'inférence.

Pour aller plus loin, il parait donc pertinent de continuer à utiliser les modèles IBD qui semblent assez bien adapté à l'inférence des processus démographiques à petite échelles géographiques et évolutives, mais aussi de considérer des approches ne faisant pas d'approximations en terme de grandes tailles de populations et/ou faibles taux de dispersion. La problématique sera alors de trouver des méthodes d'inférences qui puissent considérer ce type de modèles IBD avec de petites tailles de dèmes et de forts taux de dispersion, tout en utilisant au maximum l'information présente dans les données.

Chapitre 4

Inférences par simulation : le graal de la génétique spatiale ?

Dans le chapitre précédent, nous avons montré comment la coalescence permet d'estimer la vraisemblance d'un jeu de données génétiques, et peut donc ensuite être utilisée pour faire de l'inférence par vraisemblance des paramètres démographiques à partir de données génétiques. Nous avons ensuite montré qu'un de ces algorithme, l'IS-coa implémenté dans notre logiciel `Migraine` pour faire de l'inférence sous isolement par la distance (IBD), permettait des inférences plus précises de $D\sigma^2$ que la méthode des moments précédemment testée dans le second chapitre, et ceci malgré quelques limitations importantes. En effet, nos tests ont aussi montré que l'application de l'inférence IS-coa en IBD sur des données réelles était fortement limitée par les approximations du n-coalescent sous-jacentes à l'estimation de la vraisemblance. En effet, seul le paramètre $D\sigma^2$ est bien estimé et interprétable quand : (i) la migration est forte, et/ou (ii) la répartitions spatiale des individus dans la population ne définit pas clairement de dèmes réguliers, ce qui correspond à la majorité des situations biologiques.

Pour aller plus loin dans l'inférence de paramètres de dispersion et de densité, il est donc nécessaire de s'affranchir de ces approximations du n-coalescent, tout en essayant d'utiliser au maximum l'information des données. Cela pourrait notamment permettre (i) d'estimer d'autres paramètres du modèle, en plus de $D\sigma^2$, tels que le taux d'émigration m (ou Nm le nombre d'émigrants) et/ou des paramètres de forme de la distribution de dispersion, et (ii) mieux prendre en compte la répartition spatiale des individus en se basant sur un modèle d'IBD en habitat continu, afin de ne pas avoir à regrouper les individus en dèmes sans grande signification biologique.

Dans ce chapitre, j'introduirai tout d'abord les principes de l'inférence par simulation, en me focalisant sur les approches ABC ("Approximate Bayesian Computations", Beaumont *et al.*, 2002). Je présenterai ensuite les outils permettant de mettre en oeuvre, en pratique, l'inférence par simulation sous IBD et d'en tester les performances. Pour cela j'introduirai une nouvelle méthode alternative à l'ABC, la méthode de la vraisemblance résumée ("summary Likelihood" `SL`, Rousset *et al.*, 2017). Ces travaux étant en cours, je présenterai juste une illustration de ces inférences ainsi que quelques résultats préliminaires de tests de performance. Un des buts de ces travaux est d'utiliser l'information contenue dans le déséquilibre de liaison. Le lecteur devra cependant attendre quelques mois pour en savoir plus à ce propos car les tests concernant l'information présente dans le déséquilibre de liaison n'ont pas encore été fait :). De même, nous souhaitons comparer les performances des méthodes ABC et vraisemblance résumée. Cette comparaison ne tardera pas à

être faite mais nous n'avons à ce jour aucun résultat à vous présenter.

4.1 Inférences ABC par simulation

Comme nous l'avons vu dans le chapitre précédent, l'inférence basée sur la vraisemblance est limitée par (i) la difficulté d'estimer vraisemblance sous des modèles de populations structurées, (ii) les forts temps de calculs, et (iii) la nécessité d'utiliser des approches basées sur les approximations du n-coalescent. Les deux premiers facteurs ont stimulé, dès les années 90-2000, le développement en génétique des populations de nouvelles méthodes d'inférences se passant de la vraisemblance mais cherchant à l'approcher (Marjoram & Tavaré, 2006).

L'inférence par simulation est évoquée pour la première fois dans Diggle & Gratton (1984) dans le cadre de simulations de Monte Carlo et dans un contexte purement statistique et théorique. Elle est ensuite appliquée en génétique des populations dans Tavaré *et al.* (1997) et Pritchard *et al.* (1999), sous la forme d'un algorithme de rejet, assez éloigné du travail initial de Diggle & Gratton (1984). Cet algorithme de rejet est à la base des approches ABC. Il consiste tout simplement à (i) simuler un grand nombre de jeux de données, avec les mêmes caractéristiques que le jeu de données réelles ("observé") que l'on veut analyser (c.a.d le même nombre d'individus et de locus, et le même type de marqueurs) dans le cadre d'un scénario d'évolution hypothétique. Dans un cadre Bayesien (sur lequel repose la totalité des développements de type ABC), les paramètres des jeux de données simulés sont échantillonnés à partir de distributions a priori définies par l'utilisateur ; (ii) les données simulées sont ensuite réduites à un ensemble de statistiques, dites "statistiques résumantes", et (iii) les paramètres échantillonnés pour chaque jeu données sont acceptés ou rejetés sur la base d'une distance maximale (le seuil ϵ, souvent appelé tolérance) entre les statistiques résumantes simulées et observées ; enfin (iv) la distribution marginale à posteriori de chaque paramètre est construite à partir du sous-échantillon des valeurs acceptées de paramètres.

Les différentes approches ABC sont des versions améliorées de cet algorithme de rejet (Beaumont *et al.*, 2002; Blum & François, 2010; Marjoram *et al.*, 2003; Sisson *et al.*, 2007). Une des premières améliorations les plus significative est la correction de l'écart entre les statistiques simulées et observées en utilisant des techniques de régression locale telle que représenté en Figure 4.1 définie par Beaumont *et al.* (2002) puis par l'utilisation d'un réseau de neurones à la place du seuil de distance ϵ par Blum & François (2010). Ces deux alternatives au rejet simple engendrent une moindre influence du seuil ϵ sur les distributions a posteriori, et permettent ainsi d'obtenir de bonnes inférences à partir d'un plus petit nombre de simulations. Quelques tentatives de combinaison de l'approche ABC avec d'autres approches statistiques telles que (i) les simulations de Monte Carlo par chaînes de Markov pour explorer l'espace des paramètres de manière itérative en utilisant la distance entre les statistiques résumantes simulées et observées pour mettre à jour les valeurs actuelles des paramètres (par ex. Marjoram *et al.*, 2003; Wegmann *et al.*, 2009), mais la convergence est toujours difficile à vérifier (Marjoram & Tavaré, 2006) ; ou encore (ii) par échantillonage séquenciel (Sisson *et al.*, 2007), mais les résultat de montrent pas d'amélioration importante.

L'inférence ABC repose donc sur l'analyse incomplète des données, résumées dans un ensemble de statistiques. Ainsi, plus l'information du jeu de données sera (mal) résumée dans cet ensemble de statistiques résumantes, plus la distribution a posteriori sera différente, et notamment plus proche de la distribution a priori, que

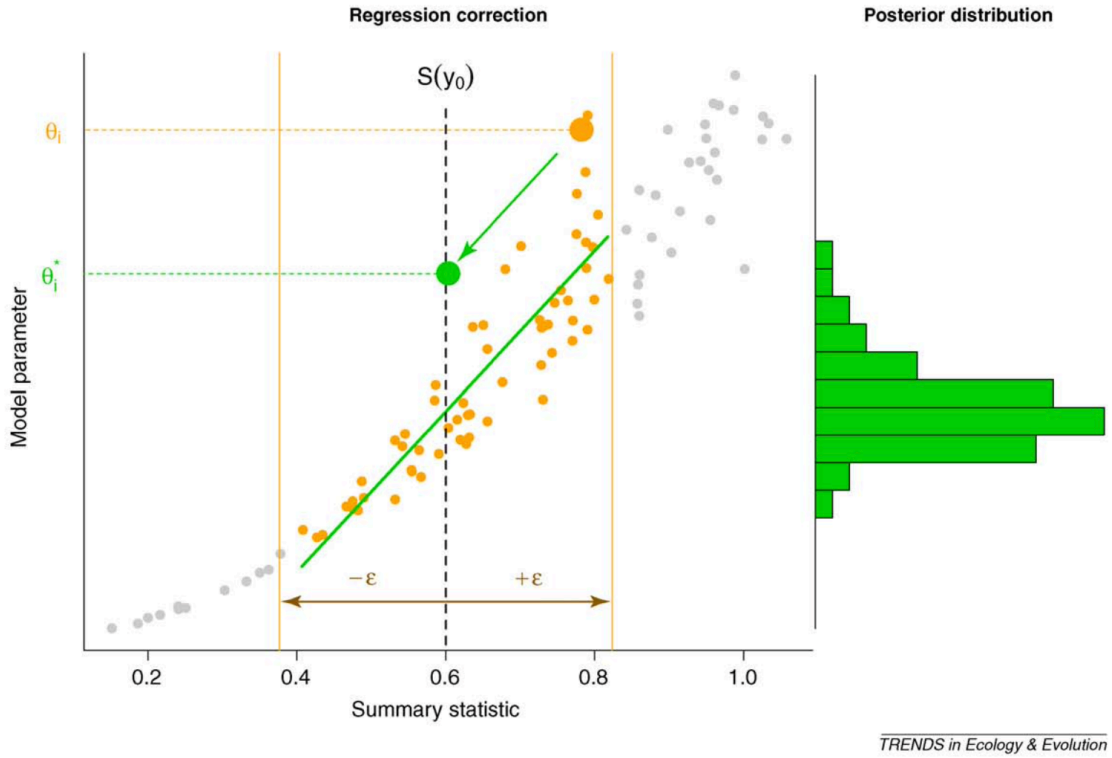


FIGURE 4.1 – Ajustement par régression linéaire dans l'algorithme ABC. Dans l'algorithme ABC, des valeurs de chaque paramètre θ_i du modèle sont tirées dans des distributions a priori spécifiées par l'utilisateur afin de simuler un ensemble de données d'apprentissage y_i dans le cadre d'un modèle donné. A partir de ces données simulées, des valeurs de statistique résumante, $S(y_i)$ (une ici mais plusieurs en pratique) sont calculées, et comparées aux valeurs calculées sur les données réelles, $S(y_0)$, à l'aide d'une mesure de distance. Si la distance entre $S(y_0)$ et $S(y_i)$ est inférieure à une certaine "tolérance" ϵ, la valeur du paramètre, θ_i, est acceptée (c.a.d sera prise en compte pour le calcul de la distribution a posteriori). Ce graphique montre comment les valeurs acceptées de θ_i (points en orange) sont ajustées selon une transformation linéaire, $\theta_i^* = \theta_i - b(S(y) - S(y_0))$ (flèche verte), où b est la pente de la droite de régression. Après ajustement, les nouvelles valeurs des paramètres (histogramme vert) forment un échantillon de la distribution a posteriori. Figure issue de Csilléry *et al.* (2010).

la distribution a posteriori attendue par vraisemblance sur l'ensemble des données. Le choix des statistiques résumantes est donc crucial en ABC et doit être réfléchi par rapport aux questions d'inférence particulières abordées. Il est intéressant de noter ici que l'inférence ABC a été initiée et largement développée en génétique des populations, puis s'est ensuite étendue à d'autre domaines. C'est, entre autre et comme nous l'avons en introduction et dans le premier chapitre, parce que la génétique des populations s'est longtemps concentrée sur le développement de statistiques informatives sur les paramètres d'intérêt des modèles considérés (tels que les F-statistiques, voir section 1.2.2 et chapitre 2).

Une façon "simple" de palier cette perte d'information des données serait d'augmenter le nombre de statistiques résumantes. Cependant, cela pose rapidement (pour plus de quelques dizaines de statistiques) des problèmes d'analyse en grande dimension, notamment pour les approches par distances, et peut rapidement réduire la précision de l'inférence (Beaumont *et al.*, 2002). En effet, la probabilité d'accepter une simulation donnée diminue exponentiellement avec la dimensionnalité. Pour contourner ce problème, divers techniques de sélection de statistiques (Joyce & Marjoram, 2008) ou de réduction de dimensions (Estoup *et al.*, 2012) ont été utilisés. Parmi celles-ci, on distingue notamment l'approche par réseaux de neurones de Blum & François (2010). En effet, les réseaux de neurones, et plus généralement les méthodes d'apprentissage (dont l'ABC fait partie...), peuvent également trouver les "combinaisons de statistiques résumantes" qui contiennent le maximum d'informations sur les paramètres d'intérêt. Nous reviendrons la dessus par la suite car c'est un champ en pleine explosion et prometteur dans le cadre général de l'inférence par simulation.

Dans ce contexte, une amélioration majeure des approches ABC est apportée par Pudlo *et al.* (2016) et Raynal *et al.* (2018) qui utilisent l'algorithme d'apprentissage automatique par forêts aléatoires ("random forests" RF, Breiman, 2001), pour sélectionner les simulations correspondant au mieux aux données observées. L'approche par forêts aléatoires (RF) proposée par Breiman est l'un des principaux algorithmes d'apprentissage automatique supervisé ("supervised machine learning", SML) pour la classification (par ex. pour faire du choix de modèle) ou la régression (par ex. pour l'estimation de paramètres continus). L'approche `ABC-RF` de Pudlo *et al.* (2016) implémente des procédure de choix de modèles et surpasse les autres méthodes ABC pour les analyses basées sur un petit nombre de simulations (Fraimout *et al.*, 2017; Pudlo *et al.*, 2016). S'appuyant sur ces résultats, Raynal *et al.* (2018) ont récemment proposé une extension de l'approche `ABC-RF` dans un cadre de régression non paramétrique pour caractériser les distributions a posteriori des paramètres d'intérêt dans le cadre d'un modèle donné. La méthode `ABC-RF` de Raynal *et al.* (2018) permet notamment : (i) une amélioration importante de la robustesse au choix des statistiques résumantes (notamment vis à vis de l'inclusion de statistiques "non informatives"), (ii) l'absence de considération d'une valeur seuil ϵ pour comparer les statistiques simulées et observées, et (iii) un bon compromis entre les temps de calcul et la précision des estimations ponctuelles des paramètres et l'exactitude des intervalles de crédibilité (qui à mes yeux reste toutefois à tester précisément :).

Assez récemment aussi, une méthode alternative d'inférence par simulation, fondée sur l'utilisation de la densité jointe des statistiques et des paramètres pour inférer la surface de vraisemblance des paramètres sachant les statistiques résumantes observées sur l'échantillon, a été développée (méthode dite par vraisemblance résumée, `summary likelihood SL`, Rousset *et al.*, 2017). Même si cette méthode a certains points communs avec l'`ABC-RF` (elle peut partir des mêmes simulations que les mé-

thodes ABC et elle utilise aussi les forêts aléatoires mais dans un autre contexte statistique), elle est philosophiquement beaucoup plus proche de l'approche originale de Diggle & Gratton (1984). Le nombre minimal de statistiques nécessaires pour l'estimation de p paramètres étant p statistiques, une première étape de la méthode `summary likelihood` consiste à réduire par RF le grand nombre de statistiques résumantes calculées sur un jeu de données en p statistiques résumantes *projetées* conçues pour conserver le maximum d'information pour chaque paramètre. La deuxième étape de la méthode `summary likelihood` diverge de l'ABC en général : plutôt que l'obtention de distributions a posteriori, `summary likelihood` vise à inférer une surface de vraisemblance des p paramètres, sachant les p statistiques projetées observées sur l'échantillon. Les techniques classiques pour l'inférence par maximum de vraisemblance peuvent alors être appliquées comme nous l'avons vu dans le chapitre 3. De plus, la vraisemblance est conçue conjointement pour tous les paramètres, contrairement à l'ABC (dont l'`ABC-RF`) qui ne permet pour l'instant d'obtenir que des distributions marginales a posteriori pour chacun des paramètres. C'est la méthode que nous présentons ci-dessous en détails pour l'inférence par simulation des paramètres de densité, dispersion et taille de population sous un modèle d'IBD.

4.2 Inférence par SL en IBD

Pour contrôler l'ensemble du processus d'inférence par simulation, nous avons développé un simulateur `GSpace`, une librairie de calcul de statistiques résumantes `GSumStat`, et couplé les deux à la librairie R `Infusion` qui implémente la méthode `SL` pour l'inférence.

4.2.1 `GSpace`

Le pipeline d'inférence par simulation développé repose donc sur un nouveau simulateur `GSpace` (Virgoulay *et al.*, 2021), qui remplacera à terme `IBDsim` (Leblois *et al.*, 2009) dont les principales limitations sont de ne pas prendre en compte la recombinaison ni la notion d'individus, et donc de ne pouvoir simuler que des marqueurs indépendants et de la migration gamétique. Dans le contexte des NGS et de l'étude de la dispersion, ce sont maintenant deux facteurs fortement limitants.

Algorithme de coalescence Comme `IBDsim`, `GSpace` est fondé sur un algorithme de coalescence génération par génération, et ne repose donc pas sur les approximations du coalescent. Il se base sur le cycle de vie vu précédemment que l'on a représenté en remontant le temps pour coller à la suite des évènements considérés dans une approche par coalescence en Figure 4.2. `GSpace` combine certaines parties de l'algorithme modifié de Hudson (Hudson 1983) pour la recombinaison et la coalescence, précédemment implémentées dans `MSprime` (Kelleher *et al.* 2016), avec de nombreuses caractéristiques de l'algorithme génération par génération d'`IBDsim` (décrit en détails dans Leblois *et al.* 2003, 2009). Une différence notable avec `IBDsim`, du point de vue la dispersion, est donc que des marqueurs non liés physiquement (c.a.d. distant sur le même chromosome, ou sur deux chromosomes différents) vont avoir des arbres de coalescence non indépendants puisque conditionnels à l'arbre généalogique réalisé des individus échantillonnés en fonction des évènements de migration et de reproduction réalisés.

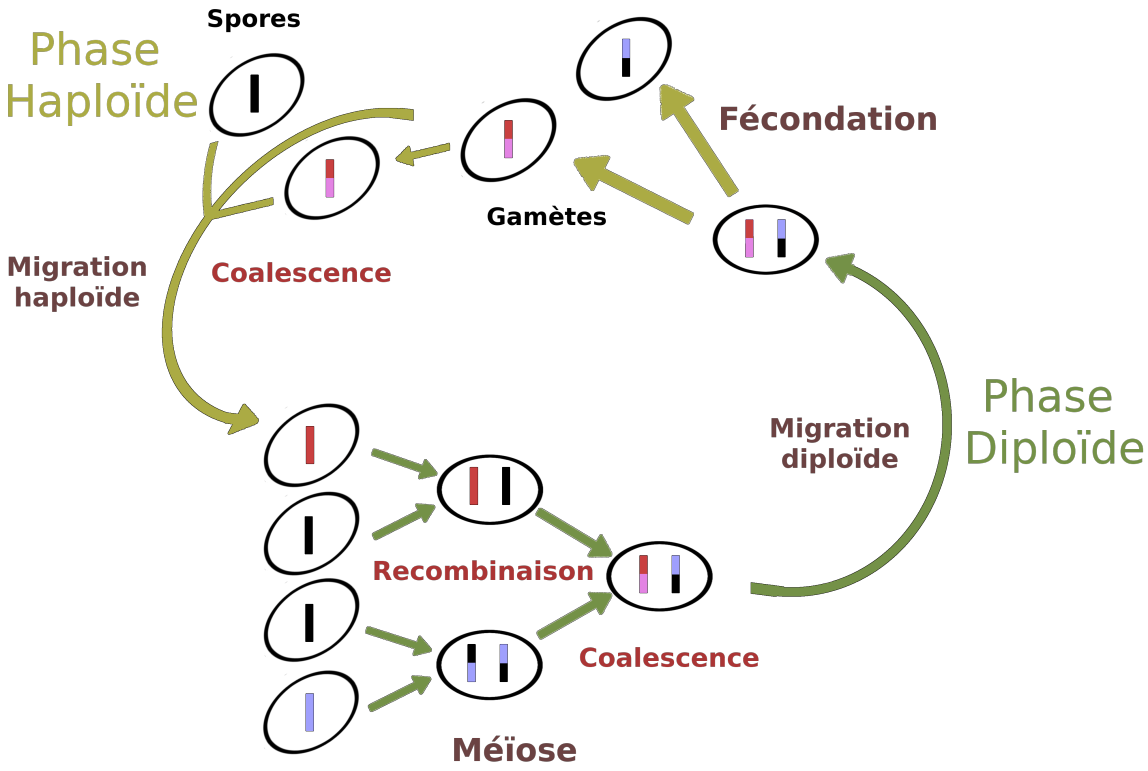


FIGURE 4.2 – Cycle de vie "haplo-diploïde" représentant, **en remontant le temps**, la recombinaison et la coalescence des lignées génétiques au cours des différentes phases possibles d'un organisme. Les lignées ancestrales échantillonnées sont représentés en rouge et bleu. Les lignées noires représentent les lignées non échantillonnées. La lignée issue de la coalescence des lignées rouge et bleu est représentée en fuchsia.

Locus, chromosomes, individus et sexes explicites Une spécificité de `GSpace` est qu'il prend en compte explicitement la notion de chromosomes homologues afin de gérer correctement les évènements multiples de recombinaison. Cette prise en compte passe notamment par la représentation explicite des chromosomes homologues dans l'algorithme en génération par génération même si un de ces chromosomes ne porte pas de lignées ancestrales (chromosome fantôme) lors de la phase diploïde. Elle implique aussi la prise en compte explicite des deux chromosomes parentaux lors de la reproduction sexuée comme indiqué sur la Figure 4.3, ce qui n'est pas nécessaire pour des algorithmes basés sur les approximations du coalescent puisqu'ils négligent tout évènement multiple (voir section 1.1). Les probabilités de coalescence des lignées ancestrales deviennent donc conditionnelles à l'origine parentale du chromosome qui les portent. Cela revient a choisir d'abord le parent ancêtre avec une probabilité de $1/N$, puis le chromosome ancestral chez ce parent avec une probabilité de $1/2$, au lieu de choisir directement le chromosome ancestral avec une probabilité de $1/2N$ comme sous le n-coalescent.

Cette prise en compte des chromosomes homologues permet de faire migrer ces derniers de manière non indépendant au moment de la phase diploïde du cycle, et donc de modéliser de façon plus pertinente la dispersion de juvéniles diploïdes, plus courante que la migration gamétique chez la majorité des espèces (animales) à phase diploïde majoritaire. Cette implémentation ouvre la voix à la considération d'autres types de dispersion, telles que la dispersion pollen-graine chez les plantes.

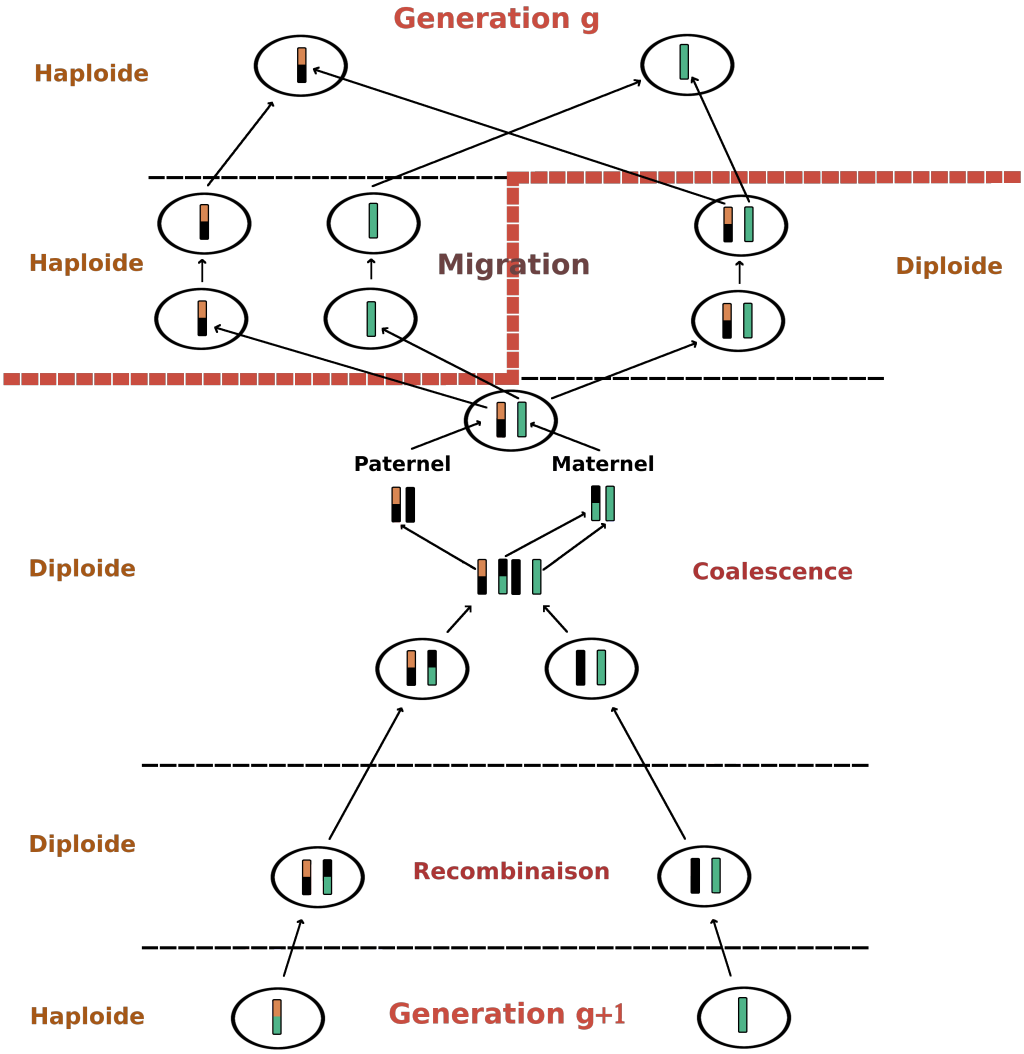


FIGURE 4.3 – Représentation détaillée du cycle de la figure 4.2 comme considéré dans l'algorithme de `GSpace`. Les lignées ancestrales échantillonnées sont représentées en vert (lignée portée par le chromosome maternel) et orange (lignée portée par le chromosome paternel).

Dispersion Comme dans `IBDsim`, l'algorithme de migration utilisé dans `GSpace` considère, qu'à chaque génération, les coordonnées du dème parental (i.e. de naissance) de chaque individu porteur de lignée(s) ancestrale(s) sont tirées aléatoirement dans la distribution de dispersion arrière donnant la probabilité de la position des parents, étant donné la position du descendant sur le réseau. Ces distributions de dispersion arrières sont calculées en fonction de la distribution avant spécifiée par l'utilisateur, avec les équations 1.1.2 et 1.1.2. `GSpace` peut considérer des distributions avant de dispersion de type Pareto/Zeta discrétisée, ainsi que la distribution de Sichel qui permet de modéliser une grande diversité de distributions de dispersion, et les distribution plus classiques : uniforme, gaussienne discrétisée, et géométrique (voir section 1.3.2 pour les détails sur ces différentes distribution et les caractéristiques attendues de la dispersion en populations naturelles).

`GSpace` a été comparé en terme de temps de calcul à d'autres simulateurs, et les résultats montrent qu'il est tout à fait compétitif (Virgoulay *et al.*, 2021).

4.2.2 `GSumStat`

Une spécificité de `GSumStat` est que l'on cherche à limiter l'influence des données manquantes afin de limiter les biais d'estimation, notamment pour de faibles tailles d'échantillons (bien que ce ne soit pas crucial dans une approche d'inférence par simulation tant que les mêmes statistiques sont calculées sur les données réelles et sur les données simulées). Les estimateurs multilocus des F-statistiques sont donc calculés selon les même formules que dans `Genepop` (Rousset, 2008), dans la lignée de la définition des F-statistiques en termes de probabilités d'identité, comme nous l'avons vu en section 1.2. Des détails sur ces estimateurs peuvent être trouvés dans l'annexe de Rousset (2001).

Statistiques spatiales Comme nous l'avons vu en section 2.2, la méthode de la régression de Rousset (1997, 2000) utilise l'inverse de la pente de la régression entre un estimateur de la différenciation ($F_{\mathrm{ST}}/(1-F_{\mathrm{ST}})$, a_{r} ou e_{r}) et de la distance géographique (ou son logarithme) comme estimateur du produit $D\sigma^2$. Cette statistique résumante paraît donc très intéressante, d'autant plus que la comparaison avec l'estimateur par vraisemblance de Rousset & Leblois (2012) montre que la pente de la régression peut préserver l'essentiel de l'information statistique concernant $D\sigma^2$ (Rousset & Leblois, 2012, et voir section 3.2.2).

Nous avons donc considéré comme statistiques résumantes les pentes et intercepts des régressions des estimateurs de $linF_{\mathrm{st}} \equiv F_{\mathrm{ST}}/(1-F_{\mathrm{ST}})$, a_{r} et e_{r} décrivant la différenciation entre dèmes ou entre individus, classiquement utilisées dans la méthode de la régression (comme vu en section 2.2, et définies dans Rousset, 1997 pour $linF_{\mathrm{st}}$, Rousset, 2000 pour a_{r}, et Watts *et al.*, 2006 pour e_{r}). En particulier e_{r}, que nous n'avons jamais détaillé jusqu'ici, est une distance génétique entre deux individus diploïdes, dans même nature que a_{r}, et calculée comme

$$e_{\mathrm{r}} = \frac{\widehat{Q}_i + \widehat{Q}_j - \widehat{Q}_{ij}}{1-\widehat{Q}_w} - \frac{\widehat{L_s}}{1-\widehat{Q}_w} \tag{4.1}$$

où $\widehat{Q}_{ij}$ est la fréquence de paires de gènes identiques sur toutes les paires d'individus (i,j) séparées par une distance géographique r ; $\widehat{Q}_i$ (resp. j) est la fréquence de

paires de gènes identiques au sein de l'individu i (resp. j) (c'est donc la fréquence de loci homozygotes au sein d'un individu) ; $\widehat{Q}_w$ est la fréquence de paires de gènes identiques au sein d'un individu sur l'ensemble des individus échantillonnés et $\widehat{L_s}$ le terme constant (par rapport à i, j) identifié par Watts *et al.* (2006) dans la statistique de Loiselle *et al.* (1995), qui est calculé comme suit :

$$\widehat{L_s} = \frac{n_p(\text{mean}(\widehat{Q}_{ij})) + n(1/2 + \widehat{Q}_w/2)}{n_p + n} \tag{4.2}$$

où n est le nombre total d'individus échantillonnés et n_p le nombre de paires d'individus.

`GSumStat` calcule aussi les probabilités d'identité Q_r entre paires de gènes pour différentes classes de distance r (voir section 1.2.1).

Statistiques de déséquilibre de liaison Dans ce chapitre, nous souhaitions notamment tester l'intérêt d'utiliser l'information du déséquilibre de liaison pour estimer les paramètres de dispersion et de densité. Les statistiques classiques de déséquilibre de liaison reposant souvent sur le calcul de fréquences de génotypes et de fréquences alléliques, elles ne sont pas directement utilisable à une autre échelle que celle de la population globale dans les modèles sans dèmes. Nous aurions pu nous intéresser aux distributions de longueurs d'haplotypes partagé entre deux haplotypes comme dans Ringbauer *et al.* (2017), et nous le ferons sans doute bientôt. Pour ce travail, nous nous sommes concentrés sur des mesures de déséquilibre de liaison analogues aux approches par probabilités d'identité décrites en section 1.2. En effet, les associations non aléatoires entre allèles à différents loci (c.a.d le déséquilibre de liaison) peuvent être décrites en terme de probabilités d'identité jointe à deux locus. Le détail de cette approche, fondée sur les travaux de Vitalis & Couvet (2001b,a), est présenté dans Virgoulay (2022). et je n'en donnerai que les grandes lignes,permettant de faire le lien avec ce que l'on a vu précédemment.

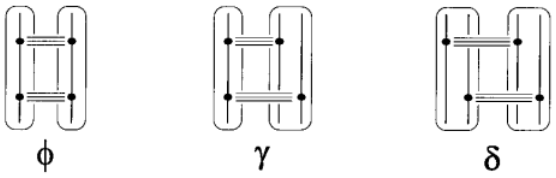


FIGURE 4.4 – Définitions des probabilités d'identités à deux locus ϕ, γ, δ impliquant respectivement deux, trois ou quatre haplotypes des deux génotypes diploides comparés. Figure issue de Vitalis & Couvet, 2001b.

Nous nous sommes inspirés de la statistique de déséquilibre d'identité entre deux marqueurs définie par Vitalis & Couvet (2001b) comme

$$\frac{\Phi_{ij} - Q_{\text{wd}_i} Q_{\text{wd}_j}}{(1 - Q_{\text{bd}_i})(1 - Q_{\text{bd}_j})} \tag{4.3}$$

où Q_{wd_i} et Q_{wd_j} représentent les probabilités d'identité aux loci i et j entre haplotypes au sein d'un dème ; et Q_{bd_i} et Q_{bd_j} pour des dèmes distincts ; et Φ_{ij} est la probabilité d'identité jointe aux loci i et j entre deux haplotypes pris dans un même dème.

Pour que ces mesures soient valides sur de données non-phasées (c.a.d. pour lesquelles on ne connaît pas les haplotypes des doubles hétérozygotes), nous avons

considéré les statistiques de déséquilibre "génotypique" définies à partir de fréquences de paires de gènes identiques qui peuvent être observées sans connaître la phase, et à la probabilité d'identité correspondante $\Phi := (\phi + 2\gamma + \delta)/4$ (voir figure 4.4 et Weir & Cockerham, 1974).

Une mesure de déséquilibre de liaison tenant compte de la distance géographique entre individus peut alors être définie comme suit :

$$\widehat{\eta}_{xy,ij} \equiv \frac{\widehat{\Phi}_{xy,ij} - \widehat{Q_{\mathrm{wd}_i}}\widehat{Q_{\mathrm{wd}_j}}}{(1 - \widehat{Q_{\mathrm{bd}_i}})(1 - \widehat{Q_{\mathrm{bd}_j}})} \tag{4.4}$$

où $\widehat{\Phi}_{xy,ij}$ est la fréquence des paires d'haplotypes, produites par tirage d'un haplotype dans le génotype diploïde de chacun de deux individus pris dans les dèmes x et y, conjointement identiques aux loci i et j ; $\widehat{Q_{\mathrm{wd}_i}}$ est la fréquence d'identité intra-dème observée au locus i, pour les individus échantillonnés dans le dème x et le dème y ; et $\widehat{Q_{\mathrm{bd}_i}}$ est la fréquence d'identité observée entre dèmes au locus i calculée sur tous les individus échantillonnés.

Pour résumer l'information contenue dans les $\widehat{\eta}_{xy,ij}$ en fonction de la distance entre marqueurs sur le chromosome et en fonction de la distance géographique entre dèmes ou individus, on calcule d'abord une estimation unique $\widehat{\eta}_{d_g,d_c}$ pour toutes les paires de loci à distance chromosomique d_c dans des paires d'individus à distance géographique d_g. Cette estimation est la somme des numérateurs des $\widehat{\eta}_{xy,ij}$ concernés, divisée par la somme des dénominateurs des mêmes $\widehat{\eta}$. On ajuste ensuite aux $\widehat{\eta}_{d_g,d_c}$ par moindres carrés un modèle de régression non linéaire de forme

$$\eta_{d_g,d_c} = a + (b - a)\exp\left[-c_g d_g - c_c \rho(d_c)\right] \tag{4.5}$$

où $\rho(d_c)$ est la probabilité de recombinaison entre les loci [1], déduite de la distance chromosomique, en inversant la fonction de Haldane (1919) : $\rho(d_c) = [1 - \exp(-2d_c)]/2$. Il y a donc quatre paramètres dans cette régression a, b, c_g, c_c qui constituent les statistiques résumantes des patrons de déséquilibre de liaison dans l'échantillon. Ce modèle de régression n'a pas de justification théorique précise, mais sa forme parait raisonnable au vu de simulations préliminaires visant à identifier un modèle de régression approprié.

A ce jour, `GSumStat` calcule donc la moyenne et la variance sur l'ensemble des locus (ou sites variables) de l'échantillon d'un ensemble des statistiques plus ou moins classiques :

- N_{a} : Le nombre d'allèles dans l'échantillon ;
- $N_{\mathrm{a\,d}}$: Le nombre moyen d'allèles dans chaque dème ;
- H_{o}, H_{e} : L'hétérozygotie observée et attendue (ou diversité génétique, Nei, 1973) ;
- Var : La variance de la taille allélique, pour les marqueurs multi-alléliques ;
- AFS/SFS : Le spectre de fréquences alléliques (voir Ewens 1972) ou, de manière équivalente, le spectre de fréquence par site pour les marqueurs bi-alléliques ;

1. la probabilité qu'un gamète pris au hasard porte à ces loci des copies de gènes des deux haplotypes parentaux

- F_{ST}, F_{IS} et F_{IT} : les F-statistiques au sein et entre individus/dème, précédemment présentés en section 1.2.2 ;
- les pentes et ordonnées à l'origine des régressions linéaires entre $linF_{\mathrm{st}} = F_{\mathrm{ST}}/(1 - F_{\mathrm{ST}})$, a_{r} et e_{r} et la distance géographique (ou son logarithme) ;
- les probabilités d'identité de paires de gènes à un locus Q_{r} pour un nombre de classes de distance géographiques fixées ;
- et les 4 paramètres de la régression exponentielle de η avec les distances chromosomiques et géographiques.

4.2.3 Le pipeline d'inférence `gspace2infr`

La librairie R `gspace2infr` interface et relie les outils dédiés aux différentes étapes de l'inférence par simulation, notamment les programmes de simulation d'échantillons génétiques `IBDsim` et `GSpace`, le calcul de statistiques résumantes par la librairie `GSumStat`, et l'inférence statistique proprement dite par la méthode `summary likelihood` présentée ci-dessus, et implémentée dans la librairie R `Infusion`. Du code pour intégrer la méthode `ABC-RF` de Pudlo *et al.* (2016) et Raynal *et al.* (2018) sera aussi bientôt intégré pour comparer ses deux approches d'inférence par simulation. Enfin, la libraire `gspace2infr` permet de faire des tests de performances sur un grand nombre d'inférences sur données simulées à paramètres constants pour évaluer ces procédures d'inférence, comme nous l'avons vu dans les chapitres précédents (voir section 4.2.4).

Un exemple d'inférence par `summary likelihood`

Cette section présente un exemple d'inférence par `summary likelihood` utilisant le pipeline implémenté dans `gspace2infr`, afin de donner une idée plus concrète des méthodes utilisées et de la nature des résultats produits.

Scénario biologique Le scénario biologique considéré repose sur le cycle de vie avec dispersion juvénile et les modèles IBD vu précédemment (p.4 et section 2.1, respectivement). La simulation est effectuée par `GSpace` ou `IBDsim` en considérant un habitat homogène carré modélisé par une grille de 70×70, avec un unique couple d'individus diploïdes par nœud. Le taux d'émigration est de 0.25 et la distribution de dispersion suit une loi géométrique de paramètre $g = 0.558615$ et de distance maximum de 20 pas (ce qui donne $\sigma^2 = 4$, voir section 3.2.2 pour la dispersion géométrique).

L'échantillon est constitué de l'ensemble des couples d'une zone centrale carrée de 10×10. Pour chaque individu, les génotypes de 500 loci bialléliques sont simulés. Ces loci sont répartis sur 10 chromosomes, à raison de 50 loci par chromosome, à intervalles réguliers en termes de probabilité de recombinaison ($r = 10^{-5}$ entre chaque locus). La probabilité de mutation par locus et par gamète est $\mu = 5.10^{-5}$.

La librairie R `Infusion` est ensuite utilisé pour l'inférence par simulation. Le modèle de simulation et d'inférence ne reposant pas sur les approximations du n-coalescent, l'inférence peut se faire en terme de paramètres canoniques du modèles n_x, μ, m, g, les autres paramètres étant constants car supposé connus. Cependant, il existe jusqu'à maintenant une incertitude totale sur le fait que l'on puisse estimer indépendamment et correctement chacun ces paramètres (typiquement non estimables

		Bornes d'estimation	
Paramètre	Valeur simulée	minimum	maximum
n_x	70	10	300
μ	5.10^{-5}	5.10^{-7}	5.10^{-2}
Densité	...	...	...
m	0.25	0.01	0.99
g	0.558615	0.01	0.99

TABLE 4.1 – Valeurs des paramètres variables utilisées dans la simulation et explorées dans l'inférence.

Paramètre	Valeur
Dist disp max	20
Indiv par dème (N)	2
Ploïdie	2
Nb dèmes échant	10x10
Indiv échant par dème	2
Nb chr par indiv	10
Nb marq par chr	50
Modèle mutation	KAM, $K = 2$
ρ	5.10^{-5}

TABLE 4.2 – Paramètres fixés dans la simulation et dans l'inférence.

sous les hypothèses du n-coalescent). Ainsi, nous considérerons aussi l'inférence des paramètres composites $\theta = 2N_{\mathrm{T}}\mu = 2n_x^2 N\mu$, le nombre $2Nm$ d'émigrants et le produit $D\sigma^2$. S'il n'y a pas d'information pour un paramètre, l'inférence devrait nous l'indiquer sous la forme d'un profil de vraisemblance (résumée) plat pour ce paramètre.

La `summary likelihood` est une méthode d'inférence consistant à approximer la surface de vraisemblance des statistiques résumantes calculées sur des données simulées par rapport aux paramètres du modèle, puis à utiliser cette surface de "vraisemblance résumée" afin d'en trouver le maximum, de calculer des intervalles de confiance et de représenter des profiles de vraisemblance en une et deux dimension. L'inférence se déroule de la manière suivante.

Construction du tableau de référence On appellera tableau de référence les données d'apprentissage. Il est donc composé d'une ligne par simulation, résumée par ses valeurs de paramètres et les statistiques résumantes associées. La construction du tableau de référence initial s'effectue par tirages uniformes (ou log-uniforme si l'on veut explorer l'espace d'un paramètre sur une échelle logarithmique) indépendants dans les bornes de l'espace des paramètres défini par l'utilisateur (tableau 4.1).

Réduction du nombre de statistiques résumantes La méthode `summary likelihood` construit ensuite par projection une statistique synthétique pour chacun des paramètres à inférer. On cherche donc ici à construire un prédicteur de chaque paramètre à partir des statistiques résumantes par apprentissage de la relation entre valeurs de

paramètres et statistiques résumantes, à partir du tableau de référence. La méthode d'apprentissage utilisée par défaut dans `Infusion` est la régression non paramétrique par forêts aléatoires ("Random Forest", RF), implémentée dans la librairie R `ranger` (Wright & Ziegler, 2017). On exploite donc ici comme dans la méthode `ABC-RF` la facilité d'utilisation et l'efficacité de cette méthode RF à faire le tri entre les statistiques porteuses d'informations sur le paramètre à estimer et celles qui n'apportent que peu ou pas d'information. La statistique synthétique résultant de cette projection est ainsi censée résumer avec peu de perte l'information disponible sur chaque paramètre dans l'ensemble des statistiques résumantes.

L'information conservée dans chaque statistique synthétique peut être visualisée par la dispersion de la relation entre valeurs de paramètres et valeurs prédites, qui donne une première idée de la capacité à inférer chaque paramètre avec précision. La figure 4.5 présente ces relations pour les données du tableau de référence (toujours en utilisant les prédictions *out-of-bag*). L'on y voit par exemple que m devrait être estimé assez précisément, et inversement que n_x ne devrait pas l'être.

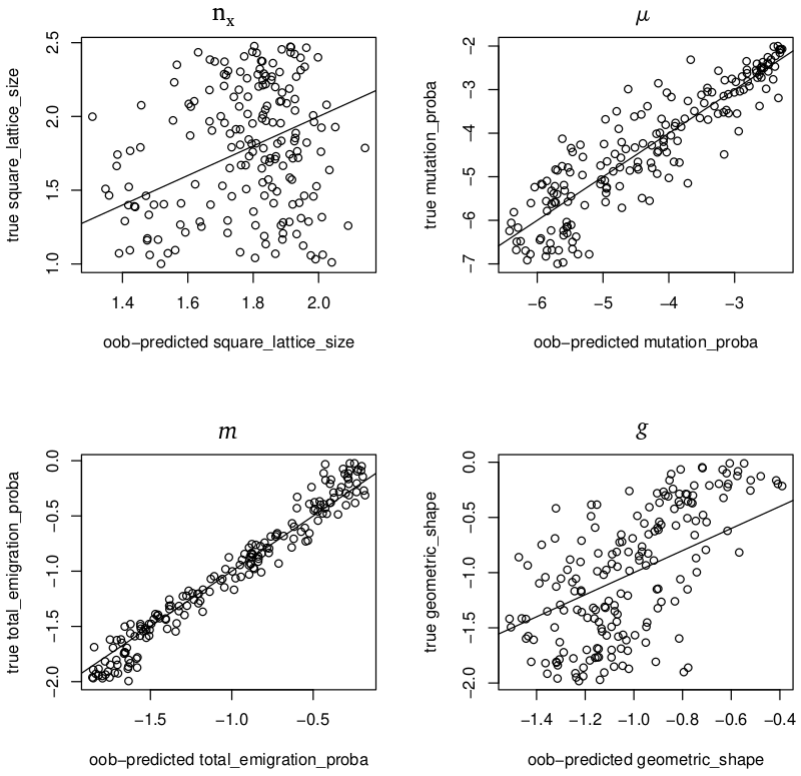


FIGURE 4.5 – Valeurs prédites des paramètres (de haut gauche à bas droite : n_x, μ, m et g) par régression non paramétrique par forêts aléatoires. La droite est la diagonale $y = x$.

Estimation de la surface de vraisemblance résumée Cette étape consiste à estimer, à partir du tableau de référence dans lequel les statistiques résumantes originales ont été remplacées par les p statistiques projetées, la distribution jointe des paramètres et des statistiques projetées, et d'en déduire une estimation de la surface de vraisemblance pour le jeu de données analysé réduit à ses statistiques résumantes projetées. La distribution jointe (en dimension 8 dans cet exemple) est inférée comme un mélange de gaussiennes multivariées (en dimension 8, donc). La librairie R `Rmixmod` (Lebret *et al.*, 2015) est utilisée par `Infusion` pour ajuster ces modèles de mélange sur une gamme de valeurs candidates du nombre d'éléments gaussiens considérés, sélectionné par comparaison des AICs de ces différents ajustements.

A partir de cette distribution jointe estimée, la densité estimée des "données" (résumées dans les statistiques projetées) pour chaque valeur de paramètres est calculée par `Infusion` par application de la formule de Bayes, en divisant la densité jointe par la densité marginale des paramètres, elle même déduite de la densité jointe. Cette densité estimée des données pour chaque valeur du vecteur de paramètres, vue comme une fonction des paramètres, constitue une estimation de la surface de vraisemblance des paramètres sachant les statistiques résumantes projetées calculées sur l'échantillon analysé.

Inférences à partir de la surface de vraisemblance résumée A partir de cette surface de "vraisemblance résumée" estimée, `Infusion` détermine son maximum, et calcul des intervalles de confiance par profilage [2] (comme nous l'avons vu pour `Migraine` en section 3.2.2). `Infusion` peut aussi effectuer une forme de bootstrap pour estimer l'incertitude sur ces différents résultats (MSE du $\mathcal{P}_{\text{MLE}}$ et des bornes des IC), ainsi que produire différentes représentations de la surface de vraisemblance, notamment des représentations des profils de rapport de vraisemblance de chaque paramètre (voir figure 4.6) et des profils de rapport de vraisemblance de paires de paramètres (voir figure 4.7).

On peut remarquer ici que le profil de vraisemblance est très plat pour le paramètre n_x contrairement à celui de m qui est beaucoup plus piqué. Ceci est cohérent avec les figures diagnostiques des projections.

Affinage de l'estimation de la surface de vraisemblance résumée Selon le même principe que celui implémenté dans `Migraine` (voir section 3.2.2), `Infusion` fonctionne par itérations successives pour affiner l'inférence. Ainsi, il est possible (même automatique) de relancer les étapes précédentes (i) en tirant de nouvelles valeurs de paramètres dans des zones d'intérêt choisies automatiquement par `Infusion`, (ii) en simulant de nouveau le processus biologique pour ces nouvelles valeurs de paramètres, et (iii) en les ajoutant au tableau de référence existant. Ces zones d'intérêt se trouvent notamment à proximité du maximum de vraisemblance, aux abords des bornes des intervalles de confiance, ou dans des zones où la vraisemblance est mal estimée. Il s'ensuit que la densité simulée des paramètres évolue au cours de l'analyse et est uniquement définie dans un but d'estimation précise de la région haute de la surface de vraisemblance (et en aucun comme une distribution a priori).

Ces itérations successives permettent :

— d'entraîner la méthode de projection avec un plus grand nombre de données la rendant potentiellement plus précise, particulièrement dans la région de paramètres la plus intéressante (Fig. 4.8) ;
— de mieux estimer la surface de vraisemblance et donc de potentiellement améliorer les estimations des paramètres.

Au fur et à mesure des itérations d'affinage, l'estimation de la surface de vraisemblance s'améliore, et on observe que les profils de vraisemblance deviennent plus piqués même si le profil de n_x reste beaucoup plus plat que celui de m (Fig. 4.9, notez les différences d'échelles des ordonnées).

Inférence de paramètres composites L'inférence peut aussi s'effectuer sur le même tableau de référence en modifiant l'espace des paramètres explorés, combinant des

2. le profilage ("profiling") désigne le fait de s'intéresser à la variation de la vraisemblance sur un ou deux axes de l'espace des paramètres, en maximisant la vraisemblance sur les autres dimensions. On s'intéresse donc à la vraisemblance d'un ou deux paramètre en fixant la valeur des autres paramètre à leur optimum.

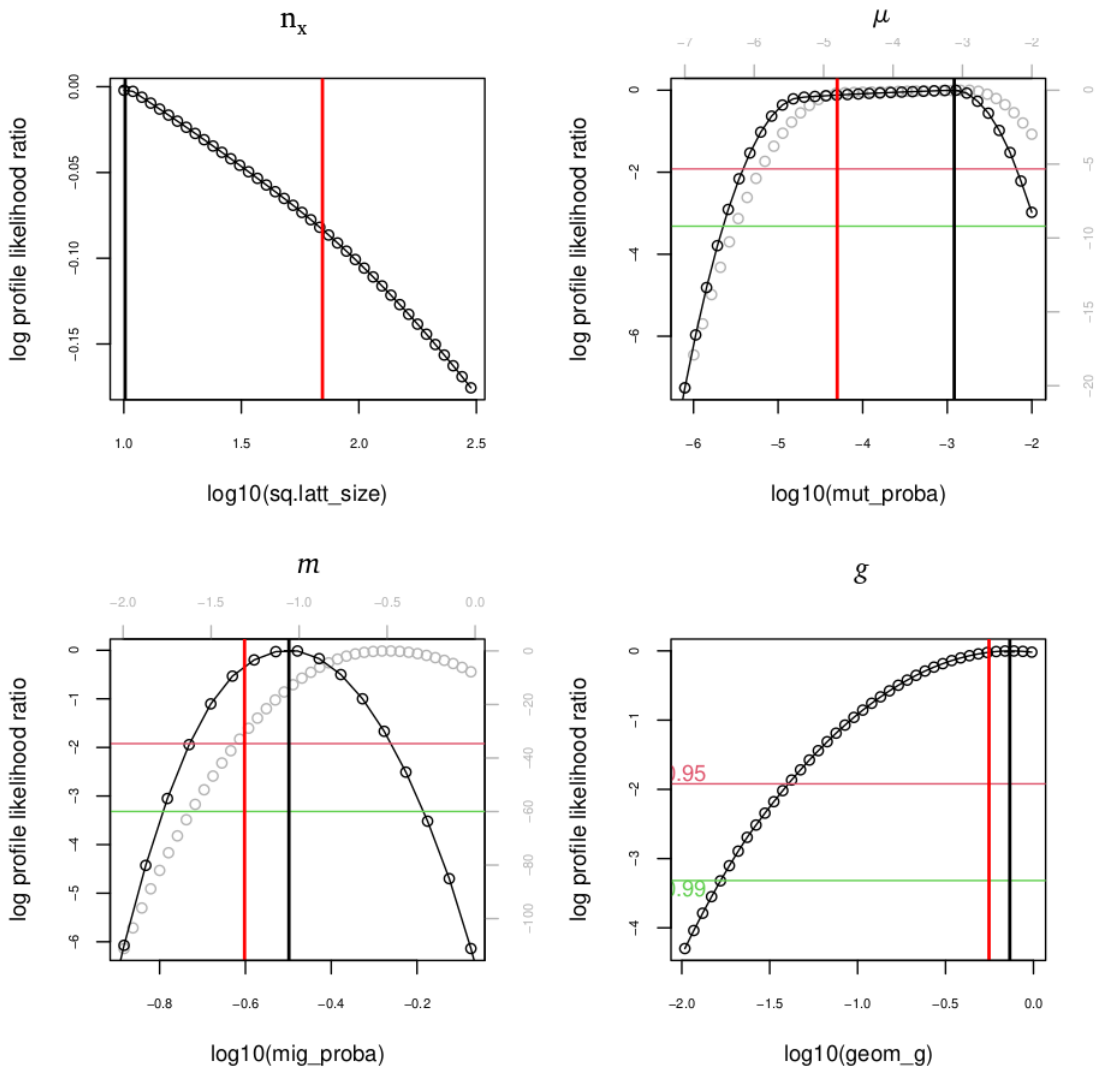
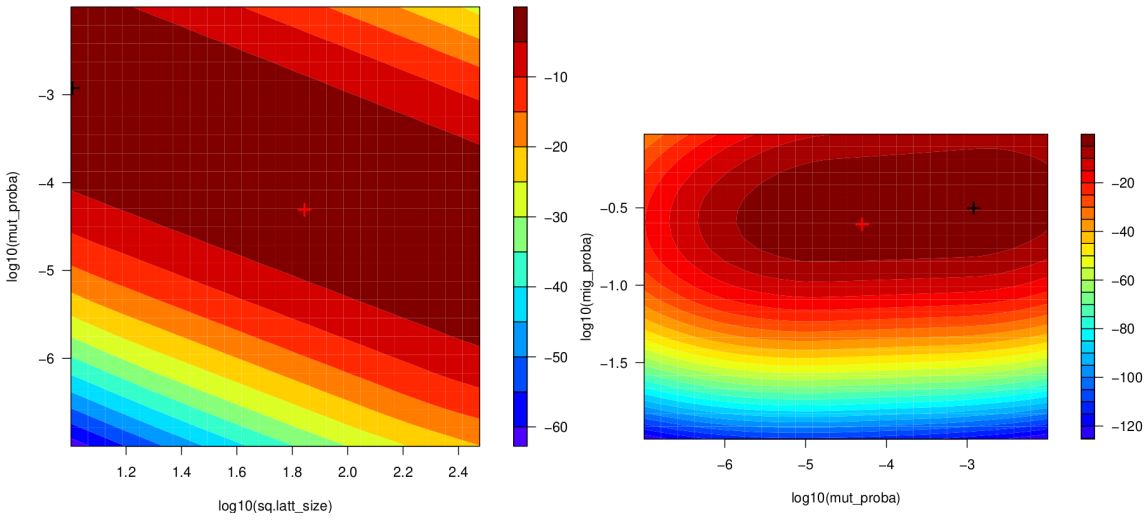
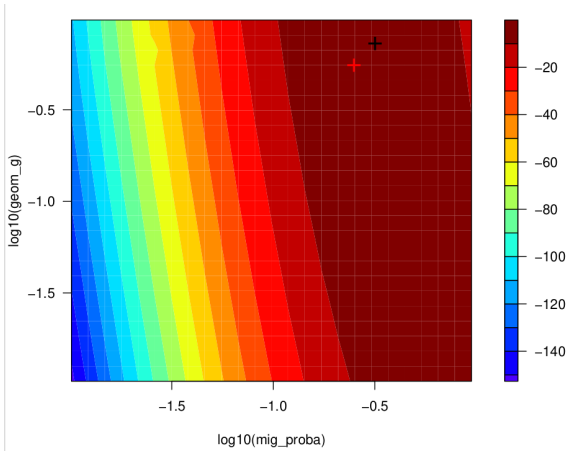


FIGURE 4.6 – Profiles LRT 1D pour chacun des paramètres inférés. Les points gris et l'échelle grise correspondante représentent l'ensemble de l'intervalle exploré. Les points noirs et l'échelle noire correspondante représentent un zoom sur la zone autour du maximum de vraisemblance estimé. Les traits rouges et verts représentent (quand ils sont calculables) respectivement les intervalles de confiance à 95% et à 90%. Ils ne s'appliquent qu'à la zone agrandie de maximum de vraisemblance. Le trait rouge vertical représente la valeur réelle de chaque paramètre utilisée pour simuler le jeu de données test, et le trait noir la valeur estimée par maximisation de cette vraisemblance.

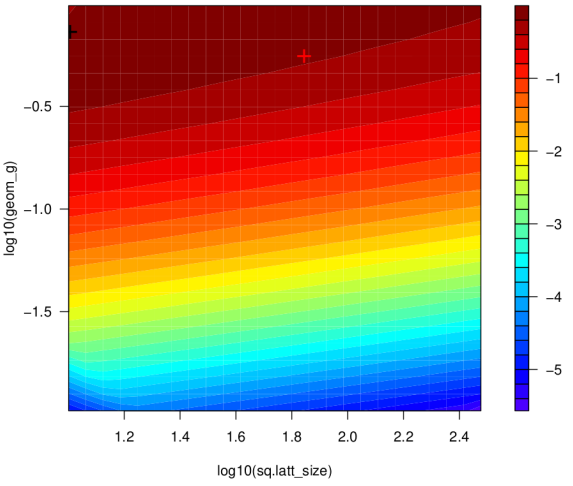


(a) n_x et μ

(b) μ et m

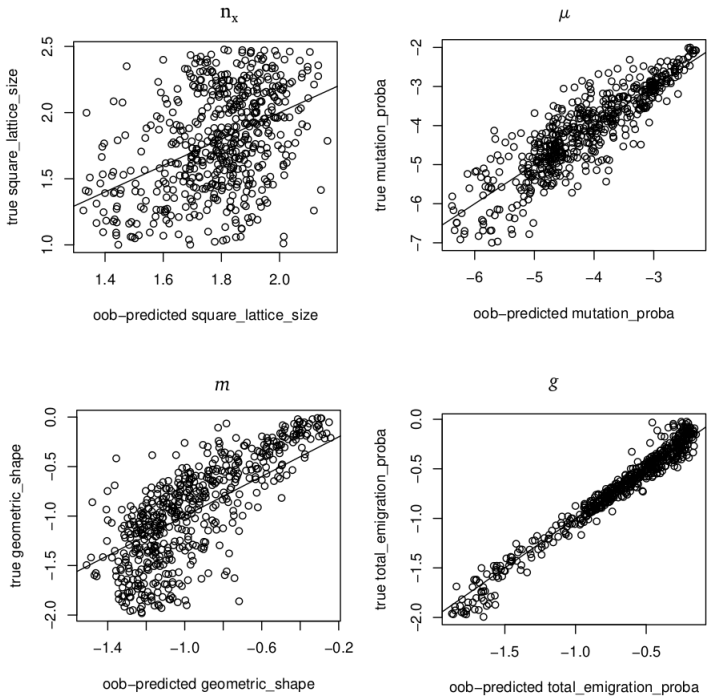


(c) m et g

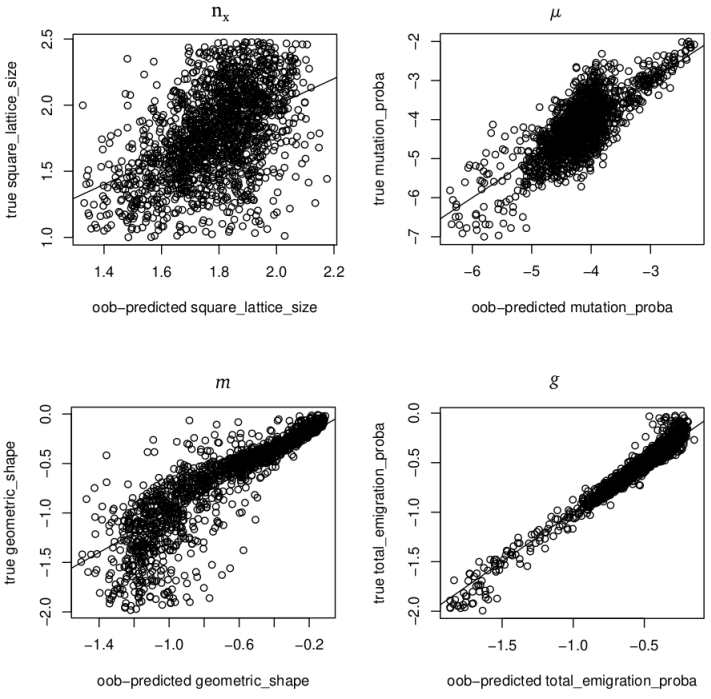


(d) n_x et g

FIGURE 4.7 – Profiles LRT 2D pour chaque paire de paramètres. La croix rouge représente les valeurs réelles de la paire de paramètres utilisées pour simuler le jeu de données test, et la croix noire les valeurs estimées par maximisation de la vraisemblance.

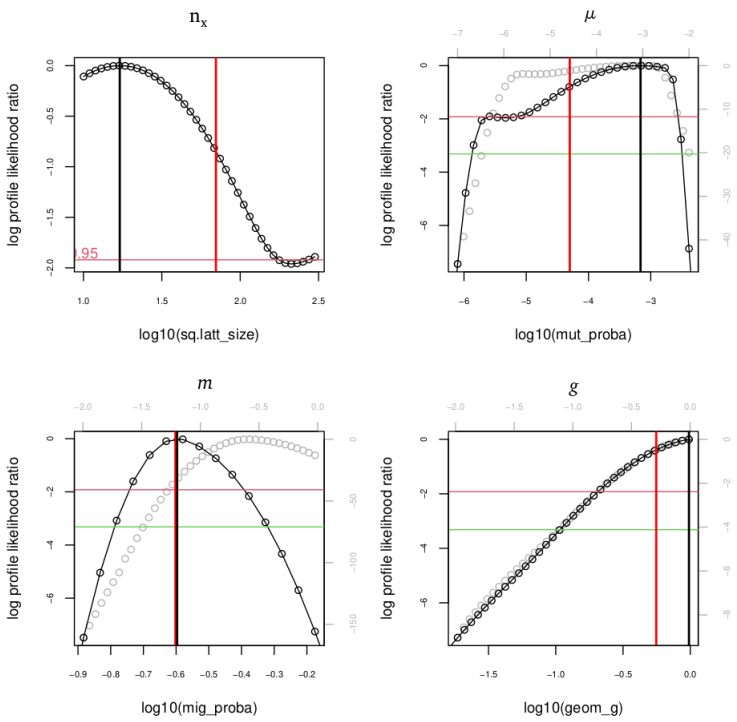


(a) 3ème itération

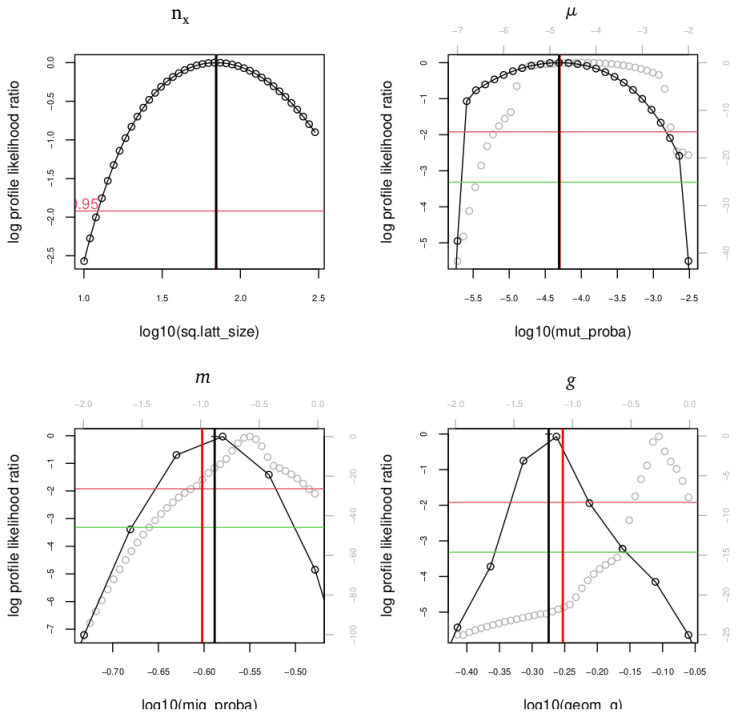


(b) 8ème itération

FIGURE 4.8 – Valeurs prédites des paramètres par régression RF en fonction des itérations d'affinage. La droite est la diagonale $y = x$.



(a) 3ème itération



(b) 8ème itération

FIGURE 4.9 – Profiles LRT 1D selon les itérations d'affinage.

paramètres canoniques et des paramètres composites (à nombre total de paramètres constant, et en ne considérant pas de paramètres trop corrélés). On peut par exemple considérer l'inférence des paramètre composite θ et $D\sigma^2$, en combinaison avec n_x et m. Les intervalles de confiance pour les paramètres composites sont calculés par rapport de vraisemblance comme pour les paramètres non composites.

Comme le montre le profil de vraisemblance bidimensionnel pour la taille de l'habitat et la probabilité de mutation (figure 4.10), le produit $\theta = 2n_x^2 N\mu$ (constant le long de la diagonale descendante sur cette figure en échelle des logarithmes des paramètres) est plus facilement estimable que chaque paramètre indépendamment. La crête de la surface de vraisemblance, visible le long d'une telle diagonale, se traduit en effet par des profils plats pour chaque valeur des paramètres du produit, alors qu'en se déplaçant orthogonalement à cette crête, on obtiendrait un profil bien plus variable pour le produit. C'est ce que l'on voit sur le profile de θ et μ sur la Figure 4.11.

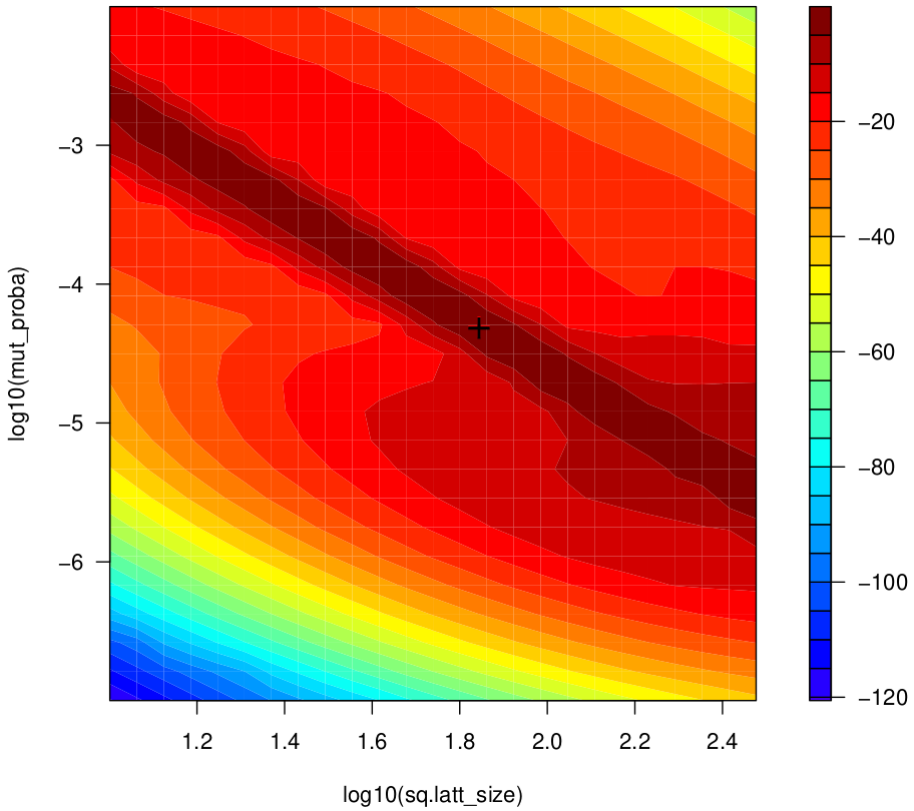


FIGURE 4.10 – Profiles LRT 2D pour n_x et μ à la $8^{\text{ème}}$ itération d'affinage.

4.2.4 Premiers tests de performance

Tests préliminaires à densité fixe

Les premiers tests d'inférence ont été réalisés dans un modèle ou seuls les paramètres (canoniques) n_x, μ, m, g varient, comme dans l'exemple que l'on vient de voir (d'où les ... pour la densité dans le tableau 4.1 :). C'est à dire que l'on suppose que la densité est fixe et connue, ce qui complètement irréaliste en pratique. Ces premiers tests ont toutefois permis de mettre en avant quelques points intéressants.

Déséquilibre gamétique entre marqueurs non-liés, Un premier résultat intéressant est apparu en comparant les inférences basées sur la simulation par `GSpace` et par

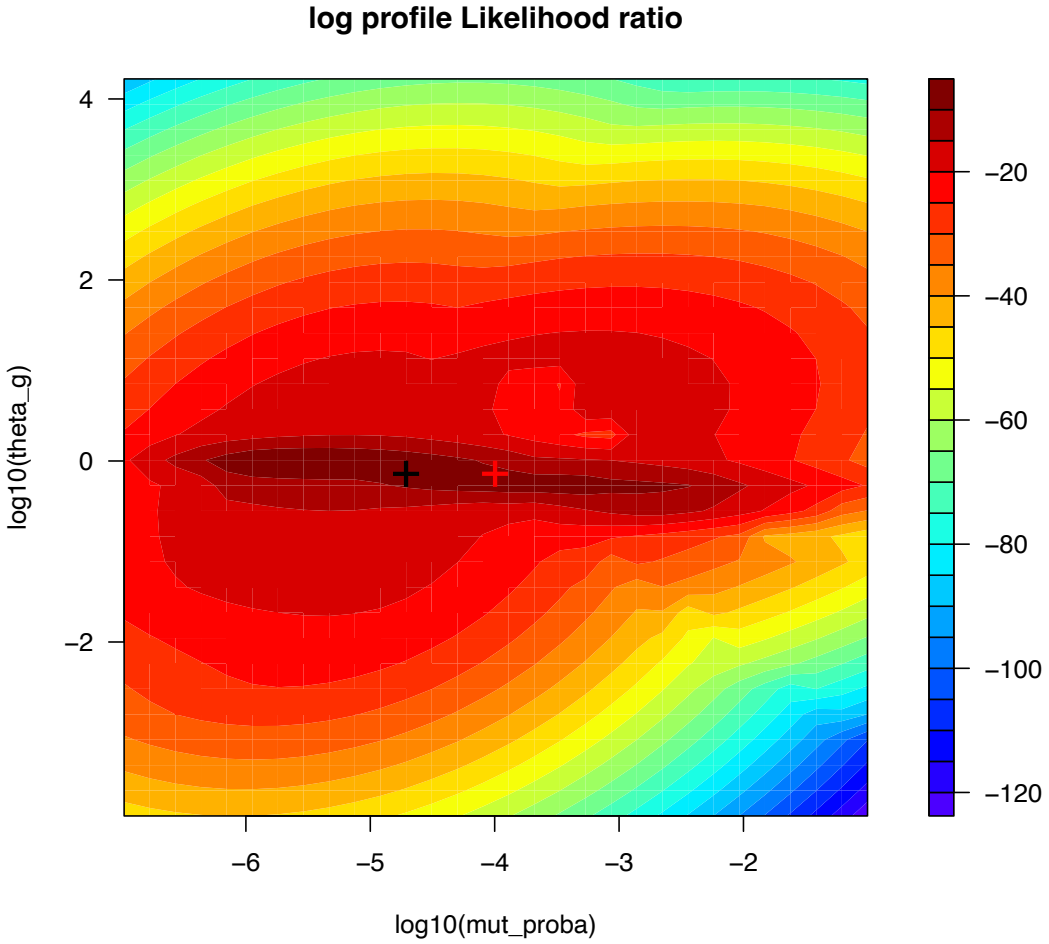


FIGURE 4.11 – Profiles LRT 2D pour θ et μ à la 8$^{\text{ème}}$ itération d'affinage.

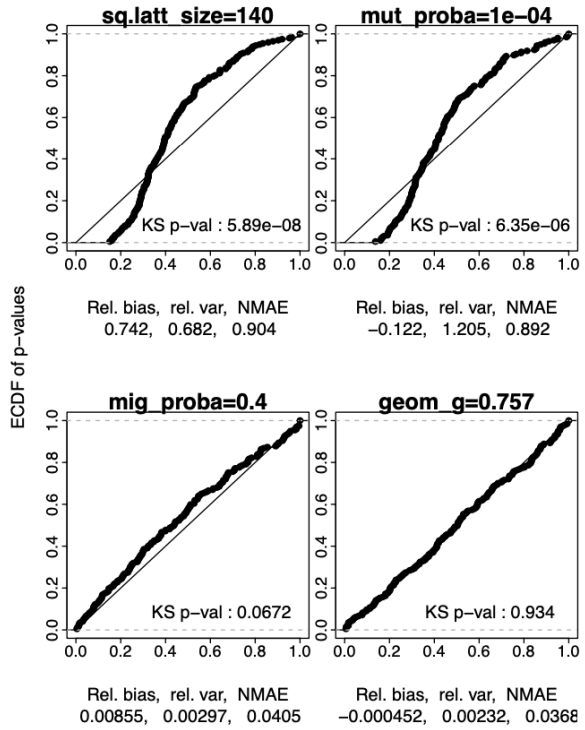
`IBDsim`. Comme expliqué précédemment, un certain niveau de déséquilibre de liaison peut s'observer même lorsque que l'on considère des paires de marqueurs non liés physiquement (c.a.d. sur des chromosomes différents) du fait que les arbres de coalescence de tous les marqueurs sont liés par l'arbre généalogique réalisé des individus de la population (pour cette raison, on parle souvent de "déséquilibre gamétique" plutôt que de "déséquilibre de liaison"). Ce déséquilibre gamétique se retrouve dans les données simulées par `GSpace` car il simule des génomes individuels et de la migration juvénile (voir section 4.2.1) mais pas dans celles simulées par `IBDsim` car ce dernier simule indépendamment chaque locus et considère de la migration gamétique. On peut donc tester l'impact de l'absence de déséquilibre gamétique dans les données simulées par `IBDsim` sur l'inférence des paramètres en comparant les performances des inférences dans un même modèle mais dans un cas en utilisant `GSpace` pour simuler les données et dans un autre cas en utilisant `IBDsim`. Pour des raisons de temps de calcul, nous avons simulé 20 locus microsatellites (SMM à 20 alleles, voir section 1.2.1) au lieu des 500 SNPS précédemment considérés (donc moins de marqueurs mais plus informatifs car plus polymorphes). Dans `GSpace`, un génome de 20 chromosomes portant chacun un seul locus est donc simulé.

Les résultats présentés en Figures 4.12 montre que l'estimation de tous les paramètres est moins précise, voir beaucoup moins précise pour σ^2, avec `GSpace` qu'avec `IBDsim`. C'est un résultat attendu car la corrélation des histoires généalogiques entre marqueurs en déséquilibre gamétique diminue l'information des données par rapport à l'information portée par des marqueurs "artificiellement" indépendants. C'est un résultat important et peu illustré en génétique des population alors que de nombreuses méthodes d'inférence et de nombreux simulateurs considèrent, comme `IBDsim` et toutes les approches étudiées dans les chapitres précédents, des marqueurs totalement indépendants. C'est le cas par exemple lorsque l'on multiplie les vraisemblances par locus pour obtenir la vraisemblance d'un jeu de données multilocus. Les intervalles de confiance obtenus en considérant les marqueurs comme indépendants seront donc trop étroits, et non conservatifs. Ceci sera bien sûr amplifié si les marqueurs considérés sont aussi liés physiquement de part leur positions proches sur le génome.

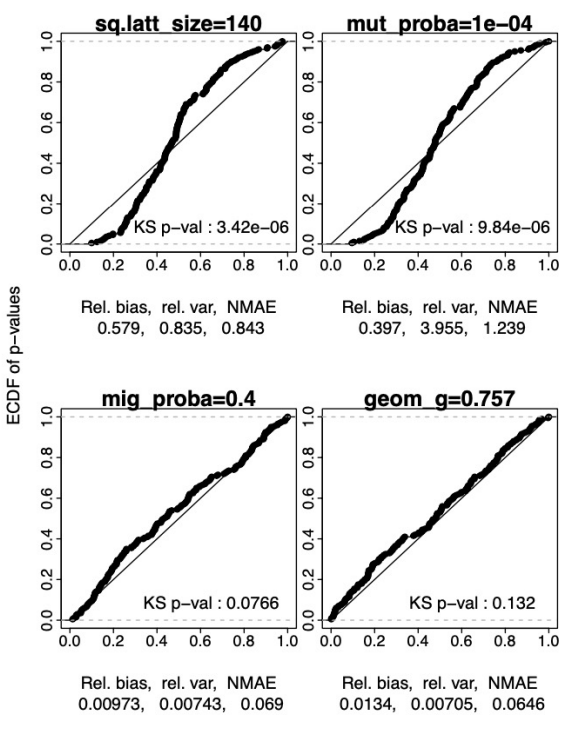
Croyant au départ à une bug, ces résultats ont été vérifiés en simulant marqueur par marqueur et en considérant de la migration gamétique avec `GSpace` et on retrouve bien dans ce cas les mêmes performances (artificiellement meilleures) d'inférence qu'avec `IBDsim`.

On peut aussi observer sur ces figures que les distribution de LRT-P-valeurs sont proches de la diagonale pour les paramètres bien estimés (m, g, θ, $/sigma^2$) mais montrent quelques imperfections. Les paramètres pour lesquels il n'y a pas (ou très peu) d'information montrent eux des distributions de LRT-P-valeurs très éloignnées de la diagonale. C'est un problème dû aux profiles de vraisemblance à tendance plate, difficile a bien prendre en compte lors de l'ajustement par mélange de gaussiennes, et difficile à corriger (mais les prochains résultats sont meilleurs et des améliorations récentes et notables ont été implémentées).

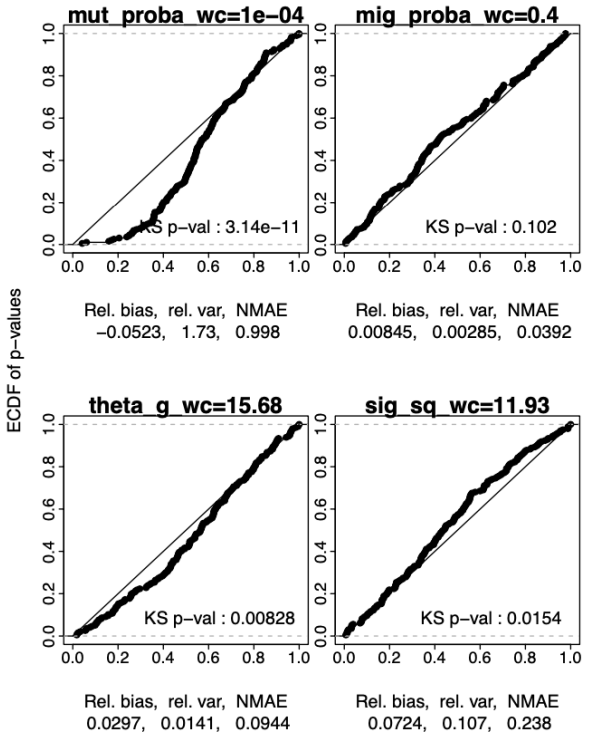
Information du déséquilibre de liaison Le second résultat intéressant concerne le test de l'apport de la statistique η et de l'information de déséquilibre de liaison qu'elle contient. Pour cela nous avons (i) inféré les paramètres (canoniques ou composites) uniquement à partir des statistiques de déséquilibre de liaison, et (ii) comparé la précision des inférences, avec ou sans ces statistiques, mais en incluant les autres statistiques présentées au chapitre 4.2.2. L'inférence est effectuée sur un tableau de référence de 1800 lignes générées lors de 7 affinages successifs. Les valeurs de



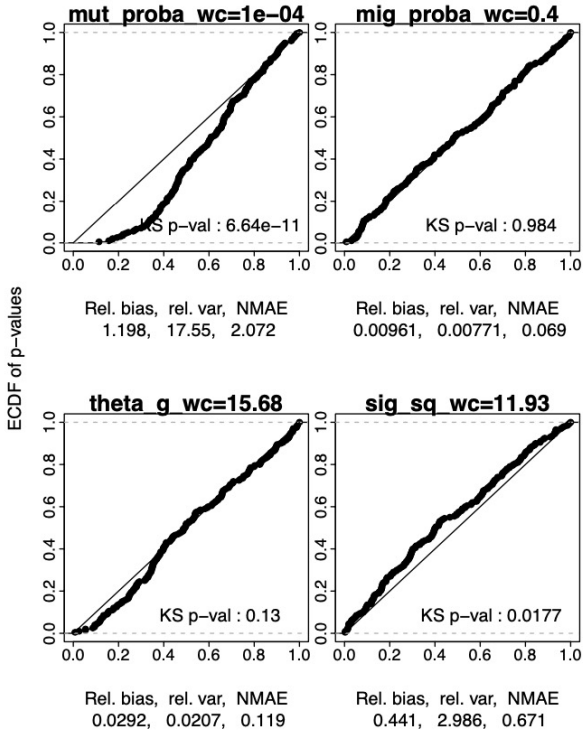
(a) IBDSim, canoniques



(b) GSpace, canoniques



(c) IBDSim, composites



(d) GSpace, composites

FIGURE 4.12 – Comparaison des performances de l'inférence par `SL` basées sur `IBDsim` et `GSpace` pour des valeurs de $\sigma^2 = 12$. Tous les autres paramètres sont ceux décrit dans le texte. Voir Fig. 3.7 et suivantes dans le chapitre 3 pour la description de ce type de figures.

paramètres choisies ainsi que les bornes de l'espace des paramètres exploré pour l'estimation sont résumées dans le tableau 4.1.

Dans le scénario décrit précédemment avec $N_T = 19600$ individus diploïdes et 50 marqueurs SNPs par chromosome, $\rho = 10^{-5}$ semble être un choix intéressant, le nombre d'évènements de recombinaison pour chaque évènement de coalescence variant entre $[0.196; 9.6]$ selon la distance entre marqueurs [3].

Les Figure 4.13, 4.14 et 4.15 montrent bien que, même si la statistique η semble apporter de l'information sur les paramètres de dispersion (m, g et par extension σ^2) et un peu sur θ dans le scénario testé, cette information semble être déjà portée par d'autres statistiques spatiales de manière bien plus efficace (le biais, la variance et le NMAE de l'estimation de ces paramètres est systématiquement plus faible dans les inférences utilisant toutes les statistiques). Ainsi, les variations du taux de recombinaison considérées affectent les performance de l'inférence basées sur η seulement, mais pas du tout celles basées sur toute les statistiques résumantes. Ces résultats nous paraissent décevants et méritent d'être explorés plus profondément avec d'autres configurations de marqueurs, d'autres taux de recombinaison et d'autres statistiques résumantes.

Tests à densité variable

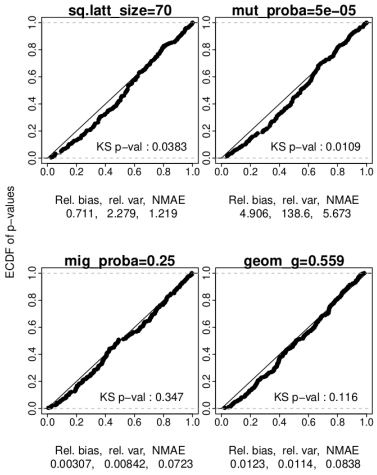
Les tests suivants ont été réalisés avec `IBDsim` uniquement car `GSpace` était en débugage intense suite aux observations mal interprétées ci-dessus de différences avec `IBDsim`.

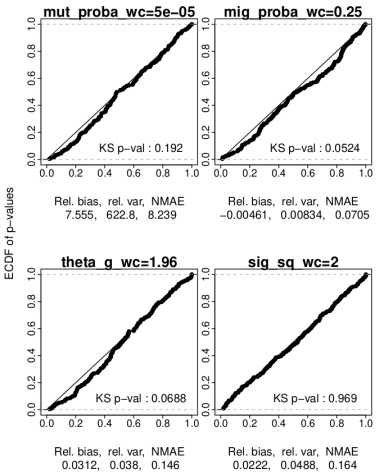
Quelques difficultés avec les modèles en réseau Une inférence complète et réaliste sous ce modèles d'IBD individuel en habitat continu nécessite d'estimer le paramètre D de densité du modèle, ou de manière équivalente la distance représentée par la maille du réseau. Estimer la densité dans ces modèles en réseau revient effet à ajuster l'échelle et les dimensions du réseau à la taille de l'habitat. Comme pour la plupart des analyses de jeux de données réelles, nous supposons ici qu'aucun paramètre du modèle n'est connu, ni même la taille de l'habitat (qui n'est pas un paramètre canonique de notre modèle). Nous devons donc envisager toutes les configurations de densités et tailles de réseau possible pour l'échantillon comme illustré sur la Figure 4.16.

Ceci impose des contraintes sur les combinaisons de valeurs de paramètres, et nous avons dû définir un paramètre "canonique pour l'inférence" lié à la densité, sous la forme $\frac{n_x}{D}$, afin de pouvoir explorer tous l'espace des paramètres facilement (c.a.d qu'il n'y pas de zone de l'espace des paramètres "impossible" à simuler, par exemple avec un échantillon plus grand que l'habitat). Dans les analyses ci-dessous, la densité D bien qu'elle soit un paramètre canonique du modèle d'IBD apparaît comme un paramètre composite de l'inférence. D'autres paramètres composites comme la taille de l'habitat pourrait être ajouté à l'inférence pour faciliter l'interprétation des estimations sur des jeux de données réelles. De plus, il est toujours possible d'exprimer la dispersion en unité de maille de réseau comme précédemment, mais il serait peut être plus "biologiquement interprétable" de l'exprimer en distance métrique. Enfin, il est impossible de définir des distributions de dispersion discrètes avec les caractéristiques voulues quand la densité est trop faible, impliquant de grandes mailles de réseau et donc peu de points pour définir une distributions discrètes. Ceci entraîne

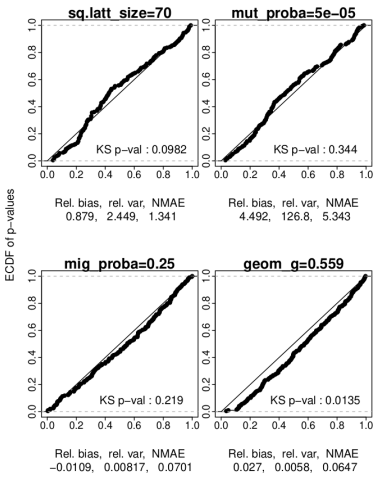
3. En connaissant approximativement le temps de coalescence moyen entre deux lignées (qui est $\approx N_T$ avec N_T le nombre de copie de gène dans la population, voir équation 1.2), la distance entre les marqueurs et ρ il est possible de trouver le nombre moyen d'évènements de recombinaison pour toutes les paires de marqueurs.

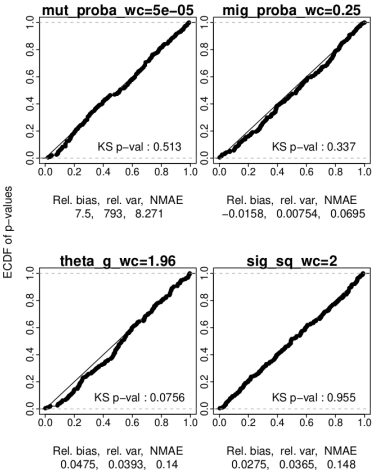
$\rho = 10^{-7}$



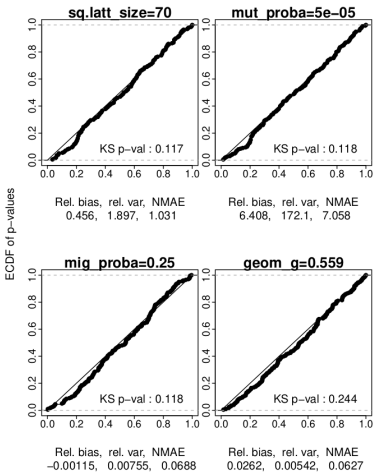


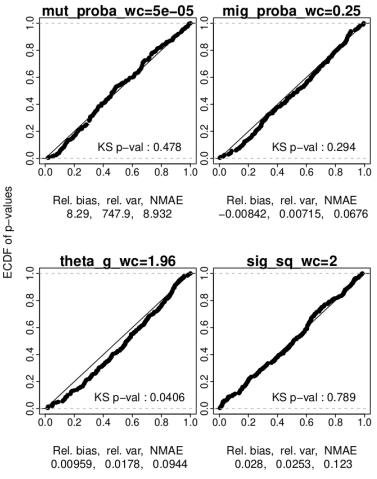
$\rho = 10^{-5}$





$\rho = 10^{-3}$

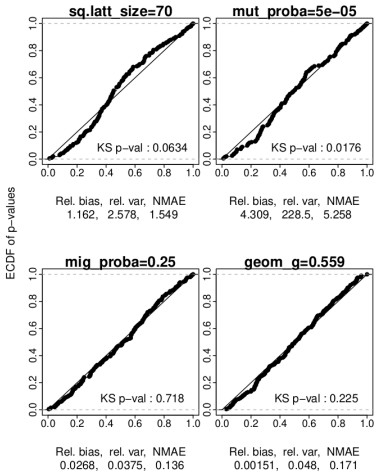


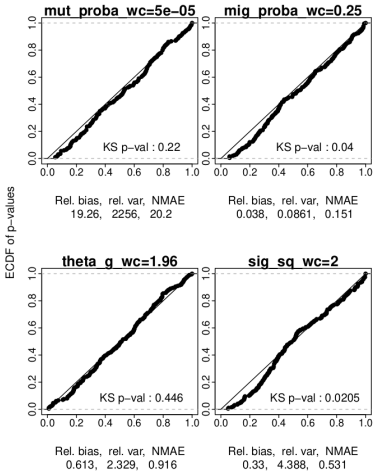


(a) Paramètres canoniques uniquement (b) Paramètres canoniques et composites

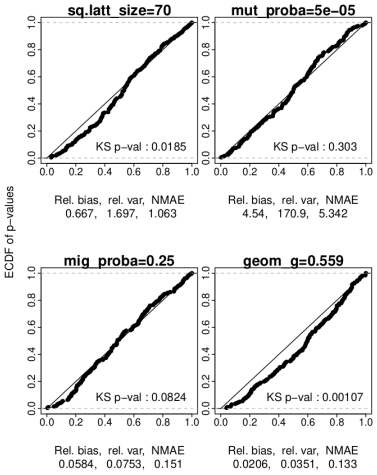
FIGURE 4.13 – Résultats des tests de performances sur des inférences effectuées avec l'ensemble des statistiques résumantes calculables par `GSumStat`(voir section 4.2.2).

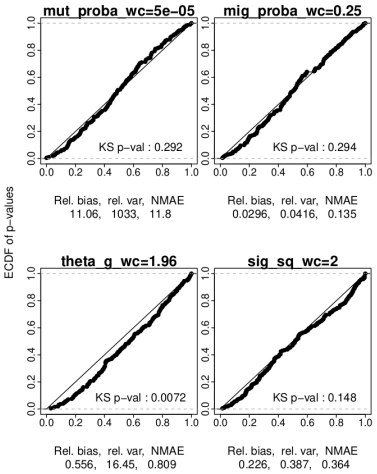
$\rho = 10^{-7}$



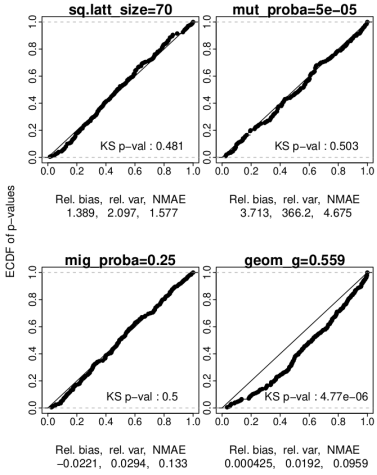


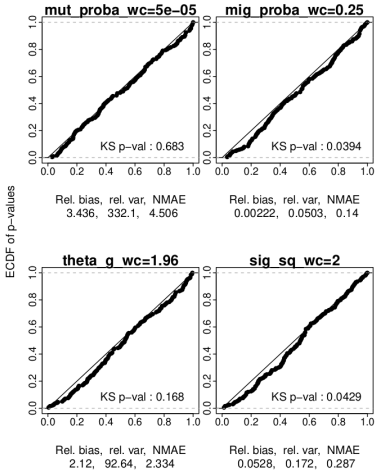
$\rho = 10^{-5}$





$\rho = 10^{-3}$

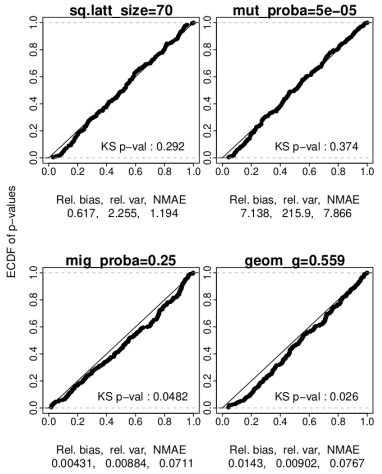


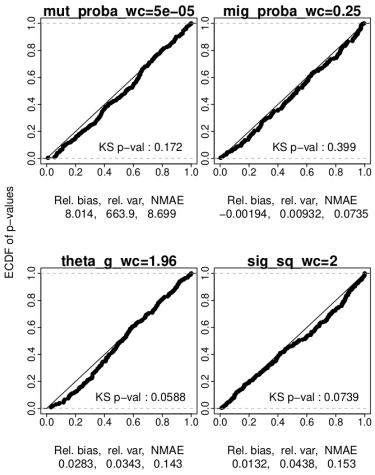


(a) Paramètres canoniques uniquement (b) Paramètres canoniques et composites

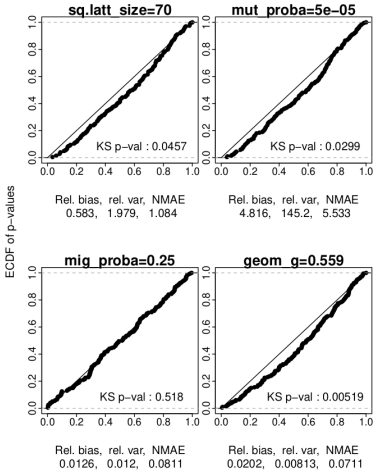
FIGURE 4.14 – Résultats des tests de performances sur des inférences effectuées avec uniquement les quatre statistiques résumantes qui décrivent les variations de η.

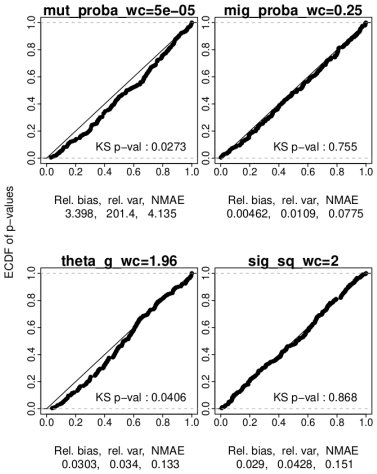
$\rho = 10^{-7}$



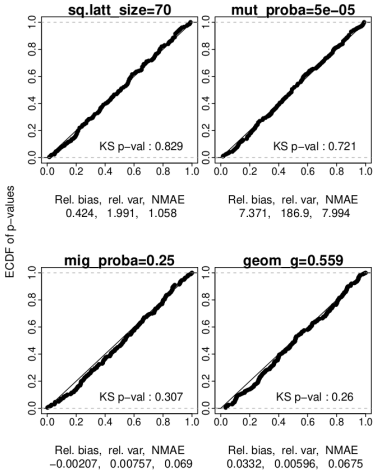


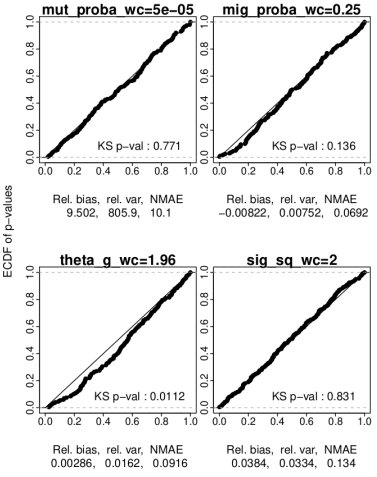
$\rho = 10^{-5}$





$\rho = 10^{-3}$





(a) Paramètres canoniques uniquement (b) Paramètres canoniques et composites

FIGURE 4.15 – Résultats des tests de performances sur des inférences effectuées avec l'ensemble des statistiques résumantes calculables par `GSumStat` mais sans les quatre statistiques résumantes qui décrivent les variations de η.

de grosses approximations sur la forme de la distribution de dispersion et implique que la méthode devrait être moins performante quand la densité réelle n'est pas au centre de l'intervalle exploré (qui correspondrait alors à la situation idéale avec une densité de 1 dans nos scénario, que nous explorons entre 0.1 et 10, et difficilement plus largement). Ce changement d'échelle des modèles réseau correspondant à des changement de densité est un point limitant qui se comprend assez bien sur la Figure 4.16, et qu'il reste aussi à explorer.

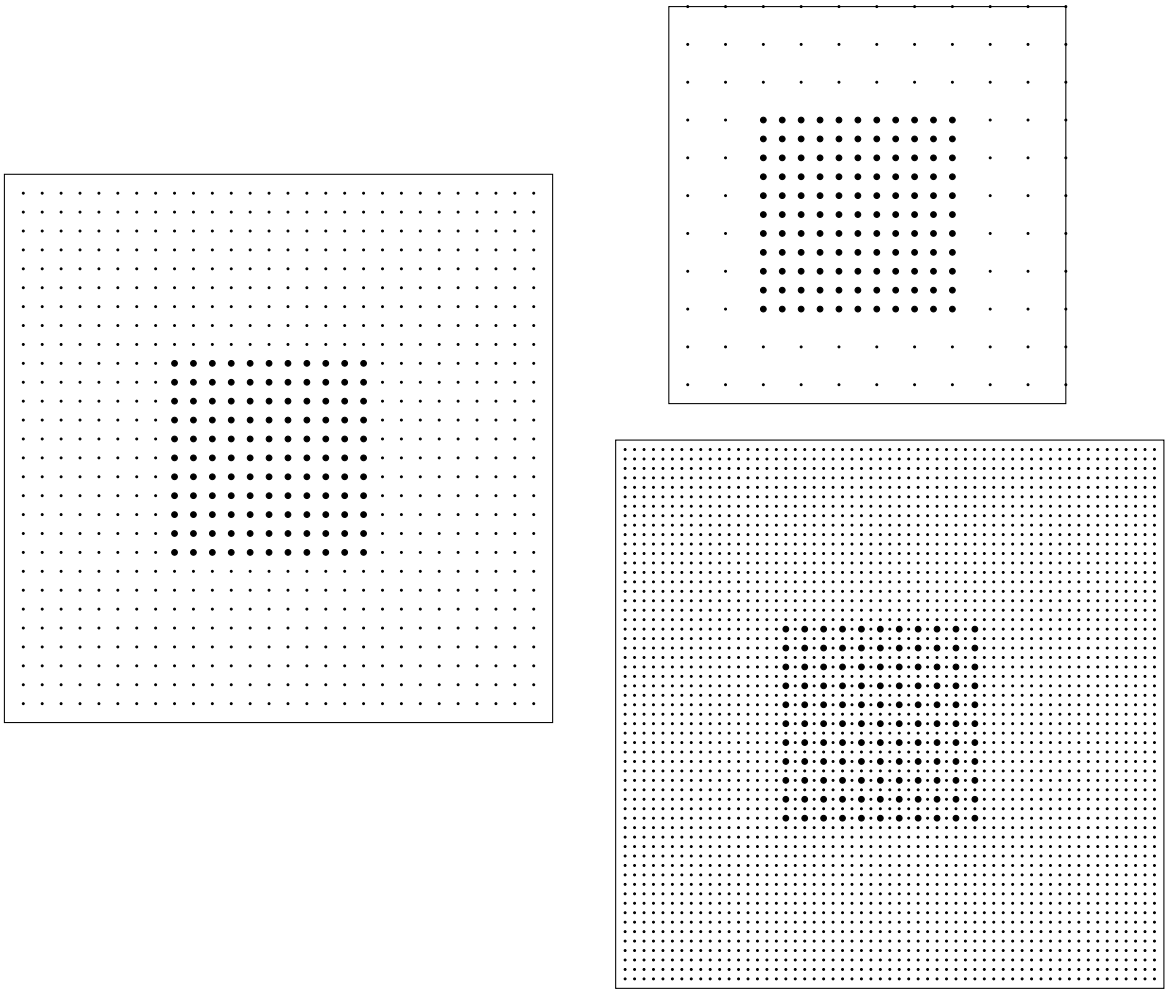


FIGURE 4.16 – Représentation graphique de trois modèle d'IBD individuel en habitat continu appliqués à un même échantillon. A gauche, le cas évoqué jusqu'ici où la maille d'échantillonnage correspond à la maille de la grille du réseau. En haut à droite, un cas où la taille de l'habitat a été réduite et où la maille du réseau est 2x plus large que celle de l'échantillon, correspondant à une densité 4 fois plus faible. En bas à droite, un cas où la taille de l'habitat est la même qu'à gauche, mais ou la maille du réseau est 2x plus petite que celle de l'échantillonnage, correspondant à une densité 4 fois plus forte. Les petits points représentent les individus du réseau, les plus gros points les individus échantillonnés, et le cadre représente l'habitat.

Les résultats avec l'inférence de la densité, présentés en Figure 4.17, sont très récents et nous n'avons pas encore eu le temps de bien les analyser. Je souhaitais toutefois les montrer car il suggèrent que l'on peut estimer très précisément la plu-

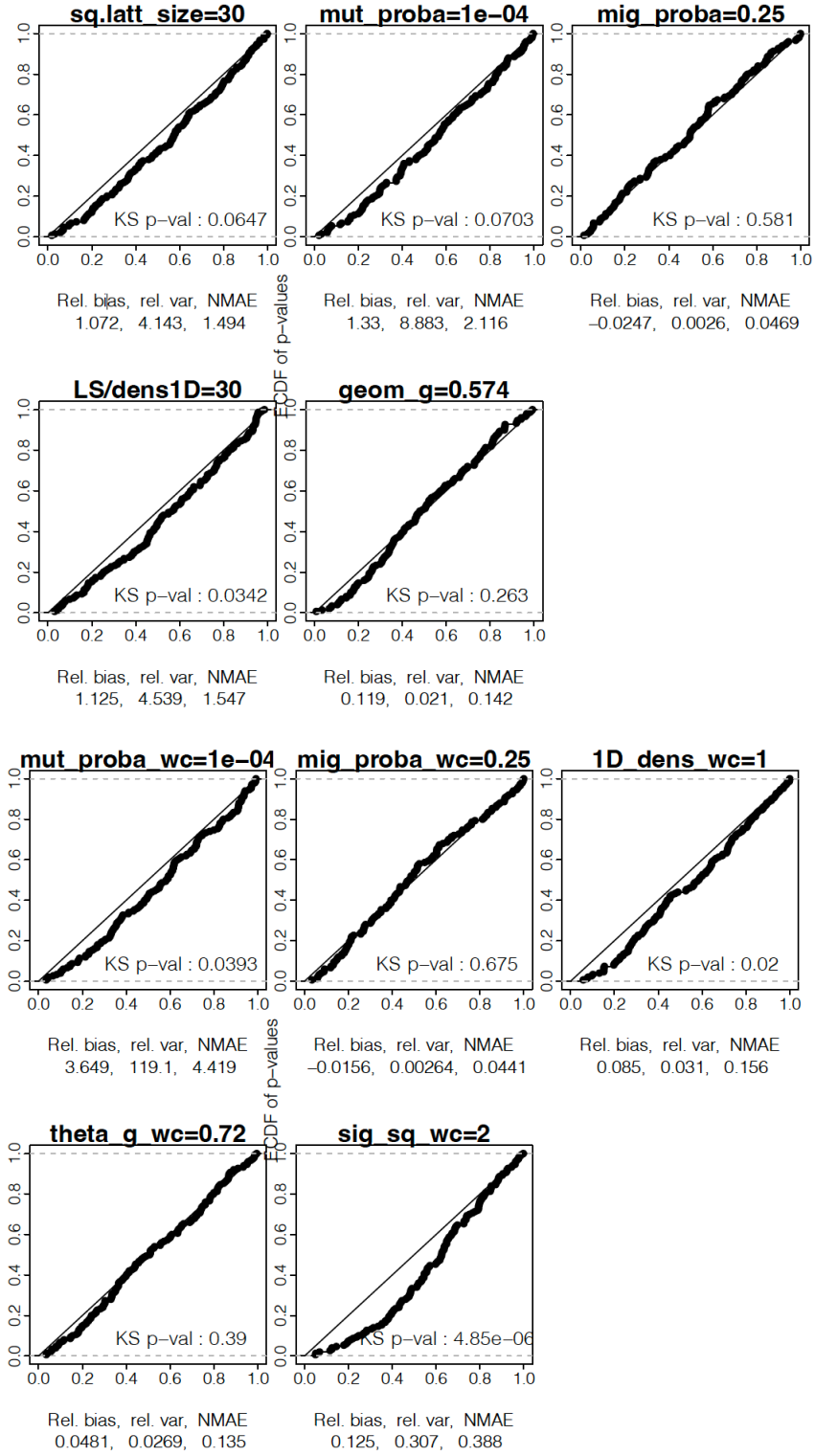


FIGURE 4.17 – Performances de l'inférence sous le modèles d'IBD individuel complet avec variation de la densité. $LS = sq.latt_size$ correspond au paramètres de taille du réseau n_x, et dens1D correspond à la densité en une dimension par rapport à la densité de référence de 1 de l'échantillon. L'analyse est basée sur 20 locus SMM et les valeurs de paramètres indiquée en haut de chaque graphique, tous les autres paramètres sont ceux décrit dans le texte. Je n'ai pas eu le temps de refaire une figure correcte...

part des paramètres biologiquement intéressant du modèle sauf la taille du réseau (et donc la taille totale de la population) et le taux de mutation qui ne s'estime correctement que sous la forme du produit $\theta_g = N_{tot}\mu = 4n_x^2\mu$. Il est notamment très intéressant de voir que l'on peut estimer D, m et g avec précision ce qui était le but de ces développements d'inférence par simulation sous IBD. Il semblerait aussi qu'il y ait un peu d'information sur n_x et μ indépendamment mais cela reste à vérifier.

Au delà de ces premiers tests très préliminaires, il convient de souligner que les distributions des LRT P-valeurs sont approximativement uniformes dans la plupart des situations pour les paramètres inférés avec précision. Ceci montre qu'il est possible de calculer des intervalles de confiance par "rapport de vraisemblance résumée" ayant la couverture attendue, et ceci, sur la base de petits tableaux de référence d'environ 1800 simulations pour l'inférence de quatre paramètres et 4200 simulations pour 5 paramètres. Bien que les performances puissent être légèrement améliorées en considérant plus de simulations, ce résultat est important car les inférences par simulation sont généralement très coûteuses en terme de temps de simulation, surtout sous les modèles spatialisés que nous avons utilisés. En comparaison, Raynal *et al.* (2018) recommandent d'utiliser 100 000 simulations par défaut pour une bonne estimation des distribution a posteriori marginales par paramètre pour l'`ABC-RF`, qui ne fait pas d'inférence multidimensionnelle.

4.3 Conclusions

Dans ce chapitre, nous avons vu comment les méthodes d'inférence par simulation couplées à des simulations par coalescence exacte permettent d'estimer les paramètres de dispersion, densité et taille de populations dans un modèle d'IBD individuel en habitat continu. Ces développements sont très récents et les performances de l'inférence doivent encore être testées en profondeur comme décrit dans les chapitres précédents pour les autres types d'inférences. Malgré tout, ces premiers tests ont montré que notre objectif est quasiment atteint : la combinaison de méthodes d'inférence par simulation avec le modèle réaliste d'IBD individuel en habitat continu est possible et permet l'estimation de beaucoup plus de paramètres que les approches développées précédemment.

Avant d'affirmer que ces méthodes peuvent bien estimer la densité, le taux de dispersion et la forme de la distribution de dispersion qui correspondent bien aux valeurs actuelles et locales des populations naturelles, il faudra notamment s'assurer de la robustesses des inférences vis à vis des processus mutationnels (notamment bien gérer la mutation pour les marqueurs SNPs), des hétérogénéités démographiques spatiales et temporelles, et que l'on puisse considérer des distributions de dispersion plus flexible que la loi géométrique telles que la Sichel évoquée dans les sections 1.3.2 et 3.2.2.

Quoiqu'il en soit, ces résultats préliminaires sont extrêmement pertinents puisqu'ils montrent, pour la première fois en génétique des populations, que l'inférence de certains paramètres canoniques des modèle de Wright-Fisher "étendus" (ici D, m, g, σ^2) peuvent être estimés à partir de données de polymorphisme génétique. Comme nous l'avons vu dans les chapitres précédents, il était jusqu'alors considéré que seuls les paramètres définissant les modèles basés sur les approximations du coalescent (ou de la diffusion), et donc les produits des taux d'événements par les tailles de population ($N\mu$, Nm, $D\sigma^2$, mais aussi T/N et $N\rho$ dont nous n'avons pas parlé), pouvaient être estimé à partir de données génétiques. Ceci est possible avec notre

méthode car les inférences sont basées sur des simulations par coalescence génération par génération, simulant de manière exacte le processus de Wright-Fisher. Mais aussi parce que les statistiques résumantes considérées contiennent de l'information sur ces différents paramètres. Notons que l'estimation de m et de σ^2 correspond de fait à l'estimation de Nm et de $D\sigma^2$, puisque $N = D = 1, 2$ lorsque σ^2 est exprimé en distance inter-individu/couple. C'est donc essentiellement l'estimation de D qui est "intéressante", même si tout cela doit être pris avec des pincettes et plus largement étudié. Les approches spatiales sans approximations du coalescent pourrait donc permettre d'avoir de l'information sur les processus de dérive locales à petite échelle évolutive (via la densité) vs. la dérive globale à plus grande échelle (via la taille des population, ou peut être uniquement le produit $N\mu$). Si ces résultats sont vérifiés, cela ouvre de nombreuses perspectives pour l'inférence des paramètres démographiques locaux des populations naturelles puisqu'il parait possible d'estimer la densité et la dispersion locale des populations indépendamment de la taille globale de la population et du taux de mutation.

Chapitre 5

Conclusions

J'espère avoir réussi à travers ce document à vous démontrer la démarche sous-jacente à mon parcours de recherche qui de la question de l'inférence de la dispersion chez le crapaud de la canne à sucre m'a amené à tester et développer différentes approches pour l'inférence des paramètres démographiques, notamment de dispersion, à petites échelles spatiales et temporelles à partir de données génétiques issues de population naturelles. je souhaitais démontrer l'importance de bien comprendre la pertinence mais aussi les limites des méthodes d'inférence en génétique de populations à travers des tests poussés de performance des estimations des différents paramètres du modèle et surtout de leur robustesse par rapports aux hypothèse des modèles sous-jacents. Dans le contexte de l'analyse de données issues de populations naturelles, cette robustesse est primordiale, et j'ai beaucoup insisté sur ce point car il me semble que de nombreux généticiens des populations l'oublient régulièrement, que ce soit lors de l'inférence sur données réelles par des empiristes (et j'avoue que j'ai aussi tendance à l'oublier quand ca m'arrange...) ou de la conception des méthodes d'analyse de données, ce qui me semble plus problématique.

Ce document m'a aussi permis de brosser un tableau peut être un peu atypique, et avec un petit retard à l'allumage pour les NGS..., de l'évolution de la génétique des populations de ces début dans les années 30 à aujourd'hui, et des différentes approches d'inférence qui ont été développées.

Les quatre résultats les plus pertinents de tous ces travaux sont, à mon sens, (i) la validation des modèles d'IBD, et notamment la version sans dèmes de l'IBD individuel en habitat continu, grâce aux comparaisons entre les estimations démographiques et génétiques sur de nombreux jeux de données. Ces comparaisons sont a étendre, et il parait maintenant intéressant de savoir si ces comparaisons tiennent encore quand on comparera indépendamment les inférences des densité et de la dispersion par les deux approches ; (ii) la possibilité d'estimer plus de paramètres du modèle IBD qu'uniquement le produit $D\sigma^2$ comme c'était le cas très récemment encore ; La limitation de l'inférence en IBD à $D\sigma^2$ a toujours été assez frustrante pour de nombreux empiristes car leur intérêt est généralement d'estimer séparément D et d'autres paramètres de la dispersion, plus "biologiquement parlant" que le σ^2 ; plus globalement (iii) la capacité à inférer avec précision les paramètres de densité et de dispersion en population naturelles à partir d'un petit échantillon génétique ; et notamment (iv) la capacité des méthodes d'inférence par simulation à inférer chaque paramètre indépendamment les uns des autres, et non sous forme de produit $N\mu$, Nm, $D\sigma^2$, lorsque l'on s'affranchi des approximations classiques de grandes tailles de populations et de petits taux d'évènements. Ce dernier point est à explorer et pourrait ouvrir de nombreuses perspectives en génétique spatiale des populations.

Il y a donc encore une infinité de questions à poser ; de nouveaux modèles, de nouvelles approches d'inférence, et de nouvelles statistiques résumantes à explorer ; et donc de nombreux tests à faire ; et je n'énoncerai donc pas ici toutes ces perspectives infinies. Je finirai sur un point que je n'ai pas abordée et que le lecteur pourrait avoir en tête si il a déjà fait des analyses sous IBD.

Pendant longtemps, même étonnamment longtemps après la publication de la méthode de la régression, la plupart des analyses empiriques d'isolement par la distance se sont contentées de tester la présence d'un patron spatial d'IBD dans les données génétiques pour mettre en évidence une dispersion limité dans l'espace ou non dans les populations étudiées. Ces études n'utilisent donc pas l'inférence quantitative de $D\sigma^2$ par la régression mais uniquement la corrélation de la différenciation avec la distance géographique. Or comme nous l'avons vu, le modèle d'IBD inclus le modèle île, et il n'y a donc pas de limite arbitraire entre un modèle d'IBD avec une dispersion peu localisée et un modèle en île. Ces approches de tests de présence d'un patron IBD dans les données génétiques se sont concentrées sur la seul méthode disponible dés les premières analyses IBD, le test de Mantel (Mantel, 1967). Or, entre autre à cause de la forte variance d'estimation de la différenciation et de par sa conception, le test de Mantel est très peu puissant avec les tailles d'échantillons classiques de génétique des populations, notamment sur l'analyse d'un échantillon issu de quelques dèmes (45 points pour un échantillon de 10 dèmes, représentant donc les 45 paires de dèmes). Il est un peu plus puissant sur les analyses individuelles du fait du plus grand nombre de points (4950 paires pour un échantillon de 100 individus). Même si il semble donc recommandé de faire les tests de Mantel sur le données individuelles, la non-significativité du test de Mantel pour l'IBD est souvent un résultat non-informatif puisqu'il ne détectera des patrons d'IBD que si la dispersion est très fortement limitée dans l'espace. Une première alternative au test de Mantel a été de calculer des intervalles de confiance sur la pente de la régression et cela permet en effet une meilleure détection des patrons d'IBD malgré un comportement non-idéal des intervalles de confiance par bootstrap sur une paramètre ayant une distribution très asymétrique ((Leblois *et al.*, 2003)). C'est maintenant possible avec les approches d'inférence par simulation, qui devraient donner de meilleurs intervalles de confiance sur $D\sigma^2$, D et g, mais aussi permettent d'implémenter des procédures de choix de modèles, et donc d'avoir in fine une meilleur mesure de l'IBD.

Cette conclusion est un peu courte mais il est 11H30 et je dois soumettre avant 12H.....

Mon projet de recherche

Le développement d'approches agro-écologiques pour la gestion des ravageurs et des auxiliaires, ainsi que de leurs vecteurs et antagonistes, nécessite une meilleure compréhension du fonctionnement démographique local de leurs populations. De même, la gestion des populations menacées nécessite une connaissance fine du statut démographique et génétique de ces populations : effectifs, fragmentation, dispersion, consanguinité... Parmi les facteurs clés à caractériser, les densités/tailles des populations et les caractéristiques de dispersion, à une petite échelle géographique, ainsi que leurs variations dans le passé récent, sont souvent mal connues alors que ces facteurs s'avèrent cruciaux pour mieux comprendre la dynamique de ces populations à l'échelle d'un paysage ou d'un bassin de production agricole (Lewis *et al.*, 1997; Veres *et al.*, 2013). Ces paramètres démographiques peuvent théoriquement être estimés par des approches démographiques de type capture - marquage- recapture, mais elles impliquent un investissement humain très important et ne donnent pas d'information sur les variations passées. Une alternative est d'utiliser des estimations "indirectes" à partir de données de polymorphisme révélées sur une partie ou l'ensemble du génome, qui contiennent de l'information sur les paramètres démographiques des populations.

L'explosion, ces vingt dernières années, des données génomiques et de la puissance de calcul disponible a été suivie d'une multitude de développements en inférence démo-génétique, notamment pour retracer les dynamiques historiques des populations (Beichman *et al.*, 2018). Cependant, peu de ces méthodes prennent en compte (i) la structuration spatiale des populations, un facteur pourtant crucial puisqu'il biaise fortement les histoires démographiques inférées (Leblois *et al.*, 2006; Chikhi *et al.*, 2010) et (ii) que la majorité des espèces a une dispersion limitée dans l'espace (Endler, 1977). Les rares méthodes considérant une structuration spatiale (Petkova *et al.*, 2016; Al-Asadi *et al.*, 2019) se sont intéressées à de grandes échelles spatio-temporelles, et sont basées sur des hypothèses peu réalistes à des échelles plus fines (dispersion peu réaliste du modèle en îles ou du stepping-stone, sous-populations panmictiques). Elles démontrent toutefois que le déséquilibre de liaison apporte de l'information sur les variations temporelles des paramètres de dispersion et/ou de tailles de populations (Al-Asadi *et al.*, 2019; Boitard *et al.*, 2016).

En continuité avec mes travaux passés et actuels, l'objectif de mon projet de recherche est de combler cette lacune méthodologique en développant et testant de nouveaux outils méthodologiques pour estimer les paramètres démographiques locaux, ainsi que leurs variations récentes, à partir de données génomiques en utilisant des modèles démo-génétiques spatialisés. La possibilité de générer de gros jeux de données spatialisées (des génomes complets sur beaucoup d'individus), couplée au développement récent de méthodes d'inférence très performantes en génétique des populations, permet de s'intéresser maintenant à des signaux génétiques faibles

et complexes laissés par des processus démographiques de plus en plus fins (Novembre & Peter, 2016). Dans la continuité de mes travaux actuels sur les modèles démo-génétiques spatialisés et le développement de méthodes d'inférence basées sur la simulation, j'étendrai les outils que j'ai développé de manière collaborative ces dernières années : (i) le simulateur génomique spatialisé `GSpace` (Virgoulay *et al.*, 2021) pour pouvoir prendre en compte des hétérogénéités spatiales et temporelles des paramètres démographique ; ii) la librairie de calcul de statistiques résumantes et l'intégration d'outils d'intelligence artificielle (IA) pour mieux prendre en compte l'information des données génétiques spatiales pertinente pour l'inférence des paramètres d'intérêt ; (3) la librairie `gspace2infr` qui couple ces deux outils à de puissantes méthodes d'inférence par simulation comme le calcul bayésien approché basé sur les techniques de Random-Forest (`ABC-RF`, Pudlo *et al.*, 2016; Raynal *et al.*, 2018) ou les développements récents de la méthode par vraisemblance résumée (`SL`, Rousset *et al.*, 2017), afin de faciliter les tests de performances et leur diffusion. Une partie importante de mon projet consiste donc à tester les performances (précision et robustesse) de ces développements, notamment le niveau de complexité spatio-temporelle qui peut être considéré en fonction du type et de la quantité de données disponibles.

Afin de valider leur intérêt pratique, ces développements sont pensés, testés et appliqués dans deux contextes bien différents : (1) l'étude et la gestion des organismes d'intérêt agronomique, à travers des collaborations internes au CBGP ; et (2) la biologie de la conservation, à travers le projet DevOCGen et ses collaborations ("Développement et applications de nouveaux outils pour la gestion et la conservation des populations naturelles à partir de données génomiques", financé par la Région Occitanie 2022-2026).

5.1 Objectifs

Mon objectif est de démontrer la pertinence de l'utilisation de la génomique des populations pour extraire des informations pertinentes sur le fonctionnement démographique local et récent des populations en répondant aux questions suivantes :

- Les méthodes d'inférence spatiales de génomique des populations appliquées à des données génomiques de grande dimension permettent-elles d'accéder aux paramètres démographiques d'intérêt et grâce à quel jeu de statistiques résumantes ?
- Quelle précision sur les différents paramètres démographiques d'intérêt peut-on atteindre selon la quantité et le type de données génomiques analysées ? Ces estimations sont elles robustes vis à vis des écarts les plus probables au modèle ?
- Quel niveau de résolution dans la modélisation spatiale et temporelle peut-on raisonnablement utiliser pour l'inférence à partir de données génomiques ?
- Ces niveaux de complexité spatiale et temporelle, et de compromis entre précision des estimations et coûts des données génomiques sont-ils pertinents pour des applications en biologie de la conservation ou en agronomie ?
- Quels sont les domaines d'application sachant que les organismes d'intérêt

agronomiques ont souvent de fortes densités, de grandes tailles de populations et de fortes capacités de dispersion, et qu'au contraire, les espèces menacées permettent rarement d'avoir accès à de très gros échantillons ?

A travers ces objectifs, je souhaite : (1) proposer une meilleure prise en compte des aspects spatiaux en génétique des populations ; (2) valoriser au maximum l'information présente dans les données génomiques, et notamment le déséquilibre de liaison (c.a.d. l'information liée aux événements de recombinaison, en plus des mutations et coalescences) pour répondre à des questions de biologie évolutive et d'écologie moléculaire à différentes échelles spatio-temporelles.

5.2 Méthodes

Mon projet repose principalement sur le développement de méthodes d'inférence innovantes basées sur l'utilisation de deux approches dont mes collaborateurs et moi sommes spécialistes : (1) des modèles démo-génétiques spatialisés d'isolement par la distance (IBD) ; et (2) des méthodes d'inférence par simulation (IPS) utilisant des approches d'intelligence artificielle (IA), et plus spécifiquement d'apprentissage automatique supervisé. Les modèles IBD considèrent des distributions de dispersion modélisant de façon très flexible n'importe quel type de dispersion plus ou moins limitée dans l'espace (Rousset, 1997; Guillot *et al.*, 2009) et peuvent considérer des populations en habitat continu dans lesquels les individus ne forment pas de sous populations panmictiques (Rousset, 2000; Leblois *et al.*, 2003, 2004). Ces deux facteurs rendent les modèles d'IBD beaucoup plus réalistes à petite échelle spatio-temporelle que les modèles classiques de génétique des populations (modèle en île et stepping stone). Les méthodes d'IPS que nous utiliserons (Approximate Bayesian Computations using Random Forest, ABC-RF (Pudlo *et al.*, 2016; Raynal *et al.*, 2018) ; et Summary-Likelihood, SL (Rousset *et al.*, 2017)) utilisent l'apprentissage automatique par la méthode des Forêts Aléatoires permettant d'obtenir de bonnes estimations à partir d'un grand nombre de statistiques résumantes (SR) et d'un petit nombre de simulations (par ex. : 100 fois moins pour l'ABC-RF que l'ABC classique, Raynal *et al.*, 2018). Elles sont donc particulièrement adaptées à l'analyse de gros jeux de données et aux modèles IBD dont la simulation est plus lente que les modèles classiques.

Ces approches seront ensuite validées en analysant les jeux de données disponibles, en lien avec les chercheurs les ayant générés. Ceci nécessite d'avoir accès à des données conformes aux attendus des méthodes : idéalement des génomes complets de qualité, sinon un grand nombre de marqueurs avec des informations de déséquilibre de liaison (génomes dits "phasés"), sur un grand nombre d'individus géo-référencés, à des échelles géographiques locales.

La production de génomes entiers sur un grand nombre d'individus est encore très coûteuse, si bien que de tels jeux de données sont encore assez rares (à l'exception de quelques espèces modèles). Cependant, dans le projet "DevOcGen", nous testons actuellement une nouvelle approche de séquençage de bonne qualité de génomes complets sur beaucoup d'individus à bas coût (l'HaploTagging Meier *et al.*, 2021), qui pourrait permettre de séquencer 200 génomes complets phasés pour environ 5 000€. Plusieurs jeux de données vont bientôt être générés dans ce projet, tels des données génomiques spatialisées d'une espèce de lézard endémique des Pyrénées et menacée par le réchauffement climatique, *Iberolacerta bonnali*, et/ou des quelques

populations restantes de la centaurée de la Clape, *Centaurea corymbosa*, pour estimer les caractéristiques de dispersion, et de connectivité entre populations, ainsi que les variations récentes de densités. Des données seront aussi disponibles à petites échelles géographiques pour le goujon occitan *Gobio occitaniae*, l'ophrys d'Aymonin *Ophrys aymononii*, l'hippocampe moucheté *Hippocampus guttulatus*, et le lézard vivipare du Mont Lozère, *Zootoca vivipara vivipara*.

Par ailleurs, de tels jeux de données sont déjà disponibles au CBGP. Nous avons par exemple des génomes complets de *Drosophila suzukii*, espèce pour laquelle il est important d'estimer la dynamique populationnelle récente pour aider à déterminer le nombre et les localisations de lâchers dans l'optique de lutte par la technique de l'insecte stérile (TIS). D'autres jeux de données d'organismes d'intérêt agronomique sont et seront produits dans les années à venir au CBGP (*Bactrocera dorsalis*, *Philaenus spumarius*, *Thaumetopoea pityocampa*) et par des collaborations extérieures.

5.3 Collaborations

En continuité de mon parcours, ce projet s'inscrit dans le cadre de collaborations de long terme avec des chercheurs montpelliérains ayant une expertise en génétique statistique, statistiques inférentielles et mathématiques appliqués, autant sur les méthodes d'inférences statistiques (François Rousset CNRS ISEM, Arnaud Estoup INRAE CBGP, Jean-Michel Marin UM IMAG) que sur les modèles spatialisés de génétique des populations (Arnaud Estoup, François Rousset, Stéphane Guidon CNRS LIRMM). Je souhaite aussi collaborer avec John Novembre (Univ. de Chicago), spécialiste à la fois en génétique spatiale des populations et en statistiques, et Asger Hobolth (Univ. de Aarhus), mathématicien et généticien des populations, avec qui je collabore depuis quelques années.

Enfin, les nombreux échanges dans les groupes (informels) thématiques "génomique statistique" (Simon Boitard, Arnaud Estoup, Mathieu Gautier, Miguel de Navascués, 4 doctorants et 2 postdoctorant), et "génétique spatiale" (Karine Berthier, Marie-Pierre Chapuis, Sylvain Piry, 3 doctorants) du CBGP m'apporte une vraie motivation et un regard critique sur mes développements. Je souhaite également lancer de nouveaux projets permettant de poursuivre les collaborations avec les personnes impliquées dans le projet "DevOcGen" (BiodivOc). Ce projet a en effet créé une forte synergie, extrêmement motivante, de la communauté régionale travaillant sur l'inférence démo-génétique théorique et appliquée grâce à ses nombreuses collaborations (plus de 20 chercheurs dans 11 UMR, 6 plateformes bio-informatique et de séquençage, 5 non-permanent.e.s) et animations (discussion d'articles mensuelles, workshops).

Bibliographie

Abdo, Z., Crandall, K. A. & Joyce, P. (2004) Evaluating the performance of likelihood methods for detecting population structure and migration. *Molecular Ecology* **13** : 837–851.

Abramovitz, M. & Stegun, I. A., éds. (1972) *Handbook of mathematical functions*. Dover, New York.

Al-Asadi, H., Petkova, D., Stephens, M. & Novembre, J. (2019) Estimating recent migration and population-size surfaces. *PLoS genetics* **15** : e1007908.

Arredondo, A., Mourato, B., Nguyen, K., Boitard, S., Rodríguez, W., Mazet, O. & Chikhi, L. (2021) Inferring number of populations and changes in connectivity under the n-island model. *Heredity* **126** : 896–912.

Bahlo, M. & Griffiths, R. C. (2000) Inference from gene trees in a subdivided population. *Theoretical Population Biology* **57** : 79–95.

Balding, D. J., Bishop, M. & Cannings, C. (2007) *Handbook of Statistical Genetics*. John Wiley & Sons, Chichester, UK, 3rd édition.

Barton, N., Etheridge, A. & Véber, A. (2010) A New Model for Evolution in a Spatial Continuum. *Electronic Journal of Probability* **15** : 162 – 216.

Barton, N. H., Etheridge, A. M., Kelleher, J. & Véber, A. (2013) Inference in two dimensions : Allele frequencies versus lengths of shared sequence blocks. *Theoretical Population Biology* **87**

Barton, N. H. & Gale, K. S. (1993) Genetic analysis of hybrid zones. In *Hybrid zones and the evolutionary process*, édité par Harrison, R. G., pp. 13–45. Oxford University Press, Oxford.

Bateman, A. J. (1950) Is gene dispersion normal ? *Heredity* **4** : 353–363.

Beaumont, M. (1999) Detecting population expansion and decline using microsatellites. *Genetics* **153** : 2013–2029.

Beaumont, M. A. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** : 1139–1160.

Beaumont, M. A., Nielsen, R., Robert, C., Hey, J., Gaggiotti, O., Knowles, L., Estoup, A., Panchal, M., Corander, J., Hickerson, M., Sisson, S. A., Fagundes, N., Chikhi, L., Beerli, P., Vitalis, R., Cornuet, J.-M., Huelsenbeck, J., Foll, M., Yang, Z., Rousset, F., Balding, D. & Excoffier, L. (2010) In defence of model-based inference in phylogeography. *Molecular ecology* **19** : 436–446.

Beaumont, M. A., Zhang, W. & Balding, D. J. (2002) Approximation Bayesian computation in population genetics. *Genetics* **162** : 2025–2035.

Beerli, P. (2004) Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology* **13** : 827–827.

Beerli, P. (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22** : 341–345.

Beerli, P. (2009) How to use MIGRATE or why are Markov chain Monte Carlo programs difficult to use ? In *Population Genetics for Animal Conservation*, pp. 42–79. Cambridge University Press.

Beerli, P. & Felsenstein, J. (1999) Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152** : 763–773.

Beerli, P. & Felsenstein, J. (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences* **98** : 4563–4568.

Beichman, A. C., Huerta-Sánchez, E. & Lohmueller, K. E. (2018) Using genomic data to infer historic population dynamics. *Annual Review of Ecology, Evolution, and Systematics* **49** : 433–456.

Bell, G. & Koufopanou, V. (1991) The architecture of the life cycle in small organisms. *Philosophical Transactions of the Royal Society of London. Series B : Biological Sciences* **332** : 81–89.

Blum, M. & François, O. (2010) Non-linear regression models for Approximate Bayesian Computation. *Statistical Computing* **20** : 63–73.

Boileau, M. G., Hebert, P. D. N. & Schwartz, S. S. (1992) Non-equilibrium gene frequency divergence : persistent founder effects in natural populations. *Journal of Evolutionary Biology* **5** : 25–39.

Boitard, S., Rodríguez, W., Jay, F., Mona, S. & Austerlitz, F. (2016) Inferring population size history from large samples of genome-wide molecular data– an approximate Bayesian computation approach. *PLoS genetics* **12** : e1005877.

Breiman, L. (2001) Random Forests. *Machine Learning* **45** : 5–32.

Broquet, T., Johnson, C. A., Petit, É., Burel, F. & Fryxell, J. M. (2006) Dispersal kurtosis and genetic structure in the American marten, *Martes americana*. *Molecular Ecology* **15** : 1689–1697.

Bullock, J. M., Mallada González, L., Tamme, R., Götzenberger, L., White, S. M., Pärtel, M. & Hooftman, D. A. P. (2017) A synthesis of empirical plant dispersal kernels. *Journal of Ecology* **105** : 6–19.

Burban, C., Petit, E. & Boursot, P. (2004) From sympatry to parapatry : a rapid change in the spatial context of incipient allochronic speciation. *Journal of evolutionary biology* **17** : 818–827.

Chakraborty, R. & Danker-Hopfe, H. (1991) Analysis of population structure : a comparative study of different estimators of Wright's fixation indices. In *Handbook of Statistics*, édité par Rao, C. R. & Chakraborty, R., vol. 8, pp. 203–254. Elsevier.

Chesson, P. & Lee, C. T. (2005) Families of discrete kernels for modeling dispersal. *Theoretical Population Biology* **67** : 241–256.

Chikhi, L., Sousa, V. C., Luisi, P., Goossens, B. & Beaumont, M. A. (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186** : 983–995.

Clark, J. S., Silman, M., Kern, R., Macklin, E. & HilleRisLambers, J. (1999) Seed dispersal near and far : patterns across temperate and tropical forests. *Ecology* **80** : 1475–1494.

Cockerham, C. C. (1969) Variance of gene frequencies. *Evolution* **23** : 72–84.

Cockerham, C. C. (1973) Analyses of gene frequencies. *Genetics* **74** : 679–700.

Cockerham, C. C. & Weir, B. S. (1987) Correlations, descent measures : drift with migration and mutation. *Proceedings of the National Academy of Science* **84** : 8512–8514.

Cornuet, J. M. & Beaumont, M. A. (2007) A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theoretical Population Biology* **71** : 12–19.

Cornuet, J. M. & Luikart, G. (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144** : 2001–2014.

Cox, D. R. (2006) *Principles of statistical inference*. Cambridge University Press, Cambridge, UK.

Cox, D. R. & Hinkley, D. V. (1974) *Theoretical statistics*. Chapman & Hall, London.

Crawford, T. (1984) The estimation of neighborhood parameters for plant populations. *Heredity* **52** : 273–283.

Crow, J. F. (1954) Breeding structure of populations. II. Effective population number. In *Statistics and mathematics in biology*, édité par Kempthorne, O., Bancroft, T. A., Gowen, J. W. & Lush, J. L., pp. 543–556. Iowa State University Press, Ames.

Crow, J. F. & Kimura, M. (1970) *An introduction to population genetics theory*. Harper & Row, New York.

Csilléry, K., Blum, M. G., Gaggiotti, O. E. & François, O. (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution* **25** : 410–418.

de Iorio, M. & Griffiths, R. C. (2004a) Importance sampling on coalescent histories. *Advances in Applied Probability* **36** : 417–433.

de Iorio, M. & Griffiths, R. C. (2004b) Importance sampling on coalescent histories. II. Subdivided population models. *Advances in Applied Probability* **36** : 434–454.

de Iorio, M., Griffiths, R. C., Leblois, R. & Rousset, F. (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology* **68** : 41–53.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. & Weissenbach, J. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380** : 152–4.

Diggle, P. J. & Gratton, R. J. (1984) Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society : Series B (Methodological)* **46** : 193–212.

Dobzhansky, T. & Wright, S. (1941) Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of *Drosophila pseudoobscura*. *Genetics* **26** : 23–51.

Donnelly, P. (1999) The coalescent and microsatellite variability. In *Microsatellites : Evolution and Applications*, édité par Goldstein, D. & Schlotterer, C., pp. 116–128. Oxford University Press, Oxford.

Endler, J. A. (1977) *Geographical variation, speciation, and clines*. Princeton University Press, Princeton.

Estoup, A. & Angers, B. (1998) Microsatellites and minisatellites for molecular ecology : theoretical and empirical considerations. In *Advances in molecular ecology*, édité par Carvalho, G., pp. 55–86. IOS Press, Amsterdam.

Estoup, A. & Cornuet, J.-M. (1998) Microsatellite evolution : inferences from population data. In *Microsatellites : evolution and applications*, édité par Goldstein, D. B. & Schlötterer, C., pp. 49–65. Oxford University Press, Oxford.

Estoup, A., Jarne, P. & Cornuet, J. M. (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* **11** : 1591–1604.

Estoup, A., Lombaert, E., Marin, J.-M., Robert, C., Guillemaud, T., Pudlo, P. & Cornuet, J.-M. (2012) Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. *Molecular Ecology Resources* **12** : 846–855.

Estoup, A., Wilson, I. J., Sullivan, C., Cornuet, J. M. & Moritz, C. (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics* **159** : 1671–1687.

Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3** : 87–112.

Ewens, W. J. (2004) *Mathematical population genetics I. Theoretical introduction*. Springer Verlag, New York, second édition.

Excoffier, L. (2001) Analysis of population subdivision. In *Handbook of statistical genetics*, édité par Balding, D. J., Bishop, M. & Cannings, C., pp. 271–307. Wiley, Chichester, U.K.

Excoffier, L. & Foll, M. (2011) fastsimcoal : a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27** : 1332–1334.

Excoffier, L. & Ray, N. (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology & Evolution* **23** : 347–351.

Faubet, P., Waples, R. S. & Gaggiotti, O. E. (2007) Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Molecular Ecology* **16** : 1149–1166.

Fearnhead, P. & Donnelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics* **159** : 1299–1318.

Felsenstein, J. (1975) A pain in the torus : some difficulties with models of isolation by distance. *American Naturalist* **109** : 359–368.

Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer Associates.

Felsenstein, J., Kuhner, M. K., Yamato, J. & Beerli, P. (1999) Likelihoods on coalescents : a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics*, édité par Seillier-Moiseiwitsch, F., vol. 33, pp. 163–185. Institute of Mathematical Statistics, Hayward, California.

Fenster, C. B., Vekemans, X. & Hardy, O. J. (2003) Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). *Evolution* **57** : 995–1007.

Fraimout, A., Debat, V., Fellous, S., Hufbauer, R. A., Foucaud, J., Pudlo, P., Marin, J.-M., Price, D. K., Cattel, J., Chen, X., Depra, M., Duyck, P. F., Guedot, C., Kenis, M., Kimura, M. T., Loeb, G., Loiseau, A. & Estoup, A. (2017) Deciphering the routes of invasion of Drosophila suzukii by means of ABC Random Forest. *Molecular biology and evolution* **34** : 980–996.

Fu, Y.-X. (1997) Coalescent theory for partially selfing populations. *Genetics* **146** : 1489–1499.

Gandon, S. & Rousset, F. (1999) Evolution of stepping stone dispersal rates. *Proceedings of the Royal Society of London B* **266** : 2507–2513.

Gelman, A. & Meng, X.-L. (1998) Simulating normalizing constants : From importance sampling to bridge sampling to path sampling. *Statistical Science* **13** : 163–185.

Gillespie, J. H. (2004) *Population Genetics : A Concise Guide*. JHU Press, Baltimore, MD.

Girod, C., Vitalis, R., Leblois, R. & Fréville, H. (2011) Inferring population decline and expansion from microsatellite data : a simulation-based evaluation of the Msvar method. *Genetics* **188** : 165–179.

Gonser, R., Donnelly, P., Nicholson, G. & Di Rienzo, A. (2000) Microsatellite mutations and inferences about human demography. *Genetics* **154** : 1793–1807.

Griffiths, R. & S.Tavaré (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society (London) B* **344** : 403–410.

Griffiths, R. & Tavaré, S. (1994a) Ancestral inference in population genetics. *Statistical Science* **9** : 307–319.

Griffiths, R. & Tavaré, S. (1994b) Simulating probability distributions in the coalescent. *Theoretical Population Biology* **46** : 131–159.

Guillot, G., Leblois, R., Coulon, A. & Frantz, A. C. (2009) Statistical methods in spatial genetics. *Molecular Ecology* **18** : 4734–4756.

Guindon, S., Guo, H. & Welch, D. (2016) Demographic inference under the coalescent in a spatial continuum. *Theoretical Population Biology* **111**

Haldane, J. B. S. (1919) The combination of linkage values and the calculation of distance between the loci fo linked factors. *Hournal of Genetics* **8** : 299–309.

Hardy, O. J. (2003) Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes unsing dominant genetic markers. *Molecular Ecology* **12** : 1577–1588.

Hardy, O. J. & Vekemans, X. (1999) Isolation by distance in a continuous population : reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83** : 145–154.

Hartl, D. L. & Clark, A. G. (2007) *Principles of Population Genetics*. Sinauer Associates, 4th édition.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** : 97–109.

Hein, J., Schierup, M. H. & Wiuf, C. (2005) *Gene genealogies, variation and evolution*. Oxford University Press, Oxford, UK.

Hey, J. (2010) Isolation with migration models for more than two populations. *Molecular Biology and Evolution* **27** : 905–920.

Hey, J., Chung, Y. S. & Sethuraman, A. (2015) On the occurrence of false positives in tests of migration under an isolation-with-migration model. *Molecular ecology* **24** : 5078–5083.

Hey, J. & Nielsen, R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167** : 747–760.

Hey, J. & Nielsen, R. (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Science* **104** : 2785–2790.

Hudson, R. R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23** : 183–201.

Hudson, R. R. (1990) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, édité par Antonovics, D. F. & J., vol. 7, pp. 1–43. Oxford University Press, Oxford.

Hudson, R. R. (1993) The how and why of generating gene genealogies. In *Mechanisms of molecular evolution*, édité par Takahata, N. & Clark, A. G., pp. 23–36. Sinauer, Sunderland, MA.

Hudson, R. R. (1998) Island models and the coalescent process. *Molecular Ecology* **7** : 413–418.

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18** : 337–338.

Jakobsson, M., Edge, M. D. & Rosenberg, N. A. (2013) The Relationship Between FST and the Frequency of the Most Frequent Allele. *Genetics* **193** : 515–528.

Joyce, P. & Marjoram, P. (2008) Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology* **7** : 26.

Juhel, A.-S., Barbu, C., Valantin-Morison, M., Gauffre, B., Leblois, R., Olivares, J. & Franck, P. (2019) Limited genetic structure and demographic expansion of the Brassicogethes aeneus populations in France and in Europe. *Pest Management Science* **75** : 667–675.

Kelleher, J., Etheridge, A. M. & McVean, G. (2016) Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology* **12** : e1004842.

Kempthorne, O. (1954) The correlation between relatives in a random mating population. *Proceedings of the Royal Society of London B* **143** : 103–113.

Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61** : 893–903.

Kimura, M. & Crow, J. F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* **49** : 725–738.

Kimura, M. & Weiss, G. H. (1964) The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. *Genetics* **49** : 561–576.

Kingman, J. F. C. (1982a) The coalescent. *Stochastic Processes and their Applications* **13** : 235 – 248.

Kingman, J. F. C. (1982b) On the Genealogy of Large Populations. *Journal of Applied Probability* **19** : 27–43.

Koenig, W. D., Vuren, D. V. & Hooge, P. N. (1996) Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology and Evolution* **11** : 514–517.

Kot, M., Lewis, M. A. & Driessche, P. v. d. (1996) Dispersal data and the spread of invading organisms. *Ecology* **77** : 2027–2042.

Kuhner, M. K. (2006) LAMARC 2.0 : maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22** : 768–770.

Kuhner, M. K. (2009) Coalescent genealogy samplers : windows into population history. *Trends in Ecology & Evolution* **24** : 86–93.

Laval, G. & Excoffier, L. (2004) SIMCOAL 2.0 : a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics* **20** : 2485–2487.

Leblois, R. (2004) *Estimation de paramètres de dispersion en populations structurées à partir de données génétiques.* Thèse, École Nationale Supérieure Agronomique de Montpellier.

Leblois, R., Estoup, A. & Rousset, F. (2003) Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. *Molecular Biology and Evolution* **20** : 491–502.

Leblois, R., Estoup, A. & Rousset, F. (2009) IBDSim : A computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources* **9** : 107–109.

Leblois, R., Estoup, A. & Streiff, R. (2006) Habitat contraction and reduction in population size : Does isolation by distance matter ? *Molecular Ecology* **15** : 3601–3615.

Leblois, R., Pudlo, P., Néron, J., Bertaux, F., Beeravolu, C. R., Vitalis, R. & Rousset, F. (2014) Maximum likelihood inference of population size contractions from microsatellite data. *Molecular Biology and Evolution* **31** : 2805–2823.

Leblois, R., Rousset, F. & Estoup, A. (2004) Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics* **166** : 1081–1092.

Leblois, R., Rousset, F., Tikel, D., Moritz, C. & Estoup, A. (2000) Absence of evidence for isolation by distance in an expanding cane toad (Bufo marinus) population : an individual-based analysis of microsatellite genotypes. *MOLECULAR ECOLOGY* **9** : 1905–1909.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G. & Govaert, G. (2015) Rmixmod : The R Package of the Model-Based Unsupervised, Supervised, and Semi-Supervised Classification Mixmod Library. *Journal of Statistical Software* **67** : 1–29.

Lewis, W., van Lenteren, J., Phatak, S. & Tumlinson, J. (1997) A total system approach to sustainable pest management. *Proceedings of the National Academy of Sciences* **94** : 12243–12248.

Li, H. & Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature* **475** : 493–496.

Li, N. & Stephens, M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165** : 2213–2233.

Li, W. H. (1976) Effect of Migration on Genetic Distance. *American Naturalist* **110** : 841–847.

Lippens, C., Estoup, A., Hima, K., Loiseau, A., Tatard, C., Dalecky, A., Bâ, K., Kane, M., Diallo, M., Sow, A., Piry, S., Leblois, R., Duplantier, J.-M. & Brouat, C. (2017) Genetic structure and invasion history of the house mouse (*Mus musculus domesticus*) in Senegal : a legacy of colonial times ? *Heredity* **119** : 64–75.

Liu, J. S. (2004) *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Loiselle, B. A., Sork, V. L., Nason, J. & Graham, C. (1995) Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). *American journal of botany* **82** : 1420–1425.

Long, J. C., Naidu, J. M., Mohrenweiser, H. W., Gershowitz, H., Johnson, P. L. & Wood, J. W. (1986) Genetic characterization of Gainj- and Kalam-speaking peoples of Papua New Guinea. *American Journal of Physical Anthropology* **70** : 75–96.

Luikart, G., England, P., Tallmon, D., Jordan, S. & Taberlet, P. (2003) The power and promise of population genomics : From genotyping to genome typing. *Nature Reviews Genetics* **4** : 981–994.

Malécot, G. (1948) *Les mathématiques de l'hérédité*. Masson, Paris.

Malécot, G. (1950) Quelques schémas probabilistes sur la variabilité des populations naturelles. *Annales de l'Université de Lyon A* **13** : 37–60.

Malécot, G. (1966) *Probabilité et Hérédité*, vol. 47 of *Travaux et Documents*. Institut national d'études démographiques, presses universitaires de france édition.

Malécot, G. (1967) Identical loci and relationship. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, édité par Le Cam, L. M. & Neyman, J., vol. 4, pp. 317–332. University of California Press, Berkeley.

Malécot, G. (1972) Génétique des populations naturelles dans le cas d'un seul locus II — Etude du coefficient de parenté. *Annales de Génétique et de Sélection Animale* **4** : 385–409.

Malécot, G. (1975) Heterozygosity and relationship in regularly subdivided populations. *Theoretical Population Biology* **8** : 212–241.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* **27** : 209–220.

Marchi, N., Schlichta, F. & Excoffier, L. (2021) Demographic inference. *Current biology : CB* **31** : R276–R279.

Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* **100** : 15324–15328.

Marjoram, P. & Tavaré, S. (2006) Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet* **7** : 759–770.

Maruyama, T. (1972) Distribution of gene frequencies in a geographically structured finite population. 1. Distribution of neutral genes and of genes with small effect. *Annals of Human Genetics* **35** : 411–423.

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. (2016) On the importance of being structured : instantaneous coalescence rates and human evolution—lessons for ancestral population size inference ? *Heredity* **116** : 362–371.

McVean, G. A. & Cardin, N. J. (2005) Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B : Biological Sciences* **360** : 1387–1393.

Meier, J. I., Salazar, P. A., Kučka, M., Lohse, K., Guerrero, P. C., Nadeau, N. J., Morrison, C. R., Zhang, W., Papa, R., Martin, S. H. & Chan, Y. F. (2021) Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proceedings of the National Academy of Sciences* **118** : e2015005118.

Merle, C., Leblois, R., Rousset, F. & Pudlo, P. (2017) Resampling : an improvement of Importance Sampling in varying population size models. *Theoretical Population Biology* **114** : 70–87.

Nagylaki, T. (1980) The strong migration limit in geographically structured populations. *Journal of Mathematical Biology* **9** : 101–114.

Nagylaki, T. (1986) Neutral models of geographical variation. In *Stochastic Spatial Processes*, pp. 216–237. Springer.

Nagylaki, T. (1989) Gustave Malécot and the transition from classical to modern population genetics. *Genetics* **122** : 253–268.

Nath, H. B. & Griffiths, R. C. (1996) Estimation in an island model using simulation. *Theoretical Population Biology* **50** : 227–253.

Nathan, R., Klein, E. K., Robledo-Arnuncio, J. J. & Revilla, E. (2012) Dispersal kernels. In *Dispersal and Spatial Evolutionary Ecology*, édité par Clobert, J., Baguette, M., Benton, T. & Bullock, J., pp. 123–135. Oxford University Press.

Nei, M. (1973) Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences* **70** : 3321–3323.

Neuhauser, C. & Krone, S. M. (1997) The genealogy of samples in models with selection. *Genetics* **145** : 519–534.

Nielsen, R. (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**

Nielsen, R. & Sltakin, M. (2013) *An Introduction to Population Genetics : Theory and Applications*. Sinauer Associates, Sunderland, MA.

Nielsen, R. & Wakeley, J. (2001) Distinguishing migration from isolation : a Markov chain Monte Carlo approach. *Genetics* **158** : 885–896.

Nordborg, M. (1997) Structured coalescent processes on different time scales. *Genetics* **146** : 1501–1514.

Nordborg, M. (2001) Coalescent theory. In *Handbook of statistical genetics*, édité par Balding, D. J., Bishop, M. & Cannings, C., pp. 179–212. Wiley, Chichester, U.K.

Nordborg, M. & Krone, S. M. (2002) Separation of time scales and convergence to the coalescent in structured populations. In *Modern developments in theoretical population genetics*, édité par Slatkin, M. & Veuille, M., pp. 194–232. Oxford University Press, Oxford.

Nordborg, M. & Tavare, S. (2002) Linkage disequilibrium : what history has to tell us. *Trends in Genetics* **18** : 83–90.

Notohara, M. (1993) The genealogical process of neutral genes with mutation in geographically structured populations. *Journal of Mathematical Biology* **31** : 123–132.

Novembre, J. & Peter, B. M. (2016) Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics & Development* **41** : 98–105.

Ohta, T. & Kimura, M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetics* **22** : 201–204.

Patil, G. P. & Joshi, S. W. (1968) *A dictionary and bibliography of discrete distributions*. Oliver & Boyd, Edinburgh.

Peter, B., Wegmann, D. & Excoffier, L. (2010) Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Molecular Ecology* **19** : 4648–4660.

Petkova, D., Novembre, J. & Stephens, M. (2016) Visualizing spatial population structure with estimated effective migration surfaces. *Nature genetics* **48** : 94–100.

Portnoy, S. & Willson, M. F. (1993) Seed dispersal curves : behavior of the tails of the distribution. *Evolutionary Ecolgy* **7** : 25–44.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. (1999) Population growth of human Y chromosomes : a study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16** : 1791–8.

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M. & Robert, C. P. (2016) Reliable ABC model choice via random forests. *Bioinformatics* **32** : 859–866.

Raynal, L., Marin, J.-M., Pudlo, P., Ribatet, M., Robert, C. P. & Estoup, A. (2018) ABC random forests for Bayesian parameter inference. *Bioinformatics* **35** : 1720–1728.

Ringbauer, H., Coop, G. & Barton, N. H. (2017) Inferring Recent Demography from Isolation by Distance of Long Shared Sequence Blocks. *Genetics* **205** : 1335–1351.

Robledo-Arnuncio, J. J. & Rousset, F. (2010) Isolation by distance in a continuous population under stochastic demographic fluctuations. *Journal of Evolutionary Biology* **23** : 53–71.

Rousset, F. (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics* **142** : 1357–1362.

Rousset, F. (1997) Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance. *Genetics* **145** : 1219–1228.

Rousset, F. (1999a) Genetic differentiation in populations with different classes of individuals. *Theoretical Population Biology* **55** : 297–308.

Rousset, F. (1999b) Genetic differentiation within and between two habitats. *Genetics* **151** : 397–407.

Rousset, F. (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology* **13** : 58–62.

Rousset, F. (2001) Inferences from spatial population genetics. In *Handbook of statistical genetics*, édité par Balding, D. J., Bishop, M. & Cannings, C., pp. 239–269. Wiley, Chichester, U.K.

Rousset, F. (2004) *Genetic structure and selection in subdivided populations*. Princeton University Press, Princeton, New Jersey.

Rousset, F. (2006) Separation of time scales, fixation probabilities and convergence to evolutionarily stable states under isolation by distance. *Theoretical Population Biology* **69** : 165–179.

Rousset, F. (2008) GENEPOP007 : a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8** : 103–106.

Rousset, F. (2013) Exegeses on maximum genetic differentiation. *Genetics* **194**

Rousset, F., Beeravolu, C. R. & Leblois, R. (2018) Likelihood computation and inference of demographic and mutational parameters from population genetic data under coalescent approximations. *Journal de la société Française de Statistique* **159** : 142–166.

Rousset, F., Gouy, A., Martinez-Almoyna, C. & Courtiol, A. (2017) The summary-likelihood method and its implementation in the Infusion package. *Molecular Ecology Resources* **17** : 110–119.

Rousset, F. & Leblois, R. (2007) Likelihood and approximate likelihood analyses of genetic structure in a linear habitat : performance and robustness to model misspecification. *Molecular Biology and Evolution* **24** : 2730–2745.

Rousset, F. & Leblois, R. (2012) Likelihood-based inferences under a coalescent model of isolation by distance : two-dimensional habitats and confidence intervals. *Molecular Biology and Evolution* **29** : 957–973.

Roze, D. & Rousset, F. (2003) Diffusion approximations for selection and drift in subdivided populations : a straightforward method and examples involving dominance, selfing and local extinctions. *Genetics* **165** : 2153–2166.

Sacks, J., Welch, W. J., Mitchell, T. J. & Wynn, H. P. (1989) Design and analysis of computer experiments. *Statistical Sciences* **4** : 409–435.

Sanchez, T., Cury, J., Charpiat, G. & Jay, F. (2021) Deep learning for population size history inference : Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources* **21**

Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics* **28** : 35–42.

Saunders, I. W., Tavare, S. & Watterson, G. A. (1984) On the genealogy of nested subsamples from a haploid population. *Advances in Applied Probability* **16** : 471–491.

Sawyer, S. (1977) Asymptotic properties of the equilibrium probability of identity in a geographically structured population. *Advances in Applied Probabilities* **9** : 268–282.

Sawyer, S. & Felsenstein, J. (1981) A continuous migration model with stable demography. *Journal of Mathematical Biology* **11** : 193–205.

Sisson, S. A., Fan, Y. & Tanaka, M. (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **104** : 1760–1765.

Slatkin, M. (1991) Inbreeding coefficients and coalescence times. *Genetics* **58** : 167–175.

Slatkin, M. (1993) Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47** : 264–279.

Slatkin, M. (1994) Gene flow and population structure. In *Ecological Genetics*, édité par Real, L. A., pp. 3–17. Princeton University Press, Princeton, New Jersey.

Slatkin, M. (2005) Seeing ghosts : the effect of unsampled populations on migration rates estimated between sampled populations. *Molecular Ecology* **14** : 67–73.

Smith, C. C. R., Tittes, S., Ralph, P. L. & Kern, A. D. (2022) Dispersal inference from population genetic variation using a convolutional neural network. *bioRxiv*

Spong, G. & Creel, S. (2001) Deriving dispersal distance from genetic data. *Proceedings of the Royal Society of London B* **268** : 2571–2574.

Stein, M. L. (1999) *Interpolation of spatial data : some theory for kriging*. Springer.

Stephens, M. & Donnelly, P. (2000) Inference in molecular population genetics (with discussion). *Journal of the Royal Society of Statistics* **62** : 605–655.

Storz, J. & Beaumont, M. (2002) Testing for genetic evidence of population expansion and contraction : An empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56** : 154–166.

Sumner, J., Estoup, A., Rousset, F. & Moritz, C. (2001) 'Neighborhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. *Molecular Ecology* **10** : 1917–1927.

Swofford, D., Olsen, G., Waddelland, P. & Hillis, D. (1996) Phylogentic inference. In *Molecular Systematics*, édité par Hillis, D., Moritz, C. & Mable, B., pp. 407 – 514. Sinauer Associates, Sunderland, MA.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105** : 437–460.

Takahata, N. (1991) Genealogy of neutral genes and spreading of selected mutations in a geographically structured population. *Genetics* **129** : 585–595.

Takezaki, N. & Nei, M. (1996) Genetic distances and reconstruction of phylogenetic tree from microsatellites DNA. *Genetics* **144** : 389–399.

Tataru, P., Simonsen, M., Bataillon, T. & Hobolth, A. (2016) Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data. *Systematic Biology* **66** : e30–e46.

Tavaré, S., Balding, D. J., Griffiths, R. & Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145** : 505–518.

Tournayre, O., Pons, J.-B., Leuchtmann, M., Leblois, R., Piry, S., Filippi-Codaccioni, O., Loiseau, A., Duhayer, J., Garin, I., Mathews, F., Puechmaille, S., Charbonnel, N. & Pontier, D. (2019) Integrating population genetics to define conservation units from the core to the edge of *Rhinolophus ferrumequinum* Western range. *Ecology and Evolution* **9** : 12272–12290.

Vekemans, X. & Hardy, O. J. (2004) New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13** : 921–934.

Veres, A., Petit, S., Conord, C. & Lavigne, C. (2013) Does landscape composition affect pest abundance and crop quality ? *Agriculture, Ecosystems & Environment* **166** : 110–117.

Virgoulay, T. (2022) *Inférences démographiques à partir de données génomiques sous des modèles spatialisés*. Thèse, Université de Montpellier.

Virgoulay, T., Rousset, F. & Leblois, R. (2021) GSpace : an exact coalescence simulator of recombining genomes under isolation by distance. *Bioinformatics* **37** : 3673–3675.

Vitalis, R. & Couvet, D. (2001a) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics* **157** : 911–925.

Vitalis, R. & Couvet, D. (2001b) Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. *Genetics* **77** : 67–81.

Vitalis, R., Gautier, M., Dawson, K. J. & Beaumont, M. A. (2014) Detecting and Measuring Selection from Gene Frequency Data. *Genetics* **196** : 799–817.

Véber, A. & Wakolbinger, A. (2015) The spatial Lambda-Fleming - Viot process : An event-based construction and a lookdown representation. *Annales de l'institut Henri Poincare (B) Probability and Statistics* **51**

Wakeley, J. (2008) *Coalescent Theory : An Introduction.* Roberts & Company Publishers.

Wakeley, J. (2010) Natural selection and coalescent theory. In *Evolution since Darwin : the first 150 years*, édité par Bell, M. A., Futuyama, D. J., Eanes, W. F. & Levinton, J. S., pp. 371–390. Sinauer Associates, Sunderland, MA.

Watts, P. C., Rousset, F., Saccheri, I. J., Leblois, R., Kemp, S. J. & Thompson, D. J. (2006) Compatible genetic and ecological estimates of dispersal rates in insect (Coenagrion mercuriale : Odonata : Zygoptera) populations : analysis of 'neighbourhood size' using a more precise estimator. *Molecular Ecology* **16** : 737–751.

Wegmann, D., Leuenberger, C. & Excoffier, L. (2009) Efficient Approximate Bayesian Computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182** : 1207–1218.

Weir, B. & Cockerham, C. (1974) Behavior of pairs of loci in finite monoecious populations. *Theoretical Population Biology* **6** : 323–354.

Weir, B. S. & Cockerham, C. C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* **38** : 1358–1370.

Whitlock, M. C. (2003) Fixation probability and time in subdivided populations. *Genetics* **164** : 767–779.

Whitlock, M. C. & McCauley, D. E. (1999) Indirect measures of gene flow and migration : $F_{\mathrm{ST}} \neq 1/(4Nm+1)$. *Heredity* **82** : 117–125.

Wilkins, J. F. (2004) A separation-of-timescales approach to the coalescent in a continuous population. *Genetics* **168** : 2227–2244.

Wilkins, J. F. & Wakeley, J. (2002) The Coalescent in a Continuous, Finite, Linear Population. *Genetics* **161** : 873–888.

Wilson, G. A. & Rannala, B. (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163** : 1177–1191.

Wilson, I. J. & Balding, D. J. (1998) Genealogical inference from microsatellite data. *Genetics* **150**

Winters, J. B. & Waser, P. M. (2003) Gene dispersal and outbreeding in a philopatric mammal. *Molecular Ecology* **12** : 2251–2259.

Wood, J. W., Smouse, P. E. & Long, J. C. (1985) Sex-specific dispersal patterns in two human populations of highland New Guinea. *American Naturalist* **125** : 747–768.

Wright, M. N. & Ziegler, A. (2017) ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **77** : 1–17.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics* **16** : 97–159.

Wright, S. (1943) Isolation by distance. *Genetics* **28** : 114–138.

Wright, S. (1946) Isolation by distance under diverse systems of mating. *Genetics* **31** : 39–59.

Wright, S. (1951) The genetical structure of populations. *Annals of Eugenics* **15** : 323–354.

Wright, S. (1969) *Evolution and the genetics of populations. II. The theory of gene frequencies.* University of Chicago Press, Chicago.

.1 Fiche de Synthèse

I – Fiche de synthèse (1 page RV)

Raphaël Leblois E-mail : raphael.leblois@inrae.fr
Né le 26 mai 1977, à Paris 14e

ORCHID : 0000-0002-3051-4497 idHAL : raphael-leblois
CNU : Section 67 - Biologie des populations et écologie

Chargé de Recherche Classe Normale INRAE, au laboratoire Centre de Biologie pour la Gestion des Populations (CBGP), Montpellier.

Diplôme de Doctorat soutenu le 3 mai 2004
Spécialité : Biologie de l'Évolution ;
Formation Doctorale : Biologie de l'Évolution et Écologie ;
École Doctorale : Biologie Intégrative

Principales collaborations :

Principal collaborateur depuis mon master : F. Rousset
Au CBGP,
en génétique des pops théorique : A. Estoup, S. Boitard, M. Gautier, M. Navascues, R. Vitalis ;
sur des données et encadrement : C. Brouat, M.-P. Chapuis, C. Kerdelhué, N. Charbonnel ;
en bio-info : E. Ortega, S. Piry,
en biologie moléculaire : L. Benoit, A. Loiseau, S. Nidelet, L. Sauné.
A l'INRAE, sur de la génétique spatiale des populations : B. Gauffre, K. Berthier.
A l'IMAG, en statistique inférentielle : J.-M. Marin, P. Pudlo, P. Bastide.
Au CEFE, sur données reptiles : P.-A. Crochet.
A l'international : génétique des populations théorique : A. Hobolth (Danemark), J.-B. Ledoux (Portugal)

Principaux contrats de recherche :
2022-2025: BioDivOC, 430 k€, co-porteur
2020-2024: ANR, 517 k€, co-resp. d'un Work Package
2017-2021: ANR, 140 k€, co-porteur
2012-2017: PIA ANR "IBC" (Institut de Biologie Computationnelle), 2.5 M€, participant
2010-2014: ANR, 500 k€, resp. d'un Work Package
2010-2014: ANR, 809 k€, resp. d'un Work Package
2009-2013: ANR, 372 k€, co-resp. d'un Work Package

Encadrement :
6 étudiant.e.s en Master 1, 16 étudiant.e.s en Master 2, 3 étudiant.e.s en thèse, 2 post-doctorant.e.s, 1 ATER

Publications & communications :
Nombre de publications : 64 (au 27/02/2023)
Nombre de citations : 3467 / 5370 (*indices d'impact SCI / google scholar*)
H-index : 27 / 30
dont 11 en premier auteur, 5 en dernier, et 6 avec des masters, 2 avec doctorant.e.s, 5 avec des post-doctorant.e.s, ainsi que plus de 45 communications orales.

Principales publications A choisies en lien avec la thématique du rapport, jointes au dossier :

Virgoulay, Rousset, Leblois. **2021**. GSpace: an exact coalescence simulator of recombining genomes under isolation by distance. *Bioinformatics.*

Rousset, Beeravolu, Leblois. **2018**. Likelihood computation and inference of demographic and mutational parameters from population genetic data under coalescent approximations. *J. Soc. Fr. Stat.*

Leblois, .., Rousset. **2014**. Maximum likelihood inference of population size contractions from microsatellite data. *MBE*.

Rousset, Leblois. **2012**. Likelihood-based inferences under a coalescent model of isolation by distance: two-dimensional habitats and confidence intervals. *MBE.*

Rousset, Leblois. **2007.** Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *MBE.*

Leblois, Rousset, Estoup. **2004**. Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population from microsatellite data. *Genetics*.

Leblois, Estoup, Rousset. **2003**. Influence of mutational and sampling factors on the estimation of demographic parameters in a continuous population under isolation by distance. *MBE*.

Résumé des thématiques de recherche et des projets :

Durant toute ma carrière, j'ai travaillé sur l'inférence de paramètres démographiques et historiques des populations, à partir de données génétiques, en me plaçant à l'interface entre la génétique des populations théoriques, la statistique inférentielle et la bio-informatique au sens large, d'un côté ; et la génétique des populations empirique, la biologie évolutive et l'écologie (moléculaire), de l'autre. Je me suis spécialisé dans la génétique spatiale des population et l'inférence de la dispersion, de la densité et des tailles de populations sous des modèles d'isolement par la distance. Dans ce contexte, avec François Rousset, nous avons développé et testé de nombreuses méthodes d'inférence, dans le but d'aller toujours vers des estimations plus complètes, plus précises, et toujours robustes, à l'aide des différentes approches d'inférence statistique. Mon projet est la suite directe de tout ce travail puisqu'il consiste à étendre nos méthodes actuelles d'inférence par simulation sous des modèles spatialisé, pour en améliorer le réalisme, et en élargir le champ d'application, notamment en prenant en compte les hétérogénéités spatiales et temporelles des paramètres démographiques, mais aussi en perfectionnant l'utilisation de l'information présente dans les données génomiques spatialisées (déséquilibre de liaison et information spatiale).

.2 CV

Lien vers mon CV HAL : https://cv.archives-ouvertes.fr/raphael-leblois

Raphaël Leblois
43 ans (26 mai 1977), nationalité française
Marié, deux enfants

E-mail : raphael.leblois@inrae.fr

ORCHID : 0000-0002-3051-4497

idHAL : raphael-leblois

Adresse personnelle :

17 route de Montferrier
34790 Grabels
Tél : + 33 (0)7 81 38 35 03

Adresse professionnelle :

Centre de Biologie pour la Gestion des Populations
UMR CBGP 1062 (INRAE-IRD-CIRAD-Supagro)
Campus International de Baillarguet CS 30016
34980 Montferrier-sur-Lez cedex, France
tél : +33 (0)4 99 62 33 31 fax : +33 (0)4 99 62 33 45

SITUATION ACTUELLE

Sept 2010- Aujourd'hui

Chercheur (CRCN) au Centre de Biologie pour la Gestion des Populations (INRAE, Montpellier).

FORMATION & ACTIVITES DE RECHERCHE

2012 – 2019
Membre de l' "Institut de Biologie Computationnelle" IBC, Montpellier, France

2006-2010
Maître de Conférence au Muséum National d'Histoire Naturelle (MNHN) à Paris.

2005-2006
PostDoc : « *Inférences démographiques à partir de données génétiques avec applications aux populations Humaines d'Asie Centrale* » avec Pr. Evelyne Heyer et Dr. Renaud Vitalis, Musée de l'Homme - MNHN, Paris, France.

2004-2005
PostDoc, Lauréat Bourse Lavoisier: « *Détection de goulet d'étranglement et estimation du nombre d'individus fondateurs à partir de marqueurs SNPs* **»**, avec Pr. Montgomery Slatkin, Université de Californie, Berkeley, Etats-Unis.

2000-2004
Thèse de Doctorat : « *Estimation de paramètres de dispersion à partir de données génétiques en populations subdivisées* **»** dirigé par Dr François Rousset et Dr Arnaud Estoup. Supagro-Montpellier - Université Montpellier II.

1997-2000
École Nationale Supérieure d'Agronomie de Montpellier (Supagro), diplôme d'ingénieur agronome et DEA en Écologie et Évolution.

COMPETENCES & SAVOIR FAIRE

Outils statistiques et informatiques :

Divers systèmes d'exploitation (Windows, Linux, Mac)
Bureautique (Office, OpenOffice, Latex),
Programmation plus ou moins avancée en C, C++, R, bash, awk, Perl, Python et Mathematica,

Techniques moléculaires :

Extractions d'ADN, PCR, Développement et Génotypage de marqueurs microsatellites.

Activités de terrain :

Echantillonnage amphibiens/reptiles/insectes/palmiers (spécimens et tissus)
Gestion du travail de terrain et d'échantillonnage en France et à l'étranger.

Domaine d'expertise :

Génétique des populations théorique et appliquée, génétique de la conservation, estimation de la dispersion, modélisation, théorie de la coalescence.
Relecteur pour les journaux suivants : *Cladistic, Conservation Biology, Ecography, Ecology Letters, eLife, Evolutionnary Applications, Evolutionary Ecology Research, Forest Ecology and Management, Genetics, Heredity, Human Biology, Mitochondrial DNA, Molecular Ecology,* and *Molecular Ecology Resources, New Phytologist, Plant Ecology, PLosOne, Proceedings of the Royal Society London.* Member of the *Peer Community in Evolutionary Biology*.

Langues :

Anglais (bonne maitrise orale et écrite).

ENSEIGNEMENT ET ENCADREMENT

Voir ci-dessous, sections 8 et 9

PRINCIPAUX FINANCEMENTS

Voir ci-dessous, section 4

RESUME DES PUBLICATIONS

Voir ci-dessous, section 7

.3 Tâches collectives

3- Participation à des tâches collectives

Depuis 2022, je suis **référent Développement Durable** du CBGP. Cette tache me plait énormément et j'y consacre 10 à 20% de mon temps de travail.

Je suis **co-organisateur des conférences MCEB** 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2021, 2022 : Colloque international « **Mathematical and Computational Evolutionary Biology** » (Mathématiques et sciences computationnelles pour la biologie évolutive) organisé chaque année de 2013 à aujourd'hui, ayant lieu en alternance St Martin de Londres (Hérault) ou Porquerolles (Var) et attirant de 75 à 90 participants dont 50 à 70% d'étrangers. Ce colloque est co-organisé par le CBGP, L'ISEM et le LIRMM, et a été initié dans le cadre de l'IBC. Deux numéros spéciaux de Systematic Biology consacrés à MCEB sont parus en 2014 et 2017.

Au Museum, puis au CBGP, j'ai souvent participé à la **gestion des plateformes locales de calcul intensif**, notamment en tant que responsable scientifique (voir CV HAL).

Enfin, j'ai régulièrement été **membre des conseils d'unité** du CBGP et de l'OSEB.

Résumé par périodes :

2021-aujourd'hui : Membre du comité pour l'enseignement du CBGP.

2022-aujourd'hui : Référent Développement Durable au CBGP (10%).

2013-aujourd'hui : Membre du comité scientifique et d'organisation des congrès MCEB (Mathematical and Computational Evolutionary Biology).

2014-2019 : Membre du comité scientifique des utilisateurs de la plateforme INRAE de calcul et de bio-informatique Migale

2014-2015 : Membre du GAS "Groupe d'Animation Scientifique" du CBGP.

2014-2016 et 2018-2020 : Membre du conseil d'unité CU du CBGP.

2011-2015 : Responsable scientifique du cluster informatique du CBGP (132 nœuds, plateforme technique du CBGP).

2008-2010 : Organisation de séminaires internes pour le département "systématique et évolution" du MNHN.

2008-2010 : Organisation de séminaires internes au laboratoire "Origine, Structure et Évolution de la biodiversité" (OSEB, MNHN, UMR 7205).

2008-2010 : Responsable scientifique du cluster informatique du Muséum (76 nœuds, MNHN, UMS 2700).

2007-2010 : Membre du conseil d'unité CU du laboratoire "Origine, Structure et Evolution de la biodiversité" (OSEB, MNHN, UMR 7205).

2006-2010 : participation à l'initiative "DNA Barcoding" pour développer le Barcode ADN en France.

2006 : membre du comité d'organisation du colloque international "DNA sampling : Strategy & Design" au MNHN (15-16 mars 2007).

.4 Contrats de recherche

4- Contrats de recherche

Titre, organisme, année, montant, resp.

Depuis 2006, j'ai participé à la rédaction des projets de recherche acceptés suivants :

2023-2025: **LabEx CeMEB**, co-porteur avec P.-A. Gagnaire (ISEM-CNRS) et S. Boitard (CBGP-INRAE), "**DEVHAPSEQ:** *DEVelopper l'HAPlotagging pour SEQuencer de grands échantillons de génomes phasés*". 20 k€

2022-2026: call **BioDivOc** (Region Occitanie), co-porteur avec S. Boitard (CBGP-INRAE), "**DevOcGen:** *Development and applications of new tools for the managment and conservation of natural populations from genomic data*". 430 k€

2021-2025: **ANR**, M. P. Chapuis (CBGP-CIRAD), resp. Work Package, "**Disland***: Inferring pest dispersal in agricultural landscapes to improve management strategies*". 287 k€

2020-2024: **ANR**, P.-A. Crochet (CEFE-CNRS), co-resp. d'un Work Package avec F. Rousset (ISEM-CNRS), "**INTROSPEC**: *Genomic consequences and evolutionary causes of introgression in the late stages of speciation*". 517 k€

2020-2022: **LabEx CeMEB**, co-porteur avec M.-P. Chapuis (CBGP-Cirad), "**proLag:** *Proof of concept in Landscape genomics: from NGS data production to demographic inferences*". 20 k€

2018-2021: **LabEx CeMEB**, porteuses C. Brouat (CBGP-IRD) & C. Smadja (ISEM-CNRS), co-resp. d'un Work Package, "**SPEED:** *DoeS PErsonality Explain spatial spreaD of invasive wild mice in Senegal? Behavioral ecology and population genomics approaches*". 26 k€

2017-2020: **LabEx CeMEB**, porteuses G. Ganem (ISEM-CNRS) & C. Brouat (CBGP-IRD), co-resp. d'un Work Package, "**D-RANGE:** *Environmental and evolutionary Drivers of species distributions and RANGE limits*". 26 k€

2017-2021: **ANR**, co-porteur avec S. Guindon (LIRMM-CNRS), "**GenoSpace:** *Improved statistical approaches for the analysis of biodiversity using genetic and spatial data*". 140 k€

2014: **Projet Jeunes Chercheurs de l'IBC**, co-porteur avec P. Pudlo (IMAG-UM), "*Using haplotype lengths and linkage disequilibrium to infer the demographic history of populations from genomic data.*". 10 k€

2013: **Projet Exploratoire Premier Soutien PEPS de l'Université de Montpellier**, co-porteur avec P. Pudlo (IMAG-UM), "*Understanding emerging diseases and epidemics: models, evolution, history and society*", 10 k€

2012-2017: **participant au PIA ANR "IBC"** (Projet Investissement d'Avenir "**Institut de Biologie Computationelle**") porté par O. Gascuel (LIRMM-CNRS). 2.5 M€

2010-2014: **ANR**, porteur S. Planes (CORAIL-EPHE-CNRS), resp. d'un Work Package "**IM-MODEL@CORALFISH:** *An isolation-migration model of the history of coral reef fish communities: theory and data*". 500 k€

2010-2014: **ANR**, porteuse C. Denys (OSEB-MNHN), resp. d'un Work Package, "**MOHMIE:** *Modern Human installation in Morocco Influence on the small terrestrial vertebrate biodiversity and its Evolution*". 809 k€

2009-2013: **ANR**, porteur J. -M. Cornuet (CBGP-INRA), co-resp. d'un Work Package, "**EMILE:** *Inference methods and software's for Evolution*". 372 k€

2008-2011: **CNRS Amazonia**, porteuse H. Fréville (MNHN) and C. Scotti-Saintagne (INRA Kourou), "**CLIPS:** *Past climate change and ecological specialization in tropical species: input from population genetic approaches*". 100 k€

2008-2010: **ANR**, porteur P. Grandcolas (OSEB-MNHN), resp. d'un Work Package, "**BioNeoCal:** *Endemism in New Caledonia: phylogenetic and population study of its emergence*". 940 k€

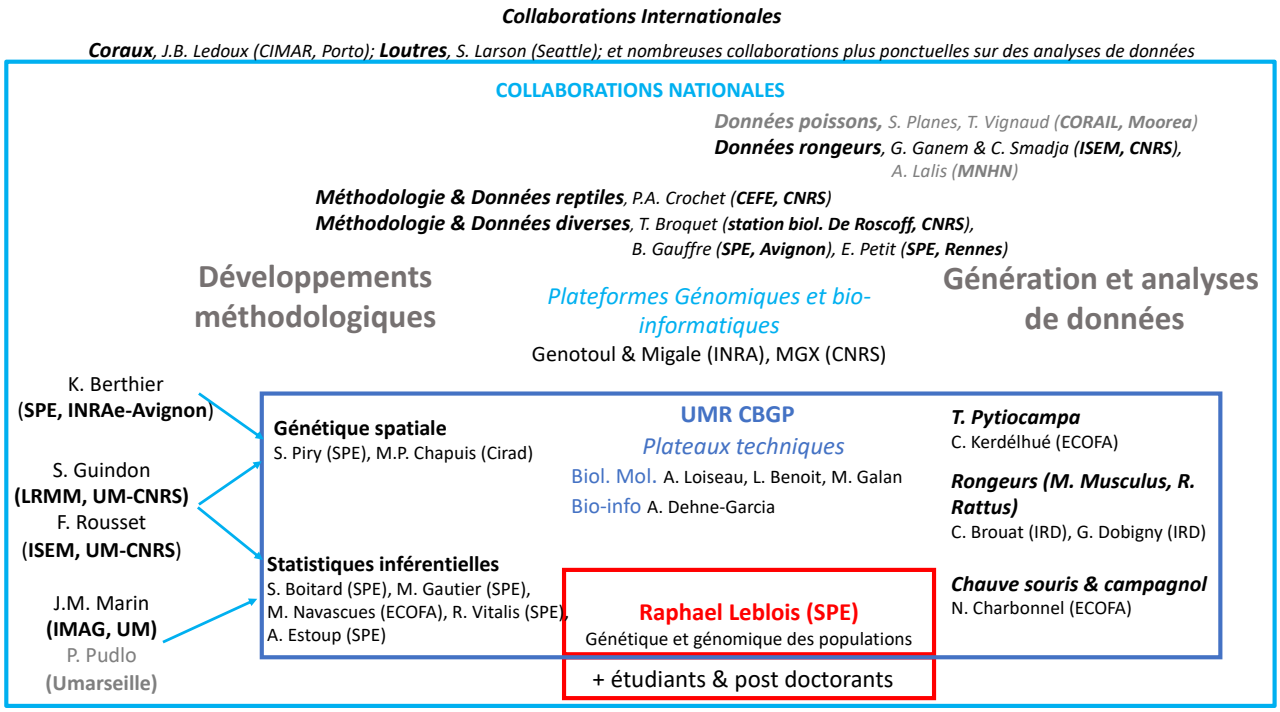
2007: **MNHN - Bonus Qualité Recherche**, porteur R. Leblois : « *Computer station for the development of inference methods from genetic data* ». 4k€

2007-2009: **MNHN - Bonus Qualité Recherche**, co-porteur avec E. Porcher (MNHN) : « *Using DNA Barcode data to measure phylogenetic biodiversity in plant communities* ». 32k€

Durant cette période, J'ai participé à la rédaction d'environ **30 autres projets** (ATIP, ANR, IFB, INRAE, MNHN, CNRS, CIRAD, UM, Région, …), dont **12 en tant que (co-)porteur** et **20 en tant que (co-) responsable d'un Work Package**, qui n'ont **pas été acceptés**. C'est un taux de réussite correct mais quel gâchis de temps tout de même…

.5 Collaborations

5- Collaborations “productives“



En gris : collaborations en sommeil ces derniers temps…

De 2006 à 2010, dans l'unité YSEB (ex-OSEB) au Museum National d'Histoire Naturelle, j'ai collaboré au sein de l'unité avec :

- **Michel Veuille** (EPHE) et **Thierry Wirth** (EPHE) notamment sur des projets ANR (IFORA, IM@Coral.Fish) et sur les techniques de DNA-barcoding (2006-2012). *Publications : Austerlitz et al. 2009, Frezal & Leblois 2008, Leblois et al. 2011, Morelli et al. 2010, Nubel et al. 2010, Lalis et al. 2012, David et al. 2012*

- **Aude Lalis** (MNHN), **Violaine Nicolas** (MNHN) et **Christiane Denys** (MNHN) sur le projet ANR MOHMIE et sur la thèse d'A. Lalis (2007). *Publications : Stoetzel et al. 2019, Lalis et al. 2016a, Lalis et al. 2016b, Lalis et al. 2012.*

Et plus largement au MNHN :

- **Marie-Catherine Boisselier** et **Sarah Samadi** sur le DNA-Barcoding, et l'analyse de données tortues et broméliacées (2006-2009). *Publications : Perez et al. 2012, Boisselier-Dubayle et al. 2010.*

Depuis 2010, au sein du CBGP, je collabore avec :

- les généticiens des populations "méthodologistes", ponctuellement sur des analyses de données ou de façon plus poussées sur certains projets de développements méthodologiques (**Simon Boitard** (SPE), **Arnaud Estoup** (SPE), **Mathieu Gautier** (SPE), **Miguel Navascués** (ECODIV), **Renaud Vitalis** (SPE)). Projets ANR EMILE, IntroSpec, IBC, projet BioDivOC DevOCGEn. *Publications: Verdu et al. 2010, Girod et al. 2011, Cornuet et al. 2014, Leblois et*

al. 2014, Burban et al. 2016, Navascués et al. 2017, Berthier et al. 2016, Leblois et al. 2017, Lippens et al. 2017, Hivert et al. 2018.

- **Sylvain Piry** (SPE) et plus récemment **Marie-Pierre Chapuis** (Cirad) sur des analyses de données en génétique des populations spatialisées sur des criquets, des rongeurs ou des chauve-souris (2010-aujourd'hui). *Publications : Berthier et al. 2016, Lippens et al. 2017, Tournayre et al. 2019, Larson et al. 2020*.

- **Carole Kerdelhué** (ECODIV) sur la génération de données de processionnaire du pin et leurs analyses (2012-aujord"hui). *Publications : Burban et al. 2016, Leblois et al. 2017, Pestopoulos et al. 2018, Burban et al. 2019.*

- **Carine Brouat** (IRD) sur l'échantillonnage, la génération de données génomiques et l'analyse de celles-ci, sur la souris domestique et de la souris rayé (2015-aujourd'hui, projets SPEED & D-Range). *Publications : Lippens et al 2017.*

- **Nathalie Charbonnel** (ECOFA) sur l'analyse de données campagnol et chauve-souris, notamment lors d'encadrement de thèse (Julie Pisano 2013-2015, et Orianne Tournayre 2016-2019). *Publications : Tournayre et al. 2019*.

Plus largement au sein de l'INRAe, je collabore avec :

- **Bertrand Gauffre** et **Karine Berthier** (SPE - Avignon) sur des analyses de données et sur des applications méthodologiques en génétiques des populations spatialisées (2015-aujourd'hui). *Publications : Gauffre et al. 2015, Berthier et al. 2016, Coleman et al. 2018, Juhel et al. 2019, Gauffre et al. 2021.*

Ma principale collaboration est bien sûr avec **François Rousset** (ISEM-CNRS, 1999-aujourd'hui) sur quasiment tous mes développements méthodologiques. J'ai plus récemment commencé à collaborer avec **Jean-Michel Marin, Pierre Pudlo**, puis **Paul Bastide** (IMAG-UM, 2011-aujourd'hui) sur ces mêmes aspects de statistique inferentielle. Projets EMILE, IM-Model@coral.fish, GenoSpace, IntroSpec, PEPS, DevOCGen… *Publications : Leblois et al. 2000, Leblois et al. 2003, Leblois et al. 2004, Delorio et al. 2005, Watts et al. 2007, Rousset & Leblois 2007, Leblois et al. 2009, Rousset & Leblois 2012, Leblois et al. 2014, Merle et al. 2017, Bonnet et al. 2017, Rousset et al. 2018, Virgoulay et al. 2021*

Ma seconde principale collaboration se fait avec **Pierre-André Crochet**, sur des échantillonnages, la génération de données et l'analyses, ainsi que sur quelques développements méthodologiques (2009-aujourd'hui). Projet ANR IntroSpec. *Publications: Ferchaud et al. 2015, Crochet et al. 2015, Bonnet et al. 2017, Miralles et al. 2020, Toyama et al. 2020.*

Je n'ai jamais eu de collaborations internationales importantes, sauf des discussions méthodologiques avec **Asger Hobolth** (Centre de Bioinformatique, Aarhus, Danemark) que j'ai invité en 2013 puis en 2022 un mois au CBGP, et que je revois régulièrement. Par contre, **je collabore régulièrement avec des laboratoires étrangers sur des analyses ponctuelles de données**. *Publications : Alberto et al. 2010, Nubel et al. 2010, Morelli et al. 2010, Leblois et al. 2011, Wereszczuk et al. 2017, Coleman et al. 2018, Macedo et al. 2019.* J'ai cependant développé deux collaborations plus poussées, surtout sur des analyses de données de coraux avec **Jean-Baptiste Ledoux** (CIMAr, Porto, 2016-aujourd'hui), et sur des données de loutres de mer avec **Shawn Larson** (Aquarium de Seattle, 2017-aujourd'hui). *Publications : Ledoux et al. 2018, 2021, Larson et al. 2020.*

.6 Publications

6- Publications

Etudiant.e.s/thésard.e.s/post-docs encadré officiellement en gras, officieusement en gris.

Charbonnel E., C. Daguin, L. Caradec, E. Moittié, O. Gilg, M. V. Gavrilo, H. Strøm, M. L. Mallory, R. I. Guy Morrison, H. Grant Gilchrist, R. Leblois, C. Roux, J. M. Yearsley, G. Yannic, T. Broquet. **2022**. Searching for genetic evidence of demographic decline in a long-lived seabird: beware of overlapping generations. *Heredity* (IF scopus 3.562)128 : 364–376. https://doi.org/10.1038/s41437-022-00515-3

Larson S., R.B. Gagne, J. Bodkin, M. J. Murray, K. Ralls, L. Bowen, R. Leblois, S. Piry, M. T. Tinker, H. Ernest. **2021**. Translocations maintain genetic diversity and increase connectivity in sea otters, Enhydra lutris. *Marine Mammal Science* (IF scopus 2.128) 37: 1475-1497. https://doi.org/10.1111/mms.12841

Gauffre B., A. Boissinot, V. Quiquempois, R. Leblois, P. Grillet, S. Morin, D. Picard, C. Ribout, O. Lourdais. 2021. Agricultural intensification alters marbled newt genetic diversity and gene flow through density and dispersal reduction. *Molecular Ecology* (IF scopus 5.549) 31: 119-133. https://doi.org/10.1111/mec.16236

Ledoux J.-B., R. Ghanem, M. Horaud, P. Lopez-Sendino, V. Romero Soriano, A. Antunes, N. Bensoussan, D. Gómez-Gras, C. Linares, A. Machordom, O. Ocaña, J. Templado, R. Leblois, J. Ben Souissi, J. Garrabou. **2021**. Gradients of genetic diversity and differentiation across the distribution range of a Mediterranean coral: patterns, processes and conservation implications. *Diversity & Distributions* (IF scopus 5.522) 27:2104–2123. https://doi.org/10.1111/ddi.13382

Virgoulay T., F. Rousset, C. Noûs, R. Leblois. **2021**. GSpace : an exact coalescence simulator of recombining genomes under isolation by distance. *Bioinformatics* (IF scopus 6.644), btab261, https://doi.org/10.1093/bioinformatics/btab261

Toyama K. S., P.-A. Crochet, R. Leblois. **2020**. Sampling schemes and drift can bias admixture proportions inferred by structure. *Molecular Ecology Resources* (IF scopus 8.638) 20: 1769-1785. https://doi.org/10.1111/1755-0998.13234

Miralles A., P. Geniez, M. Beddek, D. Mendez Aranda, J. C. Brito J. C., R. Leblois, P.-A. Crochet. **2020**. Morphology and multilocus phylogeny of the Spiny-footed Lizard (*Acanthodactylus erythrurus*) complex reveal two new mountain species from the Moroccan Atlas. *Zootaxa* (IF scopus 0.959) 4747: 302-326. http://dx.doi.org/10.11646/zootaxa.4747.2.4

Burban C., S. Rocha, R. Leblois, J.-P. Rossi, L. Sauné, M. Branco, C. Kerdelhué. **2019**. From sympatry to parapatry: a rapid change in the spatial context of incipient allochronic speciation. *Evolutionary Ecology* (IF scopus 1.819) 34: 101–121. https://doi.org/10.1007/s10682-019-10021-4

Macedo D., I. Caballero, M. Mateos, R. Leblois, S. McCay, L. A. Hurtado. **2019**. Population genetics and historical demographic inferences of the blue crab *Callinectes sapidus* in the US based on microsatellites. *PeerJ* (IF scopus 3.014) 7:e7780 https://doi.org/10.7717/peerj.7780

Tournayre O., J.-B. Pons, M. Leuchtmann, R. Leblois, S. Piry, O. Filippi-Codaccioni, A. Loiseau, J. Duhayer, I. Garin., F. Mathews, S. Puechmaille, N. Charbonnel & D. Pontier. **2019**. Integrating population genetics to define conservation units from the core to the edge of *Rhinolophus ferrumequinum* Western range. *Ecology and Evolution* (IF scopus 3.058) 9:12272–12290. http://dx.doi.org/10.1002/ece3.5714

Stoetzel E., **A. Lalis**, V. Nicolas, S. Aulagnier, T. Benazzou, Y. Dauphin, M.A. El Hajraoui, A. El Hassani, S. Fahd, M. Fekhaoui, E.M. Geigl, F.J. Lapointe, R. Leblois, A. Ohler, R. Nespoulet & C. Denys. **2019**. Quaternary terrestrial microvertebrates from Mediterranean northwestern Africa : state-of-the-art focused on recent multidisciplinary studies. *Quaternary Science Review* (IF scopus 4.163) 224:105966.https://doi.org/10.1016/j.quascirev.2019.105966

Juhel A. S., C. M. Barbu, M. Valantin-Morison, B. Gauffre, R. Leblois, J. Olivares, P. Franck. **2019**. Limited genetic structure and demographic expansion of the *Brassicogethes aeneus* populations in France and in Europe. *Pest Management science* (IF scopus 4.85) 75: 667-675. https://doi.org/10.1002/ps.5162

Ledoux J.-B., M. Frleta-Valić, S. Kipson°, A. Antunes, E. Cebrian, C. Linares, P. Sánchez, R. Leblois, J. Garrabou. **2018**. Postglacial range expansion shaped the spatial genetic structure in a marine habitat-forming species: implications for conservation plans in the Eastern Adriatic Sea. *Journal of Biogeography* (IF scopus 4.808) 45: 2645-2657. https://doi.org/10.1111/jbi.13461

Hivert V., R. Leblois, E. Petit, M. Gautier, R. Vitalis. **2018**. Measuring genetic differentiation from Pool-seq data. *Genetics* (IF scopus 3.686) 210 : 315-330. https://doi.org/10.1534/genetics.118.300900

Petsopoulos D., R. Leblois, L. Sauné, K. Ipekdal, F. A. Aravanopoulos, C. Kerdelhué, D. N. Avtzis. **2018**. Crossing the Mid-Aegean Trench: vicariant evolution of the Eastern pine processionary moth, *Thaumetopoea wilkinsoni* (Lepidoptera: Notodontidae), in Crete. *Biological Journal of the Linnean Society* (IF scopus 2.17) 124 : 228–236. https://doi.org/10.1093/biolinnean/bly041

Coleman R., B. Gauffre, A. Pavlova, L. Beheregaray, J. Kearns, J. Lyon, M. Sasaki, R. Leblois, C. Sgro, P. Sunnucks. **2018**. Artificial barriers prevent genetic recovery of small isolated populations of a low mobility freshwater fish. *Heredity* (IF scopus 3.562) 120:515–532. https://doi.org/10.1038/s41437-017-0008-3

Rousset F., **C. R. Beeravolu**, R. Leblois. **2018**. Likelihood computation and inference of demographic and mutational parameters from population genetic data under coalescent approximations. *Journal de la Société Française de Statistique* (no IF) *159 : 142-166*.

Wereszczuk A., R. Leblois, A. Zalewski. **2017**. Genetic diversity and structure related to expansion history and habitat isolation: stone marten populating rural-urban habitats. *BMC ecology* (IF scopus 3.133) 17:46. https://doi.org/10.1186/s12898-017-0156-6

Leblois R., M. Gautier, A. Rohfritsch, J. Foucaud, C. Burban, M. Galan, A. Loiseau, L. Saune, M. Branco, K. Gharbi, R. Vitalis, C. Kerdelhue. **2017**. Deciphering the evolutionary history of allochronic differentiation in the pine processionary moth Thaumetopoea pityocampa. *Molecular Ecology* (IF scopus 5.549) 27: 264-278. https://doi.org/10.1111/mec.14411

Bonnet T., R. Leblois, F. Rousset, P.-A. Crochet. **2017**. A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *Evolution* (IF scopus 3.281) 71:2140-2158. https://doi.org/10.1111/evo.13296

Navascués M., R. Leblois, C. Burgarella. **2017**. Demographic inference through approximate-Bayesian-computation skyline plots. *PeerJ* (IF scopus 3.014) 5:e3530. https://doi.org/10.7717/peerj.3530

Lippens C., A. Estoup, K. Hima, A. Loiseau, C. Tatard, A. Dalecky, K. Bâ, M. Kane, M. Diallo, A. Sow, S. Piry, R. Leblois, J.-M. Duplantier, C. Brouat. **2017**. Genetic structure and invasion history of the house mouse (Mus musculus domesticus) in Senegal : a legacy of colonial times? *Heredity* (IF scopus 3.562) 119 :64-75. https://doi.org/10.1038/hdy.2017.18

Merle C., R. Leblois, F. Rousset, P. Pudlo. **2017**. Resampling: an improvement of Importance Sampling in varying population size models. *Theoretical Population Biology* (IF scopus 1.106) 114:70-87. https://doi.org/10.1016/j.tpb.2016.09.002

Burban C., M. Gautier, R. Leblois, **J. Landes**, H. Santos, M.-R. Paiva, M. Branco, C. Kerdelhué. **2016**. Evidence for low-level hybridization between two allochronic populations of the pine processionary moth, Thaumetopoea pityocampa (Lepidoptera: Notodontidae). *Biological Journal of the Linean Society* (IF scopus 2.17) 119:311–328. https://doi.org/10.1111/bij.12829

Berthier K., M. Garba, R. Leblois, M. Navascues, C. Tatard, P. Gauthier, S Gagaré, S. Piry, A. Dalecky, A. Loiseau , G. Dobigny. **2016**. Black rat invasion of inland Sahel: insights from interviews and population genetics in South Western Niger. *Biological Journal of*

the Linean Society (IF scopus 2.17) 119:748–765. https://doi.org/10.1111/bij.12836

Lalis A., R. Leblois, E. Stoetzel, T. Benazzou, K. Souttou, C. Denys, V. Nicolas. **2016**. Phylogeography and demographic history of Shaw's Jird (Meriones shawii complex) in North Africa. *Biological Journal of the Linean Society* (IF scopus 2.17) 118:262–279. https://doi.org/10.1111/bij.12725

Zenboudji S., M. Cheylan, V. Arnal, A. Bertolero, R. Leblois, G. Astruc, G. Bertorelle, J. L. Pretus, M. Lo Valvo, G. Sotgiu, C. Montgelard. **2016**. High genetic structure and contrasting demographic history in the endangered Mediterranean tortoise *Testudo hermanni hermanni. Biological Conservation* (IF scopus 7.137) 195:279–291. https://doi.org/10.1016/j.biocon.2016.01.007

Lalis A., R. Leblois, S. Liefried, A. Ouarour, **C. Reddy Beeravolu**, J. Michaux, A. Hamani, C. Denys, V. Nicolas. **2016**. New molecular data favor an anthropogenic introduction of the wood mouse (Apodemus sylvaticus) in North Africa. *Journal of Zoological Systematics and Evolutionary Research* (IF scopus 2.564) 54:1-12. https://doi.org/10.1111/jzs.12111

Crochet P.-A., R. Leblois, J. P. Renoult. **2015**. New reptile records from Morocco and Western Sahara. *Herpetology Notes* (IF scopus 5.549) 8:583-588.

Laporte M., R. Leblois, A. Coulon, F. Bonhomme, P. Magnan, P. Berrebi. **2015**. Genetic structure of a vulnerable species, the freshwater blenny (*Salaria fluviatilis*). *Conservation Genetics* (IF scopus 2.958) *Déc.* 1-11. https://doi.org/10.1007/s10592-014-0682-0

Gauffre B., S. Mallez, M.-P. Chapuis, R. Leblois, I. Litrico, S. Delaunay, I. Badenhausser. **2015**. Spatial heterogeneity in landscape structure influences dispersal and genetic structure: empirical evidence from a grasshopper in an agricultural landscape. *Molecular Ecology* (IF scopus 5.549) 24:1713-1728. https://doi.org/10.1111/mec.13152

Ferchaud A.-L., R. Eudeline, V. Arnal, M. Cheylan, G. Pottier, R. Leblois and P.-A. Crochet. **2015**. Congruent signals of population history but radically different patterns of genetic diversity between mitochondrial and nuclear markers in a mountain lizard. *Molecular Ecology* (IF scopus 5.549) 24:192–207 . https://doi.org/10.1111/mec.13011

Vignaud T. M., J. Mourier, J. A. Maynard, R. Leblois, J. Spaet, E. Clua, V. Neglia, S. Planes. **2014**. Blacktip reef sharks, *Carcharhinus melanopterus*, have high genetic structure and varying demographic histories in their Indo-Pacific range. *Molecular Ecology* (IF scopus 5.549) 23:5193-5207. https://doi.org/10.1111/mec.12936

Vignaud T.M., J.A. Maynard, R. Leblois, M.G. Meekan, R. Vazquez-Juarez, D. Ramirez-Macias, S.J. Pierce, D. Rowat, M.L. Berumen, C. Beeravolu, S. Baksay, S. Planes. **2014**. Genetic structure of populations of whale sharks among ocean basins and evidence for their historic rise and recent decline. *Molecular Ecology* (IF scopus 5.549) 23:2590-2601. https://doi.org/10.1111/mec.12754

Leblois R., P. Pudlo, **J. Néron**, **F. Bertaux**, **C. Reddy Beeravolu**, R. Vitalis, F. Rousset. **2014**. Maximum likelihood inference of population size contractions from microsatellite data. *Molecular Biology and Evolution* (IF scopus 6.493) 31:2805-2823. https://doi.org/10.1093/molbev/msu212

Cornuet J.-M., P. Pudlo, J. Veyssier, A. Dehne-Garcia, M. Gautier, R. Leblois, J.-M. Marin, A. Estoup. **2014**. DIYABC v2.0: a software to make Approximate Bayesian Computation inferences about population history using Single Nucleotide Polymorphism, DNA sequence and microsatellite data. *Bioinformatics* (IF scopus 6.644) 30:1187-1189. https://doi.org/10.1093/bioinformatics/btt763

Lalis A.*, R. Leblois*, E. Lecompte*, C. Denys, J. Ter Meulen, T. Wirth. **2012**. The Impact of Human Conflict on the Genetics of *Mastomys natalensis* and Lassa Virus in West Africa. *PLoS One* (IF scopus 3.582), 7:e37068.https://doi.org/10.1371/journal.pone.0037068 * co-first authors.

Rousset F., R. Leblois. **2012**. Likelihood-based inferences under a coalescent model of isolation by distance: two-dimensional habitats and confidence intervals. *Molecular*

Biology and Evolution (IF scopus 6.493) 29:957-973. https://doi.org/10.1093/molbev/msr262

Perez M., R. Leblois, B. Livoreil, R. Bour, J. Lambourdiere, S. Samadi, M.C. Boisselier. **2012**. Effects of landscape features and demographic history on the genetic structure of Testudo marginata populations in the southern Peloponnese and Sardinia. *Biological Journal of the Linnean Society* (IF scopus 2.17) 105:591-606. https://doi.org/10.1111/j.1095-8312.2011.01805.x

David O., C. Laredo, R. Leblois, B. Shaeffer, N. Vergne. **2012**. Coalescent-based DNA barcoding: multilocus analysis and robustness. *Journal of Computational Biology* (IF scopus 1.535) 19:271-278. (Author list is alphabetical). https://doi.org/10.1089/cmb.2011.0122

Sagnard F., M. Deu, D. Dembélé, R. Leblois, L. Touré, M. Diakité, C. Calatayud, M. Vaksmann, S. Bouchet, Y. Mallé, S. Togola, P. Traoré. **2011**. Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild-weedy-crop complex in a western African region. *Theoretical and Applied Genetics* (IF scopus 5.457) 123:1231-1246. https://doi.org/10.1007/s00122-011-1662-0

Leblois R*, K. Kuhls*, O. Francois, G. Schönian, T. Wirth. **2011**. Guns, germs and dogs: On the origin of *Leishmania chagasi*. *Infections, Genetics and Evolution* (IF scopus 4.393) 188:165-179. https://doi.org/10.1016/j.meegid.2011.04.004 * co-first authors.

Girod C., R. Vitalis, R. Leblois, H. Fréville. **2011**. Inferring population decline and expansion from microsatellite data: a simulation-based evaluation of the MSVAR method. *Genetics* (IF scopus 3.686) 188:165-179. https://doi.org/10.1534/genetics.110.121764

Morelli G., Y. J. Song, C. J. Mazzoni, M. Eppinger, P. Roumagnac, D. M. Wagner, M. Feldkamp, B. Kusecek, A. J. Vogler, Y. J. Li, Y. J. Cui, N. R. Thomson, T. Jombart, R. Leblois, P. Lichtner, L. Rahalison, J. M. Petersen, F. Balloux, P. Keim, T. Wirth, J. Ravel, R. F. Yang, E. Carniel, M. Achtman. **2010**. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics* (IF scopus 23.977) 42:1140. https://doi.org/10.1038/ng.705

Verdu P., R. Leblois, A. Froment, S. Thery, S. Bahuchet, F. Rousset, E. Heyer, R. Vitalis. **2010**. Limited dispersal in mobile hunter-gatherer Baka Pygmies. *Biology Letters* (IF scopus 3.553) 6:858-861. https://doi.org/10.1098/rsbl.2010.0192

Nubel U., J. Dordel, K. Kurt, B. Strommenger, H. Westh, S. K. Shukla, H. Zemlickova, R. Leblois, T. Wirth, T. Jombart, F. Balloux, W. Witte. **2010**. A Timescale for Evolution, Population Expansion, and Spatial Spread of an Emerging Clone of Methicillin-Resistant *Staphylococcus aureus*. *Plos Pathogens* (IF scopus 7.01) 6:e1000855. https://doi.org/10.1371/journal.ppat.1000855

Alberto F., P. T. Raimondi, D. C. Reed, N. C. Coelho, R. Leblois, A. Whitmer, E. A. Serrão. **2010**. Habitat continuity and geographic distance predict population genetic differentiation in giant kelp. *Ecology* (IF scopus 5.009) 91:49-56. https://doi.org/10.1890/09-0050.1

Boisselier-Dubayle M. -C., R. Leblois,S. Samadi, J. Lambourdière, C. Sarthou. **2010**. Genetic structure of a xerophilious bromeliad in a fragmented habitat and the forest refuge hypothesis: *Pitcairnia geyskeysii* on inselbergs in French Guyana. *Ecography* (IF scopus 7.113) 33:175-184. https://doi.org/10.1111/j.1600-0587.2009.05446.x

Guillot G., R. Leblois, A. Coulon, A. C. Frantz. **2009**. Statistical methods in spatial genetics. *Molecular Ecology*(IF scopus 5.549) 18:4734-4756. https://doi.org/10.1111/j.1365-294X.2009.04410.x

Austerlitz F., O. David, B. Schaeffer, K. Bleakley, M. Olteanu, R. Leblois, M. Veuille, C. Laredo. **2009.** Comparing phylogenetic and statistical classification methods for DNA barcoding. *BMC Bioinformatics* (IF scopus 3.385) 10: *Suppl. 14: S10*. https://doi.org/10.1186/1471-2105-10-s14-s10

Leblois, R., A. Estoup, F. Rousset. **2009**. IBDSim: a computer package for coalescent simulations under isolation by distance with temporal and spatial heterogeneities. *Molecular Ecology Resources* (IF scopus 8.638) 9:107–109. https://doi.org/10.1111/j.1755-0998.2008.02417.x

Frezal L., R. Leblois. **2008**. 4 years of DNA barcoding: current advances and prospects. *Infection, Genetics & Evolution* (IF scopus 4.393) 8:727–736. http://doi.org/10.1016/j.meegid.2008.05.005

Rousset, F., R. Leblois. **2007.** Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Molecular Biology and Evolution* (IF scopus 6.493) 24:2730–2745. http://doi.org/10.1093/molbev/msm206

Leblois, R.**,** M. Slatkin. **2007**. Estimating the number of founder lineages from haplotypes of closely linked SNPs. *Molecular Ecology* (IF scopus 5.549) 16:2237-2245. https://doi.org/10.1111/j.1365-294X.2007.03288.x

Watts P.C., F. Rousset, I.J. Saccheri, R. Leblois, S.J. Kemp, D.J. Thompson**. 2007.** Compatibility of genetic and demographic estimates of 'neighborhood size' in insect populations: analysis of *Coenagrion mercuriale* (Odonata: Zygoptera) using an improved estimator of genetic divergence. *Molecular Ecology* (IF scopus 5.549) 16:737–751. http://doi.org/10.1111/j.1365-294X.2006.03184.x

Leblois, R., Estoup, A., Streiff, R. **2006**. Habitat contraction and reduction in population size: Does isolation by distance matter? *Molecular Ecology* (IF scopus 5.549) 15:3601–3615. https://doi.org/10.1111/j.1365-294X.2006.03046.x

De Iorio M., Griffiths R., Leblois R., Rousset F. **2005**. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoretical Population Biology* (IF scopus 1.106) 68:41-53. (Author list is alphabetical). http://doi.org/10.1016/j.tpb.2005.02.001

Leblois R., Rousset F., Estoup A. **2004**. Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population from microsatellite data. *Genetics* (IF scopus 3.686) 166:1081-1092. https://doi.org/10.1534/genetics.166.2.1081

Brouat C., Sennedot F., Audiot P., Leblois R., Rasplus J. -Y. **2003**. Fine-scale genetic structure of two carabid species with contrasted levels of habitat specialization. *Molecular Ecology* (IF scopus 5.549) 12:1731 - 1745. https://doi.org/10.1046/j.1365-294X.2003.01861.x

Leblois R., Estoup A. **2003**. Invited commentary on the article by Wilson IJ, Weale ME, Balding DJ (2003): Inferences from DNA data: population histories, evolutionary processes, and forensic match probabilities, *Journal of the Royal Statistical Society* (IF scopus 4.829) *A* 166:1-33.

Leblois R., Estoup A., Rousset F. **2003**. Influence of mutational and sampling factors on the estimation of demographic parameters in a continuous population under isolation by distance. *Molecular Biology and Evolution* (IF scopus 6.493) 20: 491-502. https://dx.doi.org/10.1534/genetics.166.2.1081

Tikel D., Peatkau D., Cortinas N.**,** Leblois R., Moritz C., Estoup A. **2000**. Microsatellite loci in the invasive toad species *Bufo marinus*. *Molecular Ecology* (IF scopus 5.549) 9:1927-1929. https://doi.org/10.1046/j.1365-294x.2000.01074-6.x

Leblois R., Rousset F., Tikel D., Moritz C., Estoup A. **2000**. Absence of evidence for isolation by distance in an expanding cane toad (*Bufo marinus*) population: an individual-based analysis of microsatellite genotypes. *Molecular Ecology* (IF scopus 5.549) 9:1905-1909. https://doi.org/10.1046/j.1365-294x.2000.01091.x

Leblois, R. 2004. Estimation de paramètres de dispersion en populations structurées à partir de données génétiques. Thèse de doctorat de l'Agro-M (ENSA-Montpellier), 233 pp.

Leblois, R. 2000. Etude par simulation de l'influence de facteurs mutationnels et démographiques sur l'estimation de paramètres démographiques à partir de génotypes individuels. Diplôme d'Etudes Approfondies (DEA Biologie de l'Evolution et Ecologie, Agro-M et Université Montpellier II). 30pp.

.7 Logiciels

7- Logiciels mis à disposition de communautés scientifiques

- **gspace2infr**, package R pour faire de l'inférence par simulation sous isolement par la distance. License CeCILL-2.

- **GSpace**, Simulateur de données génomiques avec recombinaison sous des modèles spatialisés avec hétérogénéités spatiales et temporelles, License CeCILL.

- **IBDSim**, Simulateur de données génétiques sous des modèles spatialisés avec hétérogénéités spatiales et temporelles, License CeCILL.

- **Migraine**, Inférences démographiques et historiques par maximum de vraisemblance à partir de données génétiques, License CECILL.

.8 Encadrement-Enseignement

8- Encadrement

Noms, année, titre, publis, devenir

2023 : **une étudiante en Master2**

- Sudeshna Chakraborty (M2, Université de Montpellier, Erasmus Mundus Master Programme in Evolutionary Biology, MEME) " *Invasion history of house mice in Senegal* " (co-sup. C. Brouat, E. Ortega). Finit son M2, puis veut faire une thèse.

2023 : **un étudiant en Master2** (ERJ CeMeB : équipe de recherche junior, co-encadré par Jules Romieu**)**

- Ghislain Camarata (M2 Biodiversité Écologie Évolution, Parcours DARWIN : Biologie Évolutive & Écologie, Université de Montpellier) "*Test par simulation des méthodes existantes pour inférer l'introgression adaptative*" (co-sup. J. Romieu, and F. Rousset, M. Navascues). Veut continuer en thèse.

2022-2024 : **un étudiant en thèse**

- Jules Romieu (Université de Montpellier) "*Introgression adaptative : Comment identifier la part de l'introgression due à la sélection ?*" (co-sup. F. Rousset, M. Navascues, P.-A. Crochet). Continue sa thèse.

2021 : **deux étudiant.e.s en Master1**

- Matis Bagarre (M1 Sciences et Numerique pour la Sante Parcours Bioinformatique, Connaissances et Donnes, Université de Montpellier) "*Intégration de changemnets temporels dans le simulation de données génomiques sptaialisées GSpace*". En M2.
- Marine Vue (M1 biologie informatique, ingénierie de plateforme en biologie, Université Paris-Sud) "*Comparaison de l'inférence de la dispersion sous différents modèles de génétique des populations et de phylogéographie*" (co-sup. S. Guindon). En M2.

2020 : **deux étudiant.e.s en Master2**

- Fanny Touchard (M2 Biodiversité Écologie Évolution, Parcours DARWIN : Biologie Évolutive & Écologie, Université de Montpellier) "*Impact de l'histoire des populations sur la distribution spécifique : le cas des souris striées (Rhabdomys spp.) en Afrique du Sud*" (co-sup. G. Ganem, C. Brouat, A.-S. Fiston-Lavier), financement LabEx CeMEB. En Thèse à l'ISEM.
- Paul Doniol-Valcroze (M1 Biodiversité Écologie Évolution, Parcours DARWIN : Biologie Évolutive & Écologie, Université de Montpellier) "*Caractérisation d'une zone de contact à l'aide de marqueurs génomiques : l'exemple d'Acanthodactylus erythrurus (Lacertidae).*" (co-sup. p.-A. Crochet, L. Rancilhac, C. Dufresne), financement CNRS. En thèse au CEFE.

2019-2023 : **un étudiant en thèse**

- Thimothée Virgoulay (Université de Montpellier) "*Inférences démographiques et historiques à partir de données génomiques sous des modèles spatialisés*" (co-sup. F. Rousset), bourse ED-GAIA-Univ-Montpellier. En recherche de post-doc.

2018 : **deux étudiant.e.s en Master2**

- Thimothée Virgoulay (M2 Sciences et Numerique pour la Sante Parcours Bioinformatique, Connaissances et Donnes, Université de Montpellier) "*Inférences démographiques et historiques à partir de données génomiques sous des modèles spatialisés réalistes : vers une prise en compte du paysage*" (co-sup. F. Rousset), financement ANR.

- Camille Vernier (M2 Biostatistique, Université de Montpellier) "*Estimation de la dispersion en population continue à partir de données génomiques : que permettent les nouvelles méthodes d'inférences basées sur la simulation ?*" (co-sup. J.-M. Marin, F. Rousset), financement IBC/ANR. Vient de finir sa thèse au CBGP.

2017 : **un étudiant en Master2**
- Loïs Rancilhac (M2 Biodiversité Écologie Évolution, Parcours DARWIN : Biologie Évolutive & Écologie, Université de Montpellier) "*Histoire évolutive récente du complexe d'Acanthodactylus erythrurus*" (co-sup. P.-A. Crochet), financement CNRS. En post-doc.

2016 : **un étudiant en Master2**
- Ken Toyama (M2, Erasmus Mundus Master Programme in Evolutionary Biology, MEME) "*Effects of sampling schemes and genetic drift on the admixture proportions calculated by the software STRUCTURE*" (co-sup. P.-A. Crochet), financement IBC. En these.

2015: **un étudiant en Master2**
- Valentin Hivert (M2 Modélisation en Ecologie, Université de Rennes) "*Test par simulation de l'utilisation des RADseq pour l'estimation de la dispersion* " (co-sup. M. Gautier, R. Petit), financement IBC. En post-doc.

2013-2016 : **une étudiante en thèse**
- Coralie Merle (Université de Montpellier) "*Nouvelles méthodes d'inférence de l'histoire démographique à partir de données génétiques*" (co-sup. P. Pudlo, J.-M. Marin, F. Rousset), bourse mixte LabEx CEMEB & LabEx NUMEV. Professeur de Mathématiques en Lycée.

2013 : **deux étudiantes en Master2**
- Coralie Merle (M2 Magistère de Mathématiques, Université de Paris-Sud) "*Accélération des méthodes d'échantillonnage préférentiel pour le calcul de vraisemblances en génétique des populations*" (co-sup. P. Pudlo), financement IBC.
- Marine Ranger (M2 Biologie Géosciences Agroressources Environnement, Parcours Biodiversité Écologie Évolution,, Université de Montpellier, Supagro) "*Phylogeographie d'un lézard Marocain: Acanthodactylus erythrurus*" (co-sup. P.-A. Crochet), financement INRA. Devenir inconnu.

2012 : **trois étudiant.e.s en Master2**
- Timothée Bonnet (M2 Biologie Géosciences Agroressources Environnement, Parcours Biodiversité Écologie Évolution, Université de Montpellier, Supagro) "*Modelisation de l'introgression Mitochondriale et nucléaire dans une zoone de contact secondaire*" (co-sup. P.-A. Crochet, F. Rousset), financement INRAE. Chercheur CNRS.
- Julie Landes (M2 Biologie Géosciences Agroressources Environnement, Parcours Biodiversité Écologie Évolution, Université de Montpellier) "*Différentiation allochronique chez la processionnaire du pin : apport des marqueurs neutres pour bâtir des scénarios évolutifs* " (co-sup. C. Kerdelhué), financement INRAE. En post-doc.
- Sébastien Ravel (M1 Génétique et Physiologie, Spécialité : Analyse et Modélisation des Données, Université de Clermont-Ferrand) " *Développement d'une interface graphique en Python et PyQt4 pour le logiciel IBDSim*" (co-sup. A. Dehne-Garcia), financement IBC. Bio-informaticien au CIRAD.

2011-2014 : **deux post-doc**

- Champak Reddy Beravolu "*Inference par maximum de vraisemblance à partir de données de séquences sous des modèles d'isolement avec migration IM* ", financement mixte ANR – INRA SPE. Ingénieur de recherche en Suisse.
- Aude Lalis "*MOHMIE: Modern Human installation in Morocco Influence on the small terrestrial vertebrate biodiversity and its Evolution*" (co-sup. C. Denys), financement ANR. Maitre de conference au MNHN.

2010: **une étudiante en Master2**

- Natacha Luximon (M2 Sciences en comportement, évolution et conservation, Université de Lausanne, UNIL) "*Inférence de la dispersion sous isolement par la distance : comparaison des approches individus centrées vs. En dèmes.* " (co-sup. E. Petit, T. Broquet), financement INRAE. Devenir inconnu.

2009: **deux étudiant.e.s en Master1 et 2**

- François Bertaux (equiv. M1, Ecole Polytechnique Paris) "*Inférence de changements passés de tailles de populations à partir de données génétiques* ", financement ANR. Ingénieur R&D.
- Stéphanie Barthe (M2 Biologie Géosciences Agroressources Environnement, Parcours Biodiversité Écologie Évolution, Université de Montpellier, Supagro) "*Etude de l'impact de la diversité des séquences encadrant les microsatellites sur la répartition de la diversité génétique entre allèles et la divergence entre populations*" (co-sup. I. Scotti, C. Scotti-Saintagne), financement INRAE. Devenir inconnu.

2008: **un étudiant de 2eme année polytechnique**

- Joseph Néron (equiv. M1, Ecole Polytechnique Paris) "*Inférence de changements passés de tailles de populations à partir de données génétiques*", financement MNHN. Ingénieur à la SNCF.

2007-2008: **une ATER**

- Lise Frezal (EPHE) "*Barcode ADN multigene chez la drosphile*" (co-sup. M. Veuille), financement EPHE. Ingénieure de recherche à l'Institut Pasteur (?).

2007: **deux étudiantes en Master2**

- Sandrine Bérot (M2 SDUEE, spécialité EBE, MNHN): "*Inférence des temps de divergences et des taux de migratioon par échantilloannge pondéré* " (co-sup. R. Vitalis), financement MNHN. Animatrice-formatrice chez Les Petits Débrouillards.
- Camille Madec (M2 SDUEE, spécialité EBE, MNHN): "*Inférence des taux de dispersion sexe-spécifique par ABC* " (co-sup. R. Vitalis), financement MNHN. Devenir inconnu.

Depuis 2006, j'ai participé à une trentaine de **comités de thèse** :
Anne-Laure Ferchaud (EPHE, Montpellier 2008, 2010), Christophe Girod (MNHN Paris, 2008, 2009), Camille Roux (Université Lille I, 2008), Romain Nattier (MNHN Paris, 2008, 2009), Erhan Yalcindag (Université Montpellier II, 2009, 2010), Stéfanie Wagner (INRA Bordeaux – Munich University, 2011), Axelle Bouiges (EPHE-MNHN, 2011, 2012), Ivan Paz (Univ. Paul Sabatier, 2012), Nadine Ali (CBGP-Supagro, 2013), Odrade Nougue (CEFE, 2013, 2014), Julie Pisano (CBGP-FNRS, 2014, 2015), Saliha Zenboudji (EPHE-CEFE, 2015), Timothée Bonnet (2013), Fernando Seixas (CIBIO-ISEM, 2014, 2015), Orianne Tournayre (CBGP-LBBE, 2017), Pierre Lesturgie (MNHN-OSEB, 2021, 2022), Emeline Charbonel (CBGP, 2021, 2022), Cécile Caumette (CBGP, 2022).

J'ai été **rapporteur** à l'automne 2017 de la thèse de Maria Simonsen Speed, sur le sujet " Population Genetic Models for Allele Fraction Data". Cette thèse a été soutenue Arhus, Danemark, le 30 novembre 2018.
J'ai participé, en tant qu'examinateur, aux **jurys de thèse** de (1) Marguerite Lapierre, sur le sujet " Extensions du modèle standard neutre pertinentes pour l'analyse de la diversité génétique", soutenue le 25 septembre 2017 à l'Université Pierre et Marie Curie ; de (2) Louis Raynal, sur le sujet " Inférence statistique bayésienne pour les modélisations donnant lieu à un calcul de vraisemblance impossible", soutenue le 9 septembre 2019 à l'Université de Montpellier ; et de (3) Gonche Danesh, sur le sujet « Phylodynamique des virus évoluant rapidement : approches par calcul bayésien approché et par vraisemblance » , soutenue le 06 juillet 2021, à l'Université de Montpellier.

9- Enseignement

Depuis 2006, je co-organise **1 à 4 modules d'enseignements** et donne environ **20 à 50 heures de cours** selon les années, **niveau Master1, Master2 et modules de thèses** sur les thèmes : Génétique des populations, estimation de paramètres démographiques, théorie de la coalescence, Barcode ADN.

Je suis régulièrement **co-organisateur** avec R. Vitalis (CBGP) **de 1 à 3 modules de 5 jours** "Analyse de données en génétique des populations" ayant (eu) lieu à **l'Ecole Doctorale du MNHN,** puis à **l'ED SIBAGHE de Montpellier,** au **Master 2 Biologie de l'évolution et écologie parcours DARWIN de l'UM,** en 2de année de **Supagro** ou encore au **Master 1 européen MEME de l'UM**. Je co-organisateur aussi avec F. Rousset (ISEM, UM) **un module de 5 jours** « Modèles et inférences en génétique des populations » dispensé dans le cadre du **Master 2 Biostatistique de l'UM**.

Depuis 2006: environ **30 heures de cours par an**, niveau **Master1, Master2 et modules de thèses** sur les thèmes : Génétique des populations, estimation de paramètres démographiques, théorie de la coalescence, Barcode ADN.

Depuis 2007: Coorganisateur, avec R. Vitalis, de **1-3 modules de 5 jours "Analyse de données en génétique des populations"** de l'Ecole Doctorale du MNHN, puis de l'ED SIBAGHE de Montpellier, master 2 Biologie de l'évolution et écologie, 2de année de Supagro et au master 1 européen MEME.

Depuis 2013: Coorganisateur, avec F. Rousset d'un **module de 5 jours « Modèles et inférences en génétique des populations »** dispensé dans le cadre du **Master 2 Biostatistique de l'université de Montpellier**.

.9 7 publications significatives

Influence of Mutational and Sampling Factors on the Estimation of Demographic Parameters in a "Continuous" Population Under Isolation by Distance

Raphaël Leblois,† Arnaud Estoup,* and François Rousset†*

*Laboratoire Modélisation et Biologie Evolutive, CBGP-INRA, Montferrier sur Lez, France; and †Laboratoire Génétique et Environnement, CNRS-UMR 5554, Montpellier, France

In numerous species, individual dispersal is restricted in space so that "continuous" populations evolve under isolation by distance. A method based on individual genotypes assuming a lattice population model was recently developed to estimate the product $D\sigma^2$, where D is the population density and σ^2 is the average squared parent-offspring distance. We evaluated the influence on this method of both mutation rate and mutation model, with a particular reference to microsatellite markers, as well as that of the spatial scale of sampling. Moreover, we developed and tested a nonparametric bootstrap procedure allowing the construction of confidence intervals for the estimation of $D\sigma^2$. These two objectives prompted us to develop a computer simulation algorithm based on the coalescent theory giving individual genotypes for a continuous population under isolation by distance. Our results show that the characteristics of mutational processes at microsatellite loci, namely the allele size homoplasy generated by stepwise mutations, constraints on allele size, and change of slippage rate with repeat number, have little influence on the estimation of $D\sigma^2$. In contrast, a high genetic diversity (≈0.7–0.8), as is commonly observed for microsatellite markers, substantially increases the precision of the estimation. However, very high levels of genetic diversity (>0.85) were found to bias the estimation. We also show that statistics taking into account allele size differences give unreliable estimations (i.e., high variance of $D\sigma^2$ estimation) even under a strict stepwise mutation model. Finally, although we show that this method is reasonably robust with respect to the sampling scale, sampling individuals at a local geographical scale gives more precise estimations of $D\sigma^2$.

Introduction

Dispersal rates and population sizes or densities are important demographic parameters in evolutionary processes. Many studies have attempted to estimate such parameters using either direct methods (e.g., mark-recapture methods) or indirect methods (e.g., genetic markers). A number of indirect methods for demographic parameter estimation using genetic data at neutral loci or clines of selected markers have been defined (see Slatkin (1994) and Rousset (2001*b*) for reviews). Discrepancies between estimations made with direct and indirect methods have often been attributed to inadequacies of the assumptions of the genetic models made in indirect methods (Hastings and Harrison 1994; Koenig et al. 1996; Slatkin 1994). The kinds of assumptions usually considered to be inadequate are those related to (1) the modalities of dispersal (e.g., the island model), (2) the demographic stability in space and time, (3) the mutation rates and mutation processes of genetic markers, and (4) the selective neutrality of genetic markers.

In numerous species, individual dispersal is restricted in space. This means that there is a higher probability that individuals mate with individuals born in close proximity to themselves than to individuals born far away. Several studies on animals or plants have shown such restricted dispersal (e.g., for plant data, see Crawford 1984; and for animal data, Rousset 1997, 2000; Spong and Creel 2001; Sumner et al. 2001). Isolation by distance models taking into account this biological feature were introduced by Wright (1943 and 1946). Under these models the genetic differentiation at neutral loci is expected to increase with geographical distance (e.g., Malécot 1950, 1967; Sawyer 1977). Empirical data indicate that such a relationship holds for many species (Endler 1977; Slatkin 1993). Recently, a method of analysis was developed based on the increase, at a local scale, of genetic differentiation between individuals with geographical distance in a "continuous" population evolving under isolation by distance (Rousset 2000). The method makes use of the regression of estimators of a parameter analogous to the parameter $F_{ST}/(1 - F_{ST})$, calculated between individuals, and the logarithm of the geographical distance, to estimate the product $D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance. It is expected to perform better than previous methods for several reasons. First, the demographic model on which the method is based makes weak assumptions about the shape of the distribution of dispersal distances. In particular, the method is valid for leptokurtic distributions of dispersal distance (Rousset 2000), a feature commonly observed in natural populations (for review and data, see Endler 1977; Portnoy and Willson 1993; Clark et al. 1999). Second, analysis of genetic differentiation is made at a small (local) geographical scale so that heterogeneity of demographic parameters such as dispersal or density is reduced and hence its influence on genetic differentiation is also reduced (Slatkin 1993; Rousset 2001*b*). In a similar way, influence of non-neutrality of the genetic markers may be less problematic for studies at local scale because selection parameters may be less heterogeneous at a small geographical scale. On the other hand, the theory on which the method is based shows that only estimations from analysis over short distances will be accurate (Rousset 1997). These expectations have been confirmed by several comparisons of direct and indirect estimates of $D\sigma^2$ (Rousset 1997, 2000; Sumner et al. 2001). Although the geographical scale at which the sampling has been done is

Key words: coalescence, dispersal, isolation by distance, microsatellite DNA, nonparametric ABC bootstrap.

E-mail: leblois@isem.univ-montp2.fr.

Mol. Biol. Evol. 20(4):491–502. 2003
DOI: 10.1093/molbev/msg034

expected to influence the quality of the estimation of $D\sigma^2$, very few analytical or simulation studies have formally addressed this question.

Since their discovery in the 1980s, microsatellite loci have been increasingly used as genetic markers. Rapid progress in molecular biology technologies, especially the development of the polymerase chain reaction, and attractive evolutionary features (e.g., high level of polymorphism), explain why this category of markers are progressively replacing, or at least complementing, classical markers such as allozymes for numerous applications in molecular systematics, population genetics, and ecology (reviewed in Estoup and Angers 1998; Estoup, Jarne, and Cornuet. 2002). However, the mutation processes (i.e., the nature of mutations) at microsatellite loci are complex and not yet well understood (e.g., Estoup and Cornuet 1999). The effect of the mutation processes on evolutionary inferences depends in large part on the method, the statistics, and the evolutionary time scale considered (e.g., Estoup, Jarne, and Cornuet 2002). Some authors have discussed the effect of the nature of the mutation on F_{ST} values (Slatkin 1995; Rousset 1996). Because a stepwise mutation process occurs at microsatellite loci, several statistics taking into account the allele size have been proposed (Goldstein et al. 1995; Slatkin 1995; Michalakis and Excoffier 1996). Their utility, however, has often been criticized (e.g., Takezaki and Nei 1996; Gaggiotti et al. 1999). Overall, the potential interest of the different statistics has never been addressed in the context of the estimation of demographic parameters under isolation by distance.

In this study, we developed an original simulation algorithm based on the coalescent theory in order to study the sensitivity of the estimation of $D\sigma^2$ to different factors: (1) the sampling scale of individuals, (2) the mutation model of markers and (3) their mutation rate, with particular reference to microsatellite markers for the two latest points. This algorithm was also used to test a nonparametric ABC bootstrap procedure allowing the construction of confidence intervals on the $D\sigma^2$ estimation. Finally, we draw guidelines that could be useful for empirical investigators using the individual-based method of Rousset (2000).

Models and Methods

Demographic Model and Population Cycle

The model that we considered for "continuous" populations is the lattice model with each lattice node corresponding to one diploid individual. This model without demic structure is viewed as an approximation for truly continuous populations with infinite local competition (Malécot 1975; Rousset 2000). More realistic continuous models would incorporate the feature that individuals could settle in any position in a continuous space. Although such models have been formulated (e.g., Malécot 1967; Sawyer 1977), it is known that they do not follow a well-defined set of biological assumptions (Maruyama 1972; Felsenstein 1975; see Barton et al. 2002 for an alternative approach for continuous populations). Individuals are assumed to be diploids by a model with two independent genes per node. To avoid edge effects, the lattice is represented on a circle for a one-dimensional model or a torus for a two-dimensional model. Edge effects have little influence on local differentiation when the habitat area (i.e., the lattice size) is large when compared to the mean dispersal. Finally, we considered that dispersal occurs through gametes only.

The life cycle is divided into four steps: (1) at each reproductive event, each individual gives birth to a great number of gametes, and then dies; (2) gametes undergo the effect of mutations; (3) gametes disperse; (4) diploid individuals are formed, and (5) competition brings back the number of adults in each deme to one.

Coalescent Algorithm

The genealogical tree of a sample of n genes taken from a panmictic population of constant size N can be modeled using a stochastic process known as the n-coalescent. This process was introduced by Kingman (1982*a*, 1982*b*) as an approximation of a gene genealogy under the "Wright-Fisher" neutral model (see also Hudson 1990, Tajima 1983). More sophisticated models have since been developed for analysis of more complex evolutionary scenarios with recombination, selfing, and variable population size (reviewed in Nordborg 2001).

The n-coalescent approximation can be used in the same context as diffusion equations (Nordborg 2001). It is thus valid for a restricted numbers of models of population structure, e.g., panmictic populations or the infinite island model. In the present work, we focused on isolation by distance. For this category of models, no analytical treatment of coalescence time or coalescence probabilities has been done for more than two genes. Algorithms such as those developed for likelihood estimation by Griffiths and collaborators (see Nath and Griffiths 1996; Bahlo and Griffiths 2000) could in principle deal with continuous models; however, they are not ready for demographic inferences (De Iorio and Griffiths, personal communication). The coalescent algorithm we developed is not based on the n-coalescent theory; rather it is an algorithm for which coalescence and migration events are considered "generation by generation" until the common ancestor of the sample has been found. The idea of tracing lineages back in time generation by generation is fundamental in the coalescence theory, and is well described in Nordborg (2001). At least one study already used this simple concept for simulations (i.e., Pope, Estoup, and Morris 2000). Although such a generation-by-generation algorithm leads to less efficient simulations in terms of computation time than those based on the n-coalescent theory, it is much more flexible when complex demographic and dispersal features are considered. The algorithm described below and the program used in this study were checked at every step during elaboration by comparison with exact analytical results for probabilities of identity in models of isolation by distance on finite lattice (e.g., Malécot 1975 for the lattice model, adapted to different mutation models following Rousset 1996). These comparisons show that estimates of identity probabilities from our program and

analytical expectations differ by less than one per thousand for sufficiently long runs.

Let us consider, at a given time and on a two-dimensional lattice, a sample of $n(0)$ genes numbered 1 to $n(0)$. The position of each gene on this lattice is given by a pair of coordinates (x,y). The set of coordinates of sampled genes is given by the two vectors $X(0) = [x_1(0), \ldots, x_{n(0)}(0)]$, $Y(0) = [y_1(0), \ldots, y_{n(0)}(0)]$, where $x_i(0)$ and $y_i(0)$ are the coordinates of the gene i at $G = 0$, with G corresponding to the number of generations since sampling.

This algorithm goes backward in time, generation by generation (considering discrete generations). At $G = 1$, parents of our $n(0)$ sampled genes have coordinates $x_i(1) = x_i(0) + dx$, $y_i(1) = y_i(0) + dy$, where dx and dy are random variables representing dispersal distance in one dimension, expressed in number of steps on the lattice. Under a two-dimensional model, the density function of the random variable (dx,dy) is given by $b_{dx,dy}$, the "backward" dispersal function. The term *backward* is used because the position of the parental gene is determined knowing the position of its descendant gene. This function is calculated using $f_{dx,dy}$, the forward dispersal density function describing where descendants go. The dispersal functions are detailed in the next section. We assume that dispersal is independent in each direction, so that $f_{dx,dy} = f_{dx} \times f_{dy}$. Considering that density is homogenous in space, backward dispersal functions are equal to forward dispersal functions, so that $b_{dx,dy} = f_{dx,dy} = f_{dx} \times f_{dy}$.

Once the position of the parents on the lattice is known, the coalescence events occurring at $G = 1$ are assessed. In other words, we determine whether some genes share a common parent at $G = 1$. This step corresponds to the idea of "individuals picking their parents at random from the previous generation" (Nordborg 2001). A coalescence event occurs if genes are both on the same lattice node and if they originate from the same parental gene. Multiple coalescences are allowed. The probability for a coalescence of k genes in a given parental gene is $1/2^{k-1}$ under the model with one individual per lattice node. In this case, the remaining j genes from the same lattice node coalesce in the other parental gene. For convenience, we keep the numbering ($i \in [1, \ldots, n(0)]$) of descendant genes for their parents when these genes do not coalesce and attribute new numbers ($i \in [n(0) + 1, \ldots, n(1)]$) for the parents of the coalesced genes. A gene i at $G = 0$ and its parent at $G = 1$ have the same number if there was no coalescence event between the gene i and another gene at $G = 0$. Thus our numbering refers more to the branches of the coalescent tree than to the genes themselves. This particular numbering of branches, nodes, and genes is illustrated in figure 1. At G = 1, we have $X(1) = (x_1(1), ., x_{n(1)}(1))$, $Y(1) = (y_1(1), ., y_{n(1)}(1))$, the $n(1)$ geographic coordinates at $G = 1$ for each branch corresponding to a lineage of our sample. We keep in memory the ages of the tree "nodes" (corresponding to coalescence events) and the labels of the branches descending from this "node." The entire process is repeated over generations until the most recent common ancestor of our entire gene sample has been found.

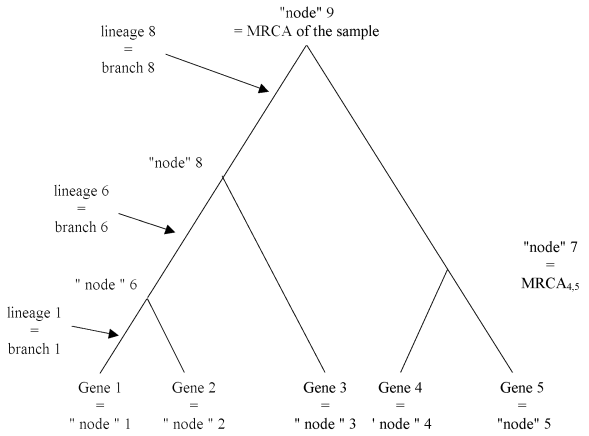


FIG. 1.—Numbering of branches, genes, and nodes of a genealogical tree for a sample of five genes as described by our coalescence algorithm.

Dispersal Functions

Biologically realistic dispersal functions often have a high kurtosis (Endler 1977; Kot, Lewis, and van den Driessche 1996). Forward dispersal distributions for which the probability of moving k steps (for $0 < k \leq K_{max}$) in one direction is of the form $f_k = f_{-k} = M/k^n$ were considered, with parameters M and n controlling the total dispersal rate and the kurtosis, respectively.

By suitable choice of the two parameter values, large kurtosis can be obtained with high migration rates (Rousset 2000). For all of our simulations, we used a dispersal distribution with a moderate σ^2 value ($\sigma^2 = 4$), corresponding to a dispersal distribution with parameters:

$$f_1 = f_{-1} = 0.06, \quad f_2 = f_{-2} = 0.03 \quad \text{and for}$$
$$2 < k < 49, \quad M = 0.802 \quad \text{and} \quad n = 2.518. \quad (1)$$

With such a dispersal distribution the product $4\pi D\sigma^2$ is 50.26. This value corresponds to a relatively strong isolation by distance, which appears biologically reasonable for many species (see references cited in the *Introduction*).

Mutation Processes

One interesting feature of the coalescent-based approach is that, for neutral loci, genealogical and mutation processes are totally independent, so that the effects of mutation are simply superimposed on the genealogical tree obtained for the gene sample.

Two theoretical mutation models, the infinite allele model (IAM: Kimura and Crow 1964) and the K-allele model (KAM: Crow and Kimura 1970), have sometimes been used for microsatellite loci. However, the most widely adopted model for microsatellite mutation is the stepwise mutation model (SMM: Ohta and Kimura 1973) in which the mutant allele differs from its parent by one repeat. Direct and indirect studies have shown that mutations of several repeats also occurred, indicating that a strict one-step model is inappropriate (Estoup and Angers 1998; Gonser et al. 2000; Ellegren 2000). In

practice, modeling assumptions are commonly limited to the SMM (e.g., Reich and Goldstein 1998; Wilson and Balding 1998), and sensitivity of the final inferences to this assumption may be substantial, although this is rarely investigated. In several studies (e.g., Pritchard et al. 1999), a generalization of the SMM was adopted in which the change in the number of repeat units forms a geometric random variable. This generalization was named the GSM (generalized stepwise mutation) model. The geometric distribution in our GSM model refers to a change expressed in an (absolute) number of repeat units subsequently added or withdrawn to the mutating allele with equal probability. Under this model, the large data set of microsatellite mutations of Dib et al. (1996) in humans suggests an estimate of the variance of the geometric distribution near 0.36 (Estoup et al. 2001). The GSM does not capture all the complexity of the mutation process at microsatellite loci. In particular, constraints on allele size occur at some microsatellite loci (reviewed in Amos 1999; Estoup and Cornuet 1999; Ellegren 2000) and potentially affect various statistics in population genetics (Estoup et al. 2002). This evolutionary feature, particular to microsatellite loci, was thus tested on our method. Allele size constraints were included in our simulations by imposing reflecting boundaries to the allele size range (e.g., Feldman et al. 1997; Estoup et al. 1999). Another outstanding feature of the microsatellite mutation process is that within-loci mutation rate increases with allele length (Ellegren 2000; Huang et al. 2002). Whether this increase is linear with the number of repeats remains subject to further investigation (Schlötterer 2000; Stumpf and Goldstein 2001; Brohede et al. 2002). In our simulations, we considered a linear model in which (1) the mutation rate was fixed to 5×10^{-4} for the allelic state of the root of the tree (fixed at 100 repeats units and considered the "middle size allele"); (2) a decrease in mutation rate with allele size of 0.1% or 1% per repeat unit for a weak or a strong variation, respectively is simulated for alleles shorter than 100 repeat units; (3) a similar increase is simulated for alleles longer than 100 repeat. In other words, this leads to the linear form: $\mu(L) = \mu_0 + s^*L$, where $\mu(L)$ is the mutation rate for an allele of size L, μ_0 the mutation rate for the smallest allele, and s the increase per repeats unit. We set $s = 0.1\%$ or 1% for a weak or a strong variation, respectively, to be close to the value given in Brohede et al. (2002).

Interlocus variability in the mutation rate potentially decreases the precision of parameter estimation in population genetics (Takezaki and Nei 1996; Gonser et al. 2000). The effect of variable mutation rate was thus tested as well. Little information is available on the interlocus variance of the mutation rate at microsatellite loci. Several pedigree studies show that the mutation rates can differ across loci in important respects (reviewed in Schlötterer 2000). Without more information, we modeled variable mutation rates at microsatellite loci by drawing single locus mutation rate values in a gamma distribution with parameters (shape, scale) being (2, 2.5 10^{-4}). This distribution has a mean equal to 5×10^{-4}, a value considered as the average mutation rate in many species (reviewed in Estoup and Angers 1998), and 2.5% and 97.5% quantiles equal to 6×10^{-5} and 1.4×10^{-3}, respectively. These values are similar to the mean and 95% confidence interval values typically considered for autosomal microsatellites in humans (Weber and Wong 1993).

The following step-by-step procedure was used to add mutations to the genealogical tree. Take at random two genes i, j and their most recent common ancestor, the gene l, and let $state_i$, $state_j$, $state_l$ be their respective allelic states. The number of mutations that occurred in lineage i is proportional to the length L_i (expressed in number of generations) of branch i (from l to i) and is given by a binomial distribution with parameters (μ, L_i), which can be approximated by a Poisson process with parameter μL_i. Let m_i be the number of mutations that occurred on branch i. One can easily deduce $state_i$ from $state_l$ through m_i successive steps, each step corresponding to a mutation event under the chosen mutation model. The allelic states of the various genes of the sample were obtained starting from a given state for the common ancestor of the sample (root of the genealogical tree) and going forward in time on each branch.

Method of Analysis

Each simulation iteration gave the genotypes at l polymorphic loci for ($n \times n$) individuals denoted by their coordinates on the lattice. l independent coalescent trees were used to simulate multi-locus genotypes. This process was repeated 1,000 times giving 1,000 multilocus samples sharing the same demographic conditions. We computed estimates of the parameter

$$a_r \equiv \frac{Q_w - Q_r}{1 - Q_w}$$

for each pair of individuals, where Q_w is the probability of identity in state for two genes taken from the same individual, and Q_r the probability of identity in state for two genes at geographical distance r (Rousset 2000). The statistic a_r is a parameter analogous to the parameter $F_{ST}/(1 - F_{ST})$, calculated between individuals (and not between populations, as in Rousset 1997). An estimator of a_r for a pair π of individuals taken from the P different possible pairs is:

$$\hat{a} \equiv \frac{SS_{b(\pi)}P}{\sum_{k=1}^{P} SS_{w(k)}} - \frac{1}{2}$$

with

$$SS_{b[etween](\pi)} \equiv \sum_{i,u} \left(X_{i.:u} - X_{..:u}\right)^2$$

and

$$SS_{w[ithin](\pi)} \equiv \sum_{i,j,u} \left(X_{ij:u} - X_{i.:u}\right)^2,$$

where $X_{ij:u}$ is an indicator variable taking the value 1 if gene i of individual j is of allelic type u and the value 0 otherwise (Rousset 2000).

To test the effect of using a statistic that takes into account the allele length differences (and hence the stepwise mutational process occurring at microsatellite

loci), we defined another parameter b_r, equivalent to a_r, except that it is defined in terms of squared differences in microsatellite allele lengths (*SD*) instead of probabilities of non- identity in state ($1 - Q$). Thus, we have

$$b_r \equiv \frac{SD_r - SD_w}{SD_w},$$

where SD_r is the expectation of the squared length differences between two genes at geographical distance r and SD_w is the expectation of the squared length differences between two genes taken in the same individual. b_r was estimated for a pair π of individuals taken from the P different possible pairs in a way similar to a_r:

$$\hat{b} \equiv \frac{SSD_{b(\pi)}P}{\sum_{k=1}^{P} SSD_{w(k)}} - \frac{1}{2}$$

with

$$SSD_{b[etween](\pi)} \equiv \sum_i (S_{i.} - S_{..})^2$$

and

$$SSD_{w[ithin](\pi)} \equiv \sum_{i,j} (S_{ij} - S_{i.})^2,$$

where S_{ij} is a variable representing the size of gene i of individual j, expressed in number of repeat units.

For each of the 1,000 repetitions, the value of the slope of the regression line between $\hat{a}$ (or $\hat{b}$) and the logarithm of geographical distance was computed. In the limit of low mutation rates, the inverse of the slope is an estimate of the product $4\pi D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance (Rousset 1997). It is worth noting that high mutation rates should not result in an asymptotic bias as long as the focus is on local processes involving distances between sampled individuals

$$r \ll \frac{\sigma}{\sqrt{2\mu}}.$$

Beyond this limit, the linear relationship between a_r (or b_r) and the logarithm of the distance holds less well (for details, see Rousset 1997). Thus, if the analysis is done at a small geographical scale, the use of highly variable loci such as microsatellite loci should not bias the estimation. However, the effect of mutation on small sample properties of the estimator needs to be tested. The quality of an estimator is usually assessed through the computation of its bias and its mean square error (MSE). These measures are suitable when estimates have approximately a normal distribution but not when the estimate is sometimes infinite. In the present case, a negative slope should be interpreted as an infinite estimate of $D\sigma^2$. Therefore we chose to work on the slope values and not on $D\sigma^2$ estimates. The following statistics were estimated over all repetitions: (1) the mean relative bias between the value of the slope and the expected value $1/(4\pi D\sigma^2)$; (2) the standard error on this relative bias; and (3) the mean square error (MSE = Bias^2 + var). The bias and the MSE are relative values, as they are computed from the ratio of the estimate to the value to be estimated, $1/(4\pi D\sigma^2)$. We also computed the proportion of negative slopes found and the probability that the estimate was within a factor of 2 from $1/4\pi D\sigma^2$. Note that the latest measure is strictly equivalent to the probability that the $D\sigma^2$ estimate was within a factor of 2 from the expected $D\sigma^2$ value.

Table 1
Coverage Probability of 95% Confidence Intervals Around the Regression Slope Using an ABC Bootstrap Procedure

	Bootstrap Sample Size		
	7 loci	13 loci	25 loci
Coverage probability	0.842	0.885	0.90
Proportion of intervals below the slope value	0.020	0.030	0.030
Proportion of intervals above the slope value	0.138	0.085	0.070

An accurate estimate of the uncertainty associated with parameter estimates is important to avoid misleading inferences. The nonparametric ABC bootstrap procedure described in DiCiccio and Efron (1996) was adapted to compute 95% confidence intervals around the regression slope. ABC bootstrap is a procedure that generates approximated bootstrap confidence intervals without real resampling. It is useful for estimation methods with high computation time needs. In this procedure, we considered genotypic data at each locus as independent replicates of the genealogical process. Tests of this procedure were performed using the same simulation program described above by calculating probability coverage of the confidence intervals for 1,000 simulated data sets. We choose arbitrarily a dispersal distribution with $\sigma^2 = 4$ [parameters given in equation (1)]. For each repetition, 100 individuals were sampled every two lattice nodes within an area of $(10\sigma \times 10\sigma)$ on a (100×100) lattice. Estimates of a_r and 95% confidence intervals were calculated for 7, 13, or 25 loci evolving under a SMM with a mutation rate equal to 5×10^{-4}.

Results

ABC Bootstrap

Table 1 shows that the non parametric ABC bootstrap procedure gives inaccurate 95% confidence intervals in terms of coverage probability even for large number of loci (e.g., coverage probability is 0.90 instead of 0.95 for 25 loci). The inaccuracy mostly concerns the lower bound of the confidence intervals for the regression slope (i.e., the proportion of intervals above the slope value is 0.07 instead of 0.025 for 25 loci; table 1). This may reflect the asymmetrical shape of the distribution with a long tail for small values (i.e., large $D\sigma^2$, data not shown). The effect of asymmetrical distribution on ABC bootstrap was tested on a simpler statistical model. ABC confidence intervals were computed for the mean of a random sample drawn in a bivariate student distribution with density

$$\Pr(r) = 2\pi r \frac{\Gamma[1+p]}{\pi u \Gamma[p]} (1 + r^2/u)^{-1-p}$$

and parameters (p,u) being (1,1). This distribution is asymmetrical with an infinite kurtosis and an infinite skewness. Even for very large sample sizes (5000

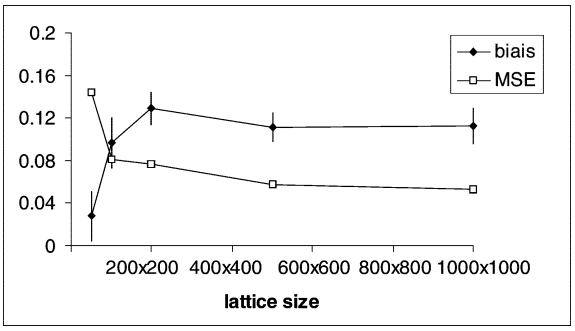


FIG. 2.—Influence of the lattice size on the estimation of the product $1/4\pi D\sigma^2$. NOTE—Only 500 iteration were done for each case. Vertical bars represent standard errors on the bias.

replicates, results not shown), the ABC procedure gives an inaccurate upper bound, resulting in underestimated confidence intervals (results not shown). In the case of the regression slope, the inaccuracy increases for small sample size (e.g., 0.842 instead of 0.95 for seven loci; table 1).

Because of the important computation time needed to construct ABC confidence intervals, this procedure was not used for evaluating the influence of the sampling scale and mutational factors on the estimation of $D\sigma^2$ (see *Models and Methods*).

Influence of the Sampling Scale

Previous simulations with two-allele loci suggested that the regression method would be efficient if one can sample all individuals within an area of about $10\sigma \times 10\sigma$, giving a sample size of $100D\sigma^2$ individuals (Rousset 2000). It is worth noting that if $D\sigma^2$ is greater than say 5, it becomes difficult in practice to sample and genotype all individuals (>500 individuals). Hence, since the number of individuals to sample is necessarily limited, the method should be less efficient when $D\sigma^2$ increases. In practice, biologists collect samples of a reasonably large number of individuals (say 100) within an area larger or smaller than the recommended ($10\sigma \times 10\sigma$) area when $D\sigma^2$ is small or large respectively. In order to assess the effect of such practical "non-scaled sampling," we simulated a distribution of dispersal with $\sigma^2 = 4$ [parameters given in expression (1)] and four different sampling schemes. One hundred individuals were taken: (1) every lattice node within an area of ($5\sigma \times 5\sigma$), for the first sampling scheme; (2) every two lattice nodes within an area of ($10\sigma \times 10\sigma$), for the second one; (3) every five lattice nodes within an area of ($25\sigma \times 25\sigma$) for the third one; and (4) every ten lattice nodes within an area of ($50\sigma \times 50\sigma$) for the last one. For each repetition the parameter estimated is a_r for 13 loci evolving under a SMM with a mutation rate equal to 5×10^{-4}. We considered that a set of 13 loci represents a reasonable number of loci in empirical studies using microsatellites. A two dimensional lattice of (200×200) individuals was considered for the first three sampling schemes and of (500×500) individuals for the last one, to avoid edge effects on the estimations when considering samples larger than half the length of the lattice. Figure 2 shows that lattice size has no major effect on the estimation, except if it is less than ten times the mean dispersal distance (simulation parameters are those used in this paragraph). Unless the lattice size is very small (50*50), the bias and the MSE do not differ notably from those for a very large lattice size (1000*1000).

The sampling scale seems to have only a limited effect on the MSE of the $D\sigma^2$ estimation (table 2). Whatever sampling scale is considered (i.e., smaller or larger than the recommended area) the MSE is low (values between 5% and 12% in the studied cases). In contrast, the sampling scale has a great effect on the bias. A sample taken from an area two times smaller than the recommended area (first column of table 2) gave a large and positive bias (22%). The bias decreases when the sampling area increases and becomes negative when the sampling area is larger than the recommended area, reaching high values (e.g., −21%, fifth column of table 2). However, it is worth noting that even for extreme sampling situations, estimates of $D\sigma^2$ are not very different from the expected value, as shown by the large proportion of estimated values falling within a factor of two from $D\sigma^2$ (>93%).

Influence of the Mutation Model

The following mutation models were considered: (1) the infinite allele model (IAM); (2) the *K*-allele model (KAM) with an arbitrary choice of $K = 10$ possible allelic states; (3) the stepwise mutation model (SMM); (4) the generalized stepwise model (GSM) with variance of the geometric distribution equal to 0.36; and (5) the GSM with constraints on allele size (bounded GSM). In the bounded GSM, the number of possible allelic states was equal to 10 or 20, each allelic state being separated by a single repeat unit.

Simulations were run considering a sample of 100 individuals for 13 loci evolving in a two-dimensional lattice of (100×100) individuals. For each repetition of the simulation process the parameter estimated is a_r. As it is often not easy in practice to sample most individuals from a small area, we considered a sample of (10×10) individuals taken every two nodes from an area of (20×20) nodes in the lattice. By doing so, we approximated the sampling scheme typically used in empirical studies. We also chose a dispersal distribution with a relatively large σ^2 value [i.e., $\sigma^2 = 4$, parameters given in equation (1)]. The logic underlying this choice is that the method may be inaccurate in this case and that it is more relevant to distinguish differences in efficiency when the method does not perform extremely well, than when it performs well, whatever the mutation model.

The mutation rate was first fixed at 5×10^{-4} for all loci for each mutation model. Our results show that the nature of the mutation model has little influence on the estimation of the product $D\sigma^2$ (table 3). Whatever mutation model is considered, the bias is positive and around 10%. Although the precision of the method is maximum under the IAM (MSE of 6%) and minimum under the GSM with strong constraints ($K = 10$, MSE = 0.11), these differences are small. For all mutation models more

Table 2
Influence of Sampling Scale on the Estimation of $1/4\pi D\sigma^2$

	Sampling Scale (Sampling Area)			
	1 (10 × 10)	2 (20 × 20)	5 (50 × 50)	10 (100 × 100)
Bias	0.219	0.130	−0.056	−0.205
(standard error)	(0.0077)	(0.0077)	(0.0072)	(0.0064)
MSE	0.106	0.0763	0.0554	0.082
2× coverage	0.999	0.996	0.967	0.93
Negative slope	0	0	0	0

NOTE—Sampling area is expressed in lattice node unit (see text for details). 2× coverages correspond to the probability that the estimate was within a factor of 2 from $1/4\pi D\sigma^2$.

than 97% of the estimations are within a factor 2 from the expected $D\sigma^2$ value.

For a given mutation rate, level of genetic diversity varies according to the mutation model considered. Because the level of genetic diversity is likely to have an important effect on the estimation of the product $D\sigma^2$, we studied the influence of different mutational models for the same level of diversity. The genetic diversity can be expressed in terms of probability of identity by $(1 - Q_w)$, where Q_w is the probability of identity in state of two genes taken in the same individual. This corresponds to the fraction of heterozygous individuals in the population. The influence of mutation models was thus studied with the same Q_w value for all mutation models. The conclusions are similar to those obtained with a mutation rate fixed at the same value for all mutation models (table 3). For a given value of genetic diversity, the bias and the MSE of $D\sigma^2$ estimates shows little variation among mutational models.

Influence of the Mutation Rate

The influence of the mutation rate (or the genetic diversity) has been studied for the GSM, a mutation model considered as more realistic for microsatellite loci than the SMM, the KAM, or the IAM (e.g., Estoup and Cornuet 1999). All other simulation parameters are those used for evaluating the influence of the mutation model. Our simulations showed that the mutation rate has a substantial effect on the bias and the MSE (fig. 3 and table 4). The MSE is more strongly influenced by the mutation rate than the bias. For "low" genetic diversities (i.e., H = 0.5), the observed bias is positive and never greater than 12%. In contrast, for genetic diversity lower than 0.6, the MSE is greater than 20% and increases relatively rapidly when the genetic diversity decreases. However, even for a genetic diversity lower than the mean genetic diversity observed in most microsatellite studies (e.g., about 0.5), 85% of the estimations are within a factor of two from $D\sigma^2$, but 15 negative slopes were found (table 4).

It is worth mentioning that the observed bias may be of two types: (1) the bias, inherent in the method, that is due to the effect of high mutation rate on the parameter value (we will name it the "parametric bias"); and (2) the bias due to the deviation of the estimates in relation to the parameter value considering a finite sample of individuals and loci (which we will name "small sample bias"). The method is expected to perform poorly for very high mutation rates because distances between some pairs of sampled individuals are then larger than

$$\frac{\sigma}{\sqrt{2\mu}}$$

(Rousset 1997). In such a case, the parametric bias is expected to be negative because the slope of the regression line will be underestimated (for details, see Rousset 1997). In our simulations, we have $\sigma = 2$ and the maximal distance between individuals equals $20\sqrt{2}$ lattice units, which is within

$$\frac{\sigma}{\sqrt{2\mu}}$$

for mutation rates lower than 0.001. However, our results show that for a genetic diversity of 0.8 (corresponding to a mutation rate of c. 0.005 in our model) the bias and the MSE are very low. The low values of the bias and the MSE in this case are likely to result from some compensatory effects between a positive "small sample bias" and a negative "parametric bias." When higher genetic diversity is considered (i.e., H = 0.85 corresponding to mutation rates of c. 0.05 in our model), the bias becomes large and negative and the MSE rapidly increases (table 4). This result is in agreement with the above prediction: for very high mutation rates the "parametric bias" becomes more important than the "small sample bias," so that the global bias observed for high mutation rates is negative.

It is sometimes considered that the large variation between loci of the mutation rate decreases the precision of parameter estimation in population genetics (e.g., Takezaki and Nei 1996; Gonser et al. 2000). To address this question, we considered 13 loci evolving under the GSM with mutation rates drawn for each locus in a gamma distribution of mean 5×10^{-4} (see earlier under *Models and Methods: Mutation Model*), all other simulation parameter values being the same as those used in the previous section. Our simulation results show that variable mutation rates for microsatellite loci have little effect on the estimation of $D\sigma^2$ (table 4). The bias and the MSE values are 11% and 11%, respectively, which does not differ much from the values of 10% and 9% obtained with a fixed mutation rate of 5×10^{-4}. More than 98% of the estimations are within a factor of 2 from $D\sigma^2$ and no negative estimates were found. Finally, our simulation results show that a linear increase in mutation rates with allele length has little effect on the estimation of $D\sigma^2$ (table 4). Strong or weak

Table 3
Influence of Mutational Processes on the Estimation of $1/4\pi D\sigma^2$ with Constant Mutation Rate or Constant Genetic Diversity for All Mutation Models

	Mutation Model											
	Constant Mutation Rate						Constant Genetic Diversity					
	IAM	KAM ($K = 10$)	SMM	GSM	Bounded GSM ($K = 10$)	Bounded GSM ($K = 20$)	IAM	KAM ($K = 10$)	SMM	GSM	Bounded GSM ($K = 10$)	Bounded GSM ($K = 20$)
Genetic diversity	0.787	0.711	0.703	0.772	0.679	0.720	0.68	0.68	0.68	0.68	0.68	0.68
Mutation rate	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0001	0.000218	0.000342	0.00012	0.0005	0.0002
Bias	0.109	0.0919	0.0917	0.104	0.0997	0.128	0.111	0.104	0.118	0.121	0.0997	0.108
(standard error)	(0.0067)	(0.0088)	(0.0093)	(0.00863)	(0.0101)	(0.009)	(0.01)	(0.01)	(0.015)	(0.0120)	(0.0101)	(0.010)
MSE	0.057	0.0853	0.0953	0.0852	0.112	0.098	0.119	0.109	0.119	0.159	0.112	0.121
2× coverage	0.998	0.982	0.975	0.987	0.976	0.984	0.96	0.97	0.96	0.938	0.976	0.962
Negative slope	0	0	0	0	0	0	0.001	0.001	0	0.001	0	0.002

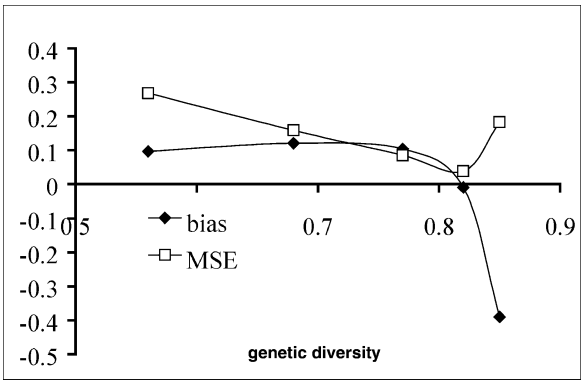


FIG. 3.—Influence of the mutation rate on the estimation of the product $1/4\pi D\sigma^2$. The mutation model is a GSM.

variations give similar results. The bias and the MSE values are about 10%–11% and 8%, respectively, which again does not differ much from the values of 10% and 9% obtained with a fixed mutation rate of 5×10^{-4}. No negative estimates were found, and more than 99% of the estimations are within a factor of 2 from $D\sigma^2$.

Test for a Statistic Taking into Account Allele Size Differences

The behavior of the statistic b_r, an equivalent of a_r based on allele sizes, has been studied under both the SMM (i.e., the mutation model under which this statistic is expected to perform optimally) and the GSM with a mutation rate fixed at 5×10^{-4}. All other simulation parameters values are those used in the two previous sections. Table 5 shows that the method of estimation of $D\sigma^2$ performs poorly when b_r is used. Under both the SMM and GSM, the increase in MSE as well as the number of negative slopes is spectacular. For instance the MSE goes from about 10% when using the classical measure a_r to values greater than 100% when using b_r. In contrast, the bias is only slightly increased compared to estimations using a_r. Although slight, the bias increase appears higher under the GSM than the SMM (+ 9% versus + 4%).

Discussion

A first general conclusion of this study is that the mutation model of the markers has little influence on the efficiency of the method of estimation of $D\sigma^2$ based on individual genotypes and allelic identity. Hence, the allele size homoplasy typically produced under stepwise mutation models (SMM and GSM), and specifically of microsatellite markers (reviewed in Estoup, Jarne, and Cornuet 2002 for different population genetics statistics), is not a feature prejudicial for the method described in this article. Our results dealing with constraints on allele sizes, an evolutionary feature also specific to microsatellite markers and known for substantially increasing size homoplasy, show that even extremely strong constraints (e.g., $K = 10$) have little effect on the estimation of $D\sigma^2$. These results can be interpreted in the context of

Table 4
Influence of the Mutation Rate on the Estimation of the Product $1/4\pi D\sigma^2$

	Mutation Rate					Interloci Variability (*)	Intraloci Variability (**)	
	0.00005	0.00012	0.0005	0.005	0.05		Weak	Strong
Genetic diversity	0.56	0.68	0.77	0.82	0.85	0.77	0.77	0.77
Bias	0.0972	0.121	0.104	0.00946	−0.390	0.114	0.0965	0.111
(standard error)	(0.01609)	(0.0120)	(0.00863)	(0.00616)	(0.0055)	(0.0096)	(0.00846)	(0.0081)
MSE	0.268	0.159	0.0852	0.0380	0.182	0.105	0.0808	0.0778
2× coverage	0.844	0.938	0.987	0.996	0.761	0.983	0.991	0.993
Negative slope	0.015	0.001	0	0	0	0	0	0

NOTE.—The mutation model is a GSM. (*) Mutation rate drawn in a gamma (2, 2.5 10^{-4}) distribution. (**) Variation in mutation rate with allele length is 0.1% and 1% per repeat unit for weak and strong variation, respectively (see text under *Influence of Mutation Rate* for details).

coalescent theory. Values of F-statistics, under the assumption of low mutation rate, can be deduced from the comparison between the distributions of coalescence probability for different pairs of genes (e.g., pairs from the same deme and pairs from different demes) (Rousset 1996, 2002). These distributions differ essentially by an "excess" of coalescence probability for the most related genes, this excess being concentrated in a brief period in the recent past. Under isolation by distance, the more distant the demes are, the more the "recent past" is extended to the distant past, permitting more mutations to act and thus to increase the sensitivity to variation in the mutation process. By contrast, sensitivity to range constraints has been observed for statistics that are not related to differences of distribution of coalescence times (e.g., genetic distances, Nauta and Weissing 1996) or for F-statistics when the excess probability of coalescence is not concentrated in a recent enough past (large subpopulation sizes and low dispersal rates, Gaggiotti et al. 1999). Because the method of Rousset (2000) focuses on local differentiation and thus on recent evolutionary processes corresponding to a narrow recent past zone, it is no surprise that mutation processes (including allele size constraints) have little influence on the estimation of $D\sigma^2$.

A second major conclusion of this study is that the mutation rate, or the genetic diversity (the latest being largely dependent on the mutation rate), has a strong influence on the estimation of $D\sigma^2$. This is in agreement with previous studies demonstrating that mutation rate is a more important feature than mutation processes for the estimation of demographic parameters through F-statistics (reviewed in Rousset 2001*a*; Estoup, Jarne, and Cornuet 2002). Interestingly, the heterozygosities at microsatellite loci are typically between 0.5 and 0.8 (reviewed in Estoup and Angers 1998), a range of values corresponding to the level of genetic diversity that was found to maximize the efficiency of the estimation of $D\sigma^2$. Moreover, the potential effect on the estimation of interlocus and intralocus variability in the mutation rate seems to be weak. Therefore microsatellites are more appropriate to estimate the product $D\sigma^2$ than less polymorphic markers such as allozymes. The importance of the level of variability of the loci used to estimate population parameters has been illustrated by several theoretical and empirical studies. For example, Robertson and Hill (1984) showed that precision in estimates of heterozygote deficiency (F_{is}) increases with the level of variability of the markers. Goudet et al. (1996) also showed that the power of statistical tests of differentiation increases with the number of alleles. In practice, although precise information on mutation rate is difficult to obtain, it is straightforward to calculate a genetic diversity index for a set of markers from which a level of efficiency can be inferred for the estimation of $D\sigma^2$. Our simulations also indicate that future studies should avoid loci with a very high level of genetic diversity (higher than, say, 0.85), because those loci were found to strongly bias negatively the estimations of $D\sigma^2$.

Many studies emphasize that traditional F_{ST} does not make use of the additional information provided by the difference in the number of repeat units at microsatellite loci. However, statistics developed for this purpose often have higher variance than statistics based on allele frequencies (e.g., Gaggiotti et al. 1999). In agreement with this finding, estimates computed using a statistic taking into account allele size differences increases by at least a factor of 10 the MSE compared to a statistic based on identity in state. This result parallels those of Gaggiotti et al. (1999), which showed that in many cases, especially when sample size and number of loci are "small" (i.e., under the conditions of most empirical studies), population structure measures based on allele frequencies alone are more reliable than measures specifically designed for microsatellite loci. Takezaki and Nei (1996) also showed that even for loci evolving under a strict SMM, genetic distances taking into account allele size differences are less efficient for phylogenetic inference than those based on identity in state, especially for short to moderate divergence times. The poor efficiency of this category of statistics appears to be a general feature of studies of evolutionary events, especially those referring to fine geographical and temporal scales.

The effects of the mutation processes and high mutation rates on the estimation of $D\sigma^2$ are expected to be more important at large geographical scales (Rousset 1997). In agreement with this expectation, our results showed that sampling at large distance leads to an underestimation of the regression slope and thus to an overestimation of $D\sigma^2$. Therefore sampling at large distance makes it less likely to detect a pattern of isolation by distance. In contrast, sampling from too small an area leads to an overestimation of the regression slope and thus

Table 5
$D\sigma^2$ Estimation Using a Statistic Taking into Account the Differences in Allele Length (B_r)

	Mutation Model[a]			
	SMM	SMM	GSM	GSM
Parameter estimated	a_r	b_r	a_r	b_r
Bias	0.0917	0.128	0.104	0.19
(standard error)	(0.0093)	(0.036)	(0.00863)	(0.034)
MSE	0.0953	1.13	0.0852	1.25
2× coverage	0.975	0.518	0.987	0.497
Negative slope	0	0.154	0	0.141

NOTE.—Mutation rate is 5.10^{-4}.

[a] SMM: stepwise mutation model; GSM: generalized stepwise mutation model.

to an underestimation of the product $D\sigma^2$. A possible explanation for this overestimation is that the linear relationship between estimates of a_r and the logarithm of the geographical distance is expected to hold less well over very short distances (Rousset 1997). However, using a sample not exactly appropriate to the biological case studied [i.e., a few times larger or smaller than the recommended area of $(10\sigma \times 10\sigma)$] still gives reasonably robust estimations because, in most cases, the estimated $D\sigma^2$ fell within a factor of 2 from the expected $D\sigma^2$ value.

Given our result on bootstrap confidence intervals, we alert biologists using this method on a standard-sized data set (10 loci and 150 individuals, e.g., Sumner et al. 2001) that ABC confidence intervals overestimate the lower bound for the regression slope and thus underestimate the upper bound for $D\sigma^2$. Construction of reliable confidence intervals based on the bootstrap is an ongoing problem for which a satisfactory solution has not yet been found, especially when the number of replications is limited computationally (DiCiccio and Efron 1996). Nevertheless, the ABC bootstrap procedure evaluated here should give an idea of the uncertainty of the $D\sigma^2$ estimate, namely a correct lower bound for $D\sigma^2$ and a minimal value for the upper bound. This procedure will be implemented in the next version of the population genetics package Genepop (Raymond and Rousset 1995).

Conclusion

Three conclusions inferred from our simulation study have important consequences for empirical investigations. First, we recommended using loci with high levels of polymorphism (genetic diversity around 0.7), although loci with too high genetic diversity, e.g., more than 0.85, should be avoided. Because the mutational processes, specifically size homoplasy and allele size constraints, have little influence on $D\sigma^2$ estimations, microsatellite markers seem to be the best choice at the present time. Second, using statistics based on allele size differences at microsatellite loci gives unreliable estimations of $D\sigma^2$ because of the very high variance of those estimations. Third, it is important to restrict the sampling design to a relatively small geographical area in order to work at a local geographical scale; however, it is necessary to sample on a relatively large scale when σ is high. Optimizing the method studied here requires a previous knowledge of σ, and we therefore recommended using a preliminary estimate of σ to allow subsequent design of an appropriate sampling scheme. In the absence of a preliminary estimate of σ, a rough estimate of this parameter deduced from consideration of known dispersal mechanisms should be useful to define the minimal scale of the study (e.g., Leblois et al. 2000). If these aspects are approximately satisfied, the method should give estimates of the product $D\sigma^2$ with low bias and low mean square error. Finally, the ABC bootstrap procedure, as implemented in the package Genepop (Raymond and Rousset 1995), should be useful to estimate a 95% confidence interval on $D\sigma^2$, although the upper bound of this interval is likely to be underestimated.

Acknowledgments

We thank R. Streiff, B. Danforth, and three anonymous reviewers for constructive comments on an earlier version of the manuscript. This work was supported financially by the AIP no. 00202 "biodiversité" from the Institut Français de Biodiversité. This is paper 2003-002 of the Institut des Sciences de l'Evolution.

Literature Cited

Amos, W. 1999. A comparative approach to the study of microsatellite evolution. Pp. 66–79 *in* D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford.

Bahlo, M., and R. C. Griffiths. 2000. Inference from GeneTree in a subdivided population. Theor. Pop. Biol. **57**:79–95.

Barton, N. H., F. Depaulis, and A. M. Etheridge. 2002. Neutral evolution in spatially continuous populations. Theor. Popul. Biol. **61**:31–48.

Brohede, J., C. Primmer, A. Møller, and H. Ellegren. 2002. Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. Nucleic Acids Res. **30**:1997–2003.

Clark, J. S., M. Silman, R. Kern, E. Macklin, and J. HilleRis-Lambers. 1999. Seed dispersal near and far: patterns across temperate and tropical forests. Ecology **80**:1475–1494.

Crawford, T. J. 1984. The estimation of neighborhood parameters for plant populations. Heredity **52**:273–283.

Crow, J. F., and M. Kimura, 1970. An introduction to population genetics theory. Harper & Row, New York.

Dib, C., S. Faure, C. Fizames et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380**:152–154.

DiCiccio, T. J., and B. Efron. 1996. Bootstrap confidence intervals (with discussion). Stat. Sci. **11**:189–228.

Ellegren, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24**:400–402.

Endler, J. A. 1977. Geographical variation, speciation, and clines. Princeton University Press, Princeton, N.J.

Estoup, A., and B. Angers. 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. Pp. 55–86 *in* G. Carvalho, ed. Advances in molecular ecology. NATO ASI series. IOS Press, Amsterdam.

Estoup, A., and J.-M. Cornuet. 1999. Microsatellite evolution: inferences from population data. Pp. 49–65 *in* D. B. Goldstein and C. Schlötterer, eds. Microsatellites: evolution and applications. Oxford University Press, Oxford.

Estoup, A., P. Jarne, and J.-M. Cornuet. 2002. Homoplasy at microsatellite loci and its consequences for population genetics analysis. Mol. Ecol. **11**:1591–1604.

Estoup, A., I. J. Wilson, C. Sullivan, J.-M. Cornuet, and C. Moritz. 2001. Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. Genetics **159**:1671–1687.

Felsenstein, J. 1975. A pain in the torus: some difficulties with models of isolation by distance. Am. Nat. **109**:359–368.

Gaggiotti, O. E., O. Lange, K. Rassmann, and C. Gliddon. 1999. A comparison of two methods for estimating average levels of gene flow using microsatellites data. Mol. Ecol. **8**:1513–1520.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92**:6723–6727.

Gonser, R., P. Donnelly, G. Nicholson, and A. Di Rienzo. 2000. Microsatellite mutations and inferences about human demography. Genetics **154**:1793–1807.

Goudet, J., M. Raymond, T. de Meeüs, and F. Rousset. 1996. Testing differentiation in diploid populations. Genetics **144**:1931–1938.

Hastings, A., and S. Harrison. 1994. Metapopulation dynamics and genetics. Annu. Rev. Ecol. Syst. **25**:167–188.

Huang, Q.-Y., F.-H. Xu, H. Shen, H.-Y. Deng, Y.-J. Liu, Y.-Z. Liu, J.-L. Li, R. R. Becker, and H.-W. Deng. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70**:625–634.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyama and J. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics **49**:725–738.

Kingman, J. F. C. 1982*a*. The coalescent. Stochast. Proc. Appl. **13**:235–248.

———. 1982*b*. On the genealogy of large populations. J. Appl. Prob. **19A**:27–43.

Koenig, W. D., D. Van Vuren, and P. N. Hooge. 1996. Detectability, philopatry, and the distribution of dispersal distances in vertebrates. Trends Ecol. Evol. **11**:514–517.

Kot, M., M. A. Lewis, and P. van den Driessche. 1996. Dispersal data and the spread of invading organisms. Ecology **77**:2027–2042.

Leblois, R., F. Rousset, D. Tikel, C. Moritz, and A. Estoup. 2000. Absence of evidence for isolation by distance in expanding cane toad (*Bufo marinus*) population: an individual-based analysis of microsatellite genotypes. Mol. Ecol. **9**:1905–1909.

Malécot, G. 1950. Quelques schémas probabilistes sur la variabilité des populations naturelles. Ann. Univ. Lyon A **13**:37–60.

———. 1967. Identical loci and relationship. Pp. 317–332 *in* L. M. Lecam and J. Neyman, eds. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 4. California University Press, Berkeley.

———. 1975. Heterozygoty and relationship in regularly subdivided populations. Theor. Popul. Biol. **8**:212–241.

Maruyama, T. 1972. Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. Genetics **70**:639–651.

Michalakis, Y., and L. Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special interest to microsatellite loci. Genetics **142**: 1061–1064.

Nath, H. B., and R. C. Griffiths. 1996. Estimation in an island model using simulation. Theor. Pop. Biol. **50**:227–253.

Nauta, M. J., and F. J. Weissing. 1996. Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics **143**:1021–1032.

Nordborg, M. 2001. Coalescent theory. Pp. 179–208 *in* D.A. Balding, M. Bishop and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, Chichester, U.K.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**:201–204.

Pope, L. C., A. Estoup, and C. Moritz. 2000. Phylogeography and population structure of an ecotonal marsupial, *Bettongia tropica*, determined using mtDNA and microsatellites. Mol. Ecol. **9**:2041–2053.

Portnoy, S., and M. F. Willson. 1993. Seed dispersal curves: behavior of the tail of the distribution. Evol. Ecol. **7**:25–44.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. 1999. Population growth of human Y chromosome microsatellites. Mol. Biol. Evol. **16**:1791–1798.

Raymond, M., and F. Rousset. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. **86**:248–249.

Reich, D. E., and D. B. Goldstein. 1998. Genetic evidence for a paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA **95**:8119–8123.

Robertson, A., and W. G. Hill. 1984. Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. Genetics **107**:703–718.

Rousset, F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142**:1357–1362.

———. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145**:1219–1228.

———. 2000. Genetic differentiation between individuals. J. Evol. Biol. **13**:58–62.

———. 2001*a*. Genetic approaches to the estimation of dispersal rates. Pp. 18–28 *in* J. Clobert, E. Danchin, A. A. Dhondt, and J. D. Nichols, eds. Dispersal: individual, population and community. Oxford University Press, Oxford.

———. 2001*b*. Inferences from spatial population genetics. Pp. 239–265 *in* D. A. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics. John Wiley & Sons, Chichester, U.K.

Sawyer, S. 1977. Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv. Appl. Prob. **9**:268–282.

Schlötterer, C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma **109**:365–371.

Slatkin, M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47**:264–279.

———. 1994. Gene flow and population structure. Pp. 3–17 *in* L. A. Real, ed. Ecological genetics. Princeton University Press, Princeton, N.J.

———. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457–462.

Spong, G., and S. Creel. 2001. Deriving dispersal distances from genetic data. Proc. R. Soc. Lond. Ser. B **268**:2571–2574.

Stumpf, M. P. H., and D. B. Goldstein. 2001. Genealogical and evolutionary inference with the human Y chromosome. Science **291**:1738–1742.

Sumner, J., F. Rousset, A. Estoup, and C. Moritz. 2001. "Neighborhood" size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. Mol. Ecol. **10**:1917–1927.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

Takezaki, N., and M. Nei. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellites DNA. Genetics **144**:389–399.

Weber, J. L., and C. Wong. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.

Wilson, I. J., and D. J. Balding. 1998. Genealogical inference from microsatellite data. Genetics **150**:499–510.

Wright, S. 1943. Isolation by distance. Genetics **28**:114–138.

———. 1946. Isolation by distance under diverse systems of mating. Genetics **31**:39–59.

Pierre Capy, Associate Editor

Accepted October 11, 2002

Influence of Spatial and Temporal Heterogeneities on the Estimation of Demographic Parameters in a Continuous Population Using Individual Microsatellite Data

Raphael Leblois,*,†,[1] François Rousset† and Arnaud Estoup*

**Centre de Biologie et de Gestion des Populations, Campus International de Baillarguet CS 30 016, 34988 Montferrier sur Lez, France and* †*Laboratoire Génétique et Environnement, Centre National de la Recherche Scientifique-UMR 5554, 34095 Montpellier, France*

Manuscript received July 4, 2003
Accepted for publication October 18, 2003

ABSTRACT

Drift and migration disequilibrium are very common in animal and plant populations. Yet their impact on methods of estimation of demographic parameters was rarely evaluated especially in complex realistic population models. The effect of such disequilibria on the estimation of demographic parameters depends on the population model, the statistics, and the genetic markers used. Here we considered the estimation of the product $D\sigma^2$ from individual microsatellite data, where D is the density of adults and σ^2 the average squared axial parent-offspring distance in a continuous population evolving under isolation by distance. A coalescence-based simulation algorithm was used to study the effect on $D\sigma^2$ estimation of temporal and spatial fluctuations of demographic parameters. Estimation of present-time $D\sigma^2$ values was found to be robust to temporal changes in dispersal, to density reduction, and to spatial expansions with constant density, even for relatively recent changes (*i.e.*, a few tens of generations ago). By contrast, density increase in the recent past gave $D\sigma^2$ estimations biased largely toward past demographic parameters values. The method was also robust to spatial heterogeneity in density and estimated local demographic parameters when the density is homogenous around the sampling area (*e.g.*, on a surface that equals four times the sampling area). Hence, in the limit of the situations studied in this article, and with the exception of the case of density increase, temporal and spatial fluctuations of demographic parameters appear to have a limited influence on the estimation of local and present-time demographic parameters with the method studied.

DISPERSAL rates and population sizes or densities are important demographic parameters in evolutionary processes. Many studies have attempted to estimate those parameters, using direct methods (*e.g.*, mark-recapture methods) or indirect methods (genetic markers). Discrepancies between estimations based on direct and indirect methods have often been attributed to inadequacies of the assumptions of the genetic models in indirect methods (Hastings and Harrison 1994; Slatkin 1994; Koenig *et al.* 1996). The assumptions that have usually been considered inadequate are those related to the modalities of dispersal (*e.g.*, the island model), the mutation rates and processes of genetic markers, the selective neutrality of genetic markers, and the demographic stability in time and space. The latter assumption raises the question of the exact meaning of demographic parameter estimations in biological systems for which temporal and/or spatial fluctuations of demographic parameters have occurred. With a few exceptions (*e.g.*, Stone and Sunnucks 1993; Beebee and Rowe 2001; Spong and Hellborg 2002), population geneticists usually consider that contemporary spatial patterns of diversity reflect the past more than the present-time population dynamics of a species. Whitlock and McCauley (1999) recently concluded that estimates of the number of migrants between subpopulations from *F*-statistics under the assumption of an island model at equilibrium were "likely to be correct within a few orders of magnitude" only because assumptions of the genetic model (*i.e.*, equal migration, no selection, and demographic stability) are often violated in biological systems. This degree of precision is of little value for understanding the present-time demographic processes of populations. This is particularly worrying in a practical context since reliable estimates of present or at least recent migration rates, dispersal distances, or densities are increasingly demanded as integral elements of applied management and conservation decisions.

The effect of temporal and spatial fluctuations on the estimation of demographic parameters strongly depends on the type and intensity of the fluctuation encountered. However, it also strongly depends on the population models assumed, the statistics computed, and the genetic markers used. Most studies dealing with disequilibrium situations referred to the classical island model or to the Wright-Fisher population model and

[1]*Corresponding author:* Laboratoire Génétique et Environnement, Institut des Sciences de l'Evolution, UMR 5554-CC065, Université des Sciences et Techniques du Languedoc, Pl. E. Bataillon, 34095 Montpellier, France. E-mail: leblois@isem.univ-montp2.fr

only a few of them have considered more sophisticated and realistic models (but see Slatkin 1993). In numerous species, individual dispersal is restricted in space (see references in Leblois *et al.* 2003). A method of analysis adapted to a "continuous" population evolving under isolation by distance was developed to estimate the product $D\sigma^2$, where D is the density of adults and σ^2 the average squared axial parent-offspring distance (Rousset 2000). This method uses a regression of estimators of a parameter a_r to the geographical distances or the logarithm of the geographical distances in one or two dimensions, respectively. The parameter a_r, defined in Rousset (2000), is analogous to the parameter $F_{ST}/(1 - F_{ST})$ but is calculated between individuals (see *Method of analysis* for details about this parameter and its estimator). The inverse of the slope of the regression line gives an estimate of $4\pi D\sigma^2$ (Rousset 1997). The method is valid for leptokurtic distributions of dispersal distance (Rousset 2000; Leblois *et al.* 2003), a feature commonly observed in natural populations (review and data in Endler 1977; Portnoy and Willson 1993). Because analysis of genetic differentiation is made at a small (local) geographical scale, heterogeneity of demographic parameters such as dispersal or density is reduced and hence its influence on genetic differentiation is also reduced (Slatkin 1993; Rousset 2001). The good properties of this method have been confirmed by comparisons of direct and indirect estimates of $D\sigma^2$ (Rousset 2000; Sumner *et al.* 2001).

As for any population genetics method of demographic parameter estimation, the quality of the estimation of $D\sigma^2$ using this method may be affected by local and temporal spatial heterogeneities in demographic parameters. In this study, we adapted the coalescence-based simulation algorithm of Leblois *et al.* (2003) to study the effect of temporal and spatial fluctuations of demographic parameters on the estimation of present-time $D\sigma^2$. Although one can imagine many scenarios dealing with demographic heterogeneities in space and time, we have chosen to focus our study on demographic scenarios often met in empirical surveys in conservation biology and in the study of introduced invading species. In this context, we assessed the effect on the estimation of the present-time $D\sigma^2$ of (i) a temporal change of the dispersal feature, (ii) a density reduction (bottleneck) or increase (flush) in time, (iii) a spatial expansion with constant density, and (iv) a sample of individuals taken from a high-density zone within a lower-density area.

MODELS AND METHODS

Spatial model and population cycle: The model that we considered for "continuous" populations is the lattice model with each lattice node corresponding to one diploid individual. This model without demic structure is viewed as an approximation for truly continuous populations with infinitely strong density regulation (Malécot 1975; Rousset 2000). More realistic continuous models would incorporate the feature that individuals could settle in any position in a continuous space. Although such models have been formulated (*e.g.*, Malécot 1967; Sawyer 1977), it is known that they do not follow a well-defined set of biological assumptions (Maruyama 1972; Felsenstein 1975; see Barton *et al.* 2002 for an alternative approach for continuous populations). To avoid edge effects, a two-dimensional lattice is represented on a torus. Edges and lattice size have little effect on local differentiation when the habitat area (*i.e.*, the lattice size) is large compared to the mean dispersal (Leblois *et al.* 2003). Finally, we considered diploid individuals with dispersal through gametes only. The life cycle is divided into five steps: (i) at each reproductive event, each individual gives birth to a great number of gametes and dies; (ii) gametes undergo the effect of mutations; (iii) gametes disperse; (iv) diploid individuals are formed; and (v) competition brings back the number of adults in each deme to N (usually $N = 1$ but see *Spatial and temporal heterogeneities*). We assume here random assortment of gametes present after dispersal at a given node. This is akin to random selfing in a population of N diploids without spatial structure, by which selfing occurs with frequency $1/N$. How alternative assumptions would affect the analysis is discussed below.

Coalescent algorithm: In this work, we focused on isolation by distance. For this category of models, no analytical treatment of coalescence time or coalescence probabilities has been done for more than two genes. The coalescent algorithm used in this study is thus not based on the large-N approximation of the n-coalescent theory; rather it is an exact algorithm for which coalescence and migration events are considered *generation by generation* until the common ancestor of the sample has been found. The idea of tracing lineages back in time generation by generation is fundamental in the coalescence theory, and is well described in Nordborg (2001). Such a *generation-by-generation* algorithm leads to less efficient simulations in terms of computation time than do those based on the n-coalescent theory (Kingman 1982a,b; Nordborg 2001). However, this algorithm is much more flexible when complex demographic and dispersal features are considered. Note that, since multiple coalescent events are taken into account by considering the probability of a coalescence event of k genes in a given parental node ($= 1/2^{k-1}$ under the model with one individual per lattice node), it allows us to build an exact coalescent tree under very small population size. The entire *generation-by-generation* algorithm that gives the coalescent tree for a sample of n genes evolving under isolation by distance, with density and dispersal homogenous in space and time, is detailed in Leblois *et al.* (2003). The algorithm and the program used in this study were checked at every step during its elaboration by comparing simulated values of probabilities of

identity of two genes under models of isolation by distance on finite lattices with their exact analytically computed values (*e.g.*, Malécot 1975 for the lattice model) with adaptation to different mutation models following general methods valid for any assumption about dispersal and density (Rousset 1996). These comparisons show that estimates of identity probabilities from our program and analytical expectations differ by less than one per thousand for sufficiently long runs.

Dispersal functions: Let (dx, dy) be the parent-offspring axial distance, backward in time, expressed in number of steps on the lattice. Under a two-dimensional model, the probability distribution of the random variable (dx, dy) is given by $b_{dx,dy}$, the "backward" dispersal function. The term backward is used because the position of the parental gene is determined knowing the position of its descendant gene. This function is calculated using $f_{dx,dy}$, the forward dispersal density function describing where descendants go. Biologically realistic dispersal functions often have a high kurtosis (Endler 1977; Kot *et al.* 1996). As previously explained (Rousset 2000), the commonly used discrete probability distributions for dispersal are not appropriate here because high kurtosis can be achieved only by assuming a low dispersal probability, *i.e.*, that most offspring reproduce exactly where their parents reproduced. Thus we used forward dispersal distributions for which the probability of moving k steps (for $0 < k \leq K_{\max}$) in one direction is of the form

$$f_k = f_{-k} = M/k^n, \tag{1}$$

with parameters M and n controlling the total dispersal rate and the kurtosis, respectively. This distribution corresponds to a truncated variant of the discrete Pareto, or ζ, distribution (see, *e.g.*, Patil and Joshi 1968). By suitable choice of the two parameter values, large kurtosis can be obtained with high migration rates (Rousset 2000). For some distributions, the first p terms were arbitrarily fixed:

$$f_1 = f_{-1} = M_1, \quad f_2 = f_{-2} = M_2, \ldots, \quad f_p = f_{-p} = M_p,$$
$$\text{and for } p < k \leq K_{\max}, \quad f_k = f_{-k} = M/k^n. \tag{2}$$

Dispersal was assumed to be independent in each direction, so that $f_{dx,dy} = f_{dx} \times f_{dy}$. When density is homogenous in space, backward dispersal functions are equal to forward dispersal functions, so that $b_{dx,dy} = f_{dx,dy} = f_{dx} \times f_{dy}$.

Mutation processes: The number of mutations on each branch of the coalescent tree follows a binomial distribution with parameter (μ, L), where μ is the mutation rate and L the length of the branch. The allelic states of each gene of the sample were obtained starting from the common ancestor of the sample (root of the genealogical tree) from an allelic state determined according to a probability distribution determined by the mutation model and then going forward in time adding mutations one by one on each branch of the tree. The study of Leblois *et al.* (2003) stressed the interest in using loci with high levels of polymorphism for $D\sigma^2$ estimation. Therefore, microsatellite markers were simulated in the present study. On the basis of direct observations of mutations at human microsatellite loci (Dib *et al.* 1996; Ellegren 2000), the generalized stepwise model (GSM) in which the change in the number of repeat units forms a geometric random variable was adopted (Pritchard *et al.* 1999; Estoup *et al.* 2001). The variance of the geometric distribution was fixed at 0.36 (Estoup *et al.* 2001), a value computed from the mutation data in Dib *et al.* (1996). The mutation rate was equal to 5×10^{-4}, a value considered as the average mutation rate in many species (reviewed in Estoup and Angers 1998). The GSM does not capture all the complexity of the mutation process at microsatellite loci (reviewed in Ellegren 2000; Schlötterer 2000). However, Leblois *et al.* (2003) have shown that exact mutation processes, and in particular the occurrence of constraints on allele size and increase of mutation rate with allele length, have little influence on $D\sigma^2$ estimations.

Method of analysis: Each simulation iteration gives the genotypes at 10 polymorphic loci of 100 (*i.e.*, 10×10) individuals characterized by their coordinates on the lattice. Ten loci and 100 individuals were considered as representative of the number of loci and individuals commonly analyzed in empirical studies based on microsatellites. Independent coalescent trees were used to simulate multilocus genotypes at independent loci. In practice it is difficult to sample all individuals in a small area. Simulations were run for a sample of (10×10) individuals taken every two nodes from an area of (20×20) nodes in the lattice. In this we aimed to roughly mimic a sampling scheme commonly achieved in empirical studies. This process was repeated 1000 times giving 1000 multilocus samples of 100 individuals sharing the same demographic history.

For each simulated multilocus sample, estimates of the parameter $a_r = (Q_w - Q_r)/(1 - Q_w)$ were computed for each pair of individuals, with Q_w the probability of identity in state for two genes taken from the same individual and Q_r the probability of identity in state for two genes at geographical distance r (Rousset 2000). The parameter a_r is a parameter analogous to $F_{ST}/(1 - F_{ST})$ calculated between individuals (not between populations as in Rousset 1997). An estimator of a_r for a pair ξ of individuals taken from the P different possible pairs is

$$\hat{a} \equiv \frac{SS_{b(\xi)}P}{\sum_{k=1}^{P} SS_{w(k)}} - \frac{1}{2}, \tag{3}$$

where $SS_{b[etween](\xi)} \equiv \Sigma_{i,j}(X_{i:u} - X_{i:u})^2$ measures divergence between genes taken from two different individuals and $SS_{w[ithin](\xi)} \equiv \Sigma_{i,j,u}(X_{ij:u} - X_{i.:u})^2$ measures divergence between genes within the same individual ($X_{ij.u}$ is an indica-

tor variable taking the value 1 if gene i of individual j is of allelic type u and the value 0 otherwise; Rousset 2000). Thus, $\hat{a}$ compares the genetic divergence of individuals at distance r (numerator) to the divergence of the two-gene copy within the individual (denominator), which is essentially what the parameter a_r does. Because stepwise mutations occur at microsatellite loci, a statistic taking into account the allele size might appear to be attractive. However, Leblois *et al.* (2003) have shown that incorporation of allele size into the estimate of a_r gives unreliable results due to the high variance of the estimates. Therefore, only the parameter a_r described in Equation 3 was used in this study.

The generalized random selfing assumption made in this article implies that the identity within individuals is identical to the identity between juveniles competing for a site. More generally, $D\sigma^2$ is related to the parameter

$$\frac{\rho_r}{1-\rho_r} = \frac{Q_0 - Q_r}{((1-Q_w)/2) - Q_0}, \quad (4)$$

where Q_w is the probability of identity of genes within individuals, Q_r is the probability of identity of two genes in different individuals at distance r, and Q_0 is the probability of identity of two genes in different individuals in the same node (Rousset 2004, Equation 8.12). Without random selfing, $\hat{a}_r$ is not the most relevant statistic. Rather one should estimate not only Q_w but also Q_0. Since there is only one adult per node of the lattice, Q_0 cannot be estimated directly from adults: it must be approximated as the identity between close adults or (better) between close juveniles before competition (see Rousset 2004, Chap. 8, for further discussion). In this way, it is easy to adapt the methods considered in this article, but this is not considered further.

For each simulated data set, the value of the slope of the regression line between $\hat{a}$ and the logarithm of geographical distance was computed. In the limit of low mutation rates, the inverse of the slope is an estimate of the product $4\pi D\sigma^2$ (Rousset 1997). High mutation rates should not result in a large sample bias as long as one focuses on local processes involving distances between sampled individuals, $r \ll \sigma/\sqrt{2\mu}$. Beyond this limit, the linear relationship between a_r and the logarithm of the distance holds less well (see Rousset 1997 for theoretical details). Thus, if the analysis is done on a small geographical scale, the use of loci with high mutation rates such as microsatellites does not bias the estimation. This is illustrated by Leblois *et al.* (2003), using simulations.

The quality of an estimator is usually assessed through the computation of its bias and its mean square error (MSE). These measures are suitable when estimates have an approximately normal distribution but not when estimates are sometimes infinite. In the present case, a negative slope should be interpreted as an infinite estimate of $D\sigma^2$. Therefore, we present the bias and the MSE for the slope values of the regression lines and not for $D\sigma^2$ estimates. Thus, the following statistics were estimated over all repetitions: (i) the mean relative bias between the value of the slope and the expected value, $1/(4\pi D\sigma^2)$ [*i.e.*, (observed slope − expected slope)/expected slope]; (ii) the standard error on this relative bias; and (iii) the mean square error [*i.e.*, MSE = ((observed slope − expected slope)/expected slope)^2]. The bias and the MSE are relative values since they are computed from the ratio of the observed to the expected value. We also computed the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

Spatial and temporal heterogeneities: One important advantage of the generation-by-generation algorithm is that virtually any demographic model including those with variations in time and space of demographic parameters can be easily implemented.

Temporal change in dispersal: We first studied the effect of a simple decrease of dispersal capabilities in time. Decrease in dispersal under isolation-by-distance models can be modeled in various ways (*i.e.*, changing various parameters in the dispersal distributions). Here we considered a decrease over time of the average squared axial parent-offspring distance (σ^2). Two different dispersal distributions with different σ^2 values were used, while all other parameters of the distribution (*i.e.*, the global shape of the distribution) remained unchanged. This situation corresponds to a change in a landscape (*e.g.*, a fragmentation) resulting in modifying the ability of a species to move within this landscape (*e.g.*, Brooker and Brooker 2002). Simulations were run with a two-dimensional lattice of (500 × 500) nodes with one individual per node. A first dispersal distribution, given in expression (2) with parameters

$$M = 0.555 \text{ and } n = 2.744 \quad \text{for } 0 < k \leq 48, \quad (5)$$

has a moderate σ^2 value ($\sigma^2 = 4$ in lattice units) and is the dispersal distribution from the present until the time of change, G_c. A second dispersal distribution, with parameters $M = 0.187$ and $n = 1.246$ for $0 < k \leq 48$ corresponds to a very high σ^2 value ($\sigma^2 = 100$) and is the dispersal distribution from the time of change G_c until the time of the most recent common ancestor (TMRCA). Four simulations were run with $G_c = 10$, $G_c = 20$, $G_c = 100$ generations (going backward in time), and G_c infinite as baseline (*i.e.*, no change in dispersal features over time).

Temporal change in density: A second category of fluctuations is temporal variations in density of individuals. We studied two simple situations: (i) a decrease in density from past to present (population bottleneck) and (ii) an increase in density from past to present (population flush). Such bottleneck or flush events are expected to occur in endangered or invasive populations, respectively. These situations were implemented in our simula-

TABLE 1

Models used to study the effects of density variation in time on the estimation of $1/(4\pi D\sigma^2)$

Demographic change	Density (no. of individuals per lattice node): From sampling time to G_c	Density (no. of individuals per lattice node): From G_c to the TMRCA	Factor
Bottleneck			
Weak decrease	1	10	10
Strong decrease	1/9	10	90
Flush			
Weak increase	1	1/9	9
Strong increase	1	1/100	100

The number of generations, G_c, indicates the moment in the past when the density variation occurred. TMRCA corresponds to the time of the most recent common ancestor of the sampled genes.

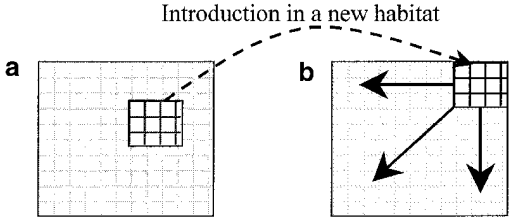


FIGURE 1.—Schema of a demographic expansion with constant density as modeled in this study. (a) The source population from which a subpopulation (dark gray grid) is introduced in an empty habitat (dotted arrow). (b) The empty habitat on which the introduced population spreads within a few generations (solid arrows). In our simulations, two-dimensional habitats are represented on a torus and not on a plane square as in this figure.

tions by changing the number of individuals per lattice node over time. Four different lattice models were used: one with 1 individual per node, one with 10 individuals per node, one with 1 individual every 3 nodes in each direction, and one with 1 individual every 10 nodes in each direction. These models correspond to densities of 1, 10, 1/9, and 1/100, respectively. Having less than 1 individual per node avoids the consideration of models with a too high number of individuals per node (*i.e.* >10) before or after a change in density, which would strongly deviate from the concept of continuous population to which the method of estimation applies. For easier coding, we modeled densities lower than 1 individual per node, considering that a given proportion of nodes of the lattice are always "empty" (*e.g.*, for a density of 1/9, 8/9 of the nodes are empty). This is equivalent to a model with a larger lattice unit (*e.g.*, a lattice unit three times larger in each dimension for a density of 1/9 compared to the lattice unit for a density of 1). A summary of the different density changes studied is presented in Table 1.

For the model with 1 individual every 9 nodes, we adapted the dispersal distribution to keep a constant $\sigma^2 = 4$. Since dispersal may occur only between "non-empty" nodes, the dispersal distribution parameters are then $M = 0.299$ and $n = 4.159$ for $0 < k \leq 48$. For the model with 1/100, 1, or 10 individuals per node, the dispersal distribution parameters are those used in the previous section [*cf.* expression (5)]. We have not adapted the dispersal distribution to keep a constant $\sigma^2 = 4$ for the model with 1 individual every 100 nodes because it was mathematically impossible to adjust this distribution with a too small number of points in the distribution (*i.e.*, in this case, there are only five possible moves in each direction between "suitable" nodes, which are located at 0, 10, 20, 30, and 40 lattice units). However, additional simulations with a 90-fold density increase (from 1/9 to 10 individuals per node) and a dispersal distribution adapted to keep a constant σ^2 gave similar results (results not shown).

For each case of density change considered, four simulations were run, using a two-dimensional habitat of (500×500) nodes with $G_c = 10$, $G_c = 20$, $G_c = 100$ generations, and G_c infinite as baseline. For each bottleneck and flush case, we simulated a weak density variation (10 and 9 times density change, respectively) and a strong density variation (90 and 100 times density change, respectively). In the case of bottleneck, the low-density models (1 and 1/9 individuals per node for weak and strong variations, respectively) were implemented from sampling time to G_c and the high-density models (10 individuals per node) from G_c to the TMRCA. In the case of density flush, the high-density models (1 individual per node) were implemented from sampling time to G_c and the low-density models (1/9 and 1/100 individuals per node for weak and strong variations, respectively) from G_c to the TMRCA (Table 1).

Spatial expansion with constant density: The third type of studied situation is a population expansion in space with constant density of individuals (Figure 1). The population introduced into an empty habitat is composed of individuals that have evolved in a source population at equilibrium with some demographic features (*i.e.*, density and dispersal distribution). The introduced population spreads within a few generations on an empty two-dimensional habitat with the same demographic features as the source population. This situation corresponds to the case of an introduced species that colonizes a new territory with similar ecological features to that of its native territory. Before expansion (*i.e.*, at generation G_c), the introduced population is composed of 100 individuals located on a (10×10) area, which were sampled from a (10×10) area in the source population, which itself evolved on a (160×160) lattice. From generation G_c to present, the introduced population spreads over a lattice of (160×160) nodes. The

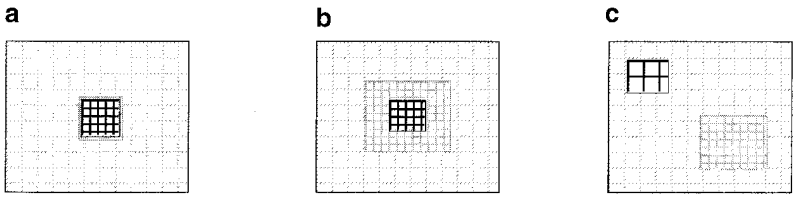


FIGURE 2.—Schema of the spatial density heterogeneities as modeled in this study. (a) A small high-density zone (dark gray grid) strictly corresponds to the sampling area (black grid) on a two-dimensional habitat with a lower density (light gray grid). (b) A large high-density zone (dark gray grid) includes the sampling area (black-grid) on a two-dimensional habitat with a lower density (light gray grid). (c) A large high-density zone (dark gray grid) is present on a two-dimensional habitat with a lower density (light gray grid); the sampling area (black grid) is located outside the high-density zone. In our simulations, two-dimensional habitats are represented on a torus and not on a plane square as in this figure.

entire (160 × 160) matrix is potentially occupied in two generations. At sampling time, as in the previous sections, 100 individuals were taken from an area of (20 × 20) nodes located outside the area of introduction, the distance between the introduction area and the sampling area being equal to 50 nodes. The forward dispersal distribution parameters are those given in expression (5) and correspond to a $\sigma^2 = 4$. Four simulations were run with $G_c = 10$, $G_c = 20$, $G_c = 100$, and G_c infinite as baseline.

Spatial density heterogeneities: The situations we choose to study reflect the fact that biologists usually collect individual samples in localities where they are easy to collect, that is, in high-density areas. Hence, we considered a lattice model with homogenous density except on a squared area where the density of individuals is higher (Figure 2). In such models with density heterogeneities in space, backward and forward dispersal differ. Each lattice node has a backward distribution that depends on the density of each surrounding node (*e.g.*, each node being at a distance less or equal to the $K_{\max}$ step). Those surrounding nodes correspond to all locations from which genes could have come in one generation (forward in time). Since those nodes are occupied by different numbers of individuals and because nodes occupied by more individuals contribute potentially more to the number of immigrants that reach a given node, we have to weight each term of the backward dispersal distribution by the number of individuals of the node from where immigrants have come. Let $N_{x,y,G}$ be the number of individuals at node (x, y) at generation G. Then for any node (x, y) the probability $b_{dx,dy}$ for a gene to move backward dx steps in one direction and dy in the other is equal to

$$b_{dx,dy} = \frac{N_{(x+dx),(y+dy),G} \cdot f_{dx,dy}}{\sum_{dx,dy \le K_{\max}} N_{(x+dx),(y+dy),G} \cdot f_{dx,dy}}. \quad (6)$$

Simulations were run for a sample of 100 individuals taken every two nodes from an area of (20 × 20) nodes evolving in a (160 × 160) lattice. Density is one individual per node, except on a $(n \times n)$ zone including the sample area where density is 10 individuals per node. Two cases were considered: (i) a small high-density zone of (20 × 20) nodes, which strictly corresponds to the sample area (Figure 2a), and (ii) a larger high-density zone of (40 × 40) nodes, which includes the (20 × 20) nodes sample area (Figure 2b). We were particularly interested in assessing whether the estimated density corresponds to the density on the sampling area (*i.e.*, the local density) or whether the estimation is influenced largely by the density surrounding the sampling area (*i.e.*, the neighboring density). This was performed by alternatively considering that the expected $D\sigma^2$ value corresponded to a density of 10 (local density) and 1 (surrounding density) individuals per node. An additional simulation was run with a single large high-density zone of (40 × 40) nodes located outside the sampling area, the distance between the high-density and sampling zones being equal to 50 nodes (Figure 2c).

RESULTS

Interpretation of observed bias: Observed bias in our simulations might be attributable to (i) a bias, inherent to the method, due to the effect of a high mutation rate on the parameter value (this we call "mutational bias"), (ii) a bias due to the deviation of the estimates relative to the parameter value considering a finite sample of individuals and loci (this we name "small sample bias"), and (iii) a bias introduced by the demographic fluctuations studied. Additional details on the small sample and mutational biases can be found in LEBLOIS *et al.* (2003). All results in the present study should be interpreted taking into account the small sample and mutational biases that can be observed in the simulations without demographic fluctuations that were included in all situations studied as baseline (G_c infinite). For example, in the case of a reduction of density (bottleneck, Table 3), the mutational and small sample bias is large when considering an intermediate-density model (baseline simulation for a weak reduction) and much lower when considering a low-density model (baseline simulation for a stronger reduction). This difference is due partly to the different densities of individuals in the two baseline simulations, which influence the global level of genetic diversity in the sample. LEBLOIS *et al.* (2003) indeed showed that differences in genetic diversity have a substantial effect on the estimation of $D\sigma^2$.

TABLE 2

Effect of a temporal reduction of dispersal on the estimation of $1/(4\pi D\sigma^2)$

G_c	Infinite	100	20	10
Bias (standard error)	0.444 (0.0062)	0.0923 (0.0081)	−0.0795 (0.0076)	−0.234 (0.0074)
MSE	0.228	0.0743	0.0642	0.109
2× coverage	0.995	0.989	0.965	0.876

The number of generations, G_c, indicates the moment in the past when the dispersal reduction occurred. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

Temporal change in dispersal: Simulation results show that the bias due to a reduction of dispersal is negative (Table 2) and thus corresponds to an overestimation of the present time $D\sigma^2$. This result is in agreement with a transition from a high $D\sigma^2$ value ($\sigma^2 = 100$) during the past generations (*i.e.*, before G_c) to a much lower value after G_c ($\sigma^2 = 4$). In other words, the method of $D\sigma^2$ estimation has a memory of temporal changes in dispersal. However, this memory is short term since a reduction of dispersal 100 generations ago gave only a slight negative bias compensated by the positive small sample and mutational biases (*cf.* first column of Table 2). Moreover, even for a recent reduction of dispersal ($G_c = 10$), the bias is <25% (*i.e.*, <0.25), a relatively low value compared to the high amplitude of the dispersal change. Standard error of the estimation also remains low for all G_c values, and for changes older than 20 generations, >95% of the estimations are within a factor of two of the present-time $D\sigma^2$. Hence, our simulations generally show that the precision of the present-time $D\sigma^2$ estimation is relatively robust to temporal changes in dispersal.

Temporal reduction of density (bottleneck): The negative bias observed in Table 3 (*i.e.*, overestimation of $D\sigma^2$) reflects the higher population density from generation G_c until the TMRCA. For a 10 times reduction of density, the method is quite robust when the density change occurred 20 or more generations ago. The bias and the MSE are low (<10%) and almost 99% of the estimations are within a factor of two of the present-time $D\sigma^2$ value. For very recent density change (*e.g.*, $G_c = 10$) the bias is substantial. However, the MSE remains low and >90% of the estimations are still within a factor of two of the present-time $D\sigma^2$ value.

The effect of reduction of density is more marked for a stronger change in density (*i.e.*, 90 times density reduction). For a very recent density reduction (*i.e.*, 10 generations ago), the negative bias reaches 50% and only 24% of the estimations are within a factor of two of the present-time $D\sigma^2$ value. For $G_c = 100$, the bias and the MSE become similar to the baseline. Note that all estimations are within a factor of two of the present-time $D\sigma^2$ for $G_c \geq 20$. Therefore, even for large recent density reductions, the method appears to be relatively robust.

Temporal increase in density (demographic flush): The positive bias observed in Table 4, which corresponds to an underestimation of the present-time $D\sigma^2$, reflects the lower population density from generation G_c until the TMRCA. For a small increase in density (10

TABLE 3

Effect of a weak (10 times density reduction) and strong (90 times density reduction) bottleneck on the estimation of $1/(4\pi D\sigma^2)$

Intensity	G_c	Infinite	100	20	10
Weak	Bias (standard error)	0.444 (0.0062)	0.0990 (0.0070)	−0.0625 (0.0064)	−0.222 (0.0061)
	MSE	0.228	0.0588	0.0449	0.0868
	2× coverage	0.995	0.997	0.989	0.915
Strong	Bias (standard error)	−0.0138 (0.0042)	−0.0743 (0.0027)	−0.330 (0.0017)	−0.526 (0.0012)
	MSE	0.0175	0.0128	0.115	0.278
	2× coverage	1	1	1	0.238

The number of generations, G_c, indicates the moment in the past when the density reduction occurred. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

TABLE 4

Effect of a weak (9 times density increase) and strong (100 times density increase) density flush on the estimation of $1/(4\pi D\sigma^2)$

Intensity	G_c	Infinite	100	20	10
Weak	Bias (standard error)	0.444 (0.0062)	0.315 (0.040)	0.685 (0.043)	1.4 (0.046)
	MSE	0.228	1.72	2.33	4.07
	2× coverage	0.995	0.45	0.381	0.238
Strong	Bias (standard error)	0.432 (0.00644)	0.648 (0.0094)	2.24 (0.015)	3.91 (0.0193)
	MSE	0.228	0.508	5.27	15.8
	2× coverage	0.999	0.89	0.00262	0

The number of generations, G_c, indicates the moment in the past when the density increase occurred. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

times), the bias and the MSE are high even for a relatively ancient flush (*e.g.*, $G_c = 100$). The proportion of estimations being within a factor of two of $D\sigma^2$ remains small (<50%) even for $G_c = 100$. The effect of the flush also increases substantially with the intensity of the density change. For a 100-fold density change and for $G_c = 10$, the bias reaches 391% and none of the estimations are within a factor of two of $D\sigma^2$ (Table 4). Hence, although the bias and the MSE decrease when G_c increases, the estimation remains unreliable for both 100- and 10-fold density change. These results contrast sharply with those pertaining to bottlenecks and dispersal changes.

Spatial increase in population size with constant density (demographic expansion): All measures (bias, MSE, and proportion of estimates within a factor of two) indicate that the estimation of the present-time $D\sigma^2$ is good when the spatial expansion occurred 20 or more generations ago (Table 5). For $G_c = 10$ only, an 8% negative bias is observed, which corresponds to an overestimation of the present-time $D\sigma^2$ (Table 5). However, the MSE is very small (10%) and 97% of the estimations are within a factor of two of the expected $D\sigma^2$ value. Hence, a spatial expansion as modeled here has only a short-term and limited influence on the present-time $D\sigma^2$ estimation; the method is precise even for very recent expansions.

Spatial heterogeneity in density (sampling within a high-density zone): Table 6 shows that $D\sigma^2$ estimation is not robust when the high-density zone is small and strictly corresponds to the sampling area. The bias and MSE values indicate that in this case the low-density area surrounding the sampling area strongly influences the $D\sigma^2$ estimation, which becomes a bad measure of both local density (*i.e.*, the density on the sampling area) and surrounding density (*i.e.*, the density surrounding the sampling area). It can be seen, however, that two times coverage probabilities, although globally low, are higher when referring to the local rather than to the surrounding area density as expected ($D\sigma^2$ value 0.018 *vs.* 0.001). This suggests that there is a tendency for the method to measure the local rather than the surrounding density. This trend becomes obvious when looking at results for a larger high-density zone (Table

TABLE 5

Effect of a spatial expansion

G_c	Infinite	100	20	10
Bias (standard error)	0.430 (0.0076)	0.387 (0.0126)	0.133 (0.0111)	−0.0824 (0.0101)
MSE	0.243	0.23	0.08	0.0581
2× coverage	0.989	0.98	0.996	0.972

The number of generations, G_c, indicates the moment in the past when the spatial expansion occurred. The expansion occurred without density and dispersal changes. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

TABLE 6

Effect of spatial heterogeneities in density

	Local density		Surrounding density	
Spatial heterogeneity	Estimation	Control	Estimation	Control
Small high-density zone				
Bias (standard error)	2.11 (0.017)	0.45 (0.025)	−0.689 (0.0017)	0.430 (0.0076)
MSE	4.76	0.83	0.477	0.243
2× coverage	0.018	0.65	0.001	0.989
Large high-density zone				
Bias (standard error)	0.393 (0.013)	0.45 (0.025)	−0.861 (0.0013)	0.43 (0.0076)
MSE	0.330	0.83	0.743	0.243
2× coverage	0.9	0.65	0	0.989
Large high-density zone outside sampling area				
Bias (standard error)	0.447 (0.00752)	0.43 (0.0076)	13.5 (0.0752)	0.45 (0.025)
MSE	0.256	0.243	187	0.83
2× coverage	0.99	0.989	0	0.65

Sampling was done on a small or large high-density zone of (20 × 20) and (40 × 40) nodes, respectively. Local density, the expected density is the local density (*i.e.*, density in the sampling area); surrounding density, the expected density is the surrounding density (*i.e.*, around the sampling area). Controls correspond to a homogenous lattice with density being the local or the surrounding density for the local and surrounding estimation cases, respectively. Bias is the mean of relative bias of each run [(observed slope − expected slope)/expected slope]; MSE is the mean of the square error of each run [((observed slope − expected slope)/expected slope)^2]; 2× coverage corresponds to the probability that the estimate of $1/(4\pi D\sigma^2)$ was within a factor of two from the expected value (*i.e.*, in the interval [expected slope/2; 2 × expected slope]).

6). In this case, the bias and the MSE are much lower when considering the local rather than the surrounding zone for the $D\sigma^2$ value. About 90% of the estimates are within a factor of two of the local $D\sigma^2$ value, while none of them are within a factor of two of the surrounding $D\sigma^2$ value. The third case of a large high-density zone located outside the sampling area (*i.e.*, 50 nodes away) confirms this result (Table 6). Hence, our simulations generally show that the method estimates local demographic parameters and is robust for such measurement when the density is relatively homogenous around the sampling area (*e.g.*, over an area equal to four times the sampling area).

DISCUSSION

This work is the first one focusing on the study of evolutionary disequilibrium situations in the complex but realistic population model of a continuous population evolving under isolation by distance. Within the limits of the situations studied in this article, and with the exception of the case of a density flush, we found that temporal and spatial fluctuations of demographic parameters, if not too strong and not too recent (*i.e.*, more than, say, 20–50 generation in the past), have a limited influence on the estimation of local and present-time demographic parameters with the method of Rousset (2000). It is worth noting that we are talking about changes on timescales of a few tens of generations in the past, which may be very recent by standards in population genetics, but not for lots of species undergoing demographic changes due to ongoing human impact. Moreover, the numbers of generations defining the time of demographic change in this study should be considered as indicative of only the length of the effect of the demographic changes studied rather than as absolute reference numbers. As a matter of fact, the persistence in time of the effect of demographic fluctuations strongly depends on various features of the demographic model (*e.g.*, σ^2 values) and disequilibrium situations. It is thus preferable to consider general trends rather than precise numbers for each situation. For clarity, those trends have been summarized in Table 7.

The robustness of the method of Rousset (2000) to several temporal and spatial demographic fluctuations somewhat contradicts previous studies dealing with the study of evolutionary disequilibrium. In their review, Koenig *et al.* (1996) concluded that estimations of dispersal parameters from genetic data give ideas about past rather than present dispersal and gene flow, so that direct methods, such as mark-recapture methods, should give a better estimation of actual dispersal parameters. Boileau *et al.* (1992) similarly showed that hundreds or thousands of generations are required to erase the effects of colonization processes on "F_{ST}-like estimates" from allozyme data in large populations, con-

TABLE 7

Qualitative summary of the effects of different temporal and spatial heterogeneities

		Effect on $D\sigma^2$ estimation			
		Bias			
Demographic change		Sign	Intensity	2× coverage	Duration
Temporal	Dispersal increase (25 times)	Positive	Medium	Good	Short
	Density decrease (10–90 times)	Positive	Low to medium	Good to poor	Short
	Density increase (9–100 times)	Negative	High	Poor	Medium
Spatial	Local high-density zone (10 times)	Negative	Low (local) to high (surrounding)	Good (local) to poor (surrounding)	NA
Temporal and spatial	Spatial expansion	Negative	Low	Good	Short

Low intensity, mean relative bias <50%; high intensity, mean relative bias >100%; good, 2× coverage >85%; poor, 2× coverage <85%; short duration, few (10–20) generations; medium duration, >100 generations; NA, not appropriate.

cluding that estimates of gene flow from genetic data should be taken with care. We fully agree that temporal demographic fluctuations in a population are likely to have a strong and persistent effect on some population genetics statistics and methods. However, the present study shows that some indirect methods and genetic markers give accurate estimations of present-time density and dispersal features even when the demographic history includes relatively recent demographic changes.

The general robustness to spatial and temporal heterogeneities of the present *F*-statistic-based method can be interpreted using arguments from the coalescence theory and analytical treatment available in this field. Values of *F*-statistics, under the assumption of low mutation rate, can be deduced by comparing the distributions of coalescence probability for different pairs of genes (*e.g.*, pairs from the same deme and pairs from different demes; *e.g.*, Rousset 2002). These distributions differ essentially by an excess of coalescence probability for the most related genes, this excess being concentrated in a brief period τ in the recent past. *F*-statistics thus depend mainly on differences between the distributions of coalescence probability for different pairs of genes in recent generations. As the sensitivity of *F*-statistics values to past demographic fluctuations is also related to this recent time period, past demographic fluctuations have less effect when the time period τ is short. This recent time period τ is shorter when high dispersal rates and/or low deme size are considered (Rousset 2004). Hence, if models with small deme size and high migration rates, such as isolation by distance between individuals where each deme is of size two genes, are considered the influence of past demographic fluctuations on the estimation of demographic parameters from *F*-statistics is limited. By contrast, under the classical island model with large deme size and low migration rates, the effect of past demographic fluctuations is expected to be more problematic. Moreover, under isolation-by-distance models, the more distant the demes are on the lattice, the more the period τ is expanding to the past, increasing the effect of past demographic parameter fluctuations (Slatkin 1994; Rousset 2004). Because the present method focuses on local differentiation and thus on recent evolutionary processes corresponding to a narrow recent past zone, it is again logical that past demographic fluctuations have limited effects on the estimation of the present-time and local $D\sigma^2$ with this method. The same reasoning can be used to understand why the method gives estimates of the local demographic parameter values rather than estimates of the surrounding demographic parameter values. As the period τ is short in the models considered, *F*-statistics depend mainly on genetic events (migration, coalescence, mutation) that occurred in a recent past and, because dispersal is localized, at a local geographical scale. Therefore, the estimate of $D\sigma^2$ by the present method should correspond to the local demographic parameter values on the sampling area and should not be much influenced by demographic features of zones that are far away from the sampling area.

Close examination of our results brings up several issues. Our simulations showed that, for the study of invading species, the present method should give precise estimates of the present-time $D\sigma^2$ provided that no demographic flush occurred during the expansion process. This is an interesting feature of the method, which makes it appropriate to study invasive organisms for which demographic features are similar in the newly founded population and in the original source population. Our simulations further showed that if a change in dispersal occurred during the invasion process, this new dispersal feature should translate quickly in the estimation of the present-time $D\sigma^2$. On the other hand, density flushes (and to a much lower extent population

bottlenecks) may strongly affect present-time $D\sigma^2$ estimation. Invading species populations often experience complex demographic fluctuations that may include both bottlenecks (*i.e.*, founder events) and density flushes during their spreading (*e.g.*, Williamson 1996; Estoup *et al.* 2001). Therefore, it seems necessary to run additional simulations adapted to those complex demographic scenarios to thoroughly evaluate the robustness of the estimation of the present-time $D\sigma^2$.

Our simulations also show that for conservation biology studies dealing with bottlenecked populations the estimation of $D\sigma^2$ is potentially biased toward past demographic parameter values. However, the memory of past demographic parameter values is short so that this bias is important for only a strong and recent decrease in density. A major genetic consequence of a population bottleneck is that the number of alleles decreases much faster than the heterozygosity (Nei *et al.* 1975; Luikart and Cornuet 1998). One might have expected the precision of the method to be reduced due to the lower number of alleles in the bottlenecked population. However, standard error on the bias was weak whatever the strength of the bottleneck. One possible explanation for this result is that the method's precision depends more on the mean heterozygosity level than on the average number of alleles.

Our simulations indicate that surrounding densities considerably influence the estimation of local $D\sigma^2$ when the sample is taken on a small high-density zone. In this case, the estimates correspond neither to the $D\sigma^2$ values on the sampling area nor to the surrounding $D\sigma^2$ values. However, if sampling is done in a sufficiently large high-density zone (*e.g.*, on a surface equals to four times the sampling area), the estimates correspond more to the local density (*i.e.*, the density in the sampling area). Our simulations allowed us to study the case of a high-density zone in the middle of a large homogenous zone with low density. This situation is realistic for various demographic systems and mimics a classical experimental bias (*i.e.*, the fact that biologists generally collect their samples in high-density areas). However, many biological situations with spatial density heterogeneities would correspond rather to random density fluctuations on each lattice node. It is expected that differentiation in such scenarios will be a function of some "effective" density and dispersal rate. The lack of analytical formulas for these effective parameters limits the interpretation of a simulation study of the performance of estimators. Nevertheless, there is no obvious reason to believe that the estimation of the effective $D\sigma^2$ would be affected more by such random fluctuations than by previously studied spatial heterogeneities.

We thank Thomas Lenormand and Franck Shaw for constructive comments on the manuscript. This work was financially supported by the Action Incitative Programmée no. 00202 "biodiversité" from the Institut Français de la Biodiversité and grant no. D4E/SRP/01118 "biological invasion" from the Ministère de l'Ecologie et du Développement Durable. This is paper ISEM 2004-007.

LITERATURE CITED

Barton, N. H., F. Depaulis and A. M. Etheridge, 2002 Neutral evolution in spatially continuous populations. Theor. Popul. Biol. **61:** 31–48.

Beebee, T., and G. Rowe, 2001 Application of genetic bottleneck testing to the investigation of amphibian declines: a case study with natterjack toads. Conserv. Biol. **15:** 266–270.

Boileau, M. G., P. D. N. Hebert and S. S. Schwartz, 1992 Non-equilibrium gene frequency divergence: persistent founder effects in natural populations. J. Evol. Biol. **5:** 25–39.

Brooker, L., and M. Brooker, 2002 Dispersal and population dynamics of the blue-breasted fairy-wren, *Malurus pulcherrimus*, in fragmented habitat in the Western Australian wheatbelt. Wildlife Res. **29:** 225–233.

Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Ellegren, H., 2000 Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. **24:** 400–402.

Endler, J. A., 1977 Geographical variation, speciation, and clines. Princeton University Press, Princeton, NJ.

Estoup, A., and B. Angers, 1998 Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations, pp. 55–86 in *Advances in Molecular Ecology NATO ASI Series*, edited by G. Carvalho. IOS Press, Amsterdam.

Estoup, A., I. J. Wilson, C. Sullivan, J.-M. Cornuet and C. Moritz, 2001 Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. Genetics **159:** 1671–1687.

Felsenstein, J., 1975 A pain in the torus: some difficulties with models of isolation by distance. Am. Nat. **109:** 359–368.

Hastings, A., and S. Harrison, 1994 Metapopulation dynamics and genetics. Annu. Rev. Ecol. Syst. **25:** 167–188.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. **13:** 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Koenig, W. D., D. Van Vuren and P. N. Hooge, 1996 Detectability, philopatry, and the distribution of dispersal distances in vertebrates. Trends Ecol. Evol. **11:** 514–517.

Kot, M., M. A. Lewis and P. van den Driessche, 1996 Dispersal data and the spread of invading organisms. Ecology **77:** 2027–2042.

Leblois, R., A. Estoup and F. Rousset, 2003 Influence of mutational and sampling factors on the estimation of demographic parameters in a 'continuous' population under isolation by distance. Mol. Biol. Evol. **20:** 491–502.

Luikart, G., and J.-M. Cornuet, 1998 Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. Conserv. Biol. **12:** 228–237.

Malécot, G., 1967 Identical loci and relationship, pp. 317–332 in *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and Probability*, Vol. 4, edited by L. M. Lecam and J. Neyman. University of California Press, Berkeley, CA.

Malécot, G., 1975 Heterozygosity and relationship in regularly subdivided populations. Theor. Popul. Biol. **8:** 212–241.

Maruyama, T., 1972 Rate of decrease of genetic variability in a two-dimensional continuous population of finite size. Genetics **70:** 639–651.

Nei, M., T. Maruyama and R. Chakraborty, 1975 The bottleneck effect and genetic variability in populations. Evolution **29:** 1–10.

Nordborg, M., 2001 Coalescent theory, pp. 179–208 in *Handbook of Statistical Genetics*, edited by D. A. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Patil, G. P., and S. W. Joshi, 1968 *A Dictionary and Bibliography of Discrete Distribution.* Oliver & Boyd, Edinburgh.

Portnoy, S., and M. F. Willson, 1993 Seed dispersal curves: behavior of the tail of the distribution. Evol. Ecol. **7:** 25–44.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun and M. W. Feldman, 1999 Population growth of human Y chromosome microsatellites. Mol. Biol. Evol. **16:** 1791–1798.

Rousset, F., 1996 Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics **142:** 1357–1362.

Rousset, F., 1997 Genetic differentiation and estimation of gene

flow from *F*-statistics under isolation by distance. Genetics **145:** 1219–1228.

Rousset, F., 2000 Genetic differentiation between individuals. J. Evol. Biol. **13:** 58–62.

Rousset, F., 2001 Inferences from spatial population genetics, pp. 239–265 in *Handbook of Statistical Genetics*, edited by D. A. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Rousset, F., 2002 Inbreeding and relatedness coefficients: What do they measure? Heredity **88:** 371–380.

Rousset, F., 2004 *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Sawyer, S., 1977 Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv. Appl. Probab. **9:** 268–282.

Schlötterer, C., 2000 Evolutionary dynamics of microsatellite DNA. Chromosoma **109:** 365–371.

Slatkin, M., 1993 Isolation by distance in equilibrium and non-equilibrium populations. Evolution **47:** 264–279.

Slatkin, M., 1994 Gene flow and population structure, pp. 3–17 in *Ecological Genetics*, edited by L. A. Real. Princeton University Press, Princeton, NJ.

Spong, G., and L. Hellborg, 2002 A near-extinction event in lynx: Do microsatellite data tell the tale? Conserv. Ecol. **6** (1): 15.

Stone, G. N., and P. Sunnucks, 1993 Genetic consequences of an invasion through a patchy environment: the cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). Mol. Ecol. **2:** 251–268.

Sumner, J., F. Rousset, A. Estoup and C. Moritz, 2001 'Neighborhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. Mol. Ecol. **10:** 1917–1927.

Williamson, M., 1996 *Biological Invasions*. Chapman & Hall, London.

Whitlock, M. C., and D. E. McCauley, 1999 Indirect measure of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. Heredity **82:** 117–125.

Communicating editor: L. Excoffier

Likelihood and Approximate Likelihood Analyses of Genetic Structure in a Linear Habitat: Performance and Robustness to Model Mis-Specification

François Rousset and Raphaël Leblois†*

*Université, Montpellier 2, CNRS, Institut des Sciences de l'Évolution, France; and †Unité Origine, Structure et Évolution de la Biodiversité, Museum National d'Histoire Naturelle, Paris, France

We evaluate the performance of maximum likelihood (ML) analysis of allele frequency data in a linear array of populations. The parameters are a mutation rate and either the dispersal rate in a stepping stone model or a dispersal rate and a scale parameter in a geometric dispersal model. An approximate procedure known as maximum product of approximate conditional (PAC) likelihood is found to perform as well as ML. Mis-specification biases may occur because the importance sampling algorithm is formally defined in term of mutation and migration rates scaled by the total size of the population, and this size may differ widely in the statistical model and in reality. As could be expected, ML generally performs well when the statistical model is correctly specified. Otherwise, mutation rate estimates are much closer to mutation probability scaled by number of demes in the statistical model than scaled by number of demes in reality when mutation probability is high and dispersal is most limited. This mis-specification bias actually has practical benefits. However, opposite results are found in opposite conditions. Migration rate estimates show roughly similar trends, but they may not always be easily interpreted as low-bias estimates of dispersal rate under any scaling. Estimation of the dispersal scale parameter is also affected by mis-specification of the number of demes, and the different biases compensate each other in such a way that good estimation of the so-called neighborhood size (or more precisely the product of population density and mean-squared parent–offspring dispersal distance) is achieved. Results congruent with these findings are found in an application to a damselfly data set.

Despite lasting efforts, estimating dispersal rates from genetic data remain a challenging problem. Many uncertainties remain about the various complicating factors that may invalidate inferences. It is not clear how many parameters can be estimated accurately and whether the results will be robust to various factors such as the mode of evolution of the markers, ancestral history of the species, and populations unaccounted for in the statistical model (e.g., Slatkin 1994; Arbogast et al. 2002; Rousset 2007 for reviews).

Nevertheless, in spatially subdivided populations, some statistical patterns depend mainly on the recent history of the population. This makes it possible to develop statistical methods that specifically exploit these patterns, and therefore could be robust to various uncontrolled factors (e.g., Slatkin 1993, 1994). For example, previous works on moment-based methods (i.e., methods based on Wright's F_{ST} and similar measures) have shown that reliable estimation of some dispersal parameters is possible under isolation by distance because such estimation may be based on genetic patterns independent of unsampled populations, of mutation models, and robust to past demographic fluctuations (Slatkin 1993; Rousset 1997; Leblois et al. 2004).

On the other hand, moment methods may throw out too much of the information in the data. Much recent efforts have been focused on developing maximum likelihood (ML) methods (Beerli and Felsenstein 1999, 2001; Bahlo and Griffiths 2000; Stephens and Donnelly 2000; Beerli 2004; de Iorio and Griffiths 2004b; de Iorio et al. 2005), which in principle use more information in the data than moment methods. Although ML methods could allow to estimate more parameters and to estimate them more accurately, the same general robustness concerns arise as for any other method. The effect of populations that are connected by dispersal to the sampled ones, but that are not accounted for in the statistical model, has received some attention (Beerli 2004; Slatkin 2005). Beerli (2004) investigated the effect of a third population on the estimation of dispersal between 2 populations, for sequence data (100,000 bp per individual sampled). Slatkin (2005) considered predicting the magnitude of these effects from a simple algebraic argument based on expected coalescence times of pairs of genes. This argument predicts a bias when there is some estimator bias in Beerli's simulations, but the predicted bias is an overestimate of the observed bias. Slatkin also notes that the method cannot be applied to all possible dispersal patterns and sampling designs, in particular in a linear array as will be considered below.

So far, performance of ML methods has been analyzed in models with up to 4 demes (Beerli and Felsenstein 2001; Beerli 2006). The ultimate aim of the present work is the application and assessment of ML methods in much larger networks of subpopulations. The type of data considered here are allelic counts at high mutation rate loci such as many microsatellites.

Estimation is expected to be less precise when the number of parameters increases, and this effect is already apparent in a 4-demes model (Beerli 2006). We therefore focus on dispersal models with few parameters, such as the stepping stone/isolation by distance models on a homogeneous lattice, rather than the whole migration matrix approach implemented, for example, in Migrate (Beerli and Felsenstein 1999, 2001). We have implemented the algorithm of de Iorio and Griffiths (2004a, 2004b) to handle the case of localized dispersal (isolation by distance) in a linear habitat. This scenario has been chosen because it is a relatively simple starting point for a larger simulation project, yet it is realistic enough to have allowed reasonably accurate statistical analyses on real data sets. Although the isolation by distance models neglect spatial heterogeneities, these do not appear to be a major concern in a number

Key words: dispersal, maximum likelihood, coalescence, isolation by distance, microsatellites.

E-mail: Rousset@isem.univ-montp2.fr.

Mol. Biol. Evol. 24(12):2730–2745. 2007
doi:10.1093/molbev/msm206
Advance Access publication September 24, 2007

of applications, for example, allowing good estimation of "neighborhood size" by nonlikelihood methods (Rousset 1997, 2000; Sumner et al. 2001; Fenster et al. 2003; Watts et al. 2007), and a similar result will be achieved here using the data of Watts et al. (2007). We attempt to analyze samples from large sets of subpopulations, not only because of the problem of unaccounted populations but because, as shown in these references, many natural populations may be described as a large network of small subpopulations connected by a large amount of dispersal, even up to the point where no subpopulations are distinguished from individuals or mating pairs ("continuous" populations). Moreover, the same works confirm the theoretical expectation that such conditions are favorable to the reliable estimation of dispersal rates.

We will see that although ML estimation under models of 10–15 populations is easy based on the algorithms of de Iorio and Griffiths, it becomes progressively more difficult as the number of subpopulations increases, and analysis of an average data set would require weeks on most desk computers when more than 40 subpopulations are considered. Hence, after a check of the method and assessment of numerical factors that may affect the precision of the estimates in simple conditions, we will consider the effects of unaccounted subpopulations on the analyses and a fast approximation to ML.

Under a nearest neighbor stepping stone model, unaccounted populations will be found to have little effect, but if dispersal distance follows a geometric distribution, stronger mis-specification effects will be obtained. Irrespective of mis-specification, the shape of the dispersal distribution will appear difficult to estimate. We will also test less thoroughly the effect of some other deviations from the model, which will appear to have less impact on performance.

A fast heuristic approximation, product of approximate conditional (PAC) likelihood (Li and Stephens 2003; Cornuet and Beaumont 2007), will be shown to yield results very close to those based on likelihood itself and will allow a more thorough investigation of possible causes of poor performance as well as of a wider range of parameter values, in particular higher dispersal among smaller demes, and lower mutation rates.

Methods

Design of Simulation Study

Population Models

As a first approximation, we may consider many species as collections of clusters of subpopulations (or of "demes") with abundant dispersal within each cluster and relatively much less dispersal among clusters. We consider the analysis of one such cluster. Typical values for the biological scenarios envisioned here would be deme size ≈1–100, dispersal probability ≈0.5, and therefore expected number of immigrants ≈0.5–50 per deme. In order to maintain enough genetic variability within the total population, we must also consider a large array of demes and/or large deme size and small migration rates.

These different requirements somehow conflict with each other and with constraints on computation times. Thus, we first consider large haploid deme size (400), small dispersal probability (0.01), and high mutation probability (10^{-3} per gene copy per generation), so that we can check performance in a small network of populations; then we will take benefit of the fast PAC likelihood method and will increase lattice size, reduce deme size, increase dispersal, and decrease mutation probability. In all cases the probability of dispersal to signed distance $k \neq 0$ can be described as

$$\frac{m}{2}(1-g)g^{|k-1|}, \tag{1}$$

for given g. g is thus a shape parameter which describes dispersal distances. The stepping stone model is the limit case $g \to 0$.

Some effects of dispersal on population processes, such as cline shape, are well quantified by the axial mean-squared parent–offspring distance, σ^2 (e.g., (Barton and Gale 1993). In the geometric dispersal model,

$$\sigma^2 = \frac{m(1+g)}{(1-g)^2}. \tag{2}$$

The product $D\sigma^2$, where D is population density, also determines spatial variation in the probability of identity of genes (isolation by distance: Sawyer 1977 for the most accurate results). Although different views have been held about the reasonable magnitude of σ^2 and $D\sigma^2$, these parameters can be low in natural populations. In such organisms, as *Dipodomys* rodents (Rousset 2000; Winters and Waser 2003), humans in the rainforest (Wood et al. 1985; Rousset 1997), *Chamaecrista fasciculata* (Fabaceae; (Fenster et al. 2003), American marten (Broquet et al. 2006), and *Gnypetoscincus queenslandiae* skinks (Sumner et al. 2001), concurrent genetic and demographic estimates of $D\sigma^2$ were $2.5 \leq \cdot \leq 40$, and such is σ^2 when measured in unit of interindividual distance (such that $D = 1$). In the latter units, σ^2 will be $7.5 \leq \cdot \leq 840$ in our simulations. Note that in the linear habitats considered in this work, $D\sigma^2$ values cannot be compared unless they are measured in the same spatial unit because density scales as distance^{-1} hence $D\sigma^2$ scales as distance. In this work, it will be reported as $N\sigma^2$, that is, in units of array step (except for the actual data analysis).

The assumed mutation probability is 10^{-3} or 10^{-4}, which is not unrealistic for microsatellite markers (reviewed in Ellegren 2000; see also e.g., Vigouroux et al. 2002; Gusmão et al. 2005). At high mutation rate loci, the allelic type of rare immigrants from distant populations should be uncorrelated to that of resident individuals, so the mutation events can also represent immigration from distant clusters of populations (Kimura and Weiss 1964). Only the 10^{-3} mutation probability will be considered in the smallest populations simulated as it is required to maintain substantial variation. Both mutations rates will be considered in larger populations, where this 10-fold variation in mutation probability will have notable consequences for the interpretation of the results.

The K-allele mutation model will be assumed in the data-generating simulations, except for a few cases where

Table 1
Notation

Population (data-generating) model	
N	Deme size (gene copies or haploid individuals)
σ^2	Mean-squared parent–offspring distance
n_d	Number of demes
$v_{\alpha\beta}$	Dispersal probability between demes α and β in de Iorio and Griffiths (2004b)
m	Total dispersal probability (this work)
Additional parameters of statistical model	
n_m	Number of demes in statistical model
N_T	Total haploid size in statistical model (N in de Iorio and Griffiths, 2004b)
N_e	Effective size in statistical model (haploid equivalent)
$\theta \equiv 2N_T\mu$	Scaled mutation rate
$m_{\alpha\beta} \equiv 2N_T v_{\alpha\beta}$	Scaled dispersal rate in between demes α and β in de Iorio and Griffiths (2004b)
Numerical parameters	
n_t	Number of ancestral trees (IS algorithm) or sequences (PAC likelihood algorithm)
n_p	Number of parameter points in which likelihood or other statistics are computed

a 10-allele–bounded stepwise mutation model (SMM) will be considered in order to test the robustness of the analyses when the marker mutational process deviates from the one assumed in the statistical model.

Statistical Models

In its most general form the statistical model considered here allows estimation of 3 parameters: a mutation rate, a migration rate (scaled probability of immigration), and the g parameter describing the geometric distribution of dispersal distances. We also investigated the performance of the estimator obtained by plugging the ML estimates of Nm and g in the parametric expression for $N\sigma^2$ in terms of these parameters (eq. 2).

Most simulations assume localized dispersal on a linear array of populations, with absorbing boundaries, much as in the population models under which samples are simulated. However, in practice this either assumes that the positions of the sampled demes in the linear array are known or this forces the user to make assumptions about this position. Hence, estimation under a circular lattice model will also be considered.

Sampling Design

We assume that samples are taken as follows: in the 4-demes model, in each deme; in the 100-demes models, at positions $50, 52, \ldots, 50 + 2(n_s - 1)$ where n_s is the number of demes sampled; in the 1,000-demes models, at positions $500, 502, \ldots, 500 + 2(n_s - 1)$ for $n_s = 4$ or 10 and at positions $500, \ldots, 519$ for $n_s = 20$.

The Algorithms and their Implementation

Notation

Some notation is summarized in table 1. Note that de Iorio and Griffiths (2004a, 2004b) denote N, the total size of the population, which we here denote N_T.

Computation of the Likelihood

The detailed features of the algorithm have been described in de Iorio and Griffiths (2004a, 2004b) and are not repeated here, although some guidance is given. In this section, their notation is followed, unless indicated otherwise. Their algorithm computes likelihood under the structured coalescent models described by Notohara (1990) and Herbots (1997), which are limit processes, for large deme size and low migration rates, of the classical migration matrix models (e.g., Nagylaki 1983; Rousset 2004, p. 54 sqq). In these algorithms, one considers an absorbing Markov chain over the state $\mathbf{n}$ of the set of ancestral lineages of a sample of genes from the time of sampling up to the most recent common ancestor. $\mathbf{n}$ is characterized by the allelic type and the geographic position of the lineages. A sample can be represented as the sequential addition from 0 to n genes, where the probability that any additional gene is of a given type will depend on the state of genes already present. The likelihood can thus be written in terms of any given sequence of gene states s_l (allelic type and geographic position) leading to the observed sample, as

$$\binom{n}{\mathbf{n}} \prod_{l=1}^{l=n} \pi(s_l|\mathbf{n}_{l-1}), \tag{3}$$

where $\pi(s_l|\mathbf{n}_{l-1})$ is the probability that an additional sampled gene s_l is of a given type, given the configuration $\mathbf{n}_{l-1}$ already generated by the sequence, and $\binom{n}{\mathbf{n}}$ is the multinomial coefficient in terms of the allelic counts $\mathbf{n}$ (de Iorio et al. 2005). de Iorio and Griffiths define an importance sampling (IS) algorithm considering the successive events (mutation, migration or coalescence) that may affect the ancestral lineages of the sample. $\pi(\cdot)$ terms can be defined for migration and mutation events from the above ones (e.g., the π for a migration event leading from some configuration "$\mathbf{n}$ + migrating gene in deme 1" to "$\mathbf{n}$ + migrating gene in deme 2" is defined from the ratio of the $\pi(\cdot|\mathbf{n})$ for addition of the migrating allele to deme 1 and of the $\pi(\cdot|\mathbf{n})$ for its addition to deme 2). The IS algorithm is defined in terms of approximations $\hat{\pi}$ to the π's. If $\hat{\pi}=\pi$, one iteration of this algorithm (i.e., one ancestral history) is enough to compute the likelihood. In general, the π's cannot be computed exactly so that $\hat{\pi} \neq \pi$. In this case, the IS algorithm may still allow consistent estimation of the likelihood, and fewer iterations

of this algorithm should be needed the closer the $\hat{\pi}$'s are to the π's. Poor choice of $\hat{\pi}$ may result in inefficient estimation of the likelihood (requiring too many iterations of IS for practical applications) or even in inconsistent estimation (Stephens and Donnelly 2000). To overcome these limitations, de Iorio and Griffiths proposed to use the following $\hat{\pi}$. Denote $\hat{\pi}(j|\alpha, \mathbf{n})$ the coefficients considered when a lineage of allelic type j in deme α is affected by some event. They are obtained as solutions of linear equations of the form:

$$[n_\alpha q_\alpha^{-1} + m_\alpha + \theta]\hat{\pi}(j|\alpha, \mathbf{n}) = n_{\alpha j} q_\alpha^{-1} + \theta \sum_i P_{ij} \hat{\pi}(i|\alpha, \mathbf{n}) + \sum_{\beta \neq \alpha} m_{\alpha\beta} \hat{\pi}(j|\beta, \mathbf{n}) \quad (4)$$

(de Iorio and Griffiths 2004b, eq. 2.11) for each j and α. Here q_α is the deme size relative to the total population size, and (P_{ij}) is a matrix of relative forward mutation rates. In the present work, we assume a 1-dimensional lattice, with n_d demes of equal size, so that $q_\alpha = 1/n_d$.

Solving a system of Z linear equations typically requires approximately $O(Z^3)$ computations (e.g., Press et al. 1988; Golub and van Loan 1996). The $\hat{\pi}(j|\alpha, \mathbf{n})$ are the solutions of a system of Kn_m equations of the form (4) and most of the computation time is spent solving such systems of equations. This is the limiting step in considering scenarios with large numbers of alleles or of demes. Ways of dealing with a large number of alleles are discussed below. The increase in computation time with the number of demes actually scales higher than n_m^3 because the number of events in the history of a sample increases as n_m increases. Iterative methods of solution of linear systems of equations can speed up the computations with a negligible loss of accuracy when compared with the direct solvers. A preconditioned conjugate gradient method (e.g., Golub and van Loan 1996) was found useful for $n_m \geq 60$ in this study.

Computation time can be substantially reduced if the above system of equations can be broken down in disjunct subsystems of equations. This occurs in particular in the symmetric K-allele model (KAM) that was the only model assumed in the estimation procedure in this work. In the KAM, $P_{ij} = 1/K$ (for $i = j$ included if we follow de Iorio and Griffiths' convention). $\sum_i \hat{\pi}(i|\alpha, \mathbf{n}) = 1$, so that the mutation term on the right-hand side of equation (4) simplifies: for each allele type j, the recursion (4) can be written as

$$[n_\alpha q_\alpha^{-1} + m_\alpha + \theta]\hat{\pi}(j|\alpha, \mathbf{n}) - \sum_{\beta \neq \alpha} m_{\alpha\beta} \hat{\pi}(j|\beta, \mathbf{n}) = n_{\alpha j} q_\alpha^{-1} + \theta/K. \quad (5)$$

Hence, for each allele j, the $\hat{\pi}(j|\alpha, \mathbf{n})$ are obtained as a solution of the system of n_d linear equations for $\alpha = 1, \ldots, n_d$. The system of Kn_d equations separates in K disjunct systems of n_d equations, only one of which is solved once a given allelic type has been chosen. Systems of n_d linear equations also arise for more complex mutation models if the mutation and genealogical processes are independent (as is usually assumed for neutral genetic variation): for each right eigenvector $\mathbf{r}_k \equiv (r_{kj})$ of (P_{ij}), one has to deduce equations for $\sum_j r_{kj} \hat{\pi}(j|\alpha, \mathbf{n})$ from equation (4). For each eigenvector $\mathbf{r}_k$, there are n_d such equations. Solutions of such a system can then be back transformed to obtain solutions of equation (4). This procedure is illustrated for an unbounded SMM, where it amounts to Fourier analysis, in de Iorio et al. (2005), but it increases computation time in comparison to the KAM and was not considered here. Instead, the performance of KAM-based estimation on data following a bounded SMM will be presented.

Another potential solution to reduce the computation time is bridge sampling (Meng and Wong 1996; Fearnhead and Donnelly 2001), in which the proposal distributions (hence the $\hat{\pi}$'s) are not computed independently for each parameter point, but only for a few driving values. There could be a trade-off between the cost of computing the $\hat{\pi}$'s for each point and the potential loss in efficiency of likelihood estimation when suboptimal proposal distributions are used, and the efficiency of de Iorio and Griffiths' proposal distribution has initially drawn us away from methods such as bridge sampling. However, this could prove useful in later applications.

The PAC Likelihood Heuristics

Cornuet and Beaumont (2007) proposed to use de Iorio and Griffiths's $\hat{\pi}$ directly as a substitute to π in equation (3) and to average over different sequences of genes leading to the sample (see also RoyChoudhury and Stephens 2007). This follows a similar suggestion by Li and Stephens (2003) who described this procedure as PAC likelihood and as maximum PAC likelihood, the procedure of maximizing this product with respect to parameters.

There is no general result showing that the PAC likelihood algorithm consistently estimates the likelihood. It clearly does so when $\hat{\pi} = \pi$, in which case simulation is not necessary. π is known in particular when the stationary joint distribution of allele frequencies in different demes is known. This occurs in the n-coalescent with parent-independent mutation (Stephens and Donnelly 2000; de Iorio and Griffiths 2004a) and can be extended to the island model with the same mutation model and a large number of islands. On the other hand, the distribution is unknown for stepwise mutation and/or for isolation by distance. Nevertheless, simulation results of Cornuet and Beaumont and RoyChoudhury and Stephens for stepwise mutation suggest that PAC likelihood may be used as a practical substitute to likelihood, and it is much faster to compute because the number of systems of linear equations to be solved for each sequence is the sample size, whereas in the IS algorithm, the number of linear systems will be increased beyond this in proportion to the number of mutation and migration events in an ancestral history.

As will be seen, the PAC likelihood statistic is not a consistent estimator of the likelihood. On the other hand, the variance of estimation of the PAC likelihood for a given number of iterations of the PAC likelihood algorithm is lower than the variance of estimation of likelihood from the same number of iterations of the IS algorithm, as was already observed by Cornuet and Beaumont and RoyChoudhury and Stephens. This reduced variance more

than compensates for the small bias in estimating likelihood. So, even if maximizing likelihood is better, in terms of mean square error (MSE), than maximizing the expectation of PAC likelihood, the estimation of demographic parameters by maximum PAC likelihood may appear better than ML estimation when both the likelihood and the PAC likelihood are estimated with some error. We will indeed find that maximum PAC likelihood estimation is at least as good, if not slightly better than ML estimation by the IS algorithm.

Likelihood Surface and ML Estimation

The likelihood in any given parameter point is estimated with some error rather than computed. This prevents the straightforward application of most algorithms for finding the maximum of a function. A convenient way to address this problem is to interpolate the likelihood surface from the estimated points by predicting it under some probabilistic model for the shape of the surface. If the surface is assumed to be the realization of a Gaussian process, this prediction can be achieved by techniques known as kriging (e.g., Cressie 1993). Both prediction uncertainty and prediction bias, when the Gaussian assumption does not hold, are expected, but in most cases this appears to be a minor source of inaccuracy, as can be tested by increasing the density of points on which interpolation is based. We use kriging as in de Iorio et al. (2005; see also Sacks et al. 1989; Welch et al. 1992): for each point in parameter space, the likelihood is estimated by simulations of n_t trees. This is repeated at n_p points in parameter space. Kriging fits a surface to the estimated likelihood values in the different parameter points. This is an estimate of the likelihood surface, of which the maximum may be sought by any of the usual algorithms. We used the package fields (Fields Development Team 2006) in the R statistical environment (R Development Core Team 2004) for the kriging computations. The Nelder–Mead algorithm as implemented in the R optim function was used to find the maximum of the estimated likelihood surface.

As in de Iorio et al. (2005), points are selected by Latin hypercube sampling (a form of stratified random sampling). The number of points is adjusted as function of time constraints and efficient use of hypercube sampling. The range of parameter space explored has to be provided by the user. In general, one should first explore a wide parameter space then focus the search around the first estimate obtained. Here, preliminary work (not shown) helped select parameter ranges within which all estimates would be found, and only results for these ranges are presented because they give the relevant information about the performance of ML estimation per se. Alternatively, a more automated iterative procedure has been used where a wide parameter range is used in the first iteration and the parameter range used in the later iterations is narrowed around the previous estimate obtained for each sample. In particular, a 2-steps procedure has been used recurrently, where estimates were first deduced from 512 points; 512 additional points were sampled from an approximately 10-fold smaller parameter space around the first estimates for each sample, and final estimates are deduced from the 1024 points thus obtained.

Computer Implementation

The C++ program used in all data analyses will be distributed as a free software, MIGRAINE, available through URL http://kimura.univ-montp2.fr/~rousset/Migraine.htm. It writes the required R code and can call R interactively to perform the above iterative procedure. It has been run on PCs under Windows and Linux, a Sun workstation, SGI Origin 3800 and IBM Power4 parallel computers of the CINES (www.cines.fr), and several Linux PC clusters. Some representative computation times are given in table legends.

Programs Checks

The likelihood estimation procedure was checked against standard formulas for probability of identity of pairs of genes (e.g., Maruyama 1970a; Malécot 1975) adapted to the KAM (e.g., Crow and Aoki 1984; Rousset 2004) and taken in the limit $N \to \infty$ for $N\mu$ and Nm fixed as in the coalescent algorithm. The simulation program generating samples has been previously described (Leblois et al. 2003, 2004) and has been checked as described in these papers.

Comparison of Performance of Different Implementations

There are 2 sources of inaccuracy of estimates. One is the inaccuracy of the ML estimate relative to the parameter value. The other is the inaccuracy of the numerical method in locating the ML estimate, which may be due to considering not enough replicate trees per point in the IS computation or not enough points. It is possible to evaluate the inaccuracy due to the numerical method by comparing independent runs on the same data (de Iorio et al. 2005). However, it would have been too time consuming to do so in all cases. Rather, the impact of numerical settings on performance will be checked in several cases.

Distributions of estimators (or distributions of differences in cases of paired simulations) were compared primarily through the estimation of differences in MSE or relative MSE, and further by estimation of differences in bias and variance. Maximum differences consistent with the data were deduced from 95% confidence intervals (CIs) for effects on MSE, bias, and variance constructed by the "bootstrap corrected and accelerated" (BC_a) method of DiCiccio and Efron (1996). However, because it would be inconvenient to report all CIs, synthetic bounds on maximum absolute effect size and/or P values derived from the confidence curves are reported when they bring the main information together with estimates reported in the tables.

Results

Numbers in brackets refer to the numbered cases in the different tables. For the parameters Nm, $N\mu$, and $N\sigma^2$, we present relative bias and relative root MSE ($\sqrt{\text{MSE}}$) as this may be more important than absolute bias and MSE in practice. These relative error measures cannot apply for g (in particular in the nearest neighbor stepping stone model, $g = 0$), for which bias and MSE are directly computed.

Table 2
Performance of Estimation in a Nearest Neighbor Stepping Stone Model

	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)
Linear array of 4 demes of 400 individuals		
$K = 4$		
[1]	0.04 (0.28)	0.11 (0.36)
[2] (n_p = 5,000)	0.04 (0.31)	0.11 (0.33)
[3] (n_p = 5,000, 10 loci)	−0.01 (0.22)	0.09 (0.25)
[4] (n_p = 1,000, 50 loci)	0.11 (0.19)	−0.03 (0.12)
[5] PAC	−0.002 (0.28)	0.09 (0.33)
$K = 10$		
[6]	0.52 (0.60)	−0.04 (0.25)
[7] (n_t = 300)	0.53 (0.61)	−0.08 (0.23)
[8] (50 loci)	0.47 (0.48)	−0.11 (0.13)
Linear array of 100 demes of 400 individuals		
$K = 4$		
[9] (n_p, n_t) = (512, 10)	0.33 (0.72)	0.10 (0.53)
[10] (n_p, n_t) = (512, 30)	0.27 (0.67)	0.12 (0.51)
[11][a] (n_p, n_t) = (5,000, 30)	0.25 (0.63)	0.16 (0.53)
[12] n_m = 100	0.30 (0.50)	−0.04 (0.32)
[13] PAC(n_p, n_t) = (512, 10)	0.26 (0.73)	0.06 (0.48)
[14] PAC(n_p, n_t) = (512, 30)	0.29 (0.75)	0.08 (0.49)
[15] PAC(n_p, n_t) = (512, 300)	0.28 (0.74)	0.07 (0.49)
[16] PAC(n_p, n_t) = (5,000, 30)	0.25 (0.69)	0.10 (0.51)
[17] PAC, n_m = 100	0.19 (0.42)	−0.12 (0.24)
$K = 10$		
[18]	0.40 (0.51)	−0.005 (0.29)
[19] Bounded SMM	−0.15 (0.25)	−0.06 (0.29)

NOTE.—Sixty samples were analyzed except 30 for case [4] and 120 for cases [9] and [18].

[a] Likelihood computations for case [11] took ≈84 min per sample on 2.66 GHz processors, whereas the otherwise identical PAC likelihood analysis took less than 7 min per sample.

Stepping Stone Dispersal

Numerical results are presented in table 2. For all cases in this table, $\mu = 0.001$ and $m = 0.01$ ($N\mu = 0.4$, $Nm = 4$). The following values apply to all cases unless noted otherwise: sample sizes were 5 loci, 4 demes sampled, and 60 genes sampled per deme; sampled ranges of parameter values were $2N\mu \in [0.125, 5]$ and $2Nm \in [0.125, 20]$; $n_m = 4$, $n_p = 512$, and $n_t = 10$.

The precision would be excellent for most practical purposes. For 5 different analyses of the same data sets [1]–[5] (4-allele model), the analysis with the largest number of loci (50, case [4]) stands out as the one with the lowest MSE for both estimates (as could be expected) as well as the lowest Nm bias, but also the highest $N\mu$ bias. For an identical total computation effort, reducing the number of loci and increasing the number of points analyzed is less efficient (cases [3] vs. [4], all $P < 0.039$ except for $N\mu$ MSE). Similar observations are made for a population of 100 demes. For the 10-allele model, MSEs and variances are likewise reduced after a 10-fold increase in the number of loci (case [8] vs. [6]); all $P < 0.006$).

For a population of 4 demes, a stronger bias and MSE in $N\mu$ estimation is observed for data generated under the 10-allele model (case [6]) than under the 4-allele model. For 100 demes, estimation is markedly improved in the 10-alleles model relative to the 4-alleles one (cases [18] vs. [9]). The latter observation is more in keeping with the frequent observation that highly polymorphic markers allow more powerful inferences, including about structured populations (power of tests of differentiation, Goudet et al. 1996; estimators of F_{ST}, Raufaste and Bonhomme 2000; assignment, Estoup et al. 1998 and Waples and Gaggiotti 2006; $D\sigma^2$ estimation, Leblois et al. 2003). However, the estimation biases are reduced by less than 15% (CI bound on bias reduction; $P = 0.023$ for Nm bias, 0.24 for $N\mu$). One reason for persistent $N\mu$ biases with increased sample size (most notably when the number of loci is increased) is that the observed number of alleles k in a 1-locus sample is often lower than K, so the program analyzes the data under a k-alleles model rather than under the correct 4- or 10-alleles model. This also readily explains the comparatively poor performance in the 4-demes, 10-alleles cases as some alleles are more likely to be absent in smaller populations. There is no obvious way to avoid the resulting biases, unless external information is provided by the user. Thus, this must be taken as an inherent bias of the method. Further, this will be less of a problem in later applications with larger total population sizes (so that k approaches K), so no attempt was done to correct for this problem in the analyses.

Beyond the number of loci, the number of parameter points considered may also set a limit to the precision that can be reached whatever the number of loci is. However, increasing the number of points from 512 to 5,000 (case [2] vs. [1]) reduces all relative measures of performance by at most 4.5% (upper CI bound, significant only for $N\mu$ MSE and variance). Finally, the performance of maximum PAC likelihood is at least as good as the ML analysis (case [5] vs. [1], all relative effects in favor of PAC likelihood and <0.065; $P = 0.003$ for $N\mu$ bias and >0.24 otherwise). Another test of a numerical factor (the number of trees sampled by the IS algorithm, case [7] vs. [6]) shows weak effect (for MSEs, at most a 6% reduction for Nm, $P > 0.22$, although this hides a bias-variance trade-off, with maximum absolute bound 7.8% and $P = 0.001$ on Nm bias).

Thus, the performance of estimation appears limited more by sample size than by numerical aspects of the algorithms and not worsened by the use of the PAC likelihood approximation. These 2 observations will recur in the sequel.

Estimation of Scaled Parameters under Mis-Specification

We now focus on the effect of unaccounted populations. We assume that unsampled demes in between the sampled ones are known and properly accounted for. Otherwise, serious mis-specification effects would occur, but these should be relatively easy to anticipate and/or avoid. By contrast, we will consider the less trivial effect of unaccounted populations outside the spatial range of sampled populations. Indeed, we will consider the effect of populations that hardly exchange any migrant directly with the sampled populations.

First, we should make clear which parameters are to be estimated when some demes are unaccounted. The coalescent algorithm is based on approximations in terms of scaled parameters, $N_T m$ and $N_T \mu$ in a stepping stone model, and is expected to perform well (in the sense of asymptotic

efficiency, at least) when the statistical model and the true population structure match each other, that is, when $n_d = n_m$. But it is not obvious how it will perform when these do not match. Consider, for example, that 4 demes have been sampled out of a large population of $n_d = 100$ demes, and that only $n_m = 10$ demes are considered in the statistical model, so that likelihood is computed for different values of $Nn_m m$. Will ML estimates of $Nn_m m$ be close to $Nn_m m$, close to $N_T m = Nn_d m$, or show a more erratic behavior? In the latter case, the analysis could be useless. In the second case, it would not be possible to infer the dispersal probability m (if N is known) or the number of immigrants Nm, unless there is additional information about n_d. Indeed, n_d is often little more than a convenient abstraction as total population sizes fluctuate over the time span of coalescence of gene lineages. The inferences will therefore be most informative in the first case, when estimates approach $Nn_m m$, and likewise for $Nn_m\mu$ so that estimators of Nm and $N\mu$ can be deduced (and given moderate demographic information, N can be taken out).

This conclusion is a bit caricatural. One could argue that the most informative scaling for mutation and for migration differ from each other. However, for the parameter values of cases [11] and [18], the simulations indeed show that estimates of $Nn_d m$ and $Nn_d\mu$ approach $Nn_m m$ and $Nn_m\mu$. To make this clear, estimates of $Nn_d m$ and $Nn_d\mu$ will be both divided by the (necessarily known) n_m and the resulting values will be compared with Nm and $N\mu$ values in order to assess performance. When we take these values as the estimands, the mis-specification bias appears low (as long as $\mu = 10^{-3}$, as later simulations will emphasize). This turns a substantial mis-specification bias into a benefit of the method.

Under this interpretation, some mis-specification effects remain apparent by comparison with the correctly specified analysis (cases [12] vs. [9]; with both reduced variance and bias of Nm estimates and reduced variance of $N\mu$ estimates, all $P < 0.032$). The correctly specified PAC likelihood analysis also yields clearly better results than all incorrectly specified ML and PAC likelihood analyses (cases [17] vs. [9]–[16]). PAC likelihood estimates of Nm are less biased than ML ones ($P < 0.043$ in all 4 comparisons [9] vs. [13], [10] vs. [14], [11] vs. [16], and [17] vs. [12]), except for Nm in the latter case. However, effect sizes are <10% overall. Numerical settings again appear to be a comparatively minor source of error in estimation, the most notable effect being a reduction in variance of $N\mu$ estimates when a higher number of points is computed (CI for this reduction is 1.1–26.5% for case [9] vs. [11], less clear cut for case [13] vs. [16]). No further attempt was made to further sort out the diverse effects of model mis-specification, small sample size, and PAC likelihood approximation and their interactions as they all appear small.

Comparison of cases [18] and [9] also show, as in previous 10-alleles/4-alleles comparisons, a lower MSE and variance of estimators with 10 alleles than with 4 (all $P < 0.017$), yet the $N\mu$ bias is not reduced (CI for relative effect −0.07 to 0.19), which was explained as an effect of mis-specification of the number of alleles in the mutation model. Because this mis-specification should be less of a concern in sample's larger populations, and given our focus on microsatellite data, all further simulations in this paper are for $K = 10$ alleles.

Table 3
Performance of Estimation for Geometric Dispersal in a Linear Array of 100 Demes

	n_m	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g bias ($\sqrt{\text{MSE}}$)
Samples of 5 loci				
$g = 0$				
[20]	10	0.42 (0.54)	−0.17 (0.30)	0.12 (0.14)
[21] PAC	—	0.26 (0.39)	−0.21 (0.28)	0.16 (0.21)
$g = 0.2$				
[22]	—	0.52 (0.65)	0.04 (0.34)	−0.01 (0.14)
[23] PAC	—	0.39 (0.49)	0.03 (0.34)	−0.02 (0.16)
[24]	25	0.30 (0.46)	0.08 (0.38)	−0.02 (0.18)
[25] PAC	40	0.04 (0.25)	0.01 (0.31)	−0.02 (0.18)
[26] PAC	100	0.05 (0.24)	0.001 (0.30)	−0.01 (0.18)
$g = 0.5$				
[27]	10	1.01 (1.13)	0.23 (0.40)	−0.15 (0.29)
[28] PAC	—	0.86 (0.97)	0.16 (0.33)	−0.13 (0.27)
[29]	16	0.77 (0.90)	0.21 (0.39)	−0.21 (0.32)
[30] PAC	—	0.68 (0.85)	0.13 (0.31)	−0.14 (0.27)
[31][a]	25	0.65 (0.84)	0.20 (0.42)	−0.18 (0.30)
[32] PAC	—	0.48 (0.68)	0.15 (0.35)	−0.16 (0.29)
[33] PAC	40	0.20 (0.61)	0.12 (0.25)	−0.12 (0.25)
[34] PAC	100	0.20 (0.61)	0.09 (0.27)	−0.12 (0.26)
$g = 0.5$, 20 loci				
[35] PAC	10	0.76 (0.79)	0.07 (0.17)	−0.07 (0.19)
[36] PAC	100	−0.02 (0.31)	0.07 (0.20)	−0.04 (0.20)

NOTE.—Thirty multilocus samples of 4 sampled demes and 60 genes sampled per deme were analyzed, except 60 samples for cases [20] and [27]–[32].

[a] Case [31] required about 24 h 30 min per sample on 2.66 GHz CPUs.

Effect of Mutation Model

We have considered a KAM for fast computation, but of course this may not be realistic. Because estimating parameters under a more general mutation model appears unpractical, we evaluated the impact of a bounded stepwise mutation process on the estimation procedure (case [19]). The total population is also mis-specified as in the KAM (case [18]). Compared with the latter, there is some reduction in MSE of $N\mu$ estimates due to an ≈55% change in mean value (this is highly significant due to the low variance of estimates, $P < 5.10^{-4}$). The dispersal estimates are robust to mis-specification of the mutation model. Similar results will be obtained for other data generated under the SMM.

Geometric Dispersal

So far only the nearest neighbor stepping stone model has been considered, and mis-specification of the number of demes seemed to have little effect. We now consider whether we can estimate the parameters of a more general dispersal distribution. For these analyses, we assume that dispersal follows a geometric dispersal model described by equation (1). The performance of estimators of Nm, $N\mu$, and g when $g = 0$, 0.2, and 0.5 is presented in table 3. For all cases in this table, $N = 400$ haploid individuals

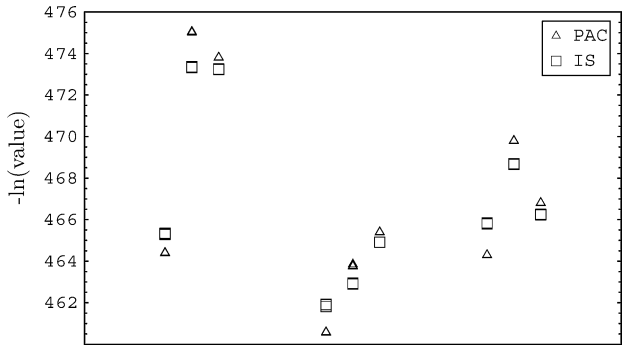


FIG. 1.—Differences between PAC likelihood and IS estimation. Each of the 9 columns shows 3 independent estimates of the PAC likelihood (△) and 3 estimates of the likelihood (□) for one given parameter point and one given sample. For each of them, the 3 replicates are barely distinguishable because the variance of estimation is very low. The 9 columns represent groups of 3 parameters points for each of 3 samples. Samples and PAC likelihood analyses are as in case [34] but with $n_t = 1{,}000$. IS analysis are for the same statistical model and same n_t.

per deme; samples were generated assuming a 10-alleles model; $n_t = 30$ and $n_p = 2197$ or 5,000; and sampled ranges were $2N\mu \in [0.125, 5]$, $2Nm \in [2.5, 20]$, and $g \in [0, 0.8]$.

The estimation of the parameter g appears very poor when $n_m = 10$. For $g = 0$, the estimator is biased upward (case [20]). A bias is expected for a MLE as the parameter value is at the boundary of the feasible parameter range, but in the present case this bias is high. For $g = 0.5$, the estimator of g is biased downward (case [27]). The weak bias observed for $g = 0.2$ may be the midpoint between the positive bias for lower values of g and the negative bias for higher value of g. For $g = 0$, there is also a slightly higher absolute bias of Nm estimates (CI 0.093–0.255) relative to analyses of the same data under the stepping stone model (cases [18] vs. [20]).

These biases may result both from model mis-specification, small sample size, and inaccurate estimation of likelihood. Effects of model mis-specification can be evidenced only by comparison with analyses assuming the true number of demes (100), and such analyses can be done routinely only with PAC likelihood. However, the value of the likelihood statistic under the "true" model was compared with the PAC likelihood in a few points, and there are demonstrable differences (fig. 1). Yet, maximum PAC likelihood performance appears at least as good as ML performance (cases [20]–[32] in table 3) as the differences are mostly in the direction of lower MSE by maximum PAC likelihood. For maximum PAC likelihood when $n_m = 100$ and $g = 0.5$ (case [34]), Nm is relatively well estimated, $N\mu$ estimates remain biased upward, and g estimation is not precise enough to be worth considering in practice. Thus, there is little information about g in the data. Accordingly, we will increase sample size (in particular, the number of demes sampled) in later simulations. With the present sampling design, performance is as good with 40 as with 100 demes, but there is evidence of mis-specification on Nm and $N\mu$ estimation as the bias and MSE of their estimators decrease with n_m increasing from 10 to 40.

Not only the number of demes but also the position of samples relative to the total habitat may be mis-specified.

Table 4
Edge Effects

	n_m	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g bias ($\sqrt{\text{MSE}}$)
Position effect, $g = 0.5$, 5 loci				
[37] PAC[a]	25	0.20 (0.59)	0.11 (0.25)	−0.12 (0.25)
Analyses under circular array model				
$g = 0$, 5 loci				
[38]	10	0.32 (0.43)	−0.17 (0.29)	0.06 (0.08)
$g = 0.2$, 5 loci				
[39]	10	0.44 (0.59)	−0.06 (0.34)	−0.07 (0.14)
[40]	25	0.21 (0.35)	0.14 (0.41)	−0.07 (0.14)
[41] PAC	40	0.04 (0.25)	0.01 (0.31)	−0.02 (0.18)
[42] PAC	100	0.04 (0.25)	0.01 (0.31)	−0.01 (0.19)
$g = 0.5$, 5 loci				
[43]	10	0.86 (0.96)	0.17 (0.32)	−0.28 (0.35)
[44][b] $n_t = 120$	10	0.91 (1.04)	0.17 (0.31)	−0.31 (0.36)
[45]	25	0.53 (0.76)	0.19 (0.31)	−0.22 (0.29)
[46] PAC	25	0.22 (0.59)	0.10 (0.25)	−0.13 (0.25)
[47] PAC	40	0.20 (0.61)	0.12 (0.25)	−0.12 (0.25)
$g = 0.5$, 20 loci				
[48][b]	10	0.84 (0.87)	0.11 (0.25)	−0.28 (0.31)
[49] PAC	100	0.02 (0.29)	0.06 (0.19)	−0.06 (0.17)

NOTE.—Samples and simulation conditions as in Table 3.

[a] Samples set in positions 10, 12, 14, 16 of the array versus 3, 5, 7, 9 in other analyses with $n_m = 25$.

[b] The analysis of each sample from cases [44] and [48] (5,000 points) takes 12 CPU hours on 2.66 GHz processors.

Estimation performance may differ whether samples are set close to the assumed edge of the habitat or in its center as shown in one example (case [37] vs. [32], differing in particular through the $N\mu$ mean, CI on relative effect 0.16–0.39). Analyses under a circular array model were investigated as a practical alternative to having to choose the position of samples on a linear lattice (table 4 and fig. 2). Overall, the performance depends somewhat on whether a circular or linear array is assumed, but this does not affect the previous conclusions (including the consistently smaller bias of maximum PAC likelihood estimates relative to MLEs across simulation conditions, and the improvement in $N\mu$ estimation when n_m is increased). As could be expected, the highest discrepancies between linear and circular analyses are observed for the lowest n_m and highest g value and should be generally negligible relative to other causes of error. As before, strong biases may be observed for the lowest n_m values, and increasing the number of replicate ancestral trees ([44] vs. [43]) or the number of loci (case [48]) has little effect, confirming that the biases are mostly due to mis-specification. The differences between circular and linear models are very small for $n_m \geq 40$ (with most CI widths for effects on means narrower than 0.01 in the PAC likelihood analyses; see fig. 2 legend).

Larger Samples

The previous results show that unaccounted populations become important when there is some "long-distance" dispersal ($g > 0$; keeping in mind that $g = 0.5$ implies only limited long-distance dispersal, compared with many biological studies). Further, even with a correctly specified model, there is less information about g in the data

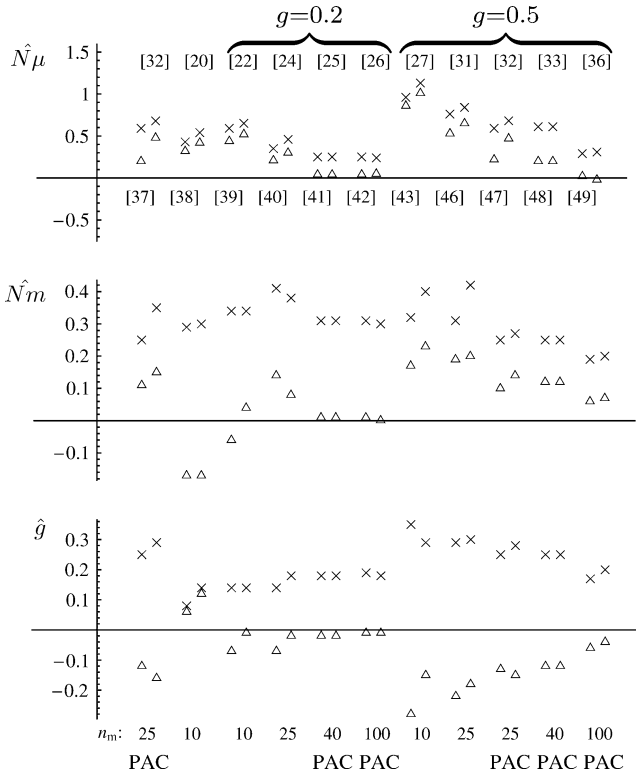


FIG. 2.—Interaction of edge effects with n_m. This figure compares selected results from tables 3 and 4. (Relative) bias (△) and $\sqrt{\text{MSE}}$ (×) are shown for the 3 estimated parameters. Simulation conditions were identical in the paired linear and circular analyses, including the sequence of the random number generator (for cases [20] and [32], this may involve only a subset of all replicates considered in table 3). Hence, with PAC likelihood, the same sequences of $\hat{\pi}$'s were computed, the only difference being the equations that the $\hat{\pi}$'s solve; whereas in the ML analysis, a $\hat{\pi}$ value at some step affects the further sequence of $\hat{\pi}$ values computed. This results in a much smaller variance of differences between paired PAC likelihood analyses compared with paired ML analyses (paired PAC likelihood analyses may remain statistically different even when the differences are not visible).

than about the migration rate. Overall, Nm was relatively well estimated in most cases, estimation of mutation rate was affected by mis-specification, and estimation of g was affected both by mis-specification and by lack of power.

To increase power, both larger number of demes sampled and of loci were considered (table 5). As previously, there is no evidence of mis-specification with 40 demes in the statistical model. The performance of maximum PAC likelihood (with 5 loci, case [51]) or of ML estimation (with 10 loci, case [50]) is excellent. Similar results are obtained with $n_t = 10$ or 100 sequences (case [52] vs. case [51], all $P > 0.059$, all CI bounds on effects <0.074) for PAC likelihood computation, confirming that a low n_t is enough. Good performance is confirmed in the case $g = 0.2$ (case [56]). However, if samples come from a 1,000-demes array, a slight bias reappears for $N\mu$ estimates (case [53]), whereas a more substantial bias appears if the true mutation rate is reduced to 10^{-4} (case [54]). The latter phenomenon will be investigated more thoroughly below. Finally, analysis of data generated under a bounded SMM shows an ≈50% reduction in mutation rate estimates but no notable effect on dispersal estimation (case [55]).

Table 5
Performance of Estimation for Geometric Dispersal

	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g	Bias ($\sqrt{\text{MSE}}$)
[50][a] IS	0.04 (0.25)	0.08 (0.17)	0.5	−0.05 (0.14)
[51] PAC	0.02 (0.29)	0.01 (0.13)	—	−0.02 (0.15)
[52] PAC, $n_t = 100$	0.01 (0.24)	0.03 (0.13)	—	−0.04 (0.14)
[53] $n_d = 1{,}000$, PAC	0.17 (0.33)	−0.01 (0.16)	—	−0.05 (0.15)
[54] PAC, $\mu = 10^{-4}$	0.77 (0.99)	0.03 (0.18)	—	−0.09 (0.13)
[55] SMM, PAC	−0.54 (0.57)	0.11 (0.25)	—	0.006 (0.14)
[56] PAC	−0.08 (0.17)	0.10 (0.3)	0.2	0.04 (0.12)

NOTE.—For each sample analyzed, 60 genes were sampled at each of 5 loci (10 in case [50]) in each of 10 demes out of a linear array of 100 demes (except 1,000 for case [53]). The mutation probability was 10^{-3} (except 10^{-4} for case [54]). $n_m = 40$ and $n_t = 10$ except 100 for case [52]. Other simulations settings were as in table 3. 2197 points were analyzed, except in case [50].

[a] In case [50], 2 steps of 512 points were computed as described in the text.

Smaller Demes with Higher Dispersal

Our aim is to test the performance of the algorithms in scenarios more representative of spatial structure at small spatial scale. In this section, we consider samples from a population of 1,000 demes, with fewer individuals per deme and higher dispersal rate. Twenty demes are sampled, so the number of demes in the statistical model is always larger than 20, and then ML analyses based on the IS algorithm become extremely time consuming. Hence, only maximum PAC likelihood is considered in all but one simulation.

The results are presented in table 6 and fig. 3. For $g = 0.5$ and $n_m = 60$, there are strong biases, in particular for $N\mu$. Mutation and migration rates are overestimated, whereas g is slightly underestimated. As before, simulation conditions were varied to understand these biases. For $n_m = 60$, increasing the number of loci yields reductions of variance of estimators (all >50% for both case [58] vs. [62] and case [61] vs. [60]) but no significant, or even consistent, improvement in biases. Varying the number of points (cases [57] vs. [59], [58] vs. [61], and [60] vs. [62]) or of replicate sequences (cases [57] vs. [58] and [59] vs. [61]) has no detectable effect, except for significant but still small effects (a few percents at most on biases) for cases [60] vs. [62]. Only increasing n_m to 200 demes does result in improved performance, with reduction of $N\mu$ bias (CI on relative reduction 0.17–0.40) and of g bias (CI on reduction 0.007–0.06; $N\sigma^2$ bias is likewise reduced). Improvement in bias of the same parameters for an identical increase in n_m is also apparent for a higher level of dispersal ($g = 0.75$, case [68] vs. [71]; $N\mu$ relative reduction 1.05–1.35, g reduction 0.002–0.07), although all biases are more moderate and less affected by n_m for lower level of dispersal ($g = 0.2$, case [66] vs. [65], all $P > 0.59$ for biases). Thus, the mis-specification problems previously encountered are met again, but at higher n_m values, when the total dispersal rate is increased.

Whether poor performance is due in part to the PAC likelihood heuristics can only be assessed by comparison with the ML analysis. One comparison was conducted for $n_m = 60$ and $g = 0.75$ (cases [67] vs. [68]), and both analyses yield very similar results, except that the PAC likelihood estimates of $N\mu$ are slightly less biased (0.03–0.15 relative reduction; MSE is reduced too). For both IS and

Table 6
Performance of Estimation under High Dispersal

	n_m	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm relative bias (relative $\sqrt{\text{MSE}}$)	g bias ($\sqrt{\text{MSE}}$)	$N\sigma^2$ relative bias (relative $\sqrt{\text{MSE}}$)
Samples from an array of 1,000 demes of 40 haploid individuals, with $m = 0.25$ ($N\mu = 0.04$, $Nm = 10$)					
$g = 0.5$, $N\sigma^2 = 60$					
[57]	60	0.58 (0.74)	0.91 (1.05)	−0.19 (0.23)	−0.12 (0.29)
[58] ($n_t = 30$)	—	0.58 (0.76)	0.88 (1.01)	−0.19 (0.21)	−0.10 (0.29)
[59] ($n_p = 2197$)	—	0.61 (0.79)	0.92 (1.05)	−0.18 (0.21)	−0.08 (0.27)
[60] ($n_t = 30$, $n_p = 2{,}197$, 20 loci)	—	0.56 (0.60)	0.86 (0.90)	−0.18 (0.19)	−0.09 (0.17)
[61] ($n_t = 30$, $n_p = 2{,}197$)	—	0.56 (0.74)	0.94 (1.10)	0.18 (0.22)	−0.08 (0.28)
[62] ($n_t = 30$, 20 loci)	—	0.53 (0.55)	0.95 (1.00)	−0.20 (0.21)	−0.12 (0.03)
[63]	200	0.27 (0.69)	0.83 (0.99)	−0.16 (0.19)	−0.05 (0.28)
[64]	40	1.26 (1.34)	0.96 (1.08)	−0.19 (0.22)	−0.06 (0.26)
$g = 0.2$, $N\sigma^2 = 18.75$					
[65]	200	0.38 (0.52)	0.41 (0.52)	−0.07 (0.08)	0.08 (0.22)
[66]	60	0.35 (0.44)	0.39 (0.45)	−0.08 (0.08)	0.09 (0.21)
$g = 0.75$, $N\sigma^2 = 280$					
[67] IS	—	1.51 (1.64)	0.77 (0.86)	−0.04 (0.13)	2.10 (4.01)
[68]	—	1.42 (1.55)	0.74 (0.84)	−0.03 (0.13)	2.45 (4.30)
[69][a] IS	—	1.57 (1.70)	0.75 (0.86)	−0.07 (0.12)	0.59 (2.63)
[70][b]	—	1.47 (1.60)	0.70 (0.78)	−0.07 (0.11)	0.53 (2.58)
[71]	200	0.27 (0.61)	0.54 (0.59)	−0.03 (0.07)	0.46 (1.05)
[72] SMM	—	−0.17 (0.44)	0.78 (0.86)	−0.08 (0.11)	0.16 (0.70)

NOTE.—Estimation is by maximum PAC likelihood except for cases [67] and [69]. Except as noted, $n_t = 10$, $n_p = 512$ and sample sizes were 5 loci per sample, 20 demes sampled, and 20 genes sampled per deme. Thirty such samples were analyzed in each case, except 120 samples for cases [67]–[70]. The parameter ranges explored were $2N\mu \in [0.0125, 0.5]$, $2Nm \in [2.5, 60]$ (except for $g = 0.2$ where $2Nm \in [2.5, 40]$ was sufficient), and $g \in [0.05, 0.8]$ except $g \in [0.2, 0.999]$ when true $g = 0.75$.

[a] Second step of 512 points after case [67].

[b] Second step of 512 points after case [68]. It took ≈ 20 min per sample on 2.6 GHz 64-bit processors. First and second iterations required about 1,000 and 110 more time, respectively, for IS computation than for PAC likelihood.

PAC likelihood methods, increasing the number of parameter points (cases [69] and [70]) markedly improved $N\sigma^2$ estimation. For the other parameters, the distribution of estimates are shown in fig. 4; PAC likelihood was again only slightly, though consistently, better than ML. Thus, the PAC likelihood heuristics again appears as an excellent substitute to likelihood estimation.

The effect of a bounded stepwise mutation process was tested again (case [72]), and it was again found that it yielded a reduction in mutation rate estimates and little effect on other parameters.

Lower Mutation Rate

Performance was assessed for a lower mutation rate ($\mu = 10^{-4}$), still with relatively high dispersal rates and small deme sizes (table 7 and fig. 3). For all cases in the table, samples were simulated for an array of 1,000 demes of 40 haploid individuals, with $m = 0.25$, 0.5, or 0.75 and $\mu = 10^{-4}$ ($N\mu = 0.004$, $Nm = 10$–30). Estimation was by maximum PAC likelihood with $n_t = 10$, and $D\sigma^2$ estimates were compared with those obtained by the moment method. Except as noted, sample sizes were 5 loci per sample, 20 demes sampled, and 40 genes sampled per deme; sampled ranges were $2N\mu \in [0.00125, 0.25]$; $2Nm \in [2.5, 60]$, $g \in [0.05, 0.8]$ when true $g = 0.5$ and $g \in [0.2, 0.999]$ when true $g = 0.75$. Genetic diversity remains high in these simulations. For example, in case [78], the probability of identity in the samples was 0.245.

With respect to $N\mu$ estimation, for $n_m = 60$, the bias appears high, but only following the previous decision to measure biases relative to $n_m N\mu$ rather than relative to $n_d N\mu$, the mutation rate scaled by the true total size of the population. For the highest dispersal, the bias is much weaker when assessed relative to $n_m N\mu$, and thus estimation performs more in accordance to the general definition of the algorithm. Expectedly, the biases are reduced when n_m is increased to 200, though still large in the highest dispersal case. Again, data simulated under a bounded SMM (case [84]) yield lower estimates of mutation rate.

Dispersal rate estimates are also substantially biased but can be interpreted as low-bias estimates neither of $n_m Nm$ nor of $n_d Nm$. In particular, they do not scale as n_d in the highest dispersal case. F_{ST}-based estimates of Nm could well be better than those derived under $n_m = 60$. The biases on Nm and g seem to compensate each other, yielding low relative biases on $N\sigma^2$ estimation. Most biases are reduced when the number of demes is increased, but additionally increasing the number of loci has little effect, which again indicates that large biases are mostly due to mis-specification.

$N\sigma^2$ estimates could be compared with those obtained by a moment method (Rousset 1997). Following the classical bias-variance trade-off of likelihood estimators, the PAC likelihood estimator generally has lower variance but higher bias than the moment estimator. The PAC likelihood estimator may have lower MSE overall, as one might expect under well-specified models, but the trend is not clear cut, which leaves room to speculate what would be the "best" method in practical conditions. The comparison could have been more favorable to the likelihood method in simulation conditions with relatively large μ/m as mutation is expected to bias results of the moment method in that case.

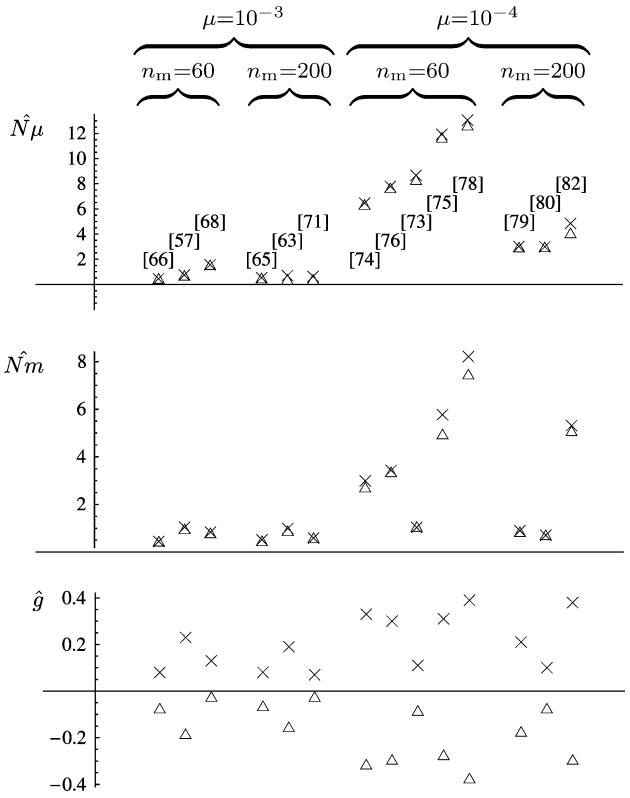


FIG. 3.—Interaction of mutation with n_m. This figure compares selected results from tables 6 and 7. (Relative) bias and $\sqrt{\text{MSE}}$ are shown as in Fig. 2. For each (μ, n_m) combination, cases are ranked by increased $N\sigma^2$ values.

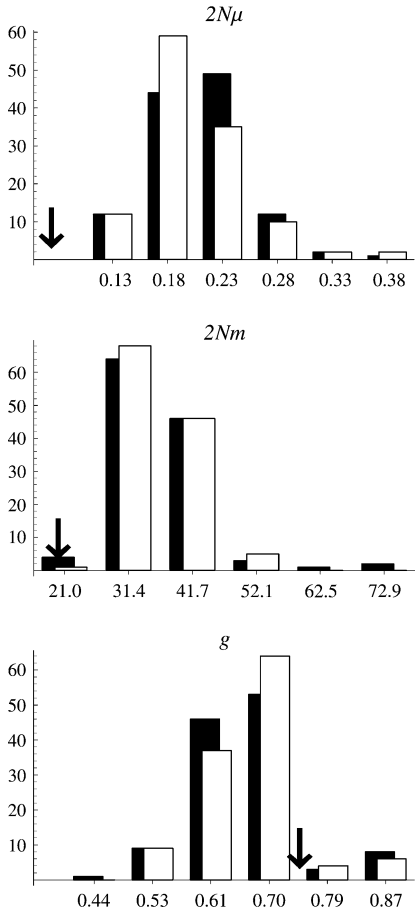


FIG. 4.—Distributions of estimates by PAC likelihood and IS estimation. The PAC likelihood distributions (case [70]) are laid over the likelihood distributions (case [69]). The arrows mark the position of the parameter values.

$N\sigma^2$ estimation is better ceteris paribus with $n_m = 60$ (case [78]) than with $n_m = 200$ (case [82]) for the same data, although the opposite effect of mis-specification holds for the other parameters. This suggests that specifying a large number of demes is less important for good estimation of $N\sigma^2$ than for other parameters.

As usual, it is not a priori obvious to which extant a relatively poor performance is due to numerical issues. In particular, for high g values, a minor error in finding the g MLE results in a high error on $N\sigma^2$ estimates. This effect was already apparent in cases [67]–[70], where increasing the density of points analyzed markedly improved $N\sigma^2$ estimation. We have further tested the effect of numerical parameters in cases showing the poorest $N\sigma^2$ estimation relative to the moment method. For $n_m = 60$, increasing n_t to 100 (case [77] vs. [76]) had no notable effect on the conclusions, whereas when $n_m = 200$, increasing n_t to 50 (case [83] vs. [82]) substantially reduced the MSE. Thus, mis-specification is the main determinant of poor $N\sigma^2$ estimation for low n_m, whereas a higher number of replicates become necessary for $N\sigma^2$ estimation with high n_m. The latter conclusion was confirmed when the mutation model is also mis-specified (cases [84]–[86]). The performance notably improves when n_t is increased from 10 to 50, mainly due to improvement of a few outlying estimates. As before, PAC likelihood performs as least as well as ML in this case; MLEs for $N\mu$ and $N\sigma^2$ actually have higher MSE ($P < 0.027$).

Application to Real Data

Watts et al. (2007) have compared genetic and demographic estimates of $D\sigma^2$ in the damselfly *Coenagrion mercuriale* along a linear habitat. The demographic estimate of $4D\sigma^2$ (for D here being a density of diploid individuals) derived from a mark–recapture study and corrected for variance in reproductive success was 277,894 individuals.m. Indirect estimates obtained from a sample of 240 individuals and 14 loci by several variants of the regression method based on pairwise comparison of individual genotypes (Rousset 2000) ranged within 179,058–242,816, with a synthetic CI 66,015–392,866.

In simulation conditions fitted to these data with respect to sampling design, gene diversity, dispersal distribution, and total population size (Watts et al. 2007, case σ = 130), the relative bias and root MSE of $D\sigma^2$ estimates yielded by the $\hat{e}$ regression estimator were 0.93 and 2.55 (reduced to 0.67 and 1.31 when 3 outliers are taken out of 200 replicates), and the other regression estimator considered yielded some negative estimates (for ease of comparison and as previously discussed in Leblois et al. 2003, bias and MSE of the more Gaussian-distributed $1/(D\sigma^2)$ were instead reported in Watts et al.). These biases are

Table 7
Performance of Estimation under Lower Mutation Rate

	$N\mu$ relative bias (relative $\sqrt{\text{MSE}}$)	Nm	Relative bias (relative $\sqrt{\text{MSE}}$)	g	Bias ($\sqrt{\text{MSE}}$)	$N\sigma^2$	Relative bias (relative $\sqrt{\text{MSE}}$) PAC	Relative bias (relative $\sqrt{\text{MSE}}$) Regression
$n_m = 60$								
[73][a]	8.19 (8.66)	10	1.01 (1.04)	0.75	−0.09 (0.11)	280	0.16 (0.54)	0.003 (0.50)
[74][a]	6.22 (6.46)	20	2.66 (2.99)	0.5	−0.32 (0.33)	120	0.05 (0.27)	−0.08 (0.44)
[75][a]	11.56 (11.94)	—	4.89 (5.77)	0.75	−0.28 (0.31)	560	0.13 (0.43)	−0.04 (0.44)
[76][a]	7.54 (7.78)	30	3.30 (3.43)	0.5	−0.30 (0.30)	180	0.36 (0.50)	−0.08 (0.20)
[77][b]	7.95 (8.19)	—	3.82 (3.97)	—	−0.34 (0.34)	—	0.31 (0.46)	−0.08 (0.20)
[78][a]	12.53 (13.05)	—	7.41 (8.21)	0.75	−0.38 (0.39)	840	0.09 (0.45)	−0.06 (0.55)
$n_m = 200$								
[79]	2.83 (3.01)	—	0.79 (0.90)	—	−0.18 (0.21)	60	−0.12 (0.26)	−0.08 (0.37)
[80]	2.85 (2.99)	10	0.66 (0.71)	—	−0.08 (0.10)	280	0.02 (0.44)	0.003 (0.50)
[81] (20 loci)	2.84 (2.89)	—	0.61 (0.63)	—	−0.07 (0.08)	—	−0.06 (0.16)	−0.04 (0.25)
[82][a]	3.96 (4.84)	30	5.02 (5.31)	—	−0.30 (0.38)	840	0.18 (0.70)	−0.06 (0.55)
[83][c]	3.79 (4.07)	—	3.66 (3.68)	—	−0.24 (0.25)	—	0.12 (0.46)	−0.06 (0.55)
Samples generated under SMM								
[84]	2.37 (2.85)	10	0.73 (0.80)	—	−0.08 (0.11)	280	0.18 (0.92)	−0.09 (0.57)
[85][d]	1.95 (2.45)	—	0.74 (0.79)	—	−0.07 (0.10)	—	0.17 (0.80)	−0.09 (0.57)
[86][e]	2.12 (2.71)	—	0.71 (0.76)	—	−0.08 (0.10)	—	0.13 (0.67)	−0.09 (0.57)
[87][f] IS	1.73 (1.82)	—	0.87 (0.91)	—	−0.06 (0.08)	—	0.34 (0.72)	
Versus [86] (subset)[g]	1.39 (1.51)	—	0.84 (0.87)	—	−0.07 (0.09)	—	0.18 (0.45)	

NOTE.—In each case, 200 samples were analyzed by the moment method, and 30 samples were analyzed by maximum PAC likelihood, except for cases [84] and [86] (60 samples) $n_p = 512$, except as noted.

[a] For analyses with $n_m = 60$, as well as case [82], where large biases were observed and difficult to anticipate, 2 steps of 512 points were computed as described in the text. In the first step, estimates were deduced from 512 points in the range $2N\mu \in [0.00125, 0.25]$, $2Nm \in [2.5, 450]$, and $g \in [0.05, 0.999]$.

[b] Two-steps procedure, $n_t = 100$ after first step of analysis of case [76].

[c] Two-steps procedure, $n_t = 50$ after first step of analysis of case [82].

[d] Two-steps procedure, the first step being case [84].

[e] As case [85] but with $n_t = 50$ in the second step.

[f] In case [87], the first 10 samples of case [86] were analyzed by ML with $n_p = 256$ and $n_t = 50$. This computation takes about 13 CPU years on 2.8-GHz processors.

[g] Same 10 samples as in case [86].

largely small-sample ones as could be seen by comparison with an estimate from 40,000 loci. More important though, the CI deduced jointly from the 2 regression estimators were little affected by such biases and had good coverage properties (Watts et al. 2007). To analyze the same simulated data by PAC likelihood, individual genotypes have to be binned in artefactual demes. Here, 80 such demes were defined exactly as described below for the actual data analysis. Geometric dispersal is still assumed in the statistical model, which now implies mis-specification of the dispersal distribution. Despite this, estimation performance is substantially improved as the maximum PAC likelihood has a lower relative root MSE of 0.54 (bias is 0.8%; from 60 replicates).

We have reanalyzed the damselfly data by PAC likelihood. Here, the linear habitat was divided in n_m spatial units of width $3500/(n_m - 1)$ m, the patch of habitat (the "Lower Itchen Complex" in Watts et al. 2007, fig. 1) being about 3500 m long. For $n_m = 80$, several gradually more focused (in parameter space) analyses led to $4D\sigma^2 \hat{=} 2159$, the unit being individuals.(bin width). When translated back to individuals.m, this is $4D\sigma^2 \hat{=} 95,645$. Several independent, less focused replicate analyses yielded likelihood ratio CI ≈50,000–140,000 (fig. 5 shows one such computation, where the Nb estimate is 92,039). Similar computations yielded $4D\sigma^2 \hat{=} 123,676$ and $113,303$ individuals.m for $n_m = 5$ and 20, respectively. Thus, as did n_m in the simulations, the bin width has little effect on $D\sigma^2$ estimates, although it affects more the other estimators.

Discussion

Performance of Estimation

In this work, we have investigated the performance of ML estimation of mutation and dispersal parameters under isolation by distance in a linear habitat, using de Iorio and Griffiths' IS algorithm. We have focused on the effect of mis-specification of the number of demes. In the same conditions, we have also found that the maximum PAC likelihood approximation is practically as efficient as ML analysis.

Beyond the simulation results reported in this ms, we have considered some additional approximations that would ease computations for large arrays of demes. In particular, approximation of the remote ancestry of a sample by Kingman's coalescent has been considered in 2-dimensional models (Cox 1989; Cox and Durrett 2002; Zähle et al. 2005), but for ancestors of genes uniformly sampled on the lattice, rather than in a small part of it as considered here. Even for uniform sampling, both analysis (Cox 1989) and simulations (Wilkins 2004) suggest it is not appropriate for linear habitats. In agreement with these results, we could not achieve good performance by such approximations while simultaneously reducing computation time by a notable extent (details not shown).

Expectedly, there is good performance of ML in favorable conditions (no model mis-specification, large sample size). In less favorable conditions, performance is affected differentially for different parameters. Estimation of g is often very poor (e.g., fig. 4). In general, the dispersal rate

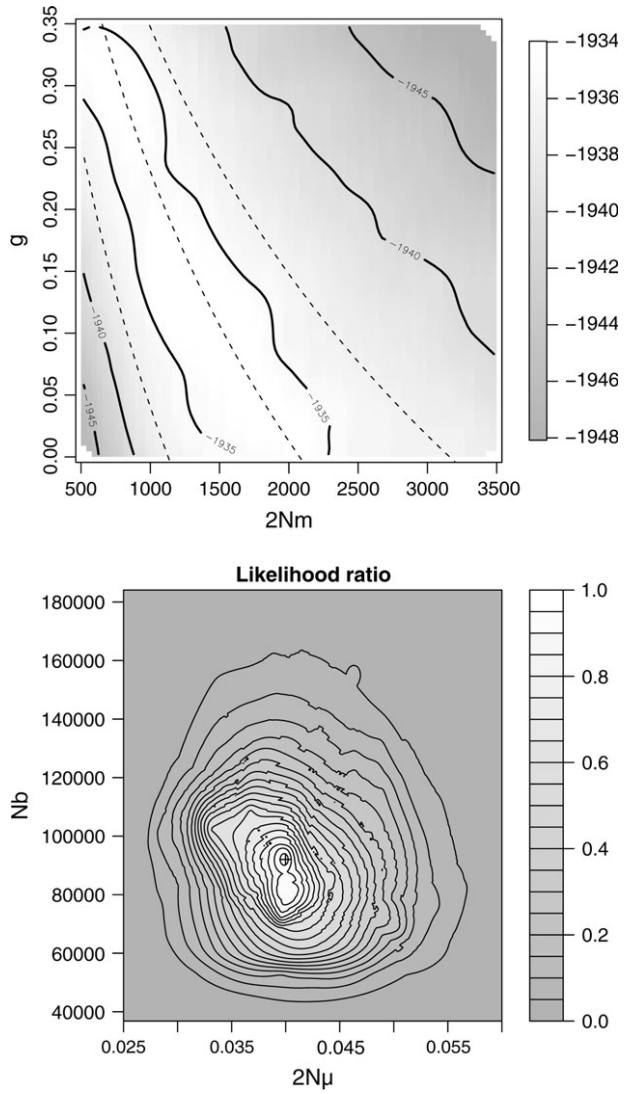


FIG. 5.—PAC likelihood surface for the *Coenagrion mercuriale* data set. Top: a contour plot of likelihood for $2N\mu = 0.4$. Dashed lines are lines of equal Nb values (50,000, 92,039, and 140,000 from left to right). PAC likelihood analysis was computed for 2197 points with $n_t = 100$ (requiring 52 CPU hours on 2.66 GHz processors). Bottom: a profile likelihood ratio plot from the same PAC likelihood computation, derived by profiling over (*Nm*, *g*) values for given neighborhood size (Nb). The likelihood ratio relative to the maximum is shown. Confidence regions are given by the χ^2 approximation for the profile likelihood ratio (Cox and Hinkley, 1974, pp. 322 sqq). With 2 df, the 95% confidence region for ($2N\mu$, Nb) is simply bounded by the 0.05 level for the profile likelihood ratio (i.e., the outer level in this plot), whereas the 95% CI for Nb is given by the 0.1465 level of the 1-dimensional profile. The point estimate is marked by a +.

Nm appears easier to estimate to the extent that ~60% relative biases are deemed acceptable, but larger biases can again result from mis-specification. Large biases occur even though sampled demes exchange essentially no migrants with demes not accounted in the statistical model. Positive biases of *Nm* estimates and negative biases of *g* estimates compensate each other to yield $N\sigma^2$ estimates with low relative MSE. Thus, the largest $N\sigma^2$ MSEs are obtained when *g* is well estimated (i.e., for $g = 0.75$ in tables 6 and 7). Figure 5 also shows that there is more information about $N\sigma^2$ than about *Nm* and *g* separately.

Estimates of the mutation rate $N\mu$ are generally biased upward, to some extent by small sample bias, but in particular when fewer subpopulations are considered in the statistical model than in simulation of the data and when the mutation rate is low. Local diversity contains information about the mutation rate scaled by the total population size (Nagylaki 1983; Slatkin 1987; Strobeck 1987), so the likelihood must depend on the total size of the population. However, the dependence of diversity on total size may be only perceptible for small mutation rates. For high mutation rates and low dispersal, the probability of identity within demes (or a few demes apart) depends little on the total size of the array (see, e.g., comparison of Maruyama's (1970b) finite lattice results to Nagylaki's (1974) infinite lattice results in fig. 2 of Cox and Durrett 2002), so that there may be little statistical information to distinguish between a 40-demes and a 1,000-demes array. This may explain why distribution of $N_T\mu$ estimates appear closer to $Nn_m\mu$ than to the true $N_T\mu$ value for the higher mutation rate (10^{-3}) and lower dispersal, whereas the reverse holds for higher dispersal and lower mutation rate (10^{-4}, table 7).

Similar trends are observed for *Nm* estimates but are not always so easily understood. For high mutation ($\mu = 10^{-3}$), relatively good estimation of *Nm* can be achieved. For $\mu = 10^{-4}$ and large dispersal, no migration rate appears well estimated when deme number is mis-specified. This does show how easily likelihood methods could be misused in realistic conditions.

There would be considerable difficulties, both conceptual and statistical, in trying to estimate the number of demes itself. Over the timescale of genealogical processes, assuming a fixed number of demes is often no more than a convenient device. Even in the ideal case considered in the simulations, comparing the PAC likelihoods of the fitted parameter values under models with different number of demes does not point to good estimates of this number. In the 3 cases from table 7 were the comparison was possible, the data were equally well fitted under the 60-demes model as under the 200-demes model; the fitted PAC likelihoods were, if anything, slightly lower for larger number of demes, thus pointing away from the true value of 1,000 demes.

Because any pure simulation study may miss important factors affecting the performance of estimation, comparisons with demographic estimates are also important to evaluate the possible impact of factors ignored in the simulations and eventually to force us to consider additional factors. In the present case, the maximum PAC likelihood estimate is ~3 times lower than the demographic estimate, and its CI excludes this estimate. As discussed by Watts et al. (2007), the demographic estimate reported in that study is a worst-case overestimate for comparisons with genetic estimates, in that no attempt was made to correct for variations in population density over years. The genetic point estimates differ in a manner consistent with the expected bias of the regression estimators under simulation conditions fitted to the conditions of the population studied, but more importantly the CI obtained by the regression method overlaps widely with the one given by PAC

likelihood. Hence, one explanation consistent with all available evidence is that both genetic methods estimate, with different small-sample biases, the same effective $D\sigma^2$ and that the demographic estimate was too high. Although discrepancies between the different methods (in particular, asymptotic bias) could still be sought, they would be of the order of differences in confidence limits (20% for the lower bound, 145% for the upper bound). Further comparisons would be necessary to demonstrate systematic differences of this magnitude.

We have assumed the same mutation rate for all loci. It is unclear how variation in mutation rate would affect the analyses, and estimating one mutation parameter per locus would be both highly impractical and would increase the MSE of the other estimates. A tentative solution to this problem could be to use a random effect model for mutation, that is, to integrate the likelihood over a distribution of mutation rates, of which some parameters would be estimated.

Predicting Mis-Specification Biases

Although our analysis has highlighted the biases resulting from mis-specification of the number of demes, some of these biases appear small compared with the accuracy sometimes expected (Whitlock and McCauley 1999) from analyses of spatial genetic structure. How far this conclusion will remain true when a wider range of biological scenarios is considered? It would be helpful to be able to predict biases by relatively simple arguments.

In an idealized world, spatial patterns would contain no information about mutation rates, and dispersal rates could be estimated independently of mutation. To some extent, this is what occurs with moment methods based on probabilities of identity of pairs of genes: local diversity depends on the mutation rate but F_{ST} and related quantities are relatively independent of mutation (Crow and Aoki 1984; Slatkin 1991), particularly at a local geographical scale (Rousset 1996). Thus, it is to some extent possible to estimate dispersal rate without good estimates of mutation rates. The present results, as those of Beerli (2004), suggest a similar behavior, in that mutation rate estimation is more affected by mis-specification. However, cases where mis-specification also notably affects estimation of dispersal were also pointed out.

Attempts to estimate simultaneously dispersal and mutation by moment methods could also result in biased estimates of both parameters (an example will be presented below). It is therefore tempting to try to predict the biases of MLEs from the analytical theory for pairs of genes, and the number of demes to be considered in the statistical model might be predicted from such theory. Bias prediction was considered by Slatkin (2005), but the approximations of diversity by expected coalescence times he considered do not describe well genetic identity at loci with high mutation rates. In addition, it may not be possible to fit exactly all probabilities of identity in a large array of demes to a model with few parameters. Validating any prediction procedure is bound to be complex.

Nevertheless, the simple example of the island model can be used to support such a logic. A way of predicting biases in the island model is to compute expected values of within- and among-deme probabilities of identity for the actual number of population and to find the numerical values of the mutation and migration rates which would give the same probabilities of identity for the assumed number of demes in the estimation model. These computations are straightforward (e.g., Nagylaki 1983; Rousset 2004, pp. 27, 224).

Thus, in the demographic conditions of case [51] ($n_d = 100$ demes, $N = 400$, $m = 0.01$), but for an island model of dispersal, if the mutation probability is 10^{-4} (for a KAM with 10 alleles) the probabilities of identity within and among demes are 0.269 and 0.181, and the mutation and migration probabilities which yield the same probabilities in a 10-demes model are 9.1×10^{-4} and 0.0083. The predicted relative biases are therefore 8.1 and −0.17. The observed biases were close: 8.8 and −0.20 out of 60 replicate samples of 5 loci (further simulation details not shown). The observed biases are thus well predicted and close to those of a moment method using the information contained in probabilities of identity. Likewise, if the mutation probability is 10^{-3}, the probabilities of identity within and among demes are 0.196 and 0.108, and the mutation and migration probabilities which yield the same probabilities in a 10-demes model are 0.0053 and 0.0048, yielding predicted relative biases 4.3 and −0.52, respectively. The observed biases were again close: 4.04 and −0.46 out of 60 replicate samples of 5 loci.

Beyond illustrating a case where the probabilities of identity provide good prediction of MLE biases, these examples also illustrate the simple expectation, consistent with the other simulation results, that the relative bias on mutation estimation will be of the order of the n_d/n_m ratio when the mutation rate is low and lower for higher mutation rates. In the latter case, however, the bias on the migration rate can be large and not so easily interpreted.

Therefore, a higher number of demes might need to be considered for lower mutation rates, which could be a serious practical problem for some types of markers. The comparison of $N\mu$ biases in table 6 versus table 7 supports this idea. Local diversity is more sensitive to total size when dispersal is less localized (higher m or g values). So, by the same logic, a higher number of demes should be considered when dispersal rates are higher, which is indeed observed in our simulations. Mis-specification effects could be important in 2-dimensional applications and more generally when the probability of identity is more dependent on the total size of the population than in a linear habitat.

Finally, the variation in local diversity in KAM versus SMMs is at most that resulting from a 2-fold variation in mutation rate (Rousset 1996), so one could expect the mutation model to have little impact on estimator performance beyond an at most 2-fold effect on mutation rate estimation, which is indeed what was observed when stepwise mutation data were analyzed under a KAM statistical model.

Conclusion

The present work has shown that ML can be applied to allelic type data from moderately large networks of populations. Maximum PAC likelihood is of potential utility for

larger networks. Its performance was practically identical to that of ML estimation and even superior in most cases for an identical computation effort. The current implementation effectively allows ML analyses of systems of $n_m = 10$ demes in a few hours, and maximum PAC likelihood analyses of larger arrays (up to 200 demes in this study) can yield reasonably accurate estimates within a week. When the true number of demes is unknown, the assessment of performance yields mixed results. The number of demes that has to be considered in the statistical model to achieve good performance depends on the scale of dispersal and the mutation rate, which may limit the range of realistic applications. Mis-specification biases for mutation rates are relatively easily understood but less so for dispersal parameters. The composite parameter $D\sigma^2$ was relatively little affected by mis-specification of the number of demes, but it may be difficult to overcome mis-specification biases in the estimation of other dispersal parameters.

Acknowledgments

This study was made possible by access first to the computing facilities of the CINES (Montpellier, France), then to a PC cluster of the University of Montpellier 2, and finally to the ISEM cluster. We thank J.-B. Ferdy for substantial help in using this cluster, as well as V. Ranwez, K. Belkhir, and J. Maizi. The MNHN cluster was also used. We thank J.-M. Cornuet for access to his unpublished work. This is publication ISEM 07-119.

Literature Cited

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Ann Rev Ecol Syst. 33:707–740.

Bahlo M, Griffiths RC. 2000. Inference from gene trees in a subdivided population. Theor Popul Biol. 57:79–95.

Barton NH, Gale KS. 1993. Genetic analysis of hybrid zones. In: Harrison RG, editor. Hybrid zones and the evolutionary process. Oxford: Oxford University Press. p. 13–45.

Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol Ecol. 13:827–836.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics. 22:341–345.

Beerli P, Felsenstein J. 1999. Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. 152:763–773.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. Proc Natl Acad Sci USA. 98:4563–4568.

Broquet T, Johnson CA, Petit É, Burel F, Fryxell JM. 2006. Dispersal kurtosis and genetic structure in the American marten, *Martes americana*. Mol Ecol. 15:1689–1697.

Cornuet JM, Beaumont MA. 2007. A note on the accuracy of PAC-likelihood inference with microsatellite data. Theor Popul Biol. 71:12–19.

Cox DR, Hinkley DV. 1974. Theoretical statistics. London: Chapman & Hall.

Cox JT. 1989. Coalescing random walks and voter model consensus times on the torus in $\mathbb{Z}^d$. Ann Probab. 17:1333–1366.

Cox JT, Durrett R. 2002. The stepping stone model: new formulas expose old myths. Ann Appl Probab. 12:1348–1377.

Cressie NAC. 1993. Statistics for spatial data. New York: Wiley.

Crow JF, Aoki K. 1984. Group selection for a polygenic behavioural trait: estimating the degree of population subdivision. Proc Natl Acad Sci USA. 81:6073–6077.

de Iorio M, Griffiths RC. 2004a. Importance sampling on coalescent histories. Adv Appl Probab. 36:417–433.

de Iorio M, Griffiths RC. 2004b. Importance sampling on coalescent histories. II. Subdivided population models. Adv Appl Probab. 36:434–454.

de Iorio M, Griffiths RC, Leblois R, Rousset F. 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theor Popul Biol. 68:41–53.

DiCiccio TJ, Efron B. 1996. Bootstrap confidence intervals (with discussion). Stat Sci. 11:189–228.

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. 16:551–558.

Estoup A, Rousset F, Michalakis Y, Cornuet JM, Adriamanga M, Guyomard R. 1998. Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). Mol Ecol. 7:339–353.

Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. Genetics. 159:1299–1318.

Fenster CB, Vekemans X, Hardy OJ. 2003. Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). Evolution. 57:995–1007.

Fields Development Team. 2006. Fields: tools for spatial data. Boulder (CO): National Center for Atmospheric Research. http://www.cgd.ucar.edu/software/fields.

Golub GH, van Loan CF. 1996. Matrix computations. Baltimore (MD): John Hopkins University Press. 3rd ed.

Goudet J, Raymond M, de Meeüs T, Rousset F. 1996. Testing differentiation in diploid populations. Genetics. 144:1931–1938.

Gusmão L, Sánchez-Diz P, Calafell F, et al. (42 co-authors). 2005. Mutation rates at Y chromosome specific microsatellites. Hum Mutat. 26:520–528.

Herbots HM. 1997. The structured coalescent. In: Donnelly P, Tavaré S, editors. Progress in population genetics and human evolution. New York: Springer-Verlag. pp. 231–255.

Kimura M, Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. Genetics. 49:561–576.

Leblois R, Estoup A, Rousset F. 2003. Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. Mol Biol Evol. 20:491–502.

Leblois R, Rousset F, Estoup A. 2004. Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. Genetics. 166:1081–1092.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 165:2213–2233. Correction: 167: 1039.

Malécot G. 1975. Heterozygosity and relationship in regularly subdivided populations. Theor Popul Biol. 8:212–241.

Maruyama T. 1970a. Effective number of alleles in a subdivided population. Theor Popul Biol. 1:273–306.

Maruyama T. 1970b. Stepping stone models of finite length. Adv Appl Probab. 2:229–258.
Meng XL, Wong WH. 1996. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Stat Sin. 6:831–860.
Nagylaki T. 1974. The decay of genetic variability in geographically structured populations. Proc Natl Acad Sci USA. 71:2932–2936.
Nagylaki T. 1983. The robustness of neutral models of geographical variation. Theor Popul Biol. 24:268–294.
Notohara M. 1990. The coalescent and the genealogical process in geographically structured population. J Math Biol. 29:59–75.
Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1988. Numerical recipes in C. Cambridge: Cambridge University Press.
R Development Core Team. 2004. R: A language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing. [Internet]. [cited 2007 Oct 19]. http://www.r-project.org.
Raufaste N, Bonhomme F. 2000. Properties of bias and variance of two multiallelic estimators of F_{ST}. Theor Popul Biol. 57:285–296.
Rousset F. 1996. Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics. 142:1357–1362.
Rousset F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics. 145:1219–1228.
Rousset F. 2000. Genetic differentiation between individuals. J Evol Biol. 13:58–62.
Rousset F. 2004. Genetic structure and selection in subdivided populations. Princeton (NJ): Princeton University Press.
Rousset F. 2007. Inferences from spatial population genetics. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics. Chichester, UK: Wiley. pp. 945–979.
RoyChoudhury A, Stephens M. 2007. Fast and accurate estimation of the population-scaled mutation rate, θ, from microsatellite genotype data. Genetics. 176:1363–1366.
Sacks J, Welch WJ, Mitchell TJ, Wynn HP. 1989. Design and analysis of computer experiments. Stat Sci. 4:409–435.
Sawyer S. 1977. Asymptotic properties of the equilibrium probability of identity in a geographically structured population. Adv Appl Probab. 9:268–282.
Slatkin M. 1987. The average number of sites separating DNA sequences drawn from a subdivided population. Theor Popul Biol. 32:42–49.
Slatkin M. 1991. Inbreeding coefficients and coalescence times. Genet Res. 58:167–175.
Slatkin M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution. 47:264–279.
Slatkin M. 1994. Gene flow and population structure. In: Real LA, editor. Ecological Genetics. Princeton (NJ): Princeton University Press. pp. 3–17.
Slatkin M. 2005. Seeing ghosts: the effect of unsampled populations on migration rates estimated between sampled populations. Mol Ecol. 14:67–73.
Stephens M, Donnelly P. 2000. Inference in molecular population genetics (with discussion). J R Stat Soc. 62:605–655.
Strobeck C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics. 117:149–153.
Sumner J, Estoup A, Rousset F, Moritz C. 2001. 'Neighborhood' size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. Mol Ecol. 10:1917–1927.
Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC, Doebley J. 2002. Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol. 19:1251–1260.
Waples RS, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol. 15:1419–1439.
Watts PC, Rousset F, Saccheri IJ, Leblois R, Kemp SJ, Thompson DJ. 2007. Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. Mol Ecol. 16:737–751.
Welch WJ, Buck RJ, Sachs J, Wynn HP, Mitchell TJ, Morris MD. 1992. Screening, prediction, and computer experiments. Technometrics. 34:15–25.
Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration: $F_{st} \neq 1/(4Nm + 1)$. Heredity. 82: 117–125.
Wilkins JF. 2004. A separation-of-timescales approach to the coalescent in a continuous population. Genetics. 168: 2227–2244.
Winters JB, Waser PM. 2003. Gene dispersal and outbreeding in a philopatric mammal. Mol Ecol. 12:2251–2259.
Wood JW, Smouse PE, Long JC. 1985. Sex-specific dispersal patterns in two human populations of highland New Guinea. Am Nat. 125:747–768.
Zähle I, Cox JT, Durrett R. 2005. The stepping stone model, II: genealogies and the infinite sites model. Ann Appl Probab. 15:671–699.

Marcy Uyenoyama, Associate Editor

Accepted September 18, 2007

Likelihood-Based Inferences under Isolation by Distance: Two-Dimensional Habitats and Confidence Intervals

François Rousset*[,1] and Raphaël Leblois[2]

[1]Institut des Sciences de l'Evolution (UM2-CNRS), Université Montpellier 2, Montpellier, France

[2]Institut National de la Recherche Agronomique, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, Montferrier-sur-Lez, France

***Corresponding author:** E-mail: francois.rousset@univ-montp2.fr.

Associate editor: Noah Rosenberg

Abstract

Likelihood-based methods of inference of population parameters from genetic data in structured populations have been implemented but still little tested in large networks of populations. In this work, a previous software implementation of inference in linear habitats is extended to two-dimensional habitats, and the coverage properties of confidence intervals are analyzed in both cases. Both standard likelihood and an efficient approximation are considered. The effects of misspecification of mutation model and dispersal distribution, and of spatial binning of samples, are considered. In the absence of model misspecification, the estimators have low bias, low mean square error, and the coverage properties of confidence intervals are consistent with theoretical expectations. Inferences of dispersal parameters and of the mutation rate are sensitive to misspecification or to approximations inherent to the coalescent algorithms used. In particular, coalescent approximations are not appropriate to infer the shape of the dispersal distribution. However, inferences of the neighborhood parameter (or of the product of population density and mean square dispersal rate) are generally robust with respect to complicating factors, such as misspecification of the mutation process and of the shape of the dispersal distribution, and with respect to spatial binning of samples. Likelihood inferences appear feasible in moderately sized networks of populations (up to 400 populations in this work), and they are more efficient than previous moment-based spatial regression method in realistic conditions.

Key words: dispersal, maximum likelihood, coalescence, isolation by distance, microsatellites.

Introduction

Accurate estimation of dispersal in natural populations by demographic observations is difficult, which has led to the development of many methods to infer dispersal from genetic information. Recent developments include some applications of assignment techniques (Wilson and Rannala 2003; Paetkau et al. 2004; Faubet and Gaggiotti 2008), methods based on simulation of the distribution of summary statistics (such as so-called approximate Bayesian computation, e.g., Beaumont 2007 applied to dispersal estimation in Hamilton et al. 2005 and Becquet and Przeworski 2007) and likelihood methods (Rannala and Hartigan 1996; Beerli and Felsenstein 1999, 2001; de Iorio and Griffiths 2004b) that aim to use all information in the data.

These methods seem to perform well for low migration rates between a small number of populations, but their performance is more generally uncertain. For example, the evaluation of likelihood remains time consuming, so that likelihood methods have been tested only for small networks of populations, and the reliability of the computations is sometimes debated (Abdo et al. 2004; Beerli 2006). Further, all methods may rest on questionable assumptions. For example, it has been found that "ghost" populations unaccounted in the statistical model can affect maximum-likelihood (ML) estimation of dispersal and mutation parameters of sampled populations (Beerli 2004; Rousset and Leblois 2007). Thus, perennial questions (e.g., Cox 2006, p. 170) about the benefits of likelihood analyses relative to alternative methods remain pending.

Application of full-likelihood methods to the scenario of localized dispersal or "isolation by distance," relevant for many ecological studies, has only been considered in Rousset and Leblois (2007), and alternative methods are still being developed (e.g., rare allele methods, Novembre and Slatkin 2009). Rousset and Leblois (2007) described the properties of point estimates of dispersal and mutation parameters in linear habitats. The evaluation of likelihood was based on the algorithms of de Iorio and Griffiths (2004b). As evaluation of likelihood performance was time consuming, a fast heuristic approximation known as product of approximate conditional likelihood (PAC-likelihood, Li and Stephens 2003) was also considered. Inferences from PAC-likelihood surfaces appeared practically as efficient (precise) as full-likelihood inferences, even though the PAC-likelihood is a biased estimate of the likelihood for each parameter point. In the present work, these results are extended to two-dimensional habitats. Further, the performance of likelihood-based confidence intervals is analyzed.

The following general features are shared with and further discussed in Rousset and Leblois (2007): Allelic type data will be considered, with microsatellite data as the intended subject of application. We envision many species as spatial clusters of subpopulations with large immigration probabilities within each cluster, and less dispersal among

clusters, and we are interested in the analysis of one such cluster. Its structure is described as a regular array of demes of size N for which we estimate the following parameters: a mutation rate, a number of immigrants per deme (Nm), and a dispersal scale parameter (that of a geometric distribution). We also consider the neighborhood size or equivalently the product of population density and mean square dispersal distance, the latter being a function of the two previous parameters. We will compare performance of neighborhood estimation to that of variants of the method based on regression of F_{ST} estimates to geographical distance (e.g., Rousset 1997; Watts et al. 2007).

Evaluation of performance involves both evaluation under ideal conditions where the data are generated under the model used as a basis for statistical analysis and evaluation of robustness under model misspecification (e.g., Casella and Berger 2002, p. 481). In this paper, we consider both steps. We first evaluate performance under nearly ideal conditions (known mutation model, known dispersal distribution), in particular to demonstrate that we have an effective implementation of likelihood inferences. Overall, the estimation performance may be considered excellent, with good coverage of the confidence intervals, and generally small biases and small mean square errors. We nevertheless obtain some nonideal results and show that they are inherent to the statistical method rather than a feature of our implementation. More specifically, the algorithm used to estimate likelihood is based on coalescent approximations, that is, approximations for large deme size, small migration, and small mutation probability. When applied to samples from finite-sized populations, the statistical model thus always appears misspecified except in the case of vanishing migration rate between arbitrarily large populations, a case that may be of limited practical interest. The coalescent approximation affects the results, as estimates of dispersal parameters (number of migrants and the shape parameter of the dispersal distribution) are biased when the dispersal probability is large. Neighborhood estimation may be more robust in this respect. We also compare strict likelihood and PAC-likelihood inferences and find that their performance are practically equivalent.

In a second step, we evaluate performance of PAC-likelihood inferences under conditions including misspecification of the dispersal distribution and of mutation model, and otherwise designed to approximate realistic conditions, based on the study of damselfly populations by Watts et al. (2007). We consider the effect of spatial binning of samples, as such binning is necessary to fit data from individuals that can be sampled from anywhere in continuous space, to the framework of the statistical model that assumes a regular grid of demes. As computations are also faster for small arrays of demes, a coarse-grained spatial binning of samples can also reduce the computation load compared with a fine-grained one. But it can also induce biases or results that are difficult to interpret. Finally, we compare neighborhood size estimation to that achieved by previous methods and conclude that likelihood-based estimation can perform better in practical conditions.

Methods

For each simulated data set, the analysis goes through three main steps, implemented in the software Migraine and further described in the Appendix. First, likelihoods are estimated, with some error, for a number of parameter points. Next, a likelihood surface is inferred from the likelihood points by a classical smoothing method (Kriging). Third, parameter values of interest (the mutation and dispersal parameters used to generate the data) are tested by profile likelihood ratio tests (profile LRTs, e.g., Cox and Hinkley 1974; Severini 2000). Profile LRTs also allow the construction of profile likelihood confidence intervals. Ideally, the main measures of the quality of inference are the coverage properties of such confidence intervals for given parameter values. Note that this differs from coverage averaged over a prior distribution of parameter values, as measured in some studies (Beerli 2006; Hey 2010; Peter et al. 2010). Only the demonstration of good coverage for fixed parameter values ensures good average coverage for any imperfectly known prior distribution or for any prior information in the form of a likelihood surface. The coverage properties of confidence intervals, for given parameter values, can be assessed through the distribution of the P value of the corresponding profile LRTs. Ideally, this distribution is uniform; but this comfortable ideal is rarely attained in practice and then some consideration of the practical importance of the biases is useful in assessing the method.

In this section, we detail the basic assumptions of the sample simulation model and of the statistical model. In the Appendix, we further detail the implementation of the statistical model and the method of inference of likelihood surfaces.

Dispersal Models for Sample Simulation

Samples have been simulated by the IBDsim program (Leblois et al. 2009). Two dispersal distributions have been considered, a geometric dispersal model similar to the one of the statistical model and the Poisson reciprocal gamma model (Chesson and Lee 2005). The latter distribution is Gaussian-looking at short distances, but power-tailed, and can therefore have a high kurtosis. Its two parameters $\gamma < 0$ and κ determine the power $\gamma - 1$ of the tail, and the second moment $\sigma^2 = -\kappa/[2(1+\gamma)]$. We vary σ^2 in our simulations by varying κ for fixed $\gamma = -2.15$, whereby the axial kurtosis varies between 20.1 and 22.5.

Exact Control of Number of Migrants

Absorbing boundaries are assumed, so that the demes near edges typically receive fewer immigrants since they have fewer close neighboring demes. The actual number of immigrants thus differs from the number of emigrants deduced from the forward distribution. Such discrepancies are easily detected by the statistical estimation of number of immigrants. Then, one needs to control the number of immigrants in the sample simulation rather than simply let it be a complex function of the forward distribution and of habitat edge effects. Hence, in both sample simulations based on the geometric distribution and in statistical analysis, the

Nm parameter is defined to give the maximal (over demes) expected number of immigrants in a deme whatever the edge effects. This differs from simulations in Rousset and Leblois (2007) and will be included as an option in future versions of the IBDsim program. On the other hand, no attempt was made to control *Nm* values in our sample simulations based on the Poisson reciprocal gamma distribution. In the latter simulations, the immigration rate in most demes was > 0.8, a situation where no demic structure would be recognized in practice.

Geometric Dispersal

The scale parameter g describes the geometric decrease, with distance between demes, of the pairwise forward immigration probabilities. In two dimensions, forward probabilities decrease according to relative values of $g^{|x|+|y|}/[(1+\delta_{x0})(1+\delta_{y0})]$, x and y being the axial dispersal distances in each dimension (not both zero), and $\delta_{ij} = 1$ if $i = j$ and $= 0$ otherwise (Kronecker's notation). As described above, the forward dispersal probability is adjusted such that the maximal expected number of immigrants in a deme has a known preset value and that the deme of origin of immigrants is chosen according to the relative values of the forward dispersal probabilities.

The Neighborhood Parameter

The classical neighborhood size parameter is defined as $\mathrm{Nb} \equiv 2D\sigma^2$ in linear habitats and $\mathrm{Nb} \equiv 2D\pi\sigma^2$ in two-dimensional habitats, where in this paper, D is a density of haploid equivalents (in the same way as N is a number of haploid equivalents). For geometric dispersal, the $D\sigma^2$ term is deduced from Nm and g, as $2Nm\sigma^2_{\mathrm{cond}}$, where σ^2_{cond} is the second moment, in unbounded space, of axial dispersal distance conditional on dispersal. Thus, for two-dimensional habitats,

$$\sigma^2_{\mathrm{cond}} = \frac{1}{2}\frac{\sum_{x=0}^{x=\infty}\sum_{y=0}^{y=\infty}(x^2+y^2)g^{x+y}}{\sum_{x=0}^{x=\infty}\sum_{y=0}^{y=\infty}g^{x+y}-1} = \frac{1+g}{(2-g)(1-g)^2} = \frac{\mathrm{Nb}}{2\pi Nm} \quad (1)$$

and for linear habitats,

$$\sigma^2_{\mathrm{cond}} = \frac{\sum_{x=1}^{x=\infty}x^2g^x}{\sum_{x=1}^{x=\infty}g^x} = \frac{1+g}{(1-g)^2} = \frac{\mathrm{Nb}}{2Nm}. \quad (2)$$

Note that Nb depends on the distance unit in linear habitats and that the above equation only holds if the distance unit is one lattice step in the statistical model.

Dispersal in the Statistical Model

A geometrical dispersal model is assumed in likelihood computations. Its exact meaning differs from that of the geometrical dispersal model assumed in sample simulations. In the likelihood computations, g describes the decrease of the expected number of immigrants with distance, whereas in the sample simulation, g describes the decrease in forward immigration rates. Such discrepancies cannot be generally avoided because the likelihood computations are based on a limit process where all dispersal probabilities among different demes are infinitesimally small and considers only one parameter Nm where the sample simulation considers the two parameters N and m separately.

In particular, the edge effects cannot be treated identically in both algorithms. In the likelihood computations, we assumed the number of immigrants between pairs of demes is a function of their relative position only and not of their position relative to the edge of the habitat; whereas in the sample simulation algorithm, it is determined by computation of backward dispersal probabilities from forward probabilities (as is usual), and this depends on the position of the two demes relative to the edges of the lattice. Further details and a numerical example are given in the Appendix, illustrating that the discrepancies between the two algorithms may be small.

Mutation Models

The default mutation model considered in sample simulations was a symmetric K-alleles model (KAM) with 10 alleles. A one-step stepwise mutation model (SMM), also with 10 alleles, was also considered in some sample simulations. The KAM was assumed in all likelihood computations.

Sampling Design

Two-hundred data sets are analyzed for each simulation condition. Each data set includes 10 independent loci. In the two-dimensional case, square habitats of 4×4 or 10×10 populations are simulated, and 10 diploid individuals are sampled at each of 8 demes, two in each corner, that is, at positions (1,1) and (2,2) in one corner and symmetrically in the other corners. In this way, both adjacent and distant populations are sampled, which should facilitate estimation of the scale of dispersal. On the other hand, this design may highlight edge effects. In linear-habitat cases, samples of 10 individuals were taken in each subpopulation for arrays of four populations; at positions 2–8 (or 2–16, cases [5], [6], [25], [26]) by steps of 2 for 16 populations; and at positions 40–58 by steps of 2 for 100 populations.

Results

Minimal Misspecification

Full Likelihood

The correctness of the confidence intervals can be examined graphically by looking whether the empirical cumulative distribution function of P values aligns (or not) with a 1:1 diagonal line. These distributions are shown for all simulation conditions in the Supplementary Material online for details. Figure 1 (left) illustrates a good result. Deviations from the diagonal are tested by the Kolmogorov–Smirnov test ("KS" inset in each subplot). Four subplots are presented, one for each of the canonical parameters and one for $D\sigma^2$. Also shown below each subplot are the relative (except for g) bias and root mean square error (RMSE) of each ML estimator (the same numbers are reported in table 1, case [1]). It may be observed, and this will also be true when confidence intervals have incorrect coverage that the bias and RMSE of $N\mu$ and Nm are small by practical standards. The $D\sigma^2$

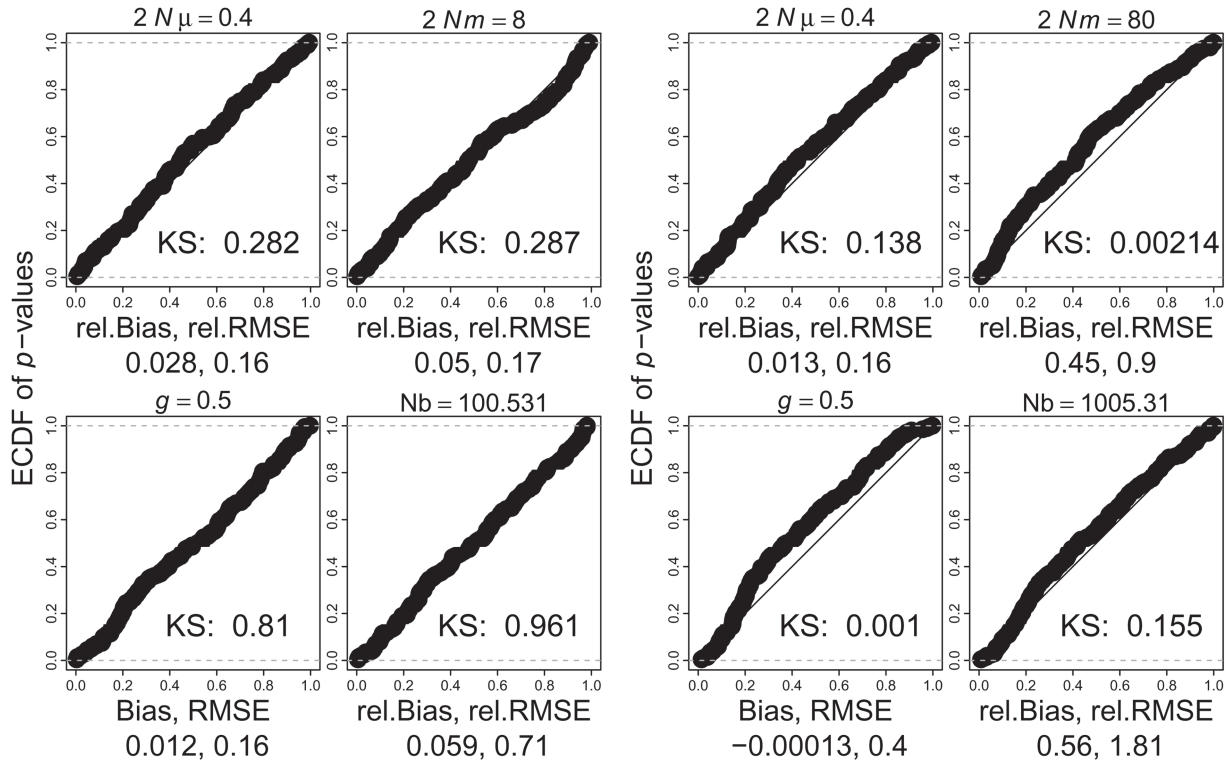


FIG. 1. Distributions of *P* values of likelihood ratio tests in cases [1] (left) and [2] (right).

relative bias and RMSE can be very large. This will typically occur when the data show no evidence of isolation by distance (and therefore when arbitrarily large $D\sigma^2$ estimates may be obtained). In general, the distribution of $1/(D\sigma^2)$ estimates is much closer to Gaussian, which make comparisons of bias and RMSE more meaningful. For this reason, the relative bias and relative RMSE reported in figures and tables are those of 1/Nb.

Figure 1 (right; case [2] in table 1) presents a less satisfying result. The only difference with the previous example is that *m* is 0.1 rather that 0.01. In this and further simulations, there are three possible sources of nonideal performance inherent to the statistical model: 1) departure from coalescence assumptions (*m* being large or *N* being small); 2) spatial edge effects: They are expected when *m* is large, and *g* is intermediate (for low *g*, immigration probabilities are affected only in the outermost demes; for $g = 1$, the sample simulation model and the statistical model are the island model, both with the same immigration rate, so the edge effects are correctly specified). For given number of demes, edge effects should also be most visible in two-dimensional lattices because a higher fraction of populations are at the edge of the habitat; 3) estimates are at the boundary of the parameter space. This can occur for *g* and then the expected distribution of LRT *P* values is not uniform. Not only LRTs for *g* but also for other parameters can be affected (Self and Liang 1987).

The first two effects should disappear as *N* increases and *m*, μ decrease for fixed *Nm*, $N\mu$. The first effect (departure from coalescent assumptions for high *m*) is best singled out under an island model, that is, when *g* is fixed to 1 in sample simulation and in statistical analyses. These simulations clearly show better inferences of a fixed *Nm* value with *N* increasing from 80 to 40, 000 and *m* decreasing from $m = 0.5$ to $m = 0.001$ (cases [3] vs. [4]). To illustrate what these changes in RMSE mean, fig. 2 shows the likelihood surfaces for the samples that yielded departures from parameter values closest to the RMSE values and of the same sign as the bias.

Under isolation by distance, the effect of the coalescent approximation is illustrated by comparison of cases [14] and [2] (*N* increasing from 40 to 40,000) and by comparison of cases [5] and [6] (*N* increasing from 400 to 40,000), although in both comparisons the third effect (*g* estimates at the boundary) may also affect performance more strongly when *m* is larger. Figure 3 shows the convergence of distributions of *P* values to uniform distributions in the last comparison. The same convergence is observed is the two previous comparisons (see Supplementary Material online for distributions of *P* values).

We can roughly rank different simulations according to the expected magnitude of the different effects from lowest to highest. Low *m* values are illustrated by cases [7]–[13] and [19], and the estimator biases are indeed small.

For $g = 0$ (stepping stone model, cases [12] and [13]), the distribution of the LRT for $g = 0$ is expectedly not uniform. The theoretical asymptotic distribution of the LRT *P* value is a mixture 1:1 of a χ^2 with 1 degree of freedom and of a probability mass at 0. The observed mass at 0 actually departs from 1/2 (see cases [12] and [13] in Supplementary Material online for details), which is a general phenomenon (e.g., Pinheiro and Bates 2000, p. 87; Hey 2010). The profile

Table 1. Performance of Estimation by Strict Likelihood.

	Parameters					Relative $N\mu$			Relative Nm			g			Relative 1/Nb		
	Array	N	m	g	μ	Bias	RMSE	KS Test	Bias	RMSE	KS Test	Bias	RMSE	KS Test	Bias	RMSE	KS Test
[1]	4 × 4	400	0.01	0.5	5×10^{-4}	0.028	**0.16**	0.28	0.05	0.17	0.29	0.012	0.16	0.81	0.059	0.71	0.96
[2]	4 × 4	400	0.1	0.5	5×10^{-4}	0.013	**0.16**	0.14	0.45	0.9	0.0021	−0.00013	0.4	0.001	0.56	1.81	0.15
[3]	4 × 4	40,000	0.001	1	5×10^{-6}	0.023	**0.16**	0.61	0.18	0.59	0.75						
[4]	4 × 4	80	0.5	1	0.0025	0.02	**0.16**	0.79	2.26	2.6	0						
[5]	**16**	40	0.25	0.25	0.001	0.0091	**0.17**	0.52	0.94	1.09	0	−0.16	0.2	0	−0.07	0.2	0.29
[6]	**16**	40,000	0.00025	0.25	1×10^{-6}	0.011	**0.18**	0.21	0.15	0.41	0.99	−0.023	0.14	0.78	−0.0038	0.25	0.89
[7]	**16**	400	0.01	0.25	0.001	0.018	**0.15**	0.2	0.061	0.26	0.05	−0.0055	0.14	0.39	0.013	0.33	0.62
[8]	**16**	400	0.01	0.5	0.001	0.041	**0.17**	0.63	0.069	0.19	0.058	−0.017	0.11	0.097	0.074	0.45	0.38
[9]	4 × 4	400	0.01	0.25	0.001	0.055	**0.19**	0.022	0.04	0.19	0.072	0.0066	0.16	0.31	0.04	0.5	0.83
[10]	4 × 4	400	0.01	0.5	0.001	0.045	**0.18**	0.2	0.034	0.17	0.42	−0.0021	0.16	0.79	0.18	0.89	0.71
[11]	4 × 4	400	0.01	0.75	0.001	0.022	**0.17**	0.11	0.051	0.18	0.44	0.0027	0.19	0.0052	0.75	2.51	0.0029
[12]	**4**	400	0.01	1×10^{-4}	1×10^{-4}	-9×10^{-4}	**0.17**	0.31	−0.0079	0.22	0.58	0.056	0.12	0	−0.096	0.27	0.83
[13]	4 × 4	400	0.01	1×10^{-4}	0.001	0.039	**0.18**	0.18	−0.028	0.14	0.2	0.048	0.096	0	−0.098	0.23	0.34
[14]	4 × 4	40,000	0.001	0.5	5×10^{-6}	0.02	**0.16**	0.57	0.22	0.68	0.58	0.007	0.35	0.32	0.52	1.66	0.31
[15]	4 × 4	40	0.05	0.25	0.001	0.023	**0.18**	0.53	0.11	0.23	0.56	0.016	0.15	0.49	−0.064	0.44	0.066
[16]	4 × 4	40	0.05	0.5	0.001	0.029	**0.19**	0.21	0.14	0.25	0.0098	0.0062	0.17	0.85	0.034	0.69	0.81
[17]	**16**	400	0.01	0.75	0.001	0.025	**0.18**	0.017	0.029	0.14	0.16	−0.0056	0.081	0.76	0.13	0.67	0.96
[18][a]	**100**	40	0.5	0.5	5×10^{-4}	0.045	**0.17**	0.29	2.39	2.73	0	−0.28	0.31	1×10^{-9}	−0.06	0.24	0.53
[19]	10 × 10	400	0.01	0.5	5×10^{-5}	0.027	**0.16**	0.75	0.0092	0.15	0.019	0.0034	0.093	0.53	0.033	0.4	0.99
[20]	4 × 4	400	0.01	0.99999	0.001	0.041	**0.19**	0.34	0.073	0.17	0.076	−0.051	0.1	2.5×10^{-8}	1.3×10^{8}	4.4×10^{8}	5.9×10^{-9}

[a]For case [18], only 30 samples were analyzed.

LRTs for the other parameters appear unaffected, but this is not a general expectation (Self and Liang 1987).

When g approaches 1 (and neighborhood size approaches infinity), the same parameter boundary effects on g are encountered (case [20]; similar results are also obtained with 50 loci by PAC-likelihood, not shown). Further, numerical issues affect tests of large values of the neighborhood size ($\sim 10^{11}$ for $g = 0.99999$). A way to circumvent this problem is to change the scale of uniform sampling of parameter points and of Kriging (i.e., uniform sampling of σ^2_{cond}, see Appendix). Although this solves most of the numerical issues, the distribution of P values for Nb is distorted in the same way as that for g.

Conversely, the highest biases are expected for high m values (fig. 4). The largest Nm bias in tables 1 and 2 is for $m = 0.5$ in a linear array of 100 demes (case [18]), and other cases with $m \geqslant 0.1$ show large distortions of the distribution of P values. For intermediate m values ($0.01 \leqslant m < 0.1$) relatively large Nm biases may still be observed, but distorsions of P value distributions are generally less obvious, except in some cases where misspecification of spatial edge effects can also contribute (in particular, case [16]).

PAC-likelihood

The PAC-likelihood approximation can easily be compared with the likelihood analysis when the latter is feasible (cases [21]–[40] in the same order as comparable strict likelihood analyses [1]–[20]). In all cases, their performance is very similar, except that PAC-likelihood estimates of the mutation rate appear unbiased or downward biased while strict likelihood ones show a slight positive bias (fig. 5). Some additional PAC-likelihood analyses were considered for 10 × 10 lattices (cases [41]–[43]) and demonstrate good performance. Case [42] is identical to case [22] except that a larger array was considered. Expectedly, the spatial edge effects are reduced and indeed no longer apparent in this case.

Misspecification Effects and Comparison with Moment-Based Method

In this section, we consider three sources of misspecification: the spatial binning of samples, the mutation process at marker loci, and the shape of the dispersal distribution. We also compare the performance of likelihood-based inference to a simple regression method for estimation of neighborhood in such conditions of misspecification.

The algorithms considered in this work rest on the definition of distinct demes. However, in natural populations, individuals are not clearly clustered in demes. It is tempting to analyze such populations as made of a large number of small breeding patches though there are computational limits to the number of demes that can be considered in practice. A straightforward method of clustering is according to regular spatial bins. It is therefore necessary to know how such a clustering can affect inferences. In particular, it is not necessarily obvious what are the parameter values to be estimated (the estimands) from the binned data.

For samples from a regular array, a putative estimand for Nm is the number of immigrants in each spatial bin, that is,

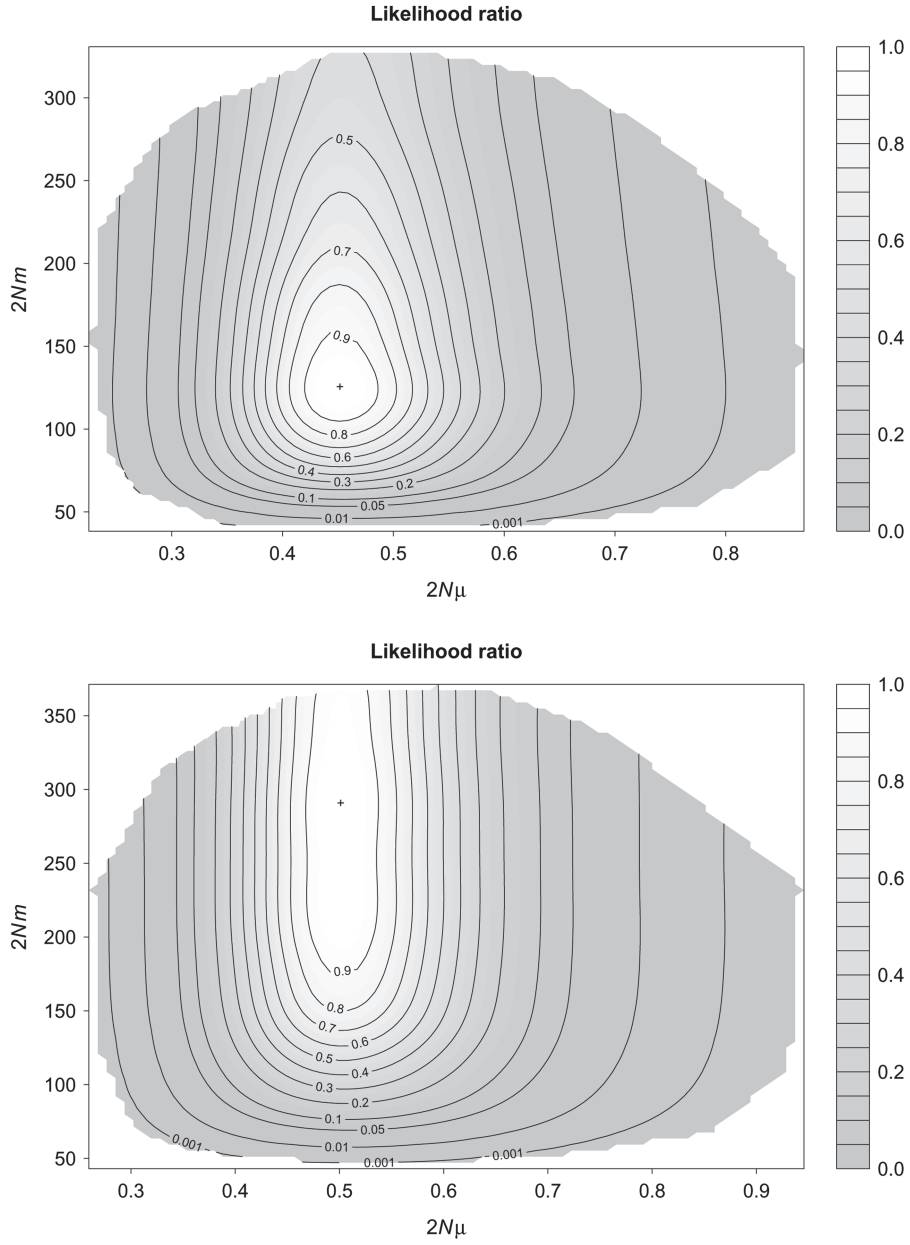


FIG. 2. Examples of likelihood surfaces for cases [3] (top) and [4] (bottom). The surfaces are inferred from 1,024 points as described in the Appendix. In both cases, the parameter values where $2N\mu = 0.4$ and $2Nm = 80$, but the bottom case illustrates the much higher RMSE of $2Nm$ estimates for low N, large m cases. The likelihood surface is shown only for parameter combinations that fell within the envelope of parameter points for which likelihoods were estimated. The cross denotes the maximum.

the sum of the numbers of immigrants within each deme, reduced by the number of immigrants exchanged among demes within a bin. The estimand neighborhood size could be invariant with respect to bin size (in linear habitats, this holds provided that spatial distance is still measured in the original units not in number of bin widths). For mutation, one may assume that the estimand is the bin population size times mutation probability. In the Appendix, we show that such predictions do not always work well, in particular for Nm and g, and that the effects of binning may also depend on the distribution of samples among bins. In general, it may be difficult to make sense of Nm and g estimates.

To evaluate performance in a biologically relevant setting, we considered conditions broadly similar to those of two-dimensional analyses of the damselfly metapopulation described by Watts et al. (2007). This damselfly scenario can also serve as a basis for a realistic comparison between likelihood- and moment-based methods of inference. We have first simulated data sets with samples taken along four lines in a rotationally symmetric pattern forming the four tips of a cross. This mimics sampling along small streams in the original study. The neighborhood value $Nb = 200\pi$ and mutation rate $2N\mu = 0.01$ approximate the moment and likelihood estimates from the damselfly data. An array of 40×40 demes is simulated and analyzed as arrays of 20×20, 10×10, or 5×5 spatial bins (table 3, cases [46]–[48]). Even by PAC-likelihood, the analysis for the larger array is computationally intensive, so the sample size considered

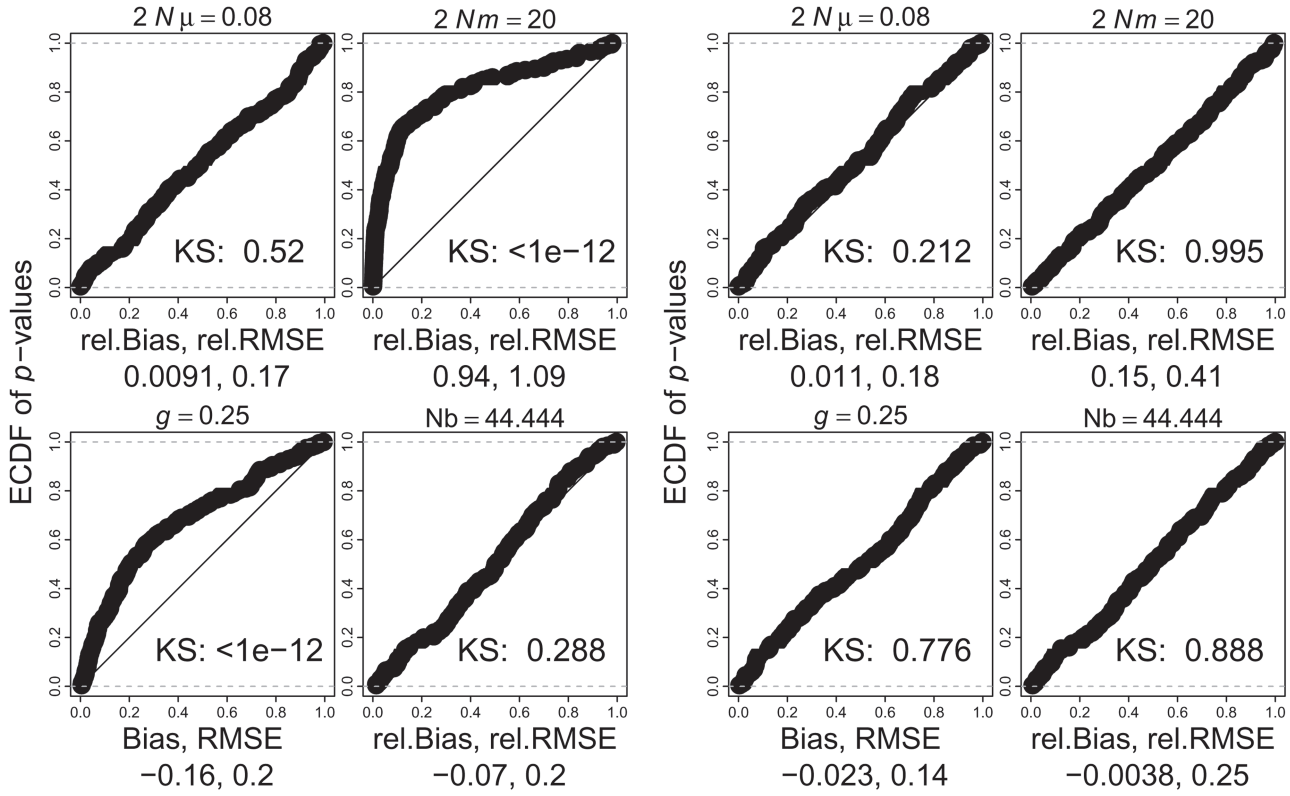


FIG. 3. Convergence of distributions of *P* values for increased *N*. The two cases differ only in *N*, *m*, and μ values for identical *Nm* and $N\mu$. $N = 40$ (case [5]) on the left and 40,000 (case [6]) on the right.

(10 loci genotyped in 200 individuals) is smaller than in the original study. This still requires about 15 CPU days per sample on ~2.5 GHz core processors (i.e., 7.5 CPU years in total for case [46]). The values tested by likelihood ratio are the estimands, that is, the true Nb value, and mutation probability times bin population size for $N\mu$.

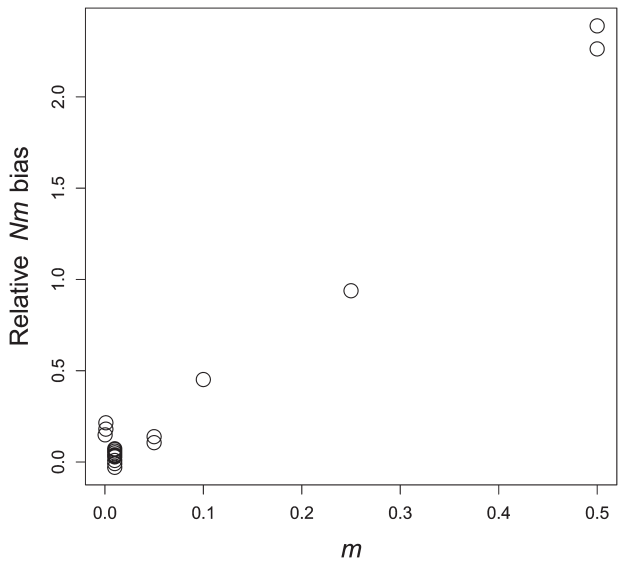


FIG. 4. Relationship between dispersal probability and bias of estimated number of migrants for all cases in table 1.

Effects of Binning

For 20 × 20 binning (case [46]), estimator performance is consistent with expectations, with good coverage of the confidence intervals. The same conclusions are supported by analyses as 10 × 10 and 5 × 5 arrays (cases [47] and [48]). In the latter case, a distorsion of *P* values becomes more apparent as well as a relative bias of 0.14 for 1/Nb. This distorsion may be in part due to the fact that many *g* estimates are at the boundary. This, and the high RMSE of *g* estimates (see case [65] in table 4, Appendix), may itself be due to the difficulty of estimating spatial effects when only a small range of distances are represented in the binned data.

Additional Effect of the Dispersal Distribution

To assess the effect of the dispersal distribution, the Poisson reciprocal Gamma distribution (Chesson and Lee 2005) is now used for the simulation of samples as described in the Methods section. Deme size was $N = 50$ as in case [46]. Cases [49]–[51] illustrate three different values of neighborhood size, the intermediate one being 200π as in case [46]. Good estimation of the neighborhood is achieved in all three cases (fig. 6, top, shows a typical profile likelihood surface in case [46]). However, for the largest neighborhood value the distribution of the LRT departs from ideal behavior if the spatial scale of sampling is not extended. In that case, more distant samples taken from a 80 × 80 lattice were also simulated (case [52]).

Table 2. Performance of Estimation by PAC-likelihood.

	Parameters					Relative $N\mu$			Relative Nm			g			Relative $1/Nb$		
	Array	***N***	***m***	***g***	μ	**Bias**	**RMSE**	**KS Test**	**Bias**	**RMSE**	**KS Test**	**Bias**	**RMSE**	**KS Test**	**Bias**	**RMSE**	**KS Test**
[21]	4×4	400	0.01	0.5	5×10^{-4}	−0.026	0.16	0.1	0.039	0.17	0.64	−0.004	0.16	0.78	0.18	0.82	0.77
[22]	4×4	400	0.1	0.5	5×10^{-4}	0.0042	0.16	0.05	0.43	0.9	0.0012	−0.015	0.41	0.00015	0.69	1.94	0.014
[23]	4×4	40, 000	0.001	1	5×10^{-6}	0.015	0.15	0.38	0.14	0.56	0.94						
[24]	4×4	80	0.5	1	0.0025	0.016	0.16	0.84	2.2	2.55	0						
[25]	**16**	40	0.25	0.25	0.001	−0.016	0.17	0.96	0.85	1	0	−0.17	0.2	0	−0.012	0.2	0.16
[26]	**16**	40, 000	0.00025	0.25	1×10^{-6}	−0.022	0.17	0.24	0.11	0.38	0.3	−0.034	0.14	0.89	0.072	0.27	0.95
[27]	**16**	400	0.01	0.25	0.001	−0.034	0.14	0.068	0.03	0.23	0.41	−0.018	0.14	0.53	0.088	0.38	0.71
[28]	**16**	400	0.01	0.5	0.001	−0.015	0.16	0.54	0.043	0.17	0.39	−0.022	0.11	0.25	0.12	0.5	0.39
[29]	4×4	400	0.01	0.25	0.001	−0.0022	0.17	0.15	0.027	0.18	0.067	−0.0045	0.16	0.89	0.11	0.55	0.28
[30]	4×4	400	0.01	0.5	0.001	−0.0049	0.17	0.33	0.023	0.16	0.89	−0.011	0.17	0.94	0.27	1.01	0.79
[31]	4×4	400	0.01	0.75	0.001	−0.022	0.17	0.066	0.04	0.19	0.43	−0.00036	0.2	0.0034	0.92	2.9	0.0043
[32]	**4**	400	0.01	1×10^{-4}	1×10^{-4}	−0.01	0.17	0.51	−0.034	0.22	0.61	0.055	0.12	0	−0.068	0.27	0.94
[33]	4×4	400	0.01	1×10^{-4}	0.001	−0.023	0.16	0.0042	−0.053	0.14	0.46	0.041	0.088	0	−0.054	0.22	0.021
[34]	4×4	40, 000	0.001	0.5	5×10^{-6}	0.012	0.16	0.66	0.2	0.67	0.62	0.0025	0.36	0.56	0.63	1.83	0.21
[35]	4×4	40	0.05	0.25	0.001	−0.048	0.17	0.23	0.096	0.23	0.56	−0.0097	0.15	0.25	0.053	0.49	0.7
[36]	4×4	40	0.05	0.5	0.001	−0.036	0.18	0.2	0.13	0.24	0.024	−0.011	0.18	0.78	0.16	0.81	0.27
[37]	**16**	400	0.01	0.75	0.001	−0.023	0.17	0.0052	0.014	0.13	0.21	−0.008	0.081	0.55	0.17	0.69	0.72
[38]	**100**	40	0.5	0.5	5×10^{-4}	−0.13	0.22	5.3×10^{-13}	2.37	2.57	0	−0.29	0.32	0	−0.029	0.23	0.43
[39]	10×10	400	0.01	0.5	5×10^{-5}	−0.16	0.2	1.1×10^{-16}	−0.0092	0.14	0.011	−0.029	0.12	0.12	0.28	0.64	0.19
[40][a]	4×4	400	0.01	0.99999	0.001	−0.0027	0.18	0.022	0.06	0.16	0.16	−0.052	0.1	5×10^{-7}	1.3×10^{8}	4.2×10^{8}	2.9×10^{-7}
[41]	10×10	400	0.01	0.5	5×10^{-4}	0.11	0.47	0.72	−0.032	0.14	0.56	−0.026	0.15	0.44	0.52	1.82	0.38
[42]	10×10	400	0.1	0.5	5×10^{-4}	0.014	0.22	0.5	0.14	0.47	0.26	0.011	0.25	0.53	0.17	0.9	0.99
[43]	10×10	400	0.01	0.75	5×10^{-4}	0.12	0.49	0.21	−0.026	0.2	0.59	−0.011	0.13	0.83	0.47	1.41	0.9
[44]	**100**	400	0.025	0.5	5×10^{-4}	−0.035	0.23	0.37	0.1	0.25	0.059	−0.025	0.1	0.46	0.068	0.38	0.16
[45]	10×10	40	0.25	0.5	5×10^{-4}	0.055	0.25	0.59	0.56	1.24	1.4×10^{-8}	0.058	0.42	3.1×10^{-10}	0.16	1.32	5.6×10^{-5}

[a] The large relative bias and RMSE of $1/\mathrm{Nb}$ estimates in case [40] is due to a number of low Nb estimates, compared to the parameter value 5.03×10^{11}.

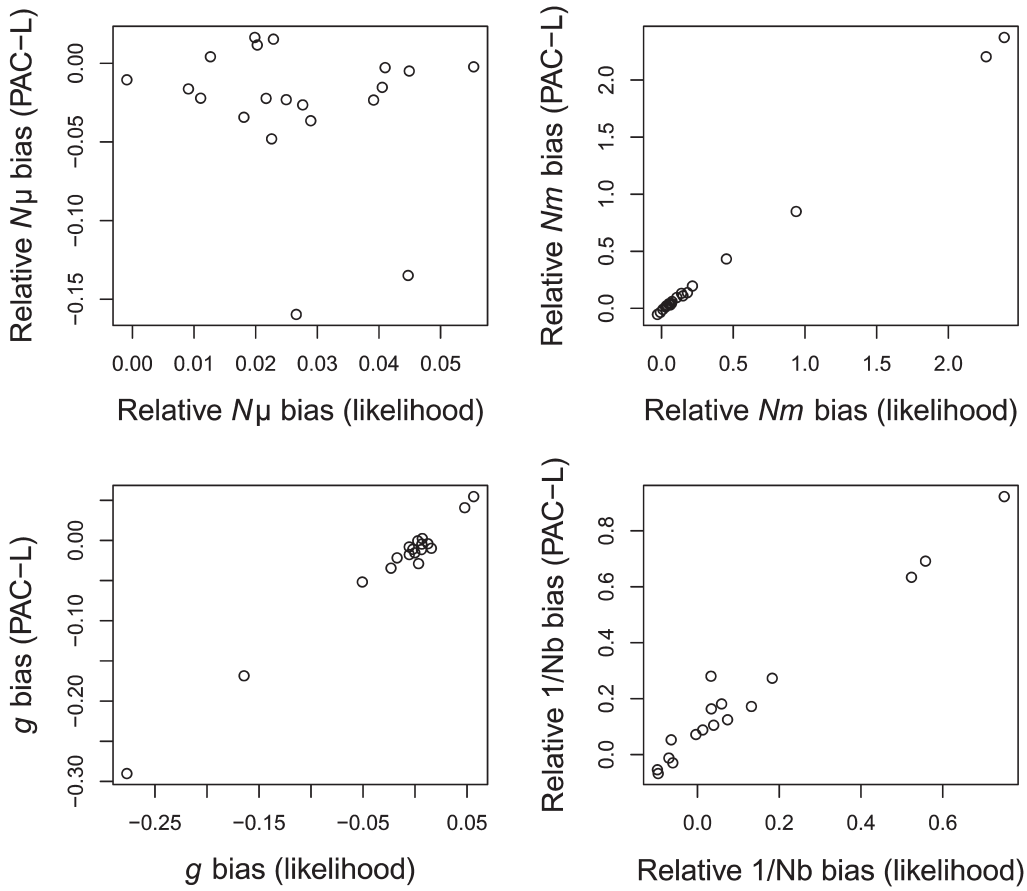


FIG. 5. Comparison of biases by strict likelihood (all cases in table 1) and PAC-likelihood (first 20 rows of table 2). A point (case [20] by likelihood, [40] by PAC-likelihood) with huge 1/Nb bias is not shown in the last panel.

Additional Effect of Stepwise Mutation

Finally, the same demographic simulation conditions were also considered for markers evolving under a SMM (cases [54]–[57]). $N\mu$ estimates are roughly halved, as previously observed for the SMM in Rousset and Leblois (2007) or even lower (case [57]; fig. 6, bottom, shows a typical profile likelihood surface in this case). Accordingly, the gene diversity is low. This implies that there may be little

Table 3. Alternative Dispersal and Mutation Models.

	Parameters					Bins	Relative $N\mu$			Relative 1/Nb		
							Bias	RMSE	KS Test	Bias	RMSE	KS Test
	40 × 40 **array,** $N = 50$, **m=0.5,** $g = 0.5$, $\mu = 1 \times 10^{-4}$ **(Geometric distribution)**											
[46]						20 × 20	0.0069	**0.15**	0.42	−0.036	0.35	0.19
[47]						10 × 10	0.0054	**0.15**	0.39	0.017	0.42	0.058
[48]						5 × 5	0.0011	**0.14**	0.079	0.14	0.56	0.00011
	Array	***N***	κ	**Nb**	μ	**(Reciprocal Poisson Gamma distribution)**						
[49]	40 × 40	**50**	0.92	126	1×10^{-4}	10 × 10	−0.032	**0.16**	0.14	0.048	0.14	0.86
[50]	40 × 40	**50**	4.6	628	1×10^{-4}	10 × 10	−0.0073	**0.16**	0.32	−0.068	0.27	0.12
[51]	40 × 40	**50**	**23**	3140	1×10^{-4}	10 × 10	−0.0056	**0.15**	0.8	−0.46	0.9	8×10^{-4}
[52]	80 × 80	**50**	**23**	3140	1×10^{-4}	10 × 10	0.016	**0.19**	0.22	−0.18	0.83	0.013
[53]	80 × 80	**50**	**23**	3140	1×10^{-4}	20 × 20	0.015	**0.18**	0.21	−0.23	0.79	0.44
Stepwise mutation												
[54]	40 × 40	**50**	0.92	126	1×10^{-4}	10 × 10	−0.54	**0.55**	**ND**	0.067	0.16	0.13
[55]	40 × 40	**50**	4.6	628	1×10^{-4}	10 × 10	−0.54	**0.55**	**ND**	−0.11	0.34	0.34
[56]	80 × 80	**50**	**23**	3140	1×10^{-4}	10 × 10	−0.72	**0.73**	**ND**	−0.4	0.83	0.56
[57]	80 × 80	**50**	**23**	3140	1×10^{-4}	20 × 20	−0.72	**0.73**	**ND**	−0.37	0.79	0.32
[58]	80 × 80	**50**	**23**	3140	5×10^{-4}	10 × 10	−0.75	**0.75**	**ND**	−0.28	0.73	0.4
[59]	80 × 80	**50**	**23**	3140	5×10^{-4}	20 × 20	−0.75	**0.75**	**ND**	−0.28	0.69	0.18

NOTE.—In the 40 × 40 array, samples were taken at positions (6,20) to (10,20), and in rotationally symmetric positions (20 samples of 10 individuals in total). In the 80 × 80 array, samples were at positions (11,40) to (19,40) by steps of two, and at rotationally symmetric positions. "ND" (not done) tests means that tests would be highly significant but were not performed as they would have required estimating the likelihood of points far from the top of the likelihood surface at the detriment of computations for inference about Nb.

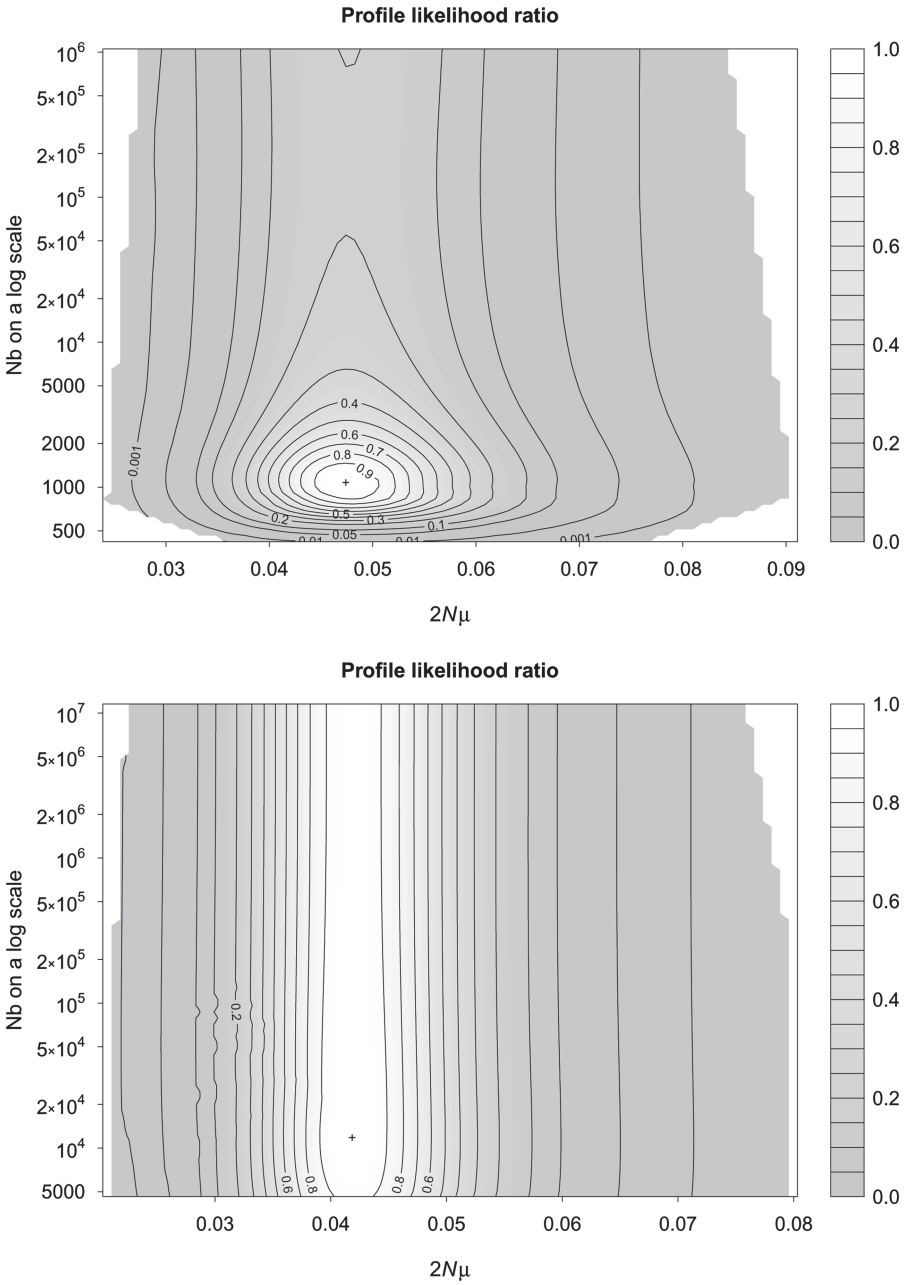


FIG. 6. Examples of profile likelihood surfaces for cases [46] (top) and [57] (bottom). The surfaces are inferred from 1,024 points as described in the Appendix. In each case, the sample that yielded estimation errors closest to the RMSE values and of the same sign as the bias were selected (hence, they exhibit positive Nb estimation error since 1/Nb estimates are negatively biased, table 3). In both cases, $2N\mu = 0.01$; Nb = 628 (top) or 3,140 (bottom). The likelihood profile surface is shown only for parameter combinations that fell within the envelope of parameter points for which likelihoods were estimated. The cross denotes the maximum.

information for other parameters, contributing to the Nb bias and also to as many as two-third of g estimates being at the boundary (not shown). In additional simulations (cases [58] and [59]), the true mutation rate was 5-fold increased, and the number of loci increased to 20, resulting in slightly improved Nb estimation.

Comparison with Moment-Based Estimates

Alternative estimators of 1/Nb are obtained as the regression slope of estimates of pairwise $F_{ST}/(1-F_{ST})$ to logarithm of distance (Rousset 1997) or of the pairwise statistic $\hat{e}$ comparing pairs of individuals as described in Watts et al. (2007). We consider only the $\hat{e}$ method below but similar results were obtained with $F_{ST}/(1-F_{ST})$. The 1/Nb estimators can be compared in terms of the ratio of their mean square errors and, as expected from a likelihood method, PAC-likelihood has lower error. Moreover, this discrepancy persists when alternative dispersal distributions and mutation models are considered.

For case [46], the ratio of MSEs is 0.66. Accordingly, the moment-based confidence intervals should be wider (fig. 7). However, they tend to be conservative (being often too

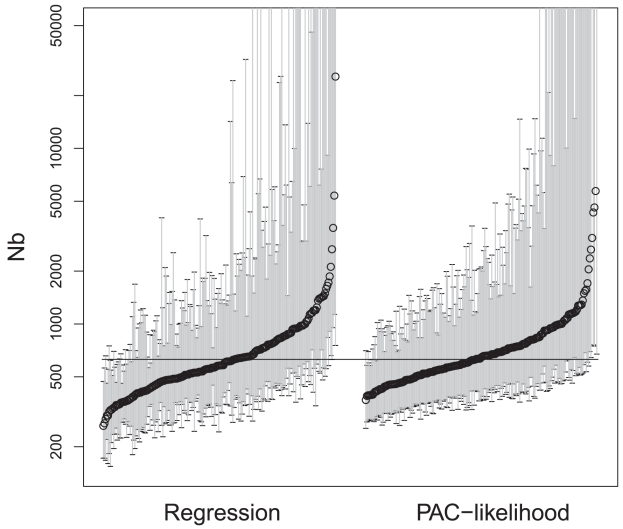


FIG. 7. Distributions of estimates and confidence intervals for Nb, by the spatial regression method and by PAC-likelihood, for case [46]. The horizontal line marks the true parameter value.

short when the Nb estimate is small), as previously shown for this and related methods (Leblois et al. 2003; Watts et al. 2007) and accentuated by the present small sample sizes.

With the alternative dispersal model, the ratios are 0.27, 0.36, 0.34, and 0.46 for the four cases [49]–[52], so that the moment method appears comparatively worse for more restricted dispersal.

In case [58], where stepwise mutation is also considered, the ratio is 0.55.

According to these results, the PAC-likelihood analysis of the original damselfly data for the two-dimensional habitat should provide a more accurate and reliable estimate and confidence interval for Nb than previous moment-based analyses. Still, the results are not very different from previous estimates: 1,110 (interval 600–3,625) by PAC-likelihood (analyzed as a 24×14 array) versus 753 (interval 319–3,162) by F_{ST}-based methods (Watts et al. 2007). They concur with the previous conclusion that the genetic estimates are only slightly higher that the demographic estimate (Nb = 555, Watts et al. 2007).

Discussion

We have presented an effective software implementation of likelihood inference under a two-dimensional model of isolation by distance and investigated the performance of inferences based on likelihood ratios in both the one- and the two-dimensional spatial models. Our results illustrate both the strengths and imperfections of such inferences: In most cases, estimators have low bias and, given the relatively small sample sizes considered, low MSE. These results are consistent with those of Rousset and Leblois (2007). When compared with a preexisting method for estimation of neighborhood, the likelihood-based estimation of neighborhood appears to be substantially more efficient and its confidence intervals to be more reliable, even when complicating factors such as the misspecification of the dispersal distribution and the binning of samples are taken into account.

However, considering the distributions of *P* values of LRTs underlines small but statistically detectable effects such as the small negative bias of PAC-likelihood estimates of mutation rate. Further, the assumptions inherent in the statistical model (low m, large N, and an approximate accounting of spatial edge effects) affect estimation of the Nm and g parameters. For $m = 0.5$, we found more than 2-fold relative bias in number of migrants. This could be expected from consideration of the infinite island model. In this simple case, the expected F_{ST} for $N = 80$ and $m = 0.5$ is $(1-m)^2/[(1-m)^2+N\{1-(1-m)^2\}] \approx 0.004$, whereas the classical low-Nm approximation $1/(1+2Nm)$ (for haploid N) is ≈ 0.012. The coalescent approximation fits the actual F_{ST} for a higher Nm value than the true one, so that Nm estimates derived from the coalescent model should be biased upward. Under isolation by distance, short-distance differentiation can be approximated by island model expectations, and we again expect, and observe, upward-biased Nm estimates. Since programs such as Migrate (Beerli and Felsenstein 1999, 2001) or Lamarc (Kuhner 2006) are based on the same coalescent approximations as de Iorio and Griffiths' algorithms, the same biases should be encountered, at least when the same type of molecular markers is considered. Inference methods based on a Dirichlet distribution for allele frequencies, as follows from Wright's (1937) diffusion formula, should be affected by the same type of biases. This was observed by Faubet et al. (2007, p. 1160) when assessing the method of Wilson and Rannala (2003) on samples drawn from populations with small N and large m.

In order to better identify other possible causes of nonideal performance, we have first assumed that the dispersal distribution and the mutational process were known. We have then relaxed these assumptions and have also considered the effects of the spatial binning of samples. Both the misspecification of the dispersal distribution and spatial binning can bias the estimation of the dispersal parameters in complex ways that may render such estimates practically meaningless. However, in general neighborhood size estimation appeared robust (see also Rousset and Leblois 2007 for a linear habitat), except when the subpopulations that are binned together already exhibit a substantial fraction of the differentiation found among the most distant subpopulations. In simulations jointly considering misspecification of the dispersal distribution, of the mutation model, and the effect of a milder but realistic spatial binning, the mutation process mainly affected $N\mu$ estimation but not Nb estimation.

The fact that neighborhood estimation appears robust implies that likelihood inference performs in the same way as a spatial regression method that would simultaneously estimate the neighborhood size from the increase in differentiation with distance (which does not rest on a coalescent approximation) and that would estimate Nm from the level of small-scale differentiation. Likelihood inference of Nb may actually be more robust than the regression method as the latter does not account for spatial edge effects.

For example, in case [39] (a 10×10 array with samples taken in the corners), the regression method has an approximately 3-fold bias (details not shown), whereas the likelihood method has correct coverage.

We did not consider the effect of a so-called continuous population structure, where individuals can settle anywhere in a continuous habitat (Felsenstein 1975; Barton et al. 2002). However, in such a case, the neighborhood parameter is best defined by considering the random walk of ancestral lineages over the finite or countable positions of ancestors rather than over continuous space, so that continuous-space models can actually be understood as discrete-space models (Robledo-Arnuncio and Rousset 2010), akin to the lattice models considered in the present work. In this respect, we do not expect important differences between the estimands in the two classes of models. In both discrete- and continuous-space models, the neighborhood parameter depends on the product of an effective mean square dispersal distance σ_e^2 and of an effective population density parameter D_e. σ_e^2 is defined as the asymptotic increase in mean square displacement per unit of time of a particle performing this random walk, and D_e is defined from the asymptotic rate of encounter of ancestral lineages that each perform the same random walk and do not coalesce when they meet each other. The estimand neighborhood size defined in this way is a good predictor of the moment method performance (Robledo-Arnuncio and Rousset 2010).

A corollary of robust neighborhood estimation and nonrobust *Nm* estimation is that algorithms based on coalescent approximations are not most appropriate to infer the shape of the dispersal distribution. A dedicated study of inference of the shape of the dispersal distribution in a wider family of distributions would either be plagued by the effects of the coalescent approximation or should confine itself to scenarios of low dispersal probability, compared with our focal population scenarios, which would strongly restrict its usefulness.

This study has been focused on isolation by distance as it is a widespread phenomenon that has been little considered in a likelihood framework. However, this is a computationally challenging problem, and simpler problems can very easily be handled within the current software implementation. Estimation of the mutation rate parameter for a single population can be performed in seconds, and remarkably even the single locus confidence interval have practically perfect coverage in this case (not shown). Analyses under an island model are also fairly straightforward.

From the present results, likelihood inferences appear feasible in moderately sized networks of populations, and they are more efficient than moment-based method in some realistic conditions. Nevertheless, the validity of inferences is affected in complex ways by many factors and may need to be analyzed in a case-by-case basis. Further progress in algorithms and refined approximation techniques would be necessary to raise full-likelihood techniques as a general-purpose method of analysis of spatial genetic data, in particular if accurate confidence intervals are sought. This will surely encourage consideration of alternative methods to derive estimates and confidence intervals. A general alternative is the one based on simulation of summary statistics, more or less similar to currently developed ABC techniques (Beaumont et al. 2002; Marjoram and Tavaré 2006). In the latter perspective, it is worth emphasizing that coalescent approximations matter, and thus sample simulation programs based on such approximations may be misleading. More speculatively, the PAC-likelihood estimators could be considered as efficient summary statistics, though improvements in computation power and in the processing of simulated distributions will be necessary to make this a practical option.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank K. Belkhir, G. Dugas, A. Weisseldinger, and V. Ranwez for management and assistance in relation to the computing grids of ISEM and UFR, and J.-M. Cornuet, R. Vitalis, and two reviewers for comments on the manuscript. This is paper ISEM 2011-154. This work was supported by the Agence Nationale de la Recherche (ANR EMILE NT09-611697).

Appendix

Algorithms and Implementation

Likelihood and PAC-Likelihood Computation

The likelihood for individual parameter points is estimated as the average value of a statistic over independent realizations of possible ancestral histories of a sample, which distribution is generated by an absorbing Markov chain with the allelic state in the common ancestor as the absorbing state (e.g., Griffiths and Tavaré 1994; de Iorio and Griffiths 2004a; de Iorio and Griffiths 2004b). In particular, the likelihood computations rest on the computation of de Iorio and Griffiths' $\hat{\pi}$ terms, which are approximations for the probability π that an additional gene sampled from a population is of a given allelic type, conditional on the allelic counts of a previous sample. The $\hat{\pi}$s may be viewed as biased estimates of the πs, and importance sampling techniques, where the $\hat{\pi}$s also affect the distribution of realizations of the Markov chain, are used to obtain an unbiased estimate of the likelihood (see de Iorio and Griffiths 2004b for a more detailed description).

This algorithm differs from those based on long runs of a recurrent Markov chain, which is the better known Markov Chain Monte Carlo type of algorithms considered in Migrate (Beerli and Felsenstein 1999, 2001, Lamarc (Kuhner 2006) or in the IM suite of programs (Hey and Nielsen 2004; Hey 2010). Therefore, the lingering issue of assessing the convergence of a recurrent Markov chain does not arise. On the other hand, the estimator of likelihood may not be consistent for certain choices of the absorbing Markov chain (Stephens and Donnelly 2000), but this problem is not apparent in the present work.

An elementary modification of the likelihood estimation algorithm, which however saves a substantial fraction of computation time, is to perform the final likelihood computation when the ancestral history reaches the two-lineages states by using standard formulas for identity in state in migration matrix models, taken in the coalescent limit (e.g., de Iorio et al. 2005) rather than by the Markov chain method. This was implemented here.

The average computational burden for each ancestral history increases with the number of events (mutation, migration, coalescence) in each history and with the amount of computation for each event. Both increase with the number of demes. The $\hat{\pi}$s can be obtained as the solution of a linear system $\mathbf{A}\hat{\boldsymbol{\pi}} = \mathbf{b}$, where each dimension of the $\mathbf{A}$ matrix is the number of subpopulations. In large arrays of demes, the computation time of this step can be reduced by using an iterative method (preconditioned conjugate gradient algorithm) for solving the linear system, provided approximate initial values are easy to compute. For linear habitats, the solution of the system defined by a pentadiagonal subset of elements of $\mathbf{A}$ was used as the starting point (Rousset and Leblois 2007). In two-dimensional habitats, the solution for the $\mathbf{A}$ matrix for the island model ($g = 1$, other parameters unchanged) is used as a starting point as it can be efficiently computed using the Sherman–Morrison formula (Bartlett 1951).

Although networks of more that 100 populations were considered in this study, one CPU year or more may be necessary to analyze a typical sample in this context. Because the estimation of likelihood in different parameter points proceeds independently, such an analysis can easily be performed in a much shorter time on a computer grid. Still, it is unpractical to analyze hundreds of samples in such conditions. A fast alternative in such cases is the PAC-likelihood method. In the incarnation previously described in Cornuet and Beaumont (2007) and Rousset and Leblois (2007), PAC-likelihood uses de Iorio and Griffiths' $\hat{\pi}$ as estimates for the corresponding π, without any recourse to the importance sampling methodology to correct for any resulting bias. The computation time of PAC-likelihood increases more slowly with the number of demes because it is independent of the number of possible events in the history of the sample. In particular, for fixed number of demes, the differences in computation time between likelihood and PAC-likelihood estimation increases when a large number of migration events occur in the realized ancestral histories that is when Nm increases. Rousset and Leblois (2007) found that PAC-likelihood could by 500 to 1,000 times faster than likelihood, but the difference can be larger depending on Nm values. In many of the simulations of table 3, parameter points with Nm values of 500 or more had to be considered, and only PAC-likelihood computation was feasible for such points.

Likelihood Surface Estimation

Smoothing of Likelihood Estimates. A likelihood surface is inferred from likelihood estimates in different points. The smoothing technique known as Kriging (e.g., Cressie 1993; Zimmerman and Stein 2010) was used in de Iorio et al. (2005) and Rousset and Leblois (2007) for that purpose and is still used in this paper. However, compared to previous works, the implementation of Kriging had to be optimized in order to yield good confidence intervals. The Kriging predictor function depends on covariances between response values at different distances in parameter (predictor variables) space, and these covariances are described by a covariance family and some covariance parameters.

The covariance parameters are now estimated by so-called generalized cross-validation (Golub et al. 1979), using the Matérn covariance family, which includes a smoothness parameter ν. In general, the estimated ν was the maximum allowed value (i.e., 4) in our estimation procedure, which generates smooth likelihood surfaces, as expected. In cases [52]–[59] (table 3), a high minimal allowed value ($\nu = 3.9$) had to be imposed to obtain consistently good results. However, it is always wise to begin with less constrained ν estimation, as this might reveal problems with the points subject to smoothing.

We do not present specific checks of the accuracy of the smoothing step here, as in the end what matters are the properties of confidence intervals. When these had poor properties, it was repeatedly checked that the Kriging steps were not the cause of concern by increasing the density of likelihood points considered. Failure of Kriging can also easily be detected on individual data sets from a diagnostic plot of the residual errors of prediction, provided by the program.

All algorithms used for Kriging and cross-validation are described in Nychka (2000) and implemented in the fields package (Fields Development Team 2006) of the R statistical software (R Development Core Team 2004). However, numerical issues related to the inversion of nearly singular matrices led us to independently reimplement these algorithms. A C++ library interfaced with R is distributed along our main software to perform Kriging. Likelihood surface prediction may be very poor when extrapolation is made out of the range of parameter values of Kriged points. For this reason, the predictor was only applied to parameter points within the convex envelope of the Kriged points. Convex envelope computations were performed using the rcdd package (Geyer and Meeden 2008). All the analyses described in this study (estimation of likelihood in individual points by de Iorio and Griffiths' algorithms, Kriging, graphical output of likelihood and profile likelihood surfaces as shown in figs. 2 and 6) can be performed with the Migraine software, a C++ executable, without knowledge of Kriging nor of the R language. Migraine is free open source software. Its current distribution page is http://kimura.univ-montp2.fr/∼rousset/Migraine.htm. Multiple parameter tests are implemented in this software but not further discussed here.

Computation Settings

The settings described in this section apply to all simulations unless mentioned otherwise.

Exploration of Parameter Space. Final likelihoods are estimated from 1,024 points, obtained in two steps. In the first step, 512 parameter points are sampled uniformly. For samples simulated under the geometric dispersal model, the initial range of parameter values explored is one-third to three times the parameter value for $N\mu$ and Nm and 0–0.999 for g. For samples simulated under a SMM, the initial $N\mu$ bounds were further halved. For samples simulated under the Poisson reciprocal gamma dispersal model, the initial $2Nm$ range was one-third to three times Nb/π (which coincides with the initial $2Nm$ range under the geometric model when $g = 0.5$).

Likelihoods are estimated for the first 512 points, and for one every 30 of them, a second replicate estimate is computed. A likelihood surface is inferred from these likelihoods by Kriging (including a cross-validation step), and a convex envelope putatively including the whole $P = 0.001$ confidence region (and possibly extending beyond the original parameter ranges) is constructed. The parameter space in which the convex envelope is defined is the same as for Kriging. Another envelope extending z times as far from the barycenter of the original envelope is defined for given z. In most simulations, 512 additional points were sampled approximately uniformly within the extended envelope with $z = 2$. In the latest simulations (in particular, cases [52]–[59]), a slightly more involved procedure was used, where half of the points are sampled uniformly within the envelope with $z = 1.1$ and the other half in the envelope with $z = 2$. Either way, these procedures appears very efficient in that most of the points sampled in this way are indeed on the "top" of the likelihood surface and contribute to the computation of final likelihood ratio tests and confidence regions.

In the second step, the likelihood for the 512 additional points are estimated, with again replicates for one every 30 of them, and a likelihood surface is inferred by Kriging from all 1,024 points (including a new cross-validation step). The effect of additional points from a third iteration was repeatedly checked and found to have no impact on the conclusions.

Other Settings. In most cases, for each locus and each parameter point, the likelihood estimate is obtained from 30 replicates of the absorbing Markov chain (i.e., 30 possible ancestral histories) or 30 replicates of the PAC-likelihood algorithm. In cases [52]–[59], only five replicates of the IS or PAC-likelihood algorithms were computed for each locus and each parameter point, as preliminary simulations suggested that this was sufficient.

Specific Settings for Large g. If the true g value is 0.99999, uniform sampling of hundreds of g values is unlikely to generate g values large enough, so that ultimately no predicted likelihood value will be available for the true g value or for the true neighborhood value. Various ad hoc corrections of the sampling of parameter points could be considered. Here, we performed uniform sampling of $\ln(\sigma^2_{\mathrm{cond}})$ rather than g. Kriging was performed on the same variable. This parameterization could be more generally useful when there is a plateau of high likelihood values for large values of the neighborhood size, which is expected for samples simulated under high neighborhood values.

Comparison with Moment-Based Inferences. Likelihood-based inferences of neighborhood size were compared with moment-based ones (e.g., Rousset 1997; Vekemans and Hardy 2004; Watts et al. 2007) as implemented in the software Genepop, version 4.1 (Rousset 2008), wherein confidence intervals are constructed by the ABC bootstrap method (DiCiccio and Efron 1996).

Misspecification Effects Due to the Coalescent Approximation

Samples are simulated under an exact backward generation-by-generation algorithm, where no "large N" approximation is used. Deme size N, forward migration probabilities, and mutation probability u are all distinct parameters, whereas the estimation algorithm is based on limit results for large N, small backward immigration probabilities, and small u. In the sample simulation program, edge effects can be accounted for in a simple mechanistic way by computing the backward dispersal distribution in a focal deme as the relative forward migration probabilities from every deme (including the nonimmigration probability from the focal deme itself), where the forward probabilities are identical from any deme. But this cannot be done in the estimation algorithm, as the coalescent model does not depend on the nonimmigration probability but only of number of immigrants (product of deme size and immigration probabilities) from other demes. To put it another way, in the coalescent limit, the forward nonmigration rate is the limit value of $N(1-m)$ as $N \to \infty$ and $m \to 0$; this is infinitely larger than any immigration rate from other demes and cannot be used to define a backward probability distribution.

This means that the statistical model is intrinsically misspecified when applied to samples generated by the exact backward algorithm. One way to overcome this discrepancy is to simulate samples under coalescent assumptions, and this case has been considered. However, an extended assessment of performance under such conditions would not give any idea of the implications of misspecification for analyses of data from populations where dispersal probabilities are not vanishingly low. Therefore, we more generally controlled the number of immigrants according to the rules described in the main text.

These rules have the following effects under the geometric dispersal model. In the estimation algorithm, the expected number of immigrant genes (haploid deme size times dispersal probability) from any given subpopulation to some focal deme d is a given Nm value times $g^{|x|+|y|}i(x,y)/G$, where $i(x,y) = 1/[(1+\delta_{x0})(1+\delta_{y0})]$, and G is the maximum value, over all demes each taken as the focal one d, of $\sum_{k\neq d} g^{|x|+|y|}i(x,y)$. For example, in case [1] (a 4×4 array of demes of haploid size $N = 400$, $m = 0.01$, and $g = 0.5$), the expected numbers of migrant genes within each deme are 4, 2.477, or 3.169, depending on whether the deme is in the central square, in the corners, or in another edge position, respectively. In

Table 4. Effects of Binning.

	Array	Parameters or Estimands				Relative $N\mu$			Relative Nm			g			Relative $1/\mathrm{Nb}$		
	or bins	$2N\mu$	$2Nm$	g	Nb	Bias	RMSE	KS Test	Bias	RMSE	KS Test	bias	RMSE	KS Test	Bias	RMSE	KS Test
	10×10	0.04	8	0.5	100.531												
[60]	5×5	0.16	16	0.357	100.531	−0.11	**0.17**	0.00063	0.055	0.45	5.1×10^{-7}	−0.014	**0.35**	8.1×10^{-8}	0.48	1.22	4.8×10^{-8}
[61]	5×5	0.16	8	0.357	100.531	−0.17	**0.23**	1.1×10^{-16}	−0.17	0.23	1.1×10^{-10}	0.053	**0.21**	0.45	1.34	1.9	6.4×10^{-15}
	100	0.4	20	0.5	120												
[62]	**25**	1.6	20	0.131	30	−0.082	**0.24**	0.32	0.13	0.26	0.19	−0.0033	**0.13**	0.68	−0.015	0.37	**0.71**
	40×40	0.01	50	0.5	628.319												
[63]	20×20	0.04	252.467	0.12791	628.319	0.0069	**0.15**	0.42	0.062	0.46	0.093	0.032	**0.19**	0.0027	−0.036	0.35	**0.19**
[64]	10×10	0.16	297.086	0.083559	628.319	0.0054	**0.15**	0.39	0.014	0.37	0.32	0.045	**0.24**	1.1×10^{-6}	0.017	0.42	**0.058**
[65]	5×5	0.64	210.365	0.176318	628.319	0.001	**0.14**	0.079	0.35	0.6	2.4×10^{-9}	−0.036	**0.32**	0	0.14	0.56	**0.00011**

NOTE.—Header lines give the three sets of true sample simulation parameters. All analyses are by PAC-likelihood. Similar results were obtained by strict likelihood for cases [60] and [62] (not shown). For easy reference, cases [63]–[65] reproduce the results for Nb and $N\mu$ already given as cases [46]–[48] in Table 3.

the sample simulation, the expected number of immigrant genes from any given subpopulation to some focal deme d is deme size times the backward immigration probability. The latter probability can be written in the form

$$\frac{g^{|x|+|y|}i(x,y)m/G}{\sum_{k\neq d} g^{|x|+|y|}i(x,y)m/G + (1-m)}, \qquad (3)$$

where m is the forward dispersal probability, and $\sum_{k\neq d}(\cdot)$ denotes a sum over source demes k distinct from the focal deme d. For the maximizing focal deme, the denominator is 1 and the number of immigrants from each other deme is $Nmg^{|x|+|y|}i(x,y)/G$ as in the estimation algorithm. For case [1], these numbers of immigrants are 4, 2.486, or 3.176 in each of the three types of demes defined above. If no correction were applied in order to control the maximal Nm, they would be 2.427, 1.506, or 1.925 (i.e., $G = 2.427/4$).

Effects of Binning

A good understanding of the effects of binning on inferences is obtained when a rule is given to generate estimands that are shown to be estimated with low bias and ideally good coverage of confidence intervals. For example, bin population size times mutation probability is a good $N\mu$ estimand as the estimates have low bias relative to this value. Departures from this rule in the simulations can be attributed to the PAC-likelihood bias rather than to binning per se. However, for the dispersal parameters, no rule was found that correctly predicts all estimands in all cases investigated. For example, the expected number of immigrants in a bin is not always the correct Nm estimand. Nevertheless, we can deduce estimands for binned data from the estimands of the moment-based regression method: the estimand Nm is deduced from the inferred F_{ST} between the nearest bins, and this appears to work well. Indeed, this may not only account for the effects of binning but also for deviations from the large N, low m approximation. Likewise, the Nb estimand can be deduced from the increase of differentiation with distance in the binned data, and the g estimand can then be deduced from Nm and Nb.

However, there are several drawbacks with these predictions. First, they can be derived from simple analytical arguments in some cases, but must otherwise be generated by a regression analysis of binned data with a large number of loci and are not uniquely related to the true parameter values only but may also depend on the sampling design as shown below. Second, the Nb estimand derived from the slope of the regression may not be a valid prediction in conditions where the regression method is expected to poorly estimate Nb. For example, the regression method does not account for edge effects in contrast to the likelihood method. Peripheral demes receive fewer immigrants and thus are more differentiated than central demes, which biases regression Nb estimates downward when samples are taken from peripheral demes. In fact, it is both more easily interpretable and overall a more accurate prediction to assume that the estimand Nb is the true Nb value. The following examples illustrate these conclusions.

Under an island model, different results are expected whether only populations are binned or whether samples are binned too. As an illustration of the first case, consider an array of 10 × 10 populations binned into a 5 × 5 array, but samples come from nonadjacent populations and therefore go into different bins. A standard F_{ST} analysis of the binned data will yield the same F_{ST} and Nm estimates as that of the original data because the binned samples are indistinguishable from the original samples, and F_{ST} estimation per se does not use any extra information. By contrast, when (say) pairs of samples are binned too, the F_{ST} estimates are halved, so that Nm estimates are roughly $1/2$ plus twice the original value. As an illustration, we reconsidered the simulation conditions of case [39] ($2Nm = 8$, $g = 0.5$, and Nb = 100.531 in a two-dimensional habitat). For bins covering two lattice units, wherein pairs of samples are binned, the estimand is $2Nm = 16$ (or more exactly 16.5) according to the island model argument, and then from equation (1), $g = 0.357$. By contrast, if four of the eight sampled populations are taken at position (3,3) and rotationally symmetrical positions rather than at position (2,2) and rotationally symmetric positions, samples are no longer binned when populations are binned. Simulations conditions are otherwise identical to the previous ones, but the estimand is $2Nm = 8$. Simulation results (case [60] vs. [61]) confirm this predicted contrast. For Nb, if true values are taken as the estimands, estimation is poor as shown in the table. However, performance is also poor when estimands are deduced from the regression analysis (not shown). In this case, regression estimates are affected by edge effects, as samples come from peripheral demes even in the absence of binning.

This example shows that the effects of binning may be difficult to predict as they are affected by the sampling design, and all the more so as in real data analyses, different number of samples may fall in different bins. The following simulations and all those reported in table 3 incorporate the latter feature.

For the simulation conditions of case [44] (m=0.025, $2Nm = 20$, $g = 0.5$, and Nb= 120 in a linear habitat), a regression analysis of a 2,000-loci data set shows that the fitted differentiation between adjacent bins of four demes is only slightly lower than that between adjacent demes (estimated $F_{ST} \approx 0.035$ vs. 0.040). The slope of the regression against geographical distance (in bin width units) is roughly 4-fold increased, which is indeed expected from the mere effect of the change of spatial scale (equivalently, the Nb estimand is invariant if distance is always measured in the same spatial units). For simplicity, in the analysis of likelihood performance, we assumed that the Nm estimand was unchanged by binning (thereby expecting a small positive bias) and that Nb estimand is the true Nb value, only 4-fold reduced by the change of scale. All these predictions are well supported (case [62]).

A similar analysis was conducted in conditions closer to the damselfly example. Estimands were deduced from a regression analysis of a 2,000-loci data set for three binning levels. The estimand Nb inferred in these three cases deviated at most by 43% from the true value (in particular, in two dimensions, the regression is relative to logarithm of distance and a change of spatial scale has no effect on the regression slope). On the other hand, there was an almost 6-fold variation of the Nm estimands from the true Nm value (up to 297 vs. 50, as shown on the left of table 4). As above, in the analysis of likelihood performance, we varied the Nm estimand as given by the regression analysis but fixed the Nb estimand to the true value as shown in the table.

The predictions are again well supported (cases [63]–[65]), although some distorsion of P values is observed for g (maybe because many estimates are at the boundary) and becomes evident for the other dispersal parameters under the highest level of binning.

These results show that the effects of binning on likelihood inferences of dispersal parameters are largely predictable from its effect on spatial regression analyses when the latter are meaningful. However, the effects on Nm and g are complex. In the main text, we consider only Nb and $N\mu$ estimates, where Nb estimands are taken to be the true values.

References

Abdo Z, Crandall KA, Joyce P. 2004. Evaluating the performance of likelihood methods for detecting population structure and migration. *Mol Ecol.* 13:837–851.

Bartlett MS. 1951.An inverse matrix adjustment arising in discriminant analysis. *Ann Math Stat.* 22:107–111.

Barton NH, Depaulis F, Etheridge AM. 2002. Neutral evolution in spatially continuous populations. *Theor Popul Biol.* 61:31–48.

Beaumont MA. 2007. Conservation genetics. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics. 3rd ed. Chichester (UK): Wiley. p. 1021–1066.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximation Bayesian computation in population genetics. *Genetics* 162:2025–2035.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* 17:1505–1519.

Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol Ecol.* 13:827–836.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* 22: 341–345.

Beerli P, Felsenstein J. 1999. Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152:763–773.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A.* 98: 4563–4568.

Casella G, Berger RL. 2002. Statistical inference. Pacific Grove (CA): Duxbury.

Chesson P, Lee CT. 2005. Families of discrete kernels for modeling dispersal. *Theor Popul Biol.* 67:241–256.

Cornuet JM, Beaumont MA. 2007. A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theor Popul Biol.* 71:12–19.

Cox DR. 2006. Principles of statistical inference. Cambridge (UK): Cambridge University Press.

Cox DR, Hinkley DV. 1974. Theoretical statistics. London: Chapman & Hall.

Cressie NAC. 1993. Statistics for spatial data. New York: Wiley.
de Iorio M, Griffiths RC. 2004a. Importance sampling on coalescent histories. *Adv Appl Prob*. 36:417–433.
de Iorio M, Griffiths RC. 2004b. Importance sampling on coalescent histories. II. Subdivided population models. *Adv Appl Prob*. 36:434–454.
de Iorio M, Griffiths RC, Leblois R, Rousset F. 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor Popul Biol*. 68: 41–53.
DiCiccio TJ, Efron B. 1996. Bootstrap confidence intervals (with discussion). *Stat Sci*. 11:189–228.
Faubet P, Gaggiotti OE. 2008. A new Bayesian method to identify the environmental factors that influence recent migration. *Genetics* 178:1491–1504.
Faubet P, Waples RS, Gaggiotti OE. 2007. Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Mol Ecol*. 16:1149–1166.
Felsenstein J. 1975. A pain in the torus: some difficulties with models of isolation by distance. *Am Nat*. 109:359–368.
Fields Development Team. 2006. Fields: tools for spatial data. Boulder (CO): National Center for Atmospheric Research. Available from: http://www.image.ucar.edu/Software/Fields/
Geyer CJ, Meeden GD. 2008. R package rcdd (C double description for R). Version 1.1. Twin Cities (MN): University of Minnesota. Available from: http://www.stat.umn.edu/geyer/rcdd.
Golub GH, Heath M, Wahba G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21:215–223.
Griffiths RC, Tavaré S. 1994. Ancestral inference in population genetics. *Stat Sci*. 9:307–319.
Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, Excoffier L. 2005. Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* 170:409–417.
Hey J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol*. 27:905–920.
Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.
Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22: 768–770.
Leblois R, Estoup A, Rousset F. 2003. Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. *Mol Biol Evol*. 20:491–502.
Leblois R, Estoup A, Rousset F. 2009. IBDSim: a computer program to simulate genotypic data under isolation by distance. Mol Ecol Resources. 9:107–109.
Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165:2213–2233.
Marjoram P, Tavaré S. 2006. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet*. 7: 759–770.
Novembre J, Slatkin M. 2009. Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. *Evolution* 63:2914–2925.
Nychka D. 2000. Spatial process estimates as smoothers. In: Schimek MG, editor. Smoothing and regression. Approaches, computation and application. New York: Wiley. p. 393–424.
Paetkau D, Slade R, Burden M, Estoup A. 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. *Mol Ecol*. 13:55–65.
Peter BM, Wegmann D, Excoffier L. 2010. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol Ecol*. 19:4648–4660.
Pinheiro JC, Bates DM. 2000. Mixed-effects models in S and S-PLUS. New York: Springer Verlag.
R Development Core Team. 2004. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. http://www.R-project.org.
Rannala B, Hartigan JA. 1996. Estimating gene flow in island populations. *Genet Res*. 67:147–158.
Robledo-Arnuncio JJ, Rousset F. 2010. Isolation by distance in a continuous population under stochastic demographic fluctuations. *J Evol Biol*. 23:53–71.
Rousset F. 1997. Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics* 145:1219–1228.
Rousset F. 2008. GENEPOP'007: a complete reimplementation of the GENEPOP software for Windows and Linux. *Mol Ecol Resour*. 8:103–106.
Rousset F, Leblois R. 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Mol Biol Evol*. 24: 2730–2745.
Self SG, Liang KY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc*. 82:605–610.
Severini TA. 2000. Likelihood methods in statistics. Oxford: Oxford University Press.
Stephens M, Donnelly P. 2000. Inference in molecular population genetics (with discussion). *J R Stat Soc*. 62:605–655.
Vekemans X, Hardy OJ. 2004. New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol Ecol*. 13:921–934.
Watts PC, Rousset F, Saccheri IJ, Leblois R, Kemp SJ, Thompson DJ. 2007. Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Mol Ecol*. 16:737–751.
Wilson GA, Rannala B. 2003. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163: 1177–1191.
Wright S. 1937. The distribution of gene frequencies in populations. *Proc Natl Acad Sci U S A*. 23:307–320.
Zimmerman DL, Stein M. 2010. Classical geostatistical methods. In: Gelfand AE, Diggle PJ, Fuentes M, Guttorp P, editors. Handbook of spatial statistics. Boca Raton (FL): CRC Press. p. 29–56.

MBE Advance Access published August 11, 2014

Maximum-Likelihood Inference of Population Size Contractions from Microsatellite Data

Raphaël Leblois,*[1,2,3] Pierre Pudlo,[1,3,4] Joseph Néron,[2] François Bertaux,[2,5] Champak Reddy Beeravolu,[1] Renaud Vitalis,[1,3] and François Rousset[3,6]

[1]INRA, UMR 1062 CBGP (INRA-IRD-CIRAD-Montpellier Supagro), Montpellier, France

[2]Muséum National d'Histoire Naturelle, CNRS, UMR OSEB, Paris, France

[3]Institut de Biologie Computationnelle, Montpellier, France

[4]Université Montpellier 2, CNRS, UMR I3M, Montpellier, France

[5]INRIA Paris-Rocquencourt, BANG Team, Le Chesnay, France

[6]Université Montpellier 2, CNRS, UMR ISEM, Montpellier, France

***Corresponding author:** E-mail: raphael.leblois@supagro.inra.fr.

Associate editor: Asger Hobolth

Abstract

Understanding the demographic history of populations and species is a central issue in evolutionary biology and molecular ecology. In this work, we develop a maximum-likelihood method for the inference of past changes in population size from microsatellite allelic data. Our method is based on importance sampling of gene genealogies, extended for new mutation models, notably the generalized stepwise mutation model (GSM). Using simulations, we test its performance to detect and characterize past reductions in population size. First, we test the estimation precision and confidence intervals coverage properties under ideal conditions, then we compare the accuracy of the estimation with another available method (MSVAR) and we finally test its robustness to misspecification of the mutational model and population structure. We show that our method is very competitive compared with alternative ones. Moreover, our implementation of a GSM allows more accurate analysis of microsatellite data, as we show that the violations of a single step mutation assumption induce very high bias toward false contraction detection rates. However, our simulation tests also showed some limits, which most importantly are large computation times for strong disequilibrium scenarios and a strong influence of some form of unaccounted population structure. This inference method is available in the latest implementation of the MIGRAINE software package.

Key words: **demographic inference, maximum likelihood, coalescent, importance sampling, microsatellites, bottleneck, population structure, mutation processes, population contraction.**

Introduction

Understanding the demographic history of populations and species is a central issue in evolutionary biology and molecular ecology, for example, for understanding the effects of environmental changes on the distribution of organisms. From a conservation perspective, a severe reduction in population size, often referred to as a "population bottleneck," increases rate of inbreeding, loss of genetic variation, fixation of deleterious alleles, and thereby greatly reduces adaptive potential and increases the risk of extinction (Lande 1988; Keller and Waller 2002; Frankham et al. 2006; Reusch and Wood 2007). However, characterizing the demographic history of a species with direct demographic approaches requires the monitoring of census data, which can be extremely difficult and time consuming (Williams et al. 2002; Schwartz et al. 2007; Bonebrake et al. 2010). Moreover, direct approaches cannot give information about past demography from present-time data. A powerful alternative relies on population genetic approaches, which allow inferences on the past demography from the observed present distribution of genetic polymorphism in natural populations (Schwartz et al. 2007; Lawton-Rauh 2008).

Until recently, most indirect methods were based on testing whether a given summary statistic (computed from genetic data) deviates from its expected value under an equilibrium demographic model (Cornuet and Luikart 1996; Schneider and Excoffier 1999; Garza and Williamson 2001). Because of their simplicity, these methods have been widely used (see, e.g., Comps et al. 2001; Colautti et al. 2005, and the reviews of Spencer et al. 2000 and Peery et al. 2012). But they estimate neither the severity of the contraction nor its age or duration.

Although much more mathematically difficult and computationally demanding, likelihood-based methods outperform these moment-based methods by considering all available information in the genetic data (see Felsenstein 1992; Griffiths and Tavaré 1994a; Emerson et al. 2001, and the review of Marjoram and Tavaré 2006). Among others, the software package MSVAR (Beaumont 1999; Storz and Beaumont 2002) has been increasingly used to infer past

demographic changes. MSVAR assumes a demographic model consisting of a single isolated population, which has undergone a change in effective population size at some time in the past. It is dedicated to the analysis of microsatellite loci that are assumed to follow a strict stepwise mutation model (SMM, Ohta and Kimura 1973). In a recent study, Girod et al. (2011) evaluated the performance of MSVAR by simulation. They have shown that MSVAR clearly outperforms moment-based methods to detect past changes in population sizes, but appears only moderately robust to misspecification of the mutational model: Deviations from the SMM often induce "false" contraction detections on simulated samples from populations at equilibrium. Chikhi et al. (2010) also found a strong confounding effect of population structure on contraction detection using MSVAR. Thus, departures from the mutational and demographic assumptions of the model appear to complicate the inference of past population size changes from genetic data.

This work extends the importance sampling (IS) class of algorithms (Stephens and Donnelly 2000; de Iorio and Griffiths 2004a, 2004b) to coalescent-based models of a single isolated population with a unique past change in population size. Such a model is rather simple compared with complex demographic scenarios occurring in natural populations but inferences based on it can easily be tested by simulation and compared with existing methods. Furthermore, we provide explicit formula for a generalized stepwise mutation model (GSM; Pritchard et al. 1999), following de Iorio et al. (2005).

We have conducted three simulation studies to test the efficiency of our methodology on past contractions (i.e., bottlenecks) and its robustness against misspecifications of the model. The first study aims at showing the ability of the algorithm to detect contractions and to recover the parameters of the model (i.e., the severity of the population size change and its age) on a wide range of contraction scenarios. In the second study, we compared the accuracy of our IS implementation with the Monte Carlo Markov Chain (MCMC) approach implemented in MSVAR. The third study tests the robustness of our method against misspecification of the mutation model, and against the existence of a population structure not considered in the model. Finally, we have applied our methodology on the orangutan data set of Goossens et al. (2006) and compared our results with those obtained with MSVAR. All analyses in these studies were performed using the latest implementation of the MIGRAINE software package, available at http://kimura.univ-montp2.fr/~rousset/Migraine.htm (last accessed July 28, 2014).

New Approaches

Our goal is to obtain maximum-likelihood (ML) estimates for single population models with a past variation in population size as described in the next section. To this end, we describe hereafter the successive steps of the inference algorithm.

Demographic Model

We consider a single isolated population with a unique past size change (fig. 1). The method and our implementation in MIGRAINE are quite general, in the sense that discrete (i.e., sudden), linear or exponential population size contractions or expansions can be considered. However, in agreement with Girod et al. (2011), we found in preliminary tests that parameter inference is less precise for expansions, especially for the time parameter. For this reason, we focused on contraction scenarios to test our method on smallish data sets with reasonable computation times (but see the Discussion section and supplementary fig. S5, Supplementary Material online, for the analysis of an expansion scenario). We denote by $N(t)$ the population size, expressed as the number of genes, t generations away from the sampling time $t = 0$. Population size at sampling time is $N \equiv N(0)$. Then, going backward in time, the population size changes according to a deterministic function until reaching an ancestral population size N_{anc} at time $t = T$. Then, $N(t) = N_{anc}$ for all $t > T$. More precisely,

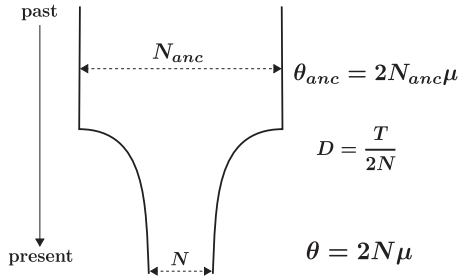


FIG. 1. Representation of the demographic model used in the study. N is the current population size, N_{anc} is the ancestral population size (before the demographic change), T is the time measured in generation since present, and μ is the mutation rate of the marker used. Those four parameters are the canonical parameters of the model. θ, D, and θ_{anc} are the inferred scaled parameters.

$$N(t) = \begin{cases} N\left(\frac{N_{anc}}{N}\right)^{\frac{t}{T}}, & \text{if } 0 < t < T, \\ N_{anc}, & \text{if } t \geq T. \end{cases} \quad (1)$$

To ensure identifiability, the parameters of interest are scaled as $\theta \equiv 2N\mu$, $\theta_{anc} \equiv 2\mu N_{anc}$, and $D \equiv T/2N$, where μ is the mutation rate per locus per generation. We are often interested in an extra composite parameter $N_{ratio} = \theta/\theta_{anc}$, which is useful to characterize the strength of the contraction. Finally, we also consider an alternative parametrization of the model using θ, θ_{anc}, and $D' \equiv \mu T$ in a few situations, for comparison between these two possible parameterizations.

Computation of Coalescent-Based Likelihood with IS

Because the precise genetic history of the sample is not observed, the coalescent-based likelihood at a given point of the parameter space is an integral over all possible histories, that is, genealogies with mutations, leading to the current genetic data. Following Stephens and Donnelly (2000) and de Iorio and Griffiths (2004a), the Monte Carlo scheme computing this integral is based here on IS. The set of possible past

histories is explored through an importance distribution depending on the demographical scenario and the parameter values. The best proposal distribution to sample from is the importance distribution leading to a zero variance estimate of the likelihood. Here this distribution would be the distribution of gene history conditional on the current genetic data, which corresponds to all backward transition rates between successive states of the histories. As computation of these backward transition rates is often too difficult, we substitute this conditional distribution with an importance distribution, and introduce a weight to correct the discrepancy. Like the best proposal distribution, the actual importance distribution is a process describing changes in the ancestral sample configuration backward in time using absorbing Markov chains. However, it does not lead to a zero variance estimate of the likelihood. Better efficiency of the IS proposals allows to accurately estimate likelihoods by considering fewer histories for a given parameter value. Stephens and Donnelly (2000), de Iorio and Griffiths (2004a, 2004b), and de Iorio et al. (2005) suggested efficient approximations that are easily computable. However, the efficiency of the importance distribution depends heavily on the demographic model and the current parameter value.

The first main difference between our algorithm and those described in de Iorio and Griffiths (2004a, 2004b) is the time inhomogeneity induced by the disequilibrium of our demographic model. Demographic models considered in de Iorio and Griffiths (2004a, 2004b) and in Rousset and Leblois (2007, 2012) suppose equilibrium and do not include indeed any change in population sizes. To relax the assumption of time homogeneity in de Iorio and Griffiths (2004b), we modify their equations (see tables 1 and 2 of de Iorio and Griffiths 2004b), so that all quantities depending on the relative population sizes now vary over time because of the population size changes. Thus, we must keep track of time in the algorithm to assign the adequate value to all time-dependent quantities. To see how this is done, consider that the genealogy has been constructed until time T_k, the time of occurrence of the kth event, and that, at this date, n ancestral lineages remain. Under the coalescent with mutations, the expected rate of a mutation event is then $n\theta/2$, and $n(n-1)\lambda(t)/2$ for a coalescence, where $\lambda(t) = N/N(t)$ is the relative population size function describing demographic disequilibrium. $\lambda(t)$ corresponds to parameter $1/q$ in de Iorio and Griffiths (2004b). The total jump rate (i.e., occurrence rate of some event) at time $t \geq T_k$ is then

$$\Gamma(t) = n\big((n-1)\lambda(t) + \theta\big)/2$$

and the next event in the genealogy occurs at time T_{k+1} whose distribution has density

$$\hat{P}(T_{k+1} \in [t, t+\mathrm{d}t]) = \Gamma(t)\exp\left(-\int_{T_k}^{t}\Gamma(u)\mathrm{d}u\right)\mathrm{d}t \quad \text{for } t \geq T_k.$$

Apart from these modifications that follow from the work of Griffiths and Tavaré (1994b), the outline of the IS scheme from de Iorio and Griffiths (2004b) is preserved (see section A1 in the supplementary material, Supplementary Material online, for more details).

We also develop specific algorithms to analyze data under the GSM, with infinite or finite number of alleles. This more realistic mutation model considers that multistep mutations occur and the number of steps involved for each mutation can be modeled using a geometric distribution with parameter p. The original algorithm of Stephens and Donnelly (2000) covers any finite mutation model but requires numerical matrix inversions to solve a system of linear equations, (see, e.g., eqs. 18 and 19 in Stephens and Donnelly 2000). Time inhomogeneity requires matrix inversions each time the genealogy is updated by the IS algorithm. To bypass this difficulty, de Iorio et al. (2005) have successfully replaced the matrix inversions with Fourier analysis when considering an SMM with an infinite allele range. We extended this Fourier analysis in the case of a GSM with an infinite allele range. However, contrarily to the SMM, the result of the Fourier analysis for the GSM is a very poor approximation if the range of allelic state is finite as soon as p is not very small (e.g., <0.1). To consider a more realistic GSM with allele ranges of finite size, we propose to compute the relevant matrix inversions using a numerical decomposition in eigenvectors and eigenvalues of the mutation process matrix, P. Because the mutation model is not time-dependent, this last decomposition is performed only once for a given matrix P. See section A4 in the supplementary material, Supplementary Material online, for details about the GSM implementation.

Finally, several approximations of the likelihood, using products of approximate conditional likelihoods (Cornuet and Beaumont 2007) once the ancestral stable population is reached, and analytical computation of the probability of the last pair of genes, have been successfully tested to speed up computation times (see section A2 in the supplementary material, Supplementary Material online). In what follows, all analyses considering a GSM use these approximations, unless otherwise specified.

Inference Method

Following Rousset and Leblois (2007, 2012), we first define a set of parameter points through a stratified random sample on the range of parameters provided by the user. Then, at each parameter point, the multilocus likelihood is the product of the likelihoods for each locus, which are estimated through the IS algorithm described above. The likelihood inferred at the different parameter point is then smoothed by a Kriging scheme (Cressie 1993). After a first analysis of the smoothed likelihood surface, the algorithm can be repeated a second time to increase the density of the grid in the neighborhood of a first ML estimate. Finally, one- and two-dimensional profile likelihood ratios are computed, to obtain confidence intervals (CI) and graphical outputs (e.g., fig. 2). Section A3 in the supplementary material, Supplementary Material online, explains how we tuned the parameters of the algorithm, namely the range

Table 1. Effects of the Number of Loci and Mutation Processes on the Performance of Estimations for Our Baseline Simulation with $\theta = 0.4$, $D = 1.25$, and $\theta_{anc} = 40.0$ under an SMM, a GSM with $p = 0.22$ and $p = 0.74$, a KAM and Two Situations with Variable Mutation Processes as Described in the Materials and Methods Section.

Case	n_ℓ	p			θ			D			θ_{anc}			CDR (FEDR)
		Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	
SMM														
[0]	10	NA	NA	NA	0.035	0.56	0.056	0.062	0.27	0.068	0.046	0.47	0.46	1 (0)
[A]	25	NA	NA	NA	0.0066	0.31	0.35	0.0079	0.16	0.73	0.0055	0.30	0.84	0.986 (0)
[B]	50	NA	NA	NA	0.015	0.23	0.62	0.0016	0.12	0.69	−0.0071	0.22	0.51	0.982 (0)
GSM 0.22														
[C]	10	0.26	0.91	0.16	0.033	0.51	0.12	0.16	0.66	0.14	0.22	1.33	0.657	0.990 (0)
[D]	50	0.17	0.47	0.12	0.059	0.25	0.44	0.012	0.14	0.75	−0.085	0.39	0.082	1.0 (0)
GSM 0.74														
[E]	10	0.016	0.14	0.0.094	0.137	0.52	0.11	0.42	0.67	$<10^{-12}$	2.46	3.4	$<10^{-12}$	0.965 (0)
[F]	50	0.045	0.081	3.8×10^{-5}	0.34	0.44	$<10^{-12}$	0.40	0.49	$<10^{-12}$	1.6	2.4	$<10^{-12}$	1.0 (0)
KAM														
[G]	10	NA	NA	NA	−0.070	0.64	0.011	0.14	0.71	0.000034	2.11	4.8	0.012	0.84 (0)
[H]	25	NA	NA	NA	−0.027	0.49	0.54	−0.058	0.69	0.54	0.61	2.6	0.041	0.97 (0)
[I]	50	NA	NA	NA	−0.084	0.32	0.085	−0.22	0.51	0.19	0.402	2.74	0.0675	1.0 (0)
var. mut. processes														
[J]	10	0.18	0.91	0.00070	0.12	0.65	9.8×10^{-5}	0.31	0.92	0.020	0.67	2.3	0.014	0.96 (0)
[K]	50	0.097	0.49	0.020	0.083	0.27	0.0055	0.040	0.18	0.99	−0.22	0.45	4.3×10^{-7}	0.97 (0)

NOTE.—n_ℓ, number of loci; Rel. Bias, relative bias; KS, P value of the Kolmogorov–Smirnov test for departure of ECDF of LRT P values from uniformity; CDR, contraction detection rate; FEDR, false expansion detection rate; RRMSE, relative root mean square.

Table 2. Effects of Scaling the Time by the Mutation Rate Instead of Population Size for Different Timings, $\theta = 0.4$ and $\theta_{anc} = 40.0$.

True D or D'	Case	Scaling	θ			D or D'			θ_{anc}		
			Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS
0.125 (0.05)	[3]	$D = T/2N$	2.6	5.7	$< 10^{-12}$	0.30	0.65	2.3×10^{-4}	0.040	0.24	0.21
	[17]	$D' = T\mu$	2.3	4.5	1.4×10^{-9}	2.6	5.61	3.6×10^{-10}	0.0045	0.26	0.016
1.25 (0.5)	[0]	$D = T/2N$	0.035	0.56	0.056	0.062	0.27	0.068	0.046	0.47	0.46
	[18]	$D' = T\mu$	0.053	0.54	0.056	0.14	0.82	0.127	−0.0026	0.46	0.857
3.5 (1.4)	[7]	$D = T/2N$	−0.026	0.38	0.82	0.0038	0.50	0.51	0.32	1.7	0.098
	[19]	$D' = T\mu$	−0.013	0.37	0.91	0.020	0.71	0.12	0.389	2.11	0.40
5.0 (2.0)	[8]	$D = T/2N$	−0.107	0.36	0.33	−0.11	0.42	0.58	0.46	2.4	0.46
	[20]	$D' = T\mu$	−0.088	0.31	0.50	−0.16	0.52	0.68	0.49	2.5	0.60

NOTE.—Computations are done considering only data sets with a significant contraction detection. No effect of such scaling is detected on CDRs nor on FEDRs. Rel. Bias, relative bias; KS, P value of the Kolmogorov–Smirnov test for departure of ECDF of LRT P values from uniformity.

of parameters, the size of parameter points, and the number of genealogical histories explored by the IS algorithm.

A genuine issue, when facing genetic data, is to test whether the sampled population has undergone size changes or not. Thus, we derived a statistical test from the methodology presented above. It aims at testing between the null hypothesis that no size change occurred (i.e., $N = N_{anc}$) and alternatives such as a population decline or expansion (i.e., $N \neq N_{anc}$). At level α, our test rejects the null hypothesis if and only if 1 lies outside the $1 - \alpha$ CI of the ratio $N_{ratio} = N/N_{anc}$.

All those developments are implemented in the `MIGRAINE` software package. A detailed presentation of the simulation settings and validation procedures used to test the precision and robustness of the method is given in the Materials and Methods section.

Results

Two Contrasting Examples

We first draw a contrast between two typical simulation outputs (figs. 2*a* and *b*), which must be kept in mind to understand further simulation results. The first one (case [0]), corresponding to our baseline simulation ($\theta = 0.4$, $D = 1.25$, and $\theta_{anc} = 40.0$), is an ideal situation in which the inference algorithm performs well due to the large amount of information in the genetic data, resulting in a likelihood surface with clear peaks for all parameters around the ML values. The contraction signal is highly significant and is clearly seen in the (θ, θ_{anc}) plot on figure 2*a*, as the ML peak is above the 1:1 diagonal. The second example (case [10]) is a more difficult situation, where the population has undergone a much weaker contraction ($\theta = 0.4$, $D = 1.25$, and $\theta_{anc} = 2.0$) that does not leave a clear signal in the genetic data. In such a

situation, there is not much information on any of the three parameters, resulting in much flatter funnel- or cross-shaped two-dimensional likelihood surfaces. A contraction signal is visible on the cross-shaped (θ, θ_{anc}) plot on figure 2*b*, but is not significant.

Implementation and Efficiency of IS on Time-Inhomogeneous Models

Simulation tests show that our implementation of de Iorio–Griffiths' IS algorithm for a model of a single population with past changes in population size and stepwise mutations is

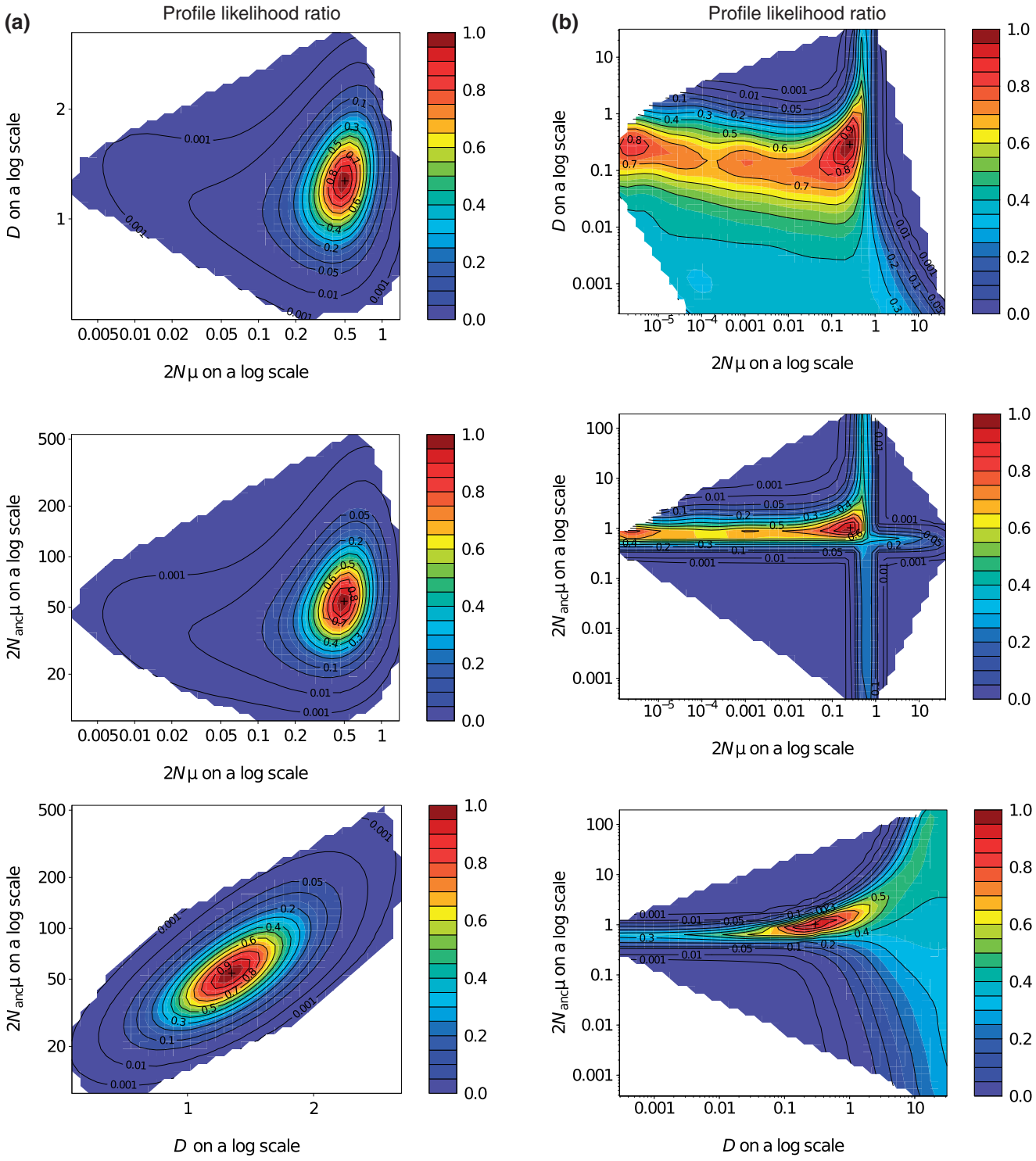


Fig. 2. Examples of typical two-dimensional profile likelihood ratios for two data sets generated with (*a*) $\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 40.0$ (case [0]) and (*b*) $\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 2.0$ (case [10]). The likelihood surface is inferred from 1,240 points in two iterative steps (*a*), and 3,720 points in three iterative steps (*b*) as described in section A3 in the supplementary material, Supplementary Material online. The likelihood surface is restricted to the region of the parameter space where the likelihood was actually estimated.

very efficient under most demographic situations tested here. Similar results are obtained for two different approximations of the likelihood (see section A2 in the supplementary material, Supplementary Material online). First, computation times are reasonably short: For a single data set with hundred gene copies and ten loci, analyses are carried out within few hours to 3 days on a single processor, even for the longest analyses with four parameters under the GSM. Second, likelihood ratio test (LRT) *P* value distributions generally indicate good CI coverage properties (see the Materials and Methods section). Empirical cumulative distribution functions (ECDF) of LRT *P* values for all scenarios, shown in section F in the supplementary material, Supplementary Material online, are most of the time close to the 1:1 diagonal as shown in figure 3*a* for our baseline scenario.

Exceptions to those global trends are of two types: 1) For scenarios in which there is not much information on one or more parameters, such as the example of a weak contraction described in the previous section, likelihood surfaces are flat on the corresponding axes (fig. 2*b*). Such scenarios with very few information on one or more parameters are discussed in the next section. In such situations, asymptotic LRT *P* value properties were not always reached (e.g., fig. 3*b*) because of the small number of loci (i.e., 10) considered. Analyzing more loci should improve CI coverage properties in those situations. 2) The more recent and the stronger contractions are, the less efficient are the IS proposals, because they are computed under equilibrium assumptions as detailed in the New Approaches section and section A1 in the supplementary material, Supplementary Material online. Contrarily to the first situation, likelihood surfaces are then too much peaked, and MLs are located in the wrong parameter region. The main defect we observed is thus a positive bias minimizing the contraction strength and bad CI coverage properties for θ, when the number of explored ancestral histories is too small (results not shown). Consideration of 2,000 ancestral histories per parameter point (as for most simulations in this study, see section A1 in the supplementary material, Supplementary Material online) ensures good CI coverage properties, except for some extreme situations. For a very recent and strong past contraction ($\theta = 0.4$, $D = 0.25$, and $\theta_{anc} = 400.0$), increasing the number of ancestral histories sampled for each point up to 200,000 decreases relative bias and relative root mean square error (RRMSE) on θ but does not provide satisfactory CI coverage properties (supplementary fig. S64, Supplementary Material online). Increasing the number of loci decreases the bias and RRMSE for D and θ_{anc} but not for θ. Such results have however only been observed in those few extreme situations with $\theta/\theta_{anc} \leqslant 0.001$ and $D \leqslant 0.25$. Figure 3*c* illustrates a more realistic situation of a very recent but not too strong population size contraction where the two defects described above are cumulated (case [3], with $\theta = 0.4$, $D = 0.125$, and $\theta_{anc} = 40.0$).

ECDF of LRT *P* values for all parameters also more often depart from the 1:1 diagonal when the mutation model moves away from a strict stepwise model and when a low number of loci (i.e., 10 or 25) is used for inference (e.g., for a GSM with $p = 0.74$, where p is the parameter of the geometric distribution of mutation step sizes, and for a K-allele model [KAM]: See table 1, cases [E], [G], and [H]). In those situations, ECDF of LRT *P* values (supplementary figs. S10, S12, and S13, Supplementary Material online) indicate slightly too narrow CI, especially for parameters for which there is not much information (e.g., θ_{anc} and D). Considering a larger number of loci (i.e., 50) restores good CI coverage properties for the KAM but not for the GSM with $p = 0.74$ (cf. perfect LRT *P*-value distributions for case [I] but not for [F]: See table 1 and supplementary figs. S11 and S14, Supplementary Material online). This suggests that the above incorrect ECDF of LRT *P* values are partly due to the small amount of information carried by a low number of loci but also due to slight misspecifications of the mutation model (i.e., the number of possible allelic states in the GSM, see section A2 in the supplementary material, Supplementary Material online).

Power and Precision under Ideal Conditions

Results for the power of the contraction detection test and for the precision of the estimates under ideal conditions (i.e., same model used for simulations and analyses) with ten loci are presented in figures 4 and 5.

Contraction detection rates (CDRs) are highest when contractions are not too recent, nor too old or too weak: A contraction is detected at a 5% level in more than 95% of the data sets when the contraction occurred more than 25 generations ago but less than 1,400 generations ago ($0.0625 < D < 3.5$, fig. 5) and when the ancestral population size is at least 20 times the actual size. Detection rates are then decreasing for more recent, older or weaker contractions, but stay high ($>$50%) in many of those situations. In this first simulation set, only extremely weak ($\theta_{anc} = 5\theta$, case [10], fig. 4) or extremely old ($D = 7.5$, case [9], fig. 5) contractions show CDRs below 50%.

Precision of parameter inference is highly dependent on the scenario considered. First, global precision on all parameters increases with the strength of the contraction. Reasonable precision, for example, a relative bias between −20% and 100% and RRMSE below 100%, is only obtained when the ancestral population size is larger than 20 times the actual population size (fig. 4). However, for weaker contractions, estimates of the order of magnitude for some but not all parameters can often be obtained. Second, precision of the inference of each parameter is strongly dependent on the timing of the population size change and this is well represented on figure 5. Parameter θ is inferred with good precision when the contraction is not too recent, for example, older than 200 generations in our simulation ($D > 0.5$). For more recent contractions, relative biases are at least 130% and RRMSE larger than 300%. On the other hand, θ_{anc} is well estimated for recent and intermediate contractions. For old contraction, for example, older than 1,000 generations ($D > 2.5$), relative bias and RRMSE are often greater than 100%. Inference of D shows an intermediate pattern, with more precise inferences for intermediate timings. Relative bias and RRMSE on D first decrease with time for contractions that occurred from 10 to 500 generations ago

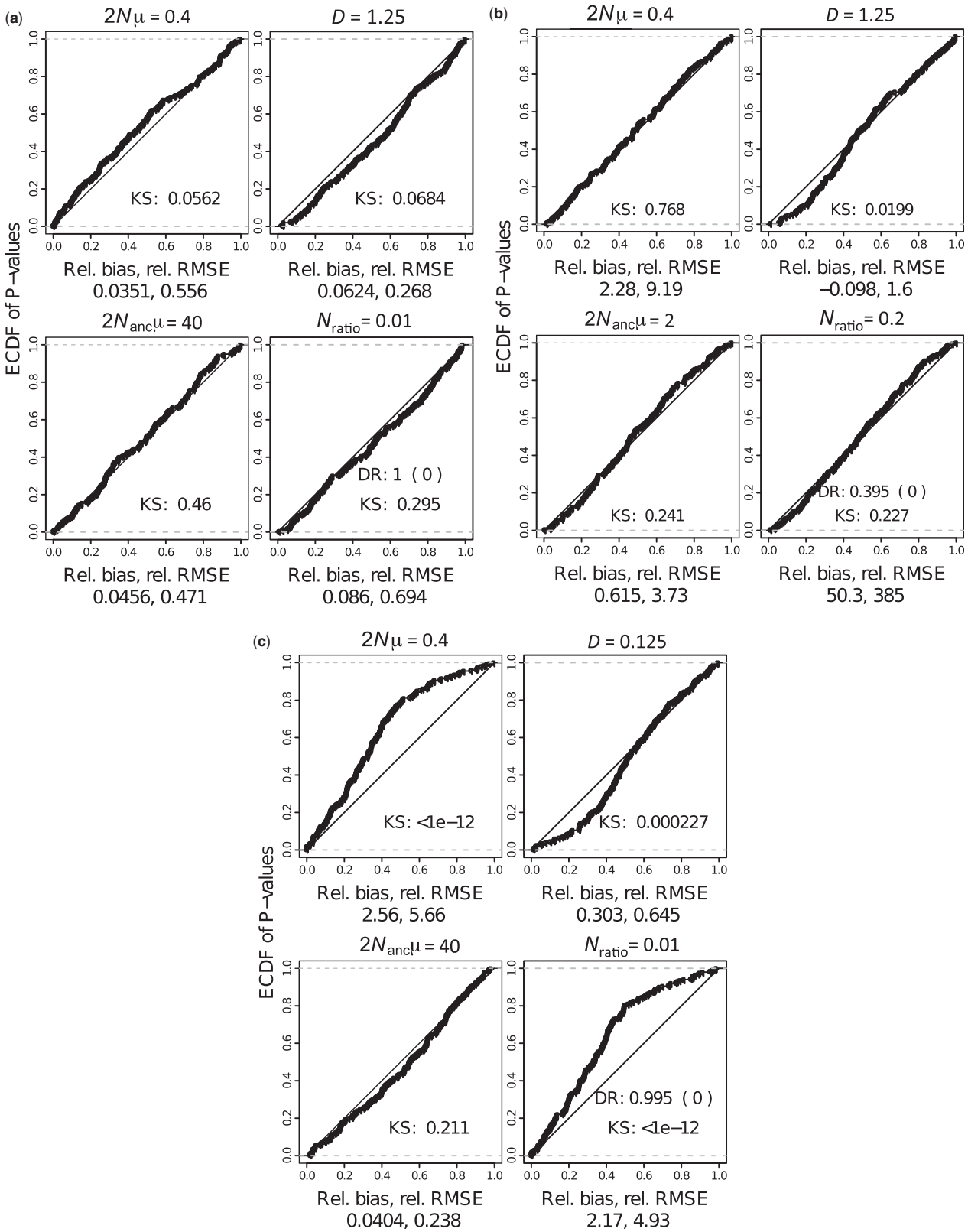


FIG. 3. ECDF of *P* values of LRTs for (*a*) the baseline scenario (case [0]), with $\theta = 2N\mu = 0.4$, $D = T/2N = 1.25$ and $\theta_{anc} = 2N_{anc}\mu = 40.0$; (*b*) a very weak contraction scenario (case [10]), with $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 2.0$; and (*c*) a recent contraction scenario (case [3]), with $\theta = 0.4$, $D = 0.125$ and $\theta_{anc} = 40.0$. Mean relative bias (rel. bias, computed as $\sum$(observed value − expected value)/expected value) and relative root mean square error (rel. RMSE, computed as $\sum[(\text{observed value} - \text{expected slope})/\text{expected value}]^2$) are reported as well as the contraction detection rate (DR) and FEDR in parentheses after DR. KS indicate the *P* value of the Kolmogorov–Smirnov test for departure of LRT *P* values distributions from uniformity.

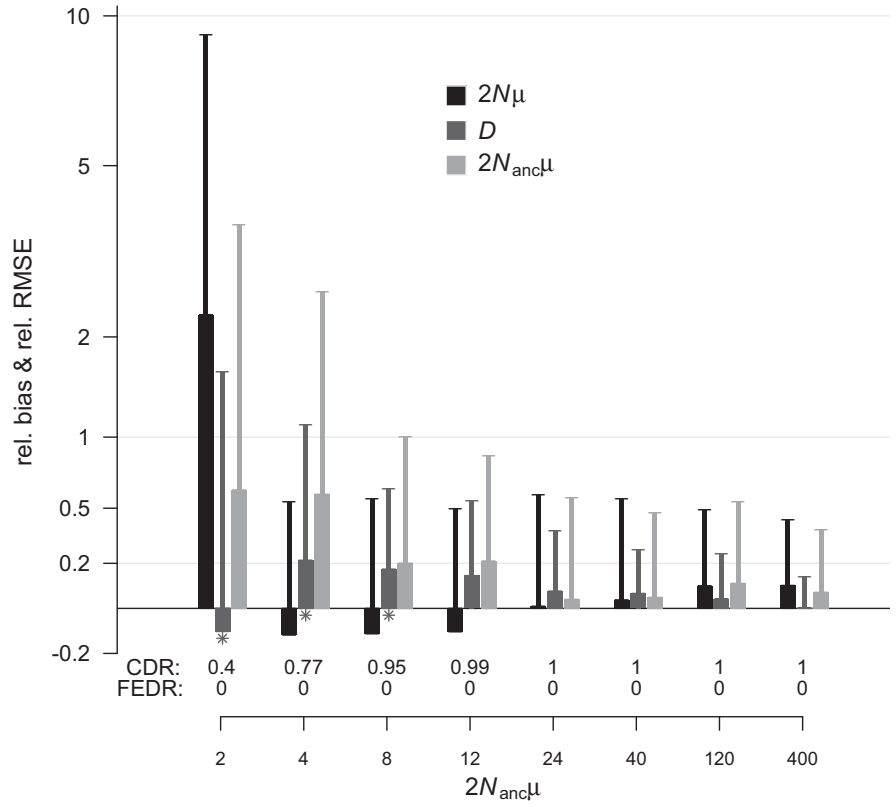


Fig. 4. Effect of the strength of the population size contraction on the inference of each parameter of the model (cases [0] and [10]–[16]: 2N\μ=0.4). Relative bias is indicated by the large bars, and RRMSE by the thin lines. Stars indicate low *P* values of the Kolmogorov–Smirnov test on the distribution of LRT *P* values (i.e., <0.05). CDR, contraction detection rate; FEDR, false expansion detection rate.

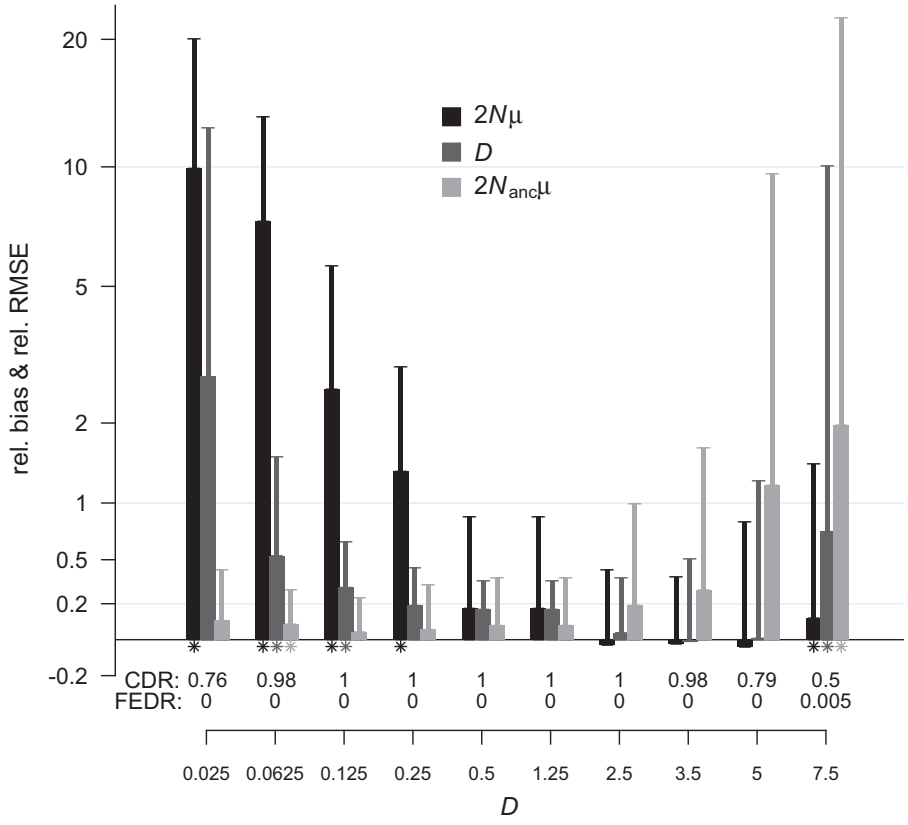


Fig. 5. Effect of the timing of the population size contraction on the inference of each parameter of the model (cases [0]–[9]). See figure 4 for details.

($0.025 < D < 1.25$), and then increase with time for older contractions.

Our baseline scenario (case [0]) with $D = 1.25$, thus seems to be the most favorable situation, for which inference of all parameters is relatively good given the small number of loci considered (ten loci; figs. 3*a*, 4, and 5). Relative biases are only about a few percent, but RRMSE values vary from 20% to 60% indicating different precision levels for the different parameters. D is the most precisely inferred parameter, followed by θ_{anc} and then by θ. The large RRMSE values are expectedly reduced when considering a larger number of loci, and reach 10–22% for all parameters when 50 loci are used (table 1).

A few simulations have been analyzed by inferring the parameter $D' = T\mu$ instead of $D = T/2N$. For those simulations, we considered $\theta = 0.4$ and $\theta_{anc} = 40.0$ as in the baseline situation and four different timings ($D = \{0.0125; 1.25; 3.5; 5.0\}$, cases [17]–[20]). Our results show that scaling time by the mutation rate globally decreases the precision of the estimation of the time parameter and does not have much effect on the other parameters θ and θ_{anc} (table 2). Relative bias and RRMSE are always higher, and sometimes much higher, on D' than on D. No effect of such scaling is detected on CDRs nor on the false expansion detection rate (FEDR, results not shown).

Effect of Mutational Processes

To test the robustness to mutational processes, we first analyzed under a strict SMM samples simulated under a stable population model with $\theta = 2.0$ and a GSM with $p = 0.22$ and 0.74 for the ten loci considered (cases [21] and [22]). For $p = 0.22$, 67% of the data sets show significant signals of false contraction. This false contraction detection rate (FCDR) increases up to 100% for $p = 0.74$. Among all simulations analyzed for these two situations, a false expansion is detected in a single data set, out of 200, with $p = 0.22$. The same simulations analyzed under a GSM show detection of false contractions in 6% and 5% of the data sets, as well as detection of false expansion in 7.5% and 6% of the data sets, for $p = 0.22$ and 0.74, respectively (cases [23] and [24]).

Next, we simulated and analyzed data under a GSM or a KAM with a past contraction corresponding to our baseline scenario with $\theta = 0.4$, $D = 1.25$, and $\theta_{anc} = 40.0$, and with $p = 0.22$ and 0.74 for the GSM (with ten loci: Cases [C], [E] for the GSM and [G] for the KAM; with 50 loci: Cases [D], [F] for the GSM and [I] for the KAM, table 1). Compared with analyses under an SMM, CDRs slightly decrease when p increases but still remain very high (e.g., $\geqslant 95\%$ with ten loci) for $p \leqslant 0.74$. On the other hand, precision of the estimations strongly differs between different parameters. Inference of p globally shows large relative bias and RRMSE for $p = 0.22$ but is very precise for $p = 0.74$. For θ, using different mutation models does not change much the precision of the estimations. For D and θ_{anc}, the mutation model has much stronger effects, showing less precise estimations for increasing p values, as well as more departure from the diagonal of the ECDF of LRT P values. However, increasing the number of loci from 10 to 50 restores good precision for the estimation of all parameters, except for the KAM, as well as good LRT P value distributions. Finally, unaccounted variation in mutation processes and mutation rates across loci slightly increases biases and RRMSE, and induces poor CI coverage properties for the mutation parameters θ and θ_{anc} (cases [J] and [K] in table 1, mutation processes detailed in the Materials and Methods section). The effect is similar but weaker for D, for which good precision and good CI coverage are observed with 50 loci.

Effect of Population Structure

We first considered the presence of a local population structure by analyzing samples generated under stable continuous populations with various levels of dispersal and different spatial scale of sampling (table 3). All data sets were simulated under a GSM with $p = 0.22$. Our results show that isolation-by-distance (IBD) structure induces high FCDRs, strongly depending on the strength of IBD as well as the spatial scale of sampling (cases [25]–[29], table 4). The stronger the IBD structure is, the higher FCDR is, varying from 15% for weak IBD with $\sigma^2 = 100$ to almost 70% for strong IBD with $\sigma^2 = 1$, for a small sampling scale (σ^2 is the mean squared parent–offspring dispersal distance and is inversely related to the strength of IBD). Considering larger sampling scales by sampling on the whole population area not only strongly decreases FCDRs to values less than or equal to 16% for all levels of IBD but also induces false expansion detection in 4%, at most, of the data sets.

Using a second set of simulations under IBD with past reductions in population size, we mimic a reduction in habitat area for organisms with limited dispersal (table 3). We show in table 5 that the presence of IBD slightly decreases CDRs, for example, from 99.5% down to 90% for very strong IBD. Strong IBD associated with small scale sampling also induces negative relative bias on θ, large positive biases on p, D, and θ_{anc}, as well as bad CI coverage properties as shown by KS values (table 5) and ECDF of LRT P values (supplementary figs. S46 and S48, Supplementary Material online). Weaker IBD structure shows similar but weaker effects (supplementary figs. S50 and S52, Supplementary Material online). Increasing sample scale increases CDRs for situations under very strong IBD only, but strongly decreases relative biases and RRMSE on all parameters except θ_{anc}. For all situations and all parameters, considering a large sample scale allows better CI coverage properties (table 5 and supplementary figs. S47, S49, S51, and S53, Supplementary Material online).

We finally tested the influence of an island population structure with varying levels of migration and population sizes (tables 6 and 7 and supplementary tables S3 and S4, Supplementary Material online). Our results first show that sampling a single island from a stable-structured population induces high FCDRs, from 11% to 52% depending on the level of population structure. With such local sampling scheme, the relationship between FCDRs and the level of population structure is complex. Increasing sampling scale by sampling three to ten populations instead of a single one has two major antagonistic effects: It strongly increases FCDRs up to 100%

Table 3. Simulated Data Sets with Population Structure.

Local Population Structure

The simulated IBD populations are composed of individuals set at the nodes of a regular lattice, whose size can vary. A past reduction in population size is thus modeled as a reduction of the habitat area keeping a constant density of individuals. Various levels of localized dispersal were simulated through truncated Pareto distributions with mean squared parent-offspring dispersal distance, say σ^2, varying in {1; 4; 10; 20; 100}.

Parameters of the IBD Populations:

- At equilibrium: $\theta = 4.0$ with a 32×31 lattice (hence $N = 1{,}984$ genes)
- Including an habitat contraction: $(D, \theta, \theta_{anc}) = (1.25, 0.4, 40.0)$ with lattices of sizes from 10×10 ($N = 200$) to 100×100 ($N_{anc} = 20{,}000$) backward in time

Simulated Sampling Schemes:

100 genes sampled

- on a 5×10 lattice in the center of the population [small sample scale], or
- regularly on the whole area (i.e., one individual every four nodes) [large sample scale].

Island Population Structure

We considered models with $d = 10$ demes of equal size N_d genes, varying in {20; 200; 1,000; 2,000}, and exchanging migrants at rate m between pairs of demes, varying in {0.000025; 0.00025; 0.0025; 0.005; 0.025; 0.075; 0.25}. The model is fully characterized by the scaled parameters $\theta = 2dN_d\mu$ and $M = 2N_dm$. When past contractions occurred, deme sizes N_d decreased forward in time but migration rates m are kept constant in time. Values of M reported below correspond to scaled migration rates at sampling time $t = 0$.

Parameters of the Island Populations:

- $\theta \in \{4.0, 20.0\}$ and $M \in \{0.01, 0.1, 1.0, 10.0, 30.0, 100.0\}$ without population size changes
- $\theta = 0.4$, $M \in \{0.01, 1.0, 100.0\}$ and a contraction with parameters $D = 1.25$, $\theta_{anc} = 40.0$

Simulated Sampling Schemes:

Samples of 100 genes picked at random

- from a single deme [small sample scale], or
- from three demes [large sample scale], or
- from all demes [very large sample scale].

Table 4. Effects of IBD on the Detection of False Contraction and Expansion Signals in Constant-Size Populations with $\theta = 4.0$.

IBD Strength (σ^2)	Case	Sampling Scale	
		Small (10×5) FCDR/FEDR	Large (28×28) FCDR/FEDR
1	[25]	0.67/0.0	0.16/0.005
4	[26]	0.54/0.0	0.095/0.010
10	[27]	0.49/0.0	0.080/0.010
20	[28]	0.41/0.005	0.090/0.020
100	[29]	0.145/0.005	0.11/0.040

NOTE.—Sample scales correspond to the area (expressed as the number of lattice nodes) from which a spatially homogeneous sample is taken. σ^2 is the mean squared parent–offspring dispersal distance and is inversely related to the strength of IBD. See the Materials and Methods section for details.

for highly structured situations (i.e., $M \leqslant 1.0$, with $M \equiv 2N_dm$); but it also reduces FCDRs for less structured populations, down to values around 10% for the larger sampling scale. Interestingly, sampling a single gene per deme strongly reduces the effect of population structure and decreases FCDR values to 2% in a situation with intermediate migration rates (i.e., $M = 1.0$, supplementary table S3, Supplementary Material online). Note that a few false expansions are also detected among all those simulations but always in less than 10% of the data sets. Finally, we can also note that, as shown in table 6 and supplementary table S3, Supplementary Material online, increasing the total diversity θ or considering an SMM also modifies the effect of population structure but more simulations are needed to find global trends for effect of genetic diversity and mutational processes.

Our last set of simulations under an island model with past reduction of population sizes shows an extremely strong effect of the sampling scale and the level of population structure (table 7 and supplementary table S4, Supplementary Material online). Compared with unstructured situations with CDR = 99% (case [G]), sampling a single deme strongly reduces CDRs to values from 88% for weak population structure with $M = 100.0$ down to 0.5% for highly structured populations with $M = 0.01$. Intermediate structure with $M = 1.0$ also leads to small CDR of 7% in our simulations. With such a small sampling scale, parameter estimation is clearly inaccurate when population structure is not very weak (e.g., $M \ll 100.0$), showing strong bias, large RMSE, and bad coverage properties of CI (table 7 and supplementary figs. S54–S57, Supplementary Material online). This is observed whether the parameters considered are local values for one deme or global values for the whole population (results not shown). The effect however decreases with higher levels of migration, and parameter inference is relatively accurate with $M = 100.0$ for all parameters except p, and shows reasonable CI coverage properties (supplementary figs. S58 and S59, Supplementary Material online). With a single exception concerning samples with a single gene per deme (supplementary table S4, Supplementary Material online), increasing sampling scale generally increases CDRs. However, contrarily to the results obtained for IBD, sampling at a large scale, even with a single gene per deme, does not improve all parameter inferences nor all CI coverage properties. Sampling at a larger scale seems to often allow better estimation of θ_{anc}, but the effect on all other parameters is highly dependent on the demographic scenario considered, and no clear conclusion can thus be drawn from our simulations.

Table 5. Effects of IBD Structure on the Detection and Characterization of a Past Contraction.

IBDLevel (σ^2)	Case	Sample Scale	p			θ			D			θ_{anc}			CDR(FEDR)
			Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	
1	[30]	Small	0.71	1.2	6.6×10^{-9}	−0.30	0.43	1.4×10^{-10}	0.90	1.2	$<10^{-12}$	0.25	1.2	0.061	0.9 (0)
	[31]	Large	0.20	0.88	0.14	−0.0577	0.46	0.057	0.46	0.79	$<10^{-12}$	0.51	1.4	0.22	0.99 (0)
4	[32]	Small	0.50	1.1	5.1×10^{-9}	−0.29	0.45	2.7×10^{-7}	0.43	0.78	6.3×10^{-9}	0.25	1.4	0.46	0.96 (0)
	[33]	Large	0.22	0.89	0.74	−0.12	0.49	0.020	0.27	0.54	4.4×10^{-4}	0.37	1.3	0.93	0.96 (0)
10	[34]	Small	0.41	1.0	1.4×10^{-5}	−0.19	0.42	3.4×10^{-6}	0.39	0.72	1.9×10^{-6}	0.40	2.15	0.0031	0.98 (0)
	[35]	Large	0.23	0.89	0.12	−0.11	0.44	0.15	0.22	0.52	1.4×10^{-4}	0.31	1.2	0.33	0.97 (0)
100	[36]	Small	0.35	0.96	1.6×10^{-4}	−0.094	0.41	0.11	0.26	0.55	6.0×10^{-4}	0.22	1.2	0.71	0.97 (0)
	[37]	Large	0.19	0.86	0.40	−0.017	0.48	0.67	0.13	0.46	0.14	0.26	1.3	0.52	0.96 (0)

NOTE.—Samples are simulated from a single continuous population under IBD that has undergone a past contraction with $\theta = 0.4$, $D = 1.25$, and $\theta_{anc} = 40.0$. The small sampling scale corresponds to 100 genes sampled on a 5×10 area, expressed in lattice nodes, in the center of the population; the large sampling scale corresponds to a spatially homogeneous sample of 100 genes taken on the whole population area (i.e., one individual every four nodes). See the Materials and Methods section for details.

Table 6. Effects of an Island Population Structure on the Detection of False Contraction or Expansion Signals in Constant-Size Populations.

Island Model Settings		Case	Sampling Scale		
θ	M		Small One Island FCDR/FEDR	Large Three Islands FCDR/FEDR	Very Large All Ten Islands FCDR/FEDR
4	0.01	[38]	0.11/0.025	1.0/0.0	1.0/0.0
	0.1	[39]	0.32/0.02	1.0/0.0	1.0/0.0
	1.0	[40]	0.21/0.0	0.84/0.0	0.76/0.0
	10.0	[41]	0.52/0.0	0.32/0.0	0.10/0.010
	30.0	[42]	0.38/0.0	0.18/0.01	0.10/0.015
	100.0	[43]	0.19/0	0.085/0.015	0.11/0.026
20	1.0	[44]	0.78/0.0	0.91/0.0	0.64/0.0

NOTE.—Samples are simulated from a stable island model with n_d demes, $\theta = 2n_d N_d \mu$ and scaled migration rate $M \equiv 2N_d m$. Sampling scale corresponds to the number of sampled demes. See the Materials and Methods section for details. FCDR, false contraction detection rate; FEDR, False expansion detection rate.

Table 7. Effects of an Island Population Structure on the Detection and Characterization of a Past Contraction.

Gene Flow Level (M)	Case	Sampling Scale	p			θ			D			θ_{anc}			CDR (FEDR)
			Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	
0.01	[45]	Small	−0.081	1.0	0.026	−0.61	0.80	7.0×10^{-10}	−0.24	1.2	0.060	−0.99	1.0	$<10^{-12}$	0.005 (0.040)
	[46]	Very large	2.2	2.3	$<10^{-12}$	−0.77	0.81	$<10^{-12}$	−0.66	0.67	$<10^{-12}$	−0.17	0.62	8.5×10^{-4}	1.0 (0)
1.0	[47]	Small	1.6	1.9	$<10^{-12}$	−0.32	0.83	1.7×10^{-5}	0.41	1.8	9.9×10^{-8}	−0.94	0.96	$<10^{-12}$	0.070 (0.070)
	[48]	Very large	−0.0027	0.61	4.2×10^{-8}	−0.72	0.80	$<10^{-12}$	−0.60	0.60	$<10^{-12}$	−0.020	0.50	5.7×10^{-5}	1.0 (0)
100	[49]	Small	0.69	1.2	4.2×10^{-6}	−0.070	0.43	0.64	0.31	0.80	0.0085	0.043	1.1	0.25	0.88 (0)
	[50]	Very large	0.014	0.83	0.0085	0.13	0.55	0.013	0.041	0.39	0.033	0.24	1.1	0.78	0.99 (0)

NOTE.—Samples are simulated from a 10-island model in which each subpopulation has undergone a past contraction, with $\theta \equiv 2dN_d\mu = 0.4$, $D \equiv T/(2dN_d) = 1.25$, and $\theta_{anc} \equiv 2dN_{d,anc}\mu = 40.0$, and varying scaled migration rate $M \equiv 2N_d m$. See the Materials and Methods section for details. Mean relative bias and RRMSE are reported as well as the CDR and the FEDR. Sampling scale corresponds to the number of sampled demes: 1) Small for one sampled deme and 2) very large for ten sampled demes. KS indicate the P value of the Kolmogorov–Smirnov test for departure of ECDF of LRT P values from uniformity.

Inferences on the Orangutan Data Set

The orangutan analyses with `MIGRAINE` show consistent results for the three pooled samples RS1, RS2 and RS1+RS2 (table 8 and supplementary fig. S4, Supplementary Material online), and for each subsample separately (supplementary table S5, Supplementary Material online). With one exception for subsample S9, all analyses detect a strong and recent past contraction of population size and allow concordant estimation of the model parameters. However, as expected due to lower sample sizes, analyses of each subsample give less precise inferences than for the pooled sampled, and we will thus focus on the pooled sample results. First, θ_{anc}'s inference is extremely consistent across analyses, and shows a high precision level with point estimates around 7.5 and narrow CI (i.e., around $[5 - 12]$). Second, the time when the contraction started in the past, D, is inferred with slightly less precision but all analyses support a relatively recent contraction with upper bounds of CI below 1.0. Third, in agreement with our simulation results, there is much less information about θ because the inferred contraction is recent. High θ values are

Table 8. Point estimates and 95% CI for All Model Parameters Obtained from the Analyses of the Orangutan Data Set.

Sample (size)	p	θ	D	θ_{anc}	N_{ratio}
RS1 (106)	0.40 [0.15 − 0.61]	0.0048 [10^{-5} − 0.36]	0.30 [0.17 − 0.80]	7.7 [5.2 − 11.5]	0.00063 [10^{-6} − 0.049]
RS2 (89)	0.42 [0.20 − 0.65]	0.00035 [10^{-5} − 0.41]	0.31 [0.19 − 0.48]	7.4 [5.4 − 9.9]	4.8×10^{-5} [1.2×10^{-6} − 0.058]
RS1+RS2 (195)	0.37 [0.15 − 0.59]	0.013 [8×10^{-5} − 0.67]	0.14 [0.045 − 0.52]	7.8 [5.3 − 11.8]	0.0016 [10^{-8} − 0.090]
		$N/2$	T_{years}	$N_{anc}/2$	
RS1		3 [1 − 180]	90 [17 − 14,400]	3,850 [2, 600 − 5,750]	
RS2		1 [1 − 205]	31 [19 − 9,840]	3,700 [2, 700 − 4,950]	
RS1+RS2		7 [1 − 335]	98 [5 − 17,420]	3,900 [2,650 − 5,900]	

NOTE.—More detailed results and a figure showing profile likelihood ratios are available in section D in the supplementary material, Supplementary Material online. The lower part of the table presents the estimates of population sizes expressed as numbers of diploid individuals ($N/2$ and $N_{anc}/2$) and times in years (T_{years}) obtained after a conversion of MIGRAINE results using a fixed mutation rate of 5×10^{-4} mutation per locus per generation and a generation time of 25 years. The "confidence" intervals reported for T_{years} is likely to be much larger than the true 95% CI, because we used the 95% CI bounds of D and θ successively to compute this interval. Sample sizes are given in number of diploid individuals. See the Materials and Methods section for details.

rejected, with upper bounds of CI between 0.4 and 0.7, but the likelihood profile is always very flat for low θ values (e.g., supplementary fig. S4, Supplementary Material online). As a result, many ML estimates were inferred at the lower bound of the explored parameter space for θ in preliminary analyses. Moreover, supplementary figure S4, Supplementary Material online, shows a clear trade-off between θ and D for high likelihood with low θ values associated with high values for D. Unrealistically, low θ values (e.g., 10^{-9}) were thus associated with high D values (e.g., about 1.0) in the first preliminary analyses. For the final analyses, we thus restricted the parameter space to biologically realistic θ values (i.e., $\theta \geq 10^{-5}$, as the latter value corresponds to a population size of a single individual when considering a very small mutation rate for microsatellite markers of 2.5×10^{-6}). Because of this uncertainty in the estimation of θ, inference of N_{ratio} clearly shows a population size ratio below 1, but the CI are very large, even when θ values are constrained. Finally, the GSM parameter p is also inferred with intermediate precision but, in all analyses, its point estimates are relatively high, for example, around 0.4, compared with values generally found in the literature.

Discussion

In this study, we adapted de Iorio–Griffiths' IS algorithm to consider a single population model with varying size and different mutation models. We investigated its performance in detecting past contractions of population size as well as estimating the model parameters. We did not explore expansion scenarios because preliminary simulations showed that parameter inference is less precise for expansions than for contractions (as shown in section E in the supplementary material and fig. S5, Supplementary Material online, and in Girod et al. 2011). For a majority of expansion scenarios, good precision is only obtained when considering large sample sizes, for example, 500 haploid individuals genotyped at 50 loci, which implies large computation times (results not shown). Likewise, preliminary tests under a model with a founder event followed by a demographic expansion showed that correct parameter inference could not be obtained with the current version of our IS algorithms within reasonable computation times. This is so because the strong disequilibrium of this model induces high variance of the likelihood estimation, which therefore requires the computation of a very large number of ancestral histories. We thus focused on a model with a single past contraction event to test the effect of the timing and amplitude of the past demographic change, and study the robustness of inferences to misspecifications of the mutational and population structure models. Our results allow us to illustrate both the strengths and the imperfections of the method.

Performances under Ideal Conditions

First, over all simulations considered in this study, LRTs for CI coverage indicate that our implementation is correct and produces accurate estimates of the likelihood surface with reasonable computation times, except in a few situations with extremely strong demographic disequilibrium (i.e., for recent, e.g., $D \leqslant 0.25$, and strong contractions, e.g., $N_{anc}/N \geqslant 1,000$). For the later situations, much longer runs are needed to obtain good CI coverage. This shows that the efficiency of de Iorio–Griffiths' IS algorithm, based on time-homogeneous demographic assumptions, strongly depends on the extent of the demographic disequilibrium considered, which can be roughly quantified by the ratio of the amplitude of the population size change divided by its duration. Our results also show that inference based on time-homogeneous IS algorithms is practically intractable for the most extreme situations.

Second, our simulations show very good performances in terms of detection of past decreases in population size. CDRs are larger than 95% for most demographic situations. Even very recent (e.g., $T = 10$ generations, $D = 0.025$), relatively ancient (e.g., $T = 2,000$ generations, $D = 5.0$) or relatively weak

contractions (e.g., population size ratio of 10) are detected in more than 50% of the data sets. Third, our results suggest that using only ten microsatellite markers allows detecting past contraction with a high power, but more markers are required for precise inferences of scaled population sizes and timing under a wide range of demographic situations. However, precision of the inference of the different parameters strongly depends on the scenario considered (see also Girod et al. 2011). This is not surprising because the ability of the method to infer past demography depends on the genetic information available in the data and this information varies as a function of the timing of the past contraction. This can easily be understood and predicted from the timing of events in the ancestry of a sample. Recent contractions result in more precise inference for the ancestral population size than for the actual population size, because much of the coalescent and mutation events occur in the ancestral population. The opposite is true for old contractions. Precise inference of current and ancestral population sizes is thus only expected for past contractions that occurred neither too recently nor too far in the past because in such scenarios coalescent and mutation events are more homogeneously distributed over all demographic phases. Finally and for the same reason, inference of contraction time is expected to be more precise for intermediate timings. This is exactly what is observed in figures 4 and 5. One important result of our study is that high CDRs as well as good inference precision for both the time and the actual population size parameters are still expected for relatively ancient contraction (e.g., $1.25 \leqslant D \leqslant 5.0$).

Finally, many recent software packages that make demographic inferences from genetic data, such as MIGRATE (Beerli and Felsenstein 2001), IM (Hey and Nielsen 2004, 2007; Hey 2010), or LAMARC (Kuhner 2006), do not use the classical coalescent parameter scaling by population size (i.e., $4Nm$ and $T/2N$) but rather use scaling by the mutation rate (i.e., m/μ and $T\mu$), or propose both options, as MIGRAINE does. Our simulations show that there is not much interest to scale time by mutation rate for inferences of past contractions. For the demographic scenarios considered here, such scaling always reduces inference precision for the time parameter. Beside those scaling issues, independent information about mutation rates of the markers can be incorporated as prior information in the analyses to allow inference of canonical parameters (i.e., N, T, and N_{anc}) instead of scaled ones (e.g., as done in MSVAR). This is an attractive possibility for practical inferences, however, it has been shown in Girod et al. (2011) and Faurby and Pertoldi (2012) that such parametrization allows precise inference of canonical parameters only if precise prior information on mutation rate is used. This is so because single-locus population genetics models in general (and Kingman's coalescent model in particular) depend upon scaled, not canonical, parameters.

Comparison with Previous Methods

In the past decade, the use of likelihood-based methods to analyze genetic data under a single population model with past variation in population size emerged with the release of the MSVAR software (Beaumont 1999; Storz and Beaumont 2002). As expected, this coalescent-based MCMC method has been shown to be much more powerful than using summary statistics in detecting past contractions or expansions (Girod et al. 2011; Peery et al. 2012). Moreover, model-based approaches can also infer model parameters, such as current and past population sizes, and the timing of the demographic change. In this study, we compared performances in terms of CDRs, parameter inference precision, and computation times of MSVAR and our IS method. Our simulations globally show similar behavior of the two methods, with slight but clear advantages for MIGRAINE in terms of power of contraction detection, parameter estimation, and computation times. For example, both methods perform well for intermediate contraction strength and timing. On the contrary, they both are inefficient when population size contraction is too strong and too recent. The MCMC algorithm of MSVAR shows strong convergence issues for very recent and strong contractions (see fig. 1 in Girod et al. 2011) and gives biased point estimates as well as bad CI for θ_{anc} (supplementary fig. S3, Supplementary Material online). For the same demographic scenarios, IS algorithms implemented in MIGRAINE are not efficient and, even with large computation times, MIGRAINE shows high relative biases and RRMSE, as well as bad coverage properties of CI for θ as discussed above. For both methods, computation times thus greatly increase with the strength of the contraction, and accurate parameter inference considering very recent and strong contractions may be difficult to achieve. However, MIGRAINE appears 1) more adapted to the analyses of microsatellite markers because of the implementation of the GSM, as detailed in the next section; 2) slightly more powerful than MSVAR as our simulations show higher CDRs (e.g., often more than 20% higher CDRs depending on the demographic situation considered) and fewer false expansion detections (i.e., 0 vs. 1 false expansion detected among all simulations with MIGRAINE and MSVAR, respectively); and 3) faster than MSVAR as computation times were always higher for MSVAR than for MIGRAINE for equivalent demographic scenarios (e.g., two to ten times faster). Finally, a certain advantage of MIGRAINE over MSVAR is that it can easily use parallel computation, thereby decreasing computation times by the number of available cores.

Robustness to Mutational Processes

Although many models have been developed to describe microsatellite mutation processes (Bhargava and Fuentes 2010), most programs that analyze microsatellite data use the SMM (e.g., IM, MIGRATE, LAMARK, see references above, but see DIYABC, Cornuet et al. 2008, and BEAST, Drummond et al. 2012). However, it has been recognized that violations of the SMM assumptions might induce severe bias in the inference of demographic history (Gonser et al. 2000). Indeed, mutations of more than one step of the GSM can produce gaps in allele length distribution, which are typically often observed after a population decline under an SMM (Garza and Williamson 2001). Peery et al. (2012)

recently showed that identification of past contractions using the summary statistic-based BOTTLENECK (Cornuet and Luikart 1996) and M-RATIO softwares (Garza and Williamson 2001) is highly biased by deviations from the mutation models implemented in those softwares, often leading to significant contraction detection in samples simulated under a stable population model. Faurby and Pertoldi (2012) also showed that estimation of current and past population sizes with MSVAR is unreliable when realistic deviations from the SMM occur. In the previous study of Girod et al. (2011), it was also shown that MSVAR was moderately robust to deviations from the SMM, which leads to false contraction detections in samples simulated from stable populations. However, this conclusion was presumably overoptimistic due to the small number of data sets analyzed. In this study, we clearly show a strong impact of violations of the SMM assumptions: Even small deviations from the SMM induce large FCDRs in samples simulated under a stable demography. We adapted our algorithm by implementing a GSM in MIGRAINE to allow inference of past population size variations under this more complex and more realistic mutational model. Our simulations first show that, in samples from stable populations, using a GSM successfully decreases the rate of false contraction detections due to mutations of more than one step. Second, for samples that effectively experienced a past contraction, our simulations show that using a GSM leads to CDRs similar to the one observed under the SMM, and also show that parameter inference precision is only slightly affected by the additional parameter p. Computation times are of course higher than for the SMM, but are still reasonable, as all data sets with 100 genes genotyped at 50 loci or less can be analyzed in a few days on a desktop computer with four cores. It is clear however that the GSM is not a perfect description of microsatellite mutation processes (Bhargava and Fuentes 2010). Many other factors such as single nucleotide insertions/deletions, asymmetric mutations, variation of the mutation rate with the length of the alleles and/or constraints on allele sizes may often occur (Sun et al. 2012), and potential additional biases in the inference of past population size changes due to these factors remain to be tested. However, the main cause of the confounding effects between mutation processes and past changes in population sizes is likely to be the presence of gaps in the sample allelic distributions, and factors others than multistep mutations should have less effects than those described above. Finally, MIGRAINE considers that mutation rates are constant across loci and we showed that unaccounted variation in mutation processes across markers 1) increases estimation biases essentially when a low number of loci are considered (i.e., 10 vs. 50 loci); but 2) also slightly deteriorates CI coverage properties, principally for θ and θ_{anc} and regardless of the number of loci.

Robustness to Population Structure

In addition to the strong effect of mutational processes, we also found that inferences of past population size changes can be drastically affected by population structure. First, at small spatial scales, IBD often occurs within populations due to spatially limited dispersal (see Guillot et al. 2009 for a review). Our simulations show that ignoring such local population structure induces large FCDRs when individuals are sampled at a small spatial scale from stable populations, even for relatively weak IBD. However, sampling individuals at a larger scale, that is, over the whole population area, efficiently reduces FCDRs. Parameter estimation, and to a lesser extent CDRs, obtained from samples coming from a population that effectively went through a contraction is also affected by IBD, and again sampling individuals at a large geographical scale efficiently reduces the impact of IBD. Parameter inference thus appears robust unless IBD is very strong and sampling scale is small.

Second, at larger spatial scales, population structure also arises due to limited gene flow within a set of discrete demes as described by the island model (Wright 1951). Such island population structure has stronger and more complex effects than IBD within populations. Our simulations show that samples coming from a single deme of stable island-structured populations show large FCDRs unless gene flow is extremely limited. Considering larger sampling scales by sampling individuals from all the demes of the total structured population reduces FCDRs but only for situations with important levels of gene flow (i.e., $M \geqslant 10.0$). Contrarily to IBD situations, enlarging sampling scale when gene flow is more limited, that is, $M \leqslant 1.0$, often increases FCDRs. Our simulations finally show that, when a contraction did occur in the past, ignoring island population structure also often strongly decreases CDRs and greatly biases parameter estimation. Moreover, both contraction detection and parameter estimation are sensitive to sampling scale. Unless gene flow is very high between demes ($M \geqslant 100.0$), small scale samples show low CDRs below 10% and accurate CDRs are only obtained using large sampling scales. Parameter inference appears highly biased for all levels of gene flow considered in this study. As for CDRs, best precision is also obtained when gene flow is high and sampling scale is large. Nevertheless, for all other situations, relative biases and RRMSEs are high suggesting that in most situations, limited gene flow between geographically distinct demes will always lead to erroneous inferences of current and past population sizes, and of the timing of the demographic change. Moreover, our results show complex interactions between levels of population structure, total genetic diversity, mutation processes, and sampling scales that strongly limit practical recommendations for the detection and characterization of past changes in population size in the presence of unaccounted population structure.

Such confounding effects of population structure and past changes in population sizes have already been observed. First, the effect of small-scale IBD population structure on CDRs obtained with the BOTTLENECK and M-RATIO softwares has been tested by simulations in Leblois et al. (2006). Our results are globally in agreement with this previous study, except that they found large FEDRs when using BOTTLENECK on IBD samples and that considering large scale samples makes FEDRs even larger. Such results showing that fine scale population structure induces false expansion signals has also been

previously stressed by Ptak and Przeworski (2002) in the context of sequence data analysis based on the Tajima's *D* statistics. Our simulations on the contrary show nonnull but small FEDR in the presence of small scale IBD structure.

Second, the effect of island population structure on past population size inference was first highlighted by simulation in Nielsen and Beaumont (2009). More recently, Peter et al. (2010), Chikhi et al. (2010), and Heller et al. (2013) also showed that analyzing samples drawn from a single deme of an island model with low to intermediate migration rates (i.e., $Nm < 5$) leads to false signals of contraction. Such erroneous imputations can be understood by considering the genealogical processes in an island model and in a single population with varying size. In a subdivided population with relatively small deme sizes and small migration rates, the genealogy of a sample taken from a single deme will show 1) many short branches for genes that rapidly coalesce within the deme in which they were sampled (i.e., before any migration event), this corresponds to the "scattering phase" described in Wakeley (1999); and 2) a few much longer branches for genes that coalesce after any emigration or immigration event from the deme sampled, this is the "collecting phase" of Wakeley (1999). The result is a genealogy with an excess of short terminal branches, as expected after a recent contraction in population size. However, if only one individual is taken from different demes, and/or if deme size or migration rates are large, the genealogical process becomes closer to the one expected under a Wright–Fisher (WF) population. Similarly, when gene flow is very limited, the ancestry of a sample coming from a single deme will also be very similar to the one expected under the WF model. Thus, except for limit cases, structured and declining population scenarios may result in more or less similar genealogies, depending on deme sizes, migration rates, and sampling scale. This expected influence of these three factors may strongly complicate the study of the effect of population structure on the inference of past population size. This can be noticed in the heterogeneity of the results of the different simulation studies available. All those comparisons based on different simulations of structured populations show that the effect of population structure is generally complex and will be quite difficult to predict except in a few simple cases. Those results also show that verbal argumentation based on oversimplified past genealogical processes may not always give the right prediction. Nevertheless, two main points arise from those simulation studies and can serve as guidelines for empirical studies: 1) Using a large sample scale strongly limits the influence of population structure on the inference of past population size variations, as advocated by Chikhi et al. (2010), but allows correct inference of past demographic changes only when migration rates are relatively high, that is, $M \geqslant 10.0$. Sampling a single gene per deme efficiently prevents false contraction detections in stable-structured populations but does not allow precise characterization of past contractions in the presence of intermediate to strong population structure; 2) for all other demographic situations, detection of past population size changes and parameter inferences based on panmictic models may often be misleading.

Such results finally imply that models themselves should be improved. First, model choice procedures should be developed to evaluate whether observed patterns of genetic diversity can be better explained by a model of population size change or by a model of subdivided populations. For example, Peter et al. (2010) used an Approximate Bayesian Computation model choice approach to distinguish between structured populations and panmictic population that undergone past changes in size. However, they show by simulation that their model choice procedure has relatively limited power to assign simulated data sets to the correct evolutionary model, even with a relatively large number of loci (e.g., 60–85.5% with 10–200 loci, respectively). An alternative is to develop models accounting for both population structure and population size changes that would probably be more realistic for most species/populations but the only available method (Hey and Nielsen 2007; Hey 2010) has never been tested for scenarios with both structured populations and past changes in population sizes.

Analysis of the Orangutan Data Set

Our analyses of the orangutan data set show that 1) all sampled sites, except S9, exhibit a clear signal of a strong and recent population size contraction; and 2) parameter inferences are extremely consistent among sites, and among the different pooled samples. Those results are in good agreement with equivalent analyses using `MSVAR` published in Goossens et al. (2006) and Sharma et al. (2012): All analyses indicate 1) ancestral population sizes of about a few thousand individuals (i.e., [2,600–5,900] for `MIGRAINE` and [3,100–13,400] for `MSVAR`), 2) much smaller current population sizes (i.e., [1–335] individuals for `MIGRAINE` and [22–1,400] with median values between 60 and 200 individuals for `MSVAR`), and 3) a relatively recent timing of the contraction (i.e., less than about 15,000 years ago for both `MIGRAINE` and `MSVAR`). However, considering the same generation time of 25 years, `MSVAR` results show less support for a very recent event (i.e., <200 years) than `MIGRAINE` ones. For this reason, Sharma et al. (2012) conclude that `MSVAR` analyses suggest that the main historical factor explaining the inferred contraction is habitat destruction due to the arrival of farmers 4–5 ka, whereas we cannot exclude from our `MIGRAINE` analyses that habitat loss through more recent deforestation for the development of massive agriculture and logging in the last 150–200 years is also a likely cause of the decline of orangutan populations in Borneo.

As discussed in the above sections, the two major problems with such inference of past changes in population size are misspecification of the mutation processes and unaccounted population structure. First, `MIGRAINE` analyses used a combination of GSM and SMM models for di- and tetranucleotide microsatellite loci, respectively, because mutations at tetranucleotide loci have been shown to be principally single steps whereas dinucleotide markers show more multistep mutations (Sun et al. 2012). Note that this adequation between marker types (di- vs. tetranucleotides) and mutation models (GSM vs. SMM) was empirically validated in

preliminary simulations using estimations of the parameter p of the GSM for the different markers. As MSVAR only considers a strict SMM, MIGRAINE analyses are thus expected to give more accurate results. However, only three markers among the 14 used are dinucleotide loci, which likely explains the weak effect of mutation process misspecification on MSVAR analyses and thus the observed good agreement between MIGRAINE and MSVAR results.

Second, despite the fact that Goossens et al. (2005) showed a weak but significant population structure among sites, especially across the Kinabatangan River (e.g., between RS1 and RS2 pooled samples, see the Materials and Methods section), the strong similarity observed among all our results for site-specific and pooled samples suggests that this factor does not have a strong effect on the estimations. Sharma et al. (2012) similarly conclude of no major effect of population structure on their inferences.

Conclusion

This work shows that our new inference method seems very competitive compared with alternative methods, such as MSVAR. However, our simulation tests also showed some limits, which most importantly are large computation times for strong disequilibrium scenarios and a strong influence of some form of unaccounted population structure. One first major improvement would thus be to speed up the analyses. Among the different possibilities, a relatively simple improvement would be to more efficiently choose the number of explored histories for each point of the parameter space. A more attractive improvement would be to design more efficient IS algorithms for time-inhomogeneous models. However, various unsuccessful attempts suggest that it may be a difficult task (not shown). A second major improvement would be to include population structure in the demographic model for simultaneous inference of migration rates and past population size change or to develop model choice procedures.

Finally, given the current revolution in genetic data production due to next generation sequencing technologies (NGS), it seems crucial to allow for the analysis of different types of independent markers, such as small DNA sequences without intralocus recombination, or single nucleotide polymorphisms (SNPs). The current version of MIGRAINE can consider three mutation models for allelic data (KAM, SMM, and GSM). It does not allow SNP data analysis, except under a KAM model with $K = 2$ allelic states. However, such a mutation model may not be adapted for SNP data because of the possibility of recurrent and backward mutations. For this reason, we did not test such analyses but a mutation model for SNPs is currently being implemented in MIGRAINE.

Given the relatively large computation times of our method, all analyses will clearly only be tractable for a limited number of markers (e.g., <10,000), but could nevertheless give very precise inferences. However, considering only independent markers is probably not the optimal approach as NGS make it possible to apply new class of methods based on the analyses of linkage disequilibrium for past demographic inferences. Such methods are based on the computation of the distribution of nonrecombining haplotype block length (e.g., Meuwissen and Goddard 2007; Albrechtsen et al. 2009; Gusev et al. 2012; Palamara et al. 2012; Theunert et al. 2012) or explicitly model the spatial dependence of markers using hidden Markov models (e.g., Dutheil et al. 2009; Mailund et al. 2012). They will probably play a major role in the future of population genetic demographic and historical inferences.

Materials and Methods

Simulation Study

A first set of simulations aims at testing the power of the algorithm to detect contractions and the accuracy of the parameters estimates when the duration (D) or the strength of the contraction (θ_{anc}) varies. The mutation process considered is an SMM over a range of 200 alleles. These experiments are presented in table 9. We also reanalyzed the 60 simulated data sets from Girod et al. (2011) to compare the results obtained with MSVAR and our own estimates. The latter simulated data sets are described in supplementary table S2, Supplementary Material online, and the comparison results are presented in section B in the supplementary material, Supplementary Material online.

A second set of simulations concerns robustness and accuracy related to mutation processes of microsatellites that are known to be highly complex (Ellegren 2000, 2004; Sun et al. 2012). This second set of simulations is thus based on a GSM with either $p = 0.22$ or $p = 0.74$, that are, respectively, the value commonly considered as a realistic average value in the literature (Dib et al. 1996; Ellegren 2000, 2004; Estoup et al. 2001), and the largest, ever reported value (Fitzsimmons 1998; Peery et al. 2012). We have also added data sets drawn with the KAM to those simulations, which might be seen as a GSM with $p = 1.0$. A first set of analyses tests the robustness to

Table 9. Simulated Demographic Scenarios with an SMM.

Case	D (T)	θ (N)	θ_{anc} (N_{anc})
[0]	1.25 (200)	0.4 (200)	40.0 (20,000)
[1]	0.025 (10)	0.4 (200)	40.0 (20,000)
[2]	0.0625 (25)	0.4 (200)	40.0 (20,000)
[3]	0.125 (50)	0.4 (200)	40.0 (20,000)
[4]	0.25 (100)	0.4 (200)	40.0 (20,000)
[5]	0.5 (200)	0.4 (200)	40.0 (20,000)
[6]	2.5 (1,000)	0.4 (200)	40.0 (20,000)
[7]	3.5 (1,400)	0.4 (200)	40.0 (20,000)
[8]	5 (2,000)	0.4 (200)	40.0 (20,000)
[9]	7.5 (3,000)	0.4 (200)	40.0 (20,000)
[10]	1.25 (200)	0.4 (200)	2.0 (1,000)
[11]	1.25 (200)	0.4 (200)	4.0 (2,000)
[12]	1.25 (200)	0.4 (200)	8.0 (4,000)
[13]	1.25 (200)	0.4 (200)	12.0 (6,000)
[14]	1.25 (200)	0.4 (200)	24.0 (16,000)
[15]	1.25 (200)	0.4 (200)	120.0 (60,000)
[16]	1.25 (200)	0.4 (200)	400.0 (200,000)

misspecification of the mutation process. Indeed, we have simulated under a GSM but inferred under an SMM. A second set of analyses tests the accuracy of the estimates when the inference algorithm is based on a GSM with unknown value of p. Because mutation processes generally differ between loci, we also considered a situation with a combination of loci with different mutation models and different mutation rates: 80% of the loci are simulated under a GSM with $p = 0.22$ and 20% under an SMM, and each locus has a mutation rate drawn from a Gamma distribution with parameters 2 and 0.0005, which gives a mean mutation rate of 0.001. For this situation with variable mutation processes, inference is also done using a GSM with unknown value of p.

The aim of the third set of simulations is to test robustness against a population structure that is ignored by the inference algorithm. All the data sets in this last series, described in table 3, were simulated under a GSM with $p = 0.22$. A first group of data sets simulates local within-population structure according to an IBD model. It thus aims at testing the robustness of the inferences to the assumption of panmixia by considering nonrandom mating due to spatially localized parent–offspring dispersal. The second group of data sets simulates both within- and among-population structure at a larger spatial scale according to an island model. Under the island model of population structure, few additional simulations were run to test 1) the influence of the mutation model and 2) the effect of sampling a single gene per deme. Those simulations are presented in section C in the supplementary material, Supplementary Material online, and discussed in the main text.

For each scenario, we simulated 200 multilocus data sets. Each simulated data set is a sample of $n_g = 100$ genes (or haploid individuals), genotyped at $n_l = 10$ unlinked microsatellite loci, except for a few situations where we indicate that 25 or 50 instead of ten loci are used. The mutation rate per gene per generation, say μ, is assumed to be constant for all loci, equal to 10^{-3}. All simulated samples, except data sets from Girod et al. (2011) (see section B in the supplementary material, Supplementary Material online), have been produced with a new version of the IBDSIM software (Leblois et al. 2009) that considers continuous changes of population sizes.

Validation

In all simulation experiments, the true (simulated) values of the parameters of interest are compared with the estimated values. The estimation bias and error, assessed by the relative mean bias and RRMSE, are reported, as well as the proportion of data sets for which a contraction or a false expansion signal is significantly detected (CDR and FEDR, respectively). Furthermore, the accuracy of the inference methodology is assessed by mean of profile LRTs (Cox and Hinkley 1974; Severini 2000). The coverage properties of the CI computed from the smoothed likelihood surface are tested through the ECDF of LRT P values, which should be asymptotically uniform. The departure from uniformity is tested by Kolmogorov–Smirnov tests, notably to check the validity of the implementation of the inference method and to assess the different factors that can affect likelihood surface inference.

Orangutan Data Set Analysis

Finally, we analyzed an orangutan (*Pongo pygmaeus*) data set from Goossens et al. (2006), sampled in 2001 in the Lower Kinabatangan food plain in Eastern Sabah, Malaysia from feces and hairs, and genotyped at 14 microsatellite loci. Our method was first applied on each subsamples S1–S6, S8, and S9 as described in Goossens et al. (2005). Those samples correspond to different more or less isolated sampling sites on each side of the Kinabatangan River (see fig. 1 in Goossens et al. 2005). Subsample S7 was not analyzed because it only contains seven individuals, whereas all others contain between 16 and 33 individuals. Goossens et al. (2005) showed a weak differentiation level within each side of the river ($0.01 \leqslant F_{ST} \leqslant 0.04$, with a mean of 0.02) but a stronger one between both sides of the river ($0.03 \leqslant F_{ST} \leqslant 0.09$ and a mean of 0.06). For this reason, and because our simulation study shows an important effect of population structure on the detection and characterization of past population size contractions, we also analyzed three larger data sets by pooling some subsamples together: 1) RS1 is the pool of subsamples coming from the south side of the river (i.e., S1+S3+S6+S8); 2) RS2 the pool of subsamples from the north of the river (S2+S4+S5+S9), and 3) RS1+RS2 is the total sample from both sides of the river.

To compare our results with those from Goossens et al. (2006) and Sharma et al. (2012) that used MSVAR on the same data sets, we need to convert our estimates of scaled parameters θ, D, and θ_{anc} into canonical parameters (i.e., N, T, and N_{anc}), because only values of canonical parameters are reported in these two publications. We did this conversion by considering a fixed mutation rate of 5×10^{-4} mutation per locus per generation, a value commonly considered as a realistic average value in the literature (Dib et al. 1996; Ellegren 2000; Sun et al. 2012). Furthermore, to express the timing of the contraction in years, we used a generation time of 25 years as in Sharma et al. (2012), a value in better agreement with long-term field studies than 8 years, the value considered in Goossens et al. (2006).

All sample sizes, results for all subsample analyses as well as details about the parametrization used for the inferences are given in section D in the supplementary material, Supplementary Material online. Results for the pooled data sets RS1, RS2, and RS1+RS2 are presented in the main text, in the Results section and illustrated in supplementary figure S4, Supplementary Material online.

The MIGRAINE software, with the implementation of the above described methods, can be downloaded from the web page http://kimura.univ-montp2.fr/~rousset/Migraine.htm (last accessed July 28, 2014).

Supplementary Material

Supplementary material, figures S1–S65, and tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to L. Chikhi, J.-M. Cornuet, A. Estoup, J.-M. Marin, and three anonymous reviewers for their constructive discussions about this work. This study was supported by the Agence Nationale de la Recherche (EMILE 09-blan-0145-01 and IM-Model@CORAL.FISH 2010-BLAN-1726-01 projects) and by the Institut National de Recherche en Agronomie (Project INRA Starting Group "IGGiPop"). Part of this work was carried out by using the resources of the Computational Biology Service Unit from the MNHN (CNRS Unité Mixte de Service 2700), the INRA MIGALE and GENOTOUL bioinformatics platforms, and the computing grids of ISEM and CBGP labs.

References

Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol.* 33:266–274.

Beaumont M. 1999. Detecting population expansion and decline using microsatellites. *Genetics* 153:2013–2029.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A.* 98:4563–4568.

Bhargava A, Fuentes F. 2010. Mutational dynamics of microsatellites. *Mol Biotechnol.* 44:250–266.

Bonebrake T, Christensen J, Boggs C, Ehrlich P. 2010. Population decline assessment, historical baselines, and conservation. *Conserv Lett.* 3: 371–378.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. 2010. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* 186:983–995.

Colautti RI, Manca M, Viljanen M, Ketelaars HAM, Bürgi H, Macisaac HJ, Heath DD. 2005. Invasion genetics of the Eurasian spiny waterflea: evidence for bottlenecks and gene flow using microsatellites. *Mol Ecol.* 14:1869–1879.

Comps B, Gömöry D, Letouzey J, Thiébaut B, Petit RJ. 2001. Diverging trends between heterozygosity and allelic richness during postglacial colonization in the European beech. *Genetics* 157:389–397.

Cornuet JM, Beaumont MA. 2007. A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theor Popul Biol.* 71: 12–19.

Cornuet JM, Luikart G. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144:2001–2014.

Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, Guillemaud T, Estoup A. 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* 24:2713–2719.

Cox DR, Hinkley DV. 1974. Theoretical statistics. London: Chapman & Hall.

Cressie NAC. 1993. Statistics for spatial data. New York: Wiley.

de Iorio M, Griffiths RC. 2004a. Importance sampling on coalescent histories. *Adv Appl Probab.* 36:417–433.

de Iorio M, Griffiths RC. 2004b. Importance sampling on coalescent histories. II. Subdivided population models. *Adv Appl Probab.* 36: 434–454.

de Iorio M, Griffiths RC, Leblois R, Rousset F. 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor Popul Biol.* 68:41–53.

Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380: 152–154.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29: 1969–1973.

Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183:259–274.

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* 16:551–558.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 5:435–445.

Emerson B, Paradis E, Thébaud C. 2001. Revealing the demographic histories of species using DNA sequences. *Trends Ecol Evol.* 16: 707–716.

Estoup A, Wilson IJ, Sullivan C, Cornuet JM, Moritz C. 2001. Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics* 159:1671–1687.

Faurby S, Pertoldi C. 2012. The consequences of the unlikely but critical assumption of stepwise mutation in the population genetic software, MSVAR. *Evol Ecol Res.* 14:859–879.

Felsenstein J. 1992. Estimating effective population size from sample sequences—inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet Res.* 59:139–147.

Fitzsimmons NN. 1998. Single paternity of clutches and sperm storage in the promiscuous green turtle (*Chelonia mydas*). *Mol Ecol.* 7:575–584.

Frankham R, Lees K, Montgomery M, England P, Lowe E, Briscoe D. 2006. Do population size bottlenecks reduce evolutionary potential? *Anim Conserv* 2:255–260.

Garza JC, Williamson EG. 2001. Detection of reduction in population size using data from microsatellite loci. *Mol Ecol.* 10:305–318.

Girod C, Vitalis R, Leblois R, Fréville H. 2011. Inferring population decline and expansion from microsatellite data: a simulation-based evaluation of the Msvar method. *Genetics* 188:165–179.

Gonser R, Donnelly P, Nicholson G, Di Rienzo A. 2000. Microsatellite mutations and inferences about human demography. *Genetics* 154: 1793–1807.

Goossens B, Chikhi L, Ancrenaz M, Lackman-Ancrenaz I, Andau P, Bruford MW. 2006. Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biol.* 4:e25.

Goossens B, Chikhi L, Jalil MF, Ancrenaz M, Lackman-Ancrenaz I, Mohamed M, Andau P, Bruford MW. 2005. Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (*Pongo pygmaeus*) populations from Sabah, Malaysia. *Mol Ecol.* 14:441–456.

Griffiths RC, Tavaré S. 1994a. Ancestral inference in population genetics. *Stat Sci.* 9:307–319.

Griffiths RC, Tavaré S. 1994b. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344: 403–410.

Guillot G, Leblois R, Coulon A, Frantz AC. 2009. Statistical methods in spatial genetics. *Mol Ecol.* 18:4734–4756.

Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, Gregersen P, Pe'er I. 2012. The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol.* 29:473–486.

Heller R, Chikhi L, Siegismund HR. 2013. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS One* 8:e62992.

Hey J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol.* 27:905–920.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A.* 104:2785–2790.

Keller LF, Waller DM. 2002. Inbreeding effects in wild populations. *Trends Ecol Evol.* 17:230–241.

Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22:768–770.

Lande R. 1988. Genetics and demography in biological conservation. *Science* 241:1455–1460.
Lawton-Rauh A. 2008. Demographic processes shaping genetic variation. *Curr Opin Plant Biol.* 11:103–109.
Leblois R, Estoup A, Rousset F. 2009. IBDSim: a computer program to simulate genotypic data under isolation by distance. *Mol Ecol Resour.* 9:107–109.
Leblois R, Estoup A, Streiff R. 2006. Habitat contraction and reduction in population size: does isolation by distance matter? *Mol Ecol.* 15: 3601–3615.
Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, Lunter G, Prüfer K, Scally A, Hobolth A, et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related Great Ape species. *PLoS Genet.* 8:e1003125.
Marjoram P, Tavaré S. 2006. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet.* 7: 759–770.
Meuwissen TH, Goddard ME. 2007. Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. *Genetics* 176:2551–2560.
Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol Ecol.* 18:1034–1047.
Ohta T, Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res.* 22:201–204.
Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet.* 91:809–822.
Peery MZ, Kirby R, Reid BN, Stoelting R, Doucet-Bër E, Robinson S, Vàsquez-Carrillo C, Pauli JN, Palsbøll PJ. 2012. Reliability of genetic bottleneck tests for detecting recent population declines. *Mol Ecol.* 21:3403–3418.
Peter B, Wegmann D, Excoffier L. 2010. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol Ecol.* 19:4648–4660.
Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosome microsatellites. *Mol Biol Evol.* 16:1791–1798.
Ptak SE, Przeworski M. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.* 18: 559–563.
Reusch TBH, Wood TE. 2007. Molecular ecology of global change. *Mol Ecol.* 16:3973–3992.
Rousset F, Leblois R. 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Mol Biol Evol.* 24: 2730–2745.
Rousset F, Leblois R. 2012. Likelihood-based inferences under a coalescent model of isolation by distance: two-dimensional habitats and confidence intervals. *Mol Biol Evol.* 29:957–973.
Schneider S, Excoffier L. 1999. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates very among sites: application to human mitochondrial DNA. *Genetics* 152:1079–1089.
Schwartz M, Luikart G, Waples R. 2007. Genetic monitoring as a promising tool for conservation and management. *Trends Ecol Evol.* 22: 25–33.
Severini TA. 2000. Likelihood methods in statistics. Oxford: Oxford University Press.
Sharma R, Arora N, Goossens B, Nater A, Morf N, Salmona J, Bruford MW, Van Schaik CP, Krützen M, Chikhi L. 2012. Effective population size dynamics and the demographic collapse of Bornean orangutans. *PLoS One* 7:e49429.
Spencer CC, Neigel JE, Leberg PL. 2000. Experimental evaluation of the usefulness of microsatellite DNA for detecting demographic bottlenecks. *Mol Ecol.* 9:1517–1528.
Stephens M, Donnelly P. 2000. Inference in molecular population genetics (with discussion). *J R Stat Soc Series B Stat Methodol.* 62: 605–655.
Storz J, Beaumont M. 2002. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* 56: 154–166.
Sun J, Helgason A, Masson G, Ebenesersdottir S, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet.* 44: 1161–1165.
Theunert C, Tang K, Lachmann M, Hu S, Stoneking M. 2012. Inferring the history of population size change from genome-wide SNP data. *Mol Biol Evol.* 29:3653–3667.
Wakeley J. 1999. Nonequilibrium migration in human evolution. *Genetics* 153:1863–1871.
Williams B, Nichols J, Conroy M. 2002. Analysis and management of animal populations: modeling, estimation, and decision making. San Diego (CA): Academic Press.
Wright S. 1951. The genetical structure of populations. *Ann Eugen.* 15: 323–354.

Journal de la Société Française de Statistique
Vol. 159 No. 3 (2018)

Likelihood computation and inference of demographic and mutational parameters from population genetic data under coalescent approximations

Titre: Calcul de la vraisemblance et inférence des paramètres démographiques et mutationnels à partir de la variation génétique des populations

François Rousset[1,4], Champak Reddy Beeravolu[2] and Raphaël Leblois[3,4]

Abstract: Likelihood methods are being developed for inference of migration rates and past demographic changes from population genetic data. We survey an approach for such inference using sequential importance sampling techniques derived from coalescent and diffusion theory. The consistent application and assessment of this approach has required the re-implementation of methods often considered in the context of computer experiments methods, in particular of Kriging which is used as a smoothing technique to infer a likelihood surface from likelihoods estimated in various parameter points, as well as reconsideration of methods for sampling the parameter space appropriately for such inference. We illustrate the performance and application of the whole tool chain on simulated and actual data, and highlight desirable developments in terms of data types and biological scenarios.

Résumé : Diverses approches ont été développées pour l'inférence des taux de migration et des changements démographiques passés à partir de la variation génétique des populations. Nous décrivons une de ces approches utilisant des techniques d'échantillonnage pondéré séquentiel, fondées sur la modélisation par approches de coalescence et de diffusion de l'évolution de ces polymorphismes. L'application et l'évaluation systématique de cette approche ont requis la ré-implémentation de méthodes souvent considérées pour l'analyse de fonctions simulées, en particulier le krigeage, ici utilisé pour inférer une surface de vraisemblance à partir de vraisemblances estimées en différents points de l'espace des paramètres, ainsi que des techniques d'échantillonage de ces points. Nous illustrons la performance et l'application de cette série de méthodes sur données simulées et réelles, et indiquons les améliorations souhaitables en termes de types de données et de scénarios biologiques.

Keywords: demographic history, coalescent processes, échantillonnage pondéré, polymorphisme génétique

Mots-clés : histoire démographique, processus de coalescence, importance sampling, genetic polymorphism

AMS 2000 subject classifications: 92D10, 62M05, 65C05

[1] Institut des Sciences de l'Evolution, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France.
E-mail: francois.rousset@umontpellier.fr
[2] Biology Department, City College of New York, New York, NY 10031, USA.
E-mail: champak.br@gmail.com
[3] CBGP UMR 1062, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier, Montpellier, France.
E-mail: Raphael.Leblois@supagro.inra.fr
[4] Institut de Biologie Computationnelle, Université de Montpellier

 ISSN: 2102-6238

1. Introduction

Since the advent of genetic markers, there have been many efforts to infer demographic parameters (e.g., population sizes and dispersal) from observed genetic variation. These efforts serve to better understand the forces affecting the evolution of natural population, and also appear to fulfill a distinct fascination for the history of past human migrations and population admixtures. Early statistical approaches have considered descriptions of genetic variation that can be understood as analyses of variance in allele frequencies among different groups of individuals (Cockerham, 1973). In particular, Wright's F-statistics (Wright, 1951) can be expressed as functions of frequencies of pairs of gene copies that are of identical allelic state, and then viewed as estimators of the corresponding functions of probabilities that pairs of gene copies are identical, under a given model. As there are theoretical expectations for these probabilities in simple models of evolution, a quantitative process-based interpretation of the descriptors is possible, to infer dispersal parameters among different subpopulations, or the demographic history of natural populations.

For the same objectives, likelihood analyses attempt to extract information from the joint allelic types of more than two genes copies. These attempts have been hampered by the increasing difficulty in computing the probability distribution of such joint configurations as the number of gene copies increases. For this reason, stochastic algorithms have been developed to estimate the likelihood of a sample of arbitrary size. These algorithms view a sample as incomplete data, where the missing information is the genealogy of all gene copies in the sample. If the "complete-data" likelihood, that is the probability of the sample given the genealogy, is easy to evaluate, the evaluation of the sample likelihood can be formulated as the evaluation of a marginal likelihood, obtained by integration of this complete-data likelihood over a probability distribution of genealogies consistent with the data. A classic recurrent Markov chain Monte Carlo approach has been used to sample from this distribution (Beerli and Felsenstein, 1999; Nielsen and Wakeley, 2001; Hey, 2010). However, the slow convergence of such methods has prompted both the development of alternative algorithms for computing the marginal likelihood, and also explains the persistence of the older methods and the development of other methodologies based on simulation of samples, such as Approximate Bayesian Computation (Beaumont, 2010).

In this paper we review an approach to perform likelihood-based inferences, using a class of importance sampling algorithms derived from the work of Griffiths and collaborators (de Iorio and Griffiths, 2004a,b; see also Stephens and Donnelly, 2000). We first explain the importance sampling algorithm defined in this work to obtain estimates of the likelihood of given parameter points. Next we discuss the additional steps required to derive reliable inferences from such likelihood estimates. A distinctive feature of the latter work, when compared to most of the literature on alternative methods of inference, is the emphasis on evaluating the inference in terms of coverage properties of likelihood-based confidence intervals. For such purposes, one has to infer a likelihood surface from estimated likelihoods in different parameter points. Kriging has classically been used for inference of response surfaces (e.g., Sacks et al., 1989), and our efforts to obtain good coverage has led us to reimplement such methods as part of a set of software tools to explore likelihood surface in an automatic way.

The methods described in this paper are all implemented in free software: the MIGRAINE software, written in C++, implements the algorithms for likelihood estimation in each given param-

 ISSN: 2102-6238

eter point, and calls R code that performs inference of the likelihood surface from the likelihood points, plots various representations of this surface and other diagnostics, evaluates likelihood ratio confidence intervals, and designs new parameter points whose likelihood should be computed in a next iteration of MIGRAINE. Most of this R code has been incorporated in a standard R package, `blackbox`, which can also be used on its own to perform optimization of simulated functions. MIGRAINE writes all the required calls to R functions so that no understanding of them is required from the user.

2. Lihelihood inference using importance sampling algorithms

2.1. Demographic scenarios

We consider several classical models in population genetics. Informally, the simplest model considers a single population of N (haploid) individuals, N being constant through time. In each generation the genes received by descendants are drawn, independently for each descendant, with equal probability from each possible parent (the so-called Wright-Fisher model). Therefore, the population size determines the probability $1/N$ that two descendants receive their genes from the same parent, and more generally characterizes the joint distribution of number of descendants of all parents for a sample of n descendants, a distribution which is a building block of the recursions we will consider later. Mutations (i.e. changes in allelic types) may occur, independently for each transmitted gene lineage in each generation. The more general demographic scenarios consider changes in population size through time, or dispersal of individuals among a set of subpopulations, or divergence of two populations from a single ancestral population. We aim to use the genotypes $\mathbf{S}$ of a sample of individuals to infer the parameters of the ancestral process, including current and ancestral population sizes, mutation rates, dispersal rates, and times of population divergence events.

In the following we consider a sample $\mathbf{S}$ of genotypes at a single locus. When analysing several loci, the information is considered independent at each locus (log likelihoods for each locus are summed). It is still a pending issue to develop likelihood methods that take into account the statistical non-independence of genetic variation at different loci, a dependence which is expected for loci located close to each other on a chromosome.

2.2. Inferring the likelihood for a parameter point by importance sampling

2.2.1. Sequential importance sampling formulation

Sequential importance sampling algorithms are importance sampling algorithms where the basic quantities (the proposal distribution and the weights) are built sequentially (Liu, 2004). They have for example been elaborated to perform likelihood-based inference in state-space models, defined in terms of an hidden Markov process, and of an emission process. The proposal and the weights refer to the states of the hidden process (e.g. Andrieu et al., 2010). Here there is no distinct emission process: the observations are viewed as the terminal value of an hidden sequential process starting from the common ancestor of the sampled gene copies, and defined as follows. Given a current sample $\mathbf{S}$, we consider the ancestral states (i.e. allelic types) of the

 ISSN: 2102-6238

gene lineages ancestral to $\mathbf{S}$ at any time t, called the "ancestral sample", $\mathbf{S}(t)$. These ancestral states are considered at any time until the time t_τ where a common ancestor of the sample is reached. We consider transition probabilities $\hat{p}$ for $\mathbf{S}(t_k)$ over successive time steps $t_0, t_1, \ldots, t_\tau$, and importance sampling weights $\hat{w}$ defined such that the likelihood of a sample can be written as

$$q(\mathbf{S}) = \mathrm{E}_{\hat{p}}\left(\prod_{k=0}^{\tau} \hat{w}[\mathbf{S}(t_k)]\right), \tag{1}$$

where the expectation is taken over the distribution of sequences $(\mathbf{S}(t_k))$ of ancestral samples generated by the transition probabilities $\hat{p}$. These transition probabilities define a Markov chain over ancestral states, with absorbing states being reached at time t_τ when a single common ancestor is reached. Each realization of this Markov chain records a sequence of coalescence, mutation and migration events until the common ancestor is reached. Estimation of $q(\mathbf{S})$ is then performed by averaging $\prod_{k=0}^{\tau} \hat{w}[\mathbf{S}(t_k)]$ over independent realizations of this Markov chain (2000 such independent ancestral histories in the following applications, unless mentioned otherwise).

de Iorio and Griffiths (2004a,b) propose $\hat{p}$ and $\hat{w}$ based on approximations for the ratio $\pi \equiv q(\mathbf{S})/q(\mathbf{S}')$ of the probabilities of samples differing by one event (mutation, migration, or coalescence event). We will detail how these approximations are constructed. For that purpose we will first consider recursions over a time interval, relating the current sample to an ancestral sample $\mathbf{S}(t)$ taken (say) a generation before.

These recursions are obtained by a coalescent argument. That is, we represent the events leading to the current sample of n genes as the realizations of two processes: a coalescent process determining the marginal distribution of ancestral genealogies of n genes, independent of the current allelic types; and given a genealogy, a mutation process that changes the allelic types along the branches of the genealogical tree. For developments of coalescent methods see Tavaré (1984), Hein et al. (2005), or Wakeley (2008).

In this perspective, the relationship between a current sample probability and the parental sample probability can be conceived as the joint realizations of two processes in addition to those leading to the parental sample: the marginal genealogical process over the latest generation, and the mutation process over this generation. In the following we consider samples from subdivided populations, where sample size is defined as a vector $\mathbf{n}$ of sample sizes in distinct subpopulations, and samples are characterized by the counts of different alleles in each sampled subpopulations. For example the sample $\mathbf{S} = ((0,4,5),(5,4,0))$ describes the counts of three alleles among $n = 18$ individuals sampled in two subpopulations ($n = (9,9)$), with the first allele only found in the second subpopulation, and so on. The recursion between a current sample $\mathbf{S}'$ and all possible parental samples $\mathbf{S}$ takes the form

$$q(\mathbf{S}') = \sum_{\mathbf{S}} \Pr(\mathbf{n}) q(\mathbf{S}) \Pr(\mathbf{S}'|\mathbf{S}), \tag{2}$$

where $q(\mathbf{S}') \equiv \Pr(\mathbf{S}'|\mathbf{n}')$ is the stationary probability that the descendant sample is $\mathbf{S}'$, given the descendant sample size $\mathbf{n}'$; $q(\mathbf{S}) \equiv \Pr(\mathbf{S}|\mathbf{n})$ is likewise the stationary probability of sample $\mathbf{S}$ given parental sample size $\mathbf{n}$; $\Pr(\mathbf{n}) \equiv \Pr(\mathbf{n}|\mathbf{n}')$ is the stationary probability that, given the descendant size $\mathbf{n}'$ (but not given $\mathbf{S}'$), the parental lineages form a sample of $\mathbf{n}$ genes. This probability depends on the stationary probability of coalescence and migration events in the latest generation, but

 ISSN: 2102-6238

the occurrence of mutations does not change $\mathbf{n}$; and $\Pr(\mathbf{S}'|\mathbf{S}) \equiv \Pr(\mathbf{S}'|\mathbf{S},\mathbf{n}')$ is the probability (given $\mathbf{n}'$) that mutation events led to the descendant sample $\mathbf{S}'$ given the parental sample $\mathbf{S}$ and the descendant $\mathbf{n}'$.

This recursion suggests the following inefficient importance sampling algorithm. We rewrite the recursion by discarding the case where $\mathbf{S}' = \mathbf{S}$ on the right-hand sum. The resulting equation can be written as

$$q(\mathbf{S}') = \sum_{\mathbf{S}\neq\mathbf{S}'} \tilde{w}(\mathbf{S}')\tilde{p}(\mathbf{S}|\mathbf{S}')q(\mathbf{S}), \tag{3}$$

where

$$\tilde{w}(\mathbf{S}') \equiv \frac{\sum_{\mathbf{S}\neq\mathbf{S}'} \Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})}{1-\sum_{\mathbf{S}\neq\mathbf{S}'} \Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})} \tag{4}$$

and

$$\tilde{p}(\mathbf{S}|\mathbf{S}') \equiv \frac{\Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})}{\sum_{\mathbf{S}\neq\mathbf{S}'} \Pr(\mathbf{n})\Pr(\mathbf{S}'|\mathbf{S})}. \tag{5}$$

The probabilities $\tilde{p}(\mathbf{S}|\mathbf{S}')$ define transition probabilities of a Markov chain such that

$$q(\mathbf{S}) = \mathrm{E}_{\tilde{p}}\left(q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1}\tilde{w}[\mathbf{S}(t_k)]\right) \tag{6}$$

where $\mathbf{S}(t_0) = \mathbf{S}$ represents the allelic counts in the current sample, and $\mathbf{S}(t_\tau)$ the allelic type of the most recent common ancestor of $\mathbf{S}(t_0)$. Thus, the $\tilde{w}$s (or their product) are importance sampling weights in a sequential importance sampling algorithm of which the proposal distribution is the distribution of ancestral histories generated by $\tilde{p}$.

A good pair (p,w) is such that $q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1} w[\mathbf{S}(t_k)]$ has low variance over realizations of p. The above pair is inefficient in this respect. An optimal IS algorithm can be defined as yielding a zero variance, and Stephens and Donnelly (2000) characterized the optimal pair (p,w) in terms of successive samples and their stationary probabilities. To derive a feasible algorithm from this characterization, de Iorio and Griffiths (2004a,b) reformulated it in terms of the probabilities $\pi(j|d,\mathbf{S})$, for any j and d, that an additional gene taken from subpopulation d is of type j. Then, approximations for the optimal (p,w) can be defined from approximations for the πs.

2.2.2. *Optimal p and w*

Rewrite

$$q(\mathbf{S}') = \sum_{\mathbf{S}\neq\mathbf{S}'} w(\mathbf{S}')p(\mathbf{S}|\mathbf{S}')q(\mathbf{S}) \tag{7}$$

as

$$q(\mathbf{S}') = \sum_{\mathbf{S}\neq\mathbf{S}'} \hat{w}(\mathbf{S}',\mathbf{S})\hat{p}(\mathbf{S}|\mathbf{S}')q(\mathbf{S}) \tag{8}$$

for some transition probabilities $\hat{p}(\mathbf{S}|\mathbf{S}')$ forming a Markov transition matrix, and for

$$\hat{w}(\mathbf{S}',\mathbf{S}) \equiv w(\mathbf{S}')\frac{p(\mathbf{S}|\mathbf{S}')}{\hat{p}(\mathbf{S}|\mathbf{S}')}. \tag{9}$$

 ISSN: 2102-6238

Then $q(\mathbf{S}) = \mathrm{E}_p\left(q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1} w[\mathbf{S}(t_k)]\right)$ becomes $q(\mathbf{S}) = \mathrm{E}_{\hat{p}}\left(q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1} \hat{w}[\mathbf{S}(t_k),\mathbf{S}(t_{k+1})]\right)$.

Consider the Markov chain defined by the transition probabilities

$$\hat{p}(\mathbf{S}|\mathbf{S}') \equiv w(\mathbf{S}')p(\mathbf{S}|\mathbf{S}')\frac{q(\mathbf{S})}{q(\mathbf{S}')} \tag{10}$$

for any pair $\mathbf{S}',\mathbf{S}$. Then $\hat{w}[\mathbf{S}(t_k),\mathbf{S}(t_{k+1})] = q[\mathbf{S}(t_{k+1})]/q[\mathbf{S}(t_k)]$ and any realization of this Markov chain over ancestral states gives the exact likelihood ("perfect simulation"):

$$q(\mathbf{S}(t_\tau))\prod_{k=0}^{\tau-1}\hat{w}[\mathbf{S}(t_k),\mathbf{S}(t_{k+1})] = q(\mathbf{S}(\tau))\prod_{k=0}^{\tau-1}\frac{q[\mathbf{S}(t_{k+1})]}{q[\mathbf{S}(t_k)]} = q(\mathbf{S}(t_0)), \tag{11}$$

which shows that $(\hat{p},\hat{w})$ is optimal.

2.2.3. *Formulation of efficient p and w*

We can rewrite the optimal importance sampling algorithm in terms of the probability $\pi(j|d,\mathbf{S})$ that an additional gene taken from deme d is of type j (such that the sum over all possible types $\sum_j \pi(j|d,\mathbf{S}) = 1$). We write the stationary probability $q(\mathbf{S})$ as an expectation over the joint distribution of frequencies X_{di} for all alleles i in all subpopulations d,

$$q(\mathbf{S}) = \mathrm{E}\left(\prod_d \binom{n_d}{(n_{di})}\prod_i X_{di}^{n_{di}}\right). \tag{12}$$

Then for any d and j, $\pi(j|d,\mathbf{S})$ is related to the stationary sample probabilities by

$$\pi(j|d,\mathbf{S})q(\mathbf{S}) = \mathrm{E}\left(X_{dj}\prod_d \binom{n_d}{(n_{di})}\prod_i X_{di}^{n_{di}}\right) = \frac{n_{dj}+1}{n_d+1}q(\mathbf{S}+\mathbf{e}_{dj}) \tag{13}$$

where the expectation is taken over the stationary density of joint allele frequencies $\mathbf{x}$ in the different demes considered. Thus if two successive samples differ by the addition of a gene copy of type j in deme d, the corresponding term $\hat{w}[\mathbf{S}(t_k),\mathbf{S}(t_{k+1})]$ in eq. 11 can be written as

$$\pi(j|d,\mathbf{S}(t_k))\frac{n_d(t_k)+1}{n_{dj}(t_k)+1} = \pi(j|d,\mathbf{S}(t_{k+1}))\frac{n_d(t_{k+1})}{n_{dj}(t_{k+1})}. \tag{14}$$

If two successive samples differ by a mutation from i to j in deme d, then

$$\hat{w}[\mathbf{S}(t_k),\mathbf{S}(t_{k+1})] = \frac{\pi(j|d,\mathbf{S}(t_{k+1}))}{\pi(i|d,\mathbf{S}(t_{k+1}))}\frac{n_{di}(t_{k+1})+1}{n_{dj}(t_{k+1})}, \tag{15}$$

as mutation can be represented as the removal of one gene copy and the addition of another gene copy of another type in the same deme. Likewise, a migration from deme d to deme d' yields

$$\hat{w}[\mathbf{S}(t_k),\mathbf{S}(t_{k+1})] = \frac{\pi(j|d',\mathbf{S}(t_{k+1}))}{\pi(j|d,\mathbf{S}(t_{k+1}))}\frac{n_{d'}(t_{k+1})n_{dj}(t_{k+1})+1}{(n_d(t_{k+1})+1)n_{d'j}(t_{k+1})}. \tag{16}$$

 ISSN: 2102-6238

Coalescent methods typically consider that only one event (coalescence, mutation, or migration) distinguishes the successive samples. Thus, in informal terms, the mutation and migration rates are assumed small, and subpopulation sizes are assumed large, so that it is unlikely that more than one coalescence event occurs in a generation (see the Appendix for a somewhat more formal statement). Then, the product of sequential weights in eq. 11 can be written, for any sequence of ancestral samples, as a product of terms given in the last three equations. Any approximation for the πs then defines an approximation for the optimal weights in an importance sampling algorithm.

The Appendix details the approximation defined by de Iorio and Griffiths (2004a,b). This approximation recovers the true πs and thus allows "perfect simulation" in a few cases where the stationary distribution of allele frequencies in populations is known, and it is otherwise very efficient for other time-homogeneous models that have been investigated (de Iorio et al., 2005; Rousset and Leblois, 2007, 2012). The previous arguments also yield importance sampling algorithms for time-inhomogeneous models where the rates of events depend on time-variations in parameter values, when random times are attached to the successive events in the ancestral history (Griffiths and Tavaré, 1994). The $\hat{\pi}$ approximation of de Iorio and Griffiths (2004a,b) has been used to extend the inference method to models with changing population size over time (Leblois et al., 2014) and models with population divergence events (divergence with migration between two populations, unpublished work). However, the $\hat{p}$ proposal at any step t only takes into account the rates at time t, not the more ancestral rate variations that also affect sample probabilities at time t, and this results in a loss of efficiency of the IS algorithm. Resampling methods (Liu, 2004) have been investigated to provide some relief to this inefficiency (Merle et al., 2017).

A large part of the computational burden stems from the computation, independently for each parameter point, of the $\hat{\pi}$ terms of de Iorio and Griffiths (2004b), which is required for the determination of the proposal distribution and of the importance sampling weights. Bridge sampling (e.g., Gelman and Meng, 1998) may be used to tentatively reduce the amount of such computation. To use bridge sampling in the present context, one first estimates likelihood as previously described for one or a few driving parameter values. Estimates of likelihood in any new given parameter value are then deduced from estimates of the likelihood ratio between driving and new values, using only the realized path of importance sampling in driving value(s), and the ratio, for driving and new parameter values, of the IS weights for such paths. This can bring computational gains if sampling from the proposal distribution is costly, but the ratio of IS weights is easy to evaluate. In early steps of this project (Leblois, 2004), we investigated the performance of bridge sampling in combination with the IS algorithm of Nath and Griffiths (1996), whose IS weights are indeed simple to evaluate, but whose proposal distribution is also much less efficient than that of de Iorio and Griffiths (2004b). Bridge sampling did not bring any improvement comparable to that brought by de Iorio and Griffiths's algorithm. In the context of de Iorio and Griffiths's algorithm, bridge sampling may be of little benefit, as in that context the ratio of IS weights depends on the $\hat{\pi}$ terms for any new parameter values (as implied by eqs. 14–16), and is thus costly to evaluate.

 ISSN: 2102-6238

2.2.4. *The PAC-likelihood heuristics*

Eq. 11 holds for any sequence $(\mathbf{S}(t_k))$, even if this sequence is not a biologically coherent sequence of ancestral states. Thus it holds for any sequence S_l defined as the sequential addition of all constituent gene copies g_l $(l = 1,\ldots,n)$ of the final sample $\mathbf{S}$, in any order. For such a sequence eq. 11 takes the form

$$q(\mathbf{S}) = \prod_{l=1}^{n} \pi(j(g_l)|d(g_l),\mathbf{S}_{l-1}) \frac{n_d(l-1)}{n_{dj}(l-1)} = \binom{n}{\mathbf{n}} \prod_{l=1}^{l=n} \pi(j(g_l)|d(g_l),\mathbf{S}_{l-1}). \tag{17}$$

where $j(g_l)$ and $d(g_l)$ represent respectively the allelic type of gene copy g_l and the subpopulation where it is added. Using some approximation for the πs in this expression yields a Product of Approximate Conditional (PAC) approximation to the likelihood (Li and Stephens, 2003). It is heuristic, in the sense that it is generally not a consistent estimator of the likelihood. However, we can use the same approximations to the πs as in the importance sampling algorithm (Cornuet and Beaumont, 2007), and in that case likelihood inference based on PAC-likelihood has proven practically equivalent to that based on likelihood (Rousset and Leblois, 2007, 2012; Leblois et al., 2014). The main drawback of the approximation is that, since there is no ancestral time attached to the successive $\mathbf{S}(l)$, this PAC-likelihood approximation cannot substitute the IS approach in models with time-varying rates. However, some models with time-varying rates include a ancestral stable demographic phase (e.g. the model with a contraction or an increase in population size used in Leblois et al. (2014) and illustrated in Fig. 1). Under such models, the PAC-likelihood can still be used to approximate the probability of the states of the ancestral genes lineages remaining when the stable demographic phase is reached backwards in time, and this approximation has allowed significant decreases in computation time without loss in precision (Leblois et al., 2014).

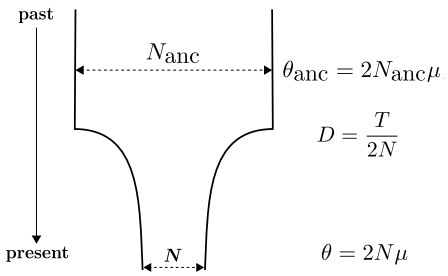


Figure 1: Representation of the time-inhomogeneous demographic model considered in Leblois et al. (2014).

N is the current population size, N_{anc} is the ancestral population size (before the demographic change), T is the time measured in generations since present and μ the mutation rate of the marker used. Those four parameters are the parameters of the finite population model. θ, D and θ_{anc} are the inferred scaled parameters of the coalescent approximations.

 ISSN: 2102-6238

2.3. Inferring the likelihood surface by smoothing

The above algorithms provide estimates of likelihood for given parameter values. A difficulty encountered in the first applications of this methodology (de Iorio et al., 2005) is that widely usable software (in particular, various R packages) were not up to the task of accurately inferring a likelihood surface from a collection of such estimates, and that the best of them still failed in a notable fraction of computations (essentially in the inversion of near-singular matrices), hampering our validation efforts. Our re-implementation of Kriging uses generalized cross-validation (Golub et al., 1979; Nychka, 2000) to obtain estimates of the smoothing parameters in reasonable time, in a way similar to the `fields` package in R (Nychka et al., 2015). We assume a Matérn correlation function, which is the most versatile model available. In particular, it can be used in Euclidean spaces of arbitrary dimension (Matérn, 1960). We estimate one scaling parameter for each parameter dimension of the coalescent model, as well as the smoothness parameter of the Matérn function. However, as the likelihood surfaces that we aim to infer are themselves smooth, a high estimate of the smoothness parameter should be obtained. Otherwise the software warns about potential problems in the input data.

We also use a complex strategy (discussed below) to sample points in parameter space in an automated way with minimal input from users. The details of the sampling strategy can substantially impact the performance, particularly as the number of parameters increases, but this impact cannot be fully assessed unless performance of the overall inference method (e.g., coverage of confidence intervals in the present case) is itself assessed.

To obtain a first estimate of the likelihood surface, one has to sample evenly in parameter space. In several dimensions, Latin square designs have been recommended (e.g., Welch et al., 1992). However, to estimate smoothing parameters, clusters of close parameters points are also useful (Zimmerman, 2006). Consistently with the latter work, our early attempts using Latin square designs were not convincing. The current implementation performs an empirical compromise between these distinct needs. From any estimate of the likelihood surface, further parameter points can then be sampled. The general resampling strategy, as detailed below, is to define a space of parameters with putatively high likelihood according to the current likelihood surface estimate, then to sample at random within this space, and to select among the sampled points those that are appropriate or best according to some additional criteria. MIGRAINE allows extrapolation beyond the parameter regions sampled in previous iterations, subject to ad hoc constraints in parameter space (such as positive mutation rates, but sometimes more complex constraints for composite parameters such as the so-called neighborhood size in models of localized dispersal).

Part of the new points are sampled uniformly in a parameter region with high predicted likelihood. But parameter regions that have yet been little sampled typically have high prediction variance, and may thus be worth sampling even if the predicted likelihood is relatively low in such regions. Expected improvement (EI) methods allow sampling of points in such regions by taking in account both point prediction and high variance in prediction (e.g., Bingham et al., 2014). The latest versions of MIGRAINE use EI to generate part of the new points, by first sampling a larger number of points (typically 100 times the target number) uniformly in a given parameter region, then retaining the ones with best EI. This approach is used to more accurately identify the ML estimates, but also the confidence limits. In the latter case, confidence limits

 ISSN: 2102-6238

(λ_-, λ_+) for any parameter λ are deduced from the profile log-likelihood ratio (LR) defined by maximization over other parameters ψ. Then EI is used to select new values of ψ given $\lambda = \lambda_-$ or $\lambda = \lambda_+$. Additional points with high EI are also selected specifically outside the parameter regions with highest predicted likelihood.

A nice feature of this iterative approach is that it is not very important to have accurate estimation of likelihood in each parameter point, because the accumulation of likelihood estimates nearby the maximum (or any other target point) over successive iterations will provide, by the infill asymptotic properties of Kriging (Stein, 1999), an accurate estimation of likelihood at the maximum. For example, under models of localized dispersal (so-called "isolation by distance" in population genetics), simulating 20 genealogies per parameter point is sufficient to obtain almost perfect coverage properties of the confidence intervals (Rousset and Leblois, 2012).

3. Examples

3.1. *Inference of a founder event in Soay sheep*

We will illustrate the whole inference procedure (i.e. likelihood computation at different points of the parameter space and likelihood surface smoothing) by analyzing available data from an isolated sheep population from the island of Hirta, previously published in Overall et al. (2005). The Hirta island was evacuated of humans and their modern domestic sheep in 1930, and 107 sheep were reintroduced in 1932 from the neighboring island of Soay. The population has since remained unmanaged and the total island population has been recently observed to reach up 2000 individuals. The data set consists in 198 individual genotypes, thus 396 gene copies, screened at 17 microsatellite markers. All genotyped individuals were born in 2007.

For this application, we first considered the model of a single population with a single past change in population (i.e. the model presented in Fig. 1), which has been thoroughly tested by simulation in Leblois et al. (2014) and applied on different data sets (e.g. Vignaud et al., 2014a; Lalis et al., 2016; Zenboudji et al., 2016). Microsatellite alleles are repeats of a very short DNA motif, and mutation models generally describe the distribution of change in number of repeats when a mutation occurs. The model assumed here is the generalized stepwise mutation model (GSM, Pritchard et al., 1999) characterized by a geometric distribution of mutation steps, with parameter p_{GSM}. Four (scaled) parameters are thus inferred under this model: p_{GSM}, $\theta = 2N\mu$, $D = T/2N$, and $\theta_{\mathrm{anc}} = 2N_{\mathrm{anc}}\mu$ (see legend of Fig. 1 for explanation of the model parameters). An additional composite parameter, $N_{\mathrm{act/anc}} = N/N_{\mathrm{anc}}$, describes past changes in population size (i.e. past contraction or expansion). The present analysis consisted in 8 iterations, each with 200 parameter points. For each point the likelihood is estimated using 2,000 genealogies. Initial parameter ranges as well as point estimates and associated confidence intervals (CI) are presented in Table 1 (lines 'Single change') and examples of one- and two-dimensional profile likelihood ratio (LR) plots are shown in Fig. 2. In all such plots, the likelihood profile are inferred only from the likelihood of parameters restricted within the convex hull of sampled parameter points, i.e. ignoring values inferred by the Kriging prediction outside this region. There is enough information on all parameters in the data set for the analysis to yield peaked likelihood profiles and relatively narrow CIs for all parameters, in particular supporting a sharp and significant past contraction signal with $N_{\mathrm{act/anc}} < 0.1$.

 ISSN: 2102-6238

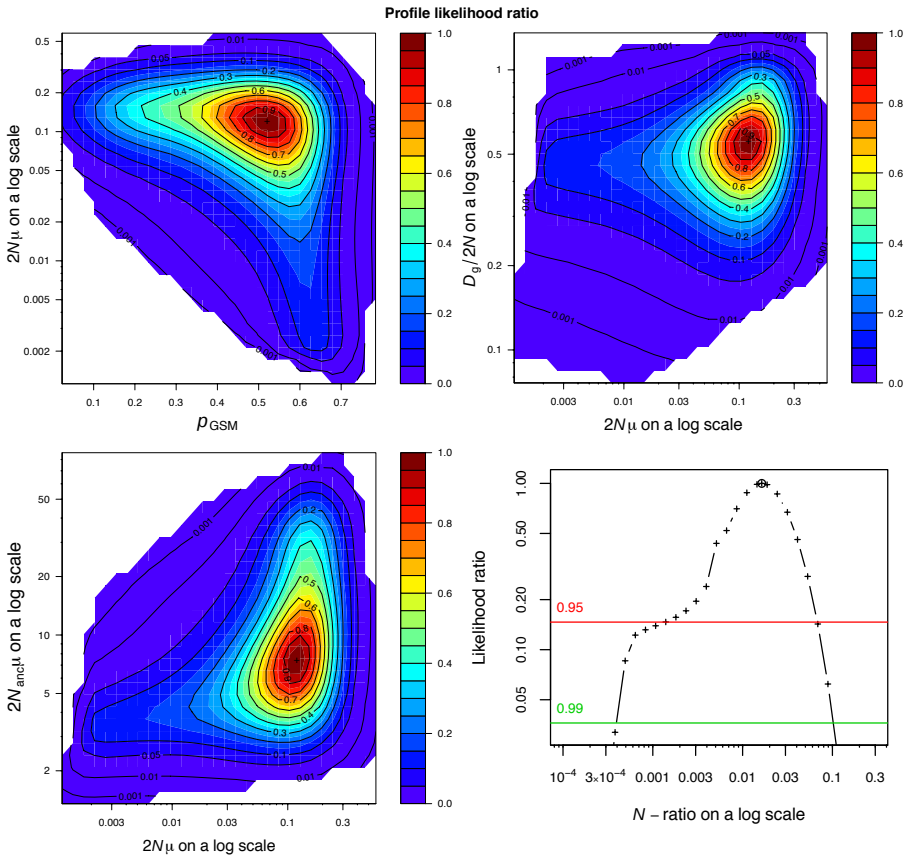


Figure 2: One- and two-dimensional profile LR plots for the sheep data set analyzed under the model with a single ancestral change in population size.

See main text and Fig. 1 for details about the model parameters and the control parameters of the iterative analysis.

In a second step, we reanalyzed the sheep data under a more complex demographic model called "Founder-Flush" (FF), illustrated in Fig. 3. The FF model is designed for the analysis of samples from an isolated population that was founded some time in the past by an unknown number of individuals coming from a stable ancestral population of unknown size, and has then grown (or declined) exponentially until present (i.e., sampling time). Such a model is well suited to study invasive, reintroduced or epidemic populations and thus seems adapted to the sheep data set from Hirta.

As for the previous analysis, we considered a GSM model for mutations but we fixed its p_{GSM} value at 0.5 (i.e. the value inferred in the previous analysis) because preliminary analyses shows flatter profile likelihood surfaces when p_{GSM} is also estimated, thus complicating the whole analysis. This is probably due to the small number of loci (i.e. 17) of the data set, resulting in a lack of information about all parameters of the model. Four (scaled) parameters are thus inferred in this analysis: $\theta = 2N\mu$, $D = T/2N$, $\theta_{\mathrm{founder}} = 2N_{\mathrm{founder}}\mu$, and $\theta_{\mathrm{anc}} = 2N_{\mathrm{anc}}\mu$ (see legend of Fig. 3

 ISSN: 2102-6238

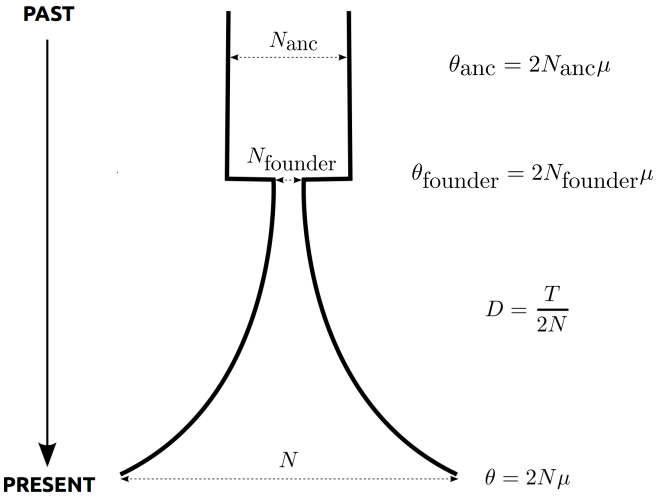


Figure 3: Representation of the Founder-Flush demographic model used for the analysis of the Soay sheep data set.

N is the current population size, N_{founder} the size of the population during the founder event, and N_{anc} is the ancestral population size (before the demographic change), T is the time measured in generation since present and μ the mutation rate of the marker used. Those four parameters are the canonical parameters of the model. θ, D, θ_{founder} and θ_{anc} are the inferred scaled parameters of the coalescent approximation.

for explanations of the model parameters). Three additional composite parameters are considered: (i) the $N_{\text{act/anc}} = N/N_{\text{anc}}$ characterizing the ratio of the current population size vs the size of the source population; (ii) $N_{\text{f/anc}} = N_{\text{founder}}/N_{\text{anc}}$ characterizing the founder event; and (iii) $N_{\text{act/f}} = N/N_{\text{founder}}$ characterizing the growth or decline of the newly founded population. The sheep data analysis under the Founder-Flush model was conducted by considering 8 iterations, with 300 points for which the likelihood is estimated using 2,000 genealogies. Initial parameter ranges as well as point estimates and associated CIs inferred after the 8 iterations are presented in Table. 1.

This second analysis under the FF model is coherent with the first analysis conducted under a simpler model: ancestral population size estimates are highly similar between the two analyses, with however narrower CI for the FF model. The FF analysis additionally detects the founding event ($N_{\text{f/anc}} = 0.00047$, CI: $[3.4 \cdot 10^{-5} - 0.0010]$). An expansion occurring after the founding event is also detected. Its estimate ($N_{\text{act/f}} = 490$, CI: $[268 - 380,000]$) is higher than expected from census sizes, which may be due to difficulties in estimating population increases that are both large and recent, but other factors, such as variance in reproductive success, may also have strongly decreased the effective size of the founder population below its census size of 107 individuals. Finally, the inferred timing of the founder event is very recent ($D = T/2N < 0.006$), but coherent with the known time of introduction (i.e. 1932, corresponding to 19 sheep generations, given a generation time of four years as reported by Coulson et al., 2010, Table 3) and the inferred current population size. This is the first time to our knowledge that a founder-flush model, characterized by two ancestral changes in population size, is fitted using microsatellite loci. This

 ISSN: 2102-6238

analysis shows that small genetic data sets, as considered here, still contain relevant information about parameters of this model.

TABLE 1. *Initial parameter ranges, point estimates and 95% CIs obtained from the three analyses of the sheep data set.*

		pGSM	θ	D	θ_{founder}	θ_{anc}	$N_{\text{act/anc}}$
Single Change	Initial range	[0.01 – 0.8]	[0.01 – 0.6]	[0.05 – 2.0]	NA	[1.0 – 80]	NA
	Final 8 iterations	0.52 [0.077 – 0.69]	0.12 [0.0052 – 0.32]	0.55 [0.23 – 1.1]	NA NA	7.36 [2.6 – 50]	0.016 [0.0014 – 0.068]
Single Change iterative procedure illustration	Initial range	[0.4 – 0.9]	[0.5 – 10.0]	[0.05 – 2.0]	NA	[1.0 – 100]	NA
	iteration 2	0.48 [0.42 – 0.60]	0.42 [*NA* – 0.54]	0.71 [0.38 – 0.99]	NA NA	12.4 [5.6 – 20]	0.034 [0.019 – 0.080]
	iteration 4	0.34 [0.25 – 0.62]	0.19 [0.17 – 0.38]	0.73 [0.38 – 1.0]	NA NA	19 [4.4 – 43]	0.010 [0.0040 – 0.074]
	iteration 10	0.54 [0.16 – 0.69]	0.13 [0.008 – 0.32]	0.52 [0.24 – 0.98]	NA NA	6.8 [2.7 – 38]	0.018 [0.0016 – 0.075]
Founder Flush	Initial range	fixed 0.5	[0.03 – 300]	$[10^{-6} – 0.5]$	$[10^{-5} – 0.1]$	[0.1 – 100]	NA
	Final 8 iterations	*NA* *NA*	1.7 [1.1 – 130]	0.0013 $[3.7 \cdot 10^{-6} – 0.0021]$	0.0034 [0.00024 – 0.0059]	7.3 [5.1 – 13]	0.23 [0.10 – 16]

See main text and Fig. 1 and 3 for details about the model parameters and the control parameters of those analyses.

3.2. Adaptive exploration of likelihood surfaces

Inference of the likelihood surface uses an iterative procedure, as described in the previous section 2.3. Here, we illustrate this iterative procedure using the sheep data and the model with a single past change in population size as before, but considering bad initial ranges for two of the four parameters (p_{GSM} and θ). This shows the capacity of MIGRAINE to automatically adjust the sampled parameter space to regions of high likelihood that were not explored in the first iterations, and to gradually increase the density of points in those regions. For that purpose, the lower bounds of initial ranges for p_{GSM} and θ were both set at higher values than the corresponding CI lower bounds obtained in the previous analysis (see Table. 1).

Expectedly, the analysis with bad initial parameter ranges required more computation than the previous analysis to get satisfactory results. We doubled the number of points (i.e. 400) for which the likelihood is estimated at each iteration compared to the previous analysis and ran 10 iterations instead of 8. All other settings are identical. Table 1 presents point estimates and associated CIs for all inferred parameters at iterations 2, 4 and 10, and Fig. 5 represents the evolution of one-dimensional LR profiles through these iterations. Those results first show that, despite the bad initial parameter ranges, MIGRAINE succeeds in generating after 10 iterations point estimates and CIs similar to those obtained in the previous analysis with better initial parameter

 ISSN: 2102-6238

ranges. Additional iterations in both analyses only marginally change the results. Second, results from intermediate iterations show how MIGRAINE progressively extend the region of high likelihood. This automatic extension of explored parameter range is apparent in Fig. 6, which shows parameter points generated at iteration 2, whose likelihoods are to be estimated in the next iteration. Therein, different points generated according to different criteria are shown in different colors, and points generated by extrapolation in a previously unexplored parameter region are shown in black. The black points are mostly located at low p_{GSM} and θ values. The same points also have high D and θ_{anc} due to correlations between p_{GSM} and these two parameters near the likelihood maximum. This parameter correlation is apparent in Fig.6 and even more visible in the two-dimensional LR profiles for (p_{GSM},D) and (p_{GSM},θ_{anc}) from the final iteration (results not shown). This diagnostic figure also shows that MIGRAINE samples points according to other criteria, aiming to ascertain the current likelihood maximum (orange points) and the CI bounds (red points), or to fill the region of high likelihood (roughly above the LR threshold of the confidence interval, but defined in two slightly different ways; green and dark blue points), or with high expected improvement outside this region (cyan points).

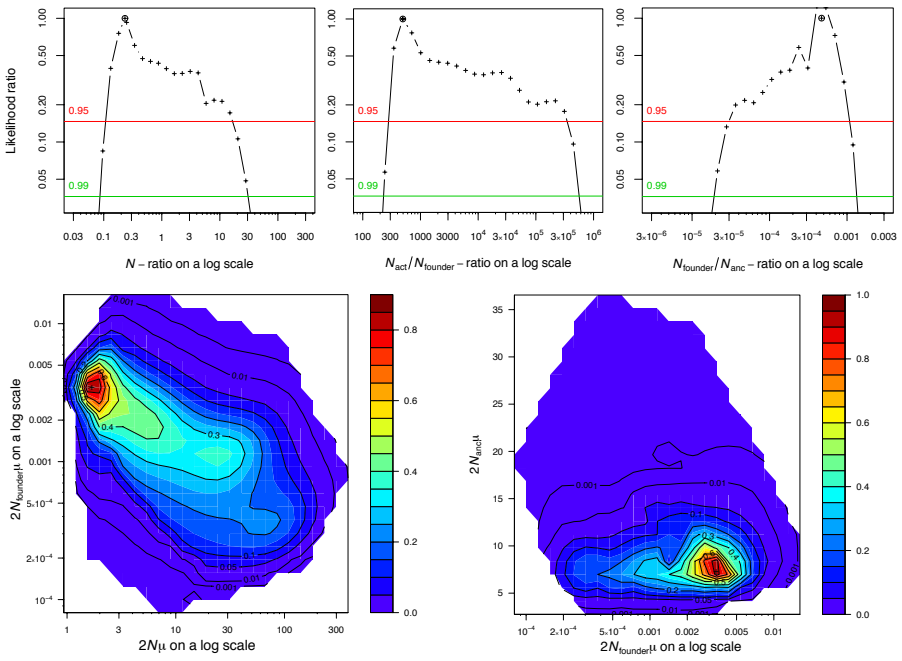


Figure 4: Examples of one- and two-dimensional profile LR plots for the sheep data set analyzed under the Founder-Flush model.

See main text and Fig. 3 for details about the model parameters and the control parameters of the analysis.

 ISSN: 2102-6238

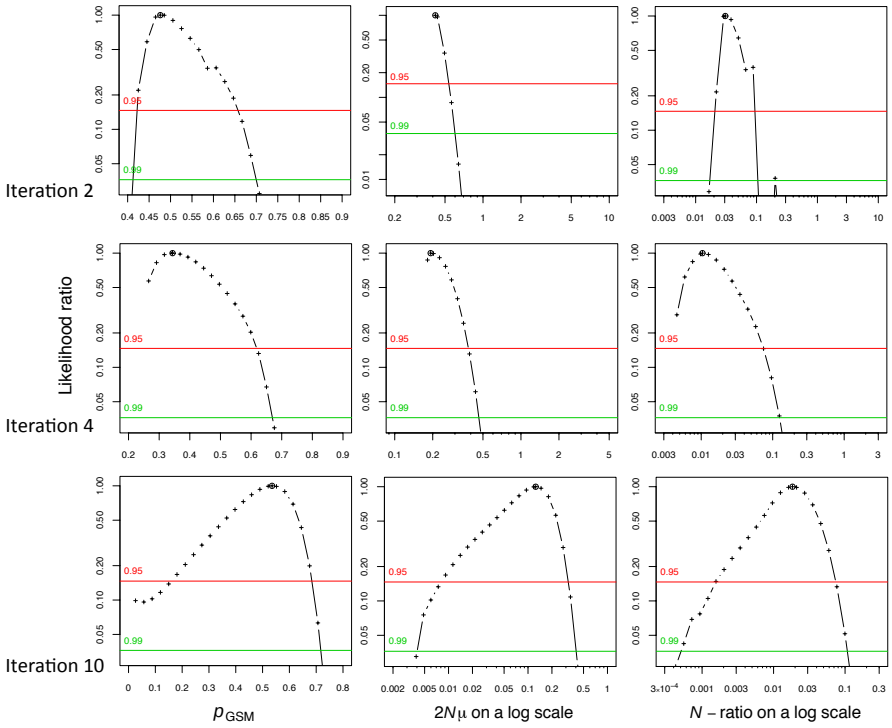


Figure 5: Illustration of the iterative procedure implemented in MIGRAINE for the sheep data set analyzed under the model with a single ancestral change in population size.

Examples of one-dimensional profile LR plots for the parameters p_{GSM}, θ, and $N_{act/anc}$ for iterations 2, 4 and 10. See main text and Fig. 1 for details about the model parameters and the control parameters of the analysis.

 ISSN: 2102-6238

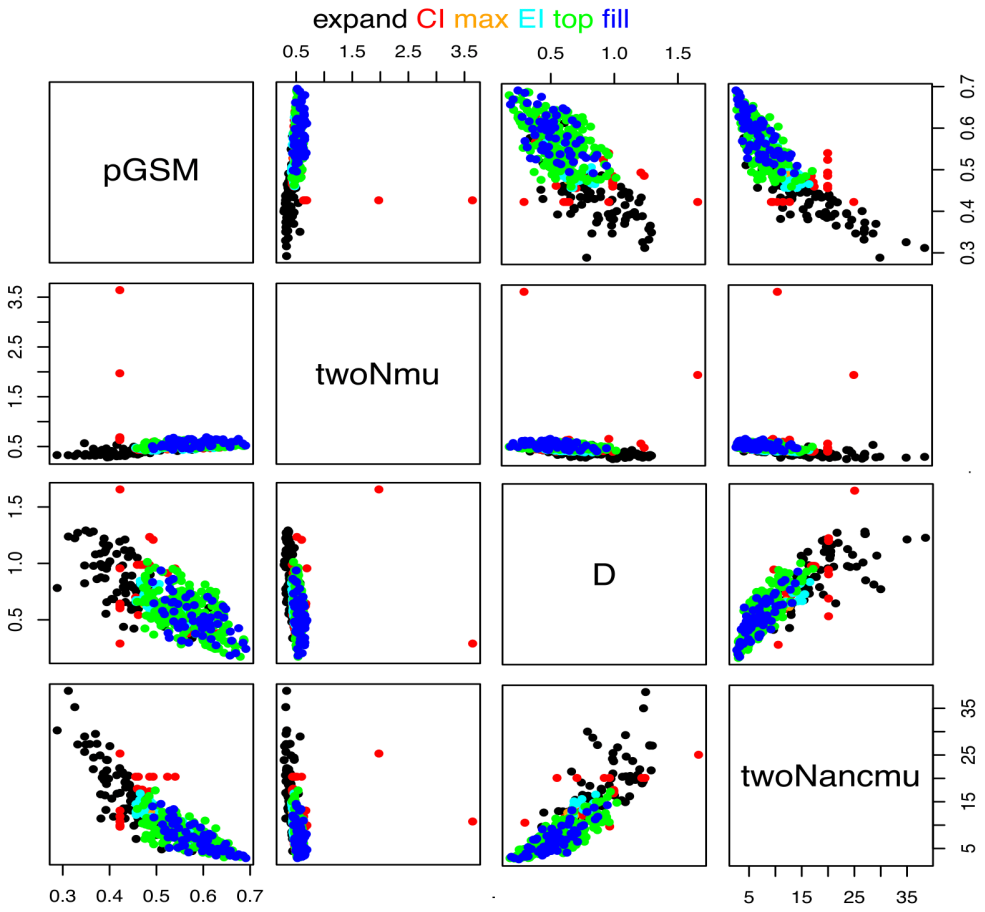


Figure 6: Diagnostic output graph illustrating the computation of new points during the iterative procedure implemented in MIGRAINE.

This graph shows the points defined at the end of iteration 2, for which the likelihood is to be estimated at iteration 3, for the sheep data set analyzed under the model with a single ancestral change in population size. See Main Text for the meaning of the different point colors. See main text and Fig. 1 for details about the model parameters and the control parameters of the analysis.

 ISSN: 2102-6238

4. Discussion

4.1. Validation

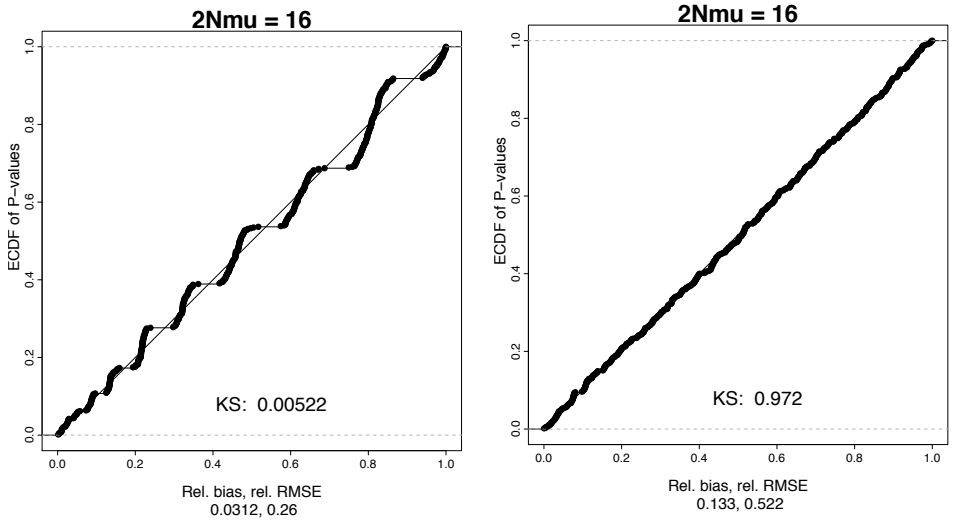
The methods reviewed in this paper have been extensively assessed, in particular in terms of coverage of confidence intervals (e.g. Fig.7 and 8, and see Rousset and Leblois, 2012; Leblois et al., 2014). Assessment of coverage is not only suitable for interval estimation, but also more useful than assessment of bias and variance to detect problems in the inference of the likelihood surface by smoothing. Such assessment would hardly deserve mention, were it not for the fact that it is not the prevalent practice in broad segments of the literature related to this work either in its objectives (inference from genetic variation) or through its methods (various stochastic methods to infer likelihoods or posterior distributions). Consequently, poorly assessed methods or software are readily available and endorsed by practitioners eager to make a story out of their data. In fact, very few publications testing methods for population genetic inference even mention confidence intervals coverage properties. Moreover, the few papers that report such information often find strong inaccuracies of the CIs (e.g. Abdo et al., 2004; Beerli, 2006; Hey, 2010; Hey et al., 2015; Appendix S3 of Peter et al., 2010).

The method defined by de Iorio and Griffiths (2004a,b) provides an approximation for the probability that a newly sampled gene is of a given type. As noted above, this approximation reduces, under a model of a single stationary population with parent-independent mutations (PIM, i.e. when the forward mutation rate from genetic type i to j is independent of i), to the true probability, and thus leads to the optimal importance sampling distribution, allowing "perfect simulation" under such a model. Under stationary models of structured populations, this approximation does not allow perfect but still very efficient importance sampling simulation.

Imperfect performance of the inferences can still result from approximations inherent in the methods. In our own work, examples include biased estimation of parameters when the analytical approximations inherent to coalescent and diffusion approximations (e.g., large population size) do not hold (Rousset and Leblois, 2012), poor robustness of some inferences with respect to details of the spatial organization of the population (Rousset and Leblois, 2007, 2012), and large variance of the importance sampling algorithm in non-equilibrium models (Leblois et al., 2014).

Poor performance could also be expected because the traditional assumptions of asymptotic likelihood theory do not hold. A first reason is the discrete nature of the data, which occasionally impacts the distribution of the likelihood ratio even in large samples. To understand how this can occur, first consider the infinite allele model (IAM), according to which each mutation generates an allele not preexisting in the population. In this model, the observed number of alleles k in a sample is a sufficient statistic for θ (Ewens, 1972), and as it is a discrete variable, the distribution of the LR is also discrete. The IAM may be seen as a limit case of the K-allele model (KAM), a model with K possible allele types and identical mutation rates between any pair of alleles. For the KAM with large K (thus approaching the IAM), but with a small mutation rate (thus with few likely values of k), steps in the distribution of the likelihood may thus become visible. This is illustrated in Fig. 7, which shows the analysis of samples of 100 gene copies generated under a 20-allele KAM. Steps are visible when these data are analyzed under a KAM with large K (400; Fig. 7a), while they disappear for small K (20; Fig. 7b). A second and more general deviation from traditional assumptions is that a sample of n genes is typically not considered as

 ISSN: 2102-6238



(a) (b)

Figure 7: Empirical cumulative distribution functions (ECDF) of P values of LR tests under a model of a single stable population

$\theta = 2N\mu = 16.0$ for a KAM with (a) $K = 400$ and (b) $K = 20$ possible alleles. Mean relative bias (rel. bias, computed as $\sum(\hat{\theta} - \theta)/\theta$) and relative root mean square error (rel. RMSE, computed as $\sum[(\hat{\theta} - \theta)/\theta]^2$ are reported. KS indicate the P value of the Kolmogorov-Smirnov test for departure of LRT P values distributions from uniformity.

resulting from n iid draws. Instead, the n genes are related through their common ancestry, and the realized ancestral genealogy can be viewed as a single draw of a latent variable. The impact of this dependence is clear for example on the inference of the mutation rate under the infinite allele model (IAM), where the variance of the ML estimator is asymptotically $O[1/\log(n)]$ rather than $O(1/n)$ (Tavaré, 1984, p. 41). Yet, in the KAM model with small K, we can achieve practically perfect coverage from small samples ($n = 30$ genes from a single locus, Fig. 7b). This observation, coupled with the fact that it is recommended to apply such methods to samples of several unlinked loci, suggests that the genealogical dependence has little impact on likelihood approximations.

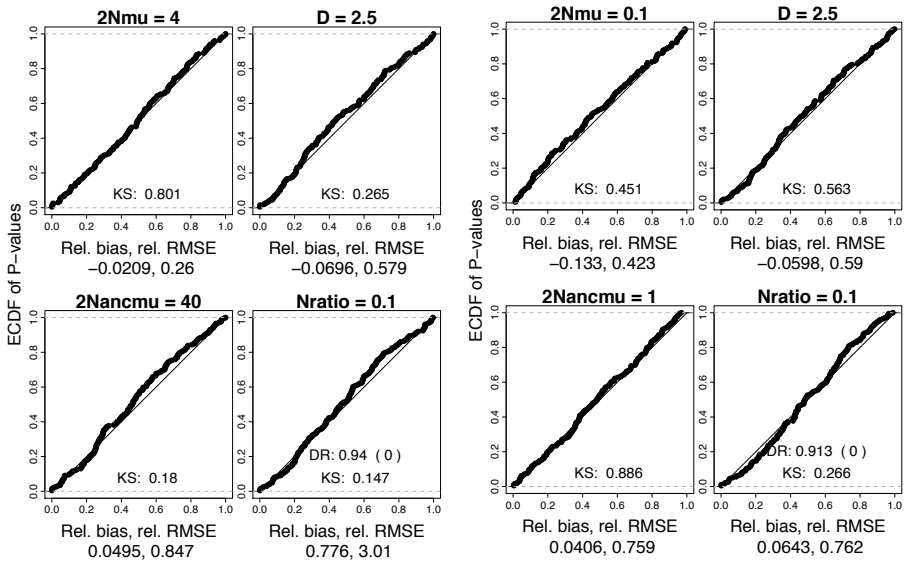
4.2. Drivers of robustness under imperfectly specified models

None of the mutation models implemented can be considered as exact representations of the actual mutation processes at the markers assayed. Thus, one typically considers a simple model such as the PIM in order to make inferences about other parameters such as dispersal rates. Robustness has been checked in this case (Rousset and Leblois, 2012). Isolation by distance analyses under a PIM model of microsatellite data simulated under a strict stepwise mutation model (SMM, Ohta and Kimura, 1973), according to which mutation results in the gain or loss of only one repeat of the DNA motif, showed that mis-specification of the mutation model has little

 ISSN: 2102-6238

impact on dispersal estimator performance, but a 50 to 75% bias in scaled mutation rate estimates is observed (Rousset and Leblois, 2007, 2012). This bias is expected because the variation in local diversity in KAM versus SMM is approximately that resulting from a 2-fold variation in mutation rate (Rousset, 1996). Similarly, inference of scaled migration rates between pairs of populations, but not of scaled mutation rate, is expected to be robust to mutational processes.

On the other hand, inferences in demographic models with time-varying parameters are much more sensitive to mutational processes. Leblois et al. (2014) showed that mis-specification of microsatellite mutational processes can induce false detection of past contraction in population sizes from samples taken from stationary populations. It can also induce biases in inferred timing and strength of a past change in population size from samples taken from a population that has indeed undergone past demographic changes. We have thus implemented variants of the importance sampling algorithms for different mutation models. Such work is illustrated for an unbounded SMM and models with one or two populations, in de Iorio et al. (2005), in Leblois et al. (2014) for a generalized stepwise mutation model (GSM) in a single population, and in Fig. 8 for the Infinitely many Site model (ISM; Kimura, 1969), a model adapted to DNA sequence markers (see next section).



(a) (b)

Figure 8: ECDF of P values of LR tests for a scenario with a single past change in population size as illustrated in Fig. 1

(a) for 30 SMM loci with $\theta = 2N\mu = 4.0$ and and $\theta_{\text{anc}} = 2N_{anc}\mu = 40.0$; (b) for 30 ISM loci with $\theta = 2N\mu = 0.1$ and and $\theta_{\text{anc}} = 2N_{anc}\mu = 1.0$; $D = T/2N = 2.5$ for both analyses. Mean relative bias (rel. bias, computed as $\sum$(observed value − parameter value)/parameter value) and relative root mean square error (rel. RMSE, computed as $\sum[$(observed value − parameter value)/parameter value$]^2$ are reported as well as the contraction detection rate (DR) and false expansion detection rate (FEDR) in parentheses after DR. KS indicate the P value of the Kolmogorov-Smirnov test for departure of LRT P values distributions from uniformity.

 ISSN: 2102-6238

4.3. *Expected developments*

All the mutation models discussed above describe allelic data, typically microsatellites, and not DNA sequences, which are also widely used genetic markers. Even if non-recombining DNA sequences can be analyzed as allelic data by considering haplotype identity only, it implies a great loss of genetic information carried by the mutations present in the different haplotypes. One mutation model adapted to DNA sequences, the infinitely-many-site model (ISM, Kimura, 1969), has been considered since the earliest developments of coalescent-based importance sampling algorithms, e.g. in the software GeneTree developed by Bahlo and Griffiths (2000) and in the approximations defined in de Iorio and Griffiths (2004b). Hobolth et al. (2008) also developed a specific proposal distribution based on an exact sampling formula from a single DNA site. Simulations showed however that the latter proposal is not more efficient than de Iorio & Griffiths' ISM specific solution derived from their general approximation (unpublished results). Nevertheless, both proposals for the ISM model have been implemented in MIGRAINE and have already allowed analysis of real data sets with sequence data (e.g., Vignaud et al., 2014b, Lalis et al., 2016). Extensive simulation tests of the ISM implementation in MIGRAINE are not yet published, but we show in Fig. 8b good performances, in terms of relative bias, relative RMSE and coverage properties of the CIs, of such analyses of DNA sequence markers evolving under the ISM compared to microsatellite markers evolving under the SMM (Fig. 8a) under a scenario with a single past change in population size (i.e., the model presented in Fig. 1).

Finally, given the explosion of single nucleotide polymorphism (SNP) data, it would be interesting to develop IS algorithms specifically adapted to SNPs, but except for de Iorio and Griffiths' suggestion to use the ISM algorithm with a single site and let θ parameters tends to 0, we are not aware of any development, application or test of IS algorithms for SNP data. SNP data may also be analyzed under a KAM with two possible alleles for the inference of dispersal between subpopulations because such inference is robust to mis-specifications of the mutation processes. On the contrary, inferences under time-inhomogeneous models may be strongly biased by such model mis-specification, especially for the timing of the different ancestral events (e.g. changes in population or divergence events).

But all algorithms dedicated to the alternative mutation models increase computation time of each replicate in comparison to the PIM, and for a given number of replicates, none has exhibited a variance as low as algorithms defined for the PIM. The current approaches for designing importance sampling algorithms are less and less efficient when mutation is more dependent on the parental type: as reviewed above, they work best for the PIM, then the GSM and the SMM, and the ISM comes last here.

4.4. *Conclusion*

The works reviewed here have shown the feasibility of likelihood-based inference for an increasing range of models of data types and demographic processes. A broader range of inferences (e.g., in demographic models with large rates of coalescence or migration) may be currently prevented by the limitations inherent to the approximations of coalescent and diffusion approaches. Analyzing a large number of loci (e.g. few thousands for typical NGS data on non-model organism) may also be challenging because of (i) the additive effect of the variance observed at

 ISSN: 2102-6238

each locus; and (ii) potentially large computation times. Such limitations underlie the persistent scope for alternative methodologies. Even within the current framework, there is still scope for substantial improvements, in particular of importance sampling algorithms for specific mutation models and time-inhomogeneous demographic models.

References

Abdo, Z., Crandall, K. A., and Joyce, P. (2004). Evaluating the performance of likelihood methods for detecting population structure and migration. *Mol. Ecol.*, 13:837–851.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *J. R. Stat. Soc. B*, 72(3):269–342.

Bahlo, M. and Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theor. Popul. Biol.*, 57:79–95.

Beaumont, M. (2010). Approximate bayesian computation in evolution and ecology. *Ann. Rev. Ecol. Evol. Syst.*, 41:379–406.

Beerli, P. (2006). Comparison of bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22:341–345.

Beerli, P. and Felsenstein, J. (1999). Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, 152:763–773.

Bingham, D., Ranjan, P., and Welch, W. J. (2014). Design of computer experiments for optimization, estimation of function contours, and related objectives. In Lawless, J. F., editor, *Statistics in Action: A Canadian Outlook*, pages 109–124. Chapman and Hall/CRC.

Cockerham, C. C. (1973). Analyses of gene frequencies. *Genetics*, 74:679–700.

Cornuet, J. M. and Beaumont, M. A. (2007). A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theor. Popul. Biol.*, 71:12–19.

Coulson, T., Tuljapurkar, S., and Childs, D. Z. (2010). Using evolutionary demography to link life history theory, quantitative genetics and population ecology. *Journal of Animal Ecology*, 79(6):1226–1240.

de Iorio, M. and Griffiths, R. C. (2004a). Importance sampling on coalescent histories. *Adv. appl. Prob.*, 36:417–433.

de Iorio, M. and Griffiths, R. C. (2004b). Importance sampling on coalescent histories. II. subdivided population models. *Adv. appl. Prob.*, 36:434–454.

de Iorio, M., Griffiths, R. C., Leblois, R., and Rousset, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor. Popul. Biol.*, 68:41–53.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87–112.

Ewens, W. J. (2004). *Mathematical population genetics I. Theoretical introduction*. Springer Verlag, New York, second edition.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13:163–185.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223.

Griffiths, R. C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying environment. *Phil. Trans. Roy. Soc. (Lond.) B*, 344:403–410.

Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene genealogies, variation and evolution*. Oxford Univ. Press, Oxford, UK.

Hey, J. (2010). Isolation with migration models for more than two populations. *Mol. Biol. Evol.*, 27:905–920.

Hey, J., Chung, Y., and Sethuraman, A. (2015). On the occurrence of false positives in tests of migration under an isolation-with-migration model. *Molecular Ecology*, 24(20):5078–5083.

Hobolth, A., Uyenoyama, M. K., and Wiuf, C. (2008). Importance sampling for the infinite sites model. *Statistical Applications in Genetics and Molecular Biology*, 7(1):1–26.

Karlin, S. and Taylor, H. M. (1981). *A second course in stochastic processes*. Acad. Press, San Diego.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903.

Lalis, A., Leblois, R., Stoetzel, E., Benazzou, T., Souttou, K., Denys, C., and Nicolas, V. (2016). Phylogeography and

 ISSN: 2102-6238

demographic history of Shaw's Jird (Meriones shawii complex) in North Africa. *Biological Journal of the Linnean Society*, 118:262–279.

Leblois, R. (2004). *Inference of dispersal parameters from genetic data in subdivided populations*. PhD thesis, Ecole Nationale Supérieure Agronomique, Montpellier, France.

Leblois, R., Pudlo, P., Néron, J., Bertaux, F., Beeravolu, C. R., Vitalis, R., and Rousset, F. (2014). Maximum likelihood inference of population size contractions from microsatellite data. *Mol. Biol. Evol.*, 31:2805–2823.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–2233.

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing*. Springer, New York.

Matérn, B. (1960). *Spatial Variation: Stochastic models and their application ·to some problems in forest surveys and other sampling investigations*. phdthesis, Forest Research Institute, Stockholm, Sweden.

Merle, C., Leblois, R., Rousset, F., and Pudlo, P. (2017). Resampling: an improvement of importance sampling in varying population size models. *Theor. Popul. Biol.*, 114:70–87.

Nath, H. B. and Griffiths, R. C. (1996). Estimation in an island model using simulation. *Theor. Popul. Biol.*, 50:227–253.

Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics*, 158:885–896.

Nychka, D. (2000). Spatial process estimates as smoothers. In Schimek, M. G., editor, *Smoothing and regression. Approaches, computation and application*, pages 393–424. Wiley, New York.

Nychka, D., Furrer, R., and Sain, S. (2015). *fields: Tools for Spatial Data*. R package version 8.2-1.

Ohta, T. and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.*, 22:201–204.

Overall, A. D. J., Byrne, K. A., Pilkington, J. G., and Pemberton, J. M. (2005). Heterozygosity, inbreeding and neonatal traits in soay sheep on st kilda. *Molecular Ecology*, 14(11):3383–3393.

Peter, B. M., Wegmann, D., and Excoffier, L. (2010). Distinguishing between population bottleneck and population subdivision by a bayesian model choice procedure. *Mol. Ecol.*, 19(21):4648–4660.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Mol. Biol. Evol.*, 16(12):1791–1798.

Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, 142:1357–1362.

Rousset, F. and Leblois, R. (2007). Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Mol. Biol. Evol.*, 24:2730–2745.

Rousset, F. and Leblois, R. (2012). Likelihood-based inferences under isolation by distance: two-dimensional habitats and confidence intervals. *Mol. Biol. Evol.*, 29:957–973.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Stat. Sci.*, 4:409–435.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for Kriging*. Springer-Verlag, New York.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics (with discussion). *J. R. Stat. Soc.*, 62:605–655.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, 26:119–164.

Vignaud, T. M., Maynard, J. A., Leblois, R., Meekan, M. G., Vázquez-Juárez, R., Ramírez-Macías, D., Pierce, S. J., Rowat, D., Berumen, M. L., Beeravolu, C., Baksay, S., and Planes, S. (2014a). Genetic structure of populations of whale sharks among ocean basins and evidence for their historic rise and recent decline. *Molecular Ecology*, 23(10):2590–2601.

Vignaud, T. M., Mourier, J., Maynard, J. A., Leblois, R., Spaet, J. L., Clua, E., Neglia, V., and Planes, S. (2014b). Blacktip reef sharks, *Carcharhinus melanopterus*, have high genetic structure and varying demographic histories in their indo-pacific range. *Molecular Ecology*, 23(21):5193–5207.

Wakeley, J. (2008). *Coalescent theory: an introduction*. Roberts and Company.

Welch, W. J., Buck, R. J., Sachs, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992). Screening, prediction, and computer experiments. *Technometrics*, 34:15–25.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugenics*, 15:323–354.

Zenboudji, S., Cheylan, M., Arnal, V., Bertolero, A., Leblois, R., Astruc, G., Bertorelle, G., Pretus, J. L., Valvo, M. L., Sotgiu, G., and Montgelard, C. (2016). Conservation of the endangered mediterranean tortoise testudo hermanni

 ISSN: 2102-6238

hermanni: The contribution of population genetics and historical demography. *Biological Conservation*, 195:279 – 291.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17:635–652.

Acknowledgements

We thank Jean-Michel Marin for inviting this contribution, and Josephine Pemberton for sharing her data from the sheep population from Hirta island. Part of this work was carried out by using the resources of the INRA MIGALE (http://migale.jouy.inra.fr) and GENOTOUL (Toulouse Midi-Pyrénées) bioinformatics platforms, the computing grid of the CBGP lab and the Montpellier Bioinformatics Biodiversity platform services. This study was supported by the Agence Nationale de la Recherche (projects IM-Model CORAL.FISH 2010-BLAN-1726-01 and GENO-SPACE ANR-16-CE02-0008) and by the Institut National de Recherche en Agronomie (Project INRA Starting Group "IGGiPop", and postdoctoral funding for C. R. Beeravolu).

5. Appendix

An efficient importance sampling algorithm has been formulated using concepts from diffusion theory. We first recall how diffusion models are used in population genetics (see Ewens, 2004 for an extensive introduction).

A process of allele frequency change in a finite population of size N with (say) mutation rate μ is approximated by the limiting process as $N \to \infty$, of a series of processes X_N with the same $N\mu$, each measured in time scaled by N. For example, consider a locus with only two possible alleles, with mutation probability μ to the other allele per gene copy per generation. The expected allele frequency in the next generation is $\mathrm{E}(x') = x(1-\mu) + (1-x)\mu$. in the classical Wright-Fisher model for a population of N haploid individuals, the allele frequency in the next generation is a binomial sample of size N with binomial frequency $\mathrm{E}(x')$ as given above. The change in allele frequency has then expectation $\mathrm{E}(x') - x = (1-2x)\mu$ and variance $\mathrm{E}(x')(1-\mathrm{E}(x'))/N$. The limiting process in scaled time is then described by the infinitesimal moments in scaled time in units of N generations, $M(x) = N\mu(1-2x)$ and $V(x) = x(1-x)$. In particular, the transition density $\phi(X_t|X_0)$ of allele frequency in the limiting process satisfies the backward Kolmogorov equation

$$\frac{\partial \phi(X_t|X_0 = x)}{\partial t} = \left(\frac{1}{2}V(x)\frac{\partial^2}{\partial x^2} + M(x)\frac{\partial}{\partial x}\right)\phi(X_t|x) \equiv \mathscr{L}\phi(X_t|x). \tag{18}$$

A backward equation holds also for the expectation of any function $f(x)$ with bounded second derivatives ("generator equation"; Karlin and Taylor, 1981, p. 215),

$$\lim_{t\to 0}\frac{\mathrm{E}(f(X_t)|x) - f(x)}{t} = \left(\frac{1}{2}V(x)\frac{\partial^2}{\partial x^2} + M(x)\frac{\partial}{\partial x}\right)f(x) = \mathscr{L}f(x). \tag{19}$$

These results are extended to models with multiple alleles and subpopulations, with the following notations. We consider deme sizes, N_d for deme d, which sum to N_T; a matrix of scaled forward mutation rates $N_\mathrm{T}\mu_{ij} \equiv N_\mathrm{T}\mu P_{ij}$ from i to j (which is row-stochastic, i.e., $\sum_j P_{ij} = 1$); and

 ISSN: 2102-6238

a matrix of scaled forward migration rates $N_T m_{dd'}$ from deme d' to d. the diffusion process is now the limit, as $N \to \infty$, of a series of processes X_N with the constant $N\mu_{ij}$, constant $Nm_{dd'}$, and constant relative deme sizes. Then the generator can be written

$$\mathscr{L} = \frac{1}{2} \sum_{\text{demes } d} \sum_{\text{allele pairs } i,j} \frac{N_T}{N_d} x_{di}(\delta_{ij} - x_{dj}) \frac{\partial^2}{\partial x_{di} \partial x_{dj}} + \sum_d \sum_i M_{di} \frac{\partial}{\partial x_{di}} \tag{20}$$

where $M_{dj} = N_T \mu \sum_j (P_{ij} - \delta_{ij}) x_{di} + N_T \sum_{d'} (x_{d'j} - x_{dj}) m_{dd'}$.

At stationary equilibrium, $\mathrm{E}[\mathscr{L} f(\mathbf{x})] = 0$ where $\mathbf{x} \equiv (x_{id})$ is the vector of frequencies of allele i in deme d, and expectation is taken over the joint stationary density $\psi(\mathbf{x})$ of these allele frequencies. Applying this result for f taken as the sample probability given $\mathbf{x}$, i.e. $f(\mathbf{x}) = \prod_d \binom{n_d}{(n_{di})} \prod_i x_{di}^{n_{di}}$ where x_{di} is the frequency of allele i in deme j, leads to a relation between probabilities of samples that differ by one coalescence/mutation/migration event:

$$\begin{aligned} N_T \left(\sum_d n_d \left(\frac{n_d - 1}{N_d} + m_d + \mu \right) \right) q(\mathbf{S}) = \\ N_T \sum_{d,j} n_d \frac{n_{dj} - 1}{N_d} q(\mathbf{S} - \mathbf{e}_{dj}) \\ + N_T \mu \sum_{d,j} \sum_i P_{ij} (n_{di} + 1 - \delta_{ij}) q(\mathbf{S} - \mathbf{e}_{dj} + \mathbf{e}_{di}) \\ + N_T \sum_{d,j} n_d \sum_{d' \neq d} m_{dd'} \frac{n_{d'j} + 1}{n_{d'} + 1} q(\mathbf{S} - \mathbf{e}_{dj} + \mathbf{e}_{d'j}). \end{aligned} \tag{21}$$

We use eq. 13 to express eq. 21 as a recursion involving ancestral samples differing by the subtraction of one gene copy relative to the descendant sample, by expressing all $q(.)$ in terms of $q(\mathbf{S} - \mathbf{e}_{dj})$s for distinct d, j:

$$\begin{aligned} N_T \sum_{d,j} \left(\frac{n_d - 1}{N_d} + m_d + \mu \right) \pi(j|d, \mathbf{S} - \mathbf{e}_{dj}) n_d q(\mathbf{S} - \mathbf{e}_{dj}) = \\ N_T \sum_{d,j} n_d \frac{n_{dj} - 1}{N_d} q(\mathbf{S} - \mathbf{e}_{dj}) \\ + N_T \mu \sum_{d,j} \sum_i P_{ij} n_d \pi(i|d, \mathbf{S} - \mathbf{e}_{dj}) q(\mathbf{S} - \mathbf{e}_{dj}) \\ + N_T \sum_{d,j} n_d \sum_{d' \neq d} m_{dd'} \pi(j|d', \mathbf{S} - \mathbf{e}_{dj}) q(\mathbf{S} - \mathbf{e}_{dj}) \end{aligned} \tag{22}$$

This provides no solution for the π's, as the closed system of equations for sample probabilities implied by this one is not simplified in any way, and remains too large. But de Iorio and Griffiths (2004b) instead considers the equations defined for each d, j by extracting the left-hand and right-hand side coefficients of $q(\mathbf{S} - \mathbf{e}_{dj})$. The system of such equations for different samples of same size as $\mathbf{S}$ over all d and j is generally inconsistent. However, the system of equations

 ISSN: 2102-6238

for identical $q(\mathbf{S}-\mathbf{e}_{dj})$ over different d, j leads to a linear system of equations of dimension the number of demes times the number of alleles. Each such equation reduces to

$$N_{\mathrm{T}}\left(\frac{n_d-1}{N_d}+m_d+\mu\right)\hat{\pi}(j|d,\mathbf{S}-\mathbf{e}_{dj}) = \frac{n_{dj}-1}{N_d} + N_{\mathrm{T}}\mu\sum_i P_{ij}\hat{\pi}(i|d,\mathbf{S}-\mathbf{e}_{dj}) + N_{\mathrm{T}}\sum_{d'\neq d} m_{dd'}\hat{\pi}(j|d',\mathbf{S}-\mathbf{e}_{dj}) \quad (23)$$

where e.g. $\sum_i P_{ij}\hat{\pi}(\ldots)$ represents a sum over different possible ancestral sample configurations with an additional i gene, cf eq. (21). The $\hat{\pi}$s solving this system are not the true πs, but they provides approximations for the πs from which importance sampling weights and a proposal distribution can be defined.

 ISSN: 2102-6238

Bioinformatics, 37(20), 2021, 3673–3675
doi: 10.1093/bioinformatics/btab261
Advance Access Publication Date: 8 May 2021
Applications Note

Genetics and population analysis

GSpace: an exact coalescence simulator of recombining genomes under isolation by distance

Thimothée Virgoulay [1,2,*], François Rousset [1], Camille Noûs[3] and Raphaël Leblois [2]

[1]Institut des Sciences de l'Evolution, Univ Montpellier, CNRS, IRD, EPHE, Montpellier, France,[2]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montferrier sur Lez, France and [3]Laboratoire Cogitamus, Univ Montpellier, Montpellier, France

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on December 17, 2020; revised on April 16, 2021; accepted on April 27, 2021editorial decision on April 20, 2021;

Abstract

Motivation: Simulation-based inference can bypass the limitations of statistical methods based on analytical approximations, but software allowing simulation of structured population genetic data without the classical *n*-coalescent approximations (such as those following from assuming large population size) are scarce or slow.

Results: We present GSpace, a simulator for genomic data, based on a generation-by-generation coalescence algorithm taking into account small population size, recombination and isolation by distance.

Availability and implementation: Freely available at site web INRAe (http://www1.montpellier.inra.fr/CBGP/software/gspace/download.html).

Contact: thimothee.virgoulay@umontpellier.fr

1 Introduction

GSpace is a program that can simulate neutral genomic data with recombination under a wide range of demographic models. It is based on a backward in time generation-by-generation (gen-by-gen) approach, coupled with an efficient recombination algorithm and flexible models for dispersal and subpopulation sizes. It simulates the ancestry of a sample of haploid or diploid individuals (in both cases following a standard haplo-diploid sexual life cycle), carrying one or more chromosomes.

Individual dispersal is generally restricted in space in natural populations (isolation by distance: Endler, 1979; Rousset, 1997; Wright, 1943). To represent this fact, GSpace considers a lattice of subpopulations of any size (down to single individuals) connected by limited dispersal, according to different possible dispersal distributions, including in particular fat-tailed distributions, such as the Zeta (discretized Pareto, Patil and Joshi, 1968) and the Sichel (Chesson and Lee, 2005).

The case where each subpopulation on the lattice hosts a single individual or a mating pair and dispersal is mostly restricted to a few steps apart is suitable to represent a range of territorial species inhabiting a continuous habitat (e.g. Rousset, 2000), but cannot be simulated when considering extensions of Kingman's (1982) *n*-coalescent which assume large sub-population sizes and small migration rates. To simulate small population sizes and high dispersal rates without biases (Nelson *et al.*, 2020), coalescence probabilities exact for small population size (Fu, 2006) must be used in a gen-by-gen simulation until the common ancestors of the whole simulated sample have been found. Such simulations are required to assess any inference framework which might for example allow separate estimation of sub-population size, mutation and migration probabilities, something that is not possible under *n*-coalescent approximations. In all these respects, GSpace retains and extends for genomic data some of the previous features of the IBDsim software (Leblois *et al.*, 2009). The current version considers only time-homogeneous models but time-heterogeneous models will be implemented in future versions similarly to IBDsim.

As gen-by-gen algorithms are expected to be slower than those involving *n*-coalescent approximations, we performed simulations to check the feasibility of simulating genomic data by such algorithms, and compared computation times with those of alternative software based on *n*-coalescent approximations, such as msprime (Kelleher *et al.*, 2016), FastSimcoal2 (Excoffier *et al.*, 2013), exact coalescence algorithms implemented as DTWF in msprime python package [back-in-time Wright-Fisher simulator, Nelson *et al.* (2020)], IBDSim (Leblois *et al.*, 2009) and forward algorithms, such as SimBit (Matthey-Doret, 2020).

The gen-by-gen algorithm in GSpace is slower than those involving *n*-coalescent approximations but much faster than IBDSim or forward simulators like SimBit in most cases (see Section 4).

2 Implementation

GSpace combines some parts of the modified Hudson's algorithm (Hudson, 1983) for recombination and coalescence implemented in msprime with previous features from IBDSim, in a new implementation in modern C++, as follows. At each generation going backward in time,

the program considers all possible migration, recombination and coalescence events, until all common ancestors have been found. Because neutral genetic data are simulated, genetic states do not affect genealogical trees and mutations can then be added downwards to the gene tree of each chromosome segment that did not recombine. Implementation details of such algorithm can be found in Kelleher *et al.* (2016) and Nelson *et al.* (2020) for the approximated and gen-by-gen algorithms for coalescence with recombination, respectively; and in Leblois *et al.* (2009) for gen-by-gen algorithms under isolation by distance. We only highlight below what can make GSpace different from other software.

At each generation t, the coordinates of the parent of each individual carrying ancestral lineages are randomly drawn in a 2D backward dispersal distribution of the position of a parent given the position of the lineage. The backward distributions are deduced by assuming that dispersal occurs independently in each dimension forward in time, and can automatically handle spatial heterogeneity (i.e. different forward migration rates and size of sub-populations on the lattice) as well as various edge effects. The program can consider (i) uniform, geometric and discretized Gaussian, Zeta and Sichel forward dispersal distributions, including the stepping stone and island models as special cases, as well as (ii) a custom forward migration rate matrix. Each chromosome harbors multiple discrete potentially recombining sites and the program handles multiple recombination events per chromosome, even in a single generation. When a recombination event occurs in a diploid genome in the backward simulation, the segments on each side of the recombination point originate from each of the two parental chromosomes and have a distinct coalescence history further backward in time. When a coalescence event occurs, ancestral segments of all descendant chromosomes have a unique parental segment and share a common coalescence and migration history until a recombination event occurs. The combination of such gen-by-gen diploid coalescence, migration and recombination algorithms simulates the exact patterns of linkage disequilibrium expected under a haplo-diploid life cycle.

Mutations are then added independently on each gene tree, going forward in time on each branch, from the common ancestor to the leaves. As the underlying algorithm assumes a finite number of mutable sites GSpace can handle numerous nucleotidic and allelic mutation models (e.g. IAM, KAM, JC69, see user manual for more models) but not the infinite site model.

3 Compilation, automated checks, inputs and outputs

The program is written in modern C++ (17) and can be compiled on any operating system with a modern compiler (g++ $\geq$ 7.5, clang $\geq$ 6.0) with simple command line arguments, or using the CMake build system (both the command line arguments and the CMake commands are provided in the manual). The CMake build includes unit tests for each part of the program and functional tests comparing simulation outcomes in terms of probability of identity of pairs of genes at one and two loci to theoretical results (see Rousset, 2004 and Vitalis and Couvet, 2001).

GSpace's runs can be controlled both by a settings file and by command-line arguments, which together allow the easy specification of many parameters, and quick changes of selected parameters between simulations. The settings file is exacxtread first, and allows the user to control all options of GSpace (detailed in the user manual). These options can then be altered by the command line arguments. Results can be saved in three different file formats for individual genetic data: Genepop for allelic data, Fasta and VCF (v4.3) for sequence data; as well as in the new binary treeSequence format (see tskit documentation) for efficient storage of trees with mutations.

4 Comparison

Gspace is, at the time of writing, the only gen-by-gen coalescence simulator specifically designed to handle recombination and allowing easy specification of various forward dispersal distributions. Thus, it cannot be compared in terms of computation time to other simulators not sharing such features, but it has been compared to algorithms from five other simulators in simpler cases: the n-coalescent approximations implemented in msprime v1.0.0a5 ('msprime') and FastSimcoal2 v2.6.0.3 ('fsc2'); the gen-by-gen algorithms implemented in msprime v1.0.0a5 ('DTWF') and IBDSim v2.0; and the forward simulator SimBit v3.9.13. Results for IBDSim are not detailed here because it cannot consider recombination and is not designed to handle long DNA sequences (e.g. $> 10^5$ bp). However, without recombination and for many allelic loci, GSpace is two to fifty time faster. Other simulations are detailed in Table 1 and show that although GSpace is not the fastest simulator, its speed approaches that of the approximate simulators rather than that of other generation-by generation ones.

Table 1. Comparison of computation times between GSpace and other simulators under three different demographic and mutational schemes.

	Method (see text for details)				
Case	msprime	fsc2	GSpace	DTWF	SimBit
A	0.320	3.959	6.231	7.096	56.781
B	14.507	5.664	7.471	36.335	52.121
C	0.048	0.016	0.028	2.459	39.055

Note: Mean run time in seconds over 100 (10 for SimBit) replicates for the simulation of a sample of 1000 haploid individuals carrying a single chromosome of 10^7 base pairs, with mutation and recombination rates of 10^{-8} per generation per site under: (A) a Wright–Fisher model with a population size of 10000 haploid individuals; (B and C) an island model with 20 subpopulations of 500 haploid individuals each and 50 sampled chromosomes of (B) 10^7 base pairs or (C) with 10^4 base pairs

Acknowledgements

We thank A. Dehne-Garcia, F.-D. Collin and M. Navascues for initial discussions on algorithms and code, as well as J. Kelleher and P. Ralph for constructive comments and help with tskit during the review process.

Funding

This work used the following HPC platforms: INRA MIGALE (http://migale.jouy.inra.fr) and GENOTOUL (Toulouse Midi-Pyrénées), Montpellier Bioinformatics Biodiversity supported by the LabEx CeMEB (ANR-10-LABX-04-01), and CBGP host platform. All authors were supported by the Agence Nationale de la Recherche (RL & TV: projects GENOSPACE ANR-16-CE02-0008 and Labex Cemeb ProLag; FR & RL: project INTROSPEC ANR-19-CE02-0011).

Conflict of Interest: none declared.

References

Chesson,P. and Lee,C.T. (2005) Families of discrete kernels for modeling dispersal. *Theor. Popul. Biol.*, **67**, 241–256.
Endler,J.A. (1979) Gene flow and life history patterns. *Genetics*, **93**, 263–284.
Excoffier,L. *et al.* (2013) Robust demographic inference from genomic and snp data. *PLoS Genet.*, **9**, e1003905.
Fu,Y.-X. (2006) Exact coalescent for the wright–fisher model. *Theor. Popul. Biol.*, **69**, 385–394.
Hudson,R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, **23**, 183–201.
Kelleher,J. *et al.* (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.*, **12**, e1004842.
Kingman,J. (1982) The coalescent. *Stoch. Process. Their Appl.*, **13**, 235–248.

Leblois,R. *et al.* (2009) Ibdsim: a computer program to simulate genotypic data under isolation by distance. *Mol. Ecol. Res.*, **9**, 107–109.
Matthey-Doret, R. (2021) SimBit: A high performance, flexible and easy-to-use population genetic simulator. *Mol Ecol Resour.* 10.1111/1755-0998.13372
Nelson,D. *et al.* (2020) Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLoS Genet.*, **16**, e1008619.
Patil,G.P. and Joshi,S.W. (1968) *A dictionary and bibliography of discrete distributions*. Published for the International Statistical Institute by Oliver and Boyd Edinburgh.
Rousset,F. (1997) Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.
Rousset,F. (2000) Genetic differentiation between individuals. *J. Evol. Biol.*, **13**, 58–62.
Rousset,F. (2004) *Genetic Structure and Selection in Subdivided Populations*. Monographs in population biology. Princeton University Press, Princeton University, New Jersey.
Vitalis,R. and Couvet,D. (2001) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, **157**, 911–925.
Wright,S. (1943) Isolation by distance. *Genetics*, **28**, 114–138.