Inférence en génétique et génomique des populations

Renaud Vitalis

Académie de Montpellier
Université Montpellier 2
Sciences et Techniques du Languedoc

École doctorale SIBAGHE

Habilitation à Diriger des Recherches

Dossier de candidature

Inférence en génétique et génomique des populations

par

Renaud VITALIS

Soutenue le 16 novembre 2012 devant le jury composé de

Christine DILLMANN	Professeur, Université Paris–Sud	Rapporteur
Joëlle RONFORT	Directrice de Recherche, INRA Montpellier	Rapporteur
Xavier VEKEMANS	Professeur, Université Lille 1	Rapporteur
Oscar GAGGIOTTI	Professeur, Université de Grenoble	Examinateur
Etienne KLEIN	Directeur de Recherche, INRA Avignon	Examinateur

Table des matières

Liste des Figures

Avant Propos

Cette habilitation à diriger des recherches m'a donné l'occasion de revenir sur onze années de recherche depuis ma thèse. Onze années qui me confortent dans l'idée que j'aime ce métier, non seulement pour la liberté de travail qu'il procure mais aussi pour la qualité des rencontres qu'il favorise. J'ai pris la liberté dans ce document de ne pas détailler de la même manière tous les travaux effectués. Je me suis rendu compte en rédigeant que je préférais mettre l'accent sur mes recherches en cours, celles pour lesquelles j'ai le plus investi d'énergie ces derniers temps, que de revenir sur des modèles ou des questions plus anciennes. C'est la raison pour laquelle je présente en Annexe deux manuscrits en préparation (Annexes F et G), mais auxquels je porte aujourd'hui une attention particulière. Tous ces travaux n'auraient pu être réalisés sans les collègues et les étudiants qui se retrouveront, j'espère, dans ce document : merci à eux. Merci aussi à ceux qui ont œuvré et qui œuvrent encore à faciliter les mobilités entre institutions : le CBGP est un lieu précieux, où il fait bon travailler, et la communauté montpelliéraine offre un environnement de recherche éminemment appréciable.

Merci enfin à Louise, Mahault et Hélène qui ont subi ces quelques semaines de rédaction. J'essaierai de me rattraper.

Première Partie

Document de synthèse

Introduction

Le développement des techniques de biologie moléculaire 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, ce qui se traduit par l'émergence d'un nouveau champ disciplinaire au croisement de la génétique des populations et de la génomique : la "génomique des populations" (Black *et al.* 2001; Luikart *et al.* 2003). Dans les prochaines années, la vitesse à laquelle de nouvelles données moléculaires vont être produites ne va cesser d'augmenter. Les généticiens des populations qui, jusqu'il y a peu, avaient à leur disposition quelques dizaines de marqueurs microsatellites, s'engagent désormais sur des programmes de séquençage de marqueurs ADN associés à des sites de restrictions (Baird *et al.* 2008; Lewis *et al.* 2007; Miller *et al.* 2007a,b), voire de séquençage de génomes *de novo*, y compris pour des espèces non-modèles. L'externalisation de ces techniques et la baisse des coûts associés, font que l'analyse des données de polymorphisme est en train de vivre une véritable révolution. La nécessité d'un traitement efficace de ces données est donc plus que jamais d'actualité.

Parallèlement à cette explosion de données génomiques, les outils d'analyse en génétique des populations ont eux aussi connu ces dernières années une évolution importante, du fait notamment de l'utilisation croissante de techniques statistiques nouvelles reposant sur le maximum de vraisemblance ou bien sur des approches bayésiennes. Il est désormais possible de développer des modèles qui permettent l'analyse de scénarios démographiques complexes, et d'utiliser ces modèles pour l'inférence statistique de paramètres démographiques à partir de données de polymorphisme génétique. Ces avancées

méthodologiques ont été rendues possibles grâce à la combinaison de développements théoriques majeurs : (*i*) la théorie de la coalescence (Kingman 1982a,b), qui offre un modèle probabiliste des généalogies de gènes dans les populations et permet ainsi de simuler des généalogies sous un grand nombre de modèles démographiques ; (*ii*) la théorie de la diffusion (Crow et Kimura 1971), qui approxime le processus discret de la dérive génétique à un processus de diffusion en temps continu, et qui permet ainsi de calculer la vraisemblance d'un échantillon de gènes ; (*iii*) les méthodes de Monte Carlo pour l'inférence statistique, qui rassemblent un grand nombre d'algorithmes comme les méthodes de Monte Carlo par chaînes de Markov (*Monte Carlo Markov Chains* ou MCMC), ou les méthodes approchées d'estimation bayésienne (*Approximate Bayesian Computation* ou ABC) (voir Beaumont 2010; Csilléry *et al.* 2010; Marjoram et Tavaré 2006). Les méthodes d'inférences basées sur ces techniques sont récentes, et ont seulement commencé à diffuser dans la communauté scientifique. Elles sont en passe de révolutionner l'analyse des données de polymorphisme génétique. Pourtant, ces méthodes qui permettent de tenir compte efficacement de l'information contenue dans les données de polymorphisme nécessitent souvent une puissance de calcul importante. Il s'agit donc désormais de réfléchir aux conditions de leur application sur des jeux de données issus, par exemple, des nouvelles technologies de séquençage.

D'une certaine manière, les recherches que j'ai conduites depuis ma thèse reflètent cette évolution de la taille et de la nature des données, ainsi que des outils statistiques afférents. Les deux premiers chapitres de cette habilitation traitent de l'inférence en génétique et génomique des populations. Dans le premier chapitre, je fais état de mes travaux passés sur l'inférence de la démographie des populations, tout en développant plus particulièrement des travaux plus récents qui concernent le développement de méthodes d'analyse bayésiennes pour retracer l'histoire démographique des populations, à partir de très gros jeux de données de polymorphisme génétique. Dans le second chapitre, je présente mes travaux visant à rechercher des signatures de la sélection naturelle ou artificielle dans les génomes. J'y détaille tout d'abord des méthodes relativement simples de détection de marqueurs "atypiques",

reposant sur la mesure d'un écart à un modèle nul de neutralité sélective. Je développe enfin mes recherches récentes qui ont permis la mise au point d'une nouvelle méthode visant à estimer l'intensité de la sélection dans les génomes. Le troisième et dernier chapitre de cette habilitation concerne un autre champ thématique, celui de l'évolution des traits d'histoire de vie en populations structurées, auquel je continue de me consacrer depuis ma thèse en marge de mes recherches sur l'inférence en génétique et génomique des populations.

Chapitre 1

Histoire démographique des populations

La génétique des populations s'intéresse aux mécanismes qui sous-tendent l'évolution de la variation génétique dans les populations. Ses développements théoriques permettent non seulement de mieux appréhender l'importance relative des différentes forces évolutives qui sont en jeu, mais aussi de proposer des outils d'analyse du polymorphisme afin d'en inférer des éléments de l'histoire démographique des populations. Dans ce chapitre, je développerai quelques uns de mes travaux qui m'ont permis de proposer ou de tester des méthodes d'inférence des biais de dispersion liés au sexe des individus, des changements de taille des populations, ou encore de l'histoire de divergence des populations.

1.1 La dispersion biaisée en faveur d'un sexe

Chez beaucoup d'espèces, la dispersion (c'est-à-dire la capacité à quitter son lieu de naissance pour aller se reproduire ailleurs) est un trait lié au sexe des individus. On dit qu'il existe chez ces espèces un *biais* de dispersion spécifique au sexe. Chez les oiseaux, ce sont en général les femelles qui dispersent plus que les mâles, tandis que chez les mammifères, ce sont en général les mâles qui dispersent plus que les femelles (Greenwood 1980). Comprendre

comment les processus démographiques sexe-spécifiques influencent la diversité génétique et sa distribution est depuis longtemps un sujet d'intérêt majeur (Disotell 1999).

1.1.1 Apport des marqueurs autosomaux

Afin de mieux comprendre les conséquences d'un biais de dispersion lié au sexe sur la distribution du polymorphisme génétique dans les populations naturelles, j'ai développé un modèle analytique de populations subdivisées chez une espèce à sexes séparés [P10] [1]. La définition de paramètres de différenciation génétique pour des marqueurs neutres (F-statistiques) spécifiques au sexe, m'a permis de montrer dans quelles conditions la différenciation génétique peut être différente entre les individus de sexe opposé. En effet, si l'on ne s'attend pas à trouver de différence de structure génétique entre sexes lorsque les individus sont échantillonnés *avant* la dispersion, la structure génétique mesurée *après* dispersion est plus forte chez les individus du sexe le moins dispersant (figure 1.1). Ce modèle m'a donc permis de proposer une méthode d'inférence des taux de migration par sexe, à partir de l'estimation du rapport des mesures de différenciation réalisées *avant* et *après* dispersion [P10]. En particulier, si l'on note m_X (respectivement m_Y) la probabilité qu'un individu de sexe X (respectivement de sexe Y) soit immigrant, et n_d le nombre de sous-populations, alors le rapport de la différenciation mesurée *après* dispersion (F_{ST}^{XY}) à la différenciation mesurée *avant* dispersion (F_{ST}^{*}) vaut :

$$\frac{F_{\mathrm{ST}}^{XY}}{F_{\mathrm{ST}}^{*}} = d_{XY} + b_{XY} \cdot F_{\mathrm{ST}}^{XY} \qquad \text{pour tout } (X,Y) \in \{♂,♀\}^2 \tag{1.1}$$

où $d_{XY} \equiv (1 - m_X\,[n_\mathrm{d}/(n_\mathrm{d}-1)])(1 - m_Y\,[n_\mathrm{d}/(n_\mathrm{d}-1)])$, et b_{XY} est la fréquence des paires d'individus tirés au hasard dans des sous-populations distinctes et qui proviennent d'une même sous-population à la génération précédente. Le dernier terme du second membre de l'équation (1.1) est négligeable devant d_{XY} et, par conséquent, F_{ST}^{*} ne diffère de F_{ST}^{XY} que par un facteur

1. Les références indiquées entre crochets correspondent aux numéros des publications listées dans l'annexe B, page 119

d_{XY}. Pour de grandes valeurs de n_{d}, $d_{XY} \approx (1 - m_X)(1 - m_Y)$. On peut donc, très simplement obtenir le taux de migration des individus de sexe X, à partir du rapport du F_{ST}^{XX} spécifique au sexe des individus échantillonnés, évalué *après* la dispersion, sur le F_{ST}^{*} évalué pour des individus échantillonnés *avant* la dispersion :

$$m_X \approx 1 - \sqrt{F_{\mathrm{ST}}^{XX} / \ F_{\mathrm{ST}}^{*}} \qquad \text{pour tout } X \in \{\male, \female\} \tag{1.2}$$

S'il existait déjà d'autres méthodes permettant de *détecter* des biais de dispersion liés au sexe (voir par exemple Favre *et al.* 1997; Goudet *et al.* 2002), cette méthode est la première à avoir permis de *quantifier* les biais de dispersion liés au sexe à partir de marqueurs nucléaires (voir aussi Fontanillas *et al.* 2004). J'ai également développé et distribué un logiciel qui permet d'appliquer cette méthode pour le calcul des taux de dispersion par sexe [L2].

J'ai par la suite étendu cette approche, développée à l'origine pour l'analyse de données de marqueurs autosomaux, pour l'appliquer à l'analyse de marqueurs uni-parentaux tels que l'ADN mitochondrial. Dans le cadre d'une collaboration avec Jennifer Cooper (qui a réalisé sa thèse à Purdue University, USA, sous la direction de Peter Waser), cette nouvelle approche a été appliquée sur le pécari à collier (*Pecari tajacu*) à partir d'échantillons provenant de 30 groupes sociaux appartenant à trois populations texanes. Dans un premier temps, nous avons adapté la méthode de ré-échantillonnage proposée par Goudet *et al.* (2002) à une analyse moléculaire de variance hiérarchisée (AMOVA, Excoffier *et al.* 1992) de la région de contrôle de l'ADN mitochondrial. Cette méthode revient simplement à générer un grand nombre de jeux de données pseudo-observés où la classe des individus (âge, sexe, etc.) est ré-assignée au hasard, à l'intérieur des groupes sociaux. Nous avons ainsi mis en évidence un signal clair de dispersion biaisée [P18] : tout d'abord, nous avons montré que la différenciation génétique mesurée entre groupes sociaux au sein des populations (F_{GP}) est plus forte qu'attendue chez les juveniles sous l'hypothèse nulle que la dispersion n'est pas âge-spécifique ($p = 0.20$), ce qui n'est pas le cas chez les adultes ($p = 0.79$) (figure 1.2A–

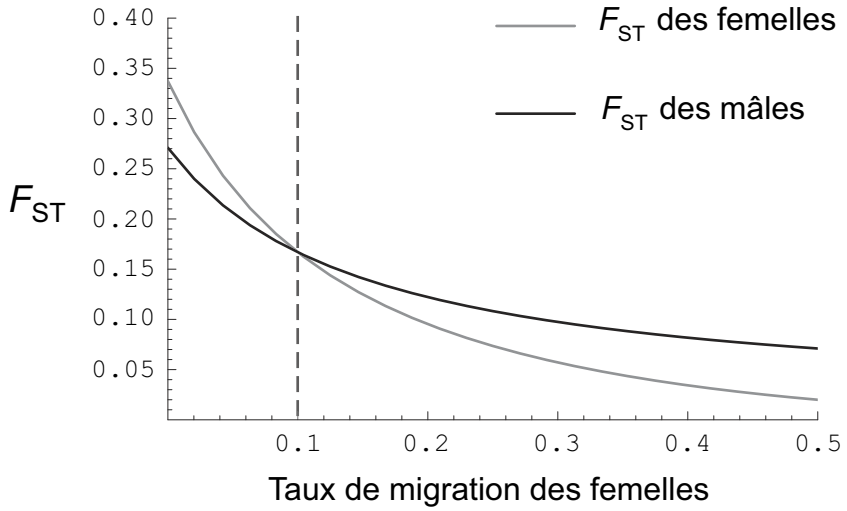


Figure 1.1: Valeurs à l'équilibre des mesures de différenciation génétique (F_{ST}) par sexe, après dispersion, dans une population subdivisée en 20 dèmes de 10 individus. Les valeurs des paramètres sont données ici en fonction du taux de migration des femelles, pour un taux de migration mâle fixé à 10%. A gauche de la ligne en pointillé, les femelles dispersent *moins* que les mâles, et la différenciation génétique mesurée chez les femelles est donc plus *forte* que chez les mâles. A droite de la ligne en pointillé, les femelles dispersent *plus* que les mâles, et la différenciation génétique mesurée chez les femelles est donc plus *faible* que chez les mâles.

B), bien que ces résultats ne soient pas significatifs. Ensuite, nous avons montré que la différenciation génétique mesurée entre groupes sociaux au sein des populations (F_{GP}) est significativement plus forte qu'attendue chez les femelles sous l'hypothèse nulle que la dispersion n'est pas sexe-spécifique ($p = 0.0002$), ce qui n'est pas le cas chez les mâles ($p = 0.98$) (figure 1.2C–D). Nos résultats plaident donc en faveur d'une dispersion biasée en faveur des mâles adultes. Nous avons donc défini, *a posteriori* deux classes d'individus : les "dispersants" (mâles adultes) et les "non-dispersants" (femelles adultes et juvéniles). Comme attendu, la différenciation génétique mesurée entre groupes sociaux au sein des populations (F_{GP}) est plus forte qu'attendue chez les "non-dispersants" sous l'hypothèse nulle ($p < 0.0001$), ce qui

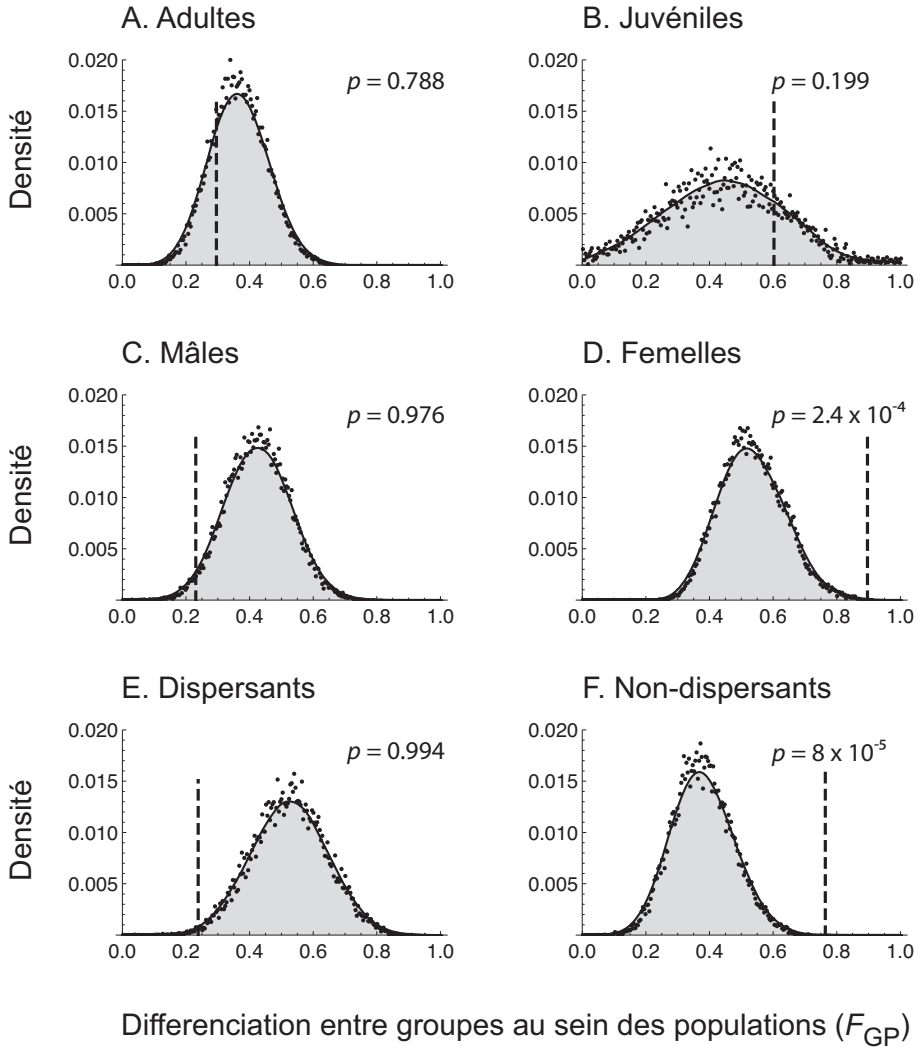


Figure 1.2: Distributions de la différenciation génétique mesurée entre groupes sociaux au sein des populations (F_{GP}) pour différentes classes d'individus (âge, sexe, statut de dispersion), attendues sous l'hypothèse nulle que la classe des individus n'affecte pas la dispersion. Ces distributions sont obtenues en réassignant au hasard la classe des individus, et reposent sur 10 000 permutations. Les valeurs observées dans le jeu de données sont représentées par les lignes verticales pointillées.

n'est pas le cas chez les "dispersants" ($p = 0.99$) (figure 1.2E–F). Enfin, nous avons également estimé à partir de ces données les taux de dispersion sexe-

spécifiques contemporains entre groupes sociaux à l'intérieur des populations, à partir de l'équation :

$$m_X \approx 1 - \sqrt{F_{\mathrm{GP}}^{XX} / \ F_{\mathrm{GP}}^{*}} \qquad \text{pour tout } X \in \{\male, \female\} \tag{1.3}$$

qui est une extension de l'équation 1.2 dans le cas s'une structure sociale hiéarchique. Nos estimations donnent un taux de dispersion entre groupes sociaux $m_{\male} = 0.37$ ($\mathrm{CI}_{95\%} = [0.32, 0.65]$) pour les mâles. Étant donnée l'absence de différence significative entre les mesures de différenciation F_{GP} avant et après dispersion pour les femelles, le taux de dispersion pour les femelles n'a pas pu être calculé. Néanmoins, ces résultats montrent bien que la distribution des haplotypes de l'ADN mitochondrial peut être utilisée pour estimer des taux de dispersion sex-spécifiques instantanés.

1.1.2 Combiner différentes catégories de marqueurs

Dispersion biaisée dans l'espèce humaine

Les biais de dispersion liés au sexe des individus existent également chez l'Homme (Cann 2001; Disotell 1999; Oota *et al.* 2001; Pennisi 2001; Seielstad *et al.* 1998; Stoneking 1998). Jusqu'à présent, c'est l'analyse de marqueurs uni-parentaux (ADN mitochondrial et portion non-recombinante du chromosome Y) qui a été privilégiée (Wilkins et Marlowe 2006). La plupart des études semblent montrer que la différenciation génétique est moins forte lorsqu'elle est mesurée sur des marqueurs mitochondriaux (à hérédité maternelle) que lorsqu'elle est mesurée sur des marqueurs du chromosome Y (à hérédité paternelle). Ces résultats ont été interprétés principalement par une plus forte migration des femmes, conséquence de la patrilocalité (une tendance pour les hommes à rester sur le lieu de leur naissance, tandis que les femmes se déplacent dans le foyer de leur mari). Or la différenciation des populations dépend principalement du produit du nombre efficace d'individus au sein de chaque dème et du taux de migration entre dèmes. Par conséquent, la dispersion biaisée en faveur d'un sexe et les différences entre effectifs efficaces pour chaque sexe ont un effet confondant sur la structure gé-

nétique mesurée sur des marqueurs hérités uni-parentalement. De plus, du fait de l'absence de recombinaison, l'ADN mitochondrial comme le chromosome Y ne constituent chacun, de fait, qu'un unique marqueur. Enfin, la présence de gènes potentiellement soumis à la sélection sur ces marqueurs peut créer, à travers des phénomènes de balayage sélectif (Maynard Smith et Haigh 1974), des biais importants pour l'inférence de paramètres démographiques sexe-spécifiques. Dans le cadre de sa thèse, que j'ai co-encadrée avec Évelyne Heyer, Laure Ségurel a donc développé une nouvelle approche pour analyser conjointement le polymorphisme mesuré sur des marqueurs autosomaux et sur des marqueurs liés au chromosome X. Cette analyse multi-locus avait pour objectif de produire des informations plus robustes, en visant à démêler les effets confondants de la dispersion biaisée et des biais de sexe-ratio sur la structure génétique. Dans la suite de cette section, je présenterai tout d'abord le terrain d'étude sur lequel ces analyses ont été conduites, avant de détailler le principe des méthodes utilisées.

Un terrain d'étude pertinent : l'Asie Centrale

L'histoire évolutive de l'Homme moderne est caractérisée, sur les 100 000 dernières années, par des phases d'expansion, de colonisation, et de migrations récurrentes entre des populations établies. Ces événements, tous intimement liés, ont façonné la variation génétique de manière complexe (Harding et McVean 2004; Przeworski *et al.* 2000). Certaines régions du monde ont probablement servi de corridors naturels pour ces mouvements de populations. Située au cœur de l'Eurasie, l'Asie Centrale est probablement l'une de ces routes de migration (Cavalli-Sforza *et al.* 1994; Nei et Roychoudhury 1993), bien que son rôle dans l'expansion de l'Homme moderne à sa sortie d'Afrique, et dans la colonisation plus récente de populations différenciées soit incertain (Cordaux *et al.* 2004; Karafet *et al.* 2001; Wells *et al.* 2001). Durant son post-doctorat, que j'ai co-encadré avec Évelyne Heyer, Begoña Martínez-Cruz a donc cherché à tester si cette région du monde a connu une "phase de maturation" au Paléolithique avant la colonisation de l'Eurasie par des vagues de migration successives, ou bien au contraire

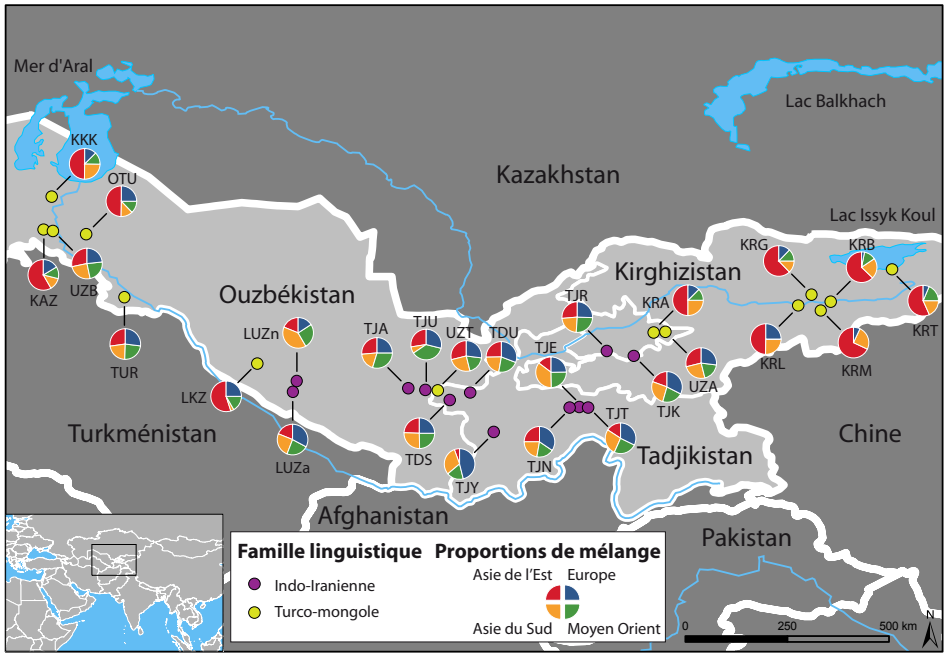


Figure 1.3: Localisation géographique des 21 populations d'Asie Centrale étudiées. Les camemberts représentent les proportions de mélange de chaque population.

si elle a été le lieu de rencontres de populations Asiatiques et Européennes différenciées après leur expansion.

Histoire du peuplement en Asie Centrale

Durant son post-doctorat, Begoña Martínez-Cruz a donc génotypé 780 individus à 27 locus microsatellites, dans 26 populations d'Asie Centrale représentant six groupes ethniques et deux familles linguistiques : indo-européenne et turco-mongole (figure 1.3). Il s'agissait de la première étude multi-locus réalisée, à si grande échelle, en Asie Centrale [P25].

Nous avons trouvé des niveaux de diversité génétique importants, et nos résultats montrent que la différenciation génétique entre populations est principalement expliquée par l'affiliation linguistique et ethnique, plutôt que par la géographie. Cette étude tend à placer l'Asie Centrale dans une position intermédiaire entre l'Europe et le Moyen-Orient, le Pakistan et l'Asie de l'Est

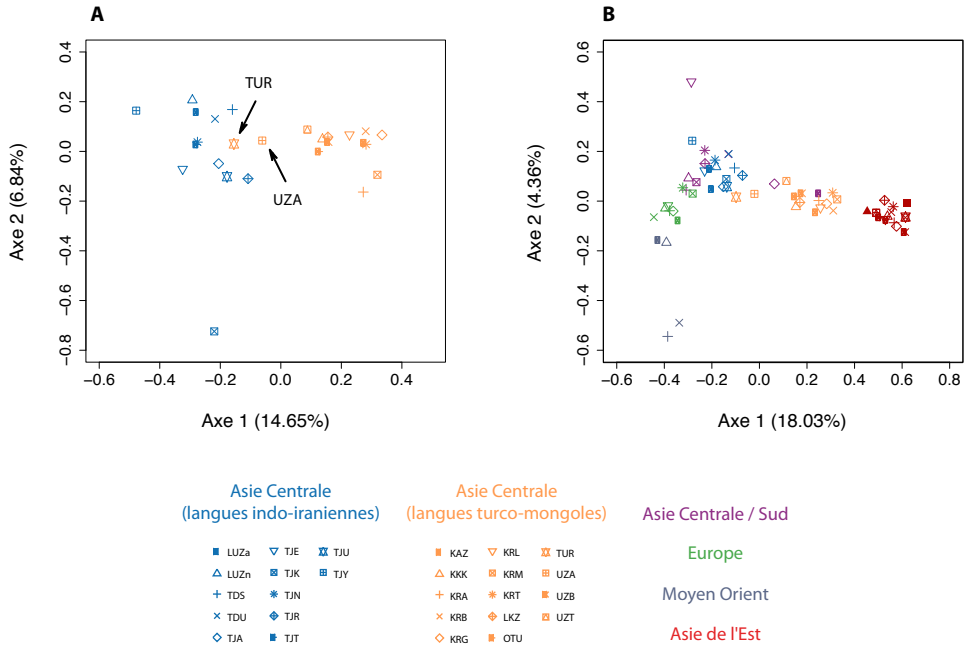


Figure 1.4: Analyse des correspondances réalisée sur les données microsatellites en Asie Centrale. L'analyse a été faite à partir de la matrice des comptages alléliques. (A) Analyse des populations d'Asie Centrale. (B) Analyse des populations d'Eurasie. Dans les deux graphes, les deux premières composantes factorielles sont représentées et leur contribution relative à l'inertie totale est indiquée.

(figures 1.4 et 1.5), suggérant que l'Asie Centrale est une région de mélange de populations venant de l'Est et de l'Ouest de l'Eurasie. Les populations de langue turco-mongole sont plus proches des populations est-asiatiques, tandis que les populations de langue indo-iranienne se regroupent avec les populations de l'Ouest de l'Eurasie. Les populations ouzbèques se répartissent entre les populations de langues turco-mongole et indo-iranienne, ce qui peut être interprété comme le témoignage de leur origine liée à la réunion de différentes tribus. Ces travaux nous ont amenés à proposer que le paysage génétique complexe de l'Asie Centrale est le résultat de mouvements récur-

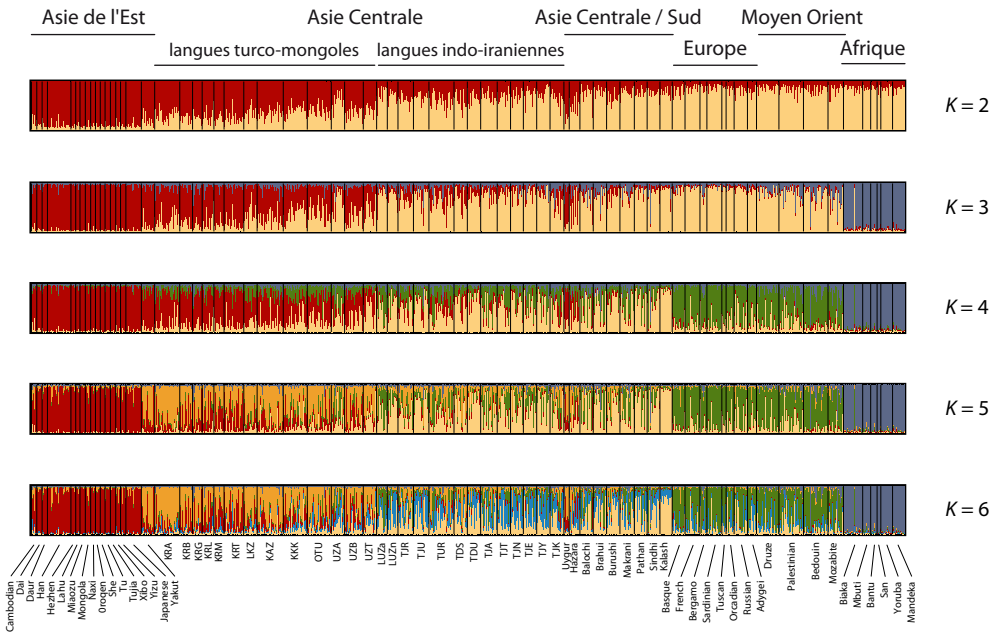


Figure 1.5: Analyse de la structure des populations d'Asie Centrale. Les analyses ont été réalisées grâce au logiciel STRUCTURE (PRITCHARD *et al.* 2000). K représente le nombre de groupes putatifs. Chaque individu est représenté par une ligne verticale, divisée en K segments au plus. La taille de chaque segment représente la proportion du génome de chaque individu ayant pour origine le groupe en question. Les analyses ont été réalisées sur 767 individus provenant de 26 populations d'Asie Centrale, génotypés à 27 marqueurs microsatellites, ainsi que 869 individus provenant de 44 populations africaines et eurasiatiques du *HGDP-CEPH Human Genome Diversity Cell Line Panel* (CANN *et al.* 2002).

rents de groupes nomades turco-mongols venant de l'Est, dans un groupe de populations sédentarisées, dont les Tadjiks et les Turkmènes seraient les représentants aujourd'hui. Nos résultats montrent également que ces invasions récurrentes n'ont pas entraîné le remplacement complet des populations locales, contrairement à ce qui est généralement avancé (ZERJAL *et al.* 2002).

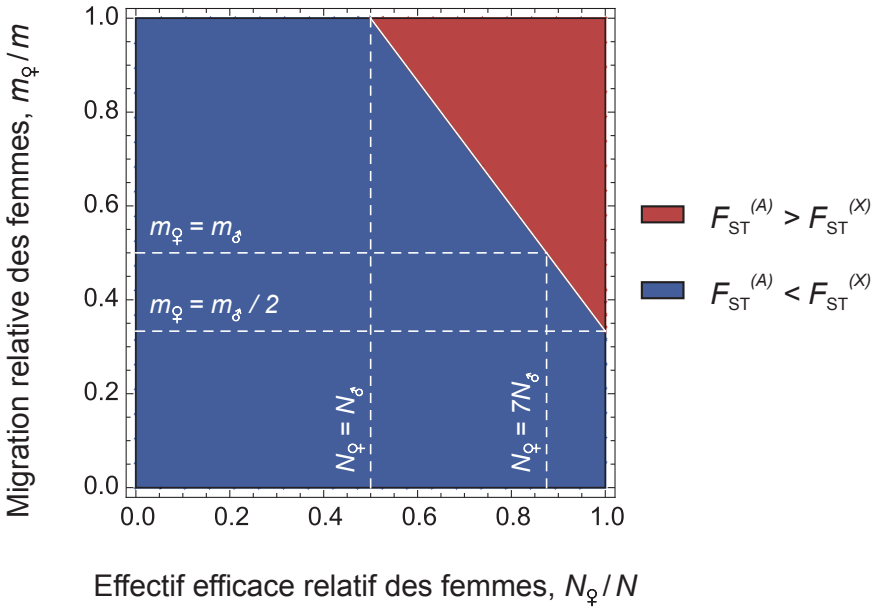


Figure 1.6: Diagramme représentant la différenciation génétique attendue mesurée sur des autosomes, et sur le chromosome X. Ce graphique résulte de l'analyse du modèle en îles étendu au cas des espèces à sexes séparés. Dans le triangle supérieur droit (en rouge), les valeurs de F_{ST} sur les autosomes sont supérieures à celles sur le chromosome X. Dans ce cas, la fraction femelle des effectifs efficaces ($N_♀/N$) est nécessairement plus grande que 0.5. Dans la région inférieure gauche (en bleu), les valeurs de F_{ST} sur les autosomes sont plus faibles que sur le chromosome X. La ligne pleine blanche représente l'ensemble des valeurs de paramètres pour lesquels les valeurs de F_{ST} sur les autosomes et sur le chromosome X sont égales. Les lignes en pointillés représentent trois cas particuliers : En $N_♀ = N_♂$, la plus faible taille efficace sur le chromosome X (qui serait 3/4 de celle sur les autosomes) peut seulement être contre-balancée par un biais complet de migration en faveur des femelles. Au contraire, si $m_♀ = m_♂$, alors un grand effectif efficace femelle compense exactement la faible taille efficace du chromosome X pour $N_♀ = 7N_♂$. Pour $m_♀ = m_♂/2$, les valeurs de F_{ST} sur les autosomes et sur le chromosome X ne peuvent être égales que si le nombre de mâles tend vers zéro.

Apport des marqueurs non-autosomaux

Dans le cadre de son Master 2 puis de sa thèse, Laure Ségurel a cherché à estimer la structure génétique sexe-spécifique en Asie Centrale,

notamment pour comparer cette structure chez les populations d'éleveurs nomades et chez les agriculteurs sédentaires. Pour cela, Laure Ségurel a développé une nouvelle approche, basée sur l'analyse de la distribution conjointe du polymorphisme génétique sur des marqueurs autosomaux et sur des marqueurs liés au chromosome X dans un modèle en îles (Wright 1931) étendu au cas des espèces à sexes séparés. Dans ce modèle, on peut montrer en effet que le niveau de différenciation attendu sur des marqueurs autosomaux est égal à :

$$F_{\mathrm{ST}}^{(A)} \approx \frac{1}{1+\left(4\frac{4N_{\mathrm{♀}}N_{\mathrm{♂}}}{N_{\mathrm{♀}}+N_{\mathrm{♂}}}\right)\left(\frac{m_{\mathrm{♀}}+m_{\mathrm{♂}}}{2}\right)} \tag{1.4}$$

tandis qu'il est égal à :

$$F_{\mathrm{ST}}^{(X)} \approx \frac{1}{1+4\left(\frac{9N_{\mathrm{♀}}N_{\mathrm{♂}}}{2N_{\mathrm{♀}}+4N_{\mathrm{♂}}}\right)\left(\frac{2m_{\mathrm{♀}}+m_{\mathrm{♂}}}{3}\right)} \tag{1.5}$$

sur des marqueurs liés au chromosome X. En réarrangeant, on obtient :

$$\frac{1-1/F_{\mathrm{ST}}^{(X)}}{1-1/F_{\mathrm{ST}}^{(A)}} = \frac{3}{4}\frac{(1+m_{\mathrm{♀}}/m)}{(2-N_{\mathrm{♀}}/N)}, \tag{1.6}$$

et, par conséquent :

$$F_{\mathrm{ST}}^{(X)} = \frac{4F_{\mathrm{ST}}^{(A)}}{4F_{\mathrm{ST}}^{(A)}-3\left(F_{\mathrm{ST}}^{(A)}-1\right)\left(\frac{1+m_{\mathrm{♀}}/m}{2-N_{\mathrm{♀}}/N}\right)}. \tag{1.7}$$

À partir de ces équations, Laure Ségurel a pu montrer que lorsque la différenciation génétique sur les autosomes est supérieure à celle mesurée sur le chromosome X, cela implique que le nombre efficace de femmes est nécessairement plus grand que celui des hommes, quel que soit le patron de dispersion sexe-spécifique (figure 1.6). Ce résultat suggère qu'il est donc possible de séparer les effets relatifs des biais de dispersion et des différences d'effectifs efficaces, à partir de l'analyse conjointe du polymorphisme sur les autosomes et le chromosome X [P15]. Il est par exemple possible de tester

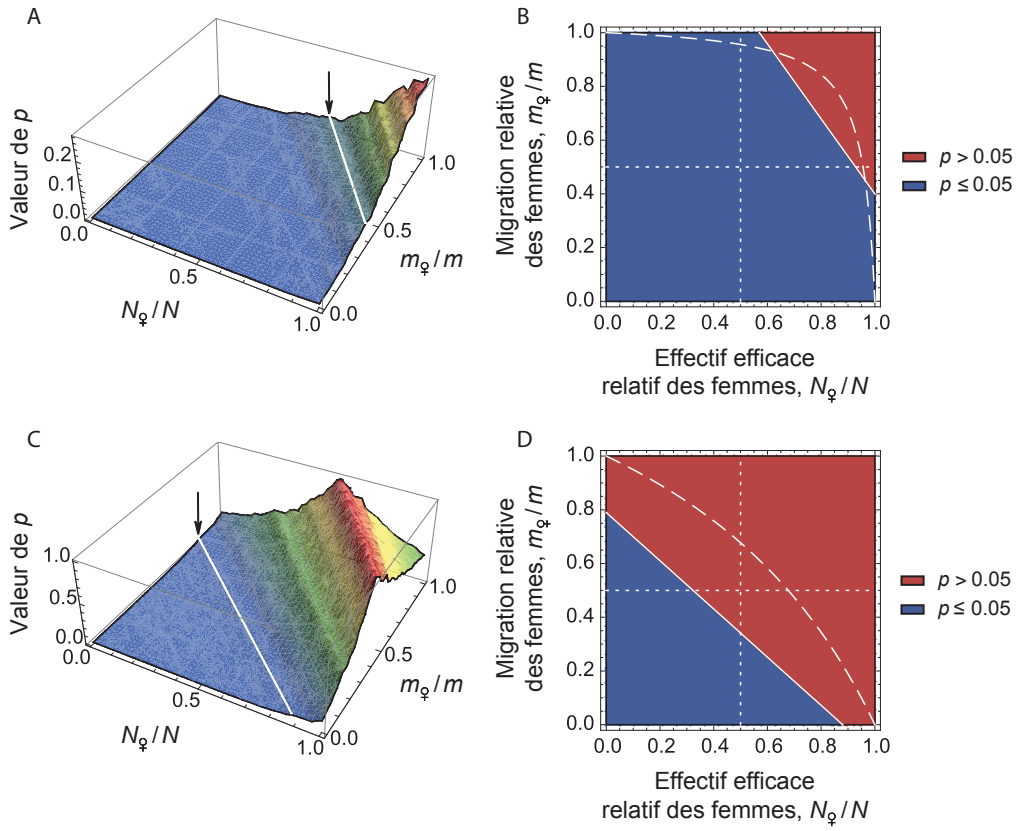


Figure 1.7: Valeurs critiques (p) des tests de Wilcoxon réalisés pour tester si les niveaux de différenciation mesurés sur les autosomes et sur le chromosome X sont compatibles avec les valeurs de paramètres $N_♀/N$ et $m_♀/m$. (A) Valeurs critiques (p) des tests de Wilcoxon, en fonction du rapport des effectifs efficaces ($N_♀/N$) et du taux relatif de migration des femmes ($m_♀/m$), pour les populations d'éleveurs nomades. La flèche indique la ligne isocline qui sépare la région où $p \leq 0.05$ de celle où $p > 0.05$. Des valeurs de p non-significatives ($p > 0.05$) correspondent à des valeurs de paramètres qui ne peuvent pas être rejetées. (B) Graphe des contours, pour les mêmes données. (C) et (D) sont équivalents à (A) et (B), respectivement, pour les populations d'agriculteurs sédentaires.

si des valeurs de paramètres ($N_♀/N$,$m_♀/m$) données sont compatibles avec les niveaux de différenciation mesurés conjointement sur les autosomes et sur le chromosome X. Pour cela, on peut comparer la distribution des valeurs

observées de $F_{\mathrm{ST}}^{(X)}$ à celle des valeurs prédites de $F_{\mathrm{ST}}^{(X)}$ d'après l'équation 1.7, grâce à un test de Wilcoxon. Ceci nous permet alors de rejeter des valeurs de paramètres incompatibles avec les données observées (figure 1.7).

En appliquant ces résultats à l'analyse des populations d'Asie Centrale, Laure Ségurel a montré que dans les groupes nomades, non seulement le taux de dispersion des femmes mais aussi leur effectif efficace relatif sont plus élevés que ceux des hommes (figure 1.7A–B). Ce biais de dispersion et d'effectif efficace n'est pas retrouvé chez les populations sédentaires dans la même région (figure 1.7B–C). Ces différences entre groupes ethniques sont interprétées comme la conséquence de pratiques sociales contrastées : les Tadjiks (agriculteurs sédentaires) sont organisés principalement en familles nucléaires ou étendues, et les mariages sont préférentiellement endogames (entre cousins). Au contraire, les Kazaks, Karakalpaks, Kirghizes et Turkmènes (traditionnellement éleveurs nomades) sont organisés en groupes d'apparentés (tribus, clans, lignées) et les mariages sont plutôt exogames, entre clans. La différence de patron de dispersion sexe-spécifique entre agriculteurs sédentaires et éleveurs nomades s'expliquerait donc par les mouvements de femmes entre clans chez les éleveurs nomades (mais pas chez les agriculteurs sédentaires) bien que ces deux groupes soient patrilocaux. La différence d'effectifs efficaces relatifs entre hommes et femmes s'expliquerait quant à elle par une réduction d'effectif efficace des hommes liée à l'organisation en groupes d'hommes d'apparentés et/ou par la transmission culturelle du succès reproducteur des hommes. Ces résultats montrent surtout l'importance de considérer les phénomènes sexe-spécifiques à une échelle locale [P15]. En cela, ces résultats contribuent à la prise de conscience que l'organisation sociale et le mode de vie ont une influence importante sur la distribution de la variation génétique dans les populations humaines. La publication [P15] qui a découlé de ce travail est reproduite en Annexe C, page 131.

1.2 Dispersion limitée dans l'espace

Les modèles d'inférence de la dispersion que j'ai présentés jusqu'ici reposent sur le modèle en îles (Wright 1931), et ne prennent donc pas en

compte l'aspect spatial de la dispersion. Or chez de nombreuses espèces, dont l'espèce humaine, la dispersion des individus est limitée dans l'espace : les individus se reproduisent en général avec des individus qui leur sont géographiquement proches. Les modèles d'isolement par la distance, qui prennent en compte ces caractéristiques de la dispersion, prédisent une augmentation de la distance génétique avec la distance géographique entre les populations et/ou les individus (Rousset 1997). Des études théoriques, mais aussi empiriques, montrent que le modèle d'isolement par la distance est le modèle le plus robuste pour l'inférence de la dispersion à partir de données de polymorphisme génétique (voir Leblois *et al.* 2003).

1.2.1 Un exemple chez l'Homme

À travers plusieurs collaborations avec Paul Verdu, dont la thèse a été dirigée par Évelyne Heyer, je me suis intéressé à l'inférence de la dispersion chez un groupe de chasseurs-cueilleurs mobiles : les Pygmées. L'objectif principal de la thèse de Paul Verdu était de reconstruire l'histoire démographique des Pygmées Baka du Cameroun, une des plus grandes populations Pygmée d'Afrique Centrale.

Histoire du peuplement des Pygmées d'Afrique Centrale

L'Ouest de l'Afrique Centrale est aujourd'hui peuplé par de nombreuses populations d'agriculteurs sédentaires, au voisinage du plus grand groupe de chasseurs-cueilleurs mobiles : les Pygmées, dont l'histoire démographique reste assez mal connue. Paul Verdu a génotypé 604 individus à 28 marqueurs microsatellites autosomaux dans 12 populations non-Pygmées et 9 populations Pygmées voisines. Ses résultats ont montré une forte différenciation entre groupes Pygmées de l'ouest de l'Afrique Centrale, ainsi que des flux de gènes asymétriques des non-Pygmées vers les Pygmées [P16]. En se basant sur des méthodes de calcul bayésien approché (*Approximate Bayesian Computation* ou ABC, voir Beaumont *et al.* 2002), Paul Verdu a montré que le scénario évolutif le plus probable du peuplement en Afrique Centrale suppose l'existence d'une population ancestrale Pygmée, qui se serait diversi-

fiée il y a environ 2 800 ans, en même temps que l'expansion des agriculteurs non-Pygmées. Un isolement récent a alors vraisemblablement conduit à une différenciation génétique rapide et substantielle entre les populations Pygmées de l'Ouest de l'Afrique Centrale [P16].

Dispersion limitée chez un peuple mobile

Les chasseurs-cueilleurs Pygmées d'Afrique Centrale sont généralement considérés comme étant extrêmement mobiles, et nous voulions donc tester l'hypothèse selon laquelle ce comportement de mobilité se traduisait par une absence de corrélation entre distances génétiques et géographiques. La dispersion limitée des individus se traduit en effet (dans un habitat à deux dimensions) par une relation linéaire entre les distances génétiques et le logarithme des distances géographiques entre paires d'individus (Rousset 2000). En considérant 87 individus Baka, échantillonnés dans trois groupes distincts et génotypés à 28 marqueurs microsatellites, nous avons calculé le moment d'ordre deux de la distance entre parents et descendants par la méthode proposée par Rousset (1997, 2000). Nous avons trouvé un très fort signal d'isolement par la distance, dû à une dispersion parents-enfants limitée [P23]. Bien qu'ils ne remettent pas en cause le fait que les Pygmées puissent avoir des mouvements fréquents dans leur aire d'activité socio-économique, nos résultats montrent qu'une forte mobilité individuelle ne reflète pas nécessairement de forte dispersion efficace entre générations. Cette dispersion efficace limitée est un facteur qui a pu contribuer à l'isolement génétique entre populations Pygmées, ce qui pourrait constituer un mécanisme clé de la forte différenciation observée entre populations Pygmées de l'ouest de l'Afrique Centrale, malgré leur divergence récente il y a 2 800 ans (voir ci-dessus). De façon intéressante, ces résultats sont assez proches de ceux obtenus chez des horticulteurs Papous de langue Gainj et Kalam en Nouvelle Guinée (Rousset 1997). Ceci suggère que, bien qu'ils soient plus mobiles, les chasseurs-cueilleurs ne dispersent pour autant pas plus efficacement que les agriculteurs. La publication [P23] qui a découlé de ce travail est reproduite en Annexe D, page 147.

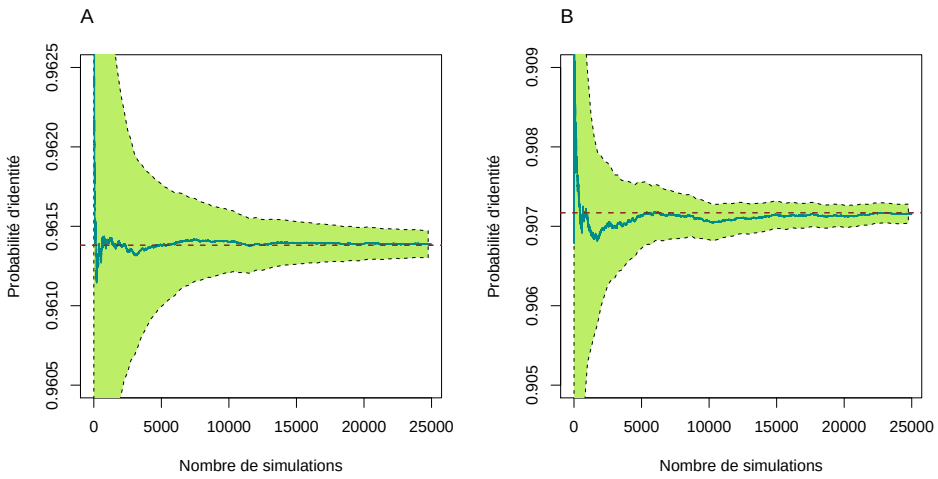


Figure 1.8: Évaluation des simulations de coalescence génération par génération, dans un modèles en îles avec les paramètres : $N_{♂} = 5$, $N_{♀} = 5$, $m_{♂} = 0.001$, $m_{♀} = 0.01$, $n_{\mathrm{d}} = 2$ et $\mu = 0.0005$. (A) Moyenne cumulée et intervalle de confiance à 95% de la probabilité d'identité entre deux gènes autosomaux, échantillonnés dans deux mâles d'un même dème. (B) Moyenne cumulée et intervalle de confiance à 95% de la probabilité d'identité entre deux gènes liés au chromosome X, échantillonnés dans deux mâles de dèmes différents. Dans les deux graphes, la valeur attendue, calculée analytiquement, est indiquée par un trait pointillé.

1.2.2 Inférence de la dispersion biaisée par ABC

Toutes les méthodes que je viens d'évoquer reposent sur le calcul de statistiques sommaires et sur l'inférence de paramètres d'intérêt par de simples méthodes des moments. Très souvent, malheureusement, la complexité des modèles ne permet pas l'évaluation explicite de la vraisemblance. Dans ce cas, le calcul bayésien approché (*Approximate Bayesian Computation* ou ABC) a été proposé comme une alternative possible (Beaumont 2010; Beaumont *et al.* 2002). Cette méthode, évoquée dans le contexte de la génétique des populations par Tavaré *et al.* (1997) et Pritchard *et al.* (1999) puis développée par Beaumont *et al.* (2002), consiste à remplacer le calcul de la vraisemblance par le calcul d'un critère de similarité entre les données ob-

servées et des données simulées sous un modèle donné. Cette similarité est mesurée comme une distance euclidienne entre un ensemble de statistiques sommaires calculées sur les données observées et sur les données simulées. Il n'y a pas encore à ce jour de règle simple pour le choix de ces statistiques et leur choix est souvent le fruit de recherches empiriques préalables. L'échantillonnage ne s'effectue donc plus dans la vraisemblance, mais dans la distribution conjointe des valeurs de paramètres et des statistiques sommaires. La distribution *a posteriori* des paramètres est obtenue à partir de l'estimation de cette distribution conjointe. L'avantage de cette méthode est qu'elle ne nécessite à aucun moment d'expression analytique de la vraisemblance : il est seulement nécessaire de pouvoir simuler des valeurs. Ces algorithmes sont donc extrêmement flexibles car ils peuvent s'appliquer à n'importe quel modèle, du moment que l'on peut simuler efficacement des données sous ce modèle, ce qui est tout à fait possible dans le cadre de la théorie de la coalescence (Kingman 1982a,b). Les valeurs de paramètres sont tirées dans des distributions *a priori* et des simulations sont réalisées selon le modèle, à partir desquelles des statistiques sommaires sont calculées. Une régression linéaire locale est ensuite calculée autour de la valeur des statistiques observées pour obtenir les distributions *a posteriori* des paramètres. Ce type d'approche a déjà prouvé son efficacité en génétique des populations pour la comparaison de scénarios évolutifs complexes et pour l'inférence de paramètres démographiques (voir, par exemple Cornuet *et al.* 2008; Fagundes *et al.* 2007).

Durant le stage de Master 2 de Camille Madec, nous avons cherché à mettre au point une nouvelle méthode d'analyse du polymorphisme, conjointement sur différents types de marqueurs (chromosomes X, Y, ADN mitochondrial et autosomes, dans un modèle "animal" à sexes séparés) afin d'estimer, dans le cadre du modèle d'isolement par la distance, les caractéristiques de dispersion propres à chaque sexe. Dans le cadre d'une collaboration avec avec Raphaël Leblois (CBGP), j'ai développé un logiciel (IbdSex [L7]) pour simuler des données dans un modèle d'isolement par la distance pour différentes catégories de marqueurs génétiques (chromosomes X, Y, ADN mitochondrial, autosomes, etc.) [p5]. Ce logiciel permet également de simuler la dispersion des graines et du pollen chez les plantes, en considérant des

marqueurs autosomaux et des marqueurs liés à des organelles à hérédité paternelle ou maternelle. Ce modèle de simulation repose sur un algorithme de coalescence "génération par génération" (voir, par exemple, Leblois *et al.* 2003), qui permet de s'affranchir des hypothèses classiques de la coalescence (figure 1.8). L'évaluation, par simulation, de la performance statistique de la méthode est encourageante, mais un effort important est encore nécessaire pour réduire les temps de calcul, avant de pouvoir ré-analyser, par exemple, les populations humaines d'Asie Centrale étudiées par Laure Ségurel dans le cadre de sa thèse.

1.3 Inférer les changements d'effectifs passés

1.3.1 Le contexte

Les effectifs des populations ne sont en général pas stables dans le temps : la plupart des populations naturelles peuvent en effet avoir traversé un goulet d'étranglement démographique ou, au contraire, un épisode d'expansion. Reconstruire l'histoire démographique des populations est donc un enjeu important en biologie évolutive, par exemple pour mieux comprendre l'impact des changements climatiques sur la distribution des espèces (Hu *et al.* 2009), ou pour la conservation des espèces menacées (Frankham *et al.* 2002). Pour cela, les approches indirectes de génétique des populations, qui permettent d'inférer la démographie passée à partir de la distribution actuelle de la variation génétique, constituent une alternative aux approches directes reposant sur des suivis démographiques (Lawton-Rauh 2008). La méthode la plus couramment utilisée dans la littérature pour détecter et dater des évènements démographiques passés de contraction ou d'expansion de populations est celle développée par Beaumont (1999) et Storz et Beaumont (2002), et implémentée dans le logiciel Msvar. Ce modèle considère une population d'effectif efficace actuel N_0, dont l'effectif efficace stable dans le passé était égal à N_1. Le changement de taille se fait de façon linéaire ou exponentielle, et a débuté T_a générations dans le passé (figure 1.9).

La méthode statistique sur laquelle repose Msvar utilise une approche

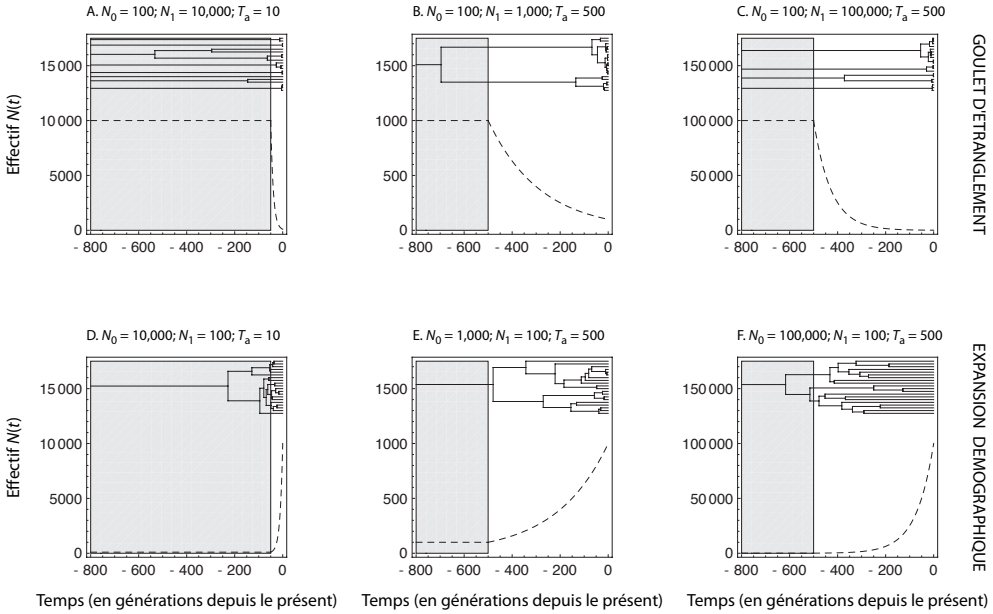


Figure 1.9: Dynamiques temporelles des changements d'effectifs $N(t)$ et généalogies de gènes correspondantes. (A)–(C) correspondent à des goulets d'étranglement démographiques et (D)–(F) à des expansions démographiques (voir les courbes en pointillés). La région grisée dans chaque graphe indique la période pendant laquelle l'effectif efficace ancestral est stable, soit $N(t) = N_1$. Au dessus de chaque courbe, une généalogie de 20 gènes est représentée, dont les longueurs de branche sont égales aux moyennes observées sur 500 000 simulations de chaque scénario.

basée sur le calcul de la vraisemblance des échantillons, c'est-à-dire sur le calcul de la probabilité d'observer les données (des comptages alléliques sur des marqueurs microsatellites) étant données les valeurs des paramètres du modèle démographique. Comme les surfaces de vraisemblance sont complexes, des méthodes ont été développées afin d'échantillonner dans ces surfaces. Une de ces méthodes, extrêmement utilisée en génétique des populations (voir Nielsen et Wakeley 2001, par exemple), est la méthode des chaînes de Markov Monte Carlo (*Markov Chain Monte Carlo*, MCMC). Celle-ci n'est utilisable que lorsqu'un calcul numérique explicite de la vraisemblance est possible, ce qui en limite la portée. Le principe de cette méthode est la

construction d'une chaîne de Markov par l'exploration sélective des paramètres, par exemple grâce à l'algorithme de Metropolis–Hastings (Hastings 1970; Metropolis *et al.* 1953). Cet algorithme explore l'espace des paramètres jusqu'à arriver à une situation stationnaire à partir de laquelle on peut estimer la distribution *a posteriori* des paramètres, proportionnelle au produit de la vraisemblance et des distributions *a priori* des paramètres. Dans le cas du logiciel Msvar, le calcul de la vraisemblance se fait conditionnellement à la généalogie de l'échantillon, grâce à la théorie de la coalescence (Kingman 1982a,b) qui offre un modèle probabiliste des généalogies de gènes dans les populations. Ce type d'approche présente l'inconvénient d'être très lourd en temps de calcul, et présente souvent des problèmes de convergence. Or la performance statistique de cette méthode n'avait jamais été évaluée autrement que par quelques analyses de jeux de données simulées (Beaumont 1999). Dans le cadre de la thèse de Christophe Girod (dirigée par Hélène Fréville et Bernard Riéra), nous avons donc étudié la performance statistique de l'approche de Beaumont (1999), en simulant des jeux de données génétiques selon différents scénarios démographiques [P28].

1.3.2 Une évaluation de la méthode Msvar

Nous avons montré que la méthode Msvar permettait de mieux détecter les évènements d'expansion que de contraction démographique, et que la puissance de détection des évènements très récents est réduite. Ce résultat n'est pas surprenant en soi : si des changements de taille se sont produits très récemment, la partie de la généalogie des gènes correspondant à cette histoire récente ne contient pas ou peu d'évènements de coalescence et/ou de mutations ; en d'autres termes, les données ne contiennent pas ou peu d'information sur cette histoire récente (voir, pour illustration, la figure 1.9). De façon générale, Msvar permet de bien caractériser des événements d'expansion ou de contraction démographiques d'intensité modérée à forte. En revanche, les paramètres du modèle ne sont pas estimés de façon précise pour les expansions démographiques, et l'estimation du temps depuis le changement de taille est biaisée dans le cas des expansions comme dans celui des

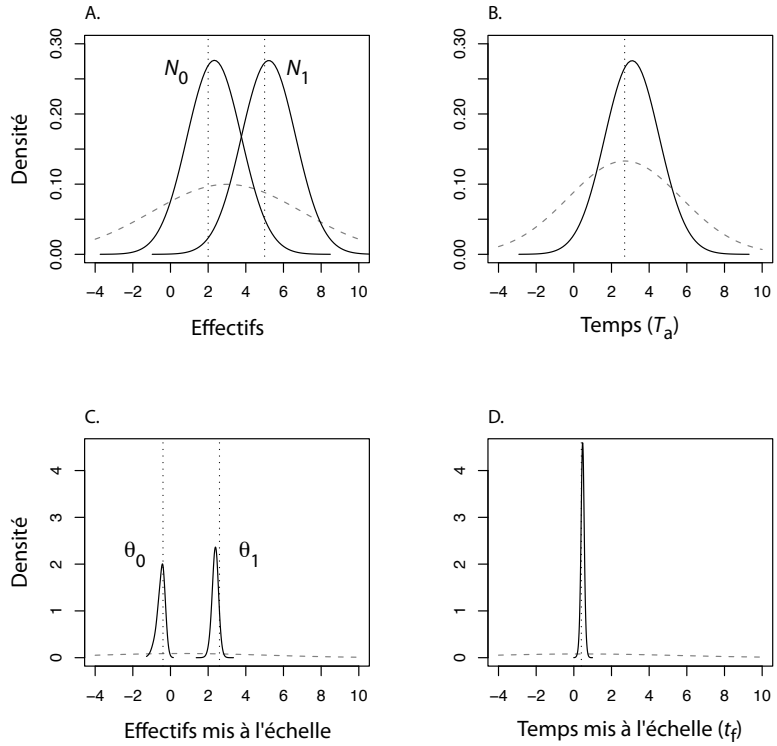


Figure 1.10: Densité marginale *a posteriori* des paramètres pour un goulet d'étranglement sévère ($N_0 = 100$, $N_1 = 100\ 000$) et ancien ($T_a = 500$). Toutes les densités sont représentées sur une échelle $\log_{10}$. (A) Effectifs présents et passés de la population, exprimés en paramètres "naturels" N_0 et N_1. (B) Temps écoulé depuis le changement démographique, exprimé par le paramètre "naturel" T_a. (C) Comme en (A), mais exprimés en paramètres "mis à l'échelle" : $\theta_0 \equiv 4N_0\mu$ et $\theta_1 \equiv 4N_1\mu$. (D) Comme en (B), mais exprimé par le paramètre "mis à l'échelle" $t_f \equiv T_a/(2N_0)$. Les vraies valeurs des paramètres sont indiquées par une ligne en pointillés sur chaque graphe. Les distributions *a priori* de chaque paramètre sont représentées par une ligne hachée grise.

contractions. Enfin, nos résultats montrent que les paramètres naturels du modèle (N_0, N_1 et T_a) sont de manière générale mal estimés, et qu'il est donc préférable de raisonner sur les paramètres mis à l'échelle : $\theta_0 \equiv 4N_0\mu$, $\theta_1 \equiv 4N_1\mu$ et $t_f \equiv T_a/(2N_0)$, où μ est le taux de mutation (voir la figure 1.10).

Ce dernier point illustre le fait que la plupart des modèles en génétique des populations dépendent de paramètres mis à l'échelle plutôt que de paramètres naturels. Estimer une taille de population, ou un temps exprimé en nombre de générations, n'est donc possible qu'à condition de spécifier une distribution *a priori* informative, par exemple sur le taux de mutation μ. La publication [P28] qui a découlé de ce travail est reproduite en Annexe E, page 153.

1.4 Retracer l'histoire des populations

1.4.1 Contexte

Une façon commode de représenter l'histoire démographique des populations peut être empruntée à l'analyse phylogénétique (Felsenstein 2003). Cette représentation repose sur l'idée selon laquelle les relations historiques entre les populations peuvent être représentées sous la forme d'un arbre. Les nœuds terminaux (encore appelés "feuilles") de l'arborescence représentent les populations échantillonnées, alors que les nœuds internes sont interprétés comme des populations ancestrales (non observées). Les longueurs de branche entre deux nœuds quelconques sont proportionnelles à la divergence génétique entre les populations. Les premières tentatives pour caractériser des arbres de populations reposaient sur des méthodes des moments pour reconstruire une topologie et estimer les longueurs de branche (Saitou et Nei 1987). Or, en principe, les techniques reposant sur le calcul de la vraisemblance (Edwards 1992) sont plus efficaces en ceci qu'elles permettent d'utiliser l'ensemble des informations contenues dans les données génétiques. Ces techniques nécessitent de définir, par le biais d'un modèle stochastique, la probabilité d'un échantillon de gènes, en fonction des paramètres qui caractérisent la topologie et les longueurs de branche. Pour calculer cette probabilité, deux familles de modèles ont été utilisées, selon que le phénomène de dérive est considéré en "remontant le temps" (théorie de la coalescence) ou "dans le sens du temps" (théorie de la diffusion).

La première approche repose donc sur le principe de la théorie de la coalescence (Kingman 1982a,b) qui fournit le cadre probabiliste pour calculer

la probabilité d'un échantillon de gènes, conditionnellement à la généalogie (inconnue) de l'échantillon (Hein *et al.* 2005; Wakeley 2008). Cette vraisemblance ne peut donc être calculée que pour une seule histoire généalogique, sachant qu'un grand nombre d'histoires généalogiques est en général compatible avec les données. Par conséquent, des algorithmes ont été développés pour intégrer le calcul de la vraisemblance sur l'espace des généalogies possibles (Hey et Nielsen 2004, 2007). Ces algorithmes sont en général sujets à des problèmes de convergence, et ceci d'autant plus que la taille de l'échantillon augmente et que les scénarios considérés sont complexes.

La deuxième approche repose sur la théorie de la diffusion (Crow et Kimura 1971) qui permet de ramener le processus discret de la dérive génétique à un processus de diffusion, en temps continu. L'application de la théorie de la diffusion à des modèles simples de génétique des populations permet de calculer une densité de probabilité des fréquences alléliques dans un échantillon de gènes (Ewens 2004; Kimura 1964), qui peut être alors utilisée pour calculer la vraisemblance de cet échantillon. C'est dans le cadre de cette seconde approche que nous collaborons avec Mathieu Gautier (CBGP), dans le but de développer une méthode pour retracer l'histoire de divergence des populations, à partir de données de comptages alléliques [p1].

1.4.2 Présentation de la méthode

La méthode que nous proposons repose sur la définition d'un modèle bayésien hiérarchique qui intègre la distribution des fréquences alléliques le long des branches d'un arbre représentant les relations historiques entre les populations échantillonnées. Les paramètres d'intérêt de ce modèle sont les longueurs de chaque branche, qui représentent l'intensité de la dérive, conditionnellement à une topologie donnée. Nous supposons donc connu l'ordre des divergences successives entre les populations. Nous verrons par la suite la stratégie que nous avons adoptée pour confronter différentes histoires possibles.

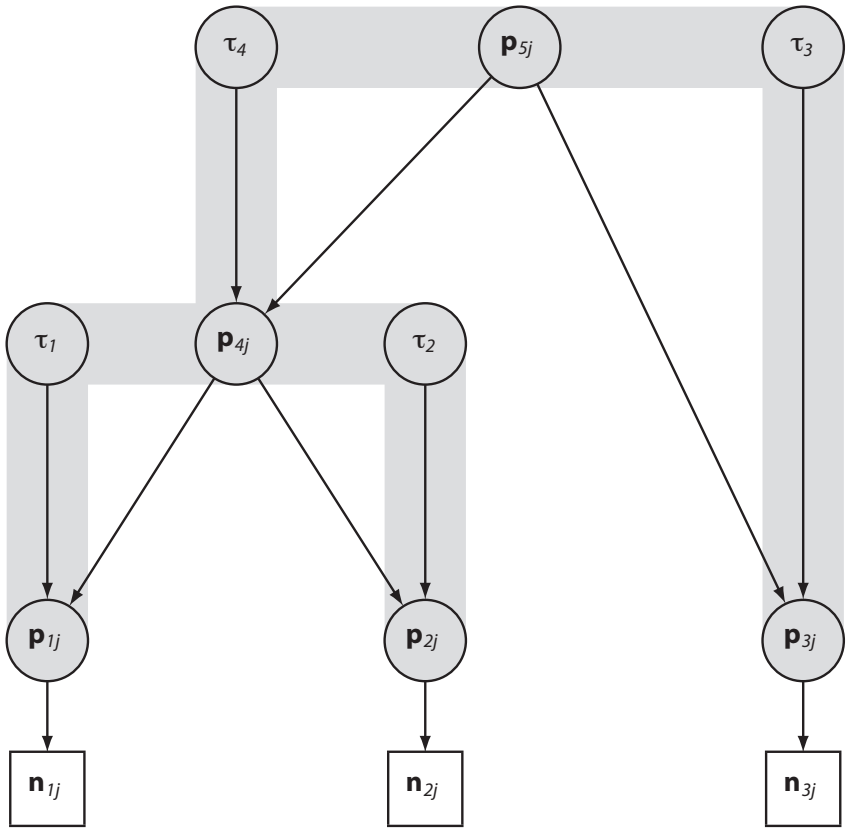


Figure 1.11: Graphe orienté acyclique (DAG) du modèle bayésien hiéarachique d'inférence des temps de divergence, basé sur les équations de diffusion de KIMURA. Une topologie reliant trois populations échantillonnées (numérotées 1, 2 et 3) est représentée en gris. Le graphe orienté acyclique est matérialisé par des flèches reliant entre eux les différents paramètres du modèle (longueurs de branche τ_j et fréquences alléliques $\mathbf{p}_{ij}$) et les données de comptages alléliques $\mathbf{n}_{ij}$.

Le modèle populationnel

Considérons un échantillon de n_{d} populations partageant une histoire commune. Chaque population possède un indice k, qui varie de 1 à n_{d} pour les populations échantillonnées (les "feuilles" de l'arbre), et de $(n_{\mathrm{d}} + 1)$ à r pour les nœuds internes de l'arbre. L'indice r représente donc la population ancestrale de l'ensemble de l'échantillon, positionnée à la racine de l'arbre.

Dans le cas d'un arbre binaire (où seules des bifurcations se produisent) on peut dénombrer $(n_{\mathrm{d}} - 1)$ nœuds internes, et donc $r = 2n_{\mathrm{d}} - 1$. Pour une phylogénie en forme d'étoile, où toutes les populations échantillonnées dérivent simultanément d'une population ancestrale unique, $r = n_{\mathrm{d}} + 1$. Nous notons $a(k)$ la population ancestrale de la population k. Le graphe orienté acyclique (*directed acyclic graph* ou DAG) du modèle est représenté dans la figure 1.11, où les notations sont données à titre indicatif dans le cas particulier d'un arbre binaire de trois populations (où $n_{\mathrm{d}} = 3$ et $r = 5$, avec $a(1) = a(2) = 4$ et $a(3) = a(4) = r = 5$). Les différents du modèle sont les fréquences alléliques $\mathbf{p}_{ij}$ et les longueurs de branches τ_j. Ces longueurs de branches τ_j sont les paramètres d'intérêt du modèle et représentent l'intensité de la dérive le long de chaque branche.

Les données

Les données sont des comptages alléliques, réalisés sur un ensemble d'individus échantillonnés parmi n_{d} populations, et génotypés à L locus. Nous considérons des marqueurs strictement bi-alléliques et notons A et a les deux allèles présents à chaque locus. On note p_{ij} la fréquence de l'allèle A dans la population i au locus j. On note n_{ij} le nombre total de gènes échantillonnés dans la $i$ème population au $j$ème locus, dont x_{ij} possèdent l'état allélique A. Le vecteur des comptages alléliques dans la population i au locus j est donc $\mathbf{n}_{ij} \equiv (x_{ij}, n_{ij} - x_{ij})$. Étant données les fréquences alléliques p_{ij} de l'allèle A, la distribution conditionnelle des comptages alléliques $\mathbf{n}_{ij}$ dans la population i au locus j est binomiale :

$$\mathcal{L}(p_{ij}; \mathbf{n}_{ij}) = \binom{n_{ij}}{x_{ij}} p_{ij}^{x_{ij}} (1 - p_{ij})^{n_{ij} - x_{ij}} \tag{1.8}$$

Considérons maintenant le deuxième niveau du modèle hiérarchique défini par le modèle illustré dans la figure 1.11, correspondant à la distribution de la fréquence p_{kj} de l'allèle A au $j$ème locus dans la population k $(k < r)$. En l'absence de mutation, en supposant que la population k a divergé de $a(k)$ pendant t_k générations discrètes et non-chevauchantes et sous l'hypothèse

que la taille efficace de la population N_k est restée constante au cours de la divergence, alors la distribution de p_{kj}, conditionnellement à la fréquence $p_{a(k)j}$ de l'allèle A dans la population parentale et à la longueur de branche $\tau_k \equiv t_k/(2N_k)$, est donnée par (pour $0 < p_{kj} < 1$) :

$$f(p_{kj} \mid p_{a(k)j}, \tau_k) = (1 - w_{kj}^2) \sum_{l=1}^{\infty} \frac{2l+1}{l(l+1)} T_{l-1}^1(w_{kj}) T_{l-1}^1(z_j) e^{-\frac{1}{2}l(l+1)\tau_k} \quad (1.9)$$

où $w_{kj} = 1-2p_{kj}$, $z_j = 1-2p_{a(k)j}$, et $T_{l-1}^1(x)$ représente le polynome de Gegenbauer qui peut être calculé par la formule récursive : $T_0^1(x) = 1$, $T_1^1(x) = 3x$, ..., et $T_n^1(x) = \left[(2x(n+\frac{1}{2})T_{n-1}^1(x) - (n+1)T_{n-2}^1(x)\right]/n$, pour $n \geq 2$. En ce qui concerne les masses en 0 et 1, la probabilité conditionnelle que $p_{kj} = 0$ est donnée par :

$$\begin{aligned} \mathbb{P}(p_{kj} = 0 \mid p_{a(k)j}, \tau_k) &= (1 - p_{a(k)j}) \\ &+ \frac{(1-z_j)^2}{2} \sum_{L=1}^{\infty} (-1)^l \frac{2l+1}{l(l+1)} T_{l-1}^1(-z_j) e^{-\frac{1}{2}l(l+1)\tau_k} \end{aligned} \quad (1.10)$$

et la probabilité conditionnelle que $p_{kj} = 1$ est donnée par :

$$\begin{aligned} \mathbb{P}(p_{kj} = 1 \mid p_{a(k)j}, \tau_k) &= p_{a(k)j} \\ &+ \frac{(1-z_j)^2}{2} \sum_{L=1}^{\infty} (-1)^l \frac{2l+1}{l(l+1)} T_{l-1}^1(z_j) e^{-\frac{1}{2}l(l+1)\tau_k} \end{aligned} \quad (1.11)$$

(voir les formules 4.9 et 4.16 dans Kimura 1964).

Pour la population ancestrale de l'échantillon complet ($k = r$), nous supposons que la distribution *a priori* des fréquences p_{rj} de l'allèle A au locus j suit une loi beta :

$$p_{rj} \sim \text{beta}(0.7, 0.7). \quad (1.12)$$

Enfin, on suppose que les longueurs de branche τ_k suivent une distribution *a priori* uniforme :

$$\tau_k \sim \mathcal{U}(0, 10) \quad (1.13)$$

1.4.3 Inférence des paramètres

L'objectif est d'échantillonner les valeurs de paramètres dans leur distribution *a posteriori*, qui est proportionnelle au produit de la vraisemblance et des distributions *a priori*. Si l'on suppose que les fréquences alléliques à différents locus sont indépendantes (équilibre de liaison) et que les fréquences alléliques dans des populations distinctes sont également indépendantes (isolement complet des populations), alors la distribution *a posteriori* du modèle complet décrit ci-dessus (voir aussi la figure 1.11) prend la forme :

$$f\left(\boldsymbol{p}, \boldsymbol{\tau} \mid \boldsymbol{n}\right) \propto \left[\prod_{i=1}^{n_{\mathrm{d}}} \prod_{j=1}^{L} \mathcal{L}(p_{ij}; \mathbf{n}_{ij})\right] \times \left[\prod_{i=1}^{r-1} f\left(\tau_i\right) \prod_{j=1}^{L} f\left(p_{ij} \mid p_{a(i)j}, \tau_i\right)\right] \prod_{j=1}^{L} f\left(p_{rj}\right) \tag{1.14}$$

Nous avons implémenté une méthode d'échantillonnage dans cette distribution, par chaîne de Markov Monte Carlo en utilisant l'algorithme de Metropolis–Hastings (voir, par exemple, Gelman *et al.* 2004). Cet algorithme peut être décrit de la façon suivante, en supposant que l'on parte d'une valeur θ du paramètre :

(*i*) on propose une nouvelle valeur du paramètre θ, notée θ^*, selon la loi $q(\theta \to \theta^*)$

(*ii*) on évalue le ratio h défini comme :

$$h = \min\left(1, \frac{\mathcal{L}(\theta^*; \mathcal{D}) f(\theta^*) q(\theta* \to \theta)}{\mathcal{L}(\theta; \mathcal{D}) f(\theta) q(\theta \to \theta^*)}\right)$$

où $\mathcal{L}(\theta; \mathcal{D})$ représente la vraisemblance, et $f(\theta)$ la distribution *a priori* du paramètre θ.

(*iii*) on accepte la valeur θ^* du paramètre avec une probabilité égale à h. Dans ce cas, on remplace la valeur courante du paramètre par sa nouvelle valeur : $\theta = \theta^*$, et l'on va à l'étape (*i*). Sinon, on reste au point θ et l'on va à l'étape (*i*).

L'ensemble de l'algorithme est répété un très grand nombre de fois. À chaque

nouvelle itération, tous les $n_{\mathrm{d}} \times L$ paramètres p_{ij} sont mis à jour (pour $1 < i \leq n_{\mathrm{d}}$), puis tous les $(r - n_{\mathrm{d}}) \times L$ paramètres p_{ij} (pour $n_{\mathrm{d}} < i \leq r$), puis tous les L paramètres p_{rj}, et enfin tous les paramètres $(r-1)\tau_i$.

Étant donné que la topologie de l'arbre nous est généralement inconnue, nous avons cherché à définir un critère pour tester la pertinence relative de différentes histoires alternatives. Pour cela, nous avons utilisé le critère d'information basé sur la déviance (*Deviance Information Criterion* ou DIC), un outil classique pour sélectionner un modèle parmi un ensemble fini de modèles possibles (voir, par exemple, Gelman *et al.* 2004). Dans ce cas, chaque modèle est défini par une topologie particulière (représentant la succession des divergences entre populations).

1.4.4 Évaluation sur des données simulées

Afin de tester la performance de ce modèle, nous avons effectué des simulations stochastiques en utilisant l'algorithme de coalescence implémenté dans le logiciel *ms* de Hudson (2002). Nous avons simulé l'histoire de trois populations dont la topologie peut s'écrire ((P1,P2),P3) au format Newick, et avons spécifié différents ensembles de longueurs de branche (figure 1.12). Chaque jeu de données simulées était constitué de 50 individus diploïdes (100 gènes) par population, génotypés à 5 000 SNPs. Cinquante réplicas ont été réalisés pour chaque histoire. Comme illustré dans la figure 1.12, notre modèle d'inférence basé sur les équations de diffusion de Kimura produit des estimations précises du temps de divergence, malgré un biais modéré, voire léger, pour la branche interne (conduisant à la population P4 dans les scénarios simulés). Ce biais est d'autant plus prononcé que le temps de divergence de cette branche interne est faible.

1.4.5 Application sur un jeu de données humaines

Nous avons ré-analysé un sous-ensemble des données humaines de Wollstein *et al.* (2010) avec notre méthode. Ce jeu de données est constitué de comptages alléliques à 815 772 SNPs autosomaux génotypés dans quatre populations humaines : les Yorubas d'Afrique occidentale (YRI, $2n = 120$),

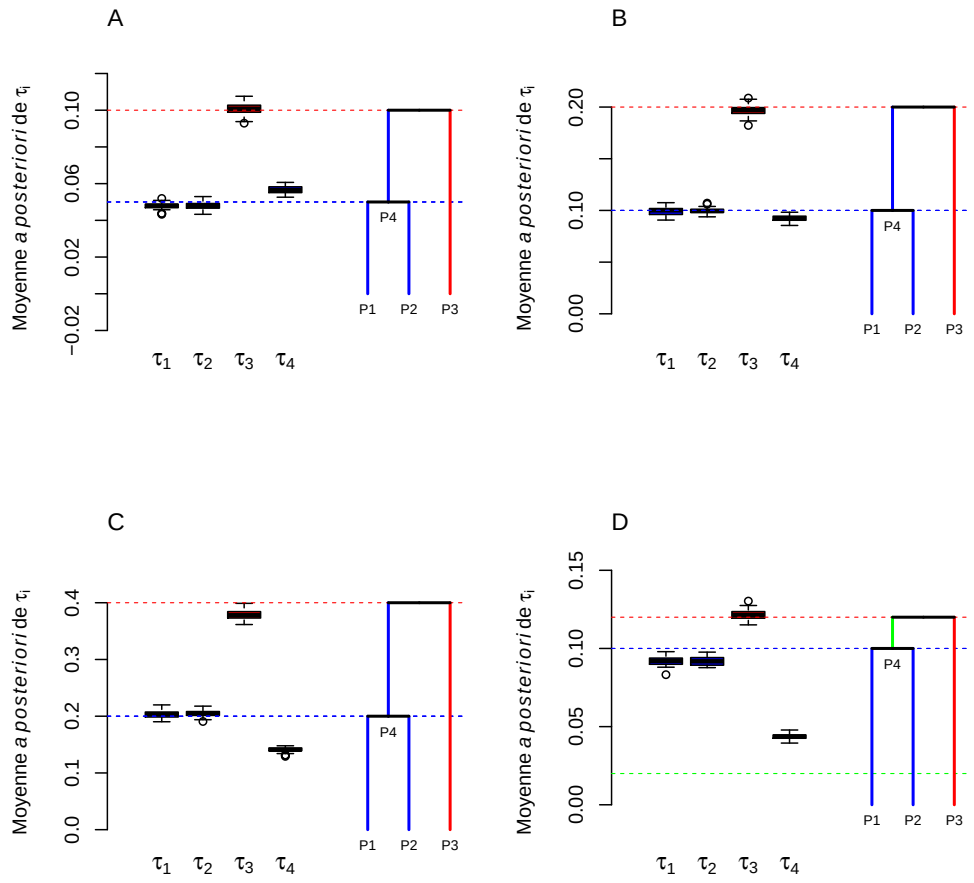


Figure 1.12: Performance du modèle d'inférence des temps de divergence. Pour chaque scénario simulé, 50 jeux de données de 5 000 SNPs ont été analysés. Les quatre scénarios (A), (B), (C) et (D) considèrent la même topologie ((P1,P2),P3), mais des ensembles de longueurs de branche différentes. Chaque scénario est représenté par un arbre à l'échelle, à droite de chaque graphe. Les distributions des moyennes *a posteriori* des longueurs τ_i de chaque branche sont représentées, ainsi que les vraies valeurs simulées (lignes horizontales pointillées).

les Américains des États-Unis d'ascendance européenne (CEU, $2n = 120$), les chinois Han de Pékin (CHB, $2n = 90$) et les "Highlanders" de Nouvelle-

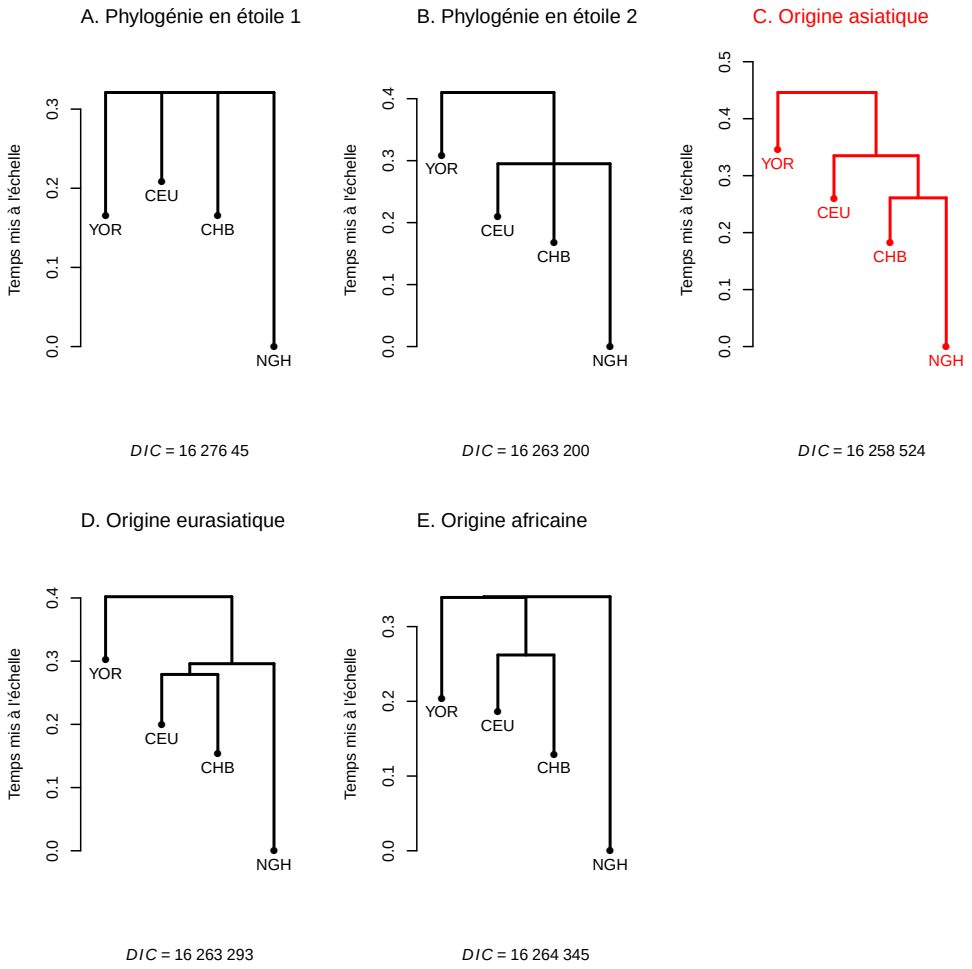


Figure 1.13: Inférence des temps de divergence pour différents scénarios évolutifs possibles de l'histoire des populations humaines YRI, CEU, CHB et NGH. Chaque analyse (conditionnellement à chaque histoire évolutive considérée ici) a été réalisée à partir des données de Wollstein *et al.* (2010), incluant 815 722 SNPs. Les valeurs de DIC sont indiquées sous chaque arbre. L'origine asiatique des Highlanders de Nouvelle-Guinée (NGH), représentée en rouge, est fortement soutenue par ce critère.

Guinée (NGH, $2n = 48$), qui sont les habitants de la région montagneuse du centre de la Papouasie–Nouvelle–Guinée.

Nous avons réalisé cinq analyses, conditionnellement à cinq topologies

possibles, dont trois avaient été analysées par Wollstein *et al.* (2010) dans leur étude originale (voir la figure 1.13). Wollstein *et al.* (2010) avaient en effet utilisé le calcul bayésien approché (Beaumont *et al.* 2002) pour tester plusieurs scénarios de l'histoire de ces populations. Sur la base du critère DIC, notre ré-analyse des données de Wollstein *et al.* (2010) place l'origine de la population de Nouvelle-Guinée en Asie de l'Est. Ce résultat contredit la propre analyse de Wollstein *et al.* (2010), qui concluait à une divergence précoce de la population de Nouvelle-Guinée à partir d'une population euro-asiatique ancestrale. Cependant, et comme souligné par Wollstein *et al.* (2010), leurs propres analyses apportaient également un "soutien appréciable" pour une origine de la population de Nouvelle-Guinée en Asie de l'Est.

Chapitre 2

Histoire adaptative des populations

DANS le chapitre précédent, je me suis uniquement intéressé à la variation génétique neutre, car les méthodes d'inférence de paramètres démographiques à partir de données de polymorphisme génétique reposent sur l'idée que seule la variation neutre peut nous renseigner sur l'histoire des populations (notamment leur taille, l'intensité des flux de gènes qu'elles échangent, etc.), tandis que la variation sélectionnée ne peut nous renseigner que sur l'histoire d'un caractère en particulier. Le développement rapide des technologies de séquençage et de génotypage à haut débit permet désormais l'accès facilité et à moindre coût à une grande quantité d'information sur la diversité génétique des populations, y compris chez des espèces pour lesquelles une connaissance fine de la structure du génome n'est pas encore disponible. Dans ce contexte, il devient désormais possible de rechercher des marqueurs moléculaires portant des signatures de sélection (qu'elle soit naturelle ou artificielle), notamment pour comprendre la dynamique de l'adaptation, ou la réponse à la sélection dans les programmes d'amélioration. Dans cette perspective, comment caractériser les régions du génome répondant à l'action de la sélection ? Comment quantifier l'intensité de la sélection dans ces régions ? N'importe quelle forme de sélection affecte certaines régions du génome plus que d'autres, tandis que l'histoire des populations, leur démographie, ou bien

encore leur structure spatiale affectent l'ensemble du génome de la même manière (Cavalli-Sforza 1966). Par conséquent, en estimant la distribution attendue du polymorphisme neutre conditionnellement aux données observées, il est possible d'identifier des marqueurs incompatibles avec l'hypothèse de neutralité (Beaumont et Nichols 1996; Bowcock *et al.* 1991; Lewontin et Krakauer 1973; Vitalis *et al.* 2001).

2.1 Détecter les signatures de sélection

2.1.1 Principe

Détecter des signatures de sélection revient *in fine* à distinguer parmi les forces agissant sur l'évolution des fréquences alléliques celles qui ont un effet global sur l'ensemble des marqueurs (migration et dérive) de celles qui ont un effet local (sélection et mutation). La plupart des méthodes actuelles appliquées aux données pangénomiques repose sur l'étude de statistiques descriptives résumant la variabilité de chaque locus au niveau intra et/ou inter populationnel (F_{ST}, homozygotie haplotypique étendue ou *EHH*, etc.) pour identifier les locus qui présentent des valeurs "atypiques" sous l'hypothèse de neutralité (Beaumont et Nichols 1996; Nielsen 2005; Storz 2005; Vitalis *et al.* 2001). Les distributions de ces statistiques sous l'hypothèse de neutralité sont alors estimées (*i*) à partir de données simulées (ce qui impose le recours à des modèles démographiques généralement simples et peu réalistes, voir Bowcock *et al.* 1991) ; (*ii*) directement à partir des données réelles (en supposant alors qu'une grande proportion des marqueurs analysés est neutre, voir Akey *et al.* 2002; Flori *et al.* 2009; Weir *et al.* 2005), ce qui rend impossible le contrôle du taux de faux positifs et de faux négatifs détectés.

Dans ce contexte, j'ai développé en 2001 un modèle décrivant la coalescence de gènes neutres dans deux populations en divergence [P9]. En définissant des paramètres de différenciation spécifiques à chacune des populations (Weir et Hill 2002) dont la valeur ne dépend que du rapport entre le temps de divergence et la taille de la population, j'ai pu montrer que la

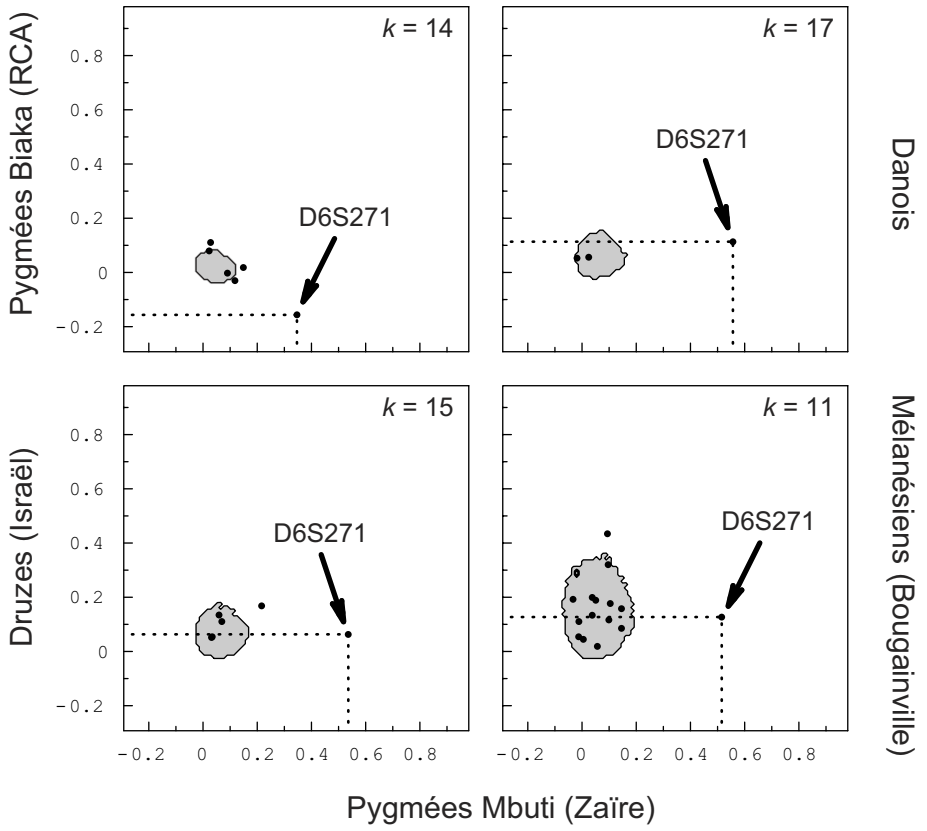


Figure 2.1: Application de la méthode de détection de marqueurs génétiques soumis à la sélection, sur un jeu de données de polymorphisme humain mesuré à 85 locus microsatellites. Dans chaque graphe, la région grisée représente 95% de la distribution conjointe des mesures de divergence pour chacune des populations, conditionnellement aux estimations multi-locus des paramètres de divergence et au nombre d'allèles dans l'échantillon total (seules sont représentées les distributions conditionnelles pour le nombre k d'allèles indiqué dans chaque graphe). Les points noirs représentent les estimations conjointes des mesures de divergence observées. Un locus particulier (D6S271) est systématiquement plus différencié qu'attendu sous l'hypothèse de neutralité dans la population de Pygmées Mbuti.

distribution conjointe des estimateurs de ces paramètres était peu sensible aux paramètres de nuisance du modèle (scénario démographique précédent la divergence, taux de mutation). Ce résultat m'a donc amené à proposer, à la manière de Beaumont et Nichols (1996), une méthode d'identification de

marqueurs génétiques soumis à la sélection [P9]. Cette approche repose sur (*i*) l'estimation des paramètres du modèle (le rapport du temps de divergence sur la taille des populations) à partir de données (réelles) de polymorphisme observé ; (*ii*) le calcul, à partir de simulations stochastiques du processus de coalescence, de la distribution attendue sous un modèle neutre de ces estimateurs ; (*iii*) l'identification des locus pour lesquels la valeur estimée des paramètres s'écarte de la distribution attendue sous le modèle neutre. Ainsi, n'importe quel locus dont le polymorphisme observé s'écarte de la distribution attendue est potentiellement sous l'influence d'un processus sélectif. Parce qu'elle repose sur l'analyse de paires de populations, cette méthode permet d'identifier des signatures de sélection *locale* sur un marqueur, contrairement aux autres méthodes existantes. Ainsi, certains marqueurs peuvent montrer un polymorphisme réduit, à cause de la sélection, dans certaines populations mais pas dans d'autres : voir l'exemple, Fig. 2.1. J'ai développé une interface logicielle conviviale (DetSel) pour appliquer cette méthode [P12]. Malgré leurs limites, DetSel tout comme Fdist (Beaumont et Nichols 1996) ont connu un certain succès. Dans la suite, je détaillerai donc quelques unes des applications qui ont été faites de ces méthodes. La première de ces applications (l'adaptation au régime alimentaire chez l'Homme) concerne un projet qui a été développé dans le cadre de la thèse de Laure Ségurel.

2.1.2 Un exemple : l'adaptation au régime alimentaire chez l'Homme

L'un des plus grands défis de la génétique évolutive humaine consiste à identifier des changements spécifiques dans les gènes qui sont à la base de l'évolution de traits biologiques. En comparaison à d'autres espèces, ces changements peuvent être catalysés et accélérés chez l'Homme, par le phénomène de transmission culturelle (Heyer *et al.* 2005). L'histoire de l'Homme est ponctuée de plusieurs changements importants, bien identifiés, où les innovations culturelles ont joué un rôle majeur. Par exemple, l'émergence de l'agriculture dans les sociétés humaines, apparue pour la première fois il y a environ 10 000 ans, s'est accompagnée d'un grand nombre de changements

culturels et démographiques. Ramenée aux 100 000 ans de l'évolution de l'Homme moderne, la transition d'un mode de vie de chasseurs-cueilleurs à un mode de vie d'agriculteurs sédentaires a été extrêmement rapide. Ce changement rapide a certainement été rendu possible par une transition culturelle majeure : la *révolution* du Néolithique. Certaines populations se sont alors spécialisées dans l'agriculture, tandis que d'autres populations se sont specialisées dans l'élevage. Entre autres choses, le régime alimentaire a été largement modifié pendant cette période, avec une augmentation significative de l'apport en céréales pour les agriculteurs, et un apport en viande accru pour les éleveurs. On pense ainsi qu'un grand nombre de gènes impliqués dans l'alimentation a été soumis à de fortes pressions de sélection lors de ce changement de régime alimentaire et que les adaptions nutritionnelles et métaboliques résultantes ont été importantes. Il existe des témoins indirects de ces adaptations passées : l'augmentation de la prévalence du diabète, de l'obésité, de l'hypertension, etc., associée au changement récent de mode de vie dans de nombreuses sociétés actuelles. Il est en effet admis que ces maladies, associées à des désordres nutritionnels, reflètent une maladaptation, c'est-à-dire une faible capacité à métaboliser la nourriture présente dans le régime alimentaire des sociétés modernes. En d'autres termes, l'architecture génétique des gènes de l'alimentation, sélectionnée pendant la transition du Néolithique, pourrait ne pas être adaptée à notre nouvel environnement nutritionnel (voir Ordovas et Corella 2004, pour une revue).

Plusieurs hypothèses ont été avancées pour expliquer quelles ont pu être les cibles de la sélection naturelle par le passé. L'une d'elle est appelée l'hypothèse de la "piste carnivore" (*carnivore connection hypothesis*) et a été proposée par Colagiuri et Brand Miller (2002). Cette hypothèse considère qu'au cours de l'aire glaciaire du Paléolithique (*i*) la viande était prépondérante dans l'alimentation, et (*ii*) l'apport de glucose était faible. En conséquence, afin de maintenir suffisamment de glucose dans le sang (glycémie de 1g/L), l'efficacité de l'insuline, responsable de la consommation et du stockage du glucose, a été contre-sélectionnée. La seconde hypothèse, celle du "génotype économe" (*thrifty genotype hypothesis*), considère que les cycles de disette et d'abondance de nourriture ont favorisé une plus grande capacité à

stocker de l'énergie sous forme de graisses pendant les phases d'abondance pour une utilisation pendant les phases de disette (Neel 1962, 1992). Bien qu'elles reposent sur des mécanismes différents, ces deux hypothèses supposent donc que la résistance à l'insuline (c'est-à-dire l'insensibilisation des récepteurs cellulaires à l'insuline) et la gluconéogénèse ont été favorisées pendant le Paléolithique, mais que les pressions de sélection s'exerçant sur ces gènes ont ensuite probablement diminué fortement dans certaines populations au cours du Néolithique.

En Asie Centrale co-existent des populations traditionnelles d'éleveurs nomades de langues turco-mongoles (Kirghizes) et des populations traditionnelles d'agriculteurs de langues indo-européennes (Tadjiks). C'est donc une zone du monde privilégiée pour comprendre l'impact sur la diversité génétique de ces modes de vie différents. Parmi les gènes candidats, connus pour être impliqués dans la résistance à l'insuline et la gluconéogénèse, se trouvent les gènes impliqués dans le diabète de type 2. Environ 10 gènes associés au diabète de type 2 ont été identifiés à ce jour, par des analyses d'association (Freeman et Cox 2006; Sladek *et al.* 2007). Dans la majeure partie des cas, il s'avère que plusieurs gènes contribuent conjointement à la maladie (diabète de type 2 complexe, polygénique, ou encore multifactoriel). Dans le cadre de la thèse de Laure Ségurel, et à travers le programme ANR Nutgenevol coordonné par évelyne Heyer, nous avons comparé la variation génétique de ces gènes candidats à celle mesurée dans des régions *a priori* neutres du génome. Il s'avère que la différenciation génétique de certains gènes liés au diabète de type 2 s'écarte significativement de l'attendu sous neutralité en Asie Centrale (*LEPR*, *KCNQ1* et *FABP2*), suggérant ainsi des pressions de sélection différentes entre ethnies aux modes de vie contrastés [R2]. Ce résultat tendrait à conforter les hypothèses du génotype économe et/ou de la piste carnivore, selon lesquelles on s'attend à observer une différenciation génétique accrue pour les gènes associés au diabète de type 2 entre groupes aux modes de vie contrastés [R2]. Mais une analyse plus précise basée sur la structure haplotypique de ces gènes révèle que ce sont des variants protecteurs du diabète de type 2 qui sont sélectionnés chez les Kirghizes et les Tadjiks, l'évènement sélectif datant du début du Néolithique : le résultat

principal de nos analyses est en effet que nous n'avons trouvé aucune signature de sélection pour les variants à risque du diabète de type 2. Ce résultat suggère donc que la transition majeure qui a eu lieu au Néolithique a été accompagnée par un nouvel avantage sélectif des variants "non-économes", ce qui n'est prédit par aucune des hypothèses évolutives avancées à ce jour [R2].

Dans le cadre de la thèse de Laure Ségurel nous nous sommes également intéressés au polymorphisme Pro11Leu du gène *AGXT*, associé à une maladie létale (l'hyperoxalurie primitive de type I) causée par le déficit d'une enzyme peroxysomale hépatique, l'alanine-glyoxylate-aminotransférase (AGT). En réalité, la mutation Pro11Leu du gène *AGXT* ne conduit pas à l'absence de l'enzyme AGT mais plutôt à sa re-localisation dans les mitochondries (alors qu'on trouve généralement cette enzyme dans les peroxysomes chez l'Homme). Or chez les mammifères, la localisation de l'AGT varie selon le régime alimentaire : ont trouve généralement l'AGT dans les peroxysomes chez les herbivores et dans les mitochondries chez les carnivores. Puisque la mutation Pro11Leu est en forte fréquence dans certaines populations humaines (de 5 à 20% chez les Caucasiens), il avait donc été proposé que cet allèle devait être avantageux dans des populations ayant une grande proportion de viande dans leur alimentation, grâce à la re-localisation d'une petite partie de l'AGT dans les mitochondries. Ainsi, l'attendu sous cette hypothèse "adaptationniste" était de trouver des fréquences plus fortes de la mutation Pro11Leu dans les populations d'éleveurs que dans celles d'agriculteurs. Or nous avons trouvé au contraire une plus faible fréquence de la mutation Pro11Leu dans les populations d'éleveurs qui consomment le plus de viande, ce qui contredit l'hypothèse d'un avantage sélectif dans ces populations. De plus, en utilisant une version modifiée de la méthode Fdist de Beaumont et Nichols (1996), nous n'avons pas trouvé un niveau de différenciation entre populations significativement plus élevé pour Pro11Leu ($F_{\mathrm{ST}} = 0.025$) que pour des régions présumées neutres du génome ($F_{\mathrm{ST}} = 0.019$, $p = 0.214$). Nos résultats ne supportent donc pas l'hypothèse que la diversité génétique de la mutation Pro11Leu diffère de celle attendue sous neutralité, mais confortent l'idée que des différences importantes de fréquences alléliques entre populations sont souvent compatibles avec des processus démographiques [P19].

2.1.3 Autres applications

Une nouvelle version du logiciel DetSel a été développée [L3]. Cette nouvelle version a été implémentée sous la forme d'un paquetage du logiciel de statistiques R. Cette nouvelle implémentation de DetSel permet le traitement de données bi-alléliques dominantes de type AFLPs (*Amplified Fragment Length Polymorphisms*) [C2]. Cette nouvelle version a été réalisée à l'occasion d'un certain nombre de collaborations : avec Claire-Lise Meyer (qui a réalisé da thèse sous la direction de Vincent Castric et Pierre Saumitou-Laprade, au Laboratoire de Génétique et Évolution des Populations Végétales, Université Lille 1), nous avons analysé des données d'un balayage génomique réalisé sur plusieurs populations d'*Arabidopsis halleri* afin d'identifier des gènes impliqués dans la tolérance aux métaux lourds [P17] ; avec Bénédicte Rhoné (qui a réalisé da thèse sous la direction d'Isabelle Bonnin, dans l'UMR de Génétique Végétale, Le Moulon), nous avons étudié la sélection agissant sur des gènes impliqués dans la date de floraison dans des populations expérimentales de blé [P21] ; avec Diego Ayala (qui a réalisé da thèse sous la direction de Frédéric Simard au laboratoire de Lutte contre les Insectes Nuisibles, Montpellier), nous avons analysé la sélection agissant sur des inversions chromosomiques chez un vecteur de paludisme *Anopheles funestus* [P24] ; avec Afiwa Midamegbe (en Master 2 sous la direction de Réjane Streiff au CBGP), nous avons analysé des données d'un balayage génomique réalisé sur plusieurs espèces de pyrale *Ostrinia spp.*, un insecte ravageur du maïs, pour identifier des gènes liés au choix de la plante-hôte [P26]. Cette étude, initialement réalisée en France, a été complétée par l'analyse de populations chinoises, lors du stage de Master 2 d'Hermine Alexandre sous la direction de Réjane Streiff, au CBGP [p4]. Enfin, la recherche de marqueurs impliqués dans l'adaptation à la plante hôte chez le puceron du pois (*Acyrthosiphon pisum*) fait l'objet de collaborations en cours avec Carole Smadja (ISE-M, Montpellier) [P31] et Jean-Christophe Simon (Institut de Génétique, Environnement et Protection des Plantes, Rennes) [R1], notamment dans le cadre du programme ANR Speciaphid coordonné par Jean-Christophe Simon.

2.2 Mesurer l'intensité de la sélection

La plupart des méthodes existantes pour détecter la sélection dans un jeu de données de polymorphisme, y compris DetSel, reposent sur l'identification de polymorphismes situés dans la zone de rejet d'un modèle nul correspondant à l'hypothèse de neutralité. La sélection n'est donc pas modélisée explicitement dans ces modèles. De plus, ce type d'approche ne prend pas en compte l'ensemble de l'information contenue dans les données, et il existe désormais des méthodes bien plus puissantes, qui reposent sur le calcul de la vraisemblance (voir, par exemple, Beaumont et Balding 2004; Foll et Gaggiotti 2008; Riebler *et al.* 2008).

Dans le cadre d'une collaboration avec Mark Beaumont, Mathieu Gautier et Kevin Dawson, j'ai développé une méthode pour estimer les coefficients de sélection associés à chaque gène dans un jeu de données de polymorphisme dans une population subdivisée. Cette approche prend le contre-pied de la plupart des méthodes existantes, en ceci qu'elle permet de modéliser explicitement l'action de la sélection sur la distribution du polymorphisme. Cette méthode utilise l'approximation de diffusion pour calculer la vraisemblance des données dans un modèle en îles. Étant donné le grand nombre de paramètres de ce modèle, il n'est malheureusement pas possible d'explorer exhaustivement l'espace des paramètres, afin par exemple de maximiser la vraisemblance. Notre modèle d'inférence statistique repose donc sur une approche bayésienne, et sur l'utilisation de chaînes de Markov Monte Carlo (MCMC) pour estimer les paramètres du modèle. Cette approche permet, à partir d'une distribution *a priori* sur les paramètres, de calculer une distribution *a posteriori* sur les paramètres, étant données les fréquences alléliques observées. La construction d'un modèle hiérarchique a permis d'obtenir des résultats encourageants, validés par simulation. Ces résultats sont l'objet d'une publication en cours de préparation [p3] et d'un logiciel (SelEstim) en cours de développement [L8]. La publication [p3] en cours de préparation est reproduite en Annexe F, page 189.

2.2.1 Présentation de la méthode

Le modèle populationnel

Notre méthode repose sur un modèle en îles infini, où le $i$ème dème contient N_i individus diploïdes, et reçoit des immigrants depuis l'ensemble de la population à un taux m_i. Le taux de migration dans le $i$ème dème est défini comme $M_i \equiv 4N_i m_i$. Nous considérons des marqueurs strictement bi-alléliques et notons A et a les deux allèles présents à chaque locus. On note p_{ij} la fréquence de l'allèle A dans le dème i au locus j, et π_j sa fréquence dans l'ensemble de la population (c'est-à-dire la fréquence de A parmi les immigrants). On peut donc définir le vecteur des fréquences alléliques dans le dème i au locus j, comme : $\mathbf{p}_{ij} \equiv (p_{ij}, 1 - p_{ij})$ et le vecteur des fréquences alléliques au locus j parmi les immigrants comme : $\boldsymbol{\pi}_j \equiv (\pi_j, 1 - \pi_j)$. En ce qui concerne la sélection, nous considérons un modèle très simple de sélection génique dans lequel, à chaque locus, l'allèle A procure un avantage sélectif. Comparés aux individus homozygotes aa, les homozygotes AA et les hétérozygotes Aa ont un gain relatif de valeur sélective égal à $1 + s_{ij}$ and $1 + s_{ij}/2$, respectivement. On définit le coefficient de sélection dans le dème i au locus j comme étant $\sigma_{ij} \equiv 2N_i s_{ij}$. Enfin, il est nécessaire de définir une variable indicatrice κ_{ij}, qui prend la valeur $\kappa_{ij} = 0$ si c'est l'allèle A qui est sélectionné, et $\kappa_{ij} = 1$ si c'est l'allèle a qui est sélectionné. Par conséquent, on peut écrire la fréquence de l'allèle sélectionné dans le dème i au locus j comme étant : $\tilde{p}_{ij} \equiv \kappa_{ij}(1 - p_{ij}) + (1 - \kappa_{ij})p_{ij}$.

Les données

Les données sont des comptages alléliques, réalisés sur un ensemble d'individus échantillonnés parmi n_d dèmes, et génotypés à L locus. On note n_{ij} le nombre total de gènes échantillonnés dans le $i$ème dème au $j$ème locus, dont x_{ij} possède l'état allélique A. Le vecteur des comptages alléliques dans le dème i au locus j est donc $\mathbf{n}_{ij} \equiv (x_{ij}, n_{ij} - x_{ij})$.

Le modèle d'inférence

Étant données les fréquences alléliques p_{ij} de l'allèle A, la distribution conditionnelle des comptages alléliques $\mathbf{n}_{ij}$ dans la population i au locus j est binomiale :

$$\mathcal{L}(p_{ij}; \mathbf{n}_{ij}) = \binom{n_{ij}}{x_{ij}} p_{ij}^{x_{ij}} (1 - p_{ij})^{n_{ij} - x_{ij}}. \tag{2.1}$$

Dans la limite de grande taille de population $N_i \to \infty$, et sous l'hypothèse que la sélection et la dérive sont de force comparable (c'est-à-dire que M_i et σ_{ij} ont une limite finie lorsque $N_i \to \infty$), alors la distribution des $\mathbf{p}_{ij}$ peut être approximée par la distribution stationnaire d'un processus de diffusion de la forme suivante :

$$\psi(p_{ij}; M_i, \sigma_{ij}, \kappa_{ij}, \boldsymbol{\pi}_j) = C^{-1} \exp(\sigma_{ij} \tilde{p}_{ij}) p_{ij}^{M_i \pi_j - 1} (1 - p_{ij})^{M_i (1 - \pi_j) - 1} \tag{2.2}$$

(Wright 1935, 1949, 1969) ; voir aussi Barton et Turelli (1989); Bürger (2000); Ethier et Nagylaki (1988). Dans l'équation (2.2), C est une constante d'intégration qui peut être calculée comme :

$$C = {}_1F_1(M_i \tilde{\pi}_{ij}; M_i; \sigma_{ij}) \frac{\Gamma(M_i \pi_j) \Gamma(M_i (1 - \pi_j))}{\Gamma(M_i)} \tag{2.3}$$

où ${}_1F_1(a; b; z)$ est la fonction hypergéometrique confluente, ou fonction de Kummer, (voir, par exemple, Abramowitz et Stegun 1965, p. 504), et où $\tilde{\pi}_{ij} \equiv \kappa_{ij}(1 - \pi_j) + (1 - \kappa_{ij})\pi_j$. Notons qu'en l'absence de sélection, ce modèle se réduit au modèle beta de Beaumont et Balding (2004), repris par Riebler *et al.* (2008) et Foll et Gaggiotti (2008).

Étant donné le modèle spécifié par les équations (2.1) et (2.2), il s'agit maintenant d'évaluer les paramètres d'intérêt du modèle, à savoir $\mathbf{M} \equiv (M_1, \ldots, \ M_i, \ldots, M_{n_\mathrm{d}})$, $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_j, \ldots, \pi_L)$, $\boldsymbol{\sigma} \equiv (\sigma_{11}, \ldots, \sigma_{ij}, \ldots, \sigma_{n_\mathrm{d} L})$ et $\boldsymbol{\kappa} \equiv (\kappa_{11}, \ldots, \kappa_{ij}, \ldots, \kappa_{n_\mathrm{d} L})$, à partir des comptages alléliques $\mathbf{n}$ mesurés sur l'ensemble des dèmes échantillonnés et des locus. Pour cela, nous supposons que les distributions *a priori* des paramètres κ_{ij} suivent une loi de Bernoulli, c'est-à-dire que $\kappa_{ij} \sim$ Bernoulli(0.5), que les distributions *a priori*

des paramètres $\boldsymbol{\pi}_j$ sont uniformes, c'est-à-dire que $\boldsymbol{\pi}_j \sim \text{Beta}(1,1)$. Nous supposons également que les distributions *a priori* des paramètres M_i sont log-uniformes (uniformes sur une échelle logarithmique) avec un support sur $[0.001, 1000]$: $\log(M_i) \sim \mathcal{U}(10^{-3}, 10^3)$.

Les distributions *a priori* des coefficients de sélection σ_{ij} (dans chaque dème, à chaque locus) sont définies de façon hiérarchique (voir, par exemple, Gelman *et al.* 2004, pp. 124-125). En particulier, nous faisons l'hypothèse que les paramètres σ_{ij} ont une distribution *a priori* de forme exponentielle, $f(\sigma_{ij}|\delta_j) \sim \exp\left(\delta_j^{-1}\right)$, qui dépend d'un hyper-paramètre locus-spécifique, noté δ_j, qui représente l'effet moyen de la sélection au locus j sur l'ensemble des dèmes. Nous faisons également l'hypothèse que ces hyper-paramètres δ_j ont une distribution *a priori* de forme exponentielle $f(\delta_j|\lambda) \sim \exp(\lambda^{-1})$ qui dépend elle-même d'un autre hyper-paramètre λ, qui représente l'effet génomique de la sélection, sur tous les locus et sur l'ensemble des dèmes. Enfin, nous faisons l'hypothèse que la distribution *a priori* de λ est $f(\lambda) \sim \exp(\Lambda^{-1})$, où $\Lambda = 0.5$. Le graphe orienté acyclique (*directed acyclic graph* ou DAG) de l'ensemble du modèle bayésien hiérarchique considéré ici est présenté dans la figure 2.2.1.

Si l'on suppose que les fréquences alléliques à différents locus sont indépendantes (équilibre de liaison) et que les fréquences alléliques dans des dèmes différents sont également indépendantes (modèle en îles infini), alors la distribution *a posteriori* des paramètres $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\kappa}, \boldsymbol{\delta}, \lambda|\mathbf{n})$, c'est-à-dire la distribution des paramètres $\mathbf{M}$, $\boldsymbol{\pi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$, et λ conditionnellement aux données $\mathbf{n}$, dépend de la vraisemblance des données et des distributions *a priori* des paramètres :

$$f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda|\mathbf{n}) \propto \prod_{i=1}^{n_\text{d}} \prod_{j=1}^{L} \mathcal{L}(p_{ij}; \mathbf{n}_{ij}) \psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij}) \times f(\mathbf{M}) f(\boldsymbol{\pi}) f(\boldsymbol{\kappa}) f(\boldsymbol{\sigma}|\boldsymbol{\delta}) f(\boldsymbol{\delta}|\lambda) f(\lambda) \quad (2.4)$$

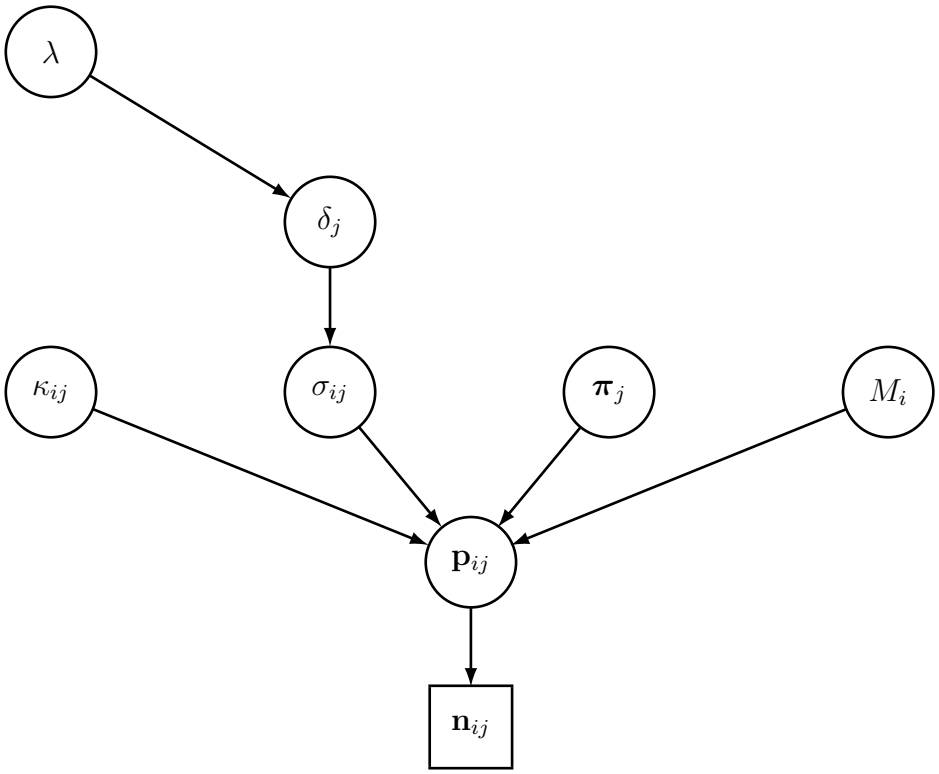


Figure 2.2: Graphe orienté acyclique (DAG) du modèle bayésien hiéarachique considéré dans SELESTIM.

Chaines de Markov Monte Carlo

La distribution *a posteriori* $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda|\mathbf{n})$ spécifiée par l'equation (2.4) est estimée grâce à un algorithme de Metropolis–Hasting (see, e.g., NTZOUFRAS 2009). En pratique, nous faisons varier les paramètres un à un, comme décrit dans la section § 1.4.3. Chaque chaîne de Markov est initialisée avec des valeurs de paramètres tirées dans les lois *a priori*, à l'exception des paramètre π_j's, pour lesquels on part des valeurs de Laplace calculées à partir des fréquences alléliques sur l'ensemble des données. Les paramètres des lois dans lesquelles sont tirées les nouvelles valeurs de paramètres $\mathbf{M}$, $\boldsymbol{\pi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$ et λ sont ajustés grâce au lancement de 25 chaînes courtes (1 000 itérations), afin d'atteindre des taux d'acceptation compris entre 0.25 et 0.40 (voir, par

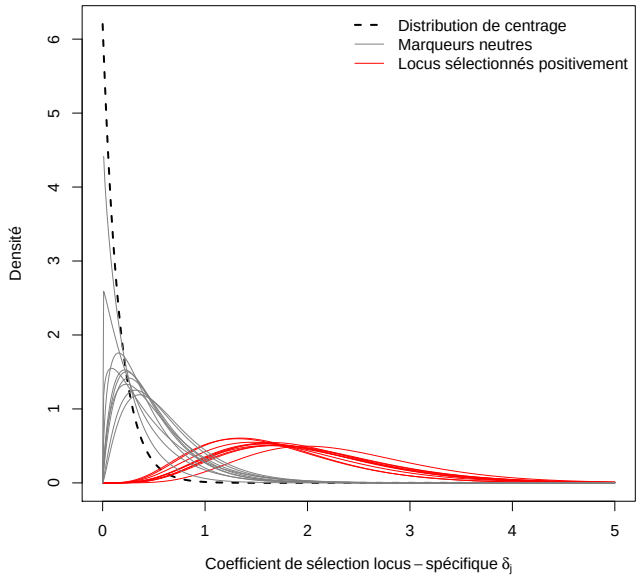


Figure 2.3: Exemple (sur des données simulées) de densités *a posteriori* de l'hyper-paramètre locus-spécifique δ_j pour des marqueurs neutres (en gris) et des locus sous sélection positive (en rouge). La divergence de Kullback–Leibler mesure l'écart entre la distribution *a posteriori* du coefficient de sélection locus-spécifique δ_j et de sa distribution "de centrage" (en pointillés), laquelle est définie par la distribution *a priori* des paramètres δ_j (de forme exponentielle) de moyenne égale à la valeur moyenne *a posteriori* du coefficient de sélection génomique λ.

exemple, Gilks *et al.* 1996).

Interprétation des résultats

Puisque nous avons fait l'hypothèse dans notre modèle que chaque locus est sélectionné, il est particulièrement pertinent de s'intéresser aux distributions *a posteriori* des hyper-paramètres locus-spécifiques δ_j : on s'attend en effet à ce que ces distributions tendent vers la valeur zéro pour des marqueurs (réellement) neutres, et que ces distributions s'écartent de la valeur zéro pour des locus (réellement) ciblés par la sélection (voir la figure 2.3). Ce-

pendant, étant donnée la structure hiérarchique de notre modèle, il ne serait pas suffisant de simplement tester si, à un locus en particulier, la distribution *a posteriori* de l'hyper-paramètre δ_j contient, ou non, la valeur zéro. En faisant cela, nous négligerions l'effet génomique de la sélection sur tous les locus et sur l'ensemble des dèmes (représenté par l'hyper-paramètre λ dans notre modèle). Puisque nous avons fait l'hypothèse dans notre modèle que les effets locus-spécifiques de la sélection δ_j sont tirés indépendamment dans une hyper-distribution *a priori* de paramètre λ, il est donc plus approprié de comparer les distributions *a posteriori* des coefficients de sélection locus-spécifiques δ_j à leur distribution de "centrage", laquelle est définie par la distribution *a priori* des paramètres δ_j (de forme exponentielle) de moyenne égale à la valeur moyenne *a posteriori* du coefficient de sélection génomique (représentée en pointillés dans la figure 2.3).

Pour comparer ces deux distributions, nous avons suivi l'exemple de Guo *et al.* (2009), et avons utilisé la divergence de Kullback–Leibler (*Kullback–Leibler divergence* ou KLD). La KLD est une mesure de dissimilarité entre deux distributions, qui est définie, pour deux densités $f(x)$ et $g(x)$, comme :

$$\mathrm{KLD}(f(x), g(x)) = \int_{-\infty}^{\infty} f(x) \log \left(\frac{f(x)}{g(x)} \right) \mathrm{d}(x). \tag{2.5}$$

La question est maintenant de savoir ce que représente la valeur de la KLD. Guo *et al.* (2009) ont proposé le raisonnement suivant pour calibrer la mesure de KLD. Si l'on considère un jeu de pile ou face avec une pièce non pipée (qui produit "face" dans 50% des lancers) et un jeu de pile ou face avec une pièce pipée (qui produit "face" dans seulement 5% des lancers), alors la KLD entre ces deux distributions de Bernoulli est égale à 0.830. Si l'on considère une pièce pipée qui produit "face" dans 1% des lancers, alors la KLD atteint 1.614. Dans la suite, nous considérerons ces valeurs limites pour distinguer des marqueurs présumés neutres de locus potentiellement sous sélection.

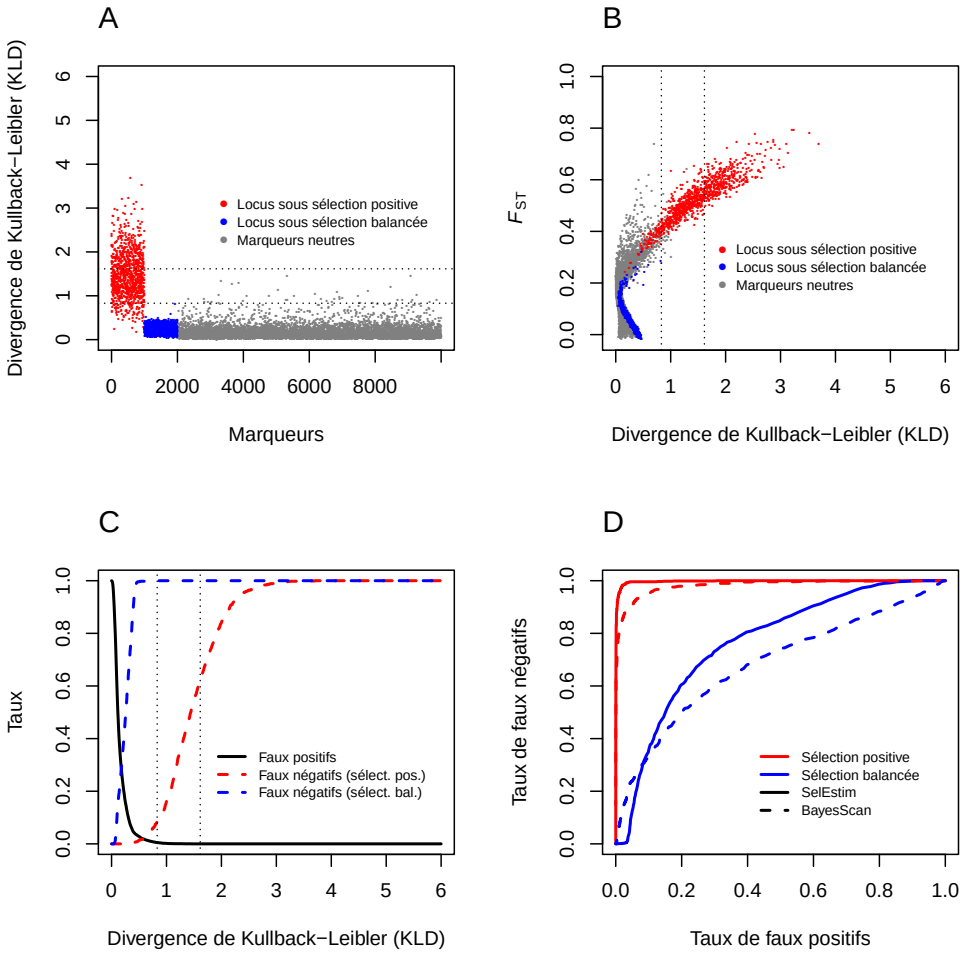


Figure 2.4: Exemple d'analyse sur un jeu de données simulées ($M \equiv 4Nm = 5$ et $\sigma \equiv 2Ns = 25$). (A) Mesure de la divergence de Kullback–Leibler (KLD) pour l'ensemble des 10 000 locus simulés. (B) F_{ST} en fonction de la mesure de KLD à tous les locus. (C) Taux de faux positifs (marqueurs neutres dont la KLD excède la valeur critique) et de faux négatifs (locus sous sélection dont la KLD est inférieure à la valeur critique), pour une large gamme de valeurs critiques de la KLD. Les deux lignes verticales en pointillés indiquent les valeurs critiques KLD = 0.830 et KLD = 1.614 (voir le texte). (D) Courbes ROC permettant de comparer les performances de SelEstim et de BayeScan.

2.2.2 Évaluation sur des données simulées

Nous avons évalué la performance de notre méthode sur des jeux de données simulés pour des valeurs de paramètres fixées. Les simulations ont été réalisées selon un modèle en îles de 50 dèmes, chaque dème étant constitué de $N = 250$ individus diploïdes. Dans ces simulations, dont le principe est largement inspiré de Beaumont et Balding (2004), on simule la distribution initiale des fréquences alléliques par tirages dans une urne de Pólya (Donnelly et Tavaré 1995), ce qui revient à considérer que la sélection agit sur la variation préexistante (*standing variation*), à la manière des modèles de Innan et Kim (2004) et Przeworsky *et al.* (2005). Ensuite, chaque génotype diploïde produit un nombre Poisson de descendants, et les processus de mutation, de dispersion et de sélection sont modélisés par des tirages multinomiaux successifs.

Dans ces simulations, pour tenir compte de l'adaptation à un environnement local, on attribue aléatoirement une "couleur" à chaque dème ("bleu", "rouge", ou "incolore"), indépendamment pour chaque locus. Pour les locus sous sélection positive, on considère un allèle B qui confère un avantage sélectif dans les dèmes "bleus" et un allèle R qui confère un avantage sélectif dans les dèmes "rouges". Les deux allèles B et R sont neutres dans les dèmes "incolores". Les homozygotes BB ont donc une valeur sélective égale à $(1+s)$ dans les dèmes "bleus" et égale à 1 dans les dèmes "rouges" et "incolores" ; les homozygotes RR ont une valeur sélective égale à $(1+s)$ dans les dèmes "rouges" et égale à 1 dans les dèmes "bleus" et "incolores" ; les hétérozygotes BR ont une valeur sélective égale à $(1+s/2)$ dans les dèmes "bleus" et "rouges" et égale à 1 dans les dèmes "incolores". Nous avons simulé également des locus sous sélection balancée (bien que cette forme de sélection ne soit pas prise en compte dans le modèle d'inférence), où les hétérozygotes BR ont un avantage sélectif (avec valeur sélective égale à $1+s$) dans les dèmes "bleus" et "rouges" ; les homozygotes BB et RR ont une valeur sélective égale à 1 dans tous les dèmes, comme les hétérozygotes BR dans les dèmes "incolores".

Par la suite, on a simulé 30% de dèmes "bleus", 30% de dèmes "rouges" et 40% de dèmes "incolores". Pour chaque locus, 50 individus diploïdes (100

gènes) ont été échantillonnés dans chaque dème. Ces échantillons de gènes ont été collectés dans 6 dèmes, dont 2 “bleus”, 2 “rouges” et 2 “incolores”. Le jeu de données simulées considéré ici consiste en 10 000 SNPs, incluant 8 000 marqueurs neutres, 1 000 locus sélectionnés positivement et 1 000 locus sous sélection balancée.

La figure 2.4 montre un exemple d'application de notre méthode sur un jeu de données simulées avec $M \equiv 4Nm = 5$ et $\sigma \equiv 2Ns = 25$. La figure 2.4A montre que la distribution de la KLD pour des marqueurs sous sélection positive s'écarte bien de la distribution de la KLD pour des marqueurs neutres et pour des marqueurs sous sélection balancée. D'autres simulations montrent que ces résultats sont en général valables pour $M \geq 5$ et $\sigma/M \geq 5$. Comme on pouvait s'y attendre, de fortes valeurs de KLD correspondent à de fortes valeurs de F_{ST} (figure 2.4B) : pour des locus sous sélection positive, en effet, un allèle (B) est sélectionné dans les populations “bleues” tandis que l'autre allèle (R) est sélectionné dans les populations “rouges”. La figure 2.4B montre en outre que la valeur critique KLD = 0.830 (voir Guo *et al.* 2009) permet de distinguer les locus sous sélection positive des marqueurs neutres. Ce dernier point est particulièrement évident sur la figure 2.4C, qui montre que cette valeur critique de la KLD permet de minimiser à la fois le taux de faux positifs (la proportion de marqueurs neutres pour lesquels on détecte une signature de sélection) et le taux de faux négatifs (la proportion de marqueurs sous sélection positive pour lesquels on ne détecte pas de signature de sélection).

Le même jeu de données simulées a été analysé avec BayeScan (Foll et Gaggiotti 2008). BayeScan repose sur un modèle beta–binomial (ou bien multinomial–Dirichlet dans le cas de marqueurs multi-alléliques) pour les fréquences alléliques dans un modèle en îles à l'équilibre migration-dérive. Dans ce modèle, à chaque locus, le paramètre F_{ST} est défini comme la variance des fréquences alléliques entre chaque dème et le patrimoine génétique des migrants. Dans BayeScan, comme dans le modèle de Beaumont et Balding (2004), le paramètre F_{ST} est décomposé en une composante locus-spécifique (α_i), partagée par l'ensemble des populations, et une composante population-spécifique (β_j), partagée par l'ensemble des locus. Dans ces modèles, une valeur de α_i “significativement différente de zéro” est considérée

comme une signature de sélection (Beaumont et Balding 2004; Foll et Gaggiotti 2008; Riebler *et al.* 2008).

Alors que Beaumont et Balding (2004) testent la significativité de α_i à partir de la distribution *a posteriori* de ce paramètre, et que Riebler *et al.* (2008) utilisent une variable indicatrice, BayeScan repose sur un algorithme de Monte Carlo par chaîne de Markov à sauts réversibles, qui permet d'estimer la probabilité *a posteriori* de deux modèles alternatifs : un modèle purement neutre (où $\alpha_i = 0$) et un modèle incluant la sélection (où $\alpha_i \neq 0$). Pour chaque analyse, un facteur de Bayes peut donc être calculé pour le modèle incluant la sélection ($\alpha_i \neq 0$). Le facteur de Bayes est un rapport où le numérateur est la probabilité *a posteriori* d'un modèle divisée par sa probabilité *a priori* et où le dénominateur est la probabilité *a posteriori* du modèle alternatif divisée par sa probabilité *a priori* (Gelman *et al.* 2004) [1].

La figure 2.4D montre la performance relative de notre approche par rapport à BayeScan. La performance relative de ces deux approches peut être appréciée par l'analyse des courbes ROC (pour *receiver operating characteristic*, voir la figure 2.4D). Dans les analyses ROC, la proportion de faux positifs et de vrais positifs est calculée pour toutes les valeurs possibles du critère de décision utilisé pour distinguer les locus sous sélection des marqueurs neutres (voir, par exemple, Fawcett 2006). Dans notre modèle, le critère de décision est la divergence de Kullback–Leibler mesurée entre les distributions *a posteriori* des coefficients de sélection locus-spécifiques δ_j et de leur distribution de "centrage". Dans le cas de BayeScan, le critère de décision est le facteur de Bayes pour le modèle incluant la sélection.

Une analyse ROC produit une courbe monotone partant du point (0,0), correspondant à une absence complète de positifs (vrais ou faux) et allant au point (1,1), correspondant à une présence exclusive de positifs (vrais et faux). Une méthode qui n'a aucune puissance de détection, qui se comporte donc comme un classificateur aléatoire, a une courbe ROC linéaire de pente égale à 1. Au contraire, un classificateur idéal a une courbe ROC qui se superpose au côté gauche et au côté supérieur du carré de côté 1.

1. Dans l'exemple présenté ici, on a supposé un rapport des probabilités *a priori* égal à 10 en faveur du modèle neutre.

En ce qui concerne les locus sélectionnés positivement, la surface sous la courbe ROC est légèrement supérieure (et plus proche de 1) pour SelEstim, par rapport à la surface sous la courbe ROC obtenue pour BayeScan (voir la figure 2.4D), ce qui suggère une meilleure performance de notre approche par rapport à BayeScan, au moins sur ces données simulées. L'ensemble de la figure 2.4 montre enfin que notre méthode n'a absolument aucune puissance statistique pour identifier des locus sous sélection balancée (bien que l'analyse ROC illustrée par la figure 2.4D tendrait à montrer une performance de notre approche légèrement plus élevée par rapport à BayeScan) . Ce résultat n'est en soit pas surprenant étant donné que le modèle de sélection considéré dans notre modèle ne considère qu'un modèle de sélection positive et pas de sélection balancée. D'autre part, d'autres études de simulation ont également montré qu'en l'absence d'un modèle de sélection explicite, les signatures de sélection balancée sont très difficiles à détecter (Beaumont et Balding 2004; Foll et Gaggiotti 2008; Riebler *et al.* 2008).

2.2.3 Inférence de l'intensité de la sélection

Nous avons ensuite examiné la distribution des moyennes *a posteriori* des paramètres κ_{ij} qui indiquent quel allèle (de B et R) confère un avantage sélectif. D'après l'hypothèse de notre modèle de simulation, $\kappa_{ij} = 0$ indique que c'est l'allèle B ("bleu") qui est sélectionné, tandis que $\kappa_{ij} = 1$ indique que c'est l'allèle R ("rouge") qui est sélectionné. La figure 2.5A montre les distributions des moyennes *a posteriori* de κ_{ij} dans chaque dème. Comme attendu, les moyennes *a posteriori* de κ_{ij} dans les dèmes 1 et 2 (les dèmes "bleus") sont décalées vers zéro, tandis que les moyennes *a posteriori* de κ_{ij} dans les dèmes 3 et 4 (les dèmes "rouges") sont décalées vers un. Dans les dèmes 5 et 6 (les dèmes "incolores"), les moyennes *a posteriori* de κ_{ij} sont centrées autour de 0.5, ce qui est cohérent avec le fait que ni l'allèle B, ni l'allèle R ne sont sélectionnés dans ces dèmes.

La figure 2.5B montre les distributions des moyennes *a posteriori* des paramètres $\sigma_{ij} \equiv 2N_i s_{ij}$, conditionnellement à κ_{ij}. En conditionnant le paramètre σ_{ij} à κ_{ij}, nous mesurons l'intensité de la sélection pour l'allèle ef-

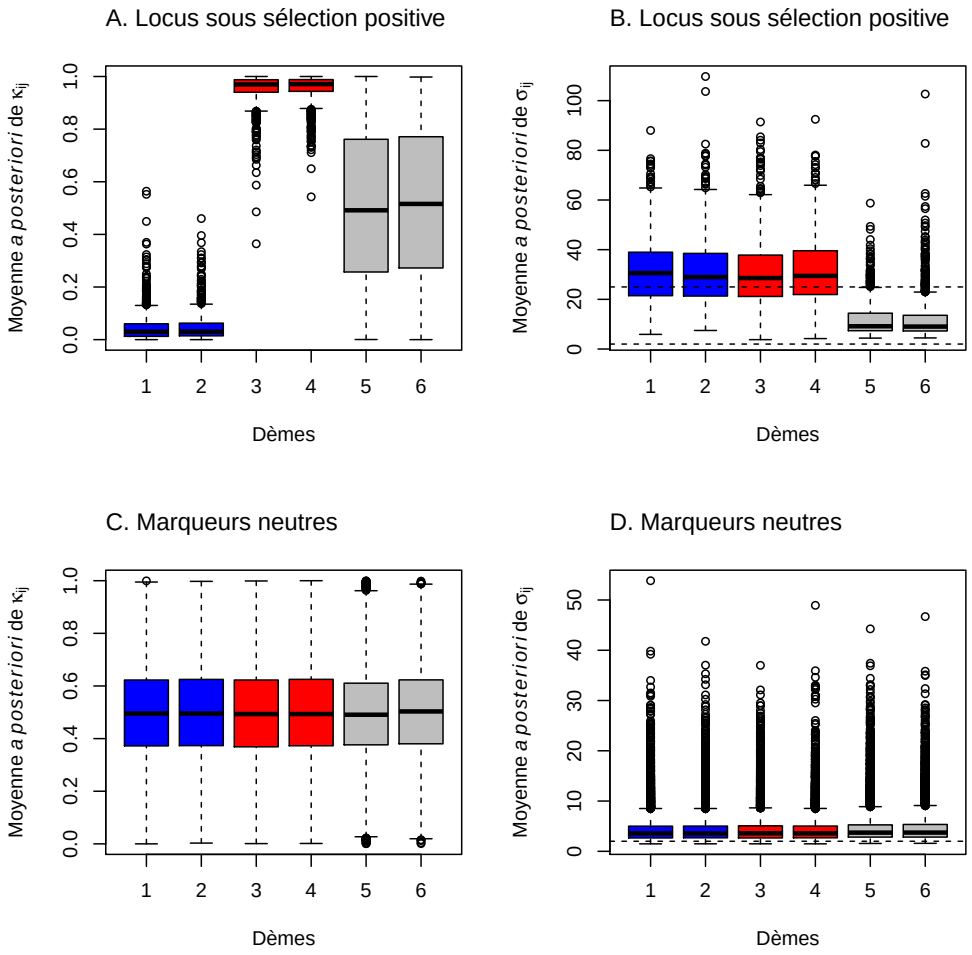


Figure 2.5: Inférence de l'intensité de la sélection sur un jeu de données simulées. (A) Moyennes *a posteriori* des paramètres κ_{ij} pour les 1 000 locus sous sélection positive, dans les dèmes "bleus" (dèmes 1 et 2), "rouges" (dèmes 3 et 4) et "incolores" (dèmes 5 et 6).(B) Moyennes *a posteriori* des coefficients de sélection σ_{ij} pour les 1 000 locus sous sélection positive. Dans les dèmes "bleus", les moyennes sont conditionnelles à $\kappa_{ij} = 0$, dans les dèmes "rouges" les moyennes sont conditionnelles à $\kappa_{ij} = 1$, dans les dèmes "incolores" les moyennes sont non-conditionnelles. (C) Idem (A) pour les 8 000 marqueurs neutres. (D) Idem (B) pour les 8 000 marqueurs neutres. Les moyennes de σ_{ij} sont non-conditionnelles.

fectivement sélectionné. La figure 2.5B montre que les moyennes *a posteriori* de $f(\sigma_{ij}|\kappa_{ij} = 0)$ dans les dèmes "bleus", et que les moyennes *a posteriori* de $f(\sigma_{ij}|\kappa_{ij} = 1)$ dans les dèmes "rouges" sont très proches des valeurs simulées (dans cet exemple, $\sigma \equiv 2Ns = 25$). En revanche, les moyennes *a posteriori* des paramètres σ_{ij}, non-conditionnelles à κ_{ij}, sont bien plus faibles, et proches de la moyenne de la distribution *a priori* de l'hyper-paramètre λ, qui représente l'effet "génomique" de la sélection, à tous les locus, sur l'ensemble des dèmes. Des résultats tout à fait similaires ont été obtenus pour des jeux de données avec $M \geq 5$ et $\sigma/M \geq 5$, et avec $M = 2$ et $\sigma/M = 10$.

Enfin, nous avons examiné la distribution des moyennes *a posteriori* des paramètres κ_{ij} pour les 8 000 marqueurs neutres du jeu de données considéré ici. La igure 2.5C montre que les moyennes *a posteriori* des paramètres κ_{ij}, qui ne dépendent pas de la "couleur" des dèmes échantillonnés, sont toutes centrées autour de 0.5. Ce résultat est cohérent avec le fait que ni l'allèle B, ni l'allèle R ne sont sélectionnés ici. De plus, les moyennes des distributions *a posteriori* des paramètres σ_{ij} pour les marqueurs neutres (non conditionnelles à κ_{ij}) sont très faibles, et proches des distributions *a priori* des l'hyper-paramètre λ (figure 2.5D). L'ensemble de la figure 2.5 montre donc que notre modèle d'inférence est capable de fournir des mesures assez précises des coefficients de sélection à un locus dans différents dèmes, et donc de mettre en évidence l'adaptation à un environnement local.

2.2.4 Application sur les données humaines du CEPH

A titre d'exemple, nous avons appliqué notre méthode sur un sous-ensemble des données de SNPs du *Stanford HGDP-CEPH Human Genome Diversity Cell Line Panel* (Cann *et al.* 2002), qui réunissent plus de 650 000 marqueurs. Nous nous sommes intéressés plus particulièrement aux marqueurs présents sur le chromosome 2 (53 765 SNPs), auxquels nous avons ajouté les données de comptages de deux SNPs (-13910C→T and -22018G→A) dont l'association à la capacité des individus adultes à digérer le lactose a été démontrée (Bersaglieri *et al.* 2004). Nous nous sommes concentrés sur les populations de l'Ancien Monde, en ne retenant que 23 populations d'Afrique

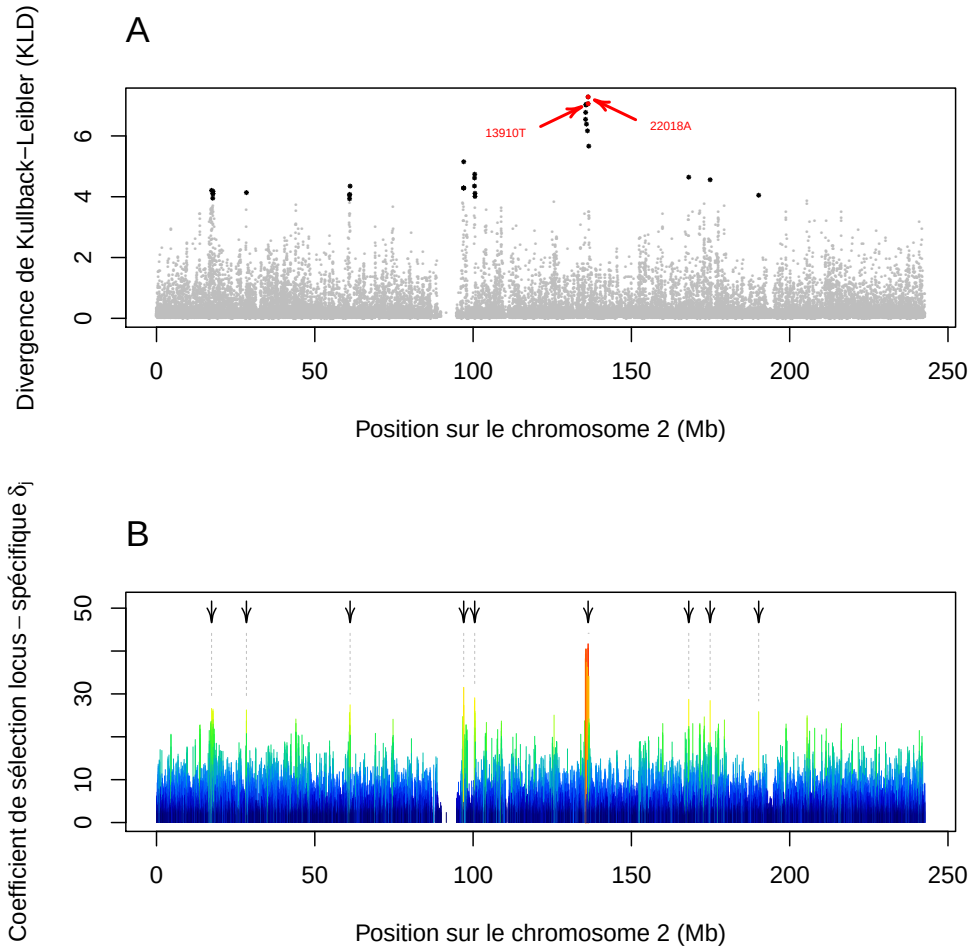


Figure 2.6: Balayage génomique le long du chromosome 2 humain pour des populations de l'Ancien Monde. (A) Mesure de la divergence de Kullback–Leibler (KLD) pour 52 633 marqueurs polymorphes le long du chromosome 2. Les allèles -13910C→T et -22018G→A, connus pour être associés à la tolérance au lactose, sont indiqués par des flèches rouges. Les points noirs correspondent aux SNPs ayant une valeur de KLD supérieure à la valeur limite KLD = 3.912, qui correspondrait à la KLD entre deux distributions de Bernoulli de paramètres 0.5 et 0.0001, respectivement. (B) Distribution, le long du chromosome 2, de la moyenne *a posteriori* du coefficient de sélection locus-spécifique δ_j.

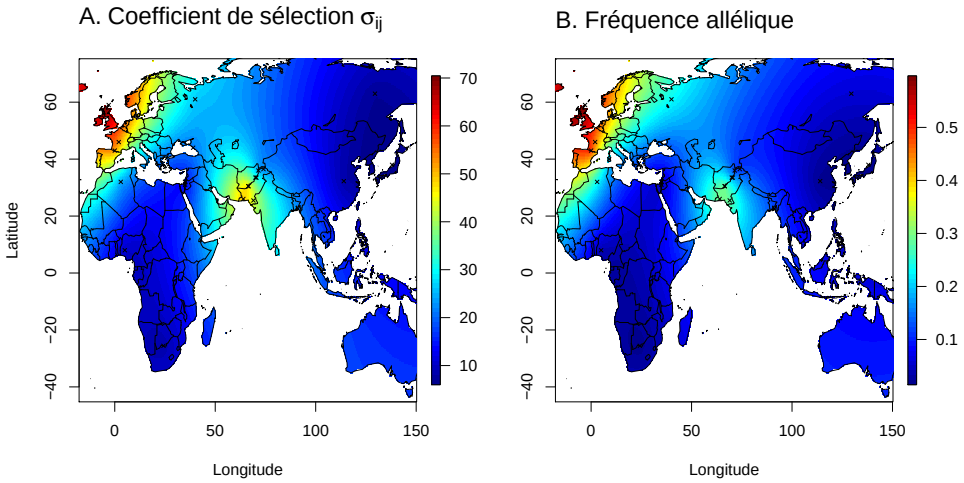


Figure 2.7: Distribution spatiale de l'intensité de la sélection pour l'allèle -13910C→T dans l'Ancien Monde. (A) Extrapolation par krigeage de la distribution spatiale de la moyenne *a posteriori* du coefficient de sélection σ_{ij} pour l'allèle -13910C→T, en Afrique et en Eurasie. Les populations échantillonnées sont représentées par une croix. (B) Extrapolation par krigeage de la distribution spatiale de la fréquence de l'allèle -13910C→T.

et d'Eurasie, pour un total de 52 633 marqueurs polymorphes (fréquence de l'allèle minoritaire > 0.01).

La figure 2.6A montre la distribution, le long du chromosome 2, de la divergence de Kullback–Leibler (KLD) mesurée, pour chaque marqueur, entre la distribution *a posteriori* du coefficient de sélection locus-spécifique δ_j et de sa distribution "de centrage" (dérivée de la distribution de l'hyper-paramètre λ qui représente l'effet génomique de la sélection). Les deux SNPs (-13910C→T et -22018G→A) sont mis en évidence. Ces deux marqueurs font partie de l'ensemble des trois marqueurs possédants les plus fortes valeurs de KLD le long du chromosome 2. En outre, les neufs marqueurs avec les valeurs de KLD les plus fortes se trouvent dans une région comprise entre 3.7 kb et 1.0 Mb en amont du gène *LCT* qui code pour l'enzyme lactase-phlorizine hydrolase dont l'activité est associée à la capacité à digérer le lactose. Ces neufs marqueurs se trouvent également à moins de 805.2 kb de -13910C→T et à moins de 813.4

kb de -22018G→A. La figure 2.6B représente la distribution des moyennes *a posteriori* des paramètres de sélection locus-spécifiques δ_j, le long du chromosome 2. Cette figure représente donc la variation de l'intensité de la sélection le long du chromosome, et montre un signal extrêmement fort de sélection positive dans le voisinage du gène *LCT*.

Enfin, la figure 2.7A illustre la distribution (obtenue par krigeage) des coefficients de sélection σ_{ij} (conditionnellement à la valeur de κ_{ij} indiquant que l'allèle -13910C→T est l'allèle sélectionné) à l'échelle des populations africaines et eurasiennes. Cette figure montre que l'intensité de la sélection est très forte en Europe et dans la vallée de l'Indus, ce qui est cohérent avec la distribution du phénotype de tolérance au lactose en Eurasie. Le seul examen de la distribution spatiale des fréquences de l'allèle -13910C→T ne permet pas de conclure à des niveaux de sélections identiques dans les deux régions (figure 2.7B). Notons qu'une étude récente (Romero *et al.* 2012) a montré que la mutation-13910C→T en Inde est associée au même haplotype étendu qu'en Europe, ce qui suggère fortement une origine de la mutation-13910C→T partagée en Europe et en Inde. Ces résultats sont cohérents avec les niveaux élevés de consommation de lait en Inde, et avec des preuves archéologiques et génétiques pour la domestication indépendante du bétail dans la vallée de l'Indus Il y a environ 7 000 ans (Romero *et al.* 2012).

De manière générale, cette approche nous semble donc prometteuse pour analyser le polymorphisme issu des données de génotypage haut-débit. Une application au jeu de données humaines POPRES (Nelson *et al.* 2008) est d'ailleurs envisagée dans le cadre du stage de Master 2 de Homa Papoli au premier semestre de l'année universitaire 2012–2013. Ce jeu de données, composé de 500 000 marqueurs SNPs caractérisés pour environ 3 000 individus d'origine européenne, devrait nous permettre de mieux caractériser la sélection agissant sur la capacité à digérer le lactose à l'âge adulte eu Europe.

Chapitre 3

Traits d'histoire de vie

Tous les travaux que je viens de présenter ont pour dénominateur commun la volonté de comprendre les conséquences de la structure (spatiale, en âge, par sexe) des populations, sur le niveau et la distribution de la variation génétique neutre ou sélectionnée. Les facteurs responsables de la mise en place de ces structures (comme par exemple la capacité de dispersion) ne sont pas des caractères fixés, mais sont au contraire susceptibles d'évoluer. Je m'intéresse donc également à l'évolution de ces caractères.

3.1 La dispersion

3.1.1 Évolution de la dispersion dans une métapopulation

Quel peut être le bénéfice adaptatif de la dispersion ? Disperser, c'est courir le risque de mourir avant d'avoir pu transmettre ses gènes à la génération suivante. Mais il peut également y avoir des avantages à la dispersion : si l'environnement est changeant par exemple, disperser peut être le moyen de diminuer le risque de faire face à un événement catastrophique local (Comins *et al.* 1980; Gandon et Michalakis 1999; Olivieri et Gouyon 1995; van Valen 1971), ou bien d'échapper à la baisse de valeur sélective due à la dépression de consanguinité (Bengtsson 1978; Gandon 1999; Motro 1991;

Perrin et Mazalov 1999, 2000; Shields 1983; Waser *et al.* 1986) ; enfin disperser, ce peut aussi être un moyen d'éviter la compétition entre des individus apparentés (Frank 1986; Gandon et Michalakis 2001; Rousset 2003; Rousset et Billiard 2000; Taylor 1988; Taylor et Frank 1996).

J'ai commencé à m'intéresser à l'évolution de la dispersion lors de mon post-doctorat avec Vincent Jansen (Royal Holloway, Université de Londres). J'ai travaillé sur des modèles analytiques qui décrivent la dynamique de stratégies de dispersion dans une *métapopulation*, c'est-à-dire dans un ensemble de populations soumises à des processus d'extinctions locales et de re-colonisations (Levins 1968). À de rares exceptions près (voir Rousset et Ronce 2004), les modèles qui prennent en compte la compétition entre individus apparentés supposent en général une taille de population constante. Afin de mieux rendre compte de l'interaction des dynamiques écologiques et évolutives, nous avons introduit dans nos modèles une dynamique écologique locale. Cette dynamique locale permet de tenir compte non seulement de la régulation des populations, mais aussi de la compétition entre individus apparentés, de la stochasticité démographique et des risques d'extinction locale [P14]. Notre modèle permet ainsi de faire la part entre la compétition locale pour les ressources, la compétition entre individus apparentés et la dynamique des extinctions et recolonisations pour expliquer l'évolution de la dispersion dans une métapopulation [P14].

Toujours lors de ce post-doctorat avec Vincent Jansen, je me suis également intéressé aux conséquences de l'existence d'un compromis ou d'une compensation (*trade-off*) entre les capacités de compétition et de colonisation des individus, sur l'évolution de la dispersion.

3.1.2 Compromis entre compétition et colonisation

Les modèles d'évolution de la dispersion qui formalisent explicitement la compétition entre individus apparentés considèrent que la compétition entre les individus est une sorte de loterie, où chaque individu a la même chance de remporter la compétition et de s'établir. Cette forme de compétition équilibrée ou équitable est raisonnable si (i) les individus ne diffèrent les uns des

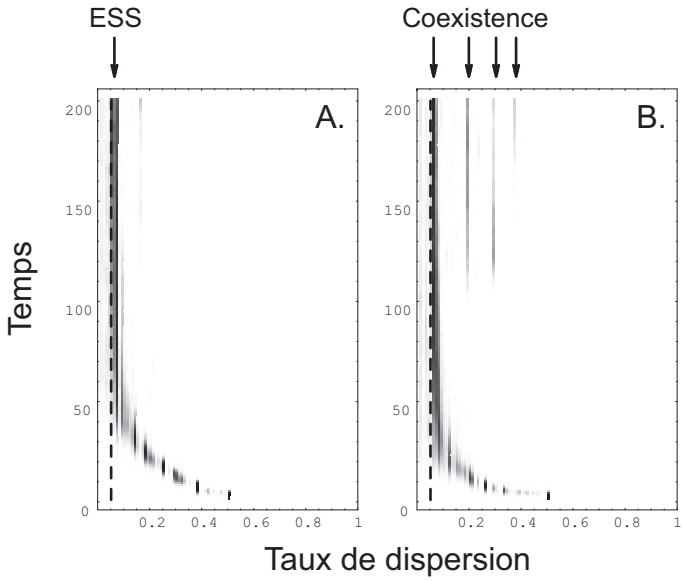


Figure 3.1: Simulations stochastiques montrant l'évolution au cours du temps du taux de dispersion dans une métapopulation lorsqu'il existe une compensation évolutive entre les capacités de compétition et de colonisation, lors de la régulation des populations. La ligne pointillée indique la valeur de la stratégie de dispersion évolutivement stable (ESS), calculée analytiquement dans le modèle. A, Lorsque les mutations ont des effets faibles, le taux de dispersion évolutivement stable envahit la métapopulation. B, Lorsque les mutations ont des effets plus forts, des stratégies alternatives peuvent apparaître par mutation et se maintenir en *coexistence* dans la population.

autres que par leur capacité de dispersion, et (*ii*) si la dispersion n'a aucune influence sur la compétitivité. Que se passe-t-il si ces hypothèses ne sont pas remplies et que des individus *bons* compétiteurs sont de *mauvais* colonisateurs (c'est-à-dire s'il existe un compromis entre les capacités de compétition et de colonisation) ? Si l'importance de cette forme de compensation a été reconnue dans le domaine de l'Écologie pour expliquer la coexistence de plusieurs espèces dans un habitat fragmenté (Lehman et Tilman 1997), les modèles de sélection de parentèle n'intègrent pas ces aspects écologiques.

J'ai donc développé un modèle, basé sur celui de Comins *et al.* (1980), pour calculer la valeur sélective d'un individu possédant une stratégie de dis-

persion légèrement déviante, dans une population composée d'individus résidents possédant tous la même stratégie de dispersion. Ce modèle considère une métapopulation composée d'un nombre infini de dèmes, chacun de taille égale et fixée. Ces dèmes ont une probabilité non-nulle de s'éteindre, chaque génération. Ce modèle considère enfin qu'il existe une relation négative, ou une compensation, entre la capacité de dispersion et la capacité de compétition des individus (un bon compétiteur est donc un mauvais colonisateur, et vice-versa).

L'analyse de ce modèle nous apprend qu'en l'absence de compensation, une seule stratégie de dispersion envahit la métapopulation. Cette stratégie ne peut être remplacée par aucune stratégie déviante. Il s'agit donc d'une stratégie évolutivement stable (Maynard Smith 1982). Mais si les bons colonisateurs sont de mauvais compétiteurs durant la phase de régulation des populations, notre modèle montre que la coexistence de différentes stratégies de dispersion est possible. En d'autres termes, l'existence d'une compensation entre les capacités de compétition et de dispersion est une condition suffisante pour que différentes stratégies de dispersion coexistent dans une métapopulation. Toutefois, ce modèle ne prédit pas l'émergence d'un polymorphisme suite à un phénomène de branchement évolutif, c'est-à-dire que notre modèle ne prédit pas que la sélection favorise l'émergence de deux stratégies à partir d'une stratégie unique. La coexistence n'est donc possible que si des stratégies alternatives de dispersion apparaissent dans la population par mutation ou par migration à partir de métapopulations voisines. Par conséquent, l'effet des mutations codant pour les stratégies de dispersion joue un rôle important dans la mise en place d'un polymorphisme pour ce caractère (Figure 3.1). Ces travaux font l'objet d'un article aujourd'hui en préparation [p6].

3.2 La dormance

Chez certaines espèces, il existe des alternatives à la dispersion spatiale. Des exemples assez frappants existent chez certaines espèces végétales ou animales qui ont mis en place au cours de l'évolution des stratégies de dormance (ou de diapause chez les insectes) qui leur permettent de produire des dia-

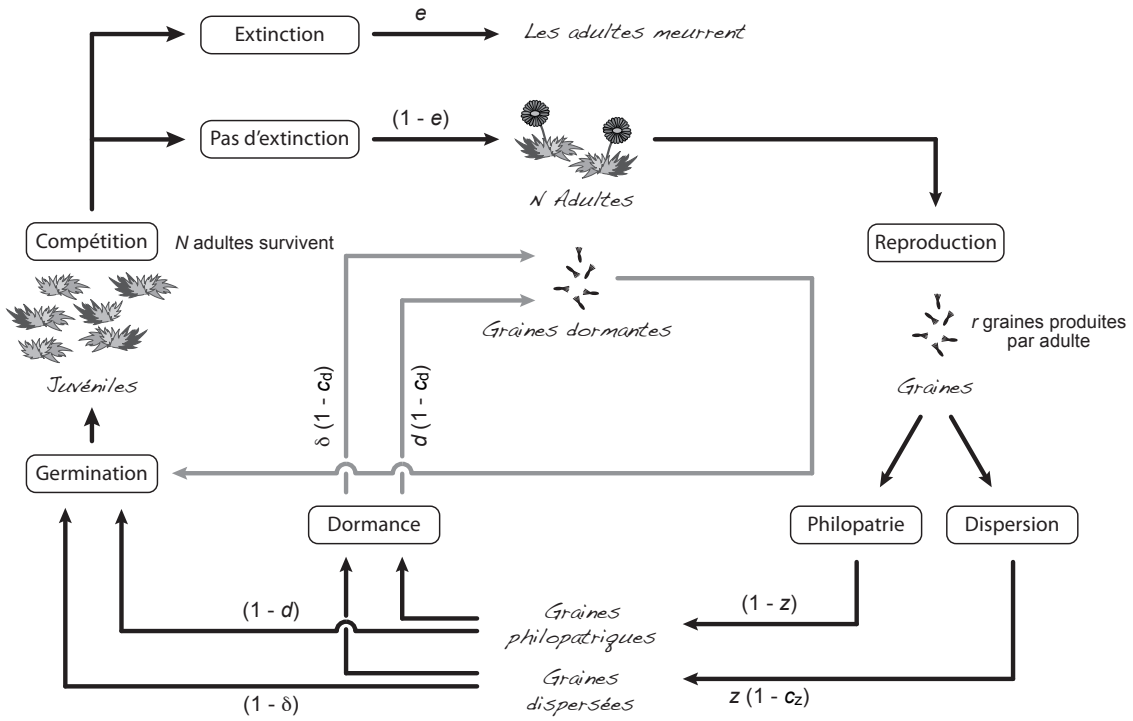


Figure 3.2: Cycle de vie du modèle d'évolution conjointe de la dormance et de la dispersion.

spores (ou des œufs) qui attendent plusieurs saisons avant de se développer (Harper 1977). C'est le cas par exemple de nombreuses espèces végétales inféodées aux habitats désertiques (Venable *et al.* 1993). Quel peut être le bénéfice adaptatif d'un tel caractère ? Comment ce caractère évolue-t-il conjointement avec la dispersion ?

3.2.1 Le contexte

La dispersion et la dormance sont des stratégies qui ont un coût, en terme de valeur sélective, au sens où ces stratégies nécessitent le développement de caractéristiques physiologiques et morphologiques spécifiques pour disperser ou pour entrer dans un stade de dormance. Il existe aussi des coûts associés à la variation des conditions environnementales : tout comme un individu qui disperse peut atteindre une parcelle d'habitat peu favorable s'il existe

une variabilité spatiale des conditions environnementales, un individu dormant peut émerger dans des conditions défavorables s'il existe une variabilité temporelle de l'environnement. La dispersion et la dormance sont également associés à des bénéfices très similaires (Venable et Brown 1988; Venable *et al.* 1993). En l'absence de densité-dépendance, la dispersion et la dormance peuvent permettre d'éviter les risques associés à la variation temporelle des conditions environnementales (Philippi et Seger 1989; Slatkin 1974). Par exemple, s'il existe une variation temporelle de la survie et/ou de la fécondité en raison de la succession de "bonnes" et de "mauvaises" années , la production de graines dormantes répartit le risque d'échec de reproduction en reportant l'émergence des propagules (Cohen 1966; Venable 2007). La dispersion peut également permettre d'éviter les risques associés à la variation temporelle des conditions environnementales (Ronce 2007; Venable et Brown 1988; Venable *et al.* 1993). Enfin, en présence de densité-dépendance, la dispersion et la dormance permettent de réduire la compétition locale entre les individus (Ellner 1985a,b; Levin *et al.* 1984), notamment la compétition entre les individus apparentés (Ellner 1986; Frank 1986; Hamilton 1964; Hamilton et May 1977; Kobayashi et Yamamura 2000; Taylor 1988).

Puisque la dispersion et la dormance répondent à des forces évolutives semblables, il est tentant de considérer que ces stratégies peuvent se substituer l'une à l'autre. Si tel est le cas, on s'attend alors à observer une corrélation négative entre ces deux traits d'histoire de vie. De ce point de vue, des études théoriques ont en effet confirmé la prédiction selon laquelle, en général, augmenter le taux de dispersion a pour effet de diminuer le taux évolutivement stable (ES) de dormance (Kobayashi et Yamamura 2000; Satterthwaite 2010). D'autres études ont également montré qu'en général l'augmentation du taux de dormance avait pour effet de sélectionner des taux de dispersion ES plus faibles (Cohen et Levin 1991; Levin *et al.* 1984; Snyder 2006). Pour autant, seuls des modèles où la dispersion et la dormance évoluent conjointement peuvent permettre de prédire si l'évolution favorise des corrélations positives ou négatives entre la dormance et dispersion. Certains modèles ont donc été développés pour étudier, numériquement,

l'évolution conjointe de la dispersion et de la dormance dans différents scénarios écologiques (Cohen et Levin 1987; Klinkhamer *et al.* 1987; McPeek et Kalisz 1998; Olivieri 2001; Tsuji et Yamamura 1992; Venable et Brown 1988; Wiener et Tuljapurkar 1994). Pourtant, aucun de ces modèles n'a considéré l'effet de la compétition entre individus apparentés sur les dynamiques évolutives des deux traits.

En collaboration avec Sylvain Gandon (CEFE, Montpellier), François Rousset, Isabelle Olivieri (ISE-M, Montpellier) et Yutaka Kobayashi (Université de Tokyo), j'ai développé un modèle mathématique afin de rechercher les stratégies évolutivement stables conjointement pour la dispersion et la dormance dans une métapopulation. Ce modèle prend explicitement en compte l'effet des extinctions locales et celui de la compétition entre individus apparentés du fait de la taille limitée des populations locales. Ce modèle fait l'objet d'une publication en cours de préparation [p2], qui est reproduite en Annexe G, page 259.

3.2.2 Le modèle

Nous considérons le cycle de vie suivant : (*i*) les adultes produisent un nombre r de graines (tiré dans une loi de Poisson) puis meurent ; (*ii*) une fraction z de graines sont dispersées, et les graines qui dispersent paient un coût noté c_{z} ; (*iii*) une fraction D des graines entrent dans un état de dormance, et toutes les graines dormantes paient un coût noté c_{d} ; (*iv*) toutes les graines non dormantes, ainsi que toutes les graines dormantes produites dans le pas de temps précédent germent ; en d'autres termes, nous supposons comme dans le modèle de Kobayashi et Yamamura (2000) que les graines dormantes passent au maximum un an dans la banque ; toutefois, nous avons relâché cette hypothèse dans des simulations individus-centrées ; (*v*) toutes les graines qui germent entrent en compétition, et seulement N d'entre elles survivent à l'âge adulte ; (*vi*) certains dèmes font face à des événements catastrophiques (extinctions) qui se produisent avec une probabilité e ; ces événements provoquent la mort de tous les individus adultes dans le dème en question. La Figure 3.2 représente ce cycle de vie. Nous considérons

également un cycle de vie alternatif, dans lequel la dormance est conditionnée à la dispersion : le taux de dormance des graines dispersées peut différer de celui des graines non-dispersées (philopatriques), comme dans Olivieri (2001). Plus précisément, nous considérons qu'à l'étape (*iii*) du cycle de vie ci-dessus, une fraction d des graines philopatriques et une fraction δ des graines dispersées entrent dans un état de dormance.

Afin d'étudier les dynamiques évolutives des taux de dispersion et de dormance, nous avons utilisé une approche de "valeur sélective directe" (*direct fitness*, voir Rousset et Billiard 2000; Taylor et Frank 1996) pour calculer la valeur sélective d'un individu focal (c'est à dire le nombre attendu de ses descendants survivants), en fonction des stratégies, ou des phénotypes de tous les individus avec lesquels il est en compétition. Nous supposons que chacun de ces traits phénotypiques est codé par un locus bi-allélique : à chaque locus, nous considérons un allèle mutant A dans une population d'individus qui portent l'allèle a. Nous supposons que l'allèle a code pour un phénotype z_a, et que l'allèle mutant A code pour un phénotype $z_A \equiv z_a + \epsilon_z$. Dans le modèle à nombre d'îles infini (Wright 1931), le changement attendu Δp de la fréquence allélique p sur une génération est donné par (voir Rousset 2004) :

$$\Delta p = p(1-p)S(z)\epsilon_z + O(\epsilon_z^2) \tag{3.1}$$

où $S(z)$ est le gradient de sélection. Déterminer une stratégie candidate pour être évolutivement stable revient à rechercher la valeur du trait pour laquelle le gradient de sélection $S(z)$ s'annule.

Dans le modèle considéré ici, tous les individus ne sont pas équivalents. Dans un dème, par exemple, les individus adultes et les graines dans la banque de graines ne sont pas en compétition les uns avec les autres. Ils doivent donc être traités différemment dans l'analyse. Tous les dèmes ne sont pas équivalents non plus. Par exemple, les dèmes qui se sont éteints une génération dans le passé ne peuvent pas contenir de graines philopatriques dormantes (c'est-à-dire de graines qui auraient été produites par des adultes résidant dans ces dèmes). Dans ces dèmes, il n'y a donc pas de compétition entre

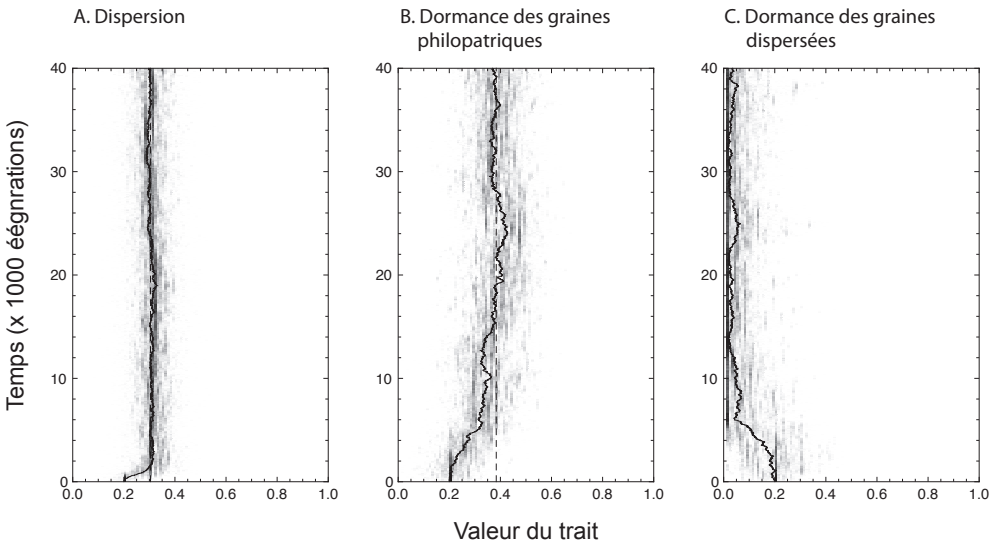


Figure 3.3: Dynamiques évolutives des taux de dormance et de dispersion dans une métapopulation de $n_{\mathrm{d}} = 2\ 000$ dèmes, chacun de taille $N = 5$. Ces dynamiques sont le résultat d'une simulation individus-centrée, où chaque individu est caractérisé par un ensemble de variables représentant son phénotype pour chaque trait d'histoire de vie. Le cycle de vie dans les simulations est le même que pour le modèle analytique (voir Figure 3.2) : chaque individu adulte produit un nombre de graines, tiré dans une loi de Poisson de moyenne $r = 100$. Pour chaque trait, les phénotypes sont altérés par des mutations qui se produisent à un taux $\mu = 0.001$. L'effet des mutations est tiré au hasard dans une distribution normale de moyenne nulle et d'écart-type SD $= 0.05$. Pour cette simulation, le coût de la dispersion était fixé à $c_{\mathrm{z}} = 0.2$, celui de la dormance à $c_{\mathrm{d}} = 0.025$, et le taux d'extinction locale à $e = 0$ (pas d'extinction). Cette simulation est partie d'un état initial monomorphe, où toutes les valeurs de traits étaient fixées à 0.2 pour l'ensemble des individus. Les lignes pointillées verticales indiquent les valeurs évolutivement stables des traits, calculées dans notre modèle.

les germinations issues de la génération précédente et les germinations de graines philopatriques dormantes. Les différents types d'individus (graines, adultes) et les différentes catégories de dèmes définissent ainsi des classes démographiques dans notre modèle. Dans les modèles structurés en classes,

les différentes classes démographiques apportent des contributions génétiques différentes à la population dans son ensemble. Il est néanmoins possible de calculer les changements de fréquence allélique en une génération, et donc le gradient de sélection, en considérant la fréquence allélique dans l'équation (3.1) comme une moyenne des fréquences alléliques dans les différentes classes, pondérées par les valeurs reproductives de chaque classe (Rousset 2004; Taylor 1990).

Il existe une difficulté supplémentaire, inhérente à notre modèle, venant du fait que la régulation de la densité se fait chez les adultes, mais pas chez les graines dormantes de la banque : voir l'étape (v) du cycle de vie ci-dessus. Le nombre de graines dans la banque est donc une variable aléatoire qui dépend du taux de dormance, c'est-à-dire du trait dont on cherche à déterminer la dynamique évolutive. Il faudrait donc, en théorie, prendre en compte l'effet du phénotype sur le changement de fréquence de l'allèle codant pour le trait, à travers l'effet du phénotype sur la démographie des populations locales (Rousset et Ronce 2004). Or la variation du nombre de graines dormantes entre populations génère un très grand nombre d'états démographiques possibles pour chacune des catégories de dèmes. Tenir compte de ces fluctuations démographiques dans le calcul est extrêmement difficile et nous empêche donc de trouver une solution analytique exacte. En revanche, il est possible d'approximer la distribution des effectifs de graines dans les banques par son espérance (qui s'exprime en fonction du taux de dormance et des autres paramètres du modèle). Les résultats de simulations individus-centrées nous montrent que cette approximation est remarquablement robuste (Figure 3.3).

Les stratégies évolutivement stables (ESS) pour chaque trait sont obtenues numériquement à partir du calcul du signe du gradient de sélection $S(z^*)$ au voisinage de z^*, en considérant les autres traits fixés. Une stratégie z^* est candidate à l'ESS si $S(z^*) = 0$. Cette stratégie est stable par convergence si $S(z^*) > 0$ pour $z < z^*$ et $S(z^*) < 0$ pour $z > z^*$. De cette façon, la population évolue jusqu'à ce qu'elle atteigne le point singulier z^* où la sélection ne s'exerce plus. Nous n'avons pas caractérisé formellement la stabilité évolutive (ce qui aurait nécessité le calcul des dérivées secondes de la valeur sélective, voir Ajar 2003; Eshel 1996; Geritz *et al.* 1998), mais nous

avons vérifié grâce à des simulations individus-centrées que les stratégies que nous trouvions étaient bien stables par convergence et évolutivement stables. S'agissant de l'évolution conjointe de tous les traits, les stratégies ESS sont déterminées en annulant les gradients de sélection simultanément pour tous les traits.

3.2.3 Évolution de la dormance

Nous avons tout d'abord pu démontrer que, dans le cas de la dormance conditionnelle, la dormance des graines dispersées n'évolue pas : les graines dispersées n'ont aucun intérêt évolutif à entrer dans une phase de dormance et ont tout intérêt au contraire à germer immédiatement. Ce résultat est en accord avec la biologie de certaines espèces végétales qui ont été décrites comme étant hétéromorphes, ce qui signifie qu'un seul individu peut produire des graines morphologiquement différenciées (McPeek et Kalisz 1998; Olivieri *et al.* 1983; Venable 1985). On trouve principalement ces espèces dans la famille des Asteraceae et Chenopodiaceae (Imbert 2002), et les données disponibles semblent soutenir notre prédiction puisqu'en général les espèces hétéromorphes produisent en effet des graines qui sont dispersées et qui germent immédiatement, ainsi que des graines qui ne sont pas dispersées et qui ont une certaine probabilité d'entrer en dormance (Olivieri 2001).

En l'absence de variation environnementale (c'est-à-dire en l'absence d'extinction) la dormance des graines diminue lorsque la dispersion augmente. Ceci est vrai pour la dormance conditionnelle (et dans ce cas, il s'agit de la dormance des graines philopatriques) et pour la dormance non-conditionnelle (et dans ce cas, il s'agit de la dormance des graines philopatriques et dispersées). Avec de la variation environnementale (c'est-à-dire en présence d'extinctions locales), et dans le cas de la dormance conditionnelle, la dormance des graines philopatriques augmente avec le taux d'extinction (Figure 3.4A). Lorsque la dispersion est faible (et que le modèle converge donc vers un modèle de population unique isolée), que les tailles des populations sont grandes (et que l'on néglige donc l'effet de la compétition entre apparentés), alors

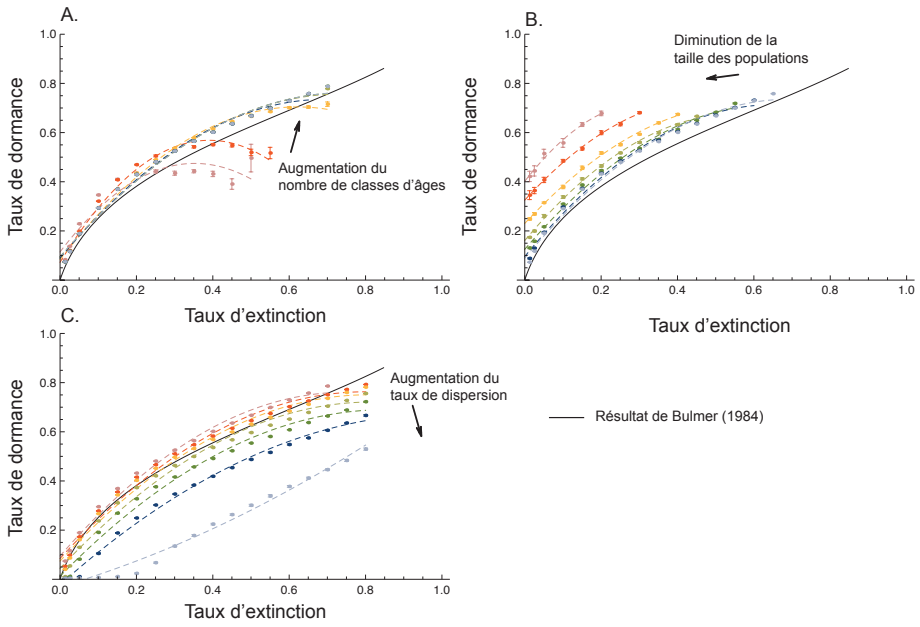


Figure 3.4: (A) Taux de dormance évolutivement stable pour les graines philopatriques (dormance conditionnelle) lorsque la dispersion est un paramètre fixé, en fonction du taux d'extinction et pour un nombre de classes d'âge dans la banque de graines variant de 1 à 100. Ces courbes ont été obtenues à partir de simulations individus-centrées, pour des grandes populations ($N = 100$) et un taux de fécondité élevé ($r = 100$). Le taux de dispersion a été fixé à une valeur très faible, de sorte que le nombre de migrants par génération est égal à 0.0001. Les valeurs des autres paramètres sont $c_{\mathrm{d}} = 0.2$ et $c_{\mathrm{z}} = 0.5$. (B) Taux de dormance évolutivement stable pour les graines philopatriques, en fonction du taux d'extinction et de la taille de la population variant de 1 à 100 (50 classes d'âge dans la banque). (C) Taux de dormance évolutivement stable pour les graines philopatriques, en fonction du taux d'extinction et du nombre de migrants par génération variant de 0.01 à 5 (50 classes d'âge dans la banque). La ligne noire indique la solution de (Bulmer 1984) pour un modèle de population unique, avec densité-dépendance, mais sans compétition entre apparentés.

notre modèle converge vers celui de Bulmer (1984) lorsque l'on considère un nombre élevé de classes d'âge dans la banque. Diminuer la taille des populations locales (et donc augmenter la compétition entre individus apparentés)

tend à augmenter le taux de dormance évolutivement stable (Figure 3.4B). Enfin, lorsque l'on fait augmenter le taux de dispersion des graines, alors le taux de dormance évolutivement stable tend à diminuer (Figure 3.4C). Dans le cas de la dormance non-conditionnelle, en revanche, les pressions de sélection antagonistes qui s'exercent sur la stratégie de dormance des graines dispersées conduisent à des résultats non triviaux : pour de faibles taux d'extinction, la dormance non-conditionnelle est sélectionnée positivement, car elle procure un moyen de recoloniser un dème vide à partir de la banque de graines. Mais lorsque les extinctions locales deviennent plus fréquentes, la dormance des graines est sélectionnée négativement, car les graines dispersées qui colonisent un dème vide n'ont aucun intérêt à retarder leur germination. Ce résultat n'est cependant pas robuste à une augmentation de la longévité des graines dans la banque.

3.2.4 Évolution conjointe de la dormance et de la dispersion

Nous nous sommes intéressés à l'évolution conjointe de la dormance et de la dispersion, et nous avons pu montrer l'émergence de corrélations négatives entre ces traits lorsque l'on fait varier leurs coûts directs : augmenter le coût de la dispersion c_z tend à diminuer le taux de dispersion évolutivement stable et à augmenter celui de la dormance. Au contraire, augmenter le coût de la dormance c_d tend à diminuer le taux de dormance évolutivement stable et à augmenter celui de la dispersion. Ceci est vérifié pour la dormance conditionnelle et non-conditionnelle, quel que soit le nombre de classes d'âge dans la banque de graines.

La Figure 3.5 montre les taux évolutivement stables de dormance (conditionnelle et non-conditionnelle) et de dispersion lorsque les tailles de population et les taux d'extinction locale varient. En l'absence d'extinctions, la corrélation entre les taux de dormance et de dispersion est positive lorsque la taille des populations varie : la dormance et la dispersion tendent à augmenter lorsque la taille des populations diminue, ce qui est la conséquence de la compétition entre individus apparentés. Toutefois, cette corrélation peut

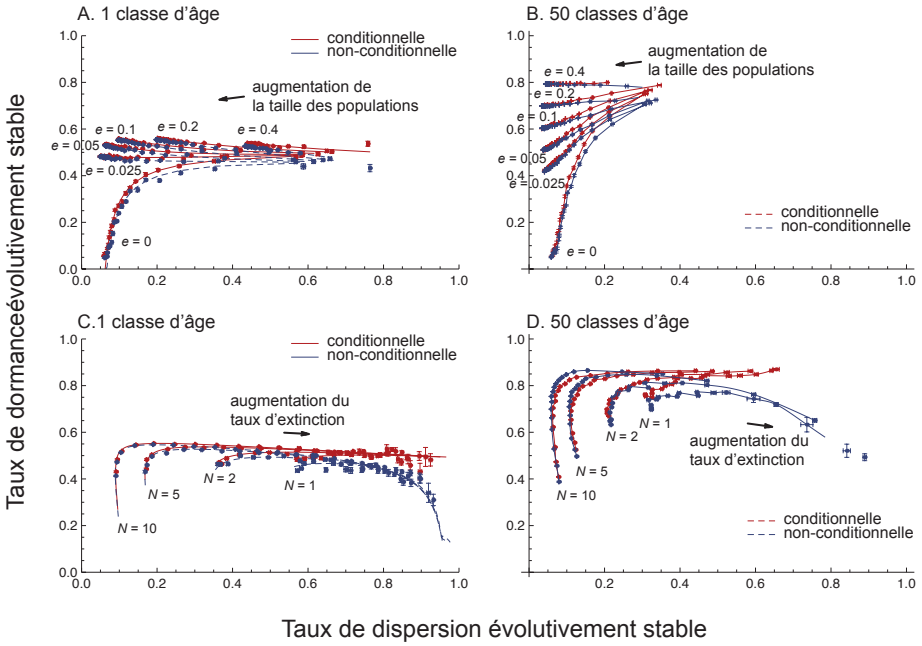


Figure 3.5: Taux évolutivement stables de dormance (conditionnelle et non-conditionnelle) et de dispersion. Les courbes rouges donnent les résultats pour le modèle avec une dormance conditionnelle, et les courbes bleues donnent les résultats pour le modèle avec une dormance non-conditionnelle. (A) Évolution conjointe en fonction du nombre d'adultes (N) qui varie de 1 à 20, pour différentes valeurs du taux d'extinction (e variant de 0 à 0.4). La flèche indique le sens de l'augmentation de N. Les autres valeurs de paramètres sont : $c_{\mathrm{d}} = 0.025$, $c_{\mathrm{z}} = 0.4$. (B) Comme en (A) avec 50 classes d'âge dans la banque de graines. (C) Évolution conjointe en fonction du taux d'extinction (e) qui varie de 0.005 à 0.995, pour différentes tailles de population (N variant de 1 à 10). La flèche indique le sens de l'augmentation de e. Les autres valeurs de paramètres sont : $c_{\mathrm{d}} = 0.025$, $c_{\mathrm{z}} = 0.4$. (D) Comme en (C) avec 50 classes d'âge dans la banque de graines.

devenir légèrement négative si le taux d'extinction augmente (Figure 3.5A). Cette dernière tendance est moins prononcée dans les simulations individus-centrées avec 50 classes d'âge dans la banque de graines (Figure 3.5B). On observe une relation "en cloche" entre les taux évolutivement stables de dormance (conditionnelle et non-conditionnelle) et de dispersion, lorsque le taux

d'extinction varie pour une population de taille donnée (Figure 3.5C). Ceci suggère que des corrélations négatives (pour de faibles taux d'extinction) et positives (pour des taux d'extinction intermédiaires) peuvent émerger. Ce résultat est confirmé par des simulations individus-centrées avec 50 classes d'âge dans la banque de graines (Figure 3.5D).

L'absence de tendances générales comme, par exemple, une corrélation négative entre taux évolutivement stables de dormance et de dispersion lorsque le taux d'extinction ou la taille de la population varient, nous laisse penser que la dispersion et la dormance ne peuvent pas être simplement considérées comme des stratégies véritablement alternatives pour réduire le risque d'extinction locale et le coût de la compétition entre individus apparentés (voir la Figure 3.5). La relation entre ces traits dépend donc des caractéristiques de l'environnement (voir aussi Cohen et Levin 1987). D'autres modèles ont montré l'existence de corrélations positives entre ces traits (Cohen et Levin 1987, 1991; Snyder 2006; Venable et Brown 1988), mais seulement si les variations de l'environnement sont corrélées dans le temps. L'intérêt de notre modèle a donc été de montrer que des corrélations positives entre ces traits peuvent émerger, en réponse à la variation de l'environnement et à la compétition entre individus apparentés, même en l'absence de corrélations temporelles de l'environnement.

Perspectives

Histoire démographique des populations

Parmi mes perspectives de recherche à court terme, j'aimerais approfondir les approches consistant à analyser le polymorphisme conjointement sur différents types de marqueurs (chromosomes X et Y, autosomes dans un modèle animal à sexes séparés). Outre la finalisation de l'approche ABC en cours de développement avec Raphaël Leblois (voir § 1.2.2), nous examinons avec Mathieu Gautier une piste consistant à analyser conjointement ce type de marqueurs dans le cadre du modèle d'inférence de l'histoire des populations, basé sur les équations de diffusion de Kimura (1964) (voir § 1.4). Nous pensons en effet que ce modèle pourrait être étendu au cas des espèces à sexes séparés dont on cherche à analyser conjointement le polymorphisme autosomal et celui lié au chromosome X. Cette approche se situerait dans la lignée des travaux effectués avec Laure Ségurel (voir l'article [P15], reproduit en Annexe C, page 131), tout en bénéficiant d'une approche bayésienne prenant en compte l'ensemble de l'information contenue dans les données de polymorphisme.

Supposons par exemple qu'une population k soit composée de $N_{♀,k}$ femelles (de caryotype XX) et $N_{♂,k}$ mâles (de caryotype XY). Le nombre total d'individus dans la population k est alors $N_k = \left(N_{♀,k} + N_{♂,k}\right)$. La taille efficace de la population k (exprimée en nombre d'individus diploïdes) vaut, pour des marqueurs autosomaux :

$$N_{e,k}^{(\text{auto})} = 2\left(\frac{4N_{♀,k}N_{♂,k}}{N_{♀,k} + N_{♂,k}}\right) \tag{3.2}$$

(voir par exemple, Rousset 2004, page 159). Si l'on note $r_k \equiv N_{\varphi,k}/N_k$ le sexe-ratio dans la population k, alors le paramètre de longueur de branche $\tau_k^{(\text{auto})}$ qui représente l'intensité de la dérive s'exerçant sur des marqueurs autosomaux le long de la branche menant à la population k prend la forme :

$$\tau_k^{(\text{auto})} \equiv \frac{t}{2N_{\text{e},k}^{(\text{auto})}} = \frac{t}{16 r_k \left(1 - r_k\right) N_k} \tag{3.3}$$

Pour des marqueurs liés au chromosome X, en revanche,

$$N_{\text{e},k}^{(\text{X})} = 2\left(\frac{9 N_{\varphi,k} N_{\sigma,k}}{2N_{\varphi,k} + 3N_{\sigma,k}}\right) \tag{3.4}$$

(voir par exemple, Rousset 2004, page 159), et le paramètre de longueur de branche $\tau_k^{(\text{X})}$ qui représente l'intensité de la dérive s'exerçant sur des marqueurs liés au chromosome X le long de la branche menant à la population k prend la forme :

$$\tau_k^{(\text{X})} \equiv \frac{t}{2N_{\text{e},k}^{(\text{X})}} = \frac{t}{6\frac{r_k(1-r_k)}{(2-r_k)} N_k} \tag{3.5}$$

Il s'agirait alors de substituer dans l'équation 1.14 (voir page 34) le produit :

$$\prod_{j=1}^{L} f\left(p_{ij} \mid p_{a(i)j}, \tau_i\right) \tag{3.6}$$

par un double produit de la forme :

$$\prod_{j=1}^{L^{(\text{auto})}} f\left(p_{ij}^{(\text{auto})} \mid p_{a(i)j}^{(\text{auto})}, \tau_i^{(\text{auto})}\right) \prod_{j=1}^{L^{(\text{X})}} f\left(p_{ij}^{(\text{X})} \mid p_{a(i)j}^{(\text{X})}, \tau_i^{(\text{X})}\right) \tag{3.7}$$

L'échantillonnage dans la distribution 1.14 par chaîne de Markov Monte Carlo permettrait alors d'estimer à la fois les temps de divergence, mais aussi les sexe-ratios dans chaque branche de l'arbre des populations, à partir de l'analyse conjointe de marqueurs autosomaux et de marqueurs liés au chromosome X.

Histoire adaptative des populations et traits d'histoire de vie

La méthode d'inférence de l'intensité de la sélection que j'ai développée ne prend pas en compte la proximité physique des marqueurs le long du chromosome, dont la conséquence est la mise en place d'un déséquilibre de liaison entre marqueurs. Or, cette dépendance spatiale des marqueurs entre eux peut être importante dans un contexte de génotypage haut-débit. Il s'agit donc désormais de réfléchir à de nouvelles approches pour prendre en compte le déséquilibre de liaison entre marqueurs dans ce type de modèles. D'autre part, la démographie (comme des expansions spatiales marquées) peut entrainer un fort taux de faux positifs (Foll et Gaggiotti 2008) et la procédure de choix de marqueurs peut également introduire des biais dans ce type d'analyse, qui restent à caractériser et à prendre en compte. Considérer un modèle de populations en divergence (tel que celui considéré au § 1.4), plutôt que considérer un modèle à l'équilibre migration–dérive, pourrait être un moyen de mieux tenir compte de l'histoire des populations. Enfin, nous pourrions également envisager d'introduire des modèles de sélection alternatifs, par exemple pour mieux caractériser la sélection balancée.

Mais au delà de la poursuite de mon activité dans le développement de méthodes d'analyse de la variabilité génétique pour inférer des paramètres démographiques, et notamment dans le contexte de l'explosion des données de génomique des populations, je souhaite également élargir mes recherches dans la caractérisation de l'architecture génomique de l'adaptation. Ces perspectives se nourrissent des collaborations que j'ai pu mettre en place depuis mon arrivée au CBGP, car l'application des méthodes que j'ai développées à divers organismes et problématiques me conduit naturellement à m'intéresser de plus près aux questions d'adaptation locale, de spécialisation, voire de spéciation, notamment chez les insectes phytophages. Dans ce contexte, je m'intéresse en particulier à deux modèles d'étude (le puceron du pois et la processionnaire du pin) à travers des collaborations en cours avec Carole Smadja (ISE-M, Montpellier) [P31] et Jean-Christophe Simon (IGEPP, Rennes) [R1] sur le puceron et Carole Kerdelhué (CBGP, Montpellier)

sur la processionnaire.

Chez le puceron du pois *Acyrthosiphon pisum*, notamment, l'adaptation et la spécialisation à différentes plantes-hôtes est l'élément principal de réduction du flux génique entre populations (Peccoud *et al.* 2010). Conséquence de cette spécialisation, les pucerons ont développé la capacité à reconnaître et préférer leur plante-hôte pour se nourrir et se reproduire, ce qui induit de l'homogamie au sein de chaque race d'hôtes et donc un renforcement de l'isolement reproducteur entre races d'hôtes (Via 1999). Il existe en Europe une série de biotypes spécialisés sur différentes plantes-hôtes présentant des degrés de divergence variés, depuis des races d'hôtes faiblement différenciées jusqu'à des espèces complètement isolées (Peccoud *et al.* 2009). Ce "continuum de spéciation" fait du puceron du pois un modèle pertinent pour estimer la taille des îlots de différenciation (les régions génomiques où la différenciation est significativement plus élevée que sous l'attendu d'un équilibre entre forces évolutives neutres) le long de ce continuum, information qui peut être utilisée pour reconstruire la dynamique de la différenciation génomique au cours d'une spéciation. Ce contexte biologique, ainsi que l'importante quantité de ressources génomiques disponibles chez ce puceron (dont un génome de référence) offrent les conditions idéales pour aborder la spéciation écologique sous l'angle de l'analyse des génomes. Plusieurs études récentes auxquelles j'ai participé [P31–R1], ont permis de révéler le rôle potentiel de gènes des récepteurs olfactifs et gustatifs dans la divergence entre races d'hôtes. L'identification de ces "gènes barrières" évoluant sous sélection divergente permet d'envisager la caractérisation de la différenciation génomique aux alentours de ces gènes, et ainsi tester l'impact de la sélection divergente sur la taille des îlots de différenciation. Parallèlement, des approches théoriques peuvent permettre de mieux comprendre comment la divergence peut progresser à partir de régions très localisées dans le génome jusqu'à l'ensemble du génome, et ainsi prédire l'étendue des fenêtres de différenciation autour de ces locus barrières. Ce projet est étroitement lié au projet ANR Speciaphid (porteur : Jean-Christophe Simon, 2011–2014) auquel je participe.

Chez la processionnaire du pin *Thaumetopoea pityocampa*, un lépidoptère ravageur des pinèdes, les variations de la phénologie peuvent limiter les flux

de gènes entre populations à travers la mise en place d'un décalage temporel de l'activité des adultes. Des différences de phénologie peuvent ainsi être impliquées dans des phénomènes de différenciation génétique en sympatrie. Or la découverte il y a 15 ans d'une population à cycle décalé, dans la forêt de Leiria au Portugal, offre une opportunité unique de détecter les régions du génome impliquées dans la phénologie et donc de mieux comprendre l'architecture génétique de la phénologie. L'existence même de cette population en sympatrie avec une population de la même espèce dont le cycle est identique aux autres populations connues (reproduction en août et développement larvaire en hiver), suggère un cas de spéciation allochronique en cours (spéciation sympatrique liée à un décalage dans le temps de la reproduction sexuée). Dans ce contexte, la caractérisation de l'histoire démographique et adaptative de ces populations du Portugal pourrait permettre de mieux comprendre la dynamique de la spéciation chez cette espèce. Pour cela, une approche basée sur la modélisation de l'évolution de la phénologie (voir, par exemple, Devaux et Lande 2008) pourrait être pertinente pour appuyer les analyses du polymorphisme génétique présumé neutre. Une telle approche pourrait être le moyen de réunir dans une même problématique les champs de recherche développés dans les chapitres 2 et 3. Ce projet est étroitement lié au projet ANR GenoPheno (porteur : Carole Kerdelhué, 2010–2014) auquel je participe.

Bibliographie

Abramowitz, M. et I. A. Stegun (1965) *Handbook of Mathematical Functions*. Dover Publication, Inc., New York.

Ajar, E. (2003) Analysis of disruptive selection in subdivided populations. *BMC Evolutionary Biology*, **3** : 22.

Akey, J. M., G. Zhang, K. Zhang, L. Jin et M. D. Shriver (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12** : 1805–1814.

Baird, N. A., P. D. Etter, T. S. Atwood, A. L. Shiver, L. Z. A., E. U. Selker, W. A. Cresko et E. A. Johnson (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3** : e3376.

Barton, N. H. et M. Turelli (1989) Evolutionary quantitative genetics : how little do we know ? *Annual Review of Genetics*, **23** : 337–370.

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

Beaumont, M. A. (2010) Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology Evolution and Systematics*, **41** : 379–406.

Beaumont, M. A. et D. J. Balding (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13** : 969–980.

Beaumont, M. A. et R. A. Nichols (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London - Series B : Biological Sciences*, **263** : 1619–1626.

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

Bengtsson, B. O. (1978) Avoiding inbreeding : at what cost ? *Journal of Theoretical Biology*, **73** : 439–444.

Bersaglieri, T., P. Sabeti, N. Patterson, T. Vanderploeg, S. Schaffner, J. Drake, M. Rhodes, D. Reich et J. Hirschhorn (2004) Genetic signatures of strong recent positive selection at the lactase gene. *American Journal of Human Genetics*, **74** : 1111–1120.

Black, W. C., C. F. Baer, M. F. Antolin et N. M. DuTeau (2001) Population genomics : genome-wide sampling of insect populations. *Annual Review of Entomology*, **46** : 441–469.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto, K. K. Kidd et L. L. Cavalli-Sforza (1991) Drift, admixture, and selection in human evolution : a study with DNA polymorphisms. *Proceedings of the National Academy of Sciences of the United States of America*, **88** : 839–843.

Bulmer, M. (1984) Delayed germination of seeds : Cohen's model revisited. *Theoretical Population Biology*, **26** : 367–377.

Bürger, R. (2000) *The Mathematical Theory of Selection, Recombination, and Mutation*. John Wiley and sons, Chichester.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang,

J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset et L. L. Cavalli-Sforza (2002) A human genome diversity cell line panel. *Science*, **296** : 261–262.

Cann, R. L. (2001) Genetic clues to dispersal in human populations : retracing the past from the present. *Science*, **291** : 1742–1748.

Cavalli-Sforza, L. L. (1966) Population structure and human evolution. *Proceedings of the Royal Society of London - Series B : Biological Sciences*, **164** : 362–379.

Cavalli-Sforza, L. L., P. Menozzi et A. Piazza (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton.

Cohen, D. (1966) Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, **12** : 119–129.

Cohen, D. et S. A. Levin (1987) The interaction between dispersal and dormancy strategies in varying and heterogeneous environments. In : *Mathematical Topics in Population Biology, Morphogenesis and Neurosciences* (eds. Teramato, E. et M. Yomaguti), pp. 110–122. Lecture notes. Biomathematics, Kyoto.

Cohen, D. et S. A. Levin (1991) Dispersal in patchy environments : the effects of temporal and spatial structure. *Theoretical Population Biology*, **39** : 36–99.

Colagiuri, S. et J. Brand Miller (2002) The 'carnivore connection'–evolutionary aspects of insulin resistance. *European Journal of Clinical Nutrition*, **56** : S30–35.

Comins, H. N., W. D. Hamilton et R. M. May (1980) Evolutionarily stable dispersal strategies. *Journal of Theoretical Biology*, **82** : 205–230.

Cordaux, R., E. Deepa, H. Vishwanathan et M. Stoneking (2004) Genetic evidence for the demic diffusion of agriculture to India. *Science*, **304** : 1125.

Cornuet, J.-M., F. Santos, M. A. Beaumont, C. P. Robert, J.-M. Marin, D. J. Balding, T. Guillemaud et A. Estoup (2008) Inferring population history with DIY ABC : a user-friendly approach to approximate bayesian computation. *Bioinformatics*, **24** : 2713–2719.

Crow, J. F. et M. Kimura (1971) *An Introduction to Population Genetics Theory*. The Blackburn Press, ? ? ?

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

Devaux, C. et R. Lande (2008) Incipient allochronic speciation due to non-selective assortative mating by flowering time, mutation and genetic drift. *Proceedings of the Royal Society of London - Series B : Biological Sciences*, **275** : 2723–2732.

Disotell, T. R. (1999) Sex-specific contributions to genome variation. *Current Biology*, **9** : R29–R31.

Donnelly, P. et S. Tavaré (1995) Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*, **29** : 401–21.

Edwards, A. W. F. (1992) *Likelihood*. The Johns Hopkins University Press, ? ? ?, 2nd edn.

Ellner, S. (1985a) ESS germination strategies in randomly varying environments. I. Logistic-type models. *Theoretical Population Biology*, **28** : 50–79.

Ellner, S. (1985b) ESS germination strategies in randomly varying environments. II. Reciprocal yield-law models. *Theoretical Population Biology*, **28** : 80–116.

Ellner, S. (1986) Germination dimorphisms and parent offspring conflicts in seed-germination. *Journal of Theoretical Biology*, **123** : 173–185.

Eshel, I. (1996) On the changing concept of evolutionary population stability as a reflection of a changing point of view in the quantitative theory of evolution. *Journal of Mathematical Biology*, **34** : 485–510.

Ethier, S. N. et T. Nagylaki (1988) Diffusion approximations of markov chains with two time scales and application to population genetics, II. *Advances in Applied Probabilities*, **20** : 525–545.

Ewens, W. J. (2004) *Mathematical Population Genetics*. Springer, ? ? ?, 2nd edn.

Excoffier, L., P. Smouse et J. M. Quattro (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes. *Genetics*, **131** : 479–491.

Fagundes, N. J. R., N. Ray, M. A. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto et L. Excoffier (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of sciences USA*, **104** : 17614–17619.

Favre, L., F. Balloux, J. Goudet et N. Perrin (1997) Female-biased dispersal in the monogamous mammal *Crocidura russula* : evidence from field data and microsatellite patterns. *Proceedings of the Royal Society of London - Series B : Biological Sciences*, **264** : 127–32.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letter*, **27** : 882–891.

Felsenstein, J. (2003) *Inferring Phylogenies*. Sinauer Associates, ? ? ?, 2nd edn.

Flori, L., F. Fritz, S. ad Jaffrézic, M. Boussaha, I. Gut, S. Heath, F.-L. Foulley et M. Gautier (2009) The genome response to artificial selection : a case study in dairy cattle. *PLoS ONE*, **4** : e6595.

Foll, M. et O. Gaggiotti (2008) A genome scan method to identify selected loci appropriate for both dominant and codominant markers : a Bayesian perspective. *Genetics*, **180** : 977–993.

Fontanillas, P., E. Petit et N. Perrin (2004) Estimating sex-specific dispersal rates with autosomal markers in hierarchically structured populations. *Evolution*, **58** : 8886–894.

Frank, S. A. (1986) Dispersal polymorphisms in subdivided populations. *Journal of Theoretical Biology*, **122** : 303–309.

Frankham, R., D. Briscoe et J. Ballou (2002) *Introduction to Conservation Genetics*. Cambridge University Press, New York.

Freeman, H. et R. D. Cox (2006) Type-2 diabetes : a cocktail of genetic discovery. *Human Molecular Genetics*, **15** : R202–209.

Gandon, S. (1999) Kin competition, the cost of inbreeding and the evolution of dispersal. *Journal of Theoretical Biology*, **200** : 345–364.

Gandon, S. et Y. Michalakis (1999) Evolutionary stable dispersal rate in a metapopulation with extinctions and kin competition. *Journal of Theoretical Biology*, **199** : 275–290.

Gandon, S. et Y. Michalakis (2001) Multiple causes of the evolution of dispersal. In : *Dispersal* (eds. Clobert, J., E. Danchin, A. A. Dhondt et N. J. D.), pp. 155–167. Oxford University Press, Oxford.

Gelman, A., J. B. Carlin, H. S. Stern et D. B. Rubin (2004) *Bayesian Data Analysis*. Chapman & Hall, New York, 2nd edn.

Geritz, S. A. H., É. Kidsi, G. Meszéna et J. A. J. Metz (1998) Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. *Evolutionary Ecology*, **12** : 35–57.

Gilks, W. R., S. Richardson et D. J. Spiegelhalter (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, New York, 2nd edn.

Goudet, J., N. Perrin et P. Waser (2002) Tests for sex-biased dispersal using bi-parentally inherited genetic markers. *Molecular Ecology*, **11** : 1103–1114.

Greenwood, P. J. (1980) Mating systems, philopatry and dispersal in birds and mammals. *Animal Behaviour*, **28** : 1140–1162.

Guo, F., D. Dey et H. K.E. (2009) A bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association*, **104** : 142–154.

Hamilton, W. D. (1964) The genetical evolution of social behavior. *Journal of Theoretical Biology*, **7** : 1–16.

Hamilton, W. D. et R. M. May (1977) Dispersal in stable habitats. *Nature*, **269** : 578–581.

Harding, R. M. et G. McVean (2004) A structured ancestral population for the evolution of modern humans. *Current Opinion in Genetics and Developments*, **14** : 667–674.

Harper, J. L. (1977) *Population Biology of Plants*. Academic Press, London.

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

Hein, J., M. H. Schierup et C. Wiuf (2005) *Gene Genealogies, Variation and Evolution : A Primer in Coalescent Theory*. Oxford University Press, USA.

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

Heyer, E., A. Sibert et F. Austerlitz (2005) Cultural transmission of fitness : genes take the fast lane. *Trends in Genetics*, **21** : 234–239.

Hu, F. S., A. Hampe et R. J. Petit (2009) Paleoecology meets genetics : deciphering past vegetational dynamics. *Frontiers in Ecology and the Environment*, **7** : 371–379.

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

Imbert, E. (2002) Ecological consequences and ontogeny of seed heteromorphism. *Perspectives in Plant Ecology, Evolution and Systematics*, **5** : 13–36.

Innan, H. et Y. Kim (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of sciences USA*, **101** : 10667–10672.

Karafet, T., L. P. Xu, R. F. Du, W. Wang, S. Feng, R. S. Wells, A. J. Redd, S. L. Zegura et M. F. Hammer (2001) Paternal population history of east asia : Sources, patterns, and microevolutionary processes. *American Journal of Human Genetics*, **69** : 615–628.

Kimura, M. (1964) Diffusion models in population genetics. *Journal of Applied Probability*, **1** : 177–232.

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*, **19A** : 27–43.

Klinkhamer, P., T. de Jong, J. A. Metz et J. Val (1987) Life history tactics of annual organisms : the joint effect of dispersal and delayed germination. *Theoretical Population Biology*, **32** : 127–156.

Kobayashi, Y. et N. Yamamura (2000) Evolution of seed dormancy due to sib competition : effect of dispersal and inbreeding. *Journal of Theoretical Biology*, **202** : 11–24.

Lawton-Rauh, A. (2008) Demographic processes shaping genetic variation. *Current Opinion in Plant Biology*, **11** : 103–109.

Leblois, R., A. Estoup et F. Rousset (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.

Lehman, C. L. et D. Tilman (1997) Competition in spatial habitats. In : *Spatial Ecology : The Role of Space in Population Dynamics and Interspecific Interactions* (eds. Tilman, D. et P. Kareiva), pp. 185–203. Princeton University Press, Princeton.

Levin, S. A., D. Cohen et A. Hastings (1984) Dispersal strategies in patchy environments. *Theoretical Population Biology*, **26** : 165–191.

Levins, R. (1968) *Evolution in Changing Environments*. Monographs in Population Biology 2. Princeton University Press, Princeton.

Lewis, Z. A., A. L. Shiver, N. Stffler, M. R. Miller, E. A. Johnson et E. U. Selker (2007) High-density detection of restriction-site-associated DNA markers for rapid mapping of mutated loci in *Neurospora*. *Genetics*, **177** : 1163–1171.

Lewontin, R. C. et J. Krakauer (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. *Genetics*, **74** : 175–195.

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

Marjoram, P. et S. Tavaré (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, **7** : 759–770.

Maynard Smith, J. (1982) *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Maynard Smith, J. et J. Haigh (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23** : 23–35.

McPeek, M. et S. Kalisz (1998) The joint evolution of dispersal and dormancy in metapopulations. *Archive für Hydrobiologie*, **52** : 33–51.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller et E. Teller (1953) Equations of state calculations bywhere fast computing machines. *Journal of Chemical Physics*, **21** : 1087–1091.

Miller, M. R., T. S. Atwood, B. F. Eames, J. K. Eberhart, Y.-L. Yan, J. H. Postlethwait et E. A. Johnson (2007a) RAD marker microarrays enable rapid mapping of zebrafish mutations. *Genome Biology*, **8** : R105.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko et E. A. Johnson (2007b) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17** : 240–248.

Motro, U. (1991) Avoiding inbreeding and sibling competition : the evolution of sexual dimorphism for dispersal. *The American Naturalist*, **137** : 108–115.

Neel, J. V. (1962) Diabetes mellitus : a "thrifty" genotype rendered detrimental by "progress" ? *American Journal of Human Genetics*, **14** : 353–362.

Neel, J. V. (1992) The thrifty genotype revisited. In : *The Genetics of Diabetes Mellitus* (eds. Kobberlong, J. et R. Tattersall), pp. 283–293. London Academic Press, London.

Nei, M. et A. K. Roychoudhury (1993) Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution*, **10** : 927–943.

Nelson, M. R., K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, P. Vollenweider, J. R. Oksenberg, S. L. Hauser, H. A. Stirnadel, J. S. Kooner, J. C. Chambers, B. Jones, V. Mooser, C. D. Bustamante, A. D. Roses, D. K. Burns, M. G. Ehm et E. H. Lai (2008) The Population Reference Sample, POPRES : a resource for population, disease, and pharmacological genetics research. *American Journal of Human Genetics*, **83** : 347–358.

Nielsen, R. (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39** : 197–218.

Nielsen, R. et J. Wakeley (2001) Distinguishing migration from isolation : a Markov chain Monte Carlo approach. *Genetics*, **158** : 885–896.

Ntzoufras, I. (2009) *Bayesian Modeling Using WinBugs*. John Wiley & Sons., Hoboken.

Olivieri, I. (2001) The evolution of seed heteromorphism in a metapopulation : interactions between dispersal and dormancy. In : *Integrating Ecology and Evolution in a Spatial Context* (eds. Silvertown, J. et J. Antonovics), pp. 245–268. Blackwell Science, Oxford.

Olivieri, I. et P.-H. Gouyon (1995) Metapopulation genetics and the evolution of dispersal. *The American Naturalist*, **146** : 202–228.

Olivieri, I., M. Swan et P.-H. Gouyon (1983) Reproductive system and colonizing strategy of two species of *Carduus* (Compositae). *Oecologia*, **60** : 114–117.

Oota, H., W. Settheetham-Ishida, D. Tiwawech, T. Ishida et M. Stoneking (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics*, **29** : 20–21.

Ordovas, J. M. et D. Corella (2004) Nutritional genomics. *Annual Review of Genomics and Human Genetics*, **5** : 71–118.

Peccoud, J., A. Ollivier, M. Plantegenest et J.-C. Simon (2009) continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proceedings of the National Academy of sciences USA*, **106** : 7495–7500.

Peccoud, J., A. Ollivier, M. Plantegenest et J.-C. Simon (2010) The pea aphid complex as a model of ecological speciation. *Ecological Entomology*, **35** : 119–130.

Pennisi, E. (2001) Tracking the sexes by their genes. *Science*, **291** : 1733–1734.

Perrin, N. et V. Mazalov (1999) Dispersal and inbreeding avoidance. *The American Naturalist*, **154** : 282–292.

Perrin, N. et V. Mazalov (2000) Local competition, inbreeding, and the evolution of sex-biased dispersal. *American Naturalist*, **155** : 116–127.

Philippi, T. et J. Seger (1989) Hedging one's evolutionary bets, revisited. *Trends in Ecology and Evolution*, **4** : 41–44.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun et M. W. Feldman (1999) Population growth of human Y chromosomes : a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16** : 1791–1798.

Pritchard, J. K., M. Stephens et P. Donnelly (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155** : 945–959.

Przeworski, M., R. R. Hudson et A. Di Rienzo (2000) Adjusting the focus on human variation. *Trends in Genetics*, **16** : 296–302.

Przeworsky, M., G. Coop et J. Wall (2005) The signature of positive selection on standing variation. *Evolution*, **59** : 2312–2323.

Riebler, A., L. Held et W. Stephan (2008) Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, **178** : 1817–1829.

Romero, I. G., C. B. Mallick, A. Liebert, F. Crivellaro, G. Chaubey, Y. Itan, M. Metspalu, M. Eaaswarkhanth, R. Pitchappan, R. Villems, D. Reich, L. Singh, K. Thangaraj, M. G. Thomas, D. M. Swallow, M. M. Lahr et T. Kivisild1 (2012) Herders of indian and european cattle share their predominant allele for lactase persistence. *Molecular Biology and Evolution*, **29** : 249–260.

Ronce, O. (2007) How does it feel to be like a rolling stone ? Ten questions about dispersal evolution. *Annual Review of Ecology, Evolution, and Systematics*, **38** : 231–253.

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. *Journal of Evolutionary Biology*, **13** : 58–62.

Rousset, F. (2003) A minimal derivation of convergence stability measures. *Journal of Theoretical Biology*, **221** : 665–668.

Rousset, F. (2004) *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton.

Rousset, F. et S. Billiard (2000) A theoretical basis for measures of kin selection in subdivided populations : finite populations and localized dispersal. *Journal of Evolutionary Biology*, **13** : 814–825.

Rousset, F. et O. Ronce (2004) Measuring selection on traits affecting metapopulation demography. *Theoretical Population Biology*, **65** : 127–141.

Saitou, N. et M. Nei (1987) The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4** : 406–425.

Satterthwaite, W. H. (2010) Competition for space can drive the evolution of dormancy in a temporally invariant environment. *Plant Ecology*, **208** : 167–185.

Seielstad, M. T., E. Minch et L. L. Cavalli-Sforza (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genetics*, **20** : 278–280.

Shields, W. M. (1983) Optimal inbreeding and the evolution of philopatry. In : *The Ecology of Animal Movements* (eds. Swingland, I. R. et P. J. Greenwood), pp. 132–159. Oxford University Press, Oxford.

Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T. J. Hudson, A. Montpetit, A. V. Pshezhetsky, M. Prentki, B. I. Posner, D. J. Balding, D. Meyre, C. Polychronakos et P. Froguel (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445** : 881–885.

Slatkin, M. (1974) Hedging one's evolutionary bets. *Nature*, **250** : 704–705.

Snyder, R. E. (2006) Multiple risk reduction mechanisms : can dormancy substitute for dispersal ? *Ecology Letters*, **9** : 1106–1114.

Stoneking, M. (1998) Women on the move. *Nature Reviews Genetics*, **20** : 219–220.

Storz, J. et M. Beaumont (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.

Storz, J. F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14** : 671–688.

Tavaré, S., D. J. Balding, R. C. Griffiths et P. Donnelly (1997) Inferring coalescence times from DNA sequence data. *Genetics*, **145** : 505–518.

Taylor, P. D. (1988) An inclusive model for dispersal of offspring. *Journal of Theoretical Biology*, **130** : 363–378.

Taylor, P. D. (1990) Allele-frequency change in class-structured populations. *American Naturalist*, **135** : 95–106.

Taylor, P. D. et S. A. Frank (1996) How to make a kin selection model. *Journal of Theoretical Biology*, **180** : 27–37.

Tsuji, N. et N. Yamamura (1992) A simple evolutionary model of dormancy and dispersal in heterogeneous patches with special difference to phytophagous lady beetles. I. Stable environments. *Researches on Population Ecology*, **34** : 77–90.

van Valen, L. (1971) Group selection and the evolution of dispersal. *Evolution*, **25** : 591–598.

Venable, D. L. (1985) Ecology of achene dimorphism in *Heterotheca latifolia*. III. consequences of varied water availability. *Journal of Ecology*, **73** : 757–763.

Venable, D. L. (2007) Bet hedging in a guild of desert annnuals. *Ecology*, **88** : 1086–1090.

Venable, D. L. et J. S. Brown (1988) The selective interactions of dispersal, dormancy, and seed size as adaptations for reducing risk in variable environment. *American Naturalist*, **131** : 360–384.

Venable, D. L., C. E. Pake et A. C. Caprio (1993) Diversity and coexistence of Sonoran desert winter annuals. *Plant Species Biology*, **8** : 207–216.

Via, S. (1999) Reproductive isolation between sympatric races of pea aphids. I. Gene flow restriction and habitat choice. *Evolution*, **53** : 1446–1457.

Vitalis, R., P. Boursot et K. Dawson (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158** : 1811–1823.

Wakeley, J. (2008) *Coalescent Theory : An Introduction*. Roberts & Company Publishers, ? ? ?

Waser, P. M., S. N. Austad et B. Keane (1986) When should animals tolerate inbreeding ? *The American Naturalist*, **128** : 529–537.

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen et W. G. Hill (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, **15** : 1468–1476.

Weir, B. S. et W. G. Hill (2002) Estimating F-statistics. *Annual Review of Genetics*, **36** : 721–50.

Wells, R. S., N. Yuldasheva, R. Ruzibakiev, P. A. Underhill, I. Evseeva, J. Blue-Smith, L. Jin, B. Suf, R. Pitchappang, S. Shanmugalakshmig, K. Balakrishnang, M. Readh, N. M. Pearsoni, T. Zerjalj, M. T. Websterk, I. Zholoshvilil, E. Jamarjashvilil, S. Gambarovm, B. Nikbinn, A. Dostievo, O. Aknazarovp, P. Zallouaq, I. Tsoyr, M. Kitaevs, M. Mirrakhimovs, A. Charievt et W. F. Bodmer (2001) The eurasian heartland : a continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences USA*, **98** : 10244–10249.

Wiener, P. et S. Tuljapurkar (1994) Migration in variable environments : exploring life-history evolution using structured poplation models. *Journal of Theoretical Biology*, **166** : 75–90.

Wilkins, J. F. et F. W. Marlowe (2006) Sex-biased migration in humans : what should we expect from genetic data ? *Bioessays*, **28** : 290–300.

Wollstein, A., O. Lao, C. Becker, S. Brauer, R. J. Trent, P. Nürnberg, M. Stoneking et M. Kayser (2010) Demographic history of oceania inferred from genome-wide data. *Current Biology*, **20** : 1983–1992.

Wright, S. (1931) Evolution in mendelian populations. *Genetics*, **16** : 97–159.

Wright, S. (1935) Evolution in populations in approximate equilibrium. *Journal of Genetics*, **30** : 257–266.

Wright, S. (1949) Adaptation and selection. In : *Genetics, Paleontology, and Evolution* (eds. Jepson, G. L., G. G. Simpson et E. Mayr), pp. 365–389. University Press, Princeton.

Wright, S. (1969) *Evolution and the Genetics of Populations. Volume II. The Theory of Gene Frequencies*. University of Chicago Press, Chicago.

Zerjal, T., R. S. Wells, N. Yuldasheva, R. Ruzibakiev et C. Tyler-Smith (2002) A genetic landscape reshaped by recent events : Y-chromosomal insights into Central Asia. *American Journal of Human Genetics*, **71** : 466–482.

Deuxième Partie

Annexes

Annexe A

Curriculum vitæ

Né le 9 février 1972, deux enfants.
Nationalité française

Situation actuelle

Sept. 2009 – présent	**Mis à disposition de l'INRA**. Centre de Biologie pour la Gestion des Populations (CBGP), Montpellier

Cursus Universitaire et Expérience Professionnelle

Oct. 2004 - Août. 2009	**Chargé de Recherche au CNRS (CR1 depuis oct. 2008)**. Unité d'Éco-Anthropologie et d'Ethnobiologie, Muséum National d'Histoire Naturelle, Paris
Fév. 2004 - Sept. 2004	**Post-doctorat** – *Bayes methods for quantifying levels of adaptive divergence between populations from gene frequency data* (Dir. Mark Beaumont). Université de Reading (Royaume-Uni)
Déc. 2001 - Nov. 2003	**Post-doctorat** – *Evolution in plant communities : modelling the co-evolution of life-history traits in interacting species* (Dir. Vincent Jansen). Royal Holloway, Université de Londres, Egham (Royaume-Uni)
Sept. 1998 - Déc. 2001	**Thèse de l'Université Montpellier 2** – *Génétique des populations subdivisées : théorie et applications* (Dir. Isabelle Olivieri et Patrick Grillas). Institut des Sciences de l'Évolution de Montpellier et Station Biologique de la Tour du Valat
Oct. 1998 - Oct. 1999	**Stage pré-doctoral** – *Détecter l'influence de la sélection* (Dir. Pierre Boursot). Laboratoire Génome, Populations, Interactions et Adaptation, Université Montpellier 2
Oct. 1997 - Sept. 1998	**Assistant d'enseignement** – *Biais de dispersion liés au sexe* (Dir. Nicolas Perrin). Institut de Zoologie et d'Écologie Animale, Université de Lausanne (Suisse)
Jan. 1996 - Sept. 1997	**Service National Civil**. Conservatoire Botanique National Méditerranéen de Porquerolles et Université Montpellier 2
Oct. 1994 - Sept. 1995	**DEA Évolution et Écologie de l'Université Montpellier 2** – *Information multigénique et structure des populations* (Dir. Denis Couvet). Institut des Sciences de l'Évolution de Montpellier
Sept. 1992 - Sept. 1995	**Magistère de Biologie et Biochimie de la Région Parisienne**, **Licence** et **Maîtrise de Biochimie**. Université Paris 7
Sept. 1990 - Juil. 1992	**DEUG Sciences de la Nature et de la Vie**. Université Paris 7
Sept. 1989 - Juil. 1990	**Baccalauréat Série C**. Lycée Hélène Boucher, Paris

Encadrement

Post-doctorats (2)

- Begoña Martinez-Cruz – *Génétique des populations humaines d'Asie Centrale* (2005–2007, co-direction avec Évelyne Heyer). Begoña Martinez-Cruz est actuellement en post-doctorat (Université de Barcelone, Espagne)
- Jean-Baptiste André – *Coopération et punition chez l'Homme* (2005–2007, co-direction avec Évelyne Heyer). Jean-Baptiste André est chargé de recherche (CR1) au CNRS depuis octobre 2007 au laboratoire "Fonctionnement et Evolution des Systèmes Ecologiques" (Paris)

Thèse (1)

- Laure Ségurel – *Mode de vie et diversité génétique dans les populations humaines d'Asie Centrale* (2006-2010, co-direction (avec Évelyne Heyer). Thèse soutenue le 13 janvier 2010. Laure Ségurel est actuellement en post-doctorat dans le laboratoire de Molly Przeworski (Dept. of Human Genetics, and Dept. of Ecology and Evolution, Chicago, USA).

Stages de Master Deuxième Année (4 stages principaux)

- (2012) Homa Papoli – *Detecting and measuring selection in gene frequency data* (Erasmus Mundus Master Programme in Evolutionary Biology, Université Montpellier 2 ; durée : 5 mois à partir de septembre 2012)
- (2007) Camille Madec – *Estimation des biais de dispersion liés au sexe* (Master Biologie Moléculaire et Cellulaire, Spécialité Génétique des Caractères Complexes, Université Paris 6, co-encadrant : Raphaël Leblois ; durée : 5 mois)
- (2007) Sandrine Bérot – *Inférences démogénétiques sous un modèle de populations en divergence avec migration* (Master Sciences de l'Univers, Environnement, Ecologie, spécialité Ecologie Biodiversité Evolution, Université Paris 11 ; co-encadrement, encadrant principal : Raphaël Leblois ; durée : 5 mois)
- (2006) Laure Ségurel – *Étude des phénomènes démographiques sexe-spécifiques dans des populations humaines d'Asie Centrale* (Magistère de Génétique, Mention Evolution et Génétique des populations, Université Paris 7, co-encadrant : Évelyne Heyer ; durée : 5 mois)
- (1997) Ambroise Dalécky – *Effet d'une banque de graines sur la structure génétique des populations* (DEA Biologie de l'Evolution et Ecologie, Université Montpellier 2, Stage Annexe du DEA ; durée : 2 semaines)

Stages de Master Première Année (3)

- (2007) Paul Theis – *Estimation des paramètres démographiques d'un modèle de divergence de populations : apport d'une approche Bayésienne* (École Polytechnique, Stage d'Option Scientifique ; durée : 3 mois). Paul Theis a reçu les félicitations de son École pour ce travail d'option
- (1998) Nevena Basic et Annick Tauxe – *Soin paternel chez une musaraigne monogame, la musaraigne musette* (Crocidura russula) (second cycle universitaire à l'Université de Lausanne, Suisse, travail de Certificat, co-encadrement avec Nicolas Perrin ; durée : 3 mois)

- (1997) Lætitia Bouchevreau – *Recherche de marqueurs microsatellites chez* Marsilea strigosa (Maîtrise de Biochimie de l'Université Montpellier 2 ; durée : 3 mois)

Stages de Premier Cycle (4)

- (1997) Vassilis Georgiadis – *Genetic structure of* Marsilea strigosa *populations using microsatellite markers* (premier cycle universitaire, Queen's Mary College of London, UK ; durée : 2 mois)
- (1997) Vuokko Vanninen – *Characterization of microsatellites loci using an enrichment protocol in* Marsilea strigosa (premier cycle universitaire, Université d'Oulu, Finlande ; durée : 2 mois)
- (1996) Sylvain Glémin – *Effet de la présence d'une banque de graines structurée en classes d'âge sur la différenciation génétique en métapopulation* (seconde année de l'Ecole Normale Supérieure de Lyon ; durée : 1 mois)
- (1996) Sylvie Muratorio – *Etude de la variabilité génétique chez* Marsilea strigosa*, fougère rare du Languedoc-Roussillon* (première année de la Formation des Ingénieurs Forestiers FIF –ENGREF, Nancy ; durée : 1 mois)

Enseignement

Présentation synthétique des activités d'enseignement

Intitulé de la formation	1997–1998	2004–2005	2005–2006	2006–2007	2007–2008	2008–2009	2009–2010	2010–2011	2011–2012
(1) Analyse des données en génét. des populations									12h
(2) Genetic Data Analysis								14h	14h
(3) Théorie de la coalescence et applications							3h	3h	3h
(4) Inférence en génétique des populations							3h		
(5) Analyse des données en génét. des populations				9h	9h	7h	9h		
(6) Introduction à la théorie de la coalescence [...]							4h		
(7) Génétique des Populations Moléculaire [...]		3h	5h	8h	14h	14h			
(8) Génétique des Populations					5h	5h			
(9) Génomique des Populations				11h	14h	14h			
(10) Anthropologie Évolutive et Primatologie					8h	8h			
(11) Université de Lausanne	90h								
Total (équivalent TD)	**90h**	**3h**	**5h**	**28h**	**50h**	**48h**	**19h**	**17h**	**29h**

Détail des formations dispensées

(1) *Analyse de données en génétique des populations*
Niveau : Master 2 (Master Écologie Biodiversité, ouvert aux étudiants de l'école doctorale SIBAGHE)
Établissement : Université Montpellier 2
Responsable de ce module (co-responsable : Raphaël Leblois)

(2) *Genetic Data Analysis*
Niveau : Master 1 (Erasmus Mundus Master Programme in Evolutionary Biology)
Établissement : Université Montpellier 2
Responsable de ce module (co-responsable : Raphaël Leblois)

(3) *Théorie de la coalescence et applications*
Niveau : Master 2 (Master Écologie, Évolution, Biométrie)
Établissement : Université Claude Bernard – Lyon 1

(4) *Inférence en génétique des populations*
Niveau : Master 1 (Master Écologie Biodiversité)
Établissement : Université Montpellier 2

(5) *Introduction à la théorie de la coalescence et inférence en génétique des populations*
Niveau : Master 1 (Master Agronomie et Agroalimentaire)
Établissement : Montpellier SupAgro

(6) *Analyse des Données en Génétique des Populations*
Niveau : École doctorale (Sciences de la Nature et de l'Homme)
Établissement : Muséum National d'Histoire Naturelle
Responsable de ce module (co-responsable : Raphaël Leblois)

(7) *Génétique des Populations Moléculaire et Coalescence*
Niveau : Master 2 (Master Environnement : milieux, techniques, sociétés)
Établissement : Muséum National d'Histoire Naturelle, Universités Paris 6, 7, 11 et École Pratique des Hautes Études
Co-responsable de ce module (responsable : Évelyne Heyer)

(8) *Génétique des Populations*
Niveau : Master 2 (Master Écologie)
Établissement : Université Lille 1

(9) *Génomique des Populations*
Niveau : Master 1 (Master de Biologie)
Établissement : École Normale Supérieure

(10) *Anthropologie Évolutive et Primatologie*
Niveau : Master 1 (Master Environnement : milieux, techniques, sociétés)
Établissement : Muséum National d'Histoire Naturelle
Co-responsable de ce module (responsable : Sabrina Krief)

(11) *Service d'assistant d'enseignement à l'université de Lausanne (Suisse), incluant :*
Génétique Moléculaire des Populations (60h de travaux pratiques en Master 1)
Biologie des Organismes (16h de travaux dirigés en Master 1)
Statistiques (8h de travaux dirigés en Master 1)
Génétique des Populations (8h de travaux dirigés en Propédeutique – Licence 2)
Zoologie (4h de travaux pratiques en Propédeutique – Licence 1)
Zoologie (8h de travaux pratiques en Médecine – Licence 1)
Zoologie (15h de travaux pratiques en Propédeutique – Licence 2)

Travaux d'expertise

Jurys de thèse (6)

- Flora Jay – *Méthodes bayésiennes pour la génétique des populations : relations entre structure génétique des populations et environnement* (Université de Grenoble, dir. Olivier François et Michaël Blum) le 14 novembre 2011 (rapporteur)
- David Enard – *Étude des points chauds de sélection positive dans les génomes de Vertébrés* (Université Paris 7, dir. Hugues Roest Crollius) le 21 septembre 2010 (examinateur)
- Laure Ségurel – *Mode de vie et diversité génétique dans les populations humaines d'Asie Centrale* (Université Paris 6, dir. Évelyne Heyer et Renaud Vitalis) le 13 janvier 2010
- Charlotte Tolleanere – *Génétique et évolution du rat noir* (Rattus rattus)*, réservoir de la peste à Madagascar* (Université Montpellier 2, dir. Jean-Marc Duplantier et Carine Brouat) le 2 décembre 2009 (examinateur)
- Mohammed Zeinedine – *Evolutionary Population Phenomena* (Royal Holloway, Université de Londres, UK, dir. Vincent AA Jansen) le 9 novembre 2005 (examinateur)
- Denis Roze – *Conséquences évolutives de la structure des populations. Probabilités de fixation, évolution de la recombinaison et des taux de dispersion* (Université Montpellier 2, dir. François Rousset et Yannis Michalakis) le 25 octobre 2004

Comités de thèse (12)

- Guillaume Laugier – *Changements évolutifs et bioinvasions* (Université Montpellier 2, dir. : Arnaud Estoup et Benoit Facon) les 18 février 2011 et 15 juin 2012
- Jonathan Romiguier – *Maladaptation moléculaire : Impact de la conversion génique biaisée sur l'évolution des protéomes mammaliens* (Université Montpellier 2, dir. : Nicolas Galtier et Vincent Ranwez) le 15 octobre 2010
- Christophe Girod – *Influence des variations climatiques passées sur la distribution des espèces forestières tropicales : l'exemple du palmier* Astrocaryum sciophilum (Université Paris 6, dir. : Hélène Fréville et Bernard Riera) le 11 décembre 2009
- Stéphanie Robert – *Routes d'expansion mondiale du champignon phytopathogène* Mycosphaerella fijiensis*. Conséquences sur la variabilité des traits liés à l'agressivité et le potentiel adaptatif* (Université Montpellier 2, dir. : Jean Carlier, Virginie Ravigné et Catherine Abadie) les 7 mai 2009 et 7 octobre 2010
- Claire-Lise Meyer – *Adaptation d'*Arabidopsis halleri *(Brassicaceae) aux milieux fortement pollués par les métaux lourds : approche de génomique des populations* (Université Lille 1, dir. : Pierre Saumitou-Laprade et Vincent Castric) le 25 septembre 2007
- Paul Verdu – *Anthropologie génétique des populations humaines d'Afrique Centrale : histoire du peuplement Pygmée* (Université Paris 6, dir. : Evelyne Heyer) le 5 décembre 2007

- Etienne Patin – *Influences du mode de vie sur la diversité génétique et la susceptibilité aux maladies chez l'Homme : l'exemple des agriculteurs et des chasseurs-cueilleurs d'Afrique Centrale* (Université Paris 6, dir. : Evelyne Heyer et Lluis-Quintana Murci) le 5 décembre 2007
- Luis-Miguel Chevin – *Effets de la sélection sur le polymorphisme neutre. Conséquences pour la détection de régions sous sélection récente* (Université Paris 11, dir. : Frédéric Hospital) le 16 juin 2006
- Cécile Edelist – *Caractérisation de l'adaptation à des marais salins d'une espèce hybride sauvage de tournesol :* Helianthus paradoxus (Université Paris 11, dir. : Christine Dillmann et Delphine Sicard) le 7 février 2005
- Florence Noël – *Étude démographique et génétique de populations d'une espèce rare et protégée en France :* Ranunculus nodiflorus *L. dans le bassin parisien* (Université Paris 6, dir. : Nathalie Machon) le 25 novembre 2003

Rôle d'arbitre dans le processus de publication pour les revues suivantes[1]

- *Annals of Human Genetics*, *Bioinformatics*, *BMC Ecology*, *Evolution*, *Evolutionary Applications*, *Evolutionary Ecology*, *Genes and Genetic Systems*, *Genetics*, *Heredity*, *Journal of Evolutionary Biology*, *Journal of Theoretical Biology*, *Molecular Ecology*, *Oikos*, *Plant Ecology*, *Proceedings of the Royal Society London Series B*, *The American Naturalist*, *Theoretical Population Biology*

Jurys de concours (2)

- 2010 : Membre du comité de sélection et rapporteur d'un dossier pour le poste MCF n°503 "Génomique de l'adaptation", affecté au Laboratoire d'Écologie Alpine de l'Université Joseph Fournier, Grenoble
- 2010 : Rapporteur de deux dossiers pour le poste MCF EPHE n°3099 "Génétique des populations moléculaire et coalescence", affecté au Laboratoire Origine structure et évolution de la biodiversité, Muséum National d'Histoire Naturelle, Paris

Jurys d'examen (6)

- 2012 : Participation au jury de soutenance des stages des étudiants de M1 Erasmus Mundus Master Programme in Evolutionary Biology, à l'Université Montpellier 2 (promotion 2011–2012)
- 2009–2012 : Participation au jury de soutenance des stages des étudiants de M1 du parcours Biodiversité, Écologie, Évolution (BEE), du Master Biologie, Géosciences, Agroressources et Environnement (BGAE) de l'Université Montpellier 2 (promotions 2009–2010, 2010–2011 et 2011–2012)
- 2010–2011 : Participation au jury de sélection des étudiants de M1 du parcours Biodiversité, Écologie, Évolution (BEE), du Master "Biologie, Géosciences, Agroressources et Environnement" (BGAE) de l'Université Montpellier 2, pour les rentrées universitaires 2010 et 2011

Autres activités d'évaluation de la recherche

- 2012 : Évaluation d'un projet de recherche dans le cadre de l'appel d'offre "Ambizione" de la Swiss National Science Foundation

1. 74 revues réalisées au 10 juillet 2012

- 2011–2012 : Évaluation d'un sujet de thèse dans le cadre de l'appel d'offre du département EFPA de l'INRA
- 2011 : Évaluation d'un projet de recherche dans le cadre de l'appel d'offre "projets innovants" du département EFPA de l'INRA
- 2008–2009 : Évaluation de projets de recherches pour le "Belgian Fund for Scientific Research" (FWO)
- 2007 : Évaluation d'actes de colloque pour le European Union for Bird Ringing (EURING)

Financements obtenus et participation à des contrats

- 2012–2016 : "Institut de Biologie Computationnelle (IBC)". Porteur : Olivier Gascuel ; participation à hauteur de 20% ; organisme financeur : ANR Programme Investissements d'Avenir (2 000 000 €)
- 2011–2014 : "Génétique de l'adaptation trophique et mécanismes d'isolement reproducteur chez les pucerons (SPECIAPHID)". Porteur : Jean-Christophe Simon ; participation à hauteur de 10% ; organisme financeur : ANR Programme Blanc 2011 (500 000 €)
- 2010–2014 : "Génomique de la phénologie chez la processionnaire du pin *Thaumetopoea pityocampa* (GENOPHENO)". Porteur : Carole Kerdelhué ; participation à hauteur de 25% ; organisme financeur : Programme Jeunes Chercheuses Jeunes Chercheurs de l'ANR (274 000 €)
- 2010–2011 : "Génétique de l'adaptation du pois à ses hôtes Fabacées (DIVAPHID)". Porteur : Jean-Christophe Simon ; participation à hauteur de 10% ; organisme financeur : Fondation de la Recherche pour la Biodiversité (FRB) appel à projet 2009 (32 000 €)
- 2010–2013 : Jeune Equipe INRA "Inférence en Génétique et Génomique des Populations (IGGiPop)". Porteur : Renaud Vitalis ; participation à hauteur de 100% ; organisme financeur : INRA (105 000 €)
- 2009–2013 : "Étude de Méthodes Inférentielles et Logiciels pour l'Évolution (EMILE)". Porteur : Jean-Marie Cornuet ; participation à hauteur de 75% ; organisme financeur : ANR Programme Blanc 2009 (371 573 €)
- 2009 : "Contribution au développement du cluster de calcul du MNHN". Porteur : Renaud Vitalis et Cyrille D'Haèse ; organisme financeur : Budget Qualité Recherche du département Systématique et Évolution du MNHN) (5 500 €)
- 2008 : "Contribution au développement du cluster de calcul du MNHN." Porteur : Renaud Vitalis ; organisme financeur : Budget Qualité Recherche du département Hommes Natures Sociétés du MNHN) (3 000 €)
- 2008–2011 : "Rôle des variations climatiques passées et de la spécialisation écologique dans la distribution des espèces tropicales : apport des approches génétiques (CLIPS)". Porteurs : Hélène Fréville et Caroline Scotti-Saintagne ; participation à hauteur de 10% ; organisme financeur : CNRS - Appel d'Offres Amazonie) (80 000 €)
- 2007–2011 : "Deciphering the complex evolution of genes involved in human adaptation to diet (NUTGENEVOL)". Porteur : Évelyne Heyer. Participation à hauteur de 25% ; organisme financeur : ANR Programme Blanc 2007 (282 564 €)

- 2010 : "Génétique de l'adaptation du puceron du pois à ses hôtes Fabacées (DIVAPHID)". Porteur : Jean-Christophe Simon ; participation à hauteur de 10% ; organisme financeur : FRB (30 000 €)
- 2007–2010 : "Génomique des populations du puceron du pois : balayage du génome pour identifier des locus impliqués dans la divergence adaptative des populations naturelles" Porteurs : Jean-Christophe Simon et Renaud Vitalis ; organisme financeur : INRA - Appel à Projets Scientifiques du Département SPE) (30 000 €)
- 2005 : "Contribution au développement du cluster de calcul du MNHN". Porteurs : Renaud Vitalis et Frédéric Austerlitz ; organisme financeur : GDR 1928 Génomique des Populations (12 000 €)
- 2004–2007 : "Histoire et diversité génétique des Pygmées d'Afrique Centrale et de leurs voisins". Porteur : Evelyne Heyer ; participation à hauteur de 20% ; organisme financeur : Ministère délégué à la Recherche – ACI PROSODIE (90 000 €)

Animation scientifique et responsabilités collectives

Participation à l'organisation de conférences

- 2012 : Membre du comité d'organisation de la Conférence Jacques Monod intitulée "Développements théoriques et empiriques en génomique évolutive", Roscoff, 31 mars – 4 avril 2012
- 2010 : Membre du comité d'organisation de la journée satellite "Biodiversité et Bioinformatique" des "Journées Ouvertes en Biologie, Informatique et Mathématiques" (JOBIM), Montpellier, 6 septembre 2010
- 2007 : Membre du comité scientifique de la conférence "DNA Sampling – Strategies and Design", Paris, 15–16 mars 2007

Animation scientifique

- 2011– : Membre du comité d'organisation des Séminaire d'Écologie et d'Évolution du Labex CeMEB
- 2009– : co-animateur du groupe de réflexion [2] "Biologie de l'Adaptation : Génétique, Génomique et Traits d'Histoire de Vie" de l'UMR 1062 (dir. F vanlerberghe)
- 2009– : co-animateur du groupe de réflexion "Génétique des Populations et Phylogéographie" du CBGP (UMR 1062)
- 2005–2006 : organisation des réunions scientifiques hebdomadaires de l'équipe "Génétique des Populations Humaines" (UMR 5145)
- 1996–1997 : organisation des réunions scientifiques hebdomadaires du laboratoire Génétique et Environnement (ISEM - UMR 5554)

Responsabilités collectives

- 2011– : Co-responsable avec Nicolas Mouquet du groupe de travail "Origine et Dynamique de la Biodiversité" au sein du Labex CeMEB

2. Le CBGP n'est pas organisé par équipes mais par projets Les grands thèmes de recherche s'appuient sur les réflexions menées dans quatre "groupes de réflexion" correspondant à des champs disciplinaires. Les "groupes de réflexion" ont essentiellement un but d'animation scientifique

- 2011– : Membre nommé du Conseil d'Unité de l'UMR 1062 (CBGP)
- 2010– : direction de la Jeune Équipe INRA "Inférence en Génétique et Génomique des Populations (IGGiPop)".
- 2006–2009 : co-responsable scientifique de la plateforme de calcul du MNHN, au sein de l'UMS 2700 (dir. Éric Pasquet) "Outils et Méthodes de la Systématique Intégrative"
- 2006–2009 : co-responsable de l'axe de recherche "Adaptation et Evolution" de l'UMR 7206 (dir. Serge Bahuchet)

Annexe B

Liste des publications

Publications

Articles publiés dans des revues internationales avec comité de lecture

[P31] Smadja C, Canbäck B, Vitalis R, Gautier M, Ferrari J, Zhou JJ and Butlin RK (2012) A novel route to the genetics of speciation : large-scale candidate gene scan for host-associated speciation genes. *Evolution*, sous presse [IF = 5.146 [1]]

[P30] Gautier M et Vitalis R (2012) rehh : An R package to detect footprints of selection in genome-wide SNP data from haplotype structure *Bioinformatics*, **28** : 1176–1177 [IF = 5.468]

[P29] Bon C, Berthonaud V, Fosse P, Gély B, Maksud F, Vitalis R, Philippe M, van der Plicht J et Elalouf, J-M (2011) Low regional diversity of late cave bears mitochondiral DNA at the time of Chauvet Aurignacian paintings. *Journal of Archaelogical Science*, **38** : 1886–1895 [IF = 1.914] [2]

[P28] [3] Girod C, Vitalis R, Leblois R et Fréville H (2011) Inferring population decline and expansion from microsatellite data : a simulation-based evaluation of the Msvar method. *Genetics*, **188** : 165–179 [IF = 4.007]

[P27] Facon B, Hufbauer RA, Tayeh A, Loiseau A, Lombaert E, Vitalis R, Guillemaud T, Lundgren JG et Estoup A (2011) Inbreeding depression is purged in the invasive insect *Harmonia axyridis*. *Current Biology*, **21** : 424–427 [IF = 9.647]

[P26] Midamegbe A [4], Vitalis R [4], Malausa T, Delava E, Cros–Arteil S et Streiff R (2011) Scanning the European corn borer (*Ostrinia* spp.) genome for adaptive divergence between host-affiliated sibling species. *Molecular Ecology*, **20** : 1414–1430 [IF = 5.522]

[P25] Martínez–Cruz B [4], Vitalis R [4], Austerlitz F, Quintana–Murci L, Aldashev A, Hegay T et Heyer E (2011) In the heartland of Eurasia : the multilocus genetic landscape of Central Asian populations. *European Journal of Human Genetics*, **19** : 216–223 [IF = 4.400]

---

1. Facteur d'impact (*Impact Factor*) de la revue, tiré du *Journal Citation Report*, *Science Edition 2011*, publié par *ISI Web of Knowledge*SM

2. Cette publication a donné lieu à un communiqué de presse du CEA : `http://www-dsv.cea.fr/dsv/themes/sciences-du-vivant-pour-les-biotechnologies/l-ours-des-cavernes-des-dessins-de-la-grotte-chauvet-a-l-extinction-de-l-espece`, ainsi qu'à plusieurs brèves dans la presse scientifique nationale : `http://www.pourlascience.fr/ewb_pages/a/actualite-la-grotte-chauvet-datee-par-l-ours-26872.php` et internationale : `http://www.newscientist.com/article/mg21028093.900-bear-dna-is-clue-to-age-of-chauvet-cave-art.html` et `http://news.discovery.com/archaeology/cavemen-bears-prehistoric-caves-110426.html`

3. J'ai pris la liberté d'indiquer, en gras, le numéro des publications correspondant soit à un travail d'encadrement de post-doctorant(e), de thésard(e) ou d'étudiant(e), soit à une collaboration étroite avec un(e) thésard(e) ou un(e) étudiant(e), dont je n'ai pas eu la responsabilité, mais qui a contribué significativement à un chapitre de sa thèse ou de son mémoire. Pour ces publications, le nom du post-doctorant, thésard ou étudiant est souligné deux fois.

4. Contribution égale des deux premiers auteurs

[P24] Ayala D, Fontaine M. C, Cohuet A, Fontenille D, Vitalis R et Simard F Chromosomal inversions, natural selection and adaptation in the malaria vector *Anopheles funestus*. *Molecular Biology and Evolution*, **28** : 745–758 [IF = 5.550]

[**P23**] Verdu P, Leblois R, Froment A, Théry S, Bahuchet S, Rousset F, Heyer E et Vitalis R (2010) Limited dispersal in mobile hunter-gatherer Baka Pygmies. *Biology Letters*, **6** : 858–861 [IF = 3.762] [1]

[P22] Alberto F, Niort J, Derory J, Lepais O, Léger V, Vitalis R, Galop D et Kremer A (2010) Population differentiation of sessile oak at the altitudinal front of migration in the French Pyrenees. *Molecular Ecology*, **19** : 2626–2639 [IF = 5.522]

[**P21**] Rhoné B, Vitalis R, Goldringer I et Bonnin I (2010) Evolution of flowering time in experimental wheat populations : a comprehensive approach to detect genetic signatures of natural selection. *Evolution*, **64** : 2110–2125 [IF = 5.146]

[P20] Beaumont MA, Nielsen R, Robert C, Hey J, Gaggiotti O, Knowles L, Estoup A, Panchal M, Corander J, Hickerson M, Sisson S, Fagundes N, Chikhi L, Beerli P, Vitalis R, Cornuet J-M, Huesenbeck J, Foll M, Yang Z, Rousset F, Balding D et Excoffier L (2010) In defense of model-based inference in phylogeography. *Molecular Ecology*, **19** : 436–446 [IF = 5.522]

[**P19**] Ségurel L, Lafosse S, Heyer E et Vitalis R (2010) Frequency of the AGT Pro11Leu polymorphism in humans : does diet matter ? *Annals of Human Genetics*, **74** : 57–64 [IF = 2.565]

[**P18**] Cooper JD, Vitalis R, Waser PM, Gopurenko D, Hellgren EC, Gabor TM et DeWoody JA (2010) Genetic detection of male-biased dispersal in a social mammal using a maternally inherited molecular marker and a novel theoretical framework. *Heredity*, **104** : 79–87 [IF = 4.597]

[**P17**] Meyer C-L, Vitalis R, Saumitou-Laprade P et Castric V (2009) Genomic pattern of adaptive divergence in *Arabidopsis halleri*, a model species for tolerance to heavy metal. *Molecular Ecology*, **18** : 250-262 [IF = 5.522]

[P16] Verdu P, Austerlitz F, Estoup E, Vitalis R, Georges M, Froment A, Lebomin S, Gessain A, Hombert J-M, Van der Veen L, Quintana–Murci L, Bahuchet S et Heyer E (2008) Origins and genetic diversity in pygmy hunter-gatherers from Western Central Africa. *Current Biology*, **19** : 312–318 [IF = 9.647] [2]

1. Cette publication a donné lieu à un communiqué de presse du CNRS : `http://www2.cnrs.fr/presse/communique/1872.htm`, du MNHN : `http://www.mnhn.fr/museum/foffice/national/national/presse/communiques/commPresse/fichePresse.xsp?ANNEE=2010&COMMPRESSE_ID=3483&idx=25&nav=liste` et de l'IRD : `http://www.ird.fr/toute-l-actualite/actualites/communiques-et-dossiers-de-presse/dispersion-genetique-limitee-dans-les-populations-mobiles-de-pygmees-baka-chasseurs-cueilleurs/%28language%29/fre-FR`

2. Cette publication a donné lieu à un communiqué de presse du CNRS : `http://www2.cnrs.fr/presse/journal/4380.htm`, à plusieurs brèves dans les revues internationales : `http://www.nature.com/news/2009/090205/full/news.2009.82.html`, `http:`

[**P15**] Ségurel L, Martínez–Cruz B, Quintana–Murci L, Balaresque P, Georges M, Hegay T, Aldashev A, Nazyrova F, Jobling MA, Heyer E et Vitalis R (2008) Sex-specific genetic structure and social organization in Central Asia : insights from a multi-locus study. *PLoS Genetics*, **4(9)** : e1000200 [IF = 8.694] [1]

[P14] Jansen VAA et Vitalis R (2007) The evolution of dispersal in a Levin's type metapopulation model. *Evolution*, **61** : 2386–2397 [IF = 5.146]

[**P13**] Vitalis R, Glémin S et Olivieri, I (2004) When genes go to sleep : the population genetics consequences of seed dormancy and monocarpic perenniality. *American Naturalist*, **163** : 295–311 [IF = 4.725]

[P12] Vitalis R, Dawson KJ, Boursot P et Belkhir K (2003) Detsel 1.0 : a computer program to detect markers responding to selection. *Journal of Heredity*, **94** : 429–431 [IF = 2.799]

[P11] Vitalis R, Riba M, Colas B, Grillas P et Olivieri I (2002) Multilocus genetic structure at contrasted spatial scales in the endangered water fern *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta). *American Journal of Botany*, **89** : 1142–1155 [IF = 2.664]

[P10] Vitalis R (2002) Sex-specific genetic differentiation and coalescence times : estimating sex-biased dispersal rates. *Molecular Ecology*, **11** : 125–138 [IF = 5.522]

[P9] Vitalis R, Dawson KJ et Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158** : 1811–1823 [IF = 4.007]

[P8] Olivieri I et Vitalis R (2001) La biologie des extinctions. *Médecine Sciences*, **17** : 63–69 [IF = 0.516]

[P7] Vitalis R et Couvet, D (2001) Estim 1.0 : a computer program to infer population parameters from one- and two-locus gene identity probabilities. *Molecular Ecology Notes*, **1** : 354–356 [IF = 3.062]

[P6] Vitalis R et Couvet D (2001) Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. *Genetical Research*, **77** : 67–81 [IF = 1.712]

[P5] Vitalis R, Dubois M-P et Olivieri I (2001) Characterization of microsatellite loci in the endangered species of fern *Marsilea strigosa* Willd. (Marsileaceae, Pteridophyta). *Molecular Ecology Notes*, **1** : 64–66 [IF = 3.062]

[P4] Vitalis R et Couvet D (2001) Estimation of effective population size and migration rate from one- and two-locus identity measures. *Genetics*, **157** : 911–925 [IF = 4.007]

//sciencenow.sciencemag.org/cgi/content/full/2009/205/3 et http://www.newscientist.com/article/dn16548-did-invading-farmers-drive-pygmy-diversity.html, ainsi qu'à plusieurs brèves dans la presse nationale : http://www.lefigaro.fr/sciences/2009/02/23/01008-20090223ARTFIG00327-l-origine-des-pygmees-revelee-par-une-etude-genetique-.php et http://www.sciencesetavenir.fr/actualite/archeo-paleo/20090206.OBS3479/un-ancetre-commun-pour-les-pygmees-d-rsquo-afrique.html

1. Cette publication a donné lieu à un communiqué de presse du CNRS : http://www2.cnrs.

[P3] Fréville H, Imbert E, Justy F, Vitalis R et Olivieri I (2000) Isolation and characterization of microsatellites in the endemic species *Centaurea corymbosa* Pourret (Asteraceae) and other related species. *Molecular Ecology*, **9** : 1671–1672 [IF = 5.522]

[P2] Godelle B, Austerlitz F, Brachet S, Colas B, Cuguen J, Gandon S, Gouyon P-H, Lefranc M, Olivieri I, Reboud X et Vitalis R (1998) Système génétique, polymorphisme neutre et sélectionné : implications en biologie de la conservation. *Genetics Selection and Evolution*, **30** : S15–S28 [IF = 2.885]

[P1] Capy P, Vitalis R, Langin T, Higuet D et Bazin C (1996) Relationships between transposable elements based upon their integrase-transposase domains : is there a common ancestor ? *Journal of Molecular Evolution*, **3** : 359–368 [IF = 2.274]

Articles en révision dans des revues avec comité de lecture

[R1] Jaquiéry J, Stoeckel S, Nouhaud P, Mieuzet L, Mahéo F, Legeai F, Bernard N, Bonvoisin A, Vitalis R et Simon J-C Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex. *Molecular Ecology*

[R2] Ségurel L, Austerlitz F, Toupance B, Gautier M, Pasquet P, Kelley, J.L, Voisin S, Lonjou C, Cruaud C, Couloux A, Hegay T, Aldashev A Vitalis R [1], et Heyer E [1] Neolithic selection favoring protective variants for type 2 diabetes, *European Journal of Human Genetics*

Articles en préparation

[p1] Gautier M et Vitalis R Inferring population histories using genome-wide allele frequency data.

[p2] Vitalis R, Rousset F, Kobayashi Y, Olivieri I et Gandon S The joint evolution of dispersal and dormancy in a metapopulation with local extinctions and kin competition.

[p3] Vitalis R, Gautier M, Dawson K et Beaumont M Detecting and measuring selection from gene frequency data.

[p4] Alexandre H, Ponsard S, Bourguet D, Vitalis R, Audiot P, Cros–Arteil S et Streiff R Exploring the genetic basis of a repeated phenotypic evolution in two related Lepidoptera species.

[p5] Vitalis R et Leblois R IbdSex : a coalescent-based computer program to simulate gene genealogies in isolation-by-distance models.

[p6] Vitalis R et Jansen VAA Coexistence of dispersal strategies in metapopulations.

`fr/presse/communique/1422.htm` et à une brève dans le *Wellcome Trust News* : `http://www.ecoanthropologie.cnrs.fr/IMG/pdf_Wellcome_trust.pdf`

1. Contribution égale des deux derniers auteurs

[p7] Vitalis R et Waller DM Mutation load and heterosis in diverging populations.

Chapitres d'ouvrages

[C2] Vitalis R DetSel : a R-package to detect marker loci responding to selection, *in* Data Production and Analysis in Population Genomics (Pompanon F et Bonin A, eds). Methods in Molecular Biology Series, Humana Press, USA

[C1] Ségurel L, Georges M, Vitalis R et Heyer E (2010) Influence du mode de vie sur la diversité génétique en Asie Centrale, *in* L'anthropologie du vivant : objets et méthodes (Chapuis-Lucciani N, Guihard-Costa A-M et Boetsch G, eds.), CNRS GDR 3267 "L'homme et sa diversité : dynamiques évolutives des populations actuelles", Paris, France

Actes de colloque avec comité de lecture

[A1] Couvet D et Vitalis R (1996) Multigenic indices and metapopulation description *in* Actes des journées du Programme Environnement, Vie et Sociétés. Résumé des Communications Orales, Session A. 15-17 janvier 1996, Cité des Sciences et de l'Industrie. Presses du CNRS, Paris, France

[A2] Vitalis R, Colas B, Riba M et Olivieri I (1998) *Marsilea strigosa* Willd. : Statut génétique et démographique d'une espèce menacée. *Ecologia Mediterranea*, **24** : 145–157

Publications universitaires (mémoires)

[M2] Vitalis R (2001) Génétique des populations subdivisées : théorie et applications. Thèse de l'Université des Sciences et Techniques du Languedoc (Montpellier 2), 170 pp. + 178 pp. d'annexes

[M1] Vitalis R (1995) Information multigénique et structure des populations. Diplôme d'Études Approfondies (DEA Écologie et Évolution) de l'Université des Sciences et Techniques du Languedoc (Montpellier 2), 24 pp. + 9 pp. d'annexes

Communications

Communications dans des colloques internationaux

2012 *Detecting and measuring selection from gene frequency data*, par Vitalis R, Gautier M, Dawson KJ et Beaumont MA. Conférence Jacques Monod "Theoretical and empirical advances in evolutionary genomics", Roscoff, France, 31 mars - 4 avril 2012 (poster)

2010 *Adaptation to industrial pollution : combining genomic and phenotypic approaches in an emergent model species*, par Castric V, Meyer C-L, Vitalis R, Frérot H et Saumitou-Laprade P. Annual Meeting of the Society of Molecular Biology and Evolution, Lyon, France, 4-8 juillet 2010 (poster)

2010 *Genetic adaptations to diet in herders and farmers from Central Asia : the case of type 2 diabetes*, par Ségurel L, Pasquet P, Georges M, Hegay T, Aldashev A, Vitalis R et Heyer E. Annual Meeting of the Society of Molecular Biology and Evolution, Lyon, France, 4-8 juillet 2010 (communication orale)

2010 *Past climate changes in the Neotropics and actual distribution of tree forest species : a phylogeographic study on the palm* Astrocaryum sciophilum, par Girod C, Leblois R, Vitalis R, Pintaud JC, Riéra B et Fréville H. International symposium on the biology of the palm family – Palms 2010, Montpellier, France, 05-07 mai 2010 (communication orale)

2010 *Looking for genetic adaptations to diet from a comparative study of herders and agriculturalists in Central Asia*, par Ségurel L, Pasquet P, Georges M, Hegay T, Aldashev A, Vitalis R et Heyer E. 79th Annual meeting of the American Association of Physical Anthropologists, Albuquerque, USA, 14-17 avril 2010 (communication orale)

2010 *Limited dispersal in mobile hunter-gatherer African Pygmies*, par Verdu P, Leblois R, Froment A, Théry S, Bachuchet S, Rousset F, Heyer E et Vitalis R. 79th Annual meeting of the American Association of Physical Anthropologists, Albuquerque, USA, 14-17 avril 2010 (poster)

2009 *Testing the carnivore connection hypothesis in Central Asia*, par Ségurel L, P Pasquet, Hegay T, Aldashev A, Vitalis R et Heyer E. 12th Annual Meeting of the European Society for Evolutionary Biology, Turin, Italie, 24-29 août 2009 (poster)

2009 *Social behavior and genetic diversity, in human populations*, par Heyer E, Chaix R, Ségurel L, Austerlitz F, Vitalis R, Hegay T et Blum. 78th Annual Meeting of the American Association of Physical Anthropologists, Chicago, USA, 31 mars-3 avril 2009 (communication orale)

2009 *Influence of dietary habits on genetic diversity in Central Asia*, par Ségurel L, Martínez-Cruz B, Georges M, Lafosse S, Hegay T, Aldashev A, P Pasquet, Vitalis R et Heyer E. Société d'Anthropologie de Paris 1859-2009 : 150 ans. Paris, 26-30 janvier 2009 (communication orale)

2008 *Chromosomal inversions, natural selection and adaptation in* Anopheles funestus, par Ayala D, Fontaine M. C, Cohuet A, Costantini C, Fontenille D, Vitalis R et Simard F. 57th Annual Meeting of the American Society of Tropical Medicine and Hygiene, New Orleans, USA, 7-11 décembre 2008 (communication orale)

2008 *Influence of social organization on the sex-specific genetic structure in Central Asia*, par Ségurel L, Heyer E, et Vitalis R. Annual Meeting of the Society of Molecular Biology and Evolution, Barcelone, Espagne, 5-7 juin 2008 (poster)

2008 *A MCMC approach for detecting selection from gene frequency data*, par Vitalis R, Dawson K et Beaumont M. Annual Meeting of the Society of Molecular Biology and Evolution, Barcelone, Espagne, 5-7 juin 2008 (poster)

2007 *Flowering time trait, gene and microsatellite differentiation among experimental populations of wheat evolving in contrasted environments*, par Rhoné B, Goldringer I, Vitalis R, Remoulé, Galic N et Bonnin I. 11th Annual Meeting of the European Society for Evolutionary Biology, Uppsala, Suède, 20-25 août 2007 (communication orale)

2007 *A MCMC approach for detecting selection from gene frequency data*, par Vitalis R, Dawson K et Beaumont M. Conférence Jacques Monod "Evolutionary Genomics", Roscoff, France, 2-6 mai 2007 (poster)

2000 *How to estimate population effective size and migration rate using two-locus identity by descent coefficients*, par Vitalis R. Conférence du TMR FRAGLAND, Montpellier, France, 29 mars-1 avril 2000 (communication orale)

2000 *The effect of diapause or dormancy on the expected genetic structure of metapopulations*, par Vitalis R. Conférence du TMR FRAGLAND, Cordoba, Espagne, 19-21 février 2000 (communication orale)

1999 *Metapopulation genetics of annual plants with a seed bank*, par Vitalis R. Conférence "Habitat loss : Ecological, evolutionary and genetic consequences", Helsinki, Finlande, 7-12 septembre 1999 (communication orale)

1997 *Population genetics of subdivided plant populations with a seed bank*, par Vitalis R et Olivieri I. 6th Annual Meeting of the European Society for Evolutionary Biology, Arnhem, Pays-Bas, 24-28 août 1997 (communication orale)

1997 *Population genetics of subdivided plant populations with a seed bank*, par Vitalis R, Glémin S et Olivieri I. 3rd Conference PhD Students in Evolutionary Biology, Facultat de Biologia, Universitat de Barcelona, Espagne, 26-28 février 1997 (communication orale)

Communications dans des colloques nationaux et séminaires

2012 *Détecter et mesurer la sélection dans les jeux de données de polymorphisme*, par VITALIS R. Réunion de l'ANR EMILE, Montpellier, 3 juillet 2012 (séminaire)

2012 *Détecter et mesurer la sélection dans les jeux de données de polymorphisme*, par VITALIS R. Séminaire de l'équipe SMILE, Collège de France. Paris 22 juin 2012 (séminaire)

2010 *Détecter et mesurer la sélection dans les jeux de données de polymorphisme*, par VITALIS R. Ecologie 2010. Montpellier 2-5 septembre 2010 (communication orale)

2010 *Estimer des changements d'effectif avec des microsatellites : MSVAR*, par VITALIS R. séminaire organisé par THOMAS LAMY (CEFE), Montpellier 12 mai 2010 (séminaire)

2009 *Mesurer la sélection dans les données de polymorphisme*, par VITALIS R. "Scan F_{ST}" : workshop organisé par NICOLAS BIERNE. Sète 6-7 juillet 2009 (communication orale)

2009 *Mode de vie et diversité génétique en Asie Centrale*, par SÉGUREL L, GEORGES M, LAFOSSE S, P PASQUET, VITALIS R et HEYER E. Réunion du Groupement des Anthropologistes de Langue Française : Biologie, environnements et comportements des populations humaines : passé, présent, futur. Bordeaux, 27-30 mai 2009 (communication orale)

2007 *Rechercher des signatures de la sélection dans les jeux de données de polymorphisme*, par VITALIS R. Réunion du Groupe de Recherche "Interaction, adaptation, spéciation" du Centre de Biologie et de Gestion des Populations, Montpellier, 20 mars 2007 (séminaire)

2007 *Identifier des signatures de la sélection dans les jeux de données de polymorphisme*, par VITALIS R. Réunion de l'ANR MAEV, Paris, 21 janvier 2007 (séminaire)

2006 *Rechercher des signatures de la sélection dans les jeux de données de polymorphisme*, par VITALIS R. Réunion annuelle du GDR 1928 "Génomique des Populations", Banyuls, 15 mars 2006 (séminaire)

2006 *Comment détecter des signatures de sélection dans les génomes ?*, par VITALIS R. Réunion du laboratoire GEPV (UMR CNRS 8016), Lille, 24 février 2006 (séminaire)

2005 *Comment détecter des signatures de sélection dans les génomes ?*, par VITALIS R. Réunion "Évolution Moléculaire", Paris, 1er juin 2005 (séminaire)

2005 *Comment détecter des signatures de sélection ?*, par VITALIS R. Réunions du Département Systématique et Évolution (MNHN), Paris, 24 mai 2005 (séminaire)

2005 *Comment détecter des signatures de sélection dans les génomes ?*, par VITALIS R. Réunion "Midi-Pile", Orsay, 1er février 2005 (séminaire)

2004 *Détecter la sélection à partir de mesures de F_{ST}*, par VITALIS R. Réunion du laboratoire d'Écologie Alpine (UMR CNRS 5553), Grenoble, 14 mai 2004 (séminaire)

2002 *Interpretation of variation across marker loci as evidence of selection*, par VITALIS R. Environment and Evolutionary Biology seminars, Royal Holloway, Université de Londres, Royaume Uni, 30 janvier 2002 (séminaire)

2002 *Biologie des populations d'espèces menacées*, par VITALIS R. Réunion du groupe de travail sur les enjeux de conservation (Life Program "Mares temporaires"), Tour-du-Valat, Arles, 26-28 juin 2000 (séminaire)

1999 *Non-symmetric measures of population divergence give new insight for evaluating homogeneity among molecular markers*, par VITALIS R, *in* Réunion du groupe de travail : Development, optimisation and validation of molecular techniques for the measurement of genetic diversity in domestic ungulates (EC funded program to GM Hewitt), Londres, Royaume Uni, 10 octobre 1999 (séminaire)

1998 *Effet de la dispersion biaisée en faveur d'un sexe sur la structure génétique des populations*, VITALIS R. Journées d'Écologie Comportementale, Université de Bourgogne, Dijon, 19-21 Mars 1998 (communication orale)

1997 *Population genetics of subdivided plant populations with a seed bank*, par VITALIS R. Laboratory of Genetics (Prof. Outi Savolainen), Université d'Oulu, Finlande, 8 Octobre 1997 (séminaire)

1997 *Statut génétique et démographique d'une espèce menacée :* Marsilea strigosa, par VITALIS R, COLAS B, RIBA M et OLIVIERI I. 19ème Colloque de Biologie et de Génétique des Populations, Perpignan, 2-5 septembre 1997 (communication orale)

1997 Marsilea strigosa*, statut génétique et démographique*, par VITALIS R, COLAS B, RIBA M et OLIVIERI I. Élaboration du plan de gestion de la reserve naturelle de Roque-Haute, Montpellier, 9 juillet 1997 (communication orale)

1996 *La plupart des éléments transposables dérivent-ils d'une même séquence ancestrale ?*, par CAPY P, VITALIS R, LANGIN T, HIGUET D et BAZIN C. 18ème Colloque de Biologie et Génétique des Populations, Grenoble, 26-30 août 1996 (communication orale)

1995 *Information multigénique et structure des populations*, par VITALIS R et COUVET D. 17ème Colloque de Biologie et de Génétique des Populations, Lyon, 29-31 août 1995 (communication orale)

1995 *Histoire évolutive des éléments transposables à partir de la comparaison des domaines intégrase-transposase*, par CAPY P, VITALIS R, LANGIN T, HIGUET D et BAZIN C. 17ème Colloque de Biologie et de Génétique des Populations, Lyon, 29-31 août 1995 (poster)

1994 *La région DDE chez les éléments transposables*, par CAPY P et VITALIS R. 3ème Réunion sur les Eléments Transposables, Clermont-Ferrand, 4-6 juillet 1994 (poster)

Logiciels

[L8] SelEstim : En cours de développement

[L7] IbdSex : En cours de développement

[L6] rehh : http://cran.r-project.org/web/packages/rehh/index.html

[L5] inferR : https://r-forge.r-project.org/R/?group_id=1291

[L4] coalesceR : https://r-forge.r-project.org/R/?group_id=1281

[L3] DetSel : http://cran.r-project.org/web/packages/DetSel/index.html

[L2] SexDispersal : http://www.ecoanthropologie.cnrs.fr/IMG/zip/SexDispersal_11.zip

[L1] Estim : http://www.ecoanthropologie.cnrs.fr/IMG/zip/ESTIM12-2.zip

Annexe C

Article 1 : organisation sociale en Asie Centrale

Ségurel L., Martínez-Cruz B., Quintana-Murci L., Balaresque P., Georges M., Hegay T., Aldashev A., Nazyrova F., Jobling M. A., Heyer E. et Vitalis R. (2008) Sex-specific genetic structure and social organization in Central Asia : insights from a multi-locus study. *PLoS Genetics* **4(9)** : e1000200

OPEN ACCESS Freely available online

PLoS GENETICS

Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study

Laure Ségurel[1]*, Begoña Martínez-Cruz[1¤], Lluis Quintana-Murci[2], Patricia Balaresque[3], Myriam Georges[1], Tatiana Hegay[4], Almaz Aldashev[5], Firuza Nasyrova[6], Mark A. Jobling[3], Evelyne Heyer[1], Renaud Vitalis[1]

1 Muséum National d'Histoire Naturelle – Centre National de la Recherche Scientifique UMR 5145 – Université Paris 7, Éco-Anthropologie et Ethnobiologie, Musée de l'Homme, Paris, France, **2** Human Evolutionary Genetics Unit, CNRS URA3012, Institut Pasteur, Paris, France, **3** Department of Genetics, University of Leicester, Leicester, United Kingdom, **4** Uzbek Academy of Sciences, Institute of Immunology, Tashkent, Uzbekistan, **5** Institute of Molecular Biology and Medicine, National Center of Cardiology and Internal Medicine, Bishkek, Kyrgyzstan, **6** Tajik Academy of Sciences, Institute of Plant Physiology and Genetics, Dushanbe, Tajikistan

Abstract

In the last two decades, mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY) have been extensively used in order to measure the maternally and paternally inherited genetic structure of human populations, and to infer sex-specific demography and history. Most studies converge towards the notion that among populations, women are genetically less structured than men. This has been mainly explained by a higher migration rate of women, due to patrilocality, a tendency for men to stay in their birthplace while women move to their husband's house. Yet, since population differentiation depends upon the product of the effective number of individuals within each deme and the migration rate among demes, differences in male and female effective numbers and sex-biased dispersal have confounding effects on the comparison of genetic structure as measured by uniparentally inherited markers. In this study, we develop a new multi-locus approach to analyze jointly autosomal and X-linked markers in order to aid the understanding of sex-specific contributions to population differentiation. We show that in patrilineal herder groups of Central Asia, in contrast to bilineal agriculturalists, the effective number of women is higher than that of men. We interpret this result, which could not be obtained by the analysis of mtDNA and NRY alone, as the consequence of the social organization of patrilineal populations, in which genetically related men (but not women) tend to cluster together. This study suggests that differences in sex-specific migration rates may not be the only cause of contrasting male and female differentiation in humans, and that differences in effective numbers do matter.

Citation: Ségurel L, Martínez-Cruz B, Quintana-Murci L, Balaresque P, Georges M, et al. (2008) Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study. PLoS Genet 4(9): e1000200. doi:10.1371/journal.pgen.1000200

Editor: Molly Przeworski, University of Chicago, United States of America

Received April 7, 2008; **Accepted** August 18, 2008; **Published** September 26, 2008

Funding: This work was supported by the Centre National de la Recherche Scientifique (CNRS) ATIP programme (to EH), by the CNRS interdisciplinary programme "Origines de l'Homme du Langage et des Langues" (OHLL) and by the European Science Foundation (ESF) EUROCORES programme "The Origin of Man, Language and Languages" (OMLL). We also thank the "Fondation pour la Recherche Médicale" (FRM) for financial support. LS is financed by the French Ministry of Higher Education and Research. MAJ is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (grant number 057559).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lsegurel@mnhn.fr

¤ Current address: Unidad de Biología Evolutiva, Departamento de Ciencias Experimentales y de la Salud, Universidad Pompeu Fabra, Barcelona, Spain

Introduction

Understanding the extent to which sex-specific processes shape human genetic diversity has long been a matter of great interest for human population geneticists [1,2]. To date, as detailed in Table 1, the focus has mainly been on the analysis of uniparentally inherited markers: mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY). A large number of studies have found that the level of differentiation was greater for the Y chromosome than for mtDNA, both at a global [3] and a local scale [4–11], for a review see [12]. This result has mainly been explained by patrilocality, a widespread tendency for men to stay in their birthplace while women move to their husband's house [13] (see Table 1 for more detailed interpretations). This hypothesis of a higher migration rate of women has been especially strengthened by the comparison of patrilocal and matrilocal populations at a local scale [14–17]. These studies have shown that in patrilocal populations, genetic differentiation is stronger among men than among women, while the reverse is observed in matrilocal populations. It is also noteworthy that the absolute difference between male and female genetic structure is more pronounced in patrilocal than in matrilocal populations [16]. Interestingly, while social practices seem to consistently influence the sex-specific demography at a local scale, the robustness of a sex-specific genetic structure at a global scale is still a challenging issue (see Table 1). A recent analysis of mtDNA and NRY variation at a global scale, which used the same panel of populations for both categories of markers (an omission that was criticized in Seielstad et al.'s [3] study [18]) showed no difference between the male and female genetic structure [19]. Consistent with this result, an analysis of the autosomal and X-linked microsatellite markers in the HGDP-CEPH Human Genome Diversity Cell Line Panel showed no major differences between the demographic history of men and women [20]. The apparent paradox between local and global trends can be resolved though, since the geographical clustering of populations with potentially

Author Summary

Human evolutionary history has been investigated mainly through the prism of genetic variation of the Y chromosome and mitochondrial DNA. These two uniparentally inherited markers reflect the demographic history of males and females, respectively. Their contrasting patterns of genetic differentiation reveal that women are more mobile than men among populations, which might be due to specific marriage rules. However, these two markers provide only a limited understanding of the underlying demographic processes. To obtain an independent picture of sex-specific demography, we developed a new multi-locus approach based on the analysis of markers from the autosomal and X-chromosomal compartments. We applied our method to 21 human populations sampled in Central Asia, with contrasting social organizations and lifestyles. We found that, in patrilineal populations, not only the migration rate but also the number of reproductive individuals is likely to be higher for women. This result does not hold for bilineal populations, for which both the migration rate and the number of reproductive individuals can be equal for both sexes. The social organization of patrilineal populations is the likely cause of this pattern. This study suggests that differences in sex-specific migration rates may not be the only cause of contrasting male and female differentiation in humans, and that differences in effective numbers do matter.

different lifestyles may minimize the differences in sex-specific demography at a global scale [21,22]. It may also be that the global structure reflects more ancient, pre-agricultural, social patterns, as patrilocality may only have increased in human societies only with the recent transition to agriculture [12].

The higher differentiation level found on the NRY as compared to mtDNA at a local scale could also be the consequence of a higher effective number of women, for example through the practice of polygyny, a tendency for men (but not for women) to have multiple mates [4,7,15,23–25], and/or through the paternal transmission of reproductive success [11]. However, the influence of such processes on genetic structure has often been considered as negligible, since realistic rates of polygyny cannot create large differences in male and female genetic structure [3,5,14]. Hence, until now, the effect of local social processes on male and female effective numbers has not been investigated directly, possibly because current methods fail to unravel the relative contribution of effective number and migration rate on the differentiation level [26]. The consequence is that the vast majority of studies fail to show whether the observed differentiation arises from sex-specific differences in migration rate, effective numbers, or both (see Table 1). New methods need therefore to be developed in order to appreciate the relative influence of sex-biased dispersal and differences in effective numbers on genetic structure.

Another limitation to the use of uniparentally inherited markers stems from the fact that each of them is, in effect, a single genetic locus. For that reason, we cannot test for the robustness of the sex-specific genetic structure on these markers. We cannot either rule out the possibility that mtDNA and NRY, which contain multiple linked genes, may be shaped by selection [27,28]. This raises the question of whether results based on uniparentally inherited markers simply reflect stochastic variation, or real differences in sex-specific demography. To answer this question, we propose a novel approach based on the joint analysis of autosomal and X-linked markers. This multi-locus analysis has the potential of providing more robust information, as these markers give an independent picture of sex-specific demography. This approach also aims to disentangle the effects of sex-biased dispersal and effective numbers on genetic structure.

In order to recognize the impact of social organization on these differences, we investigate sex-specific genetic structure in human populations of Central Asia (Figure 1), where various ethnic groups, characterized by different languages, lifestyles and social organizations, co-exist. Although all groups share a patrilocal organization, Tajiks (sedentary agriculturalists) are bilineal, i.e. they are organized into nuclear or extended families where blood links and rights of inheritance through both male and female ancestors are of equal importance, and they preferentially establish endogamous marriages with cousins. By contrast, Kazaks, Karakalpaks, Kyrgyz and Turkmen (traditionally nomadic herders) are patrilineal, i.e. they are organized into paternal descent groups (tribes, clans, lineages), and they practice exogamous marriages, in which a man chooses a bride from a different clan.

Results/Discussion

Uniparentally-Inherited Markers

We sampled 780 healthy adult men from 10 populations of bilineal agriculturalists and 11 populations of patrilineal herders from West Uzbekistan to East Kyrgyzstan, representing 5 ethnic groups (Tajiks, Kyrgyz, Karakalpaks, Kazaks, and Turkmen) (see Figure 1 and Table 2). We genotyped all bilineal populations, and 8 out of 11 patrilineal populations at the HVS-I locus of mtDNA, and at 11 microsatellite markers on the NRY (for more details on the markers used, see Table 3). The overall genetic differentiation was higher for NRY, as compared to mtDNA, both among the 10 bilineal agriculturalist populations $\left(F_{\mathrm{ST}}^{(Y)}=0.069 \text{ vs. } F_{\mathrm{ST}}^{(mtDNA)}=0.034\right)$, and among the subset of 8 patrilineal herder populations $\left(F_{\mathrm{ST}}^{(Y)}=0.177 \text{ vs. } F_{\mathrm{ST}}^{(mtDNA)}=0.010\right)$. Assuming an island model of population structure, this implies that female migration rate (m_{f}), and/or the effective number of females (N_{f}), is higher than of the corresponding parameters for males (m_{m} and N_{m}). These results also suggest that the differences in sex-specific genetic structure are much more pronounced in the patrilineal herders than in the bilineal agriculturalists. From the above F_{ST} estimates, we obtained the female-to-male ratio of the effective number of migrants per generation (see the Methods section for details): $N_{\mathrm{f}}m_{\mathrm{f}}/N_{\mathrm{m}}m_{\mathrm{m}}\approx 2.1$ for bilineal populations and $N_{\mathrm{f}}m_{\mathrm{f}}/N_{\mathrm{m}}m_{\mathrm{m}}\approx 21.6$ for patrilineal populations. The ratio in patrilineal populations is thus one order of magnitude higher than in bilineal populations. However, since each of these markers is a single genetic locus, we cannot test for the robustness of the sex-specific genetic structure on these markers. We therefore examined the amount of information contained in multi-locus data on autosomal and X-linked markers, both of which average over male and female histories.

A New Multi-Locus Approach

In the infinite island model of population structure with two classes of individuals (males and females), we obtained the following expressions of F_{ST} (see the Methods section for details):

$$F_{\mathrm{ST}}^{(A)} \approx \frac{1}{1+4\frac{4N_{\mathrm{f}}N_{\mathrm{m}}}{N_{\mathrm{f}}+N_{\mathrm{m}}}\frac{m_{\mathrm{f}}+m_{\mathrm{m}}}{2}}, \quad (1)$$

for autosomal genes, and

$$F_{\mathrm{ST}}^{(X)} \approx \frac{1}{1+4\frac{9N_{\mathrm{f}}N_{\mathrm{m}}}{2N_{\mathrm{f}}+4N_{\mathrm{m}}}\frac{2m_{\mathrm{f}}+m_{\mathrm{m}}}{3}}, \quad (2)$$

Table 1. Human sex-specific demography inferred from genetic data.

Region	Markers	Method	Social organization[a]	Differences in demographic parameters between males and females[b]			References
				Sex-biased migration		Skewed effective population size	
GLOBAL	mtDNA, NRY SNPs[c]	Genetic structure (AMOVA[d])	NA[e]	None			[19]
GLOBAL	Autosomal STRs[f], X-linked STRs	Genetic structure (AMOVA)	NA	None			[20]
GLOBAL	mtDNA, NRY SNPs[c]	Coalescent-based (TMRCA[g] estimates)	NA	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny)	[24]
GLOBAL[h]	mtDNA, NRY STRs+SNPs, Autosomal STRs+SNPs	Genetic structure (F_{ST})	NA	$m_f > m_m$ (patrilocality)		Considered as negligible[i]	[3]
GLOBAL[h]	NRY SNPs	Coalescent-based (mismatch distributions)	NA	Not considered[j]		$N_f > N_m$ (polygyny)	[23]
India	mtDNA	Genetic structure (R_{ST}, haplotype sharing)	Endogamy, patrilocality	None			[21]
	NRY STRs		Endogamy, matrilocality	None			
Sinai peninsula	mtDNA, NRY	Genetic diversity	Endogamy and rare patrilocal exogamy, polygyny	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny)	[4]
West New Guinea	mtDNA, NRY STRs+SNPs	Genetic structure and diversity (F_{ST}, R_{ST}, haplotype diversity)	Exogamy, patrilocality, patrilineality, polygyny	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny, warfare)	[7]
Sub-Saharan Africa	mtDNA, NRY STRs+SNPs	Genetic structure (AMOVA)	FPP[k]: patrilocality, high polygyny	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny)	[15]
			HGP[l]: moderate patrilocality, low polygyny	$m_f < m_m$ (multilocality)	and/or	$N_f < N_m$	
Thailand	mtDNA, NRY STRs	Coalescent-based (Approximate Bayesian Computation)	Patrilocality	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (patrilocality)	[16]
			Matrilocality	$m_f < m_m$ (matrilocality)	and/or	$N_f < N_m$ (matrilocality)	
Eastern North America	mtDNA, NRY STRs+SNPs	Genetic structure (AMOVA), coalescent-based (MIGRATE[m])	Patrilocality, patrilineality	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (patrilocality)	[17]
			Matrilocality, matriliny	$m_f < m_m$ (matrilocality)	and/or	$N_f < N_m$ (matrilocality)	
Central Asia (pastoral populations)	mtDNA, NRY STRs	Genetic structure and diversity (AMOVA, R_{ST})	Exogamy, patrilineality	$m_f > m_m$ (patrilineality, exogamy)	and/or	$N_f > N_m$ (patrilineality,VRS[n])	[11]
New Britain	mtDNA, NRY SNPs, X-linked loci	Coalescent-based (θ^o and TMRCA estimates)	No strong endogamy, ambilocality, polygyny	$m_f < m_m$	and	$N_f > N_m$ (polygyny)	[25]
Central Asia	mtDNA, NRY STRs	Genetic structure (AMOVA)	Exogamy, patrilocality, polygyny	$m_f > m_m$ (patrilocality)		Considered as negligible	[5]
Thailand	mtDNA, NRY STRs	Genetic structure and diversity (haplotype diversity, R_{ST})	Patrilocality	$m_f > m_m$ (patrilocality)		Considered as negligible	[14]
			Matrilocality	$m_f < m_m$ (matrilocality)		Considered as negligible	
Sub-Saharan Africa[h]	mtDNA, NRY SNPs	Genetic structure and diversity (haplotype diversity, AMOVA)	NA[c]	$m_f < m_m$		Not considered	[22]
Continental Asia[h]	mtDNA, NRY SNPs	Genetic structure (F_{ST})	NA[c]	$m_f > m_m$ (patrilocality)		Not considered	[6]
Russia	mtDNA, NRY SNPs	Genetic structure (F_{ST})	Patrilocality, patrilineality	$m_f > m_m$ (patrilocality)		Not considered	[8]
Caucasus	mtDNA, NRY SNPs	Genetic structure (AMOVA)	NA	$m_f > m_m$ (patrilocality)		Not considered	[9]
Turkey	mtDNA, NRY STRs+SNPs	Genetic structure (AMOVA)	NA	$m_f > m_m$ (patrilocality)		Not considered	[10]

Table 1. cont.

This table summarizes the observed patterns of sex-specific differences in demographic parameters reported in a number of recent studies. The first column lists the location of the sampled populations, or indicates whether the study is conducted at a global scale. The second column gives the markers used, and the third column indicates the statistical methods employed. The fourth column provides indications on social organization, available a priori for the populations under study. In the fifth and sixth columns, the authors' interpretations of sex-specific differences in demographic parameters are given, with respect to skewed gene flow and/or effective numbers.

[a]Indications on social organization, marriage rules, etc., as provided by the authors.

[b]The differences in demographic parameters between males and females, as inferred by the authors, are given in terms of sex-biased gene flow, and skewed effective numbers; the authors' interpretation to the observed pattern is given in parentheses, when available.

[c]Single nucleotide polymorphisms.

[d]Analysis of molecular variance [69].

[e]Not available (no detailed information given by the authors concerning social organization, marriage rules, etc.).

[f]Short tandem repeats.

[g]Time to the most recent common ancestor.

[h]mtDNA and NRY were not sampled in the same individuals or populations.

[i]The authors discussed a possible difference in demographic parameters between males and females, but considered it as negligible.

[j]The authors did not consider this pattern.

[k]Food-producer populations.

[l]Hunter-gatherer populations.

[m]Monte Carlo Markov chain method to estimate population sizes and migration rates [70].

[n]Variance in Reproductive Success.

[o]population-mutation parameter.

doi:10.1371/journal.pgen.1000200.t001

for X-linked genes. A special case of interest occurs when $F_{ST}^{(X)} = F_{ST}^{(A)}$, i.e. when the differentiation of X-linked genes exactly equals that of autosomal genes. Combining eqs (1) and (2), we find that this occurs for $\frac{m_f}{m} = \left(5 - 4\frac{N_f}{N}\right)/3$, with $N = N_f + N_m$ and $m = m_f + m_m$. Furthermore, as shown in Figure 2, if we observe a lower genetic differentiation of autosomal markers, as compared to X-linked markers (blue zone in Figure 2), this suggests that $\frac{m_f}{m} < \left(5 - 4\frac{N_f}{N}\right)/3$. This may happen, e.g., for $N_f = N_m$ and $m_f = m_m$, i.e. for equal effective numbers of males and females and unbiased dispersal. But if autosomal markers are more differentiated than X-linked markers ($F_{ST}^{(A)} > F_{ST}^{(X)}$, see the red upper-right triangle in Figure 2), this implies that $\frac{m_f}{m} > \left(5 - 4\frac{N_f}{N}\right)/3$. In this case, since m_f/m and N_f/N are ratios varying between 0 and 1, the effective number of females must be higher than that of males ($N_f > N_m$), and the female migration rate must be higher than half the male migration rate ($m_f > m_m/2$). Hence, a prediction from this model is that when $F_{ST}^{(A)} > F_{ST}^{(X)}$, the effective number of females is higher than that of males, whatever the pattern of sex-specific dispersal. This suggests that it is indeed possible to test for differences in effective numbers between males and females from the joint analysis of autosomal and X-linked data. We note however that when $F_{ST}^{(X)} > F_{ST}^{(A)}$, we cannot conclude on the relative male and female effective numbers and migration rates.

We tested the above prediction in the 10 bilineal agriculturalist populations and 11 patrilineal herder populations sampled in Central Asia by comparing the genetic structure estimated from 27 unlinked polymorphic autosomal microsatellite markers ($AR = 16.2$, $H_e = 0.803$ on average) to that from 9 unlinked polymorphic X-linked microsatellite markers ($AR = 12.6$, $H_e = 0.752$ on average) (for more details on the markers used, see Table 4). Overall heterozygosity was not significantly different between X-linked and autosomal markers, neither in the pooled sample (two-tailed Wilcoxon sum rank test; $p = 0.09$), nor in the bilineal agriculturalists ($p = 0.13$) or the patrilineal herders ($p = 0.12$). The overall population structure was significantly higher for autosomal as compared to X-linked markers among patrilineal herders: $F_{ST}^{(A)} = 0.008\ [0.006 - 0.010]$ and $F_{ST}^{(X)} = 0.003\ [0.001 - 0.006]$ (one-tailed Wilcoxon sum rank test; $H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}$; $p = 0.02$). Among bilineal agriculturalists, the result was not significant: $F_{ST}^{(A)} = 0.014\ [0.012 - 0.016]$ and $F_{ST}^{(X)} = 0.013\ [0.008 - 0.018]$ ($p = 0.36$). From these results, and following our model predictions, we conclude that in patrilineal herders (where $F_{ST}^{(A)} > F_{ST}^{(X)}$), the effective number of females is higher than that of males. This conclusion does not hold for the bilineal agriculturalists.

From our model, it is possible to get more precise indications on the sets of (N_f/N, m_f/m) values that are compatible with our data. Rearranging eqs (1–2), we get:

$$\frac{1 - 1/F_{ST}^{(X)}}{1 - 1/F_{ST}^{(A)}} = \frac{3}{4}\frac{(1 + m_f/m)}{(2 - N_f/N)}, \tag{3}$$

i.e.:

$$F_{ST}^{(X)} = \frac{4F_{ST}^{(A)}}{4F_{ST}^{(A)} - 3\left(F_{ST}^{(A)} - 1\right)\left(\frac{1 + m_f/m}{2 - N_f/N}\right)}. \tag{4}$$

For any given set of (N_f/N, m_f/m) values, we can therefore calculate from eq. (4) the expected value of $F_{ST}^{(X)}$ for each $F_{ST}^{(A)}$

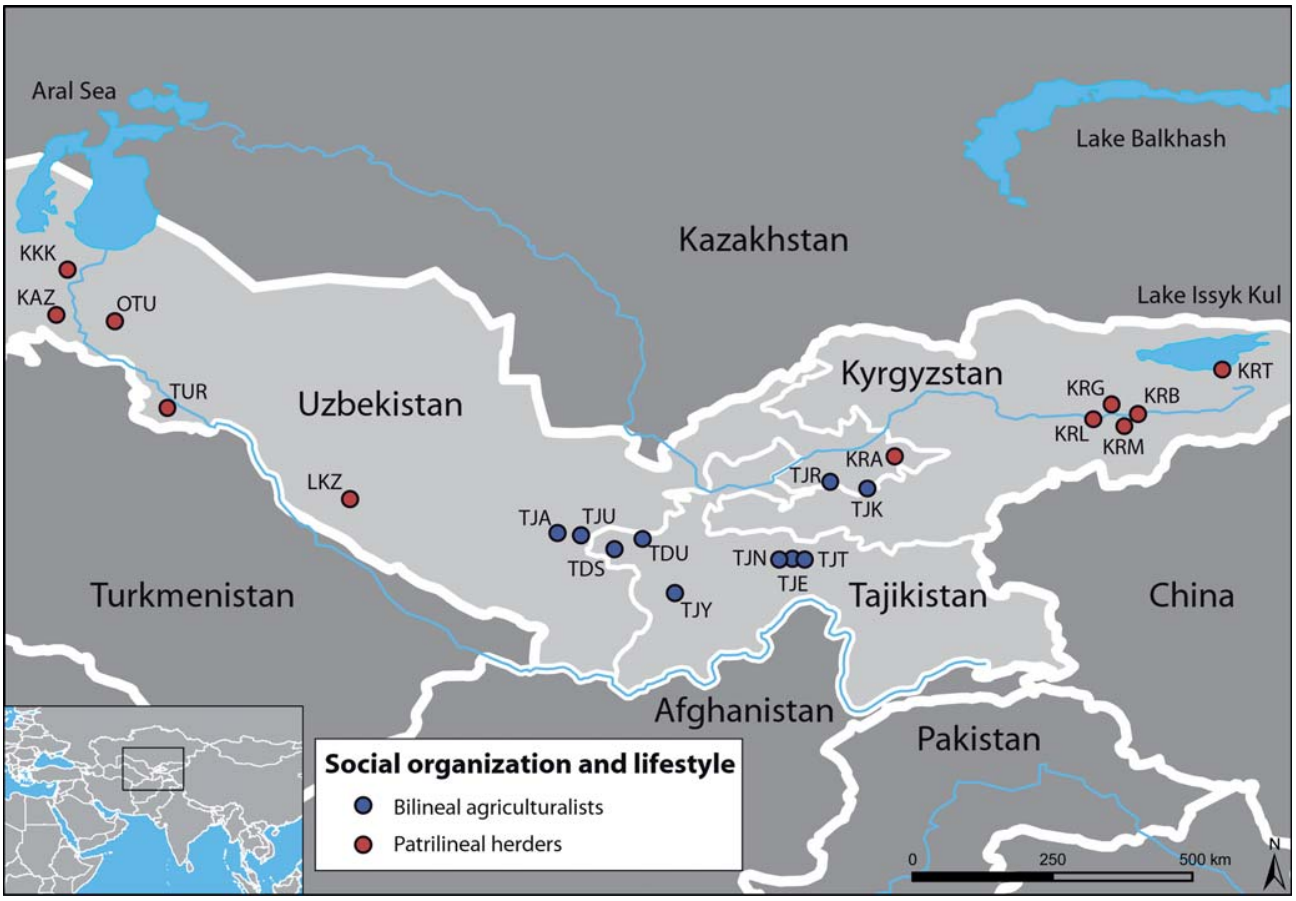


Figure 1. Geographic map of the sampled area, with the 21 populations studied. Bilineal agriculturalist populations are in blue (Tajiks); Patrilineal herders with a semi-nomadic lifestyle are in red (Kazaks, Karakalpaks, Kyrgyz and Turkmen).
doi:10.1371/journal.pgen.1000200.g001

estimate in the dataset. We can then test the null hypothesis $H_0 : F_{ST}^{(X)} = 4F_{ST}^{(A)} \Big/ \left[4F_{ST}^{(A)} - 3\left(F_{ST}^{(A)} - 1\right)\left(\frac{1+m_f/m}{2-N_f/N}\right)\right]$ by comparing the distribution of observed and expected $F_{ST}^{(X)}$ values. If the hypothesis can be rejected at the $\alpha = 0.05$ level, then the corresponding set of $(N_f/N,\ m_f/m)$ values can also be rejected. Following Ramachandran et al. [20], we varied the values of the ratios N_f/N and m_f/m (respectively, the female fraction of effective number, and the female fraction of the total migration rate) from 0 to 1, with an interval of 0.01 between consecutive values. For each set of $(N_f/N,\ m_f/m)$ values, we applied the transformation in eq. (4) to each of the 27 locus-specific $F_{ST}^{(A)}$ values observed. Thus, for each set of $(N_f/N,\ m_f/m)$ values, we obtained 27 expected values of $F_{ST}^{(X)}$, given our data. These expected values of $F_{ST}^{(X)}$ were then compared to the 9 observed locus-specific $F_{ST}^{(X)}$ in our dataset, and we calculated the p-value for a two-sided Wilcoxon sum rank test between the list of 27 expected $F_{ST}^{(X)}$ values and the 9 $F_{ST}^{(X)}$ observed in the dataset. The results are depicted in Figure 3. Significant p-values ($p \leq 0.05$) correspond to a significant difference between the observed and expected values, thus to sets of $(N_f/N,\ m_f/m)$ values that are rejected, given our data (see the blue region in Figure 3). Conversely, non-significant p-values ($p > 0.05$) correspond to sets of $(N_f/N,\ m_f/m)$ values that cannot be rejected (see the red region in Figure 3).

For the patrilineal herder populations (Figures 3A–3B), most sets of $(N_f/N,\ m_f/m)$ values are rejected, except those corresponding to larger effective numbers for females (from Figures 3A–3B: $N_f/N > 0.55$, i.e. $N_f > 1.27N_m$) and $m_f > 0.67m_m$. Because the multi-locus estimate of $F_{ST}^{(A)}$ is significantly higher than the estimate of $F_{ST}^{(X)}$, we expected to find such patterns of non-significant values (see Figure 2). For the bilineal agriculturalist populations, we could not reject the hypothesis that the effective numbers and migration rates are equal across males and females or even lower in females (see Figures 3C–3D). This is also reflected by the fact that the estimates of $F_{ST}^{(A)}$ were not significantly higher than the estimates of $F_{ST}^{(X)}$ in those populations.

Finally, we have shown that the effective number of women is higher than that of men among patrilineal herders, but not necessarily among bilineal agriculturalists. Furthermore, a close inspection of the results depicted in Figures 3A and 3B reveals that, among herders, we reject all the sets of $(N_f/N,\ m_f/m)$ values for which $m_f < m_m$ at the $\alpha = 0.10$ level. This is not true for agriculturalists. This suggests that the migration rates are also likely to be higher for women than for men in patrilineal populations, as compared to bilineal populations (compare Figures 3B and 3D). Although both groups are patrilocal, such a difference in sex-specific migration patterns might be expected, since patrilineal herders are exogamous (among clans) and bilineal agriculturalists are preferentially endogamous. For example, it was observed that in patrilocal and matrilocal Indian populations, where migrations are strictly confined within endogamous groups, sex-specific patterns were not influenced by post-marital residence [21].

Table 2. Sample description.

Sampled populations (area)	Acronym	Location	Long.	Lat.	n_X	n_A	n_Y	n_{mt}
Bilineal agriculturalists								
Tajiks (Samarkand)	TJA	Uzbekistan/Tajikistan border	**39.54**	**66.89**	26	31	32	32
Tajiks (Samarkand)	TJU	Uzbekistan/Tajikistan border	**39.5**	**67.27**	27	29	29	29
Tajiks (Ferghana)	TJR	Tajikistan/Kyrgyzstan border	**40.36**	**71.28**	30	29	29	29
Tajiks (Ferghana)	TJK	Tajikistan/Kyrgyzstan border	**40.25**	**71.87**	26	26	35	40
Tajiks (Gharm)	TJE	Northern Tajikistan	**39.12**	**70.67**	29	25	27	31
Tajiks (Gharm)	TJN	Western Tajikistan	**38.09**	**68.81**	33	24	30	35
Tajiks (Gharm)	TJT	Northern Tajikistan	**39.11**	**70.86**	31	25	30	32
Tajiks (Penjinkent)	TDS	Uzbekistan/Tajikistan border	**39.28**	**67.81**	30	25	31	31
Tajiks (Penjinkent)	TDU	Uzbekistan/Tajikistan border	**39.44**	**68.26**	40	25	31	40
Tajiks (Yagnobs from Douchambe)	TJY	Western Tajikistan	**38.57**	**68.78**	39	25	36	40
Patrilineal herders with a semi-nomadic lifestyle								
Karakalpaks (Qongrat from Karakalpakia)	KKK	Western Uzbekistan	**43.77**	**59.02**	56	45	54	55
Karakalpaks (On Tört Uruw from Karakalpakia)	OTU	Western Uzbekistan	**42.94**	**59.78**	49	45	54	53
Kazaks (Karakalpakia)	KAZ	Western Uzbekistan	**43.04**	**58.84**	47	49	50	50
Kazaks (Bukara)	LKZ	Southern Uzbekistan	**40.08**	**63.56**	20	25	20	31
Kyrgyz (Andijan)	KRA	Tajikistan/Kyrgyzstan border	**40.77**	**72.31**	31	45	46	48
Kyrgyz (Narin)	KRG	Middle Kyrgyzstan	**41.6**	**75.8**	20	18	20	20
Kyrgyz (Narin)	KRM	Middle Kyrgyzstan	**41.45**	**76.22**	21	21	22	26
Kyrgyz (Narin)	KRL	Middle Kyrgyzstan	**41.36**	**75.5**	36	22	-	-
Kyrgyz (Narin)	KRB	Middle Kyrgyzstan	**41.25**	**76**	31	24	-	-
Kyrgyz (Issyk Kul)	KRT	Eastern Kyrgyzstan	**42.16**	**77.57**	33	37	-	-
Turkmen (Karakalpakia)	TUR	Western Uzbekistan	**41.55**	**60.63**	42	47	51	51

Long., longitude; Lat., latitude. n_X, n_A, n_Y and n_{mt}: sample size for X-linked, autosomal, Y-linked and mitochondrial markers, respectively.
doi:10.1371/journal.pgen.1000200.t002

Table 3. Level of diversity and differentiation for NRY markers and mtDNA.

NRY markers			F_{ST}	
Locus name	**Allelic richness (AR)**	H_e	**Herders**	**Agriculturalists**
DYS426	4	0.500	0.3326	0.0068
DYS393	8	0.492	0.1095	0.0517
DYS390	8	0.739	0.1229	0.1253
DYS385 a/b	15	0.858	0.1414	0.0278
DYS388	9	0.531	0.3003	0.0736
DYS19	7	0.743	0.1081	0.1310
DYS392	10	0.516	0.1345	0.0701
DYS391	7	0.495	0.2533	0.0686
DYS389I	6	0.541	0.1537	0.1395
DYS439	7	0.725	0.1638	0.0291
DYS389II	8	0.763	0.1556	0.0395
mtDNA			F_{ST}	
Locus name	**Polymorphic sites**	H_e	**Herders**	**Agriculturalists**
HVS-I	121	0.0156	0.0098	0.0343

We calculated the total allelic richness (*AR*) (over all populations) and the expected heterozygosity H_e [55] using Arlequin version 3.1 [56]. Genetic differentiation among populations was measured both per locus and overall loci, using Weir and Cockerham's F_{ST} estimator [57], as calculated in GENEPOP 4.0 [58]. We calculated the total number of polymorphic sites, the unbiased estimate of expected heterozygosity H_e [55], and F_{ST} using Arlequin version 3.1 [56].
doi:10.1371/journal.pgen.1000200.t003

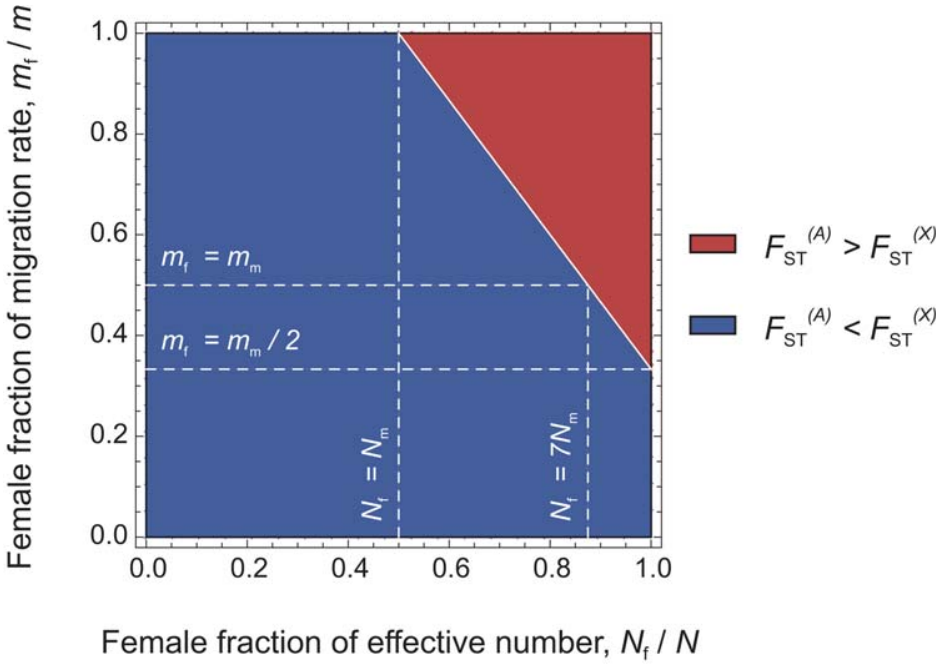


Figure 2. Diagram representing the relative values of expected genetic differentiation for autosomal markers $\left(F_{ST}^{(A)}\right)$ and for X-linked markers $\left(F_{ST}^{(X)}\right)$. In the red upper right triangle, the F_{ST} estimates for autosomal markers are higher than for X-linked markers. In this case, N_f/N is necessarily larger than 0.5. In the blue region of the figure, the F_{ST} estimates for autosomal markers are lower than for X-linked markers. The white plain line, at which $\frac{m_f}{m} = \left(5-4\frac{N_f}{N}\right)/3$, represents the set of (N_f/N, m_f/m) values where the autosomal and X-linked F_{ST} estimates are equal. In this case $\left(F_{ST}^{(X)} = F_{ST}^{(A)}\right)$, if $N_f = N_m$, then the lower effective size of X-linked markers (which would be three-quarters that of autosomal markers) can only be balanced by a complete female-bias in dispersal ($m_f/m = 1$). Conversely, if $m_f = m_m$, the large female fraction of effective numbers compensates exactly the low effective size of X-linked markers only for $N_f = 7N_m$. Last, if $m_f = m_m/2$, then the autosomal and X-linked F_{ST} estimates can only be equal as the number of males tends towards zero.
doi:10.1371/journal.pgen.1000200.g002

What Could Explain a Larger Effective Number of Females?

While an influence of post-marital residence on the migration rate of women and men has already been widely proposed [14–17] (see also Table 1), the factors that may locally affect the effective number of women, relatively to that of men, are not well recognized. As seen in Table 1, although a number of studies have compared matrilocal and patrilocal populations, few have compared contrasting groups of populations with respect to other factors as, e.g., the tendency for polygyny [15]. Furthermore, a number of these studies lack ethnological information a priori, concerning social organization, marriage rules, etc., which makes interpretation somewhat difficult (see Table 1). Here, we compared two groups of patrilocal populations with contrasting social organizations, and at least five non-mutually exclusive interpretations for a larger effective number of females can be invoked:

(*i*) *Social organization*, i.e. the way children are affiliated to their parents, can deeply affect sex-specific genetic variation. In Central Asia, herder populations are organized in patrilineal descent groups (tribes, clans, lineages). This implies that children are systematically affiliated with the descent groups of the father. Chaix et al. [11] showed that the average number of individuals carrying the same Y chromosome haplotype was much higher in patrilineal herder populations than in bilineal agriculturalist populations (where children are affiliated both to the mother and the father). These "identity cores" would be the direct consequence of the internal dynamics of their patrilineal organization. Indeed, the descent groups are not formed randomly and related men tend to cluster together, e.g. through the recurrent lineal fission of one population into new groups. This particular dynamics increases relatedness among men, and may therefore reduce the effective number of men, as compared to women.

(*ii*) Indirectly, the social organization can also deflate the effective number of men through *the transmission of reproductive success* [29] if this success is culturally transmitted exclusively from fathers to sons. Because herders are patrilineal (so that inheritance is organized along paternal descent groups), social behaviors are more likely to be inherited through the paternal line of descent only. It has recently been argued that the rapid spread of Genghis Khan's patrilineal descendants throughout Central Asia was explained by this social selection phenomenon [30]. The correlation of fertility through the patriline has also been described in patrilineal tribes in South America [31]. By contrast, in bilineal societies such as the agriculturalists of Central Asia, social behaviors that influence reproductive success are more likely to be transmitted by both sexes. Furthermore, differences of cultural transmission of fitness between hunter-gatherers and agriculturalists have already been reported [32]. Interestingly, a slightly higher matrilineal intergenerational correlation in offspring number has been observed in the Icelandic population, which suggests that in some populations, reproductive behaviors can be maternally-inherited [33].

Table 4. Level of diversity and differentiation for X-linked and autosomal markers.

			F_{ST}	
Locus name	**Allelic richness (AR)**	H_e	**Herders**	**Agriculturalists**
X-linked markers				
CTAT014	19	0.746	0.0018	0.0225
GATA124E07	15	0.847	0.0024	0.0136
GATA31D10	8	0.697	0.0069	0.0007
ATA28C05	7	0.722	0.0086	0.0179
AFM150xf10	14	0.832	−0.0021	0.0152
GATA100G03	14	0.734	−0.0019	0.0084
AGAT121P	15	0.593	−0.0016	0.0048
ATCT003	10	0.797	0.0095	0.0261
GATA31F01	11	0.804	0.0069	0.0053
Autosomal markers				
AFM249XC5	19	0.848	0.0080	0.0081
ATA10H11	13	0.680	0.0128	0.0193
AFM254VE1	14	0.837	0.0105	0.0086
AFMA218YB5	14	0.852	0.0030	0.0151
GGAA7G08	22	0.896	0.0096	0.0138
GATA11H10	16	0.776	0.0017	0.0056
GATA12A07	16	0.857	0.0001	0.0163
GATA193A07	15	0.825	0.0064	0.0087
AFMB002ZF1	11	0.820	0.0028	0.0169
AFMB303ZG9	16	0.858	0.0090	0.0148
ATA34G06	12	0.675	0.0088	0.0132
GATA72G09	18	0.884	−0.0023	0.0131
GATA22F11	21	0.897	0.0152	0.0144
GGAA6D03	13	0.831	0.0048	0.0176
GATA88H02	17	0.892	0.0063	0.0056
SE30	15	0.762	0.0084	0.0103
GATA43C11	16	0.870	0.0028	0.0093
AFM203YG9	14	0.753	0.0105	0.0084
AFM157XG3	13	0.753	0.0147	0.0196
UT2095	16	0.738	0.0032	0.0112
GATA28D01	25	0.896	0.0156	0.0139
GGAA4B09	19	0.707	0.0034	0.0208
ATA3A07	12	0.746	0.0078	0.0070
AFM193XH4	11	0.716	0.0164	0.0129
GATA11B12	26	0.896	0.0104	0.0265
AFM165XC11	13	0.785	0.0058	0.0185
AFM248VC5	20	0.620	0.0246	0.0145

We calculated the allelic richness (*AR*) and unbiased estimates of expected heterozygosity H_e [55], obtained both by locus and on average with Arlequin version 3.1 [56]. Genetic differentiation among populations was measured both per locus and overall loci, using Weir and Cockerham's F_{ST} estimator [57] as calculated in GENEPOP 4.0 [58].

(*iii*) *Polygyny*, in which the husband may have multiple wives, has often been invoked as a factor that could reduce the effective number of men [4,7,15,23–25]. While we could not find any evidence of polygyny in present-day Central Asian populations, this custom was traditionally practiced in the nomadic herder Kazak populations, although limited to the top 10 percent of men from the highest social rank [5,34]. Hence, even though we lack ethnological data to determine to what extent herders are or were practicing polygyny in a recent past, the practice of polygyny among herders in Central Asia might have influenced (at least partially) the observed differences in men and women effective numbers.

(*iv*) *Recurrent bottlenecks in men* due to a higher pre-reproductive mortality could also severely reduce the effective numbers of men. From the study of several groups in West Papua and Papua New Guinea [7,35], it appears that warfare may indeed lead to the quasi-extinction of adult men in some communities, while the mass killing of adult women is far more rarely reported. However, this differential mortality could also be balanced by potentially high death rates of women during childbirth. In any case, a differential mortality is equally likely to arise in herder and agriculturalist populations. It may therefore not be relevant in explaining why we detect higher effective numbers of women (as compared to men) in patrilineal herders and not in bilineal agriculturalists.

(*v*) Since our approach implicitly assumes equal male and female generation time, the observed higher effective number of women, relatively to that of men, could result from a *shorter generation time for women*, due to the tendency of women to reproduce earlier in life than men and the ability of men to reproduce at a later age than women. This has indeed been described in a number of populations with different lifestyles, from complete genealogical records or mean-age-at-first-marriage databases [33,36,37]. It has even been proposed to be a nearly universal trait in humans, although its magnitude varies across regions and cultures [37]. Tang et al. [38] suggested that accounting for longer generation time in males could minimize the difference between maternal and paternal demography. However, the differences in sex-specific generation times that have been reported (e.g., 28 years for the matrilines and 31 years for the patrilines in Iceland [33], 29 years for the matrilines and 35 years for the patrilines in Quebec [36]) are unlikely to explain the observed differences in male and female effective numbers [24].

Limits of the Approach

There might also be non-biological explanations of our results, however, as they are based on the simplifying assumptions of Wright's infinite island model of population structure [39]. This model assumes (*i*) that there is no selection and that mutation is negligible, (*ii*) that each population has the same size, and sends and receives a constant fraction of its individuals to or from a common migrant pool each generation (so that geographical structure is absent), and (*iii*) that equilibrium is reached between migration, mutation and drift. On the first point, we did not find any evidence of selection, for any marker, based on Beaumont and Nichols' method [40] for detecting selected markers from the analysis of the null distribution generated by a coalescent-based simulation model (data not shown). As for the second point, we tested for the significance of the correlation between the pairwise $F_{ST}/(1-F_{ST})$ estimates and the natural logarithm of their geographical distances [41]. We found no evidence for isolation by distance, either for X-linked markers ($p = 0.47$ for agriculturalists, $p = 0.24$ for herders), or for autosomal markers ($p = 0.92$ for agriculturalists, $p = 0.45$ for herders). As for the third point, the X-to-autosomes (X/A) effective size ratio can significantly deviate from the expected three-quarters (assuming equal effective numbers of men and women) following a bottleneck or an

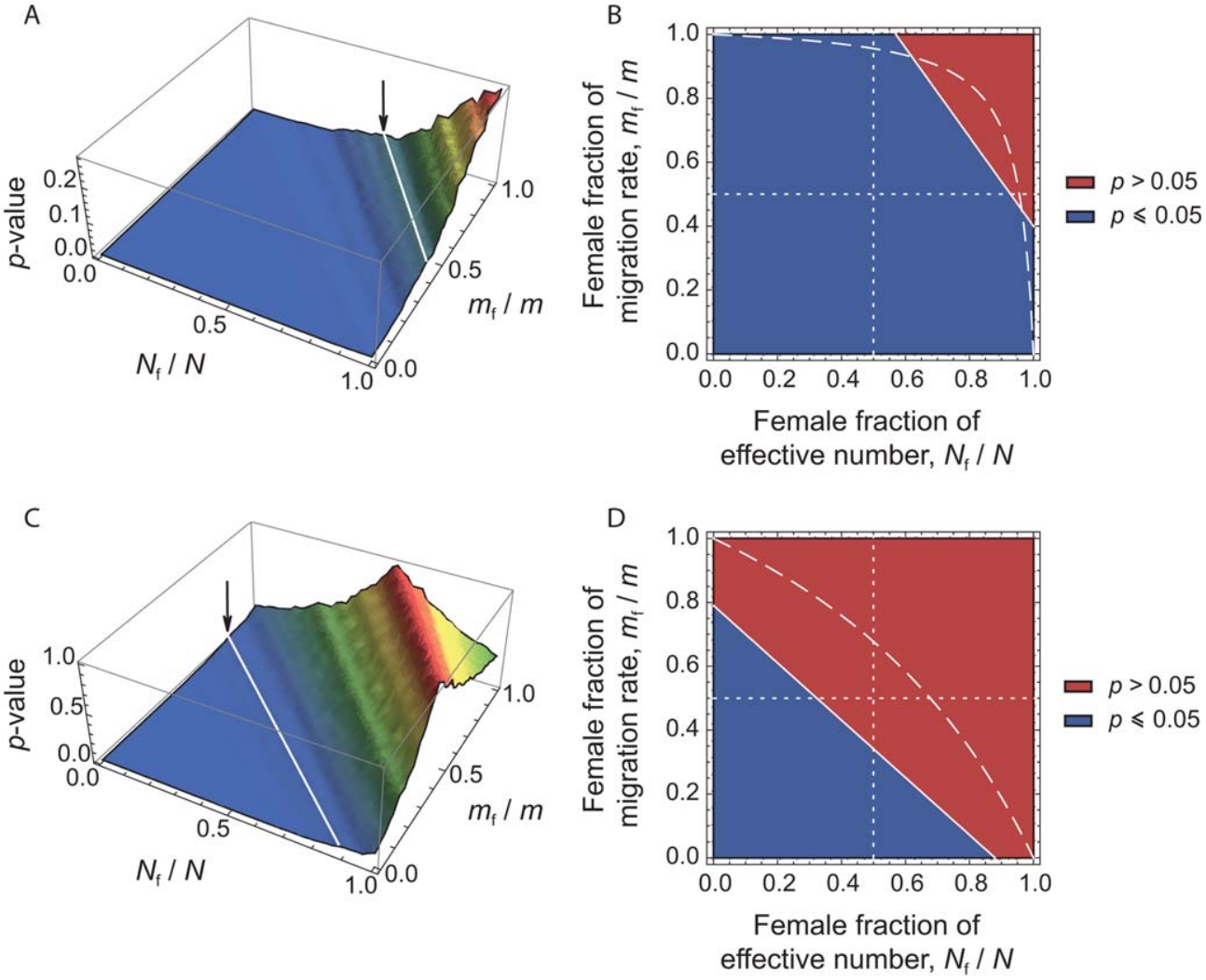


Figure 3. *p*-values of Wilcoxon tests plotted in the (N_f/N, m_f/m) parameter space. For each set of (N_f/N, m_f/m) values, we applied the transformation in eq. (4), and tested whether our data on autosomal and X-linked markers were consistent, given the hypothesis defined by the set of (N_f/N, m_f/m) values. (A) Surface plot of the *p*-values, as a function of the female fraction of effective number and the female fraction of migration rate, for the herders (11 populations). The arrow indicates the line that separates the region where $p \leq 0.05$ from that where $p > 0.05$. Non-significant *p*-values ($p > 0.05$) correspond to the values of (N_f/N, m_f/m) that could not be rejected, given our data. (B) Contour plots, for the same data. The dashed line indicates the range of (N_f/N, m_f/m) values inferred from the ratio of NRY and mtDNA population structure, as obtained from the relationship: $N_f m_f / N_m m_m = \left(1 - 1/F_{ST}^{(mtDNA)}\right) / \left(1 - 1/F_{ST}^{(Y)}\right)$. The dotted lines correspond to the cases where $N_f = N_m$ (vertical line) and $m_f = m_m$ (horizontal line). (C) and (D) as (A) and (B), respectively, for the agriculturalists (10 populations).
doi:10.1371/journal.pgen.1000200.g003

expansion [42]. This is because X-linked genes have a smaller effective size, and hence reach equilibrium more rapidly. After a reduction of population size, the X/A diversity ratio is lower than expected, while after an expansion, the diversity of X-linked genes recovers faster than on the autosomes, and the X/A diversity ratio is then closer to unity. In the latter case, $F_{ST}^{(X)}$ would be reduced and could then tend towards $F_{ST}^{(A)}$. However, neither reduction nor expansion should lead to $F_{ST}^{(X)} < F_{ST}^{(A)}$, as we found in herder populations of Central Asia. Therefore, we do not expect the limits of Wright's island model to undermine our approach.

Evaluation by Means of Stochastic Simulations

We aimed to investigate to what extent the approach proposed here is able to detect differences in male and female effective numbers. To do this, we performed coalescent simulations in a finite island model, for a wide range of (N_f/N, m_f/m) values. The simulation parameters were set to match those of our dataset: 11 sampled demes, 30 males genotyped at 27 autosomal and 9 X-linked markers per deme (for further details concerning the simulations, see the Methods section). We used 1421 sets of (N_f/N, m_f/m) values, covering the whole parameter space (represented as white dots in Figure 4B). For each set of (N_f/N, m_f/m) parameter values, we simulated 100 independent datasets. For each dataset, we calculated the estimates of $F_{ST}^{(A)}$ and $F_{ST}^{(X)}$ at all loci, and we calculated the *p*-value for a one-sided Wilcoxon sum rank test for the list of 27 $F_{ST}^{(X)}$ and 9 $F_{ST}^{(X)}$ estimates $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$. Hence, for each set of (N_f/N, m_f/m) parameter values, we could calculate the proportion of significant tests at the $\alpha = 0.05$ level, among the 100

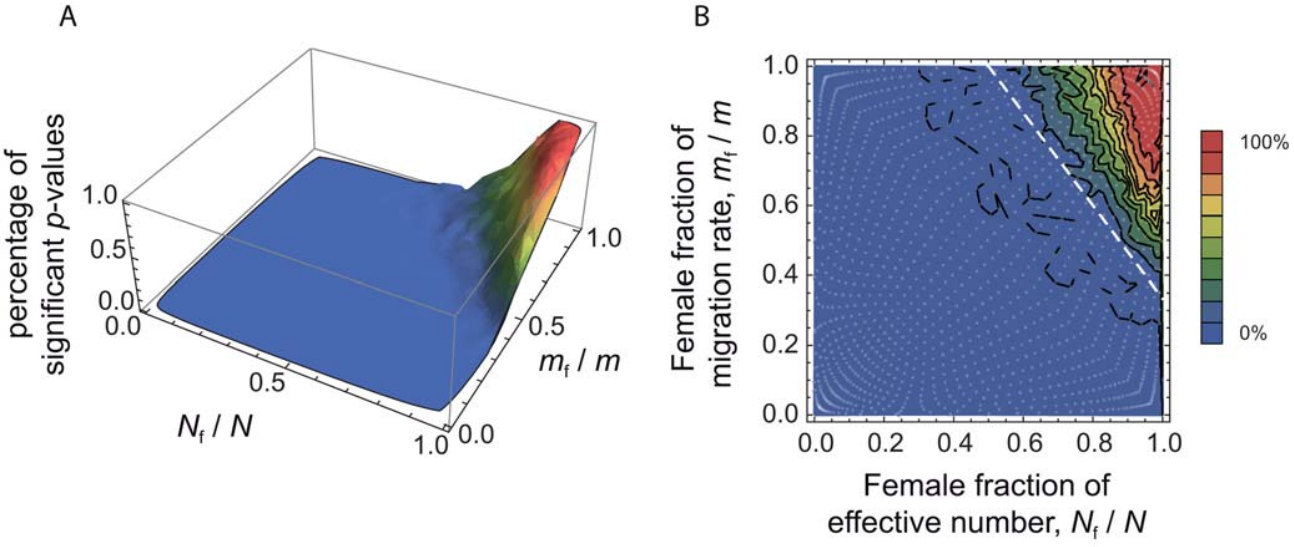


Figure 4. Percentage of significant tests in the (N_f/N, m_f/m) parameter space, for simulated data. We chose a range of 49 ($N_f m_f/N_m m_m$) ratios, varying from 0.0004 to 2401, and for each of these ratios we chose 29 sets of (N_f/N, m_f/m) values. By doing this, we obtained 1421 sets of (N_f/N, m_f/m) values, represented as white dots in the right-hand side panel B, covering the whole parameter space. For each set, we simulated 100 independent datasets using a coalescent-based algorithm, and taking the same number of individuals and the same number of loci for each genetic system as in the observed data. For each dataset, we calculated the *p*-value for a one-sided Wilcoxon sum rank test $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$, and for each set of ($N_f/N$, m_f/m) values we calculated the percentage of significant *p*-values (at the $\alpha = 0.05$ level). A. Surface plot of the proportion of significant *p*-values (at the $\alpha = 0.05$ level), as a function of the female fraction of effective number and the female fraction of migration rate. B. Contour plot, for the same data. The dotted line, at which $\frac{m_f}{m} = \left(5 - 4\frac{N_f}{N}\right)/3$, represents the set of (N_f/N, m_f/m) values where the autosomal and X-linked F_{ST}'s are equal. The theory predicts that we should only find $F_{ST}^{(A)} > F_{ST}^{(X)}$ in the upper-right triangle defined by the dotted line. Hence, the proportion of significant *p*-values for any set of (N_f/N, m_f/m) values in this upper right triangle gives an indication of the power of the method.
doi:10.1371/journal.pgen.1000200.g004

independent datasets. Figure 4 shows the distribution of the percentage of significant tests in the (N_f/N, m_f/m) parameter space. Theory predicts that in the upper-right triangle where $\frac{m_f}{m} > \left(5 - 4\frac{N_f}{N}\right)/3$, we should have $F_{ST}^{(A)} > F_{ST}^{(X)}$. One can see from Figure 4 that, given the simulation parameters used, the method is conservative: the proportion of significant tests at the $\alpha = 0.05$ level is null outside of the upper-right triangle. However, we find a fairly large proportion of significant tests for large N_f/N and m_f/m ratios which indicates (*i*) that the method presented here has the potential to detect differences in male and female effective numbers, but (*ii*) that only strong differences might be detected, for similarly sized datasets as the one considered here.

Robustness to the Sampling Scheme

We also aimed to investigate whether the results obtained here were robust to our sampling scheme, and that our results were not biased by the inclusion of particular populations. To do this, we re-analyzed both the bilineal agriculturalists and the patrilineal herders datasets, removing one population at a time in each group. For each of these jackknifed datasets, we calculated the *p*-value of a one-sided Wilcoxon sum rank test $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$, as done on the full datasets. The results are given in Table 5. We found no significant test for any of the bilineal agriculturalist groupings ($p > 0.109$), which supports the idea that, in those populations, both the migration rate and the number of reproductive individuals can be equal for both sexes. In patrilineal herders, the tests were significant at the $\alpha = 0.05$ level for 8 out of 11 population groupings. For the 3 other groupings, the *p*-values were 0.068, 0.078 and 0.073 (see Table 5). Overall, the ratio of $F_{ST}^{(A)}$ over $F_{ST}^{(X)}$ multi-locus estimates ranged from 1.7 to 3.5 in patrilineal herders (and from 0.9 to 1.2 in bilineal agriculturalists). Although in some particular groupings of patrilineal herder populations, the difference in the distributions of $F_{ST}^{(A)}$ and $F_{ST}^{(X)}$ may not be strong enough to be significant, we can clearly distinguish the pattern of differentiation for autosomal and X-linked markers in patrilineal and bilineal groups. Results from coalescent simulations (see above) suggest that this lack of statistical power might be expected for $F_{ST}^{(A)} / F_{ST}^{(X)}$ ratios close to unity. Indeed, we found that the tests were more likely to be significant for fairly large N_f/N and m_f/m ratios (the upper-right red region in Figure 4) which would correspond to $F_{ST}^{(A)} / F_{ST}^{(X)}$ ratios much greater than one.

Comparison with Uniparentally-Inherited Markers

Importantly, our results on X-linked and autosomal markers are consistent with those obtained from NRY and mtDNA (see Figures 3B–3D): in these figures, the dashed line gives all the sets of (N_f/N, m_f/m) values that are compatible with the observed $F_{ST}^{(Y)}$ and $F_{ST}^{(mtDNA)}$ estimates. These are the sets of values that satisfy $\left(\frac{N_f/N}{1-N_f/N}\right) = 2.1\left(\frac{1-m_f/m}{m_f/m}\right)$ for the bilineal populations, and $\left(\frac{N_f/N}{1-N_f/N}\right) = 21.6\left(\frac{1-m_f/m}{m_f/m}\right)$ for the patrilineal populations, since we inferred $N_f m_f/N_m m_m \approx 2.1$ and $N_f m_f/N_m m_m \approx 21.6$, respectively, for the two groups. For the bilineal agriculturalists (Figure 3D), the set of (N_f/N, m_f/m) values inferred from the $F_{ST}^{(Y)}$ and $F_{ST}^{(mtDNA)}$ estimates fall within the range that was not rejected, given our data on X-linked and autosomal markers. For the patrilineal herders (Figure 3B), the overlap is only partial: from the NRY and mtDNA data only, low N_f/N ratios associated with high m_f/m ratios are as likely as high N_f/N ratios associated with low m_f/m ratios. Yet, it is clear from this figure that a large set of (N_f/N, m_f/m) values inferred

Table 5. Autosomal and X-linked differentiation on jackknifed samples.

Sample removed	$F_{ST}^{(A)}$	$F_{ST}^{(X)}$	*p*-value	$F_{ST}^{(A)}/F_{ST}^{(X)}$
Patrilineal groups				
KAZ	0.0084	0.0050	0.068	1.7
KKK	0.0085	0.0050	0.078	1.7
KRA	0.0078	0.0027	0.022	2.9
KRB	0.0080	0.0030	0.028	2.7
KRG	0.0078	0.0035	0.037	2.2
KRL	0.0086	0.0038	0.018	2.3
KRM	0.0069	0.0023	0.018	3.0
KRT	0.0081	0.0044	0.047	1.8
LKZ	0.0088	0.0025	0.002	3.5
OTU	0.0089	0.0038	0.022	2.3
TUR	0.0054	0.0025	0.073	2.2
Bilineal groups				
TDS	0.0125	0.0109	0.443	1.1
TDU	0.0132	0.0153	0.705	0.9
TJA	0.0144	0.0123	0.109	1.2
TJE	0.0140	0.0133	0.148	1.1
TJK	0.0134	0.0131	0.457	1.0
TJN	0.0148	0.0144	0.387	1.0
TJR	0.0140	0.0141	0.401	1.0
TJT	0.0139	0.0121	0.225	1.1
TJU	0.0139	0.0127	0.283	1.1
TJY	0.0139	0.0116	0.259	1.2

For each group, we removed one sample in turn and calculated the differentiation on autosomal and X-linked markers. The *p*-value gives the result of a one-sided Wilcoxon sum rank test $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$, as performed on the full dataset.

doi:10.1371/journal.pgen.1000200.t005

from the single-locus estimates $F_{ST}^{(Y)}$ and $F_{ST}^{(mtDNA)}$ can be rejected, given the observed differentiation on X-linked and autosomal markers. All genetic systems (mtDNA, NRY, X-linked and autosomal markers) converge toward the notion that patrilineal herders, in contrast to bilineal agriculturalists, have a strong sex-specific genetic structure. Yet, the information brought by X-linked and autosomal markers is substantial, since we show that this is likely due to both higher migration rates and larger effective numbers for women than for men.

Comparison with Other Studies

Our results, based on the X chromosome and the autosomes, also confirm previous analyses based on the mtDNA and the NRY, showing that men are genetically more structured than women in other patrilocal populations [3–10,14–17] (see also Table 1). A handful of studies have also shown a reduced effective number of men compared to that of women, based on coalescent methods [23,24], but none have considered the influence of social organization on this dissimilarity (see Table 1).

In some respects, our results contrast with those of Wilder and Hammer [25], who studied sex-specific population genetic structure among the Baining of New Britain, using mtDNA, NRY, and X-linked markers. Interestingly, they found that $N_f > N_m$, but $m_f < m_m$, and claimed that a similar result, although left unexplored by the authors, was to be found in a recent study by Hamilton et al. [16]. This raises the interesting point that sex-specific proportions of migrants (m) are likely to be shaped by factors that may only partially overlap with those that affect the sex-specific effective numbers (N). Further studies of human populations with contrasted social organizations, as well as further theoretical developments, are needed to appreciate this point.

In order to ask to what extent our results generalize to other human populations, we investigated sex-specific patterns in the 51 worldwide populations represented in the HGDP-CEPH Human Genome Diversity Cell Line Panel dataset [43], for which the data on the differentiation of 784 autosomal microsatellites and 36 X-linked microsatellites are available (data not shown). By doing this, we found a larger differentiation for X-linked than for autosomal markers $\left(F_{ST}^{(X)} > F_{ST}^{(A)}\right)$. Therefore, we confirmed Ramachandran et al.'s [20] result that no major differences in demographic parameters between males and females are required to explain the X-chromosomal and autosomal results in this worldwide sample. Ramachandran et al.'s approach [20] is based upon a pure divergence model from a single ancestral population, which is very different from the migration-drift equilibrium model considered here. In real populations, however, genetic differentiation almost certainly arises both through divergence and limited dispersal, which places these two models at two ends of a continuum. Yet, importantly, if we apply Ramachandran et al.'s [20] model to the Central Asian data, our conclusions are left unchanged. In their model, the differentiation among populations is $F_{ST} \approx 1 - e^{-t/(2N_e)}$, where t is the time since divergence from an ancestral population and N_e the effective size of the populations (see, e.g., [44]). Hence, we get $F_{ST}^{(A)} \approx 1 - e^{-t/\left(2N_e^{(A)}\right)}$ and $F_{ST}^{(X)} \approx 1 - e^{-t/\left(2N_e^{(X)}\right)}$ for autosomal and X-linked markers, respectively. Therefore, our observation that $F_{ST}^{(A)} > F_{ST}^{(X)}$ implies that $N_e^{(X)} > N_e^{(A)}$, which requires that $N_f > 7N_m$ since $N_e^{(A)} = 8N_f N_m/(N_f + N_m)$ and $N_e^{(X)} = 9N_f N_m/(N_f + 2N_m)$ (see, e.g., [45]). In this case, the female fraction of effective number is larger than that of males, which is consistent with our findings in a model with migration.

The HGDP-CEPH dataset does not provide any detailed ethnic information for the sampled groups, and we can therefore not distinguish populations with different lifestyles. However, at a more local scale in Pakistan, we were able to analyze a subset of 5 populations (Brahui, Balochi, Makrani, Sindhi and Pathan), which are presumed to be patrilineal [46]. For this subset, we found a higher differentiation for autosomal $\left(F_{ST}^{(A)} = 0.003\right)$ than for X-linked markers $\left(F_{ST}^{(X)} = 0.002\right)$, although non-significantly ($p = 0.12$). This result seems to suggest that other patrilineal populations may behave like the Central Asian sample presented here. Therefore, because the geographical clustering of populations with potentially different lifestyles may minimize the differences in sex-specific demography at a global scale [21,22], and/or because the global structure may reflect ancient (pre-agricultural) marital residence patterns with less pronounced patrilocality [12], we emphasize the point that large-scale studies may not be relevant to detect sex-specific patterns, which supports a claim made by many authors.

Conclusion

In conclusion, we have shown here that the joint analysis of autosomal and X-linked polymorphic markers provides an efficient tool to infer sex-specific demography and history in human populations, as suggested recently [12,47]. This new multilocus approach is, to our knowledge, the first attempt to combine the information contained in mtDNA, NRY, X-linked and autosomal markers (see Table 1), which allowed us to test for

the robustness of a sex-specific genetic structure at a local scale. Unraveling the respective influence of migration and drift upon neutral genetic structure is a long-standing quest in population genetics [48,49]. Here, our analysis allowed us to show that differences in sex-specific migration rates may not be the only cause of contrasted male and female differentiation in humans and that, contrary to the conclusion of a number of studies (see Table 1), differences in effective numbers may also play an important role. Indeed, we have demonstrated that sex-specific differences in population structure in patrilineal herders may be the consequence of both higher female effective numbers and female effective dispersal. Our results also illustrate the importance of analyzing human populations at a local scale, rather than global or even continental scale [2,19,21]. The originality of our approach lies in the comparison of identified ethnic groups that differ in well-known social structures and lifestyles. In that respect, our study is among the very few which compare patrilineal vs. bilineal or matrilineal groups (see Table 1), and we believe that it contributes to the growing body of evidence showing that social organization and lifestyle have a strong impact on the distribution of genetic variation in human populations. Moreover, our approach could also be applied on a wide range of animal species with contrasted social organizations. Therefore, we expect our results to stimulate research on the comparison of X-linked and autosomal data to disentangle sex-specific demography.

Methods

DNA Samples

We sampled 10 populations of bilineal agriculturalists and 11 populations of patrilineal herders from West Uzbekistan to East Kyrgyzstan, representing 780 healthy adult men from 5 ethnic groups (Tajiks, Kyrgyz, Karakalpaks, Kazaks, and Turkmen) (see Table 2). The geographic distribution of the samples and information about lifestyle is provided in Figure 1. Also living in Central Asia, Uzbeks are traditionally patrilineal herders too, but they have recently lost their traditional social organization [11], and we therefore chose not to include any sample from this ethnic group for the purpose of this study. We collected ethnologic data prior to sampling, including the recent genealogy of the participants. Using this information, we retained only those individuals that were unrelated for at least two generations back in time. All individuals gave their informed consent for participation in this study. Total genomic DNA was isolated from blood samples by a standard phenol-chloroform extraction [50].

Uniparentally Inherited Markers

The mtDNA first hypervariable segment of the mtDNA control region (HVS-I) was amplified using primers L15987 (5′TCAAATGGGCCTGTCCTTGTA) and H580 (5′TTGAGGAGGTAAGCTACATA) in 18 populations out of 21 (674 individuals, see Table 2). The amplification products were subsequently purified with the EXOSAP standard procedure. The sequence reaction was performed using primers L15925 (5′TAATACACCAGTCTTGTAAAC) and HH23 (5′AATAGGGTGATAGACCTGTG). Sequences from positions 16 024–16 391 were obtained. Eleven Y-linked microsatellite markers (see Table 3) were genotyped in the same individuals, following the protocol described by Parkin et al. [51].

Multi-Locus Markers

27 autosomal and 9 X-linked microsatellite markers (see Table 4) were genotyped in the same individuals. We used the informativeness for assignment index I_n [52] to select subsets of microsatellite markers on the X chromosome and the autosomes from the set of markers used in Rosenberg et al.'s worldwide study [43]. This statistic measures the amount of information that multiallelic markers provide about individual ancestry [52]. This index was calculated among a subset of 14 populations, chosen from the Rosenberg et al.'s dataset [43] to be genetically the closest to the Central Asian populations (Balochi, Brahui, Burusho, Hazara, Pathan, Shindi, Uygur, Han, Mongola, Yakut, Adygei, Russian, Druze and Palestinian). The rationale was to infer the information provided by individual loci about ancestry from this subset of populations, and to extrapolate the results to the populations studied here. For the X chromosome data, we pooled the 'Screening Set10' and 'Screening Set52' from the HGDP-CEPH Human Genome Diversity Cell Line Panel [53] analyzed by Rosenberg et al. [43] which represented a total of 36 microsatellites. We chose 9 markers among the 11 with the highest I_n. For autosomal data, we used the 'Screening Set10', which represented a total of 377 microsatellites, and chose 27 markers among the 30 with the highest I_n. All markers were chosen at a minimum of 2 cM apart from each others [54]. PCR amplifications were performed in a 20 µl final volume composed of 1× Eppendorf buffer, 125 µM each dNTP, 0.5U Eppendorf Taq polymerase, 125 nM of each primer, and 10 ng DNA. The reactions were performed in a Eppendorf Mastercycler with an initial denaturation step at 94°C for 5 min; followed by 36 cycles at 94°C for 30 s, 55°C for 30 s, 72°C for 20 s, and 72°C for 10 min as final extension. Forward primers were fluorescently labeled and reactions were further analyzed by capillary electrophoresis (ABI 310, Applied Biosystems). We used the software package Genemarker (SoftGenetics LLC) to obtain allele sizes from the analysis of PCR products (allele calling).

Statistical Analyses

We calculated the total allelic richness (*AR*) (over all populations), the unbiased estimate of expected heterozygosity H_e [55], the total number of polymorphic sites and F_{ST} for mtDNA using Arlequin version 3.1. [56]. Genetic differentiation among populations for the autosomes, the X and the Y chromosome was measured both per locus and overall loci using Weir and Cockerham's F_{ST} estimator [57], as calculated in GENEPOP 4.0. [58]. The 95% confidence intervals were obtained by bootstrapping over loci [58], using the approximate bootstrap confidence intervals (ABC) method described by DiCiccio and Efron [59]. Isolation by distance (i.e. the correlation between the genetic and the geographic distances) was analyzed by computing the regression of pairwise $F_{ST}/(1-F_{ST})$ estimates between pairs of populations to the natural logarithm of their geographical distances, and rank correlations were tested using the Mantel permutation procedure [60], as implemented in GENEPOP 4.0. [58]. All other statistical tests were performed using the software package R v. 2.2.1 [61].

Sex-Biased Dispersal in the Island Model

Let us consider an infinite island model of population structure [62], with two classes of individuals (males and females), which describes a infinite set of populations with constant and equal sizes that are connected by gene flow. Then the expected values of F_{ST} for uniparentally inherited markers depend on the effective number N_m (resp. N_f) of adult males (resp. females) per population and the migration rate m_m (resp. m_f) of males (resp. females) per generation, as: $F_{ST}^{(mtDNA)} \approx 1/(1+2N_f m_f)$ and $F_{ST}^{(Y)} \approx 1/(1+2N_m m_m)$ (see, e.g., [63]). We can therefore calculate the female-to-male ratio of the effective number of migrants per generation as: $N_f m_f / N_m m_m = \left(1-1/F_{ST}^{(mtDNA)}\right)/\left(1-1/F_{ST}^{(Y)}\right)$.

In this model, we can also compute for the autosomes and the X chromosome the reproductive values for each class (sex), which are interpreted here as the probability that an ancestral gene lineage was in a given class in a distant past [64]. From these, we can obtain the well-known expressions of effective size N_e for autosomal and X-linked genes: $N_e^{(A)} = 8N_fN_m/(N_f+N_m)$ and $N_e^{(X)} = 9N_fN_m/(N_f+2N_m)$, respectively [45]. Note that N_e is expressed here as a number of gene copies (i.e., twice the effective number of diploid individuals for autosomes). Likewise, the effective migration rate, i.e. the average dispersal rate of an ancestral gene lineage, is given by $m_e^{(A)} = (m_f+m_m)/2$ for autosomal genes, and $m_e^{(X)} = (2m_f+m_m)/3$ for X-linked genes, respectively. Substituting these expressions into the well-known equation: $F_{ST} \approx 1/(1+2N_em_e)$ [64], we get:

$$F_{ST}^{(A)} \approx \frac{1}{1+4\frac{4N_fN_m}{N_f+N_m}\frac{m_f+m_m}{2}}, \quad (5)$$

for autosomal genes, and

$$F_{ST}^{(X)} \approx \frac{1}{1+4\frac{9N_fN_m}{2N_f+4N_m}\frac{2m_f+m_m}{3}}, \quad (6)$$

for X-linked genes.

Evaluation of the Approach through Stochastic Simulations

We performed coalescent simulations, using an algorithm in which coalescence and migration events are considered generation-by-generation until the common ancestor of the whole sample has been reached (see [65]). We simulated a finite island model with 50 demes, each made of $N = N_f+N_m = 500$ diploid individuals, with a migration parameter $m = m_f+m_m = 0.2$. Using these total values for N and m, we then varied the sex-specific parameters to cover the $(N_f/N,\ m_f/m)$ parameter space evenly. Note that the parameter m is the total migration rate, which corresponds to twice the effective migration rate for autosomal markers. Hence, for each set of $(N_f/N,\ m_f/m)$ values, the total number of individuals is 500 (although the number of females may vary from 1 to 499) and the effective migration rate for autosomal markers is $m_e^{(A)} = (m_f+m_m)/2 = 0.1$. We chose these total values for N and m such that, for a ratio $N_fm_f/N_mm_m = 21.6$ (as observed for the herder populations), the distribution of F_{ST} estimates on uniparentally-inherited markers in the simulations were close to the observations: for mtDNA, the 95% highest posterior density interval (see [66], pp. 38–39) for the distribution of F_{ST} estimates in the simulations was [0.007; 0.033] with a mode at 0.014 (estimated value from the real dataset: $F_{ST}^{(mtDNA)} = 0.010$ among the herders) while for the NRY, the 95% highest posterior density interval was [0.088; 0.374] with a mode at 0.187 (estimated value from the real dataset: $F_{ST}^{(Y)} = 0.177$).

Each simulated sample consisted in 330 sampled males from 11 populations (30 males per population), genotyped at 27 autosomal, 9 X-linked markers as well as 10 Y-linked markers and a single mtDNA locus. Each locus was assumed to follow a Generalized Stepwise Model (GSM) [67] with a possible range of 40 contiguous allelic states, except the mtDNA, which was assumed to follow an infinite allele model of mutation. The average mutation rate was 5.10^{-3}, and the mean parameter of the geometric distribution of the mutation step lengths for microsatellites was set to 0.2 [67,68].

Acknowledgments

We thank all the people who volunteered to participate in this study, or who helped us in the field. We are grateful to Sylvain Théry for valuable help in handling geographic data, to Hélène Fréville and Nicolas Perrin for helpful comments on previous versions of this manuscript, as well as to three anonymous reviewers for insightful and constructive comments. We acknowledge the "Service de Systématique Moléculaire" (SSM) at the Museum National d'Histoire Naturelle (MNHN) and the Biological Resource Center of the Foundation Jean Dausset-CEPH for genotyping facilities. Part of this work was carried out by using the resources of the Computational Biology Service Unit from the Museum National d'Histoire Naturelle (MNHN) which was partially funded by Saint Gobain.

Author Contributions

Conceived and designed the experiments: EH RV. Performed the experiments: LS BMC LQM PB MG. Analyzed the data: LS RV. Contributed reagents/materials/analysis tools: BMC TH AA FN MJ EH. Wrote the paper: LS RV. Collected the samples: LS, BMC, EH.

References

1. Disotell TR (1999) Human evolution: sex-specific contributions to genome variation. Curr Biol 9: R29–31.
2. Wilkins JF (2006) Unraveling male and female histories from human genetic data. Curr Opin Genet Dev 16: 611–617.
3. Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20: 278–280.
4. Salem AH, Badr FM, Gaballah MF, Pääbo S (1996) The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. Am J Hum Genet 59: 741–743.
5. Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, et al. (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65: 208–219.
6. Oota H, Kitano T, Jin F, Yuasa I, Wang L, et al. (2002) Extreme mtDNA homogeneity in continental Asian populations. Am J Phys Anthropol 118: 146–153.
7. Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill P, et al. (2003) Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet 72: 281–302.
8. Malyarchuk B, Derenko M, Grzybowski T, Lunkina A, Czarny J, et al. (2004) Differentiation of mitochondrial DNA and Y chromosomes in Russian populations. Hum Biol 76: 877–900.
9. Nasidze I, Ling EY, Quinque D, Dupanloup I, Cordaux R, et al. (2004) Mitochondrial DNA and Y-chromosome variation in the Caucasus. Ann Hum Genet 68: 205–221.
10. Nasidze I, Quinque D, Ozturk M, Bendukidze N, Stoneking M (2005) MtDNA and Y-chromosome variation in Kurdish groups. Ann Hum Genet 69: 401–412.
11. Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, et al. (2007) From social to genetic structures in central Asia. Curr Biol 17: 43–48.
12. Wilkins JF, Marlowe FW (2006) Sex-biased migration in humans: what should we expect from genetic data? Bioessays 28: 290–300.
13. Burton ML, Moore CC, Whiting JWM, Romney AK (1996) Regions based on social structure. Curr Anthro 37: 87–123.
14. Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29: 20–21.
15. Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, et al. (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol 21: 1673–1682.
16. Hamilton G, Stoneking M, Excoffier L (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proc Natl Acad Sci U S A 102: 7476–7480.
17. Bolnick DA, Bolnick DI, Smith DG (2006) Asymmetric male and female genetic histories among Native Americans from Eastern North America. Mol Biol Evol 23: 2161–2174.
18. Stoneking M (1998) Women on the move. Nat Genet 20: 219–220.
19. Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36: 1122–1125.
20. Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW (2004) Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Hum Genomics 1: 87–97.

21. Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, et al. (2006) Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. PLoS Genet 2: e53.
22. Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, et al. (2001) Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18: 1189–1203.
23. Dupanloup I, Pereira L, Bertorelle G, Calafell F, Prata MJ, et al. (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J Mol Evol 57: 85–97.
24. Wilder JA, Mobasher Z, Hammer MF (2004) Genetic evidence for unequal effective population sizes of human females and males. Mol Biol Evol 21: 2047–2057.
25. Wilder JA, Hammer MF (2007) Extraordinary population structure among the Baining of New Britain. In: Friedlaender JS, ed. Genes, Language, and Culture History in the Southwest Pacific. Oxford, UK: Oxford University Press. pp 199–207.
26. Seielstad M (2000) Asymmetries in the maternal and paternal genetic histories of Colombian populations. Am J Hum Genet 67: 1062–1066.
27. Langergraber KE, Siedel H, Mitani JC, Wrangham RW, Reynolds V, et al. (2007) The genetic signature of sex-biased migration in patrilocal chimpanzees and humans. PLoS ONE 2: e973.
28. Bazin E, Glemin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. Science 312: 570–572.
29. Heyer E, Sibert A, Austerlitz F (2005) Cultural transmission of fitness: genes take the fast lane. Trends Genet 21: 234–239.
30. Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, et al. (2003) The genetic legacy of the Mongols. Am J Hum Genet 72: 717–721.
31. Neel JV (1970) Lessons from a "primitive" people. Science 170: 815–822.
32. Blum MG, Heyer E, Francois O, Austerlitz F (2006) Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. PLoS Genet 2: e122.
33. Helgason A, Hrafnkelsson B, Gulcher JR, Ward R, Stefansson K (2003) A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. Am J Hum Genet 72: 1370–1388.
34. White DR (1988) Rethinking polygyny: co-wives, codes, and cultural systems. Curr Anthro 29: 529–558.
35. Heider KG (1997) Grand valley Dani: peaceful warriors. In: GS, LS, eds. Case studies in cultural anthropology. Forth Worth, Texas: Harcourt Brace College Publishers.
36. Tremblay M, Vezina H (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am J Hum Genet 66: 651–658.
37. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128: 415–423.
38. Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW (2002) Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics 161: 447–459.
39. Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. Heredity 82: 117–125.
40. Beaumont M, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond 263: 1619–1626.
41. Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145: 1219–1228.
42. Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. Evolution 61: 3001–3006.
43. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.
44. Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. Genetics 105: 767–779.
45. Wright S (1939) Statistical genetics in relation to evolution. Actualités scientifiques et industrielles 802 Exposés de Biométrie et de Statistique Biologique XIII. Paris: Hermann et Cie.
46. Tamisier JC (1998) Dictionnaire des peuples. Sociétés d'Afrique, d'Amérique, d'Asie et d'Océanie. Paris: Larousse-Bordas.
47. Balaresque P, Jobling MA (2007) Human populations: houses for spouses. Curr Biol 17: R14–16.
48. Lawson-Handley LJ, Perrin N (2007) Advance in our understanding of mammalian sex-biased dispersal. Molecular Ecology 16: 1559–1578.
49. Hurles ME, Jobling MA (2001) Haploid chromosomes in molecular ecology: lessons from the human Y. Mol Ecol 10: 1599–1613.
50. Maniatis T, Fritsh EF, S J (1982) Molecular cloning. A laboratory manual. New York: Cold Spring Laboratory.
51. Parkin EJ, Kraayenbrink T, GLvD, Tshering K, de Knijff P, et al. (2006) 26-Locus Y-STR typing in a Bhutanese population sample. Forensic Science International 161: 1–7.
52. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73: 1402–1422.
53. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.
54. Wilson JF, Goldstein DB (2000) Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. Am J Hum Genet 67: 926–935.
55. Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89: 583–590.
56. Excoffier L, Laval LG, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinfo Online 1: 47–50.
57. Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358–1370.
58. Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Res 8: 103–106.
59. DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Statistical Science 11: 189–228.
60. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27: 209–220.
61. R Development Core Team (2007) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.
62. Wright S (1931) Evolution in mendelian populations. Genetics 16: 97–159.
63. Hedrick PW (2007) Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution 61: 2750–2771.
64. Rousset F (2004) Genetic Structure and Selection in Subdivided Populations. Princeton, New Jersey: Princeton University Press.
65. 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.
66. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis. Second Edition. New York: Chapman & Hall/CRC.
67. Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol 11: 1591–1604.
68. Dib C, Faure S, Fizames C, Samson D, Drouot N, et al. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380: 152–154.
69. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–491.
70. 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.

Annexe D

Article 2 : dispersion limitée d'un peuple mobile

Verdu P., Leblois R., Froment A., Théry S., Bahuchet S., Rousset F., Heyer E. et Vitalis R. (2010) Limited dispersal in mobile hunter–gatherer Baka Pygmies. *Biology Letters* **6** : 858-861

---

biology letters
Population genetics

Biol. Lett. (2010) **6**, 858–861
doi:10.1098/rsbl.2010.0192
Published online 28 April 2010

Limited dispersal in mobile hunter–gatherer Baka Pygmies

Paul Verdu[1,2,*], Raphaël Leblois[3], Alain Froment[4], Sylvain Théry[2], Serge Bahuchet[2], François Rousset[5], Evelyne Heyer[2] and Renaud Vitalis[2,†]

[1]*Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA*
[2]*MNHN–CNRS, Université Paris 7, UMR 7206 'Ecoanthropology and Ethnobiology', Paris, France*
[3]*MNHN–CNRS, UMR 7205 'Origine, Structure et Evolution de la Biodiversité', Paris, France*
[4]*IRD–MNHN, UMR 208 'Patrimoines locaux', Paris, France*
[5]*Université Montpellier 2–CNRS, UMR 5554 'Institut des Sciences de l'Evolution', Montpellier, France*
**Author for correspondence (verdu@umich.edu).*
[†]*Present address: CNRS–INRA, UMR CBGP (INRA–IRD–CIRAD–Montpellier SupAgro), Campus International de Baillarguet, Montferrier-sur-Lez, France.*

Hunter–gatherer Pygmies from Central Africa are described as being extremely mobile. Using neutral genetic markers and population genetics theory, we explored the dispersal behaviour of the Baka Pygmies from Cameroon, one of the largest Pygmy populations in Central Africa. We found a strong correlation between genetic and geographical distances: a pattern of isolation by distance arising from limited parent–offspring dispersal. Our study suggests that mobile hunter–gatherers do not necessarily disperse over wide geographical areas.

Keywords: African Pygmies; dispersal; isolation by distance; microsatellites; population genetics

1. INTRODUCTION

The correlation between genetics and geography has been extensively studied at a worldwide scale to better characterize the distribution of human genetic diversity and estimate key parameters of modern humans' expansion worldwide (e.g. Ramachandran *et al.* 2005; Liu *et al.* 2006; DeGiorgio *et al.* 2009). However, none of these studies have characterized parent–offspring dispersal and therefore, the dispersal mechanisms underlying the geographical distribution of human genetic diversity remain unclear.

Ethnologists have shown the great variability of mobility across human populations with different lifestyles and modes of subsistence. For example, foragers have been shown to be much more mobile than farmers (MacDonald & Hewlett 1999). However, the relationship between mobility and effective dispersal (characterized by the distances between children's and parents' birthplaces) often remains unknown. To explore this aspect, we studied African Pygmies, which represent the largest hunter–gatherer group of populations worldwide. African Pygmies are described as being very mobile within the rainforest for seasonal mobility and socioeconomic activities (Bahuchet 1992). Demographic data in an Aka Pygmy population further showed large mating ranges and large distances between birthplaces and places of residence (Cavalli-Sforza & Hewlett 1982). This suggests that such mobile behaviour should translate into long distances between children's and parents' birthplaces, but this assumption has never been explicitly tested owing to limited demographic data (Cavalli-Sforza 1986). In this context, population genetics can provide indirect estimates of effective dispersal in these hunter–gatherers. Indeed, isolation by distance theory predicts that at equilibrium between drift, migration and mutation, the regression of F-statistics estimates on the logarithm of geographical distances provides a robust estimate of the neighbourhood size, which is proportional to the population density and the second moment of parent–offspring distance, if individuals are sampled at a short geographical scale (Rousset 1997).

Here, we report estimates of effective dispersal based on the relationship between genetic and geographical distances in three groups of the Baka Pygmies from Cameroon. This is, to the best of our knowledge, the only genetic data for such a small geographical scale in African Pygmy populations, which provides an opportunity to infer effective dispersal using the genetics of these mobile hunter–gatherers.

2. MATERIAL AND METHODS

The Baka from Cameroon represent one of the largest Pygmy populations with 35 000 individuals occupying 75 000 km^2 in the rainforest (figure 1). These values, which were compiled from ethnographic data (Vallois & Marquer 1976; Dhellemmes & Macaigne 1985; Cavalli-Sforza 1986; Sato 1992; Abéga 1998; Tsuru 2001), provide an estimate of the population density of $D = 0.47$ individuals $\cdot$ km^{-2}.

We consider 87 Baka adults genotyped at 28 independent tetranucleotide microsatellite loci (Verdu *et al.* 2009). We visited Baka settlements along three transects of approximately 50 km each (figure 1), asking volunteers to gather for DNA sampling at a single location for each transect. Frequent movements between places of temporary residence, as well as hospitality rules among Pygmies that assimilate visitors as residents during their visit, make it difficult to determine whether an individual met in a village is a resident or a visitor (Cavalli-Sforza 1986; Bahuchet 1992). Therefore, the sampling location, or the location where individuals were first met, are probably a poor predictor of Pygmies' location after dispersal and these data were discarded. Instead, we considered that birthplaces provided more robust data to study Pygmies' dispersal. After collecting each individual's birthplace, we went back on the road to determine the geographical coordinates of these locations. Each donor provided appropriate informed oral consent.

Geographically limited dispersal across generations in two-dimensional habitats results in a linear relationship between pairwise genetic distances a_r, an analogue to $F_{ST}/(1 - F_{ST})$ calculated between pairs of individuals (Rousset 2000), and the logarithm of geographical distance. We used the software package GenePop' 007 (Rousset 2008) to calculate the slope of this linear regression, which, at a small geographic scale, provides a robust estimator of $1/(4D\pi\sigma^2)$, where D is the effective population density and σ^2 is the second moment of parent–offspring axial distance (Rousset 1997). We tested the significance of the correlation using a Mantel test (Mantel 1967) with 30 000 permutations. First, we considered all pairs of individuals from the three geographically distant Baka groups (figure 1). Second, we considered only pairs of individuals born within the same group and discarded pairs of individuals born in different groups. To that end, we wrote a R script (R Development Core Team 2007), available upon request, which modified the Mantel test to calculate rank correlation coefficients and to permute the pairwise distances within groups only.

3. RESULTS AND DISCUSSION

Considering all possible pairs of Baka individuals, we found a significant positive correlation between pairwise genetic distance and the logarithm of geographical distance ($p = 0.010$): individuals born

Received 24 February 2010
Accepted 8 April 2010

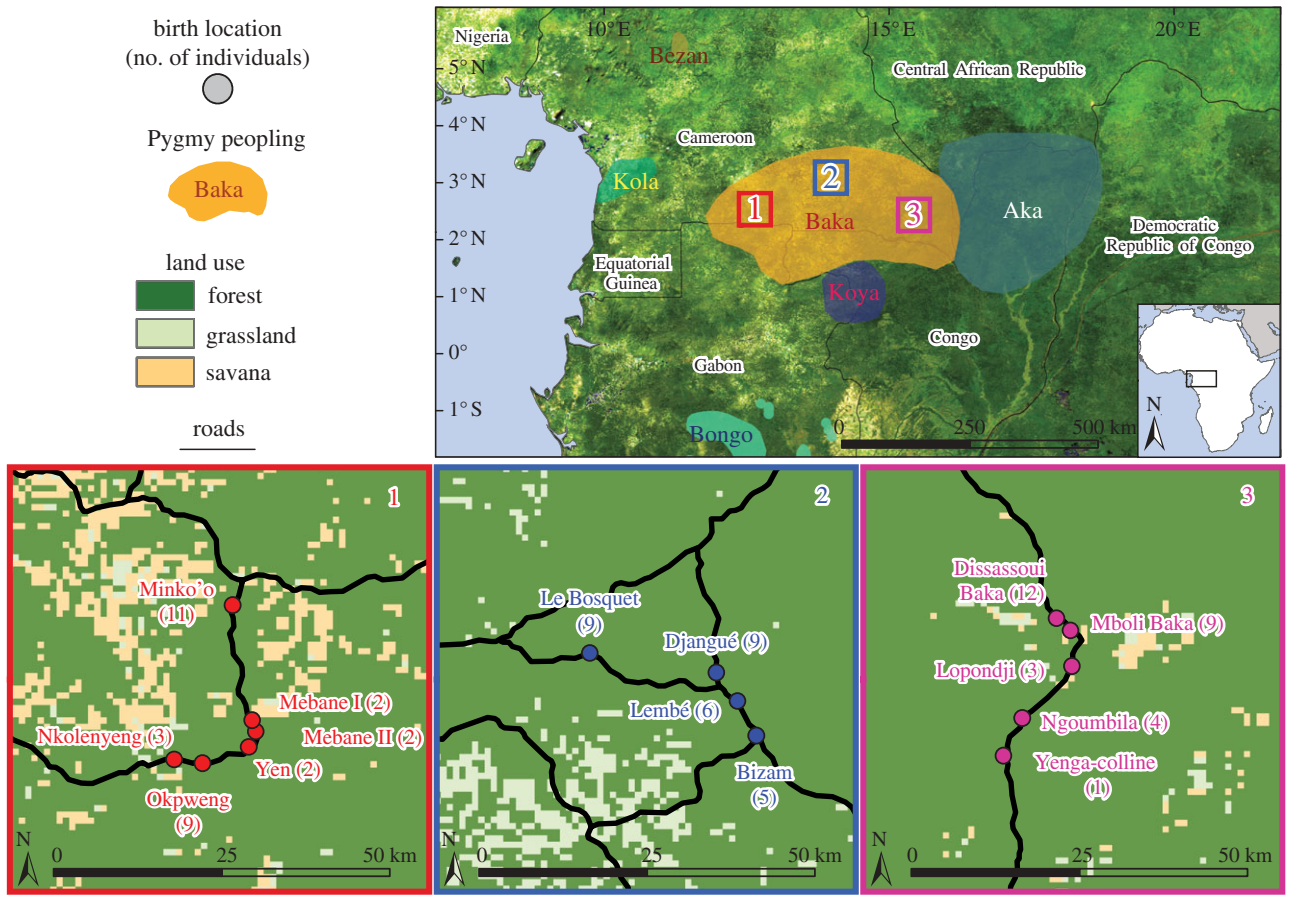


Figure 1. Geographical distribution of the 87 Baka Pygmies sampled in southeastern Cameroon. In each Baka group, birthplaces of sampled individuals are shown with sample sizes given in parentheses. Other Pygmy peopling areas inferred from our ethnographical fieldworks are shown. Map sources: Global Land Cover Facility.

nearby are more genetically related than individuals born further away from each other. The slope of regression between the genetic and geographical distances equalled 0.0027 (95% CI 0.0007–0.0046; see figure 2). We estimated from this slope that $4\pi D\sigma^2 = 373$ individuals and, using $D = 0.47$ individuals $\cdot$ km^{-2}, we obtained $\sigma^2 = 63.2$ km^2. But theory shows that mutation wipes out the linear increase in genetic differentiation with geographical distance, if the distances between samples are larger than $d_{\max} = 0.56\,\sigma/\sqrt{2\mu}$, where μ is the mutation rate (Rousset 1997, 2004). Here, the maximum distance between any two individuals' birthplace was 296 km, a value much larger than $d_{\max} = 120$ km, assuming $\mu = 7 \times 10^{-4}$ (Zhivotovsky *et al.* 2003). It is therefore likely that considering the full range of geographical distances in this first analysis leads to an overestimate of σ^2.

Therefore, in the second analysis, we discarded all pairs of individuals born in different Baka groups and considered only pairs of individuals born nearby. We then found a stronger increase of genetic differentiation with geographical distance (slope = 0.0137; 95% CI 0.0038–0.0265; $p = 0.004$; figure 2) providing an estimate of $4D\pi\sigma^2 = 73$ individuals, which gave $\sigma^2 = 12.4$ km^2. Here, the maximum distance between two sampled individuals' birthplaces (26 km) was shorter than $d_{\max} = 53$ km, indicating that this estimate of σ^2 should no longer be biased by mutation. But we cannot exclude that reducing the sampling scale potentially excluded some long-distance migrants, which would lead to an underestimate of σ^2 (Rousset 2004).

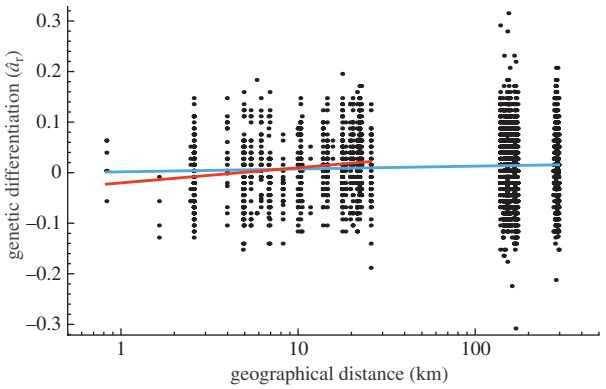


Figure 2. Correlation between genetic differentiation and the logarithm of geographical distances among Baka Pygmies. Multilocus estimates of pairwise differentiation ($\hat{a}_r$) are plotted against the logarithm of geographical distances (in kilometres). The linear regression considering all pairs of individuals is $y = 0.0027x - 0.0153$ (in blue). The linear regression considering only pairs of individuals born within the same group is $y = 0.0137x - 0.1138$ (in red).

Altogether, this suggests that the effective parent–offspring dispersal in the Baka Pygmies lies between 12.4 and 63.2 km^2. These indirect estimates of effective dispersal average over male and female genetic contributions, since they are based on autosomal

data. Assuming that parent–offspring dispersal distances are exponentially distributed (Cavalli-Sforza & Hewlett 1982), then half of the offspring disperse at distances shorter than $\sigma \ln(2)/\sqrt{2} \approx 0.49\,\sigma$. In this case, half of the Baka offspring disperse at a maximal distance between 1.7 and 3.9 km. However, using other distributions than the exponential for the dispersal distances would provide different interpretations. Therefore, dispersal distances estimated from σ^2 should be preferred here since they are independent of the (unknown) shape of the dispersal distribution.

Without quantitative demographic data on the Baka's mobility, we used the available demographic data from the Aka Pygmies from the Central African Republic (Cavalli-Sforza & Hewlett 1982) for comparison. We calculated the second moment of the distribution of distances between birthplaces and places of residence, an estimate of σ^2 from demographic data. From table 5 in Cavalli-Sforza & Hewlett (1982), we computed σ^2 as $\sum_i i^2 p_i$, where p_i is the proportion of individuals whose birthplace and place of residence are separated by i kilometres. We found $\sigma^2 = 3599\ \mathrm{km}^2$ for men and $4061\ \mathrm{km}^2$ for women ($3683\ \mathrm{km}^2$ on average). Hence, our estimates of the Baka's dispersal from genetic data are, respectively, 58.3 and 297 times lower than the average estimate of dispersal from demographic data in the Aka. How can this discrepancy be reconciled?

First, ethnologists show that mobility significantly differs across Pygmy populations from Central Africa (Bahuchet 1992). For instance, the mobility of the Bongo from Gabon probably decreased recently as opposed to the extended mobility of the Aka, since the Bongo nowadays live in permanent houses and widely practice agriculture (P. Verdu 2006–2007, unpublished results). In this context, the difference between our indirect estimates of dispersal in the Baka and the direct estimates in the Aka could result from differences in dispersal behaviour between these groups. Second, Cavalli-Sforza & Hewlett (1982) provide the distribution of distances between birthplaces and places of residence. This demographic data primarily reflects exploration behaviour rather than effective parent–offspring dispersal. To consistently compare demographic and genetic estimates, demographic estimates require an accurate knowledge of parents' and offspring's birthplaces, which is particularly challenging in Pygmies (Cavalli-Sforza 1986).

In conclusion, we found a strong signal of isolation by distance among the Baka Pygmies, a pattern owing to limited parent–offspring dispersal. Although our results do not challenge the view that hunter–gatherer Pygmies have frequent movements in their socio-economic area, we demonstrate that extended individual mobility does not necessarily reflect extended dispersal across generations. Limited effective dispersal may have reinforced genetic isolation among Pygmy populations, which could be a key mechanism explaining the strong genetic differentiation found among Western Central African Pygmies, despite their recent divergence from a single ancestral population about 2800 years ago (Verdu *et al.* 2009).

More generally, our findings also challenge the view that mobile hunter–gatherers disperse more than sedentary farmers. Hunter–gatherers are generally described as much more mobile than agricultural populations, based on mating distances (see fig. 3 in MacDonald & Hewlett (1999)). Here, we show that Baka hunter–gatherers' dispersal is strongly localized, as is the effective dispersal previously estimated in the horticulturalist Gainj- and Kalam-speakers from New Guinea ($\sigma^2 = 1.41\ \mathrm{km}^2$; Rousset 1997), the only other human genetic data collected at the appropriate geographical scale for applying Rousset's (1997) regression method (Long *et al.* 1987). This suggests that, despite their very mobile behaviour, foragers do not necessarily disperse more than farmers.

Research and sampling authorizations were obtained from the ethical committees of both the French and the Cameroonian governments.

The authors warmly thank all donors from Cameroon. We thank Frédéric Austerlitz, Michael DeGiorgio, Ethan Jewett, Trevor Pemberton, Noah Rosenberg and two anonymous reviewers for useful suggestions. This work was funded by the CNRS, the ACI Prosodie and the ANR (grant 05-BLAN-0400-01 MPRA; programme BLANC 'EMILE' NT09-611697).

Abéga, S. C. 1998 *Pygmées Baka: le droit à la différence*. Yaoundé, Cameroon: INADES-formation Cameroun.

Bahuchet, S. 1992 Spatial mobility and access to the resources among the African Pygmies. In *Mobility and territoriality* (eds M. J. Casimir & A. Rao), pp. 205–257. New York, NY; Oxford, UK: Berg.

Cavalli-Sforza, L. L. 1986 *African pygmies*. Orlando, FL: Academic Press.

Cavalli-Sforza, L. L. & Hewlett, B. 1982 Exploration and mating range in African Pygmies. *Ann. Hum. Genet.* **46**, 257–270. (doi:10.1111/j.1469-1809.1982.tb00717.x)

DeGiorgio, M., Jakobsson, M. & Rosenberg, N. A. 2009 Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl Acad. Sci. USA* **106**, 16 057–16 062. (doi:10.1073/pnas.0903341106)

Dhellemmes, I. & Macaigne, P. 1985 *Le père des pygmées*. Aventure vécue. Paris, France: Flammarion.

Liu, H., Prugnolle, F., Manica, A. & Balloux, F. 2006 A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79**, 230–237. (doi:10.1086/505436)

Long, J. C., Smouse, P. E. & Wood, J. W. 1987 The allelic correlation structure of Gainj- and Kalam-speaking people. II. The genetic distance between population subdivisions. *Genetics* **117**, 273–283.

MacDonald, D. H. & Hewlett, B. S. 1999 Reproductive interests and forager mobility. *Curr. Anthropol.* **40**, 501–523. (doi:10.1086/200047)

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

R Development Core Team 2007 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. & Cavalli-Sforza, L. L. 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15 942–15 947. (doi:10.1073/pnas.0507611102)

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. (doi:10.1046/j.1420-9101.2000.00137.x)

Rousset, F. 2004 *Genetic structure and selection in subdivided populations*. Monographs in population biology. Princeton, NJ: Princeton University Press.

Rousset, F. 2008 GenePop ' 007: a complete re-implementation of the GenePop software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–106. (doi:10.1111/j.1471-8286.2007.01931.x)

Sato, H. 1992 Notes on the distribution and settlement pattern of hunter–gatherers in Northwestern Congo. *African Study Monographs* **13**, 203–216.

Tsuru, D. 2001 Generation and transaction processes in the spirit ritual of the Baka Pygmies in Southeast Cameroon. *Afric. Stud. Monographs, Suppl.* **27**, 103–123.

Vallois, H. V. & Marquer, P. 1976 *Les Pygmées Baka du Cameroun: anthropologie et ethnographie. Mémoires du Mus. Nat. Hist. Nat., Série A, Tome C*. Paris, France: Muséum National d'Histoire Naturelle.

Verdu, P. *et al.* 2009 Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr. Biol.* **19**, 312–318. (doi:10.1016/j.cub.2008.12.049)

Zhivotovsky, L. A., Rosenberg, N. A. & Feldman, M. W. 2003 Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72**, 1171–1186. (doi:10.1086/375120)

Annexe E

Article 3 : évaluation de la méthode MsVar

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

DOI: 10.1534/genetics.110.121764

Inferring Population Decline and Expansion From Microsatellite Data: A Simulation-Based Evaluation of the Msvar Method

Christophe Girod,* Renaud Vitalis,† Raphaël Leblois‡ and Hélène Fréville*,§,1

**UMR 7179 (Centre National de la Recherche Scientifique–Muséum National d'Histoire Naturelle), F-91800 Brunoy, France, †Centre National de la Recherche Scientifique–Institut National de la Recherche Agronomique and ‡Institut National de la Recherche Agronomique, UMR Centre de Biologie pour la Gestion des Populations (INRA–Institut de Recherche pour le Développement–Centre de Coopération Internationale en Recherche Agronomique pour le Développement–Montpellier SupAgro), F-34988 Montferrier-sur-Lez Cedex, France and §Centre d'Ecologie Fonctionnelle et Evolutive, UMR 5175, Centre National de la Recherche Scientifique, F-34293 Montpellier Cedex 05, France*

Manuscript received August 4, 2010
Accepted for publication February 21, 2011

ABSTRACT

Reconstructing the demographic history of populations is a central issue in evolutionary biology. Using likelihood-based methods coupled with Monte Carlo simulations, it is now possible to reconstruct past changes in population size from genetic data. Using simulated data sets under various demographic scenarios, we evaluate the statistical performance of Msvar, a full-likelihood Bayesian method that infers past demographic change from microsatellite data. Our simulation tests show that Msvar is very efficient at detecting population declines and expansions, provided the event is neither too weak nor too recent. We further show that Msvar outperforms two moment-based methods (the *M*-ratio test and Bottleneck) for detecting population size changes, whatever the time and the severity of the event. The same trend emerges from a compilation of empirical studies. The latest version of Msvar provides estimates of the current and the ancestral population size and the time since the population started changing in size. We show that, in the absence of prior knowledge, Msvar provides little information on the mutation rate, which results in biased estimates and/or wide credibility intervals for each of the demographic parameters. However, scaling the population size parameters with the mutation rate and scaling the time with current population size, as coalescent theory requires, significantly improves the quality of the estimates for contraction but not for expansion scenarios. Finally, our results suggest that Msvar is robust to moderate departures from a strict stepwise mutation model.

INFERRING past demography is a central concern in evolutionary biology and applied ecology. Characterizing past variations in population size is crucial, *e.g.*, for understanding the impact of past climatic fluctuations on the current distribution of species (Jacobsen *et al.* 2005; Elmer *et al.* 2009; Hu *et al.* 2009) and for the conservation of endangered species (Frankham *et al.* 2002). Characterizing the demographic history of a species by direct approaches requires the monitoring of census data, which can be extremely difficult, not to say impossible, particularly in long-lived species. Yet variations in census numbers of individuals also affect the dynamics of the genes carried by these individuals. A powerful alternative to direct approaches is therefore to use the recent advances in population genetic theory, which allow inferences on past demography from the observed distribution of genetic variation in natural populations (Lawton-Rauh 2008).

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.121764/DC1.

[1]*Corresponding author:* Centre d'Ecologie Fonctionnelle et Evolutive, UMR 5175 CNRS, 1919 Rte. de Mende, 34293 Montpellier Cedex 5, France. E-mail: helene.freville@cefe.cnrs.fr

Until recently, most of these indirect methods relied on summary statistics calculated from genetic data and tests for departure from their theoretical distribution under a given demographic and mutational model (Cornuet and Luikart 1996; Schneider and Excoffier 1999; Garza and Williamson 2001). For instance, Cornuet and Luikart's (1996) approach relies on the rationale that rare alleles, which contribute marginally to the heterozygosity, are more likely to be lost following a bottleneck. A transient excess in heterozygosity, compared to that expected at equilibrium given the observed number of alleles in the sample, can therefore be used as a proxy to detect a bottleneck (Luikart and Cornuet 1998). Conversely, a transient heterozygosity deficiency may provide evidence for a population expansion (Cornuet and Luikart 1996; Leblois *et al.* 2006). In the same line of ideas, Garza and Williamson (2001) proposed a test to detect past population declines, based on the ratio (*M*) of the number of alleles to the range in allele size observed at microsatellite loci. Because they are easy to implement and do not require time-consuming computations, these moment-based methods have been used in many empirical studies

(see, *e.g.*, Spencer *et al.* 2000; Comps *et al.* 2001; Colautti *et al.* 2005). However, these methods suffer from a limited statistical power because they do not make full use of the data. Furthermore, they do not provide any estimate of the severity and the duration of the bottleneck.

Likelihood-based methods coupled with Monte Carlo sampling offer a powerful alternative to these moment-based methods (Felsenstein 1992; Griffiths and Tavaré 1994; Emerson *et al.* 2001). They rely upon the computation of the likelihood of a sample configuration, *i.e.*, the probability to observe the allele counts or the DNA polymorphic sites in that sample, given a demographic and mutational model. The parameters of interest of the underlying model are then estimated by maximizing the likelihood of the observed data. Likelihood-based methods that have been developed for inferring past demographic changes from the observed distribution of genetic variation include, *e.g.*, Batwing (Wilson and Balding 1998; Wilson *et al.* 2003), Beast (Drummond and Rambaut 2007), IM and IMa (Hey and Nielsen 2004, 2007), Lamarc (Kuhner 2006), and Msvar (Beaumont 1999). These methods differ not only with respect to the underlying demographic model, but also with respect to the markers used (microsatellites, DNA sequences, etc.). However, because the computational burden required to evaluate statistical power and accuracy is particularly high, only few studies have attempted to test these methods (Wilson *et al.* 2003; Abdo *et al.* 2004; Rousset and Leblois 2007; Chikhi *et al.* 2010; Strasburg and Rieseberg 2010).

Among those methods, the one developed by Beaumont (1999), implemented in the software package Msvar and further improved by Storz and Beaumont (2002) and Storz *et al.* (2002), has been increasingly used in the past few years to infer past demographic changes (supporting information, Table S1). Msvar assumes a demographic model consisting of a single isolated population, which has undergone a linear or exponential change in effective population size at some time in the past. This method is designed to analyze multilocus microsatellite data that evolve according to a stepwise mutation model (SMM) (Ellegren 2004). Msvar uses a Markov chain Monte Carlo (MCMC) method to sample from the posterior distribution of the model parameters (*i.e.*, the current effective population size, the ancestral effective population size before the demographic change, the time at which the latter occurred, and the mutation rate of microsatellite loci).

Although Msvar has been widely used, the statistical performance of the method has never been extensively evaluated. In his original article, Beaumont (1999) simulated a handful of data sets with known mutational and demographic parameters and then evaluated the performance of the method for detecting demographic events and its sensitivity to the shape (linear or exponential) of the demographic change. However, the precision of the estimation of the model parameters was not evaluated. Furthermore, the performance of Msvar with respect to the severity of demographic change, the time since the population started changing in size, and the mutation model has not been studied yet.

Here, we therefore aimed at evaluating the statistical performance of Msvar (i) in detecting population declines and expansions and (ii) in providing accurate estimates of the model parameters, as a function of the severity of the demographic change, the time since it occurred, and the mutation model. To that end, we performed stochastic simulations to generate microsatellite data sets under different demographic scenarios and mutation models and then analyzed these simulated data with Msvar. In light of our results, we comment upon the published empirical studies that used Msvar and provide some guidelines for future studies.

METHODS

Demographic model: The demographic model implemented in Msvar (Beaumont 1999; Storz and Beaumont 2002) considers an isolated panmictic population of size N_0 at sampling time ($t = 0$). Going backward in time, the population size $N(t)$ changes deterministically (either linearly or exponentially) to an ancestral size N_1 at time $t = T_a$ and then remains constant at N_1 for $t > T_a$ (Beaumont 1999). In the following, we will consider only an exponential change in population size, with

$$N(t) = N_0 \left(\frac{N_1}{N_0}\right)^{t/T_a},$$

for $0 < t < T_a$, and $N(t \geq T_a) = N_1$. For simplicity, the time is measured in units of generations, and population sizes are expressed as numbers of diploid individuals.

Simulation study: To test how Msvar performed depending upon the nature of the demographic change (decline or expansion), its strength, and its time of occurrence, we simulated population declines and expansions for a range of parameter values for the current population size N_0, the ancestral population size N_1, and the time T_a. The computational burden of the method prevented an exhaustive exploration of the parameter space. In a first set of simulations, we therefore concentrated on a set of parameter values that represented a range of situations characterized by weak, moderate, and strong changes in population size, with varying time of occurrence. For population declines, we fixed the current population size $N_0 = 100$ in all scenarios and varied the ancestral population size $N_1 = \{1000; 10{,}000; 100{,}000\}$ and the time since the demographic change $T_a = \{10; 50; 100; 500\}$. For population expansions, we fixed the ancestral population size $N_1 = 100$ in all scenarios

and varied the current population size $N_0 = \{1000; 10{,}000; 100{,}000\}$ and the time since the demographic change $T_a = \{10; 50; 100; 500\}$. A total of 24 sets of demographic parameters were therefore considered. For this first set of simulations, each locus evolved according to a strict SMM, as assumed in Msvar. The mutation rate μ was set at 10^{-3}, which is in agreement with estimates from the literature (Ellegren 2004).

Then, to test how Msvar performed depending upon the mutation model, we performed a second set of simulations. The mutation process of microsatellites is complex and highly heterogeneous across loci and organisms (Ellegren 2000, 2004). While some observations of spontaneous mutations support a strict SMM, others suggest that multistep mutations occur, with a frequency of multistep changes p varying from 0.04 to 0.74 (Ellegren 2000, 2004). Apart from a strict SMM, we thus simulated microsatellite data under a generalized stepwise model (GSM) with $p = 0.22$, an average value found in the literature (Dib *et al.* 1996; Ellegren 2000; Estoup *et al.* 2001; Ellegren 2004), and with $p = 0.74$, the most extreme value reported ever (Fitzsimmons 1998). The mutation rate μ was set at 10^{-3}. For that second set of simulations, we considered a population decline scenario (with $N_0 = 100$, $N_1 = 10{,}000$, and $T_a = 500$), a population expansion scenario (with $N_0 = 10{,}000$, $N_1 = 100$, and $T_a = 500$), and a stable population scenario, taking the (constant) population size as the harmonic mean of the population size change from 100 to 10,000 for $T_a = 500$ generations; *i.e.*, $N_0 = N_1 = 464$. This second set of simulations therefore consisted of seven sets of parameters: three mutation models were considered for the stable population scenario (the SMM and the two GSMs), and two mutation models were considered for each of the declining and expanding population scenarios (the two GSMs).

Microsatellite data were simulated with Simcoal2 (Laval and Excoffier 2004), which generates samples of genes under various demographic models, using a discrete-generation coalescent algorithm. Discrete-generation algorithms produce simultaneous and multiple coalescences, which are canceled out in the continuous-time approximation of the coalescent. There might therefore be a slight discrepancy between the coalescence rate in the discrete-generation algorithm and the continuous-time approximation that is assumed in Msvar, particularly for large sample sizes and small effective population size (see, *e.g.*, Figure S2 in Cornuet *et al.* 2008). However, we find it more relevant to simulate the data without relying on approximations. Each data set consisted of a sample of 50 diploid individuals, genotyped at 10 unlinked microsatellite loci. This sampling scheme is consistent with empirical studies that inferred past demographic changes using Msvar: from an exhaustive survey of the literature (Table S1), we found that the median numbers of microsatellite loci and sampled individuals across data sets were 11.5 and 30, respectively.

For each set of parameters, we simulated five microsatellite data sets to have replicates from the same underlying demographic and mutation model. We therefore obtained a total of 120 simulated data sets for the first set of simulations and 35 data sets for the second set. For each set, we calculated the mean and standard deviation over the five replicates of the expected heterozygosity H_e (Nei 1978), the observed number of alleles N_a, the range in allele size A_r, and the variance of allele range V_a, using Arlequin (Excoffier *et al.* 2005).

Parameterization of Msvar: In Msvar, the posterior distribution of the model parameters is computed by means of a MCMC method using the Metropolis–Hastings algorithm (Metropolis *et al.* 1953; Hastings 1970). The likelihood is calculated from the genealogical history of the sample of genes, represented as a sequence of events (coalescences and mutations, see Beaumont 1999).

We used version 1.3 of Msvar, which provides separate estimates for N_0, N_1, μ, and T_a (Storz and Beaumont 2002). This implementation of Beaumont's (1999) method, available at http://www.rubic.rdg.ac.uk/~mab/stuff/, relies upon a hierarchical model where demographic and mutational parameters are allowed to vary among loci. The extent of interlocus variation is set by priors, as described in File S1. The parameter values reported in the following correspond to the mean of N_0, N_1, T_a, and μ, across loci. To test whether the method could retrieve information from the data, we chose relatively flat priors on the mean parameters, including the mutation rate.

Implementation: Analyses were run on a Beowulf cluster made of 19 computer nodes, with CPUs ranging from biprocessors AMD Opteron monocore running at 1.8 GHz to biprocessors Intel Xeon quadcore running at 2.0 GHz. For each of the simulated data sets, three independent Msvar analyses were performed, with different starting values of the model parameters and different sets of seeds for the random number generator. For the first set of simulations (strict SMM with population size change), each Markov chain was initially run for 10^9 steps and was thinned to 40,000 output lines by recording parameter values every 25,000 steps. In the absence of convergence, longer chains were run (see results). For the second set of simulations (strict SMM with stable population and GSM with stable, expanding, and declining populations), each Markov chain was run for 3×10^9 steps, with parameter values recorded every 30,000 steps. The first 10% of steps of the chains were discarded as burn-in. For each data set, convergence was assessed by computing the multivariate extension of Gelman and Rubin's diagnostic (Brooks and Gelman 1998) on the three independent Markov chains. Gelman and Rubin's diagnostic is based on the computation of the ratio of the pooled-chains variance over the within-

chain variance, which should be close to 1 if the chains converge to the target distribution. The multivariate Gelman and Rubin's diagnostic was calculated from the means $\mathbf{M} \equiv \{M_{N_0}, M_{N_1}, M_{T_a}, M_\mu\}$ and standard deviations $\mathbf{V} \equiv \{V_{N_0}, V_{N_1}, V_{T_a}, V_\mu\}$ of Msvar parameters across loci, using the CODA package (Plummer *et al.* 2006) implemented in the statistical software R (R Development Core Team 2009). Although it might be recommended to run more chains to compute Gelman and Rubin's diagnostic (*e.g.*, Nettel *et al.* 2009 used eight independent chains), the computational burden prevented us from running more than three chains in the present study.

Analysis of Msvar outputs: Msvar outputs were analyzed by focusing on two issues: (i) the performance of Msvar at detecting past demographic changes and (ii) the precision of Msvar estimates of the model parameters. For each of the simulated data sets, we combined the three Markov chains before running the following analyses. The strength of evidence of population expansion *vs.* population decline (and vice versa) was evaluated using Bayes factors (Jeffreys 1961; Kass and Raftery 1995), as suggested by Beaumont (1999) and Storz and Beaumont (2002). The Bayes factor is a ratio where the numerator is the posterior probability of one model divided by its prior probability and the denominator is the posterior probability of an alternative model divided by its prior probability (Gelman *et al.* 1995). With identical priors for the population decline and the population expansion models (*i.e.*, identical priors for N_0 and N_1), the Bayes factor for, *e.g.*, a population decline is the ratio of the posterior probability of a population decline divided by the posterior probability of a population expansion. This ratio can be estimated by counting the number of states in the chain in which the population has declined (*i.e.*, $N_0/N_1 < 1$) and then dividing this by the number of states in which the population has expanded (*i.e.*, $N_0/N_1 > 1$) (see Storz and Beaumont 2002).

We estimated the marginal posterior distributions of the model parameters using the LOCFIT package (Loader 1999) implemented in R (R Development Core Team 2009). Point estimates of natural parameters N_0, N_1, T_a, and μ were computed from the mode of their marginal posterior distribution. The 90% highest probability density (HPD) intervals were computed with the CODA package. We also estimated the marginal posterior distributions of the scaled parameters $\theta_0 \equiv 4N_0\mu$, $\theta_1 \equiv 4N_1\mu$, and $t_f \equiv T_a/(2N_0)$, and we computed point estimates and 90% HPD intervals for these scaled parameters. For each demographic scenario considered, we calculated the absolute value of the bias for both natural and scaled parameters over the five replicated data sets.

Detection of population size change with Bottleneck and the *M*-ratio test: Finally, for the first set of simulations (strict SMM with population size change), we compared the performance of Msvar to detect genetic signatures of demographic changes with the two most widely used moment-based methods available for microsatellite data. First, we analyzed the data sets using the method developed by Cornuet and Luikart (1996) and implemented in the software package Bottleneck v.1.2 (Cornuet and Luikart 1996). Wilcoxon signed-rank tests were performed to determine if a data set exhibited a significant number of loci with heterozygosity excess as expected in bottlenecked populations (Luikart *et al.* 1998) or with heterozygosity deficiency as expected in expanding populations (Cornuet and Luikart 1996). Second, we calculated Garza and Williamson's (2001) M ratio on the 60 data sets corresponding to population declines. We compared empirical values of the M ratio to 95% critical values (M_c) derived from 10,000 simulations of stable populations using the program Critical_M. Simulations were performed using the true value of θ_1 ($\theta_1 = 4$, 40, and 400 in the scenarios considered) and assuming a strict stepwise mutation model. We considered that an M ratio below the critical value M_c was indicative of a population decline.

RESULTS

Genetic diversity of the simulated data sets, under a strict SMM: For the first set of simulations (strict SMM with population size change), the expected heterozygosity H_e, the number of alleles N_a, and the range in allele size A_r are reported in Table 1. For contraction scenarios, H_e ranged from 0.24 to 0.94. N_a ranged from 2.3 to 23.7 and A_r varied from 1.3 to 39.2. In agreement with theoretical expectations, H_e, N_a, and A_r increased with N_1, the genetic diversity in the current population being sustained by large ancestral populations. Furthermore, genetic diversity decreased with increasing T_a, the loss of genetic diversity being more pronounced for long contraction events. For expansion scenarios, H_e ranged from 0.30 to 0.58. N_a ranged from 2.7 to 4.9 and A_r varied from 1.8 to 4.0. In agreement with theoretical expectations, H_e, N_a, and A_r increased with increasing T_a since the number of mutations that segregate in the population increases with the age of the expansion event. We also observed a tendency for genetic diversity to increase with increasing N_0, although this trend was not clear cut.

MCMC convergence: In the following, we used Gelman *et al.*'s (2004) rule of thumb, which suggests that values of the multivariate Gelman and Rubin's convergence diagnostic between 1.0 and 1.1 indicate reasonable convergence, whereas values >1.1 indicate poor convergence. Of the 120 analyses of the first set of simulations, 67 converged after 10^9 steps (Table S2). The average computational time of these chains was 1.5 days for expansions and 3 days for contractions. The 53 nonconverged analyses were run again for 3×10^9 steps

TABLE 1

Genetic diversity in the first set of simulated data

	Contractions, with $N_0 = 100$ (SMM)											
	$N_1 = 1{,}000$				$N_1 = 10{,}000$				$N_1 = 100{,}000$			
T_a	H_e	N_a	A_r	V_a	H_e	N_a	A_r	V_a	H_e	N_a	A_r	V_a
10	0.60 (0.06)	4.5 (0.4)	3.8 (0.6)	2.7 (0.8)	0.86 (0.01)	10.8 (0.4)	11.8 (1.9)	17.2 (5.3)	0.94 (0.01)	23.7 (2.6)	34.6 (6.1)	107.3 (62.8)
50	0.61 (0.04)	4.4 (0.3)	3.7 (0.7)	2.7 (1.0)	0.85 (0.02)	10.2 (0.1)	12.7 (0.9)	21.7 (6.4)	0.93 (0.01)	20.1 (0.6)	37.5 (1.9)	138.7 (15.3)
100	0.55 (0.03)	3.9 (0.5)	3.5 (0.8)	2.8 (1.0)	0.80 (0.01)	8.1 (0.6)	10.9 (1.9)	16.4 (6.1)	0.89 (0.01)	14.8 (1.1)	39.2 (2.8)	171.6 (33.8)
500	0.24 (0.05)	2.3 (0.2)	1.3 (0.2)	0.7 (0.2)	0.56 (0.04)	4.2 (0.4)	6.9 (1.3)	13.5 (5.6)	0.66 (0.05)	5.8 (0.6)	27.1 (2.2)	165.7 (58.6)
	Expansions, with $N_1 = 100$ (SMM)											
	$N_0 = 1{,}000$				$N_0 = 10{,}000$				$N_0 = 100{,}000$			
T_a	H_e	N_a	A_r	V_a	H_e	N_a	A_r	V_a	H_e	N_a	A_r	V_a
10	0.34 (0.03)	2.7 (0.1)	1.8 (0.2)	1.0 (0.3)	0.31 (0.03)	2.8 (0.1)	1.8 (0.1)	1.0 (0.1)	0.44 (0.02)	3.4 (0.4)	2.4 (0.5)	1.4 (0.4)
50	0.30 (0.03)	3.0 (0.3)	2.0 (0.4)	1.1 (0.3)	0.30 (0.03)	3.2 (0.3)	2.2 (0.3)	1.1 (0.2)	0.36 (0.06)	3.3 (0.4)	2.3 (0.5)	1.3 (0.4)
100	0.32 (0.03)	3.0 (0.1)	2.0 (0.2)	1.1 (0.1)	0.34 (0.05)	3.6 (0.2)	2.6 (0.2)	1.4 (0.2)	0.32 (0.05)	3.5 (0.2)	2.5 (0.3)	1.4 (0.2)
500	0.43 (0.08)	3.5 (0.3)	2.6 (0.3)	1.5 (0.2)	0.52 (0.04)	4.3 (0.4)	3.3 (0.4)	2.0 (0.3)	0.58 (0.03)	4.9 (0.3)	4.0 (0.3)	2.6 (0.2)

SMM, stepwise mutation model; H_e, expected heterozygosity; N_a, number of alleles; A_r, allele size range; V_a, variance of allele size range. Estimates of genetic diversity are averaged over the five simulated data sets for each set of demographic parameters (N_0, current effective population size; N_1, ancestral effective population size; T_a, time since the population size change). Standard deviations are indicated in parentheses.

and were thinned to 120,000 output lines by recording parameter values every 25,000 steps. Of these 53 analyses, 20 converged after 3×10^9 steps, which took on average 20 days per chain. Finally, the last 33 nonconverged analyses were run for 1.5×10^{10} steps, which took 60 days per chain on average. Of these 33 analyses, 16 converged after 1.5×10^{10} steps. Therefore, a total of 17 analyses of 120 (14.2%) did not converge after 1.5×10^{10} steps. Most of these nonconverged analyses corresponded to recently and severely bottlenecked populations ($T_a < 500$ and $N_0/N_1 = 0.001$; Table S2). However, visual inspection of the three chains in the nonconverged analyses, as well as the similarity of the marginal posterior distributions, suggested that the chains were close to equilibrium. Therefore, we included the 17 nonconverged analyses in our results. The cumulative computation time for the completion of all the analyses included in our study exceeded 276×10^3 hr (33.5 years). There was a significantly positive correlation between the time of convergence and the average range in allele size A_r in the sample for both contractions (Spearman's $\rho = 0.82$; $P < 0.001$) and expansions (Spearman's $\rho = 0.49$; $P < 0.001$).

Detection of demographic events with Msvar, under a strict SMM: Bayes factors (BF) were computed for each of the 120 analyses of the first set of simulations and interpreted following Jeffreys (1961): BF $\geq$ 10 indicate strong support, BF ranging from 3 to 10 indicate substantial support, BF ranging from 0.33 to 3 indicate no support, and values <0.33 indicate false detection of contraction or expansion. In 85 analyses of 120 (70.8%), Bayes factors indicated a change in population size consistent with the simulated scenario with substantial to strong support (BF $\geq$ 3 and BF $\geq$ 10, respectively; see Figure 1). Of the 60 Markov chains corresponding to contraction scenarios, 41 (68.3%) indicated a population decline (BF $\geq$ 3), of which 40 (97.6%) showed strong support (BF $\geq$ 10). Fifteen of these 40 analyses (37.5%) did not converge. Of the 60 analyses corresponding to expansion scenarios, 44 (73.3%) indicated a population expansion (BF $\geq$ 3), of which 34 (77.3%) showed strong support (BF $\geq$ 10). Two of these 34 analyses (5.9%) did not converge. Overall, all the ancient ($T_a \geq 50$) and severe demographic changes ($N_0/N_1 \leq 0.01$ for contractions and $N_0/N_1 \geq 100$ for expansions) were detected with substantial to strong support (Figure 1). By contrast, recent declines and expansions ($T_a = 10$) were largely undetected (BF < 3), except for strong contractions ($N_0/N_1 = 0.001$). Moreover, weak contractions ($N_0/N_1 = 0.1$) were largely undetected whatever their time of occurrence and one false expansion was even detected for an ancient and weak bottleneck (BF < 0.33, $T_a = 500$, $N_0/N_1 = 0.1$).

Comparison of Msvar with moment-based methods: Because Bayes factors cannot be formally compared to *P*-values, we were not able to use the same criterion for

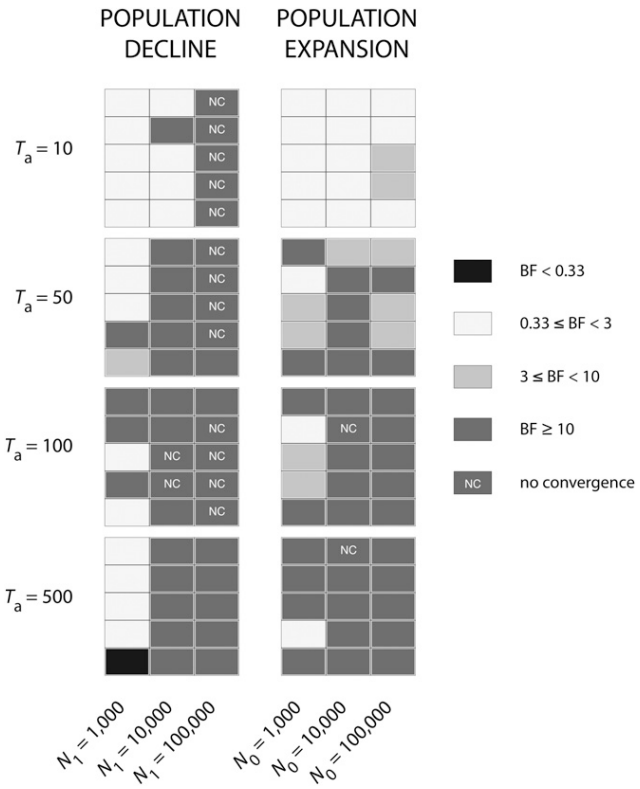


Figure 1.—Detection of change in population size with Msvar. For population declines (left) and population expansions (right), the Bayes factors (BF) are given for each set of demographic parameters N_0, N_1, and T_a for each replicated data set (lines). Following Jeffreys (1961), BF ≥ 10 indicate strong support, and BF ranging from 3 to 10 indicate substantial support. BF ranging from 0.33 to 3 indicate no support and values <0.33 indicate false detection of contraction or expansion. Nonconverged (NC) analyses are also indicated.

detecting population size change with Msvar, Bottleneck, and the *M*-ratio test. Therefore, we reported in Figure 2 the criteria that are generally used in empirical studies: BF ≥ 3 for Msvar, the result of the Wilcoxon signed-rank tests at the $\alpha = 0.05$ level for Bottleneck (Cornuet and Luikart 1996), and an *M* ratio below the critical value M_c (Garza and Williamson 2001) for the *M*-ratio test.

Bottleneck detected a significant excess of heterozygosity in only 5 of the 60 data sets corresponding to contraction scenarios (8.3%). Ancient events (T_a = 500) were never detected whatever their severity. Moreover, there was no clear relationship between the rate of detection of population decline and the severity of the event. Finally, 4 data sets corresponding to ancient contractions ($T_a \geq 100$) showed significant heterozygote deficiency, hence supporting population expansions. Contrastingly, Bottleneck detected a significant deficiency in heterozygosity in 35 of 60 data sets corresponding to expansion scenarios (58.3%). Garza and Williamson's (2001) *M*-ratio method correctly detected a signal of contraction in 32 of 60 data sets (53.3%). The rate of detection was higher for ancient ($T_a \geq 50$) and moderate-to-severe population declines ($N_0/N_1 \geq 0.01$), with 26 significant tests of 30. Recent ($T_a = 10$) and/or weak declines ($N_0/N_1 = 0.1$) were barely detected (6 significant tests of 30).

Estimation of demographic and mutational parameters with Msvar, under a strict SMM: Demographic and mutational parameters were estimated for all data sets, by combining the three Markov chains run for each data set. We assessed the quality of the estimates by examining the marginal posterior distributions of the parameters, compared to their prior distributions. We summarized these results by calculating the modes and the 90% HPD intervals for each data set (Figure S1, Figure S2, Figure S3, and Figure S4) as well as the absolute value of the bias and the average HPD range over the five replicate data sets for each of the 24 demographic scenarios (Figures 3 and 4).

Estimates of the natural parameters N_0, N_1, T_a, and μ: Overall, the marginal posterior distributions of the demographic parameters N_0, N_1, and T_a were wide and departed only slightly from the priors (*e.g.*, Figure 5, A and B). The estimated 90% HPD limits were therefore broad (Figure 3), ranging from −4 to 8 in $\log_{10}$ scale (Figure S1 and Figure S2).

For contractions, replicated data sets tended to provide more consistent results for old and severe events, compared to recent events ($T_a = 10$) or events of low severity ($N_0/N_1 = 0.1$) (Figure S1). The precision of the demographic parameter estimates tended to increase with increasing severity of the demographic change (measured by the ratio N_0/N_1) and the time of the event: (i) the 90% HPD range of the demographic parameter estimates decreased with increasing N_0/N_1 (Figure 3 and Figure S1); (ii) for moderate to strong contractions ($N_0/N_1 > 0.1$), the 90% HPD range decreased with increasing T_a; and (iii) for $N_0/N_1 = 0.1$, the 90% HPD range was the lowest for intermediate values of T_a. The absolute value of the bias of N_0 estimates tended to be lower than that of N_1 and was maximized for recent events ($T_a = 10$).

The quality of the estimates of N_0, N_1, and T_a was poorer for expansions, compared to contractions. The marginal posterior distributions were not sharply peaked and did not depart markedly from the priors. The 90% HPD limits were wide and the absolute value of the bias was high, overall (Figure 3). This was true whatever the severity of the event and its time of occurrence. It is noteworthy that, with few exceptions, all demographic parameter estimates differed markedly across replicate data sets (Figure S2). We noted that for a given expansion severity, estimates of N_0 increased with T_a, while estimates of N_1 decreased with T_a (Figure S2).

For contractions and expansions, the marginal posterior distributions of μ departed only slightly from the prior distributions, whose mean was set at $\alpha_\mu = -4$ on a $\log_{10}$ scale. Because the true mutation rate μ of the simulated data sets was set at −3 on a $\log_{10}$ scale, the

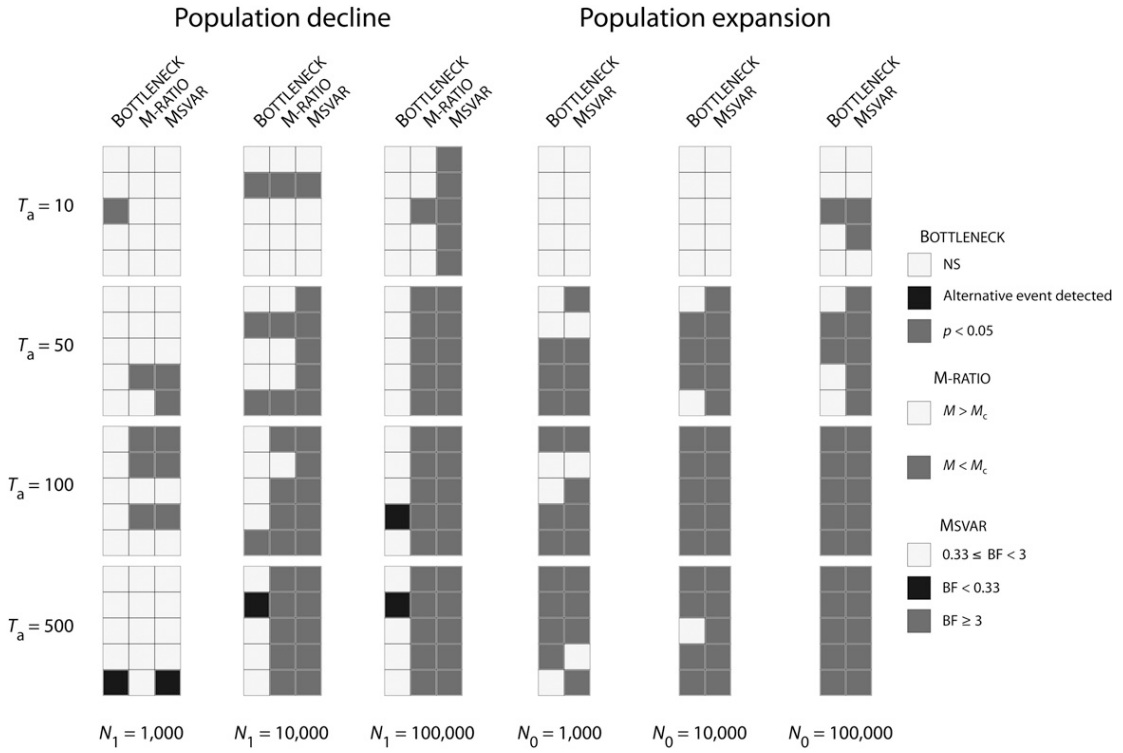


FIGURE 2.—Detection of change in population size with Bottleneck, the *M*-ratio test, and Msvar. For population declines with $N_0 = 100$ (left set of panels), Bottleneck, the *M*-ratio test, and Msvar are compared. For population expansions with $N_1 = 100$ (right set of panels), only Bottleneck and Msvar are compared. Squares with dark shading indicate detection with Msvar, the *M*-ratio test, or Bottleneck; squares with light shading indicate no detection; and solid squares indicate detection with Msvar or Bottleneck of the "wrong" event. Bottleneck: NS, not significant. *M*-ratio test: M_c is the critical value below which the test is significant. Msvar: BF, Bayes factor.

mutational parameter μ was therefore systematically underestimated, as already pointed out by MILTON *et al.* (2009). The 90% HPD intervals of the marginal posterior distributions of μ were wide (data not shown).

Finally, we examined the patterns of correlation between natural parameters to assess the performance of Msvar to estimate natural parameters separately. We observed strong correlations between natural parameters of the model. Overall, both N_0 and N_1 were negatively correlated with the mutational parameter μ and there was a positive correlation between N_0 and T_a (Figure S5). The correlations were stronger for more severe events and more ancient events. Furthermore, the correlations were more pronounced for contractions than for expansions.

Estimates of the scaled parameters θ_0, θ_1, and t_f: Scaled parameters were overall much more precisely estimated than the natural parameters for contractions, whereas they were poorly estimated for expansions. As with the natural parameters, the quality of the estimates depended upon the severity of the demographic change and its time of occurrence.

For contractions, the marginal posterior distributions of the scaled parameters θ_0, θ_1, and t_f were very peaked and departed markedly from the prior distributions (*e.g.*, Figure 5, C and D), except for contractions of low severity ($N_0/N_1 = 0.1$). The precision (low bias, narrow 90% HPD interval) increased with increasing severity of the event and time of occurrence (Figure 4 and Figure S3). In particular, estimates of θ_1 and t_f were overall very precise for moderate to severe bottlenecks ($N_0/N_1 < 0.1$), except for very recent events ($T_a = 10$). Although θ_0 was also well estimated for ancient declines ($T_a > 50$) from moderate to strong severity, the bias and the range of 90% HPD intervals were larger compared to those of θ_1. Replicate data sets provided consistent results for θ_1 and t_f, for moderate and strong contractions ($N_0/N_1 < 0.1$) that occurred >10 generations ago ($T_a > 10$). Larger variation across replicate data sets was observed for θ_0.

For expansions, the marginal posterior distributions of the scaled parameters θ_0 and θ_1 were peaked and departed markedly from the prior distributions (data not shown). However, the mode of the marginal posterior distributions for θ_0 departed markedly from the true simulated value, resulting in severe biases (Figure 4 and Figure S4). By contrast, the scaled parameter θ_1 exhibited low bias in all scenarios (Figure 4), although the 90% HPD intervals were wide, especially for weak and recent expansions ($N_0/N_1 = 10$; $T_a > 10$). We

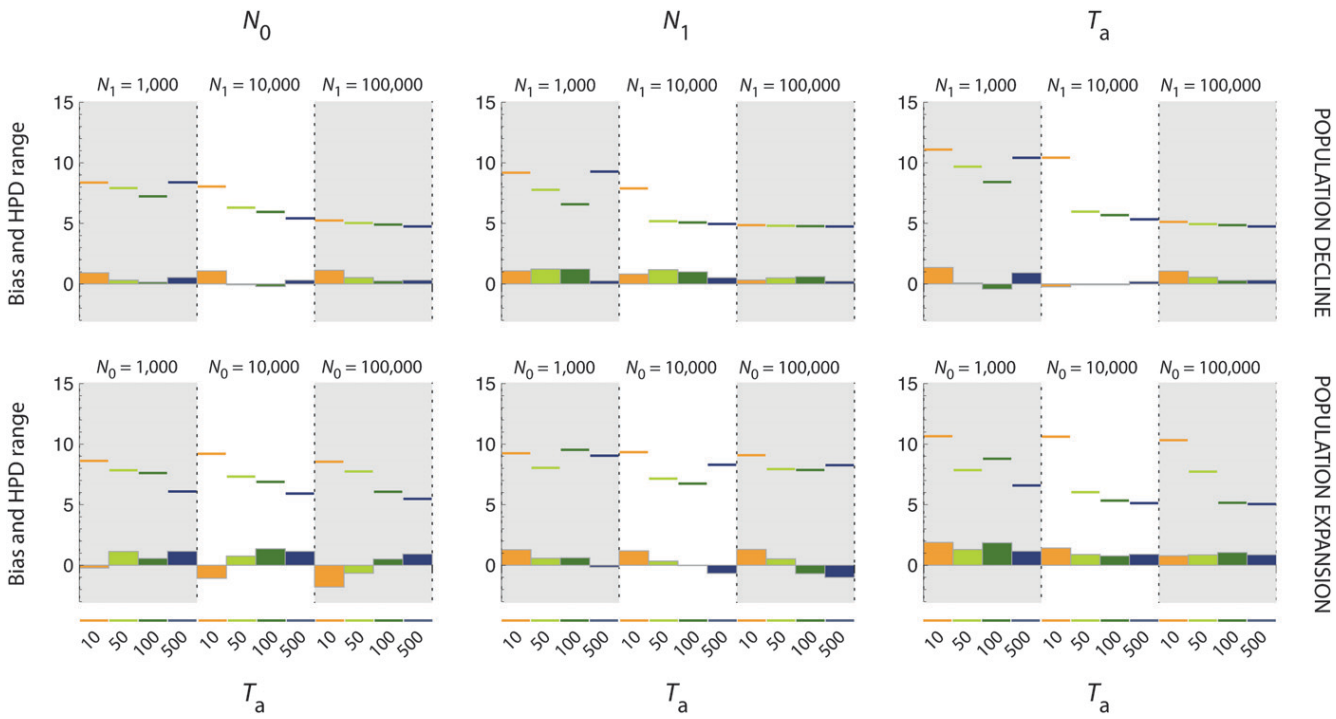


FIGURE 3.—Precision of the estimates of the natural demographic parameters N_0, N_1, and T_a. Bias (histograms) and absolute value of the range of the 90% HPD interval (horizontal colored traits) for natural demographic parameters N_0, N_1, and T_a (from left to right) are presented in a $\log_{10}$ scale. Top, population declines; bottom, population expansions. In each graph, the dotted vertical line separates scenarios of increasing severity ($N_1 = 1000$, $N_1 = 10{,}000$, and $N_1 = 100{,}000$ for population declines and $N_0 = 1000$, $N_0 = 10{,}000$, and $N_0 = 100{,}000$ for population expansions). For each severity, the time of occurrence of the demographic event T_a is represented by different colors (orange for $T_a = 10$, light green for $T_a = 50$, dark green for $T_a = 100$, and blue for $T_a = 500$).

noted that the marginal posterior distributions of θ_0 and θ_1 were skewed, respectively, to upper and lower values (Figure 4 and Figure S4). The 90% HPD intervals of the marginal posterior distributions for the time parameter t_f were wide in most conditions (Figure 4 and Figure S4) and estimates were severely biased, except for $T_a = 500$. Both the 90% HPD interval and the bias decreased with increasing T_a. We observed large variations across replicate data sets for all scaled parameters in almost all situations, particularly for t_f.

Influence of the mutation model in Msvar: Of the 30 analyses presented in Figure 6 for the GSM, 12 (40%) did not converge after 3×10^9 steps. Out of these, 9 analyses (75%) concerned the data sets generated with the strongest GSM ($p = 0.74$). With data generated under the moderate GSM ($p = 0.22$), Msvar successfully detected a population decline for the five simulated data sets and a population expansion for four simulated data sets of five. However, Msvar detected a false signal of population decline for two data sets of five that were simulated under a stable population scenario (Figure 6). Under a strong GSM ($p = 0.74$), Msvar detected a signal of population decline with strong support (BF $\geq$ 10), whatever the simulated scenario (Figure 6).

The quality of Msvar estimates of scaled parameters for the moderate GSM ($p = 0.22$) was very similar to that observed for the strict SMM, with very precise estimates of θ_1 and t_f, a slightly larger bias, and 90% HPD intervals for θ_0 compared to θ_1 in contraction scenarios and poorer estimates, with large variations across replicate data sets, in expansion scenarios (Figure S6). For stable population scenarios, both the strict SMM and the moderate GSM ($p = 0.22$) produced unbiased estimates of θ_0 and θ_1, but with very large 90% HPD intervals. Note that in the absence of population size change, estimates of t_f are meaningless. Very consistently, Msvar produced biased estimates of the model parameters, with very narrow 90% HPD intervals, for all the data sets generated under the strong GSM ($p = 0.74$) (Figure S6).

DISCUSSION

Comparing Msvar, Bottleneck, and the *M*-ratio test: Bottleneck performed poorly in detecting population declines from our simulated data sets under a SMM, with only 5 significant tests of 60. The statistical power of Bottleneck for population declines is much lower when microsatellite loci evolve under a strict SMM than under an infinite-allele model (CORNUET and LUIKART 1996) or a GSM (LEBLOIS *et al.* 2006). This may partly explain the low performance of Bottleneck in our comparative study. Our results for weak population declines ($N_1/N_0 = 10$) are in agreement with previous

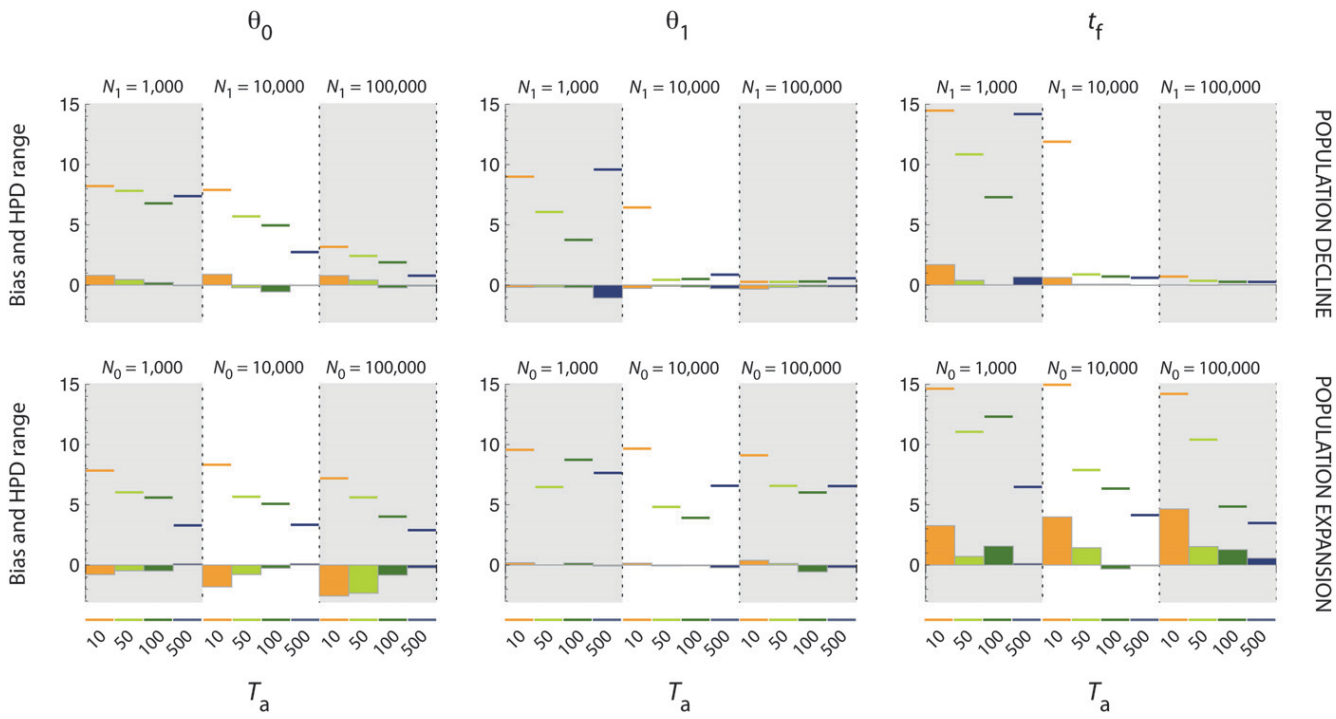


FIGURE 4.—Precision of the estimates of the scaled parameters θ_0, θ_1, and t_f. Bias (histograms) and absolute value of the range of the 90% HPD interval (horizontal colored traits) for scaled parameters θ_0, θ_1, and t_f (from left to right) are presented in a $\log_{10}$ scale. Top, population declines; bottom, population expansions. In each graph, the dotted vertical line separates scenarios of increasing severity ($N_1 = 1000$, $N_1 = 10{,}000$, and $N_1 = 100{,}000$ for population declines and $N_0 = 1000$, $N_0 = 10{,}000$, and $N_0 = 100{,}000$ for population expansions). For each severity, the time of occurrence of the demographic event T_a is represented by different colors (orange for $T_a = 10$, light green for $T_a = 50$, dark green for $T_a = 100$, and blue for $T_a = 500$).

simulation-based evaluations, given the set of demographic and mutational parameters considered here (see, *e.g.*, Figure 3B in CORNUET and LUIKART 1996). For moderate to severe population declines ($N_1/N_0 \geq 100$), however, the rate of detection was lower in our study than in CORNUET and LUIKART (1996). Two possible reasons may explain this discrepancy. First, the average heterozygosity in our simulated data sets was

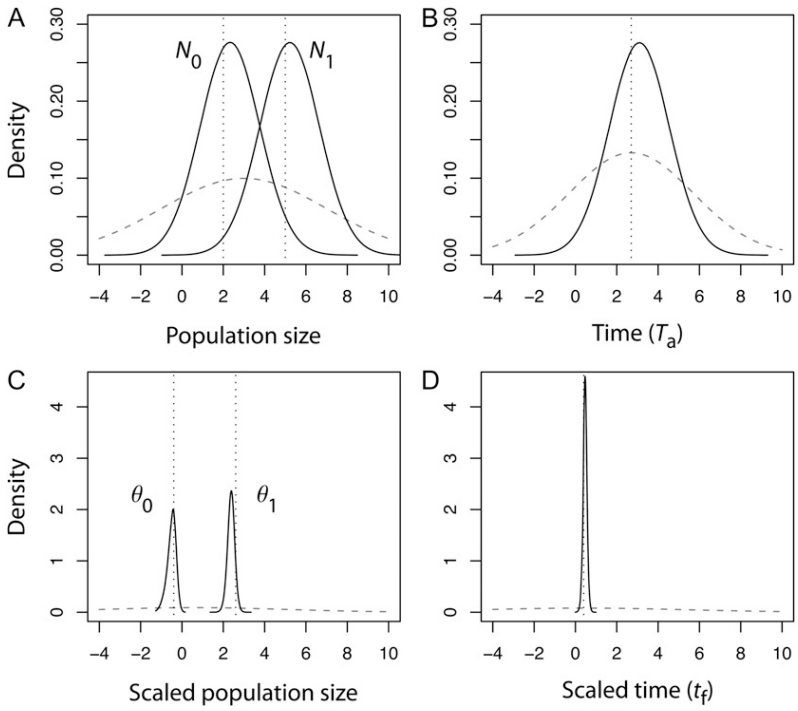


FIGURE 5.—Marginal posterior density of N_0, N_1, and T_a and θ_0, θ_1, and t_f for an ancient and severe population decline. All densities are represented in a $\log_{10}$ scale. (A) Population size natural parameters N_0 and N_1. (B) Time natural parameter T_a. (C) Scaled parameters θ_0 and θ_1. (D) Scaled parameter t_f. The scenario corresponds to an ancient ($T_a = 500$) and severe population decline ($N_0 = 100$, $N_1 = 100{,}000$). The true values of the parameters in a $\log_{10}$ scale ($N_0 = 2$, $N_1 = 5$, $T_a = 2.70$, $\theta_0 = -0.40$, $\theta_1 = 2.60$, $t_f = 0.40$) are indicated by the vertical dotted line in each graph. The prior distributions of the parameters are given by the shaded dashed curve in each graph.

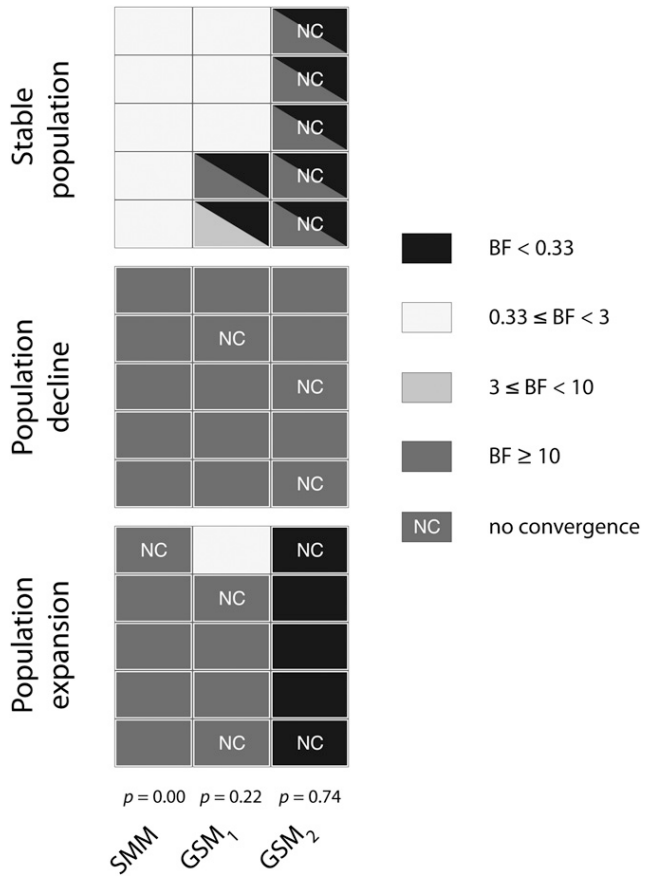


FIGURE 6.—Detection of change in population size with Msvar. The Bayes factors (BF) are given for each of the following demographic scenarios: a stable population (with $N_0 = N_1 = 464$, $T_a = 500$), a declining population (with $N_0 = 100$, $N_1 = 10{,}000$, and $T_a = 500$), and an expanding population (with $N_0 = 10{,}000$, $N_1 = 100$, and $T_a = 500$). We considered three different mutation models, which differ from each other by the value of p, the frequency of multistep mutation changes: $p = 0.00$ (stepwise mutation model, SMM), $p = 0.22$ (moderate generalized stepwise model, GSM_1), and $p = 0.74$ (strong generalized stepwise model, GSM_2). For the stable population scenario, the lower triangle provides the Bayes factor for a population decline (*i.e.*, the ratio of the posterior probability of a population decline divided by the posterior probability of a population expansion), and the upper triangle provides the Bayes factor for a population expansion (*i.e.*, the ratio of the posterior probability of a population expansion divided by the posterior probability of a population decline). Following JEFFREYS (1961), BF ≥ 10 indicate strong support, and BF ranging from 3 to 10 indicate substantial support. BF ranging from 0.33 to 3 indicate no support and values <0.33 indicate false detection of contraction or expansion. Nonconverged (NC) analyses are also indicated.

overall higher than in CORNUET and LUIKART (1996) who considered a variable mutation rate across loci and simulations, to cover a range of heterozygosities per set of parameters. Second, we simulated an exponential change of population size, whereas CORNUET and LUIKART (1996) assumed an instantaneous reduction of population size in their simulation-based tests. The impact of the shape of the demographic change on the performance of Bottleneck has not been studied yet. Consistent with our study, the simulation-based evaluation of Bottleneck by LEBLOIS *et al.* (2006) also showed a low statistical power of the method. Interestingly, Bottleneck performed largely better for expansions (58.3%) than for contractions (8.3%), given the model parameter values of our study.

We found that the *M*-ratio test was more efficient than Bottleneck, which is consistent with LEBLOIS *et al.* (2006), for retrieving signals of population declines from our simulated data sets (32 significant tests of 60). The rate of detection was higher for ancient and moderate-to-severe declines, while recent and weak declines were barely detected. These results are consistent with previous simulation-based studies that have shown that the *M*-ratio test has low statistical power for small θ_1 values (here, $\theta_1 = 4$, see GARZA and WILLIAMSON 2001; WILLIAMSON-NATESAN 2005) and for recent population declines (see WILLIAMSON-NATESAN 2005; LEBLOIS *et al.* 2006). Here, we applied the *M*-ratio test by comparing the statistic *M* estimated from the data with the expected distribution of that statistic, conditionally on the true value of θ_1. This procedure increased the statistical power of the *M*-ratio test. Since the true value of the parameter θ_1 is generally unknown in real situations, GARZA and WILLIAMSON (2001) recommended the use of $M_c = 0.68$ as a conservative threshold for the critical value. The reanalysis of our data sets with $M_c = 0.68$ (*e.g.*, as in LEBLOIS *et al.* 2006) resulted in a lower rate of detection (22 significant tests of 60), but in similar qualitative trends (higher rate of detection for severe and ancient population declines).

Given the set of demographic and mutational parameters used in our study, and using the decision criteria recommended by the developers of each method, Msvar clearly outperformed the *M*-ratio test and Bottleneck for detecting population size change. While Msvar correctly detected 68.3% of the declines, the *M*-ratio test and Bottleneck detected only 53.3% and 8.3%, respectively, of the declines. Any population decline detected by the *M*-ratio test and Bottleneck was also recovered by Msvar, apart from one case of weak recent decline that was identified only by Bottleneck ($T_a = 10$ and $N_0/N_1 = 0.1$). Therefore, our study does not support the previous claims that the *M*-ratio test and Bottleneck are best suited to detect recent population declines, whereas Msvar is more appropriate to detect ancient contractions (GARZA and WILLIAMSON 2001; WILLIAMSON-NATESAN 2005). Moreover, while Msvar detected 73.3% of the population expansions, Bottleneck detected only 58.3% of the expansions. Any expansion detected by Bottleneck was also recovered by Msvar.

Performance of Msvar: What does coalescent theory tell us? Not surprisingly, we found that the performance of Msvar to infer past demography strongly depended

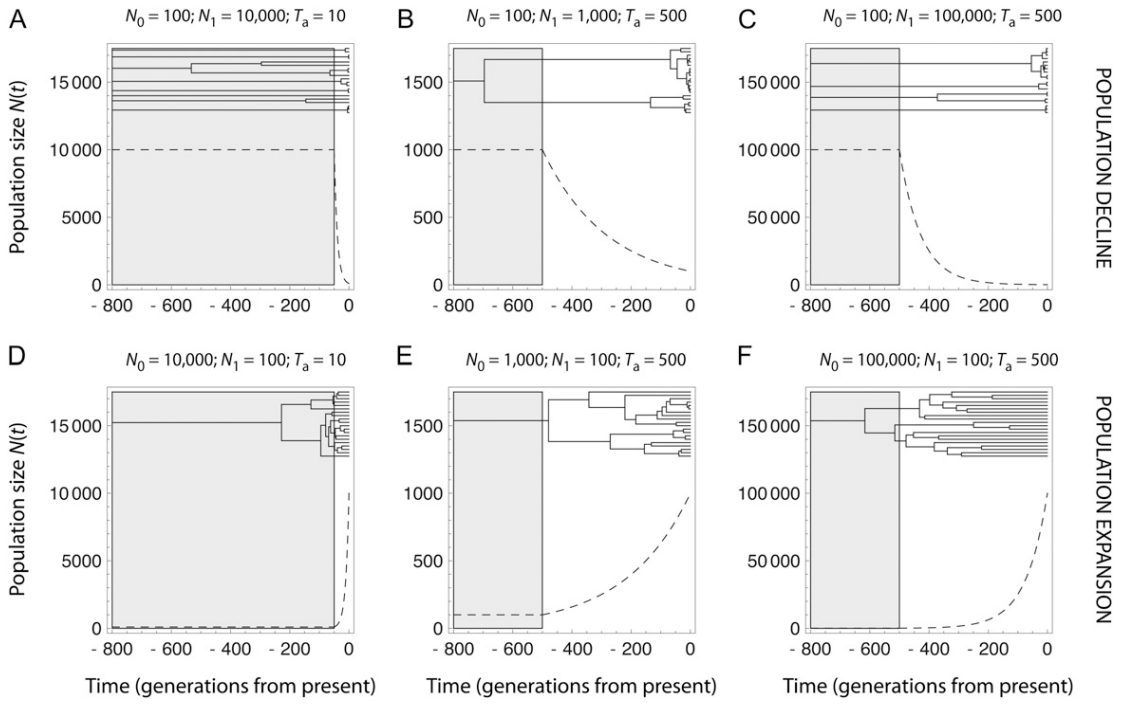


FIGURE 7.—Dynamics of population size changes $N(t)$ corresponding to simulated scenarios and expected gene genealogies. A–C corresponds to population decline and D–F to population expansion (dashed curve). The shaded area in each graph indicates when the ancestral population size is constant; *i.e.*, $N(t) = N_1$. Above each curve, the expected gene genealogy for 20 sampled lineages is represented. Expected gene genealogies were obtained by averaging coalescence times over 500,000 simulations of each demographic scenario. The simulations were based on a generation-by-generation coalescent algorithm developed by the authors. Note that some genealogies are incomplete (A and C), some lineages having not coalesced 800 generations from present.

on the information available in the data, which may be inferred from coalescent theory. Coalescent theory indeed predicts that variations in population size strongly affect the shape of gene genealogies, which are star shaped with long terminal branches in expanding populations and shallower in declining populations (Figure 7 and HEIN *et al.* 2005).

Coalescent theory further shows that only scaled parameters can be directly estimated from the data (TAVARÉ *et al.* 1997; NORDBORG 2007). Indeed, all parameters in coalescent models are scaled, and the likelihood function in Msvar makes no exception (BEAUMONT 1999). Hence, inference of unscaled quantities such as population size, or time measured in generations, requires external information. In our study, unscaled parameters were therefore much less precisely estimated than the scaled ones (Figure 5) and were also highly correlated (Figure S5; see also Figure 5 in STORZ *et al.* 2002). We deliberately chose poorly informative priors, to test the capacity of Msvar to retrieve information from the data only. In empirical studies, more informative priors of the natural parameters are usually specified. We acknowledge that Msvar offers a principled approach for providing prior information on the mutation rate, to recover posterior densities for natural parameters. Yet it should be borne in mind that precise estimates of unscaled parameters may then largely stem from the specification of the priors. Imagine that analyses were performed using a prior distribution for the mutation rate with very low standard deviation (*i.e.*, σ_μ close to zero). We would then necessarily recover the same level of precision for the natural parameters and the scaled parameters. Yet this improved precision may come at the expense of accuracy, if the prior distribution for the mutation rate departs from its true distribution.

Scenarios of population decline: Msvar was very efficient for detecting population declines. However, its performance for detecting change in population size and accurately estimating the model parameters was lowest for recent events ($T_a = 10$) of low-to-moderate severity ($N_0/N_1 \geq 0.01$), as well as for events of low severity ($N_0/N_1 = 0.1$). This is expected since, for very recent declines, the gene genealogy can barely be distinguished from that expected in a stable population with population size N_1 (Figure 7A). Interestingly, for $N_0/N_1 = 0.1$, the performance of Msvar was maximized for intermediate values of T_a, particularly with respect to the precision of θ_1 estimates. This might be easily understood by considering that for ancient events ($T_a = 500$) most coalescence events occur while $N(t)$ is close to the current population size (see, *e.g.*, Figure 7B). This might be further quantified by calculating the expected number j of ancestral lineages at (scaled) time τ, from which a sample of n genes is descending. This

number has a known distribution in a constant-size population (Tavaré 1984), and Leblois and Slatkin (2007) extended Tavaré's (1984) formula in the case of an exponentially growing or declining population. Using their Equation 2 that gives an expression for the probability $\Pr(j|n)$ that a sample of n genes has j ancestors T_a generations ago, we may compute the expected number m of lineages at the time of the population size change as

$$m = \sum_{j=1}^{n} j \Pr(j|n) = \sum_{j=1}^{n} j \sum_{i=j}^{n} \frac{(2i-1)(-1)^{i-j} j_{(i-1)} n_{[i]}}{j!(i-j)! n_{(i)}} e^{-i(i-1)\tau/2}, \quad (1)$$

where $a_{(i)} \equiv a(a+1)\ldots(a+i-1)$, $a_{[i]} \equiv a(a-1)\ldots(a-i+1)$, and $\tau = \int_0^{T_a} (dt'/2N(t')) = ((1-N_0/N_1)/N_0 \operatorname{Log}(N_1/N_0))T_a$. For a declining population with $T_a = 500$ and $N_0/N_1 = 0.1$, we get $m = 1.43$, which confirms that most coalescence events are expected to occur in the current population with this set of parameter values.

For moderate to severe contractions ($N_0/N_1 \leq 0.1$), both the bias and the 90% HPD range of θ_0 decreased with increasing T_a. Using Equation 1, we found that the expected number m of lineages at the time of the event varies between 48.49 and 2.20 for T_a varying from 10 to 500 and for $N_0/N_1 = 0.01$. This indicates that more coalescence events are expected to occur in the declining population when the event is older (see also Figure 7C). In contrast, θ_1 was overall precisely estimated (see Figure 4 and Figure S3). This is so because, for the scenarios considered here, a large part of the genealogy depends upon the ancestral history, with several lineages coalescing in the ancestral population (see, *e.g.*, Figure 7, B and C) at a rate that depends upon θ_1. Had we considered older events ($T_a > 500$), though, thereby decreasing the number of lineages in the ancestral population, it is likely that the precision of θ_1 estimates would have declined.

In summary, most scenarios of population decline result in gene genealogies with large times to the most recent common ancestor (TMRCAs). With the set of model parameters considered here, since a large part of gene genealogies depends upon θ_1, this latter parameter is generally precisely and accurately estimated. Contrastingly, θ_0 can be precisely and accurately estimated only if the demographic event is severe and ancient. If the change in population size is too recent, provided that it is not too pronounced, θ_0 estimates tend to converge to the true value of θ_1, and no change of population size is detected. If the difference in population size is weak, then the difference in coalescence rates before and after the event is not sufficient for Msvar to detect a population size change and to provide precise estimates of θ_0 and θ_1.

Scenarios of population expansion: Msvar was also very efficient for detecting expansions. Nevertheless, the estimates of the scaled current population size θ_0 were more severely biased and less precise, compared to scenarios of population decline, for the same relative severity of the event. This may be explained by the fact that expanding populations result in young genealogies with short TMRCAs (compare Figure 7, A–C, to 7, D–F) and hence rare mutation events. We found that the absolute value of the bias increased with N_0. We further found that both the 90% HPD range and the absolute value of the bias of θ_0 decreased with increasing T_a (Figure 4 and Figure S4). Using Equation 1, we found that the expected number m of lineages at the time of the event varies between 48.49 and 2.20 for T_a varying from 10 to 500 and for $N_0/N_1 = 100$. This indicates that the number of coalescence events in the expanding population is expected to increase as T_a increases (see also Figure 7F). More generally, the likelihood surface for expanding populations is complex (Beaumont 1999). In particular, as the genealogies become more star shaped, the joint posterior distribution of θ_0 and t_f reduces to a ridge along a line $\log_{10}(2\mu T_a) = k$ independent of θ_0 (Figure 4b in Beaumont 1999), which suggests that Msvar provides information on $2\mu T_a$, rather than on θ_0 in expanding populations. It is worth noting that Wakeley *et al.* (2001) found similar results in inferring demographic history from single-nucleotide polymorphims (see a comparison in Beaumont 2004).

Estimates of θ_1 had a low bias but a large 90% HPD range (Figure 4). Although the marginal posterior distributions of θ_1 were generally peaked, they were flat tailed on the left. This is so, because large ancestral population sizes are not compatible with the low polymorphism observed in the data. Instead, a large range of small values of θ_1 may be equally likely, provided the genealogy is star shaped.

Influence of the underlying demographic model: Athough Msvar equally detected population declines and expansions, inferences of the demographic parameters were in general more accurate for declines than for expansions. In addition to the above argument from coalescent theory, the exponential model assumed for population size change may partly explain this pattern. For declines, the size of the declining population $N(t)$ decreases sharply at T_a and converges rapidly to N_0 (Figure 7, A–C). Therefore, most coalescence events occurring in the declining population take place while $N(t)$ is close to N_0. For expansions, instead, the size of the expanding population $N(t)$ increases smoothly at T_a before it converges rapidly to N_0 (Figure 7, D–F). Therefore, a large proportion of the coalescence events that occur in the growing population take place while $N(t)$ is close to N_1 (compare, *e.g.*, Figure 7C to 7F). This can be expressed more formally by considering the harmonic mean of population sizes, which provides the

coalescent rate during the change in population size (Hein *et al.* 2005). For, say, $T_a = 500$, the harmonic mean of an exponentially declining population with $N_0 = 100$ and $N_1 = 10{,}000$ is 464, which is strictly equal to the harmonic mean of an exponentially growing population with $N_0 = 10{,}000$ and $N_1 = 100$. Hence, the harmonic mean of a declining population is closer to its current size (N_0) than its ancestral size (N_1), while the reverse is true for expanding populations. Therefore, given the exponential model of population growth, one might expect poor statistical properties of θ_0 estimates in expanding populations, compared to declining populations.

Robustness of Msvar to the misspecification of the mutation model: Most importantly, our results suggest that Msvar is robust to moderate departures from a strict SMM, *e.g.*, a GSM with $p \leq 0.22$, typical of those observed in the literature (see Figure 6 and Figure S6). However, severe departures from a strict SMM (here, a GSM with $p = 0.74$) led Msvar to detect a signal of population decline with strong support (BF ≥ 10), even in expanding populations (Figure 6 and Figure S6). This is not surprising since it has been recognized that violation of the assumptions of the SMM might induce severe bias in the inference of demographic history (Gonser *et al.* 2000). Indeed, mutations that arise under a strong GSM involve large changes in allele length, which produce some gaps in the distribution of allele types. The large resulting variance of allele range V_a is reminiscent of that observed with population decline (Storz and Beaumont 2002), even in expanding populations (compare Table 1 to Table S3).

Insights from empirical studies: The better performance of Msvar compared to the *M*-ratio test and Bottleneck also emerged from the empirical studies that inferred past demographic changes from microsatellite data using Msvar and at least one of the *M*-ratio or Bottleneck methods (Table S1). We found indeed that Msvar detected a population decline whenever one of the moment-based methods provided a significant test. By contrast, a large number of population declines that were not detected with any of the moment-based methods were detected with Msvar. Unfortunately, the scarcity of expansion events detected in the literature (Table S1) prevented any empirical comparison of Msvar and Bottleneck for growing populations. Importantly, the average genetic diversity measured from our simulated data sets was not substantially different from that observed in empirical studies (compare Table 1 to Table S1).

Because of the large heterogeneity of the published results, we did not attempt to analyze the quality of Msvar estimates in empirical studies. Some studies used Msvar 0.4 (Beaumont 1999), hence providing estimates for scaled parameters, and some studies used Msvar 1.3 (Storz and Beaumont 2002), hence providing estimates for unscaled parameters. Only a handful of studies used both methods, and few provided estimates of the scaled parameters using Msvar 1.3, as in the present study. Finally, credibility intervals were often not reported or calculated using different methods, which hampered any comparison among studies.

Recommendation guidelines and conclusions: Our simulation tests as well as an exhaustive survey of the literature clearly demonstrate that Msvar outperforms both the *M*-ratio test and Bottleneck for detecting population declines. Our study further shows that Msvar is also very efficient to detect population expansions and outperforms Bottleneck in that respect. However, to our knowledge, Msvar has only scarcely been applied on presumably expanding populations (see, *e.g.*, Hufbauer *et al.* 2004; Bonhomme *et al.* 2008; Wirth *et al.* 2008). Hence, we confidently recommend the use of Msvar for detecting past population size variation, even if this method is computationally demanding.

Most importantly, in contrast to the *M*-ratio test and Bottleneck, Msvar provides estimates of the parameters that characterize the population demographic history and the mutational model. Using Msvar 1.3 (Storz and Beaumont 2002), we have shown that the scaled parameters are more precisely estimated than the natural parameters. Although the latter are easier to interpret, our results clearly advocate drawing conclusions from inferences of θ_0, θ_1, and t_f. These parameters were precisely estimated for population declines, provided that the change in population size was neither too recent nor too weak, given the scenarios considered in our study. For expansions instead, both unscaled and scaled parameters were poorly estimated, although the method was efficient for detecting increase in population size. Hence, our results suggest that Msvar estimates in presumably expanding populations should be taken cautiously. We did not compare the performance of Msvar 0.4 (Beaumont 1999), which provides estimates for scaled parameters, to that of Msvar 1.3 (Storz and Beaumont 2002), which provides estimates for unscaled parameters, since the two versions differ by a number of other aspects. Msvar 0.4 assumes a basic (nonhierarchical) model, where the parameters are not allowed to vary among loci. The parameterization of Msvar 1.3 by means of a hierarchical model allows for some variation of the parameters among loci, which may provide a means of identifying aberrant loci. However, with broad priors on interlocus variation of the model parameters (**V**), Msvar 1.3 does not fully "borrow strength" from the different loci, *e.g.*, by simply pooling information (multiplying the likelihoods) across loci (Beaumont and Rannala 2004). This generally results in broader posterior distributions in the hierarchical Msvar 1.3 model, compared to the basic Msvar 0.4 model. Hence, although our results suggest that the use of scaled parameters should be preferred, further analyses are required to compare the performances of Msvar 0.4 and 1.3.

Finally, we recommend that inferences about population demographic change with Msvar should be interpreted cautiously in light of potential departures from the model assumptions. First, Msvar assumes that microsatellites evolve according to a strict SMM. Although a moderate departure from this mutation model, as classically measured with observations of spontaneous mutations (Ellegren 2000, 2004), does not seem to undermine Msvar performance (see Figure 6 and Figure S6), loci that evolve under a strong GSM may invalidate the approach by detecting false signals of population decline whatever the true demographic history. However, the hierarchical model implemented in Msvar 1.3 allows for variations in mutation and demographic parameters across loci and may thus limit the potential biases due to misspecifications of the mutation model. Indeed, Storz *et al.* (2002) argue that loci that strongly depart from the strict SMM shall be given less weight, thereby minimizing their impact on the inference made. Second, Msvar assumes that populations are isolated. Real populations, however, are in general connected by gene flow. It is now acknowledged that population structure and/or isolation by distance may result in incorrect inference of population demographic history (Pope *et al.* 2000; Leblois *et al.* 2006; Nielsen and Beaumont 2009; Chikhi *et al.* 2010; Peter *et al.* 2010). Finally, further work is needed to evaluate how Msvar performs when the demography is more complex, *e.g.*, with successions of population declines and expansions.

We are grateful to L. Chikhi for sharing an unpublished manuscript, as well as to Mark A. Beaumont and two anonymous reviewers for their constructive criticism of this article. This work was supported by the Plan Pluri-Formation "Evolution et Structure des Ecosystèmes" (2004–2008) from the Muséum National d'Histoire Naturelle (MNHN), by a Nouragues Research grant from the Centre National de la Recherche Scientifique (CNRS) programme Amazonie I (2007), by the CNRS programme Amazonie II (2008–2011), and by the Agence Nationale de la Recherche programme blanc Études de Méthodes Inférentielles et Logiciels pour l'Évolution (EMILE) NT09-611697. This work is part of Christophe Girod's Ph.D., who was supported by a grant from the French Ministry of Research (2007–2010). 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).

LITERATURE CITED

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

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

Beaumont, M. A., 2004 Recent developments in genetic data analysis: What can they tell us about human demographic history? Heredity **92:** 365–379.

Beaumont, M. A., and B. Rannala, 2004 The Bayesian revolution in genetics. Nat. Rev. Genet. **5:** 251–261.

Bonhomme, M., A. Blancher, S. Cuartero, L. Chikhi and B. Crouau-Roy, 2008 Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. Mol. Ecol. **17:** 1009–1019.

Brooks, S., and A. Gelman, 1998 General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7:** 434–455.

Chikhi, L., V. C. Sousa, P. Luisi, B. Goossens and M. A. Beaumont, 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, R. I., M. Manca, M. Viljanen, H. A. Ketelaars, H. Bürgi *et al.*, 2005 Invasion genetics of the Eurasian spiny waterflea: evidence for bottlenecks and gene flow using microsatellites. Mol. Ecol. **14:** 1869–1879.

Comps, B., D. Gömöry, J. Letouzey, B. Thiébaut and R. J. Petit, 2001 Diverging trends between heterozygosity and allelic richness during postglacial colonization in the European Beech. Genetics **157:** 389–397.

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

Cornuet, J.-M., F. Santos, M. A. Beaumont, C. P. Robert, J.-M. Marin *et al.*, 2008 Infering population history with DIY ABC: a user-friendly approach to Approximate Bayesian Computation. Bioinformatics **24:** 2713–2719.

Dib, C., S. Fauré, C. Fizames, D. Samson, N. Drouot *et al.*, 1996 A comprehensive map of the human genome based on 5,264 microsatellites. Nature **380:** 152–154.

Drummond, A., and A. Rambaut, 2007 BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7:** 214.

Ellegren, H., 2000 Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. **16:** 552–558.

Ellegren, H., 2004 Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. **5:** 435–445.

Elmer, K. R., C. Reggio, T. Wirth, E. Verheyen, W. Salzburger *et al.*, 2009 Pleistocene desiccation in East Africa bottlenecked but did not extirpate the adaptive radiation of Lake Victoria haplochromine cichlid fishes. Proc. Natl. Acad. Sci. USA **106:** 13404–13409.

Emerson, B. C., E. Paradis and C. Thebaud, 2001 Revealing the demographic histories of species using DNA sequences. Trends Ecol. Evol. **16:** 707–716.

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.

Excoffier, L., G. Laval and S. Schneider, 2005 Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinform. Online **1:** 47–50.

Felsenstein, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139–147.

FitzSimmons, N. N., 1998 Single paternity of clutches and sperm storage in the promiscuous green turtle (*Chelonia mydas*). Mol. Ecol. **7:** 575–584.

Frankham, R., D. A. Briscoe and J. D. Ballou, 2002 *Introduction to Conservation Genetics*. Cambridge University Press, New York.

Garza, J. C., and E. G. Williamson, 2001 Detection of reduction in population size using data from microsatellite loci. Mol. Ecol. **10:** 305–318.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis*. Chapman & Hall, London.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, 2004 *Bayesian Data Analysis*. Chapman & Hall/CRC, New York.

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

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. **344:** 403–410.

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

Hein, J., M. H. Schierup and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution. A Primer in Coalescent Theory*. Oxford University Press, Oxford.

Hey, J., and R. Nielsen, 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., and R. Nielsen, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Natl. Acad. Sci. USA **104:** 2785–2790.

Hu, F. S., A. Hampe and R. J. Petit, 2009 Paleoecology meets genetics: deciphering past vegetational dynamics. Front. Ecol. Environ. **7:** 371–379.

Hufbauer, R. A., S. M. Bogdanowicz and R. G. Harrison, 2004 The population genetics of a biological control introduction: mitochondrial DNA and microsatellite variation in native and introduced populations of *Aphidus ervi*, a parasitoid wasp. Mol. Ecol. **13:** 337–348.

Jacobsen, B. H., M. M. Hansen and V. Loeschcke, 2005 Microsatellite DNA analysis of northern pike (*Esox lucius* L.) populations: insights into the genetic structure and demographic history of a genetically depauperate species. Biol. J. Linn. Soc. **84:** 91–101.

Jeffreys, H., 1961 *Theory of Probability*, Ed. 3. Oxford University Press, Oxford.

Kass, R. E., and A. E. Raftery, 1995 Bayes factors. J. Am. Stat. Assoc. **90:** 773–795.

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

Laval, G., and L. Excoffier, 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.

Lawton-Rauh, A., 2008 Demographic processes shaping genetic variation. Curr. Opin. Plant Biol. **11:** 103–109.

Leblois, R., and M. Slatkin, 2007 Estimating the number of founder lineages from haplotypes of closely linked SNPs. Mol. Ecol. **16:** 2237–2245.

Leblois, R., A. Estoup and R. Streiff, 2006 Genetics of recent habitat contraction and reduction in population size: Does isolation by distance matter? Mol. Ecol. **15:** 3601–3615.

Loader, C., 1999 *Local Regression and Likelihood*. Springer-Verlag, New York.

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.

Luikart, G., F. W. Allendorf, J.-M. Cornuet and W. B. Sherwin, 1998 Distortion of allele frequency distribution provides a test for recent population bottlenecks. J. Hered. **89:** 238–247.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1092.

Milton, K., J. D. Lozier and E. A. Lacey, 2009 Genetic structure of an isolated population of mantled howler monkeys (*Alouatta palliata*) on Barro Colorado Island, Panama. Conserv. Genet. **10:** 347–358.

Nei, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89:** 583–590.

Nettel, A., R. S. Dodd and Z. Afzal-Rafii, 2009 Genetic diversity, structure, and demographic change in tanoak, *Lithocarpus densiflorus* (Fagaceae), the most susceptible species to sudden oak death in California. Am. J. Bot. **96:** 2224–2233.

Nielsen, R., and M. Beaumont, 2009 Statistical inferences in phylogeography. Mol. Ecol. **18:** 1034–1047.

Nordborg, M., 2007 Coalescent theory, pp. 843–877 in *Handbook of Statistical Genetics*, Ed. 3, edited by D. Balding, M. Bishop and C. Cannings. John Wiley & Sons, Chichester, UK.

Peter, B. M., D. Wegmann and L. Excoffier, 2010 Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. Mol. Ecol. **19:** 4648–4660.

Plummer, M., N. Best, K. Cowles and K. Vines, 2006 Coda: output analysis and diagnostics for MCMC. R News **6:** 7–11.

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.

R Development Core Team, 2009 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rousset, F., and R. Leblois, 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.

Schneider, S., and L. Excoffier, 1999 Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics **152:** 1079–1089.

Spencer, C. C., J. E. Neigel and P. L. Leberg, 2000 Experimental evaluation of the usefulness of microsatellite DNA for detecting demographic bottlenecks. Mol. Ecol. **9:** 1517–1528.

Storz, J. F., and M. A. Beaumont, 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.

Storz, J. F., M. A. Beaumont and S. C. Alberts, 2002 Genetic evidence for long-term population decline in a savannah-dwelling primate: inferences from a hierarchical Bayesian model. Mol. Biol. Evol. **19:** 1981–1990.

Strasburg, J. L., and L. H. Rieseberg, 2010 How robust are ''isolation with migration'' analyses to violations of the IM model? A simulation study. Mol. Biol. Evol. **27:** 297–310.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–165.

Tavaré, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.

Wakeley, J., R. Nielsen, S. N. Liu-Cordero and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. Am. J. Hum. Genet. **69:** 1332–1347.

Williamson-Natesan, E. G., 2005 Comparison of methods for detecting bottlenecks from microsatellite loci. Conserv. Genet. **6:** 551–562.

Wilson, I. J., and D. J. Balding, 1998 Genealogical inference from microsatellite data. Genetics **150:** 499–510.

Wilson, I. J., M. Weale and D. J. Balding, 2003 Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. Ser. A **166:** 155–188.

Wirth, T., F. Hildebrand, C. Allix-Béguec, F. Wölbeling, T. Kubica *et al.*, 2008 Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathog. **4:** e1000160.

Communicating editor: N. Takahata

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.121764/DC1

Inferring Population Decline and Expansion From Microsatellite Data: A Simulation-Based Evaluation of the Msvar Method

Christophe Girod, Renaud Vitalis, Raphaël Leblois and Hélène Fréville

DOI: 10.1534/genetics.110.121764

FILE S1

Supporting Methods

The version 1.3 of MSVAR provides separate estimates for N_0, N_1, μ and T_a (STORZ and BEAUMONT 2002). This implementation of BEAUMONT's (1999) method, available at http://www.rubic.rdg.ac.uk/~mab/stuff/, relies upon a hierarchical model where demographic and mutational parameters are allowed to vary among loci. Hence, the set of parameters of interest is given by $\mathbf{\Phi} \equiv \{N_{0i}, N_{1i}, T_{ai}, \mu_i\}_{i=1,k}$ for k loci and the prior distributions of these parameters depend upon hyper-prior distributions. Priors and hyper-priors were specified following STORZ and BEAUMONT (2002). For each locus, the prior distributions of the model parameters were assumed to be log-normal distributions, each with means (on a $\log_{10}$ scale) of $\mathbf{M} \equiv \{M_{N_0}, M_{N_1}, M_{T_a}, M_\mu\}$ and standard deviations (SDs) of $\mathbf{V} \equiv \{V_{N_0}, V_{N_1}, V_{T_a}, V_\mu\}$. Hyper-prior distributions for the means $\mathbf{M}$ were themselves assumed to be normal distributions with means α_{N_0}, α_{N_1}, α_{T_a}, α_μ and SDs σ_{N_0}, σ_{N_1}, σ_{T_a}, σ_μ. Prior means for the current and ancestral population sizes were set equal to the $\log_{10}$-transformed values $\alpha_{N_0} = \alpha_{N_1} = 3$, which amounts to consider population contraction and expansion as equally likely. To allow for uncertainty in these estimates, and to avoid a strong effect of the prior specification, we considered prior distribution with large SDs by setting $\sigma_{N_0} = \sigma_{N_1} = 4$. The prior mean for the time since the population started changing in size was set equal to 500 generations, giving $\alpha_{T_a} = 2.7$. The SD of the mean around the time of demographic change, $\sigma_{T_a} = 3$, was chosen so that more recent and more ancient dates were also supported. In order to test whether the method could retrieve information on the mutation rate from the data, we chose a relatively flat prior on the mean mutation rate per generation with $\alpha_\mu = -4$ and $\sigma_\mu = 2$. Last, the hyper-prior distributions for the SDs $\mathbf{V}$ were assumed to be normal distributions truncated at zero with means $\beta_{N_0} = \beta_{N_1} = \beta_{T_a} = \beta_\mu = 0$ and SDs $\tau_{N_0} = \tau_{N_1} = \tau_{T_a} = \tau_\mu = 0.5$. The parameter values reported in the main text correspond to the mean of N_0, N_1, T_a and μ, across loci.

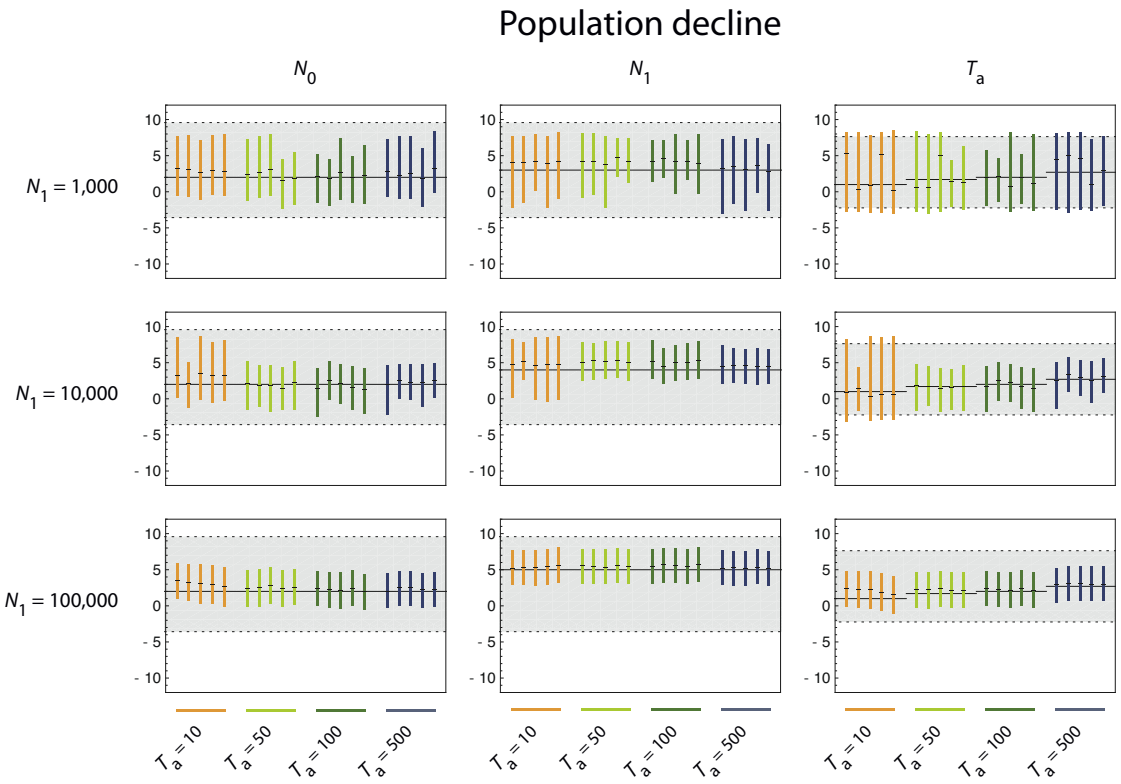


FIGURE S1.—Inference of the natural demographic parameters N_0, N_1 and T_a in declining populations. In each graph, the coloured vertical lines represent the 90% HPD interval in a $\log_{10}$ scale. The black horizontal trait over each coloured line represents the mode of the marginal posterior distribution of the parameters. Each colour stands for a value of the (unscaled) time parameter T_a (orange for $T_a = 10$, light green for $T_a = 50$, dark green for $T_a = 100$ and blue for $T_a = 500$). The grey area within the dotted lines in each graph represents the 90% support of the prior distribution of the parameters. The black continuous horizontal line gives the true value of the parameters.

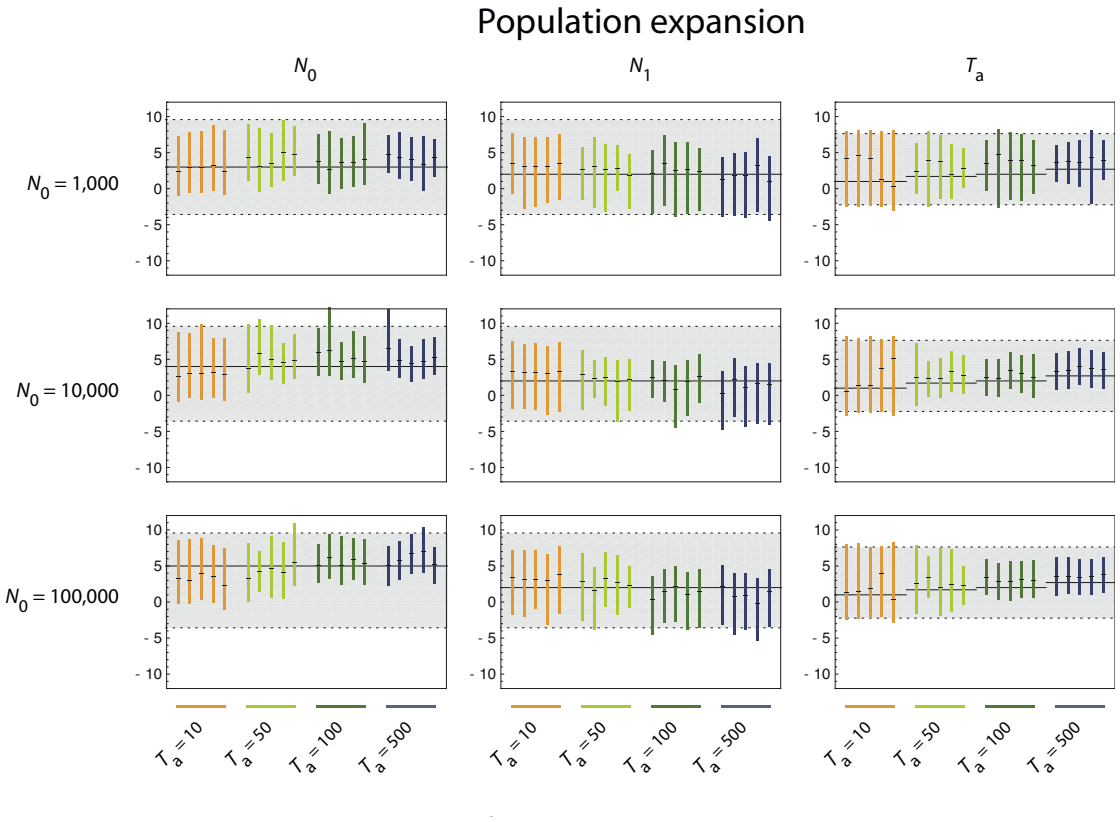


FIGURE S2.—Inference of the natural demographic parameters N_0, N_1 and T_a in expanding populations. In each graph, the coloured vertical lines represent the 90% HPD interval in a $\log_{10}$ scale. The black horizontal trait over each coloured line represents the mode of the marginal posterior distribution of the parameters. Each colour stands for a value of the (unscaled) time parameter T_a (orange for T_a = 10, light green for T_a = 50, dark green for T_a = 100 and blue for T_a = 500). The grey area within the dotted lines in each graph represents the 90% support of the prior distribution of the parameters. The black continuous horizontal line gives the true value of the parameters.

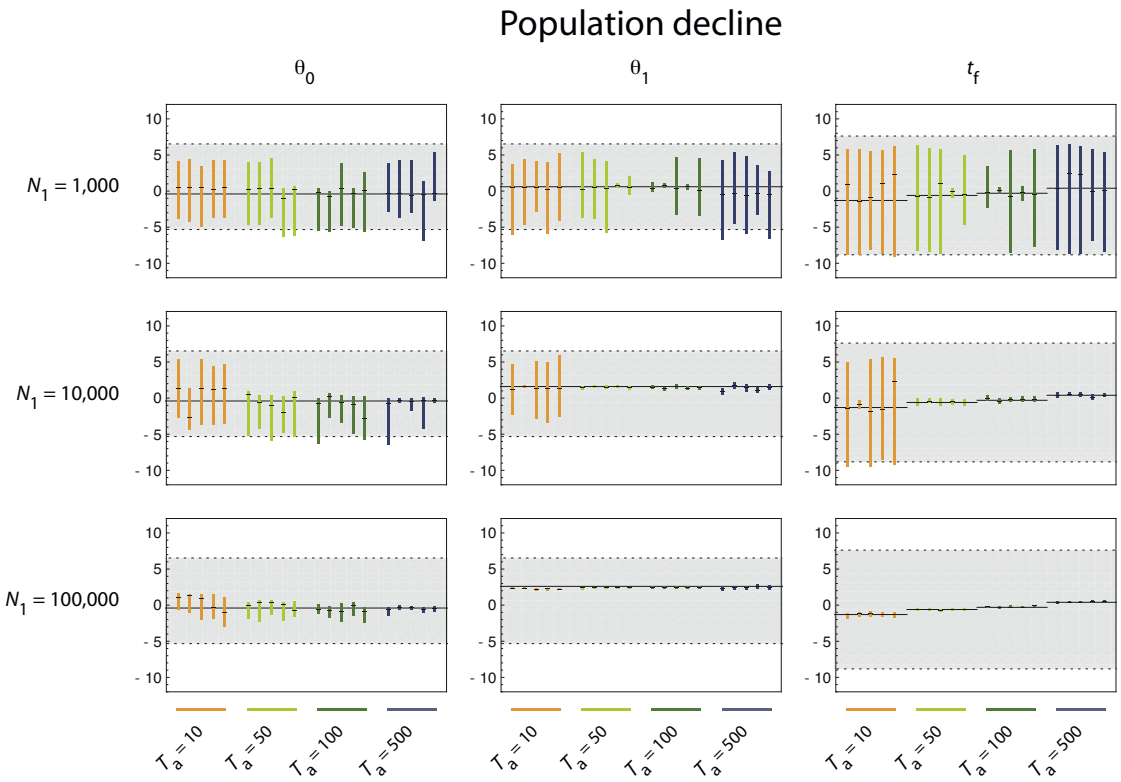


FIGURE S3.—Inference of the scaled demographic parameters θ_0, θ_1 and t_f in declining populations. In each graph, the coloured vertical lines represent the 90% HPD interval in a $\log_{10}$ scale. The black horizontal trait over each coloured line represents the mode of the marginal posterior distribution of the parameters. Each colour stands for a value of the (unscaled) time parameter T_a (orange for $T_a = 10$, light green for $T_a = 50$, dark green for $T_a = 100$ and blue for $T_a = 500$). The grey area within the dotted lines in each graph represents the 90% support of the prior distribution of the parameters. The black continuous horizontal line gives the true value of the parameters.

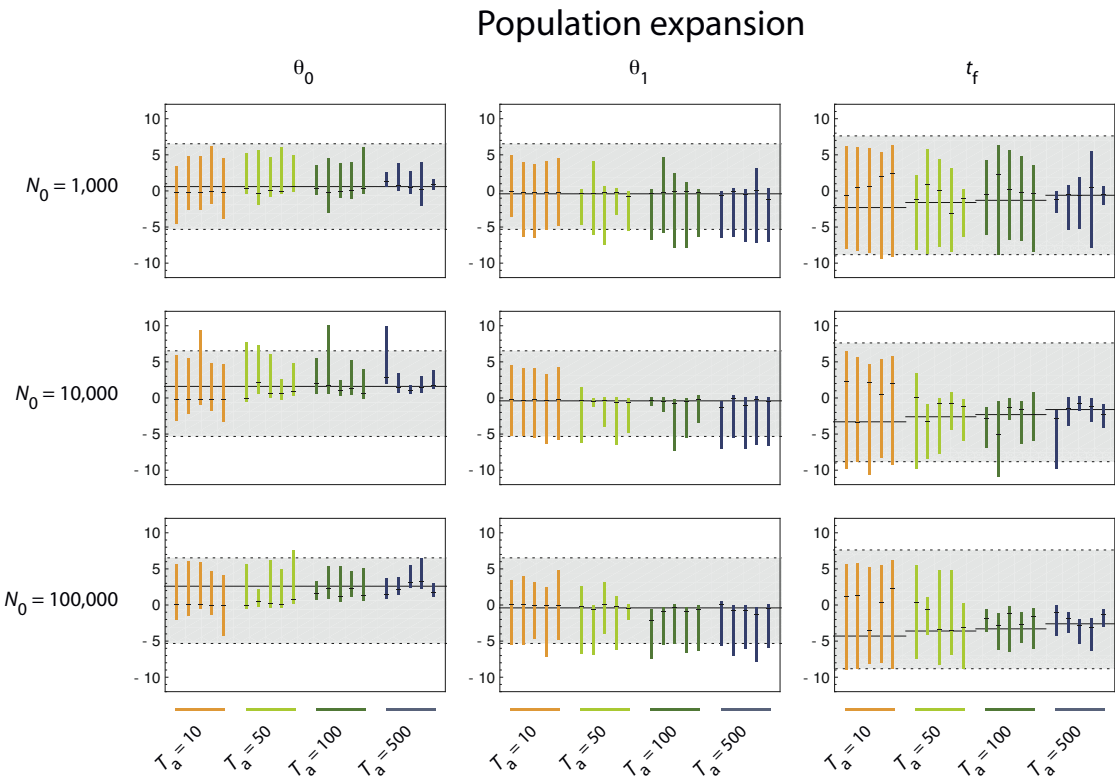


FIGURE S4.—Inference of the scaled demographic parameters θ_0, θ_1 and t_f in expanding populations. In each graph, the coloured vertical lines represent the 90% HPD interval in a $\log_{10}$ scale. The black horizontal trait over each coloured line represents the mode of the marginal posterior distribution of the parameters. Each colour stands for a value of the (unscaled) time parameter T_a (orange for $T_a = 10$, light green for $T_a = 50$, dark green for $T_a = 100$ and blue for $T_a = 500$). The grey area within the dotted lines in each graph represents the 90% support of the prior distribution of the parameters. The black continuous horizontal line gives the true value of the parameters.

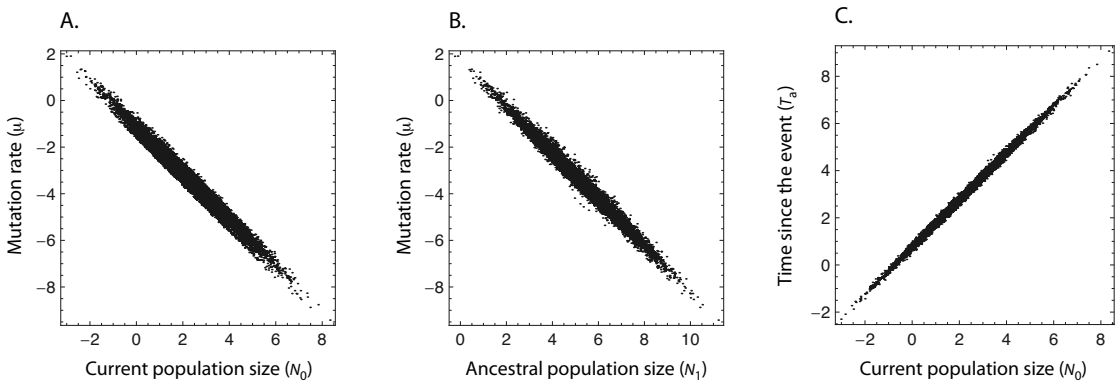


FIGURE S5.—Correlation between natural parameters: (A) current population size N_0 and mutation rate μ, (B) ancestral population size N_1 and mutation rate μ and (C) current population size N_0 and time since event T_a.

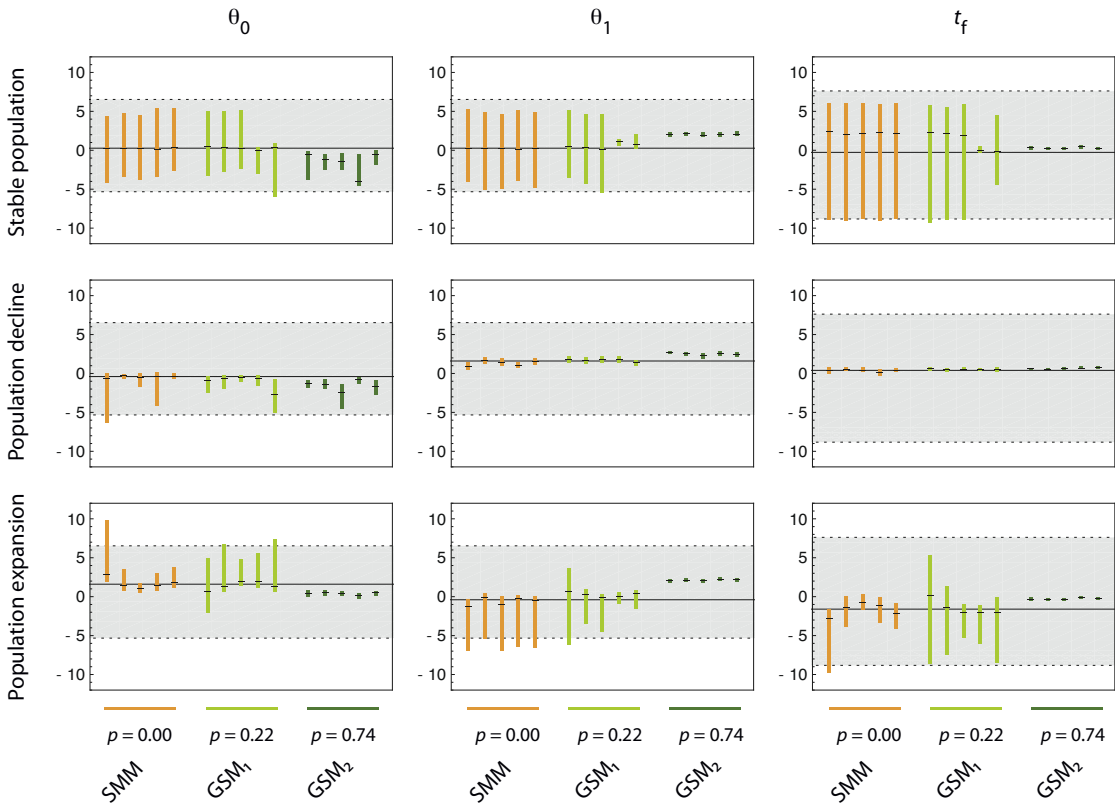


FIGURE S6.—Inference of the scaled demographic parameters θ_0, θ_1 and t_f in a stable (with $N_0 = N_1 = 464$, $T_a = 500$), a declining (with $N_0 = 100$, $N_1 = 10{,}000$ and $T_a = 500$), and an expanding (with $N_0 = 10{,}000$, $N_1 = 100$ and $T_a = 500$) populations. In each graph, the coloured vertical lines represent the 90% HPD interval in a $\log_{10}$ scale. The black horizontal trait over each coloured line represents the mode of the marginal posterior distribution of the parameters. Each colour stands for a value of p, the frequency of multi-step mutations changes: orange for $p = 0.00$ (stepwise mutation model, SMM), light green for $p = 0.22$ (moderate generalized stepwise model, GSM_1), dark green for $p = 0.74$ (strong generalized stepwise model, GSM_2). The grey area within the dotted lines in each graph represents the 90% support of the prior distribution of the parameters. The black continuous horizontal line gives the true value of the parameters.

TABLE S1

Exhaustive Review of the Studies that used MSVAR for Inference of Population Size Change

					MSVAR				
Species	*s*	*n*	No. of loci	H_e	Vers.	Res.	*M*-ratio	Bottleneck	Reference
Mexican goodeid fish	2	20 – 31	5 – 7	-	v. 0.4	↘	-	-	BAILEY *et al.* (2007)
Himalayan brown bear	1	54	6	-	v. 0.4	↘	-	-	BELLEMAIN *et al.* (2007)
Ground beetle	2	54 – 56	9	-	v. 0.4	↘	-	-	KELLER *et al.* (2005)
European grayling	14	13 – 48	8	0.62 – 0.69	v. 0.4	↘	-	-	MELDGAARD *et al.* (2003)
	4	28 – 52	17	-	v. 0.4	↘	-	-	KOSKINEN *et al.* (2002a)
Arctic grayling	1	71	-	-	v. 0.4	↘	-	-	KOSKINEN *et al.* (2002b)
	6	80 – 214	7	-	v. 0.4	↘	-	-	STAMFORD and TAYLOR (2005)
Northern pike	16	15 – 50	5	-	v. 0.4	↘	-	-	JACOBSEN et al. (2005)
	2	26	5	-		↗			
Japanese eel	1	89	6	-	v. 0.4	↘	-	-	TSENG *et al.* (2003)
American and European eels	4	100	7	-	v. 0.4	↘	-	-	WIRTH and BERNATCHEZ (2003)
Fishtail palm	1	143	9	0.67	v. 1.3	↘	-	-	CIBRIAN-JARAMILLO *et al.* (2009)
Lake Victoria Cichlid fishes	3	-	12	-	v. 0.4	↘	-	-	ELMER *et al.* (2009)
Cape Fear Shiner	2	26 – 29	18	-	v. 1.3	↘	-	-	SAILLANT *et al.* (2004)
Gray snapper	3	50	13	-	v. 1.3	↘	-	-	GOLD *et al.* (2009)
Lane snapper	6	50	13	-	v. 1.3	↘	-	-	KARLSSON *et al.* (2009)
Drosophila sp.	12	22 – 30	47	0.10 – 0.51	v. 0.4	↘	-	-	HARR AND SCHLÖTTERER (2004)
Caenorhabditis elegans	1	-	9	-	v. 0.4	↘	-	-	SIVASUNDAR AND HEY (2003)
Red deer	3	31 – 33	11	0.37 – 0.50	v. 1.3	↘	-	-	NIELSEN ET AL. (2008)
Mycobacterium complex	8	-	24	-	v. 0.4	↗	-	-	WIRTH ET AL. (2008)

Sea otter	1	40	24	-	v. 1.3	↘	- a	- a	AGUILAR ET AL. (2008)
Bornean orang-utan	2	26 – 27	14	-	both	↘	-	- a	GOOSSENS ET AL. (2006)
Tanoak	1	447	9	-	v. 0.4	↘	-	- a	NETTEL ET AL. (2009)
Rattlesnake	5	18 – 54	9	0.61 – 0.73	v. 0.4	↘	↘	-	HOLYCROSS ET AL. (2007)
Parasitoid wasp		50	5	-	v. 1.3	↘	-	- a	HUFBAUER ET AL. (2004)
		48	5	-		↗	-	- a	
Cynomolgus macaque	1	81	16	0.66	v. 0.4	=	-	↗	BONHOMME ET AL. (2008)
Madagascar fish-eagle	1	44	22	0.19	v. 0.4	↘	-	=	JOHNSON ET AL. (2009)
African elephant	5	80 – 319	20	-	v. 1.3	↘	-	↘	OKELLO ET AL. (2008)
	1	79	20	0.74		↘	-	=	
Giant panda	2	29 – 32	9	0.49 – 0.56	v. 1.3	↘	-	=	ZHANG ET AL. (2007)
	1	40	9	0.61		↘	-	↗	
Eurasian otter	2	65 – 132	10	0.53 – 0.59	v. 1.3	↘	-	=	HAJKOVA ET AL. (2007)
	2	29 – 58	11	-	v. 0.4	↘	-	=	PERTOLDI ET AL. (2001)
Ethiopian walia ibex	1	24	5	0.35	v. 1.3	↘	-	=	GEBREMEDHIN ET AL. (2009)
Persian wild ass	1	24	12	0.54	v. 1.3	↘	-	=	NIELSEN ET AL. (2007)
Rock ptarmigan	3	17 – 20	6	0.45 – 0.75	v. 1.3	↘	-	=	PRUETT *et al.* (2010)
Cyprinid fish	4	30 – 48	6	0.23 – 0.35	v. 1.3	↘	-	=	SOUSA *et al.* (2008)
	2	21 – 30	6	0.24 – 0.26		=	-	=	
	6	12 – 50	6	0.22 – 0.45	v. 1.3	↘	-	=	SOUSA *et al.* (2010)
Golden-brown mouse lemur	8	15 – 59	8	0.54 – 0.65	both	↘	-	=	OLIVIERI *et al.* (2008)
Bongolava	1	27	8	0.57	both	↘	-	=	

mouse lemur									
Danfoss' mouse lemur	1	30	8	0.71	both	↘	-	=	
Milne-Edwards's sportive lemur	2	10	14	0.53 – 0.55	both	↘	-	=	CRAUL *et al.* (2009)
Golden eagle	1	172	13	0.48	both	↘	-	=	BOURKE *et al.*(2010)
Yunnan sub-nosed monkey	1	135	10	0.70	v. 1.3	=	-	=	LIU *et al.* (2009)
Drosophila	1	13	17	0.80	v. 0.4	↘	-	↘	DIERINGER ET AL. (2005)
	4	10 – 30	17	0.47 – 0.79		↘	-	=	
	10	8 – 30	17	0.48 – 0.79		↘	-	↗	
	1	20	10	-	v. 0.4	↘	-	=	FRYDENBERG ET AL. (2002)
	1	20	10	-		↘	-	↗	
	1	20	10	-		↗	-	=	
	1	20	10	-		↗	-	↗	
	1	20	10	-		=	-	=	
Reed warblers	1	40	9	-	v. 1.3	↘	=	=	PROCHAZKA ET AL. (2008)
Tiger	1	57	30	-	both	↘	↘	↘	MONDOL ET AL. (2009)
European otter	1	6	11	0.77	v. 0.4	=	=	=	RANDI *et al.* (2003)
	7	3 – 29	11	0.45 – 0.74		↘	=	=	
European grayling	30	17 – 112	13 – 14	0.24 – 0.64	v. 1.3	↘	↘	=	SWATDIPONG *et al.* (2010)
	1	35	13	-	v. 1.3	↘	↘	↘	
Eastern red-blacked salamander	2	25 – 28	6	0.41 – 0.47	v. 0.4	↘	↘	=	JORDAN *et al.* (2009)
European wolf	2	30 – 103	18	-	v. 0.4	↘	↘	=	LUCCHINI *et al.* (2004)
	3	34 – 115	18	-		↘	=	=	
African buffalo	2	33 – 54	17	0.81 – 0.82	v. 1.3	↘	↘	=	HELLER *et al.* (2008)
Howler monkeys	1	50	10	0.58	v. 1.3	↘	=	↗	MILTON *et al.* (2009)

s, number of samples; *n*, number of individuals per sample; No. of loci, number of microsatellite loci used; H_e , expected heterozygosity; Vers., version of MSVAR used in the study (v. 0.4, v. 1.3 or both); Res., MSVAR results, as interpreted by the authors; ↘, evidence for population decline; =, no evidence of population size change; ↗, evidence of population expansion; -

, information not available; all studies using the *M*-ratio test calculated the critical value M_c (see main text), except that of MILTON *et al.* (2009), which considered $M_c = 0.68$; [a], data not available, since the *M*-ratio test and/or BOTTLENECK were not applied on the same set of populations as MSVAR. When multiple mutation models have been considered for the *M*-ratio test and/or BOTTLENECK, only the results assuming a SMM are reported.

Literature Cited

AGUILAR, A., D. A. JESSUP, J. ESTES and J. C. GARZA, 2008 The distribution of nuclear genetic variation and historical demography of sea otters. Animal Conservation **11:** 35-45.

BAILEY, N. W., C. MACIAS GARCIA and M. G. RITCHIE, 2007 Beyond the point of no return? A comparison of genetic diversity in captive and wild populations of two nearly extinct species of Goodeid fish reveals that one is inbred in the wild. Heredity **98:** 360-367.

BELLEMAIN, E., M. A. NAWAZ, A. VALENTINI, J. E. SWENSON and P. TABERLET, 2007 Genetic tracking of the brown bear in northern Pakistan and implications for conservation. Biological Conservation **134:** 537-547.

BONHOMME, M., A. BLANCHER, S. CUARTERO, L. CHIKHI and B. CROUAU-ROY, 2008 Origin and number of founders in an introduced insular primate: estimation from nuclear genetic data. Molecular Ecology **17:** 1009-1019.

BOURKE, B. P., A. C. FRANTZ, C. P. LAVERS, A. DAVISON, D. DAWSON *et al.*, 2010 Genetic signatures of population change in the British golden eagle (*Aquila chrysaetos*). Conservation Genetics **in press**.

CIBRIAN-JARAMILLO, A., C. D. BACON, N. C. GARWOOD, R. M. BATEMAN, M. M. THOMAS *et al.*, 2009 Population genetics of the understory fishtail palm *Chamaedorea ernesti-augusti* in Belize: high genetic connectivity with local differentiation. BMC Genetics **10:** 65-82.

CRAUL, M., L. CHIKHI, V. SOUSA, G. L. OLIVIERI, A. RABESANDRATANA *et al.*, 2009 Influence of forest fragmentation on an endangered large-bodied lemur in northwestern Madagascar. Biological Conservation **142:** 2862-2871.

DIERINGER, D., V. NOLTE and C. SCHLÖTTERER, 2005 Population structure in African *Drosophila melanogaster* revealed by microsatellite analysis. Molecular Ecology **14:** 563-573.

ELMER, K. R., C. REGGIO, T. WIRTH, E. VERHEYEN, W. SALZBURGER *et al.*, 2009 Pleistocene desiccation in East Africa bottlenecked but did not extirpate the adaptive radiation of Lake Victoria haplochromine cichlid fishes. Proceedings of the National Academy of Science **106:** 13404-13409.

FRYDENBERG, J., C. PERTOLDI, J. DAHLGAARD and V. LOESCHCKE, 2002 Genetic variation in original and colinizing *Drosophila buzzatii* populations analysed by microsatellite loci isolated with a new PCR screening method. Molecular Ecology **11:** 181-190.

GEBREMEDHIN, B., G. F. FICETOLA, S. NADERI, H.-R. REZAEI, C. MAUDET *et al.*, 2009 Combining genetic and ecological data to assess the conservation status of the endangered Ethiopian walia ibex. Animal Conservation **12:** 89-100.

GOLD, J. R., E. SAILLANT, N. D. EBELT and S. LEM, 2009 Conservation genetics of Gray Snapper (*Lutjanus griseus*) in U.S. Waters of the Northern Gulf of Mexico and Western Atlantic Ocean. Copeia **2:** 277-286.

GOOSSENS, B., L. CHIKHI, M. ANCRENAZ, I. LACKMAN-ANCRENAZ, P. ANDAU *et al.*, 2006 Genetic signature of anthropogenic population collapse in orang-utans. PLoS Biology **4:** 285-291.

HAJKOVA, P., C. PERTOLDI, B. ZEMANOVA, K. ROCHE, B. HAJEK *et al.*, 2007 Genetic structure and evidence for recent population decline in Eurasian otter populations in the Czech and Slovak Republics: implications for conservation. Journal of Zoology **272:** 1-9.

HARR, B., and C. SCHLÖTTERER, 2004 Patterns of microsatellite variability in the *Drosophila melanogaster* complex. Genetica **120:** 71-77.

HELLER, R., E. D. LORENZEN, J. B. A. OKELLO, C. MASEMBE and H. R. SIEGISMUND, 2008 Mid-Holocene decline in African buffalos inferred from Bayesian coalescent-based analyses of microsatellites and mitochondrial DNA. Molecular Ecology **17:** 4845-4858.

HOLYCROSS, A. T., and M. E. DOUGLAS, 2007 Geographic isolation, genetic divergence and ecologicalnon-exchangeability define ESUs in a threatened sky-island rattlesnake. Biological conservation **134:** 142-154.

HUFBAUER, R. A., S. M. BOGDANOWICZ and R. G. HARRISON, 2004 The population genetics of a biological control introduction: mitochondrial DNA and microsatellite variation in native and introduced populations of *Aphidus ervi*, a parasitoid wasp. Molecular Ecology **13:** 337-348.

JACOBSEN, B. H., M. M. HANSEN and V. LOESCHCKE, 2005 Microsatellite DNA analysis of northern pike (*Esox lucius* L.) populations: insights into the genetic structure and demographic history of a genetically depauperate species. Biological Journal of the Linnean Society **84:** 91-101.

JOHNSON, J. A., R. E. TINGAY, M. CULVER, F. HAILER, M. L. CLARKE *et al.*, 2009 Long-term survival despite low genetic diversity in the critically endangered Madagascar fish-eagle. Molecular Ecology **18:** 54-63.

JORDAN, M. E., D. A. MORRIS and S. E. GIBSON, 2009 The influence of historical landscape change on genetic variation and population structure of a terrestrial salamander (Plethodon cinereus). Conservation Genetics **10:** 1647-1658.

KARLSSON, S., E. SAILLANT and J. R. GOLD, 2009 Population structure and genetic variation of lane snapper (*Lutjanus synagris*) in the northern Gulf of Mexico. Marine Biology **156:** 1841-1855.

KELLER, I., L. EXCOFFIER and C. R. LARGIADER, 2005 Estimation of effective population size and detection of a recent population decline coinciding with habitat fragmentation in a ground beetle. Journal of Evolutionary Biology **18:** 90-100.

KOSKINEN, M. T., T. O. HAUGEN and C. R. PRIMMER, 2002a Contemporary fisherian life-history evolution in small salmonid populations. Nature **419:** 826-830.

KOSKINEN, M. T., I. KNIZHIN, C. R. PRIMMER, C. SCHLÖTTERER and S. WEISS, 2002b Mitochondrial and nuclear DNA phylogeography of *Thymallus* spp. (grayling). Molecular Ecology **11:** 2599-2611.

LIU, Z., B. REN, R. WU, L. ZHAO, Y. HAO *et al.*, 2009 The effect of landscape features on population genetic structure in Yunnan snub-nosed monkeys (*Rhinopithecus bieti*) implies an anthropogenic genetic discontinuity. Molecular Ecology **18:** 3831-3846.

LUCCHINI, V., A. GALOV and E. RANDI, 2004 Evidence of genetic distinction and long-term population decline in wolves (*Canis lupus*) in the Italian Apennines. Molecular Ecology **13:** 523-536.

MELDGAARD, T., E. E. NIELSEN and V. LOESCHCKE, 2003 Fragmentation by weirs in a riverine system: A study of genetic variation in time and space among populations of European grayling (Thymallus thymallus) in a Danish river system. Conservation Genetics **4:** 735-747.

MILTON, K., J. D. LOZIER and E. A. LACEY, 2009 Genetic structure of an isolated population of mantled howler monkeys (*Alouatta palliata*) on Barro Colorado Island, Panama. Conservation Genetics **10:** 347-358.

MONDOL, S., K. U. KARANTH and U. RAMAKRISHNAN, 2009 Why the Indian subcontinent holds the key to global tiger recovery. PLoS Genetics **5:** e1000585.

NETTEL, A., R. S. DODD and Z. AFZAL-RAFII, 2009 Genetic diversity, structure, and demographic change in tanoak, Lithocarpus densiflorus (Fagaceae), the most susceptible species to sudden oak death in California. Am. J. Bot. **96:** 2224-2233.

NIELSEN, E. K., C. R. OLESEN, C. PERTOLDI, P. GRAVLUND, J. S. F. BARKER *et al.*, 2008 Genetic structure of the Dansih red deer (*Cervus elaphus*). Biological Journal of the Linnean Society **95:** 688-701.

NIELSEN, R. K., C. PERTOLDI and V. LOESCHCKE, 2007 Genetic evaluation of the captive breeding program of the Persian wild ass. Journal of Zoology **272:** 349-357.

OKELLO, J. B. A., G. WITTEMYER, H. B. RASMUSSEN, P. ARCTANDER, S. NYAKAANA *et al.*, 2008 Effective population size dynamics reveal impacts of historic climatic events and recent anthropogenic pressure in African elephants. Molecular Ecology **17:** 3788-3799.

OLIVIERI, G. L., V. SOUSA, L. CHIKHI and U. RADESPIEL, 2008 From genetic diversity and structure to conservation: genetic signature of recent population declines in three mouse lemur species (*Microcebus* spp.). Biological Conservation **141:** 1257-1271.

PERTOLDI, C., M. M. HANSEN, V. LOESCHCKE, A. B. MADSEN, L. JACOBSEN *et al.*, 2001 Genetic consequences of population decline in the European otter (*Lutra lutra*): an assessment of microsatellite DNA variation in Danish otters from 1883 to 1993. Proceedings of the royal society B **268:** 1775-1781.

PROCHAZKA, P., E. BELLINVIA, D. FAINOVA, P. HAJKOVA, A. ELHALAH *et al.*, 2008 Immigration as a possible rescue of a reduced population of a long-distant migratory bird: Reed warblers in the Azraq Oasis, Jordan. Journal of Arid Environments **72:** 1184-1192.

PRUETT, C. L., T. N. TURNER, C. M. TOPP, S. V. ZAGREBELNY and K. WINKER, 2010 Divergence in an archipelago and its conservation consequences in Aleutian Island rock ptarmigan. Conservation Genetics **11:** 241-248.

RANDI, E., F. DAVOLI, M. PIERPAOLI, C. PERTOLDI, A. B. MADSEN *et al.*, 2003 Genetic structure in otter (*Lutra lutra*) populations in Europe: implications for conservation. Animal Conservation **6:** 93-100.

SAILLANT, E., J. C. PATTON, K. E. ROSS and J. R. GOLD, 2004 Conservation genetics and demographic history of the endangered Cape Fear shiner (*Notropis mekistocholas*). Molecular Ecology **13:** 2947-2958.

SIVASUNDAR, A., and J. HEY, 2003 Population genetics of Caenorhabditis elegans: the paradox of low polymorphism in a widespread species. Genetics **163:** 147-157.

SOUSA, V., F. PENHA, M. J. COLLARES-PEREIRA, L. CHIKHI and M. M. COELHO, 2008 Genetic structure and signature of population decrease in the critically endangered freshwater cyprinid *Chondrostoma lusitanicum*. Conservation Genetics **9:** 791-805.

SOUSA, V., F. PENHA, I. PALA, L. CHIKHI and M. M. COELHO, 2010 Conservation genetics of a critically endangered Iberian minnow: evidence of population decline and extirpations. Animal Conservation **13:** 162-171.

STAMFORD, M. D., and E. B. TAYLOR, 2005 Population subdivision and genetic signatures of demographic changes in Arctic grayling (*Thymallus arcticus*) from an impounded watershed. Canadian Journal of Fisheries and Aquatic Sciences **62:** 2548-2559.

SWATDIPONG, A., C. R. PRIMMER and A. VASEMÄGI, 2010 Historical and recent genetic bottlenecks in European grayling, *Thymallus thymallus*. Conservation Genetics **11:** 279-292.

TSENG, M.-C., W.-N. TZENG and S.-C. LEE, 2003 Historical decline in the Japanese eel *Anguilla japonica* in northern Taiwan inferred from temporal genetic variations. Zoological Studies **42:** 556-563.

WIRTH, T., and L. BERNATCHEZ, 2003 Decline of north Atlantic eels: a fatal synergy? Proceedings of the royal society of London B **270:** 681-688.

WIRTH, T., F. HILDEBRAND, C. ALLIX-BÉGUEC, F. WÖLBELING, T. KUBICA *et al.*, 2008 Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathogens **4:** e1000160.

ZHANG, B., M. LI, Z. ZHANG, B. GOOSSENS, L. ZHU *et al.*, 2007 Genetic viability and population history of the Giant Panda, putting an end to the ''evolutionary dead end''? Molecular Biology and Evolution **24:** 1801-1810.

TABLE S2

Values of the GELMAN - RUBIN Shrink Factor

A. Contractions, with N_0 = 100 (SMM)

	N_1 = 1,000			N_1 = 10,000			N_1 = 100,000		
T_a = 10	1.01	-	-	1.01	-	-	2.09	2.10	1.12 NC
	1.01	-	-	1.08*	1.09	1.01	1.85	2.03	1.47 NC
	1.03	-	-	1.09	-	-	2.73	2.56	2.18 NC
	1.01	-	-	1.09	-	-	1.56	1.32	2.14 NC
	1.02	-	-	1.02	-	-	2.24	1.26	1.25 NC
T_a = 50	1.03	-	-	2.57	1.19	1.08	1.98	1.92	1.15 NC
	1.01	-	-	1.49	1.28	1.03	1.75	1.32	2.41 NC
	1.02	-	-	1.67	1.5	1.05	2.58	4.30	1.81 NC
	1.05	-	-	1.28	1.27	1.03	1.67	1.56	1.18 NC
	1.01	-	-	2.56	1.26	1.07	2.85	1.32	1.04
T_a = 100	1.03	-	-	2.4	1.47	1.02	1.97	1.66	1.02
	1.01	-	-	1.2	1.08	-	1.53	3.10	1.29 NC
	1.01	-	-	3.17	1.32	1.47 NC	2.37	4.59	1.44 NC
	1.04	-	-	1.69	1.63	1.33 NC	1.79	3.58	2.24 NC
	1.04	-	-	1.14	1.15	1.02	2.80	2.68	1.52 NC
T_a = 500	2.18	1	-	1.04	-	-	1.70	1.04	-
	1	-	-	1.04	-	-	1.26	1.12	1.01
	1	-	-	1.07	-	-	1.17	1.01	-
	1.02	-	-	1.63	1.27	1.03	1.10	-	-
	1.06	-	-	1.03	-	-	1.16	1.04	-

B. Expansions, with N_1 = 100 (SMM)

	N_0 = 1,000			N_0 = 10,000			N_0 = 100,000		
T_a = 10	1	-	-	1.02	-	-	1.11	1.00	-
	1.01	-	-	1.02	-	-	1.06	-	-
	1.01	-	-	1.02	-	-	1.06	-	-
	1.07	-	-	1.01	-	-	1.00	-	-
	1.01	-	-	1.01	-	-	1.00	-	-
T_a = 50	1.16	1.02	-	1.06	-	-	1.03	-	-
	1.01	-	-	1.3	1.21	1.05	1.04	-	-
	1.03	-	-	1.37	1.02	-	1.18	1.04	-
	1.08	-	-	1.07	-	-	1.02	-	-
	1.62	1.09	-	1.05	-	-	1.10	1.30	1.10
T_a = 100	1.04	-	-	1.05*	1.04	-	1.10	-	-
	1.02	-	-	1.08*	1.13	1.88 NC	1.20	1.08	-
	1	-	-	1.05	-	-	1.28	1.03	-
	1.01	-	-	1.61	1.43	1.01	1.21	1.07	-

	1.11	1.08	-	1.09	-	-	3.13	1.08	-
T_a = 500	1.07	-	-	1.13	3.71	1.97 NC	1.23	1.02	-
	1.08	-	-	1.22	1.06	-	2.11	1.02	-
	1.06	-	-	1.03	-	-	1.67	1.07	-
	1.02	-	-	1.03	-	-	2.90	1.16	1.06
	1.05	-	-	1.05	-	-	1.07	-	-

N_0, ancestral effective population size; N_1, current effective population size; T_a, time since the population size change. For each value of N_1 or N_0, the first column corresponds to a run length of 10^9 steps, the second column to 3 x 10^9 steps (if performed) and the third one to 1.5 x 10^{10} steps (if performed); *, chains that have not reached stationarity using the GELMAN - RUBIN diagnostic plot, despite having GELMAN - RUBIN shrink factor < 1.10; NC, non-converged Markov chains.

TABLE S3

Genetic Diversity in the Second Set of Simulated Data

	Declining population				Stable population				Expanding population			
	H_e	N_a	A_r	V_a	H_e	N_a	A_r	V_a	H_e	N_a	A_r	V_a
SMM	0.56	4.2	6.9	13.5	0.54	3.8	3.0	2.0	0.52	4.3	3.3	1.98
$p = 0.00$	(0.04)	(0.4)	(1.3)	(5.6)	(0.07)	(0.3)	(0.2)	(0.3)	(0.04)	(0.4)	(0.4)	(0.30)
GSM	0.51	3.9	9.5	30.9	0.57	4.8	4.8	4.4	0.52	5.7	5.3	4.2
$p = 0.22$	(0.04)	(0.2)	(1.6)	(16.2)	(0.08)	(0.7)	(1.4)	(2.3)	(0.04)	(0.1)	(0.4)	(0.6)
GSM	0.56	4.44	32.9	323.0	0.62	6.7	16.9	43.5	0.62	12.1	20.5	38.2
$p = 0.74$	(0.08)	(0.59)	(14.6)	(211.8)	(0.03)	(0.7)	(1.5)	(6.2)	(0.06)	(0.8)	(1.6)	(4.9)

H_e, expected heterozygosity; N_a, number of alleles; A_r, allele size range; V_a, variance of allele size range; estimates of genetic diversity are averaged over the five simulated datasets for each set of parameters; standard deviations are indicated below the mean, into parentheses. Data were simulated for a declining population ($N_0 = 100$; $N_1 = 10{,}000$; $T_a = 500$), a stable population ($N_0 = N_1 = 464$; $T_a = 500$), and an expanding population ($N_0 = 10{,}000$; $N_1 = 100$; $T_a = 500$), under three mutation models: a strict stepwise mutation model (SMM), a moderate GSM with a frequency of multi-step changes set to $p = 0.22$, and a strong GSM with a frequency of multi-step changes set to $p = 0.74$.

Annexe F

Article 4 : mesurer l'intensité de la sélection

Vitalis R., Gautier M., Dawson K. et Beaumont M. Detecting and measuring selection from gene frequency data. (2012) Detecting and measuring selection from gene frequency data. *En préparation*

Detecting and measuring selection from gene frequency data

Renaud Vitalis[*1], Mathieu Gautier[†2], Kevin J Dawson[‡3], and Mark A Beaumont[§4]

[1]Centre National de la Recherche Scientifique – Institut National de la Recherche Agronomique, UMR CBGP (INRA – IRD – CIRAD – Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier sur Lez Cedex, France

[2]Institut National de la Recherche Agronomique, UMR CBGP (INRA – IRD – CIRAD – Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier sur Lez Cedex, France

[3]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

[4]Department of Mathematics and School of Biological Sciences, University of Bristol, Bristol BS8 1TNW, UK

[*]Corresponding author; e-mail: renaud.vitalis@supagro.inra.fr
[†]e-mail: mathieu.gautier@supagro.inra.fr
[‡]e-mail: kevin.dawson@sanger.ac.uk
[§]e-mail: m.beaumont@bristol.ac.uk

Introduction

In the new era of population genomics, surveys of genetic polymorphism ("genome scan") offer the opportunity to distinguish locus-specific from genome-wide effects at many loci (Black et al. 2001). Identifying presumably neutral regions of the genome that are assumed to be influenced by genome-wide effects only, and excluding presumably selected regions is critical to infer population demography and phylogenetic history reliably (Ross et al. 1999). Conversely, detecting locus-specific effects may help identify those genes that have been, or still are, targeted by natural selection (Luikart et al. 2003). Such genes may be involved, e.g., in the adaptation to new environments or in the arm-race with pathogens (Nielsen 2005). The applications for population genomic analyses therefore cover a wide range of disciplines. How best to identify regions, loci or single nucleotides which have been, or still are, under selection is still a challenging issue, however (Nielsen 2001).

Tests of selective neutrality have been developed for samples drawn from single populations. Most of them are based on the comparison of some summary statistics of the site-frequency spectrum (i.e., the observed distribution of gene frequencies), to their expected distribution from diffusion theory under an infinitely-many sites mutation model (Bustamante et al. 2001; Payseur et al. 2002; Nielsen et al. 2005*b*; Williamson et al. 2005). Accounting for different classes of markers (e.g., selected and neutral) is achieved by a Poisson Random Field (PRF) approximation, which assumes independent mutation and selection parameters across sites (see, e.g., Kim and Stephan 2002; Bustamante et al. 2003). In particular, Bustamante et al. (2003) developed a hierarchical PRF model that allows the estimation of selection coefficients at

a set of DNA polymorphisms sampled in a single population (see also Nielsen et al. 2005*a*). Williamson et al. (2005) used a similar approach to infer selection in a non-equilibrium demographic model. Yet, they assume *a priori* which mutations are selectively neutral, and which are not. The putatively neutral class of markers is then used to infer demographic parameters and, given these estimates, inferences regarding selection are performed on the other class of markers.

Other tests of selective neutrality are based on haplotype structure. Focusing on haplotypes at a locus of interest (referred to as "core haplotypes"), Sabeti et al. (2002) analyzed the decay of gene identity as a function of distance from the core, as measured by the extended haplotype homozygosity (EHH). Core haplotypes that have both a high population frequency and a high EHH are evidence of recent positive selection. Deriving the expected distribution of the EHH requires making strong assumptions about the underlying population history, tough, which makes it difficult to evaluate the significance of observed values. Several extensions have therefore been proposed and adapted to genome-wide scans of single nucleotide polymorphism (SNP) data, based on the empirical distribution of EHH-like statistics either for single populations (Voight et al. 2006) or for pairs of differentiated populations (Tang et al. 2007). Such approaches therefore rely on the assumption that most SNPs behave neutrally, so that the observed distribution of the statistics provides a proxy to the null distribution.

When markers are genotyped across multiple populations, it has been advocated that signatures of natural selection may simply be identified in the extreme tails of the empirical distribution of F_{ST} estimates (Goldstein and

Chikhi 2002). These model-free approaches have been applied to both the Perlegen (Hinds et al. 2005) and the HapMap (The International HapMap Consortium 2003, 2005) SNP datasets (see, e.g., Akey et al. 2002; Weir et al. 2005; Barreiro et al. 2008). Such methods are intended to be immune to arbitrary assumptions about the (unknown) demographic history of the sample. Dependence upon the unknown demography (including the geographic and historical relationship among populations) was indeed a severe criticism of Lewontin and Krakauer's (1973) tests of selective neutrality, based on the sampling distribution of the parameter F_{ST} (Robertson 1975; Nei and Maryuyama 1975). However, recent refinements of this controversial test showed that the distribution of F_{ST} estimates should be relatively robust to demographic effects, which prevents the need to model the demography explicitly (Beaumont and Nichols 1996; Vitalis et al. 2001). This robustness to the effects of demography stems from the properties of gene genealogies in structured populations (Beaumont 2005) which, in many cases, naturally tend towards a simple structure (Nordborg 1997; Wakeley 1999).

F_{ST}-based methods, which look for locus-specific effects on F_{ST} estimates are typically not designed to identify population-specific selection (Beaumont and Nichols 1996). One approach to tackle this problem has been to consider pairwise population divergence models (see Wilding et al. 2001; Vitalis et al. 2001). Weir et al. (2005) have also shown the utility of estimating population-specific F_{ST} values, rather than the population-average value. But these moment-based (Vitalis et al. 2001) or empirical (Weir et al. 2005) approaches raise the problem of multiple testing. An alternative approach has been proposed by Beaumont and Balding (2004), which consists

in decomposing F_{ST} into locus-, population-, and locus-by-population components in a hierarchical Bayesian analysis. This formulation, which provided the statistical ground to test which loci are targeted by selection, was further extended by Riebler et al. (2008), Foll and Gaggiotti (2008) and Guo et al. (2009).

A major limitation of the methods based on comparisons among population, is that they do not quantify selection. Rather, they are constructed as tests of departure from selective neutrality. While the neutral theory is a convenient null hypothesis, a proper interpretation of the observed patterns of variability, in particular the extent to which the neutral theory is applicable, requires methods that rely on non-neutral models (see, e.g., Donnelly et al. 2001). Furthermore, proper tests of selection should provide estimates the parameters of interest, i.e. the strength and the model of selection on segregating polymorphisms.

Here, we provide a new method to distinguish neutral from selected polymorphisms and estimate the intensity of selection at the latter. Our model accounts explicitly for positive selection, and we consider that all marker loci in the dataset are selected for, to some extent. The method is based on a diffusion approximation for the distribution of allele frequency in a population subdivided in a number of demes that exchange migrants (i.e., an island model, see Wright 1931). The framework for statistical inference from this model consists in a hierarchical Bayesian model (see Gelman et al. 2004). We use a componentwise Markov chain Monte Carlo (MCMC) algorithm to sample from the joint posterior distribution of the model parameters. We then test the performance of our method, by means of stochastic simulations.

Last, we re-analyze a subset of SNP data from the Stanford HGDP-CEPH Human Genome Diversity Cell Line Panel (Cann et al. 2003).

Materials and Methods

The Model

We consider an infinite island model where the ith deme is made of N_i diploid individuals, and receives immigrants from the whole population at rate m_i. We define the scaled migration parameter in the ith deme as $M_i \equiv 4N_i m_i$. We consider bi-allelic markers, i.e. that only two alleles (noted A and a) may occur at a given locus. We note p_{ij} the frequency of allele A in deme i at locus j, and π_j the frequency of allele A at the jth locus in the whole population. Since we consider that the population as a whole is made of an infinite number of islands, π_j gives the frequency of allele A in the pool of migrant individuals. The following notations will be used hereafter: the vector of allele frequencies in deme i at locus j is $\mathbf{p}_{ij} \equiv (p_{ij}, 1-p_{ij})$; and the vector of allele frequencies at locus j among migrants is $\boldsymbol{\pi}_j \equiv (\pi_j, 1-\pi_j)$. We consider a simple genic model of selection where, at each locus, the allele A provides a selective advantage. The homozygote individuals AA and the heterozygotes Aa have a relative increase of fitness of $1+s_{ij}$ and $1+s_{ij}/2$, respectively, as compared to the aa homozygotes. We define the scaled coefficient of selection in deme i at locus j as $\sigma_{ij} \equiv 2N_i s_{ij}$. We define the indicator variable κ_{ij}, which takes the value $\kappa_{ij} = 0$ if allele A is selected for, and $\kappa_{ij} = 1$ if allele a is selected for. Therefore, the frequency of the selected allele in deme i at locus j, reads $\tilde{p}_{ij} \equiv \kappa_{ij}(1-p_{ij}) + (1-\kappa_{ij})p_{ij}$.

The data consist in individuals collected in a set of n_{d} demes, and genotyped at L loci. We note n_{ij} the total number of genes sampled in the ith deme at the jth locus, out of which x_{ij} have allelic state A. The vector of allele counts in deme i at locus j therefore reads $\mathbf{n}_{ij} \equiv (x_{ij}, n_{ij} - x_{ij})$.

Hierarchical Bayesian approach

Given the frequencies p_{ij} of allele A, the conditional distribution of allele counts $\mathbf{n}_{ij}$ in population i at locus j is binomial:

$$\mathcal{L}(p_{ij};\mathbf{n}_{ij}) = \binom{n_{ij}}{x_{ij}} p_{ij}^{x_{ij}}(1-p_{ij})^{n_{ij}-x_{ij}}. \tag{1}$$

In the limit of large deme size, as $N_i \to \infty$, and assuming that selection and random genetic drift are of comparable strength (i.e., that M_i and σ_{ij} have a finite limit as $N_i \to \infty$), the distribution of the $\mathbf{p}_{ij}$ may be approximated by the stationary density of a diffusion process, which has the form:

$$\psi(p_{ij};M_i,\sigma_{ij},\kappa_{ij},\boldsymbol{\pi}_j) = C^{-1}\exp(\sigma_{ij}\tilde{p}_{ij}){p_{ij}}^{M_i\pi_j-1}(1-p_{ij})^{M_i(1-\pi_j)-1} \tag{2}$$

This equation is known as Wright's formula (Wright 1935, 1949, 1969); see also Ethier and Nagylaki (1988); Barton and Turelli (1989); Bürger (2000). In eq. (2), C is the constant that ensures that the distribution integrates to 1. This constant can be evaluated as:

$$\begin{aligned} C &= \int \exp(\sigma_{ij}\tilde{p}_{ij}){p_{ij}}^{M_i\pi_j-1}(1-p_{ij})^{M_i(1-\pi_j)-1}\mathrm{d}\mathbf{p}_{ij} \\ &= {}_1F_1(M_i\tilde{\pi}_{ij};M_i;\sigma_{ij})\frac{\Gamma(M_i\pi_j)\Gamma(M_i(1-\pi_j))}{\Gamma(M_i)} \end{aligned} \tag{3}$$

where ${}_1F_1(a;b;z)$ is the confluent hypergeometric, or Kummer's, function (see, e.g., Abramowitz and Stegun 1965, p. 504), and $\tilde{\pi}_{ij} \equiv \kappa_{ij}(1-\pi_j)+(1-\kappa_{ij})\pi_j$.

Given the model specified in eqs. (1) and (2), we are interested in evaluating the parameters of interest $\mathbf{M} \equiv (M_1,\ldots,M_i,\ldots,M_{n_\mathrm{d}})$, $\boldsymbol{\pi} \equiv (\pi_1,\ldots,\pi_j,\ldots,\pi_L)$,

$\boldsymbol{\sigma} \equiv (\sigma_{11}, \ldots, \sigma_{ij}, \ldots, \sigma_{n_{\mathrm{d}}L})$ and $\boldsymbol{\kappa} \equiv (\kappa_{11}, \ldots, \kappa_{ij}, \ldots, \kappa_{n_{\mathrm{d}}L})$, from the observed allele counts $\mathbf{n}$ over all sampled demes and loci. The directed acyclic graph (DAG) for this model is shown in Figure 1.

We assume a Bernoulli prior distribution for the parameters κ_{ij}, i.e. $\kappa_{ij} \sim$ Bernoulli(0.5), and a uniform prior for the $\boldsymbol{\pi}_j$'s, that is $\boldsymbol{\pi}_j \sim \mathrm{Beta}(1, 1)$. We further assume a log-uniform prior for the M_i's with support from 0.001 to 1000, i.e. the priors of the M_i's are uniform in log scale: $\log(M_i) \sim \mathcal{U}(10^{-3}, 10^3)$. The prior distributions for the selection coefficients σ_{ij} (at each locus, in each deme) are modelled hierarchically (see, e.g., Gelman et al. 2004, pp. 124-125). In particular, we assume that σ_{ij} has an exponential prior distribution $f(\sigma_{ij}|\delta_j) \sim \exp\left(\delta_j^{-1}\right)$ that depends upon the locus-specific hyperparameter δ_j, which represents the average effect of selection at locus j (over all demes). We further assume that this hyperparameter δ_j has an exponential prior distribution $f(\delta_j|\lambda) \sim \exp(\lambda^{-1})$ that depends, in turn, upon the hyperparameter λ, which represents the genome-wide effect of selection over all demes and loci. Last, we assume that the prior distribution of λ, is $f(\lambda) \sim \exp(\Lambda^{-1})$, with $\Lambda = 0.5$, in what follows. Assuming independence of allele frequencies among loci and populations, the posterior distribution of the parameters $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\kappa}, \boldsymbol{\delta}, \lambda|\mathbf{n})$, i.e., the conditional distribution of the parameters $\mathbf{M}$, $\boldsymbol{\pi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$, and λ given the data $\mathbf{n}$, depends upon the prior distributions of the parameters and the data as:

$$f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda|\mathbf{n}) \propto \prod_{i=1}^{n_{\mathrm{d}}} \prod_{j=1}^{L} \mathcal{L}(p_{ij}; \mathbf{n}_{ij}) \psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij}) \times f(\mathbf{M}) f(\boldsymbol{\pi}) f(\boldsymbol{\kappa}) f(\boldsymbol{\sigma}|\boldsymbol{\delta}) f(\boldsymbol{\delta}|\lambda) f(\lambda) \quad (4)$$

In what follows, the full posterior distribution of the parameters $\mathbf{M}$, $\boldsymbol{\pi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$, and λ, which is specified by equation (4), is estimated by a single-component Metropolis–Hastings (or Metropolis within Gibbs) algorithm (see, e.g., Ntzoufras 2009). In practice, we therefore update one parameter at each time, iteratively, as detailed in the Appendix. The proposal distributions for each of the $\mathbf{M}$, $\boldsymbol{\pi}$, $\boldsymbol{\kappa}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$, and λ parameters are adjusted by means of 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40 (see, e.g., Gilks et al. 1996).

Analyses of the outputs from MCMC simulations

Because the model assumes that each and every locus in a dataset is selected to a certain extent, we are particularly interested in the posterior densities of the locus-specific hyperparameters δ_j: we expect the density to be shifted toward zero for neutral markers, and to positive values for (presumably) selected loci (see Figure 2). Yet, given the hierarchical structure of our model, it would not be sufficient to simply test whether, at a particular locus, the posterior distribution of δ_j departs from zero. This approach would indeed neglect the genome-wide effects of selection. Since we assume in our model, that the δ_j's are drawn independently from a common hyperdistribution with parameter λ (that represents the genome-wide effect of selection), it is more appropriate to compare the posterior distributions of the locus-specific coefficients of selection with the "centering" distribution derived from the hyperdistribution of the genome-wide effect of selection.

Following Guo et al. (2009), we consider the following steps to detect outlier loci in a dataset: (i) approximate the posterior distributions of the

locus-specific selection parameters (δ_j), and that of the genome-wide effect of selection (λ); then (ii) compute the distance between the locus-specific selection parameters and the "centering" distribution derived from the hyperdistribution of the genome-wide effect of selection (which, broadly, describes the among-locus variation in the locus-specific effect of selection); last (iii) measure the mean of the posterior distributions of σ_{ij} for outlier loci over sampled populations, in order to appreciate the distribution of selection effects across sampling locations.

Since our preliminary analyses have shown that the posterior distribution of the parameters δ_j is unimodal, with support on $[0, \infty)$, it can be approximated by a gamma distribution. We therefore approximate the posterior distribution of δ_j with a gamma distribution $\Gamma(k_0, \theta_0)$, which has the same mean and variance as the mean $\bar{x}_{\delta_j}$ and the variance $s^2_{\delta_j}$ of the posterior distribution of δ_j, as estimated from the MCMC outputs, i.e. with $k_0 = \bar{x}^2_{\delta_j}/s^2_{\delta_j}$ and $\theta_0 = s^2_{\delta_j}/\bar{x}_{\delta_j}$. Since we assumed an exponential prior distribution for the hyperparameter λ, i.e., $f(\lambda) \sim \exp(\Lambda^{-1})$, we may further approximate the posterior distribution of λ with a gamma distribution $\Gamma(1, \theta_1)$, which has the same mean as the mean $\bar{x}_\lambda$ of the posterior distribution of λ, as estimated from the MCMC outputs, i.e. with $\theta_1 = \bar{x}_\lambda$.

We then compare the posterior distribution for each locus-specific hyperparameter δ_j, with the posterior of the hyperparameter λ, which represents the genome-wide effect of selection. The loci for which the posterior of δ_j departs substantially from that of λ are considered as good candidates for being targeted by selection. We use the Kullback–Leibler divergence (KLD) to measure the divergence between the posterior of δ_j and its centering dis-

tribution. The KLD between two densities $f(x)$ and $g(x)$ is defined as

$$\mathrm{KLD}(f(x), g(x)) = \int_{-\infty}^{\infty} f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}(x). \tag{5}$$

With little algebra, one can show that the KLD between two gamma distributions with shape and scale parameters (k_0, θ_0) and (k_1, θ_1), respectively, is given by:

$$\begin{aligned}\mathrm{KLD}\left(\Gamma(k_0, \theta_0), \Gamma(k_1, \theta_1)\right) = \log\left(\frac{\Gamma(k_1)\theta_1^{k_1}}{\Gamma(k_0)\theta_0^{k_0}}\right) + k_0\frac{\theta_0 - \theta_1}{\theta_1} \\ + (k_0 - k_1)\left[\log(\theta_0) + \Psi(k_0)\right],\end{aligned} \tag{6}$$

where $\Psi(\cdot)$ is the digamma function.

We calibrate the KLD following Guo et al. (2009): consider flipping a "fair" coin with equal probability 0.5 for head and tail versus flipping a biased coin with probability 0.05 (resp. 0.01) for head, then the KLD between these two Bernoulli distributions equal 0.830 (resp. 1.614). All the post-processing statistical analyses were performed using the R software environment for statistical computing, version 2.15.0 (R Development Core Team 2012). Posterior densities were estimated from the posterior samples using the local-likelihood method of Loader (1996), as implemented in the locfit package for R (version 1.5-8).

Simulated datasets

We evaluated the performance of the method by simulating artificial datasets for fixed parameter values. The simulations were performed according to an island model with 50 demes, each made of $N = 250$ diploid individuals.

Following Beaumont and Balding (2004), we simulated allele counts data from a Wright–Fisher model with migration and selection.

Initialization was achieved by means of a Pólya urn scheme simulation of the coalescent (Donnelly and Tavaré 1995). This amounts considering selection acting on standing variation, and makes this simulation model similar in spirit to the models considered by Innan and Kim (2004) and Przeworsky et al. (2005). At each generation (generations were discrete and non-overlapping), each individual produced a random number of offspring drawn from a Poisson distribution with mean 100. Mutations then occurred at rate 2×10^{-5}. Dispersal of the (diploid) offspring then occurred, with dispersing individuals reaching necessarilly a distinct deme. Selection of the offspring surviving to adulthood was then achieved, according to the scheme detailed below. A number N of adults was drawn from offspring, except if the number of offspring in a deme was less than N, in which case all offspring survived. This life-cycle was repeated for 25,000 generations. Samples were then taken, but only if the minimum allele frequency (the frequency of the least frequent allele) was larger than 0.01. All loci were considered as independent, so that each multilocus dataset was made of independent realizations of that process.

To account for the possibility of positive selection to local environmental conditions, the demes were arbitrarily provided with attributes ("blue", "red", or "uncolored"), which were assigned at random, independently for each selected locus. For positively selected loci, one allele B was considered as advantageous in a "blue" deme (and neutral in a "red" deme), while the other allele R was considered as advantageous in a "red" deme (and neutral in a "blue" deme). Both alleles were considered as neutral in "uncolored" demes.

Therefore, BB homozygotes had fitness $(1+s)$ in "blue" demes and 1 in "red" and "uncolored" demes; RR homozygotes had fitness $(1+s)$ in "red" demes and 1 in "blue" and "uncolored" demes; BR heterozygotes had fitness $1+s/2$ in "red" and "blue" demes and 1 in "uncolored" demes. For loci under balancing selection, only the heterozygote genotypes were selected for in the "blue" and "red" demes, with relative fitness $(1+s)$. Homozygote genotypes were neutral (relative fitness 1) in all demes, as were the heterozygote genotypes in "uncolored" demes.

A total of twelve datasets were generated using the Wright–Fisher model described above (see Table 1). In the following, we assumed that 30% of all demes were "blue" demes, 30% were "red" demes and 40% were "uncolored" demes. For each locus, 50 diploid individuals (100 genes) were sampled per deme. The details that distinguish the different datasets are given in Table 1. For example, for sets 1 to 9, the samples were taken in 6 demes: 2 "blue" demes, 2 "red" demes, and 2 "uncolored" demes, and each simulated dataset consisted in 10,000 SNPs, with 8,000 neutral markers, 1,000 positively selected loci and 1,000 loci under balancing selection. Table 1 further gives the combinations of M values and σ/M ratios used for the simulations.

For each of these twelve datasets a Markov chain Monte Carlo (MCMC) was run to sample from the joint posterior distribution of the model parameters. For each Markov chain, 50,000 updating steps were completed after 25 short pilot runs of 1,000 iteration and a burn-in of 10,000 steps. Samples were collected for all the model parameters every 25 steps (thinning) to avoid autocorrelations, yielding 2,000 observations. Estimation of the posterior densities, computation of the KLD measure, receiver operating character-

istic (ROC) analysis (see, e.g., Fawcett 2006, for further information) were all performed using the R software environment for statistical computing, version 2.15.0 (R Development Core Team 2012). The same datasets were analyzed using BayeScan version 2.1 (Foll and Gaggiotti 2008) with default option values. Each Markov chain was run for 50,000 updating steps, after 20 short pilot runs of 5,000 iteration and a burn-in of 10,000 steps. Samples were collected every 25 steps (thinning), yielding 2,000 observations.

Human data

We applied our method on the Stanford HGDP-CEPH Human Genome Diversity Cell Line Panel (Cann et al. 2003) SNP Genotyping Data, that consist in genotypes at more than 650,000 SNP loci determined with the Illumina BeadStation technology. Because we were interested in measuring the genetic signature of selection in the lactase gene, we only used the data from chromosome 2 (53,765 SNPs), and incorporated the genotyping data of the two SNPs reported to be very tightly associated with lactase persistence (-13910C→T and -22018G→A) as published by Bersaglieri et al. (2004). The data were downloaded from the HGDP-CEPH: ftp://ftp.cephb.fr/hgdp_supp1. All the populations with less than 15 genotyped individuals were discarded from the dataset. Furthermore, we removed seven populations from Oceania and Southern America, as well as three populations from Sub-Saharan Africa (the Biaka Pygmies, Mbuti Pygmies and the Mandenka) that were absent from Bersaglieri et al.'s (2004) dataset. Last, the two Bantu populations (from Kenya and South Africa) were merged, as in Bersaglieri et al. (2004). This resulted in a final dataset with 23 populations from Africa and Eurasia. We

applied a minimum allele frequency of 0.01, so that only the SNPs which frequency of the least frequent allele was larger than 0.01 were retained. This resulted in genotyped data from 52,631 marker loci from the HGDP-CEPH data, and from the two SNPs (-13910C→T and -22018G→A) as published in Bersaglieri et al. (2004).

Results

Evaluating performance on simulated data

Figure 3 shows the performance of the method on the dataset 5 (see Table 1), which corresponds to $M \equiv 4Nm = 5$ and $\sigma \equiv 2Ns = 25$. Figures S1 to S11 provide the same outputs for all other simulated datasets. Figure 3A shows that the distribution of KLD measures for positively selected loci departs from that of the neutral markers and the loci under balancing selection. This is essentially true for the datasets for which $M \geq 5$ and $\sigma/M \geq 5$ (datasets 6–11) and $M = 2$ and $\sigma/M = 10$ (dataset 3), as can be seen from Figures S1 to S10. Not surprisingly, large KLD measures correspond to large F_{ST} estimates (Figure 3B). This is so because for positively selected loci one allele is selected for in "blue" populations and the other in "red" populations, which tends to exacerbate differentiation. Figure 3B further shows that using the KLD = 0.830 threshold (see Guo et al. 2009) enables to discriminate between positively selected loci and neutral markers (see also Table 2). This point is strengthened by the examination of the false positive rate (the proportion of neutral markers that exhibit a signature of selection), and the false negative rate (the proportion of selected loci that do not exhibit a signature of selection) as a function of the Kullback–Leibler divergence measure (Figure 3C). Indeed, using KLD = 0.830 as a threshold value to discriminate between neutral and positively selected loci minimizes both the false positive and the false negative rates. Last, Figure 3 shows that our method has no statistical power to identify loci under balancing selection (see also Table 2). Although the mean and the variance of he KLD measures for loci under balancing selection (and their F_{ST} estimates) are lower as compared

to neutral markers, the KLD measures for these loci remain very low. This result is not surprising, though, since the selection scheme considered in our model of inference only accounts for positive genic selection. Furthermore, previous simulation studies have also shown that, in the absence of an explicit model of selection, similar methods generally lack power to detect balancing selection (Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler et al. 2008).

Comparison with BayeScan

All the datasets described in Table 1 were analyzed with BayeScan version 2.1 (Foll and Gaggiotti 2008). BayeScan is based on the Multinomial–Dirichlet model for allele frequencies in an island model of population structure. At each locus, the variance of allele frequency between each subpopulation and the common pool of migrants is given by a subpopulation-specific F_{ST} parameter. In BayeScan, as in Beaumont and Balding (2004) model, the parameter F_{ST} is decomposed into a locus-specific component (α_i) shared by all populations, and a population-specific component (β_j) shared by all loci. Significantly positive or negative values of α_i are taken as evidence of selection. BayeScan is based on a reversible-jump Markov chain Monte Carlo algorithm, which estimates the posterior probabilities of two alternative models, a purely neutral one ($\alpha_i = 0$) and one including selection ($\alpha_i \neq 0$). For each output, we computed the Bayes factor (BF) for the model including selection ($\alpha_i \neq 0$). The Bayes factor is a ratio where the numerator is the posterior probability of one model divided by its prior probability and the denominator is the posterior probability of an alternative model divided by

its prior probability (Gelman et al. 2004). Here we assumed a prior odd of 10 for the neutral model.

Figure 4A shows the relationship between BayeScan Bayes factor and the Kullback–Leibler divergence measure for each and every locus from dataset 5. Using Jeffreys' scale of evidence for Bayes factors (Jeffreys 1961; Kass and Raftery 1995), a Bayes factor of 3 is considered as being a "substantial" evidence for selection, and a Bayes factor less than 3 as barely worth mentioning. It is clear from Figure 4A that, for this set of simulated data, an appreciable proportion of positively selected loci are not classified as outliers using BayeScan BF criterion. Yet, the KLD measure for these loci are, for most of them, above the KLD = 0.830 threshold. Whenever the posterior probability of the model including selection ($\alpha_i \neq 0$) was equal to 1, we arbitrarily defined the $\log_{10}(\text{BF})$ as $\log_{10}(1999.5/2000) - \log_{10}(0.5/2000)$, in order to account for the chain length (2,000 iterations). The maximum value that the BF can take (BF = 4.556, see Figure 4A) is therefore arbitrary. Yet it is interesting to see that this maximum value corresponds to a wide range of variation of the KLD measure, and presumably to a wide range of selection strength.

This apparent superiority of our method to BayeScan (Foll and Gaggiotti 2008) is confirmed by the receiver operating characteristic (ROC) analysis (Figure 4B). In the ROC analysis, the proportion of false positives and true positives is computed for each possible value of the threshold that is used to classify a locus under selection (see, e.g., Fawcett 2006, for further information). For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and

its centering distribution, while in the case of BayeScan it was the Bayes factor. The ROC analysis yields a monotonic curve with no positives (true or false) at one end and all positives at the other. If a method has no classification power, the curve should be linear with slope 1, and the area under the ROC curve (AUC) should be 0.5. If a method has perfect classification power, the curve should perfectly superimpose to the left-hand and upper sides of the unit square, and the AUC should be 1. Considering positively selected loci first, the area under the ROC curve for our method is slightly larger, and closer to 1, than that obtained for BayeScan (see Figure 4B and also Figures S1 to S11). As for loci under balancing selection, our method seems slightly better than BayeScan based on the ROC analysis, although both methods lack statistical power in this set of simulated data.

Inference of selection coefficients

For dataset 5 (see Table 1), we examined the distributions of the posterior means of the parameters κ_{ij} that indicate which allele is selected for. Here, from the hypotheses of our simulation model, $\kappa_{ij} = 0$ indicates that the "blue" allele is selected for, and $\kappa_{ij} = 1$ indicates that the "red" allele is selected for. Figure 5A shows the distributions of the posterior means of κ_{ij} in each sampled deme. Consistent with our expectation, is apparent from Figure 5A that the posterior means of κ_{ij} in demes 1 and 2 ("blue" demes) are shifted towards zero, and that the posterior means of κ_{ij} in demes 3 and 4 ("red" demes) are shifted towards one. Alleles of the right "color" are therefore selected for in the right "deme". It is also reassuring to see that in demes 5 and 6 ("uncolored" demes), the posterior means of κ_{ij} are centered

around 0.5, which is consistent with the fact that neither allele should be selected for in these demes.

We further examined the posterior means of the scaled coefficients of selection $\sigma_{ij} \equiv 2N_i s_{ij}$, conditionally on κ_{ij}. By doing so, we estimate the coefficient of selection associated with the allele being effectively targeted by selection. Figure 5B shows that the posterior means of $f(\sigma_{ij}|\kappa_{ij} = 0)$ in "blue" demes, and the posterior means of $f(\sigma_{ij}|\kappa_{ij} = 1)$ in "red" demes are very close to the simulated values ($\sigma \equiv 2Ns = 5$ in dataset 5, see Table 1). On the contrary, the posterior means of σ_{ij} that was not conditioned upon κ_{ij}, is much lower, and closer to the prior distribution of the hyperparameter λ, which represents the genome-wide effect of selection over all demes and loci. Figures S12 to S16 reproduce the same outputs as in Figure 5 for datasets 1-4 and 6-11. The posterior means of the scaled coefficients of selection σ_{ij} conditionally on κ_{ij} are very close to the simulated values, for $M \geq 5$ and $\sigma/M \geq 5$ (datasets 6–11) and $M = 2$ and $\sigma/M = 10$ (dataset 3).

Last, we examined the distributions of the posterior means of κ_{ij} for the 8,000 neutral markers in dataset 5. Figures 6A shows that the posterior means of κ_{ij}, which do not depend on the "color" of the sampled demes, are all centered around 0.5. This result is consistent with the fact that neither allele should be selected for in these demes. Furthermore, the distributions of κ_{ij} for neutral markers are narrower, as compared to the posterior means of κ_{ij} for selected loci in "uncolored" demes (see Figure 5A). The distributions of κ_{ij} for neutral markers are therefore closer to the Bernoulli(0.5) prior distribution of κ_{ij}, and the discrepancy with the distributions of κ_{ij} for selected loci in "uncolored" demes certainly stems from the influence of selection occurring

for the same loci in "blue" and "red" demes. The posterior means of σ_{ij} for neutral markers (unconditionally upon κ_{ij}) are very low, and close to the prior distribution of the hyperparameter λ.

Implicitly, Figures 5 and 6 demonstrates that our model is able to give accurate measures of the scaled coefficient of selection at one locus in different demes, and therefore to provide evidence of local adaptation. This paves the way for the inference of the distribution of selection strength across populations in a landscape as will be illustrated in the next section.

Application on human data

We ran three independent Markov chains on a subset of the Stanford HGDP-CEPH Human Genome Diversity Cell Line Panel (Cann et al. 2003) SNP Genotyping Data. The data consisted in 52,631 SNPs from the HGDP-CEPH data, and two SNPs (-13910C→T and -22018G→A) known to be tightly associated with lactase persistence (Bersaglieri et al. 2004), genotyped in 23 populations from Africa and Eurasia. After 25 pilot runs of 1,000 iterations, each Markov chain was run for 100,000 updating steps, after a burn-in period of 25,000 steps. Samples were collected from the Markov chains for all the model parameters every 25 steps (thinning) to avoid autocorrelations, yielding 4,000 observations for each parameter.

Convergence was assessed by computing the multivariate extension of Gelman–Rubin's diagnostic (Brooks and Gelman 1998) on the three independent Markov chains. The Gelman–Rubin's diagnostic is based on the computation of the ratio of the pooled-chains variance over the within-chain variance, and was calculated using the coda package, version 0.14-7, (Plum-

mer et al. 2006) as implemented for R (R Development Core Team 2012). The Gelman–Rubin's diagnostic was equal to 1.01 for the hyperparameter λ and to 1.06 for the parameters θ_i, which indicates that the chains converge to the target distribution. We then combined the outputs from the three Markov chains before running the following analyses.

Figure 7A shows the distribution of the Kullback–Leibler divergence measure between the posterior distributions of the locus-specific coefficients of selection with the "centering" distribution derived from the hyperdistribution of the genome-wide effect of selection, along the chromosome 2. The two SNPs that are tightly associated with lactase persistence (-13910C→T and -22018G→A) are highlighted. These two SNPs are among the set of 3 markers with the largest KLD values along all chromosome 2. Furthermore, the nine SNPs with the largest KLD values were located 3.7 Kb and 1.0 Mb upstream of the *LCT* gene, at less than 805.2 Kb from -13910C→T and less than 813.4 Kb from -22018G→A. Figure 7B represents the distribution of the posterior means of the locus-specific selection parameter δ_j, along the chromosome 2. This figure therefore represents the variation of the strength of selection along the chromosome, and depicts a very strong signal of positive selection in the vicinity of the *LCT* gene (located from base pair 136,545,414 to 136,594,749), which encodes for the enzyme lactase-phlorizin hydrolase and is associated with adult-type hypolactasia.

Figure 8A shows shows the distribution of the scaled coefficients of selection σ_{ij} (conditionally on κ_{ij} indicating allele -13910C→T to be targeted by selection) across African and Eurasian populations. The maps from Figure 8 were extrapolated by kriging using the R package fields (Fields Development

Core Team 2006), version 6.6.3. It is obvious from Figure 8A that the intensity of selection is very strong in Europe and around the Indus valley, and attains similar levels in both geographic regions. Interestingly, there was no evidence of similar selection strength in these regions, from the examination of the spatial distribution of the frequency of allele -13910C→T (Figure 8B).

Discussion

Detection of selection

We developed a hierarchical-Bayesian method, implemented via Markov chain Monte Carlo (MCMC), that considers explicitly the effect of genic selection on the distribution of single nucleotide polymorphisms (SNPs). Previous approaches based on the Multinomial-Dirichlet model for allele frequencies in an island model of population structure (see, e.g., Beaumont and Balding 2004; Riebler et al. 2008; Foll and Gaggiotti 2008; Guo et al. 2009), decomposed population-specific F_{ST} parameters into a locus-specific component (α_i) shared by all populations, and a population-specific component (β_j) shared by all loci. Deciding whether a locus is targeted by selection amounts, in those models, in testing whether the locus-specific effects (α_i) differ significantly from zero. The tests differ among these methods: Beaumont and Balding (2004) adopted a simple informal criterion assuming that α_i is significantly different from zero at some level P if its equal-tailed $100(1-P)\%$ posterior interval excludes zero. Riebler et al. (2008) introduced a Bernoulli-distributed auxiliary variable to indicate whether or not a locus is targeted by selection. Both approaches were criticized by Foll and Gaggiotti (2008), who proposed instead to use a reversible-jump Markov chain Monte Carlo algorithm, which estimates the posterior probabilities of two alternative models, a purely neutral one ($\alpha_i = 0$) and one including selection ($\alpha_i \neq 0$). In an attempt to account for the local correlation among loci for high-resolution genomic data, Guo et al. (2009) extended the above methods using conditional autoregressive models, and proposed an approach that measures divergence between the posterior distributions of locus-specific effects and the common

F_{ST} with the Kullback-Leibler divergence measure.

Here, we did not attempt to implement Bayesian model selection within the MCMC algorithm, as originally suggested by Beaumont and Balding (2004) and implemented in Foll and Gaggiotti (2008). Instead, we considered that each and every locus in a genome are selected, to some extent. To that end, we considered a simple genic model of selection where, at each locus, one allele provides a selective advantage. Since we defined a parameter that indicate which allele is selected for, the selected allele needs not to be the same in all the sampled demes. Furthermore, the strength of selection needs not to be the same in all demes. Our approach therefore accounts for situations where selection is acting in some populations, but not all, possibly in opposite direction (with alternative alleles being selected for in different environments). It is therefore particularly relvant to detect the signatures of local adaptation in subdivided populations.

Like Beaumont and Balding (2004), Riebler et al. (2008), Foll and Gaggiotti (2008) and Guo et al. (2009) who considered population-specific effects on F_{ST}, we considered in our model that the distribution of allele frequency depends upon population-specific parameters (M_i). In an early analysis of population differentiation using the HapMap dataset, Weir et al. (2005) already showed the utility of estimating F_{ST} population-specific values. In particular, concentrating their analyses on chromosome 2, they did not find any outstanding peak of population average F_{ST} around the LCT gene, although there was a clear elevation of the population specific F_{ST} values for Caucasians of European descent and European Americans. Yet, because their analysis used moment-based estimates of F_{ST} (Weir and Hill 2002), they

could not provide a statistical criterion to decide which loci were outliers of the empirical, genome-wide distribution of F_{ST}.

Here, because we assume a hierarchical Bayesian model, where the locus- and population-specific parameters of selection depend upon a locus-specific hyperparameters δ_j that gives the population-wide effect of selection at a particular locus, it is natural to use the posterior distribution of the hyperparameters δ_j as a means to classify markers as outliers or non-outliers. We indeed expect the posterior density of δ_j to be shifted toward zero if the jth marker is neutral, and toward positive values if the jth marker is targeted by selection. To do so, it would be possible to follow Beaumont and Balding (2004) and adopt a simple informal criterion assuming that δ_j is significantly different from zero at some critical level P if its equal-tailed $100(1-P)\%$ posterior interval excludes zero. Yet, this approach would neglect the genome-wide effect of selection, which in our model is driven by the hyperparameter λ. We therefore proposed to compare the posterior distributions of the locus-specific coefficients of selection δ_j with the "centering" distribution derived from the hyperdistribution with parameter λ. To that end, we used the Kullback–Leibler divergence (KLD) to measure the divergence between these two distributions. Following Guo et al. (2009), we calibrated this measure by comparing values with those produced from the divergence between a biased and a fair coin, and found that KLD = 0.830 (corresponding to the divergence between a fair and a biased coin which gives a head with probability 0.05) generally provided low false positive and false negative rates (Table 2, Figure 3, Figures S1-S10).

We used ROC analyses, as in Riebler et al. (2008), to compare our model

with BayeScan (Foll and Gaggiotti 2008). Based on the area under the ROC curve (AUC) our method performed slightly better than BayeScan (Figure S1 to S10), except for a single dataset (dataset 4, see Figure S4). Since BayeScan was shown to outperform Beaumont and Balding's (2004) approach, as well as some other popular moment-based methods using dominant markers (Pérez-Figueroa et al. 2010), we may therefore conclude that our approach represents a substantial improvement to the population genomicist's toolbox.

Not surprisingly, we found that our method has no statistical power to identify loci under balancing selection (see Figure 3 and Table 2). Since our genic selection model only allows for positive selection, this was somewhat expected. Beaumont and Balding (2004) concluded from simulations that their method could not identify loci under balancing selection, even for very strong selection. Although Foll and Gaggiotti (2008) showed that microsatellites could be used to detect balancing selection, especially with data sets containing a large number of sampled populations, they needed 10 populations with SNPs to achieve the same rate of detection (Foll and Gaggiotti 2008).

Although the likelihood from equation (1) can be integrated analytically over the distribution of unknown population frequencies given by equations (2) and (3), we found that it increases the computational burden significantly. This is so, because additional gamma and confluent hypergeometric functions are then required to compute the posterior distribution of the model parameters.

Inference of selection

Because our model accounts explicitly for positive selection, it can not only be used to detect the genomic signatures of selection, but also to measure the strength of selection along the genome. Contrary to previous approaches that approximated selection as a locus-specific effect in a ad-hoc inverse linear regression model (Balding et al. 1996) or a reduction in migration rate (see, e.g., Bazin et al. 2010), we introduced explicitly a scaled coefficient of selection $\sigma_{ij} \equiv 2N_i s_{ij}$ for locus j in deme i, where s_{ij} represents the relative gain in fitness brought by a positively selected allele. We found that the posterior means of the scaled coefficients of selection σ_{ij} (conditionally on κ_{ij}) were close to the simulated value for positively selected loci, although slightly overestimated (Figure 5, Figures S11-S15). We also found that the variation of σ_{ij} across populations with different selection regimes was remarkably well inferred, with selected loci exhibiting large coefficients of selection in the "colored" demes, and small coefficients of selection in "uncolored" demes (Figure 5, Figures S11-S15).

Figure S16A further confirms that, in the absence of selection (dataset 12, see Table 1), the posterior means of κ_{ij} are all centered around 0.5 and narrower as compared to datasets that include positively selected loci (compare, e.g., with Figure 6). This is consistent with the posterior means of κ_{ij} being closer to the Bernoulli(0.5) prior distributions for these parameters. In the absence of selection, the posterior means of σ_{ij} for neutral markers (unconditionally upon κ_{ij}) are very low and largely below to the prior distribution of the hyperparameter λ (Figure S16B).

Selection at the LCT *gene*

The region around the *LCT* gene that allows lactose tolerance to persist into adulthood is a very-well known example of selection in humans (Sabeti et al. 2006). The first causative polymorphim described was the -13910C→T mutation (Enattah et al. 2002), which lays in the cis-acting regulatory element located in the 13th intron of a neighboring gene, *MCM6*. Although this single mutation of purported western Eurasian origin accounts for much of observed lactase persistence outside Africa, multiple independent mutations in the same region upstream of the *LCT* gene have been associated with this trait in pastoralists from Saudi Arabia (Enattah et al. 2008) and Africa (Tishkoff et al. 2007). The lactase persistence allele at the *LCT* locus lies on a haplotype that is common in Europeans but that extends largely undisrupted for more than 1 Mb, much farther than is typical for an allele of that frequency (Bersaglieri et al. 2004).

Our analyses point to a very strong signal of positive selection between 3.7 Kb and 1.0 Mb upstream of the *LCT* gene. Furthermore, we found the strongest selection coefficients in Europe and in the Indus Valley (Figure 8A), which matches the interpolated map of lactase persistence phenotype frequencies in the Old World (Itan et al. 2010). Our results therefore confirm those of Romero et al. (2012), who found that the -13910C→T mutation explains a substantial proportion of lactase persistence in the Indian subcontinent. Most interestingly, Romero et al. (2012) showed that the -13910C→T mutation in India is identical by descent to the European allele and is associated with the same extended haplotype in both populations, which strongly suggests that the origin of the -13910C→T mutation is shared in Europe and

India. These results are consistent with the high levels of present-day milk consumption in India, and with archaeological and genetic evidence for the independent domestication of cattle in the Indus valley c.a. 7,000 years ago (Romero et al. 2012).

We found strong coefficients of selection acting on the -13910C→T allele, with $\sigma \equiv 2Ns$ ranging from 58.32 (French) to 95.47 (Orcadians) in Europe and from 4.99 (Kalash) to 72.04 (Balochi). There has been previous attempts to measure the strength of selection acting at the LCT gene. For example, Aoki (1986) predicted that a selection coefficient $s > 5\%$ would be necessary to explain the observed allele frequency of the -13910C→T allele, assuming that this mutation appeared 6,000 years ago in a population of effective size 500, which would give $\sigma \equiv 2Ns = 50$.Bersaglieri et al. (2004) estimated the coefficient of selection s to be 1-15% for a new mutation arising in a population of effective size comprised between 500 and 5,000. More recently, Tishkoff et al. (2007) estimated selection intensity by matching simulated data under a coalescent framework to the observed cM span and the observed frequency of the allele targeted by selection. They found extremely recent and strong positive selection in many African populations ($\sigma \equiv 2Ns$ ranging from 800 to 1,940 assuming an effective population size N of 10,000). Modelling a geographical structuring of selection pressure by latitude, Gerbault et al. (2009) found selection coefficients in the range between 0.8 and 1.8% (Gerbault et al. 2011), also assuming a carrying capacity of 10,000. However, assuming an effective population size N of 10,000 may largely overestimates $\sigma \equiv 2Ns$ (see Tenesa et al. 2007, for more accurate estimates of effective size based on measures of linkage disequilibrium). Last, using a spatially ex-

plicit model and approximate Bayesian computation (Beaumont et al. 2002), Itan et al. (2009) estimated coefficients of selection to lay in the range of 5.2-15.9%. The difficulty in comparing these values is that strong hypotheses about the effective population size need to be made. It is clear from the stationary density of the diffusion process in eq. (2), that the two parameters s and N are not identifiable. Estimating s therefore requires informative priors on N. Furthermore, the population size considered in our model is the local effective size of a deme, not the effective size of the total population.

Last, for the purpose of comparison, we ran a BayeScan analysis of the 52,631 SNPs from the HGDP-CEPH data, using the same MCMC parameters (number of pilot runs, burn-in, chain length, etc.) as in the previous study (Figure S17A). It is clear from this figure that BayeScan Bayes factors attain their maximum (arbitrary) value for a substantial number of markers fro which the Kullback–Leibler divergence provides no evidence of selection (Figures S17B-C). Furthermore, this effect of "saturation" observed in Figure S17A may prevent the identification of genomic regions potentially targeted by selection (see Figure 7), which advocates the use of the Kullback–Leibler divergence measure.

Acknowledgements

This work was supported by a Biotechnology and Biological Sciences Research Council grant awarded to MAB and KJD. MAB was supported by a Natural Research Council Advanced Fellowship. RV was supported by the ANR programmes NUTGENEVOL 07-BLAN-0064 and EMILE 09-BLAN-0145-01.

Literature Cited

Abramowitz, M., and I. A. Stegun. 1965. Handbook of Mathematical Functions. Dover Publication, Inc., New York.

Akey, J. M., G. Zhang, L. Jin, and M. D. Shriver. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12:1805–1814.

Aoki, K. A. 1986. Stochastic model of gene-culture coevolution suggested by the ćulture historical hypothesisfor the evolution of adult lactose absorption in humans. Proceedings of the National Academy of sciences USA 83:2929–2933.

Balding, D. J., M. Greenhalgh, and R. A. Nichols. 1996. Population genetics of STR loci in caucasians. International Journal of Legal Medecine 108:300–305.

Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci. 2008. Natural selection has driven population differentiation in modern humans. Nature Genetics .

Barton, N. H., and M. Turelli. 1989. Evolutionary quantitative genetics: how little do we know? Annual Review of Genetics 23:337–370.

Bazin, E., M. A. Beaumont, and K. J. Dawson. 2010. Likelihood-free inference of population structure and local adaptation in a bayesian hierarchical model. Genetics 185:587–602.

Beaumont, M. A. 2005. Adaptation and speciation: what can F_{ST} tell us? Trends in Ecology & Evolution 20:435–440.

Beaumont, M. A., and D. J. Balding. 2004. Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology 13:969–980.

Beaumont, M. A., and R. A. Nichols. 1996. Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London Series B Biological Sciences 263:1619–1626.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. Approximate bayesian computation in population genetics. Genetics 162:2025–2035.

Bersaglieri, T., P. Sabeti, N. Patterson, T. Vanderploeg, S. Schaffner, J. Drake, M. Rhodes, D. Reich, and J. Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics 74:1111–1120.

Black, W. C., C. F. Baer, M. F. Antolin, and N. M. DuTeau. 2001. Population genomics: genome-wide sampling of insect populations. Annual Review of Entomology 46:441–469.

Brooks, S., and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7:434–455.

Bürger, R. 2000. The Mathematical Theory of Selection, Recombination, and Mutation. John Wiley and sons, Ltd, Chichester, England.

Bustamante, C. D., R. Nielsen, and D. L. Hartl. 2003. Maximum likelihood and bayesian methods for estimating the distribution of selective effects

among classes of mutations using DNA polymorphism data. Theoretical Population Biology 63:91–103.

Bustamante, C. D., J. Wakeley, S. A. Sawyer, and D. L. Hartl. 2001. Directional selection and the site-frequency spectrum. Genetics 159:1779–1788.

Cann, H. M., C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. 2003. A human genome diversity cell line panel. Science 296:261–262.

Donnelly, P., M. Nordborg, and P. Joyce. 2001. Likelihoods and simulation methods for a class of nonneutral population genetics models. Genetics 159:853–867.

Donnelly, P., and S. Tavaré. 1995. Coalescents and genealogical structure under neutrality. Annual Review of Genetics 29:401–21.

Enattah, N. S., T. G. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen, H. Rasinpera, H. El-Shanti, J. K. Seo, M. Alifrangis, I. F. Khalil, A. Natah, A. Ali, S. Natah, D. Comas, S. Q. Mehdi, L. Groop, E. M. Vestergaard, F. Imtiaz, M. S. Rashed, B. Meyer, J. Troelsen, and L. Peltonen. 2008. Independent introduction of two lactase-persistence alleles into human pop-

ulations reflects different history of adaptation to milk culture. American Journal of Human Genetics 82:57–72.

Enattah, N. S., T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Järvelä. 2002. Identification of a variant associated with adult-type hypolactasia. Nature Genetics 30:233–237.

Ethier, S. N., and T. Nagylaki. 1988. Diffusion approximations of markov chains with two time scales and application to population genetics, II. Advances in Applied Probabilities 20:525–545.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognition Letter 27:882–891.

Fields Development Core Team. 2006. fields: Tools for Spatial Data. National Center for Atmospheric Research Boulder, CO.

Foll, M., and O. Gaggiotti. 2008. A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. Genetics 180:977–993.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian Data Analysis. 2nd edition. Chapman & Hall, New York.

Gerbault, P., A. Liebert, Y. Itan, A. Powell, M. Currat, J. Burger, D. M. Swallow, and D. M. Thomas. 2011. Evolution of lactase persistence: an example of human niche construction. Philosophical Transactions of the Royal Society Series B 366:863–877.

Gerbault, P., C. Moret, M. Currat, and A. Sanchez-Mazas. 2009. Impact

of selection and demography on the diffusion of lactase persistence. PLoS ONE 4:e6369.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov Chain Monte Carlo in Practice. 2nd edition. Chapman & Hall, New York.

Goldstein, D. B., and L. Chikhi. 2002. Human migrations and population structure: what we know and why it matters. Annual Review of Human Genetics 3:129–152.

Guo, F., D. Dey, and H. K.E. 2009. A bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. Journal of the American Statistical Association 104:142–154.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and C. D. R. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

Innan, H., and Y. Kim. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of sciences USA 101:10667–10672.

Itan, Y., B. L. Jonesand, C. J. E. Ingram, D. M. Swallow, and M. G. Thomas. 2010. A worldwide correlation of lactase persistence phenotype and genotypes. BMC Evolutionary Biology 10:36.

Itan, Y., A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas. 2009. The origins of lactase persistence in europe. PLoS Computational Biology 5:e1000491.

Jeffreys, H. 1961. Theory of Probability. 3rd edition. Oxford University Press, Oxford.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90:773–795.

Kim, Y., and W. Stephan. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160:765–777.

Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. Genetics 74:175–195.

Loader, C. R. 1996. Local Regression and Likelihood. Springer, New York.

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

Nei, M., and T. Maryuyama. 1975. Lewontin–Krakauer test for neutral genes. Genetics 80:395.

Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. Heredity 86:641–647.

—. 2005. Disclosure of variation. Nature 434:288–289.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. 2005*a*. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biology 3:e170.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005*b*. Genomic scans for selective sweeps using SNP data. Genome Research 15:1566–1575.

Nordborg, M. 1997. Structured coalescent processes on different time scales. Genetics 146:1501–1514.

Ntzoufras, I. 2009. Bayesian Modeling Using WinBugs. John Wiley & Sons, Inc., Hoboken, NJ.

Payseur, B. A., A. D. Cutter, and M. W. Nachman. 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Molecular Biology and Evolution 19:1143–1153.

Pearson, J. 2009. Computation of Hypergeometric Functions. Ph.D. thesis University of Oxford.

Pérez-Figueroa, A., M. J. García-Pereira, M. Saura, E. Rolán-Alvarez, and A. Caballero. 2010. Comparing three different methods to detect selective loci using dominant markers. Journal of Evolutionary Biology 23:2267–2276.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. Coda: output analysis and diagnostics for MCMC. R News 6:7–11.

Przeworsky, M., G. Coop, and J. Wall. 2005. The signature of positive selection on standing variation. Evolution 59:2312–2323.

R Development Core Team. 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

Riebler, A., L. Held, and W. Stephan. 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. Genetics 178:1817–1829.

Robertson, A. 1975. Remarks on the Lewontin–Krakauer test. Genetics 80:396.

Romero, I. G., C. B. Mallick, A. Liebert, F. Crivellaro, G. Chaubey, Y. Itan, M. Metspalu, M. Eaaswarkhanth, R. Pitchappan, R. Villems, D. Reich, L. Singh, K. Thangaraj, M. G. Thomas, D. M. Swallow, M. M. Lahr, and T. Kivisild1. 2012. Herders of indian and european cattle share their predominant allele for lactase persistence. Molecular Biology and Evolution 29:249–260.

Ross, K. G., D. D. Shoemaker, M. J. B. Krieger, J. DeHeer, and L. Keller. 1999. Assessing genetic structure with multiple classes of molecular markers: A case study involving the introduced fire ant *Solenopsis invicta*. Molecular Biology and Evolution 16:525–543.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. Science 312:1614–1620.

Tang, K., K. R. Thornton, and M. Stoneking. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biology 5:e171.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, M. E. Goddard, and V. P. M. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Research .

The International HapMap Consortium. 2003. The international HapMap project. Nature 426:789–796.

—. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas. 2007. Convergent adaptation of human lactase persistence in africa and europe. Nature Genetics 39:31–40.

Vitalis, R., P. Boursot, and K. Dawson. 2001. Interpretation of variation across marker loci as evidence of selection. Genetics 158:1811–1823.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biology 4:e72.

Wakeley, J. 1999. Nonequilibrium migration in human history. Genetics 153:1863–1871.

Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Research 15:1468–1476.

Weir, B. S., and W. G. Hill. 2002. Estimating f-statistics. Annual Review of Genetics 36:721–50.

Wilding, C. S., R. K. Butlin, and J. Grahame. 2001. Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. Journal of Evolutionary Biology 14:611–619.

Williamson, S. H., R. Hernandez, A. Fledel-Alon, Z. L., R. Nielsen, and C. D. Bustamante. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proceedings of the National Academy of sciences USA 102:7882–7887.

Wright, S. 1931. Evolution in mendelian populations. Genetics 16:97–159.

—. 1935. Evolution in populations in approximate equilibrium. Journal of Genetics 30:257–266.

—. 1949. Adaptation and selection. *in* G. L. Jepson, G. G. Simpson, and E. Mayr, eds. Genetics, Paleontology, and Evolution Pages 365–389. University Press, Princeton.

—. 1969. Evolution and the Genetics of Populations. Volume II. The Theory of Gene Frequencies. University of Chicago Press, Chicago.

Yang, Z. 2005. Bayesian inference in molecular phylogenetics. *in* O. Gascuel, ed. Mathematics of Evolution and Phylogeny Pages 63–90. Oxford University Press, Oxford.

Table 1: Parameters of simulated data sets. All the simulations were performed according to an island model with $n_d = 50$ demes, each made of $N = 250$ diploid individuals (500 genes). The number of loci simulated under positive selection ("pos."), balancing selection ("bal.") and neutrality ("neut.") is indicated. The total number of sampled demes is also given, together with the composition of the sample: (2,2,2) indicates that the sample of six demes consisted in 2 "blue" demes, 2 "red" demes and 2 "uncolored" demes.

							Markers			Sampled demes	
Dataset	N	m	$M \equiv 4Nm$	s	$\sigma \equiv 2Ns$	σ/M	Pos.	Bal.	Neut.	number	categories
1	250	0.002	2	0.008	4	2	1000	1000	8000	6	(2,2,2)
2	250	0.002	2	0.02	10	5	1000	1000	8000	6	(2,2,2)
3	250	0.002	2	0.04	20	10	1000	1000	8000	6	(2,2,2)
4	250	0.005	5	0.02	10	2	1000	1000	8000	6	(2,2,2)
5	250	0.005	5	0.05	25	5	1000	1000	8000	6	(2,2,2)
6	250	0.005	5	0.1	50	10	1000	1000	8000	6	(2,2,2)
7	250	0.01	10	0.04	20	2	1000	1000	8000	6	(2,2,2)
8	250	0.01	10	0.1	50	5	1000	1000	8000	6	(2,2,2)
9	250	0.01	10	0.2	100	10	1000	1000	8000	6	(2,2,2)
10	250	0.005	5	0.05	25	5	1000	1000	8000	3	(1,1,1)
11	250	0.005	5	0.05	25	5	1000	1000	8000	12	(4,4,4)
12	250	0.005	5	–	–	–	0	0	10000	6	(0,0,6)

Table 2: Proportion of positively selected loci, loci under balancing selection and neutral markers that were classified as non-outliers or outliers, using the threshold KLD = 0.830 (resp. 1.614), which equals the KLD between two Bernoulli distributions corresponding to flipping a fair coin and a biased coin that gives a head with probability 0.05 (resp. 0.01)

	Positive selection		Balancing selection		Neutrality	
Dataset	non-outliers (%)	outliers (%)	non-outliers (%)	outliers (%)	non-outliers (%)	outliers (%)
1	92.1 (98.7)	7.9 (1.3)	99.8 (100)	0.2 (0.0)	99.0 (100.0)	1.0 (0.0)
2	58.7 (95.0)	41.3 (5.0)	100.0 (100)	0.0 (0.0)	98.6 (100.0)	1.4 (0.0)
3	26.6 (80.7)	73.4 (19.3)	100.0 (100)	0.0 (0.0)	99.2 (100.0)	0.8 (0.0)
4	87.9 (99.3)	12.1 (0.7)	100.0 (100)	0.0 (0.0)	99.9 (100.0)	0.1 (0.0)
5	8.3 (62.4)	97.7 (37.6)	100.0 (100)	0.0 (0.0)	99.5 (100.0)	0.5 (0.0)
6	0.2 (21.3)	99.8 (78.7)	100.0 (100)	0.0 (0.0)	99.7 (100.0)	0.3 (0.0)
7	61.5 (96.3)	38.5 (3.7)	100.0 (100)	0.0 (0.0)	100.0 (100.0)	0.0 (0.0)
8	0.4 (32.4)	99.6 (67.6)	100.0 (100)	0.0 (0.0)	99.9 (100.0)	0.1 (0.0)
9	0.0 (2.6)	100.0 (97.4)	100.0 (100)	0.0 (0.0)	100.0 (100.0)	0.0 (0.0)
10	53.0 (94.4)	47.0 (5.6)	100.0 (100)	0.0 (0.0)	99.7 (100.0)	0.3 (0.0)
11	0.1 (9.0)	99.9 (91.0)	100.0 (100)	0.0 (0.0)	99.5 (100.0)	0.5 (0.0)
12	–	–	–	–	99.3 (100.0)	0.7 (0.0)

A Appendix

Details on the componentwise Markov chain Monte Carlo algorithm

Here we provide the computational details for the componentwise Markov chain Monte Carlo updates. Our aim is to sample from the joint posterior distribution of $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda|\mathbf{n})$, which is specified by equation (4) and by the directed acyclic graph (DAG) in Figure 1. To do so, we use a combination of the Metropolis–Hastings algorithm and the Gibbs sampler for generating observations from $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda|\mathbf{n})$ using outputs from a Markov chain (see, e.g., Gelman et al. 2004).

Each Markov chain is initialized with random values of the parameters drawn from their prior densities, except for the parameters p_{ij}, for which the observed frequencies are used, and the parameters π_js, for which the Laplace values are calculated from the dataset frequencies. The updating sequence is as follows: (*i*) all $L \times n_d$ parameters p_{ij}; (*ii*) all n_d parameters M_i; (*iii*) all L parameters π_j; (*iv*) the hyperparameter λ; (*v*) all L hyperparameters δ_j; (*vi*) all $L \times n_d$ parameters σ_{ij}; (*vii*) all $L \times n_d$ parameters κ_{ij}. Since the full posterior distribution of the model can be decomposed as a product over loci and over populations (see equation 4), each update only requires the recomputation of the relevant terms of the distribution $f(\mathbf{M}, \boldsymbol{\pi}, \boldsymbol{\kappa}, \boldsymbol{\sigma}, \boldsymbol{\delta}, \lambda|\mathbf{n})$. This improves the computational efficiency of the algorithm considerably.

The confluent hypergeometric, or Kummer's, functions ${}_1F_1(a; b; z)$ (see, e.g., Abramowitz and Stegun 1965, p. 504) were computed following a procedure proposed by Pearson (2009), which is based on the power series defi-

nition of the function:

$$ {}_1F_1(a;b;z) = \sum_{j=0}^{\infty} \underbrace{\frac{(a)_j}{(b)_j}\frac{z^j}{j!}}_{A_j}, \tag{A.1} $$

where, for some parameter p, the Pochhammer symbol $(p)_j$ is defined as:

$$ (p)_0 = 1, \quad (p)_j = p(p+1)\dots(p+j-1), \quad \text{for } j = 1, 2, \dots \ . \tag{A.2} $$

The computation of the terms of the power series in equation (A.1) can then be carried out using the following procedure:

$$ \begin{aligned} &A_0 = S_0 = 1, \\ &A_{j+1} = A_j \times \frac{a+j}{b+j} \times \frac{z}{j+1}, \\ &S_{j+1} = S_j + A_{j+1}, \quad \text{for } j = 1, 2, \dots \end{aligned} \tag{A.3} $$

where A_j represents the $(j+1)$th term of the power series in equation (A.1), and S_j represents the sum of the first $(j+1)$ terms. The computation was stopped when both $|A_N|/|S_{N-1}| < 10^{-12}$ and $|A_{N+1}|/|S_N| < 10^{-12}$. This criterion is equivalent to truncating the series in equation (A.1), and require that two consecutive terms to be small compared to the sum already computed.

Updating p_{ij}: The parameters p_{ij} are updated iteratively in each deme, one locus at a time. In the ith deme, at locus j, one allele is chosen at random from a Bernoulli trial with probability 0.5. The new allele frequency p'_{ij} is chosen as a random variable drawn from a uniform distribution around the

current value p_{ij}:

$$p'_{ij} \sim U\left(p_{ij} - \Delta_p, p_{ij} + \Delta_p\right). \tag{A.4}$$

The size of the interval Δ_p is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40 (see, e.g., Gilks et al. 1996). Since p_{ij} is a frequency comprised between 0 and 1, if p'_{ij} is outside the interval $[0, 1]$, the excess is reflected back into the interval; that is, if $p'_{ij} < 0$ then p'_{ij} is reset to its absolute value $|p'_{ij}|$, and if $p'_{ij} > 1$ then p'_{ij} is reset to $2 - p'_{ij}$. This proposal is symmetric (Yang 2005). The updated allele frequency p'_{ij} is therefore accepted according to the appropriate Metropolis probability, which reads:

$$1 \wedge \frac{\mathcal{L}(p'_{ij}; \mathbf{n}_{ij})\psi(p'_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij})}{\mathcal{L}(p_{ij}; \mathbf{n}_{ij})\psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij})}. \tag{A.5}$$

Equation (A.5) can be rewritten as

$$1 \wedge \exp\left[\sigma_{ij}\left(\tilde{p}'_{ij} - \tilde{p}_{ij}\right)\right] \frac{p'^{x_{ij}+M_i\pi_j-1}_{ij}(1-p'_{ij})^{(n_{ij}-x_{ij})M_i+(1-\pi_j)-1}}{p^{x_{ij}+M_i\pi_j-1}_{ij}(1-p_{ij})^{(n_{ij}-x_{ij})M_i+(1-\pi_j)-1}}, \tag{A.6}$$

where $\tilde{p}'_{ij} \equiv \kappa_{ij}(1 - p'_{ij}) + (1 - \kappa_{ij})p'_{ij}$.

Updating M_i: The parameters M_i are updated iteratively, one deme at a time. The proposed value M'_i is drawn from a lognormal distribution with median equal to the current value M_i, i.e.:

$$q(M_i \to M'_i) = \frac{1}{M'_i \nu_M \sqrt{2\pi}} \exp\left(\frac{-\ln(M'_i/M_i)^2}{2\nu_M^2}\right), \tag{A.7}$$

where ν_M is the standard deviation on the log scale. The standard deviation ν_M is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities (which is sometimes referred to as the "Hastings term": see, e.g., Gelman et al. 2004, p. 291). This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(M_i' \to M_i)/q(M_i \to M_i')$ reduces to M_i'/M_i. In order to avoid computational problems with excessively small or large M_i values, all moves falling outside the interval $[0.001, 1000]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value M_i' is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$1 \wedge \frac{\left[\prod_{j=1}^{L} \psi(p_{ij}; M_i', \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij})\right] f(M_i') q(M_i' \to M_i)}{\left[\prod_{j=1}^{L} \psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij})\right] f(M_i) q(M_i \to M_i')}. \tag{A.8}$$

Equation (A.8) can be rewritten as

$$1 \wedge \left[\frac{\Gamma(M_i)}{\Gamma(M_i')}\right]^{L} \frac{\prod_{j=1}^{L} \Gamma(M_i\pi_j)\Gamma(M_i(1-\pi_j)) {}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij}) p_{ij}^{M_i'\pi_j}(1-p_{ij})^{M_i'(1-\pi_j)}}{\prod_{j=1}^{L} \Gamma(M_i'\pi_j)\Gamma(M_i'(1-\pi_j)) {}_1F_1(M_i'\tilde{\pi}_{ij}; M_i'; \sigma_{ij}) p_{ij}^{M_i\pi_j}(1-p_{ij})^{M_i(1-\pi_j)}} \tag{A.9}$$

Updating π_j: The parameters π_j are updated iteratively, one locus at a time. In the ith deme, at locus j, one allele is chosen at random from a Bernoulli trial with probability 0.5. The proposed allele frequency π_j' is chosen as a

random variable drawn from a uniform distribution around the current value π_j:

$$\pi'_j \sim U\left(\pi_j - \Delta_\pi, \pi_j + \Delta_\pi\right). \tag{A.10}$$

The size of the interval Δ_π is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Since π_j is a frequency comprised between 0 and 1, if π'_j is outside the interval $[0, 1]$, the excess is reflected back into the interval; that is, if $\pi'_j < 0$ then π'_j is reset to its absolute value $|\pi'_j|$, and if $\pi'_j > 1$ then π'_j is reset to $2 - \pi'_j$. This proposal is symmetric, and the updated allele frequency π'_j is therefore accepted according to the appropriate Metropolis probability, which reads:

$$1 \wedge \frac{\left[\prod_{i=1}^{n_\mathrm{d}} \psi(p_{ij}; M_i, \boldsymbol{\pi}'_j, \kappa_{ij}, \sigma_{ij})\right] f(\boldsymbol{\pi}'_j)}{\left[\prod_{i=1}^{n_\mathrm{d}} \psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij})\right] f(\boldsymbol{\pi}_j)}. \tag{A.11}$$

Equation (A.11) can be rewritten as

$$1 \wedge \frac{\prod_{i=1}^{n_\mathrm{d}} \Gamma(M_i\pi_j)\Gamma(M_i(1-\pi_j)) {}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij}) p_{ij}^{M_i\pi'_j}(1-p_{ij})^{M_i(1-\pi'_j)}}{\prod_{i=1}^{n_\mathrm{d}} \Gamma(M_i\pi'_j)\Gamma(M_i(1-\pi'_j)) {}_1F_1(M_i\tilde{\pi}'_{ij}; M_i; \sigma_{ij}) p_{ij}^{M_i\pi_j}(1-p_{ij})^{M_i(1-\pi_j)}}, \tag{A.12}$$

where $\tilde{\pi}'_{ij} \equiv \kappa_{ij}(1-\pi'_j) + (1-\kappa_{ij})\pi'_j$.

Updating λ*:* The proposed value of the hyperparameter λ' is drawn from a lognormal distribution with median equal to the current value λ, i.e.:

$$q(\lambda \to \lambda') = \frac{1}{\lambda' \nu_\lambda \sqrt{2\pi}} \exp\left(\frac{-\ln(\lambda'/\lambda)^2}{2\nu_\lambda^2}\right), \tag{A.13}$$

where ν_λ is the standard deviation on the log scale. The standard deviation ν_λ is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities. This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(\lambda' \to \lambda)/q(\lambda \to \lambda')$ reduces to λ'/λ. In order to avoid computational problems with excessively small or large λ' values, all moves falling outside the interval $[0, 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value λ' is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$1 \wedge \frac{\left[\prod_{j=1}^{L} f(\delta_j|\lambda')\right] f(\lambda'|\Lambda) q(\lambda' \to \lambda)}{\left[\prod_{j=1}^{L} f(\delta_j|\lambda)\right] f(\lambda|\Lambda) q(\lambda \to \lambda')}. \tag{A.14}$$

Equation (A.14) can be rewritten as

$$1 \wedge \left(\frac{\lambda}{\lambda'}\right)^{L-1} \exp\left[(\lambda' - \lambda)\left(\frac{\sum_{j=1}^{L} \delta_j}{\lambda\lambda'} - \frac{1}{\Lambda}\right)\right] \tag{A.15}$$

Updating δ_j: The parameters δ_j are updated iteratively, one locus at a time. The proposed value of the hyperparameters δ'_j is drawn from a lognormal distribution with median equal to the current value δ_j, i.e.:

$$q(\delta_j \to \delta'_j) = \frac{1}{\delta'_j \nu_\delta \sqrt{2\pi}} \exp\left(\frac{-\ln(\delta'_j/\delta_j)^2}{2\nu_\delta^2}\right), \tag{A.16}$$

where ν_δ is the standard deviation on the log scale. The standard deviation ν_δ is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities. This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(\delta'_j \to \delta_j)/q(\delta_j \to \delta'_j)$ reduces to δ'_j/δ_j. In order to avoid computational problems with excessively small or large δ_j values, all moves falling outside the interval $[0, 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value δ'_j is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$1 \wedge \frac{\left[\prod_{i=1}^{n_\mathrm{d}} f(\sigma_{ij}|\delta'_j)\right] f(\delta'_j|\lambda) q(\delta'_j \to \delta_j)}{\left[\prod_{i=1}^{n_\mathrm{d}} f(\sigma_{ij}|\delta_j)\right] f(\delta_j|\lambda) q(\delta_j \to \delta'_j)}. \tag{A.17}$$

Equation (A.17) can be rewritten as

$$1 \wedge \left(\frac{\delta_j}{\delta'_j}\right)^{n_\mathrm{d}-1} \exp\left[\left(\delta'_j - \delta_j\right)\left(\frac{\sum_{i=1}^{n_\mathrm{d}} \sigma_{ij}}{\delta_j \delta'_j} - \frac{1}{\lambda}\right)\right] \tag{A.18}$$

Updating σ_{ij}: The parameters σ_{ij} are updated iteratively in each deme, one locus at a time. In the ith deme, at locus j, the proposed value of the parameters σ'_{ij} is drawn from a lognormal distribution with median equal to the current value σ_{ij}, i.e.:

$$q(\sigma_{ij} \to \sigma'_{ij}) = \frac{1}{\sigma'_{ij}\nu_\sigma\sqrt{2\pi}} \exp\left(\frac{-\ln(\sigma'_{ij}/\sigma_{ij})^2}{2\nu_\sigma^2}\right), \tag{A.19}$$

where ν_σ is the standard deviation on the log scale. The standard deviation ν_σ is a constant, which is adjusted during 25 short pilot runs of 1,000 iterations, in order to get acceptance rates between 0.25 and 0.40. Because the lognormal jumping rule is asymmetric, a Metropolis–Hastings update is required in which the Metropolis ratio is weighted by the ratio of the forward and reverse proposal densities. This means that when some moves are more likely to happen (because of the asymmetry of the proposal distribution), their probability of acceptance is decreased proportionately. Here, the ratio $q(\sigma'_{ij} \to \sigma_{ij})/q(\sigma_{ij} \to \sigma'_{ij})$ reduces to σ'_{ij}/σ_{ij}. In order to avoid computational problems with excessively small or large σ_{ij} values, all moves falling outside the interval $[0, 500]$ are discarded (i.e., the chain is kept unchanged). Otherwise, the proposed value σ'_{ij} is accepted according to the appropriate Metropolis–Hastings probability, which is:

$$\frac{\psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma'_{ij}) f(\sigma'_{ij}|\delta_j) q(\sigma'_{ij} \to \sigma_{ij})}{\psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij}) f(\sigma_{ij}|\delta_j) q(\sigma_{ij} \to \sigma'_{ij})}. \tag{A.20}$$

Equation (A.20) can be rewritten as

$$\frac{\sigma'_{ij}}{\sigma_{ij}} \exp\left[(\sigma'_{ij} - \sigma_{ij})\left(\tilde{p}_{ij} - \frac{1}{\delta_j}\right)\right] \frac{{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma_{ij})}{{}_1F_1(M_i\tilde{\pi}_{ij}; M_i; \sigma'_{ij})}. \tag{A.21}$$

Updating κ_{ij}: The parameters κ_{ij} are updated iteratively in each deme, one locus at a time. In the ith deme, at locus j, the variable κ_{ij}, which indicates which of the two alleles is selected for, is updated using Gibbs sampling based on the conditional posterior distribution:

$$f(\kappa_{ij}|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]}) \propto \psi(p_{ij}; M_i, \boldsymbol{\pi}_j, \kappa_{ij}, \sigma_{ij}) f(\kappa_{ij}), \tag{A.22}$$

where $\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]}$ represents all the model parameters but κ_{ij}. Since κ_{ij} can only take two integer values (0 and 1), it can be shown that:

$$\Pr(\kappa_{ij} = 0|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]}) \propto \frac{1}{2}\left[\frac{\exp\left[\sigma_{ij} p_{ij}\right]}{{}_1F_1(M_i\pi_j; M_i; \sigma_{ij})}\right], \tag{A.23}$$

and

$$\Pr(\kappa_{ij} = 1|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]}) \propto \frac{1}{2}\left[\frac{\exp\left[\sigma_{ij}(1 - p_{ij})\right]}{{}_1F_1(M_i(1-\pi_j); M_i; \sigma_{ij})}\right]. \tag{A.24}$$

Therefore, the conditional posterior distribution of $\left(\kappa_{ij}|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]}\right)$ from equation (A.22) can be rewritten as

$$\left(\kappa_{ij}|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]}\right) \sim \text{Bernoulli}\left(\rho\right), \tag{A.25}$$

where

$$\begin{aligned} \rho &\equiv \frac{\Pr(\kappa_{ij} = 0|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]})}{\Pr(\kappa_{ij} = 0|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]}) + \Pr(\kappa_{ij} = 1|\boldsymbol{\theta}_{[-\boldsymbol{\kappa_{ij}}]})} \\ &= \left[1 + \frac{{}_1F_1(M_i\pi_{ij}; M_i; \sigma_{ij})}{{}_1F_1(M_i(1-\pi_{ij}); M_i; \sigma_{ij})} \exp\left[\sigma_{ij}(1 - 2p_{ij})\right]\right]^{-1}. \end{aligned} \tag{A.26}$$

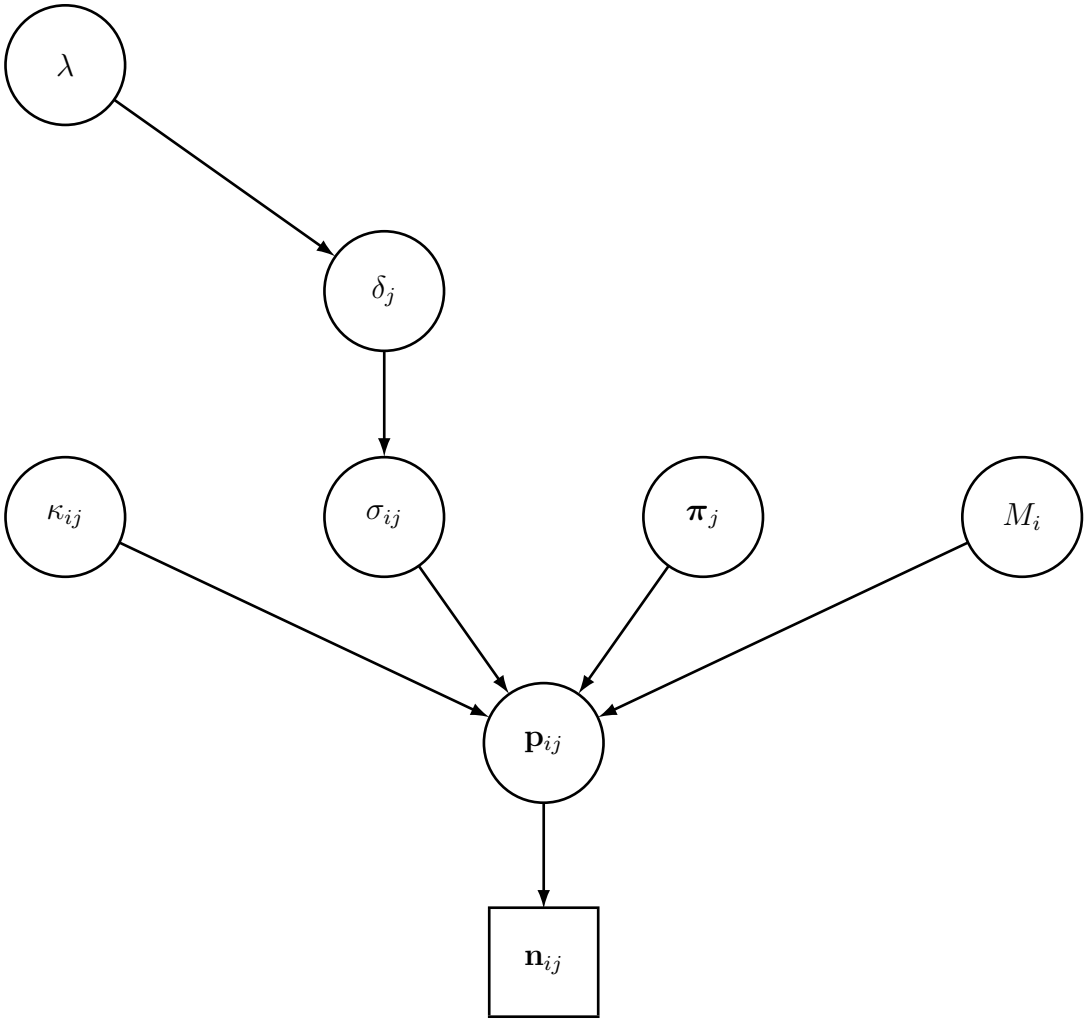


Figure 1: Directed acyclic graph (DAG) of the hierarchical Bayesian model.

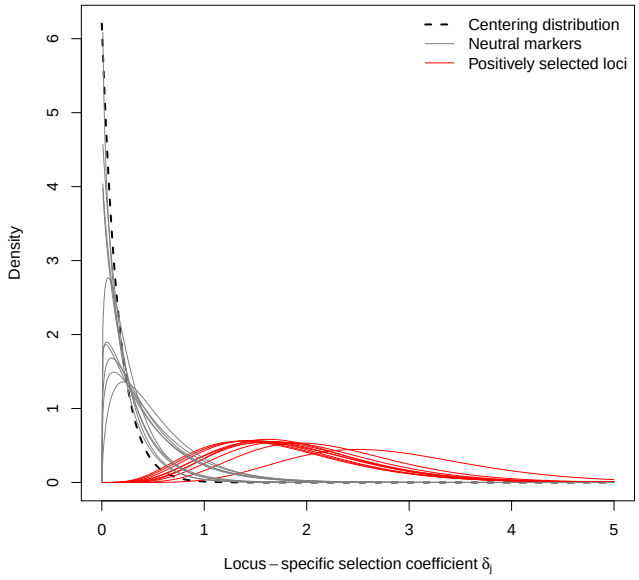


Figure 2: Posterior densities of the locus-specific hyperparameter δ_j for neutral markers (in grey) and positively selected loci (in red).

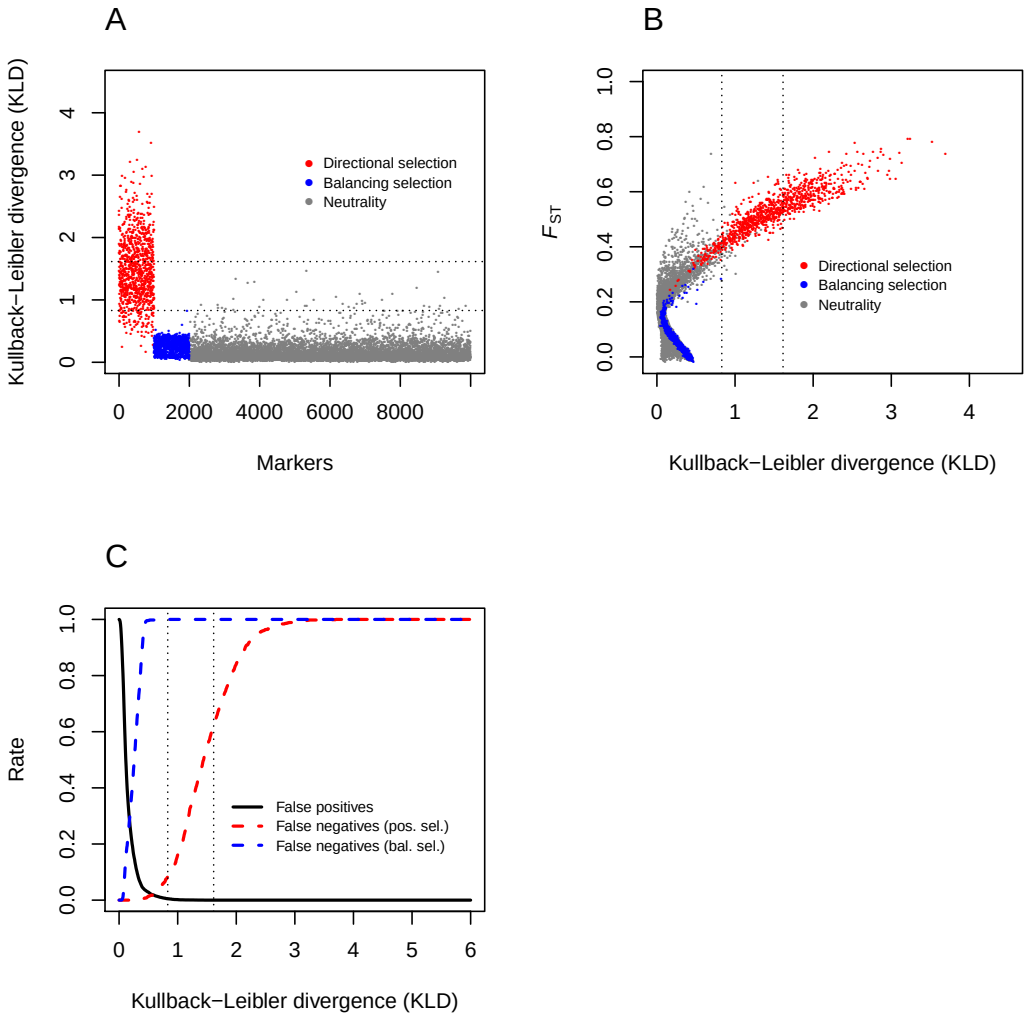


Figure 3: (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci in dataset 5. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure.

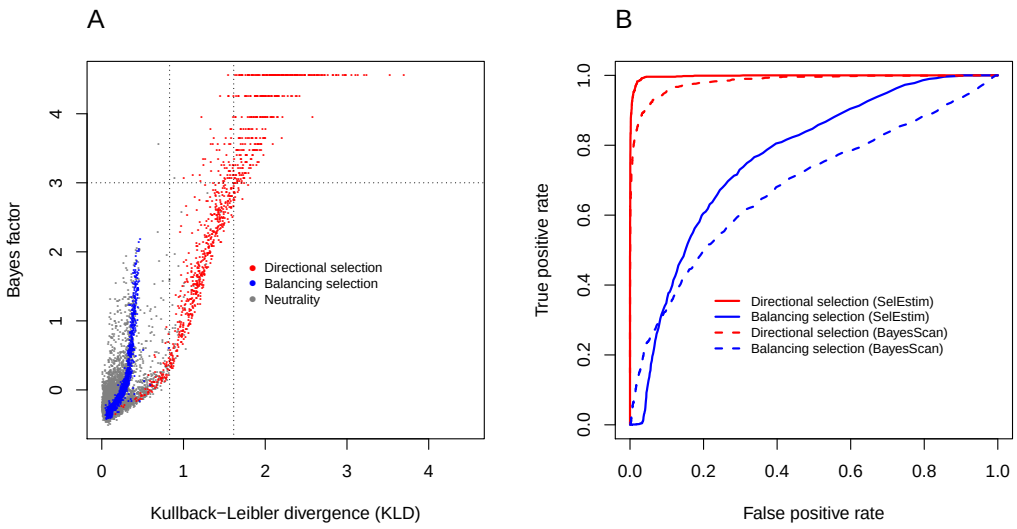


Figure 4: (A) Relationship between BayeScan Bayes factor and Kullback–Leibler divergence (KLD) for all markers in dataset 5. Positively selected markers are in red, loci under balancing selection are in blue and neutral markers are in grey. The horizontal dotted line indicates the BF = 3 threshold, and the two vertical dotted lines indicate the thresholds KLD = 0.830 (resp. KLD = 1.614), which equals the KLD between two Bernoulli distributions corresponding to flipping a fair coin and a biased coin that gives a head with probability 0.05 (resp. 0.01). (B) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.

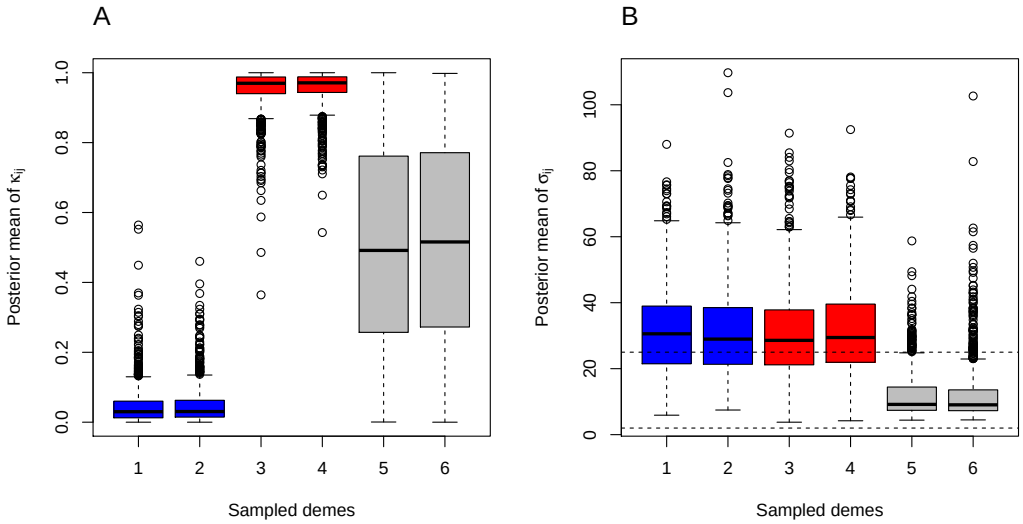


Figure 5: Analysis of the allele count data from dataset 5. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 1,000 positively selected loci in "blue" demes (1–2), "red" demes (3–4) and "uncolored" demes (5–6). (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for positively selected loci in dataset 5. For "blue" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "blue" allele being selected for ($\kappa_{ij} = 0$. For "red" demes, the posterior means of the selection coefficients σ_{ij} are conditional upon the "red" allele being selected for ($\kappa_{ij} = 1$. For "uncolored" demes, the posterior means of the selection coefficients σ_{ij} are unconditional.

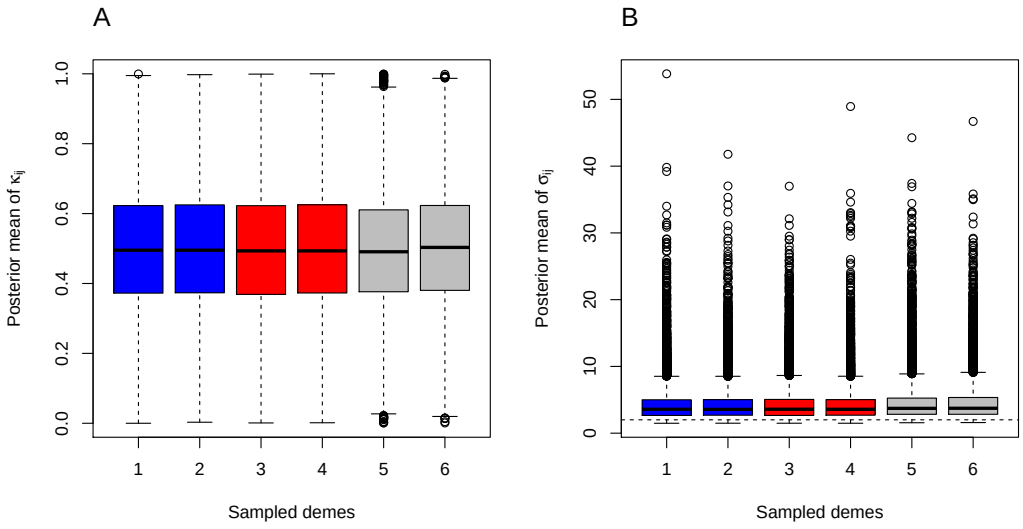


Figure 6: Analysis of the allele count data from dataset 5. (A) Boxplot representation of the posterior means of the parameters κ_{ij} (that indicate which allele is selected for) for the 8,000 neutral markers in "blue" demes (1–2), "red" demes (3–4) and "uncolored" demes (5–6). (B) Boxplot representation of the posterior means of the selection coefficients σ_{ij} for neutral markers in dataset 5. The posterior means of the selection coefficients σ_{ij} are unconditional.

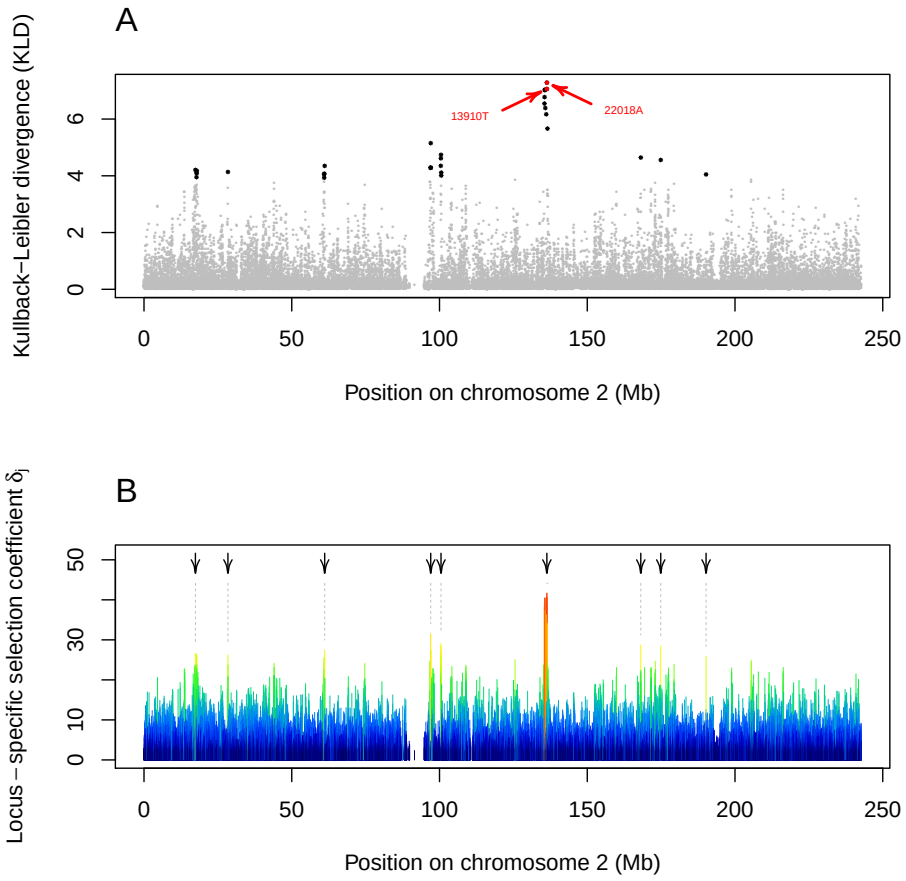


Figure 7: (A) Kullback–Leibler divergence (KLD) between the posterior of δ_j and its centering distribution. The alleles -13910C→T and -22018G→A associated with lactase persistence are indicated in red. (B) Locus-specific selection coefficient δ_j along chromosome 2. The color (from blue to red) and the width of each segment is proportional to the strength of selection

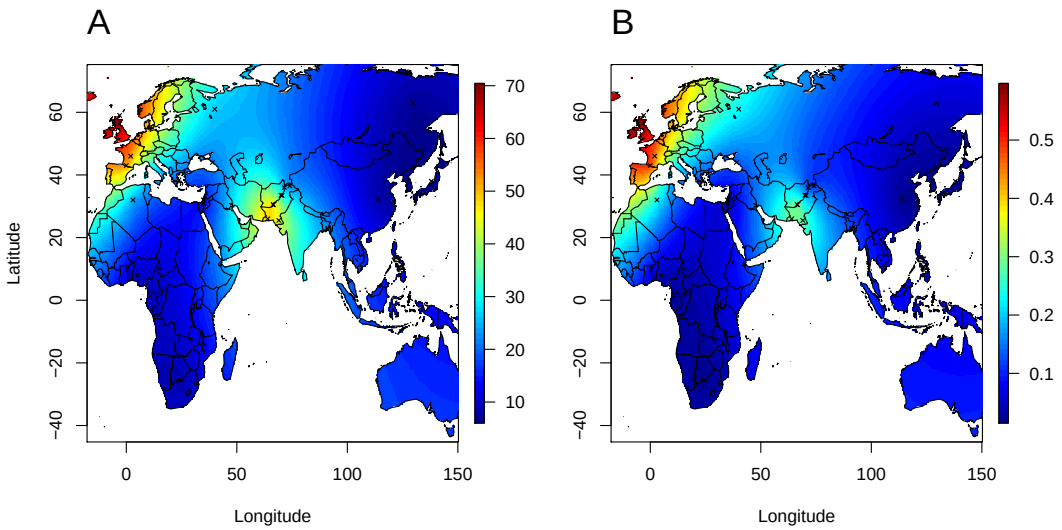
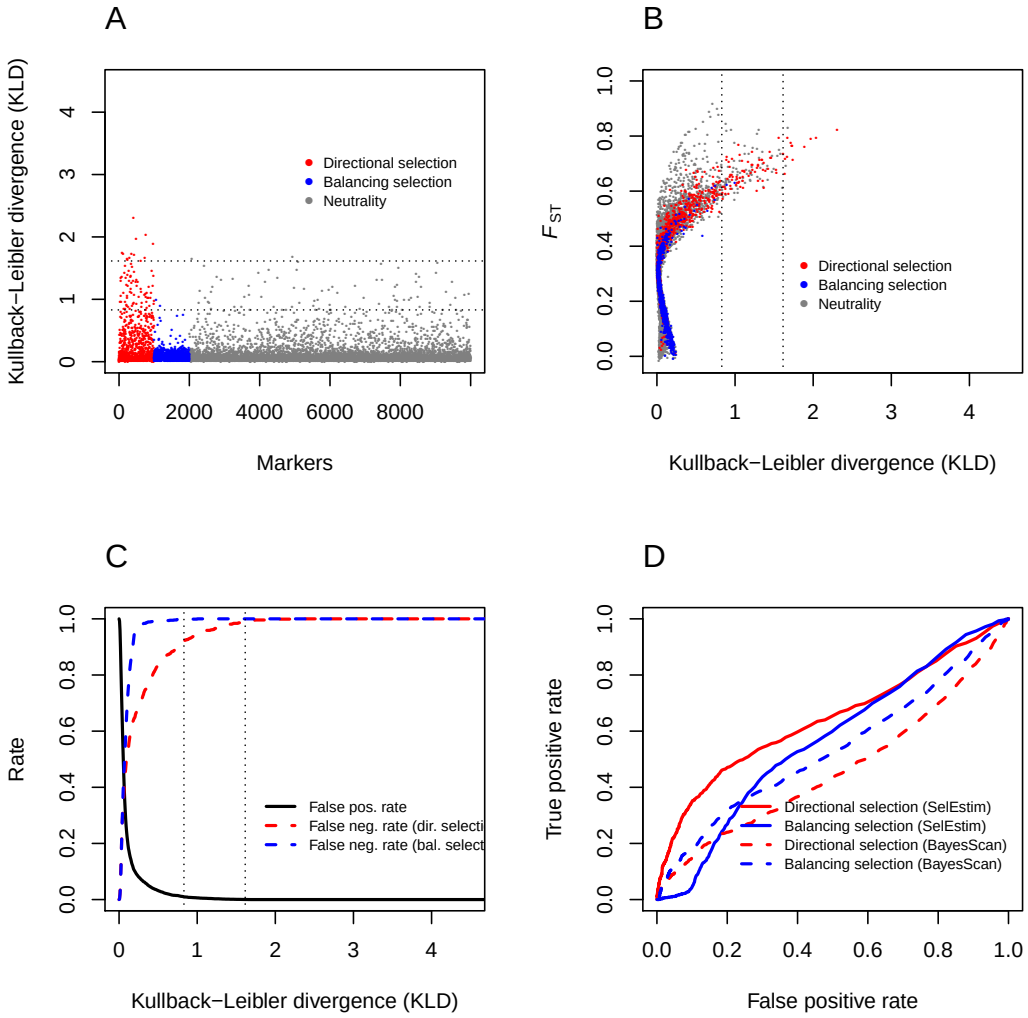
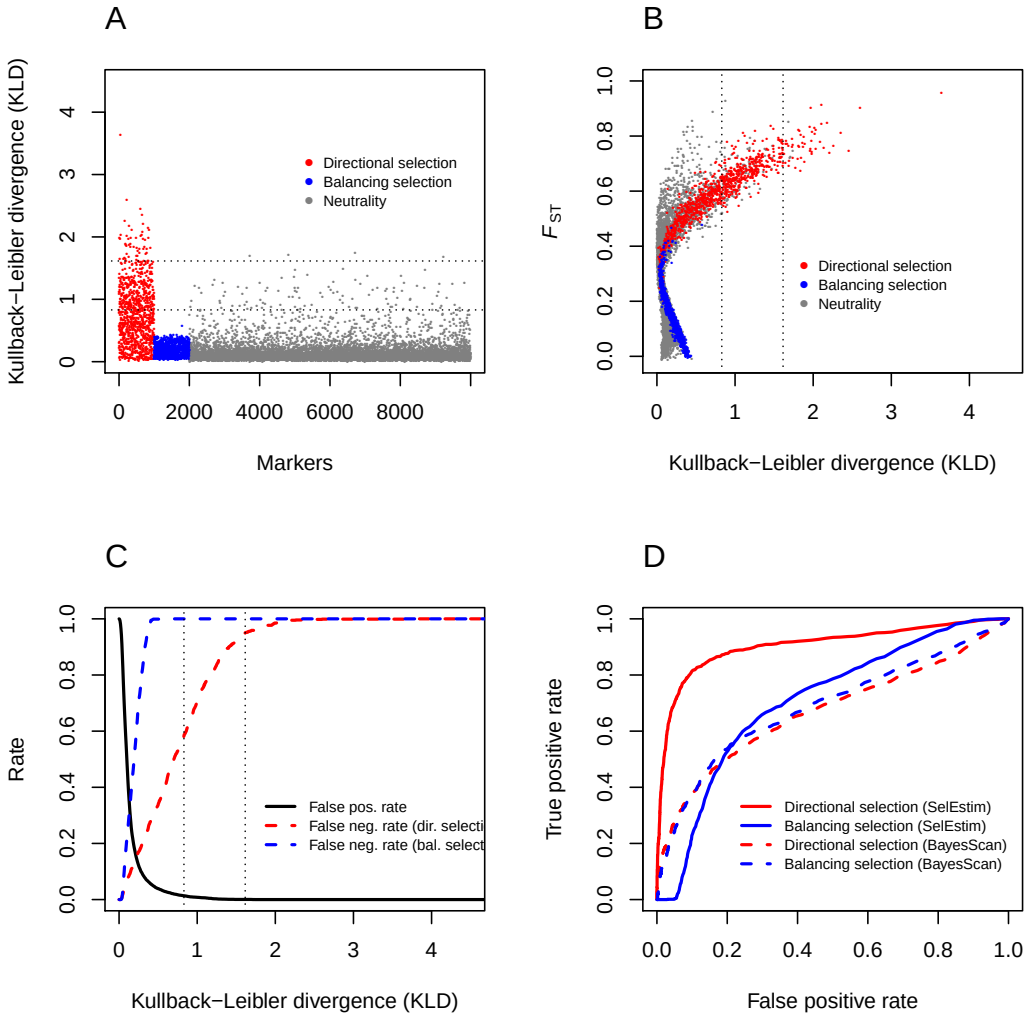


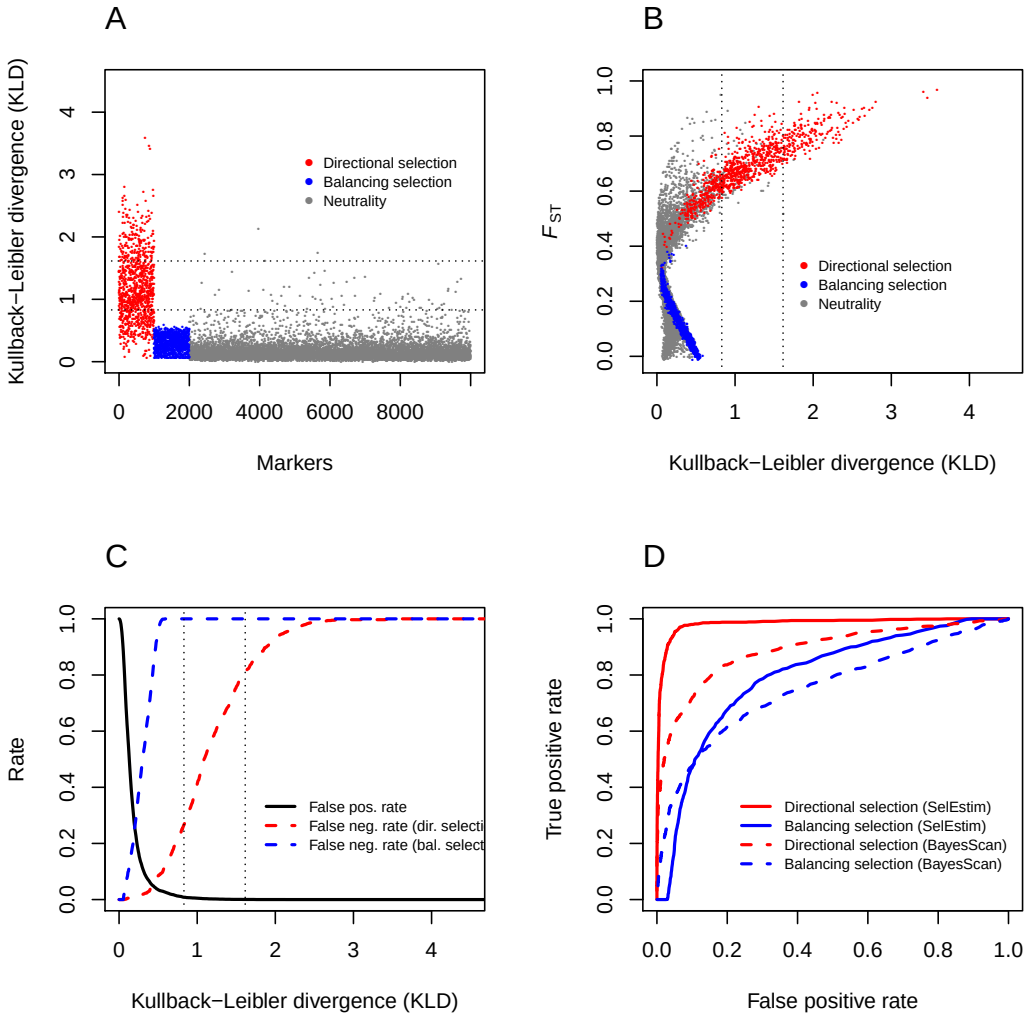
Figure 8: (A) Extrapolated spatial distribution of the selection coefficient σ_{ij} at locus -13910C→T, conditionally on allele -13910C→T being selected for, across African and Eurasian populations. (B) Spatial frequency distribution of allele -13910C→T associated with lactase persistence.



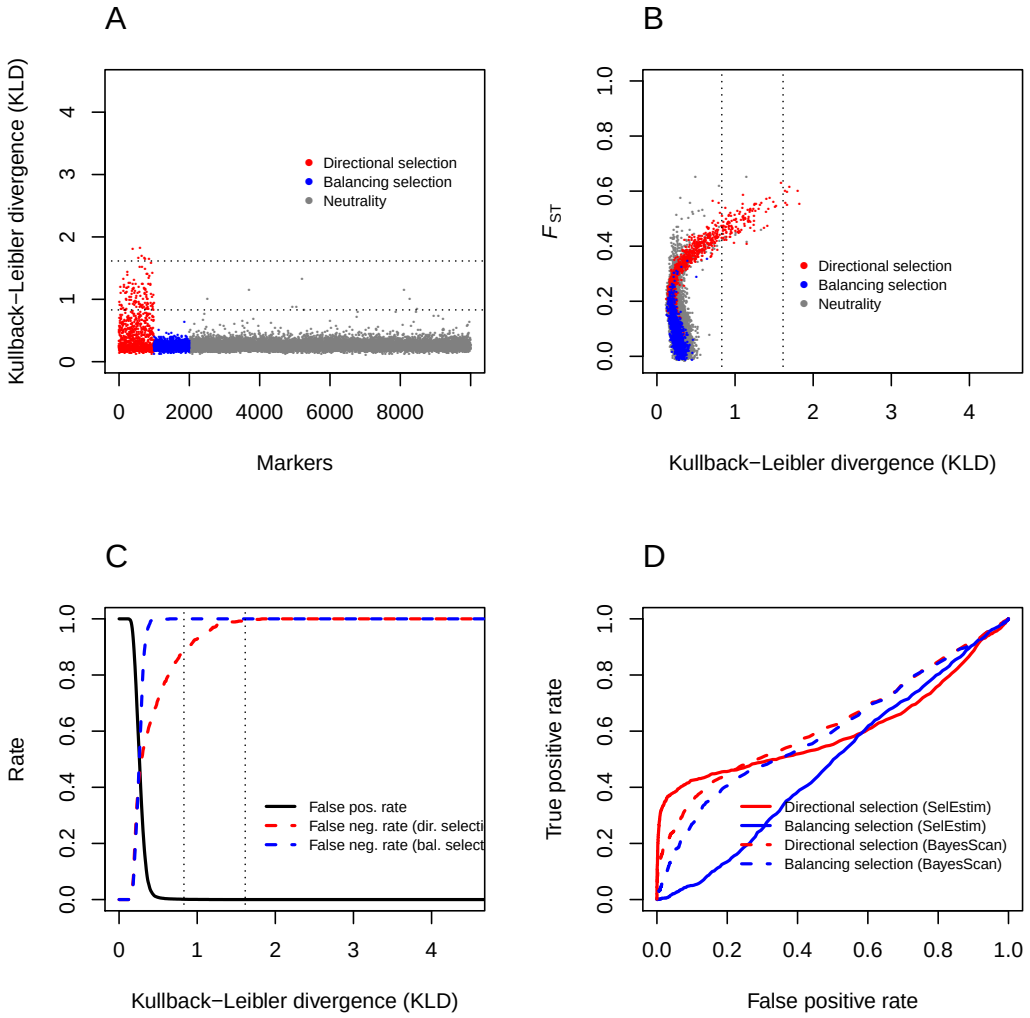
Supplementary Figure S1: Analysis of the allele count data from dataset 1. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.



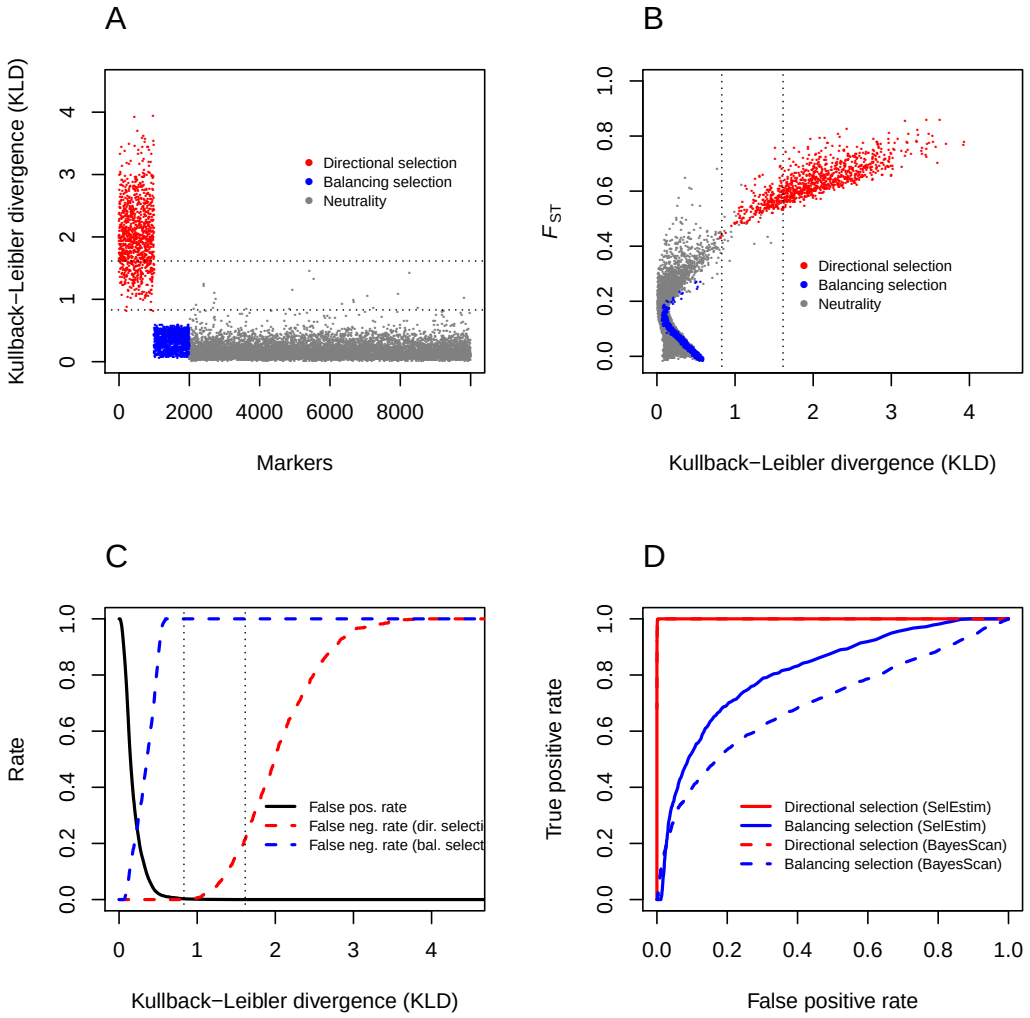
Supplementary Figure S2: Analysis of the allele count data from dataset 2. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.



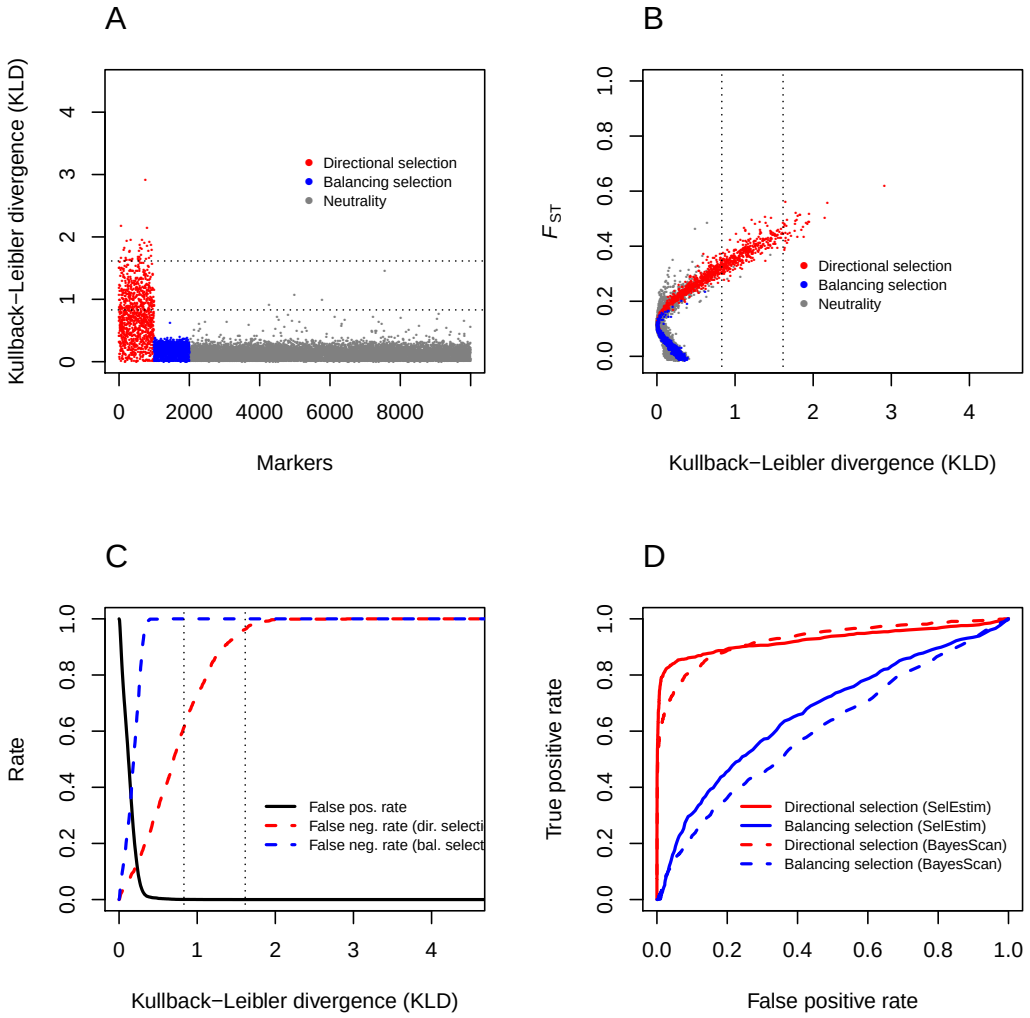
Supplementary Figure S3: Analysis of the allele count data from dataset 3. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.



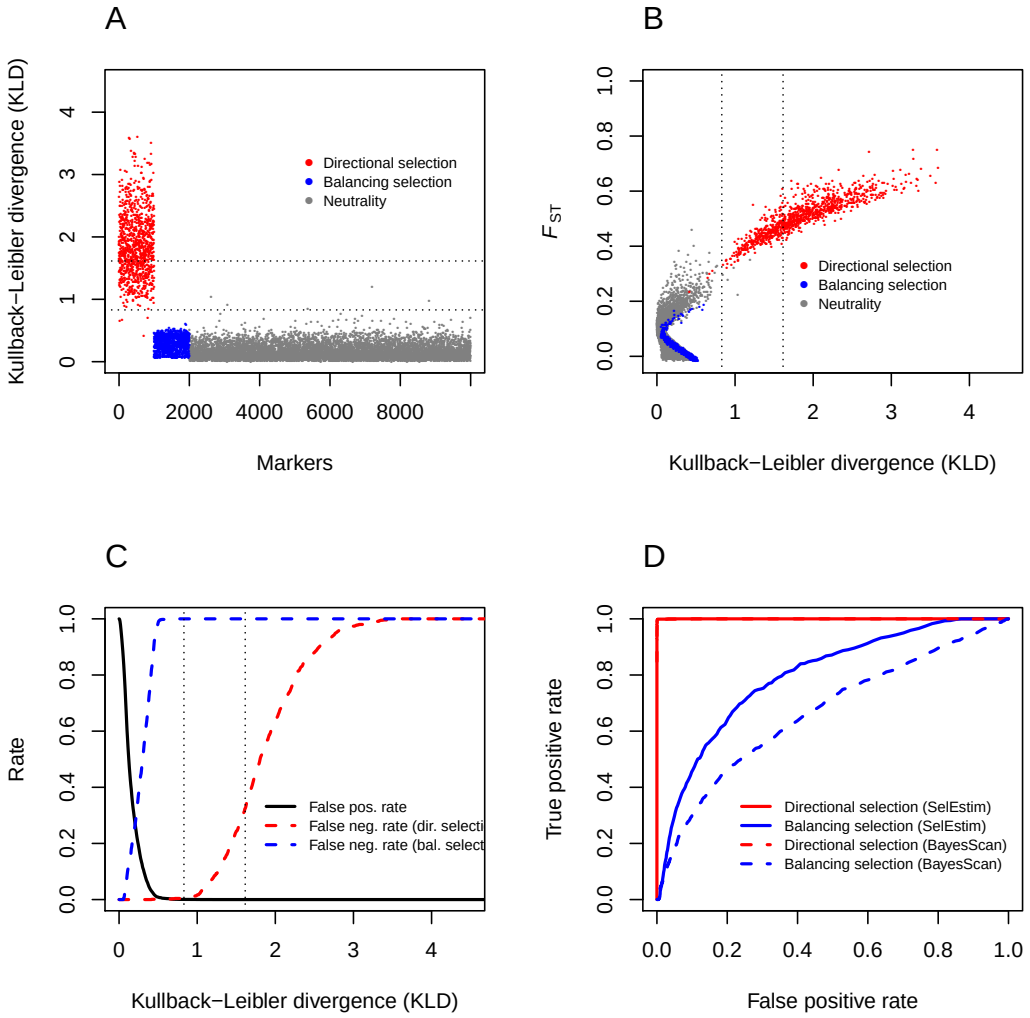
Supplementary Figure S4: Analysis of the allele count data from dataset 4. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.



Supplementary Figure S5: Analysis of the allele count data from dataset 6. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.



Supplementary Figure S6: Analysis of the allele count data from dataset 7. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.



Supplementary Figure S7: Analysis of the allele count data from dataset 8. (A) Kullback–Leibler divergence (KLD) measure between the posterior of δ_j and its centering distribution for all simulated loci. Loci under positive selection are depicted in red, loci under balancing selection in blue, and neutral markers are in grey. (B) F_{ST} as a function of the KLD measure for all loci. (C) False positive (neutral loci detected as outliers) and false negative (selected loci not detected as outliers) rates as a function of the KLD measure. (D) Receiver operating characteristic (ROC) analysis for the dataset 5. For our model, the classifying variable was the KLD between the posterior distribution of the locus-specific coefficient of selection δ_j and its centering distribution, while in the case of BayeScan it was the Bayes factor.

Annexe G

Article 5 : évolution de la dormance

Vitalis R., Rousset F., Kobayashi Y., Olivieri I. et Gandon S. The joint evolution of dispersal and dormancy in a metapopulation with local extinctions and kin competition. *En préparation*

The joint evolution of dispersal and dormancy in a metapopulation with local extinctions and kin competition

Renaud Vitalis[*1], François Rousset[†2], Yutaka Kobayashi[‡3], Isabelle Olivieri[§2], and Sylvain Gandon[¶4]

[1]Centre National de la Recherche Scientifique – Institut National de la Recherche Agronomique, UMR CBGP (INRA – IRD – CIRAD – Montpellier SupAgro), Campus International de Baillarguet, CS 30016, F-34988 Montferrier sur Lez Cedex, France

[2]Université Montpellier 2 – Centre National de la Recherche Scientifique, UMR 5554 'Institut des Sciences de l'Évolution de Montpellier', Place Eugène Bataillon, 34095 Montpellier Cedex 05, France

[3]Department of Biological Sciences, The University of Tokyo, Hongo 7-3-1, Bunkyoku, Tokyo 113-0033, Japan

[4]CEFE, UMR 5175 CNRS, F-34293 Montpellier cedex 05, France

[*]Corresponding author; e-mail: vitalis@supagro.inra.fr
[†]e-mail: francois.rousset@univ-montp2.fr
[‡]e-mail: yutaka@biol.s.u-tokyo.ac.jp
[§]e-mail: isabelle.olivieri@univ-montp2.fr
[¶]e-mail: sylvain.gandon@cefe.cnrs.fr

abstract

Dispersal and dormancy are often presented as alternative strategies to recolonize empty patches and escape kin competition. Yet our understanding of the joint evolution of these two traits remains limited. Here we analyze the evolution of dispersal and dormancy as a function of direct fitness costs, environmental variation, and competition among relatives. We consider two scenarios depending on whether the rates of dormancy for philopatric and dispersed individuals are constrained to be the same (unconditional dormancy) or allowed to be different (conditional dormancy). We show that only philopatric individuals should enter dormancy, at a rate increasing with increasing rates of local extinction and decreasing population sizes. When dormancy and dispersal evolve jointly, we observe a wide range of evolutionary outcomes. In particular, we find that the evolutionary stable rates of dispersal and dormancy are not necessarily negatively correlated, which challenges the idea that they are exchangeable strategies.

Introduction

Many plant and animal species produce seeds or eggs that do not emerge when their development is achieved and the environmental conditions are favorable (Evans and Dennehy 2005). Instead, the progagules may stay in a dormant stage, sometimes long before they hatch, thereby forming seed banks or egg banks. Such delay in early life development might be viewed as a form of temporal dispersal (Venable and Brown 1988), which suggests that the evolution of dormancy and dispersal might be driven by very similar selective forces.

Both dispersal and dormancy entail some costs, since these two strategies require the development of physiological and morphological attributes that are necessary to disperse or to enter a dormant stage. There are also costs associated with the variation of environmental conditions: just like a disperser may land in an unsuitable habitat if there is spatial variability, a dormant individual may face harsh conditions after emergence if there is temporal variability. On the other hand, both traits are associated with very similar benefits (Venable and Brown 1988; Venable et al. 1993). First, considering density independent processes only, dispersal and dormancy may provide a means to hedge one's bets, i.e. to avoid the risks associated with the temporal variation of environmental conditions (Slatkin 1974; Philippi and Seger 1989). For example, with a temporal variation in survival and/or fecundity due to the succession of good years and bad years, producing dormant seeds spreads the risk of reproductive failure by postponing the emergence of the propagules (Cohen 1966; Venable 2007). Dispersal may also evolve as a bet-hedging strategy, but in less straightforward ways. For example, although dispersal responds to the between-year variation of the rate of extinction of local populations, it may not respond to between-year local variation in fecundity (Metz et al. 1983). Both dormancy and dispersal will also respond to stochastic variation in fecundity between generations, but only if the number of patches is finite (Venable and Brown 1988; Venable et al. 1993; Ronce 2007). The second category of benefits associated with dispersal and dormancy relies on the fact that with density dependence, both strategies

allow to reduce crowding (Levin et al. 1984; Ellner 1985a,b). Third, dispersal and dormancy may help reducing the impact of local competition that occurs among relatives (Ellner 1986; Hamilton 1964; Hamilton and May 1977; Frank 1986; Taylor 1988; Kobayashi and Yamamura 2000) and avoiding reduced fitness due to inbreeding depression (Waser et al. 1986; Gandon 1999; Perrin and Mazalov 1999; Morgan 2002; Roze and Rousset 2005, 2009), as illustrated empirically (see, e.g., Richards 2000; Paland and Schmid 2003; Busch 2006; Ebert et al. 2002).

Since dispersal and dormancy presumably respond to similar evolutionary forces, it is tempting to consider that these strategies may substitute to each other. One would expect in that case to observe a negative covariation between these traits. Several theoretical studies looking at the evolution of dormancy confirmed indeed the prediction that, in general, increasing dispersal tends to decrease the evolutionary stable (ES) rate of dormancy (Kobayashi and Yamamura 2000; Satterthwaite 2010). Several studies analyzing the evolution of dispersal also found that, in general, increasing dormancy selects for lower ES rates of dispersal (Levin et al. 1984; Cohen and Levin 1991; Snyder 2006). Yet in order to predict the outcome of the evolution of dispersal and dormancy, and to characterize the emerging covariation between both traits, it is necessary to consider models where dispersal and dormancy evolve jointly. Some models have been developed to study, numerically, the joint evolution of dispersal and dormancy under various ecological scenarios (Cohen and Levin 1987; Klinkhamer et al. 1987; Venable and Brown 1988; Wiener and Tuljapurkar 1994; McPeek and Kalisz 1998; Olivieri 2001; Tsuji and Yamamura 1992). Yet none of these models considered the potential effect of kin competition on the evolutionary dynamics of both traits.

Here, we use an analytical model in order to analyze the joint evolution of dispersal and dormancy in a metapopulation with kin competition and local extinctions. Our model is based on the computation of selection gradients in a metapopulation. The formal derivation of the gradients relies on standard results for class-structured populations (see, e.g., Hamilton 1966; Charlesworth 1994; Taylor 1990) completed by the results of Rousset and Ronce (2004),

which take into account the feedback of individual behaviour on allele frequency change, through the effect of this behaviour on the demography of the local populations. However, the exact calculation of the gradient in our model is unpractical, so that we use some analytical approximations to find the convergent stable strategies for dispersal and dormancy. We show that our predictions are remarkably consistent with individual-based simulations. In the following we first detail the hypotheses of our model and derive the gradients of selection for dispersal and dormancy. Then we provide the results of our analyses for the evolution of each trait when they evolve independently from the others. Finally, since in reality selection acts simultaneously on all phenotypic traits, we examine the outcome of the joint evolution of all the traits. At each step of these analyses, we emphasize the connection with previous models devoted to the evolution of dispersal and dormancy. The originality of the present study lies in the fact that it reconciles some results obtained with simpler evolutionary scenarios, generates new quantitative and testable predictions, and paves the way towards a better understanding of the evolution of delayed emergence in variable environments.

The model

Life cycle

Our model may apply indifferently to a number of plant or animal species with delayed emergence. Yet, for the sake of simplicity, we will restrict our vocabulary to plant life cycles. We consider a metapopulation with an infinite number of local populations (or "demes"). Each population can either contain a fixed number N of haploid asexual individuals, or none, after extinction.

We consider the following life cycle: (i) adults produce a random, Poisson distributed, number of seeds and then die; (ii) a fraction z of seeds are dispersed, and the seeds that disperse incur a cost noted c_{z}; (iii) a fraction D of the seeds enter a dormant state, and all dormant seeds incur a cost noted c_{d}; (iv) all the non-dormant seeds, as well as all the dormant seeds produced in the previous time step germinate; in other words we assume a maximal age of dormant seeds of one year, as in Kobayashi and Yamamura (2000); however, this assumption is relaxed in individual-based simulations; (v) competition occurs among germinating seeds and a fixed number N of them survive to adulthood; (vi) some demes face random catastrophic events (extinctions) that arise with probability e; these events result in the death of all the standing (i.e., non-dormant) individuals in the deme. For the sake of clarity, Figure 1A depicts the above life cycle, and Table 1 summarizes the model parameters. We also consider an alternative life cycle, in which dormancy is conditional upon dispersal, i.e. where the rate of dormancy of dispersed seeds may differ from that of non-dispersed seeds, as in Olivieri (2001). More precisely, we consider that in step (iii) of the above life cycle, a fraction d of the philopatric seeds and a fraction δ of the dispersed seeds enter a dormant state. Both life cycles will be analyzed in this paper.

Gradient of selection

In order to investigate the evolutionary dynamics of the rate of dispersal and that of dormancy, we use a direct fitness approach (see Taylor and Frank 1996; Rousset and Billiard 2000) to compute the fitness of a focal individual (i.e., its expected number of surviving offspring), as a function of the strategies of all the individuals with which it competes. We assume that each of these phenotypic traits is encoded by a bi-allelic locus. Let us first consider the case of dispersal evolution alone (but the following argument holds for all traits), as in Hamilton and May (1977), Frank (1986) and Taylor (1988) : at each locus, we consider a mutant allele A in a population of individuals that bear allele a. We assume that allele a gives phenotype (here, the dispersal rate) z_a, and that the mutant allele A gives phenotype $z_A \equiv z_a + \epsilon_z$. In the infinite island model of dispersal, the expected change Δp in allelic frequency p over one generation can then be expressed as (see Rousset 2004):

$$\Delta p = p(1-p)S(z)\epsilon_z + O(\epsilon_z^2), \tag{1}$$

where $S(z)$ is the selection gradient, which is also the inclusive fitness effect under weak selection, i.e. for small ϵ_z (Hamilton 1964).

In the model considered here, all individuals are not equivalent. Within a deme, for example, standing individuals and seeds in the bank do not compete with each other. They must therefore be treated as different types. All the demes are not equivalent either. For example, the demes that have gone extinct in the previous time step cannot contain philopatric dormant seeds (i.e., seeds that would have been produced by resident adults in the previous time step). In these demes, there is therefore no competition between the offspring of standing adults and those of philopatric dormant seeds. Different categories of demes must therefore be distinguished, depending on the history of extinctions over two successive time steps (see Figure 1B). Both the individual types and the deme categories define eight demographic classes in our model (see Figure 1B).

In class-structured populations, the different demographic classes of individuals can make different contributions to the future of the population. Nevertheless, equation (1) holds if allele frequency is defined as a weighted average of allele frequencies $\mathbf{p}$ in the different demographic classes. These weights are known to be the reproductive values of each class, noted $\boldsymbol{\alpha}$, that give the relative ultimate contributions of all the gene lineages present in a class to the future pool of genes (Taylor 1990; Rousset 2004, chapter 11). The gradient of selection $S(z)$ measures the first order effect of selection on the change of this weighted sum of mutant frequency.

We consider in our model that density-dependent regulation occurs among adults, but not among dormant seeds in the bank: see the step (v) of the above life cycle. The number of seeds in the bank is therefore a random variable that depends upon trait values. This generates a large number of populations in different demographic states (i.e. with different seed bank sizes) within a particular category of deme. Taking into account such demographic fluctuations in the seed bank yields complex fitness functions (see the appendix in the online edition), which makes it very difficult to find an analytic solution. We therefore approximate the distribution of seed bank sizes with its expectation (see the appendix in the online edition). This simplification allows us to use only the eight demographic classes of individuals defined in Figure 1B. Below we show that this approximation is remarkably consistent with stochastic individual-based simulations.

The selection gradient $S(z)$ may be expressed as a weighted sum of relatedness coefficients and functions $f_{(i,k)\leftarrow(j,l)}$ that give the probability that a gene in class (i,k) is a copy of a gene from any of the A parent in class (j,l) (Rousset 2004). We define the class (i,k) for type-i individuals in demes of category k. The weights depend upon the reproductive values of each class, the transition probabilities between deme categories, and the stationary distribution of deme categories (see the appendix in the online edition). The functions $f_{(i,k)\leftarrow(j,l)}$ depend upon the fitness functions $w_{(i,k)\leftarrow(j,l)}$ that give the expected number of offspring in class (i,k) produced by a focal individual in class (j,l). The fitness functions depend upon the

phenotypes of the different individuals in competition with a focal individual (see, e.g., Frank 1998). In the following, we distinguish the value of the trait in a focal individual from the mean values of that trait in different categories of actors. The subscript "$_\bullet$" (e.g., $z_\bullet$) refers to the focal individual; the subscript "$_0$" (e.g., z_0) refers to the mean value of the trait in the focal individual's deme, and the subscript "$_1$" (e.g., z_1) refers to the mean value of the trait in the focal individual's deme, in the previous time step. Indeed in our model, competition may occur among seeds produced by adults at time t and seeds that emerge at t from the bank constituted at $(t-1)$. Hence, the fitness of a focal individual depends upon the strategies adopted by other individuals in the previous time-step. We show in the appendix in the online edition that, if we neglect demographic stochasticity, then the selection gradient S can be approximated as:

$$\begin{aligned} S(z) &= \sum_{i,k} \alpha(i,k) \sum_{l} v(l|k) \sum_{j} \left(\frac{\partial f^{P}_{(i,k)\leftarrow(j,l)}}{\partial z_\bullet} + \frac{\partial f^{P}_{(i,k)\leftarrow(j,l)}}{\partial z_0} Q^{0}_{(j,l)} + \frac{\partial f^{P}_{(i,k)\leftarrow(j,l)}}{\partial z_1} Q^{1}_{(j,l)} \right. \\ &+ \left. \sum_{m} P(m) \frac{\partial f^{D}_{(i,k)\leftarrow(j,m)}(l)}{\partial z_\bullet} \right), \end{aligned} \tag{2}$$

where $\alpha(i,k)$ is the reproductive value of class (i,k), $v(l|k)$ is the backward transition probability that a deme in category k at $t+1$ was in category l at t and $P(m)$ is the stationary distribution of deme categories. The function $f^{P}_{(i,k)\leftarrow(j,l)}$ gives the probability that a philopatric gene in class (i,k) is a copy of a gene from any of the A parent in class (j,l). Likewise, $f^{D}_{(i,k)\leftarrow(j,m)}(l)$ gives the probability that a dispersed gene in class (i,k) at $t+1$ is a copy of a gene originally in a deme of category m that has been dispersed in a deme that was in category l at t. $Q^{0}_{(j,l)}$ is the relatedness between a focal individual in class (j,l) and an adult actor in its deme; $Q^{1}_{(j,l)}$ is the relatedness between a focal individual in class (j,l) at t and an adult actor in its deme at $t-1$ (see the appendix in the online edition). The superscripts "0" and "1" stand for the number of time-step (0 or 1) that separates the focal from an adult actor in its deme.

Equation (2) gives the first order effects of different actors on the number of offspring in class (i,k) of a focal individual, weighted by the probabilities of genetic identity $Q^0_{(j,l)}$ and $Q^1_{(j,m)}$ between the focal individual's gene and the actor's one. The first and the last terms within brackets in the right-hand side of equation (2) give the effect of the focal individual on its expected number of adult offspring. The second term gives the effect of different actors in the same deme on the expected number of adult offspring of the focal individual. The third term within brackets in the right-hand side of equation (2) gives the effect of actors in the same deme in the previous time step, on the expected number of adult offspring of the focal individual. This inter-generational term provides the indirect benefit received by the focal individual, from the behavior of actors in the previous generation (see, e.g., Lehmann 2007). Expressions for the selection gradient for other traits may be obtained by replacing z with D (or d and δ in the conditional dormancy model) in equation (2).

Evolutionarily stable strategies

Candidate evolutionarily stable strategies (ESS) for each trait independently are found by numerically computing the sign of the gradient of selection, e.g., $S(z^*)$ near z^*, assuming that the other traits (e.g., D) are fixed parameters. A strategy z^* is a candidate ESS if $S(z^*) = 0$. This strategy is locally convergence stable (CS) if $S(z^*) > 0$ at $z < z^*$ and $S(z^*) < 0$ at $z > z^*$, so that the population evolves until it reaches the point z^* where there is no longer directional selection. Characterizing evolutionary stability would require the computation of second-order derivatives of the fitness (see Eshel 1996; Geritz et al. 1998; Ajar 2003). For all the results that follow, individual-based stochastic simulations have shown that the candidate ESS were indeed convergence and evolutionarily stable.

Candidate evolutionarily stable strategies (ESS) for all traits simultaneously are found by numerically computing the signs of the gradients of selection $S(z^*)$ and $S(D^*)$, and determined the joint set of strategies z^* and D^* for which the gradients of selection vanish.

With conditional dormancy, we considered instead the gradients $S(z^*)$, $S(d^*)$ and $S(\delta^*)$, simultaneously. Although we did not consider the stability conditions for the evolution of multidimensional traits suggested by Leimar (2009), we checked with individual-based stochastic simulations the candidate ESS were convergence and evolutionarily stable.

Stochastic simulations

In order to test the accuracy of the approximations we used a stochastic, individual-based simulation model. Each individual was characterized by a set of random variables representing its genotype for each phenotypic trait. The same life cycle as in the analytical model was considered (see Figure 1A), except that we relaxed the assumption that seeds cannot be older than one year in the seed bank. We therefore assumed an arbitrary number of age classes in the seed bank so that, each generation, a fraction $(1-d)$ of seeds in age class i the bank germinates, and a fraction $d(1-c_\mathrm{d})$ goes to age class $i+1$. In other words, of the seeds that do not germinate, a fraction c_d decays each year. See the appendix in the online edition for further details on the simulations.

Results

In the following, we will first consider the evolution of each phenotypic trait independently, assuming that the other traits are fixed parameters that do not evolve. Then, we will consider the joint evolution of all the traits, hence accounting for potential evolutionary feedbacks. For all the results that follow, we checked that our approximate solutions for the candidate ESS of dispersal and dormancy were in agreement with individual-based simulations. As shown in Supplementary Figure S1, we obtained a remarkable fit between the predicted evolutionarily stable (ES) rates and the equilibrium frequency of the traits in stochastic simulations, despite the approximation ignoring demographic stochasticity. The fit between the predicted ES rates and the equilibrium frequency of the traits in stochastic simulations is also evident in Figures 2 and 4–5.

Evolution of dormancy

Evolution in a constant environment

In a constant environment ($e = 0$), if we assume that the rate of dormancy is the same for philopatric and dispersed seeds (unconditional dormancy), our model reduces to Kobayashi and Yamamura (2000)'s one. Cancelling the dispersal cost c_z, as they assume, we indeed obtained the same analytical expression for the ES rate of dormancy D^* as in their haploid asexual model (equations [A.7a]–[A.7c] in Kobayashi and Yamamura 2000). In the limit case where $N = 1$, we find:

$$D^* = \frac{(1-\eta)^2(1-c_d) - c_d(2-\eta)}{[(1-\eta)(1-c_d) - c_d])\,[(2-\eta) - \eta(1-c_d)]}, \tag{3}$$

where $\eta = (1-c_z)z/(1-c_z z)$ is the backward dispersal rate (i.e., the probability that a seed sampled after dispersal is an immigrant). Evaluation of equation (3) shows, not surprisingly, that D^* decreases as the cost of dormancy (c_d) increases. Equation (3) also shows that

D^* is a decreasing function of η, which depends on both the dispersal rate z and the cost of dispersal c_z. Hence, large dispersal rates and/or small costs of dispersal both select for lower ES dormancy rate D^* (Figure 2). Here, in the absence of local extinctions, kin competition is the only force selecting for dormancy. Because kin competition is weaker in larger populations, D^* decreases as the adult population size (N) increases. If there is no cost to enter a dormant stage ($c_d = 0$), the convergent stable strategy is to put half of the seeds in the seed bank ($D^* = 1/2$). Because we consider that all dormant seeds germinate after one year, competition among offspring is strictly equivalent whether all seeds germinate ($D = 0$) or all seeds go dormant ($D = 1$). It is only if a fraction of the seeds go dormant, that competition among related individuals is spread over the generations; and with a single age class in the seed bank, competition among kin is minimized by dividing equally the offspring into a dormant and a non-dormant pool (Kobayashi and Yamamura 2000). Note that this results only holds with a single age class in the seed bank, so that ES dormancy rates $D^* > 1/2$ may evolve if dormant seeds can survive more than one year in the bank. However, Figure 2 shows that, in the absence of environmental variation, there is a very good agreement between the analytical model that only considers a single age class and the simulations run with 50 age classes in the bank.

We have considered so far that the rate of dormancy was the same for dispersed and philopatric seeds (unconditional dormancy). Yet it can be shown from our model that when the rate of dormancy of dispersed seeds may differ from that of philopatric seeds (conditional dormancy), the gradient of selection $S(\delta)$ for the rate of dormancy of dispersed seeds is strictly negative for $c_d > 0$. This means that dormancy of dispersed seeds is always selected against for $c_d > 0$, and hence that $\delta^* = 0$. Hence, dispersed seeds should never go dormant, and dormancy evolves only for philopatric seeds. If there is no cost of dormancy ($c_d = 0$), though, we get $S(\delta) = 0$, which indicates that the rate of dormancy for dispersed seeds evolves neutrally. Besides, we found that the convergent stable rate of dormancy of philopatric seeds (d^*) is always higher than that of unconditional dormancy (Figure 2). For example, in the

limit case where $N = 1$ and $e = 0$, we find:

$$d^{\star} = \frac{(1-\eta) - c_{\mathrm{d}}(2-\eta)}{(1-\eta)(2-3c_{\mathrm{d}})}, \tag{4}$$

which is always higher than the unconditional ES rate of dormancy given in equation (3). This is so not only because unconditional dormancy must balance the antagonistic selective pressures acting on dispersed and philopatric seeds, but also because dispersed dormant seeds pay the cost of both dispersal and dormancy. As for unconditional dormancy, large dispersal rates and/or small costs of dispersal both select for lower ES dormancy rate d^{*} (Figure 2).

Evolution in a varying environment

Environmental variation was introduced in our model by considering a probability e that populations go extinct. This approach is equivalent to Cohenıs (1966) model, who considered two types of year, good and bad, which occur in a random uncorrelated sequence. Cohenıs citeyearparCohen1966 model was later extended by Bulmer (1984), to include density-dependent regulation in the model. There are two main differences between Bulmerıs (1984) and ours: as in Cohen (1966), Bulmer (1984) considers a single isolated population of infinite size, and the maximal age a seed can reach in the bank is infinite. Bulmer (1984) found that the ES rate of dormancy d^{*} is the solution of (using our notations):

$$\begin{cases} \left(1 - \frac{1-d}{1-e}\right)^{1-e} = d(1-c_{\mathrm{d}}) \\ r = \frac{d(1-c_{\mathrm{d}})}{d-e} \end{cases} \tag{5}$$

In order to test whether our model converged toward Bulmerıs (1984) results, we used stochastic simulations with large population sizes (in order to reduce the effect of kin competition), a very low dispersal rate (in order to mimic the fate of isolated populations), and a large number of age classes in the seed bank. The results are presented in Figure 3A, for

conditional dormancy (but the same results hold for unconditional dormancy): despite very different ways of modelling, our model converges well toward Bulmerıs (1984) predictions as the maximal number of age classes in the bank increases. Our model therefore accounts for environmental variation as in Cohen (1966) and Bulmer (1984). The main result in Figure 3A is that environmental variation (in the form of random extinctions) selects for larger rates of dormancy d^* for philopatric seeds. Figure 3A further shows that increasing the longevity of seeds in the seed bank increases the ES rate of dormancy, although this effect is important for relatively large extinction rates. Local extinctions and prolonged dormancy yield evolutionary stable rates of dormancy that can largely exceed 0.5 (Figure 3A).

Although Bulmerıs (1984) model accounts for density-dependent regulation, it assumes, in effect, infinitely large population sizes. Our model is more realistic in the sense that populations are finite in size, which allows competition among kin to occur. Figure 3B shows the effect of population size on the ES rate of dormancy for philopatric seeds. Since the competition among kin increases in smaller populations, the ES rate of conditional dormancy increases as population size decreases (Figure 3B). If we now vary the rate of dispersal (Figure 3C), so that $N\eta$ ranges from 0.01 to 5, then we observe that increasing dispersal selects for lower rates of dormancy for the philopatric seeds.

For unconditional dormancy, we might expect that the antagonistic forces acting on philopatric and dispersed seeds (as revealed by the fact that $d^* \neq \delta^*$) should lead to non-trivial relationships between D^* and the model parameters. For a single age class in the bank, and with varying environmental conditions, we found indeed that the ES rate of unconditional dormancy D^* is a non-monotonic function of the rate of extinction e (see Supplementary Figure S2). For low extinction rates, unconditional dormancy is selected for, as a means to recolonize empty patches with philopatric dormant seeds. As local extinctions become more frequent however, seed dormancy is selected against because dispersed seeds that colonize an empty patch have no selective advantage to delay their germination: they should germinate as fast as possible to settle in this new site. Since the fraction of empty

sites increases with local extinctions, the selection against dormancy is more pronounced for large values of e (Supplementary Figure S2). Furthermore, we observed that with either frequent local extinctions or low dispersal rates, decreasing population size tends to decrease the unconditional ES dormancy rate, which contradicts the intuition that dormancy evolves to reduce competition among relatives (Supplementary Figure S2). This is because, with either frequent local extinctions or low dispersal rates, dormant seeds may often germinate in extinct patches, with few immigrant competitors. In such patches, competition occurs mainly among germinating seeds, which are all the more related when population sizes are small. If dormancy only delays competition for a single generation, it does not provide an efficient means to escape competition among relatives. Increasing the number of age classes in the bank dampens this effect, and the ES rate of unconditional dormancy tends towards a monotonic positive relationship with the extinction rate, and a monotonic negative relationship with the dispersal rate.

Evolution of dispersal

With a single age class in the seed bank ,the evolutionarily stable dispersal rate z^* is a non-monotonic function of the rate of dormancy (Figure 4). In the absence of any cost of dormancy (c_d = 0), as pointed out in the previous section, intermediate rates of dormancy minimize the competition among kin by spreading competition across successive generations. Since reducing the competition among related individuals tends to relax selection for dispersal (see Hamilton and May 1977; Frank 1986; Taylor 1988; Gandon and Rousset 1999), the evolutionary stable dispersal rate is minimal for intermediate rates of dormancy. Increasing the cost of dormancy tends to increase relatedness among competing offspring, which selects for higher dispersal (not shown).

The distinction between conditional and unconditional dormancy is important for dispersal evolution. Obviously, when only philopatric seeds can go dormant (conditional dor-

mancy), these are the only seeds that might pay the cost of dormancy. In that case, dormancy imposes an extra cost on philopatry, which may select for extreme ES dispersal rates despite high costs of dispersal. For example, with $e = 0$ and $\delta = 0$, we get $S(z = 1) = c_{\mathrm{d}}d - c_{\mathrm{z}}$, which shows that $z^* = 1$ is convergent stable for $c_{\mathrm{d}}d > c_{\mathrm{z}}$.

Joint evolution of dispersal and dormancy

Conditional dormancy

In the following, we consider the effects of the model parameters on the joint evolutionary outcomes under the assumption that dormancy is conditional. In this case, dormancy only evolves for philopatric seeds ($\delta^* = 0$, see above) and reaches a single joint evolutionary stable equilibrium (we did not find any evidence of bistable evolutionary dynamics). Since we could not find a general closed-form expression, we focused on the case with $N = 1$ and $e = 0$, which corresponds to the scenario analyzed by Hamilton and May (1977) for the evolution of dispersal only. We found that the joint ES rates of dispersal and dormancy read:

$$z^* = \frac{1 - c_{\mathrm{d}}}{2(1 + c_{\mathrm{z}})(1 - c_{\mathrm{d}}) - 1}, \tag{6}$$

and

$$d^* = \frac{1 - (1 - c_{\mathrm{d}})(1 + c_{\mathrm{z}})}{1 - (1 - c_{\mathrm{d}})(1 + 2c_{\mathrm{z}})}. \tag{7}$$

Equations (6) and (7) generalize the model considered by Kobayashi and Yamamura (2000), in which dispersal was a fixed parameter, for the case $N = 1$. A straightforward analysis of Equations (6) and (7) further shows that a negative monotonic relationship is expected between dispersal and dormancy for $N = 1$ in the absence of local extinctions. More generally, for $N > 1$, we found that increasing the cost of dormancy c_{d} selects against conditional dormancy and for dispersal, while increasing the cost of dispersal c_{z} selects against dispersal and for conditional dormancy. This may therefore lead to negative correlations between

these traits in different environments (Supplementary Figure S3).

Figure 5 shows the emerging relationships between ES conditional dormancy and ES dispersal when various parameters (representing environmental characteristics) vary. In the absence of extinctions, the correlation between the ES rates of conditional dormancy and dispersal is positive when population size is varied: both dormancy and dispersal increase as the population size decreases. Yet, the correlation between the ES rates of conditional dormancy and dispersal may become slightly negative as the extinction rates and/or the dispersal costs increases (Figure 5A). This latter tendency is less pronounced in the individual-based simulations with 50 age classes in the seed bank (Figure 5B).

A hump-shaped relationship is obtained between conditional dormancy and dispersal, when the extinction rate is varied for a fixed population size (Figure 5C). This suggests that both negative (for very low e) and positive (for intermediate e) correlations may emerge between dispersal and conditional dormancy, in contrasted extinctions regimes. This effect also emerges from the individual-based simulations run with 50 age classes in the seed bank (Figure 5D).

Unconditional dormancy

For most parameter values, we found a single solution for each trait, suggesting that the evolutionary dynamics result in a single set of ES strategies. Yet for some parameter values, we found three joint equilibria, two of which are locally stable and the third one is unstable, indicating that the joint evolution of dispersal and unconditional dormancy may sometimes result in bistable evolutionary dynamics, where the evolutionary endpoint depends on initial conditions (Supplementary Figure S4). One stable equilibrium corresponds to intermediate rates of dispersal and dormancy (equilibrium A, in Supplementary Figure S4). The unstable equilibrium corresponds to lower rates of dispersal and dormancy (equilibrium B, in Supplementary Figure S4), and the other stable equilibrium (noted C in Supplementary Figure S4) corresponds to a null rate of dormancy. The conditions for bistable dynamics are limited,

though, and this is not a general output from the model (Supplementary Figure S5).

Not surprisingly, the ES rate of unconditional dormancy is generally lower than that of conditional dormancy, for a given dispersal rate (which is reminiscent of Figures 2 and 4). Increasing the costs of dispersal and dormancy has the same effects on the evolution of unconditional dormancy as for the evolution of conditional dormancy (see Supplementary Figure S3). As with conditional dormancy, we further found that, in the absence of extinctions, the correlation between the ES rates of conditional dormancy and dispersal is positive when population size is varied and may become slightly negative for large extinction rates and/or dispersal costs (Figure 5A). It should be noted that this tendency is less pronounced when the number of age classes in the seed bank increases (Figure 5B). When the extinction rate is varied for a fixed population size (Figure 5C), we observed that both positive and negative correlations may emerge between dispersal and unconditional dormancy, in contrasted extinctions regimes. This hump-shaped relationship between unconditional dormancy and dispersal is much more pronounced than for conditional dormancy, even for a large number of age classes in the bank (Figure 5D).

Discussion

In this paper, we analyzed the evolution of both dispersal and dormancy in a metapopulation with local extinctions and kin competition. Our model follows from previous attempts (e.g., Cohen and Levin 1987; Venable and Brown 1988) to study the effect of various selective forces on the evolution of dispersal and dormancy. The novelty of our approach is that it combines the effects of temporal variability and kin competition on the joint evolution of these two traits. In the following, we first discuss our results for the evolution of conditional and unconditional dormancy, and then comment on the patterns resulting from the joint evolution of both dispersal and dormancy.

Evolution of conditional and unconditional dormancy

We have analyzed the evolution of conditional dormancy, and we have shown that dormancy of dispersed seeds is always selected against. Philopatric and dispersed seeds indeed respond to very different selective pressures. First, dispersed dormant seeds pay both the cost of dispersal and that of dormancy. Second, dispersed seeds falling in an empty site benefit from immediate germination since this allows them to colonize a new site where competition is minimized (Venable and Lawlor 1980; Olivieri 2001). Last, dispersed seeds falling in an occupied site compete with unrelated individuals; in that case, the role of dormancy as a means to escape kin competition therefore brings no further benefits.

We also observed a non-monotonic relationship between the ES rate of unconditional dormancy and the rate of local extinction (Figure 5C). In our model, the decrease of the rate of unconditional dormancy with larger rates of local extinction results from the fact that the dormancy of dispersed seeds is selected against in newly colonized patches (as we have learned from our results on conditional dormancy). As the rate of local extinctions increases, most dispersed seeds fall in empty sites, which tends to select against dormancy. Such a hump-shaped relationship between the ES rate of unconditional dormancy and the rate of local

extinctions has already been described (see Olivieri 2001). It has been interpreted as resulting from two antagonistic evolutionary forces: local extinctions, which tend to select for more dormancy, and incomplete saturation of local patches following extinction, which weakens local competition and therefore tends to select for less dormancy. Yet this interpretation, which is reminiscent of what has been observed for the evolution of dispersal (see Ronce et al. 2000), does not hold in our model because all the patches that are occupied are saturated (at a fixed population size N). The consequence of incomplete population saturation deserves further attention, though, and could be studied by means of stochastic simulations at low fecundity.

It is worth noting that other forms of conditionality for dormancy may exist in natura. Seeds may for example respond to environmental cues and germinate according the favourability of the upcoming season. In particular, there are some evidence that density-dependent germination may be a means to avoid intense competition (Tielbörger and Valleriani 2005; Tielbörger and Prasse 2009). It would therefore be interesting to extend our model and explore the consequences of kin competition on the evolution of alternative forms of conditional dormancy.

The joint evolution of dispersal and dormancy

In order to generate predictions regarding expected patterns of covariation between dispersal and dormancy, we have analyzed the joint evolution of the two traits. In most cases, we found that a single, joint evolutionary stable strategy was attained. This implies that, whatever the initial conditions, the metapopulation evolves towards this joint ESS. Yet, there were specific situations where the joint evolutionary outcome varied with initial conditions. We could only characterize these bistable equilibria (Supplementary Figure S4) in the case of unconditional dormancy, for a narrow range of parameter values (see Supplementary Figure S5). We found no evidence of bistability in the case of conditional dormancy. Interestingly, previous mod-

els already showed the existence of bistable evolutionary dynamics, but only with periodic changes of the environment (see the Figure 3 in Cohen and Levin 1987).

In addition to the direct costs associated with dispersal and dormancy, multiple factors are involved in the evolution of these traits. First, finite population size tends to increase the relatedness between competing individuals, which may generate indirect benefits for seeds to disperse or to go dormant. Second, both dispersal and dormancy might be viewed as alternative strategies for recolonizing empty patches after local extinction. Our analyses only partially confirm these predictions: while it is true that in many instances, both seed dormancy and dispersal increase with decreasing population sizes (see Figures 5A-B for conditional dormancy), or with increasing extinction rates (see Figure 5C-D for conditional dormancy), some non-trivial results also emerge from our model. The absence of general trends like, e.g., a negative correlation between the ES rates of dispersal and dormancy when the extinction rate or the population size vary, indicates that dispersal and dormancy cannot simply be considered as truly alternative strategies to reduce the risk of local extinction and the cost of kin competition (see Figure 5). The relationship between these traits depends upon the characteristics of the environment (see also Cohen and Levin 1987). The absence of any general trend or syndrome of covariation between dispersal and dormancy is reminiscent of what was found by Ronce et al. (2000) concerning the interactions between reproductive effort and dispersal.

Empirical and experimental perspectives

Measuring accurately dispersal and dormancy is notoriously difficult in many organisms. Yet some of our predictions could in principle be tested, at least in some species. For example in plants, some species have been described as heteromorphic, which means that a single individual produce morphologically differentiated seeds (Olivieri et al. 1983; Venable 1985; McPeek and Kalisz 1998). These species are most commonly found in the Asteraceae

and Chenopodiaceae (Imbert 2002). As discussed in Olivieri (2001), the available data seemingly support our prediction that with conditional dormancy, philopatric seeds are more dormant than dispersed ones. Heteromorphic species indeed produce some seeds that are dispersed and then germinate immediately, and some seeds that are not dispersed and have some probability of entering a dormant stage. This requires further investigation, though, since there might be alternative, non-adaptive interpretations for this pattern related to, e.g., developmental constraints in the formation of seeds on the capitule (but see Olivieri and Berger 1985, who provide examples of heteromorphic species with no seed dormancy, therefore suggesting that constraints are unlikely). Furthermore, some counter-examples exist, like *Bidens frondosa*, which peripheral achenes have a reduced ability to disperse and to go dormant (Brandel 2004).

A broad comparative approach might also be conducted in some clades, to test our predictions. Between-species comparisons have already been used to study the effect of perturbations on the evolution of dormancy in a guild of desert annual plants (Venable 2007), and on the evolution of dispersal in planthoppers (Denno et al. 1991). Similar data sets (see, e.g., Holmes and Newton 2004; Schurr et al. 2007) could potentially be used to test the predicted patterns of covariation between dispersal and dormancy (see Figure 5), in different ecological conditions.

Last, our predictions might also be tested by means of evolution experiments with microorganisms. Experimental evolution has already been used to explore the evolution of dispersal in bacteria (see, e.g., Nakajima and Kurihara 1994; Taylor and Buckling 2010). But some bacteria also have the ability to enter in a dormant, non-dividing state (Balaban et al. 2004; Kussell et al. 2005; Lewis 2007). These persisters may survive to temporal perturbations of their environment (e.g., by resisting to antibiotics Gefen and Balaban 2009). Since the genetic architecture of this trait is well characterized (Rotem et al. 2010), experimental evolution could be used to explore the evolution of dormancy, for various ecological scenarios.

Theoretical perspectives

As we have shown, our model extends previous studies on the evolution of dispersal and dormancy. It relies, however, on simplifying assumptions, the strongest being that environmental variation is uncorrelated in space and time. Yet, temporal and/or spatial correlations of the environment are known to affect the evolution of dispersal and dormancy (Snyder 2006; Cohen and Levin 1987, 1991). For example, periodic changes in the environment may lead to bistable evolutionary dynamics for the evolution of dormancy (Cohen and Levin 1987). Furthermore, positive temporal autocorrelation in environmental conditions has been shown to select for lower rates of dispersal and dormancy (Cohen and Levin 1987; Venable and Brown 1988; Cohen and Levin 1991; Snyder 2006), which may therefore also generate patterns of positive covariation between these traits (Cohen and Levin 1987; Venable and Brown 1988; Cohen and Levin 1991; Snyder 2006). The importance of the spatial correlation of the environment has also been explored theoretically (e.g., Venable and Brown 1988; Snyder 2006) but considering spatial correlation makes only sense if dispersal is limited by distance. Extending our theoretical framework to incorporate these various effects is particularly challenging and the analysis of more complex scenarios will certainly rely exclusively on stochastic simulations. The present model, which incorporates the classical selective forces known to affect the evolution of dispersal and dormancy, may therefore be considered as a stepping stone towards a better understanding of the joint evolution of these two traits in spatially and temporally variable environments.

A Online Appendix

A.1 Selection gradient with class-structure and demographic stochasticity

In order to investigate the evolutionary dynamics of phenotypic traits, we use a direct fitness approach (see Taylor and Frank 1996; Rousset and Billiard 2000) to compute the fitness of a focal individual, as a function of the strategies of all the individuals with which it competes. For convenience, we call an *offspring* in any class, the descendant of a parent that was in any class in the previous time step: e.g., an adult may be the offspring of a dormant seed in the bank in the previous generation, and a dormant seed is likewise the offspring of an adult. A *juvenile* is a germinating seed.

In the model considered here, not all individuals are equivalent. Within a deme, for example, standing individuals and seeds in the bank do not compete with each other. They must therefore belong to different types. Following the life cycle described in the main text, we consider three different types of individuals. Type-$\mathcal{A}$ individuals are adults, type-$\mathcal{S}_p$ individuals are philopatric dormant seeds (i.e. seeds that do not disperse and go dormant) and type-$\mathcal{S}_d$ individuals are dispersed dormant seeds. All the demes are not equivalent either. For example, the demes that have gone extinct in the previous time step cannot contain philopatric dormant seeds (i.e., seeds that would have been produced by resident adults in the previous time step). In these demes, there is therefore no competition between the offspring of standing adults and philopatric dormant seeds. Different categories of demes must therefore be distinguished, depending on the history of extinctions over two successive time steps (see Figure 1B). Demes in category "○○" did not go extinct during the last two generations. Demes in category "⊗○" went extinct two generations ago (but did not last generation). Demes in category "○⊗" went extinct last generation (but did not two generations ago). Demes in category "⊗⊗" went extinct twice in the last two generations (see

Figure 1B). Altogether, twelve demographic classes are so defined (three types of individuals in four categories of demes). Yet, because some types of individuals are absent in some categories of demes, only eight demographic classes are needed. In the following, we use the notation (i,k) for type-i individuals in demes of category k, with $i \in \{\mathcal{A}, \mathcal{S}_p, \mathcal{S}_d\}$ and $k \in \{\bigcirc\bigcirc, \otimes\bigcirc, \bigcirc\otimes, \otimes\otimes\}$.

We assume that each of the phenotypic traits considered is encoded by a bi-allelic locus. Let us first consider the case of dispersal evolution alone (but the following argument holds for all traits): at each locus, we consider a mutant allele A in a population of individuals that bear allele a. We assume that allele a gives phenotype z_a, and that the mutant allele A gives phenotype $z_A \equiv z_a + \epsilon_z$. We further distinguish the value of the trait in a focal individual from its mean value in different categories of actors (e.g., individuals in the focal individual's class, individuals in distinct classes, etc.). The subscript "$_\bullet$" (e.g., $z_\bullet$) refers to the focal individual; the subscript "$_0$" (e.g., z_0) refers to the mean value of the trait in the focal individual's deme, and the subscript "$_1$" (e.g., z_1) refers to the mean value of the trait in the focal individual's deme, in the previous time step. Let $\mathbf{z} \equiv (z_\bullet, z_0, z_1, z, D_\bullet, D_0, D_1, D)$ be the vector of the focal individual's phenotype, and of the average phenotypes of all categories of actors. With conditional dormancy, the vector reads $\mathbf{z} \equiv (z_\bullet, z_0, z_1, z, d_\bullet, d_0, d_1, d, \delta_\bullet, \delta_0, \delta_1, \delta)$.

In order to compute the selection gradient, which determines the fate of the mutant allele A, we need to evaluate the change in allele frequency from one generation to the next. For the sake of clarity, let us first consider a model without demographic stochasticity. Given the vector of allele frequencies $\mathbf{p}$ in the different classes (j,l) in the parental generation at time t, the vector of allele frequencies $\mathbf{p}'$ in the different classes (i,k) after one generation is given by:

$$\mathrm{E}\left[\mathbf{p}'|\mathbf{p}\right] = \mathbf{F}(\mathbf{z})\mathbf{p}. \tag{A.1}$$

Equation (A.1) implies that $\mathbf{F}(\mathbf{z}) \equiv \left(f_{(i,k)\leftarrow(j,l)}(\mathbf{z})\right)$ gives the probability that a gene in class (i,k) is a copy of a gene from a parent in class (j,l). This probability depends upon the fitness

function $w_{(i,k)\leftarrow(j,l)}(\mathbf{z})$ that gives the expected number of offspring in class (i,k) produced by a focal individual in class (j,l):

$$f_{(i,k)\leftarrow(j,l)}(\mathbf{z}) = \frac{N_{jl}}{N_{ik}}\, w_{(i,k)\leftarrow(j,l)}(\mathbf{z}), \tag{A.2}$$

where N_{ik} gives the number of individuals in class (i,k), and N_{jl} the number of individuals in class (j,l). The fitness functions $w_{(i,k)\leftarrow(j,l)}(\mathbf{z})$ depend upon the focal individual's strategy, and the strategies adopted by its competitors.

In a class-structured population, the different demographic classes of individuals make different contributions to the future of the population. To account for these different contributions, the allele frequency in equation (A.1) must be defined as a weighted average of allele frequencies in the different demographic classes. These weights are such that the weighted frequency remains constant over generations in the absence of selection, i.e. with $\epsilon_z = 0$ (Taylor 1990; Rousset 2004, chapter 11). The weights, denoted $\boldsymbol{\alpha}$, are known to be the reproductive values of each class, i.e. the ultimate contributions of all the gene lineages present in a class at time t to the future pool of genes. The reproductive values $\boldsymbol{\alpha}$ are given by the dominant left eigenvector of the backward transition matrix $\mathbf{F}(\mathbf{z})$ of gene lineages between classes, with elements $f_{(i,k)\leftarrow(j,l)}(\mathbf{z})$ evaluated in the absence of selection. In a spatially structured model these backward transition probabilities depend on the dispersal rates (see, e.g., Leturque and Rousset 2002), and with demographic structure they additionally depend on the transition rates between different demographic classes for non-dispersed genes (see, e.g., Rousset 1999; Rousset and Ronce 2004).

Furthermore, the demography may vary over generations and demographic fluctuations may depend upon the traits under selection. In our model, the absence of density dependence in the seed bank allows for some variation in the density of seeds in the bank that depend, among other things, on the rate of dormancy. The functions $f_{(i,k)\leftarrow(j,l)}(\mathbf{z})$ therefore depend on the demographic state of the metapopulation, which may differ from one generation

to the next. Let $\mathbf{N}$ represent the demographic state of the metapopulation at time t. $\mathbf{N}$ is characterized by the number of individuals in each class, which includes the number of adults and the size of the seed bank in the different categories of deme. The prime superscript (') indicates that the parameter is evaluated at time $t+1$. The expected number of offspring in class (i,k) of a focal individual with genotype A in class (j,l) is then given by the fitness function $w_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})$ that depends upon the focal individual's strategy, the strategies adopted by its competitors, and the demographic states of the metapopulation at t and $t+1$. Exact expressions for $w_{ik\leftarrow jl}(\mathbf{N},\mathbf{N}',\mathbf{z})$ are given below. Let N_{jl} be the number of parents in class (j,l) at t, and N'_{ik} be the number of offspring in class (i,k) at $t+1$. Then, the backward transition matrix of gene lineages between classes reads $\mathbf{F}(\mathbf{N},\mathbf{N}',\mathbf{z}) \equiv \left(f_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})\right)$ and equation (A.2) reads:

$$f_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z}) = \frac{N_{jl}}{N'_{ik}}\, w_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z}). \tag{A.3}$$

Taking expectations over all possible demographic states $\mathbf{N}'$ at time $t+1$, the expected allele frequency in the offspring generation develops as:

$$\mathrm{E}\left[\boldsymbol{\alpha}(\mathbf{N}')\cdot\mathbf{p}'|\mathbf{p},\mathbf{N}\right] = \sum_{\mathbf{N}'}\boldsymbol{\alpha}(\mathbf{N}')\Pr\left(\mathbf{N}'|\mathbf{N},\mathbf{z}\right)\mathbf{F}(\mathbf{N},\mathbf{N}',\mathbf{z})\mathbf{p}, \tag{A.4}$$

where $\Pr\left(\mathbf{N}'|\mathbf{N},\mathbf{z}\right)$ is the conditional probability that the demographic state of the metapopulation is $\mathbf{N}'$ at time $t+1$, given it was $\mathbf{N}$ at t (Rousset and Ronce 2004). $\Pr\left(\mathbf{N}'|\mathbf{N},\mathbf{z}\right)$ therefore represents the transition probability between the demographic states of the metapopulation over one generation.

The selection gradient S, which is also the inclusive fitness effect under weak selection, is then obtained by taking the derivative of the right-hand side of equation (A.4), with respect to a change in phenotypic effect ϵ_z (Hamilton 1964). The gradient of selection S measures the first order effect of selection on the weighted change of mutant frequency. Rousset and

Ronce (2004) showed that this gradient of selection reduces to two terms: $S = S_{\mathrm{f}} + S_{\mathrm{Pr}}$. The first term, S_{f}, involves derivatives of the elements of $\mathbf{F}(\mathbf{N}, \mathbf{N}', \mathbf{z})$ and gives the selection component due to allele frequency changes in descendants from each parental class, given the distribution of class sizes determined by the resident trait values. The second term, S_{Pr}, involves derivatives of the $\Pr(\mathbf{N}'|\mathbf{N}, \mathbf{z})$'s and gives the selection component due to changes in the reproductive value of gene lineages, as a consequence of changes in the probability that a descendant gene copy finds itself in a given class. In other words, this latter term measures the influence of the neighbours of the focal individual on direct fitness via their impact on the future demographic state of the populations. In models where the trait under selection does not affect the demographic dynamics of the population (e.g., Taylor 1990; Taylor and Frank 1996; Leturque and Rousset 2002) the term S_{Pr} is nil.

A.2 Approximating the selection gradient

Because the bank size can take large values, a very large number of terms should be considered in equation (A.4): if fecundity is Poisson distributed, then the number of terms in $\mathbf{N}'$ is infinite, unless some more or less arbitrary truncation is performed. Nevertheless, as in Leturque and Rousset (2004) and Lehmann et al. (2006), good approximations can be derived. In particular, if we assume that the variation of reproductive value with bank size is small, we do not need to consider the selection component due to changes in the reproductive value of gene lineages as a consequence of changes in class sizes. Then, the effect of the phenotype under selection on the bank size can be neglected. It is important to realize that this approximation neglects the second term S_{Pr} of the selection gradient, which measures the influence of the neighbours of the focal individual on direct fitness via their impact on the future demographic state of the populations (see above). In other words our analysis does not take into account the evolutionary consequences of demographic stochasticity. As shown in the main text, our approximation yields predictions that are remarkably consistent

with individual-based simulations.

However, seed bank size also affects the fitness functions $w_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})$ and the functions $f_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})$, as will be detailed below, and here too there is no easy simplification. Therefore, in the following, we neglect demographic fluctuations. Thus, the weighted change in the mutant frequency over one generation reduces from equation (A.4) to:

$$\mathrm{E}\left[\boldsymbol{\alpha}\cdot\mathbf{p}'|\mathbf{p}\right] = \boldsymbol{\alpha}\mathbf{F}(\mathbf{z})\mathbf{p}. \tag{A.5}$$

Since we neglect demographic fluctuations, the fitness functions $w_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})$ and the functions $f_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})$ may be written, for simplicity, as $w_{(i,k)\leftarrow(j,l)}(\mathbf{z})$ and $f_{(i,k)\leftarrow(j,l)}(\mathbf{z})$. In the following, we will use the shorthand notations $w_{(i,k)\leftarrow(j,l)}$ and $f_{(i,k)\leftarrow(j,l)}$ for brevity, since these functions always depend upon the phenotypes $\mathbf{z}$.

Furthermore, and because we consider an infinite island model of population structure, we assume that the demographic state of the metapopulation converges to a stationary equilibrium (Chesson and Warner 1981). In our model, where we neglect demographic fluctuations, the demographic state of the metapopulation is characterized by the distribution of deme categories, which depends upon the history of local extinctions. In order to characterize the demographic state of the metapopulation, we need to consider the forward transition probability $u(i|j)$ from demes in category j at t to demes in category i at $t+1$. It is easy to see from Figure 1B, that the matrix of forward transition probabilities $\mathbf{U}$ with (i,j)th element $u(i|j)$ reads:

$$\mathbf{U} = \begin{pmatrix} 1-e & 1-e & 0 & 0 \\ 0 & 0 & 1-e & 1-e \\ e & e & 0 & 0 \\ 0 & 0 & e & e \end{pmatrix}. \tag{A.6}$$

Then, the stationary distribution of deme categories is given by the dominant right eigen-

vector $\mathbf{P} \equiv (P(i))$ of the matrix $\mathbf{U} \equiv (u(i|j))$ (see, e.g., Taylor 1990), i.e.:

$$\mathbf{P} = \begin{pmatrix} (1-e)^2 \\ e(1-e) \\ e(1-e) \\ e^2 \end{pmatrix}. \tag{A.7}$$

Hence, demes in category "○○" are those that have not been extinct for two successive generations, and are in frequency $(1-e)^2$ in the metapopulation; demes in category "⊗⊗" are those that have faced two successive extinctions, and are in frequency e^2 in the metapopulation. It will also prove to be useful to define the backward transition probability that a deme in category k at $t+1$ was in category l at t, i.e $v(j|i) = u(i|j)P(j)/P(i)$. The matrix of backward transition probabilities $\mathbf{V}$ with (i,j)th element $v(j|i)$ reads:

$$\mathbf{V} = \begin{pmatrix} 1-e & e & 0 & 0 \\ 0 & 0 & 1-e & e \\ 1-e & e & 0 & 0 \\ 0 & 0 & 1-e & e \end{pmatrix}. \tag{A.8}$$

A.3 Formulas for computation

In the following, we distinguish the contribution of a focal individual to its deme (philopatric offspring), from its contribution to other demes (dispersed offspring): we note $w^P_{(i,k)\leftarrow(j,l)}$ the expected number of philopatric offspring in class (i,k) from a focal individual in class (j,l), and $w^D_{(i,k)\leftarrow(j,l)}$ the expected number of dispersed offspring in class (i,k) from a focal individual in class (j,l). Therefore, $f^P_{(i,k)\leftarrow(j,l)}$ (resp. $f^D_{(i,k)\leftarrow(j,l)}$) gives the probability that a philopatric (resp. dispersed) gene in class (i,k) is a copy of a gene from any of the A parent in class (j,l). Both $f^P_{(i,k)\leftarrow(j,l)}$ and $f^D_{(i,k)\leftarrow(j,l)}$ contribute to the expression $f_{(i,k)\leftarrow(j,l)}$ that gives the total probability that a gene in (i,k) is a copy of a gene in (j,l). Because the expected

number of dispersed offspring of a focal adult may depend upon the category m of the deme reached by the offspring, we get:

$$f_{(i,k)\leftarrow(j,l)} = v(l|k) f^{P}_{(i,k)\leftarrow(j,l)} + P(l) \sum_{m} v(m|k) f^{D}_{(i,k)\leftarrow(j,l)}(m). \tag{A.9}$$

The function $f_{(i,k)\leftarrow(j,l)}$ gives the total backward transition probability that a gene lineage in class (i,k) at t+1 was in class (j,l) at t. The first term in the right-hand side of equation (A.9) gives the probability that an allele A in class (i,k) at $t+1$ is the copy of a philopatric gene that was in class (j,l) at t. The second term in the right-hand side of equation (A.9) gives the probability that an allele A in class (i,k) at $t+1$ is the copy of a gene originally in a deme of category l that has been dispersed in a deme that was in category m at t.

From equation (A.5), the unweighted change of allele frequency reads:

$$\mathrm{E}\left[\boldsymbol{\alpha}\cdot\mathbf{p}'|\mathbf{p}\right] = \sum_{i,k} \alpha(i,k) p'_{ik} = \sum_{i,k} \alpha(i,k) \sum_{j,l} f_{(i,k)\leftarrow(j,l)} p_{jl}. \tag{A.10}$$

From equation (A.9), and using an appropriate change of variable to factorize the $v(l|k)$ terms, we get:

$$\mathrm{E}\left[\boldsymbol{\alpha}\cdot\mathbf{p}'|\mathbf{p}\right] = \sum_{i,k} \alpha(i,k) \sum_{l} v(l|k) \sum_{j} \left(f^{P}_{(i,k)\leftarrow(j,l)} p_{jl} + \sum_{m} P(m) f^{D}_{(i,k)\leftarrow(j,m)}(l) p_{jm} \right). \tag{A.11}$$

The first order effect of selection on the change of this weighted sum of mutant frequency $\Delta\left(\boldsymbol{\alpha}\cdot\mathbf{p}\right) \equiv \boldsymbol{\alpha}\cdot\mathbf{p}' - \boldsymbol{\alpha}\cdot\mathbf{p}$ is given by the selection gradient:

$$S(\mathbf{z}) = \frac{d\mathrm{E}\left[\Delta\left(\boldsymbol{\alpha}\cdot\mathbf{p}\right)\right]}{d\epsilon_{\mathrm{z}}}. \tag{A.12}$$

Following equation (A.12), we now take the derivative of equation (A.11) for all c-actors acting on the focal individual. In this computation, the different partial derivatives of the fitness functions with respect to each element z_c of the $\mathbf{z}$ vector, $\partial f^{P}_{(i,k)\leftarrow(j,l)} / \partial z_c$ give the

change of the focal individual's fitness due to the effects of c-actors. These terms are weighted by the extent to which the actors' strategy is affected, i.e. by the derivative of z_c with respect to the phenotypic effect, $dz_c(\mathbf{p})/d\epsilon_{\mathrm{z}}$, which is simply the allele frequency p_c among the class of individuals which phenotype is represented by z_c. These p_c's come in factor with elements p_{jl} of $\mathbf{p}$ in equation (A.5), and these products of allele frequencies $p_{jl}p_c$ may then be expressed as functions of probabilities of identity between appropriate pairs of genes. This forms the logical basis of the direct fitness method for computation of fitness gradients (Taylor and Frank 1996; Rousset and Billiard 2000). For this computation, probabilities of genetic identity at neutrality are sufficient since effects of selection on these probabilities would only contribute to higher order effects on allele frequency (for the latter computations see Ajar 2003; Roze and Rousset 2008). Overall, the approximate gradient computed from equation (A.12) then reads:

$$\begin{aligned} S(\mathbf{z}) &= \sum_{i,k}\alpha(i,k)\sum_{l}v(l|k)\sum_{j}\left(\sum_{c=\bullet,0,1}\frac{\partial f^{P}_{(i,k)\leftarrow(j,l)}}{\partial z_c}Q^{c}_{(j,l)}\right. \\ &+ \left.\sum_{m}P(m)\sum_{c=\bullet,0,1}\frac{\partial f^{D}_{(i,k)\leftarrow(j,m)}(l)}{\partial z_c}Q^{c}_{(j,m)}\right). \end{aligned} \tag{A.13}$$

Since the weighted allele frequency is by definition a function of reproductive values which are not defined as function of $\mathbf{z}$, the reproductive values in equation (A.13) are also considered at neutrality. The gradient of selection in equation (A.13) gives the first order effects of c-actors upon the number of offspring in class (i,k) of a focal individual, weighted by the probabilities of genetic identity $Q^{c}_{(j,l)}$ and $Q^{c}_{(j,m)}$ between the focal individual's gene in class (j,l) or (j,m) and a c-actor's genes. The first term in the right-hand side of equation (A.13) gives the first order effects of actors on philopatric seeds, while the second term in the right-hand side of equation (A.13) gives the first order effects of actors on dispersed seeds.

In the infinite island model considered here, the identity probabilities between genes in different demes can be considered nil, and the within-deme probabilities can be computed as

probabilities of "identity by descent" (IBD) following standard techniques (see, e.g., Crow and Kimura 1970; Rousset 2002). Therefore, the first order effects upon the offspring of a focal individual of c-actors in different demes have a null weight (and thus, all the $\partial f^P_{(i,k)\leftarrow(j,l)}/\partial z_c$ and the $\partial f^D_{(i,k)\leftarrow(j,l)}/\partial z_c$ terms vanish from the above expression).

Furthermore, the first order effects upon the focal individual's dispersed offspring of any actor but itself have a null weight. Thus, all the $\partial f^D_{(i,k)\leftarrow(j,l)}/\partial z_c$ terms with $c \neq \bullet$ vanish from the above expression. It follows that in the model presented here,

$$\begin{aligned} S(\mathbf{z}) &= \sum_{i,k}\alpha(i,k)\sum_{l}v(l|k)\sum_{j}\left(\frac{\partial f^P_{(i,k)\leftarrow(j,l)}}{\partial z_\bullet}+\frac{\partial f^P_{(i,k)\leftarrow(j,l)}}{\partial z_0}Q^0_{(j,l)}+\frac{\partial f^P_{(i,k)\leftarrow(j,l)}}{\partial z_1}Q^1_{(j,l)}\right. \\ &+ \left.\sum_{m}P(m)\frac{\partial f^D_{(i,k)\leftarrow(j,m)}(l)}{\partial z_\bullet}\right), \end{aligned} \tag{A.14}$$

where $Q^0_{(j,l)}$ is the IBD probability between a focal in class (j,l) and an adult actor in its deme; likewise, $Q^1_{(j,l)}$ is the IBD probability between a focal in class (j,l) at t and an adult actor at $t-1$ in its deme (see below). For these computations, probabilities of genetic identity at neutrality are sufficient since effects of selection on these probabilities would only contribute to higher order effects on allele frequency (for the latter computations see Ajar 2003). In the gradient computation, reproductive values are also considered at neutrality. However, both the probabilities of identity and the reproductive values are function of the resident trait value in which the derivatives are computed.

We have provided an expression for the convergence stability condition for the evolution of the dispersal fraction in the model. Expressions for the convergence stability conditions for the evolution of other traits follow by replacing z with parameters D, d and δ in the above expressions.

A.4 General expressions for fitness functions

Let us now derive the expected number of offspring in any class from parents in any class. In the following, we derive the exact expressions for the fitness functions $w_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})$ and the functions $f_{(i,k)\leftarrow(j,l)}(\mathbf{N},\mathbf{N}',\mathbf{z})$. In particular, we consider the full distributions of offspring numbers in order to compute the expected numbers of offspring in each class. Then, in the next section, we will provide the approximate expressions used in the main text.

Adults (type-$\mathcal{A}$ individuals) exist only in demes of category ○○ and ⊗○. We note r the fecundity of adults. In demes of category ○○ and ⊗○, each focal adult produces a random, Poisson distributed, number of seeds $\sim \mathcal{P}(r)$. A fraction $(1-z_\bullet)(1-d_\bullet)$ of seeds is not dispersed and germinates in the following generation. Likewise, a fraction $z_\bullet(1-c_z)(1-\delta_\bullet)$ is dispersed and germinates in the following generation. Thus, one adult in a focal deme of category ○○ or ⊗○ produces $\sim \mathcal{P}\left[r(1-z_\bullet)(1-d_\bullet)\right]$ philopatric non-dormant seeds, and $\sim \mathcal{P}\left[rz_\bullet(1-c_z)(1-\delta_\bullet)\right]$ dispersed non-dormant seeds. The adults at t produce J_0^P philopatric juveniles at $(t+1)$ in a focal deme of category ○○ or ⊗○

$$J_0^P \sim \mathcal{P}\left[Nr(1-z_0)(1-d_0)\right], \tag{A.15}$$

and J^D dispersed juveniles

$$J^D \sim \mathcal{P}\left[N(1-e)rz(1-c_z)(1-\delta)\right]. \tag{A.16}$$

Likewise, the adults in other demes of category ○○ or ⊗○ at t produce J^P philopatric juveniles at $(t+1)$

$$J^P \sim \mathcal{P}\left[Nr(1-z)(1-d)\right], \tag{A.17}$$

and J^D dispersed juveniles at $(t+1)$.

All the germinating seeds, be they issued from the bank or from the adults in the previous time-step, are in competition. In a focal deme of category $\circ\circ$ or $\otimes\circ$, G_0^P philopatric seeds, which are dormant at t, have been produced at $(t-1)$:

$$G_0^P \sim \mathcal{P}\left[Nr(1-z_1)d_1(1-c_\mathrm{d})\right], \tag{A.18}$$

and G^D dispersed seeds, which are dormant at t

$$G^D \sim \mathcal{P}\left[N(1-e)rz(1-c_\mathrm{z})\delta(1-c_\mathrm{d})\right]. \tag{A.19}$$

Each seed in the bank produces a single juvenile. Thus, the total number of seeds (both philopatric and dispersed) that germinate at $(t+1)$ from the bank, e.g. in a focal deme of category $\circ\circ$, is $G_0^P + G^D$. Likewise, the number of philopatric seeds that germinate at $(t+1)$ from the bank in another deme is

$$G^P \sim \mathcal{P}\left[Nr(1-z)d(1-c_\mathrm{d})\right], \tag{A.20}$$

and the number of dispersed seeds is G^D, as before.

In the following we distinguish the contribution of a focal individual to its deme (philopatric offspring), from its contribution to other demes (dispersed offspring). We note $w^P_{(i,k)\leftarrow(j,l)}$ the expected number of philopatric offspring in class (i,k) from a focal individual in class (j,l) and $w^D_{(i,k)\leftarrow(j,l)}$ the expected number of dispersed offspring in class (i,k) from a focal individual in class (j,l). These two functions contribute to the expression $w_{(i,k)\leftarrow(j,l)}$ that gives the total expected number of offspring in (i,k) from a focal in (j,l).

A.4.1 Adult offspring from adults

The expected number of philopatric offspring in a deme of category ○○ of a focal adult in a deme of category ○○ is given by

$$w^{P}_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)}(\mathbf{N}) = N\mathrm{E}\left[\frac{\mathcal{P}\left[r(1-z_{\bullet})(1-d_{\bullet})\right]}{G^{P}_{0}+G^{D}+J^{P}_{0}+J^{D}}|G^{P}_{0}+G^{D}\right], \tag{A.21}$$

where the expectation is conditional upon the total number $(G^{P}_{0}+G^{D}_{0})$ of seeds in the bank of the focal deme, and is taken over the distributions of all the juveniles produced. Note that the random variables in numerator and denominator of each ratio are not independent.

The expected number of dispersed offspring in a deme of category ○○ of a focal adult in a deme of category ○○ depends upon the ancestral category m of the deme reached by the offspring

$$w^{D}_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)}(\mathbf{N},m) = N\mathrm{E}\left[\frac{\mathcal{P}\left[rz_{\bullet}(1-c_{\mathrm{z}})(1-\delta_{\bullet})\right]}{G^{P}+G^{D}+J^{P}+J^{D}}|G^{P},G^{D}\right], \quad \text{if } m=\circ\circ, \tag{A.22}$$

and

$$w^{D}_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)}(\mathbf{N},m) = N\mathrm{E}\left[\frac{\mathcal{P}\left[rz_{\bullet}(1-c_{\mathrm{z}})(1-\delta_{\bullet})\right]}{G^{D}+J^{P}+J^{D}}|G^{D}\right], \quad \text{if } m=\otimes\circ\,. \tag{A.23}$$

The right-hand side of equation (A.22) represents the expected number of dispersed offspring that reach a deme of category ○○ that do not go extinct at $t+1$. There, the competition is among all the juveniles, i.e. those born from philopatric and dispersed non-dormant seeds as well as those born from philopatric and dispersed dormant seeds. The right-hand side of equation (A.23) represents the expected number of dispersed offspring that reach a deme of category ⊗○ that do not go extinct at $t+1$, where the juveniles born from philopatric dormant seeds are absent (see Figure 1B), and thus do not compete.

The expected number of philopatric offspring in a deme of category ○○ of a focal adult

in a deme of category $\otimes\circ$ is given by

$$w^{P}_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\otimes\circ)}(\mathbf{N}) = N\mathrm{E}\left[\frac{\mathcal{P}\left[r(1-z_{\bullet})(1-d_{\bullet})\right]}{G^{D}+J^{P}_{0}+J^{D}}|G^{D}\right] \tag{A.24}$$

where the expectation is conditional upon the total number G^D of seeds in the bank of the focal deme. Note that there are no philopatric dormant seeds in competition in that case. The expected number of dispersed offspring in a deme of category $\circ\circ$ of a focal adult in a deme of category $\otimes\circ$ is given by

$$w^{D}_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\otimes\circ)}(\mathbf{N},m) = w^{D}_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)}(G^{P},G^{D},m) \tag{A.25}$$

The expected number of philopatric offspring in a deme of category $\otimes\circ$ of a focal adult in a deme of category $\circ\circ$ is nil, because a deme of category $\otimes\circ$ cannot derive from a deme of category $\circ\circ$. The expected number of (dispersed) offspring in all demes of category $\otimes\circ$ of a focal adult in a deme of category $\circ\circ$ depends upon the ancestral class of the deme reached by dispersed seeds, and is given by

$$w^{D}_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{A},\circ\circ)}(\mathbf{N},m) = N\mathrm{E}\left[\frac{\mathcal{P}\left[rz_{\bullet}(1-c_{\mathrm{z}})(1-\delta_{\bullet})\right]}{G^{P}+G^{D}+J^{D}}|G^{P},G^{D}\right], \quad \text{if } m=\circ\otimes, \tag{A.26}$$

and

$$w^{D}_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{A},\circ\circ)}(\mathbf{N},m) = N\mathrm{E}\left[\frac{\mathcal{P}\left[rz_{\bullet}(1-c_{\mathrm{z}})(1-\delta_{\bullet})\right]}{G^{D}+J^{D}}|G^{D}\right], \quad \text{if } m=\otimes\otimes\,. \tag{A.27}$$

The right-hand side of equation (A.26) represents the expected number of dispersed offspring in demes of category $\circ\otimes$ that do not go extinct at $t+1$. In such demes, the competition is between juveniles born from philopatric and dispersed dormant seeds and dispersed adults only (because there was no adult in demes of category $\circ\otimes$, there can be no philopatric juveniles produced). The right-hand side of equation (A.27) represents the expected number

of dispersed offspring in demes of category $\otimes\otimes$, that do not go extinct at $t+1$. There, the competition is between juveniles born from dispersed seeds only (dormant or not). Likewise, the expected number of (dispersed) offspring in a deme of category $\otimes\circ$ of a focal adult in a deme of category $\otimes\circ$ is given by the same expression

$$w^D_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{A},\otimes\circ)}(\mathbf{N},m) = w^D_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{A},\circ\circ)}(G^P,G^D,m). \tag{A.28}$$

A.4.2 Adult offspring from dormant seeds

The number of offspring in demes of category $\circ\circ$ of a focal philopatric dormant seed in a deme in category $\circ\circ$ is

$$w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_p,\circ\circ)}(\mathbf{N}) = N\mathrm{E}\left[\frac{1}{G^P_0+G^D+J^P_0+J^D}|G^P_0,G^D\right], \tag{A.29}$$

where the expectation is taken over the distribution of $(J^P_0+J^D)$. However, the number of offspring in demes of category $\circ\circ$ of a focal philopatric seed in a deme in category $\circ\otimes$ is $w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_p,\circ\otimes)}=0$, because demes of category $\circ\circ$ cannot derive from demes of category $\circ\otimes$. The numbers of offspring of a focal dispersed seed $w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_d,\circ\circ)}(G^D)$ is given by the same expression as $w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_p,\circ\circ)}(G^P_0,G^D)$, i.e.

$$w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_d,\circ\circ)}(\mathbf{N}) = w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_p,\circ\circ)}(G^P_0,G^D). \tag{A.30}$$

Likewise,

$$w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_d,\otimes\circ)}(\mathbf{N}) = N\mathrm{E}\left[\frac{1}{G^D+J^P_0+J^D}|G^D\right]. \tag{A.31}$$

However, $w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_d,\circ\otimes)} = w^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_d,\otimes\otimes)} = 0$, because demes of category $\circ\circ$ cannot derive from demes of categories $\circ\otimes$ and $\otimes\otimes$.

The number of offspring in demes of category $\otimes\circ$ of a focal philopatric seed in a deme in

category $\circ\circ$ is $w^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_p,\circ\circ)} = 0$, because demes of category $\otimes\circ$ cannot derive from demes of category $\circ\circ$. The number of offspring in demes of category $\circ\circ$ of a focal philopatric seed in a deme in category $\circ\otimes$ is

$$w^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_p,\circ\otimes)}(\mathbf{N}) = N\mathrm{E}\left[\frac{1}{G_0^P + G^D + J^D}|G_0^P, G^D\right], \tag{A.32}$$

because there are no adults in demes of category $\circ\otimes$, there can be no philopatric juveniles produced. The number of offspring of a focal dispersed seed $w_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_d,\circ\circ)} = w_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_d,\otimes\circ)}$ are both nil, because demes of category $\otimes\circ$ cannot derive from demes of categories $\circ\circ$ and 2. However, the expected number of offspring in a deme of category $\otimes\circ$ of a focal dispersed seed in a deme of category $\circ\otimes$ is given by

$$w^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_d,\circ\otimes)}(\mathbf{N}) = w^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_p,\circ\otimes)}(G_0^P, G^D), \tag{A.33}$$

and the expected number of offspring in a deme of category $\otimes\circ$ of a focal dispersed seed of a deme of category $\otimes\otimes$ is given by

$$w^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_d,\otimes\otimes)}(\mathbf{N}) = N\mathrm{E}\left[\frac{1}{G^D + J^D}|G^D\right], \tag{A.34}$$

because competition is only between juveniles born from dispersed seeds, dormant or not.

A.4.3 Dormant seed offspring from adults

A focal adult at t in a deme of category $\circ\circ$ or $\otimes\circ$ produces $\sim \mathcal{P}\left[r(1-z_\bullet)d_\bullet(1-c_\mathrm{d})\right]$ dormant seeds in its deme. Its number of offspring (G_0') is given conditional upon the bank size in the deme in the next generation:

$$G_0' \sim \mathcal{P}\left[Nr(1-z_0)d_0(1-c_\mathrm{d}) + N(1-e)rz(1-c_\mathrm{z})\delta(1-c_\mathrm{d})\right]. \tag{A.35}$$

The number of philopatric dormant seeds of the focal adult, given G_0', has the distribution of a Poisson variable observed conditionally on a sum of independent Poisson distributed variables including itself. This is a binomial distribution $\mathcal{B}(G_0', p)$, where p is the ratio of the expectation of the number of the focal's seeds over that of G_0'.

$$\begin{aligned} w^P_{(\mathcal{S}_p,\bigcirc\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(\mathbf{N}') &= w^P_{(\mathcal{S}_p,\bigcirc\bigcirc)\leftarrow(\mathcal{A},\otimes\bigcirc)}(G_0') = \mathrm{E}\left[\mathcal{B}\left(G_0', \frac{r(1-z_\bullet)d_\bullet(1-c_\mathrm{d})}{\mathrm{E}[G_0']}\right)\right] \\ &= G_0'\frac{r(1-z_\bullet)d_\bullet(1-c_\mathrm{d})}{\mathrm{E}[G_0']}. \end{aligned} \tag{A.36}$$

Likewise, $w^P_{(\mathcal{S}_p,\bigcirc\otimes)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(\mathbf{N}') = w^P_{(\mathcal{S}_p,\bigcirc\otimes)\leftarrow(\mathcal{A},\otimes\bigcirc)}(\mathbf{N}') = w^P_{(\mathcal{S}_p,\bigcirc\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(\mathbf{N}')$. A focal adult at t in a deme of category $\bigcirc\bigcirc$ or $\otimes\bigcirc$ produces $\sim \mathcal{P}\left[rz_\bullet(1-c_\mathrm{z})\delta_\bullet(1-c_\mathrm{d})\right]$ dispersed dormant seeds at $t+1$. Its number of offspring is given conditional upon the bank size in the deme in the next generation G':

$$G' \sim \mathcal{P}\left[Nr(1-z)d(1-c_\mathrm{d}) + N(1-e)rz(1-c_\mathrm{z})\delta(1-c_\mathrm{d})\right]. \tag{A.37}$$

The number of dispersed dormant offspring of the focal in demes of category $\bigcirc\bigcirc$ or $\otimes\bigcirc$ is

$$\begin{aligned} w^D_{(\mathcal{S}_d,\bigcirc\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(\mathbf{N}') &= w^D_{(\mathcal{S}_d,\bigcirc\bigcirc)\leftarrow(\mathcal{A},\otimes\bigcirc)}(G') = \mathrm{E}\left[\Pr(G')\mathcal{B}\left(G', \frac{rz_\bullet(1-c_\mathrm{z})\delta_\bullet(1-c_\mathrm{d})}{\mathrm{E}[G']}\right)\right] \\ &= \Pr(G')G'\frac{rz_\bullet(1-c_\mathrm{z})\delta_\bullet(1-c_\mathrm{d})}{\mathrm{E}[G']}, \end{aligned} \tag{A.38}$$

in demes that are of category $\bigcirc\bigcirc$ at $t+1$. $\Pr(G')$ gives the probability that the total number of seeds in the deme attained by the focal's seeds is G'. The expected number of dispersed dormant seeds from the focal individual is the same in all categories of demes. This is so because there is no competition among the seeds in the bank and because the total number of dispersed dormant seeds is identically distributed whatever the category of the deme.

Therefore,

$$w^D_{(\mathcal{S}_d,\cdot)\leftarrow(\mathcal{A},\circ\circ)}(\mathbf{N}') \quad = \quad w^D_{(\mathcal{S}_d,\cdot)\leftarrow(\mathcal{A},\otimes\circ)}(G'). \tag{A.39}$$

It is assumed that the seeds in the bank cannot survive over one generation. Thus,

$$w^P_{(\mathcal{S}_p,i)\leftarrow(\mathcal{S}_p,j)} = w^P_{(\mathcal{S}_p,i)\leftarrow(\mathcal{S}_d,j)} = w^P_{(\mathcal{S}_d,i)\leftarrow(\mathcal{S}_p,j)} = w^P_{(\mathcal{S}_d,i)\leftarrow(\mathcal{S}_d,j)} = 0, \tag{A.40}$$

for all i's and j's.

A.5 Approximate fitness functions

In the following, we derive the approximate fitness functions that are obtained by neglecting demographic fluctuations. In particular, we replace the expectation of ratios of random variables in the previous expressions for fitness functions by the ratio of expectations of these variables. Furthermore, we consider that, in all situations, the number of seeds in the bank is equal to the expectation of that number, i.e. that the numbers of individuals in the different classes at t are $N_{\mathcal{A}\cdot} = N$, $N_{\mathcal{S}_p\cdot} = Nr(1-z_1)d_1(1-c_\mathrm{d})$, and $N_{\mathcal{S}_d\cdot} = N(1-e)rz(1-c_\mathrm{z})\delta(1-c_\mathrm{d})$. Likewise, the numbers of individuals in the different classes at $t+1$ are $N'_{\mathcal{A}\cdot} = N$, $N'_{\mathcal{S}_p\cdot} = Nr(1-z_0)d_0(1-c_\mathrm{d})$, and $N'_{\mathcal{S}_d\cdot} = N(1-e)rz(1-c_\mathrm{z})\delta(1-c_\mathrm{d})$. As we will demonstrate, approximating the distribution of seed bank sizes with its expectation yields much simpler fitness functions. From the definition of the function $f_{(i,k)\leftarrow(j,l)}$ given in equation (A.3), and from the above approximations, we get the following expressions:

$$f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)} = \frac{(1-z_\bullet)(1-d_\bullet)}{(1-z_1)d_1(1-c_\mathrm{d}) + (1-z_0)(1-d_0) + (1-e)z(1-c_\mathrm{z})(1-\delta c_\mathrm{d})}, \tag{A.41}$$

$$f^{D}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(m=\bigcirc\bigcirc)=\frac{z_\bullet(1-c_z)(1-\delta_\bullet)}{(1-z)d(1-c_d)+(1-z)(1-d)+(1-e)z(1-c_z)(1-\delta c_d)}, \tag{A.42}$$

$$f^{D}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(m=\otimes\bigcirc)=\frac{z_\bullet(1-c_z)(1-\delta_\bullet)}{(1-z)(1-d)+(1-e)z(1-c_z)(1-\delta c_d)}, \tag{A.43}$$

$$f^{P}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{A},\otimes\bigcirc)}=\frac{(1-z_\bullet)(1-d_\bullet)}{(1-z_0)(1-d_0)+(1-e)z(1-c_z)(1-\delta c_d)}, \tag{A.44}$$

$$f^{D}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{A},\otimes\bigcirc)}(m)=f^{D}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(m), \tag{A.45}$$

$$f^{P}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{S}_p,\bigcirc\bigcirc)}=\frac{(1-z_1)d_1(1-c_d)}{(1-z_1)d_1(1-c_d)+(1-z_0)(1-d_0)+(1-e)z(1-c_z)(1-\delta c_d)}, \tag{A.46}$$

$$f^{P}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{S}_d,\bigcirc\bigcirc)}=\frac{(1-e)z(1-c_z)\delta(1-c_d)}{(1-z_1)d_1(1-c_d)+(1-z_0)(1-d_0)+(1-e)z(1-c_z)(1-\delta c_d)}, \tag{A.47}$$

$$f^{P}_{(\mathcal{A},\bigcirc\bigcirc)\leftarrow(\mathcal{S}_d,\otimes\bigcirc)}=\frac{(1-e)z(1-c_z)\delta(1-c_d)}{(1-z_0)(1-d_0)+(1-e)z(1-c_z)(1-\delta c_d)}, \tag{A.48}$$

$$f^{D}_{(\mathcal{A},\otimes\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(m=\bigcirc\otimes)=\frac{z_\bullet(1-c_z)(1-\delta_\bullet)}{(1-z)d(1-c_d)+(1-e)z(1-c_z)(1-\delta c_d)}, \tag{A.49}$$

$$f^{D}_{(\mathcal{A},\otimes\bigcirc)\leftarrow(\mathcal{A},\bigcirc\bigcirc)}(m=\otimes\otimes)=\frac{z_\bullet(1-\delta_\bullet)}{(1-e)z(1-\delta c_d)}, \tag{A.50}$$

$$f^D_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{A},\otimes\circ)}(m) = f^D_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{A},\circ\circ)}(m), \tag{A.51}$$

$$f^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_p,\circ\otimes)} = \frac{(1-z_1)d_1(1-c_\mathrm{d})}{(1-z_1)d_1(1-c_\mathrm{d}) + (1-e)z(1-c_\mathrm{z})(1-\delta c_\mathrm{d})}, \tag{A.52}$$

$$f^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_d,\circ\otimes)} = \frac{(1-e)z(1-c_\mathrm{z})\delta(1-c_\mathrm{d})}{(1-z_1)d_1(1-c_\mathrm{d}) + (1-e)z(1-c_\mathrm{z})(1-\delta c_\mathrm{d})}, \tag{A.53}$$

$$f^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_d,\otimes\otimes)} = \frac{\delta(1-c_\mathrm{d})}{(1-\delta c_\mathrm{d})}, \tag{A.54}$$

$$f^P_{(\mathcal{S}_p,\circ\circ)\leftarrow(\mathcal{A},\circ\circ)} = f^P_{(\mathcal{S}_p,\circ\circ)\leftarrow(\mathcal{A},\otimes\circ)} = f^P_{(\mathcal{S}_p,\circ\otimes)\leftarrow(\mathcal{A},\circ\circ)} = f^P_{(\mathcal{S}_p,\circ\otimes)\leftarrow(\mathcal{A},\otimes\circ)} = \frac{(1-z_\bullet)d_\bullet}{(1-z_0)d_0}, \tag{A.55}$$

and

$$f^D_{(\mathcal{S}_d,\cdot)\leftarrow(\mathcal{A},\circ\circ)} = f^D_{(\mathcal{S}_d,\cdot)\leftarrow(\mathcal{A},\otimes\circ)} = \frac{z_\bullet\delta_\bullet}{(1-e)z\delta}. \tag{A.56}$$

A.6 Recurrence equations for identity probabilities

We note $Q_{X/Y} = Q_{Y/X}$ the probability of identity by descent (IBD) between one gene in class X and one gene in class Y, both at generation t. These probabilities are evaluated for pairs of genes in the same deme, just after reproduction, and depend upon IBD probabilities for pairs of genes sampled after dispersal, noted $Q'_{X,Y}$. IBD probabilities for genes sampled in individuals from the same generation obey:

$$Q_{X,Y} = Q'_{X,Y} \tag{A.57}$$

except for

$$Q_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)} = \left[\frac{1}{N} + \left(1 - \frac{1}{N}\right) Q'_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)}\right], \tag{A.58}$$

and

$$Q_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)} = \left[\frac{1}{N} + \left(1 - \frac{1}{N}\right) Q'_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)}\right]. \tag{A.59}$$

The recurrence equations for the IBD probabilities are given below.

A.6.1 Identity probabilities within generations

Since we consider an infinite island model of dispersal, all the IBD probabilities among genes from different demes cancel out. Also, IBD probabilities between one gene sampled from a dispersed seed and any other gene are all nil. The IBD probability between two genes sampled among individuals in class i and j in a deme of category n is given by

$$Q'_{(i,n)/(j,n)}(t+1) = \sum_m v(m|n) \sum_k \sum_l f^P_{(i,n)\leftarrow(k,m)} f^P_{(j,n)\leftarrow(l,m)} Q'_{(k,m)/(l,m)}(t). \tag{A.60}$$

The fitness functions $f^P_{(i,n)\leftarrow(k,m)}$ and $f^P_{(j,n)\leftarrow(l,m)}$ are evaluated in the neutral case, where all individuals adopt the same set of strategies. Equation (A.60) sums over the backward probabilities that the ancestral category of the deme was m. Then the probabilities that the gene lineages in (i,n) and (j,n) have ancestors of types k and l in one deme in category m are weighted by the IBD probability $Q_{(k,m)/(l,m)}$ of the ancestors. Equation (A.60) develops as:

$$\begin{aligned} Q'_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)}(t+1) &= v(1|1)\left[\left(f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)}\right)^2 Q'_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)}(t)\right. \\ &+ 2 f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)} f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_p,\circ\circ)} Q'_{(\mathcal{A},\circ\circ)/(\mathcal{S}_p,\circ\circ)}(t) \\ &+ \left.\left(f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_p,\circ\circ)}\right)^2 Q'_{(\mathcal{S}_p,\circ\circ)/(\mathcal{S}_p,\circ\circ)}(t)\right] \\ &+ v(2|1)\left(f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\otimes\circ)}\right)^2 Q'_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)}(t), \end{aligned} \tag{A.61}$$

$$
\begin{aligned}
Q'_{(\mathcal{A},\circ\circ)/(\mathcal{S}_p,\circ\circ)}(t+1) &= v(1|1)\Big[f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\circ\circ)} f^P_{(\mathcal{S}_p,\circ\circ)\leftarrow(\mathcal{A},\circ\circ)} Q'_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)}(t) \\
&+ f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{S}_p,\circ\circ)} f^P_{(\mathcal{S}_p,\circ\circ)\leftarrow(\mathcal{A},\circ\circ)} Q'_{(\mathcal{S}_p,\circ\circ)/(\mathcal{A},\circ\circ)}(t)\Big] \\
&+ v(2|1) f^P_{(\mathcal{A},\circ\circ)\leftarrow(\mathcal{A},\otimes\circ)} f^P_{(\mathcal{S}_p,\circ\circ)\leftarrow(\mathcal{A},\otimes\circ)} Q'_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)}(t),
\end{aligned} \tag{A.62}
$$

$$
\begin{aligned}
Q'_{(\mathcal{S}_p,\circ\circ)/(\mathcal{S}_p,\circ\circ)}(t+1) &= v(1|1)\left(f^P_{(\mathcal{S}_p,\circ\circ)\leftarrow(\mathcal{A},\circ\circ)}\right)^2 Q'_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)}(t) \\
&+ v(2|1)\left(f^P_{(\mathcal{S}_p,\circ\circ)\leftarrow(\mathcal{A},\otimes\circ)}\right)^2 Q'_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)}(t),
\end{aligned} \tag{A.63}
$$

$$
Q'_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)}(t+1) = v(3|2)\left(f^P_{(\mathcal{A},\otimes\circ)\leftarrow(\mathcal{S}_p,\circ\otimes)}\right)^2 Q'_{(\mathcal{S}_p,\circ\otimes)/(\mathcal{S}_p,\circ\otimes)}(t), \tag{A.64}
$$

and

$$
\begin{aligned}
Q'_{(\mathcal{S}_p,\circ\otimes)/(\mathcal{S}_p,\circ\otimes)}(t+1) &= v(1|3)\left(f^P_{(\mathcal{S}_p,\circ\otimes)\leftarrow(\mathcal{A},\circ\circ)}\right)^2 Q'_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)}(t) \\
&+ v(2|3)\left(f^P_{(\mathcal{S}_p,\circ\otimes)\leftarrow(\mathcal{A},\otimes\circ)}\right)^2 Q'_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)}(t).
\end{aligned} \tag{A.65}
$$

The relevant probabilities concern the identity-by-descent between a focal in class (j,l) and an adult actor in its deme. We use the short-hand notation $Q^0_{(j,l)} \equiv Q'_{(\mathcal{A},l)/(j,l)}$ for these IBD probabilities. Hence, $Q^0_{(\mathcal{A},\circ\circ)} \equiv Q'_{(\mathcal{A},\circ\circ)/(\mathcal{A},\circ\circ)}$, $Q^0_{(\mathcal{S}_p,\circ\circ)} \equiv Q'_{(\mathcal{A},\circ\circ)/(\mathcal{S}_p,\circ\circ)}$ and $Q^0_{(\mathcal{A},\otimes\circ)} \equiv Q'_{(\mathcal{A},\otimes\circ)/(\mathcal{A},\otimes\circ)}$.

A.6.2 Identity probabilities between generations

We note Q^Y_X the IBD probability between one gene in class X at t and one gene in class Y at $(t-1)$. Generally, the IBD probabilities (after dispersal) between genes among individuals at t can be expressed as the sum of IBD probabilities between genes from one individual at t and another individual at $(t-1)$, weighted by the probabilities of origin of that latter individual. For example, the IBD probability between genes in a type-$\mathcal{S}_p$ individual (individual A) and in

a type-i individual (individual B), both in a deme in category n is given by the relationship:

$$\begin{aligned} Q_{(\mathcal{S}_p,n)/(i,n)} &= \sum_m \Pr(\text{A's ancestor in a } m \text{ deme} \mid \text{A in a } n \text{ deme}) \\ &\times \Pr(\text{A has been produced in the deme} \mid \text{A's ancestor in } m) \\ &\times \Pr(\text{A's ancestor and B are IBD}), \end{aligned} \tag{A.66}$$

which gives

$$Q_{(\mathcal{S}_p,n)/(i,n)} = \sum_m v(m|n) f^P_{(\mathcal{S}_p,n)\leftarrow(\mathcal{A},m)} Q^{(\mathcal{A},m)}_{(i,n)}. \tag{A.67}$$

From this expression, and since $f^P_{(\mathcal{S}_p,n)\leftarrow(\mathcal{A},m)} = 1$ at neutrality (see equation [A.55]),we get:

$$Q_{(\mathcal{S}_p,\bigcirc\bigcirc)/(\mathcal{A},\bigcirc\bigcirc)} = v(1|1) Q^{(\mathcal{A},\bigcirc\bigcirc)}_{(\mathcal{A},\bigcirc\bigcirc)} + v(2|1) Q^{(\mathcal{A},\otimes\bigcirc)}_{(\mathcal{A},\bigcirc\bigcirc)} \equiv Q^1_{(\mathcal{A},\bigcirc\bigcirc)}, \tag{A.68}$$

$$Q_{(\mathcal{S}_p,\bigcirc\bigcirc)/(\mathcal{S}_p,\bigcirc\bigcirc)} = v(1|1) Q^{(\mathcal{A},\bigcirc\bigcirc)}_{(\mathcal{S}_p,\bigcirc\bigcirc)} + v(2|1) Q^{(\mathcal{A},\otimes\bigcirc)}_{(\mathcal{S}_p,\bigcirc\bigcirc)} \equiv Q^1_{(\mathcal{S}_p,\bigcirc\bigcirc)}, \tag{A.69}$$

and

$$Q_{(\mathcal{S}_p,\bigcirc\otimes)/(\mathcal{S}_p,\bigcirc\otimes)} = v(1|3) Q^{(\mathcal{A},\bigcirc\bigcirc)}_{(\mathcal{S}_p,\bigcirc\otimes)} + v(2|3) Q^{(\mathcal{A},\otimes\bigcirc)}_{(\mathcal{S}_p,\bigcirc\otimes)} \equiv Q^1_{(\mathcal{S}_p,\bigcirc\otimes)}. \tag{A.70}$$

Here, $Q^1_{(j,l)}$ has been defined as the IBD probability between a focal's gene in class (j,l) at t and an adult actor's gene at $t-1$ in its deme.

A.7 Stochastic simulations

At the beginning of the life cycle, each individual produces a random number of offspring, drawn from a Poisson distribution with mean $r = 100$. Mutation occurs at rate $\mu = 0.001$ for each trait, and the mutation effect is randomly drawn from a normal distribution with zero

mean and standard deviation s.d. = 0.05. Mutations giving rise to trait values outside the [0,1] interval are discarded. The fate of each individual depends upon its phenotype that determines its probability to disperse, to enter a dormant stage, to die during dispersal or in the seed bank, etc. Competition occurs among all offspring in each population, and a number N of individuals are randomly drawn to form the next generation. If the number of offspring is less than N, then all individuals survive to adulthood. At low fecundity, saturation may not be attained in each deme, and some populations may therefore go extinct because of demographic stochasticity. We considered a finite, yet large, number of populations: $n_{\mathrm{d}} = 500$.

For each set of parameter values, we ran a single simulation for 200,000 generations. We used batch means to compute Monte Carlo standard errors (Hastings 1970). The rationale is to split the Markov chain into a number of batches, which lengths are chosen so that successive batch means are practically uncorrelated, and then to calculate the variance among batches. Here, we discarded the first 40,000 generations, and we computed the batch mean estimate of Monte Carlo variance as: $\sigma^2 = \frac{b}{a-1} \sum_{k=1}^{a} (Y_k - \mu)^2$, where $a = 20$ is the number of batches of size $b = 8,000$, Y_k is the Monte Carlo estimate of the mean of the kth batch, and μ the overall mean. Standard errors were then estimated as: $s.e. = \sigma/\sqrt{n}$, where $n = 160,000$ is the total number of iterations. For each graph, error bars were computed as $\pm 1.96\sigma/\sqrt{n}$.

Literature Cited

Ajar, E. 2003. Analysis of disruptive selection in subdivided populations. BMC Evolutionary Biology 3:22.

Balaban, N. Q., J. Merrin, R. Chait, L. Kowalik, and S. Leibler. 2004. Analysis of disruptive selection in subdivided populations. Science 305:1622–1625.

Brandel, M. 2004. Dormancy and germination of heteromorphic achenes of *Bidens frondosa*. Flora 199:228–233.

Bulmer, M. 1984. Delayed germination of seeds: Cohen's model revisited. Theoretical Population Biology 26:367–377.

Busch, J. 2006. Heterosis in an isolated, effectively small, and self-fertilizing population of the flowering plant *Leavenworthia alabamica*. Evolution 60:184–191.

Charlesworth, B. 1994. Evolution in age-structured populations. 2nd edition. Cambridge University Press, Cambridge.

Chesson, P. L., and R. R. Warner. 1981. Environmental variability promotes coexistence in lottery competitive systems. American Naturalist 117:923–943.

Cohen, D. 1966. Optimizing reproduction in a randomly varying environment. Journal of Theoretical Biology 12:119–129.

Cohen, D., and S. A. Levin. 1987. he interaction between dispersal and dormancy strategies in varying and heterogeneous environments. *in* E. Teramato, and M. Yomaguti, eds. Mathematical topics in population biology, morphogenesis and neurosciences Pages 110–122. Lecture notes. Biomathematics, Kyoto.

—. 1991. Dispersal in patchy environments: the effects of temporal and spatial structure. Theoretical Population Biology 39:36–99.

Crow, J. F., and M. Kimura. 1970. An introduction to Population Genetics Theory. Burgess Publishing Company, Minneapolis.

Denno, R. F., G. K. Roderick, K. L. Olmstead, and H. G. Döbel. 1991. Density-related migration in planthoppers (homoptera: Delphacidae): the role of habitat persistence. American Naturalist 138:1513–1541.

Ebert, D., C. Haag, M. Kirkpatrick, M. Riek, J. W. Hottinger, and V. I. Pajunen. 2002. A selective advantage to immigrant genes in a *Daphnia* metapopulation. Science 295:485–488.

Ellner, S. 1985*a*. Ess germination strategies in randomly varying environments. i. logistic-type models. Theoretical Population Biology 28:50–79.

—. 1985*b*. Ess germination strategies in randomly varying environments. ii. reciprocal yield-law models. Theoretical Population Biology 28:80–116.

—. 1986. Germination dimorphisms and parent offspring conflicts in seed-germination. Journal of Theoretical Biology 123:173–185.

Eshel, I. 1996. On the changing concept of evolutionary population stability as a reflection of a changing point of view in the quantitative theory of evolution. Journal of Mathematical Biology 34:485–510.

Evans, M., and J. Dennehy. 2005. Germ banking: bet-hedging and variable release from egg and seed dormancy. The Quarterly Review of Biology 80:431–451.

Frank, S. 1998. Foundations of social evolution. Princeton University Press, Princeton.

Frank, S. A. 1986. Dispersal polymorphism in subdivided populations. Journal of Theoretical Biology 122:303–309.

Gandon, S. 1999. Kin competition, the cost of inbreeding and the evolution of dispersal. Journal of Theoretical Biology 200:345–364.

Gandon, S., and F. Rousset. 1999. Evolution of stepping-stone dispersal rates. Proceedings of the Royal Society of London B, Biological Sciences 266:2507–2513.

Gefen, O., and N. Q. Balaban. 2009. The importance of being persistent: heterogeneity of bacterial populations under antibiotic stress. FEMS Microbiology Reviews 33:704–717.

Geritz, S. A. H., É. Kidsi, G. Meszéna, and J. A. J. Metz. 1998. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. Evolutionary Ecology 12:35–57.

Hamilton, W. D. 1964. The genetical evolution of social behavior. I. Journal of Theoretical Biology 7:1–16.

—. 1966. The moulding of senescence by natural selection. Journal of Theoretical Biology 12:12–45.

Hamilton, W. D., and R. May. 1977. Dispersal in stable habitats. Nature 269:578–581.

Hastings, W. K. 1970. Monte carlo sampling methods using markov chains and their applications. Biometrika 57:97–109.

Holmes, P. M., and R. J. Newton. 2004. Patterns of seed persistence in south african fynbos. Plant Ecology 172:143–158.

Imbert, E. 2002. Ecological consequences and ontogeny of seed heteromorphism. Perspectives in Plant Ecology, Evolution and Systematics 5:13–36.

Klinkhamer, P., T. de Jong, J. A. Metz, and J. Val. 1987. Life history tactics of annual organisms: the joint effect of dispersal and delayed germination. Theoretical Population Biology 32:127–156.

Kobayashi, Y., and N. Yamamura. 2000. Evolution of seed dormancy due to sib competition: effect of dispersal and inbreeding. Journal of Theoretical Biology 202:11–24.

Kussell, E. L., R. Kishony, N. Q. Balaban, and S. Leibler. 2005. Bacterial persistence: a model of survival in changing environments. Genetics 169:1807–1814.

Lehmann, L. 2007. The evolution of trans-generational altruism: kin selection meets niche construction. Journal of Evolutionary Biology 20:181–189.

Lehmann, L., N. Perrin, and F. Rousset. 2006. Population demography and the evolution of helping behaviors. Evolution 60:1137–1151.

Leimar, O. 2009. Multidimensional convergence stability. Evolutionary Ecology Research 11:191–208.

Leturque, H., and F. Rousset. 2002. Dispersal, kin competition, and the ideal free distribution on a spatially heterogeneous population. Theoretical Population Biology 62:169–180.

—. 2004. Intersexual competition as an explanation for sex-ratio and dispersal biases in polygynous species. Evolution 58:2398–2408.

Levin, S. A., D. Cohen, and A. Hastings. 1984. Dispersal strategies in patchy environments. Theoretical Population Biology 26:165–191.

Lewis, K. 2007. Persister cells, dormancy and infectious disease. Nature Reviews Microbiology 5:48–56.

McPeek, M., and S. Kalisz. 1998. The joint evolution of dispersal and dormancy in metapopulations. Archive für Hydrobiologie 52:33–51.

Metz, J., T. J. de Jong, and P. G. L. Klinkhamer. 1983. What are the advantages of dispersing; a paper by kuno explained and extended. Oecologia 57:166–169.

Morgan, M. T. 2002. Genome-wide deleterious mutation favors dispersal and species integrity. Heredity 89:253–257.

Nakajima, T., and Y. Kurihara. 1994. Evolutionary changes of dispersiveness in experimental bacterial populations. Oikos 69:217–223.

Olivieri, I. 2001. The evolution of seed heteromorphism in a metapopulation: interactions between dispersal and dormancy. Pages 245–268 *in* J. Silvertown, and J. Antonovics, eds. Integrating Ecology and Evolution in a Spatial Context. Blackwell Science, Oxford.

Olivieri, I., and A. Berger. 1985. Seed dimorphism for dispersal: physiological, genetic and demographical aspects. Pages 413–429 *in* P. Jacquard, G. Heim, and J. Antonovics, eds. In:Genetic Differentiation and Dispersal in Plants. Springer-Verlag, Berlin.

Olivieri, I., M. Swan, and P.-H. Gouyon. 1983. Reproductive system and colonizing strategy of two species of *Carduus* (compositae). Oecologia 60:114–117.

Paland, S., and B. Schmid. 2003. Population size and the nature of genetic load in *Gentianella germanica*. Evolution 57:2242–2251.

Perrin, N., and V. Mazalov. 1999. Dispersal and inbreeding avoidance. American Naturalist 154:282–292.

Philippi, T., and J. Seger. 1989. Hedging one's evolutionary bets, revisited. Trends in Ecology and Evolution 4:41–44.

Richards, C. 2000. Inbreeding depression and genetic rescue in a plant metapopulation. American Naturalist 155:383–394.

Ronce, O. 2007. How does it feel to be like a rolling stone? ten questions about dispersal evolution. Annual Review of Ecology, Evolution, and Systematics 38:231–253.

Ronce, O., F. Perret, and I. Olivieri. 2000. Landscape dynamics and evolution of colonizer syndromes: interactions between reproductive effort and dispersal in a metapopulation. Evolutionary Ecology 14:233–260.

Rotem, E., A. Loinger, I. Ronin, I. Levin-Reisman, C. Gabay, N. Shoresh, O. Biham, and N. Q. Balaban. 2010. Regulation of phenotypic variability by a threshold-based mechanism underlies bacterial persistence. Proceedings of the National Academy of Sciences of the USA 107:12541–12546.

Rousset, F. 1999. Genetic differentiation in populations with different class of individuals. Theoretical Population Biology 55:297–308.

—. 2002. Inbreeding and relatedness coefficients: what do they measure? Heredity 88:371–380.

—. 2004. Genetic structure and selection in subdivided populations. Princeton University Press, Princeton.

Rousset, F., and S. Billiard. 2000. A theoretical basis for measures of kin selection in subdivided populations: finite populations and localized dispersal. Journal of Evolutionary Biology 13:814–825.

Rousset, F., and O. Ronce. 2004. Measuring selection on traits affecting metapopulation demography. Theoretical Population Biology 65:127–141.

Roze, D., and F. Rousset. 2005. Inbreeding depression and the evolution of dispersal rates: a multilocus model. American Naturalist 166:708–721.

—. 2008. Multilocus models in the infinite island model of population structure. Theoretical Population Biology 73:529–542.

—. 2009. Strong effects of heterosis on the evolution of dispersal rates. Journal of Evolutionary Biology 22:1221–1233.

Satterthwaite, W. H. 2010. Competition for space can drive the evolution of dormancy in a temporally invariant environment. Plant Ecology 208:167–185.

Schurr, F. M., G. F. Midgley, A. G. Rebelo, G. Reeves, P. Poschlod, and H. S. I. 2007. Colonization and persistence ability explain the extent to which plant species fill their potential range. Global Ecology and Biogeography 16:449–459.

Slatkin, M. 1974. =hedging one's evolutionary bets. Nature 250:704–705.

Snyder, R. E. 2006. Multiple risk reduction mechanisms: can dormancy substitute for dispersal? Ecology Letters 9:1106–1114.

Taylor, P. D. 1988. An inclusive fitness model for dispersal of offspring. Journal of Theoretical Biology 130:363–378.

—. 1990. Allele-frequency change in class-structured populations. American Naturalist 135:95–106.

Taylor, P. D., and S. A. Frank. 1996. How to make a kin selection model. Journal of Theoretical Biology 180:27–37.

Taylor, T. B., and A. Buckling. 2010. Competition and dispersal in *Pseudomonas aeruginosa*. American Naturalist 176:83–89.

Tielbörger, K., and R. Prasse. 2009. Do seeds sense each other? testing for density-dependent germination in desert perennial plants. Oikos 118:792–800.

Tielbörger, K., and A. Valleriani. 2005. Can seeds predict their future? germination strategies of density-regulated desert annuals. Oikos 111:235–244.

Tsuji, N., and N. Yamamura. 1992. A simple evolutionary model of dormancy and dispersal in heterogeneous patches with special difference to phytophagous lady beetles. I. Stable environments. Researches on Population Ecology 34:77–90.

Venable, D. L. 1985. Ecology of achene dimorphism in *Heterotheca latifolia*. III. consequences of varied water availability. Journal of Ecology 73:757–763.

—. 2007. Bet hedging in a guild of desert annnuals. Ecology 88:1086–1090.

Venable, D. L., and J. S. Brown. 1988. The selective interactions of dispersal, dormancy, and seed size as adaptations for reducing risk in variable environment. American Naturalist 131:360–384.

Venable, D. L., and L. Lawlor. 1980. Delayed germination and dispersal in desert annuals: escape in space and time. Oecologia 46:272–282.

Venable, D. L., C. E. Pake, and A. C. Caprio. 1993. Diversity and coexistence of Sonoran desert winter annuals. Plant Species Biology 8:207–216.

Waser, P. M., S. N. Austad, and B. Keane. 1986. When should animals tolerate inbreeding? American Naturalist 128:529–537.

Wiener, P., and S. Tuljapurkar. 1994. Migration in variable environments: exploring life-history evolution using structured poplation models. Journal of Theoretical Biology 166:75–90.

Table 1: Summary of main parameter notations.

Notation	Parameter definition
z	Dispersal rate
D	Rate of unconditionnal dormancy
d	Rate of conditionnal dormancy for philopatric seeds
δ	Rate of conditionnal dormancy for dispersed seeds
c_{z}	Cost of dispersal
c_{d}	Cost of dormancy
e	Rate of extinction
r	Fecundity
N	Number of adults in each deme
ϵ	Effect of mutation
Q	Probability of genetic identity
$w_{(i,k)\leftarrow(j,l)}$	Expected number of offspring in class (i,k) produced by a focal individual in class (j,l)
$f_{(i,k)\leftarrow(j,l)}$	Probability that a gene in class (i,k) is a copy of a gene from any of the A parent in class (j,l)
$\mathbf{U}$	Forward matrix transition for deme categories, with (i,j)th element $u(i\|j)$
$\mathbf{V}$	Backward matrix transition for deme categories, with (i,j)th element $v(j\|i)$
$\mathbf{P}$	Stationary distribution of deme categories, with ith element $P(i)$
$\mathbf{F}$	Backward transition matrix of gene lineages between classes, with (ik,jl)th element $(f_{(i,k)\leftarrow(j,l)})$
$\boldsymbol{\alpha}$	Vector of class reproductive values (dominant left eigenvector of $\mathbf{F}$), with (i,k)th element $\alpha(ik)$
J	Number of juveniles issued from non-dormant seeds
G	Number of juveniles issued from dormant seeds

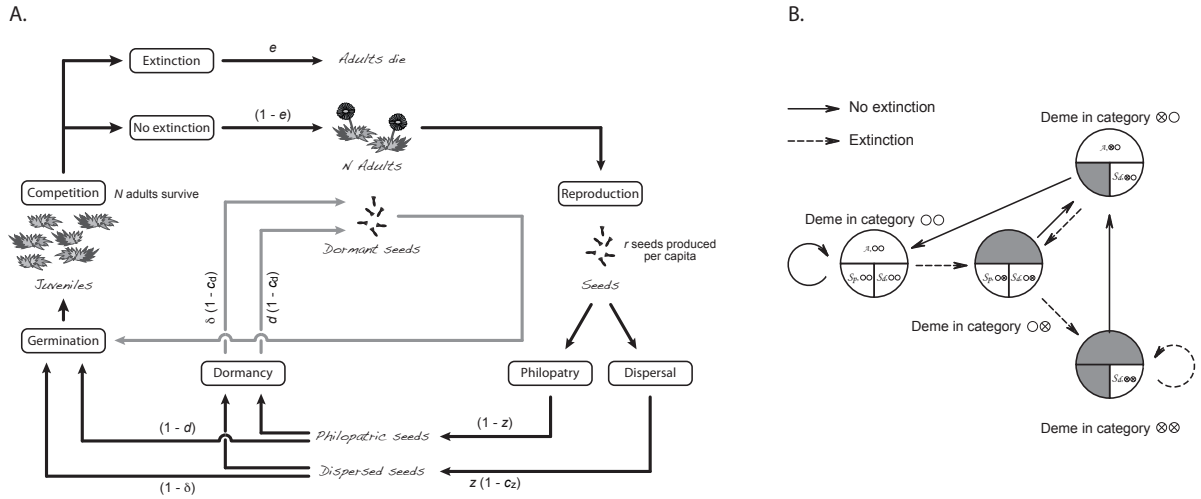


Figure 1: (A) Life cycle. (B) Definition of the demographic classes. There are four distinct categories of demes, depending on the history of extinctions. For each category of deme (circles, identified by bold numbers), the individual types are represented: type-$\mathcal{A}$ individuals are adults (top), type-$\mathcal{S}_p$ individuals are philoparic seeds (bottom left), and type-$\mathcal{S}_d$ individuals are dispersed seeds (bottom right). Non-existing types of individuals (e.g. adults in extinct demes) are figured in grey. We index each category as (i, k), for type-i individuals in a deme of category k. The transitions between deme categories are represented with arrows (see legend). For example, demes are in category ○○ at $(t + 1)$, if and only if they were in category ○○ or ⊗○ at t, and if no extinction occurred. Were the demes in category ○⊗ or ⊗⊗ at t (i.e. demes with no adults, following an extinction event), philopatric dormant seeds could not have been produced, resulting in an empty class of philopatric seeds at $t + 1$ (individual class 2), which is incompatible with the definition of category ○○ demes.

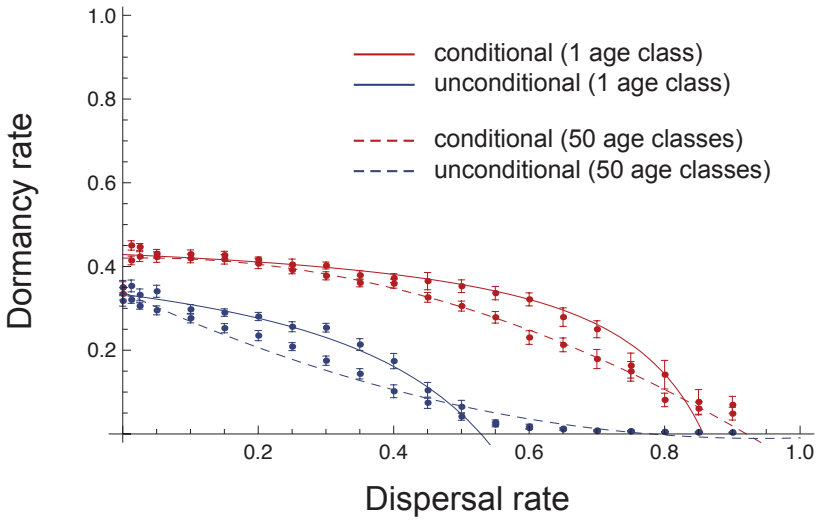


Figure 2: Evolutionary stable rate of dormancy as a function of the (fixed) dispersal rate z, with $N = 1$, $c_z = 0.5$, $c_d = 0.2$, and $e = 0$ (no extinction). Both the rate of unconditional dormancy (D^*, plain blue line) and the rate of conditional dormancy for philopatric seeds (d^*, plain red line) are shown. In the latter case, $\delta^* = 0$. The dashed lines provide the simulation results for 50 age classes in the seed bank.

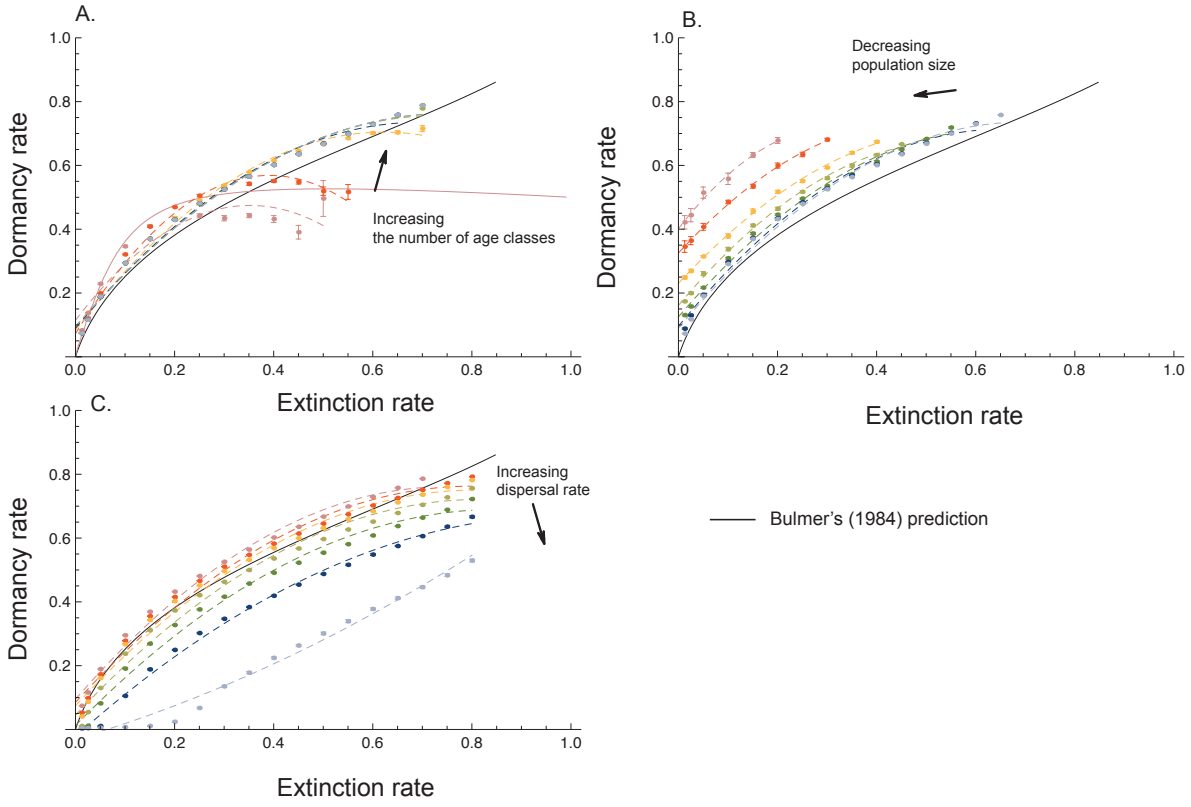


Figure 3: Evolutionary stable rate of conditional dormancy of philopatric seeds when dispersal is a fixed parameter, as a function of the extinction rate for a various number of age classes in the seed bank (varying from 1 to 100). A large population size (N = 100) and a high fecundity (r = 100) are considered. The dispersal rate was fixed at a very low value, so that $N\eta = 0.0001$. Other parameter values are c_{d} = 0.2 and c_{z} = 0.5. (B) Evolutionary stable rate of conditional dormancy of philopatric seeds when dispersal is a fixed parameter, as a function of the extinction rate for population size varying from 1 to 100 and 50 age classes in the bank. Other parameter values are as in (A). (C) Evolutionary stable rate of conditional dormancy of philopatric seeds when dispersal is a fixed parameter, as a function of the extinction rate for a number of migrants per generation varying from 0.01 to 5. Other parameter values are as in (A). The black plain line indicates the solution from Bulmer's (1984) prediction (see his equation 3). Note that, since fecundity is limited in the simulations (here, r = 100), the metapopulation as a whole may not be viable for small population sizes and high extinction rates. The metapopulation may therefore go extinct because of demographic stochasticity, for some sets of parameter values. This explains why the curves in (B) were only obtain for small extinction rates at low population size.

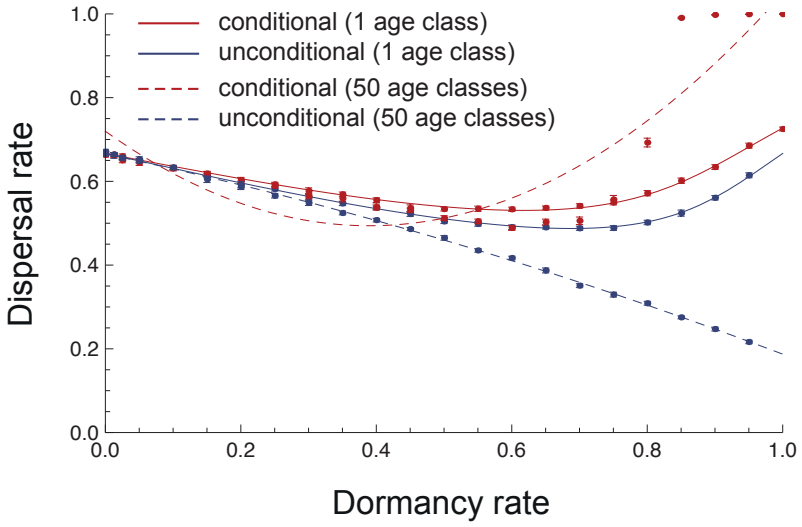


Figure 4: Evolutionary stable dispersal rate z^* as a function of the (fixed) rate of dormancy, with $N = 1$, $c_z = 0.5$, $c_d = 0.2$, and $e = 0$ (no extinction). The ES rate of dispersal is shown in the case of conditional dormancy (plain red line) and unconditional dormancy (plain blue line). The dashed lines provide the simulation results for 50 age classes in the seed bank.

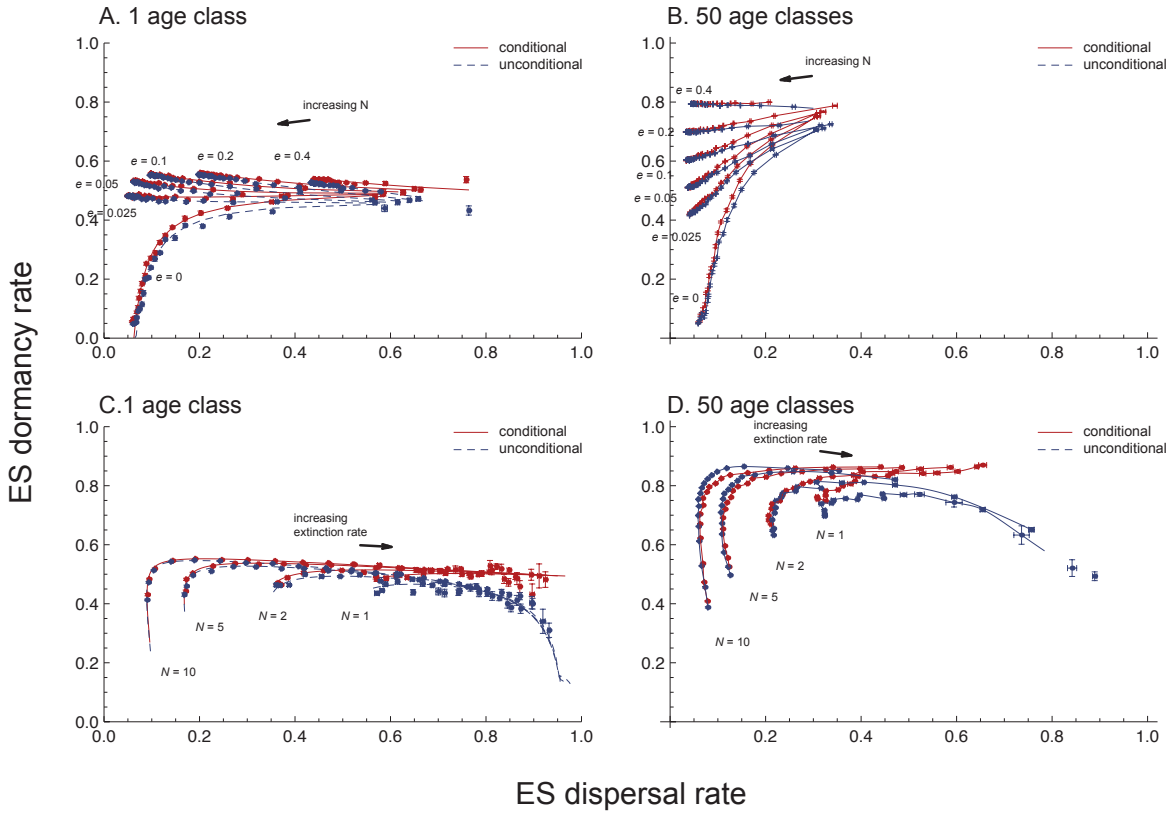
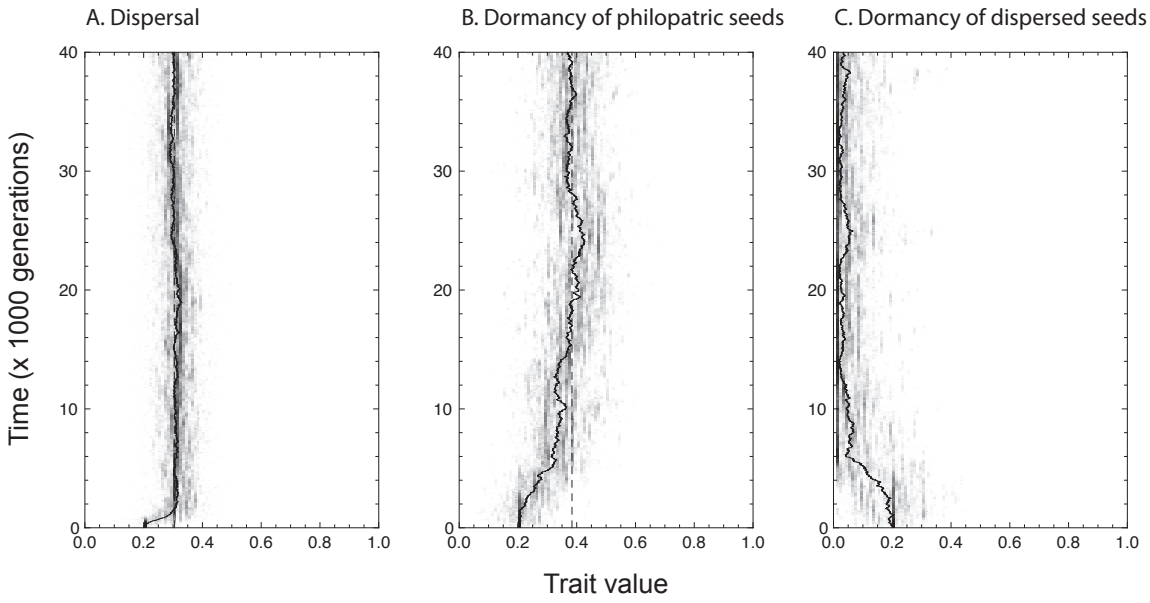
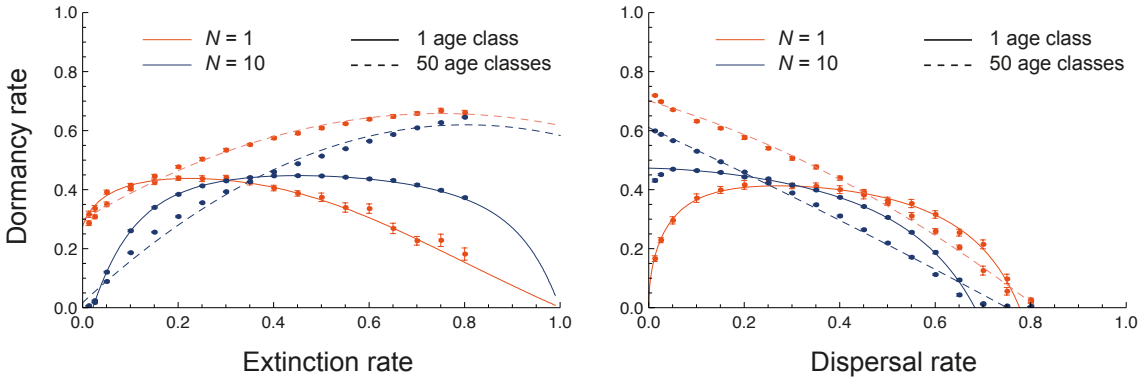


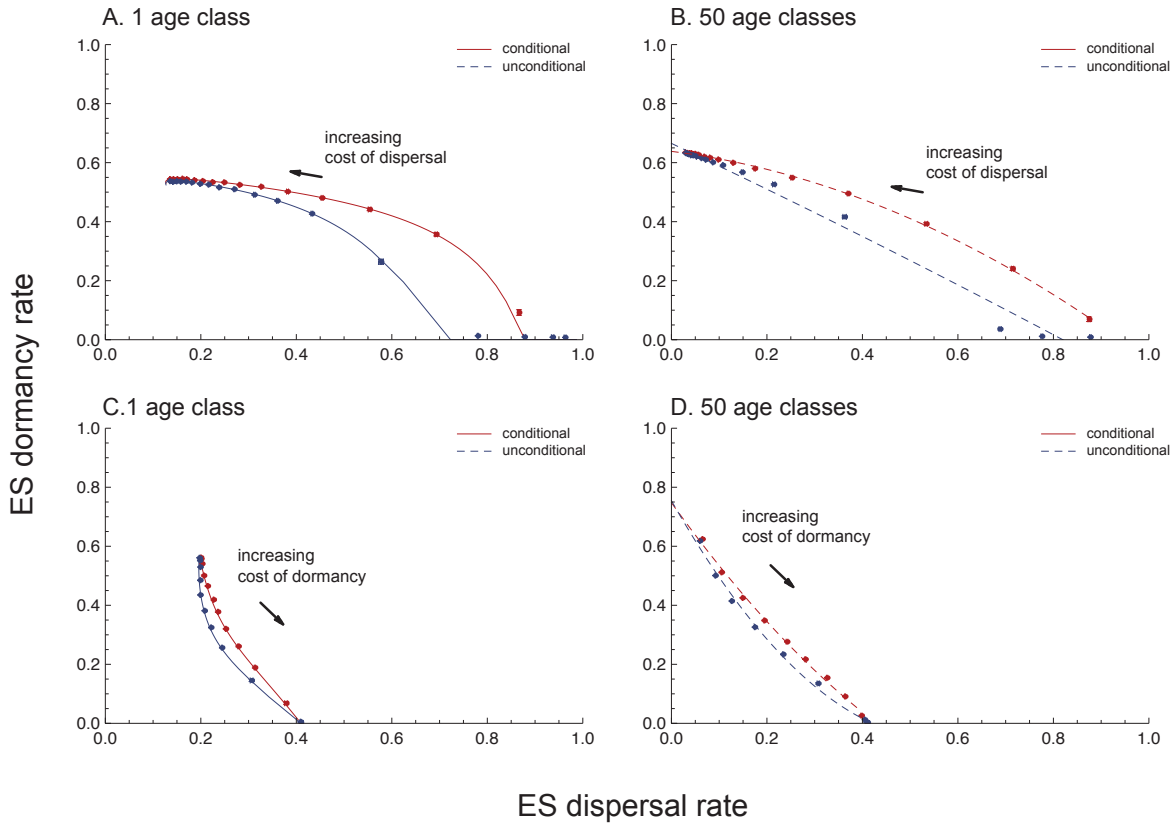
Figure 5: Joint evolutionary stable rates of dispersal and dormancy. The plain lines provide the results for the model with conditional dormancy, and the dashed line that with unconditional dormancy. (A) As a function of the number of adults (N), which varies from 1 to 20. The arrow indicates the direction of increasing N. Other parameter values are: $c_{\mathrm{d}} = 0.025$, $c_{\mathrm{z}} = 0.4$, and e varies from 0 to 0.4. (B) Idem with 50 age classes in the seed bank. (C) As a function of the rate of extinction (e), which varies from 0.005 to 0.995. The arrow indicates the direction of increasing e. Other parameter values are: $c_{\mathrm{d}} = 0.025$, $c_{\mathrm{z}} = 0.4$, and N varies from 1 to 10. (D) Idem with 50 age classes in the seed bank.



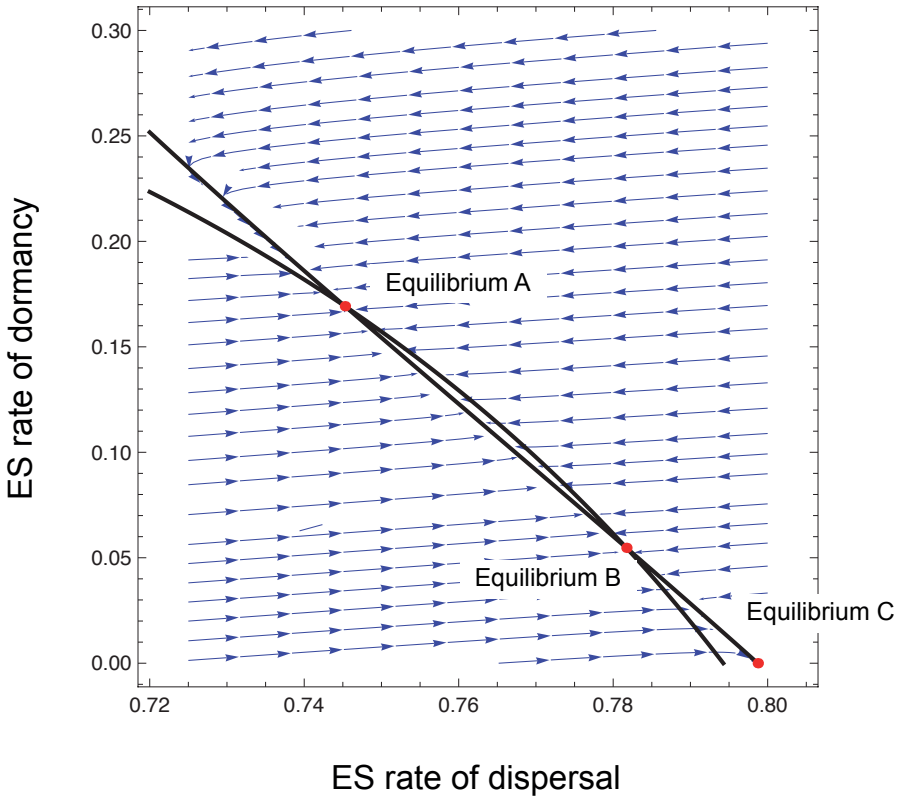
Supplementary Figure S1: Evolutionary dynamics of the traits in a large metapopulation with n_d = 2,000 demes, each of size N = 5. This figure results from a single run of an individual-based simulation model, where each individual is characterized by a set of random variables representing its genotype for each phenotypic trait. The same life cycle as in the analytical model was considered. Each individual produces a random number of offspring, drawn from a Poisson with mean r = 100. Mutation occurs at rate μ = 0.001 for each trait, and the mutation effect is randomly drawn from a Normal distribution with zero mean and standard deviation SD = 0.05. Other parameter values are: c_z = 0.2, c_d = 0.025, e = 0 (no extinction). The metapopulation was initially monomorphic, with all trait values fixed to 0.2.The dashed line gives the evolutionary stable trait value. The first 25,000 generations are shown. The rate of dormancy for philopatric seeds converge more slowly towards the equilibrium, as compared to the rate of dispersal. This suggests that the selection gradient is weaker for the rate of dormancy for philopatric seeds than for the rate of dispersal.



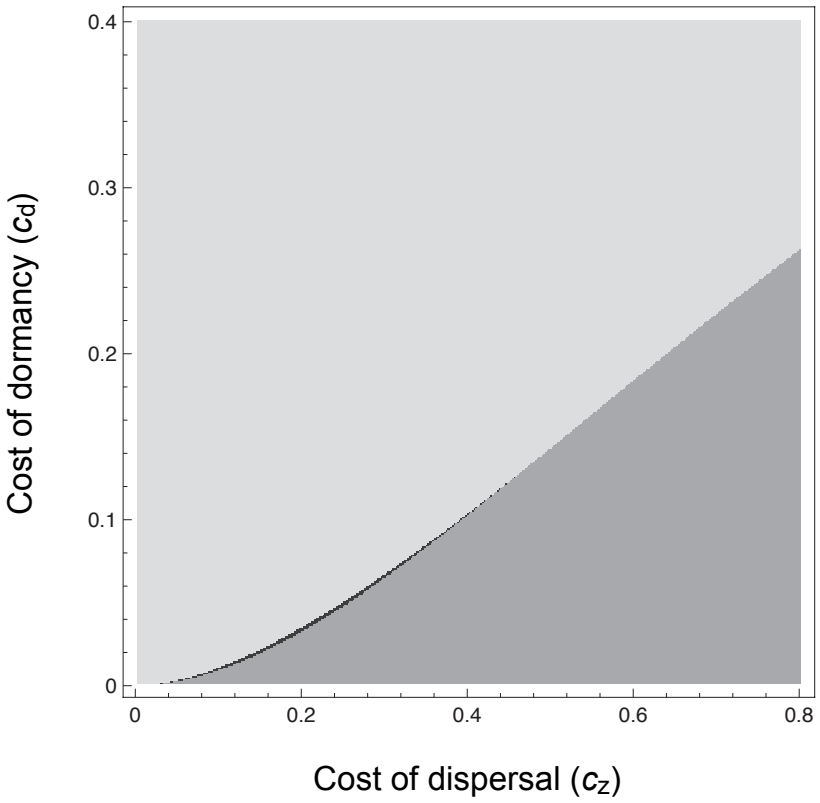
Supplementary Figure S2: (A) Evolutionary stable rate of unconditional dormancy as a function of the extinction rate for different population sizes (N = 1, N = 5, and N = 10), with c_z = 0.5, c_d = 0.2, and z = 0.2. (B) Evolutionary stable rate of unconditional dormancy as a function of the (fixed) dispersal rate for different population sizes (N = 1, N = 5, and N = 10), with c_z = 0.5, c_d = 0.2, and e = 0.4.



Supplementary Figure S3: (A) Joint evolutionary stable rates of dispersal and dormancy, as a function of the cost of dispersal (c_z), which varies from 0.0125 to 0.8, for a single age class in the bank, with local extinctions ($e = 0.2$), $N = 10$ and $c_d = 0.05$. The plain lines give the numerical solutions from the analytical model (equation 2) for unconditional dormancy (D^*, in blue) and conditional dormancy for philopatric seeds (d^*, in red). (B) As in (A) for 50 age classes in the bank. The dots and error bars give the mean values of the trait from individual-based simulations. (C) Joint evolutionary stable rates of dispersal and dormancy, as a function of the cost of dormancy (c_d), which varies from 0.0125 to 0.45 for a single age class in the bank, with local extinctions ($e = 0.2$), $N = 10$ and $c_z = 0.5$. The plain lines give the numerical solutions from the analytical model (equation 2) for unconditional dormancy (D^*, in blue) and conditional dormancy for philopatric seeds (d^*, in red). (D) As in (C) for 50 age classes in the bank.



Supplementary Figure S4: An example of bistable evolutionary dynamics for the joint evolution of dispersal and unconditional dormancy, with $N = 1$, $c_d = 0.05$, $c_z = 0.252$ and $e = 0$ (no extinction). In this gradient plot, the arrows show the direction of selection acting on dispersal and dormancy. As can be seen from the plot, two out of the three joint equilibria are stable (equilibria A and C), while equilibrium B is unstable, indicating that the evolutionary endpoint may depend upon initial conditions.



Supplementary Figure S5: Region plot of parameter space, where evolutionary bistable rates of unconditional dormancy occur (black area), with $N = 1$ and $e = 0$ (no extinction). Light grey: the ES rate of dormancy is nil. Dark grey: a single joint strategy for dispersal and dormancy exists.