Inférences en Génétique des Populations

Simon Boitard, Raphaël Leblois, Miguel de Navascués

Inférences en Génétique des Populations

Simon Boitard, Raphaël Leblois, Miguel de Navascués

Centre de Biologie pour la Gestion des Populations (CBGP), INRAE

Module de génétique des populations, Master 1 BEE, Univ Montpellier

décembre 2024

Plan

Plan

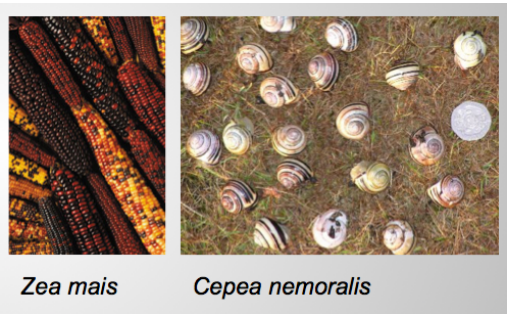
Génétique des populations

La génétique des populations vise à analyser les **processus** qui contrôlent le **polymorphisme génétique** (= variabilité) dans les populations.

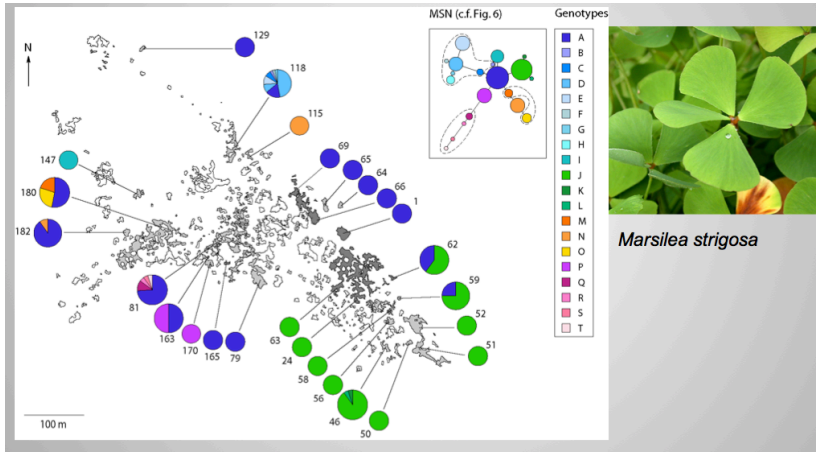
- Décrire le polymorphisme génétique et sa distribution au sein et entre les individus et les populations
- Déduire (inférer) les processus (forces évolutives) qui façonnent le polymorphisme génétique.

→ Comprendre **comment fonctionne l'évolution**

Répartition du polymorphisme génétique:



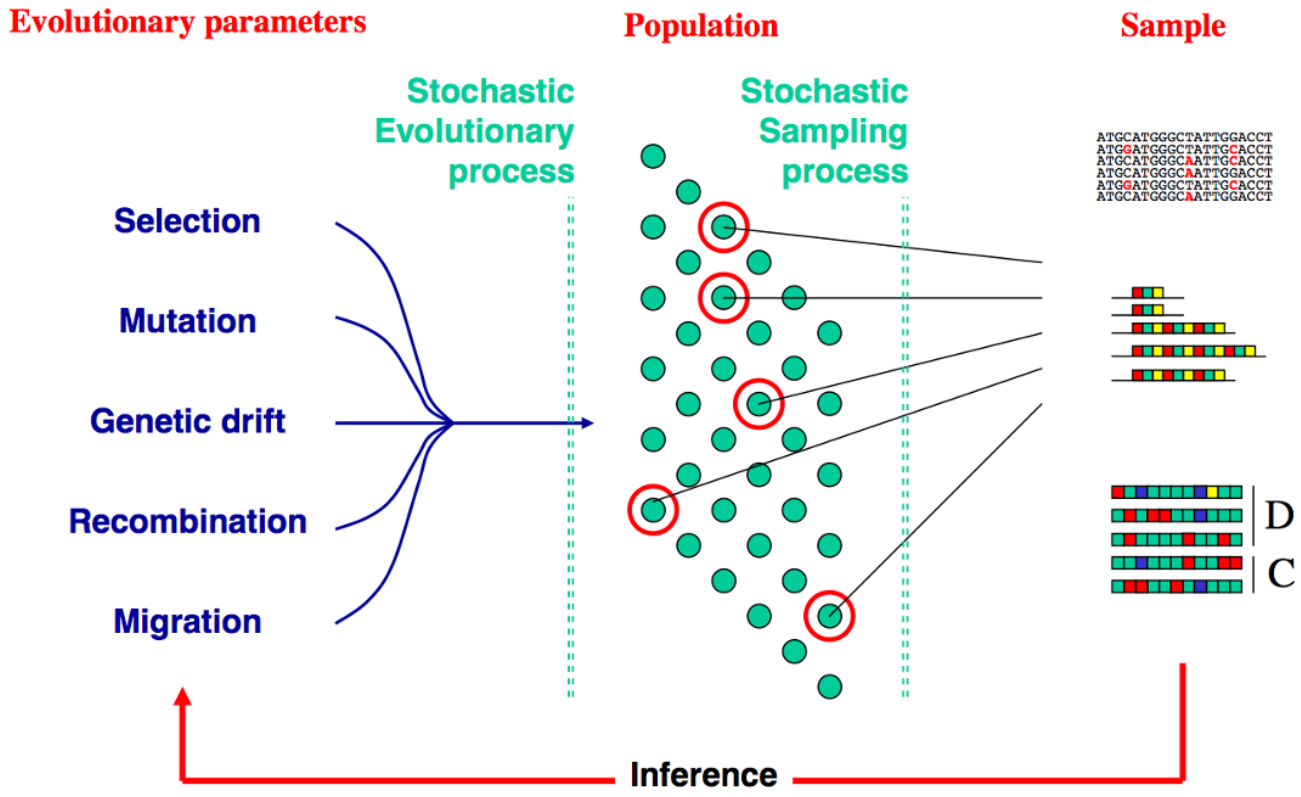
au sein des individus et entre eux



au sein des populations et entre elles

Introduction

Principe de l'inférence en génétique des populations



→ Il faut des données, des modèles démo-génétiques, des méthodes d'inférence statistique

Quelques définitions importantes en génétique des populations

Un **individu diploïde** possède **deux gènes homologues** (un de chaque parent) à un **locus autosomal**: Son **génotype** mono-locus. Ces deux gènes homologues peuvent avoir le même **état allélique** (génotype **homozygote**) ou avoir deux états différents (génotype **hétérozygote**).

Gène: copie d'une information génétique (par exemple une séquence de nucléotides, mais pas seulement : cf méthylation,)

Locus : emplacement d'un gène sur un chromosome.

Allèle (ou état allélique) : classe de gènes homologues équivalents. Deux gènes sont dans le même état allélique s'ils sont des copies exactes d'un ancêtre commun ou s'ils ont la même séquence d'ADN.

Plan

Inférence statistique

Réalité

- Une pièce avec face et pile

- Si on lance dans l'air la pièce elle tombe sur face ou sur pile...

Inférence statistique

Réalité

- Une pièce avec face et pile

- Si on lance dans l'air la pièce elle tombe sur face ou sur pile...

Modèle

- Probabilité face : p
- Probabilité pile : $1 - p$

Inférence statistique

Réalité

- Une pièce avec face et pile

- Si on lance dans l'air la pièce elle tombe sur face ou sur pile...

Modèle

- Probabilité face : p
- Probabilité pile : $1 - p$

Question : estimer p

- p =?

Inférence statistique

Réalité

- Une pièce avec face et pile

- Si on lance dans l'air la pièce elle tombe sur face ou sur pile...

Modèle

- Probabilité face : p
- Probabilité pile : $1 - p$

Question : estimer p

- $p = ?$

Expérience

- On lance dans l'air la pièce 10 fois

Inférence statistique

Réalité

- Une pièce avec face et pile

- Si on lance dans l'air la pièce elle tombe sur face ou sur pile...

Modèle

- Probabilité face : p
- Probabilité pile : $1 - p$

Question : estimer p

- p =?

Expérience

- On lance dans l'air la pièce 10 fois

Données observées

Inférence statistique: Méthode des moments

Données

Estimateur de p

nombre de tirages : $n = 10$
face : $t = 1$; pile : $t = 0$

$$M = \frac{\sum_1^n t_i}{n}$$

$$\mathbb{E}[M] = \frac{\sum_1^n \mathbb{E}[t_i]}{n} = \frac{\sum_1^n p}{n} = p$$

$$\hat{p} = \frac{\sum_1^n t_i}{n} \rightarrow \hat{p} = 0.4$$

Inférence statistique: Maximum de vraisemblance

Données

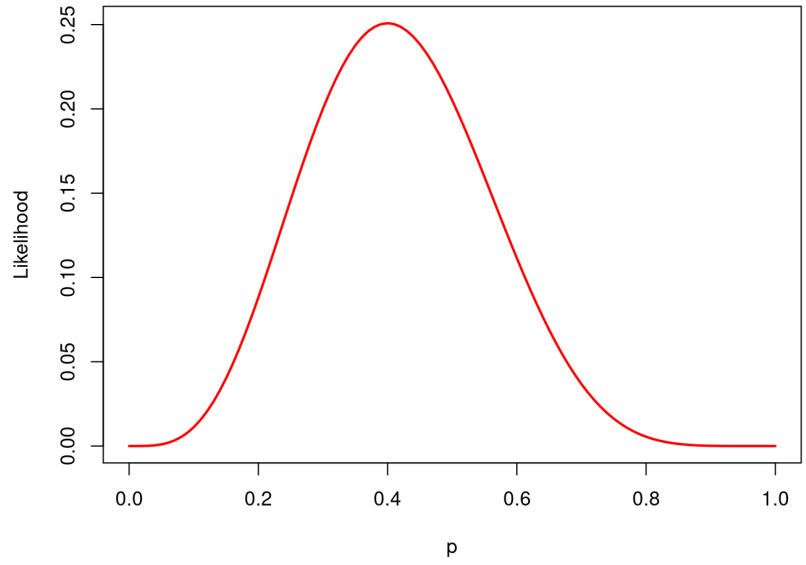
Estimateur de p

$L(p; D) = P(D|p)$

nombre de tirages : $n = 10$
nombre de tirages face : $k = 4$

vraisemblance (loi binomiale)
$P(D|p) = \binom{n}{k} p^k (1-p)^{n-k}$

$\max\{L(p; D)\} \rightarrow \hat{p}$



Inférence statistique: Inférence par simulation

Données

Estimateur de p

$L(p; D) = P(D|p) = ?$

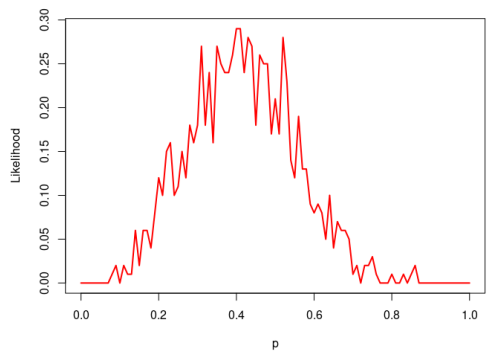
Calcul pour $p = 0.5$ à partir de $r = 100$ simulations

simulation 1 : F,P,F,F,P,P,F,F,P,F ; $k'_1 = 5 \neq k_{\text{obs}} \rightarrow s_1 = 0$

...

simulation 100 : P,P,F,P,F,F,P,F,P,P ; $k'_{100} = 4 = k_{\text{obs}} \rightarrow s_{100} = 1$

$L(p = 0.5; D) \approx \sum_i^r s_i / r$



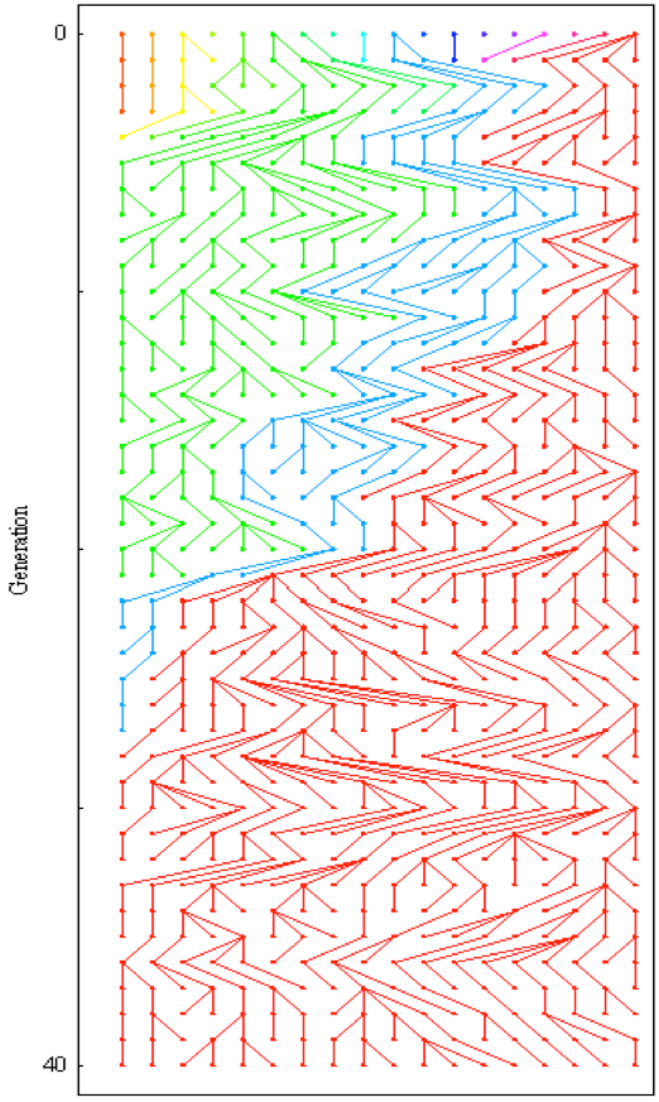
Bilan

- **Méthode des moments**
  - \+ simple d'utilisation
  - \- réduction forte des données
  - \- pas toujours possible / pas pour tous les paramètres
- **Maximum de Vraisemblance**
  - \+ utilise pleinement les données
  - \+ bonnes propriétés statistiques
  - \- vraisemblance pas toujours calculable
- **Inférence par simulations**
  - \+ très flexible (presque toujours possible)
  - \- réduction potentielle des données
  - \- approximation
  - \- temps de calcul

Plan

Modèle de Wright-Fisher pour un locus non-recombinant

sans mutation



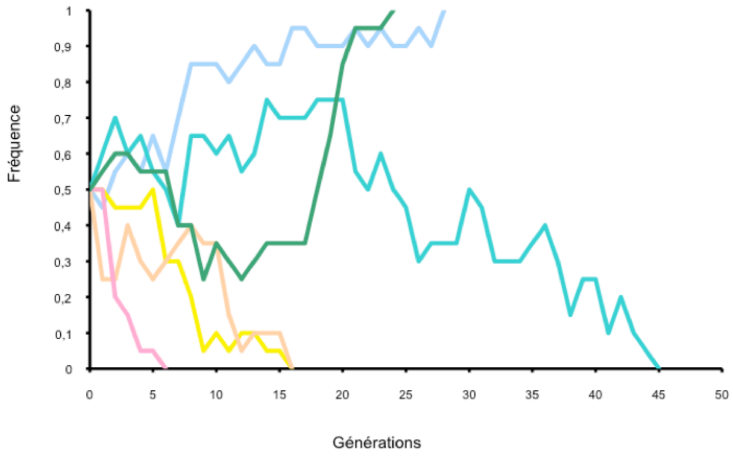
Hypothèses:

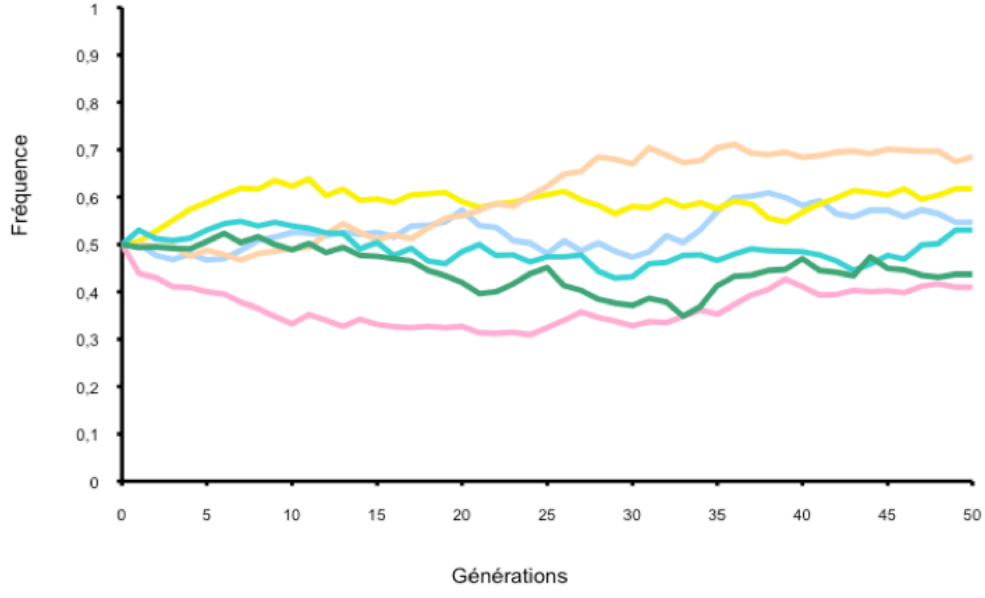
- une population haploïde isolée de taille finie et constante N.
- générations non chevauchantes
- même succès reproducteur pour chaque individu (1 descendant) en espérance mais reproduction aléatoire

Les fréquences alléliques fluctuent au cours du temps

La dérive génétique

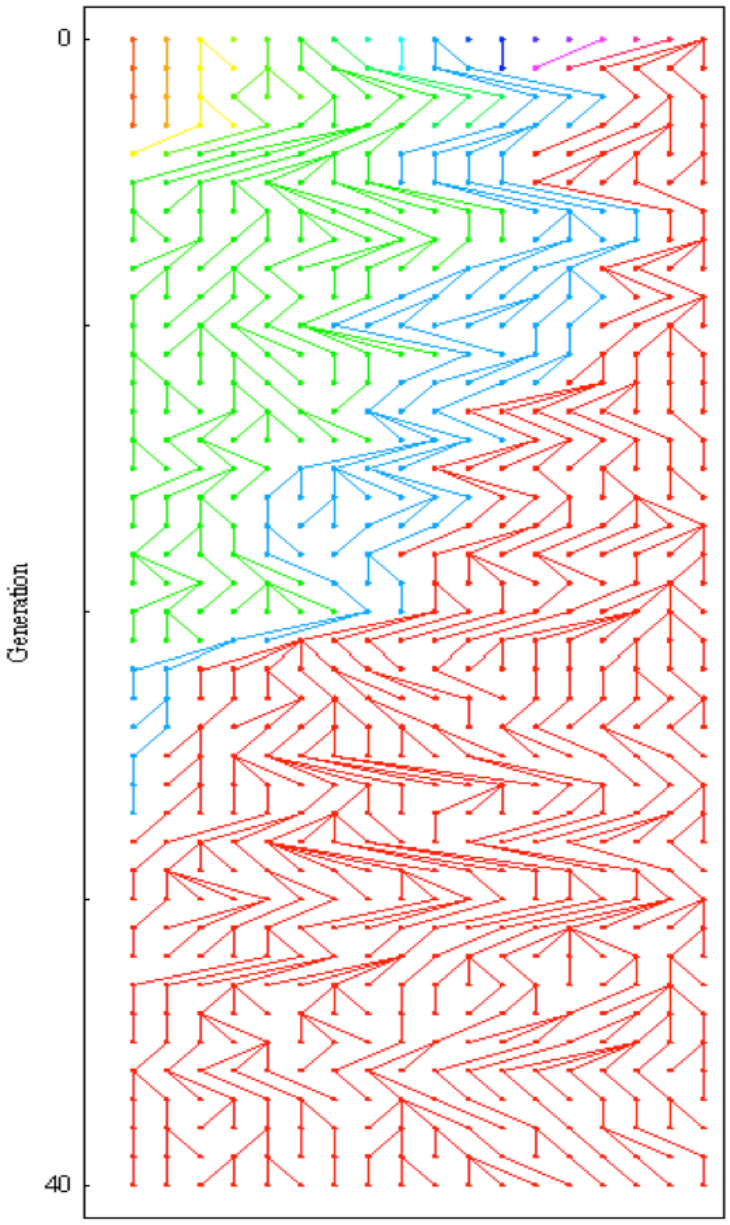
sans mutation





- les fréquences alléliques fluctuent dans le temps en fonction de la taille de la population

Modèle de Wright-Fisher pour un locus non-recombinant



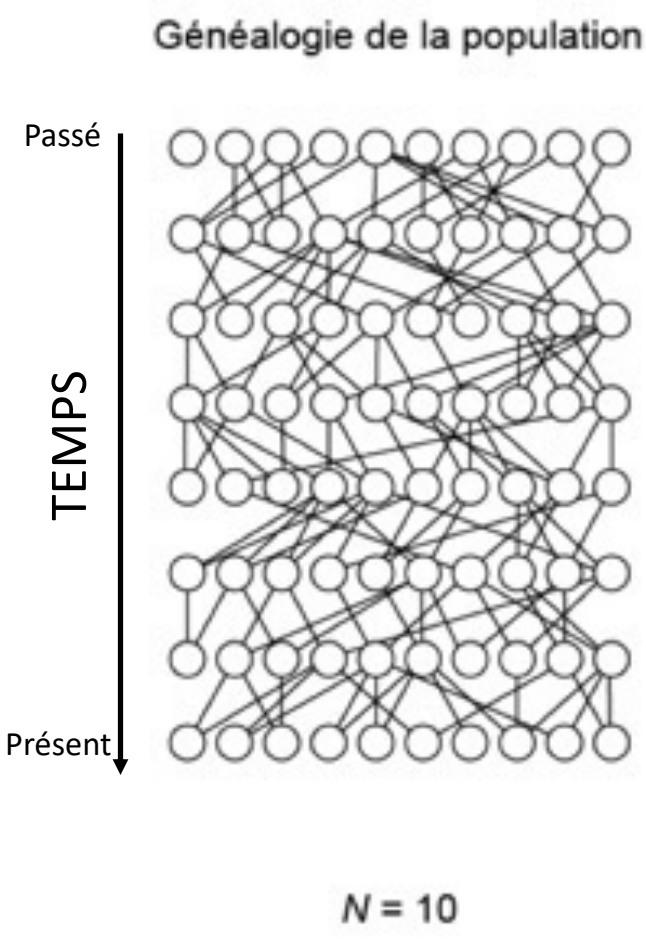
Hypothèses:

- une population haploïde isolée de taille finie et constante N.
- générations non chevauchantes
- même succès reproducteur pour chaque individu (1 descendant) en espérance mais reproduction aléatoire

On peut aussi remonter le temps

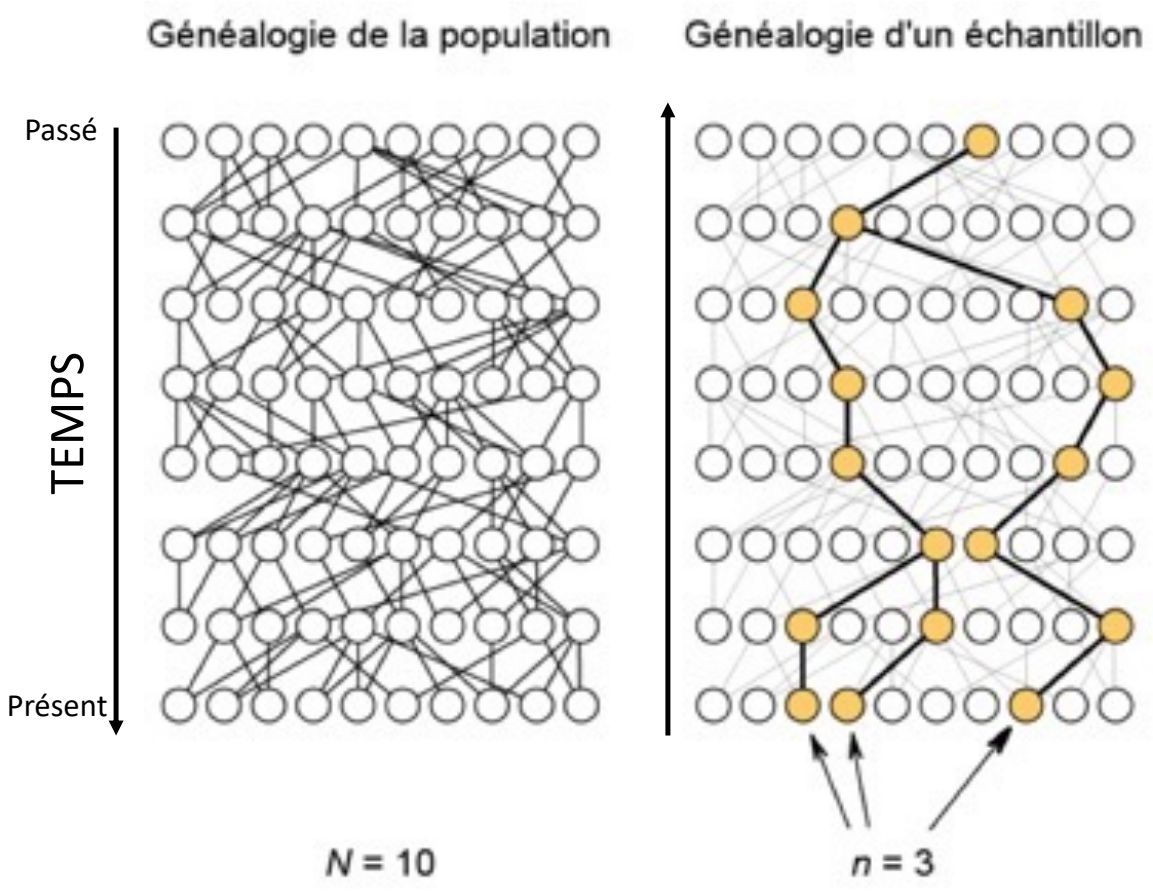
→Chaque individu de la génération $t+1$ choisit son parent uniformément au hasard et avec remplacement parmi les N adultes de la génération t (coalescence).

La coalescence



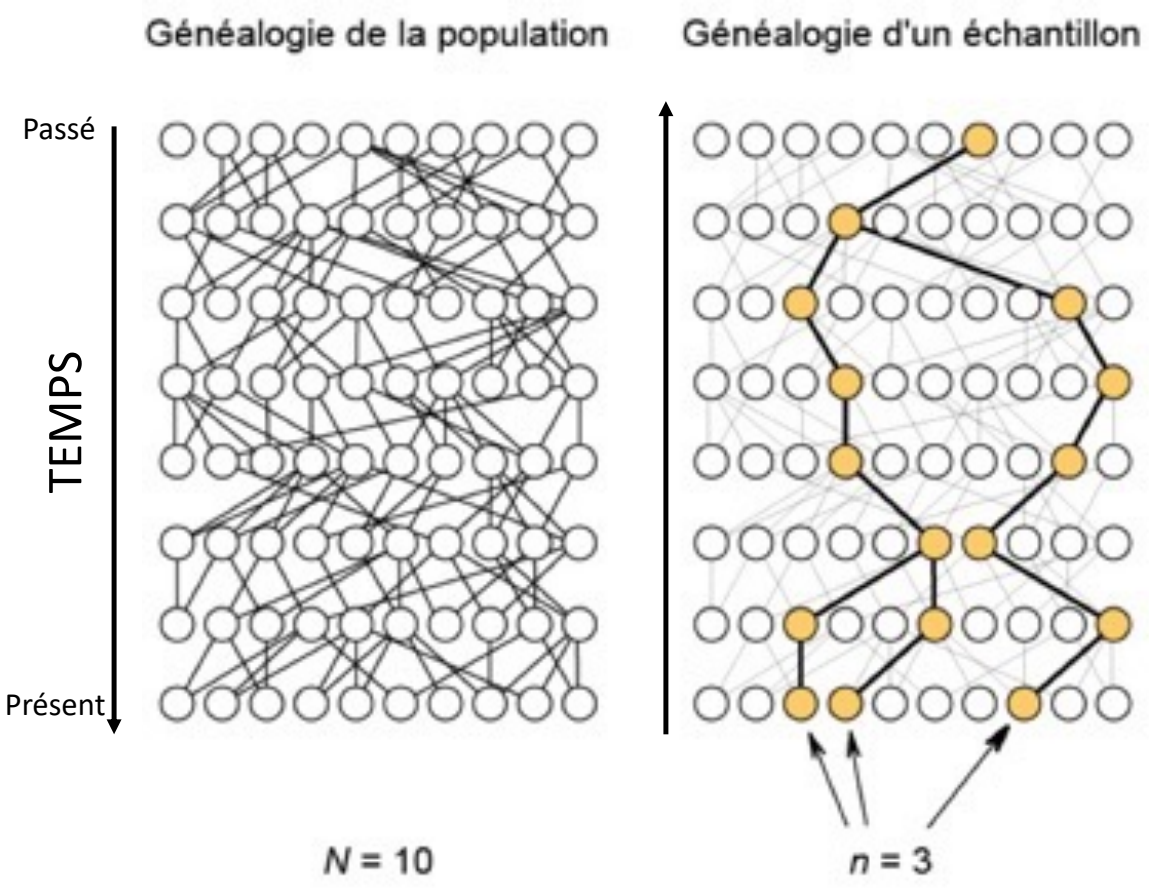
Approche classique "en avant": Reconstruire la transmission des lignées génétiques par descendance à chaque génération, dans le sens de l'évolution et dans toute la population

La coalescence génération par génération



La coalescence ("en arrière") : Reconstruire la généalogie des lignées ancestrales d'un échantillon

La coalescence génération par génération

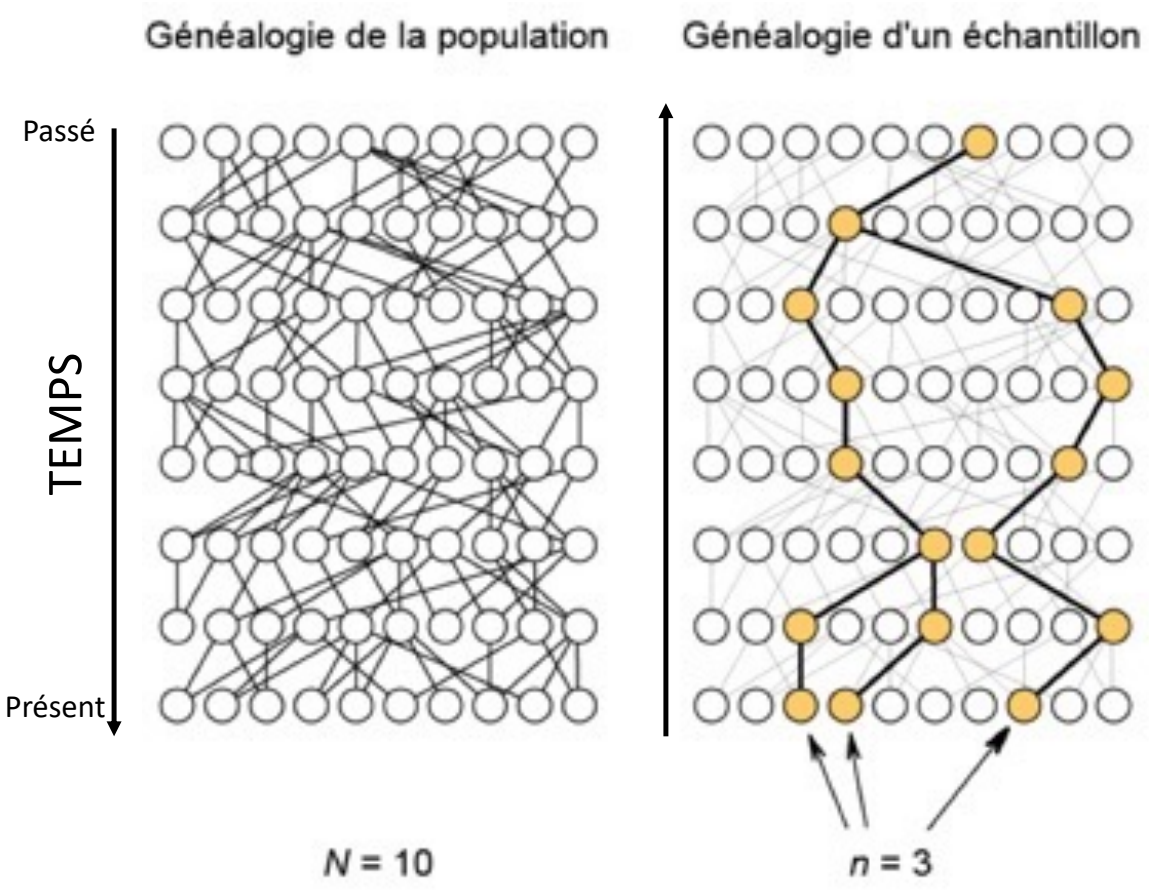


coalescence $(1/N)$?

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération

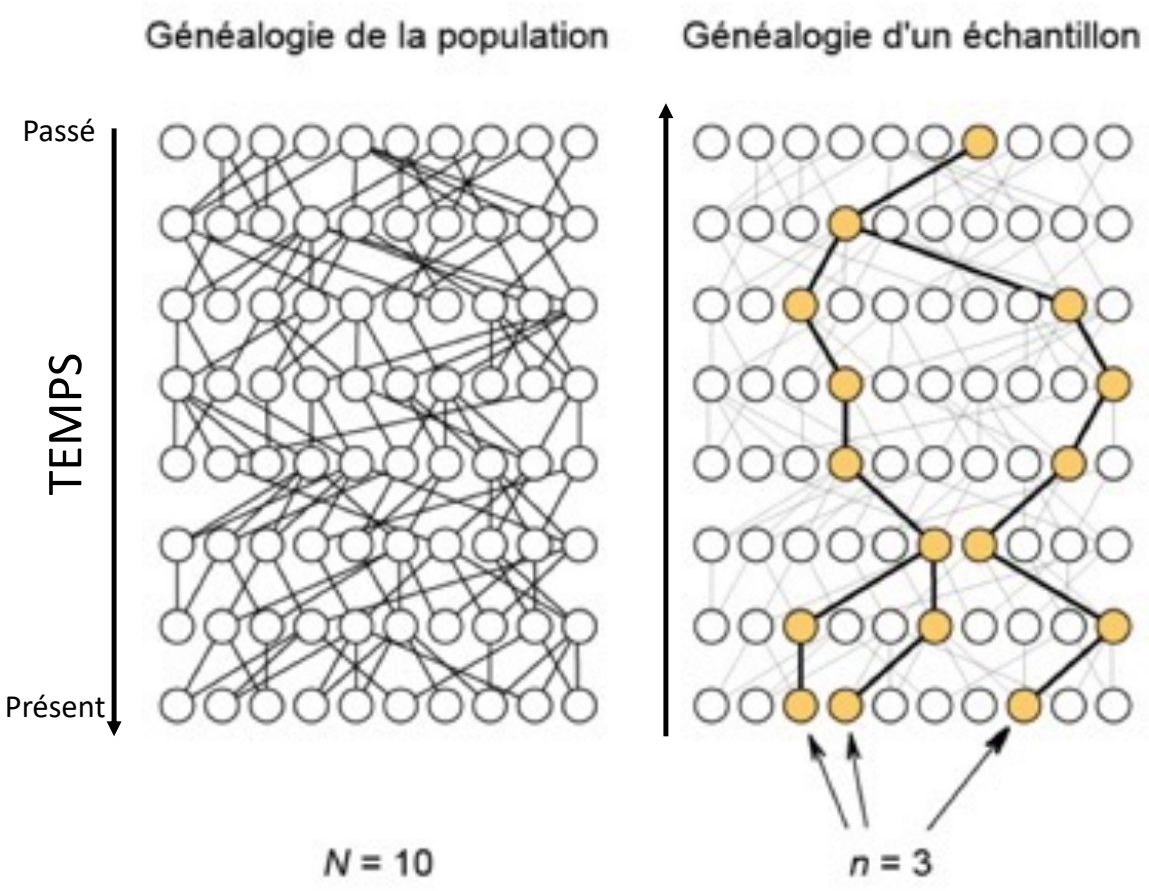


pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



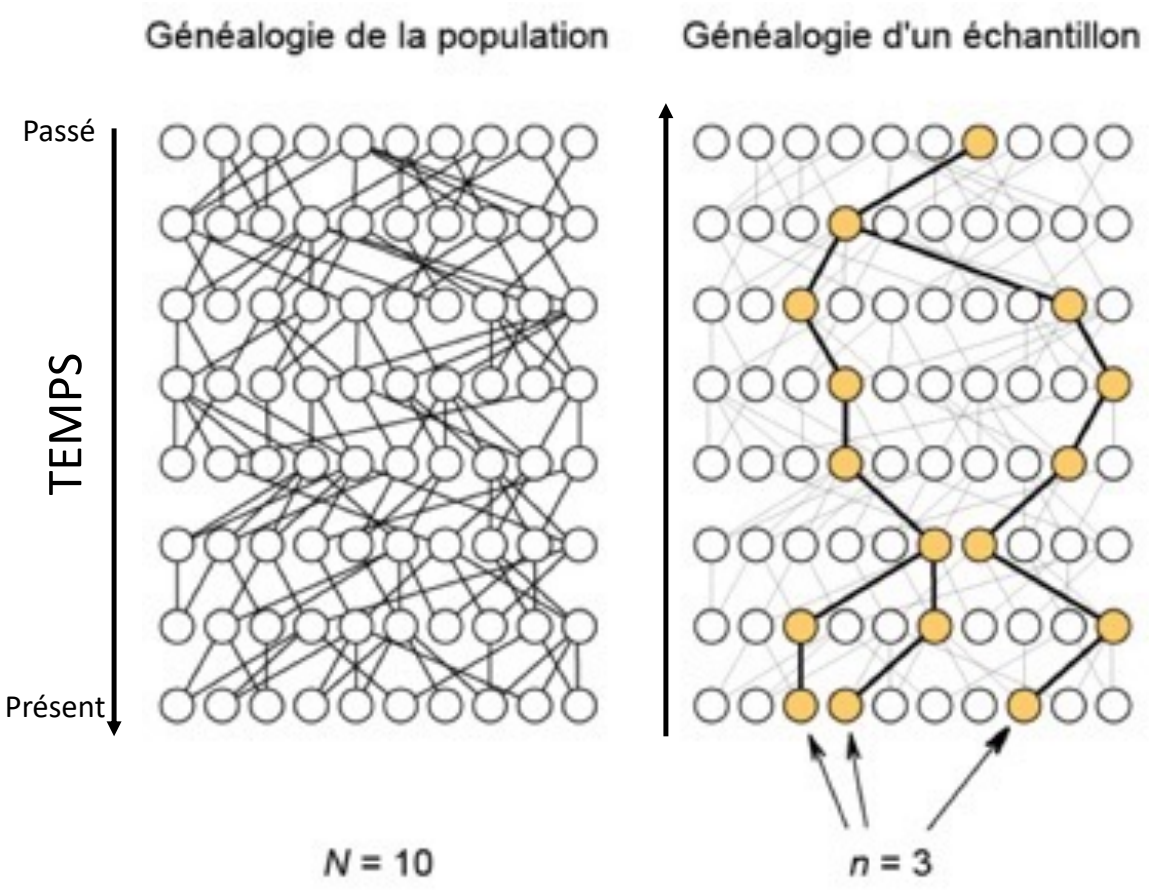
coalescence $(1/N)$?

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



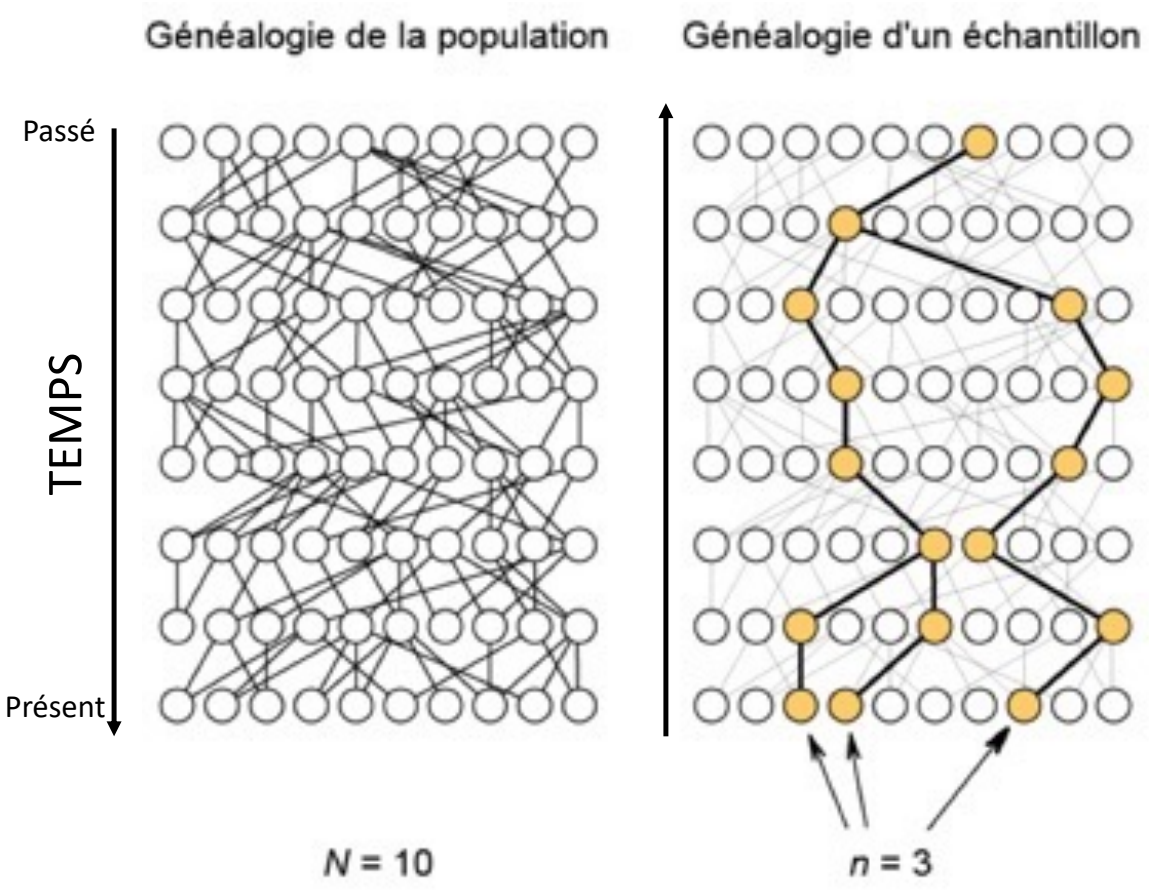
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



coalescence $(1/N)$?

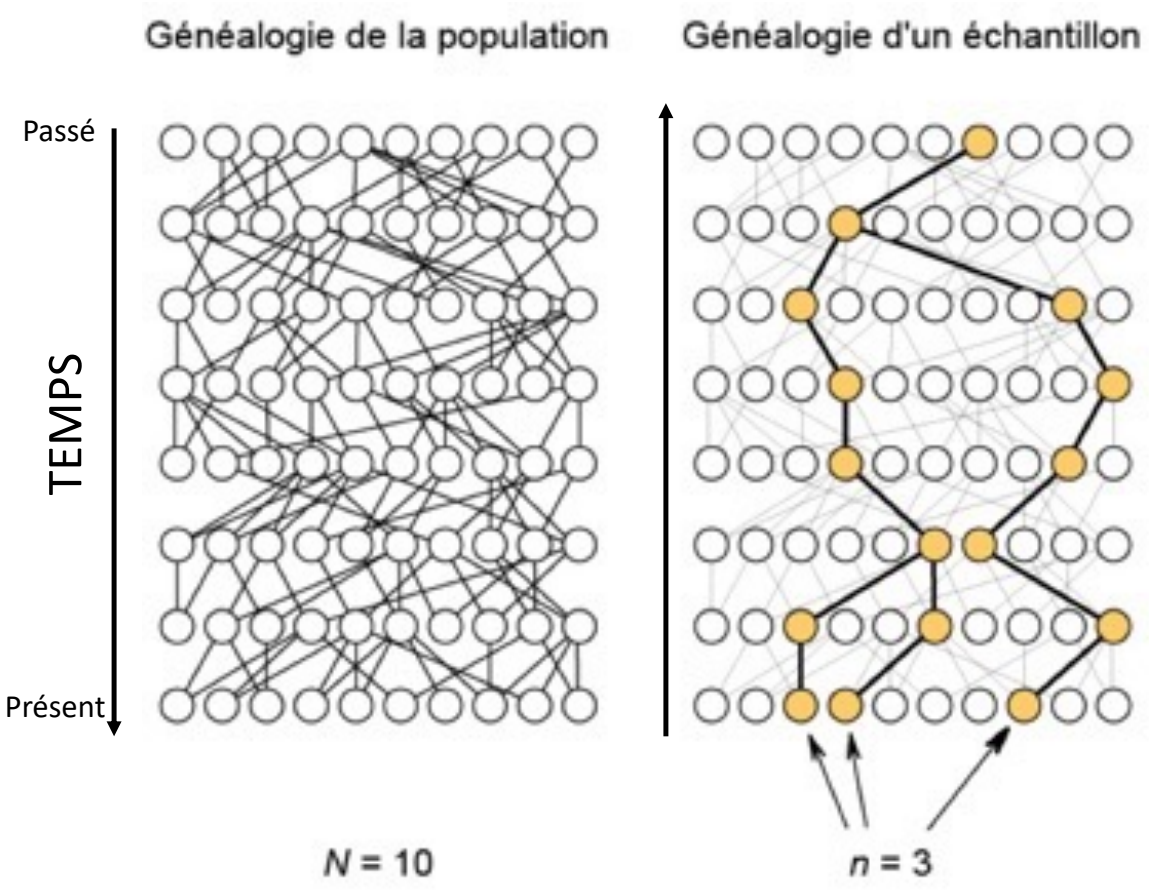
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



pas de coalescence

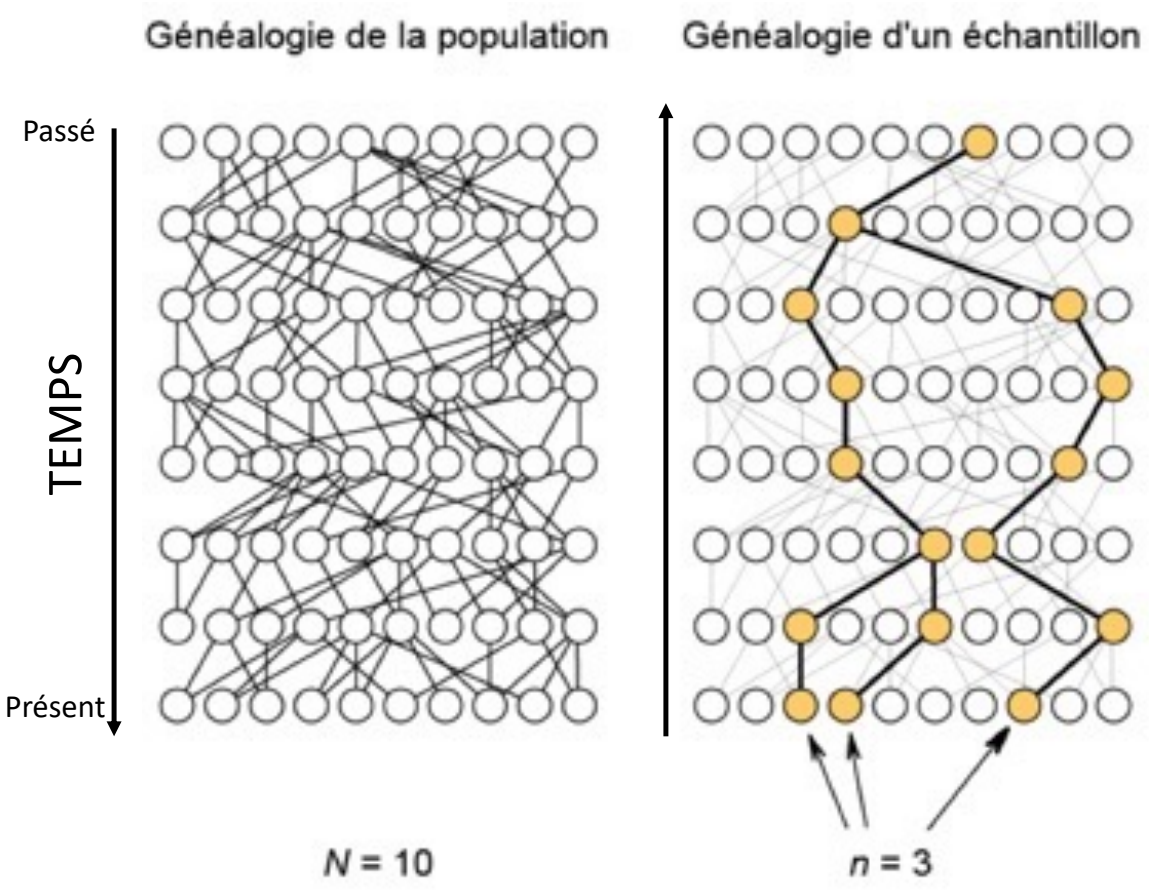
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



coalescence $(1/N)$?

pas de coalescence

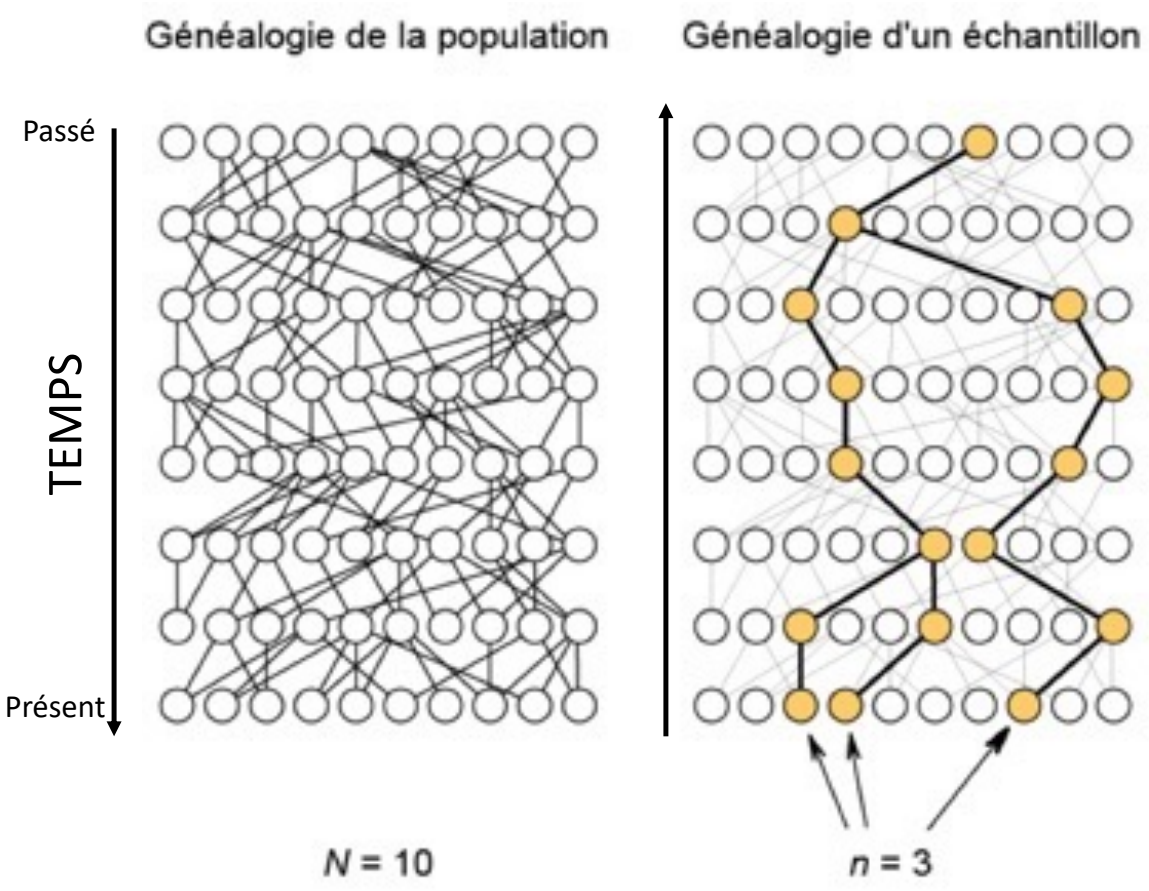
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



pas de coalescence

pas de coalescence

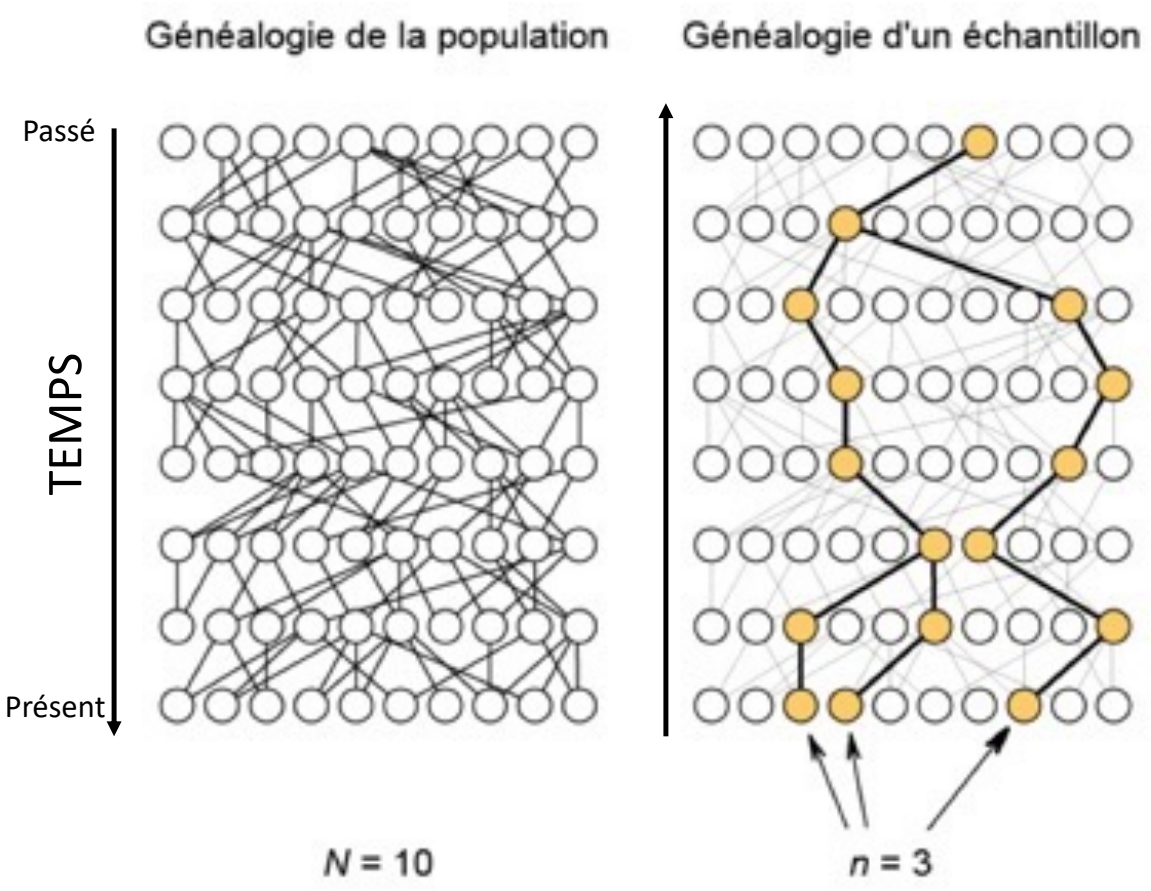
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



coalescence $(1/N)$?

pas de coalescence

pas de coalescence

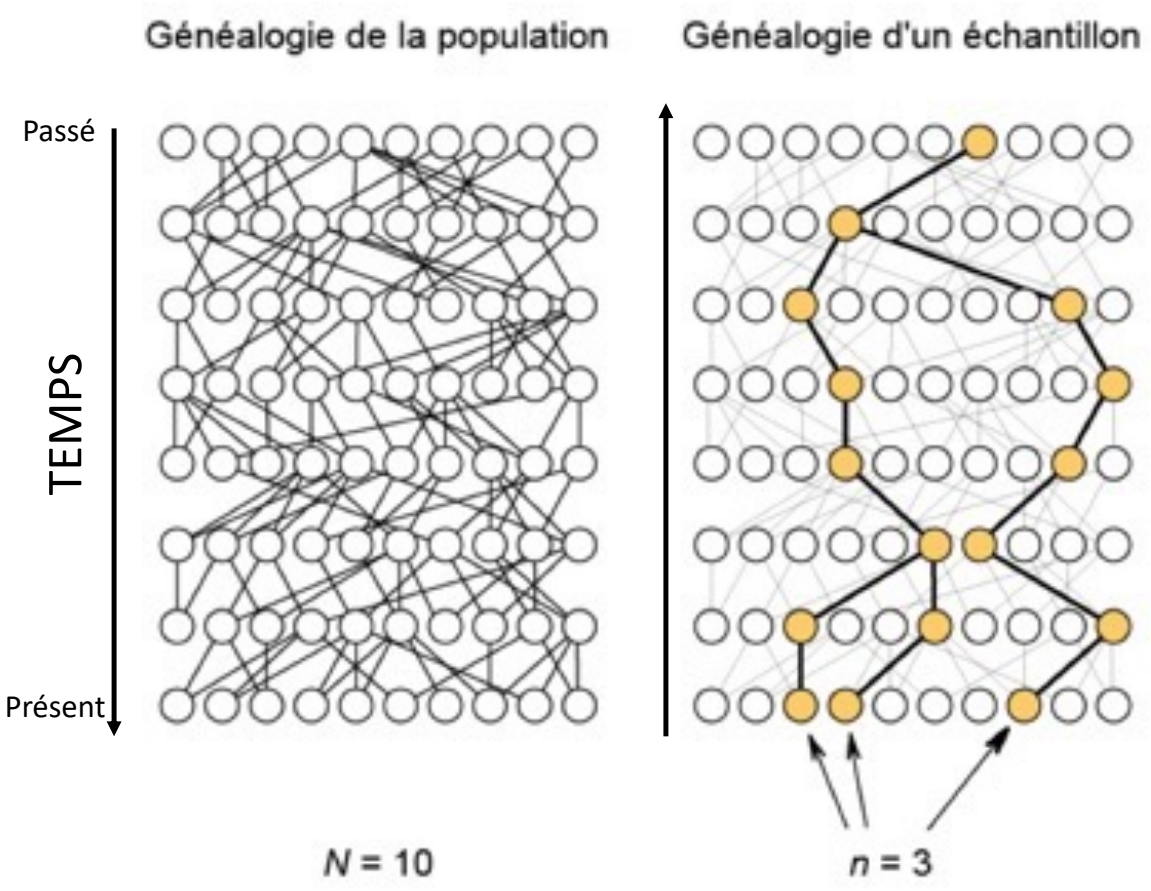
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



pas de coalescence

pas de coalescence

pas de coalescence

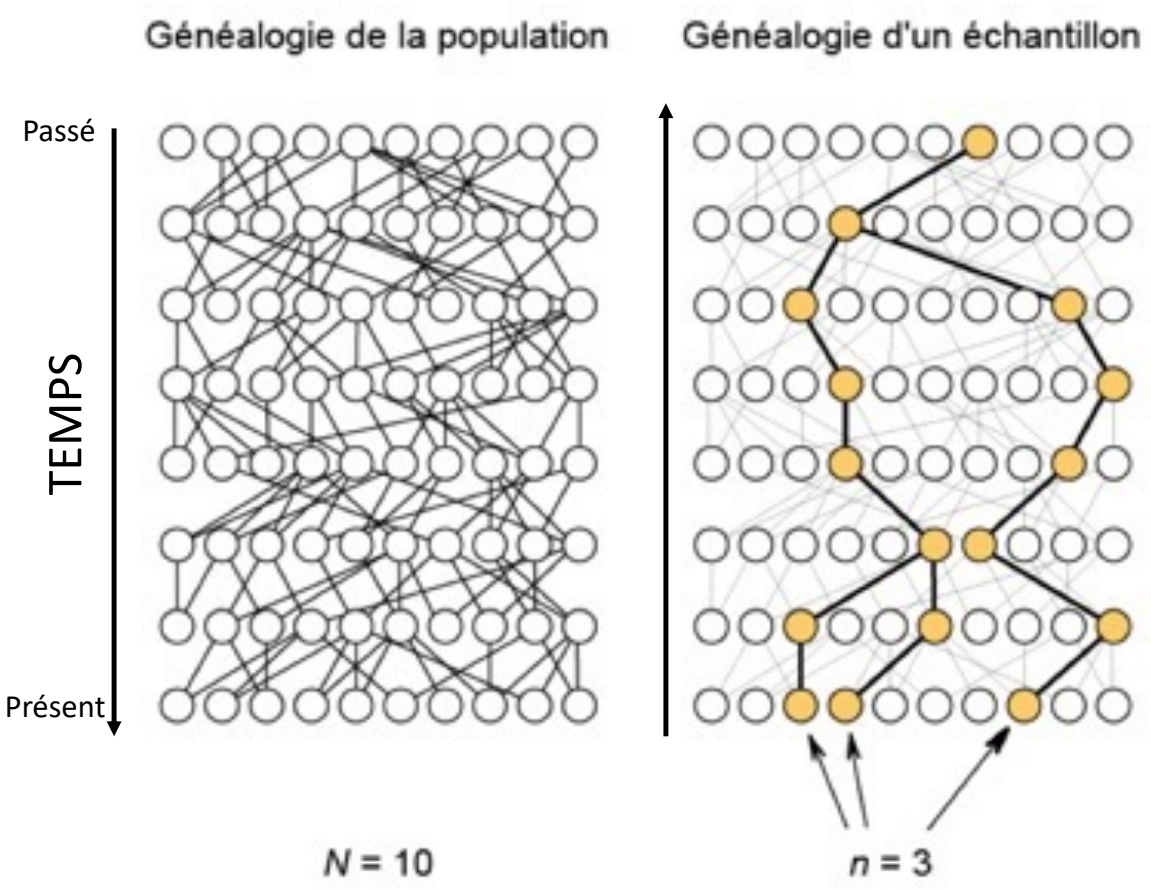
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération

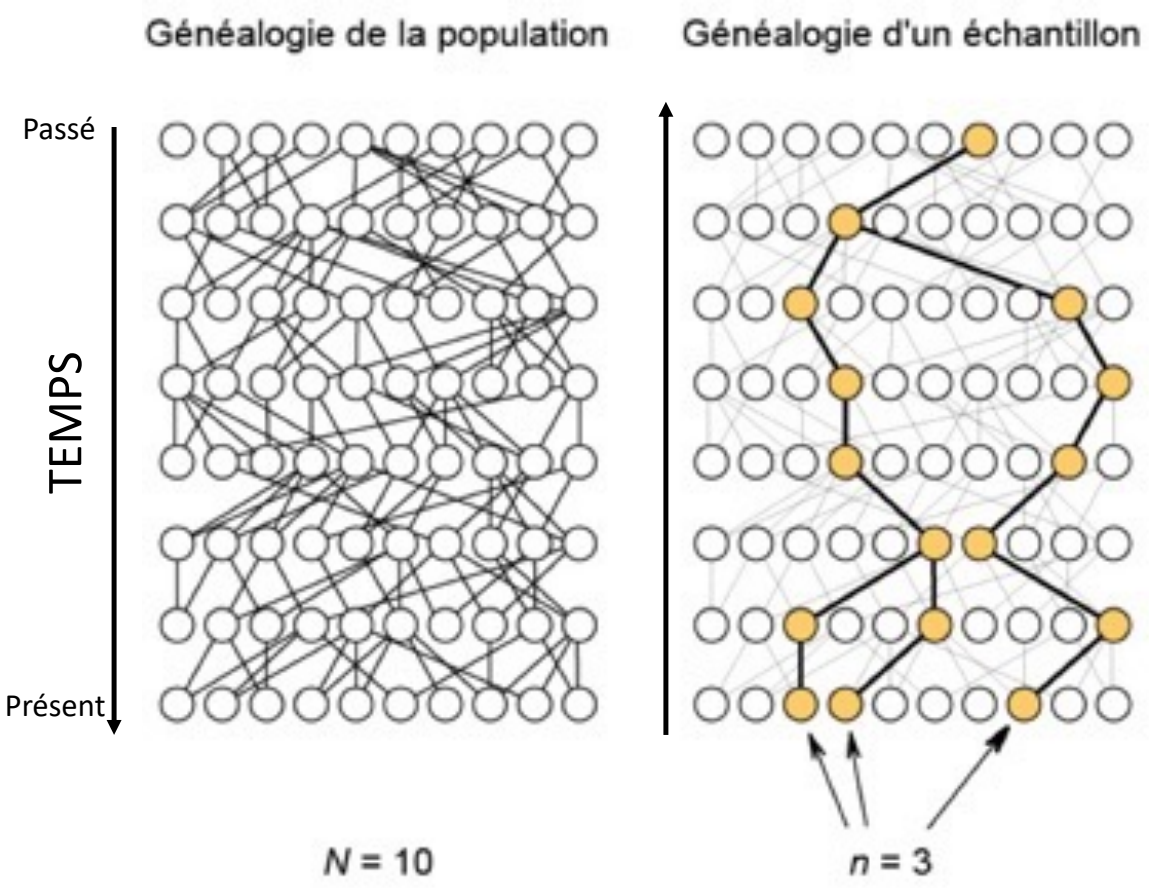


coalescence $(1/N)$?
pas de coalescence
pas de coalescence
pas de coalescence
une coalescence
pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération

La coalescence génération par génération



Ancêtre de tout l'échantillon -> STOP

une coalescence

pas de coalescence

pas de coalescence

pas de coalescence

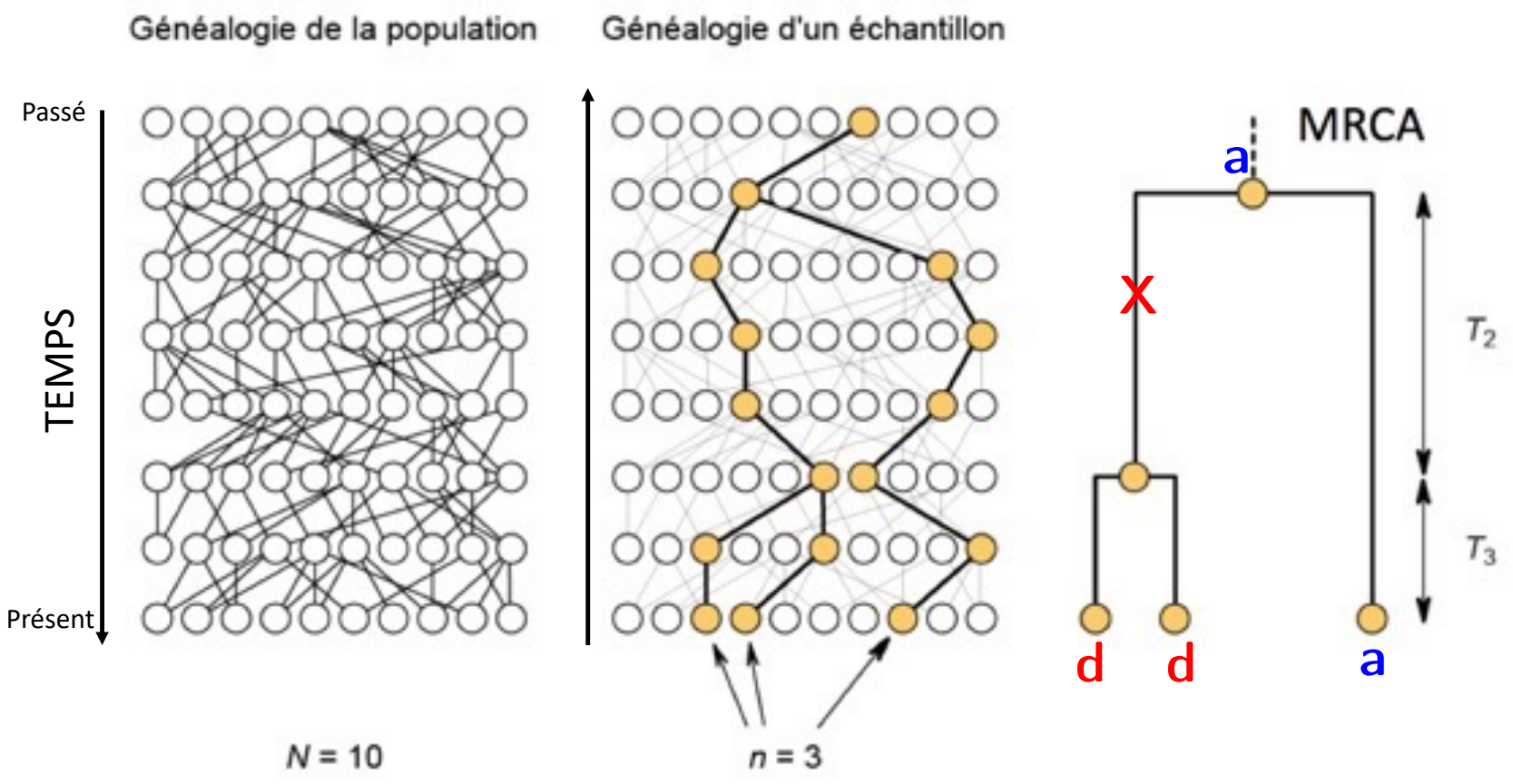
une coalescence

pas de coalescence

La coalescence : Reconstruire l'arbre des lignées ancestrales

Approche exacte génération par génération (rapide et flexible)

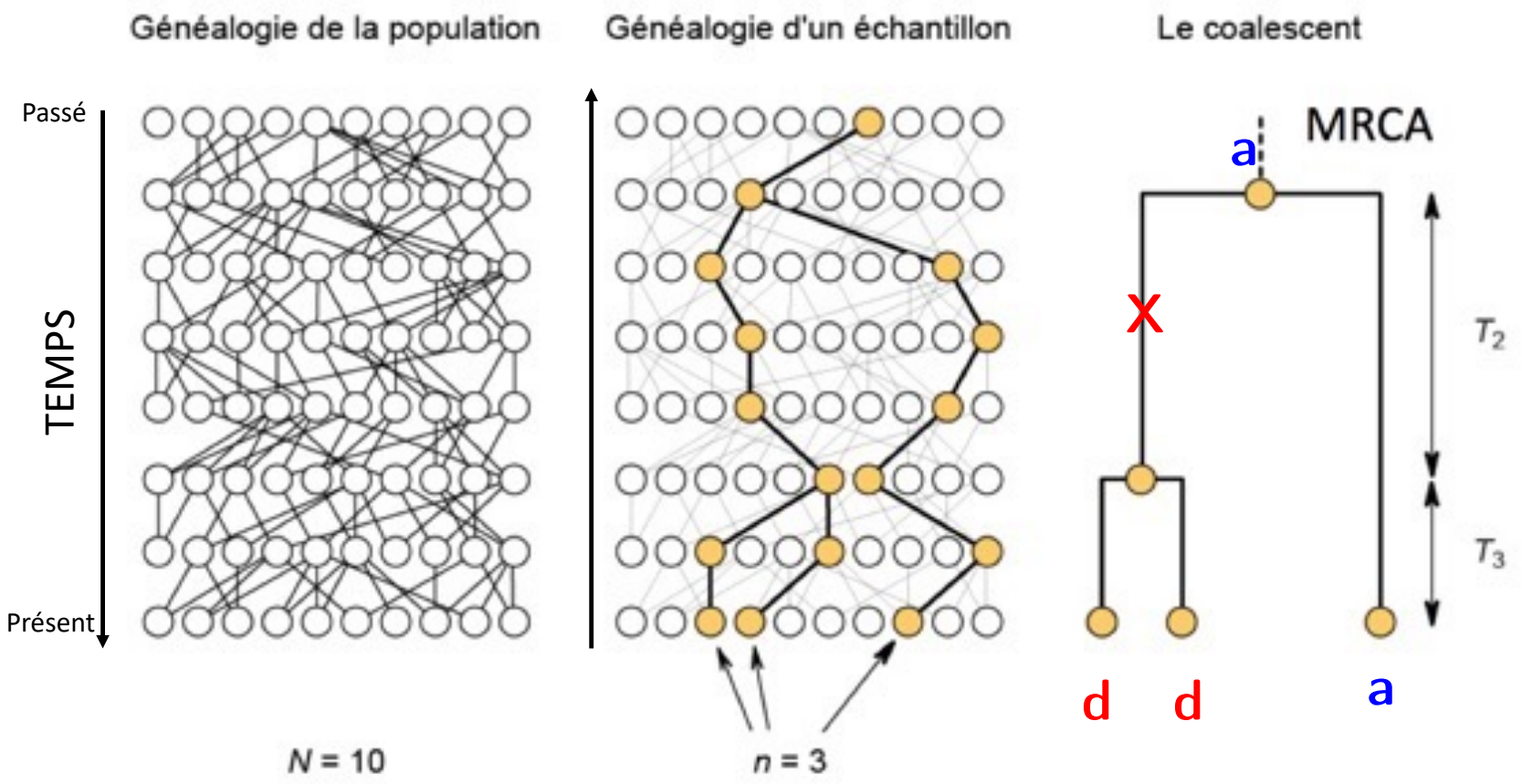
La coalescence génération par génération



La coalescence : Reconstruire l'arbre des lignées ancestrales

Ajout des mutations sur les branches de l'arbre pour obtenir un échantillon génétique (d,d,a)

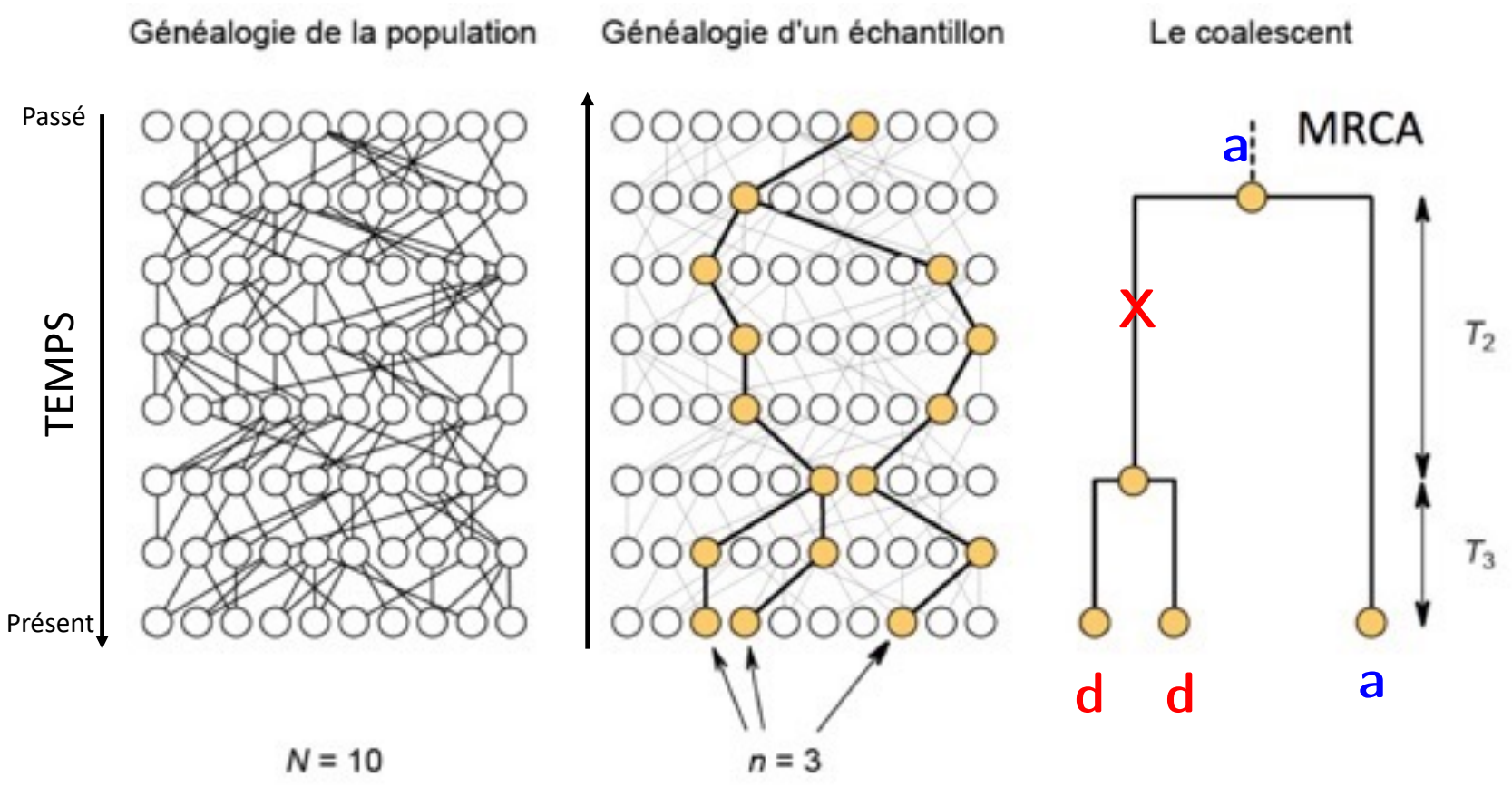
La coalescence



Le n–coalescent = modèle mathématique approché du processus de coalescence.

Le plus souvent utilisé, le plus efficace mais le moins flexible

La coalescence



Information des données = évènements de coalescence et mutations

La coalescence

La théorie de la coalescence permet donc:

- Simulations très efficaces
- intuition sur les patrons de polymorphisme et le comportement des inférences en génétique des populations

Plan

Plan

coalescence et taille de population

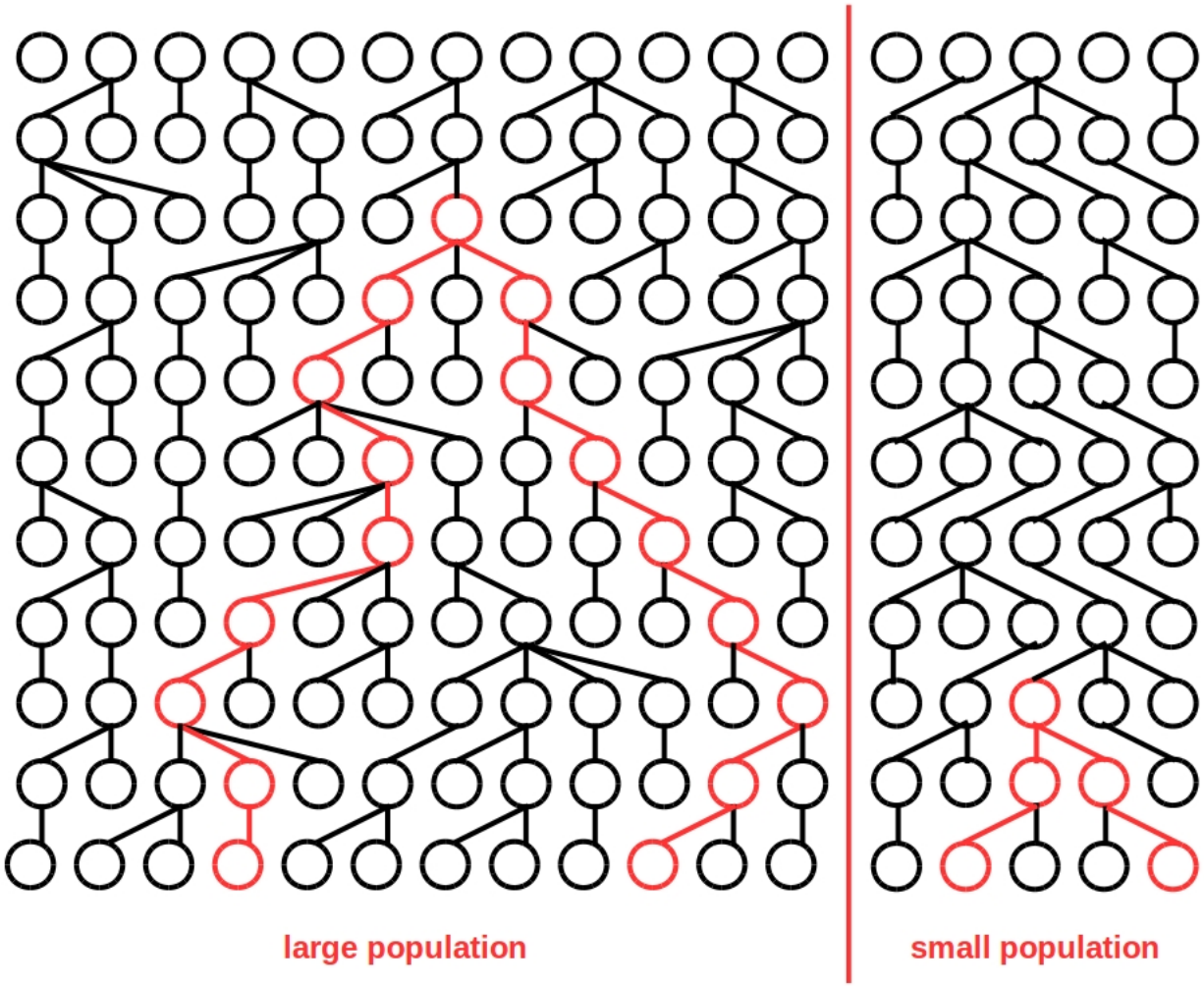
- probabilité de coalescence de 2 lignées $= 1/N$

  $\rightarrow$ Le temps de coalescence augmente avec la taille de la population

  $\rightarrow$ Le nombre de mutations sur une branche augmente avec le temps de coalescence

Intuition en population de taille constante

N grand → temps de coalescence plus longs → plus de mutations

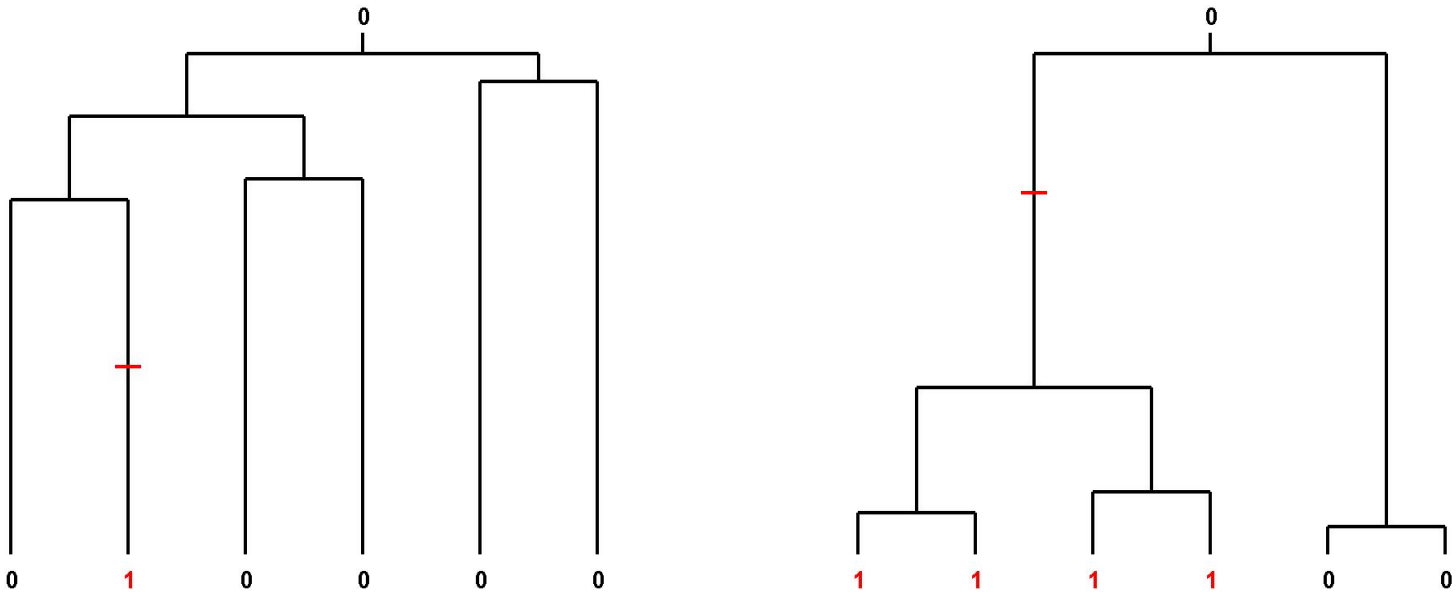


→ diversité génétique plus forte dans les grandes populations

(= moins de dérive)

Intuition en population de taille variable

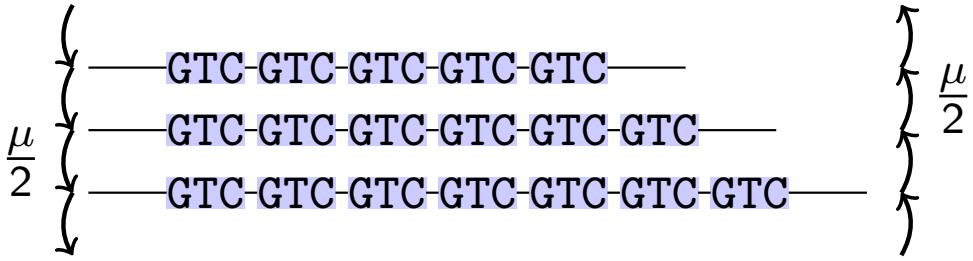
- Population en **expansion** → temps de coalescence plus longs dans un passé récent → beaucoup d'**allèles dérivés rares**
- Population en **déclin** → temps de coalescence plus longs dans un passé ancien → beaucoup d'**allèles dérivés communs**



Plan

Méthode des moments : Estimation de $\theta = 4N_e\mu$

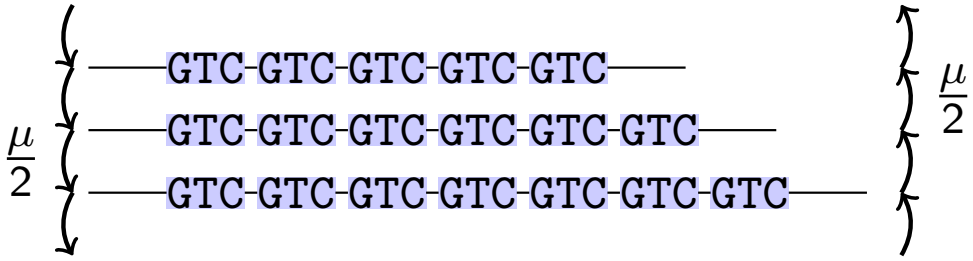
Les microsatellites et le modèle de mutation par pas



3 estimateurs de θ

Méthode des moments : Estimation de $\theta = 4N_e\mu$

Les microsatellites et le modèle de mutation par pas



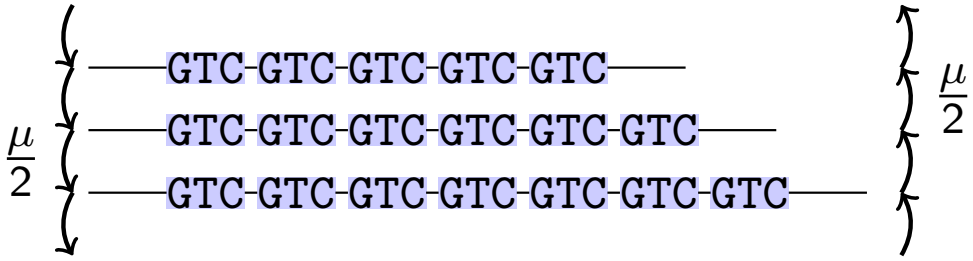
3 estimateurs de θ

- Nombre d'allèles :

  A ; $\mathbb{E}[A] = \sqrt{1 + 2\theta}$; $\hat{\theta}_A = \frac{A^2-1}{2}$

Méthode des moments : Estimation de $\theta = 4N_e\mu$

Les microsatellites et le modèle de mutation par pas



3 estimateurs de θ

- Nombre d'allèles :

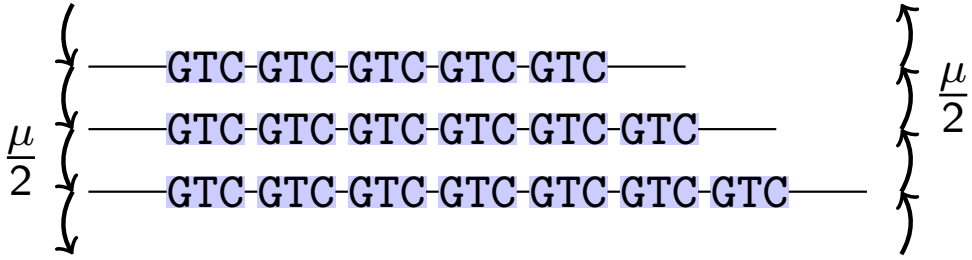
  A ; $\mathbb{E}[A] = \sqrt{1+2\theta}$; $\hat{\theta}_A = \frac{A^2-1}{2}$

- Hétérozygotie :

  $H_e = 1 - Hom$; $\mathbb{E}[Hom] = \frac{1}{\sqrt{1+2\theta}}$; $\hat{\theta}_{Hom} = \frac{1/Hom^2-1}{2}$

Méthode des moments : Estimation de $\theta = 4N_e\mu$

Les microsatellites et le modèle de mutation par pas



3 estimateurs de θ

- Nombre d'allèles :

  A ; $\mathbb{E}[A] = \sqrt{1+2\theta}$; $\hat{\theta}_A = \frac{A^2-1}{2}$

- Hétérozygotie :

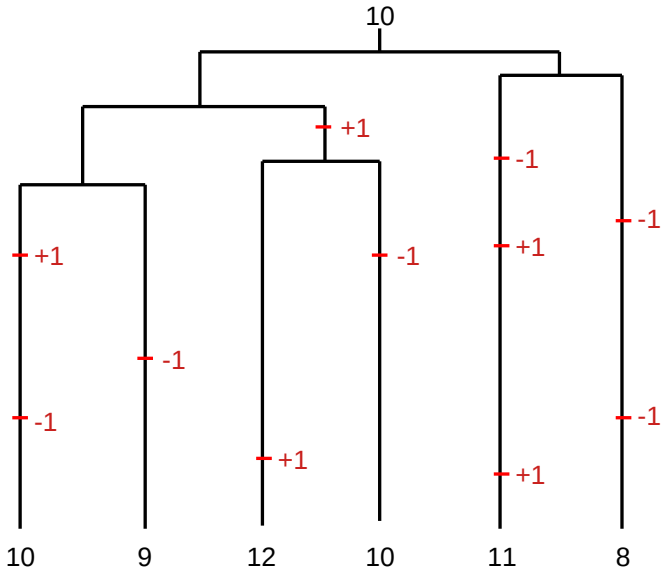
  $H_e = 1 - Hom$; $\mathbb{E}[Hom] = \frac{1}{\sqrt{1+2\theta}}$; $\hat{\theta}_{Hom} = \frac{1/Hom^2-1}{2}$

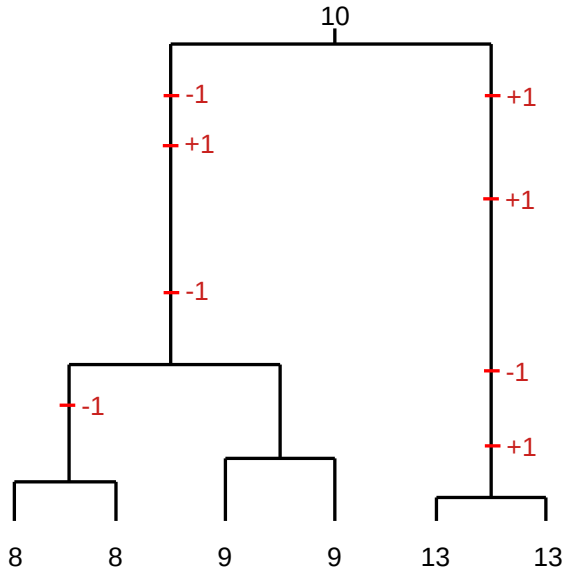
- Variance de taille allélique :

  V ; $\mathbb{E}[V] = \theta$; $\hat{\theta}_V = V$

Taille de population variable et diversité des microsatellites

A, He, V ont des dynamiques différentes suite à un changement de taille de population





Augmentation de la taille

$A = 5$; $H_e = 0.93$; $V = 2$

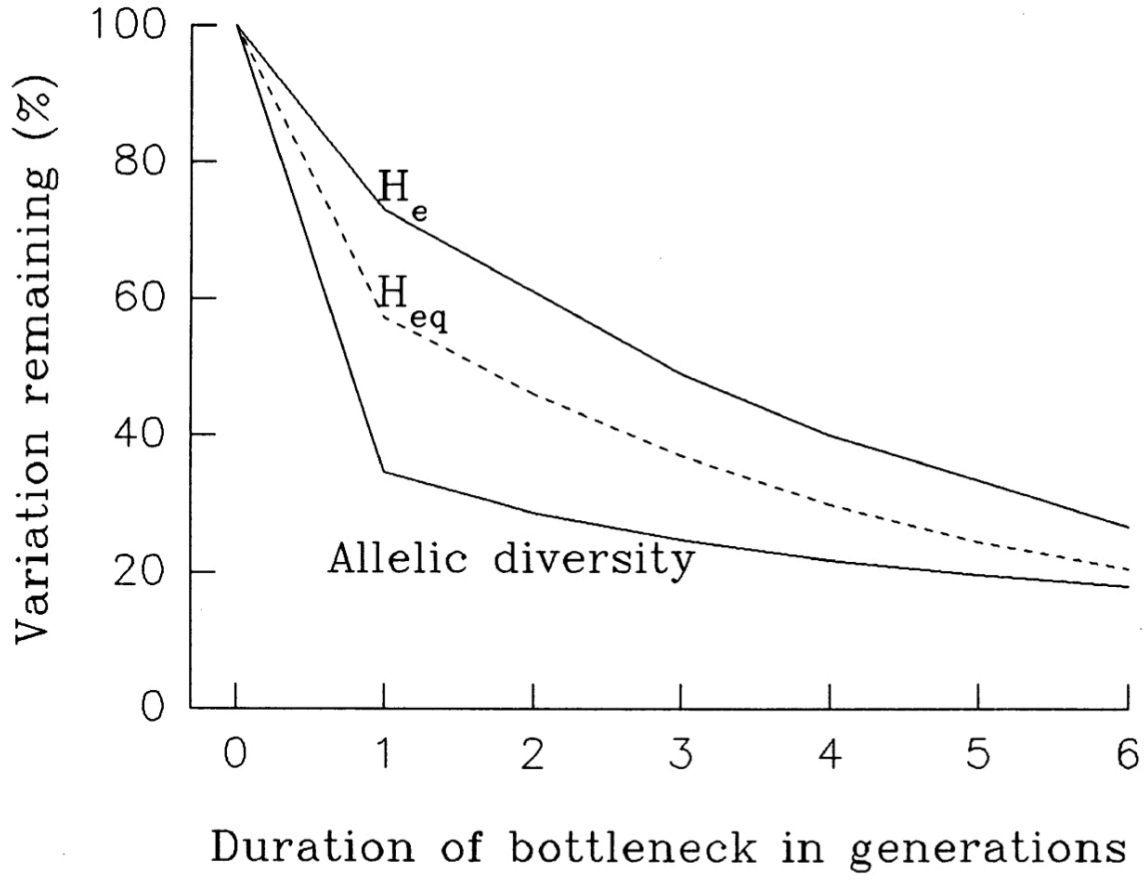
Réduction de la taille

$A = 3$; $H_e = 0.8$; $V = 5.6$

→ les 3 estimateurs de θ peuvent être utilisés pour détecter un changement de taille de population

Taille de population variable et diversité des microsatellites

Logiciel `Bottleneck` (Cornuet et al. 1996, CBGP) : détection de changements passés de taille de population à partir de données microsatellites



Test statistique sur l'excès (contraction) / déficit (expansion) de H_e par rapport à l'attendu H_{eq} sachant A dans une population à l'équilibre

Question typique : Les orangs-outans et la déforestation de Bornéo

- Le génome des orangs-outans a été façonné par son histoire démographique et on sait qu'il y a eu une diminution de la taille de l'habitat
- Goossens et al. 2006 : 200 indvidus échantillonnés sur 9 sites et génotypés à 14 locus microsatellites

- La génétique des populations peut-elle détecter une potentielle réduction passé de taille de population due à la diminution de la taille de l'habitat?

Question typique : Les orangs-outans et la déforestation de Bornéo

- Le génome des orangs-outans a été façonné par son histoire démographique et on sait qu'il y a eu une diminution de la taille de l'habitat
- Goossens et al. 2006 : 200 indvidus échantillonnés sur 9 sites et génotypés à 14 locus microsatellites

- Oui, tout les sites échantillonnés montrent un excès significatif d'H_e avec `Bottleneck`

Question typique : Les orangs-outans et la déforestation de Bornéo

- Le génome des orangs-outans a été façonné par son histoire démographique et on sait qu'il y a eu une diminution de la taille de l'habitat
- Goossens et al. 2006 : 200 indvidus échantillonnés sur 9 sites et génotypés à 14 locus microsatellites

- Oui, tout les sites échantillonnés montrent un excès significatif d'H_e avec `Bottleneck`
- mais ces données microsatellites permettent-elles de caractériser ces réductions passés de tailles de populations ?

Méthodes des moments et structuration des populations

Inférences basées sur les méthodes des moments

- une seule **statistique** de différenciation génétique, liée à un paramètre du modèle, ex. F_{st}:

$$F_{st} \approx \frac{1}{(1+4Nm)} \rightarrow \widehat{Nm} \approx \frac{1}{4}\left(\frac{1}{\widehat{F_{st}}-1}\right) \text{ (modèle de migration en îles)}$$

$$F_{st} \approx 1 - (1 - \frac{1}{(2N)})^t \approx 1 - \exp(-t/(2N)$$

$$\rightarrow \widehat{t/2N} \approx -log(1 - \widehat{Fst}) \text{ (modèle de divergence)}$$

Méthodes des moments et structuration des populations

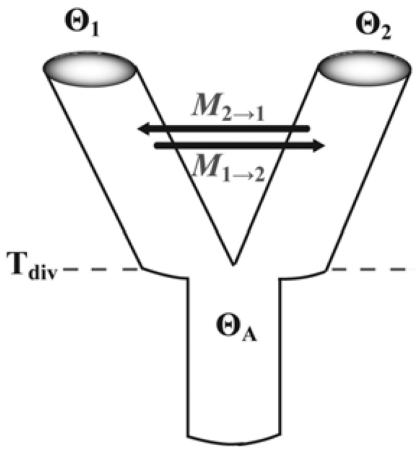
Inférences basées sur les méthodes des moments

- une seule **statistique** de différenciation génétique, liée à un paramètre du modèle, ex. F_{st}:

$$F_{st} \approx \frac{1}{(1+4Nm)} \rightarrow \widehat{Nm} \approx \frac{1}{4}\left(\frac{1}{\widehat{F_{st}}-1}\right)$$

$$F_{st} \approx 1-(1-\frac{1}{(2N)})^{t} \approx 1-\exp(-t/(2N)$$

$$\rightarrow \widehat{t/2N} \approx -log(1-\widehat{Fst})$$



forte limitation: ne peut pas prendre en compte des modèles plus complexes, par exemple divergence avec migration

Plan

Motivations et principe

L'inférence basé sur la vraisemblance est statistiquement optimale.
Par rapport à la méthode des moments, elle permet :

- Plus de puissance (tests)
- Plus de précision (estimations)
- Considérer des modèles un peu plus complexes

  Exemple : un changement de taille
  → 3 paramètres:
  taille actuelle (N), taille ancestrale (N_{anc}),
  temps du changement (T)

Motivations et principe

L'inférence basé sur la vraisemblance est statistiquement optimale.
Par rapport à la méthode des moments, elle permet :

- Plus de puissance, plus de précision, modèles plus complexes

  Exemple : un changement de taille (N, T, N_{anc})

- Démographie $N()$, vraisemblance L
  $L(N(); D) = P(D|N()) = P(D|N, N_{anc}, T)$ (p.ex.)

Motivations et principe

L'inférence basé sur la vraisemblance est statistiquement optimale.
Par rapport à la méthode des moments, elle permet :

- Plus de puissance, plus de précision, modèles plus complexes

  Exemple : un changement de taille (N, T, N_{anc})

- Démographie $N()$, vraisemblance L
  $L(N(); D) = P(D|N()) = P(D|N, N_{anc}, T)$ (p.ex.)

Motivations et principe

L'inférence basé sur la vraisemblance est statistiquement optimale.
Par rapport à la méthode des moments, elle permet :

- Plus de puissance, plus de précision, modèles plus complexes

  Exemple : un changement de taille (N, T, N_{anc})

- Démographie $N()$, vraisemblance L
  $L(N(); D) = P(D|N()) = P(D|N, N_{anc}, T)$ (p.ex.)
- L locus indépendants : $P(D|N()) = \prod_{i=1}^{L} P(D_i|N())$
- $P(D_i|N())$ implique typiquement des modèles de coalescence et des calculs déjà assez complexes (MCMC, Importance sampling . . .).

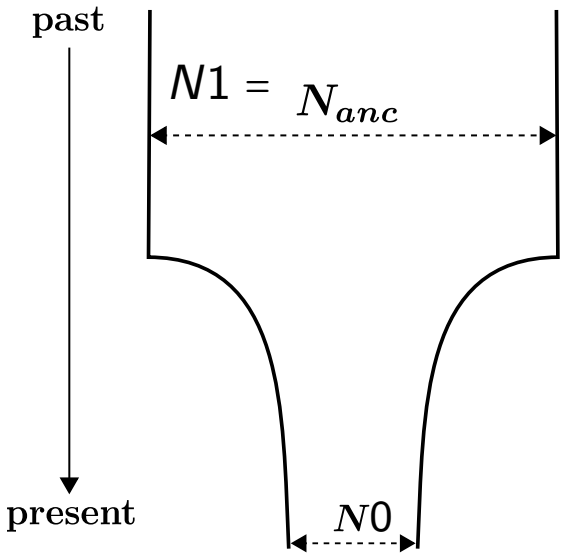
Question typique : Les orangs-outans et la déforestation de Bornéo

- Le génome des orangs-outans a été façonné par son histoire démographique et on sait qu'il y a eu une diminution de la taille de l'habitat
- Goossens et al. 2006 : 200 indvidus échantillonnés sur 9 sites et génotypés à 14 locus microsatellites

- La génétique des populations peut-elle aider à comprendre l'histoire des populations d'orangs-outans ? = caractériser les variations passés de tailles de population

Question typique : Les orangs-outans et la déforestation de Bornéo

- **Modèle démographique**: une population panmictique isolée (WF) avec une variation exponentielle de la taille de population



$\theta_{anc} = 2N_{anc}\mu$

$D = \frac{T}{2N}$

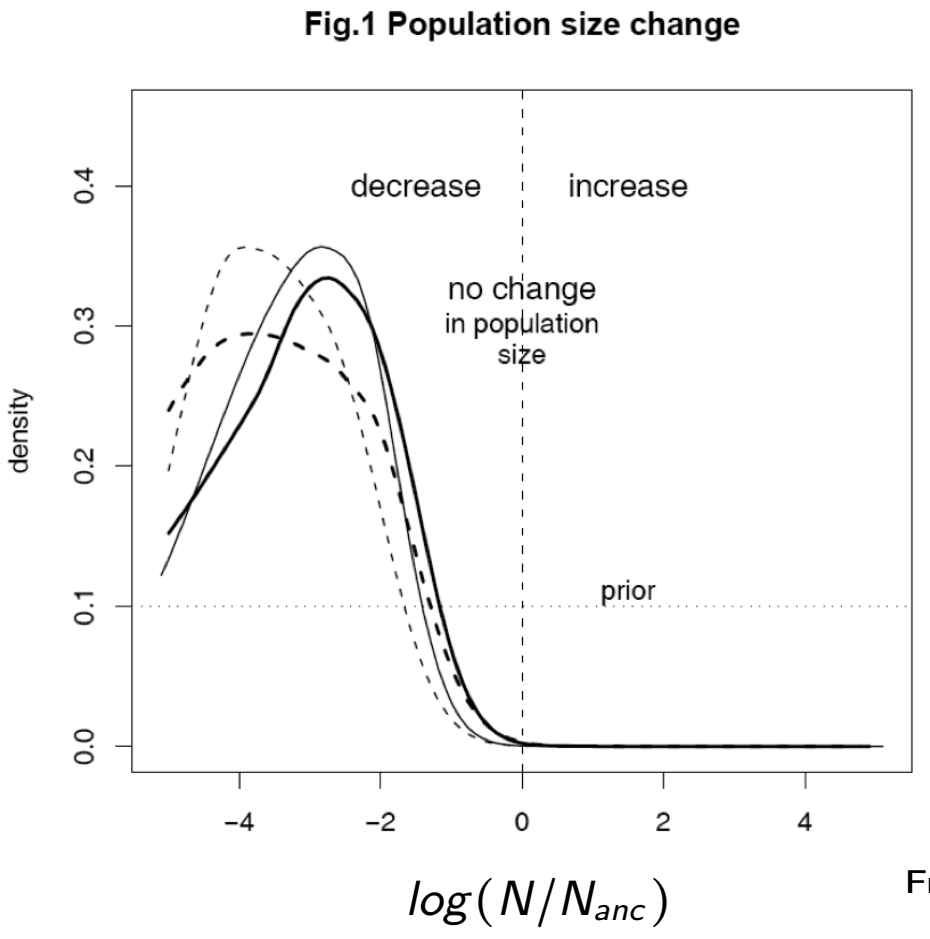
$\theta = 2N\mu$

Contraction (ou expansion) de la population

- `MSVAR` : méthode basée sur la coalescence et un algorithme MCMC pour inférer ces paramètres à partir d'un échantillon génétique unique (microsatellites, mutation par pas)

Question typique : Les orangs-outans et la déforestation de Bornéo

- Résultats de `MSVAR`



→ `MSVAR` permet de détecter efficacement une diminution antérieure de la taille de la population et de la quantifier

From Goossens et al. 2006 PLos Biology

Question typique : Les orangs-outans et la déforestation de Bornéo

- Résultats de `MSVAR`

FE: début de l'exploitation massive des forêts
F: premiers agriculteurs
HG: premiers chasseurs-cueilleurs

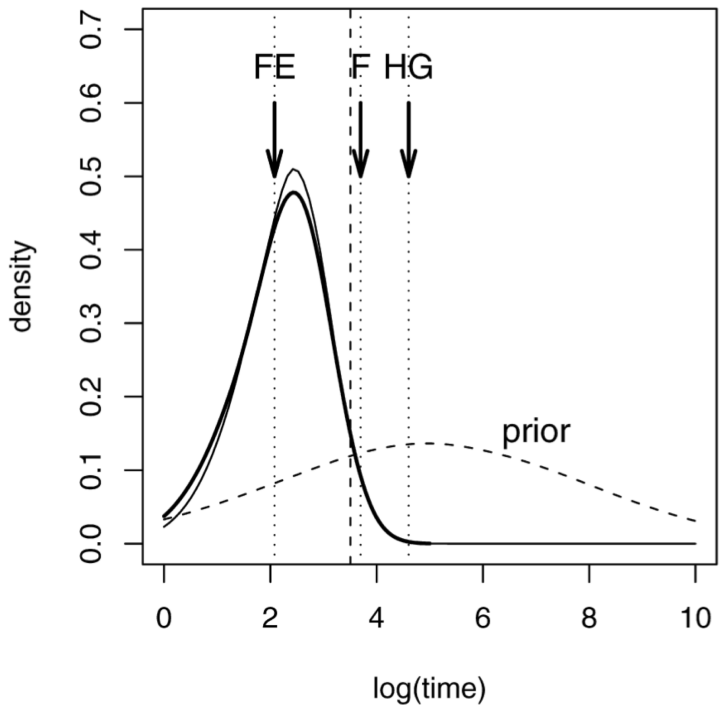


Figure 3. Time since the Population Collapse

→ `MSVAR` permet de détecter et quantifier une réduction passée de taille de population...

...et de dater cet évènement (FE)

From Goossens et al. 2006 PLos Biology

Question typique : Les orangs-outans et la déforestation de Bornéo

- Résultats de `MSVAR`

FE: début de l'exploitation massive des forêts
F: premiers agriculteurs
HG: premiers chasseurs-cueilleurs

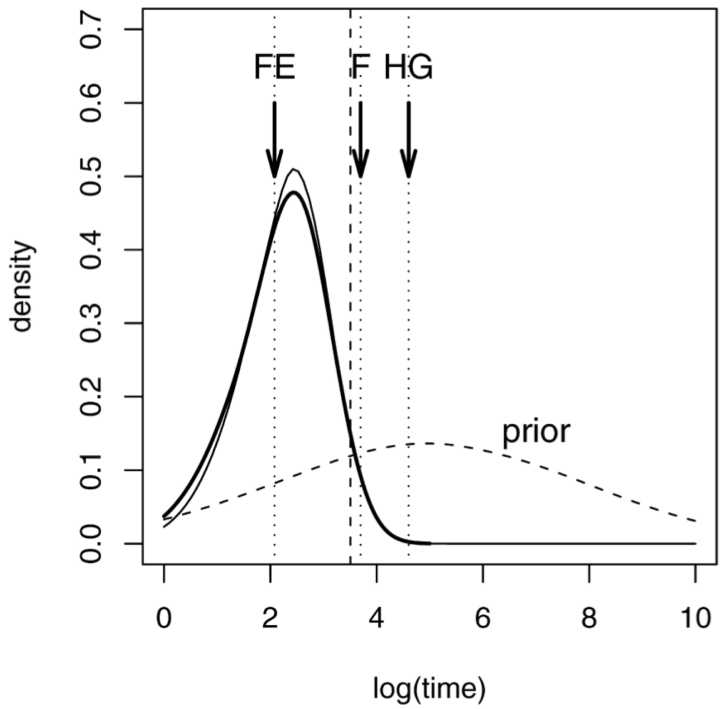


Figure 3. Time since the Population Collapse

→ `MSVAR` permet de détecter et quantifier une réduction passée de taille de population...

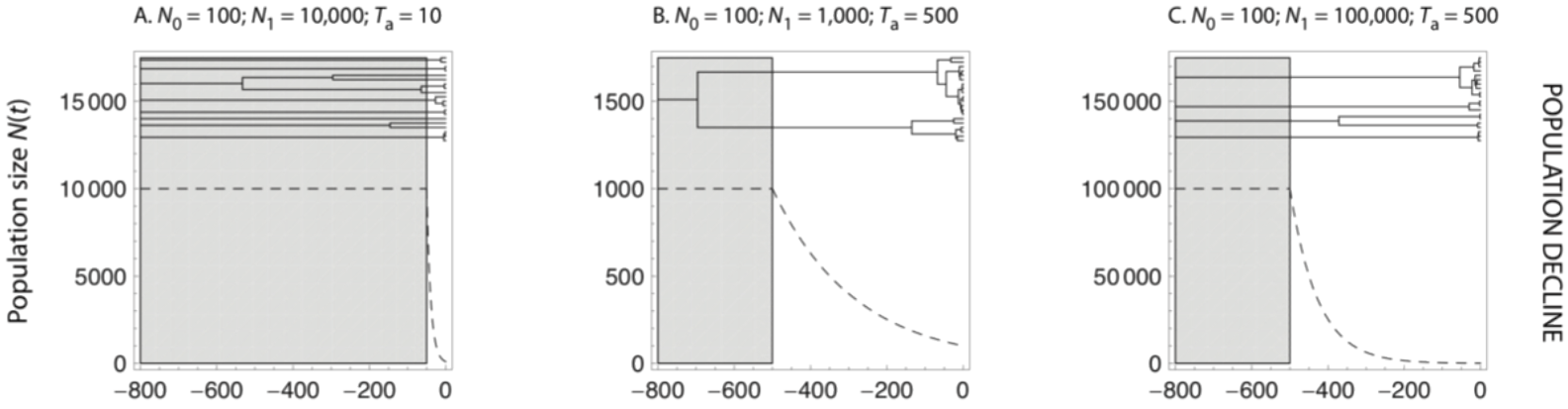
...et de dater cet évènement (FE)

La génétique des populations révèle l'histoire démographique passée à partir d'un unique échantillon actuel

From Goossens et al. 2006 PLos Biology

Intuitions basées sur la coalescence pour comprendre la précision des estimations

- généalogies façonnées par les paramètres démographiques



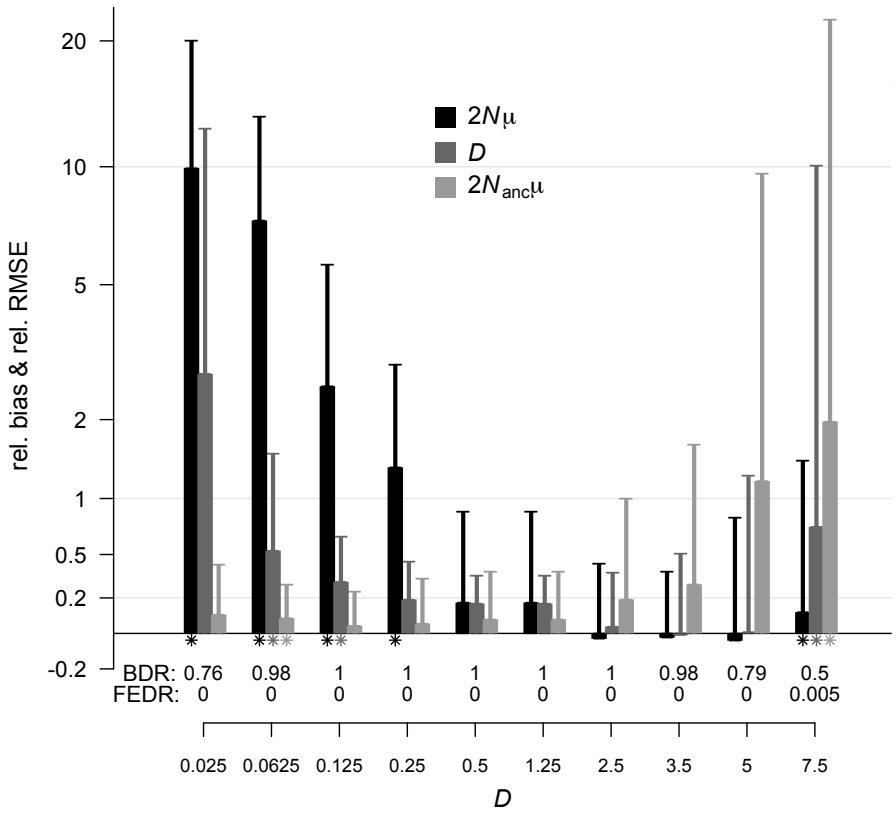
- l'information contenue dans les données dépend du nombre de mutations et d'événements de coalescence au cours des différentes phases démographiques

→ "prédire" la quantité d'informations présentes dans les données

→ "prédire" la précision attendue des estimations

Tests par simulation de la précision attendue

Biais et variance des estimations de $2N\mu$, $2N_{anc}\mu$ et $D = T_g/2N$ (MIGRAINE, Leblois et al. 2014).



Les résultats dépendent fortement de l'ancienneté ($D = T_g/2N$) de la contraction passée.

Bonne précision des estimations des 3 paramètres à condition que la contraction ne soit ni trop récente ni trop ancienne....

Si très récent, alors pas d'information sur la taille actuelle N, ni le temps $D = T_g/2N$

Si très ancien, alors pas d'information sur la taille passé N, ni le temps $D = T_g/2N$

comme prédit par la coalescence

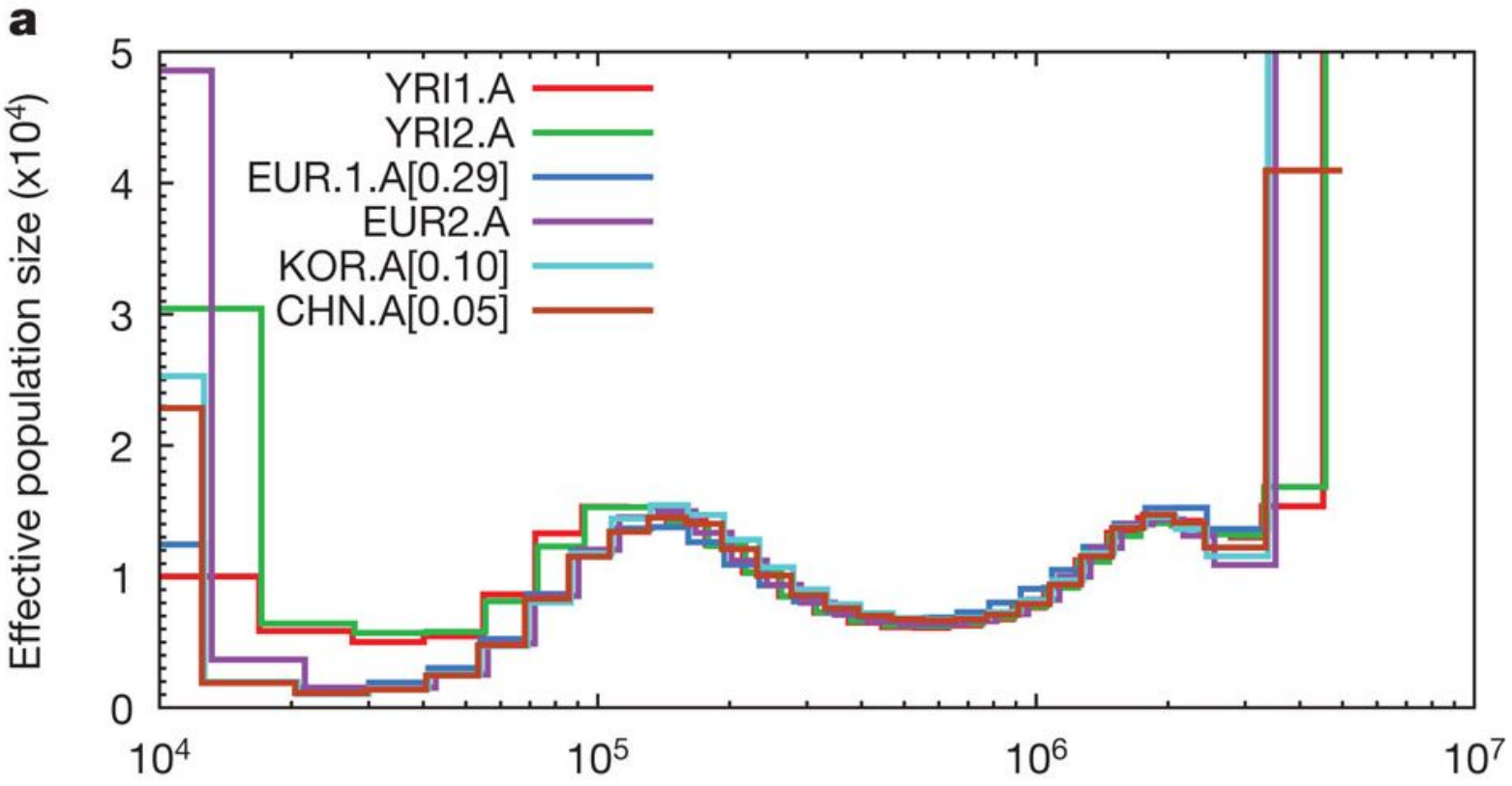
Plan

Une double révolution (2000–2020)

- **Génomique :**
  - Nouveaux types de polymorphismes (Single Nucleotide Polymorphisms, SNP) nouvelles techniques d'acquisition (puces d'hybridation, séquençage NGS).
  - \+ Beaucoup **plus de données** → meilleures estimations, modèles plus complexes.
  - \- Modéliser la recombinaison pour tenir compte de la **corrélation des marqueurs**.
- **Numérique :**
  - Capacités de calcul démultipliées (p.e. clusters).
  - \+ Approches computationnelles (vs théoriques) beacoup plus accessibles.

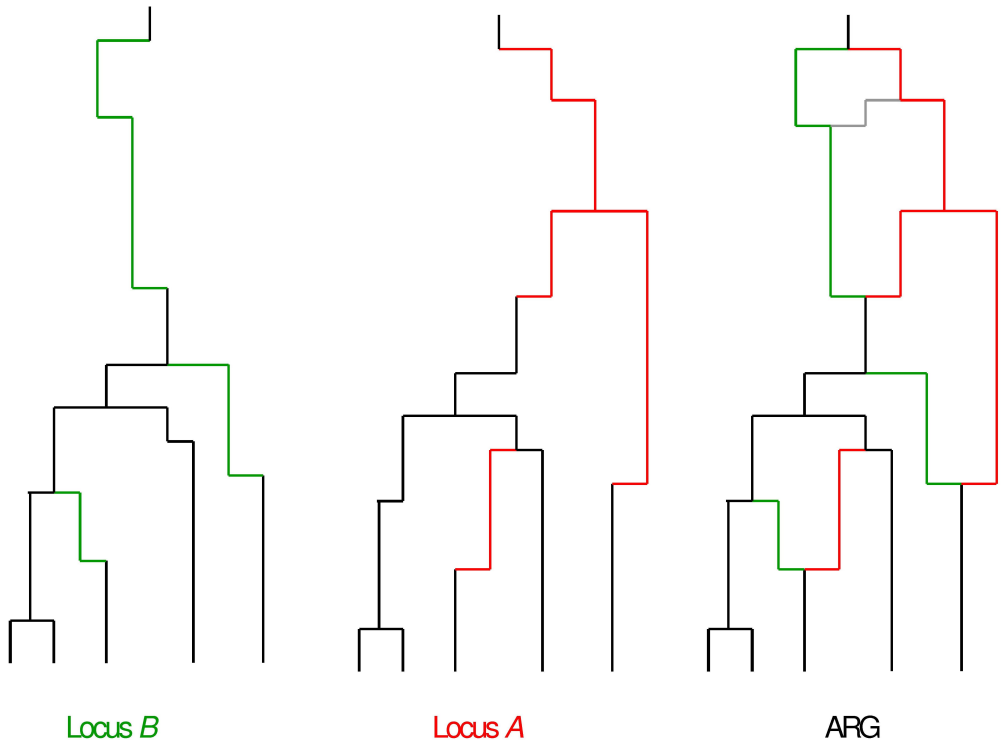
Exemple : la méthode PSMC (Li et Durbin, 2011)

Estime l'évolution passée de la taille efficace au cours du temps à partir du **génome entier d'un (seul) individu diploïde**.



Modéliser la coalescence avec recombinaison

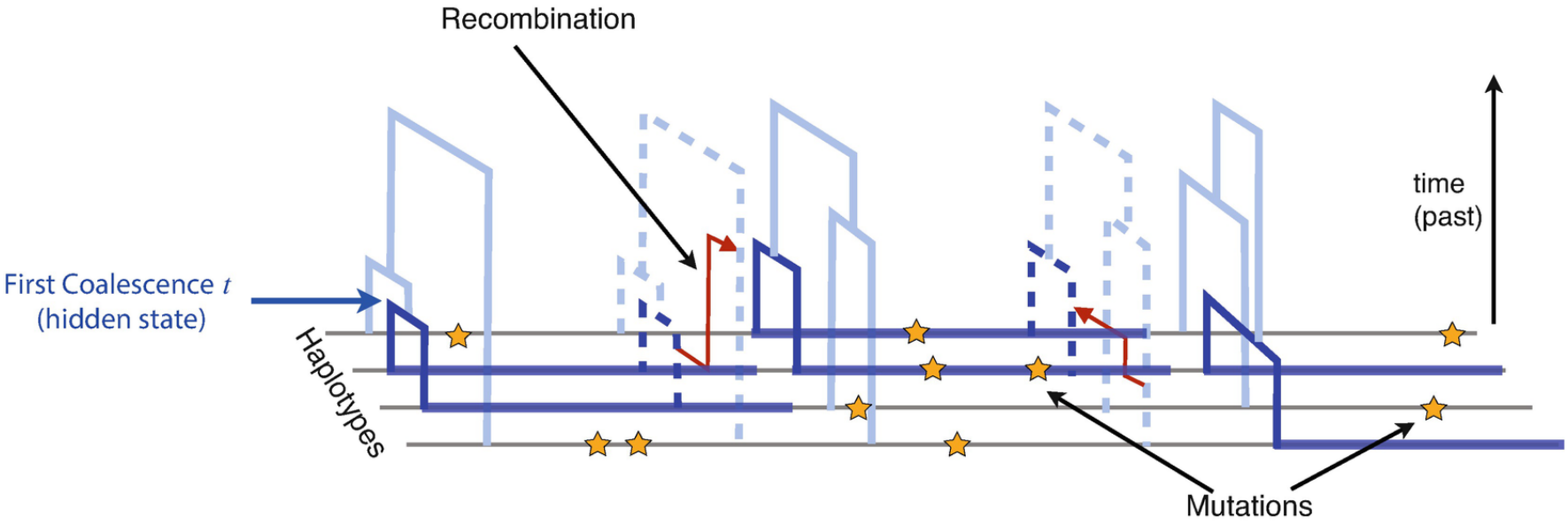
Arbres distincts mais corrélés, résumés dans une structure unique appelée Ancestral Recombination Graph (ARG).



Espace des graphes beaucoup plus complexe à explorer!

La coalescence le long du génome

Arbres distincts pour chaque portion non recombinante du génome mais corrélés le long du génome



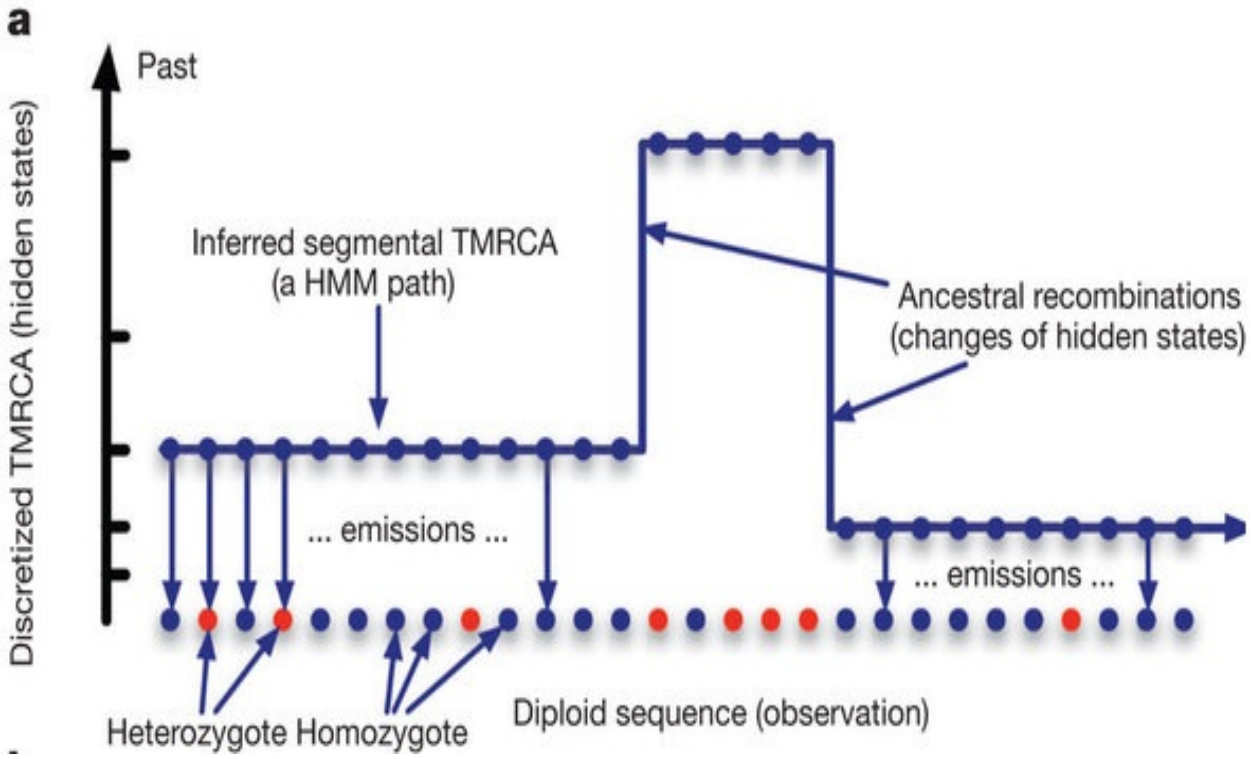
from schiffels & durbin 2014

L'espace des arbres peut être exploré le long du génome

Approche SMC (Sequentially Markovian Coalescent)

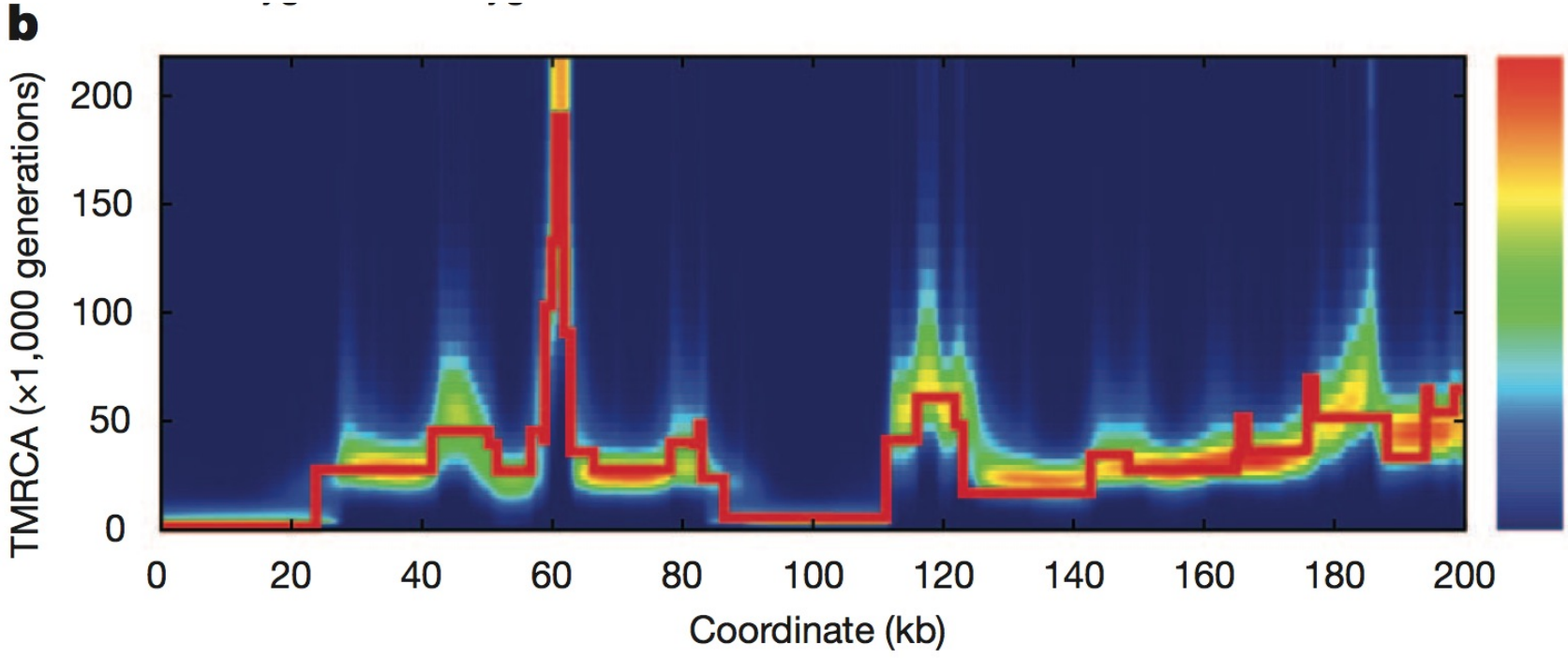
- **Simplifier l'ARG** en le considérant comme une chaîne de Markov le long du génome, qui change d'état suite à des recombinaisons.
- Calcul de la **vraisemblance** sous un modèle approché (HMM, Hidden Markov Model)

Approche SMC (Sequentially Markovian Coalescent)



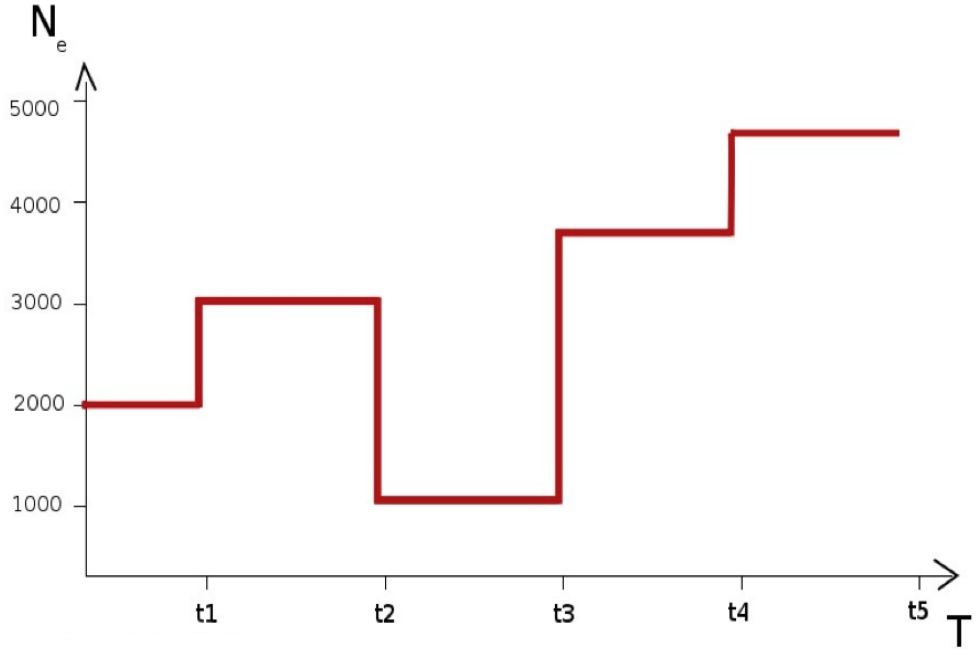
Résultat 1 : les temps de coalescence

Distribution a posteriori du temps de coalescence estimé par PSMC le long du génome



Résultat 2 : les tailles efficaces

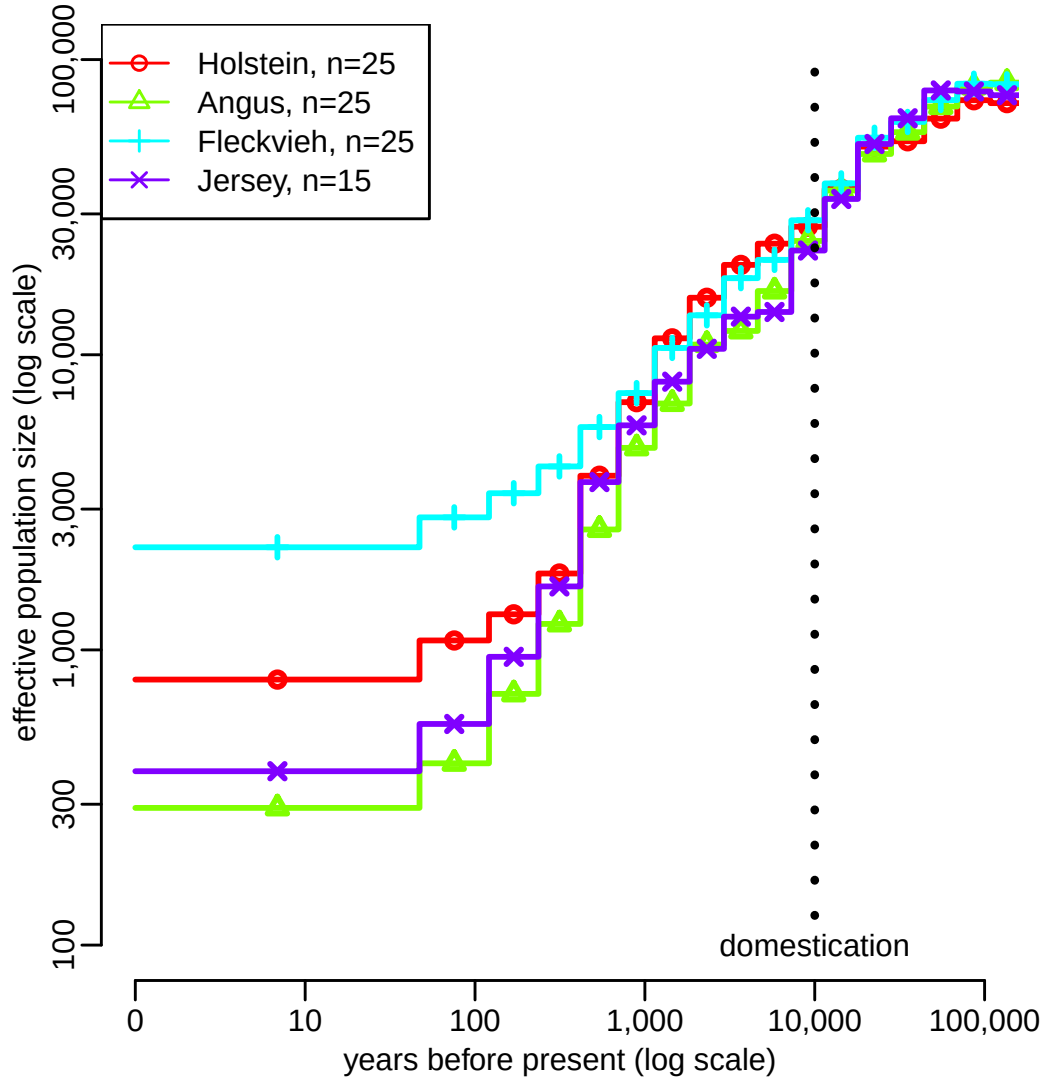
- Taille constante par morceaux avec des temps de changements fixés à l'avance ("skyline" model).



- Trouve la trajectoire de tailles expliquant le mieux la séquence des temps reconstruite (résultat 1).

Approche ABC (Boitard *et al*, 2016)

Inférence par simulation pour éviter le calcul de la vraisemblance.



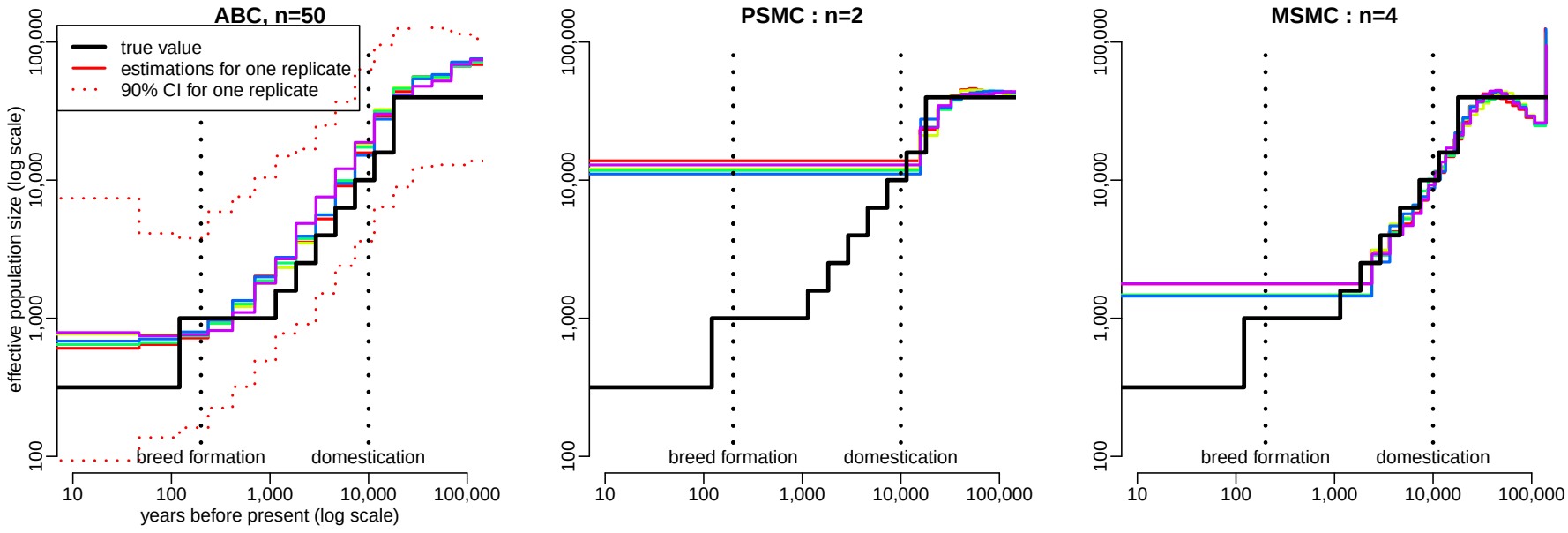
ABC (Beaumont *et al*, 2002)

- ABC = Approximate Bayesian Computation.
- 'Approximate' pour deux raisons :
  - Résume les données par des statistiques $\mathcal{S} = f(D)$
    $\rightarrow$ objectif $L(N()|\mathcal{S})$ et non $L(N()|D)$
  - Procède par simulations intensives.
- Pour un grand nombre de répétitions i:
  1. Tire au hasard $N_i()$ dans une distribution a priori.
  2. Simule un échantillon de génomes sous l'histoire $N_i()$ (ARG).
  3. Calcule les statistiques résumées $\mathcal{S}_i$ pour cet échantillon.
  4. Conserve $N_i()$ si $\mathcal{S}_i$ et $\mathcal{S}$ suffisamment proches.
- Estime $N()$ à partir des $N_i()$ conservés.

Approche ABC (Boitard *et al*, 2016)

- Dépend du choix des statistiques, qui doivent être suffisamment informatives.

\+ Permet d'analyser un grand nombre de séquences, ce qui améliore l'estimation pour les temps récents.



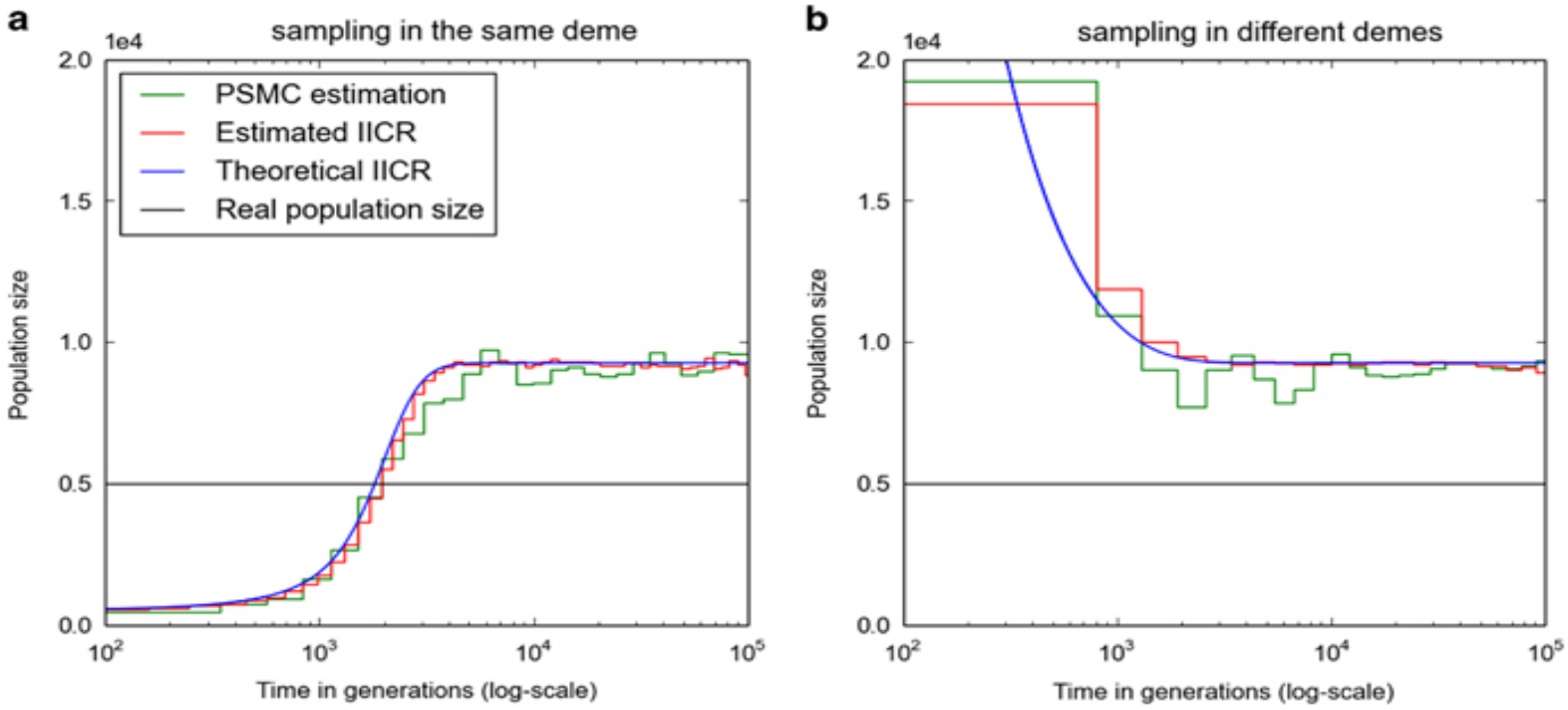
Attention à la robustesse!

- Si on analyse des données avec un modèle qui n'est pas le bon, les résultats obtenus avec ce modèle sont par définition suspects!
- La réalité biologique est complexe et inconnue et l'analyser demande forcément des approximations.

→ Explorer a minima (par simulations) comment un modèle se comporte pour les scénarios alternatifs les plus probables.

Exemple : PSMC et structure

- Vrai modèle : n **populations de taille constante** connectées par des **migrations**.



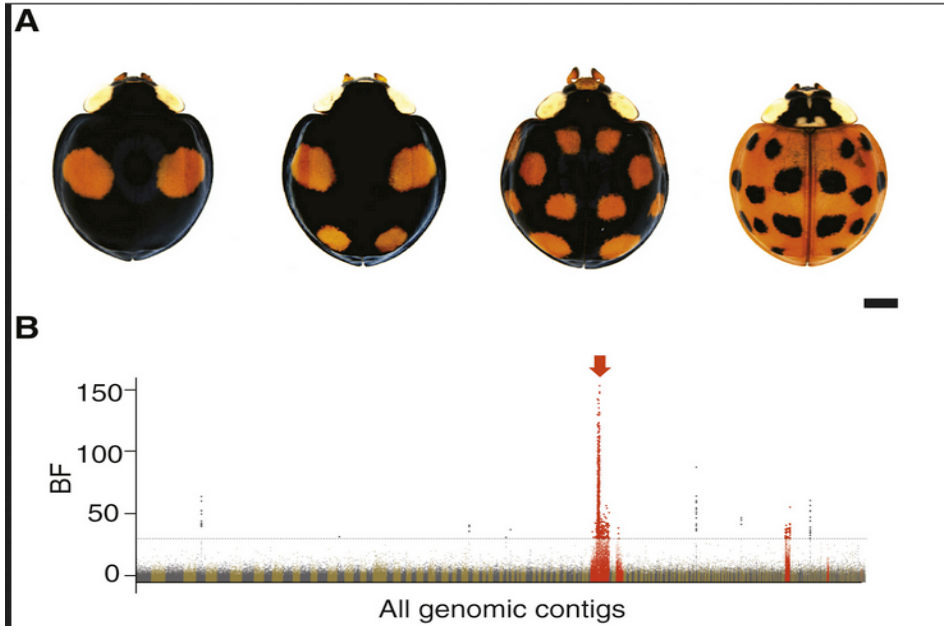
- PSMC estime un **changement de taille inexistant**.
- N'importe quel modèle structuré conduira à un problème de ce type (à des degrés différents).

Apport des données génomiques haut débit

- Reconstruction plus précise de l'histoire démographique des populations : modèles plus complexes, précision accrue.

Apport des données génomiques haut débit

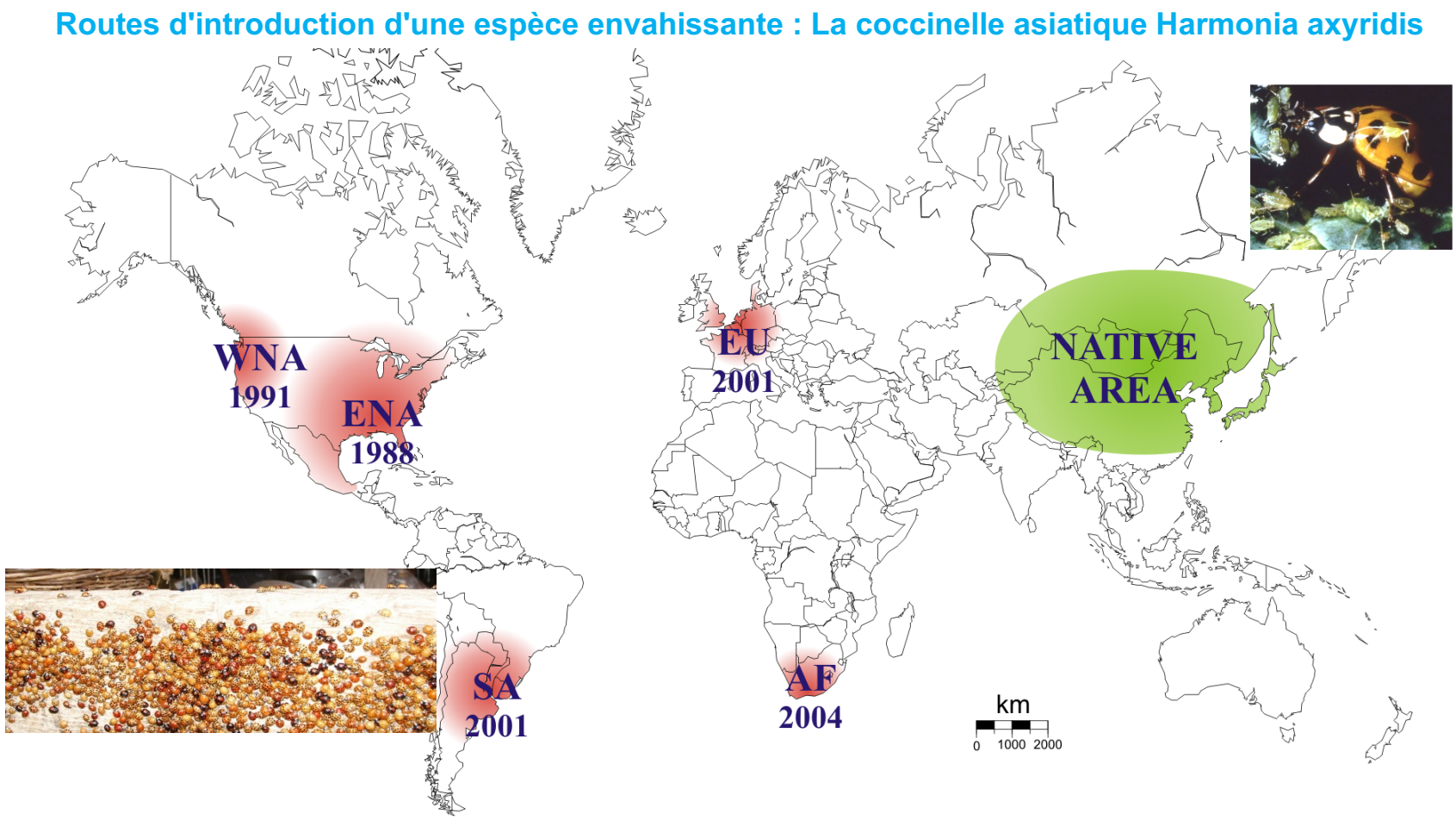
- Explorer la spécificité de chaque région du génome :
  - Variations de la recombinaison, de la mutation ...
  - Variants adaptatifs, QTL (Quantitative Trait Loci)
  - Ex : déterminisme de la coloration chez la coccinelle *Harmonia axyridis* (Gautier *et al*, 2018)



Plan

Example plus complexe : inférence des routes d'invasions

- Inférences par simulation (ABC) pour reconstruire les routes d'invasion de la coccinelle asiatique à partir d'échantillons de multiples populations



Work of Arnaud Estoup, CBGP

Contexte

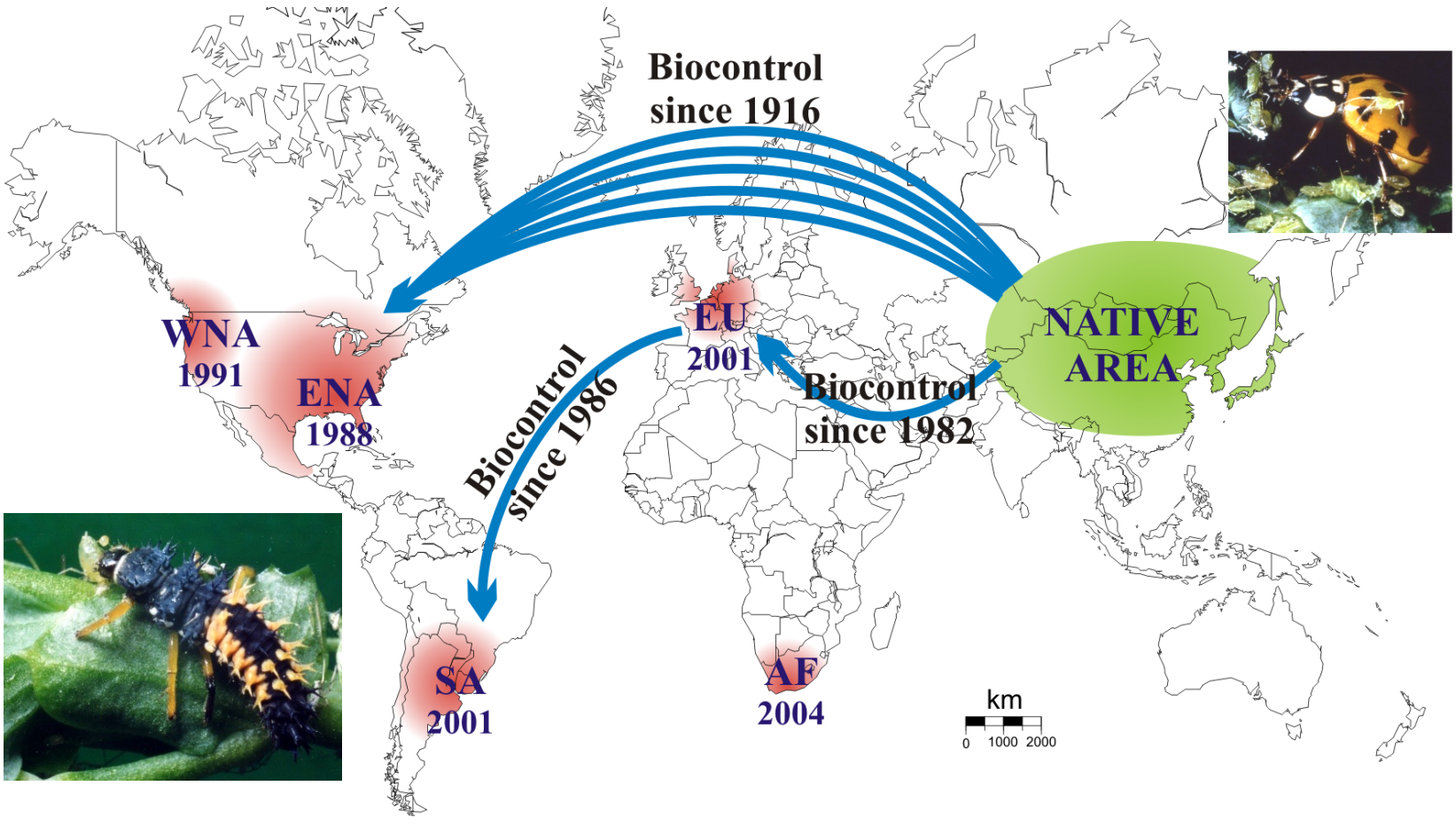
Pourquoi reconstruire les routes d'invasion

- Définir et tester des hypothèses concernant les facteurs évolutifs et environnementaux agissant lors des invasions biologiques
- Faciliter les stratégies de contrôle ou de prévention
  - surveillance et quarantaine ciblant les principales zones sources et les moyens de dispersion
  - souches d'agents de lutte biologique de même origine géographique que la population envahissante

Contexte

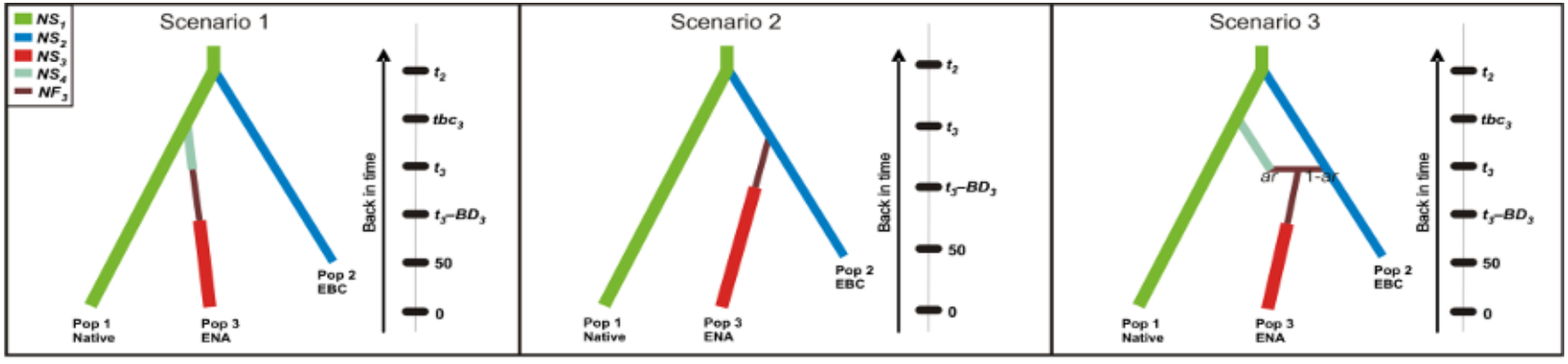
Harmonia axyridis : un bon modèle pour étudier les routes d'invasion

Données historiques sur le biocontrôle



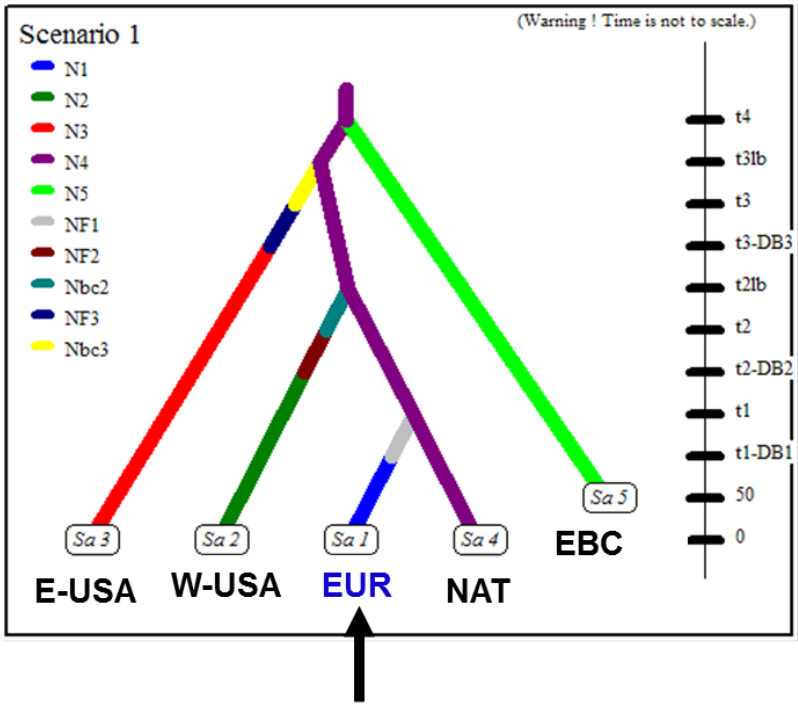
Contexte

- Les populations naturelles ont des histoires démographiques complexes : leurs tailles varient au cours du temps, mais aussi parfois leurs aires de répartition, entraînant des processus de fission (divergences) et de fusion (admixture) qui laissent des signatures sur leur composition génétique.

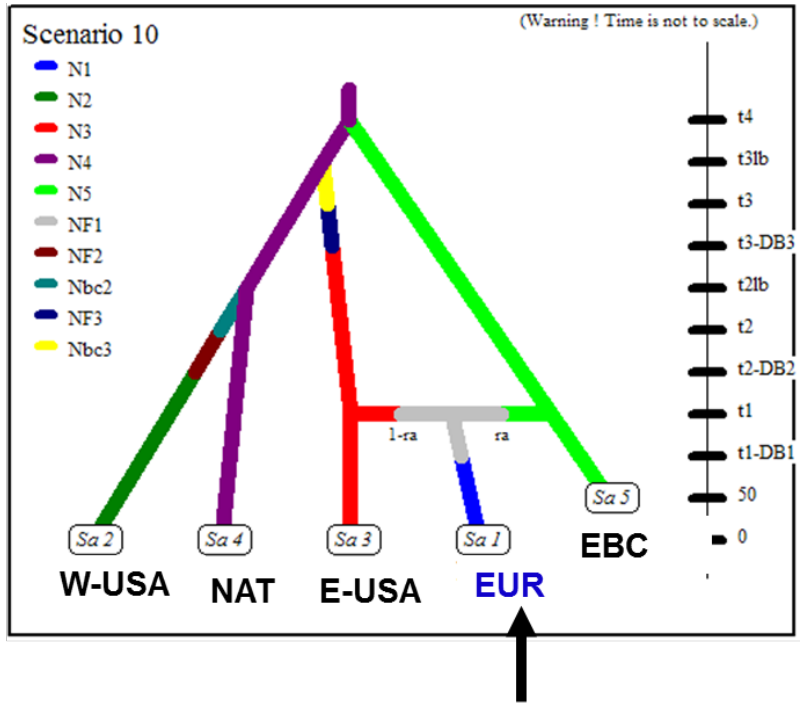


Inférence des routes d'invasion

- inférer le scenario le plus vraisemblable
- inférer les paramètres sous ce scénario



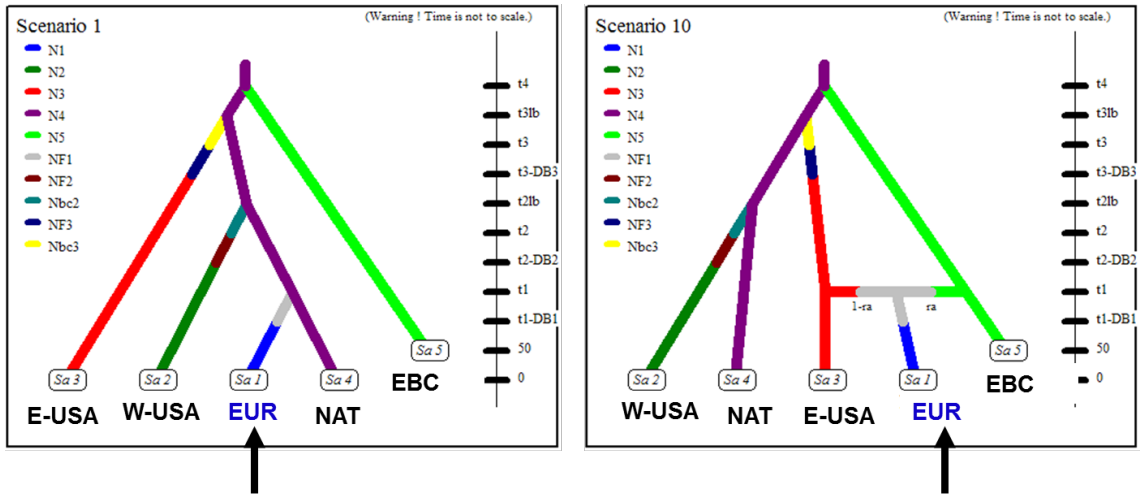
Origin: native area



Origin: admixture E-USA + EBC

Inférence des routes d'invasion

- inférer le scenario le plus vraisemblable
- inférer les paramètres sous ce scénario



Origin: native area

Origin: admixture E-USA + EBC

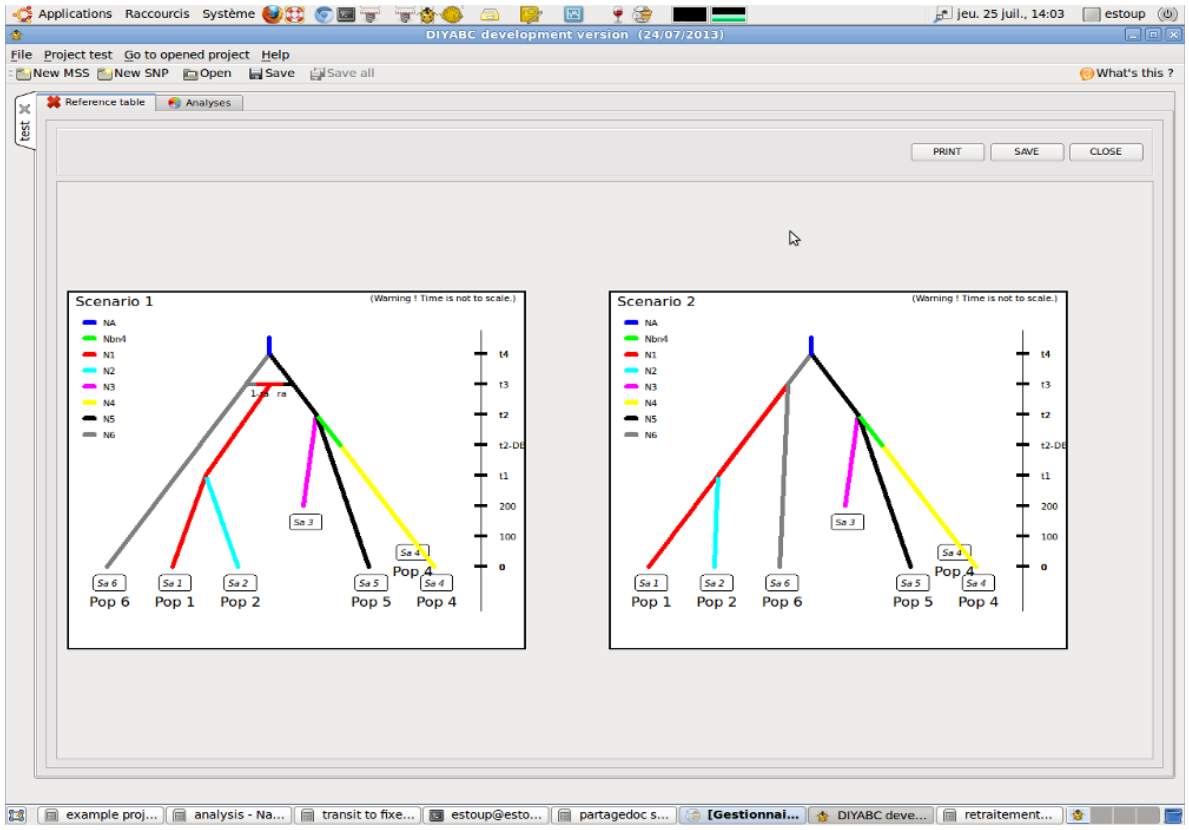
1

La vraisemblance n'est pas calculable pour ces modèles avec divergences et admixtures

→ Inférence par simulation (ABC)

Inférence des routes d'invasion

Logiciel Do It Yourself ABC (DIY-ABC, CBGP, Estoup et al)



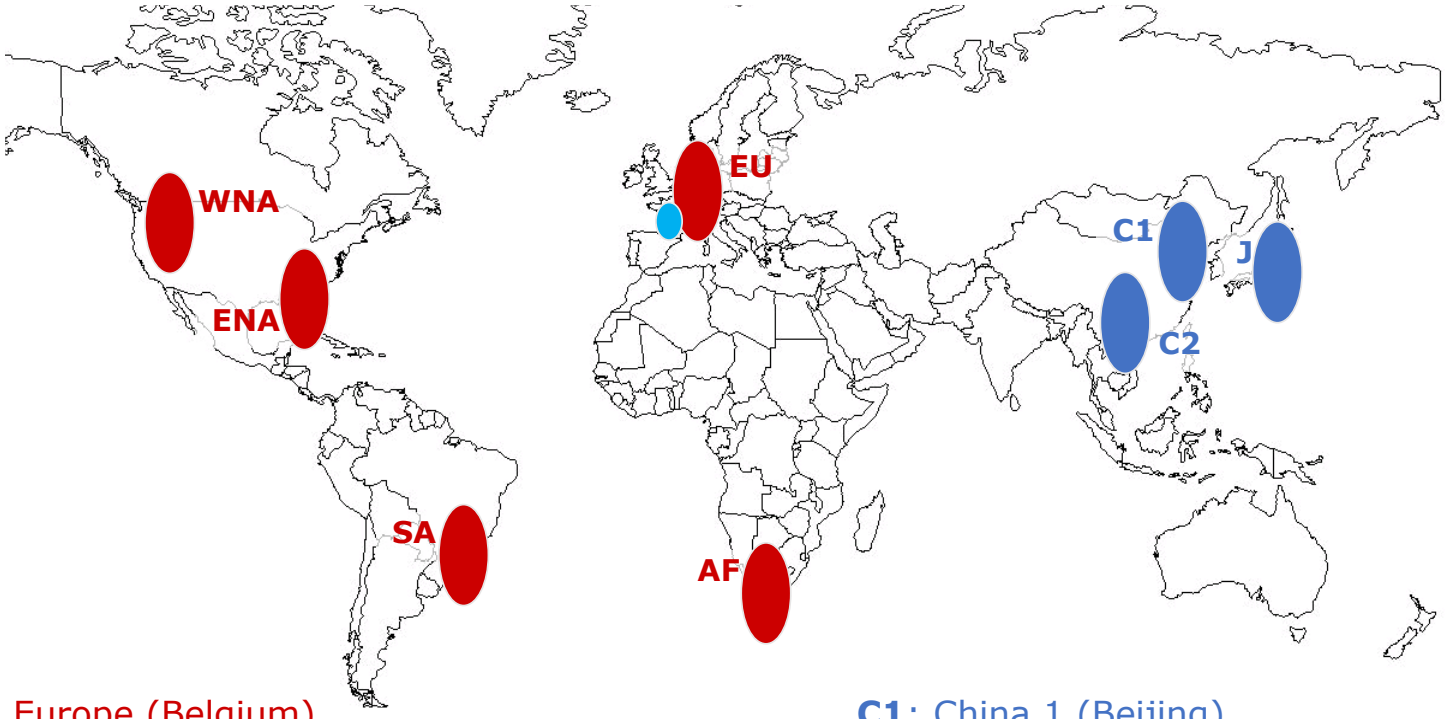
Inférence des routes d'invasion

Logiciel Do It Yourself ABC (DIY-ABC, CBGP, Estoup et al)

- simulation par coalescence
- calcul des statistique résumantes
  - Nbre d'allele, Hétérosygotie,
  - F_{ST}, distances génétiques,
- inférence par ABC

Inférence des routes d'invasion

Plus de 300 échantillons collectés dans 8 localités
+ souche biocontrole INRAE (2010)



EU: Europe (Belgium)
ENA: Eastern North America (Louisiana)
WNA: Western North America (Washington)
SA: South America (Brazil)
AF: Africa (South Africa)

(18 < sample size < 42)

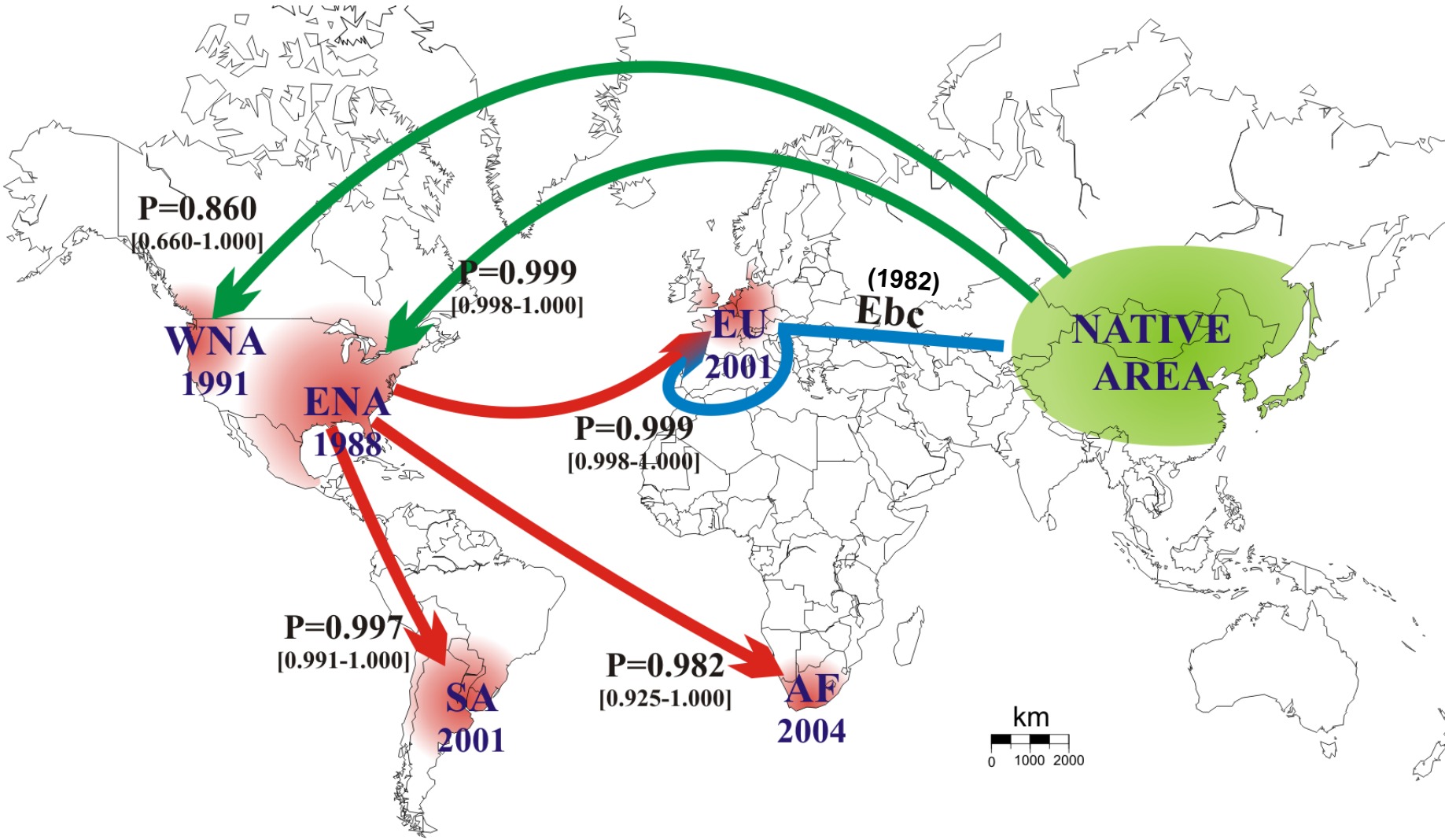
C1: China 1 (Beijing)
C2: China 3 (Yunnan province)
J: Japan

Native (West)

+ EBC: European Biocontrol Strain (INRA 1987)

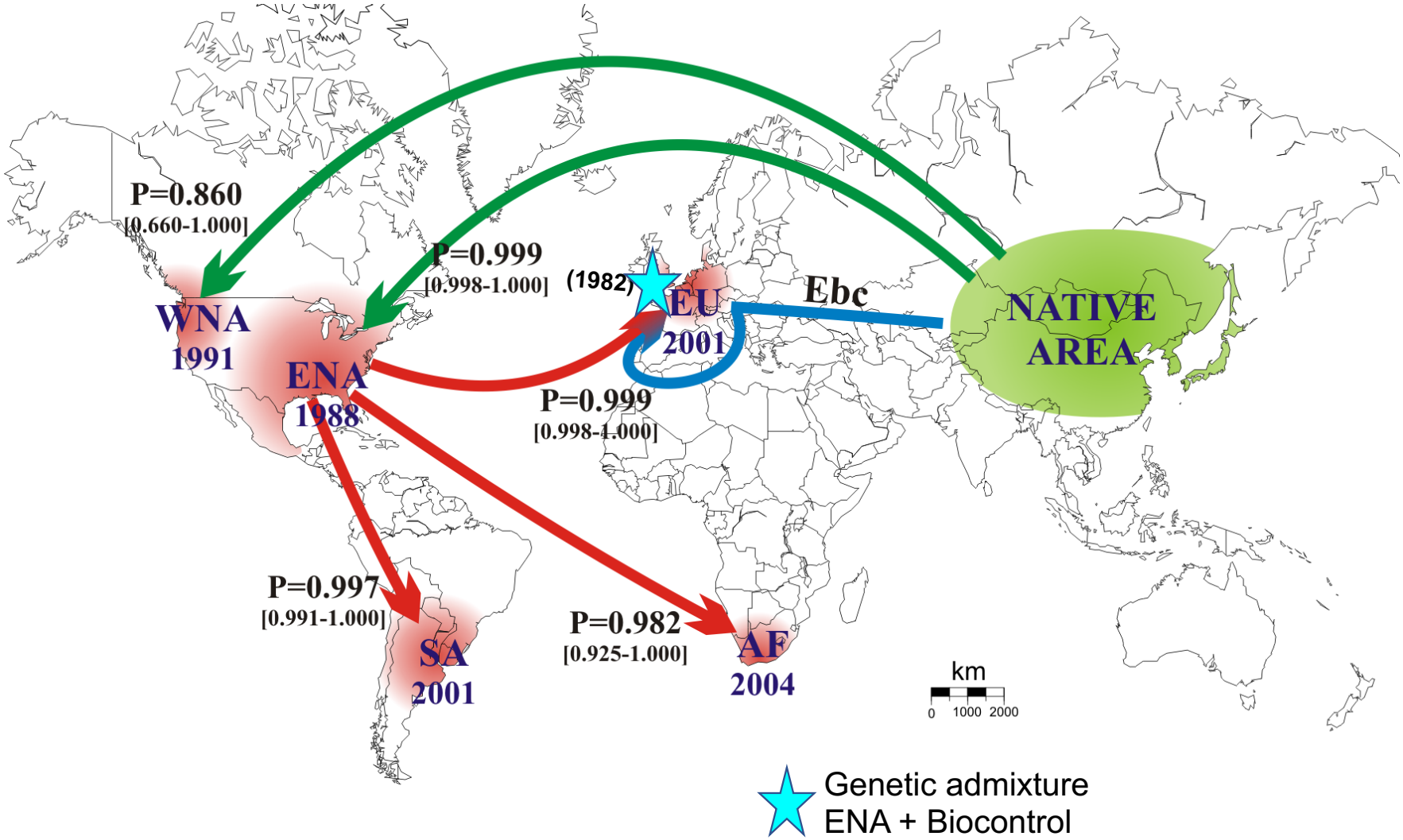
Inférence des routes d'invasion

Routes d'invasion d'Harmonia axyridis à partir de 9 sites échantillonnés



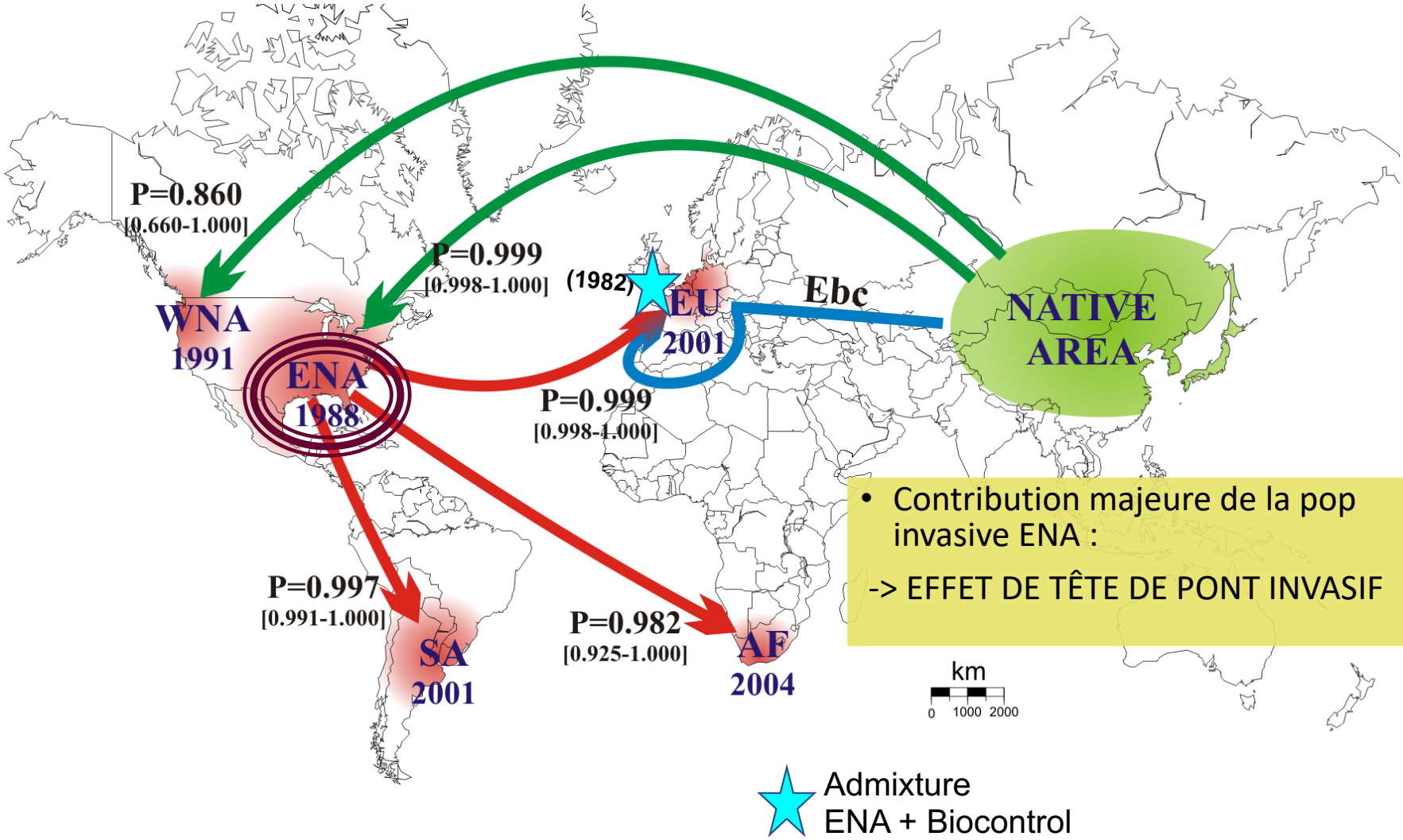
Inférence des routes d'invasion

Routes d'invasion d'Harmonia axyridis à partir de 9 sites échantillonnés



Inférence des routes d'invasion

Routes d'invasion d'Harmonia axyridis à partir de 9 sites échantillonnés



- Contribution majeure de la pop invasive ENA :

-> EFFET DE TÊTE DE PONT INVASIF

Plan

Conclusions

- Données génomiques = **information très riche** sur l'histoire évolutive des espèces.
- **Intérêt théorique** (mécanismes évolutifs) mais aussi **pratique** (gestion des populations en agriculture, conservation . . .).
- Discipline de plus en plus technique et computationnelle, influence croissante de l'IA.

- MAIS il est **important** de:
  - Continuer à **connaître les classiques** (He, Fst, WF, coalescence, diffusion . . .) pour garder une intuition sur les données et le sens des analyses plus complexes.
  - Garder en tête qu'un modèle est toujours imparfait et toujours **questionner la robustesse des résultats**.