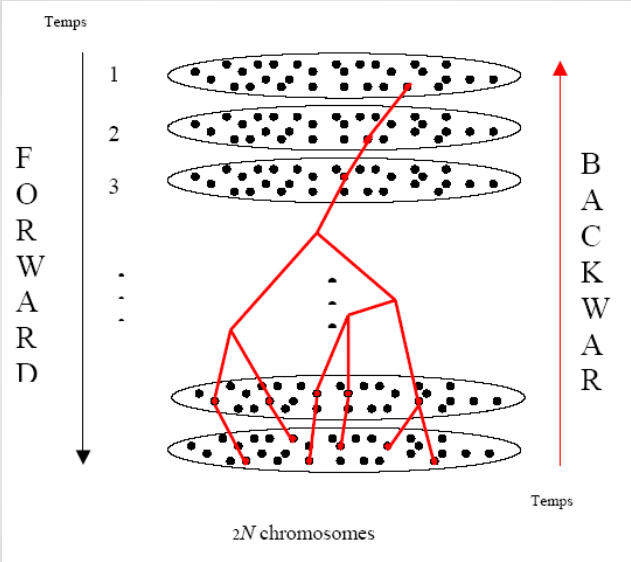
La théorie de la coalescence et ses applications



Raphael Leblois
Centre de Biologie et de Gestion des Populations , CBGP
INRA, Montpellier

Cours M1, ENS Lyon, Novembre 2010

Plan du cours

1. Principes de la théorie de la coalescence
2. Simulations d'arbres et de données : algorithmes et applications directes
3. Coalescence et inférence de paramètres démo-génétiques
4. Quelques exemples d'applications

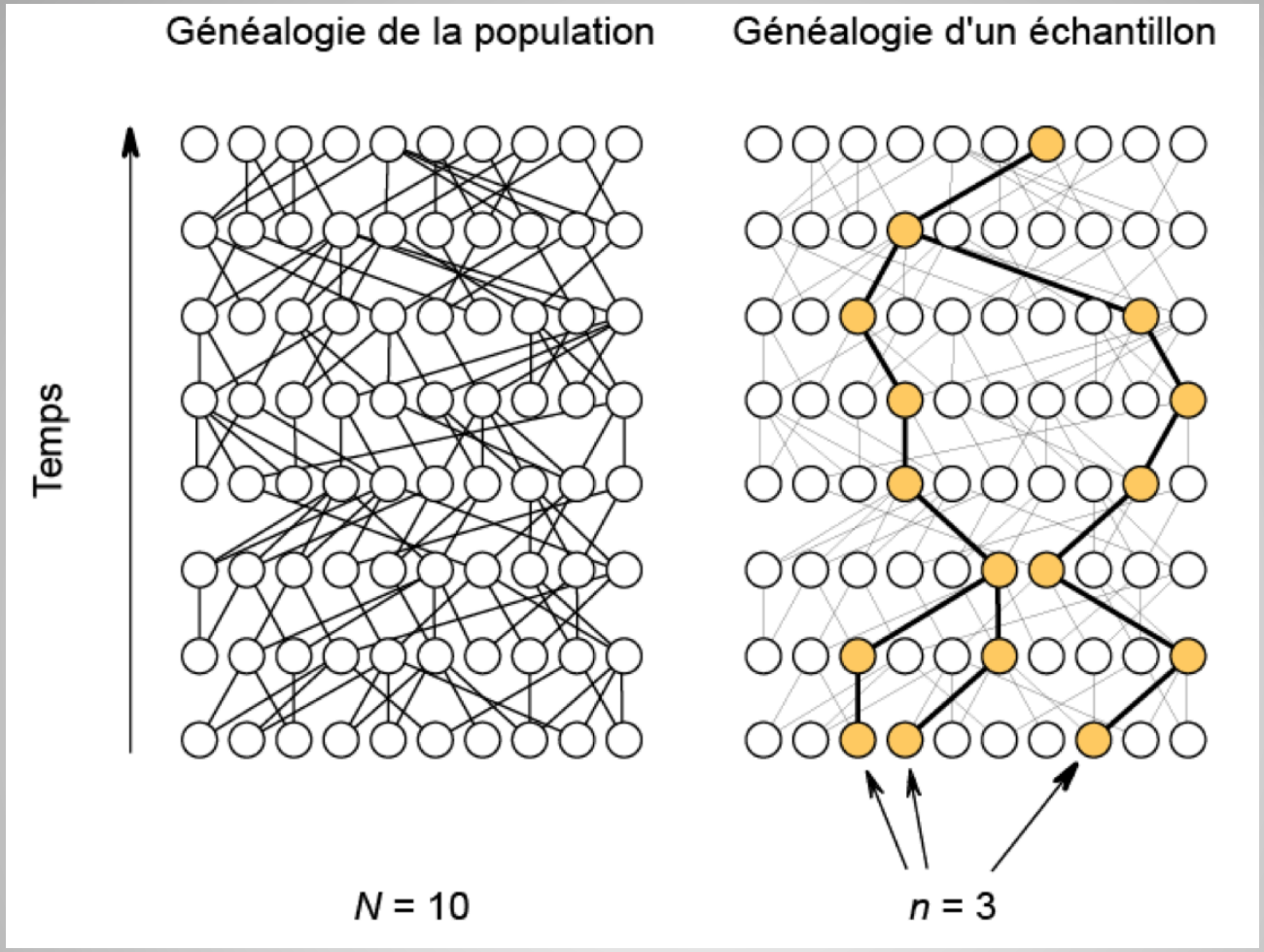
Origine de la théorie de la coalescence

- 1974 –1982 gestation (Kingman, Ewens, Watterson)

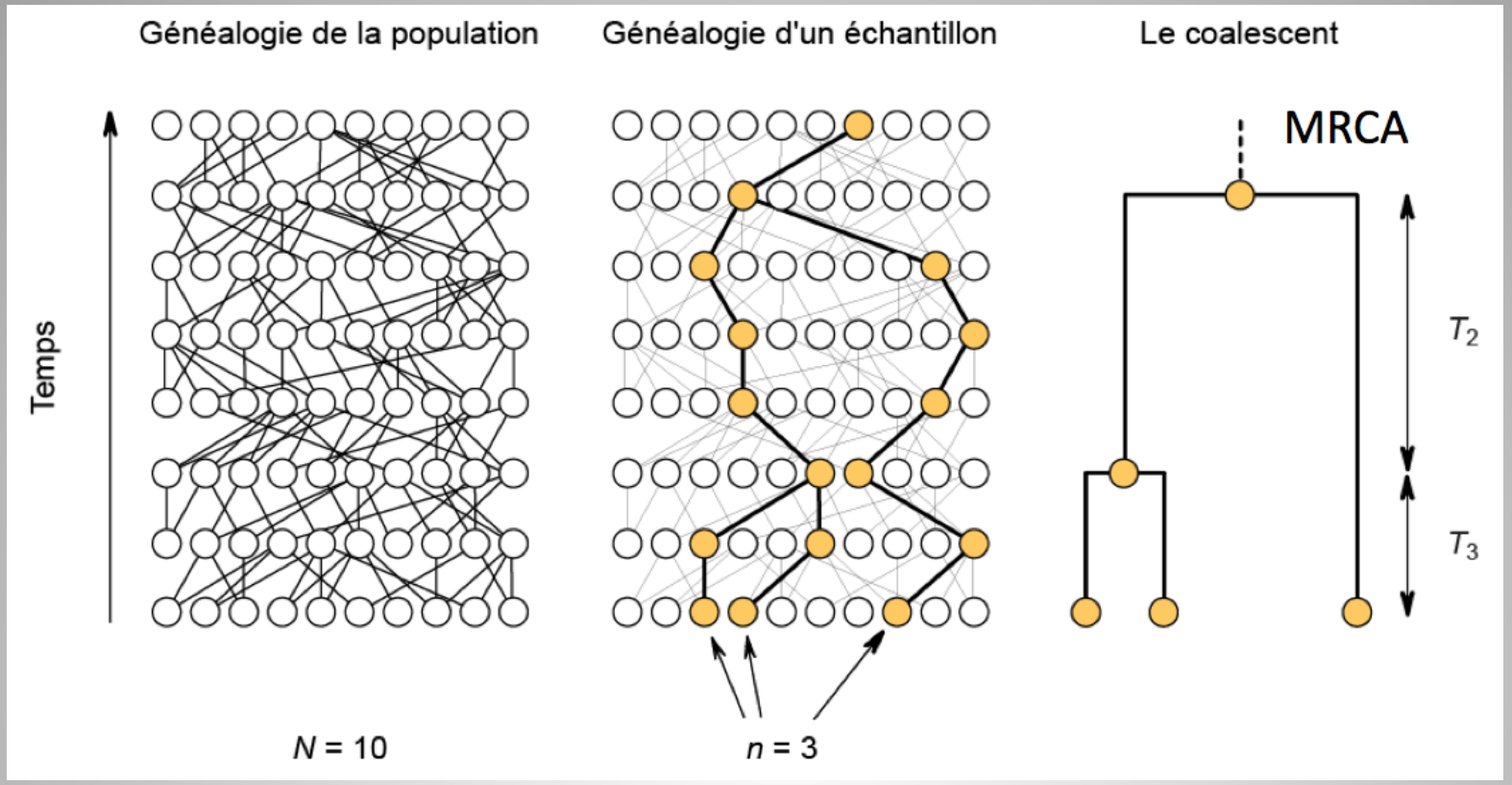
- 1982 Kingman & Tajima

- depuis 1990, nombreux développements

  par Griffiths, Tavaré, Hudson, Donnelly, Felsenstein, et beaucoup d'autres…



La coalescence s'intéresse à la généalogie d'un échantillon de gènes en remontant le temps jusqu'à l'ancêtre commun de l'échantillon



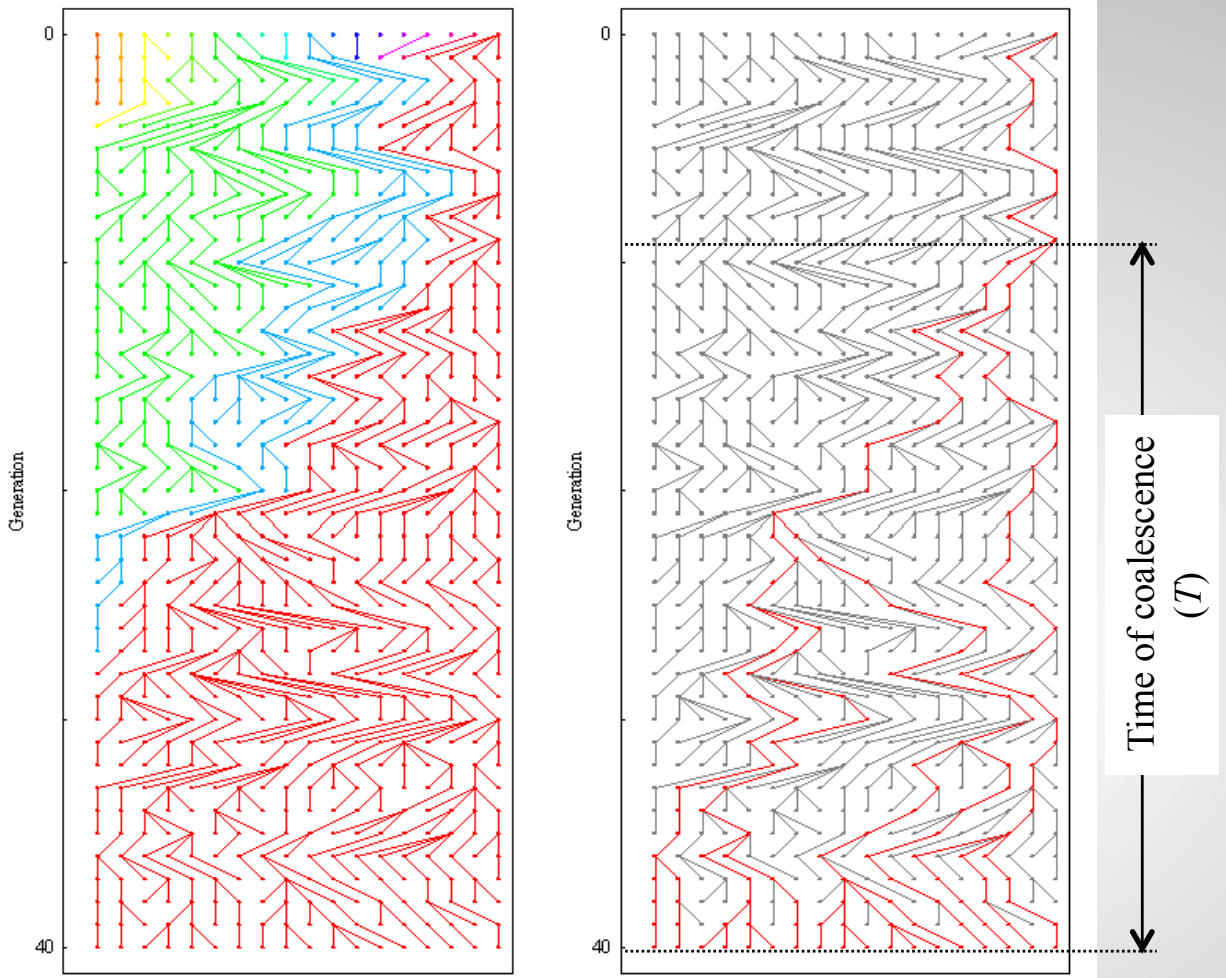
→ **Nouvelle approche de génétique des populations :**

✓ Approche classique

- POPULATION
- Fréquences
- Vision avant (Forward)

✓ Approche « coalescence »

- ECHANTILLON
- Généalogie des gènes
- Vision arrière (Backward)

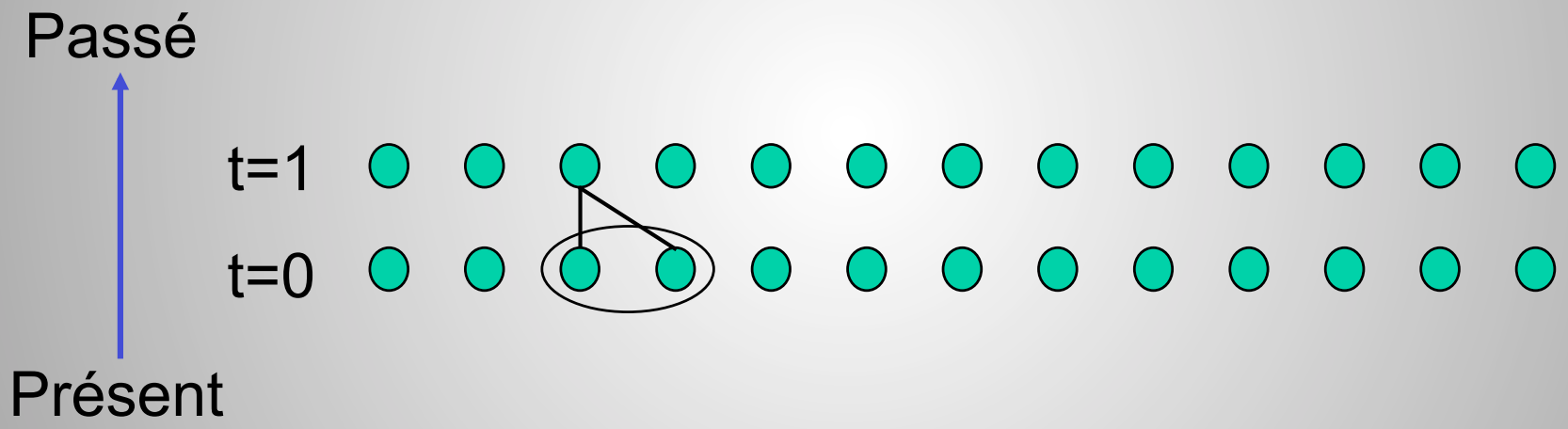


Les différentes lignées fusionnent (coalescent) au fur et à mesure que l'on remonte vers le passé

Modélisation du processus de dérive génétique en "remontant dans le temps" jusqu'à l'ancêtre commun d'un échantillon de gènes

Coalescence en une génération de 2 lignées

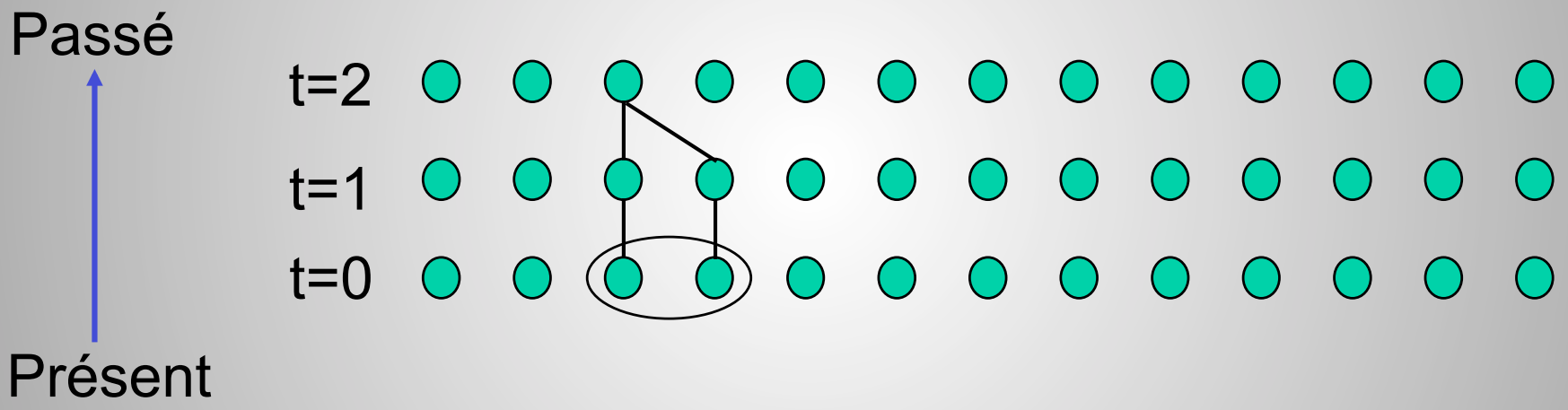
dans une population haploïde de taille N



$$P(T_2 = 1) = \frac{1}{N}$$

Coalescence en 2 générations de 2 lignées

dans une population haploïde de taille N

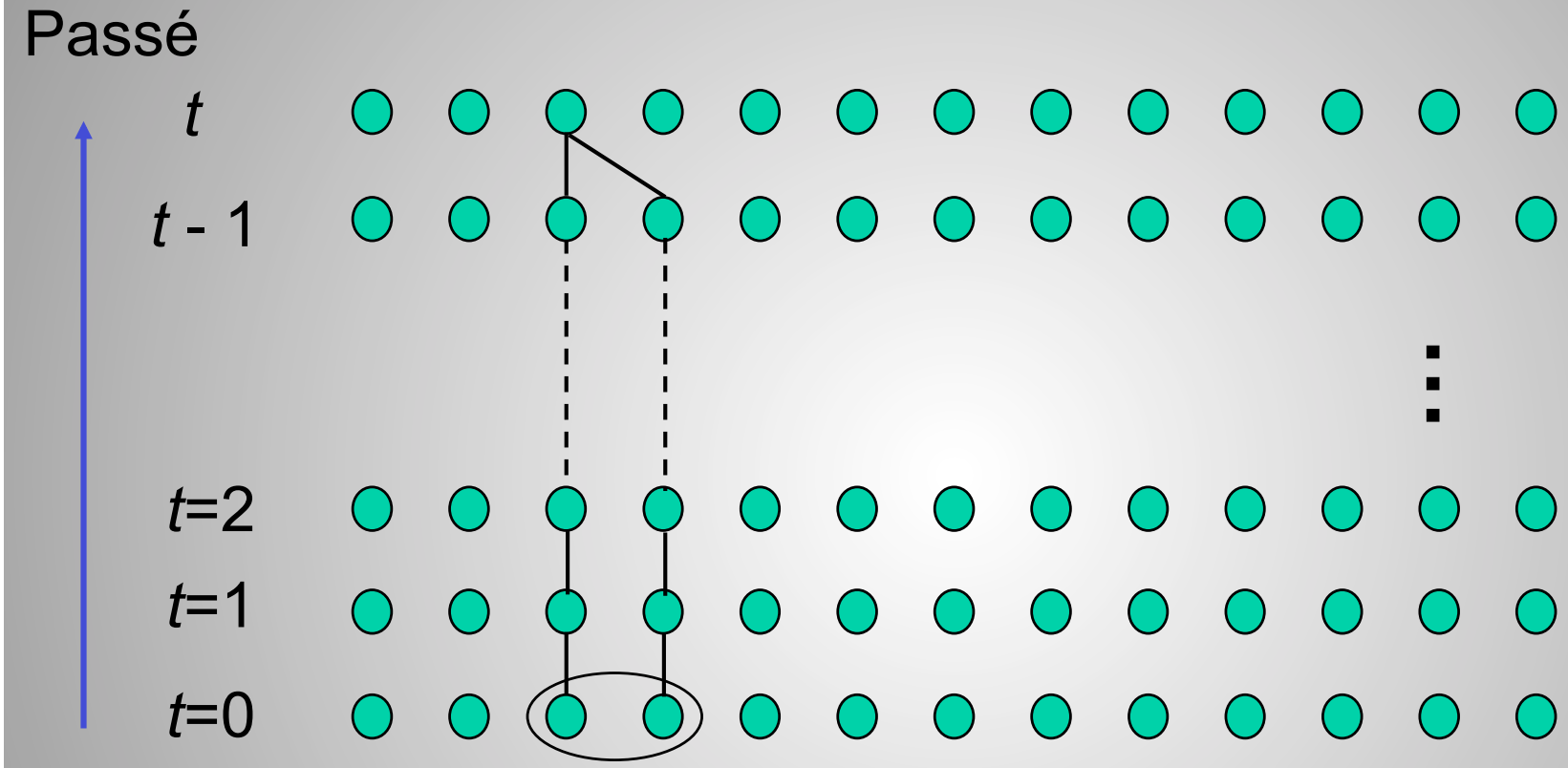


(Proba de ne pas avoir coalescé à t=1)*(proba de coalescer à t=2)

$$P(T_2 = 2) = \left(1 - \frac{1}{N}\right)\frac{1}{N}$$

Coalescence en *t* générations de 2 lignées

dans une population haploïde de taille N



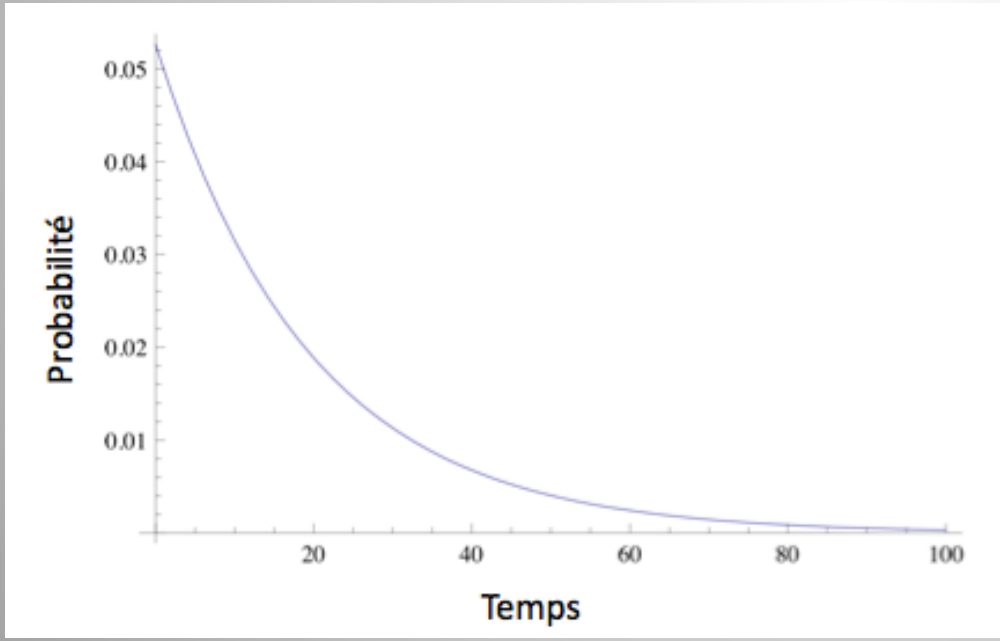
(proba de pas coalescer en t-1 générations)*(proba de coalescer à t)

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$

Coalescence en *t* générations de 2 lignées

dans une population haploïde de taille N

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$



C'est une loi géométrique de paramètre 1/*N*

Coalescence en *t* générations de 2 lignées

dans une population haploïde de taille N

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$

On peut montrer que l'espérance du temps de coalescence de 2 lignées est :

$$E[t] = \sum_{t=0}^{\infty} t * P[t] = N$$

Il faut donc remonter, en moyenne, *N* générations pour trouver l'ancêtre commun de 2 gènes

Raisonnement intuitif : il y a une chance sur 6 pour faire un 4 aux dés

➔ il faut en moyenne 6 coups de dés pour faire un 4.

Coalescence en *t* générations de 2 lignées

dans une population haploïde de taille N

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N}$$

On peut montrer que l'espérance du temps de coalescence de 2 lignées est :

$$E[t] = \sum_{t=0}^{\infty} t * P[t = 2] = N$$

Il faut donc remonter, en moyenne, N générations pour trouver l'ancêtre commun de 2 gènes. Mais il y a une très forte variance :

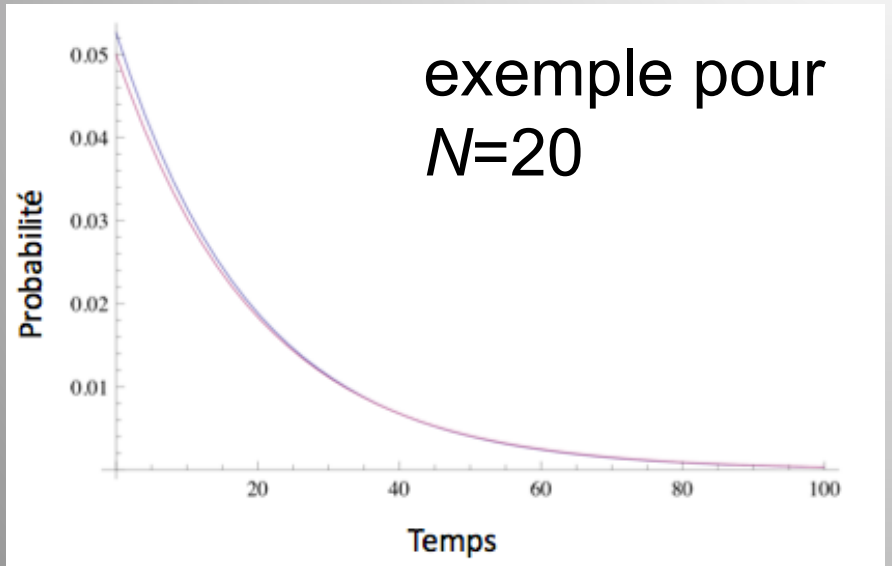
$$V[t] = N * (N - 1) \approx N^2$$

Coalescence en *t* générations de 2 lignées

dans une population haploïde de taille N

Quand $x << 1$, on a $(1-x)^t \approx e^{-xt}$
on peut donc approximer la loi géométrique (discrète) par une loi exponentielle (continue) quand *N* est grand

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \approx \frac{1}{N} e^{-Nt}$$



exemple pour *N*=20

On a donc une approximation continue d'un processus discontinu et cette approximation est très robuste

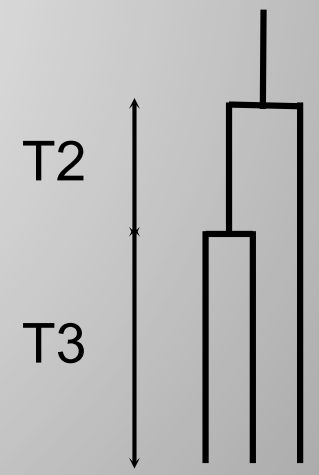
Coalescence en *t* générations de 2 lignées

dans une population haploïde de taille N

Quand $x << 1$, on a $(1-x)^t \approx e^{-xt}$
on peut donc approximer la loi géométrique (discrète) par une loi exponentielle (continue) quand *N* est grand

$$P(T_2 = t) = \left(1 - \frac{1}{N}\right)^{t-1} \frac{1}{N} \approx \frac{1}{N} e^{-Nt}$$

Le temps de coalescence de deux lignées (longueur des branches) suit une loi de distribution exponentielle d'espérance *N*



Coalescence en *t* générations de *j* lignées

dans une population haploïde de taille N

HYPOTHESE : pas de coalescence multiple quand *N* est grand.

$C_j^2 = j*(j - 1)/2$ paires de lignées peuvent coalescer avec Pr =1/*N*

$$\Pr(\text{2 lignées parmi } j \text{ coalescent à chaque génération}) = \frac{j(j-1)}{2N}$$

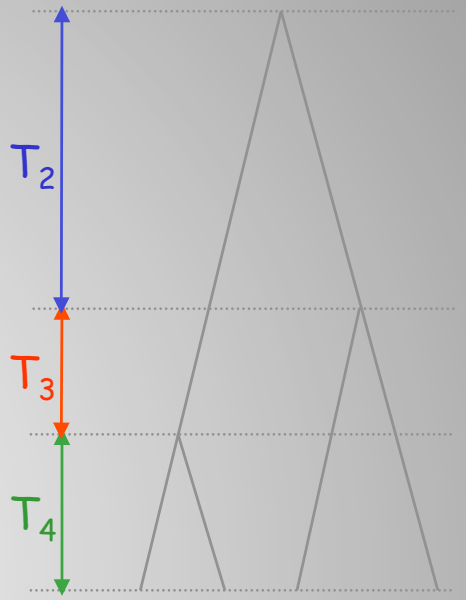
Le temps entre deux coalescences dans un ensemble de *j* lignées ancestrales suit donc une loi géométrique de paramètre *j*(j*-1)/2*N*, pouvant être approximée par une loi exponentielle d'espérance 2*N* / (*j (*j* -1))**

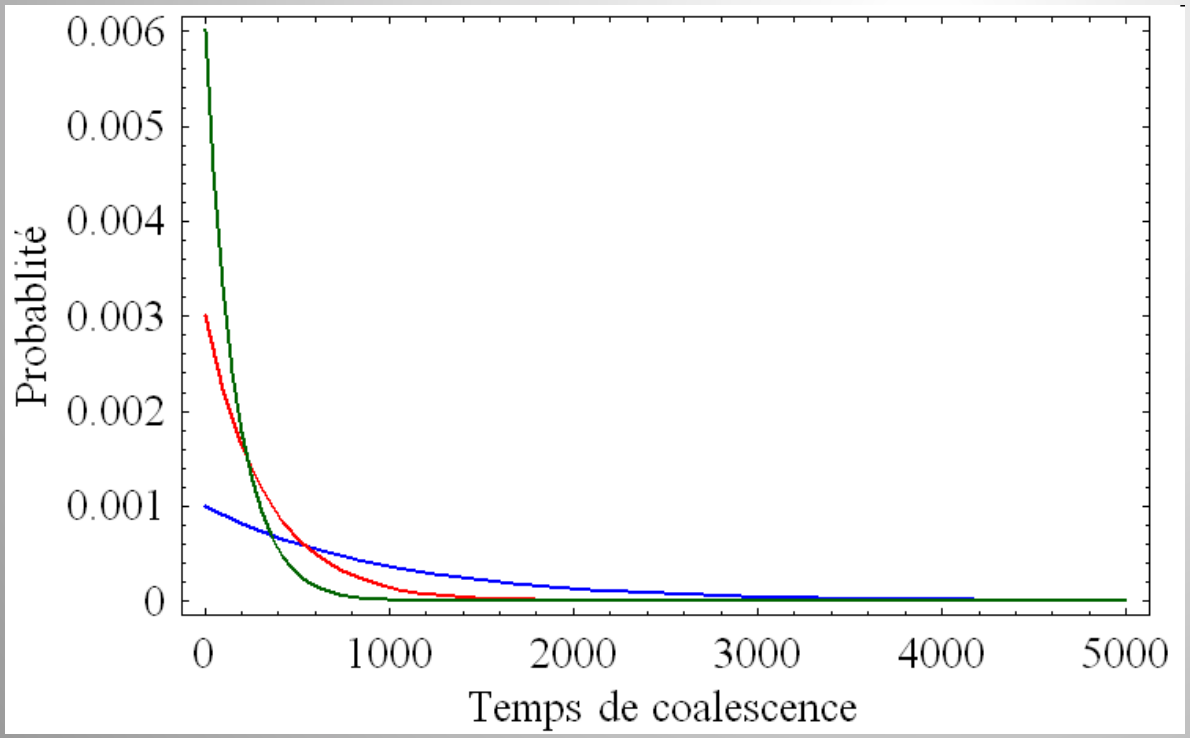
$$\Pr(T_j = t) = (1 - \frac{j(j-1)}{2N})^{t-1}(\frac{j(j-1)}{2N}) \approx \frac{j(j-1)}{2N} e^{-\frac{j(j-1)}{2N}t}$$

Coalescence en *t* générations de *j* lignées

$$\Pr(T_j = t) = (1 - \frac{j(j-1)}{2N})^{t-1}(\frac{j(j-1)}{2N})$$

$$\approx \frac{j(j-1)}{2N} e^{-\frac{j(j-1)}{2N}t}$$

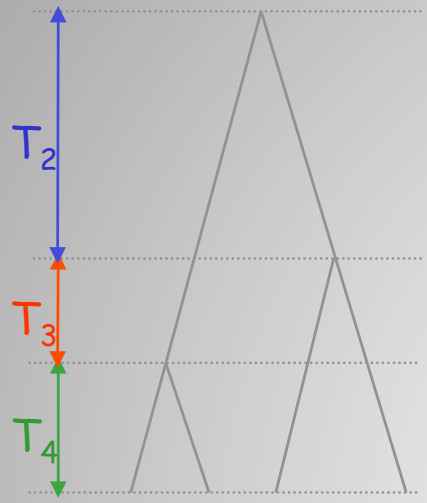


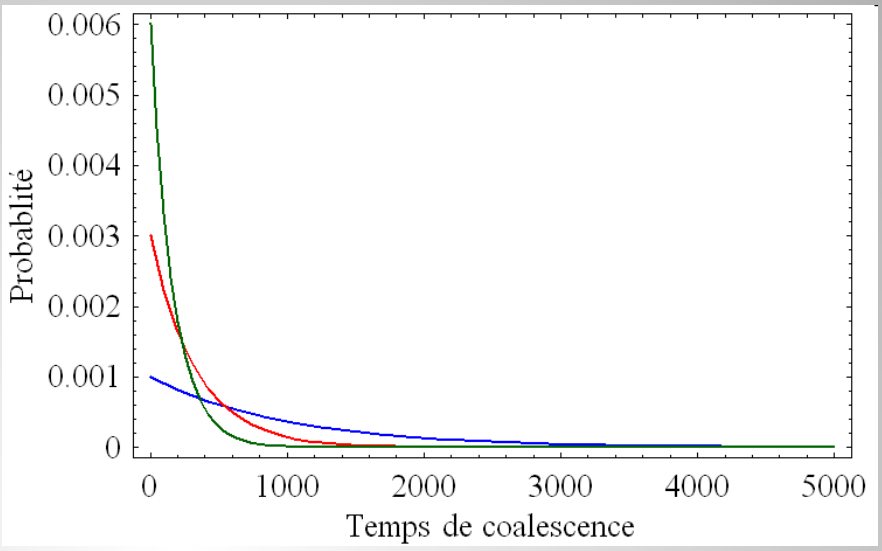


$$E(T_j) = \frac{2N}{j(j-1)}$$

$$\text{var}(T_j) = \frac{4N^2}{j^2(j-1)^2}$$

Coalescence en *t* générations de *j* lignées





$$E(T_j) = \frac{2N}{j(j-1)}$$

Le temps de coalescence est d'autant plus court que le nombre de lignées est grand

$$\mathrm{var}(T_j) = \frac{4N^2}{j^2(j-1)^2}$$

Très forte variance des temps de coalescence autour de la moyenne :
deux locus indépendants auront des temps de coalescence très différents

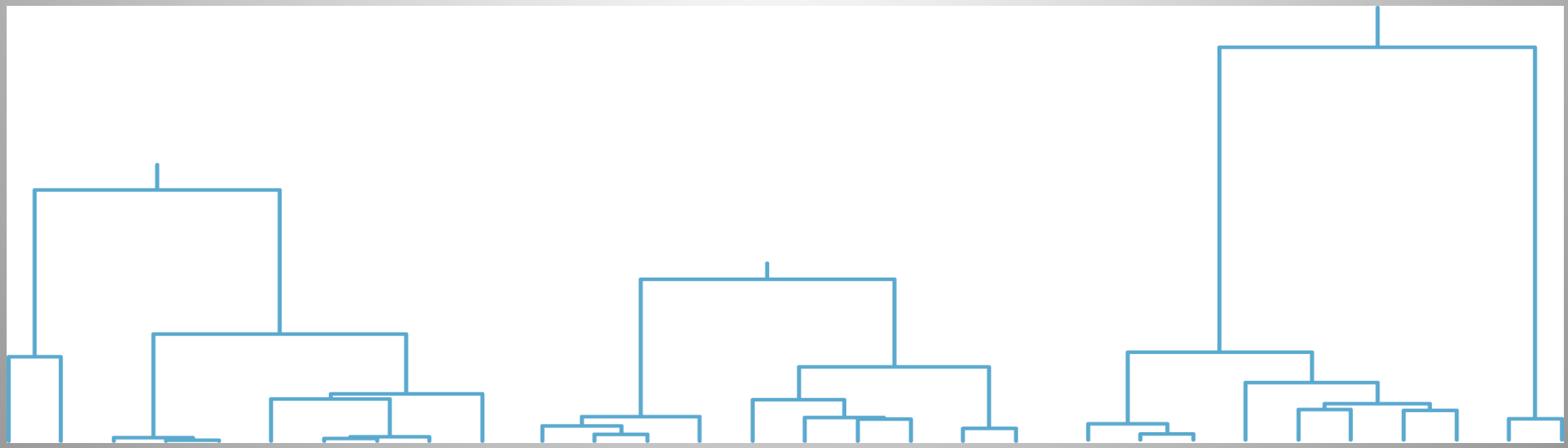
Coalescence en *t* générations de *j* lignées

$$E(T_j) = \frac{2N}{j(j-1)}$$

Le temps de coalescence est d'autant plus court que le nombre de lignées est grand

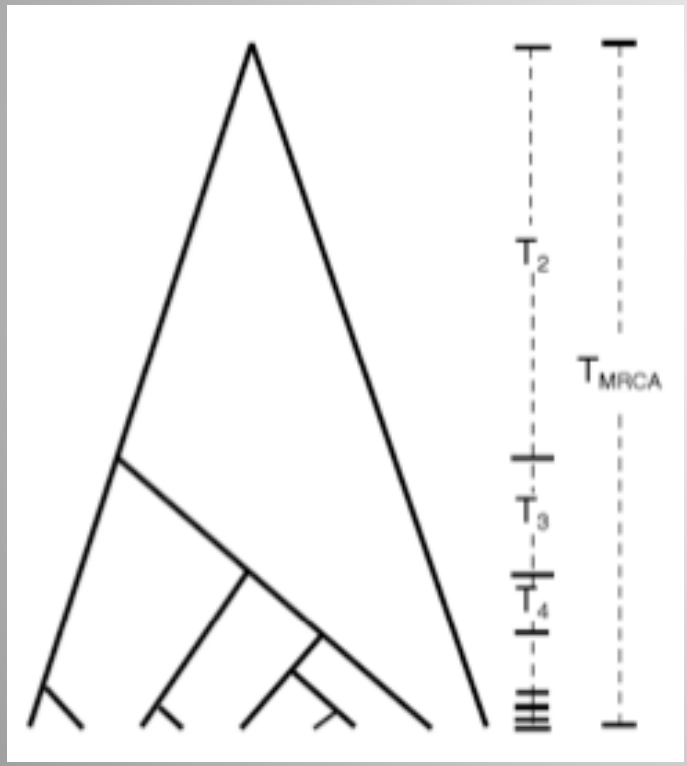
$$\mathrm{var}(T_j) = \frac{4N^2}{j^2(j-1)^2}$$

Très forte variance des temps de coalescence autour de la moyenne :
deux locus indépendants auront des temps de coalescence très différents



Taille de l'arbre de coalescence et TMRCA

TMRCA = Time to the Most Recent Common Ancestor
= âge du dernier nœud (coalescence) de l'arbre
= hauteur de l'arbre



$$E[TMRCA] = \sum_{i=2}^{j} E[T_i] = \sum_{i=2}^{j} \frac{2N}{i(i-1)}$$

$$= 2N \times \sum_{i=2}^{j} \left(\frac{1}{i-1} - \frac{1}{i}\right)$$

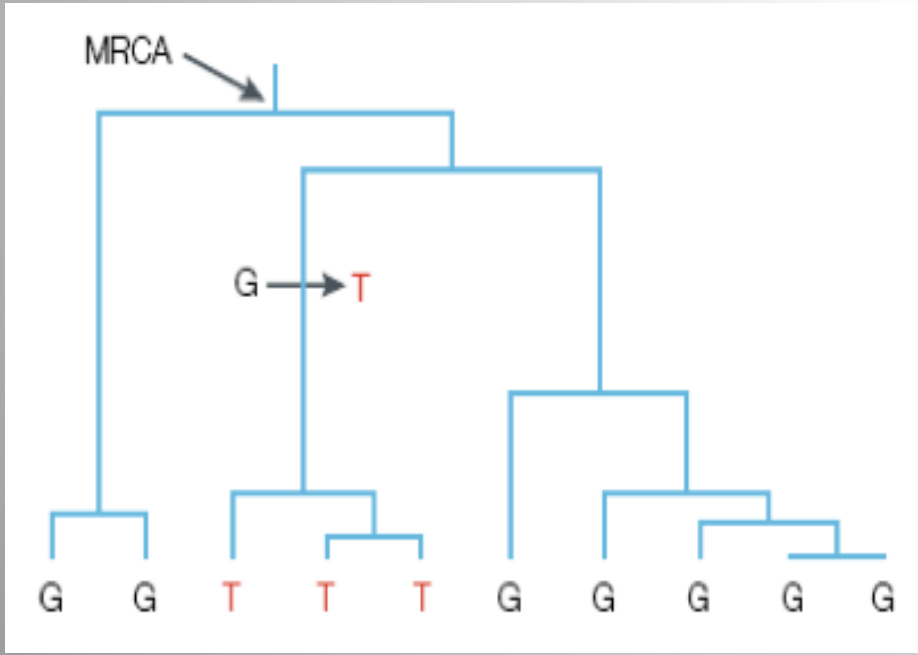
$$= 2N\left(1 - \frac{1}{j}\right)$$

✓ TMRCA tends vers 2*N* pour *j* grand
✓ TMRCA d'un petit échantillon est quasiment le même que celui de la population totale

Arbre de coalescence et mutations

Sous l'hypothèse de neutralité des marqueurs génétiques étudiés, les mutations sont indépendantes de la généalogie
i.e. la généalogie ne dépend que des processus démographiques

On construit donc la généalogie selon les paramètres démographiques (ex. N),



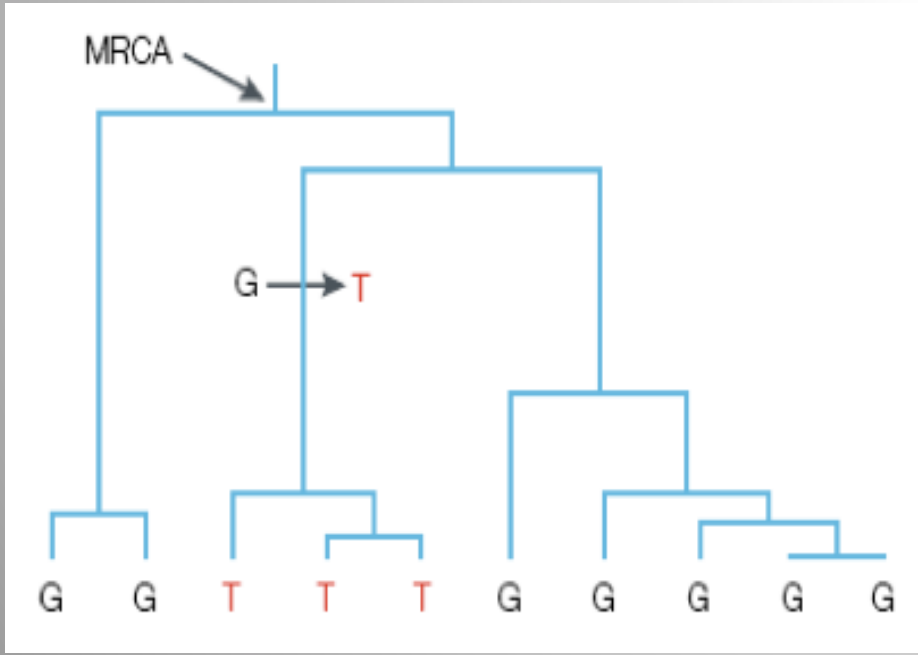
puis on ajoute a posteriori les mutations sur les différentes branches, du MRCA au feuilles de l'arbre
On obtient ainsi des données de polymorphisme sous les modèles démographiques et mutationnels considérés

Arbre de coalescence et mutations

Le nombre de mutation à appliquer sur chaque branche dépend du taux de mutation, μ, du marqueur considéré.

μ = nombre de mutation moyen par génération.

Ex: 5.10^{-4} pour les microsatellites, 10^{-6} par nucléotide pour des séquences



Sur une branche de longueur t, le nombre de mutation suit une loi binomiale de paramètres (μ,t).

Souvent approximé par une loi de poisson de paramètre (μ*t).

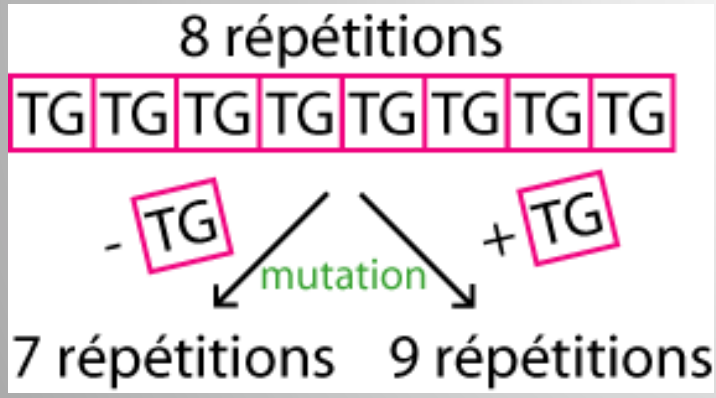
$$\Pr(k \text{ mut} | t) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

Arbre de coalescence et mutations

Il existe différents modèles mutationnels adaptés aux différents marqueurs génétiques :

✓Pour les séquences d'ADN, on utilise des modèles mutationnels nucléotidiques (Pr[A➔T], Pr[A➔C], Pr[T➔G], etc…)

✓Pour les microsatellites, il existe différents modèles de mutation par pas



Le SMM (Stepwise mutation model): la mutation ajoute ou retire une répétition à l'allèle parental

Le GSM (Generalized stepwise mutation model) : la mutation ajoute ou retire *x* répétitions (où *x* est une variable aléatoire suivant une loi géométrique).

Principaux avantages de la coalescence

- La coalescence offre un modèle probabiliste pour les généalogies de gènes

la généalogie, et plus généralement l'histoire évolutive, d'un échantillon, est la grande inconnue en évolution et ne peut pas être "refait" ⇒ la coalescence permet de bien prendre en compte cette inconnue

- La coalescence permet la simplification de l'analyse des modèles stochastiques de génétique des populations et de leur interprétation

La structure des données génétiques reflète la généalogie sous-jacente ⇒ la coalescence facilite donc l'analyse de la variabilité génétique observée et la compréhension des phénomènes évolutifs ayant façonnés le polymorphisme génétique.

Principaux avantages de la coalescence

- La coalescence permet de simuler très efficacement la variabilité génétique attendue sous différents modèles démo-génétiques (simulation d'échantillons plutôt que de populations entières)

- La coalescence permet le développement de puissantes techniques d'inférences statistiques de paramètres évolutifs populationnels (démographiques, génétiques,…), dont certaines permettent l'usage complet de l'information contenue dans les données

Construction d'arbres de coalescence et simulation de données de polymorphisme

- Principe général (Rappel) :
  - ✓ Pour des marqueurs neutres, le nombre de descendants est indépendant des types allélique porté par les parents

    ➔ les processus démographiques sont donc indépendants des processus mutationnels

  - ✓ La simulation de données de polymorphisme se fait donc en 2 temps :

    (1) Construction de l'arbre de coalescence :

    topologie + longueurs de branches

    (2) Ajout des mutation

Simulation d'arbres de coalescence

- **Modèle en urne**

  très rapide mais ne marche que dans le cas d'une population panmictique sans fluctuations démographiques

- **Approximations continues de Hudson (1991)**

  Assez rapide mais ne marche pas dans des cas complexes (petites pops, forts taux de migration, modèles complexes)

- **Génération par génération**

  Ok pour tous modèles démographiques et génétiques, mais **lent**

RAPIDITE :

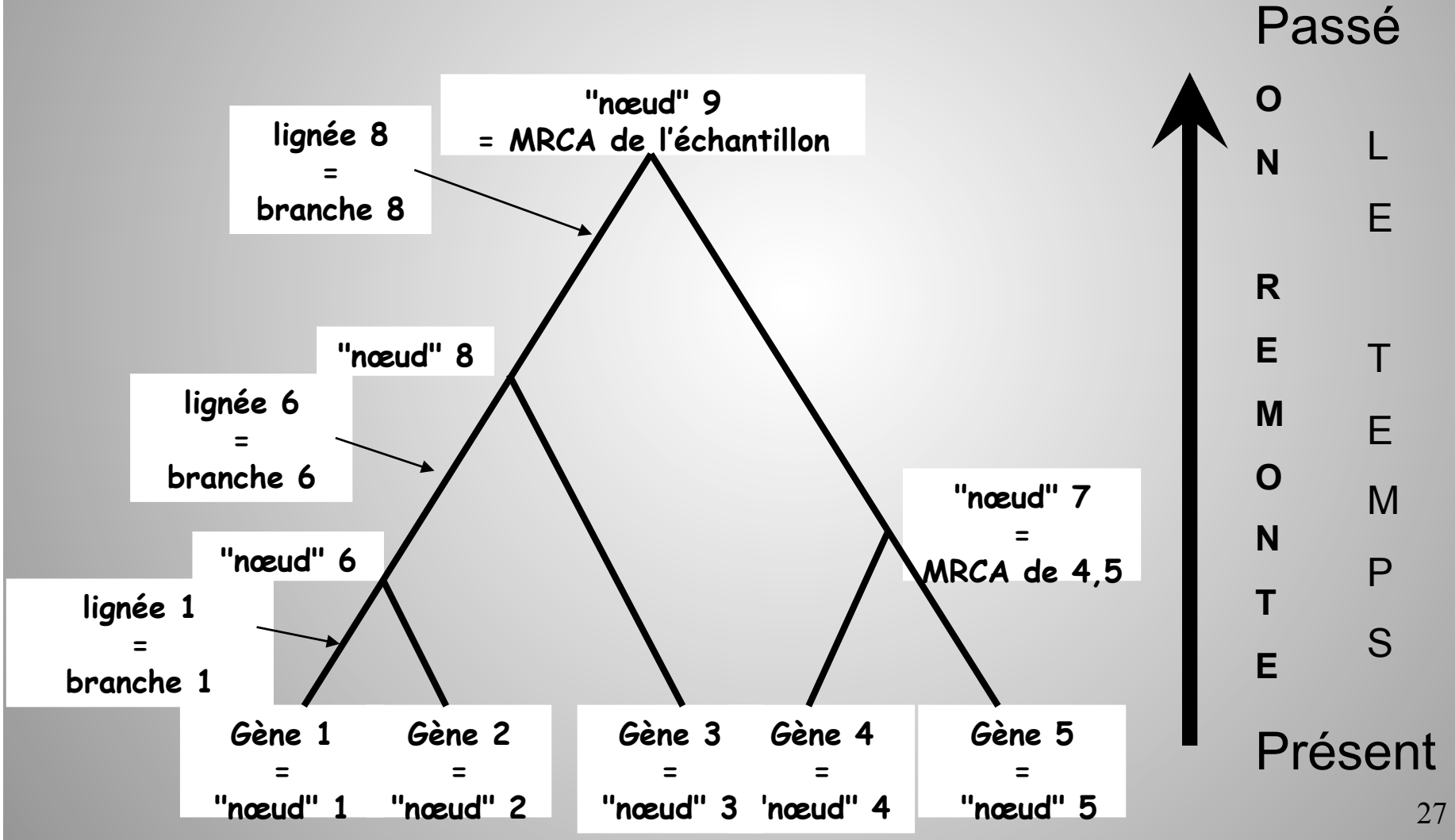
Urne > Approximation continue > Génération par génération

FLEXIBILITE :

Génération par génération >Approximation continue > Urne

Simulation d'arbres de coalescence

- Représentation de l'arbre :



Simulation d'arbres de coalescence
Génération par génération

- Principe simple, sans aucune approximation:
  - ✓ On remonte le temps génération par génération
  - ✓ À chaque génération, on recherche les éventuels évènements affectant la généalogie
    ex: coalescence, migration, recombinaison
  - ✓ On s'arrête quand on arrive a l'ancêtre commun de tous les gènes de l'échantillon
    = MRCA (Most Recent Common Ancestor)

Simulation d'arbres de coalescence
Génération par génération

- Un exemple simple :
  - ✓ échantillon de 4 gènes
  - ✓ à un locus neutre
  - ✓ ayant évolué dans une population haploïde panmictique de taille N=10

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	1	2	3	4
Nombre aléatoire entre 1 et N assigné aux nœuds				
Génération d'apparition du nœud/lignée	0	0	0	0

① ② ③ ④

Gn=0

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	1	2	3	4
Nombre aléatoire entre 1 et N assigné aux nœuds				
Génération d'apparition du nœud/lignée	0	0	0	0

Probabilité d'avoir une coalescence parmi *j* lignées à une génération

$=j(j-1)/2N$

= probabilité de tirer 2 nombre identiques entre 1 et *N* en *j* tirages

Gn=0

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

  Probabilité d'avoir une coalescence parmi *j* lignées à une génération = $j(j-1)/2N$

  = probabilité de tirer 2 nombre identiques entre 1 et *N* en *j* tirages

  En d'autres termes, on tire au hasard un parent pour chaque gène parmi N (taille stable)

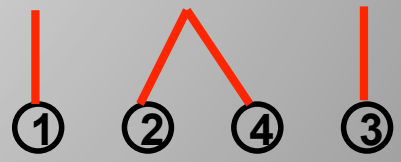
  Les gènes ayant le même parent coalescent

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	1	2	3	4
Nombre aléatoire entre 1 et N assigné aux nœuds	2	6	5	6
Génération d'apparition du nœud/lignée	0	0	0	0

Coalescence à la génération 1 des nœuds/ lignées 3 et 4



Gn=1

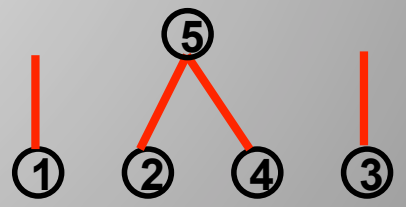
Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	1	3	5
Nombre aléatoire entre 1 et N assigné aux nœuds	2	5	6
Génération d'apparition du nœud/lignée	0	0	1

Coalescence à la génération 1 des nœuds/ lignées 3 et 4

Donne le nœud 5



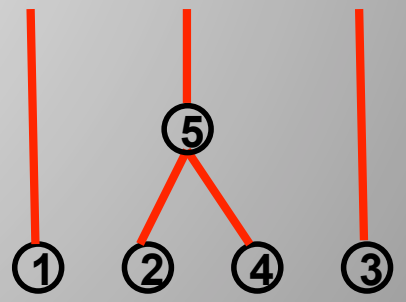
Gn=1

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	1	3	5
Nombre aléatoire entre 1 et N assigné aux nœuds	3	1	7
Génération d'apparition du nœud/lignée	0	0	1

Rien à la génération 2



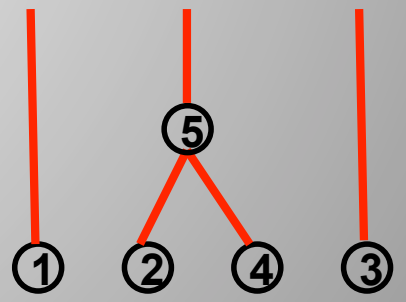
Gn=2

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	1	3	5
Nombre aléatoire entre 1 et N assigné aux nœuds	7	4	8
Génération d'apparition du nœud/lignée	0	0	1

Rien à la génération 3



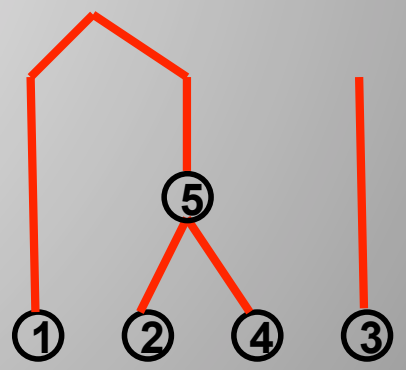
Gn=3

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	1	3	5
Nombre aléatoire entre 1 et N assigné aux nœuds	5	2	5
Génération d'apparition du nœud/lignée	0	0	1

Coalescence à la génération 4 des nœuds/ lignées 1 et 5



Gn=4

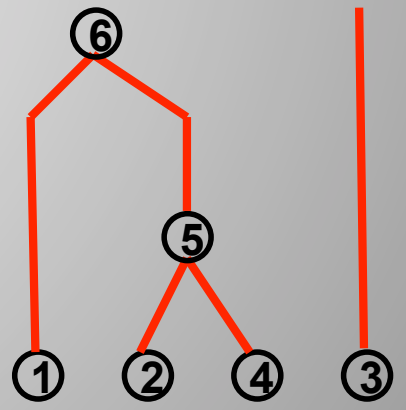
Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	3	6
Nombre aléatoire entre 1 et N assigné aux nœuds	2	5
Génération d'apparition du nœud/lignée	0	5

Coalescence à la génération 4 des nœuds/ lignées 1 et 5

Donne le nœuds 6



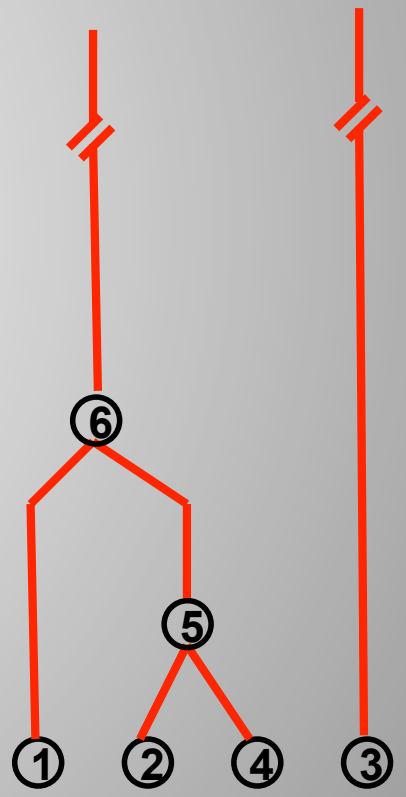
Gn=4

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	3	6
Nombre aléatoire entre 1 et N assigné aux nœuds	3	9
Génération d'apparition du nœud/lignée	0	5

Rien aux générations 5,6,…



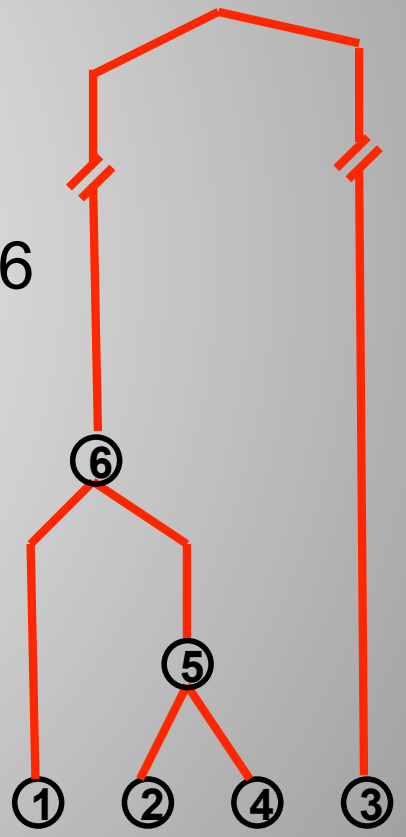
Gn=5

Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	3	6
Nombre aléatoire entre 1 et N assigné aux nœuds	7	7
Génération d'apparition du nœud/lignée	0	5

Coalescence à la génération 20 des 2 dernières lignées 3 et 6



Gn=20

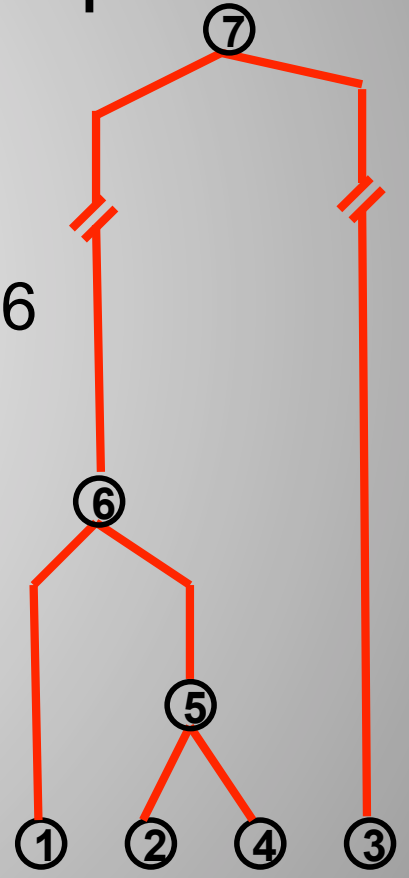
Simulation d'arbres de coalescence
Génération par génération

- Exemple : éch. 4 gènes, neutre, 1 pop. N=10

Numéro des nœuds/lignées	3	6
Nombre aléatoire entre 1 et N assigné aux nœuds	7	7
Génération d'apparition du nœud/lignée	0	5

Coalescence à la génération 20 des 2 dernières lignées 3 et 6

Donne le nœuds 7 = MRCA de l'échantillon



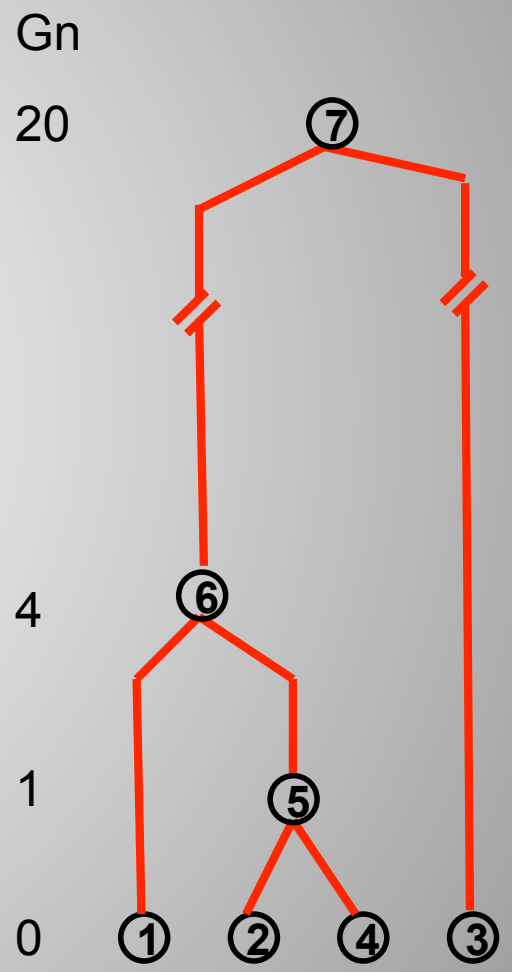
Gn=20

Simulation d'arbres de coalescence
Génération par génération

L'arbre (topologie et longueurs de branches) est construit.

C'est un processus aléatoire, donc si on fait plusieurs arbres, ils seront tous différents mais partageront certaines propriétés.

Pour obtenir des données de polymorphisme génétique, il faudra ajouter les mutations…



Simulation d'arbres de coalescence Approximations continues de Hudson

- Principe : 2 étapes successives

(1) Construction de la topologie de l'arbre en coalesçant au hasard les lignées

(2) Simulation des temps entre 2 coalescences successives = longueurs des branches

Simulation d'arbres de coalescence Approximations continues de Hudson

- Exemple : éch. 4 gènes, une pop N=10

(1) construire la topologie en coalesçant au hasard les lignées ancestrales

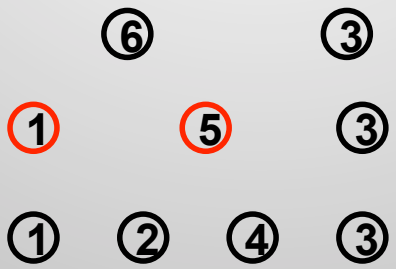
1ère coalescence = tirage au sort de 2 lignées parmi 4 -> les lignées 2 et 4 donnent la lignée 5

Simulation d'arbres de coalescence Approximations continues de Hudson

- Exemple : éch. 4 gènes, une pop N=10

(1) construire la topologie en coalesçant au hasard les lignées ancestrales

2ème coalescence = tirage au sort de 2 lignées parmi les 3 restantes -> les lignées 1 et 5 donnent la lignée 6

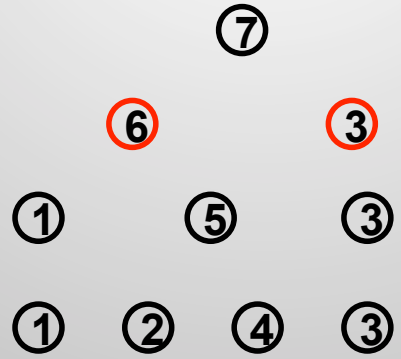


Simulation d'arbres de coalescence Approximations continues de Hudson

- Exemple : éch. 4 gènes, une pop N=10

(1) construire la topologie en coalesçant au hasard les lignées ancestrales

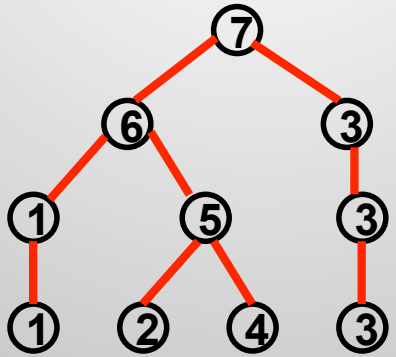
3ème coalescence = seules les 2 dernières peuvent coalescer -> les lignées 6 et 3 donnent la lignée 7



Simulation d'arbres de coalescence Approximations continues de Hudson

- Exemple : éch. 4 gènes, une pop N=10

(1) On a construit la topologie de l'arbre en coalesçant au hasard les lignées ancestrales

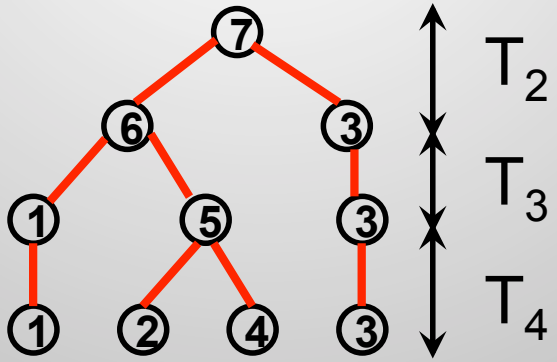


Simulation d'arbres de coalescence
Approximations continues de Hudson

- Exemple : éch. 4 gènes, une pop N=10

(2) Simulation des temps entre les coalescences successives
= longueurs des branches

3 longueurs de branches a simuler T_4, T_3, T_2



Simulation d'arbres de coalescence Approximations continues de Hudson

- Exemple : éch. 4 gènes, une pop N=10

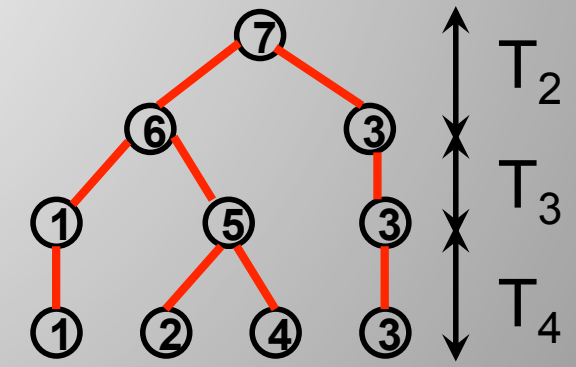
3 longueurs de branches a simuler T4, T3, T2

$$\Pr(T_j = k) = \frac{j(j-1)}{2N} e^{\frac{-j(j-1)}{2N}k}$$

T_4 tiré dans une loi exponentielle de paramètre

$j(j\text{-}1)/2N = 4*3/2*10$

(algorithmes disponibles)

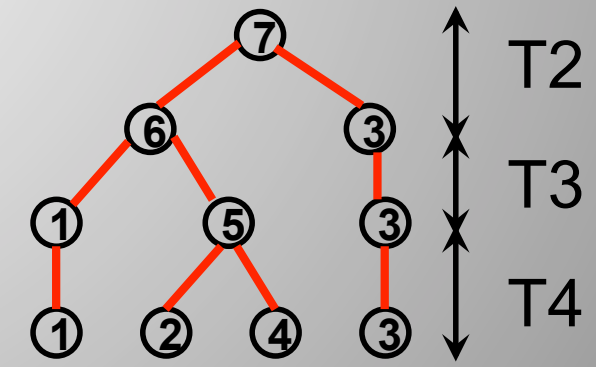


Simulation d'arbres de coalescence Approximations continues de Hudson

- Exemple : éch. 4 gènes, une pop N=10

3 longueurs de branches a simuler T4, T3, T2

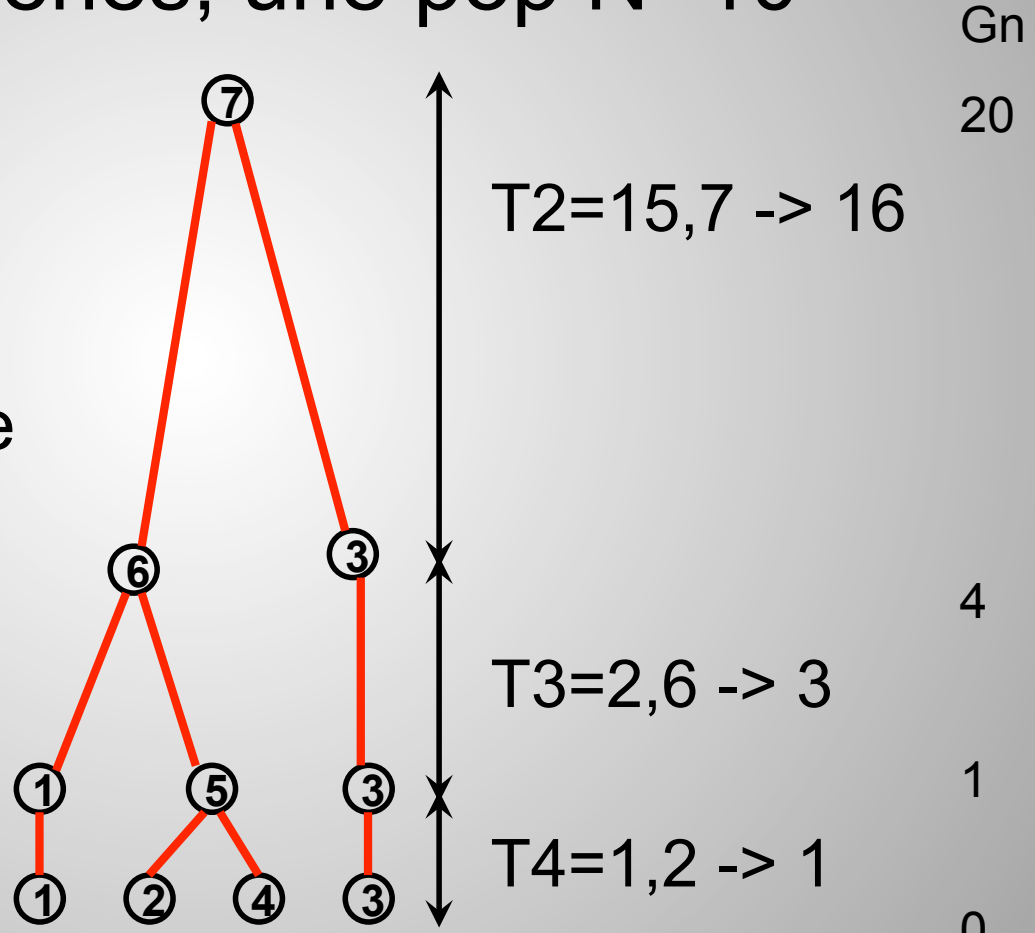
Ex:
T4 tiré dans une loi exp. (j(j-1) /2N=4*3/2*10) -> 1,2
T3 tiré dans exp. (3*2/2*10) -> 2,6
T2 tiré dans exp. (2*1/2*10) -> 15,7



Simulation d'arbres de coalescence
Approximations continues de Hudson

Exemple : éch. 4 gènes, une pop N=10

On a donc la topologie et la longueur des branches, ce qui donne l'arbre complet



Attention : Il faut connaître les distributions des temps de coalescence

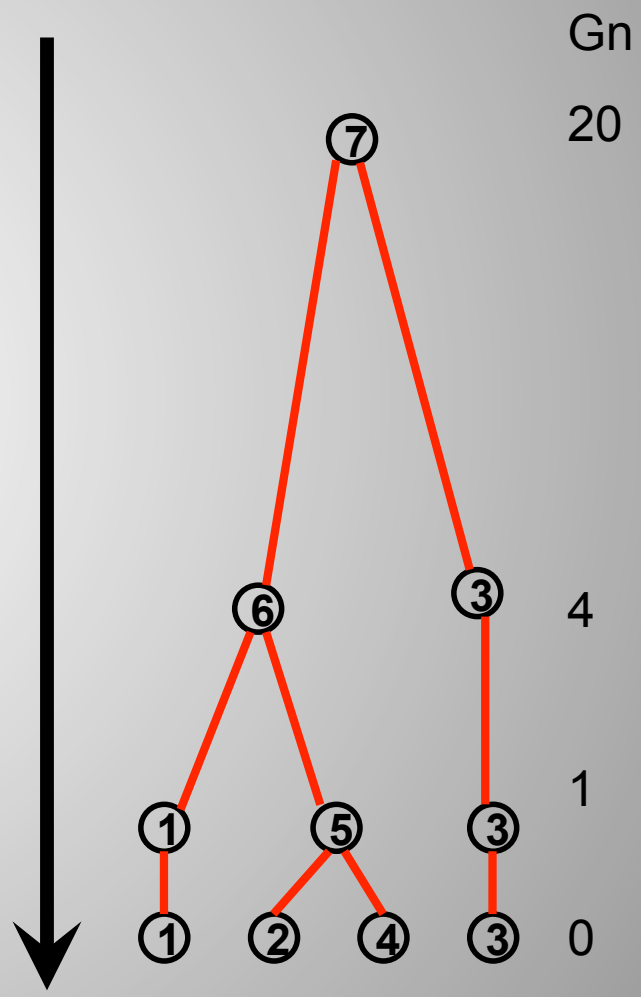
Simulation de données de polymorphisme à partir d'un arbre de coalescence

Principe général (rappel) :

On distribue les mutations sur les différentes branches de l'arbre en descendant du MRCA vers l'échantillon en fonction du taux de mutation μ

Chaque mutation induit un changement de l'état allélique du nœud descendant

Le changement se fait en fonction du modèle mutationnel choisi (qui doit refléter les processus mutationnels réels des marqueurs considérés)

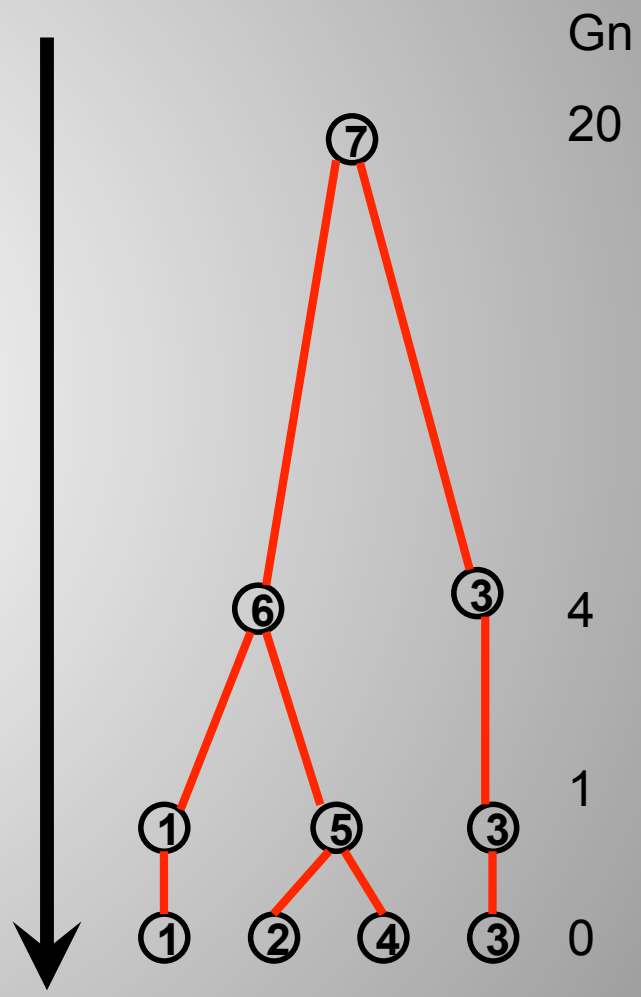


Simulation de données de polymorphisme à partir d'un arbre de coalescence

Sur une branche de longueur *t*, le nombre de mutation suit une loi binomiale de paramètres (μ,*t*)

Approximation loi de poisson de paramètre (μ**t*)

$$\Pr(k \text{ mut} | t) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$

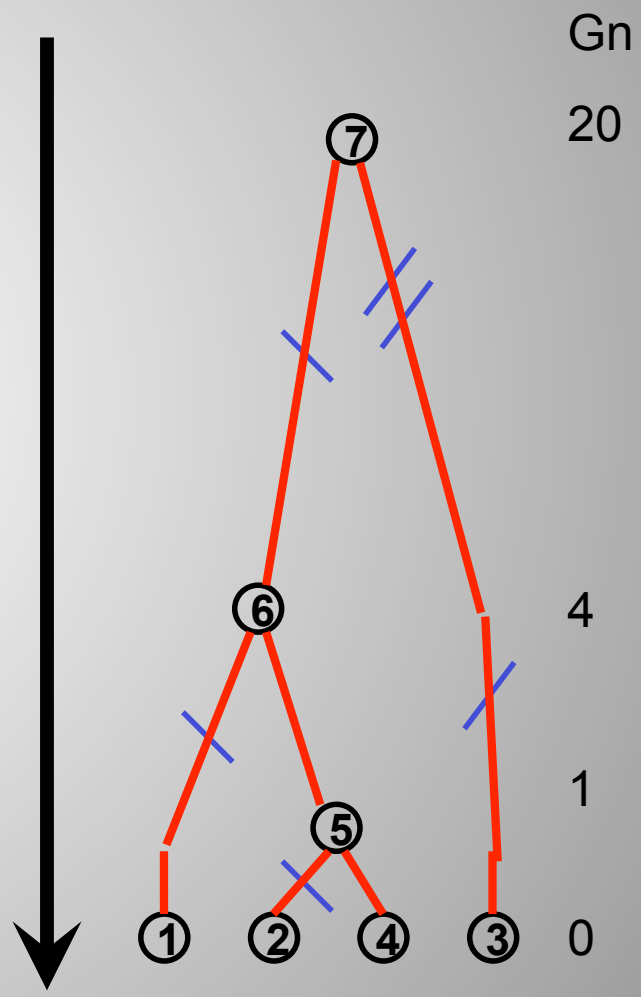


Simulation de données de polymorphisme à partir d'un arbre de coalescence

Exemple en SMM : perte ou gain d'un motif a chaque mutation

Ajout des mutations sur chaque branche (loi de Poisson)

$$\Pr(k \text{ mut} | t) = \frac{(\mu t)^k e^{-\mu t}}{k!}$$



Simulation de données de polymorphisme à partir d'un arbre de coalescence

Exemple en SMM : perte ou gain d'un motif a chaque mutation

Ajout des mutations sur chaque branche (loi de Poisson)

Choix au hasard du type du MRCA : 20

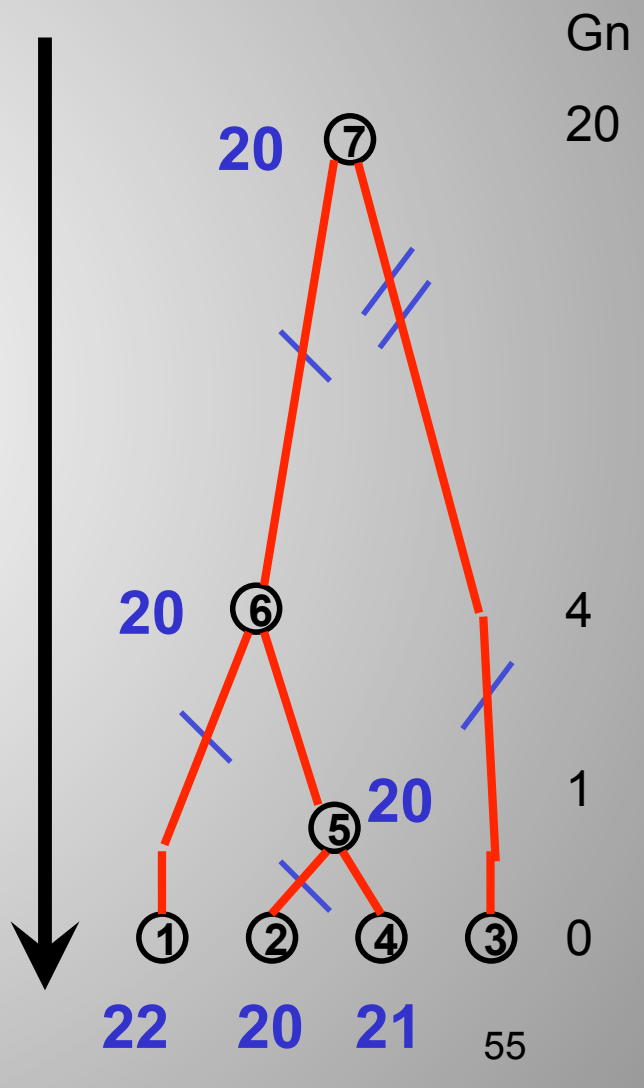
7 vers 6 : une fois ±1 ➔ 21

6 vers 1 : une fois ±1 ➔ 22

6 vers 5 : 0 fois ±1 ➔ 21

5 vers 2 : une fois ±1 ➔ 20

5 vers 4 : 0 fois ±1 ➔ 21



Simulation de données de polymorphisme à partir d'un arbre de coalescence

Exemple en SMM : perte ou gain d'un motif a chaque mutation

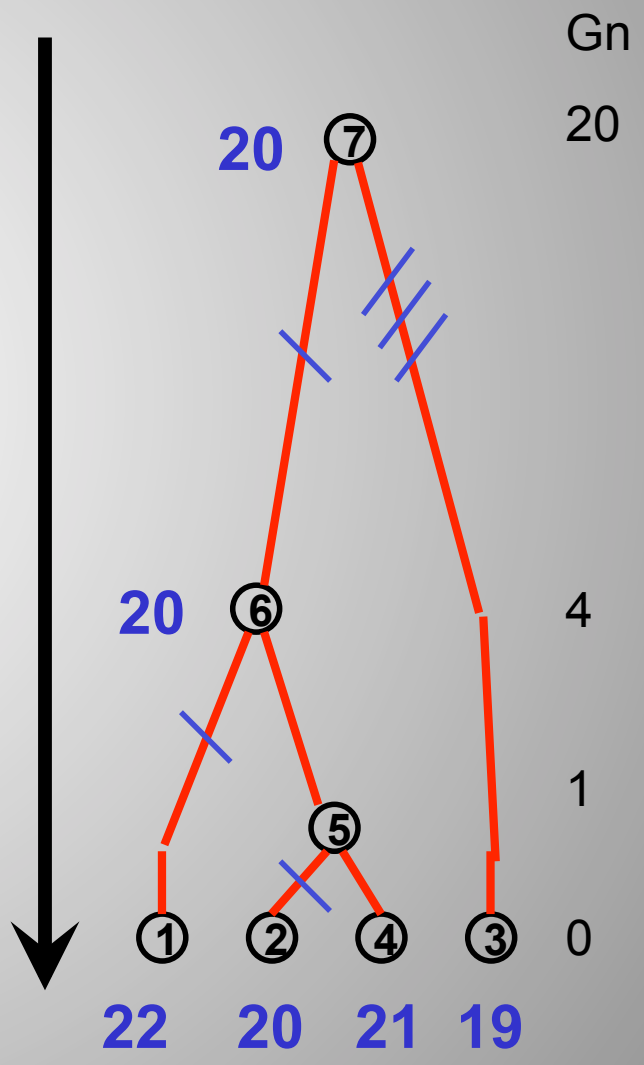
Ajout des mutations sur chaque branche (loi de Poisson)

Choix au hasard du type du MRCA : 20

7 vers 3 : 3 fois ±1 ➔ 19

On a un échantillon de polymorphisme

4 gènes différents de type 19, 20, 21, 22



Simulation de données de polymorphisme à partir d'un arbre de coalescence

Exemple pour des données de séquence de 5 nucléotides.

Séquence ancestrale (ATTGC)

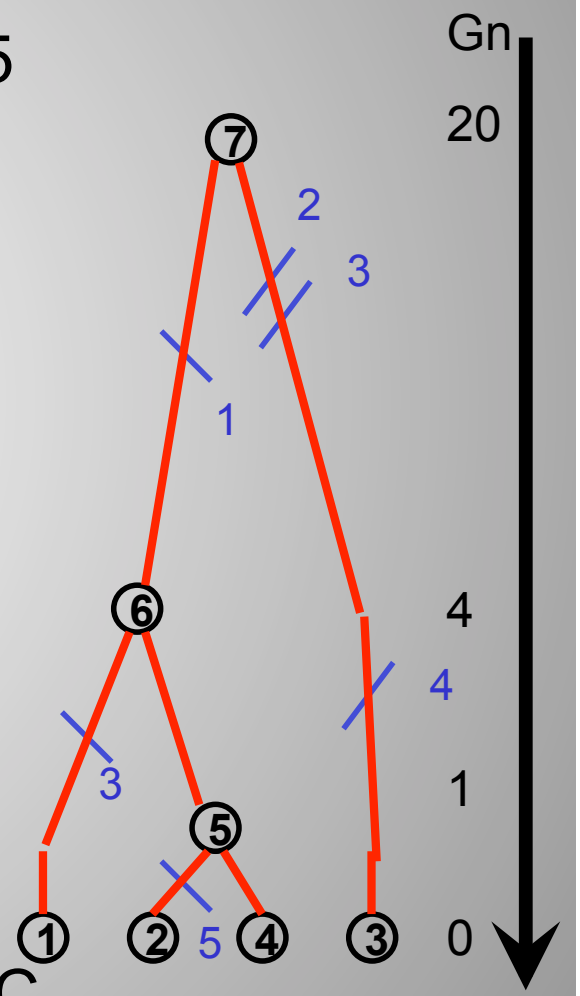
Les mutations arrivent indépendamment sur chaque nucléotide

7 vers 6 : 1 mut sur le nucl. 1 ➔ TTTGC

6 vers 1 : 1 mut sur le nucl. 3 ➔ TTAGC

5 vers 2 : 1 mut sur le nucl. 5 ➔ TTTGG

7 vers 3 : 3 mut sur les nucl. 2,3,4 ➔ AAACC



Simulation de données de polymorphisme à partir d'un arbre de coalescence

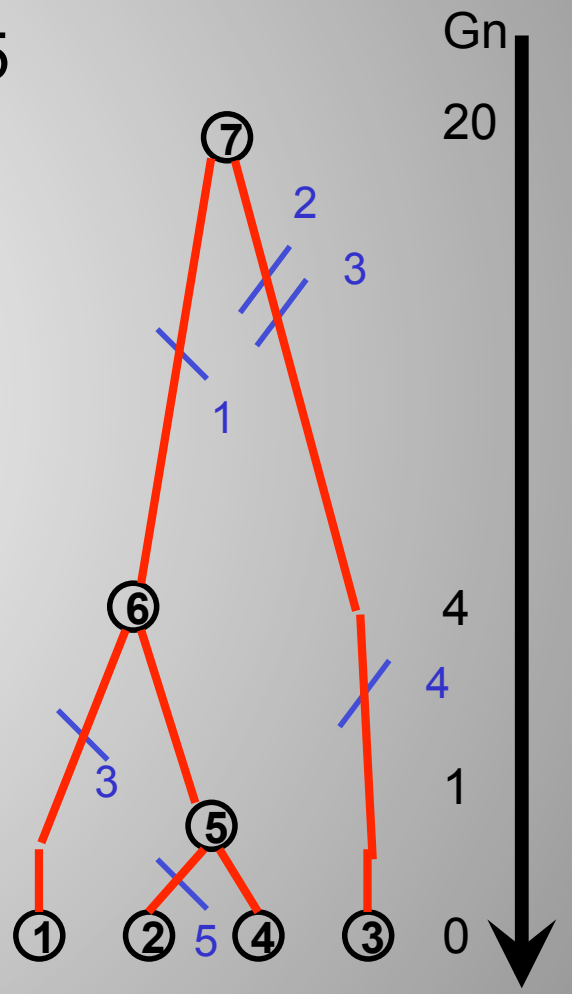
Exemple pour des données de séquence de 5 nucléotides.

Séquence ancestrale (ATTGC)

Les mutations arrivent indépendamment sur chaque nucléotide

L'échantillon de polymorphisme est donc composé de 4 séquences différentes :

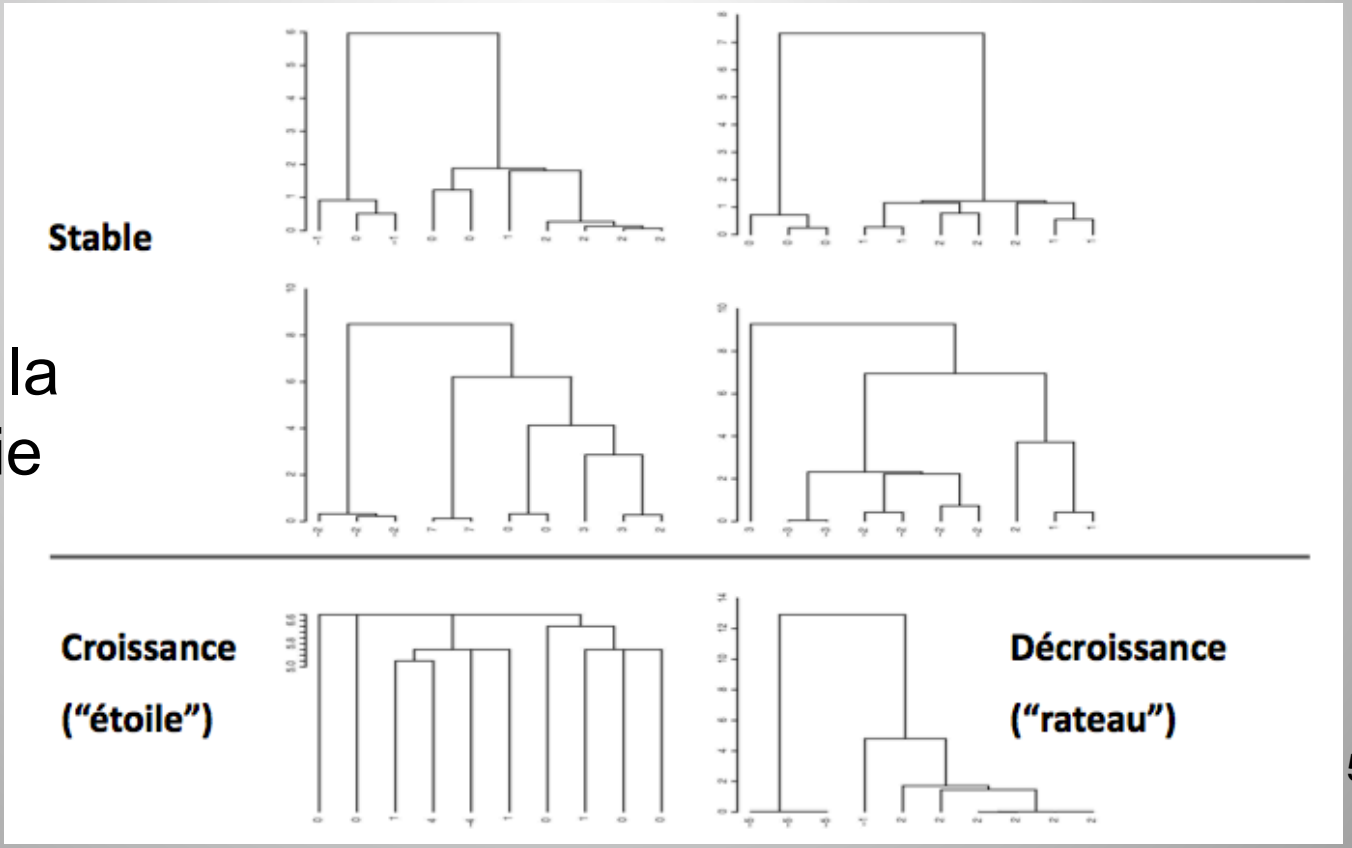
TTAGC,TTTGG,TTTGC,AAACC



A quoi servent ces arbres de coalescence et la simulation de données génétiques ?

- Approche exploratoire : Étudier l'effet de certain paramètres sur la forme de l'arbre et sur la distribution du polymorphisme au sein d'un échantillon

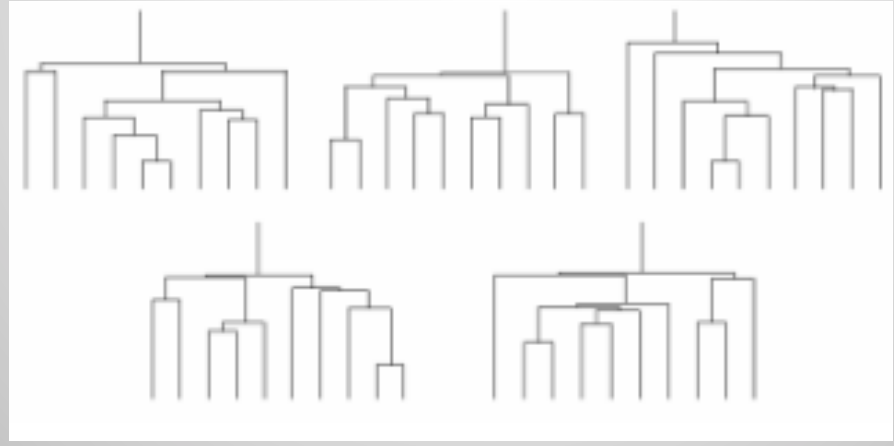
Ex: effet de la démographie



A quoi servent ces arbres de coalescence et la simulation de données génétiques ?

- Approche exploratoire : effets de la démographie
  - Croissance de la taille de population (ex: invasion)

  Il y a plus de coalescence anciennes (petit *N*) que de coalescences récentes (grand *N*), les arbres de coalescence ont donc de longues branches terminales

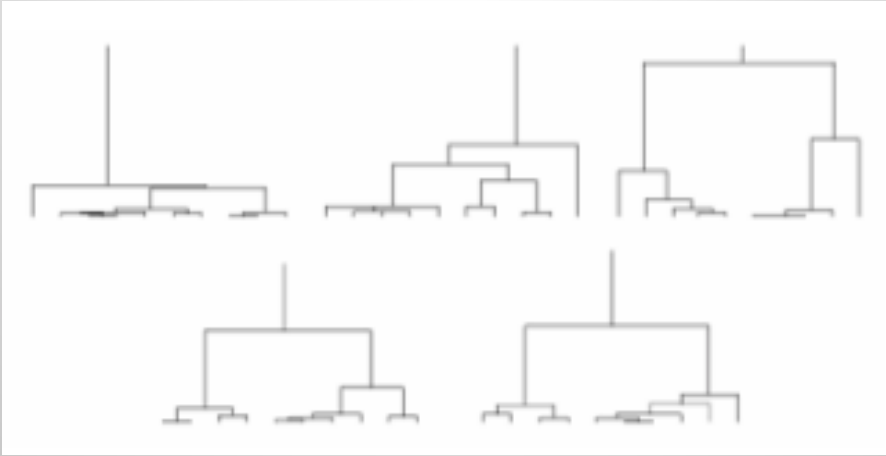


Une croissance démographique entraine un excès d'allèles à faibles fréquences (allèles rares)

A quoi servent ces arbres de coalescence et la simulation de données génétiques ?

- Approche exploratoire : effets de la démographie
  - Décroissance de la population (ex : espèce menacée)

  Il y a plus de coalescence récentes (petit *N*) que de coalescences anciennes (grand *N*), les arbres de coalescence ont donc de courtes branches terminales



Une décroissance entraine un déficit d'allèles à faibles fréquences

A quoi servent ces arbres de coalescence et la simulation de données génétiques ?

- **Approche exploratoire** : Étudier l'effet de certain paramètres sur la forme de l'arbre, sur le polymorphisme d'un échantillon et sur de statistiques résumées calculées sur un échantillon.
- **Test par simulation** : Créer des échantillons simulés pour tester la précision et la robustesse de méthodes d'estimation
- **Approche inférentielle** : Estimer des paramètres évolutifs populationnels (tailles de pops, migration, histoire démographique) à partir de données de polymorphisme

Coalescence et inférences démo-génétiques

- L'approche inférentielle repose sur la modélisation du fonctionnement des populations. Chaque modèle est caractérisé par un ensemble de paramètres démographiques et génétiques *P*
- Le but est d'estimer ces paramètres à partir d'un jeu de données de polymorphisme (échantillon génétique)
- L'échantillon génétique est alors considéré comme la réalisation d'un processus stochastique défini par le modèle démo-génétique

Coalescence et inférences démo-génétiques

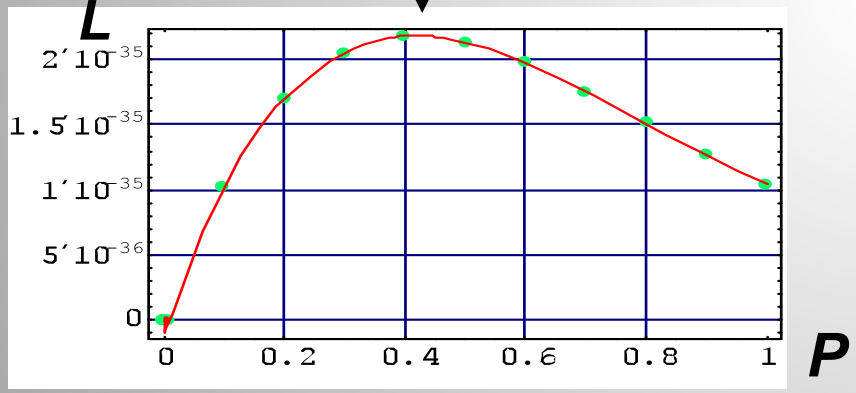
- Dans un premier temps, on cherche à calculer la probabilité Pr(D |P) d'observer les données D sachant des valeurs fixées de paramètres P, c'est la vraisemblance : L(P |D)=Pr(D |P)

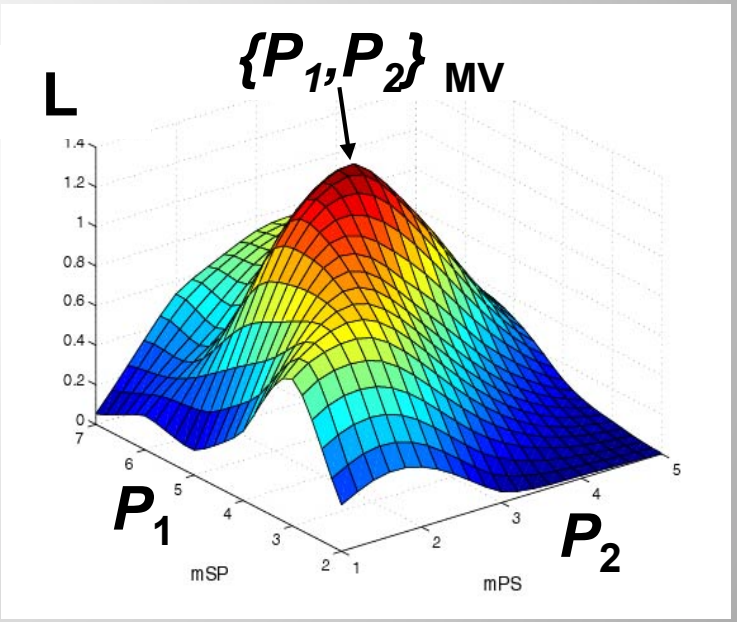
- On cherche ensuite les valeurs de paramètres qui maximise cette probabilité d'observation des données (méthode du maximum de vraisemblance)

Coalescence et inférences démo-génétiques

- Méthode du maximum de vraisemblance :

P_{MV} = estimateur du maximum de vraisemblance





!! bcp de paramètres → grand espace à explorer !!

Coalescence et inférences démo-génétiques

- Problème : la plupart du temps, on ne sait pas calculer la vraisemblance de données de polymorphisme P($D|P$) car on a pas d'expression mathématiques explicite
- Mais on sait calculer P($D|P,G_i$), la probabilité d'observer les données génétiques sachant des valeurs de paramètres et une généalogie G_i
- La vraisemblance peut donc s'écrire comme la somme des vraisemblances sur tout l'espace des généalogies possibles :

$$L(P|D) = \int_G \Pr(D|G;P)\Pr(G|P)dG$$

Coalescence et inférences démo-génétiques

- La vraisemblance peut s'écrire comme la somme des P(D|P,G_i) sur tout l'espace des généalogies possibles :

$$L(P|D) = \int_G \Pr(D|G;P)\Pr(G|P)\,dG$$

Paramètres mutationnels

Théorie de la coalescence, paramètres démographiques

- Les généalogies sont considérées comme des paramètres de nuisance (ou des données manquantes), elles sont importantes pour les calculs mais on ne cherche pas à les estimer

C'est très différent de l'approche phylogénétique

Coalescence et inférences démo-génétiques

$$L(P|D) = \int_G \Pr(D|G;P)\Pr(G|P)dG$$

Somme sur toutes les généalogies possibles

⇒ souvent impossible à faire !!!

On utilise alors la méthode de Monté Carlo dans laquelle on simule un grand nombre *K* de généalogies selon Pr(*G*|*P*) et l'on fait la moyenne sur ces *K* généalogies :

$$L(P|D) = E\left[\Pr(D|G;P)\right] \approx \frac{1}{K}\sum_{k=1}^{K}\Pr(D|G_k;P)$$

Bcp de généalogies à simuler pour avoir une bonne estimation de la vraisemblance

Coalescence et inférences démo-génétiques

$$L(P|D) = E\left[\Pr(D|G;P)\right] \approx \frac{1}{K}\sum_{k=1}^{K}\Pr(D|G_k;P)$$

La méthode de Monte Carlo seule est souvent peu efficace car beaucoup de généalogies donnent une probabilité très faible d'observer les données, on a donc recours a d'autres algorithmes pour chercher les généalogies expliquant le mieux les données.

Coalescence et inférences démo-génétiques

Il existe différents algorithmes plus efficace que les simulations de Monte Carlo seule :

- IS : L'échantillonnage d'importance (Importance Sampling)
- MCMC : les techniques de Monte Carlo couplées à des chaînes de Markov et à l'algorithme de Metropolis Hastings

permettent d'explorer les généalogies proportionnellement à leur probabilité d'expliquer les données P(D|P;G).

Coalescence et inférences démo-génétiques
ex. de l'approche de Felsenstein et al. (MCMC)

- Probabilité d'une généalogie sachant les paramètres démographiques du modèle: *N*,[N_i m_{ij} *si population structurée]*)

exemple pour une pop panmictique

$$\Pr(G|P) = \prod_{\tau=1}^{T} \left(\frac{j_\tau (j_\tau - 1)}{4N} e^{-\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{k}_\tau} \right)$$

Produit sur tous les évènements « démographiques » (coalescence ou migration si pop structurée) de la généalogie

Nombre de lignées avant l'évènement

Intervalle de temps entre cet évènement et le précédent

Coalescence et inférences démo-génétiques
ex. de l'approche de Felsenstein et al. (MCMC)

- Probabilité d'une généalogie sachant les paramètres démographiques du modèle (N, m_{ij})

$$\Pr(G|P) = \prod_{\tau=1}^{T} \left(\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{e}^{\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{k}_\tau} \right)$$

- Probabilité de l'échantillon sachant la généalogie et les paramètres mutationnels (μ, P_{mut} matrice de mutation)

$$\Pr(D|G) = \prod_{b=1}^{B} \left((P_{mut})^{i_b} \frac{(\mu L_b)^{i_b}}{i_b!} \mathrm{e}^{\mu L_b} \right)$$

Produit sur toutes les branches de l'arbre

Nombre de mutation sur la branche b

Longueur de la branche b

Loi de poisson pour la probabilité d'avoir i mutation sur un intervalle de temps Lb

Coalescence et inférences démo-génétiques
ex. de l'approche de Felsenstein et al. (MCMC)

- Probabilité d'une généalogie sachant les paramètres démographiques du modèle (N, m_{ij})

$$\Pr(G|P) = \prod_{\tau=1}^{T} \left(\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{e}^{\frac{j_\tau (j_\tau - 1)}{4N} \mathrm{k}_\tau} \right)$$

- Probabilité de l'échantillon sachant la généalogie et les paramètres mutationnels

$$\Pr(D|G) = \prod_{b=1}^{B} \left((P_{mut})^{i_b} \frac{(\mu L_b)^{i_b}}{i_b!} \mathrm{e}^{\mu L_b} \right)$$

- Par définition

$$L(P|D) \approx \frac{1}{K} \sum_{k=1}^{K} \Pr(D|G_k; P) \approx \frac{1}{K} \sum_{k=1}^{K} \Pr(D|G_k) \Pr(G_k|P)$$

Coalescence et inférences démo-génétiques

C’est un problème complexe, notamment a cause des grands espaces à explorer (Généalogie et Paramètres)

Plus il y a de paramètres, plus les généalogies sont complexes

Toujours avoir en tête que plus le modèle a de paramètres plus il faut de temps et/ou des algorithmes plus efficaces pour explorer l'espace des paramètres mais aussi celui des généalogies.

➔ essayer de considérer des modèles plus simple mais robustes

Algorithme de Metropolis-Hastings

(1) Dans l'espace des paramètres, partir de Θ

(2) Proposer un changement (Θ') selon $q\left(\Theta \rightarrow \Theta'\right)$

(3) Accepter ce changement avec la probabilité :

$$h = \min\left(1, \frac{L\left(\Theta';D\right)}{L\left(\Theta;D\right)} \frac{\mathrm{P}\left(\Theta'\right)}{\mathrm{P}\left(\Theta\right)} \frac{q\left(\Theta' \rightarrow \Theta\right)}{q\left(\Theta \rightarrow \Theta'\right)}\right)$$

(4) Aller à (1)

Cet algorithme assure que le MCMC explore les états possibles proportionnellement à la vraisemblance

Algorithme de Metropolis-Hastings

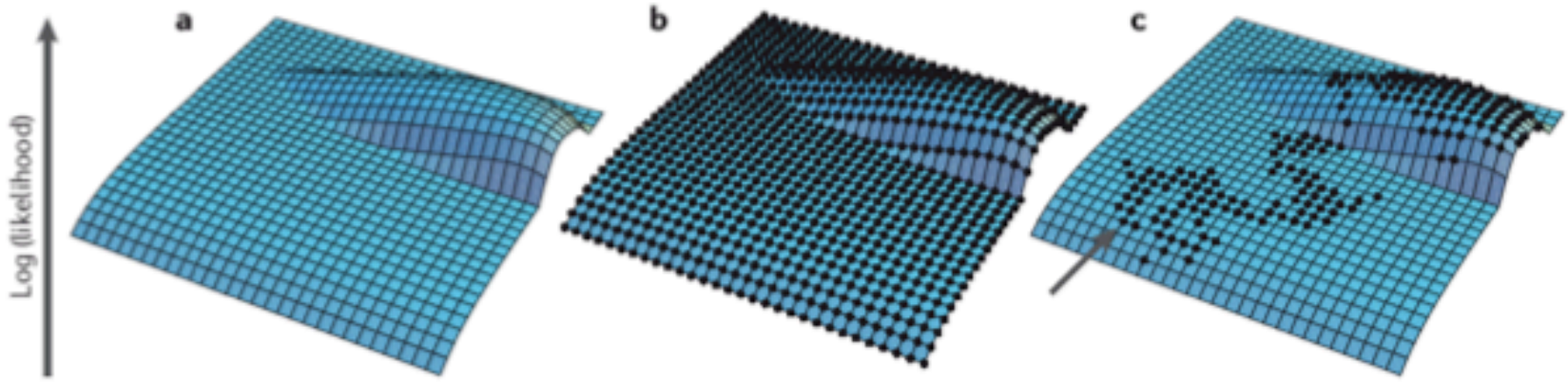


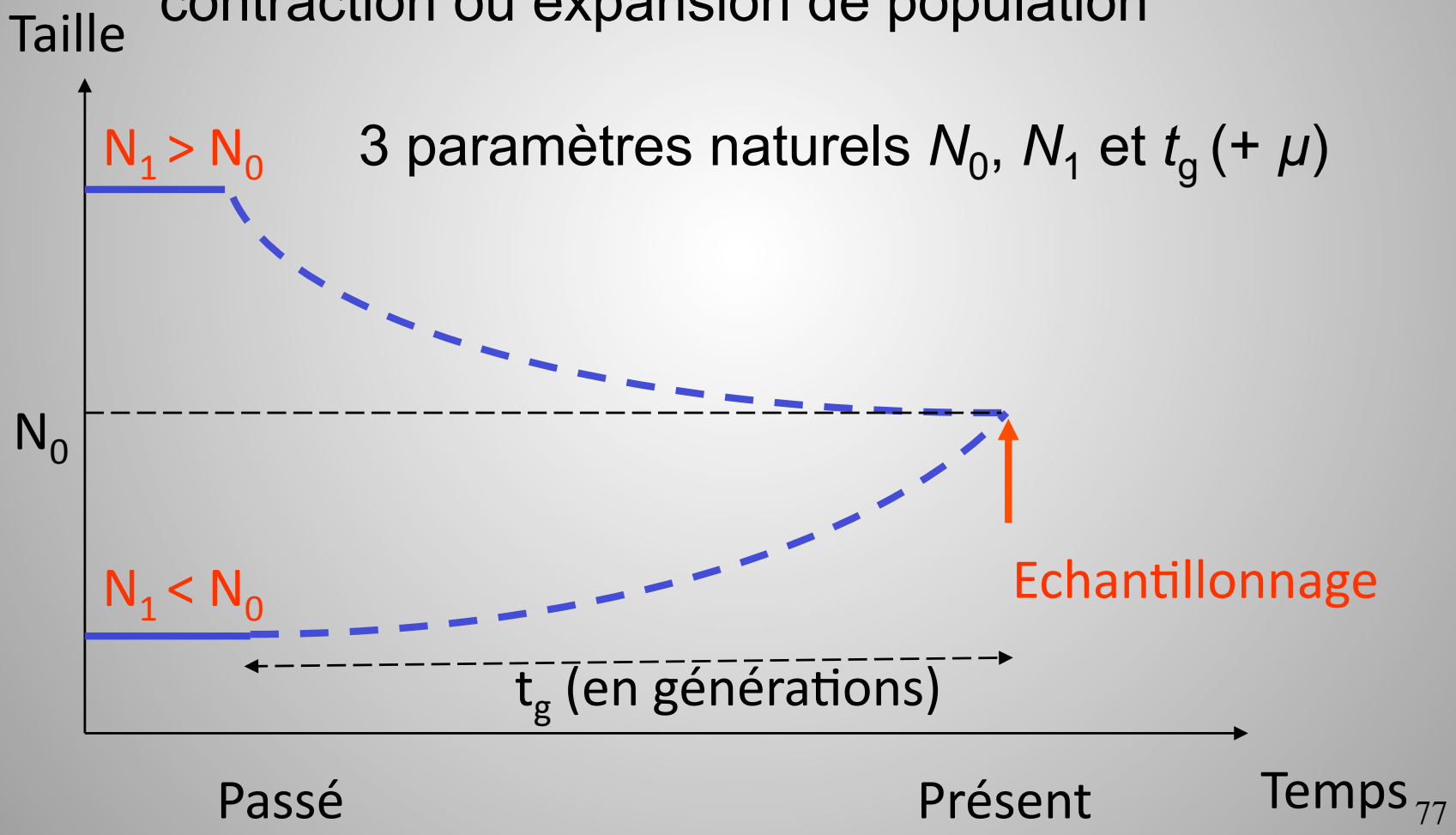
Image courtesy of Peter Beerli, Florida State University, USA.

Une manière plus efficace d'explorer l'espace…

Crédit : Excoffier et Heckel (2006) *Nature Reviews Genetics* **7** : 745-758

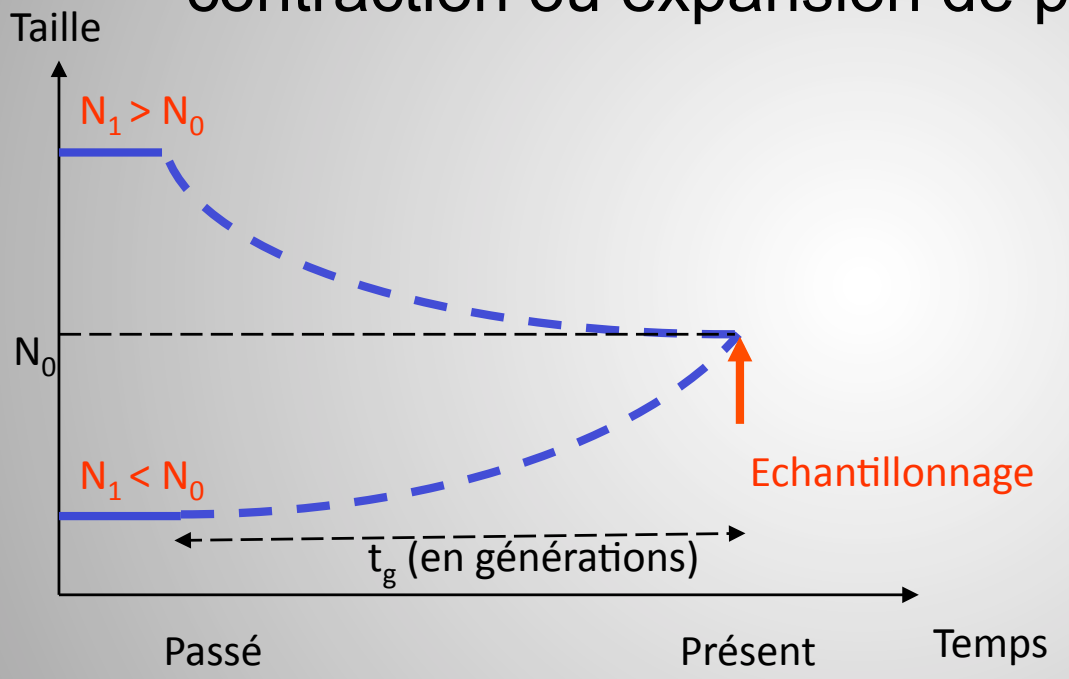
Un exemple : la méthode MsVar (Beaumont 1999)

✓Modèle de changement de taille de population passée : contraction ou expansion de population



Un exemple : la méthode MsVar (Beaumont 1999)

✓Modèle de changement de taille de population passée : contraction ou expansion de population



3 paramètres N_0, N_1 et t_g (+ μ) que l'on va estimer avec un algorithme MCMC…

Un exemple : la méthode MsVar (Beaumont 1999)

- Simulations de Monte Carlo par chaînes de Markov (MCMC)
  - ✓Pour explorer l'espace des généalogies
  - ✓ et l'espace des paramètres

Un exemple : la méthode MsVar (Beaumont 1999)

- Simulations de Monte Carlo par chaînes de Markov (MCMC)

  ✓Pour explorer l'espace des généalogies, on doit tout d'abord construire une généalogie de départ:

1- tirage d'un temps de coalescence

2- tirage d'un temps de mutation

3- choix du temps le plus court

4- si coa : on fait coalescer 2 lignées de même type allélique

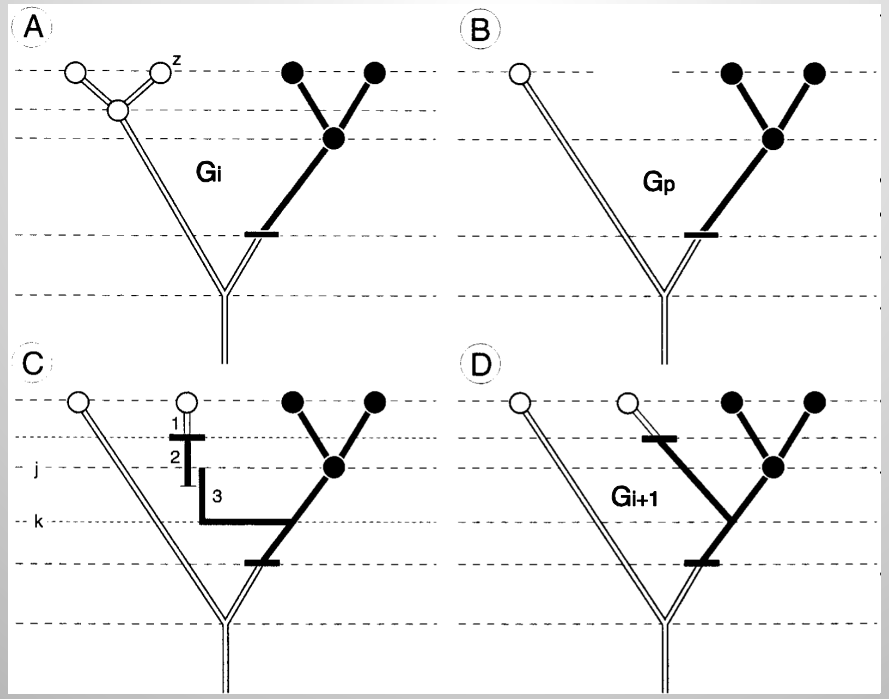
5- Si mut : on fait muter un gènes pris au hasard

Etc… jusqu'au MRCA

Un exemple : la méthode MsVar (Beaumont 1999)

- Simulations de Monte Carlo par chaînes de Markov (MCMC)

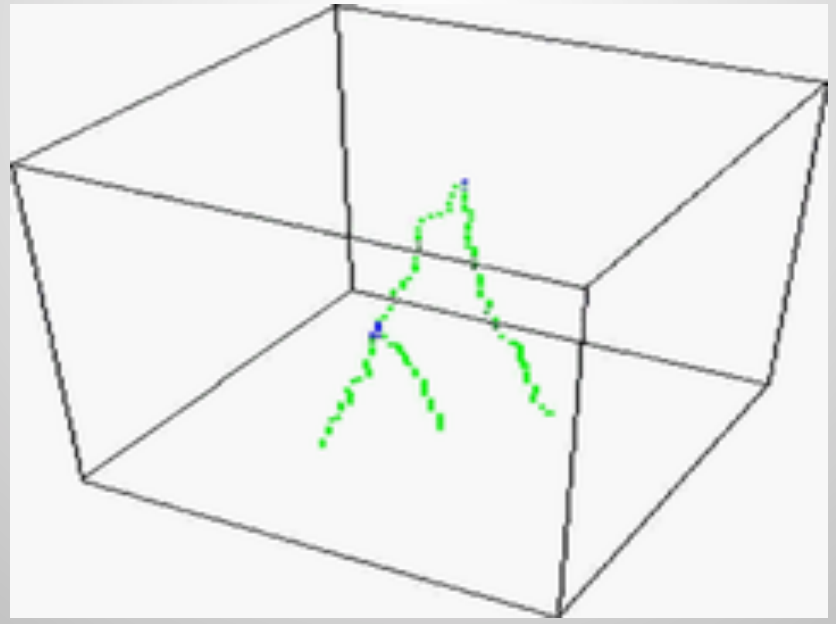
  ✓Pour explorer l'espace des généalogies, on va ensuite explorer différentes généalogies par délétion-reconstruction d'un bout de la généalogie courante:



Un exemple : la méthode MsVar (Beaumont 1999)

- Simulations de Monte Carlo par chaînes de Markov (MCMC)

  ✓Pour explorer l'espace des généalogies, on va ensuite explorer différentes généalogies par délétion-reconstruction d'un bout de la généalogie courante:



Un problème potentiel : ca donne des arbres corrélés…

Un exemple : la méthode MsVar (Beaumont 1999)

- Simulations de Monte Carlo par chaînes de Markov (MCMC)

  ✓Pour explorer l'espace des généalogies, on va ensuite explorer différentes généalogies par délétion-reconstruction d'un bout de la généalogie courante.

  ✓ parallèlement, on va aussi explorer les différentes valeurs de paramètres dans la MCMC :

  ➔À chaque pas de la MCMC:

  soit on modifie la généalogie,

  soit on modifie la valeur d'un paramètre

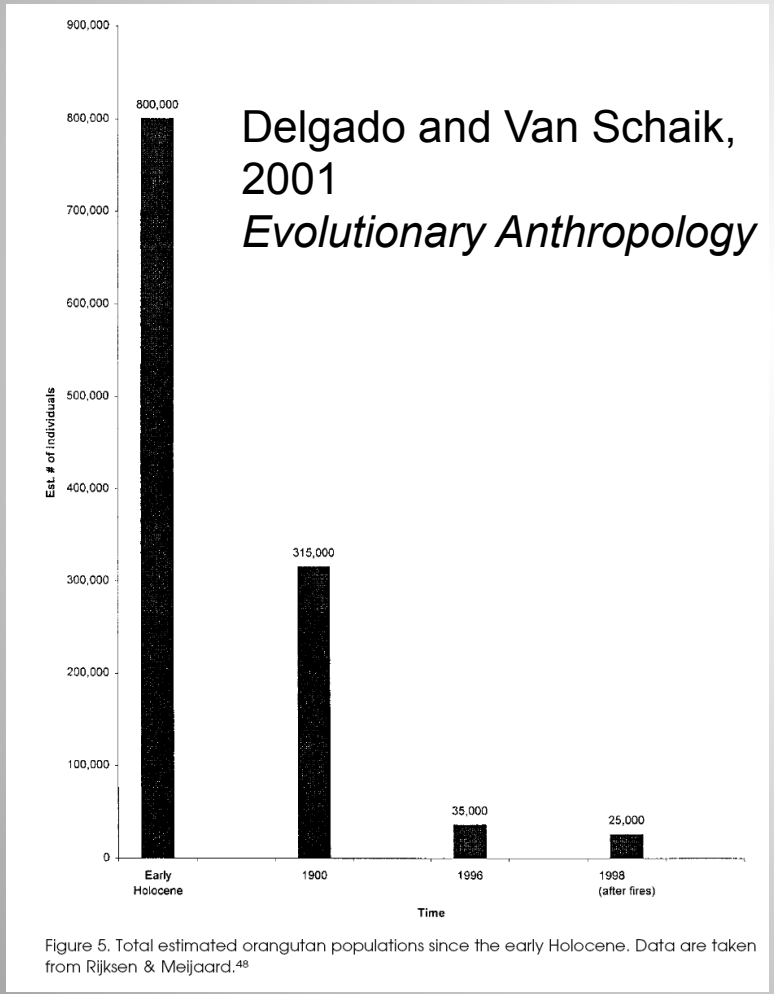
Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation

Benoît Goossens et al. (2006, Plos Biology) ont montré que le génome des Orangs-Outans est marqué par un signal d'effondrement démographique.

Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation



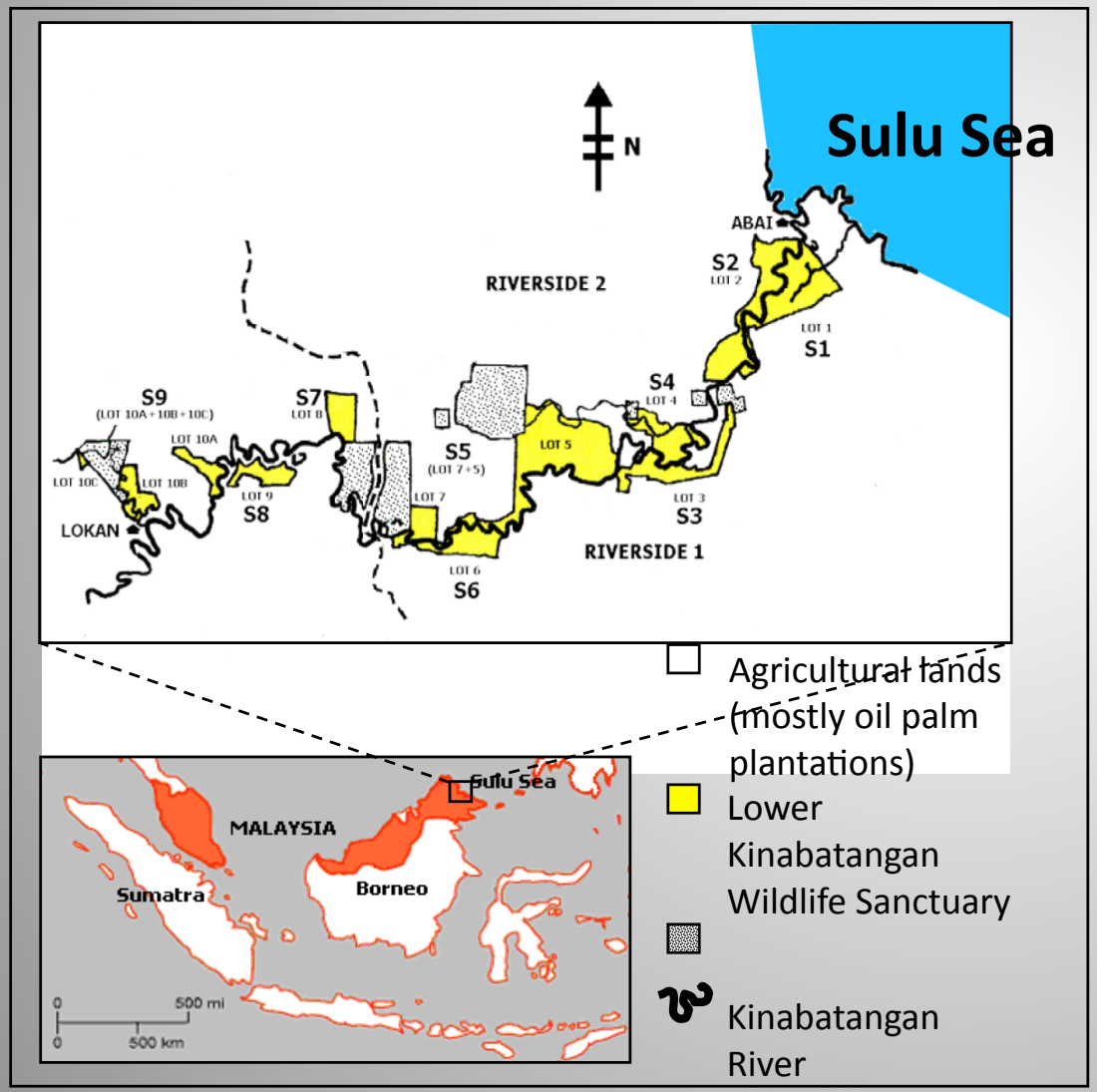
Delgado and Van Schaik, 2001
Evolutionary Anthropology

Figure 5. Total estimated orangutan populations since the early Holocene. Data are taken from Rijksen & Meijaard.[48]

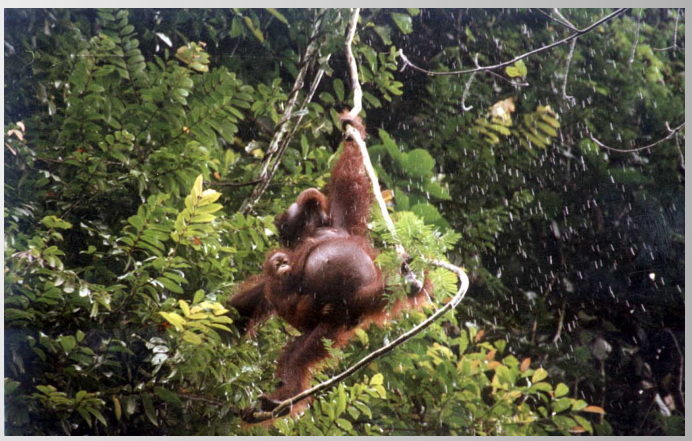
Quelle est la cause de la baisse de taille de population?
La génétique peut elle nous aider?

Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation : les données

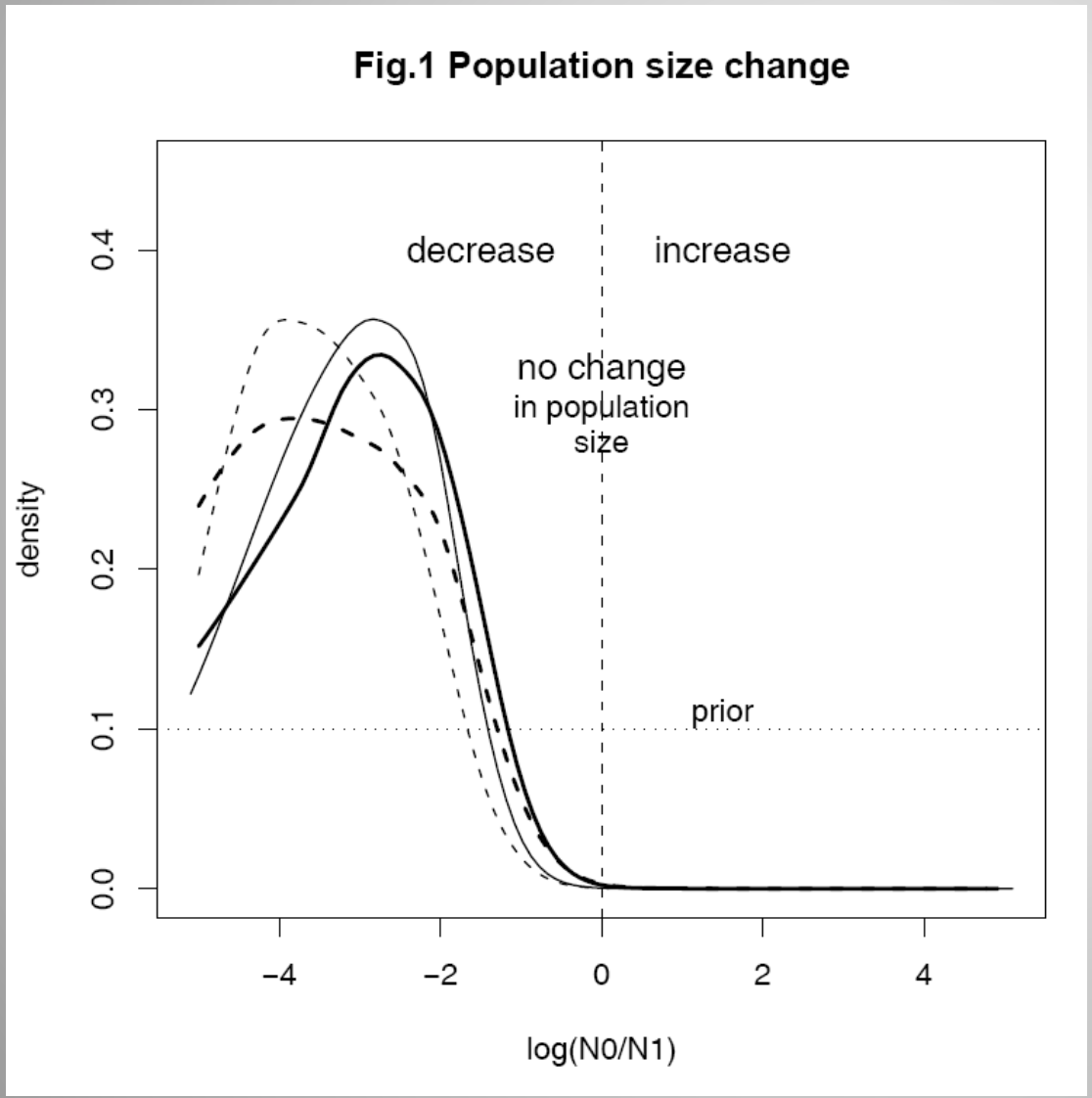


200 individus
14 locus microsatellites

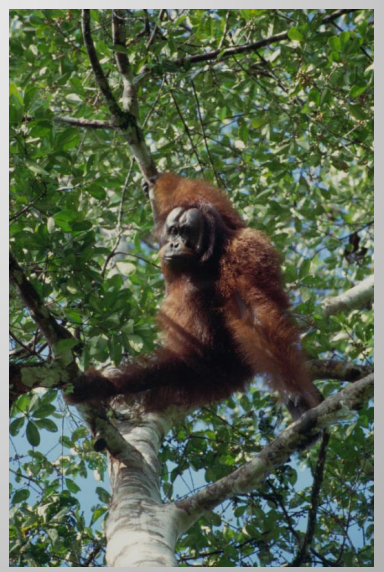


Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :



MsVar détecte bien un réduction de taille de population



Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :

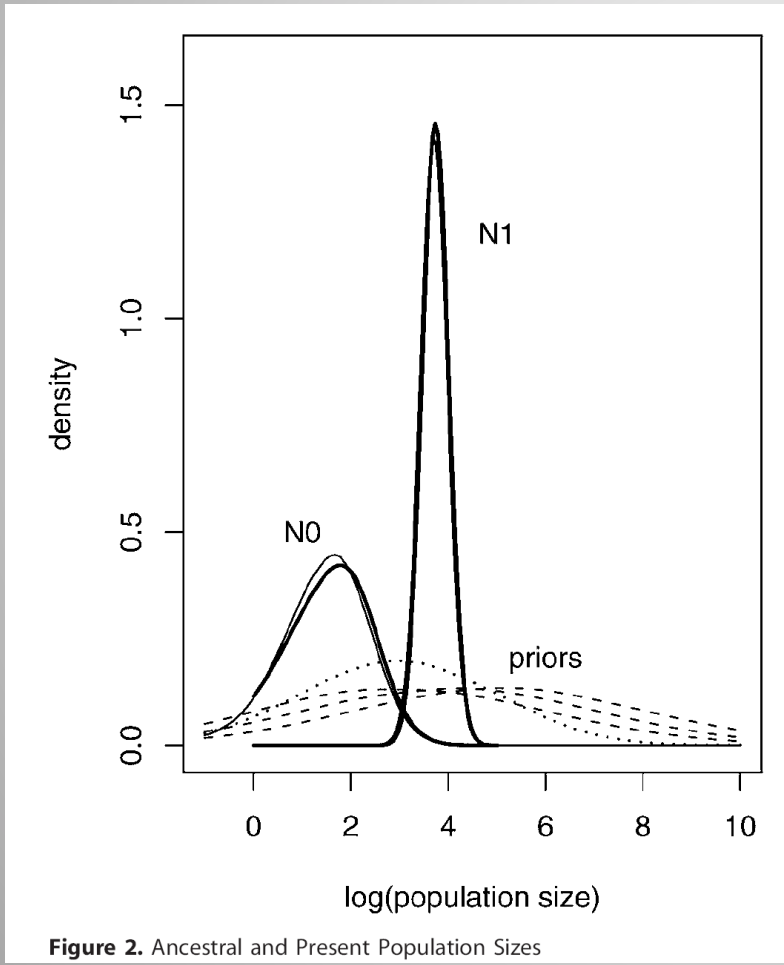
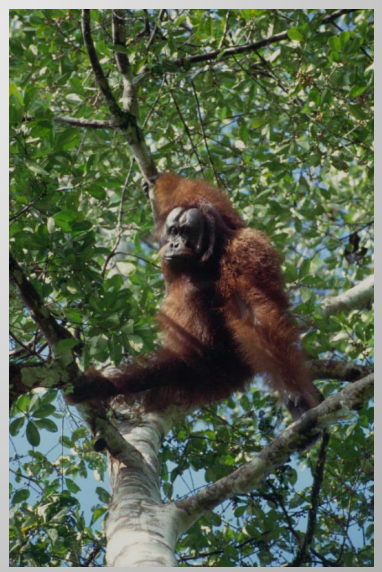


Figure 2. Ancestral and Present Population Sizes

MsVar détecte bien un réduction de taille de population



Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :

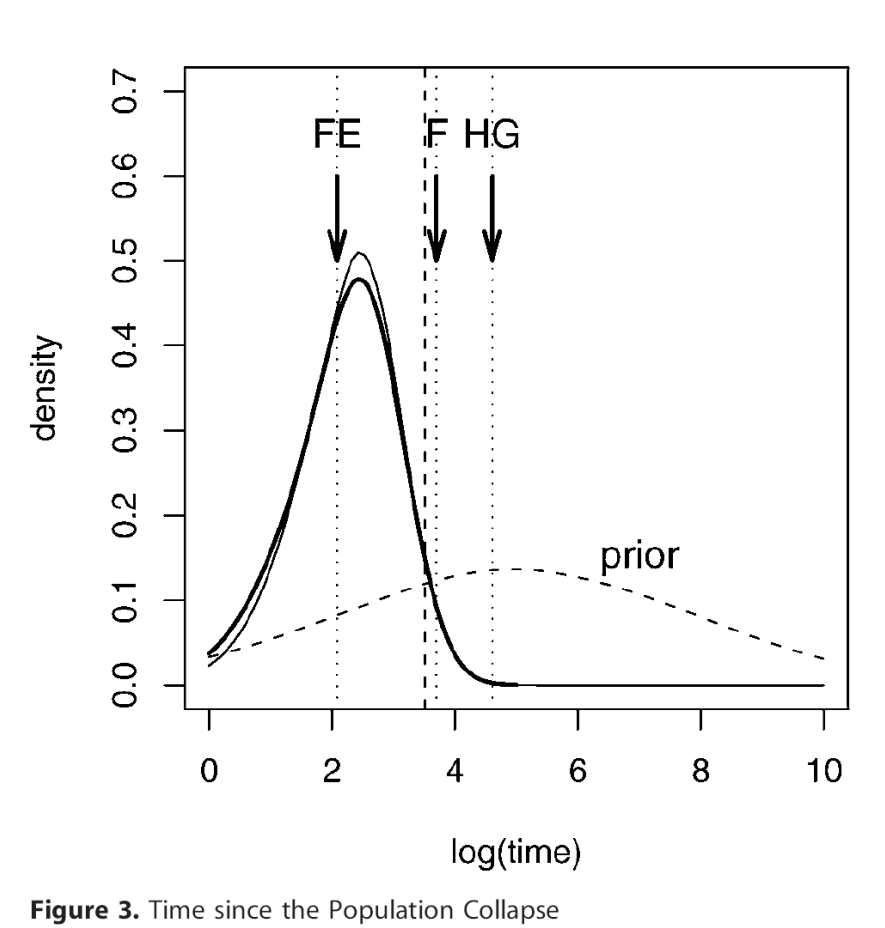


Figure 3. Time since the Population Collapse

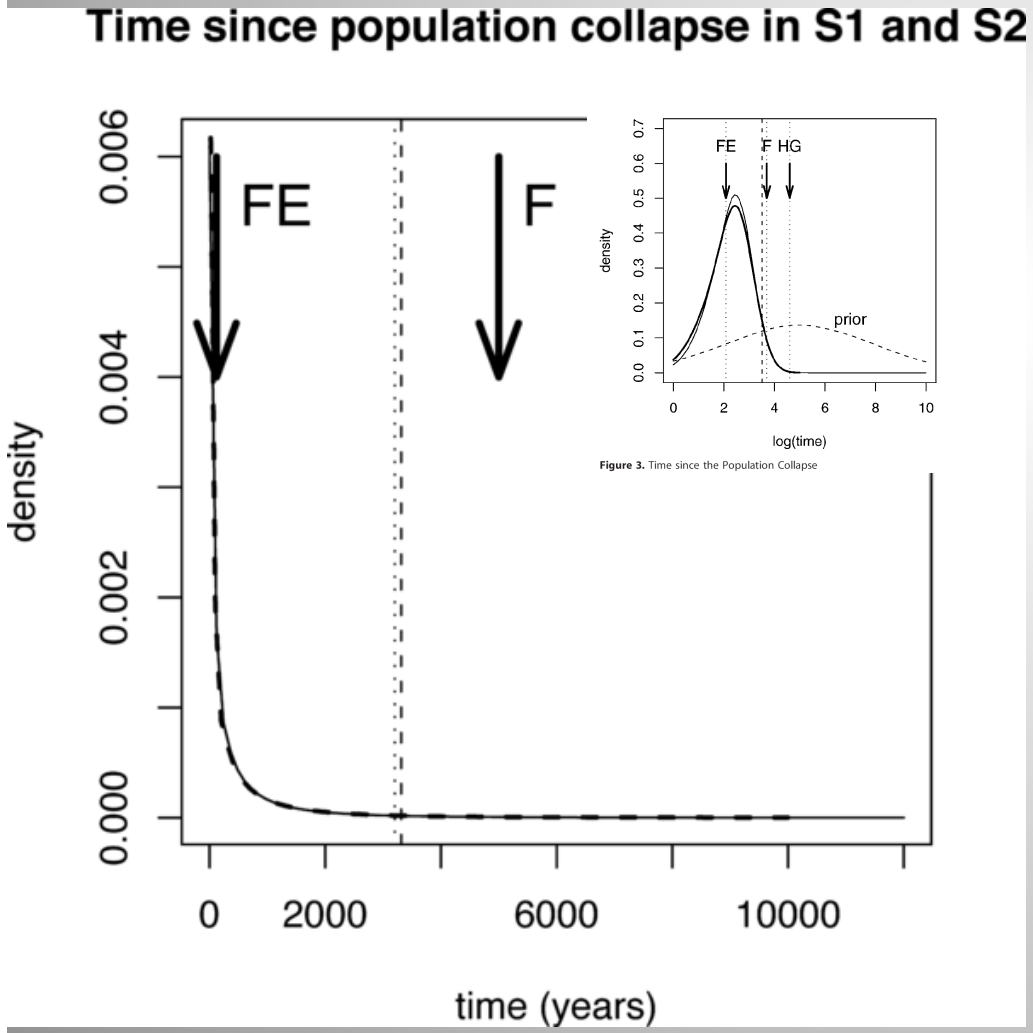
MsVar détecte bien un réduction de taille de population

et permet d'obtenir une datation

FE : Forest exploitation
F: Farmers
HG: Hunter-gatherers

Un exemple d'application de la méthode MsVar

- Les Orangs-Outans et la déforestation :



MsVar détecte bien un réduction de taille de population

et permet d'obtenir une datation: l'exploitation de la forêt semble être la cause…

FE : Forest exploitation
F: Farmers
HG: Hunter-gatherers

Approximate Bayesian Computation (ABC)

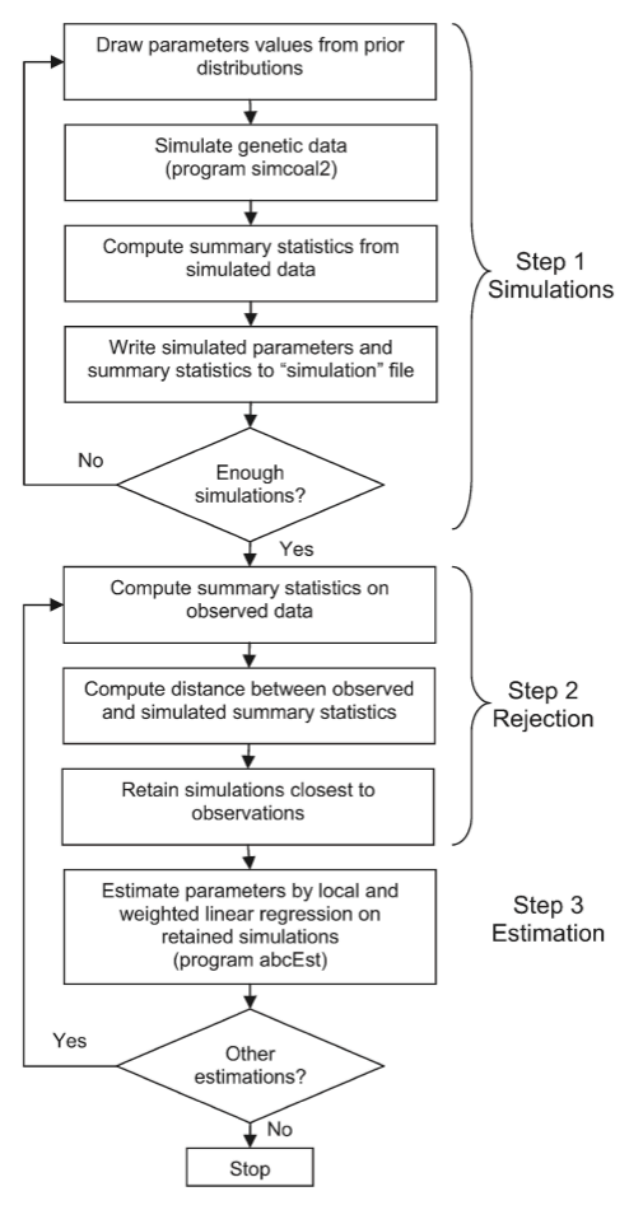
• Pour des modèles complexes, on ne sait pas estimer la vraisemblance car on a pas d'attendu pour les distributions des temps des différents évènements.

• on approxime alors P(D|P) par P(D|S) ou S est un ensemble de statistiques résumant le jeu de données.

Approximate Bayesian Computation (ABC)

• on approxime alors P(*D*|*P*) par P(*D*|*S*) ou *S* est un ensemble de statistiques résumant le jeu de données

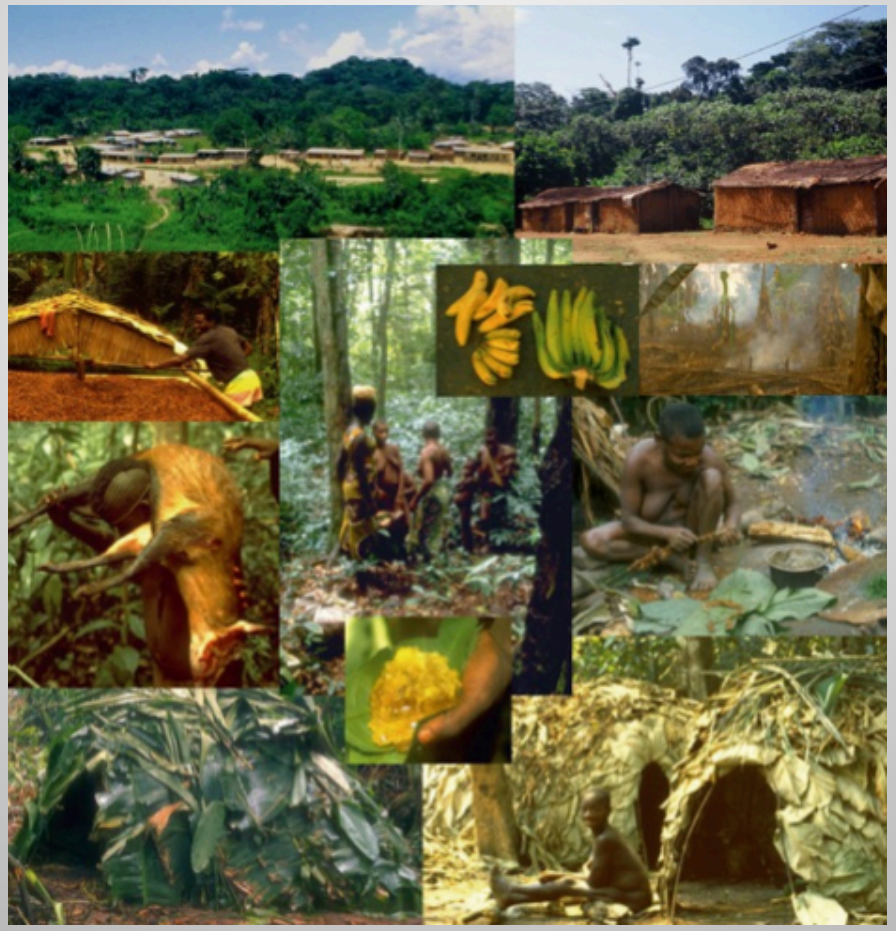
On ne calcule pas la vraisemblance mais on cherche les jeux de données simulés qui ressemblent les plus au jeu de données réel, par le biais de statistiques résumées (ex : N_a, H_e, F_{st}, etc…)



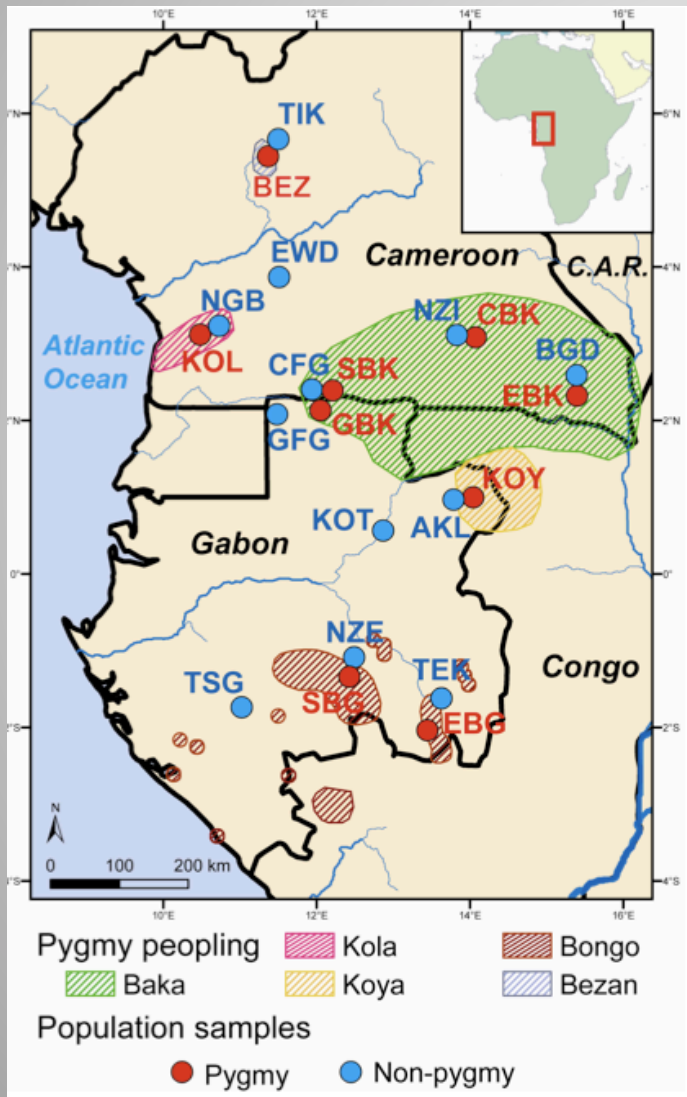
Excoffier et al. 2005 Genetics

Approximate Bayesian Computation (ABC) un exemple d'applications sur les Pygmées

Thèse de Paul Verdu (MNHN, éco-anthropologie) :
Histoire des pygmées d'Afrique de l'Ouest



Approximate Bayesian Computation (ABC) un exemple d'applications sur les Pygmées



Crédit : Serge Bahuchet

604 individus, 12 populations non-pygmées, 9 populations pygmées, 28 marqueurs microsatellites

Verdu *et al.* (2009) *Current Biology* **19**: 312-318

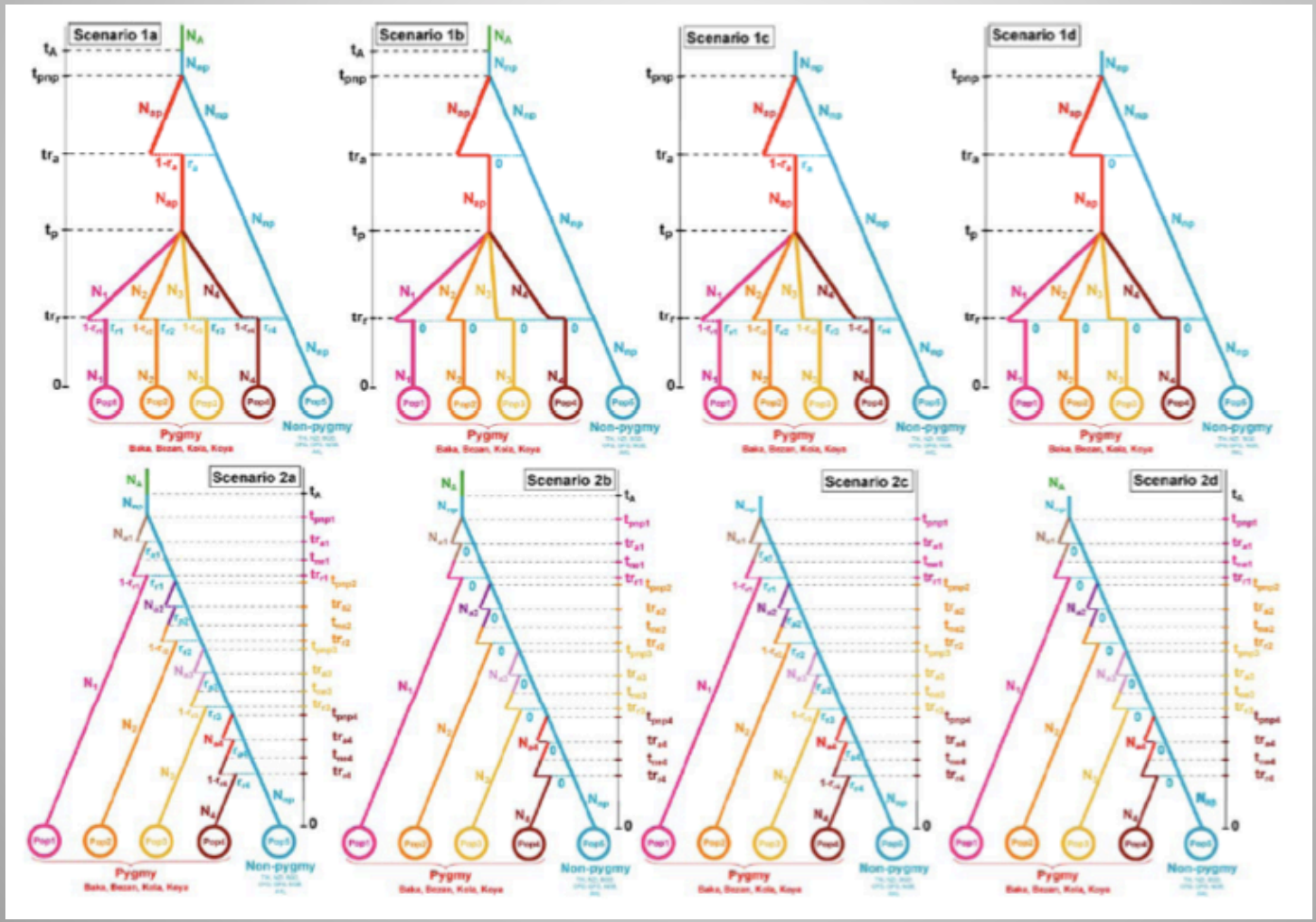
Approximate Bayesian Computation (ABC) un exemple d'applications sur les Pygmées

Les pygmées ont ils une origine commune?

Y a t il beaucoup d'échanges entre populations Pygmées et non-pygmées?

Approximate Bayesian Computation (ABC) un exemple d'applications sur les Pygmées

Différents scénarios possibles, choix de scenario par ABC

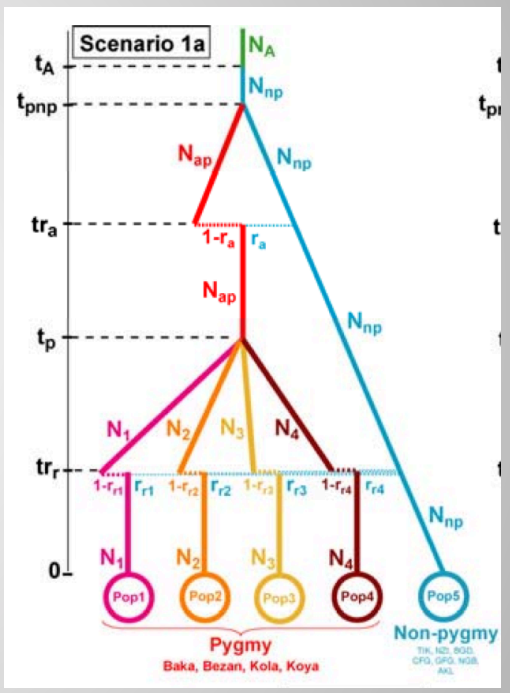


Verdu et al. 2009

Approximate Bayesian Computation (ABC) un exemple d'applications sur les Pygmées

Différents scénarios possibles, choix de scenario par ABC

Historical scenario	Prior Set 1 5,000 closest simulations	 50,000 closest simulations
Scenario 1a	0.9604 [0.9072 - 1.0000]	0.8806 [0.8518 - 0.9093]
Scenario 1b	0.0373 [0.0000 - 0.0906]	0.0994 [0.0703 - 0.1285]
Scenario 1c	0.0018 [0.0000 - 0.0036]	0.0142 [0.0111 - 0.0172]
Scenario 1d	0.0000 [0.0000 - 0.0000]	0.0010 [0.0000 - 0.0022]
Scenario 2a	0.0006 [0.0002 - 0.0009]	0.0049 [0.0041 - 0.0056]
Scenario 2b	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0000]
Scenario 2c	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0001]
Scenario 2d	0.0000 [0.0000 - 0.0000]	0.0000 [0.0000 - 0.0000]



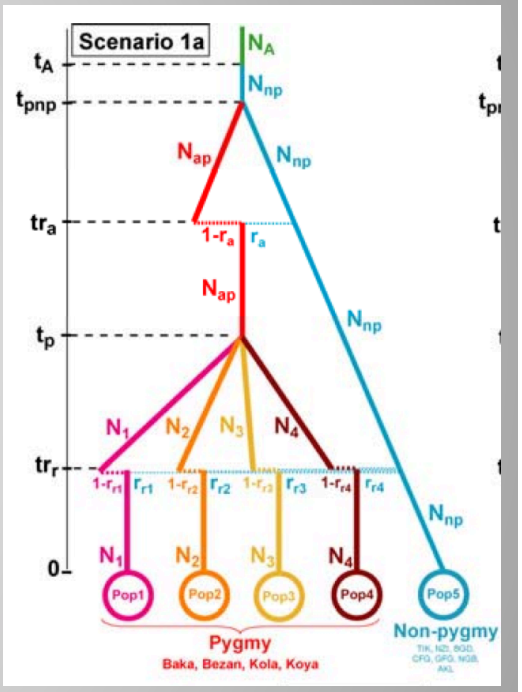
Le scenario 1a est largement soutenu par rapport aux autres ➔ plaide pour une origine commune des populations pygmées d'Afrique de l'Ouest

Verdu et al. 2009

Approximate Bayesian Computation (ABC) un exemple d'applications sur les Pygmées

Estimation des paramètres sous le scénario le plus probable

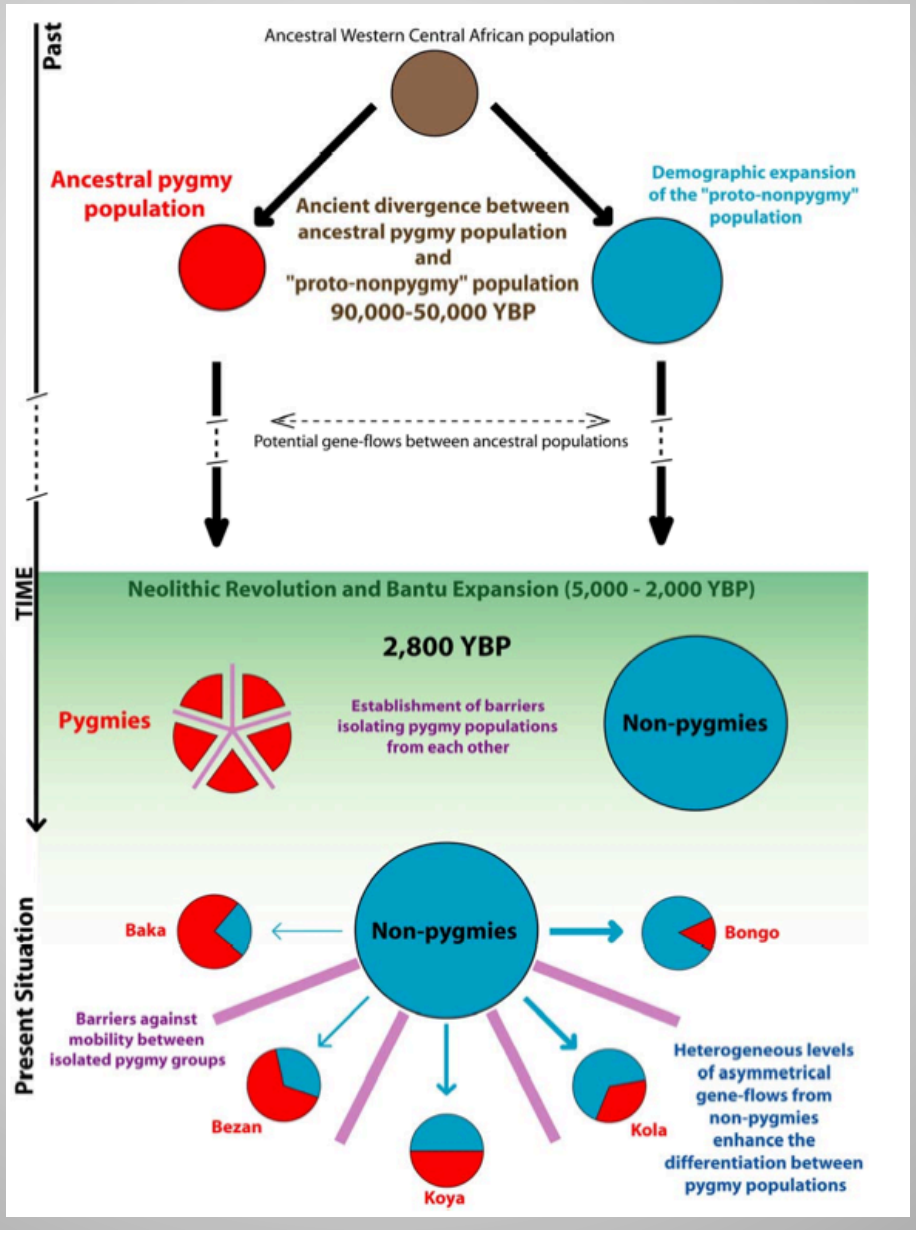
Parameter	mean	median	mode	quantile 2.5%	quantile 97.5%
Original Parameters					
N_1 (Baka)	6,164	6,368	8,137	1,347	9,824
N_2 (Bezan)	5,055	4,840	2,795	790	9,677
N_3 (Kola)	4,486	4,100	3,302	603	9,599
N_4 (Koya)	5,608	5,619	3,197	1,134	9,771
N_{np} (Non-pygmies)	66,265	67,168	77,157	27,926	97,828
N_{ap}	5,901	6,163	8,007	960	9,825
N_A	3,074	2,631	1,071	202	8,404
tr_r	115	67	8	4	485
t_p	364	256	105	29	1,371
tr_a	1,353	1118	771	212	3,749
t_{pnp}	3,101	3170	3,587	921	4,913
t_A	4,217	3,740	2,802	663	9,419
r_{r1}	0.662	0.674	0.696	0.261	0.957
r_{r2}	0.461	0.440	0.416	0.098	0.899
r_{r3}	0.647	0.662	0.672	0.219	0.955
r_{r4}	0.523	0.514	0.465	0.147	0.920
r_a	0.572	0.605	0.927	0.041	0.982
$\overline{\mu}$	0.00024	0.00021	0.00016	0.00011	0.00056
$\overline{p}$	0.11	0.11	0.10	0.10	0.16



Verdu et al. 2009

Approximate Bayesian Computation (ABC) un exemple d'applications sur les Pygmées

Scénario évolutif :
on « raconte » une histoire à partir de ces inférences



Verdu et al. 2009

Estimation de paramètres démographiques en populations subdivisées

1. Introduction

2. Robustesse d'une méthode fondée sur F_{ST}

3. Estimation par maximum de vraisemblance
   i. MCMC : Test de MIGRATE
      i. Jeu de donnée réel
      ii. Par simulation

   iii. IS : Test de la méthode de Griffiths et al.
      - Résultats préliminaires

4. Conclusions générales et perspectives

100

L'approche de Griffiths et al.

- Nath et Griffiths (1996) algorithme pour populations subdivisées

- Extension de l'algorithme de Stephens et Donnelly (2001) pour des populations subdivisées

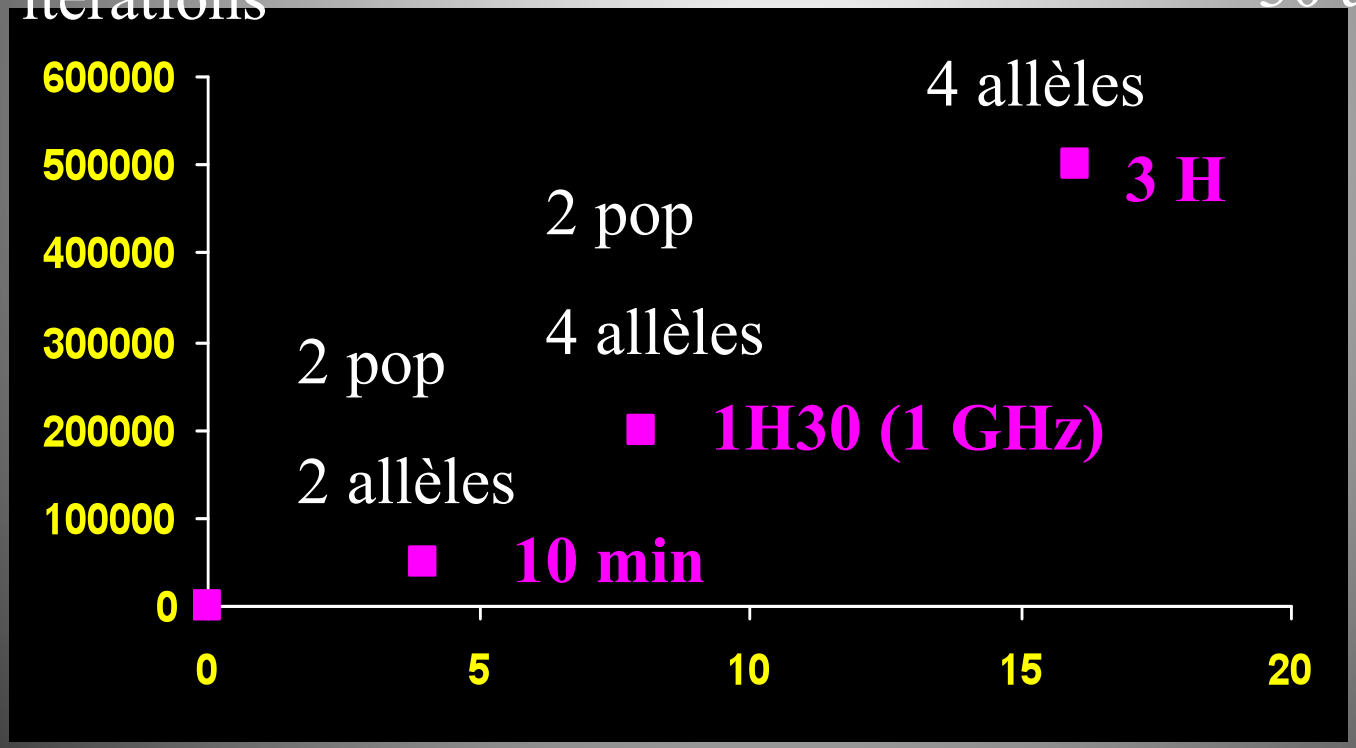
  -> De Iorio et Griffiths (2004a, b)

101

Temps de calcul et complexité des modèles avec l'algorithme de Nath et Griffiths (1996)

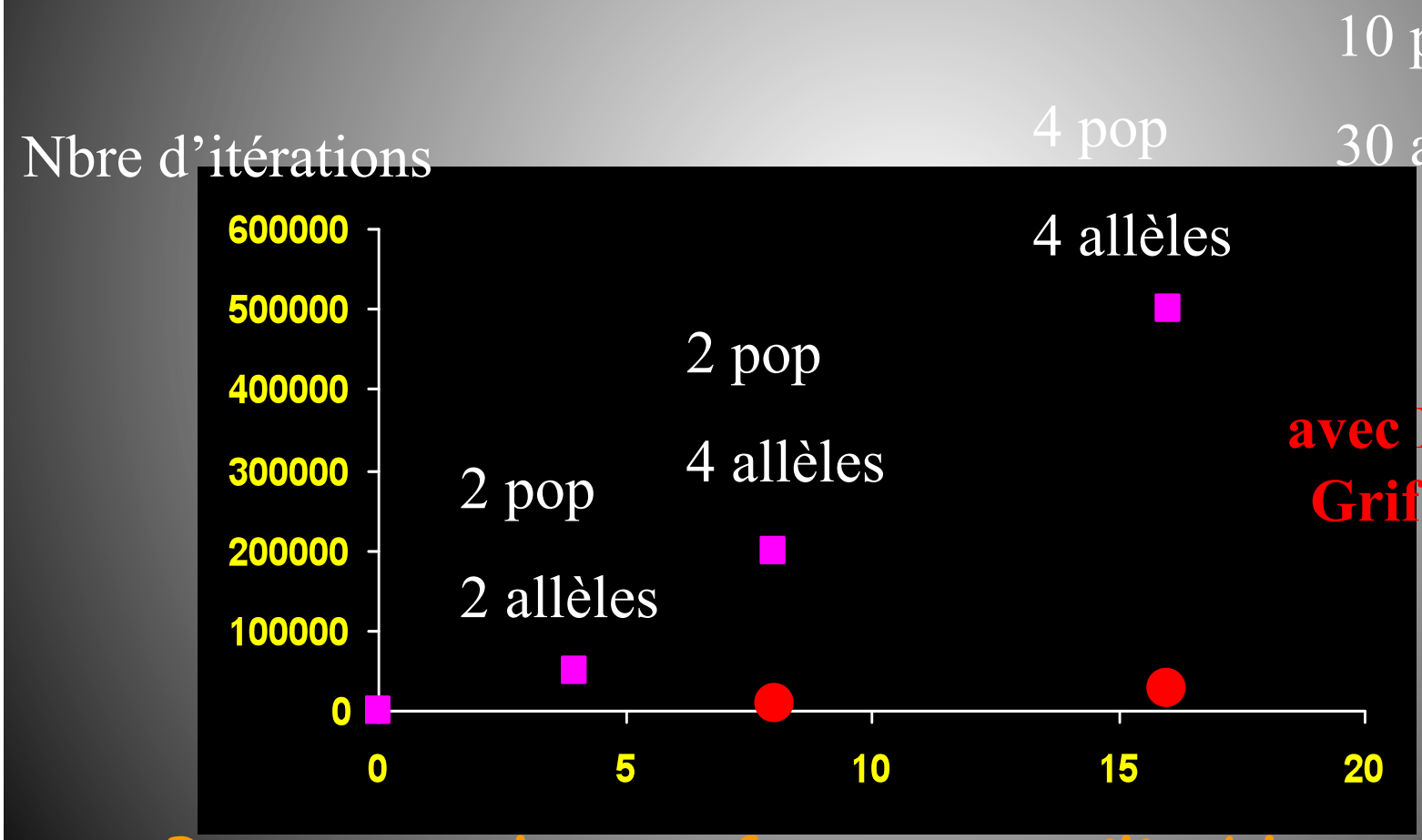
10 pop → ?

30 allèles ?



102

Temps de calcul et complexité des modèles avec l'algorithme de De Iorio et Griffiths (sous presse)



→ Beaucoup plus performant, utilisable en pratique

…mais uniquement pour mutations indépendantes du type parental (PIM=KAM) et modèle de migration simples (en îles)… plus complexe pour d'autres modèles

- "solutions actuelles":
  - N pop, modèle en îles, PIM (DG 2004b)
  - 2 pop, mutation par pas (SMM) (DeIorio et al. 2005 TPB)
  - Isolement par la distance linéaire (1 dimension), PIM (Rousset & Leblois soumis) cf 2ème partie du cours
- "en cours de développement" :
  - Divergence de 2 pop avec migration, PIM
  - Isolement par la distance en 2 dimensions, PIM

Tout cela sera implémenté dans le logiciel MIGRAINE (Rousset & Leblois)

104

2 populations, mutation par pas (SMM)

De Iorio, Griffiths, Leblois, Rousset, 2005 TPB

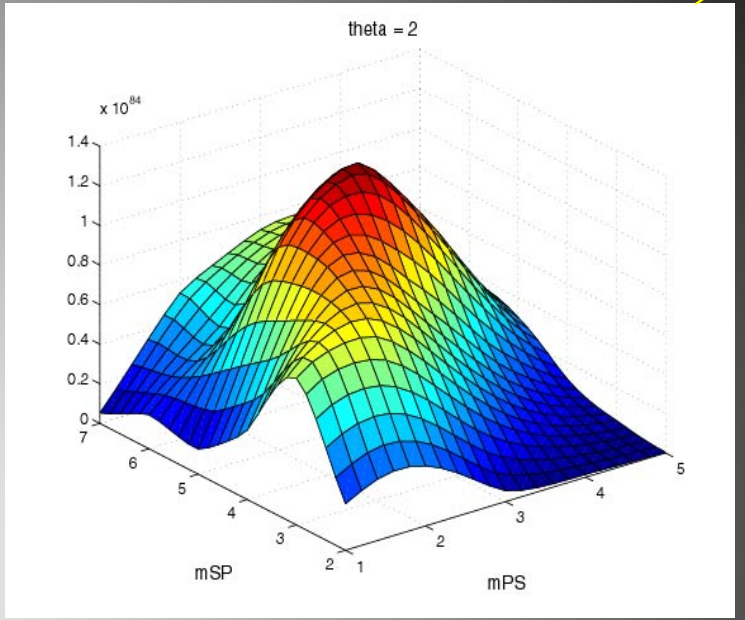
- Cas spécial de De Iorio & Griffiths (2004a): résolu par transformée de Fourier

- Résultats préliminaires :
  - Un jeu de données réel (renard)
  - Quelques simulations

Australian Red Fox (Lade *et al. 1996*)

- DATA :
  - 2 populations (Island, Mainland)
  - 7 microsatellites
- MODEL :
  - Single step mutation (SMM)
  - 3 parameter estimation ($\theta=4N\mu$, $4N_M m_{MI}$,$4N_I m_{IM}$)
  - 1 million runs for 30 parameter sets (θ_i, $4N_M m_{MIi}$, $4N_I m_{Ii}$) (~few days on 1Ghz)

Australian Red Fox (Lade *et al. 1996*) Results



- Good convergence between independent runs
- MLE : $4N_M m_{MI} = 4.0$

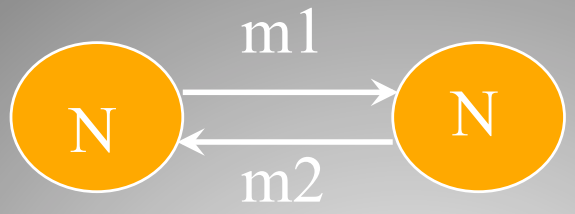
  $4N_I m_{IM} = 3.0$

- For comparison :
  - F_{ST}-> $4Nm \sim 3.0$ (R_{ST} -> $4Nm \sim 7.4$)
  - MIGRATE :

$4N_M m_{MI}$ =[2.3-3.0-3.8-1.5] $4N_I m_{IM}$ =[1.4-3.6-2.8-1.0]

(large variance between runs with different starting values)

Tests par Simulation



γ =4Nm

- ✓ 2 populations (N=1000, même θ=4Nμ=2.0)
- ✓ Migration symétrique (4Nm1=4Nm2=2.0)
- ✓ Mutation par pas (SMM)
- ✓ 30 individus pour 5 et 20 locus
- ✓ 10 jeux de données (1 mois sur 50 processeurs 1 GHz!!)

108

Résultats des simulations

estimation du paramètre de migration γ =4Nm

- IS : Griffiths et al.
- MCMC : MIGRATE

à temps de calcul comparables

- ✓5 locus
  - Biais relatif=0.6
  - MSE=2.2
- ✓20 locus
  - Biais relatif=0.5
  - MSE=1.2
- ✓5 locus
  - Biais relatif=2.38
  - MSE=12.5
- ✓20 locus
  - Biais relatif=0.5
  - MSE=2.6

109

…à ce stade, beaucoup de problèmes persistent pour le MV…

- Temps de calcul (IS et MCMC) très long mais amélioration récentes réduisant les temps de calculs pour IS
- Surestimation (inhérente aux méthodes?)

110

…à ce stade, beaucoup de problèmes persistent pour le MV…

- Temps de calcul (IS et MCMC)
- Surestimation (inhérente aux algorithmes?)

Il faut encore tester l'effet de:

- Nombre de populations échantillonnées vs nombre total de sous populations
- Processus mutationnels complexes des locus microsatellites (déviations du modèle par pas strict)
- Effet de fluctuations démographiques passées

111