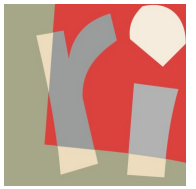
Génomique des populations
Démographie

Flora Jay
flora.jay@lri.fr

29 septembre 2016



Pourquoi étudier la diversité génétique ?

2 objectifs

(1) Décrire la diversité actuelle

- *patrons* de diversité génétique
- distance entre espèces, entre populations, …

(2) Comprendre les évolutions passées y menant

- intérêt écologique/anthropologique : histoire des populations …
- évolution d'un segment ADN en particulier, sélection ...

Rappel : forces évolutives

MUTATION
RECOMBINAISON

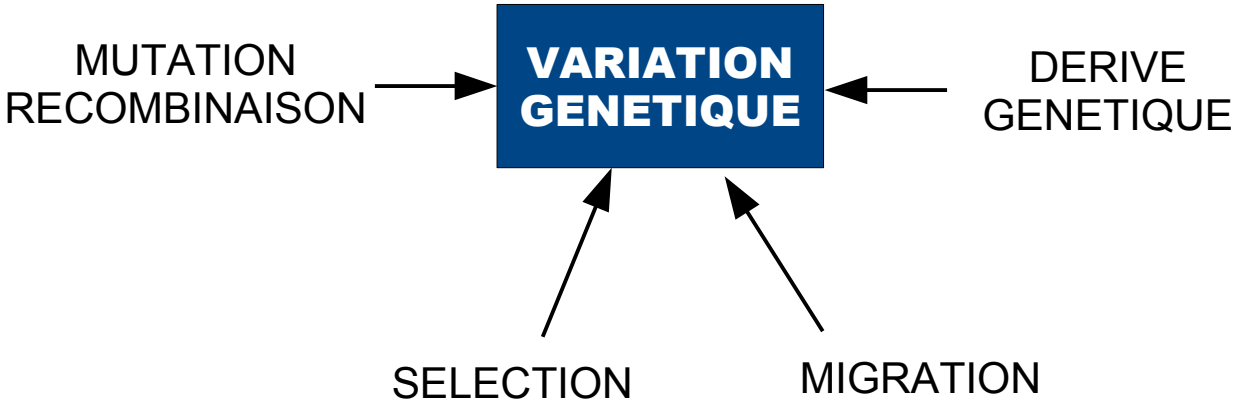
VARIATION GENETIQUE

DERIVE
GENETIQUE

SELECTION

MIGRATION

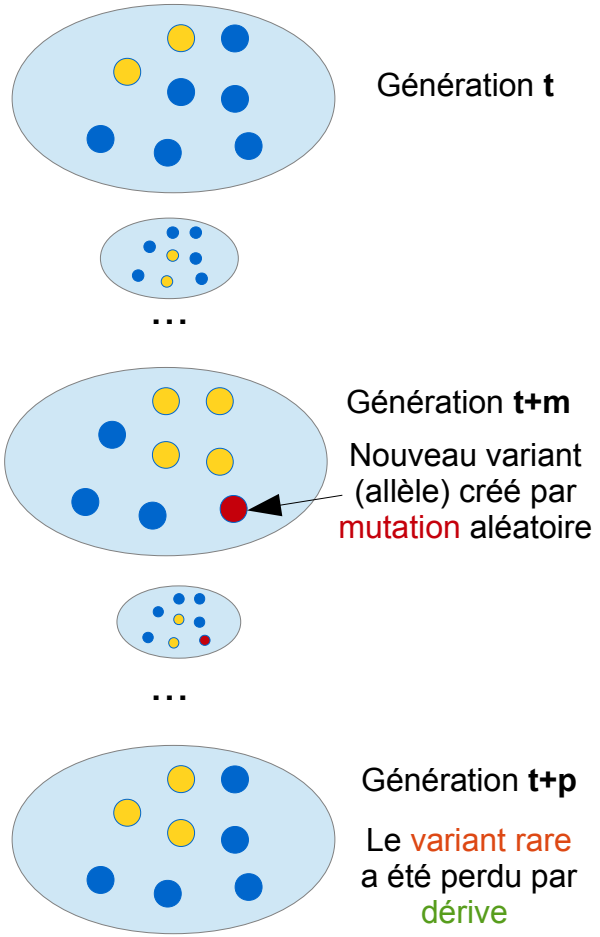
Rappel : forces évolutives



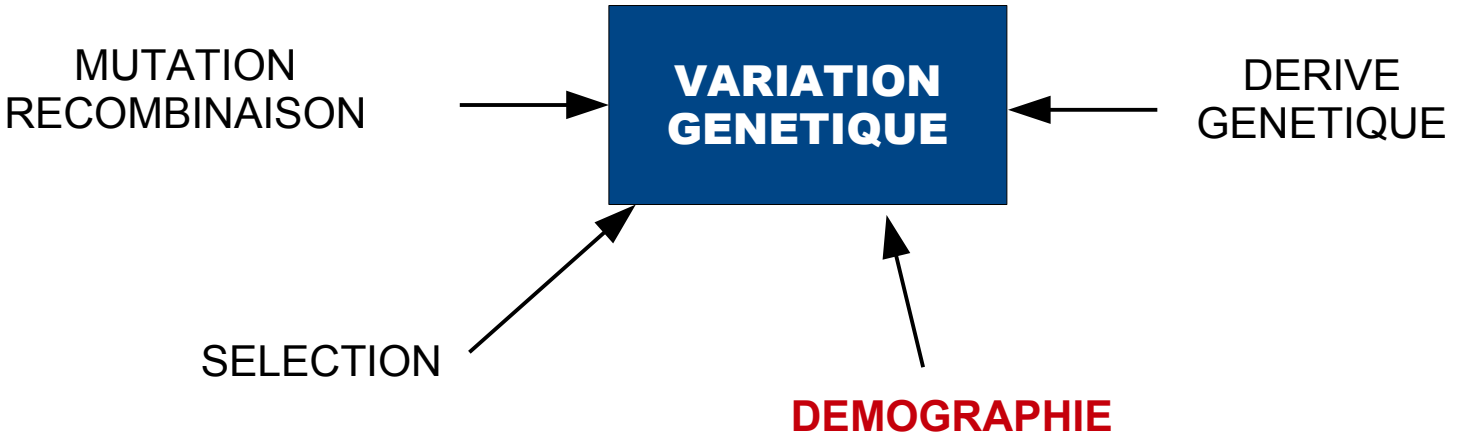
Mutation + recombinaison ↗ ⇒ Diversité génétique ↗

Dérive (processus aléatoire) ↗ ⇒ Diversité génétique ↘

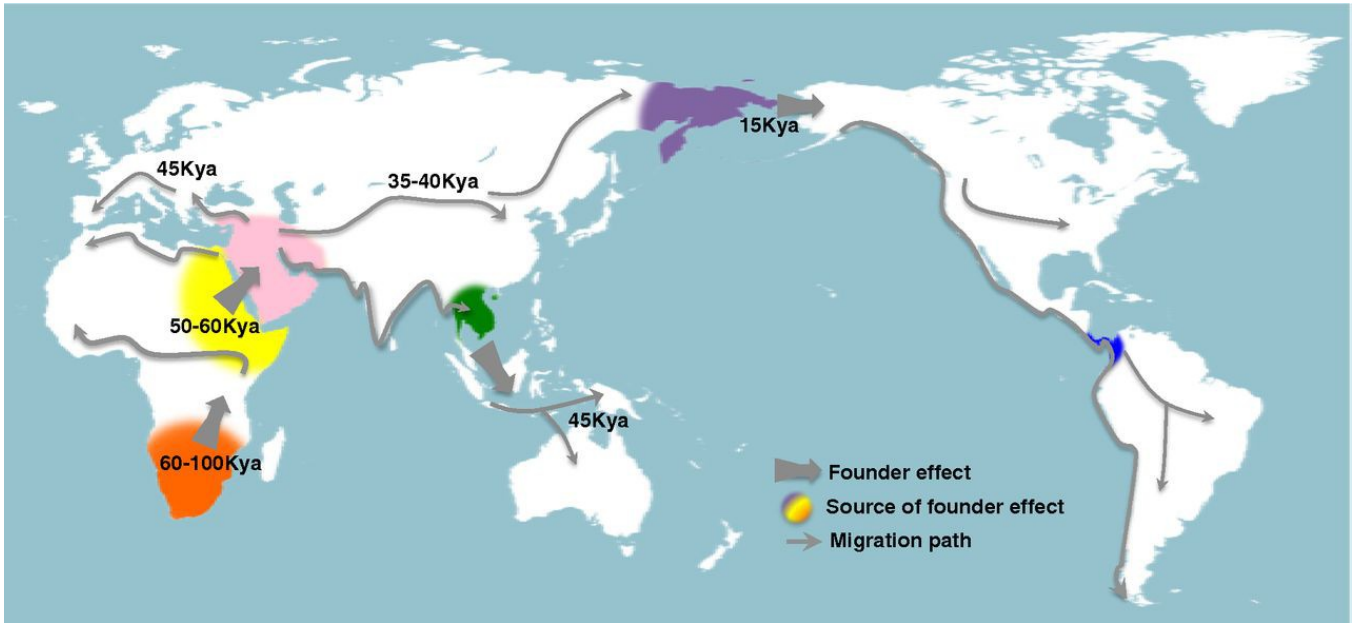
Exemple



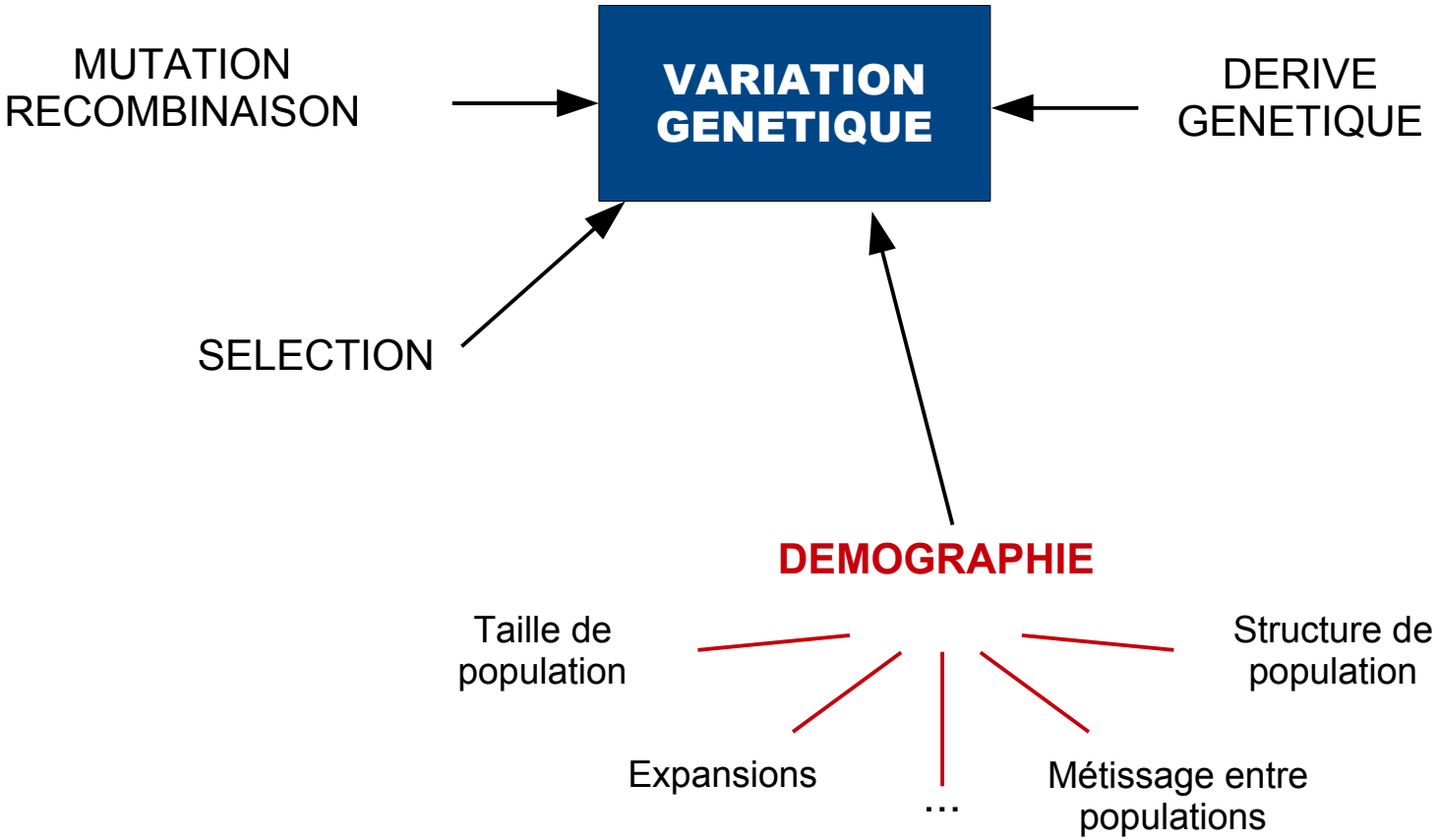
Diversité génétique et démographie



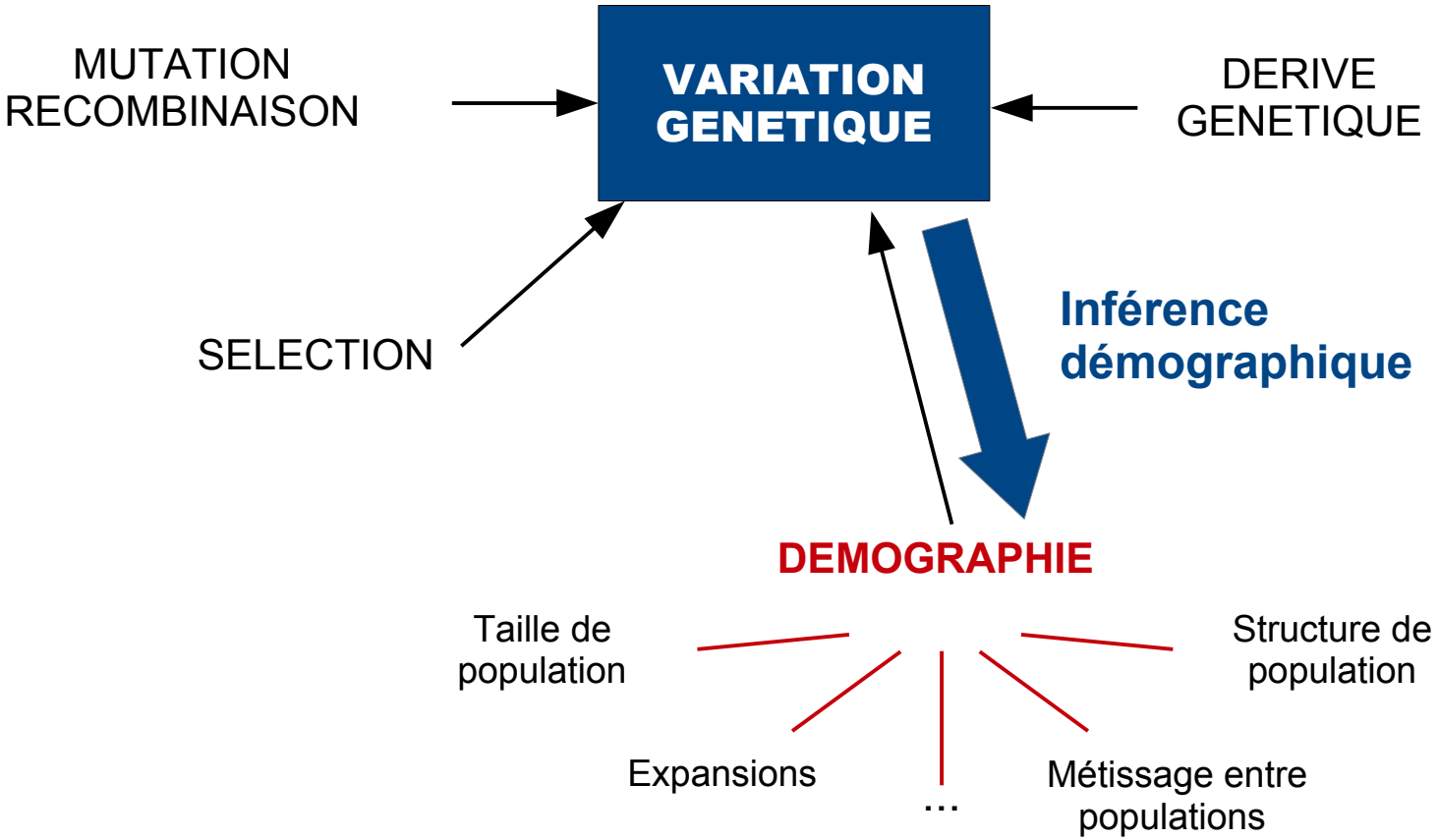
Eg. colonisation de la planète



Diversité génétique et démographie



Diversité génétique et démographie

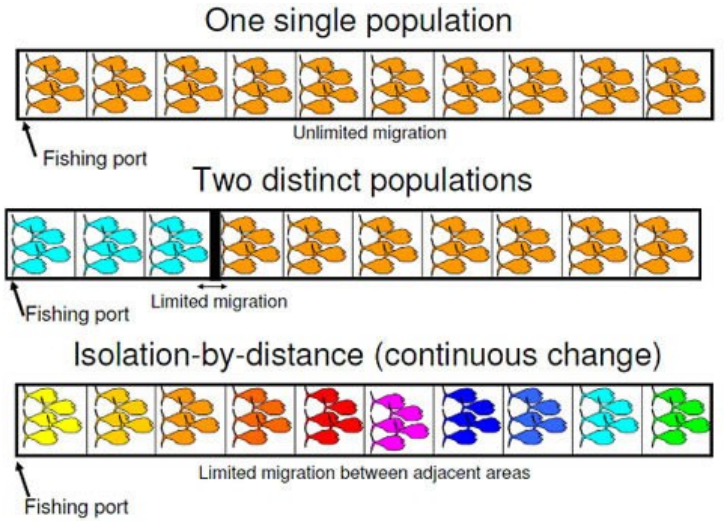


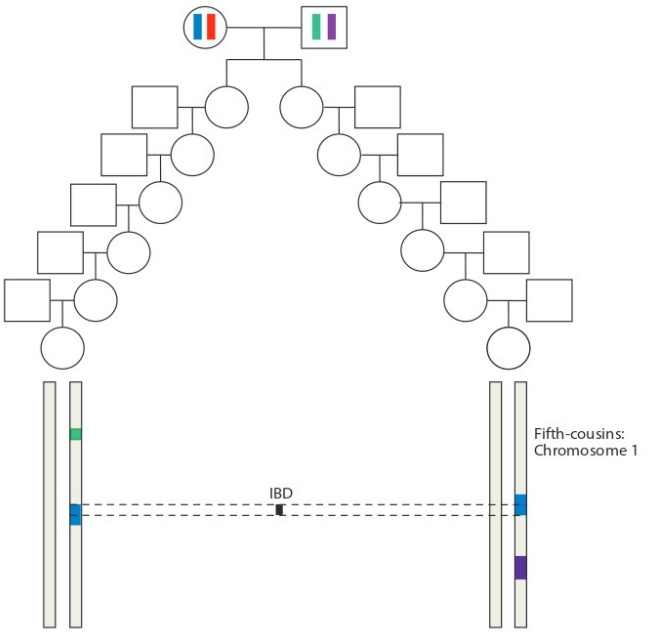
INFERENCE =
Identifier les événements,
les dater,
estimer leurs caractéristiques principales (eg : taux de migration)

Quelles questions démographiques ?

?

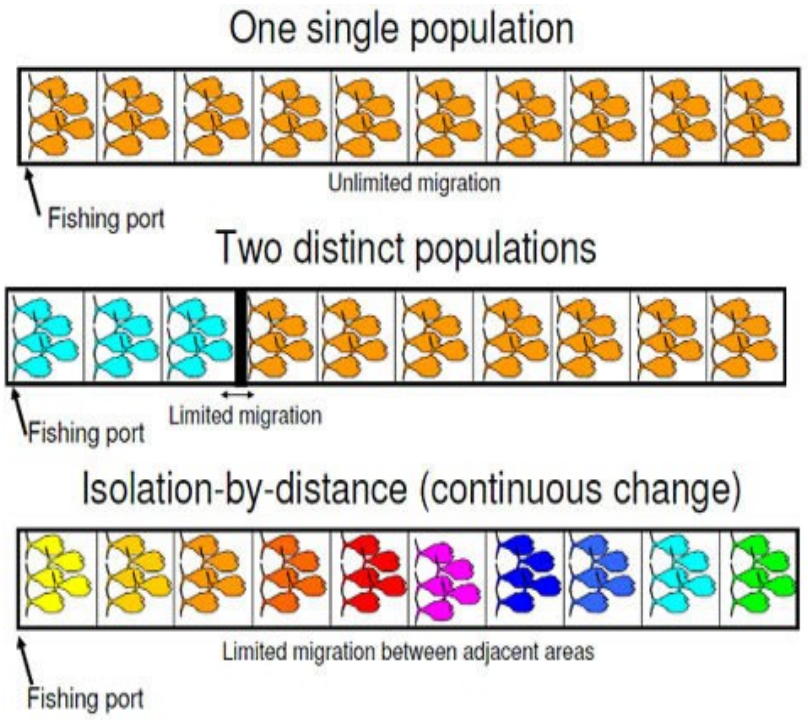
Quelles questions démographiques ?

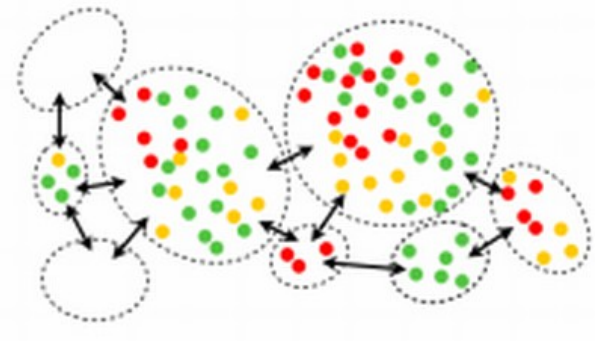


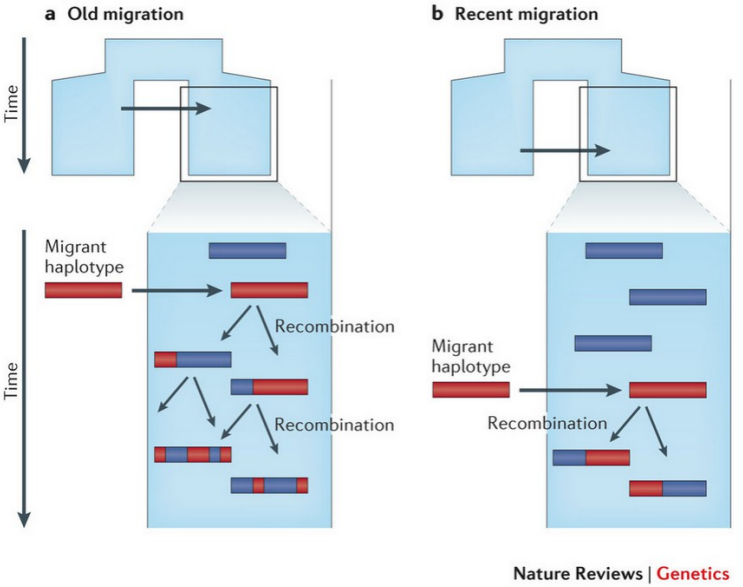


Browning & Browning Annu. Rev. Genet. 2012

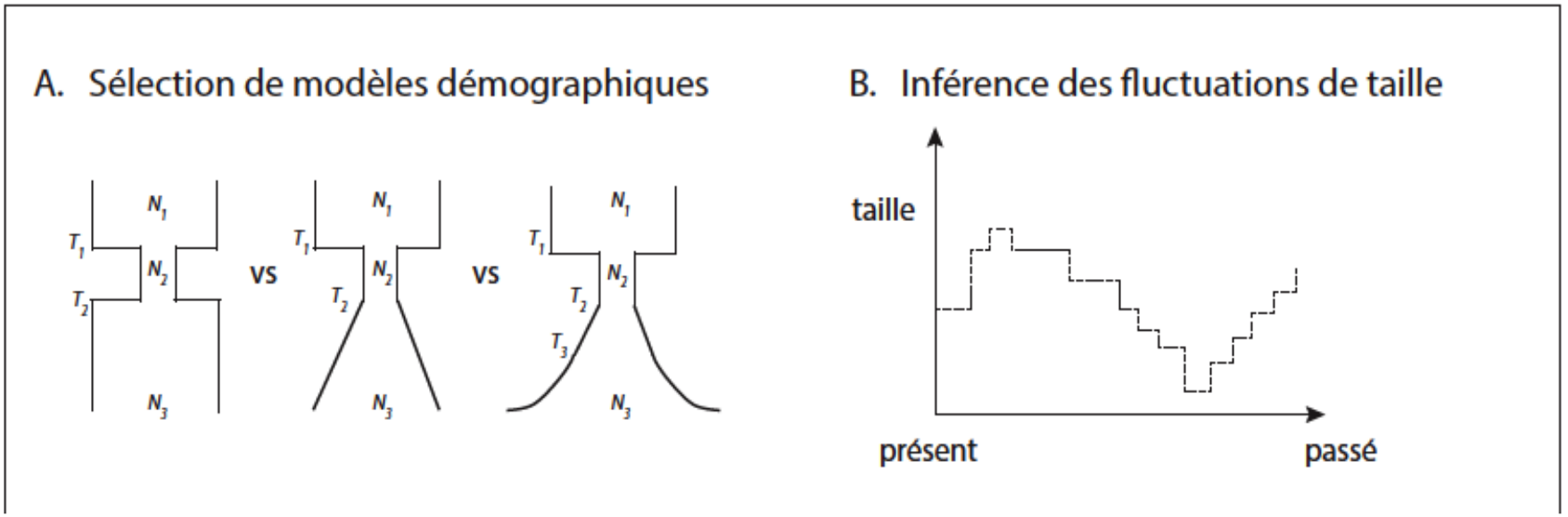
Modèles démographiques
Structuration, divergence, migration



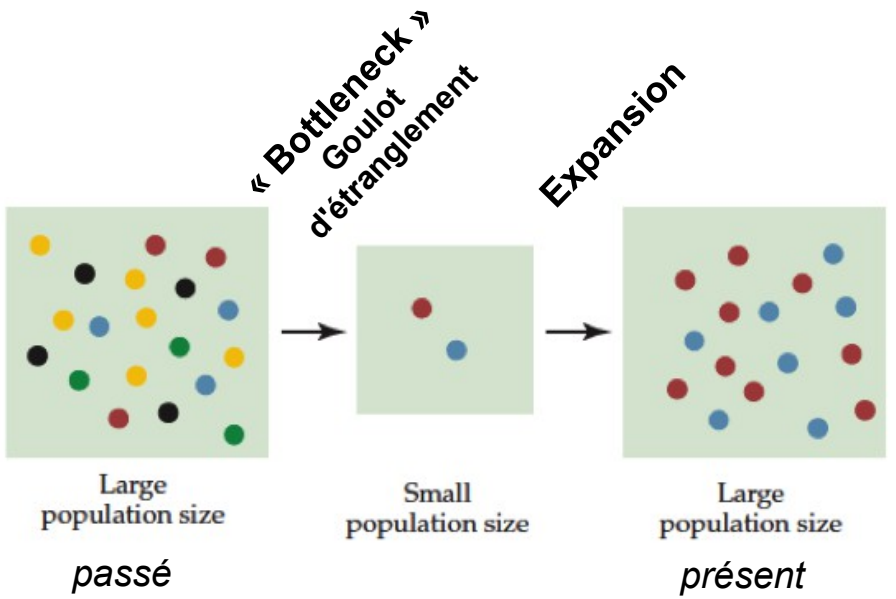




Modèles démographiques
Tailles de population



Signal génétique :



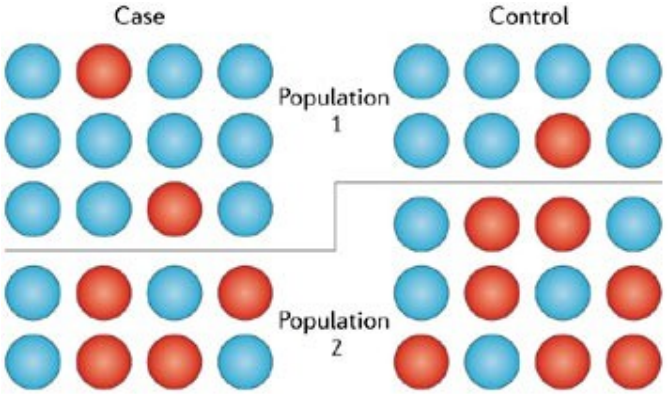
© Slatkin

Connaître la démographe, intéressant pour

- Les amoureux d'histoire

- Les fans du médical
  ex : effet confondant de la structure de population

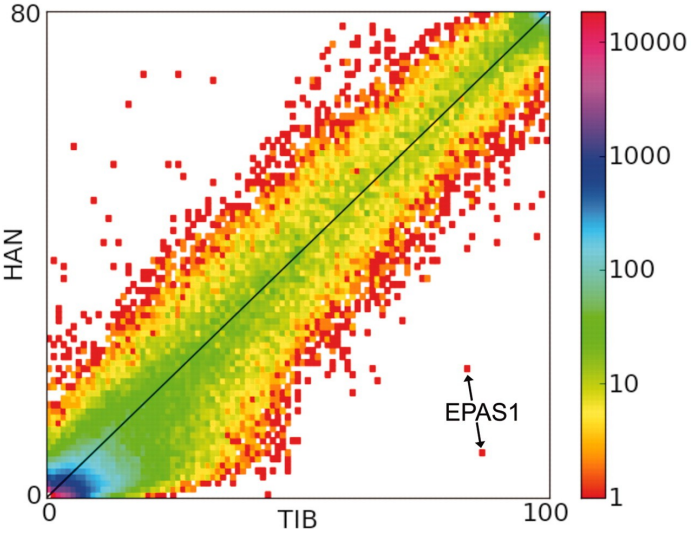
Balding



- Les adeptes de la sélection/adaptation
  démo = référence neutre
  ex : adaptation à l'altitude des Tibétains

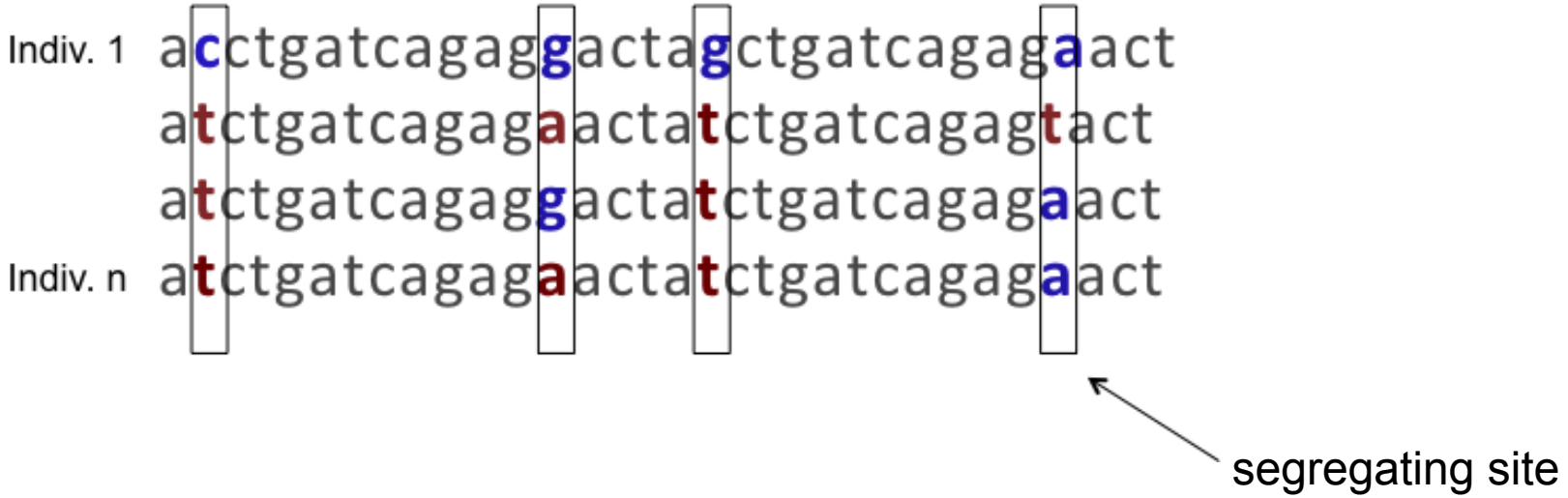
Yi et al 2010

ex : bottleneck et sélection durs à distinguer si on ne se concentre que sur un gène



ON OBSERVE

- SNP = single nucleotide polymorphism



Modèles mathématiques

(Wright-Fisher)
et le coalescent

- But : approximation du phénomène biologique de transmission du matériel génétique de génération en génération

  → modèles expliquant les observations

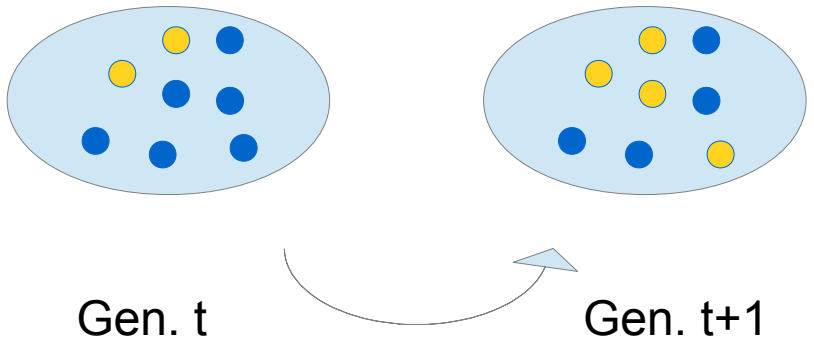
- Construire des estimateurs de la taille de population, (ou d'autres paramètres démographiques)

Modèle Wright-Fisher

Hypothèses principales

- Générations non chevauchantes
- Panmixie (reproduction aleatoire)
- Taille constante

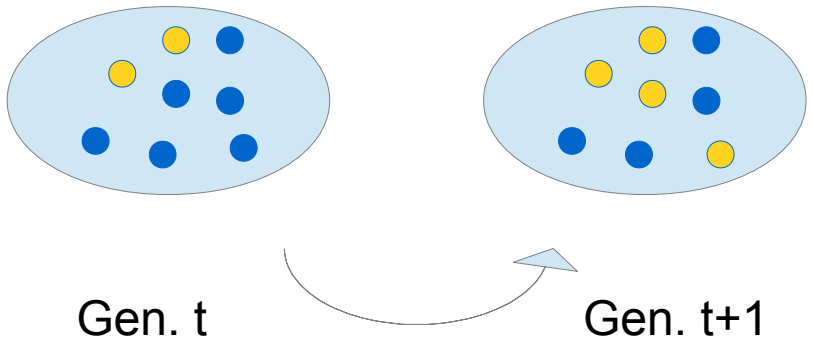
- Pas de sélection
- Pas de recombinaison



Modèle Wright-Fisher

Hypothèses principales

- Générations non chevauchantes
- Panmixie (reproduction aleatoire)
- Taille constante



X[t+1] = nb alleles jaunes
~ Binomial (N, p= fréquence à t)
~ Binomial (N, p= X[t]/N)

Exercice 1 (plus tard/chez vous)

(1) Implémenter une fonction genetic_drift qui simule la dérive génétique pour g générations à une position polymorphique ayant 2 allèles a/A

Arguments :

N taille de population

p fréquence initiale de A

g nombre de generations

Return :

vecteur de taille g+1 contenant la fréquence allélique de A à chaque pas de temps (le 1er élément étant la fréquence initiale)

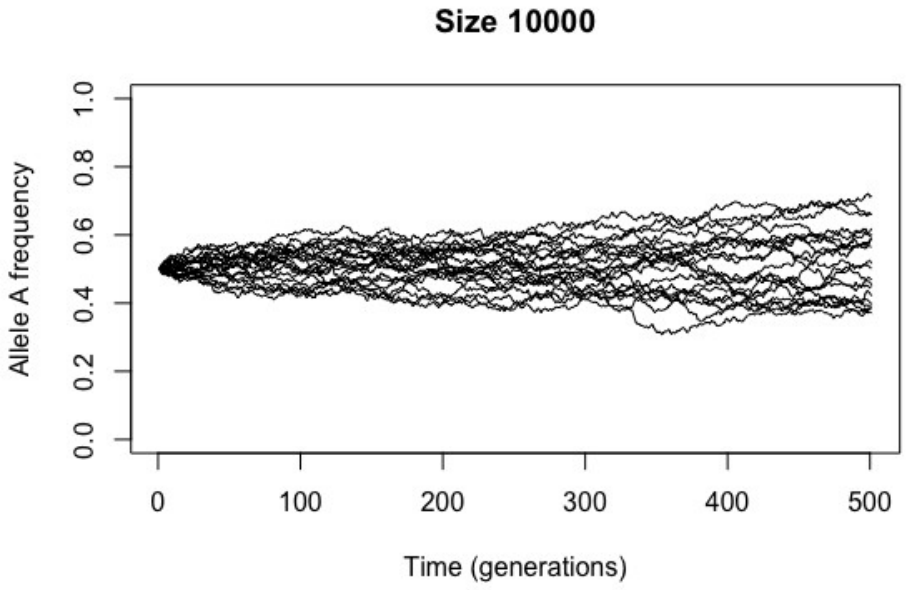
Indice ?rbinom

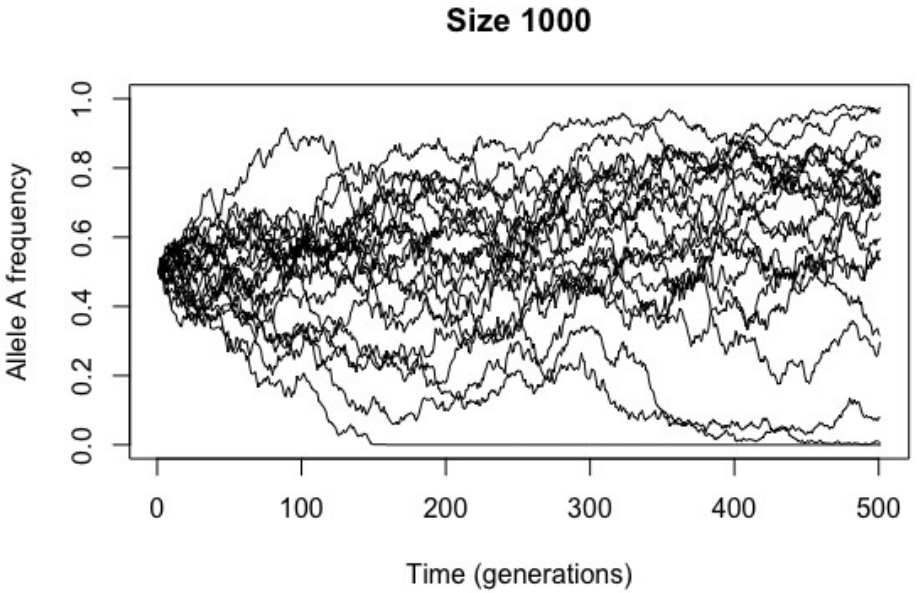
(2) a. Tracer plusieurs courbes montrant la dérive génétique pendant 500 générations avec pour fréquence initiale p=0.5 et comme taille de population n=10000
b. Idem pour n=100. Discuter la différence des résultats

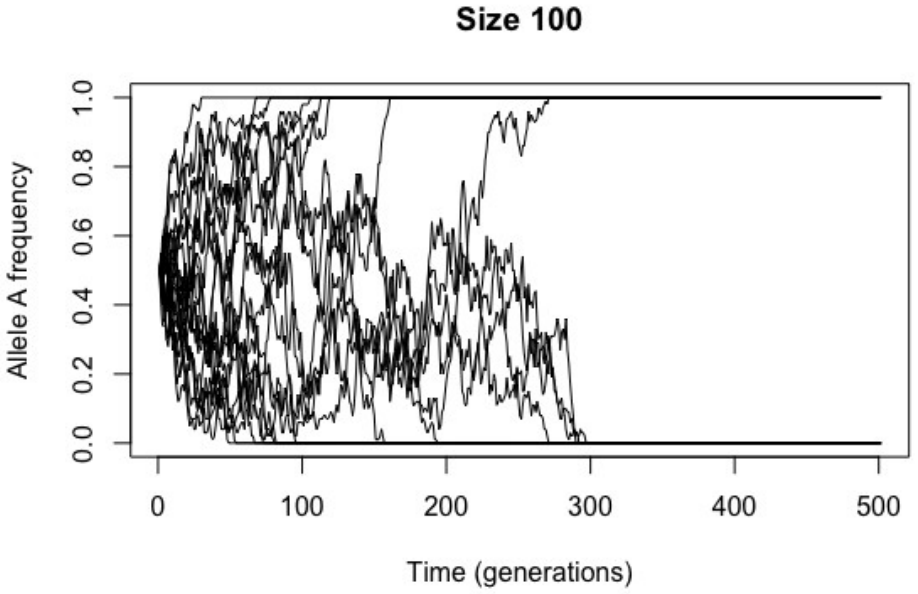
R

```
my_f_name = function(x,y,..) {
        ....
        return(res)
}

plot(vec, type= "l", ylim=c(0,1))
lines(othervec)

?rbinom
```

Exercice 1







Quelles conséquences sur le polymorphisme génétique d'une population ?

What is this N ? effective population size (N or Ne)

- ≠ census population size
- ~ nb individuals contributing to gene pool
- Size of an *idealized population (Wright-Fisher population)*

  ↳ matches with patterns of real population

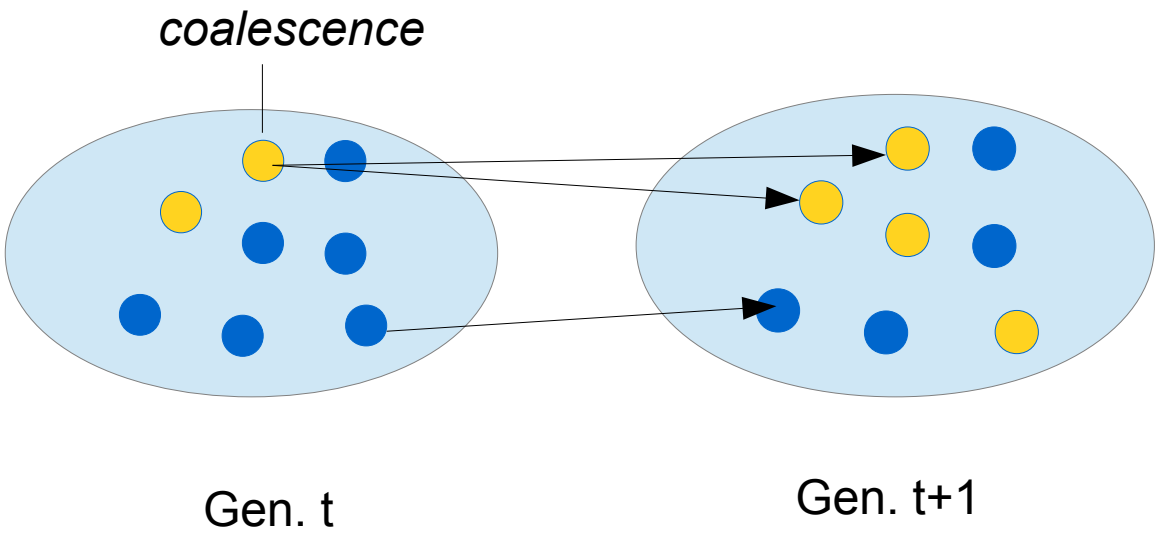
  (eg genetic diversity)

<u>Factors impacting Ne:</u>

- changes in census size
- variance in offspring number ↗ ⇒ Ne↘
  "Genghis Khan effect"
- unequal sex ratio ⇒ Ne ≠ Nmale + Nfemale
- ...

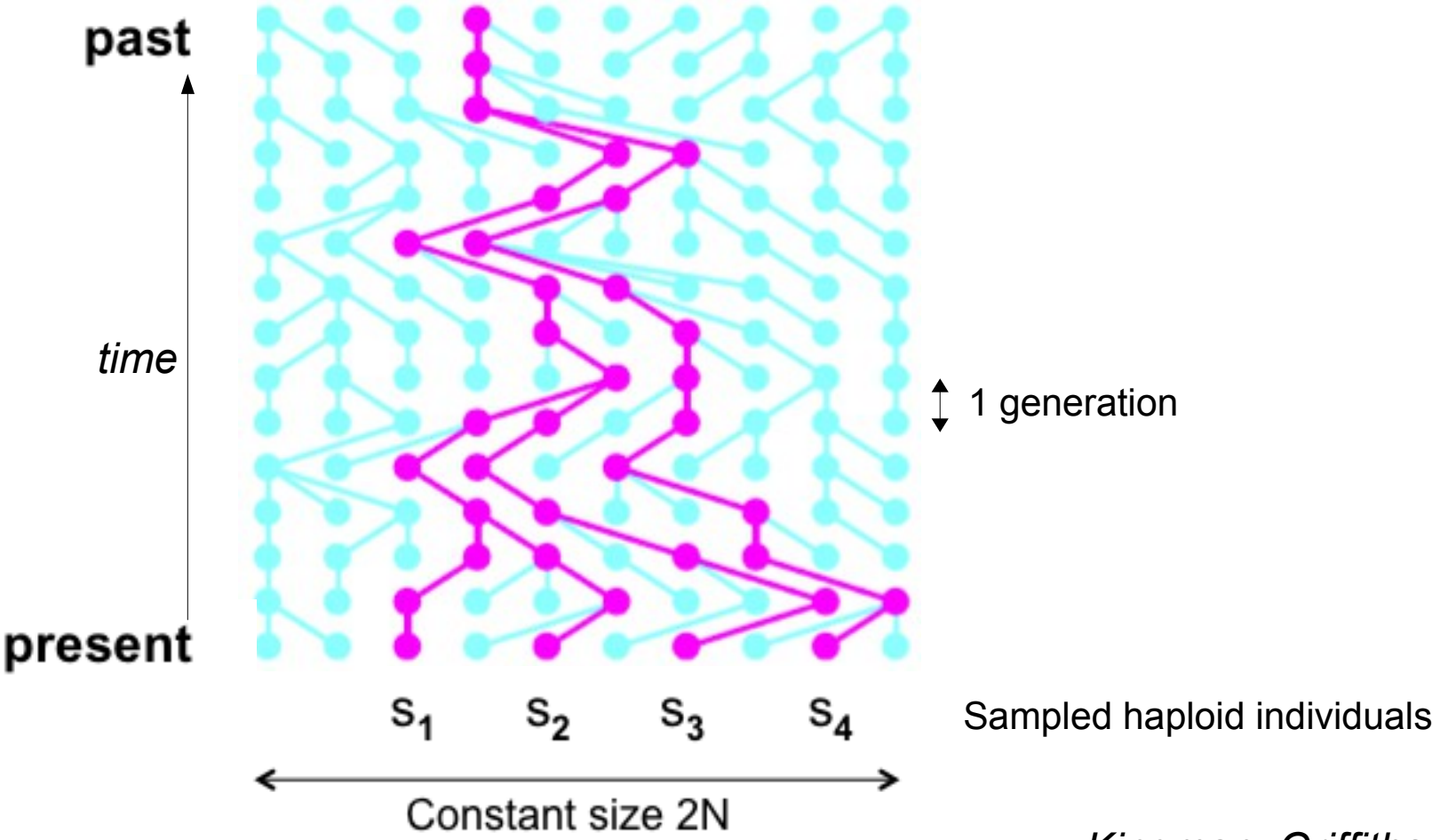
Modèle coalescent

Comprendre les relations entres individus → généalogies ?
ie tracer les ancêtres



Modèle coalescent

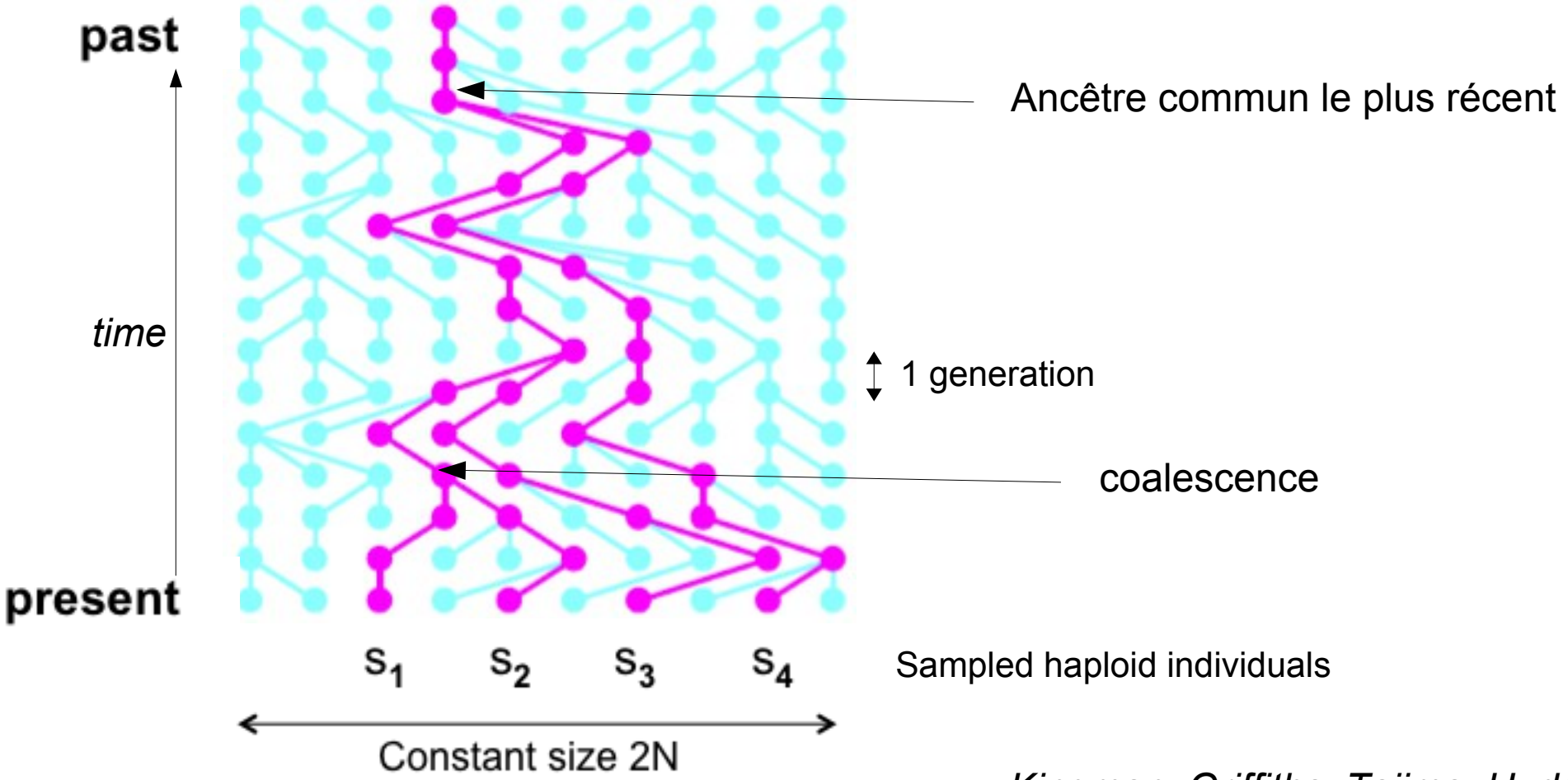
Comprendre les relations entres individus → généalogies ?
ie tracer les ancêtres



Kingman, Griffiths, Tajima, Hudson

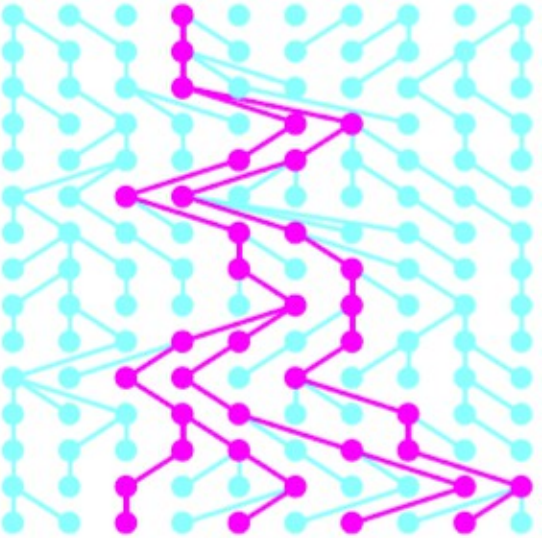
Modèle coalescent

Comprendre les relations entres individus → généalogies ?
ie tracer les ancêtres



Kingman, Griffiths, Tajima, Hudson

Modèle coalescent
Exercice 2



Taille de population = 2N
mêmes hypothèses que pour WF

- Proba que 2 lignées (gènes, individus) coalescent à la génération précédente ?
- …….
  ne coalescent pas pendant les t générations précédentes ?
- Proba que 3 individus coalescent à la génération précédente ?
- Sur feuille ou ordi, si N=1000
  Pr(2 lignées coal. à t=10) ?
  Pr(2 lignées coal. à t>10) ?

- (Chez vous) Tracer la probabilité de coalescence après t en fonction de t
  Visuellement pour quelle valeur de t cette proba tombe en dessous de 0.5 .
  Analytiquement ? (Pensez à la transformation log(a**b)=b log(a)

Modèle coalescent

Taille de population = 2N

T_2= temps d'attente avant coalescence des 2 lignées

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

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

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

- Si N large, alors p=1/(2N) petit . Approximation $(1-p)^t \to e^{-pt}\ quand\ p\to 0$
- $P(T_2>t)=e^{\frac{-t}{2N}}$
- E[T_2] = 1/p = 2N *(cf loi exponentielle)*
- Var(T_2) = $1/p^2 = 4N^2$

Le temps est mesuré continu et **la distribution géométrique remplacée par une distribution exponentielle**

Le temps d'attente moyen pour que 2 lignées précises coalescent est 2N

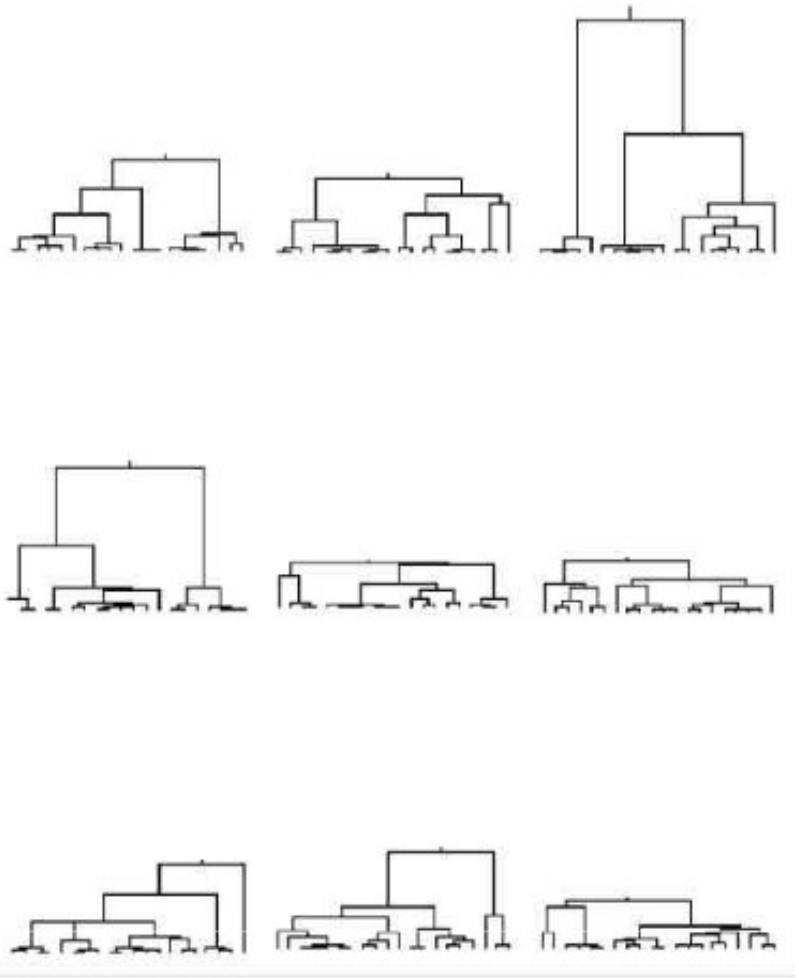
Modèle coalescent

Taille de population = 2N

T_2= temps d'attente avant coalescence des 2 lignées

$Var(T_2) = 1/p^2 = 4N^2$

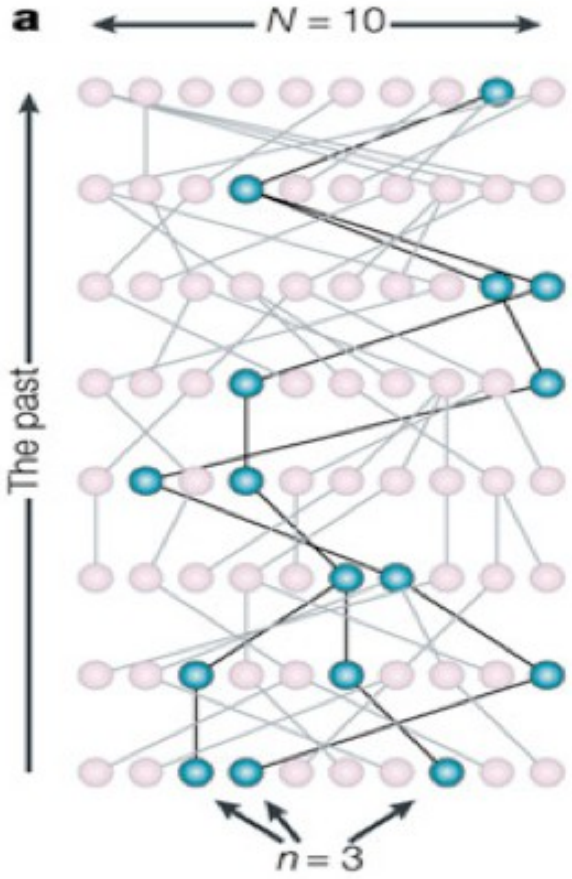
Grande variance du processus
→ un même modèle démographique
peut produire des généalogies
très différentes



S. Glémin

Modèle coalescent

- Coalescent = modèle « backward in time » (présent vers passé)
- Avantage par rapport à une simulation « forward » (passé vers présent) de type WF : On n'a pas besoin de connaître la généalogie dans toute la pop seulement pour les n lignées

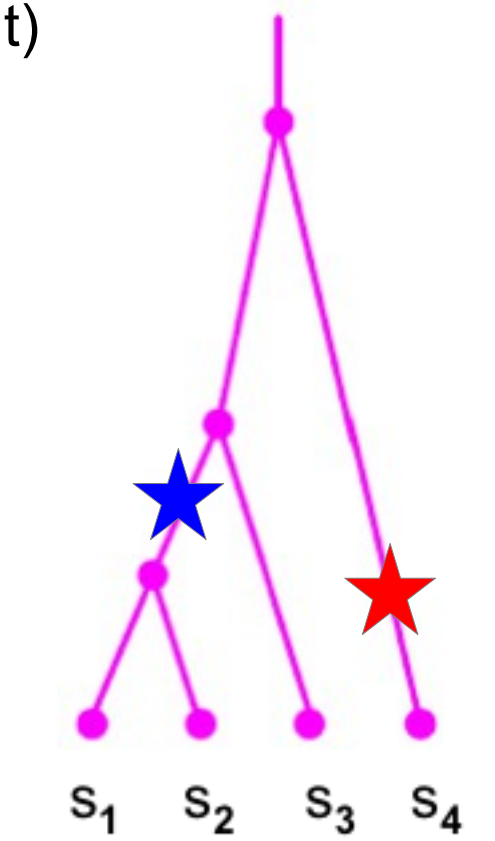


Nature Review Genetics

Ajout des mutations

= placer des événements sur les branches

Taux de mutation / gen = μ
ie en moyenne μt mutations en t générations
Nombre de mutations sur une branche ~ Poisson(μt)
infinite-site model : Chaque mutation a lieu
à un nouvelle position du gène



ctaa ctaa ctca ctct

Modèle coalescent

Taille de population = 2N $E[T_2] = 2N$

Taux de mutation / gen = μ
ie en moyenne μt mutations en t générations
chaque mutation a lieu a un nouvelle position du gène *(infinite-site model)*

Exercice 3 – Estimateur de Tajima

- Quel est le nombre de mutations moyen sur l'arbre reliant s1 et s2 en fonction de N et de μ ? (*)
- *Infinite-site model : chaque mutation touche un site différent (ie pas de mutation récurrente en un site, ie on ne voit pas A → C → G)*

  On note S le nombre de sites polymorphiques, on a donc E[S]= (*)
  Discuter l'influence de N sur S. Etait-ce intuitif pour vous ?
- Proposer une procédure pour estimer S à partir d'un échantillon de taille *n.* Si on connait μ on peut en déduire un estimateur de N.

Modèle coalescent

Taille de population = 2N $E[T_2] = 2N$

Taux de mutation / gen = μ

- Estimateur de Tajima

$$On \quad note \quad \theta = 4 N \mu$$
$$\theta = E[S]$$

$$\hat{\theta}_{Taj} = \frac{\sum_{i<j} d_{ij}}{n(n-1)/2}$$

$$Parfois \quad noté \quad \hat{\theta}_{\pi}$$

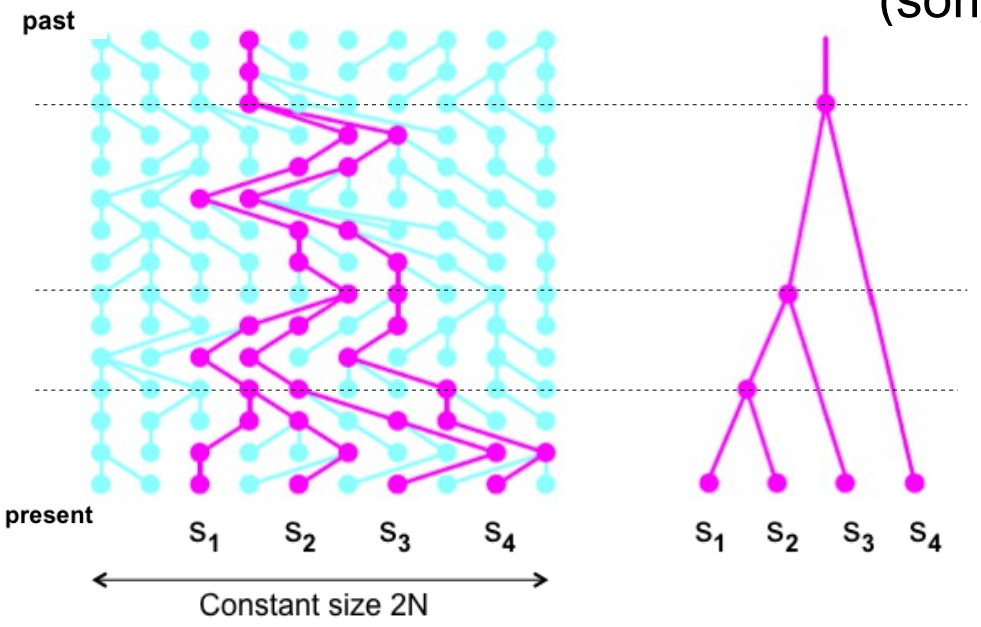
Coalescent avec n>2 lignées

Taille de pop 2N, nb lignées n

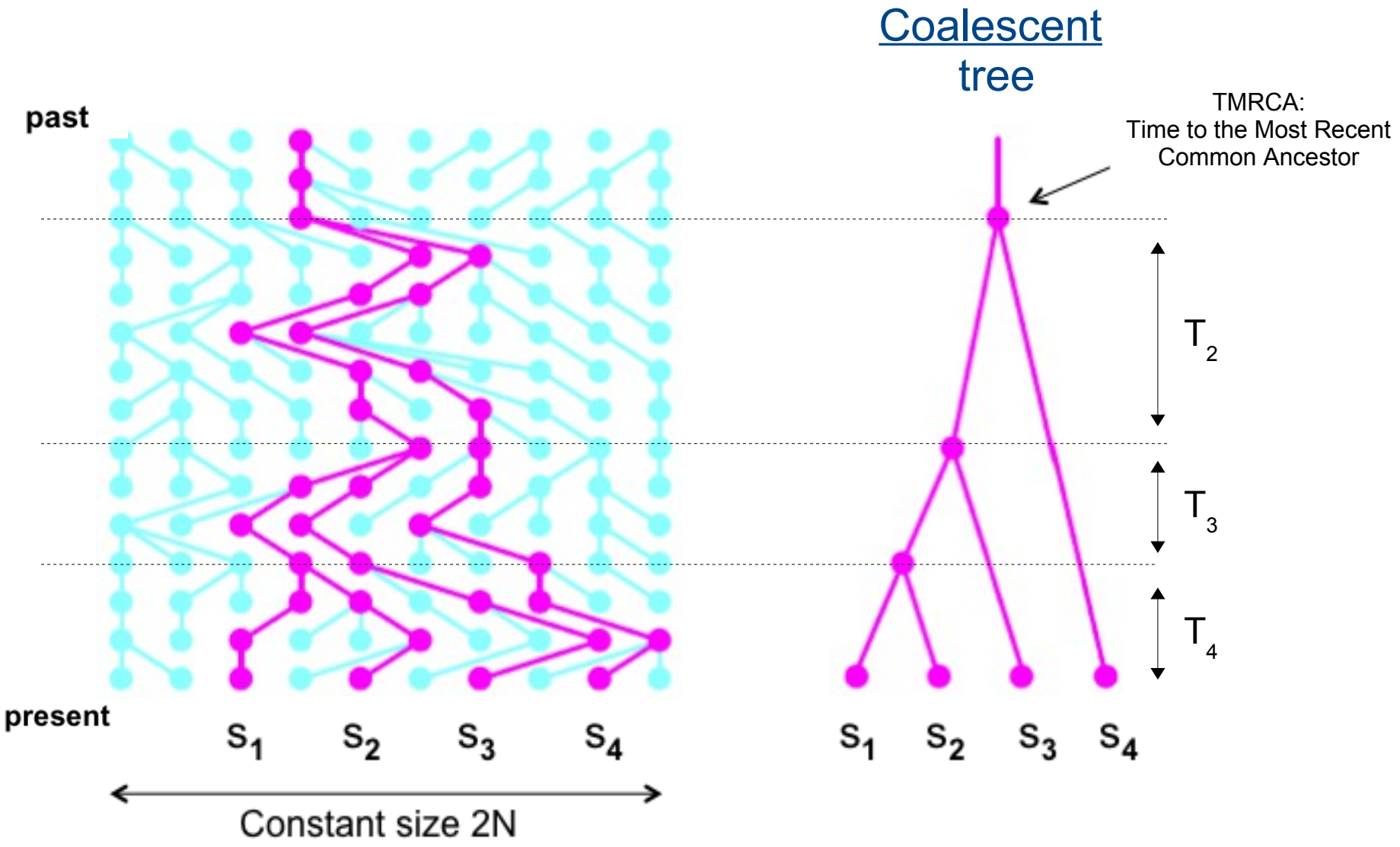
- n(n-1)/2 pairs possibles avec même proba de coalescer
- Pr(1 evt de coal à t=1) = n(n-1)/2 Pr(lignées (si, sj) coalescent à t=1)
- Taux de coalescence lorsqu'on a k lignées ?

  T_k = Temps d'attente avant coalescence lorsque l'on a k lignées. $E[T_k]$?

- Montrer T_2 , T_3 , T_4 sur cet arbre. En déduire la taille totale de l'arbre (somme des longueurs de branches)



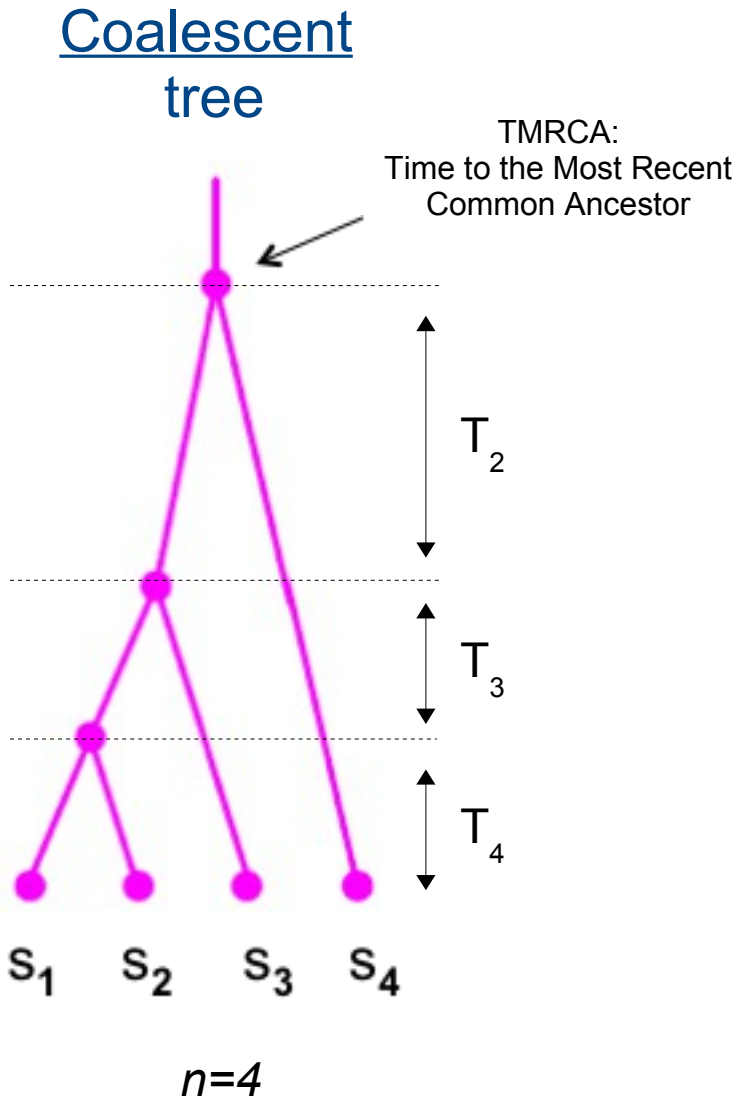
Coalescent avec n>2 lignées



Coalescent avec n>2 lignées

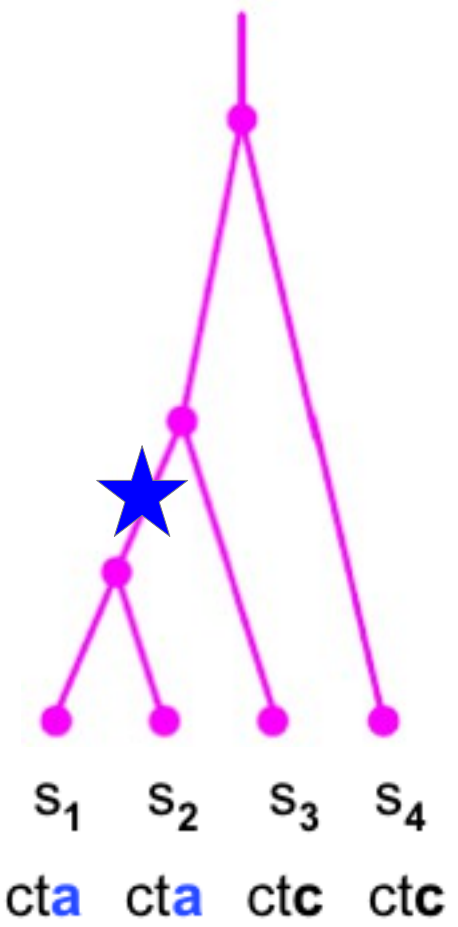
$$E[T_k] = \frac{1}{coal.rate(k)} = ¿$$

$$Total\ branch\ length = \sum_{k=2}^{n} k\,E[T_k]$$



Coalescent avec n>2 lignées

Let's introduce mutations

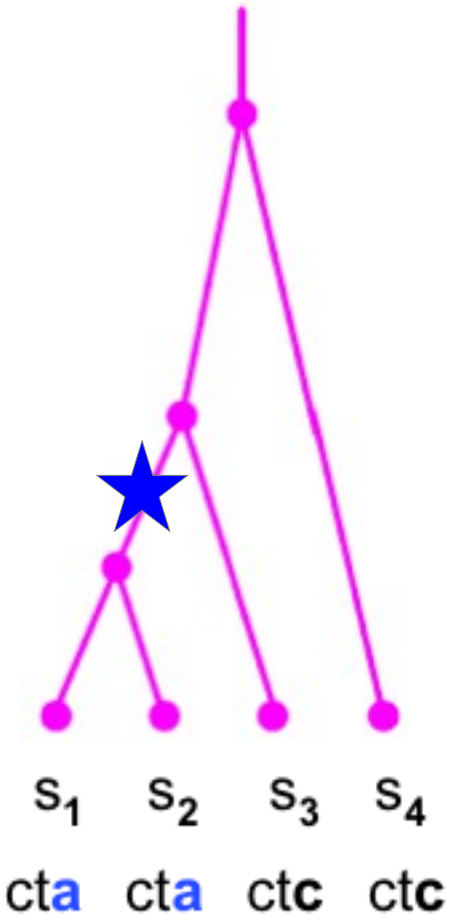


Coalescent avec n>2 lignées

Let's introduce mutations

μ : *mutation rate per generation for whole region (locus)*

$$E[\underbrace{Nb\ segregating\ sites}_{S}] = Total\ branch\ length \times \mu$$



Coalescent avec n>2 lignées

$$E[\underbrace{Nb\ segregating\ sites}_{S}] = Total\ branch\ length \times \mu \quad = \quad 4N \sum_{k=1}^{n-1} \frac{1}{k} \times \mu$$

⇒ Un deuxième estimateur de θ=4Nμ ou de N *Watterson's estimator*

$$\hat{\theta}_{Wat} = \frac{S_{observed}}{\sum_{k=1}^{n-1} \frac{1}{k}} \qquad \hat{N}_{Wat} = \frac{S_{observed}}{4 \sum_{k=1}^{n-1} \frac{1}{k}\ \mu}$$

$S_{observed} = 4$
$L = 32$
$\mu = \mu_{per.site} \times L$
$n = 4$

Exercice pour plus tard

(1)

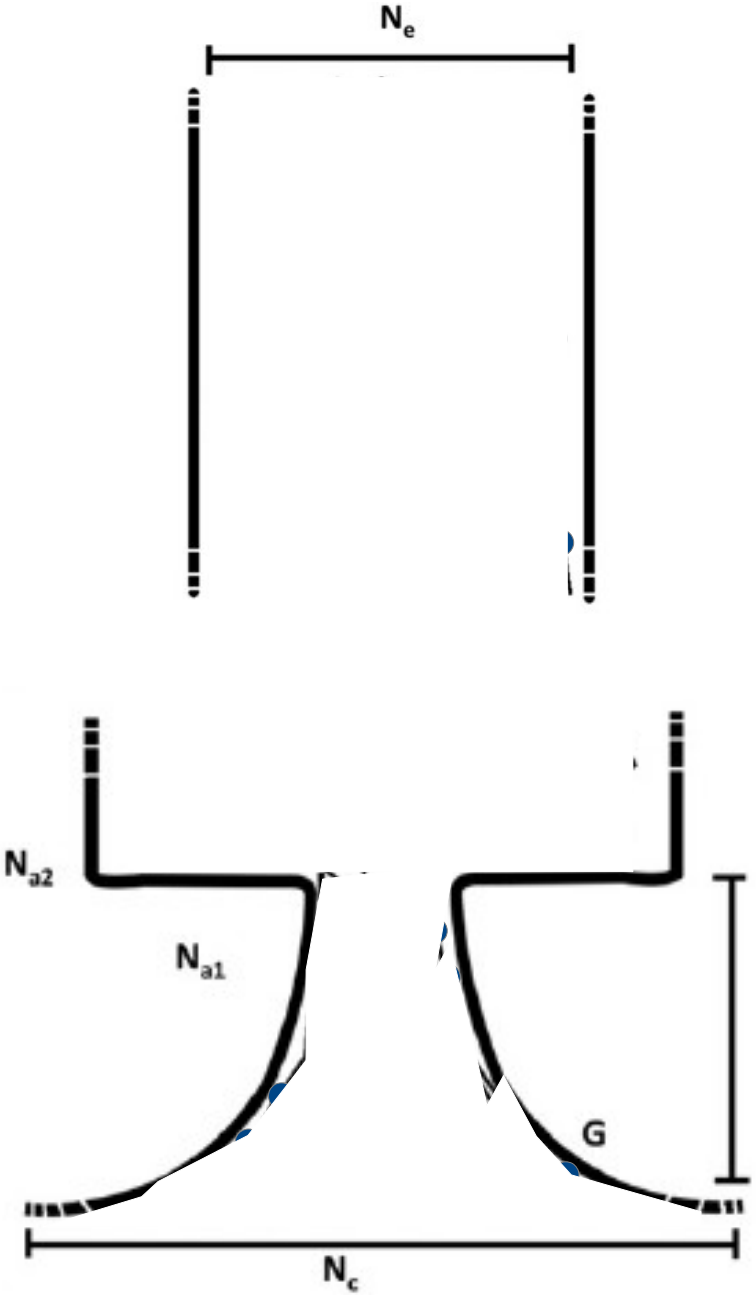
Soit un échantillon de taille 2 (s1, s2)

Hétérozygotie attendue en fonction de theta ?

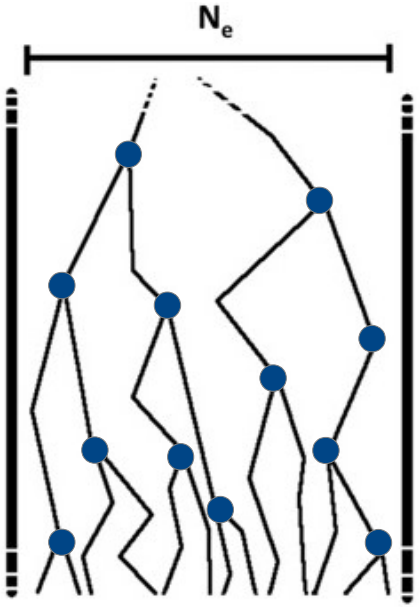
Indice : considérer la mutation et la coalescence comme deux processus concurrents. Si s1 et s2 coalescent avant qu'une mutation que peut-on dire de s1 et s2 ?

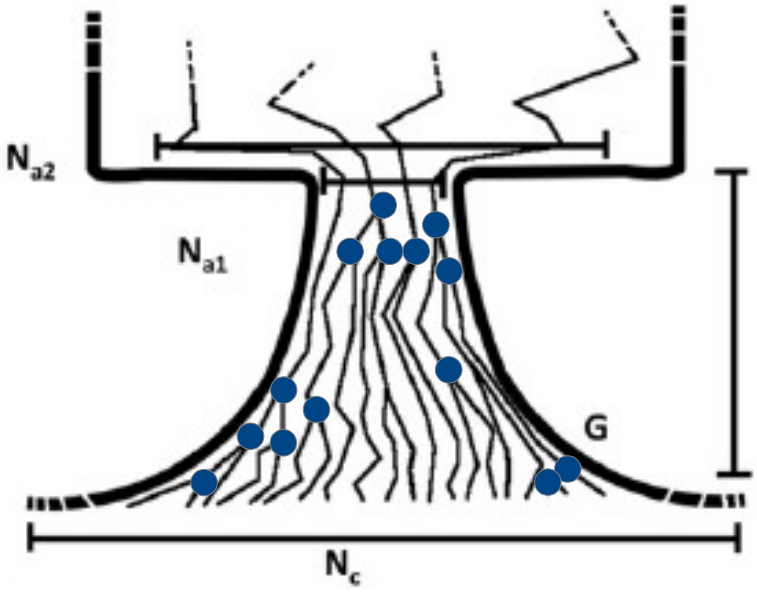
(2) E[TMRCA de n lignées]

Motivation (pour écouter le cours ?)



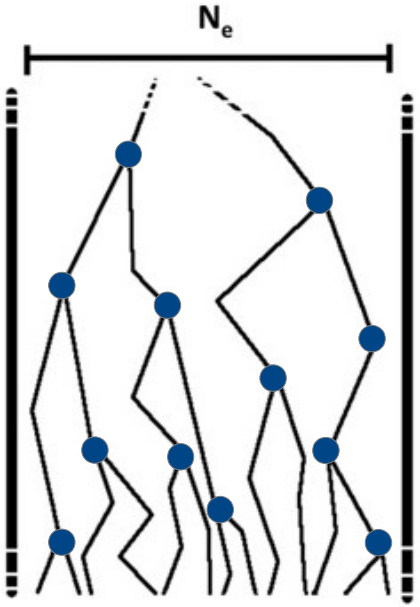
Motivation (pour écouter le cours ?)

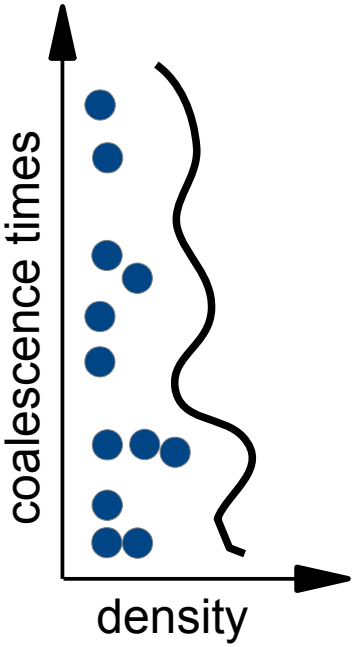


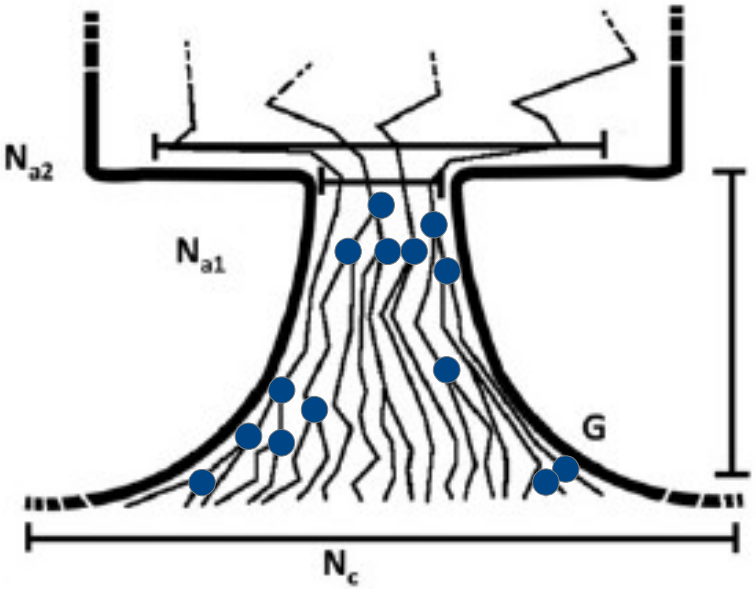


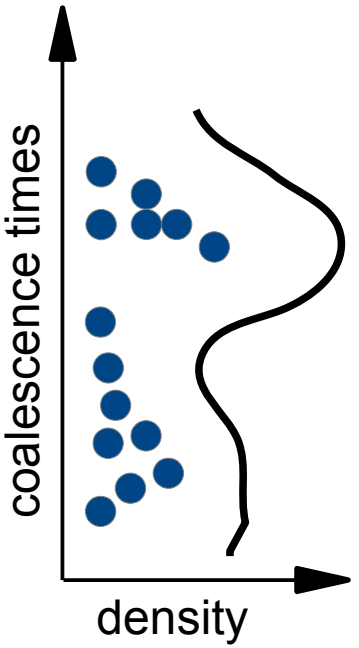
Ce coalescent est une extension du modèle à taille constante, les taux de coalescence varient en fonction du temps (et de la taille à ce temps t)

Motivation

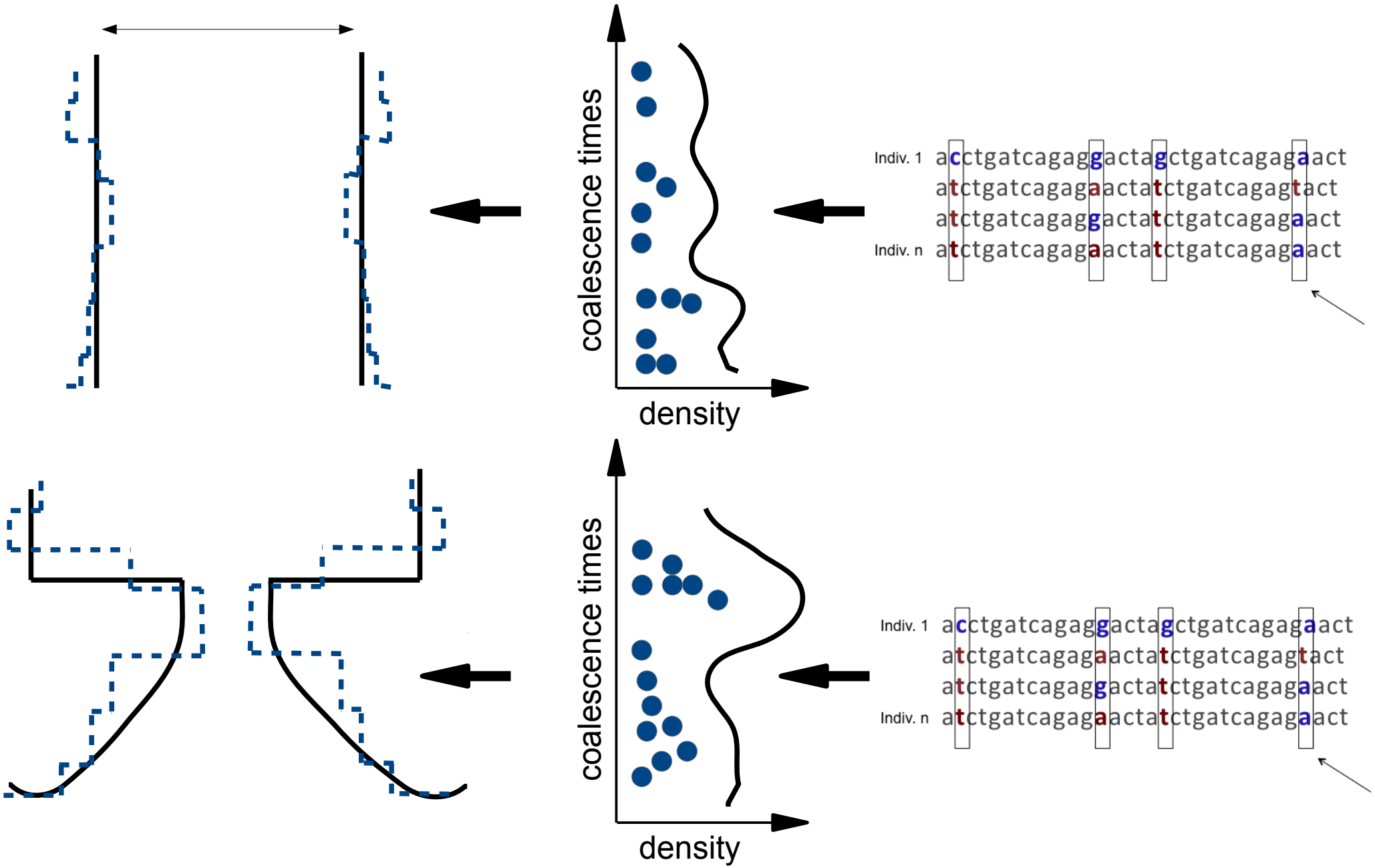

N_e


coalescence times
density


N_{a2}
N_{a1}
G
N_c


coalescence times
density

Motivation

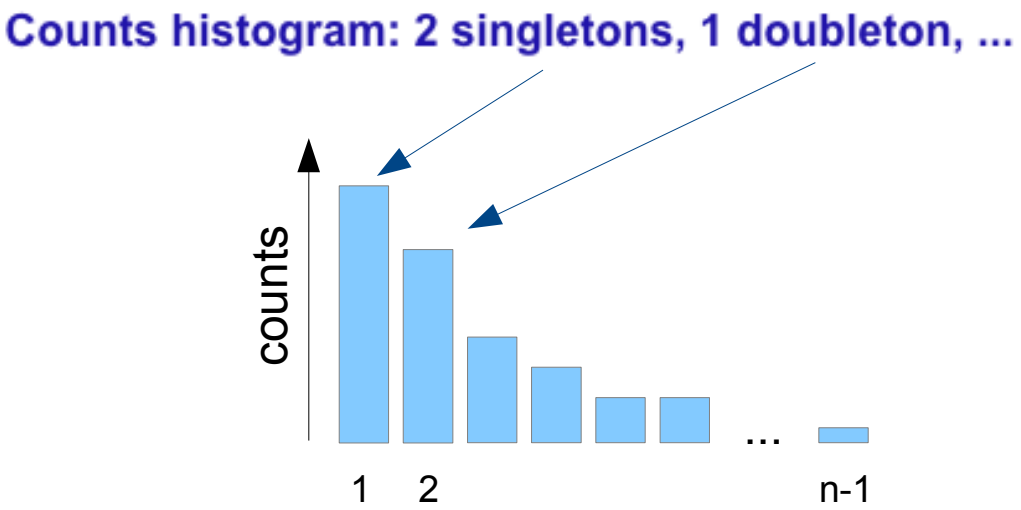

coalescence times
density
Indiv. 1
Indiv. n
acctgatcagaggactagctgatcagagaact
atctgatcagagaactatctgatcagagtact
atctgatcagaggactatctgatcagagaact
atctgatcagagaactatctgatcagagaact
coalescence times
density
Indiv. 1
Indiv. n
acctgatcagaggactagctgatcagagaact
atctgatcagagaactatctgatcagagtact
atctgatcagaggactatctgatcagagaact
atctgatcagagaactatctgatcagagaact

Le spectre de fréquences alléliques

Spectre/histogramme des fréquences

« AFS », « SFS »

Si n individuals : à une position dans le génome (un site) on peut voir
entre 0 et n allèles dérivés
SFS = Histogramme des nb d'allèles dérivés aux sites polymorphiques



SFS

f_j = proportions de sites ayant j allèles dérivés. Il a été montré que :

E[Nombre de singletons] = θ

E[Nombre total de mutation]= E[S] = $\theta \sum_{k=1}^{n-1} \frac{1}{k}$

$$E[f_1] = \frac{\theta}{\theta \sum_{k=1}^{n-1} \frac{1}{k}} = \frac{1}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

Il a été montré que pour tout j=1,...n-1

$$E[f_j] = \frac{1/j}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

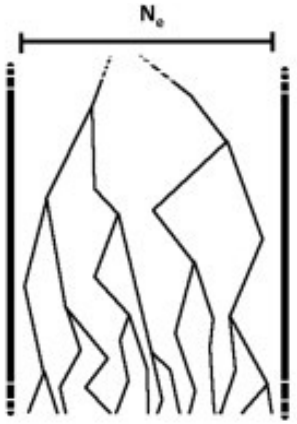
Forme du SFS

Ne dépend
ni du taux de mutation
ni de N !

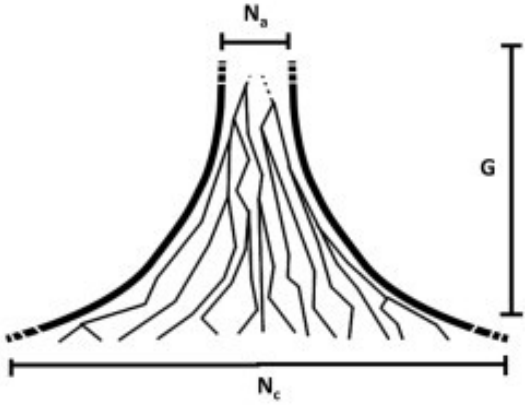
Le spectre de fréquences informe sur la démographie

Entre autre...

past



present

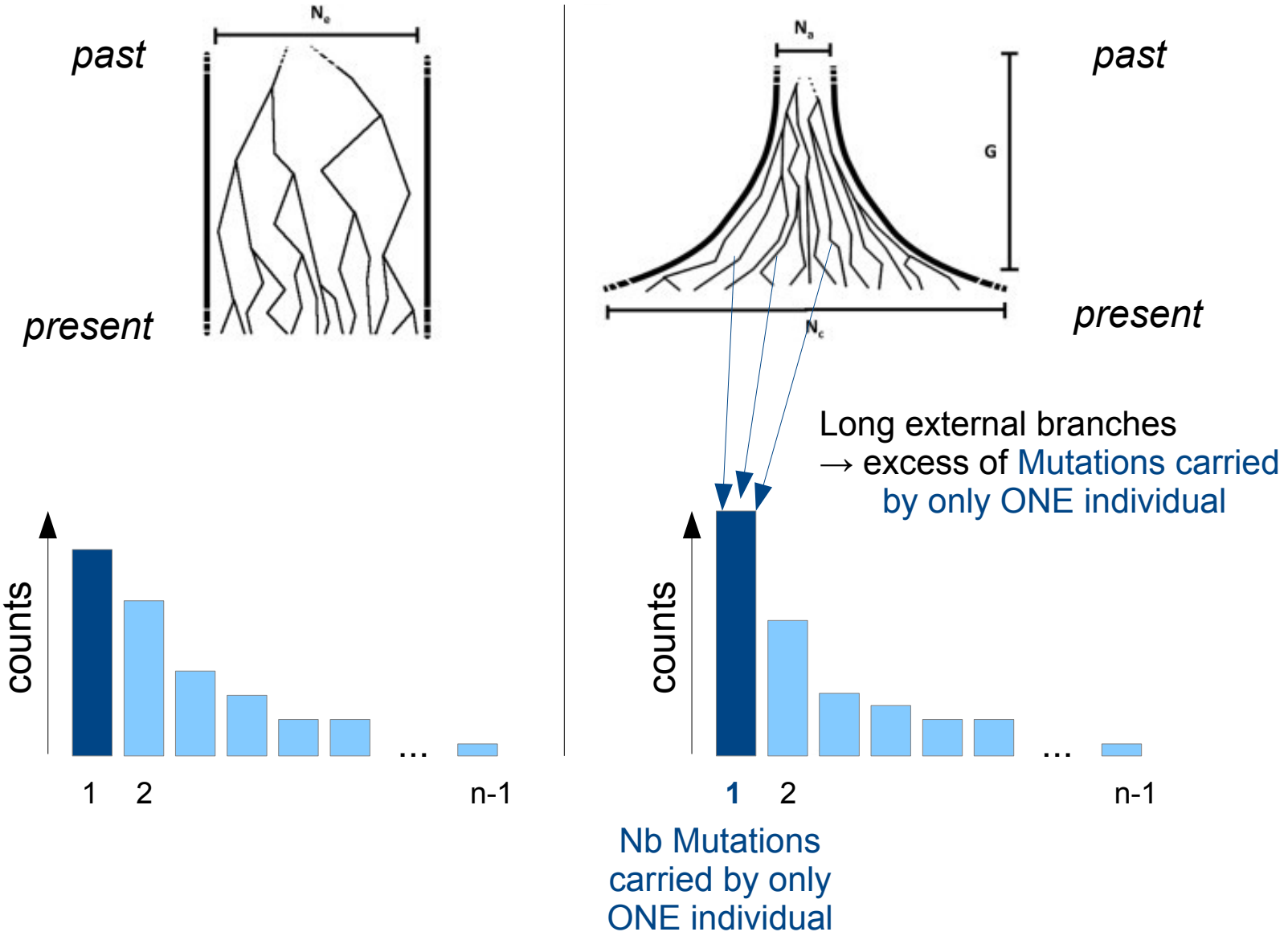


past

present

Le spectre de fréquences informe sur la démographie

Entre autre...



Pour aller plus loin...

Idée générale

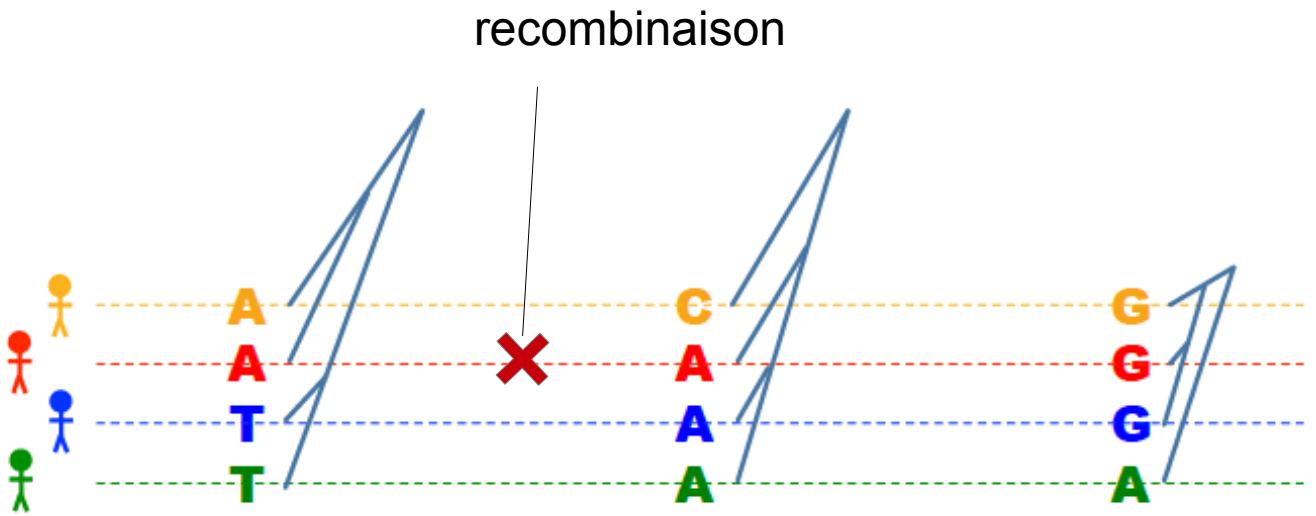
- Dériver les formules du coalescent pour des modèles démographiques plus complexes (isolation-migration → **dessin**, modèle en îles, fluctuations des tailles de population, ...)
- → Proba(données | modèle)
  ex. choisir le modèle qui explique (fit) le mieux le spectre de fréquence observé

Limitations :

- Hypothèses du coalescent (panmixie, générations distinctes)
- La recombinaison complique les dérivations
  - il n'y a pas qu'une seule généalogie pour tout le génome !
  - trouver des régions « indépendantes » → plein de généalogies « indé »
  - tenir compte de la corrélation le long du génome –> généalogies liées

Limitations - recombinaison

- Recombinason le long du génome à chaque génération
  → génome = mosaïque du matériel génétique de nos ancêtres
  → multiples généalogies changeant le long de la séquence



Adapted from
© Sarah Sheehan

Autres approches

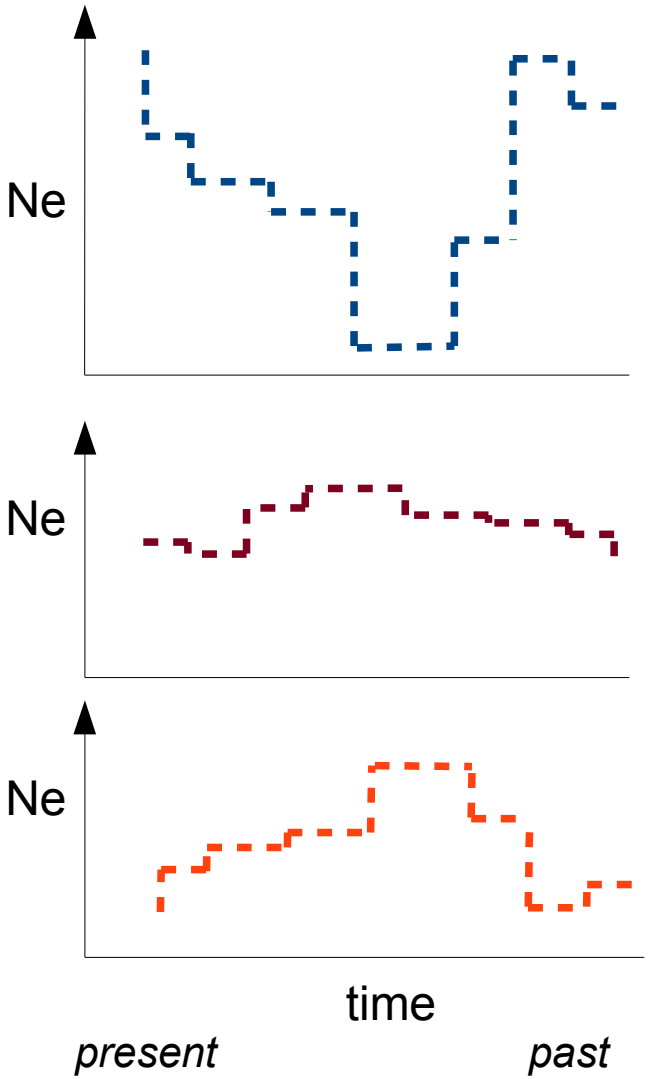
(1) Reconstruire la généalogie exacte (plutôt que le SFS par ex)

(2) SFS = une manière de résumer les données
Trouver d'autres statistiques résumées puis essayer d'identifier le modèle reproduisant au mieux cet ensemble de statistiques

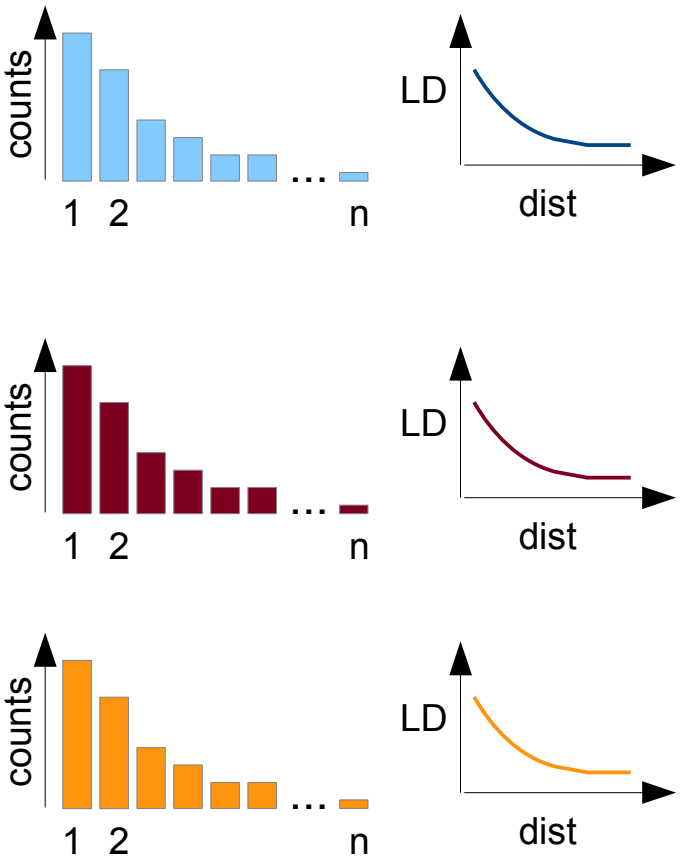
...

Exemple : plus de statistiques résumées

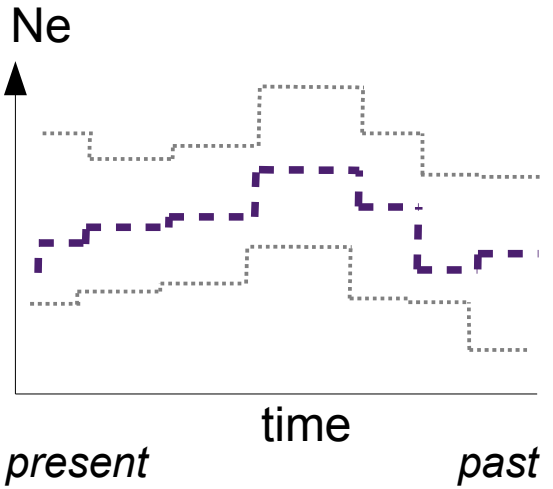
Simuler des histoires démographiques aléatoirement



Calcul des stats. résumées *SFS, Déséquilibre de Liaison*



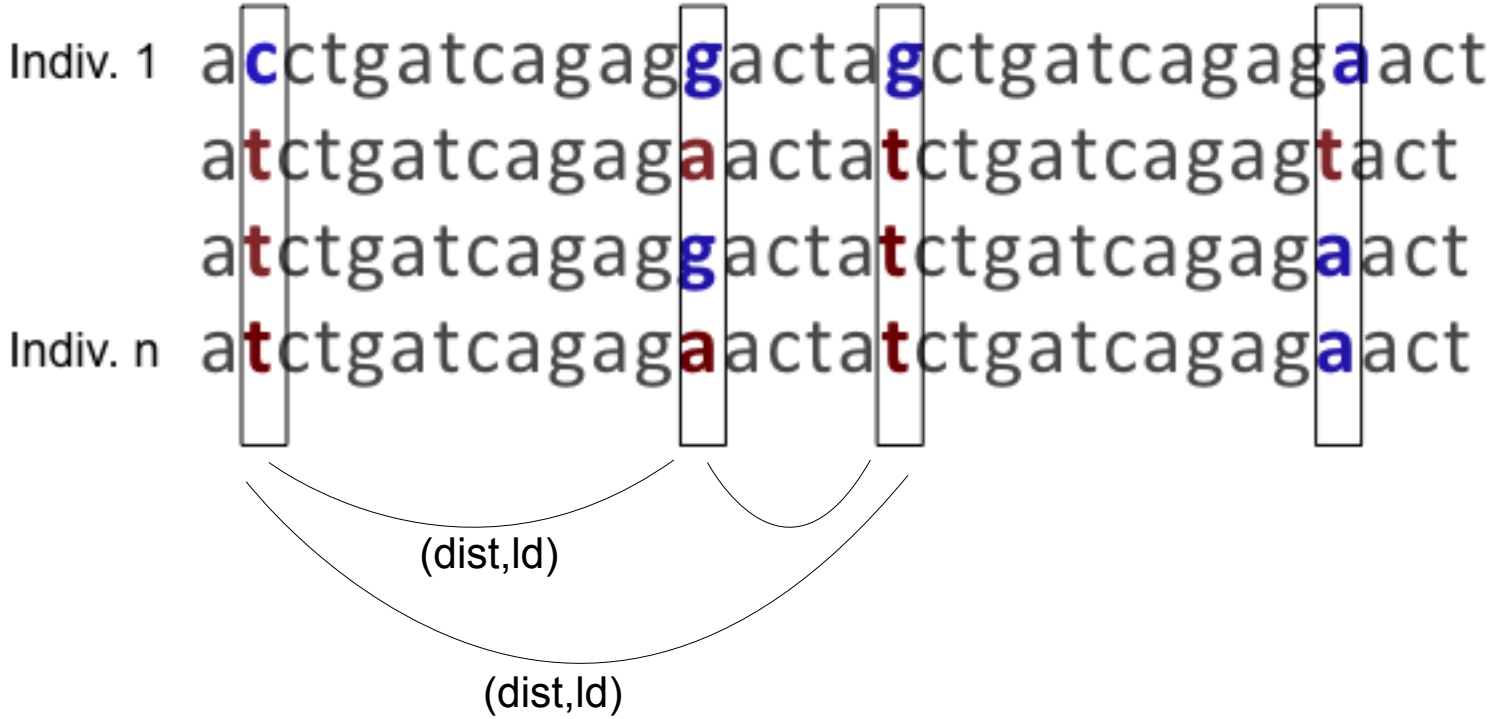
Garder les histoires qui ont produit des stats proches des stats observées dans les vraies données



Boitard et al PloS Genet 2016

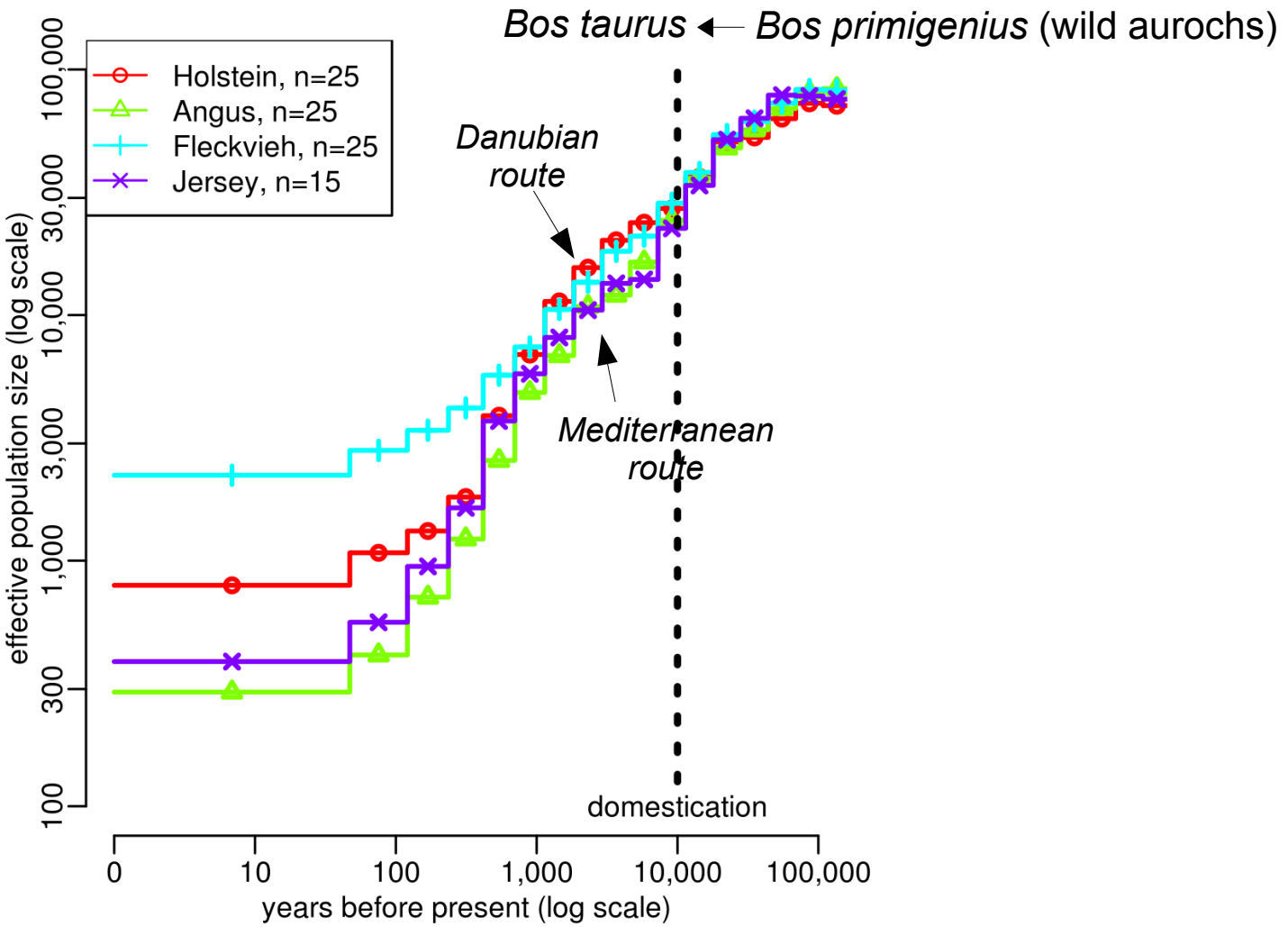
LD=Déséquilibre de liaison

LD = correlation entre SNPs



Application to cattle breeds

Boitard et al 2016



TP

Sur https://github.com/jayflora/GdP-material
Téléchargez
CG_54genomes_indiv.txt
chr22.CG_54genomes_shapeit_phased.haps.tgz (version compressée)
Et la fiche d'exercice
Tar -xzf fichier.tgz pour décompresser

Commencez par l'exercice sur l'heterozygo.