



Génétique statistique

Simon Boitard, Willy Rodríguez

► To cite this version:

Simon Boitard, Willy Rodríguez. Génétique statistique. Master. Master (Master Bioinformatique et Biologie des Systèmes), 2017. hal-02788119

HAL Id: hal-02788119

<https://hal.inrae.fr/hal-02788119>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

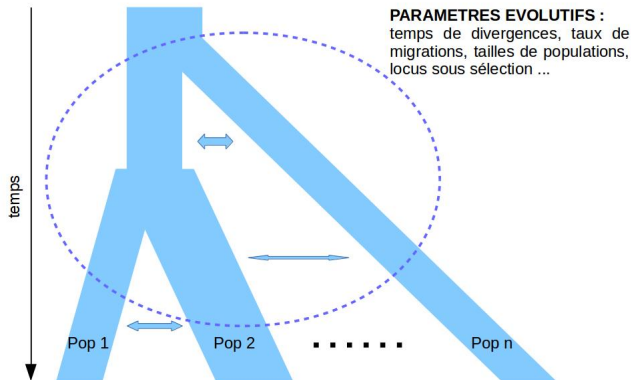
Coalescence : principe et bases théoriques

Simon Boitard

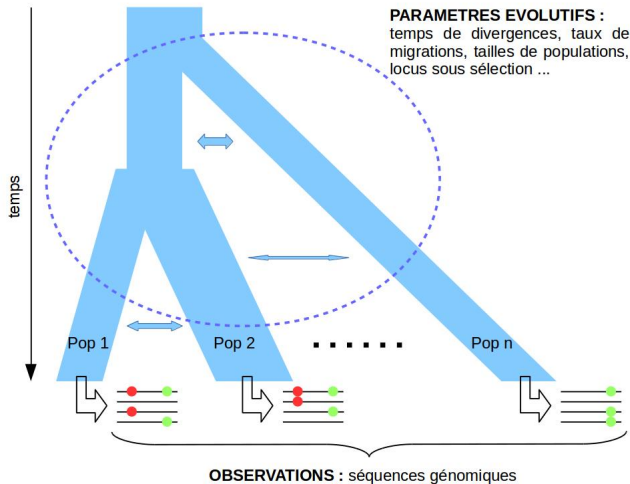
INRA, GenPhySE, Toulouse

M2 BBS, Atelier Génétique Statistique, 2017-2018

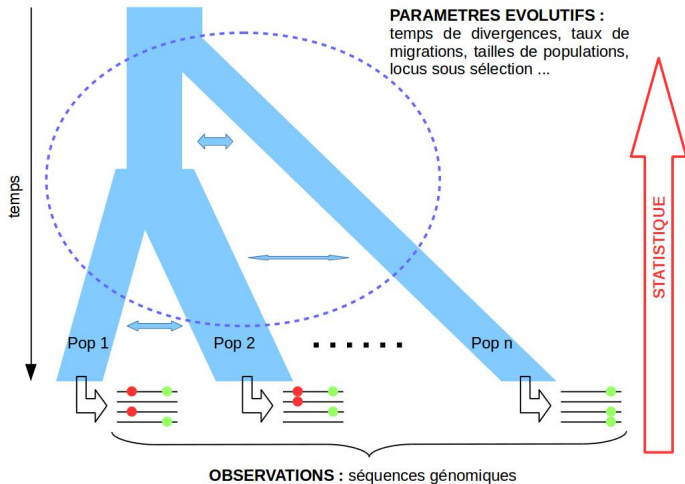
Reconstruction de l'histoire évolutive à une échelle de temps courte (intra-espèce) à partir de **données génomiques**.



Reconstruction de l'histoire évolutive à une échelle de temps courte (intra-espèce) à partir de **données génomiques**.



Reconstruction de l'histoire évolutive à une échelle de temps courte (intra-espèce) à partir de **données génomiques**.

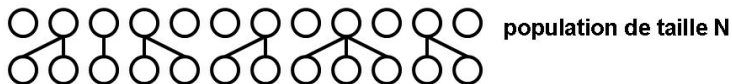


- Modèle mathématique permettant de décrire / simuler la généalogie d'un échantillon de séquences d'ADN (ou individus haploïdes).
- Outil pour comprendre l'influence du processus évolutif sur la diversité génétique.
- On **remonte** le temps au lieu de le descendre.
- Article fondateur de Kingman en 1982, très utilisé depuis.

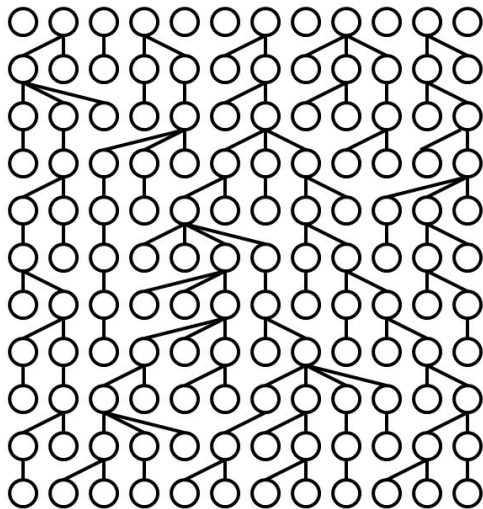
Modèle de Wright-Fisher pour un locus sans recombinaison

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ **population de taille N**

Modèle de Wright-Fisher pour un locus sans recombinaison

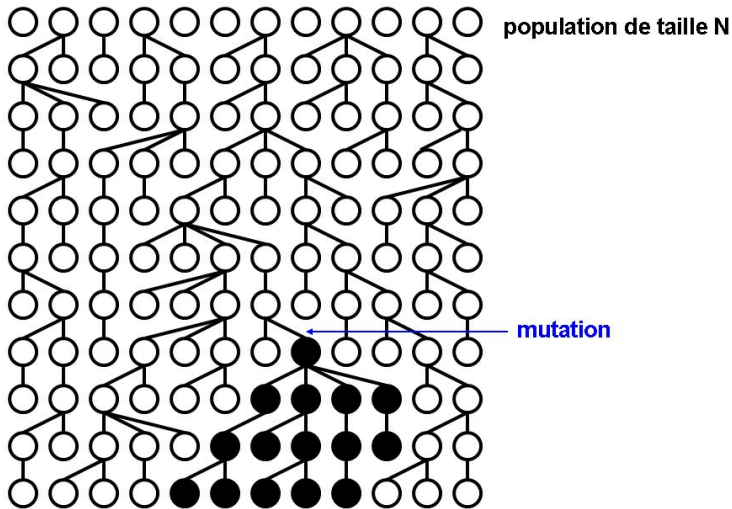


Modèle de Wright-Fisher pour un locus sans recombinaison

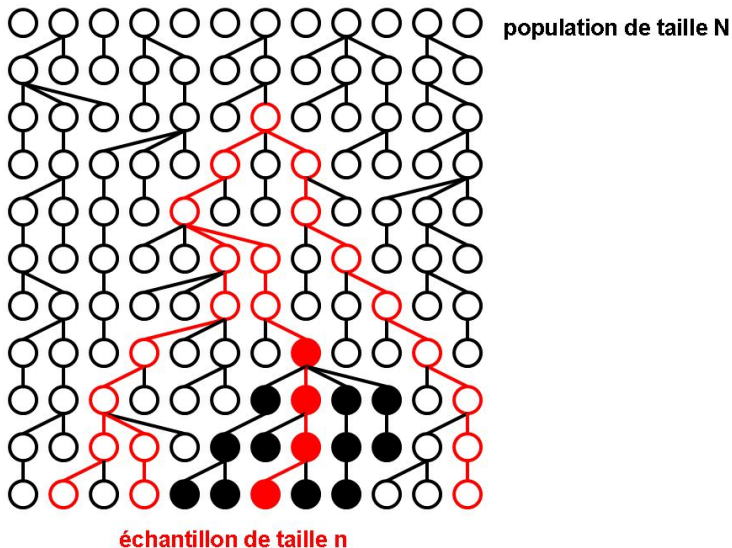


population de taille N

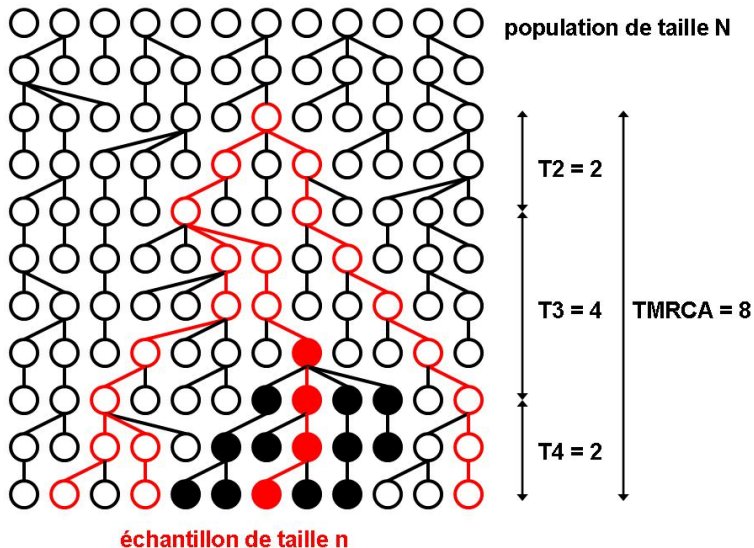
Modèle de Wright-Fisher pour un locus sans recombinaison



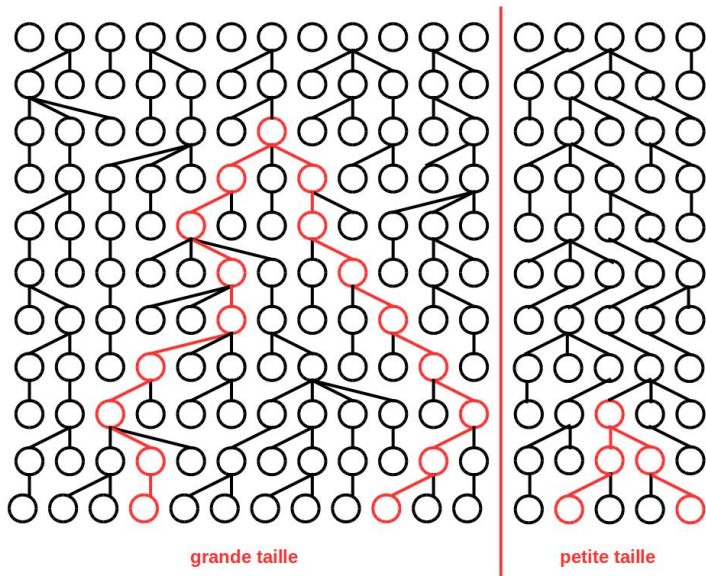
Modèle de Wright-Fisher pour un locus sans recombinaison



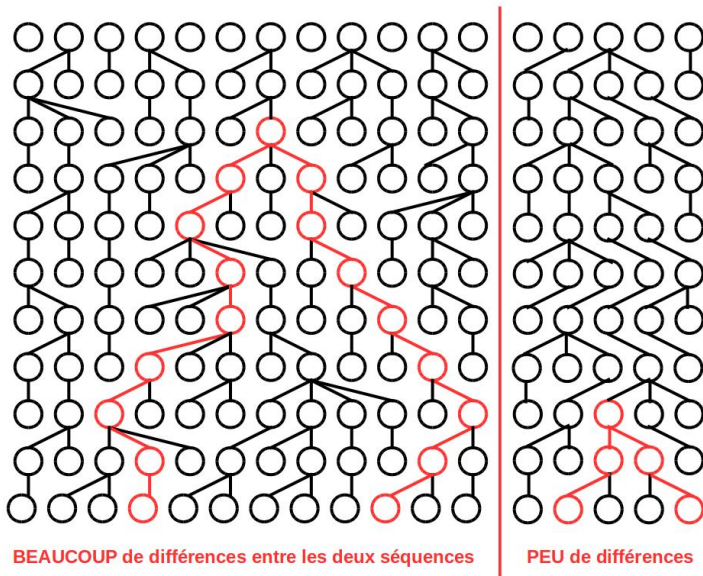
Modèle de Wright-Fisher pour un locus sans recombinaison



Généalogie et taille efficace



Généalogie et taille efficace



- 1 Modèle de base : une population de taille constante, un locus sans recombinaison
- 2 Population de taille variable au cours du temps
- 3 Population structurée
- 4 Modèle avec recombinaison
- 5 Conclusions

- 1 Modèle de base : une population de taille constante, un locus sans recombinaison
- 2 Population de taille variable au cours du temps
- 3 Population structurée
- 4 Modèle avec recombinaison
- 5 Conclusions

Temps de coalescence

- Proba qu'il n'y ait pas de coalescence à une génération donnée

$$q^N(n) = \prod_{i=1}^{n-1} \left(1 - \frac{i}{N}\right) = 1 - \frac{n(n-1)}{2N} + O\left(\frac{1}{N^2}\right)$$

- T_n^N suit une loi géométrique

$$\mathbb{P}(T_n^N > t) = (q^N(n))^t$$

- Changement d'échelle $\tau = \frac{t}{N}$

$$\mathbb{P}(T_n^N > N\tau) = (q^N(n))^{N\tau} \approx \left(1 - \frac{n(n-1)}{2N}\right)^{N\tau} \rightarrow e^{-\frac{n(n-1)}{2}\tau}$$

quand $N \rightarrow +\infty$.

$\rightarrow T_n^N$ tend vers T_n , de loi exponentielle de paramètre $\frac{n(n-1)}{2}$.

- La proba que trois individus ou plus coalescent à la même génération est en $O(\frac{1}{N^2})$
→ négligeable quand $N \rightarrow +\infty$.
- En pratique, une coalescence consiste toujours à regrouper exactement deux lignées.

$$\begin{aligned}T_{MRCA} &= \sum_{k=2}^n T_k \\ \mathbb{E}[T_{MRCA}] &= \sum_{k=2}^n \mathbb{E}[T_k] \\ &= \sum_{k=2}^n \frac{2}{k(k-1)} \\ &= 2 \sum_{k=2}^n \left(\frac{1}{k-1} - \frac{1}{k} \right) \\ &= 2 \left(1 - \frac{1}{n} \right) \approx 2\end{aligned}$$

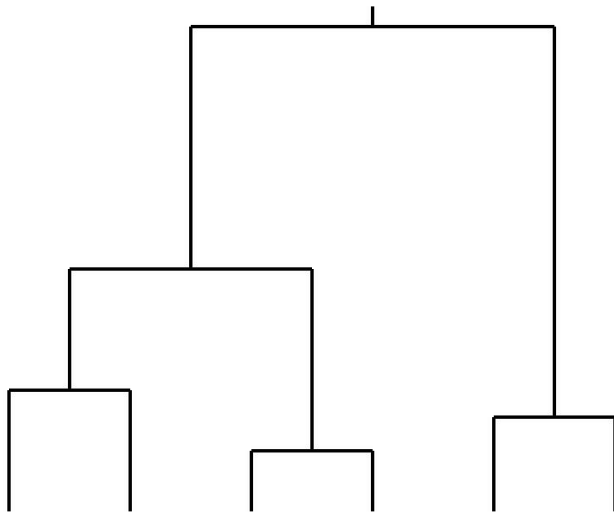
- Proba μ de mutation par site et par méiose.
- $M(t)$ nombre de mutations pour une branche de longueur $t = N\tau$ et un locus de taille L .

$$\mathbb{E}[M(t)] = \mu L N\tau = \frac{\theta}{2} L\tau$$

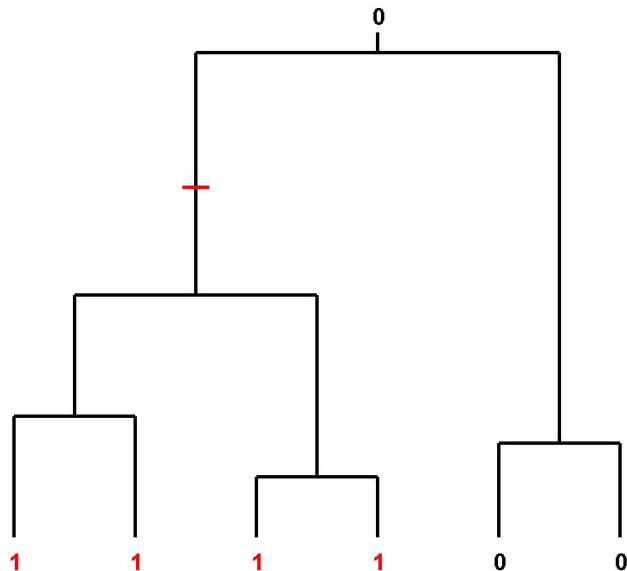
en posant $\theta = 2N\mu$.

- $M(\tau)$ loi de Poisson d'intensité $\frac{\theta}{2} L\tau$.
- On peut choisir ensuite le modèle qu'on veut pour décrire ce qui se passe quand une mutation se produit.

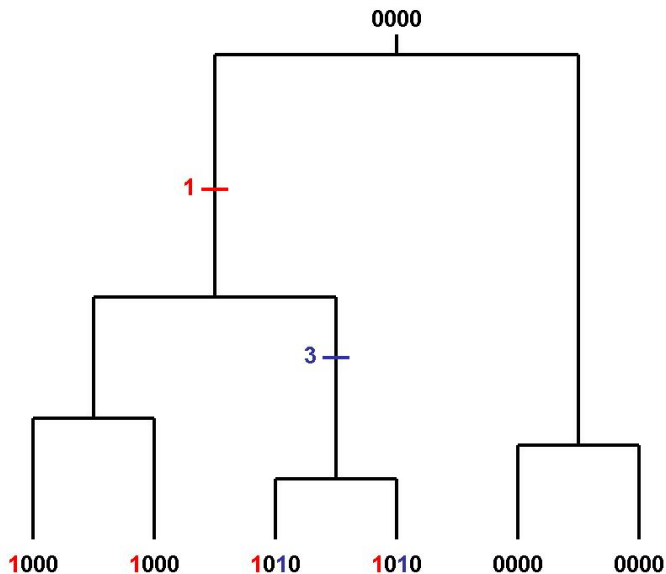
Mutations



Mutations



Infinite site model



- 1 Pour k allant de n à 2 :
 - 1 Simuler une loi exponentielle T_k de paramètre $\frac{k(k-1)}{2}$.
 - 2 Choisir uniformément celui des $\frac{k(k-1)}{2}$ couples d'haplotypes qui coalesce.
- 2 Placer les mutations indépendamment sur chaque branche selon un processus de Poisson d'intensité $\frac{\theta}{2}L$.

Beaucoup plus rapide que de simuler la population en forward!

- N taille de la population, $2N$ allèles.
- Remplacer N par $2N$ dans tout ce qui précède.
- En particulier : unité de temps $2N$ générations, $\theta = 4N\mu$.

- 1 Modèle de base : une population de taille constante, un locus sans recombinaison
- 2 Population de taille variable au cours du temps
- 3 Population structurée
- 4 Modèle avec recombinaison
- 5 Conclusions

- La taille de la population change selon une fonction déterministe du temps :
 - Croissance exponentielle : $N(t) = N_0 * e^{-\alpha t}$, $\alpha > 0$
 - Décroissance exponentielle : $N(t) = N_0 * e^{\alpha t}$, $\alpha > 0$
 - Décroissance linéaire : $N(t) = N_0 + \alpha t$, $\alpha > 0$
 - Goulot d'étranglement ("bottleneck") : $N(t) = N_0$ si $t < t^*$, N_1 si $t \geq t^*$
- La probabilité de coalescence dépend de la taille de population et donc varie au cours du temps.
- L'unité de temps est $\tau = \frac{t}{N_0}$.
- En un temps τ , on a à peu près:

$$T_n(\tau) \sim \frac{N(\tau)}{N_0} \exp\left(\frac{n(n-1)}{2}\right)$$

Simulation (approchée)

$$\tau = 0$$

1 Pour k allant de n à 2 :

1 Simuler une loi exponentielle T_k de paramètre $\frac{k(k-1)}{2}$.

2 $T_k = T_k * \frac{N(\tau)}{N_0}$.

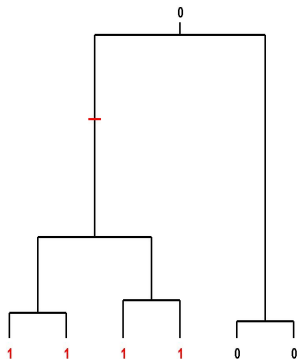
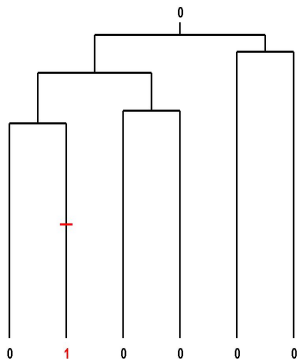
3 $\tau = \tau + T_k$.

4 Choisir uniformément celui des $\frac{k(k-1)}{2}$ couples d'haplotypes qui coalesce.

2 Placer les mutations indépendamment sur chaque branche selon un processus de Poisson d'intensité $\frac{\theta}{2}L$.

Application

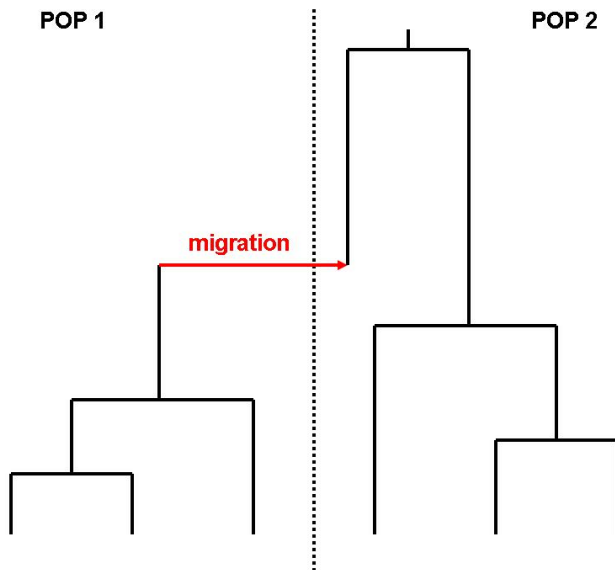
- Population croissante → temps de coalescence plus longs en bas de l'arbre → plus de fréquences alléliques extrêmes.
- Population décroissante → temps de coalescence plus longs en haut de l'arbre → plus de fréquences alléliques intermédiaires.



- 1 Modèle de base : une population de taille constante, un locus sans recombinaison
- 2 Population de taille variable au cours du temps
- 3 Population structurée**
- 4 Modèle avec recombinaison
- 5 Conclusions

- Pour chaque noeud de l'arbre, une étiquette indique sa population d'origine.
- Coalescences intra population. Par ex, pour deux pops 1 et 2, 2 évènements possibles : coalescence dans 1 et coalescence dans 2.
- Taille relative des pops importante car détermine la vitesse des évènements.
- Si migrations entre populations avec une proba par génération en $O(\frac{1}{N})$, évènements "migration de 1 vers 2" et "migration de 2 vers 1" possibles.

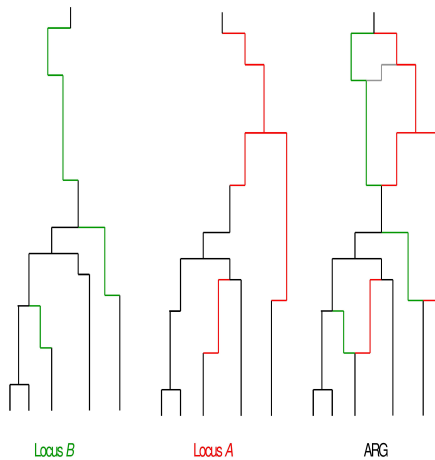
Exemple



- 1 Modèle de base : une population de taille constante, un locus sans recombinaison
- 2 Population de taille variable au cours du temps
- 3 Population structurée
- 4 **Modèle avec recombinaison**
- 5 Conclusions

- Un haplotype créé par recombinaison n'a plus un seul mais deux parents.
- En remontant dans le temps, une branche peut donc se diviser en deux. Cela se produit avec une proba c par génération, c taux de recombinaison entre les locus.
- A l'échelle $\tau = \frac{t}{N}$, et en posant $\rho = 2Nc$, le temps au bout duquel une branche se divise en deux suit une loi exponentielle de paramètre $\frac{\rho}{2}$.
- La généalogie joite des deux locus est appelée "Ancestral Recombination Graph". On peut en déduire les arbres de coalescence marginaux pour chacun des locus (qui sont corrélés).

Exemple



$k = n$. Tant que $k > 1$:

- 1 Simuler U , une loi exponentielle de paramètre $\frac{k(k-1)}{2}$.
- 2 Simuler V , une loi exponentielle de paramètre $\frac{k\rho}{2}$.
- 3
 - Si $U \leq V$, choisir au hasard celui des $\frac{k(k-1)}{2}$ couples d'haplotypes qui coalesce.
 - Si $U > V$, choisir au hasard celle des k lignées qui se sépare.

Remarque : le stade $k = 1$ finit toujours par être atteint, mais cela peut être long.

- 1 Modèle de base : une population de taille constante, un locus sans recombinaison
- 2 Population de taille variable au cours du temps
- 3 Population structurée
- 4 Modèle avec recombinaison
- 5 Conclusions**

Le coalescent en génétique des populations

- Modèle théorique permettant de faire le lien entre des données génétiques observées et des modèles démographiques.
- Outil de simulation très performant.
- Logiciels de simulation : ms, msprime, scrm ...
- En pratique, tous les aspects du modèles (démographie, structure, recombinaison) sont mélangés.

- Simon Tavaré (2001), *Ancestral inference in molecular biology*. Ecole de Probabilités de Saint-Flour.
- Handbook of statistical genetics, Wiley. Chapitres 22.6, 25, 26, 27.3.