ù

Développement d'outils d'inférences de paramètres démographiques à l'aide de données génomiques dans des modèles spatiaux à l'échelle du paysage

Thimothée Virgoulay

Institut des Sciences de lEvolution de Montpellier/Centre Biologique de Gestion des Populations

Directeur de recherche: François Rousset, Chercheur CNRS, ISEM

Co-encadrant: Raphaël LEBLOIS, Chercheur INRA, CBGP

2 mars 2021

Analyse du polymorphisme neutre dans un cadre spatial à l'échelle du paysage

Inférence de paramètres de populations structurées sur de faible échelles géographiques

- Modèles complexes (IBD, MSER, ...)
- Approches analytiques, MCMC, maximum de vraisemblance => trop lents ou impossibles
- Approche d'inférences par simulation (comparaison données réelles/données simulées ex : ABC-RF, SL)
- Difficile d'inférer séparément certains paramètres (taux de mutation/taux de dispersion et densité de la population).

Question : Avec plus d'informations (+ de marqueurs génétiques) ou de nouvelles informations (déséquilibre de liaison) possibilité d'inférer séparément ces paramètres ?

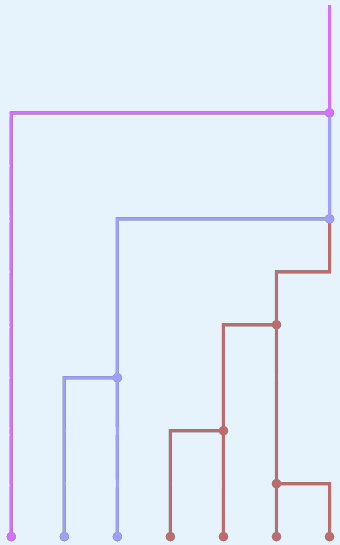
Moyens

Approche d'inférence par simulation :

- Simulateur backward pour des petites populations (libre des hypothèses du *n*-coalescent) (Gspace)
- Librairie de calcul de statistiques résumantes : grands jeux de données et DL (Glib *nom temporaire*)
- Des méthodes d'inférence peu gourmandes en simulation (ABC-RF et SL)

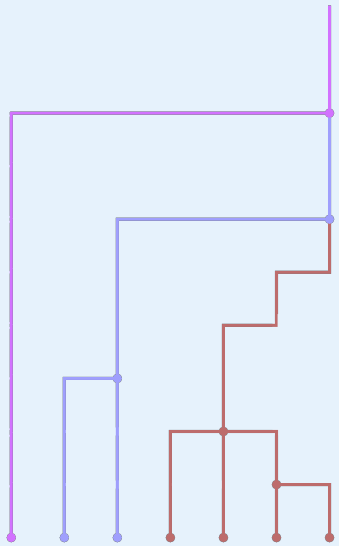
Simulateur en coalescence

- Grandes tailles de populations
- Faible taux d'évènements



Permet d'implémenter
un algorithme en temps "continue"

- Petites tailles de populations
- Forts taux d'évènements



Nécessite un algorithme
en génération par génération

Stat résum DL : Idée générale

Estimateurs : Probabilités d'identités pour 1 ou Probabilités d'identités jointes pour 2 locus.

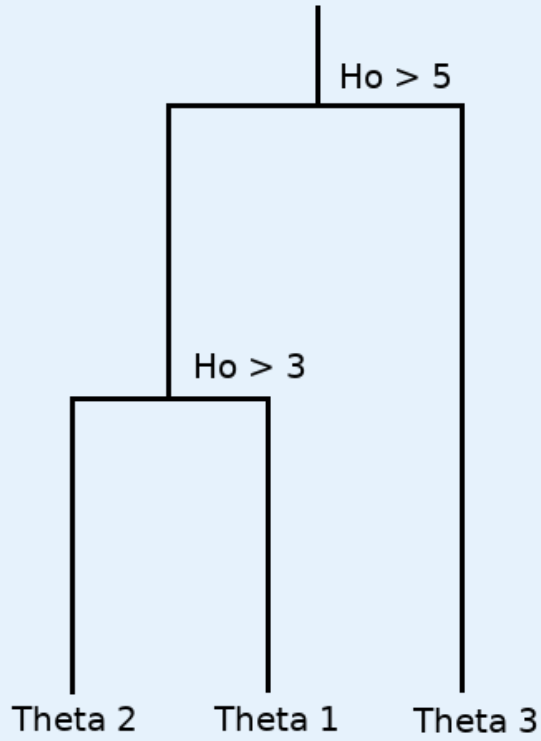
- Numérateur : Estimateurs relatifs au DL entre 2 sous-populations (spatialisé mais variance élevée)
- Dénominateur : Estimateurs relatifs à l'ensemble de l'échantillon considéré (non spatialisé mais variance faible)

$\longrightarrow$ Estimateurs spatialisés avec une variance moins forte.

Random Forest

Création d'un ensemble d'arbres binaires de choix (par bootstrap des simulations)

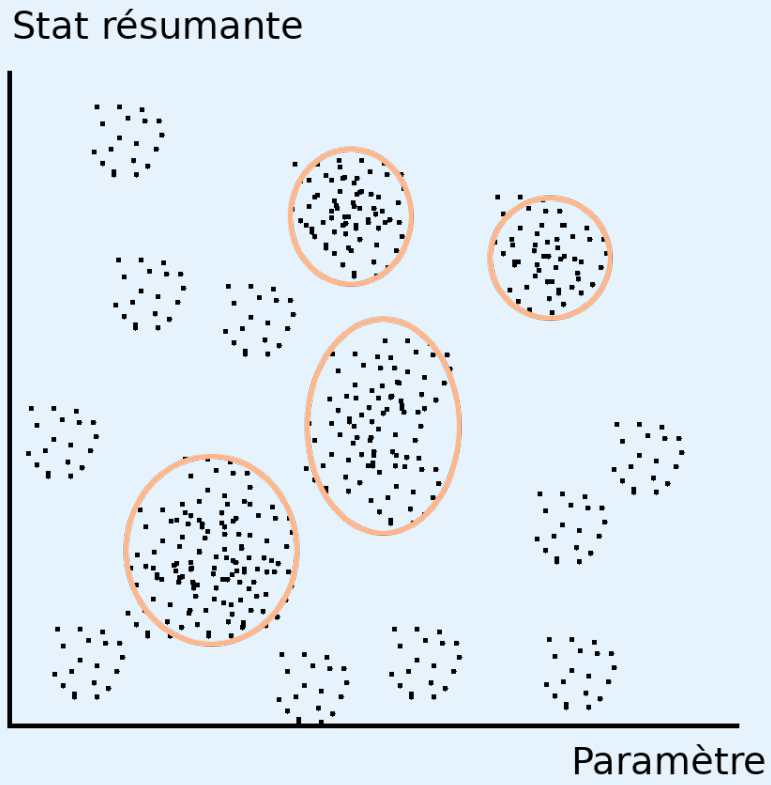
- 1 arbre = Sélection d'un sous ensemble aléatoire de covariables cov_i
- Pour chaque noeud *Si* création d'une règle de décision binaire
- Règle $X_i > t_i$ avec $X_i \in cov_i$ et t_i maximisant un critère
- Jusqu'à cas d'arrêt



Summary Likehood

Approximation de la vraisemblance grâce à des simulations

- Créer une table $\Theta_i \rightarrow S_i$ avec $i \in 0, 1, \ldots n$ simulations d'un ensemble Θ de paramètres et d'un ensemble S de statistiques résumantes
- Mixture de gaussiennes sur l'espace joint des paramètres et des stats résumantes
- Permet d'estimer la surface de vraisemblance
- Procédure itérative pour densifier les points dans la zone de maximum de vraisemblance



Librairie de calcul de statistiques résumantes

Besoins spécifiques

- Que soient calculées des statistiques
  - que l'on peut relier aux paramètres au travers des méthodes ABC-RF et SL (H_e, H_o, N_a, $Q_{r>0}$)
  - dont on peut tirer de l'information sur les paramètres sans passer par les méthodes ABC-RF et SL (F_{stat}, a_r, e_r, η)
- Qu'elles le soient rapidement (grands jeux de données)
- En minimisant les biais (dûs aux données manquantes dans les jeux de données réelles)

Calculs réalisés

- Hobs
- Hexp
- Fis/Fst
- Qr
- ar
- er
- régression linéaire et exponentielle

Stat résum DL : Idée générale

Rapport de probabilités d'identités par paire de dèmes :

- Numérateur : Les estimateurs sont relatifs à votre paire (spatialisé mais variance élevée)
- Dénominateur : Les estimateurs sont relatifs à l'ensemble de l'échantillon considéré (non spatialisé mais variance faible)

→ Estimateur spatialisé avec une variance faible.

$$\widehat{\eta}'_{xy,ij} \equiv \frac{\widehat{\phi}_{xy,ij} - \widehat{Q}_{1j}\widehat{Q}_{1i}}{(1-\widehat{Q}_{2i})(1-\widehat{Q}_{2j})}$$

$\widehat{\phi}_{xy,ij}$ proba d'identité à 2 locus calculée entre les locus i et j diploïde entre le deme x et le deme y
$\widehat{Q}_{1i}$ proba d'identité calculée pour le locus i au sein des dèmes x et y (inter-indiv)
$\widehat{Q}_{2i}$ proba d'identité calculée pour le locus i entre tous les dèmes

Statistiques d'intérêts

Dans le cadre d'un MSER :

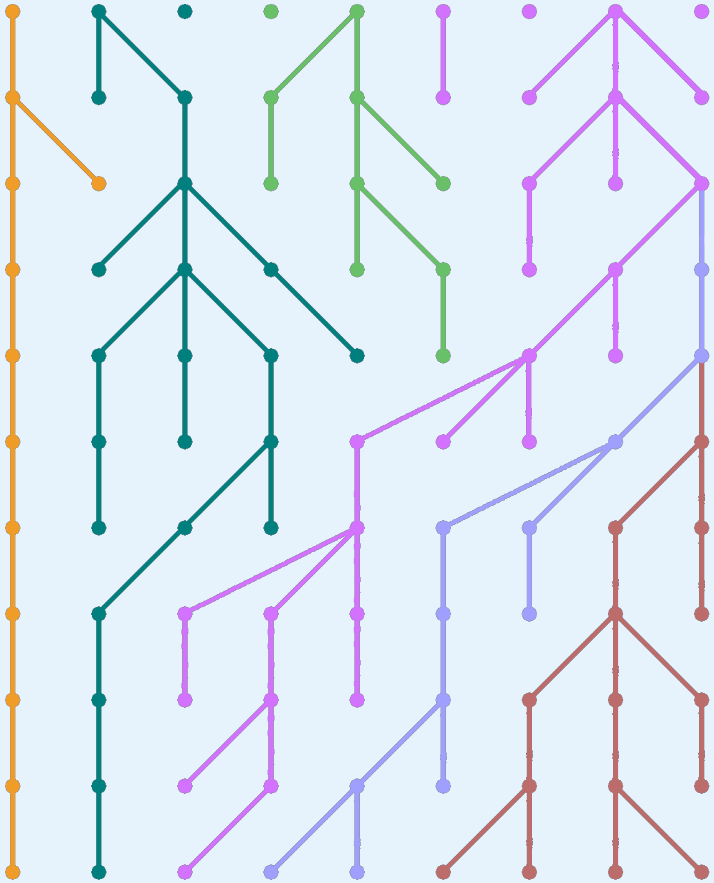
$$\widehat{\eta'}_{xy,ij} \equiv \frac{\widehat{\phi}_{xy,ij} - \widehat{Q}_{0j}\widehat{Q}_{0i}}{(1 - \widehat{Q}_{2i})(1 - \widehat{Q}_{2j})}$$

$\widehat{\phi}_{xy,ij}$ proba d'identité à 2 locus calculée entre les locus i et j diploïde entre l'indiv x et l'indiv y

$\widehat{Q}_{0i}$ proba d'identité calculée intra-indiv pour le locus i et pour l'ensemble des indiv

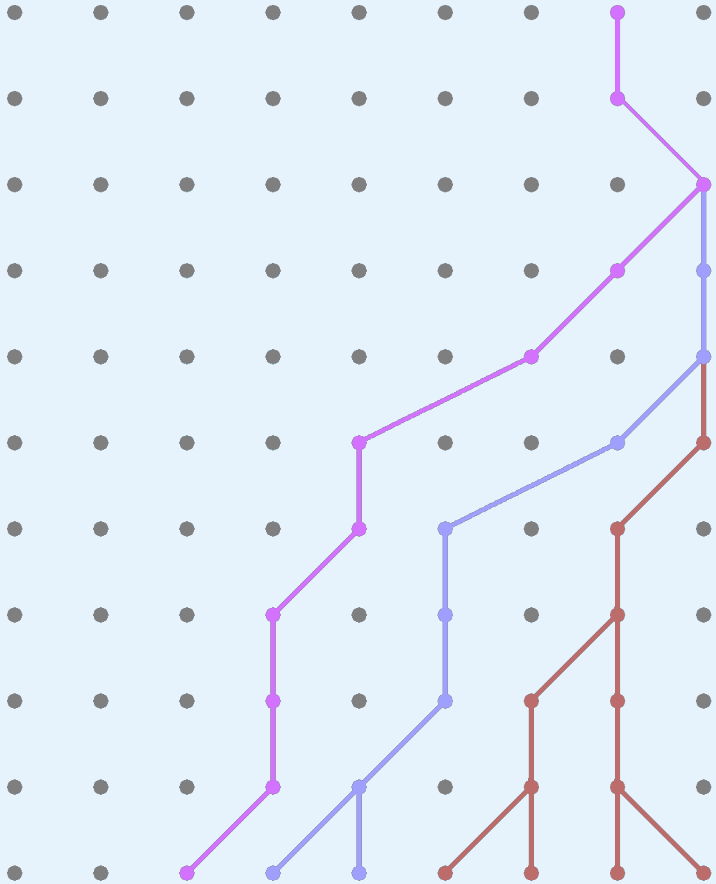
$\widehat{Q}_{2i}$ proba d'identité calculée pour le locus i entre tous les indiv

Méthode de simulation : Forward



- Simulation en avançant dans le temps
- Suivre toutes les lignées génétiques de la population
- Coûteux en ressources informatiques

Méthode de simulation : Backward

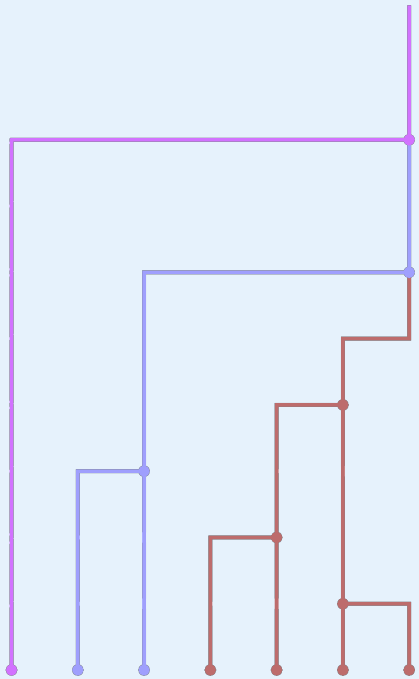


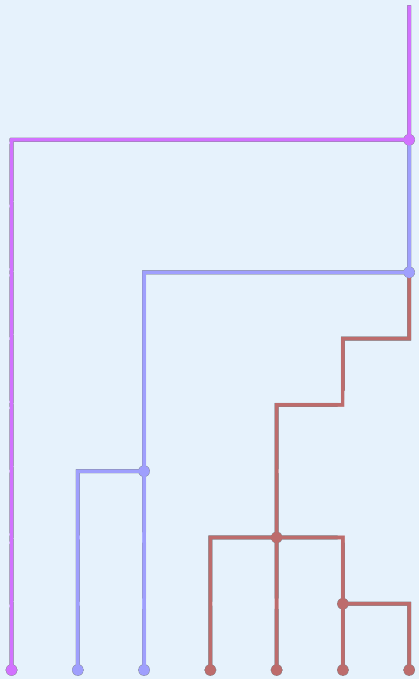
- Simulation en "coalescence"
- Simulation en reculant dans le temps
- Suivre les lignées génétiques de l'échantillon
- Moins coûteux en ressources informatiques

Coalescence

Algorithme en temps exponentiel : probabilité de 2 évènements simultanés (coalescence, recombinaison, migration) $= 0$

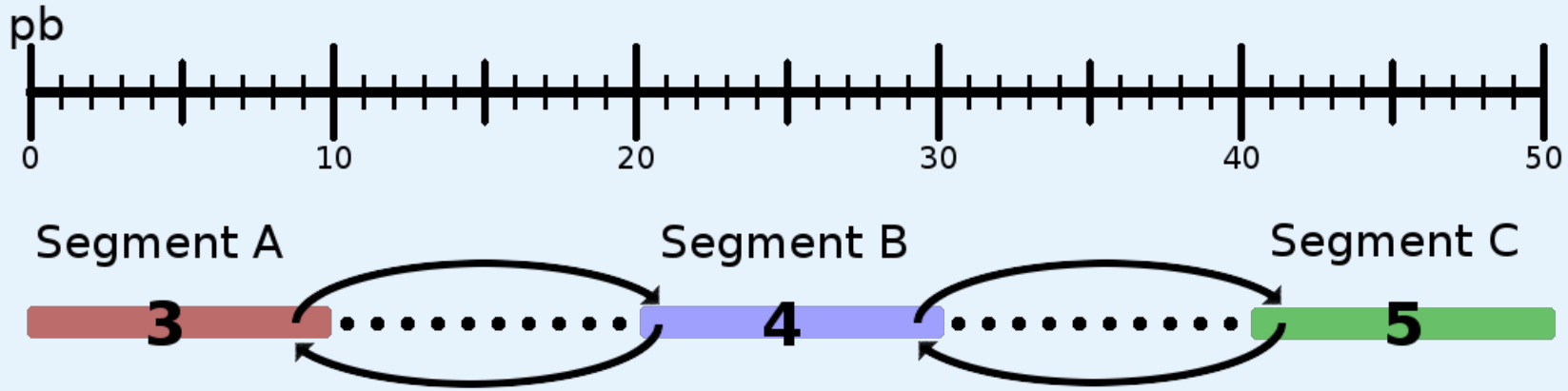
Algorithme génération/génération : Probabilité non nulle que plusieurs évènements arrivent à la même génération





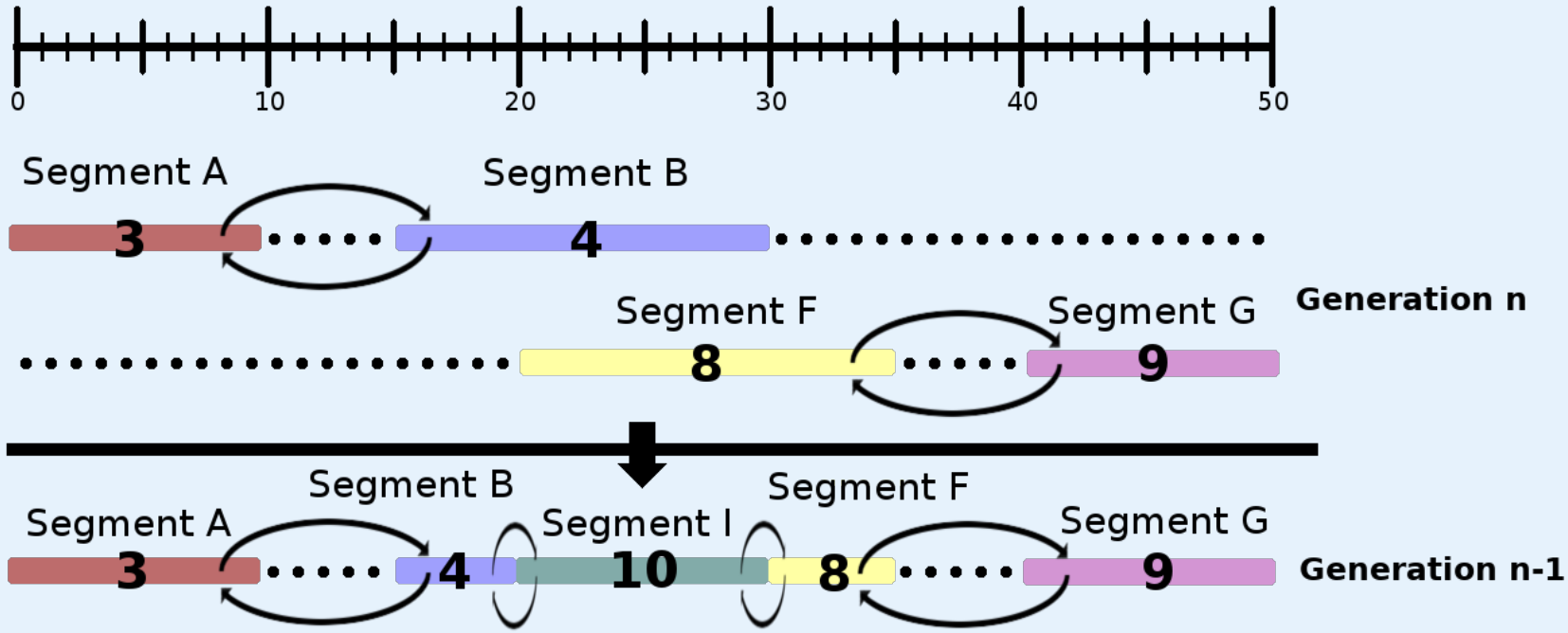
Segments

Segments : Portion physique du chromosome



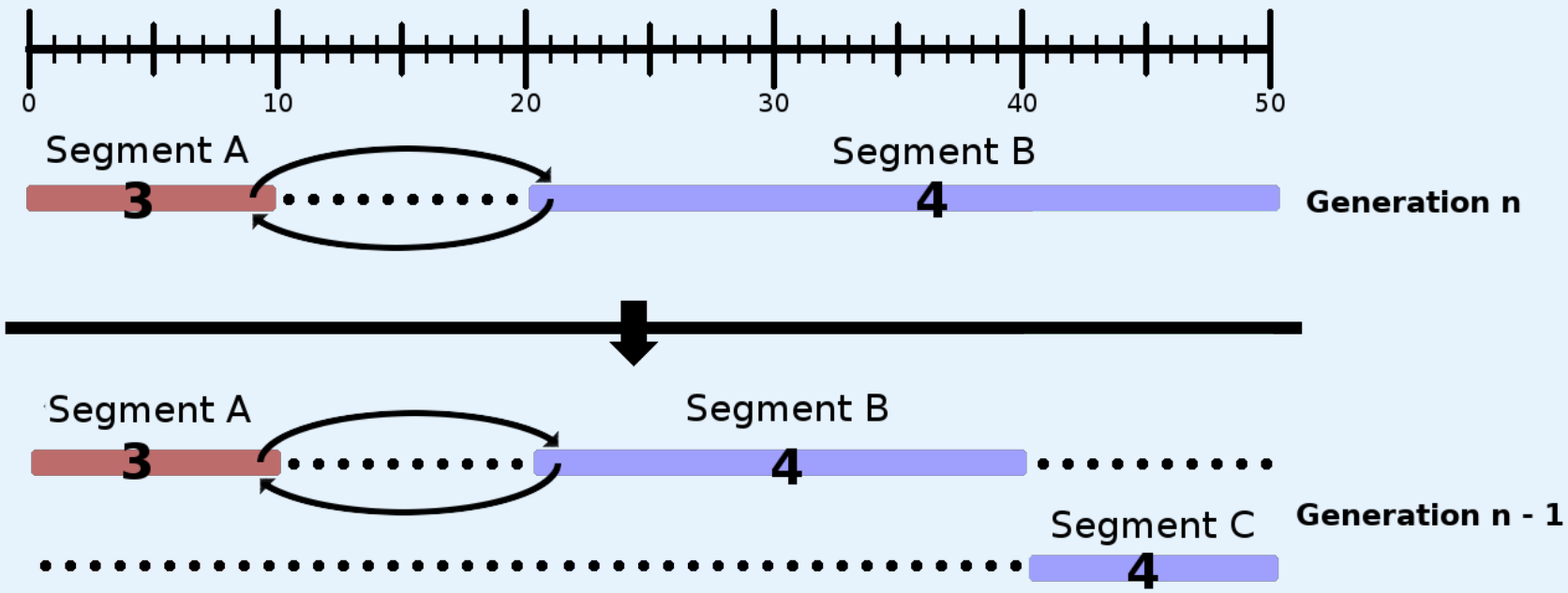
Coalescence

Fusionner 2 chromosomes (et les lignées génétiques qu'ils portent)



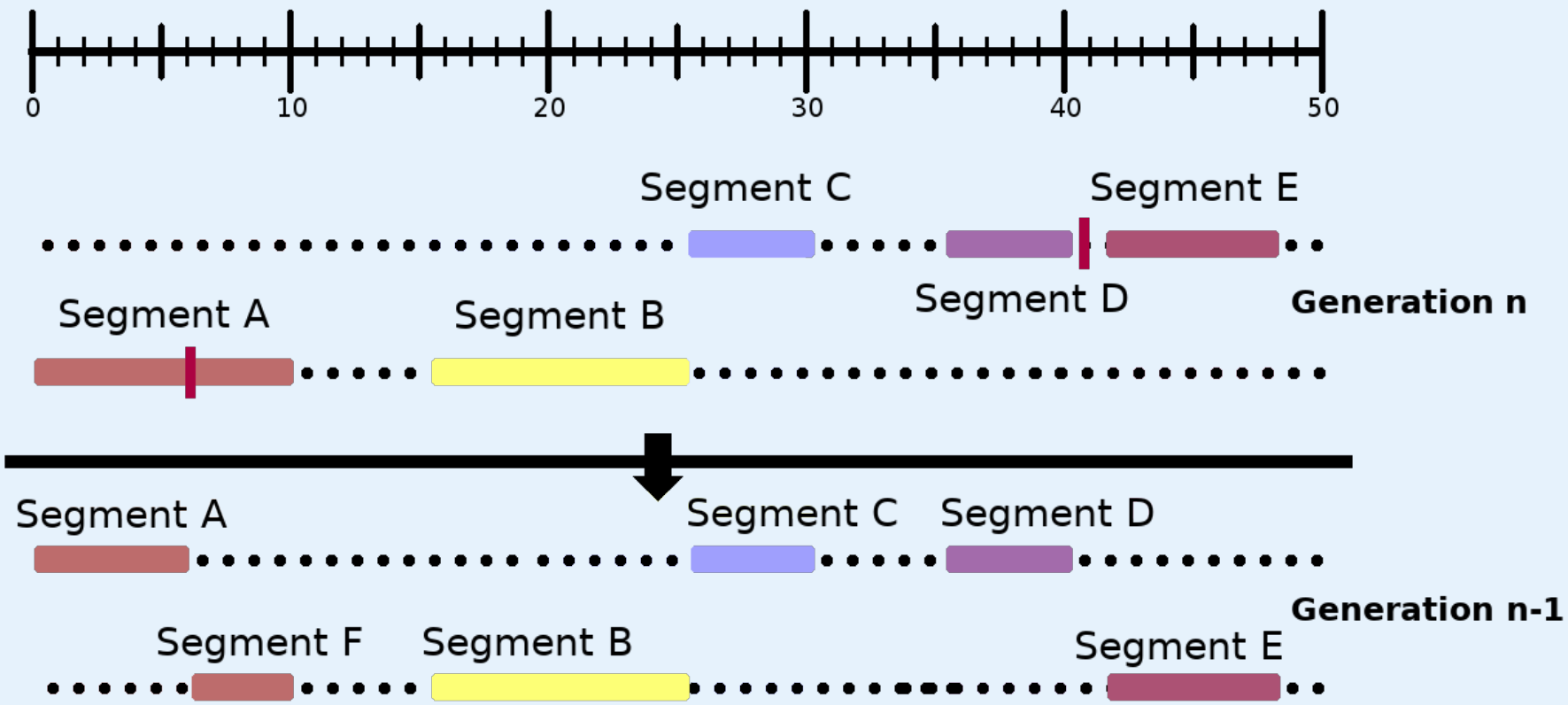
Recombinaison

Séparer un chromosome en deux (potentiellement une lignée génétique en deux)

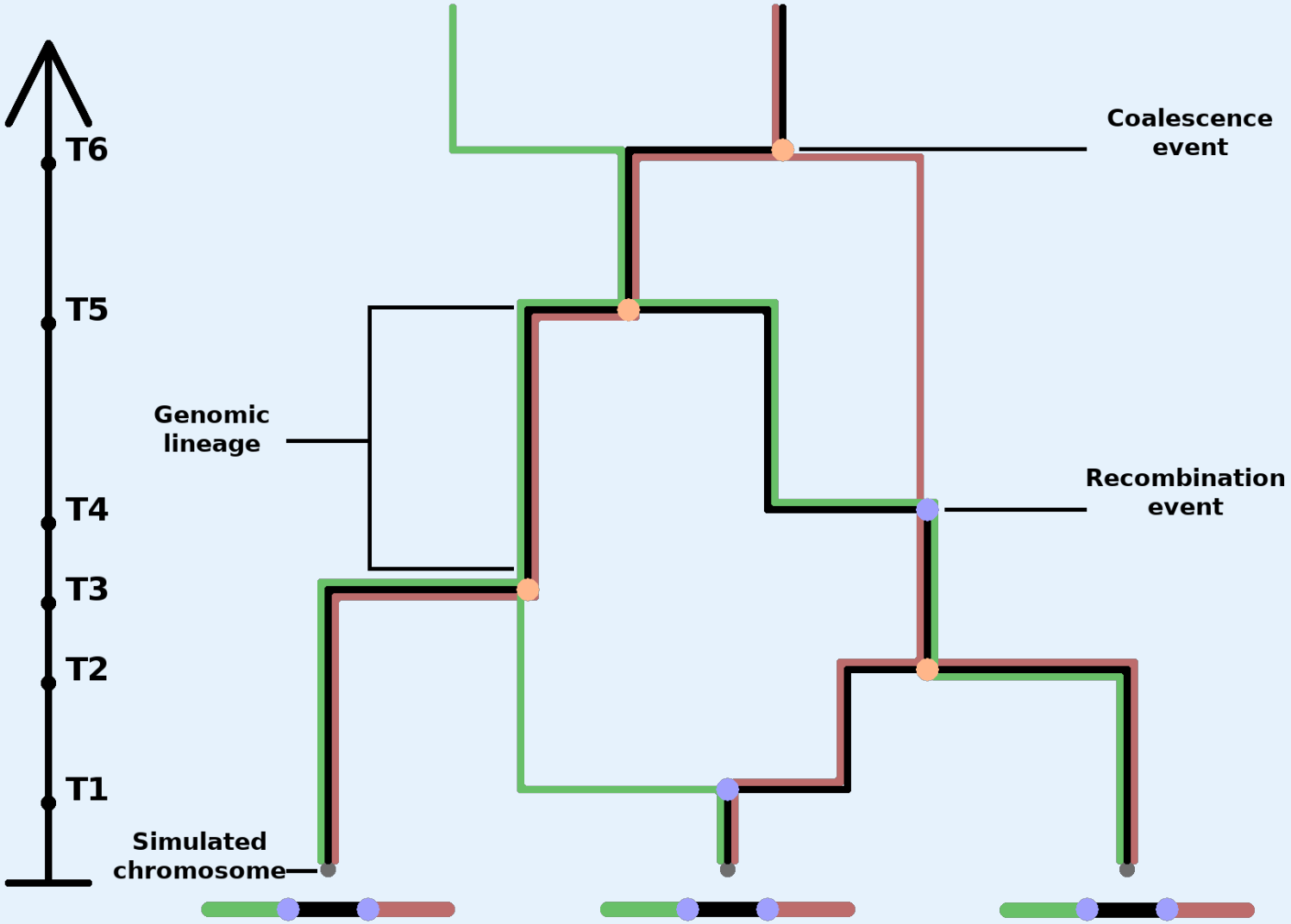


Recombinaison de 2 segments

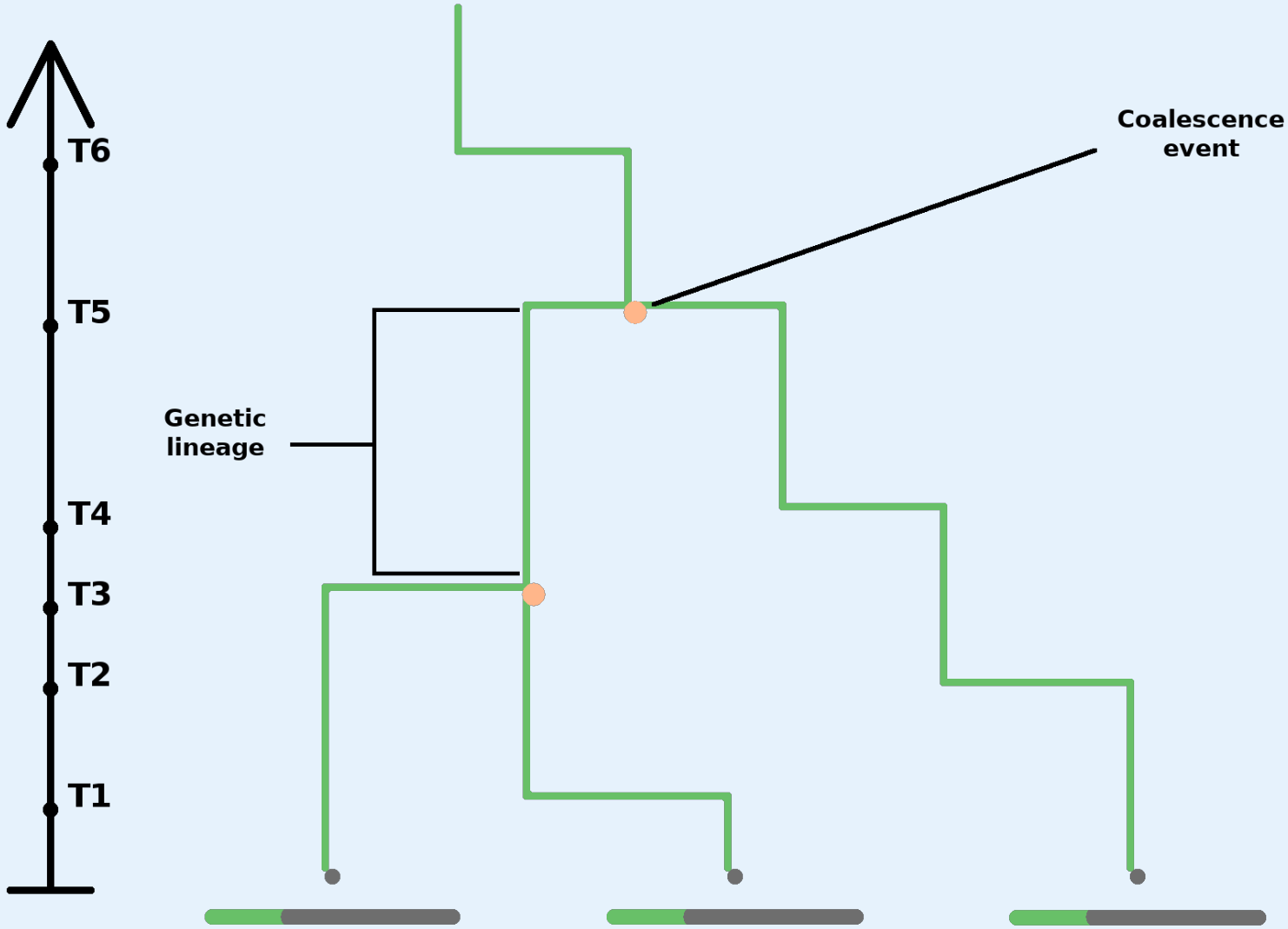
Plusieurs évènements de recombinaisons pendant la même génération : 2 segments



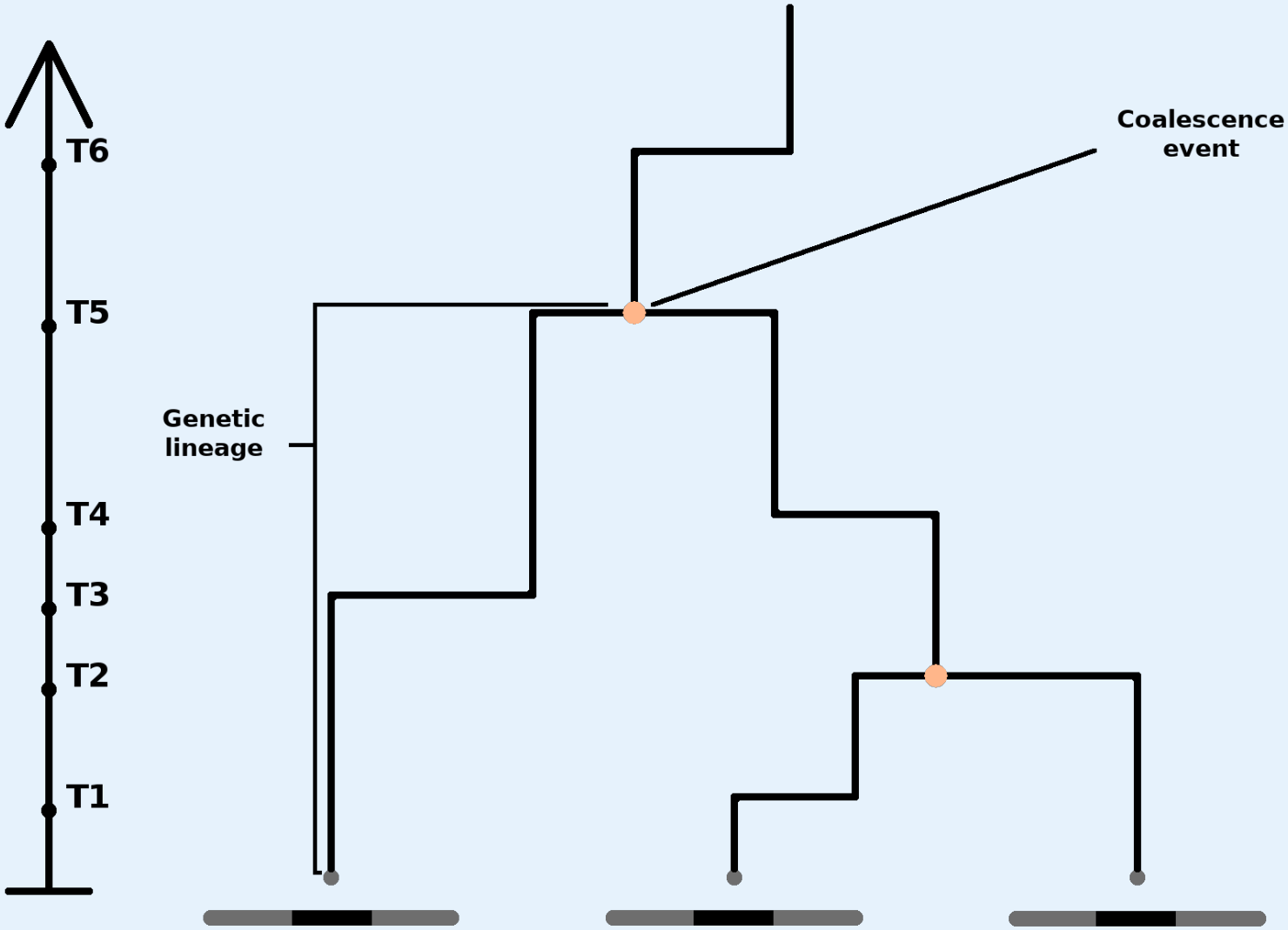
Graphe ancestral de recombinaison



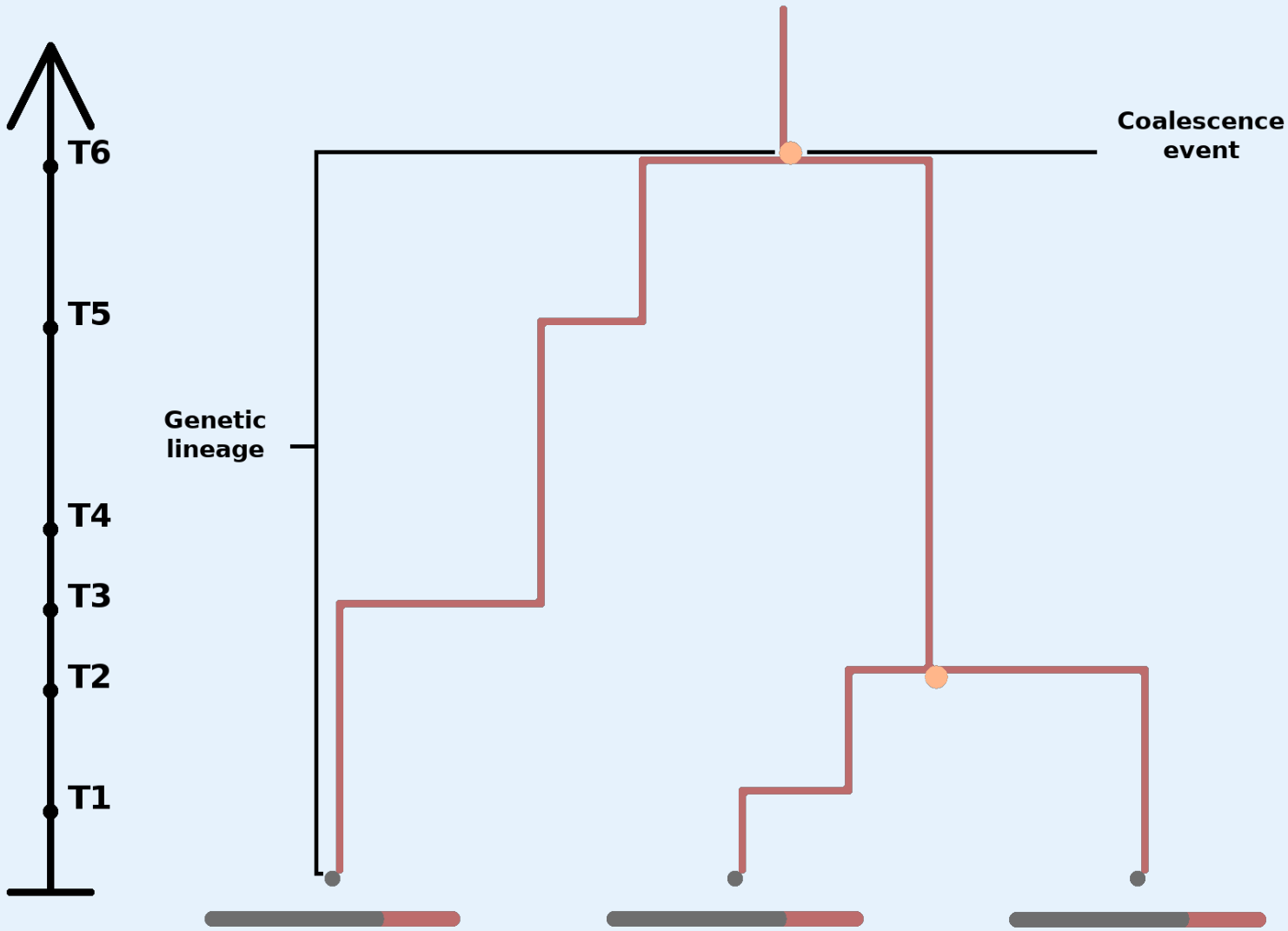
Graphe ancestral de recombinaison



Graphe ancestral de recombinaison

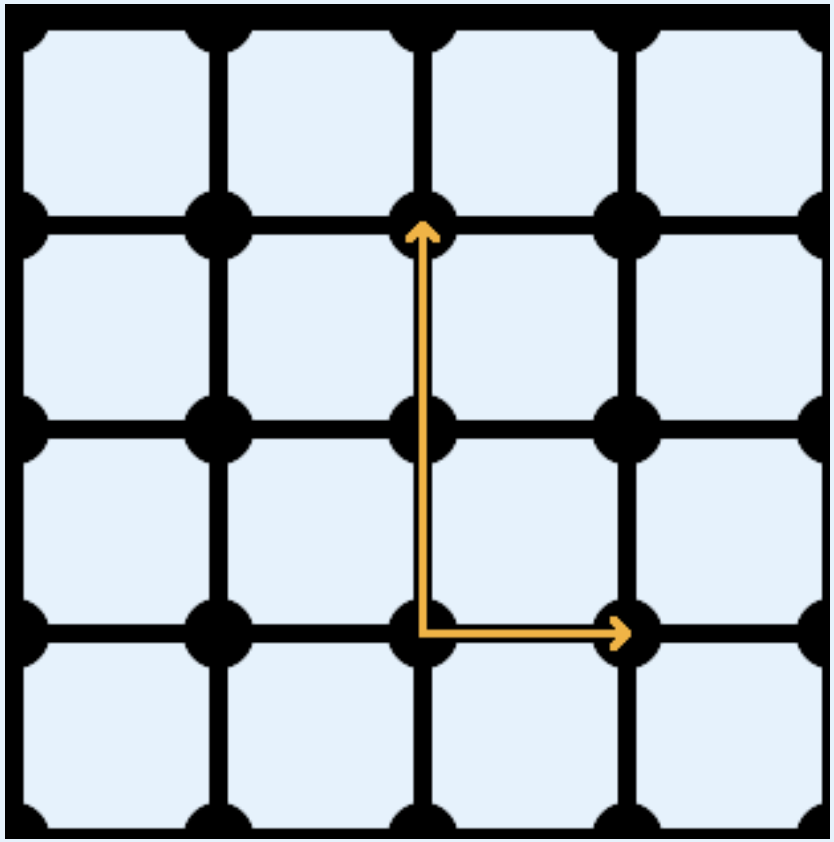


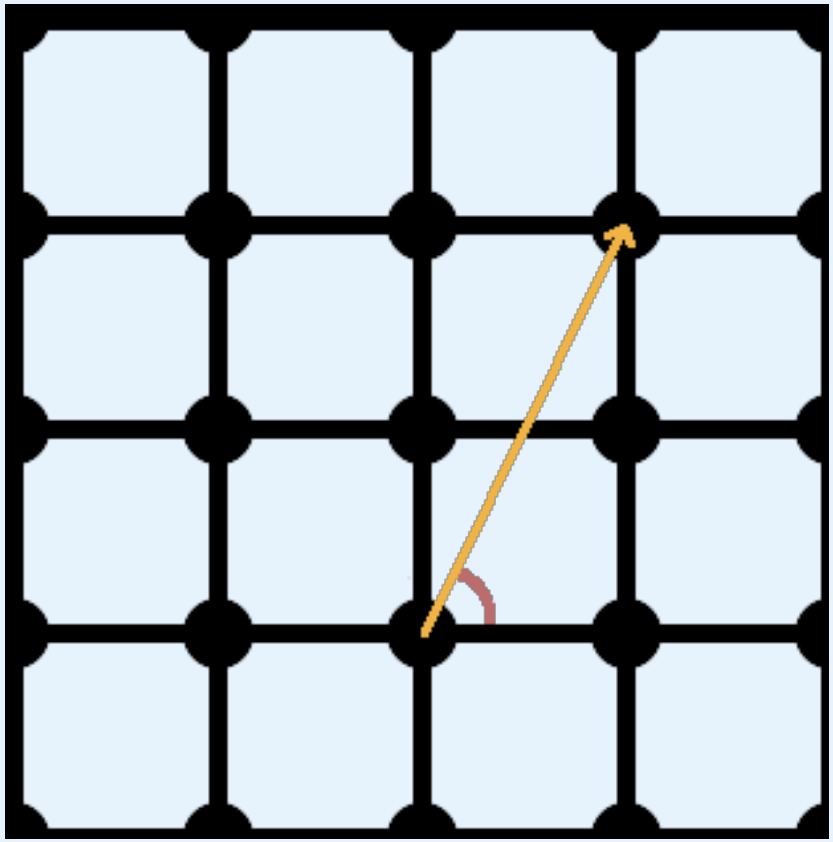
Graphe ancestral de recombinaison



Lattice et déplacement

Grille où les sous-populations sont placées aux intersections

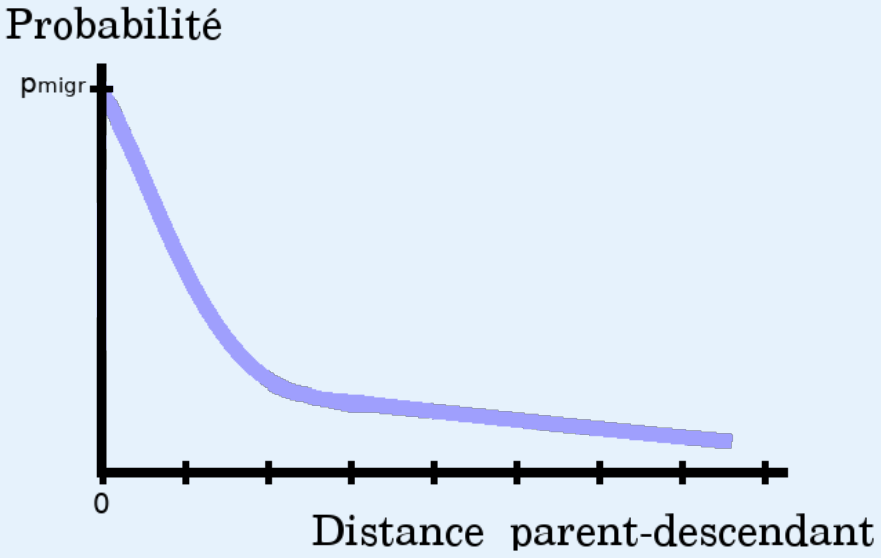




Cas homogène

Déplacement en x, y avec une probabilité p_{migr} suivant une loi géométrique de paramètre g

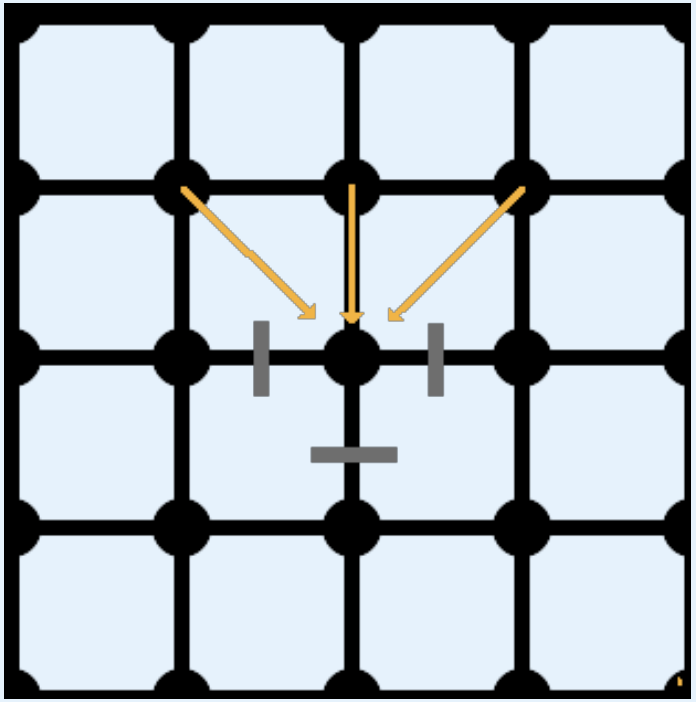
$$backward\_distrib_{dx,dy}() = forward\_distrib_{dx,dy}()$$



Cas hétérogène

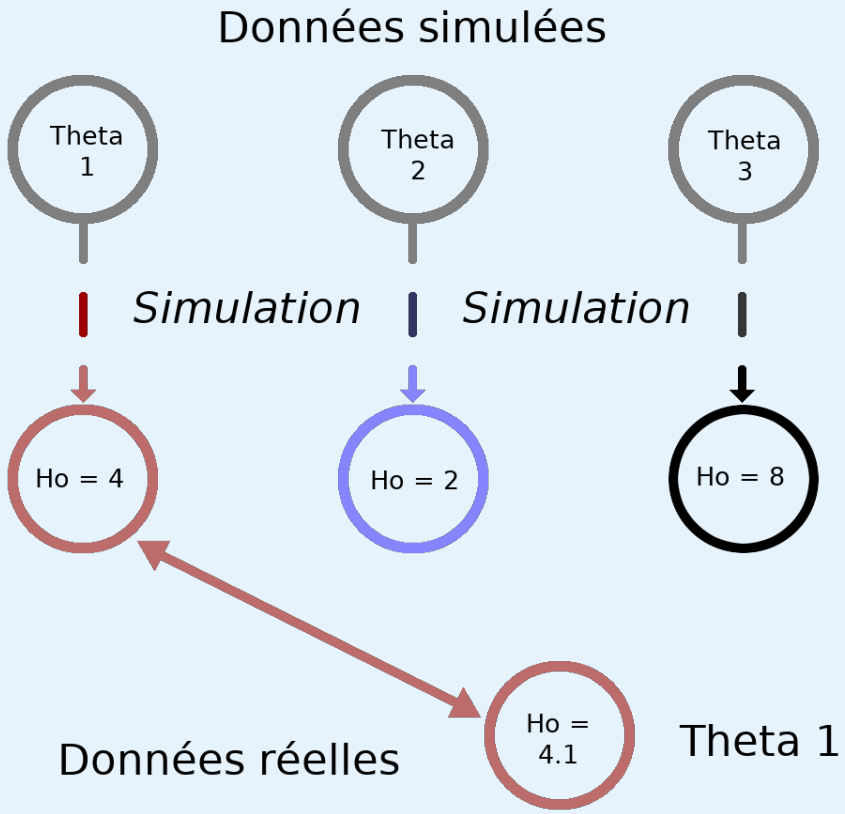
La distribution backward prend en compte les distributions forward avec un K_{max} (la distance de dispersion max) suffisant.

$$backward\_distrib_{dx,dy}() = \frac{N_{x+dx,y+dy} \cdot f_{dx,dy}}{\sum_{dx,dy \leq K_{max}} N_{x+dx,y+dy} \cdot f_{dx,dy}}$$



ABC

- "L'Approximate Bayesian Computation" (ABC) permet d'inférer des paramètres Θ grâce à la comparaison de jeux de données simulés et de données réelles à travers des statistiques résumantes



Avantages

- ABC général : Remédie aux cas où la fonction de vraisemblance n'est pas calculable
- ABC-rf : Pas sensible aux statistiques résumantes non informatives (bruit)
- ABC-rf : Nécessite beaucoup moins de simulations que l'ABC classique

Avantages

- Estimation de la surface de vraisemblance
  Profils de vraisemblance en 1D ou 2D
- Précision dépendante du nombre d'itérations
- Possibilité de réduire l'espace des stats résumantes en utilisant des stats résumantes synthétiques crées grâce à l'aide d'une projection