Estimation de l'histoire démographique d'une population à partir de données génomiques haut débit

Simon Boitard

INRA, Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Toulouse

Séminaire de l'UMR CBGP
9 novembre 2018

Plan de l'exposé

Plan de l'exposé

Parcours scientifique

1999–2002 **Ecole d'ingénieur en mathématiques appliquées et informatique** ENSIMAG, Grenoble

2002–2006 **M2 puis Doctorat en Probabilités / Statistique**
INRA MIAT & Université Paul Sabatier, Toulouse
Directeurs : Brigitte Mangin et Jean-Marc Azaïs

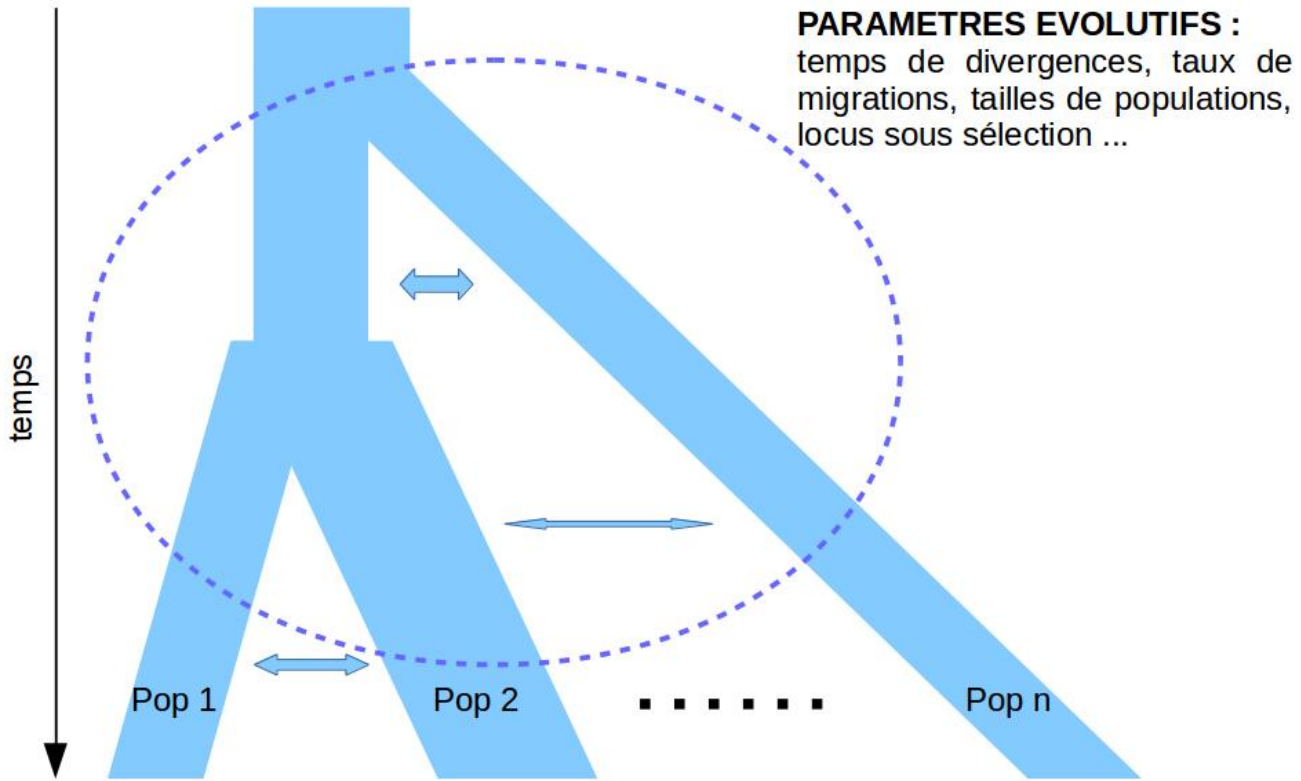
2007 **Stage postdoctoral** Université de Vienne
Directeur : Andreas Futschik

2008-2012 **Chargé de Recherches** INRA, LGC, Toulouse
Département Génétique Animale

2012-2015 **Chargé de Recherches** INRA, GABI, Jouy-en-Josas
Mise à disposition 80% au MNHN, ISYEB.

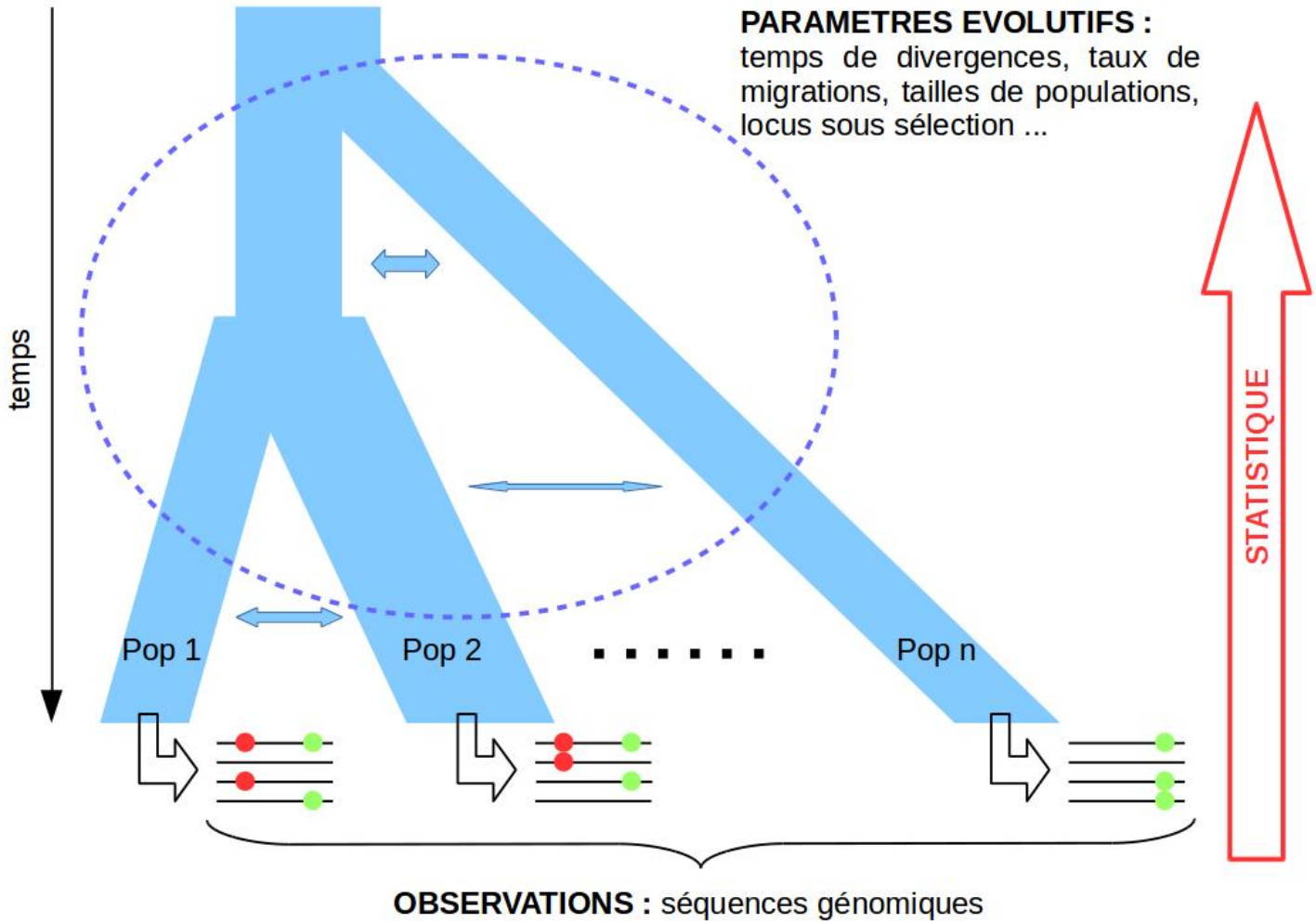
2016-auj **Chargé de Recherches** INRA, GenPhySE, Toulouse

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



Génétique des populations

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



Données génomiques

- **Fréquences des allèles** pour les positions polymorphes.

A-A-C-G-**G**-G-T-A-**T**-C-G- ....

A-A-C-G-**G**-G-T-A-**A**-C-G- ....

A-A-C-G-**C**-G-T-A-**T**-C-G- ....

- **SNP** (Single Nucleotide Polymorphism) : très nombreux, obtenus par puces ou **NGS** (Next Generation Sequencing).
- SNP proches $\rightarrow$ fréquences alléliques **corrélées** : **Déséquilibre de Liaison (LD)**.

Infinite site model

- Pour un SNP, **une seule mutation** au cours de l'évolution $\rightarrow$ deux allèles, un **ancestral** (0) et un **dérivé** (1).

0-0-0-0-**1**-0-0-0-**0**-0-0- ....

0-0-0-0-**1**-0-0-0-**1**-0-0- ....

0-0-0-0-**0**-0-0-0-**0**-0-0- ....

- y_i nombre d'allèles 1 au SNP i.

Détection de locus sous sélection positive (2007-auj)

Approche **intra population**: zones de faible diversité génétique.

- Statistiques résumant la diversité génétique d'une région : D de Tajima, F de Fu et Li ...
- **Probabilité d'un sweep** dans une région sachant les **fréquences alléliques** (Kim et Stephan, 2002; Nielsen *et al*, 2005).

1. Prise en compte de la **corrélation entre SNP** par un modèle HMM (Boitard *et al*, Genetics 2009).
2. Utilisation de données **Pool-seq** (Boitard *et al*, MBE 2012; Mol Eco Res 2013).

Détection de locus sous sélection positive (2007-auj)

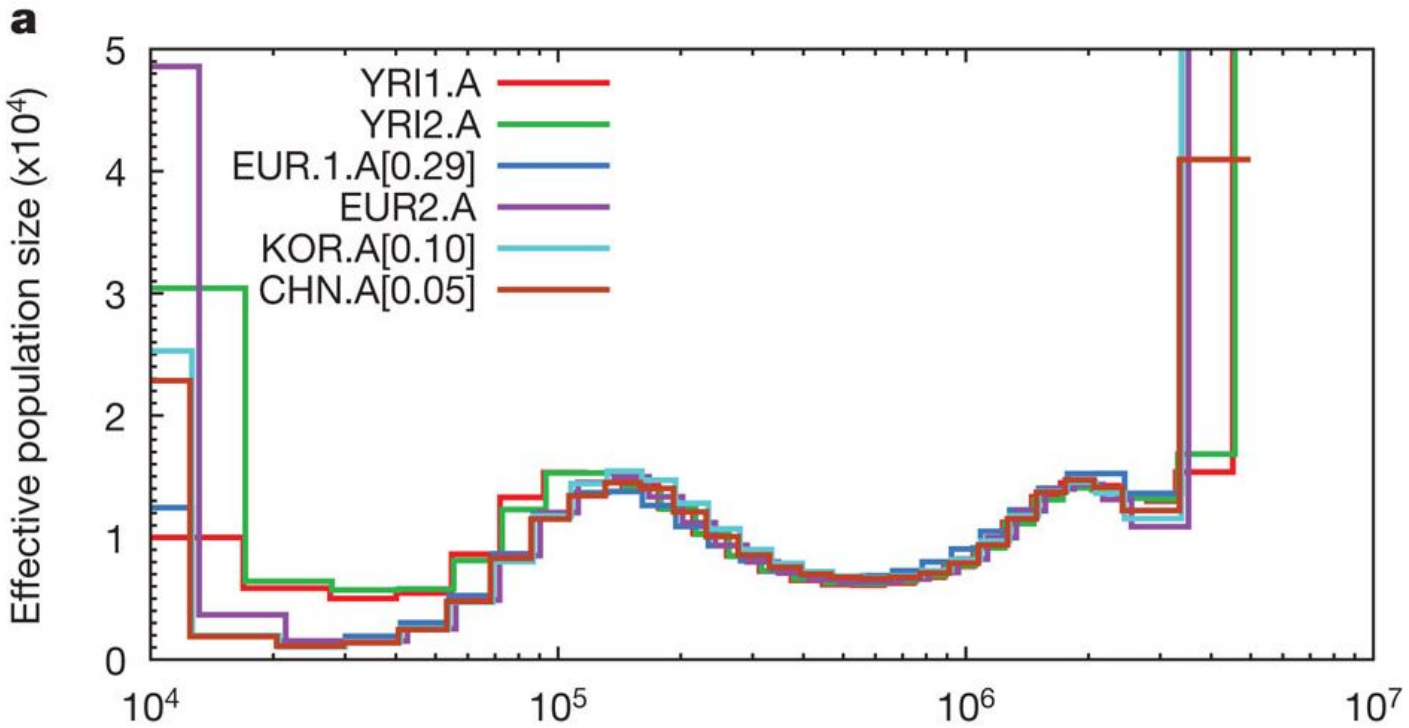
Approche **inter populations**: zones de forte différentiation génétique.

- Classiquement mesurée par le F_{ST}.

1. FLK : prise en compte de la **phylogénie** des populations (Bonhomme *et al*, Genetics 2010).
2. hapFLK : utiliser le LD via les **haplotypes** (Fariello *et al*, Genetics 2013).
3. Score local : utiliser le LD en **cumulant les tests de SNP proches** (Fariello *et al*, Mol Eco 2017).

Estimation de l'histoire démographique (2012-auj)

- **Variations de la taille d'une population** dans le passé.



Li et Durbin (2011)

- Interprétation, détection de populations menacées.

Espèces d'élevage

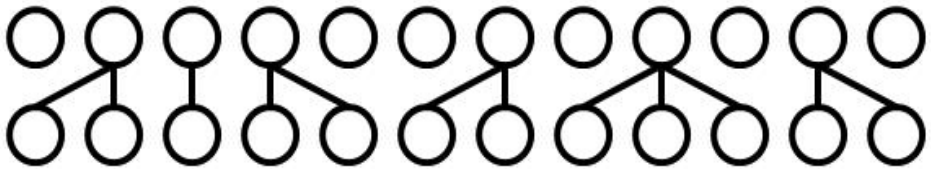
- **Intérêt agronomique** des QTL ou locus sous sélection.
- Des **questions spécifiques** : domestication, diversification et création des races, sélection intensive moderne ...
- **Données très riches** :
  - Animaux génotypés en routine pour la sélection.
  - Projets internationaux de séquençage : plus de 2,000 génomes disponibles chez la vache.

Plan de l'exposé

Plan de l'exposé

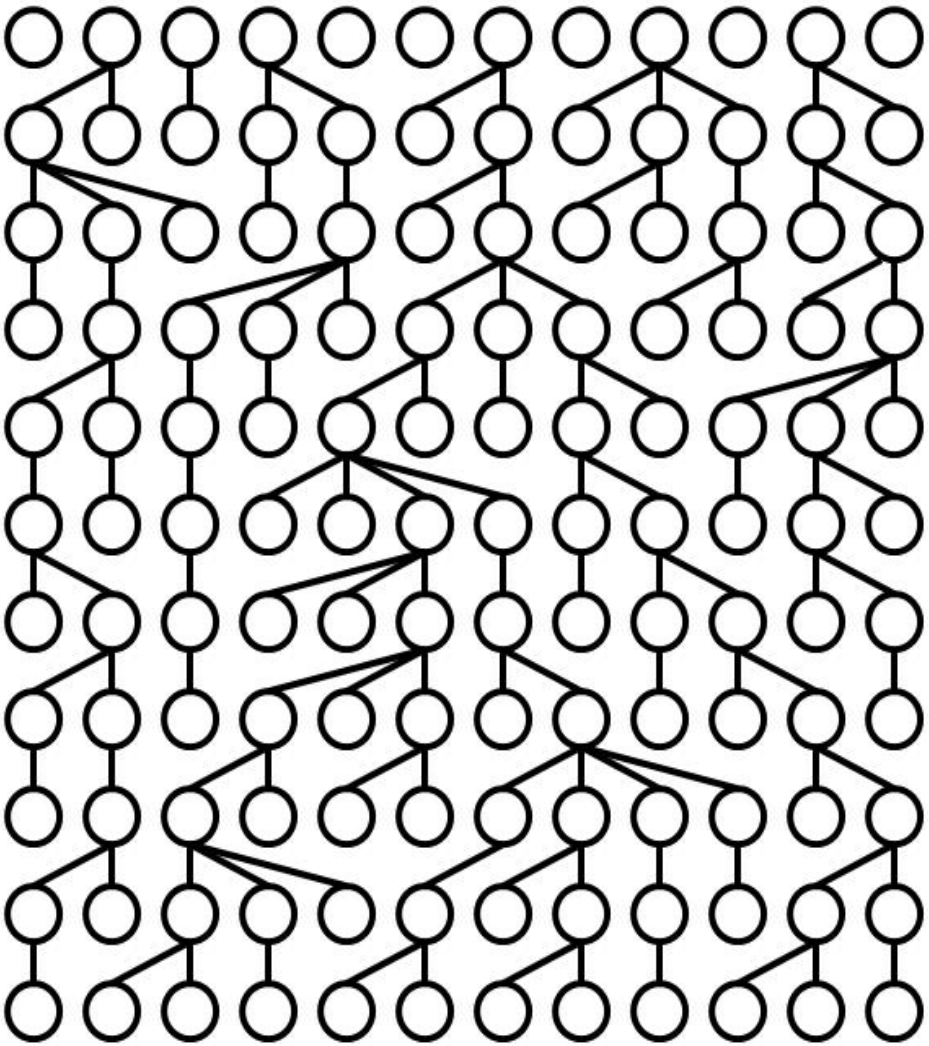
Modèle de Wright-Fisher à un locus

Modèle de Wright-Fisher à un locus



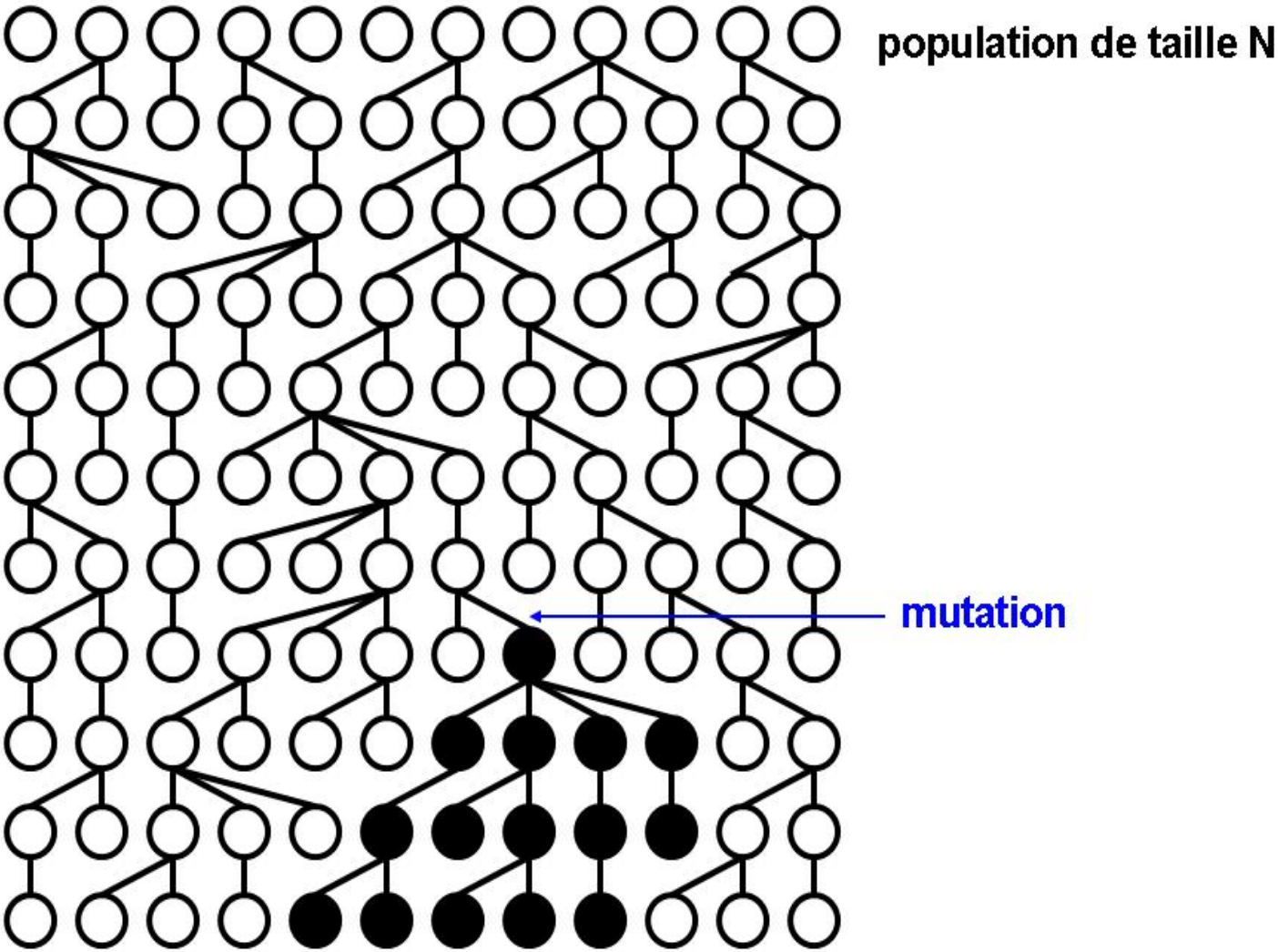
population de taille N

Modèle de Wright-Fisher à un locus

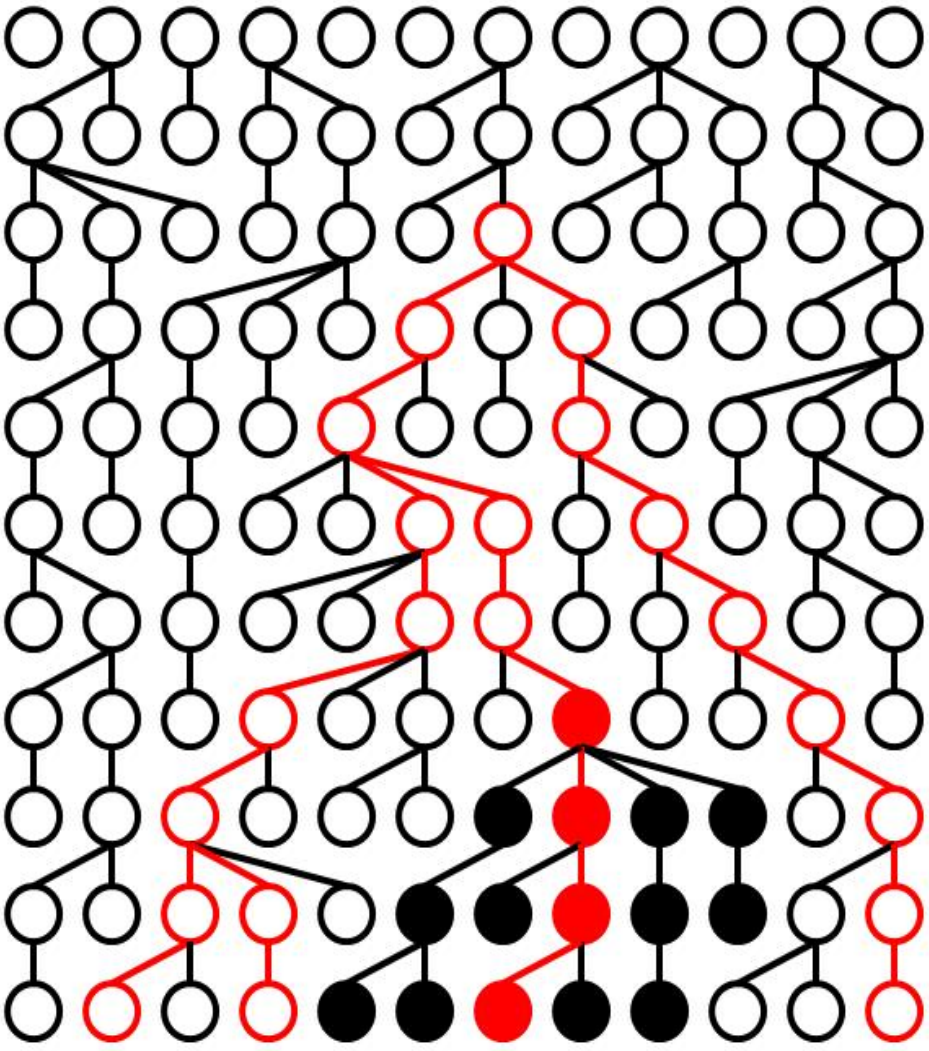


population de taille N

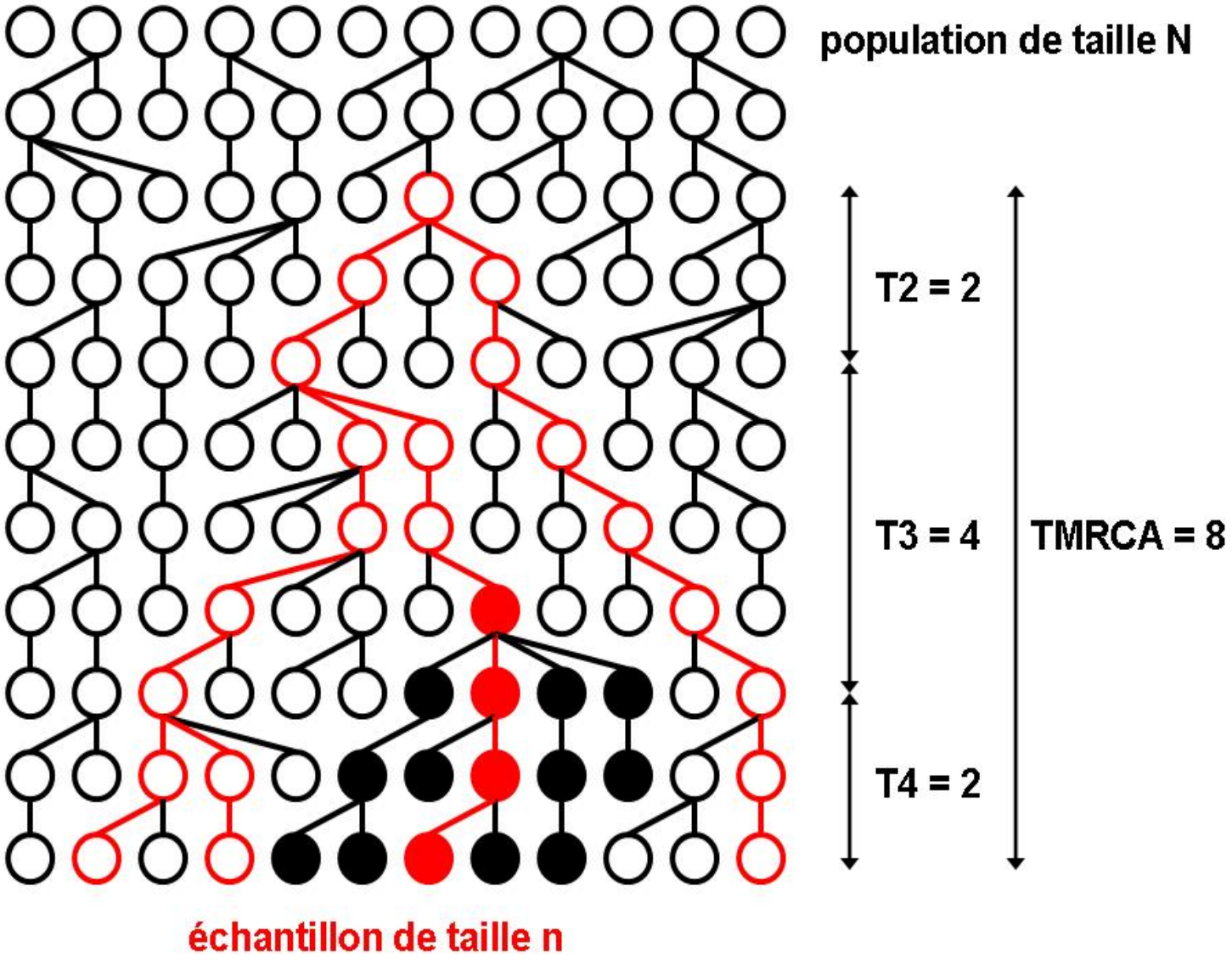
Modèle de Wright-Fisher à un locus



Généalogie et coalescence



Généalogie et coalescence

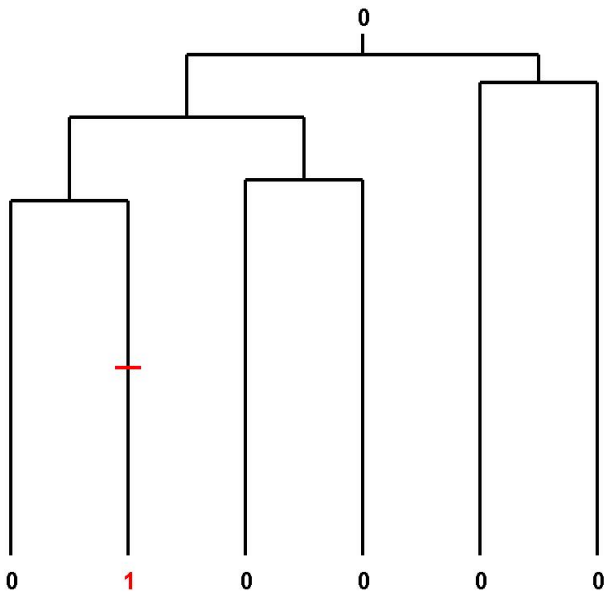


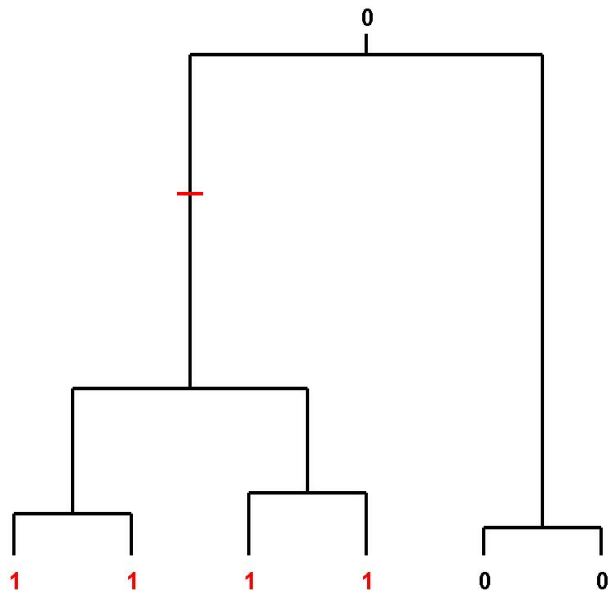
Lien avec la taille efficace

1. Le temps de coalescence augmente avec la taille efficace.
2. Le nombre de mutations sur une branche augmente avec le temps de coalescence.

Application

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





Estimation

- Pour **un locus i sans recombinaison** :

$$\mathbb{P}(\mathcal{D}_i \mid N()) = \sum_G \mathbb{P}(\mathcal{D}_i \mid G)\mathbb{P}(G \mid N())$$

  $\mathcal{D}_i$ séquences observées, $N()$ démographie, G généalogie.

- Pour p **locus indépendants** :

$$\mathbb{P}(\mathcal{D} \mid N()) = \prod_{i=1}^{p} \mathbb{P}(\mathcal{D}_i \mid N())$$

- Méthodes Msvar (Beaumont, 1999), Bottleneck (Piry *et al*, 1999), Beast (Drummond et Rambaut, 2007), VarEff (Nikolic et Chevalet, 2014) ...

Données génomiques haut-débit

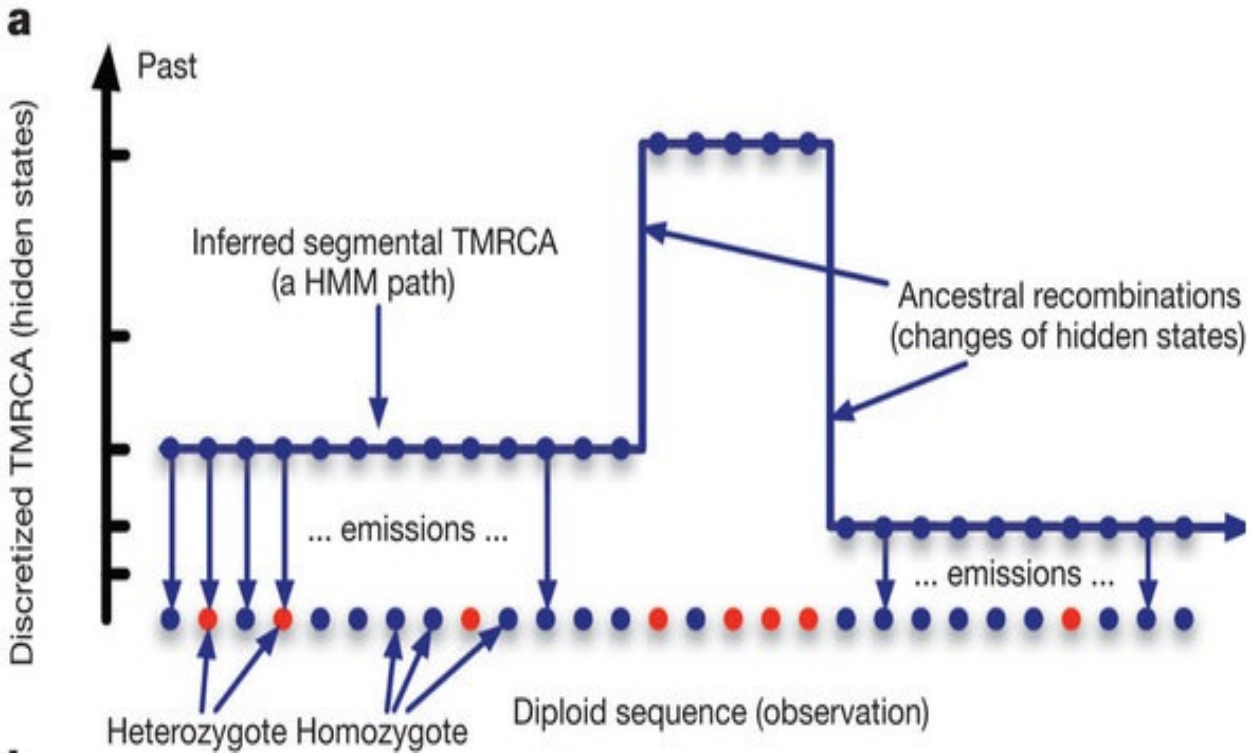
- **Intérêt :**
  - Observation (indirecte) d'un **très grand nombre de généalogies** issues du **même modèle démographique**.
  - Estimation plus précise de ce modèle.
- **Obstacles :**
  - **Généalogies** G_i et G_j pour deux locus proches différentes mais **corrélées**.
  - Corrélation difficile à modéliser, prise en compte de la **recombinaison**.
- **Besoin de nouvelles méthodes** adaptées.

SMC = Sequentially Markovian Coalescent

- **Approximation de Markov**: $G_{i+1} = f(G_i)$
  $\rightarrow$ Estimation par chaîne de Markov cachée (HMM).
- **Avantages:**
  - Données complètes, vraisemblance exacte.
- **Inconvénients:**
  - **Longs fragments** d'ADN continus (chromosomes).
  - Petit nombre d'individus ($\approx$ 5 diploïdes max).
    $\rightarrow$ **Faible précison pour la démographie récente**.
- Exemples : PSMC (Li et Durbin, 2011), dical (Sheehan *et al*, 2013), MSMC (Schiffels et Durbin, 2014).

PSMC = Pairwise SMC

- Un individu diploïde.
- Généalogie simplifiée: $G = T_2$.



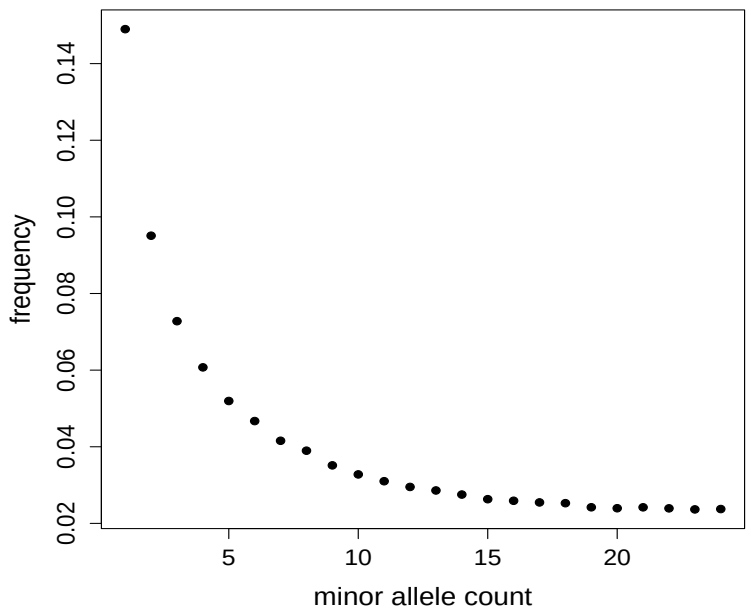
- Estime la démographie (paramètres du modèle) mais aussi la distribution de T_2 (variable cachée) associée.

Données résumées

- Remplacer les données complètes $\mathcal{D}$ par un ensemble de statistiques $\mathcal{S}$ résumant ces données.
- **Avantages:**
  - $\mathbb{P}(\mathcal{S} \mid N())$ **calculable**, $\mathbb{P}(\mathcal{D} \mid N())$ non.
  - Possible si ADN discontinu (RADseq, contigs courts).
- **Inconvénients:**
  - Perte d'information, estimation un peu moins précise.

Exemples de statistiques

- **Spectre des Fréquences Alléliques (AFS):** Bhaskar *et al* (2015), Liu *et al* (2015), Waltoft et Hobolt (2017).



- Lien théorique avec l'espérance des temps de coalescence ($\mathbb{E}[T_i], i = 2 \dots n$) (Griffiths et Tavaré, 1998).

Exemples de statistiques

- Décroissance du **Déséquilibre de Liaison (LD)** avec la distance génétique : Hayes *et al* (2003).
- Longueur des segments **Identiques par Descendance (IBD):** McLeod et al (2013), Harris et Nielsen (2013).

Plan de l'exposé

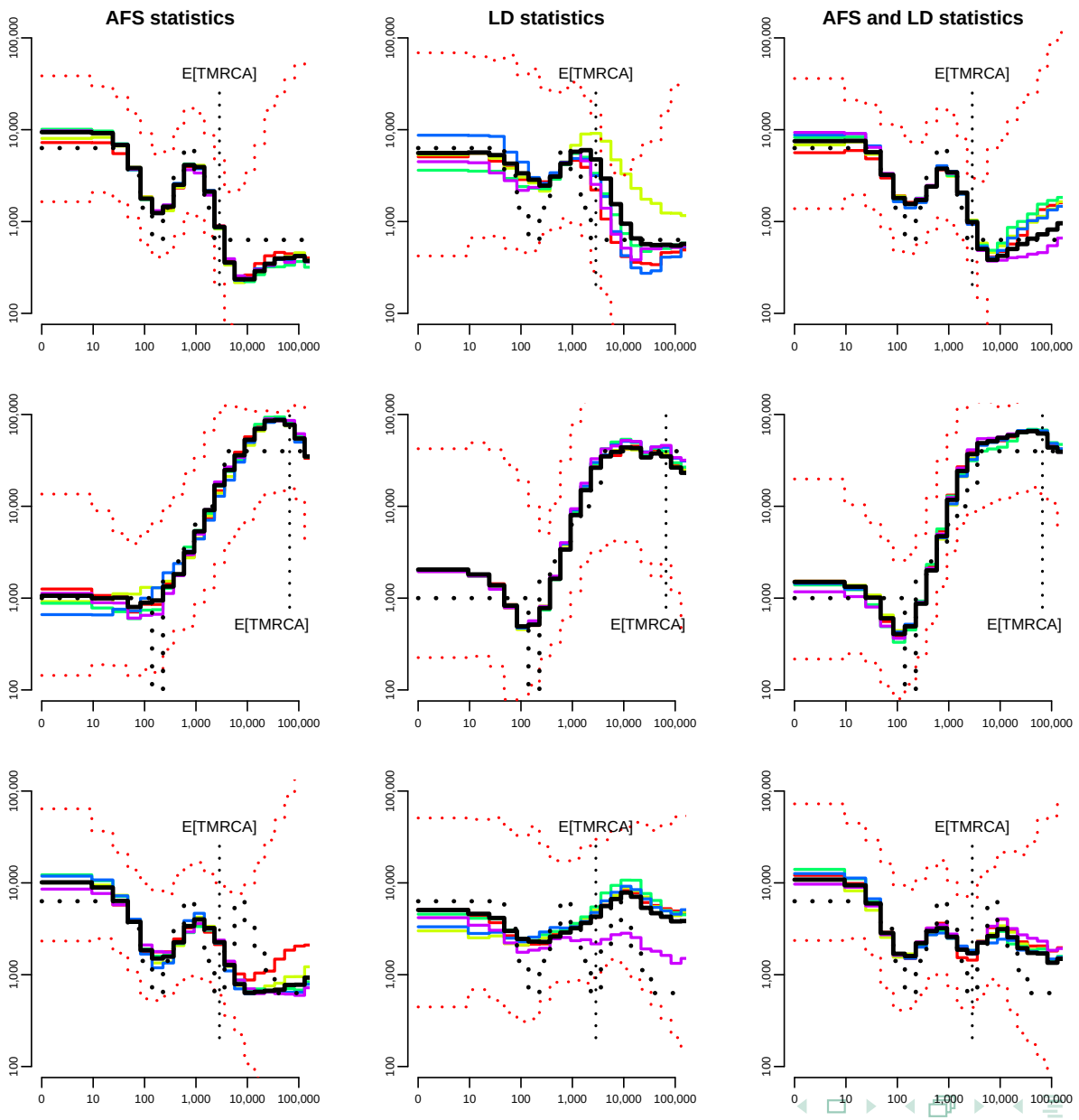
Approche ABC (Beaumont *et al*, 2002)

- ABC = Approximate Bayesian Computation.
- Estime $\mathbb{P}(N() \mid \mathcal{S})$.
- Procède par **simulations intensives**:
  1. Tirer une histoire démographique $N_i()$ selon une loi a priori.
  2. Simuler un échantillon de génomes selon cette histoire.
  3. Caculer les statistiques résumantes $\mathcal{S}_i$.
  4. Conserver les paramètres $N_i()$ si $\mathcal{S}_i$ proche de $\mathcal{S}$.

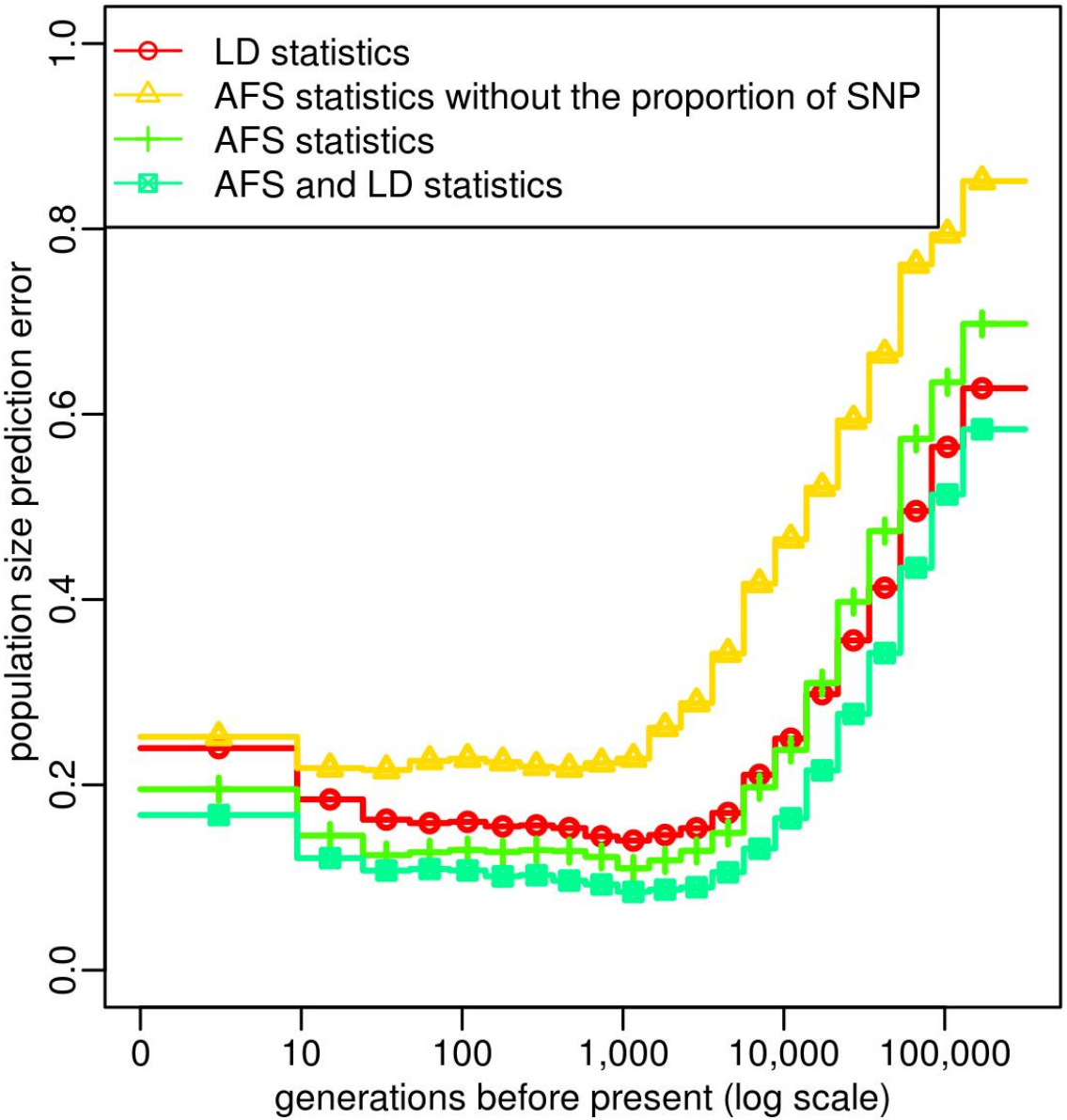
PopSizeABC (Boitard *et al*, PLoS Genetics 2016)

- **Modèle**: Taille de population constante par morceaux (21 morceaux).
- **Statistiques résumantes:** AFS et LD.
- Estimation ABC par **réseaux de neurones** (Blum et François, 2009), package R *abc*.

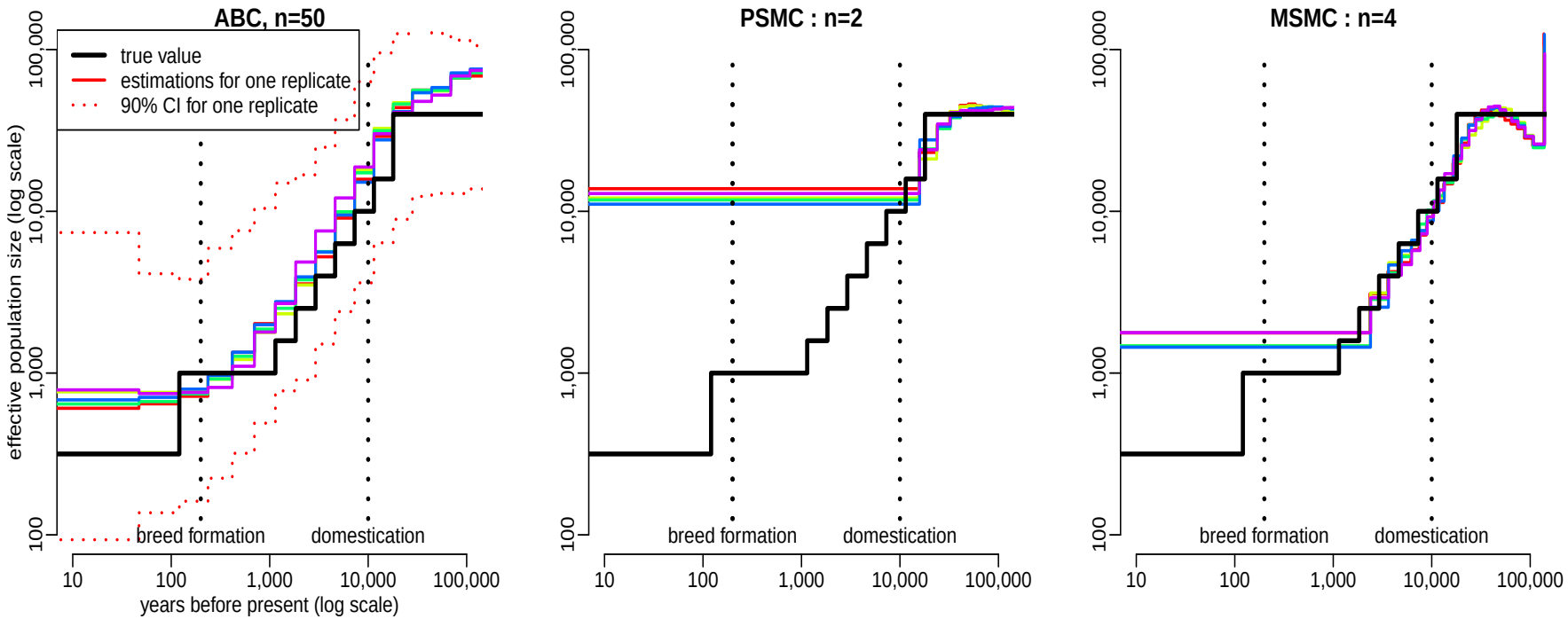
AFS et LD complémentaires



AFS et LD complémentaires

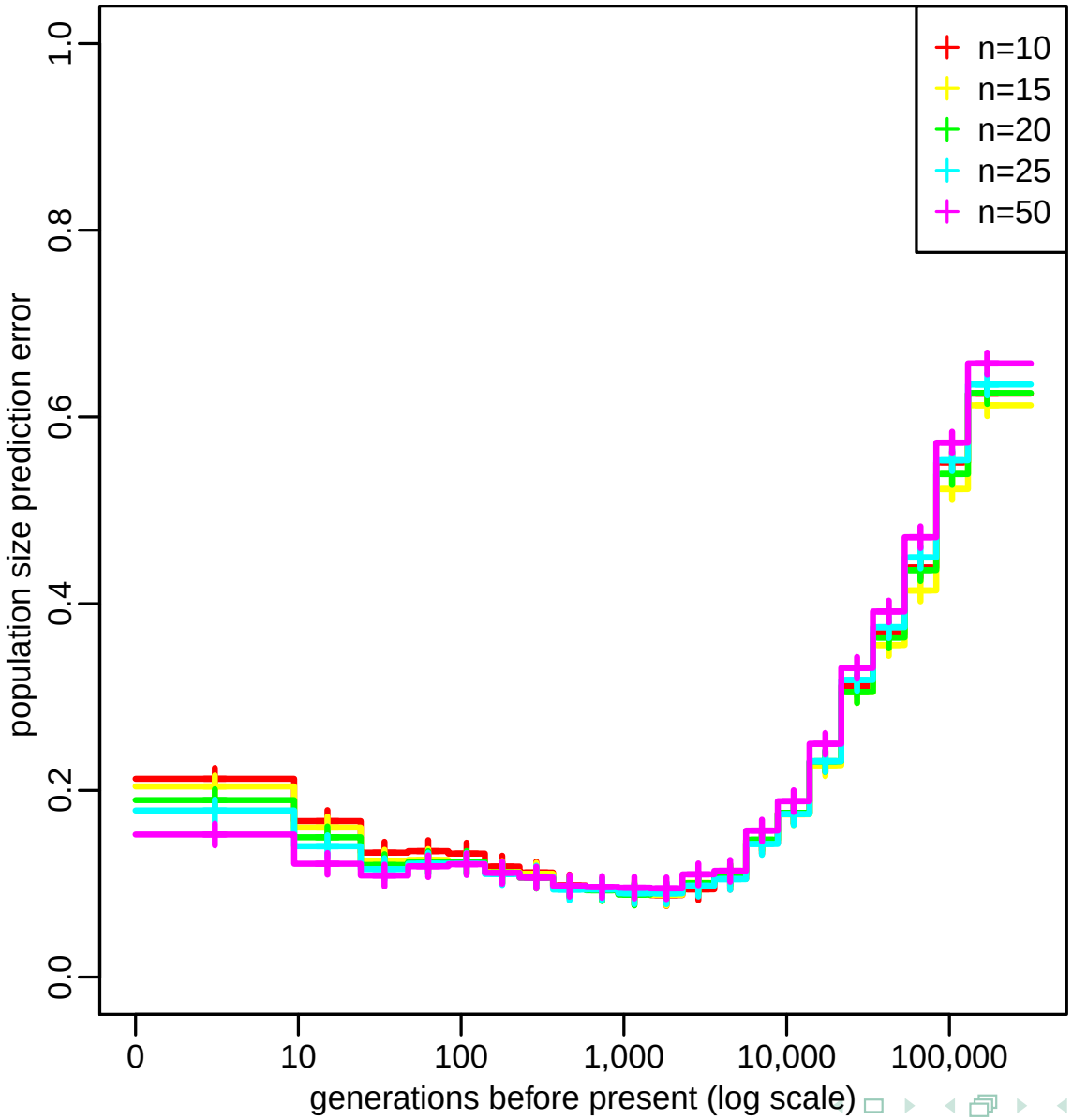


Comparaison avec les approches SMC

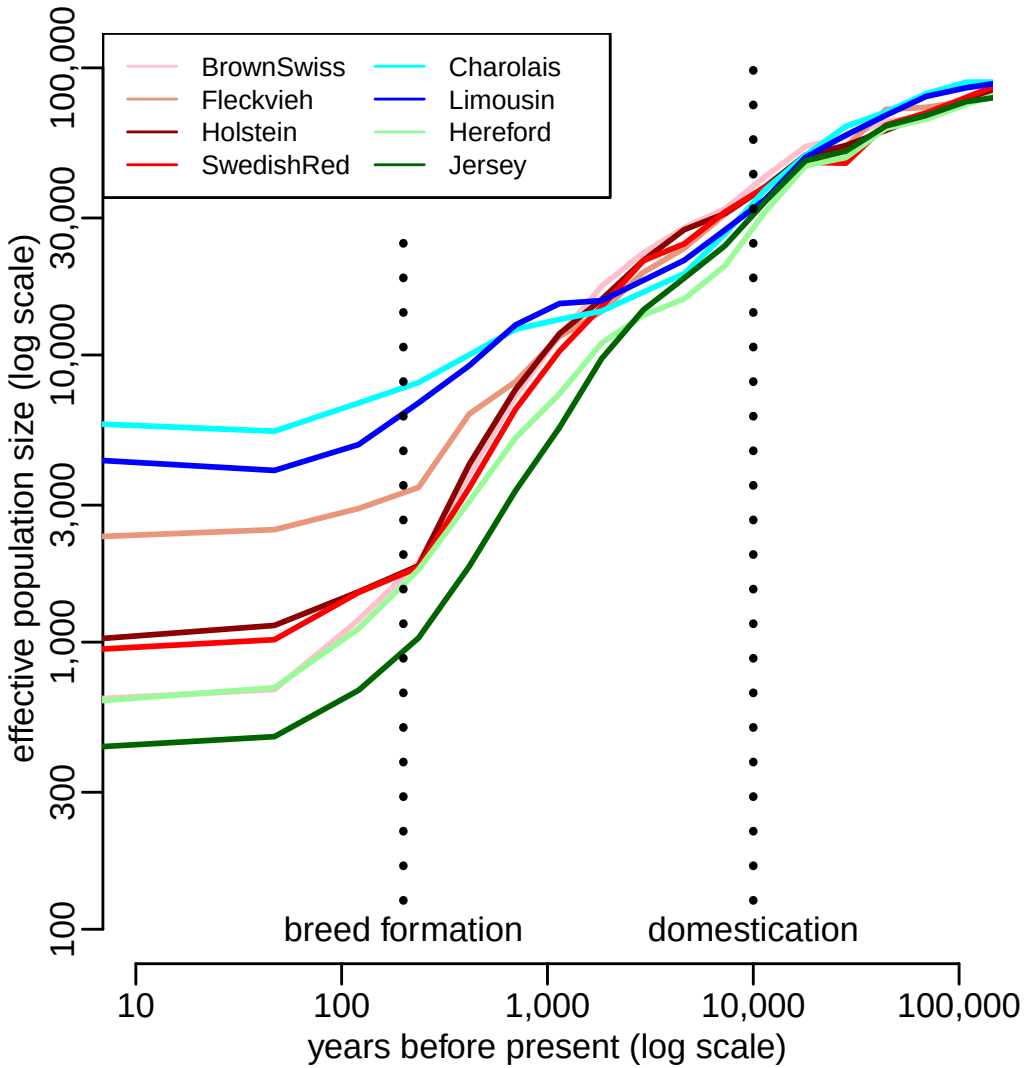


Meilleure estimation de l'**histoire récente** (grands échantillons).

Influence de la taille d'échantillon



Application chez la vache: projet 1000 génomes, run4



- Les trajectoires des races **divergent après la domestication**.
- Déclin continu débutant **avant la domestication**, cf MacLeod *et al* (2013), Gautier *et al* (2016).
- Classement des races selon taille efficace récente cohérent avec d'autres études.

Autres commentaires

- **Simulation de longs segments d'ADN avec recombinaison**, $(100 \times 2Mb)$
  - Approches ABC classiques en génétique des populations plutôt basées sur des segments très courts.
  - Permet d'inclure l'information du LD.
  - Beaucoup plus coûteux à simuler
    $\rightarrow$ limiter l'espace des histoires possibles
    $\rightarrow$ utiliser des simulateurs efficaces : *msprime* (Kelleher *et al*, 2016).
- **LD** calculé à partir des **génotypes** $\rightarrow$ non affecté par les erreurs de phasage, contrairement à MSMC (par exemple).

Plan de l'exposé

Distribution de T_2, modèle panmictique

- Soit $\lambda(t) = \frac{N(t)}{N(0)}$.

$$\mathbb{P}^{PSC}(T_2 \geq t) = e^{-\int_0^t \frac{1}{\lambda(s)} ds}$$

PSC = Population Size Change

- On a

$$\frac{f^{PSC}(t)}{\mathbb{P}^{PSC}(T_2 \geq t)} = \frac{1}{\lambda(t)}$$

$\rightarrow \lambda(t)$ **Inverse Instantaneous Coalescence Rate** (IICR).

Distribution de T_2, modèle d'évolution arbitraire

- Soit $F(t) = \mathbb{P}(T_2 \leq t)$, et $F'(t) = f(t)$.
- On introduit l'**IICR**

$$\lambda(t) = \frac{1 - F(t)}{f(t)}$$

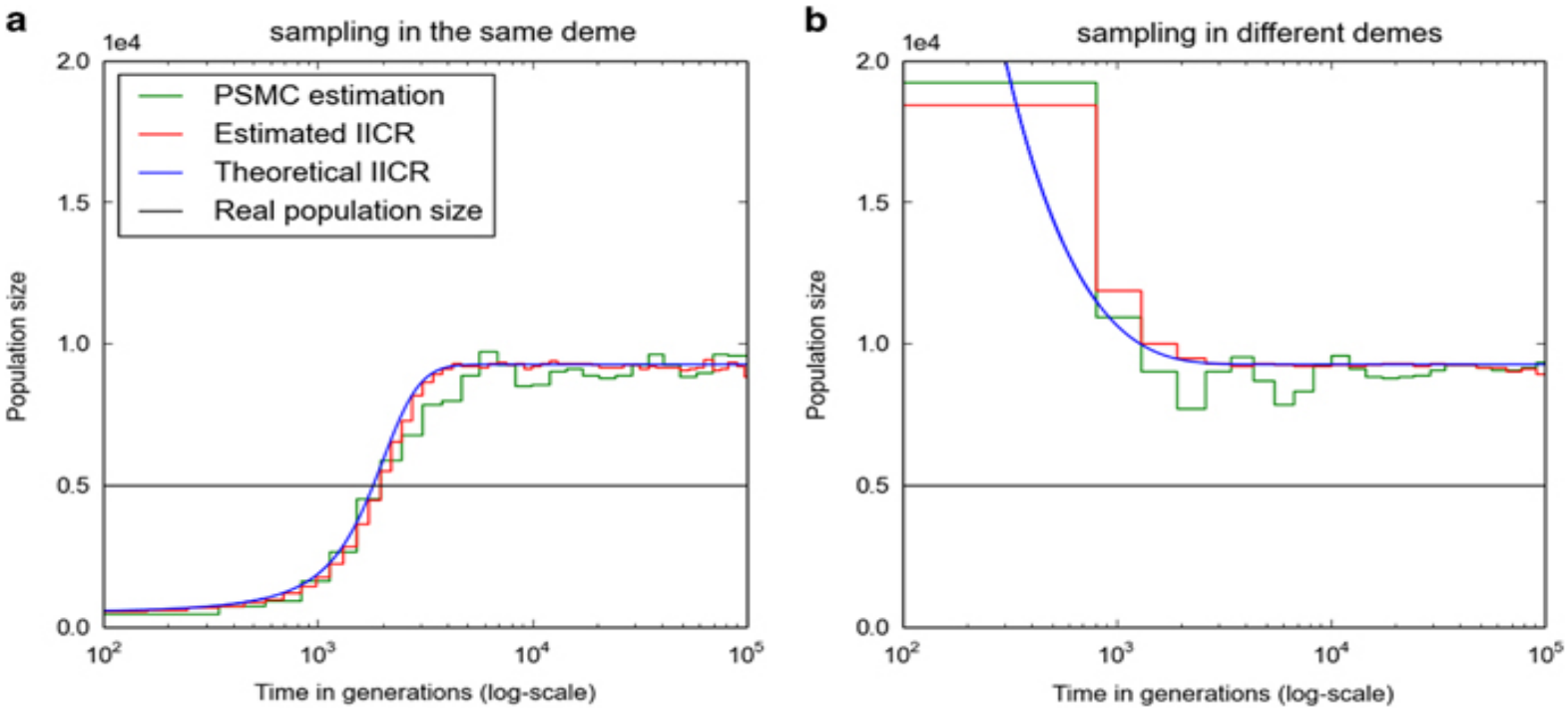
- On montre que (Mazet *et al*, Heredity 2016)

$$F(t) = 1 - e^{-\int_0^t \frac{1}{\lambda(s)} ds} = F^{PSC,\lambda}(t)$$

$\rightarrow$ **Il existe toujours un modèle de changements de taille reproduisant parfaitement la distribution de T_2 observée à partir de deux séquences.**

Exemple : modèle en îles symétrique

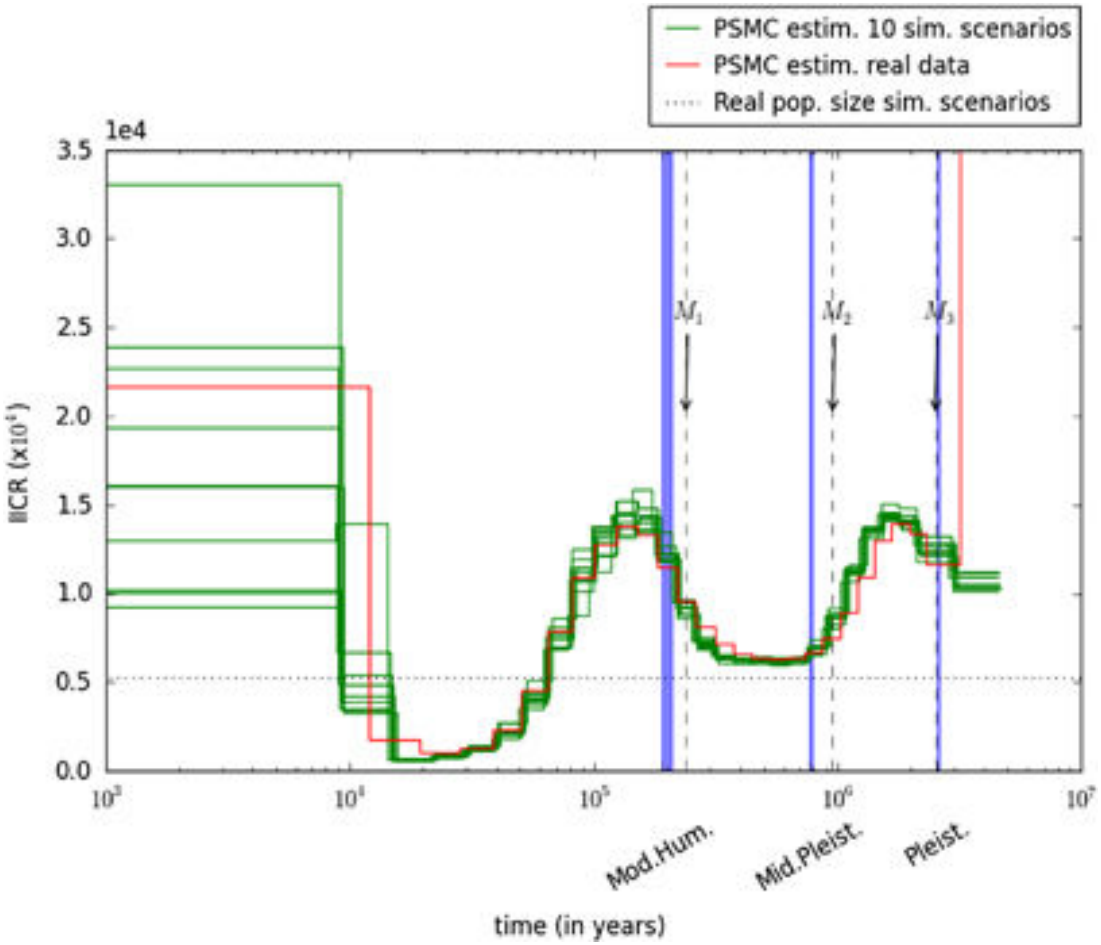
- n **populations** connectées par des **migrations**.



- PSMC estime un **changement de taille inexistant**, exactement identique à la prédiction de l'IICR.

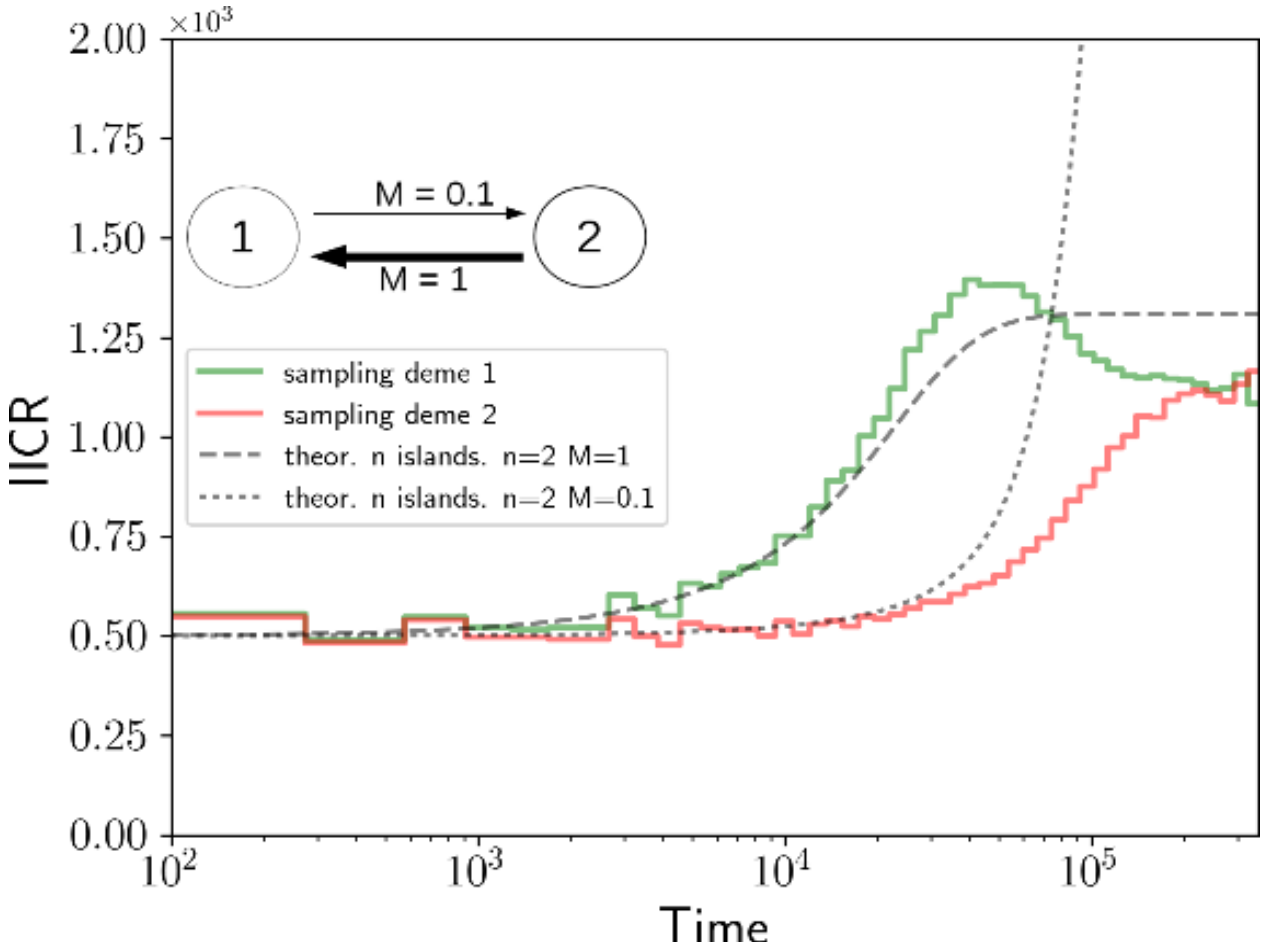
Implications

- **Remet en question** un grand nombre de **scénarios évolutifs** proposés récemment à partir de PSMC.



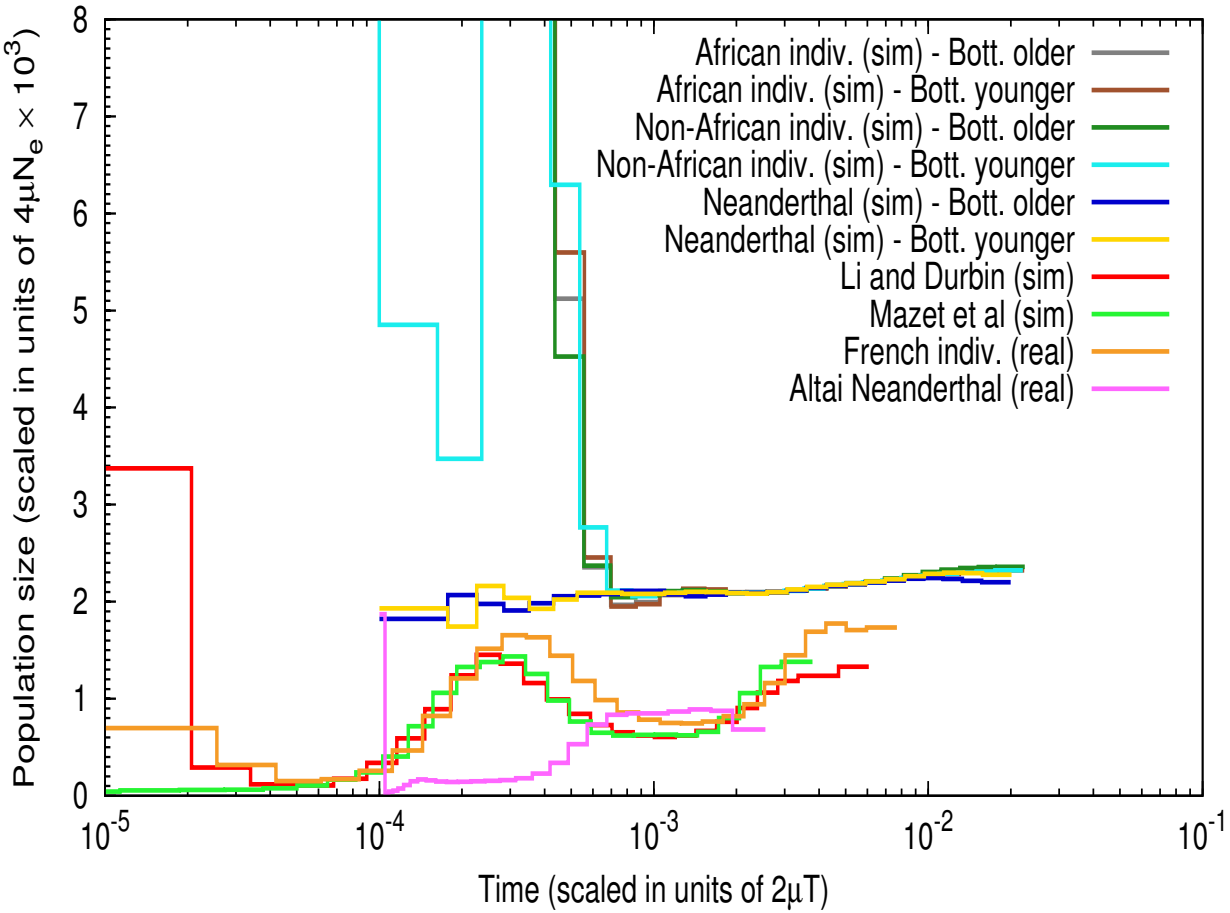
Applications possibles (Chikhi *et al*, Heredity 2018)

- **Caractériser l'IICR** de plusieurs modèles classiques, pour **aider l'interpratation** des résultats PSMC.



Applications possibles (Chikhi *et al*, Heredity 2018)

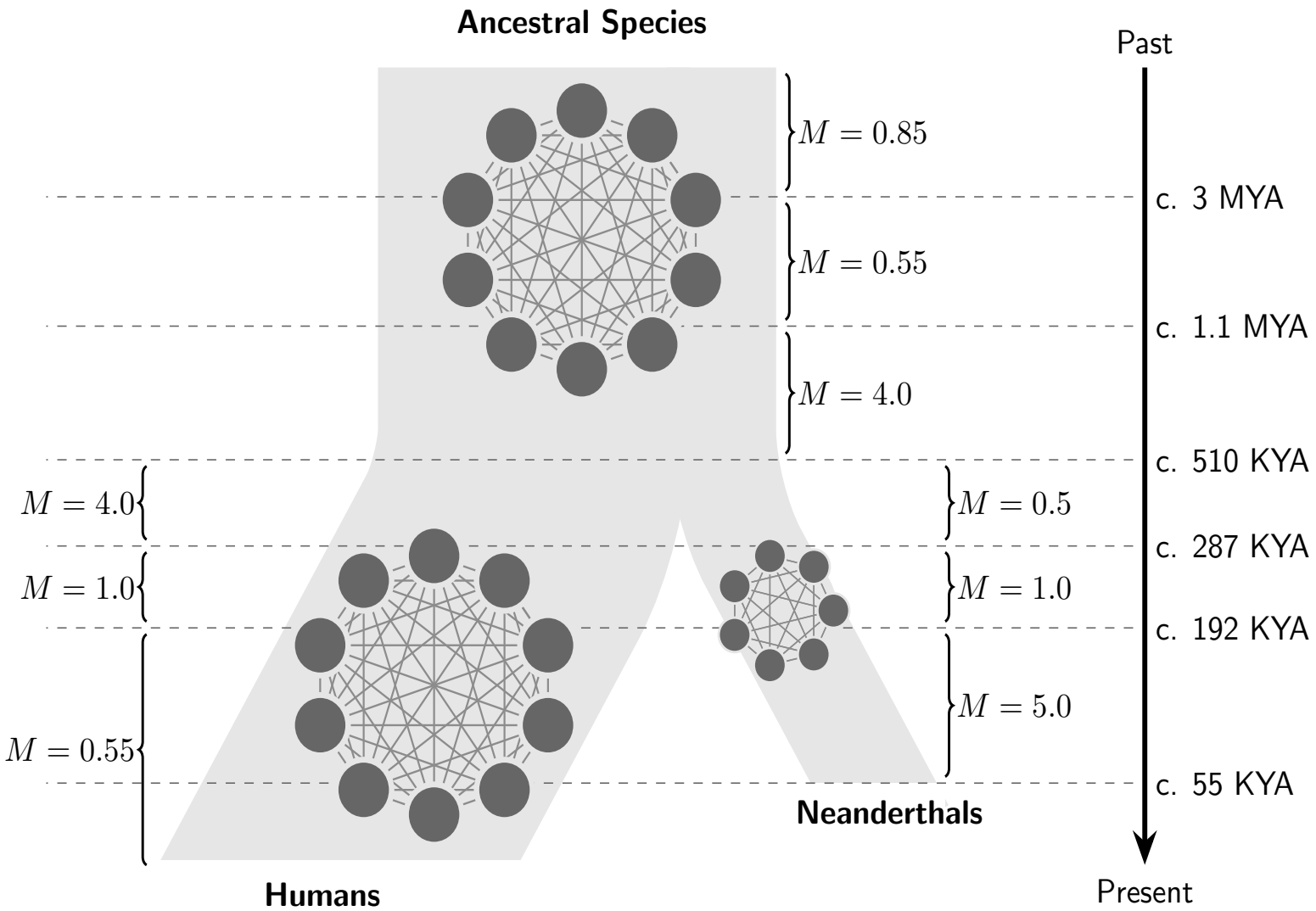
- **Utiliser l'IICR** (ou le PSMC) comme une **statistique résumante** pour accepter / rejeter des scénarios évolutifs.



Modèles d'admixture proposés par Yang *et al* (2012).

Applications possibles (Rodriguez *et al*, Heredity 2018)

- **Proposer des scenarios** compatibles avec l'IICR.



Plan de l'exposé

Conclusions

- **Diversité génétique** actuelle (haut débit) **très informative** sur histoire évolutive.
- Utiliser de **grands échantillons** permet de reconstruire **l'histire récente**.
- **AFS moyen souvent insuffisant**: utiliser le LD, la distribution des temps de coalescence, les généalogies locales (Bunnefeld *et al*, 2015; Sainudin and Veber, bioRxiv) ...
- **Tenir compte de la structure** dans l'interprétation ou l'estimation.

Perspectives

- **Influence** de la **structure** sur d'**autres méthodes** (AFS).
- Estimation de **modèles complexes** incluant structure et changements de taille par les approches **ABC** ou **IICR**.
- IICR au delà du T_2 (Grusea *et al*, J Math Biol 2018).

Echantillonage temporel

- Progrès spectaculaires dans le séquençage d'**ADN ancien**.
- Echantillons conservées en **cryobanques**.
- **Développement méthodologique** : thèse de Cyriel Paris (2017-2019).
- **Données** : projet ANR path2bos, paléogénomique bovine (2018-2021).

Remerciements

PopSizeABC:

- **Flora Jay**, Laboratoire de Recherches en Informatique, Orsay
- **Willy Rodriguez**, Institut de Mathématiques de Toulouse (IMT)
- **Stefano Mona & Frédéric Austerlitz**, Muséum National d'Histoire Naturelle, Paris

IICR:

- **Olivier Mazet, Willy Rodriguez & Simona Grusea**, IMT
- **Lounès Chikhi**, Evolution et Diversité Biologique Toulouse & Instituto Gulbenkian de Ciência, Portugal

Infrastructure de calcul:

- Plateforme bio-informatique Genotoul Toulouse Midi-Pyrénées.