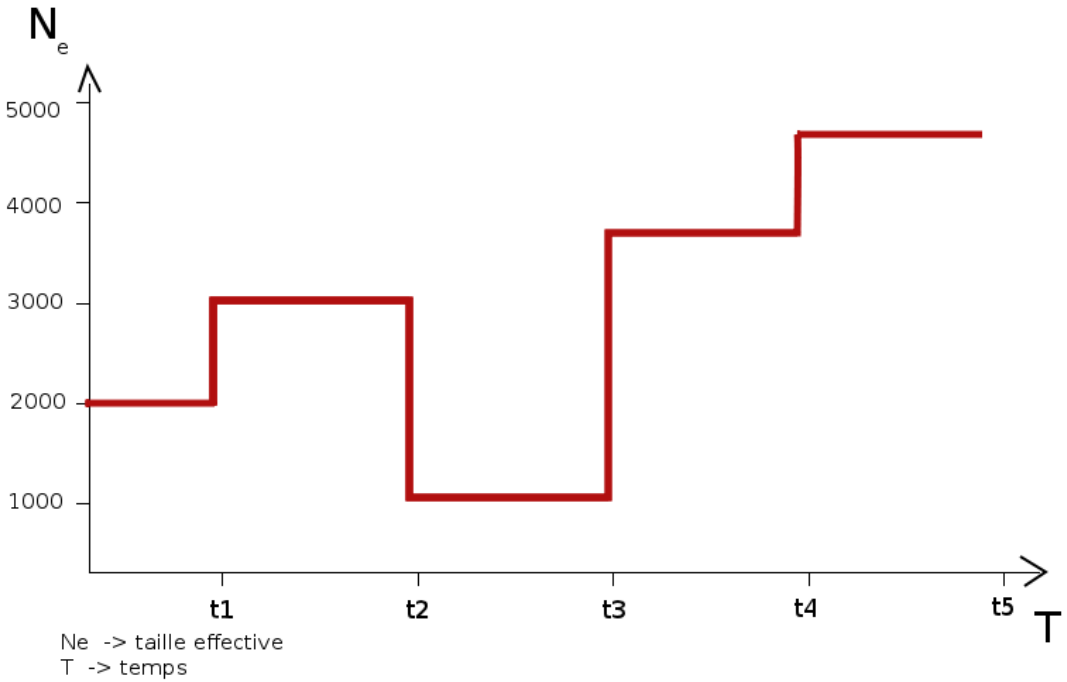
Estimation de tailles efficaces variables dans le temps à partir de données génomiques actuelles

Flora Jay[1], Simon Boitard[2]

1 : CNRS, Laboratoire de Recherche en Informatique (LRI), Orsay
2 : INRA, Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Toulouse

Réseau des Ressources Génétiques Animales
12 et 13 mai 2016

Motivations



- Vision dynamique, identifier les populations en déclin.
- Interprétation des variations de taille: évènements géologiques, climatiques, anthropomorphiques . . .
- Meilleure détection des locus sous sélection.

Plan de l'exposé

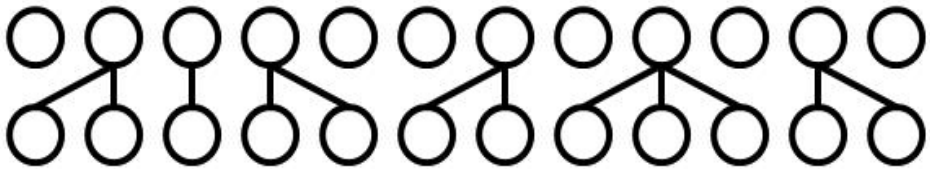
1. Estimation à partir de locus indépendants
   - Généalogie à un locus
   - Méthodes d'estimation
2. Estimation à partir de données génomiques haut débit
   - Données complètes: les modèles SMC
   - Données résumées
   - L'approche ABC (Approximate Bayesian Computation)
3. Conclusions

Plan de l'exposé

1. Estimation à partir de locus indépendants
   - Généalogie à un locus
   - Méthodes d'estimation
2. Estimation à partir de données génomiques haut débit
   - Données complètes: les modèles SMC
   - Données résumées
   - L'approche ABC (Approximate Bayesian Computation)
3. Conclusions

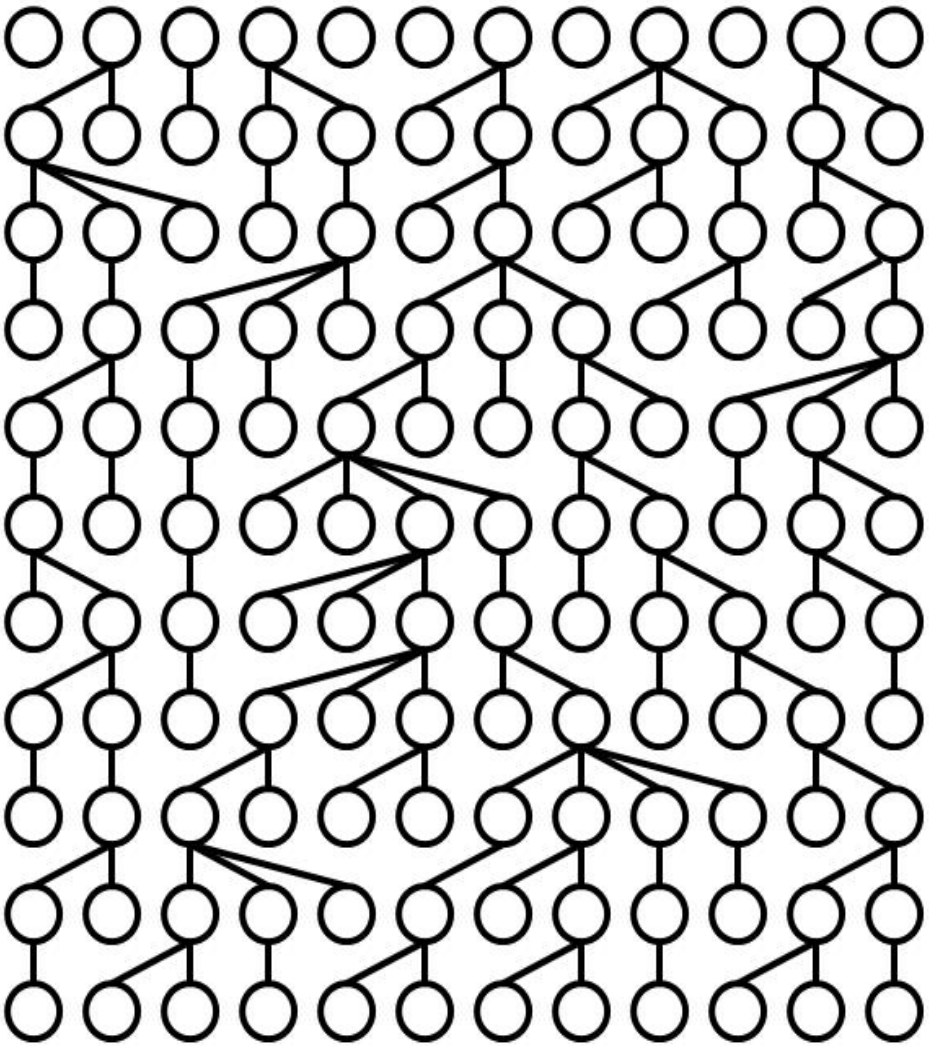
Modèle de Wright-Fisher à un locus

Modèle de Wright-Fisher à un locus



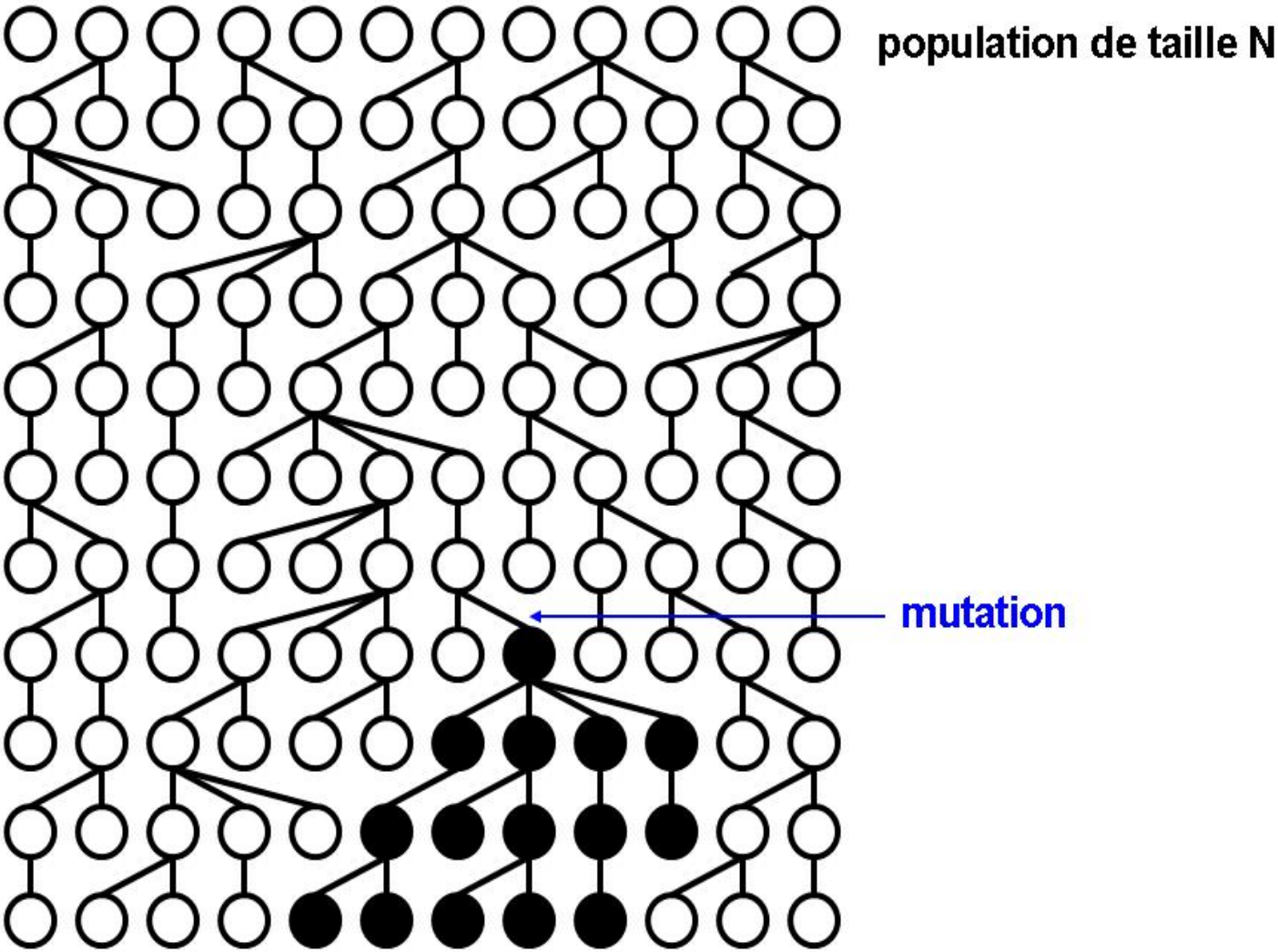
population de taille N

Modèle de Wright-Fisher à un locus

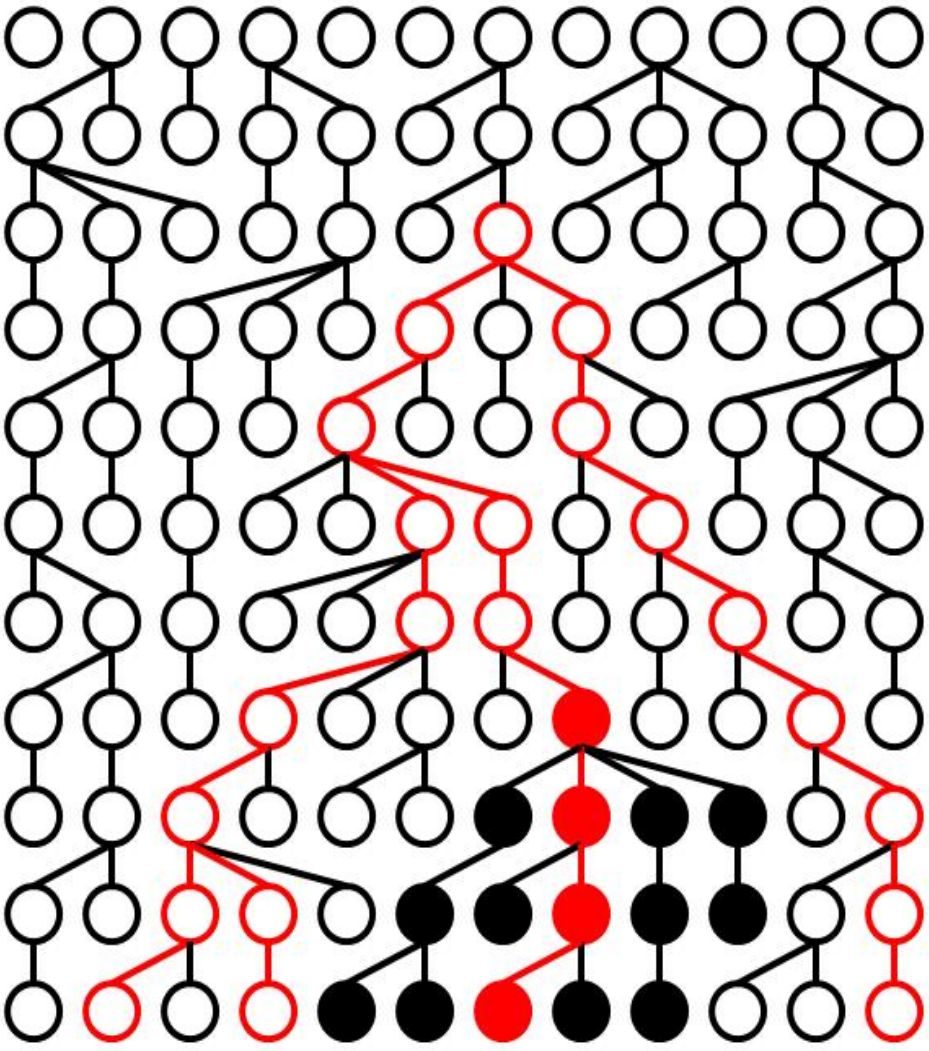


population de taille N

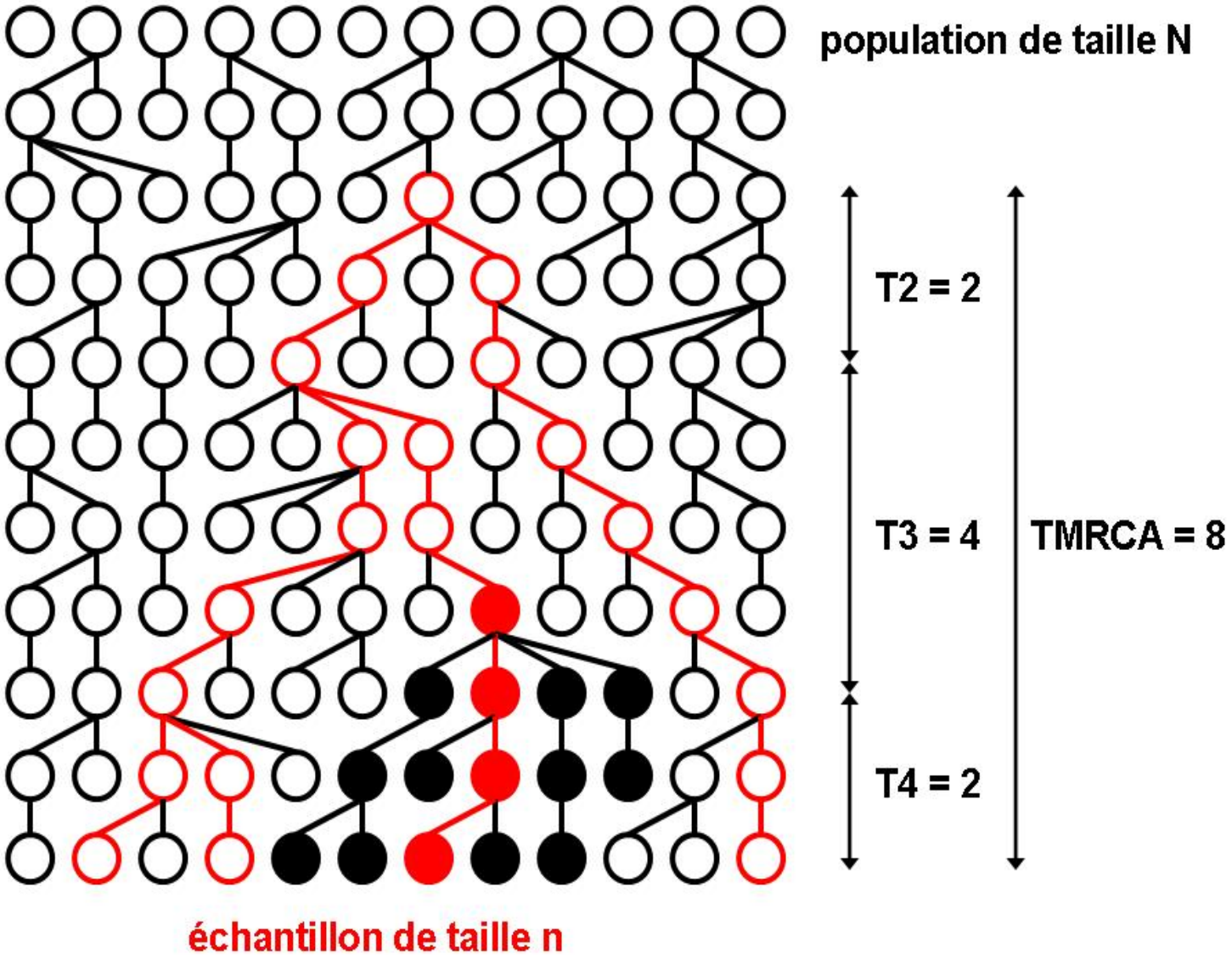
Modèle de Wright-Fisher à un locus



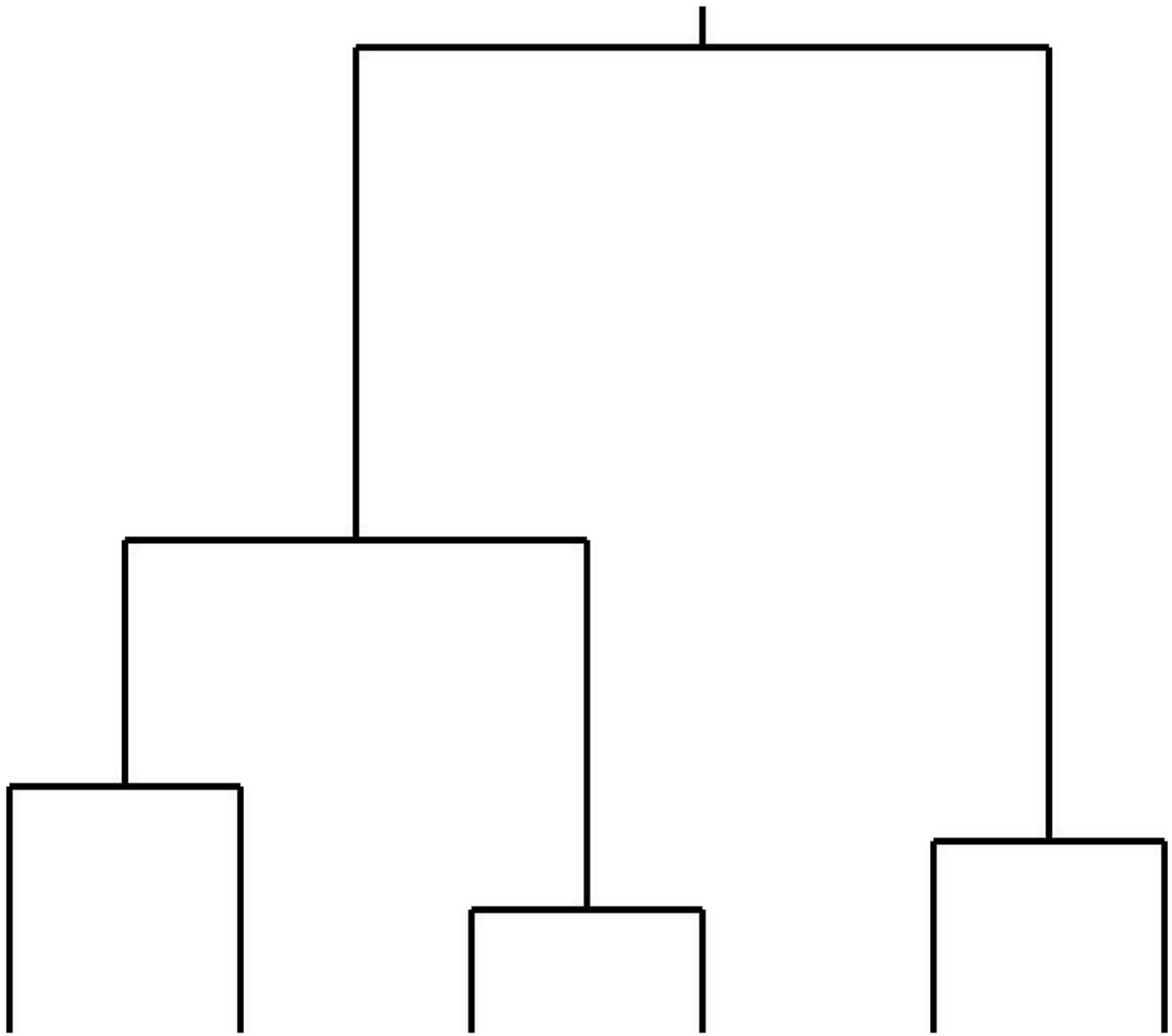
Généalogie



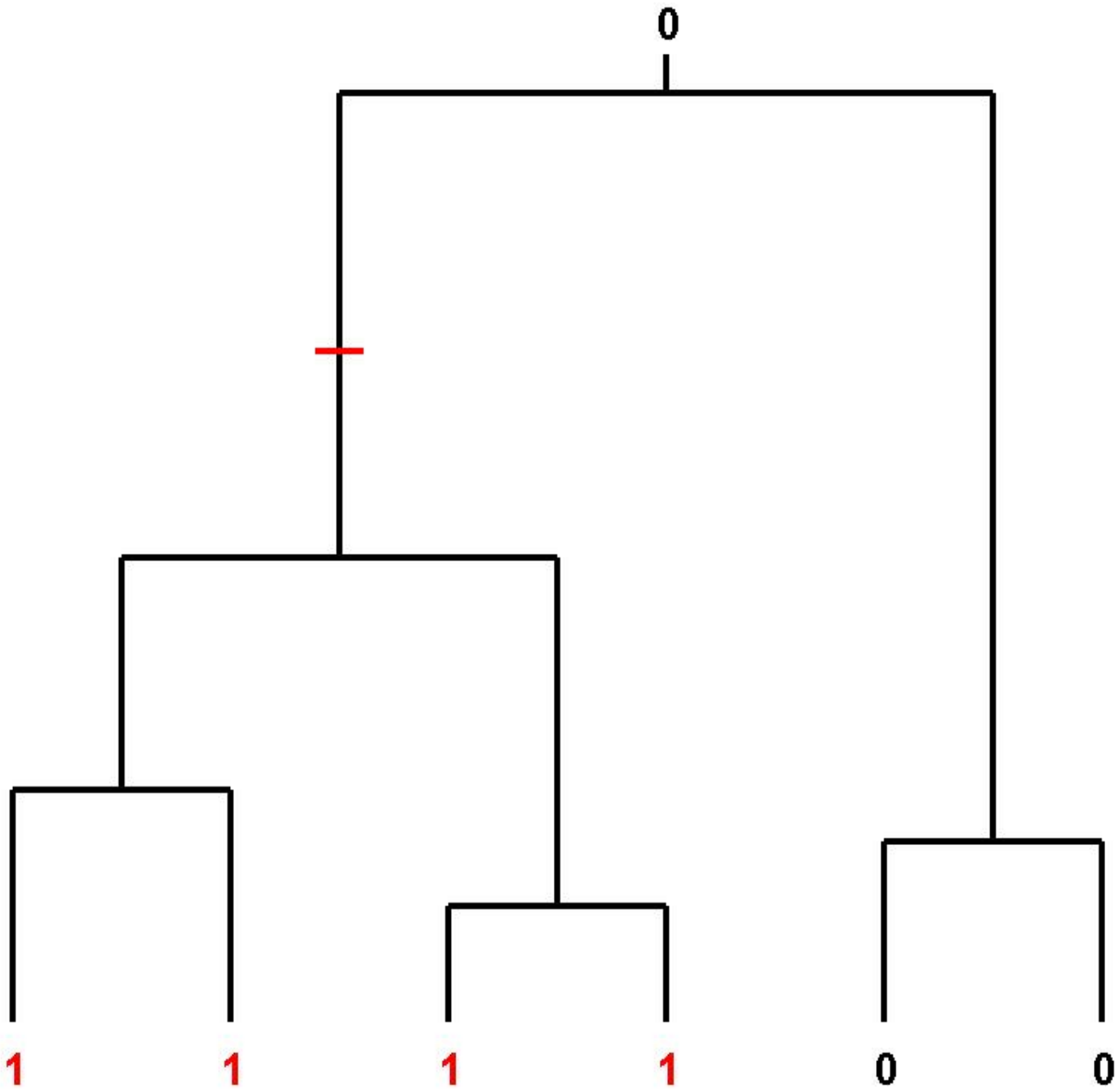
Généalogie


population de taille N
T2 = 2
T3 = 4
TMRCA = 8
T4 = 2
échantillon de taille n

Mutations



Mutations

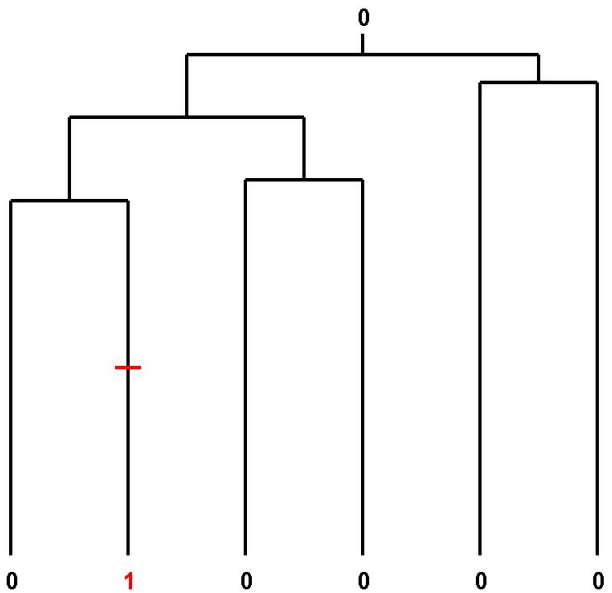


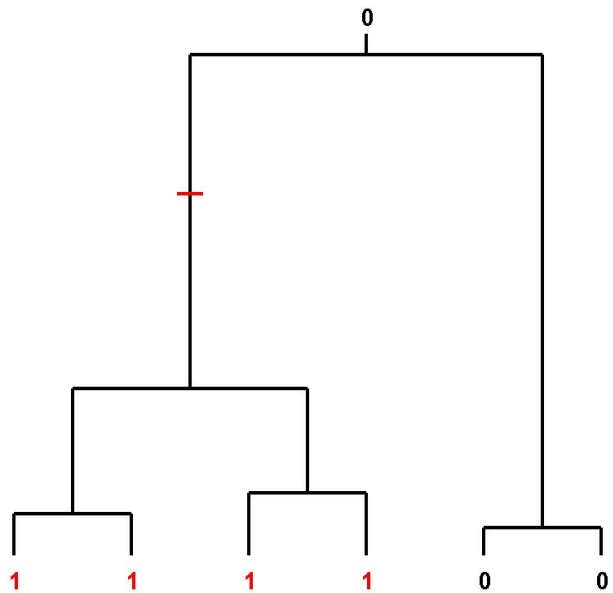
Lien avec la taille

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.





Plan de l'exposé

1. Estimation à partir de locus indépendants
   - Généalogie à un locus
   - Méthodes d'estimation
2. Estimation à partir de données génomiques haut débit
   - Données complètes: les modèles SMC
   - Données résumées
   - L'approche ABC (Approximate Bayesian Computation)
3. Conclusions

Principe général

- Pour un locus i donné:

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

  $\mathcal{D}_i$ allèles observés, $N()$ démographie, G généalogie.
- En pratique, simulation "intelligente" de généalogies car énumération impossible.
- Pour p locus indépendants:

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

Exemples

Méthode	**N()**	**données**	**approche**
Bottleneck (Cornuet et Luikart, 1996)	1 changement	microsatellites	test heuristique
Msvar (Beaumont, 1999)	1 changement	microsatellites	MCMC
Beast (Drummond et Rambaut, 2007)	continu	séquences	MCMC
VarEff (Nikolic et Chevalet, 2014)	continu	microsatellites	formule approchée

Plan de l'exposé

1. Estimation à partir de locus indépendants
   - Généalogie à un locus
   - Méthodes d'estimation

2. Estimation à partir de données génomiques haut débit
   - Données complètes: les modèles SMC
   - Données résumées
   - L'approche ABC (Approximate Bayesian Computation)

3. Conclusions

Problématique

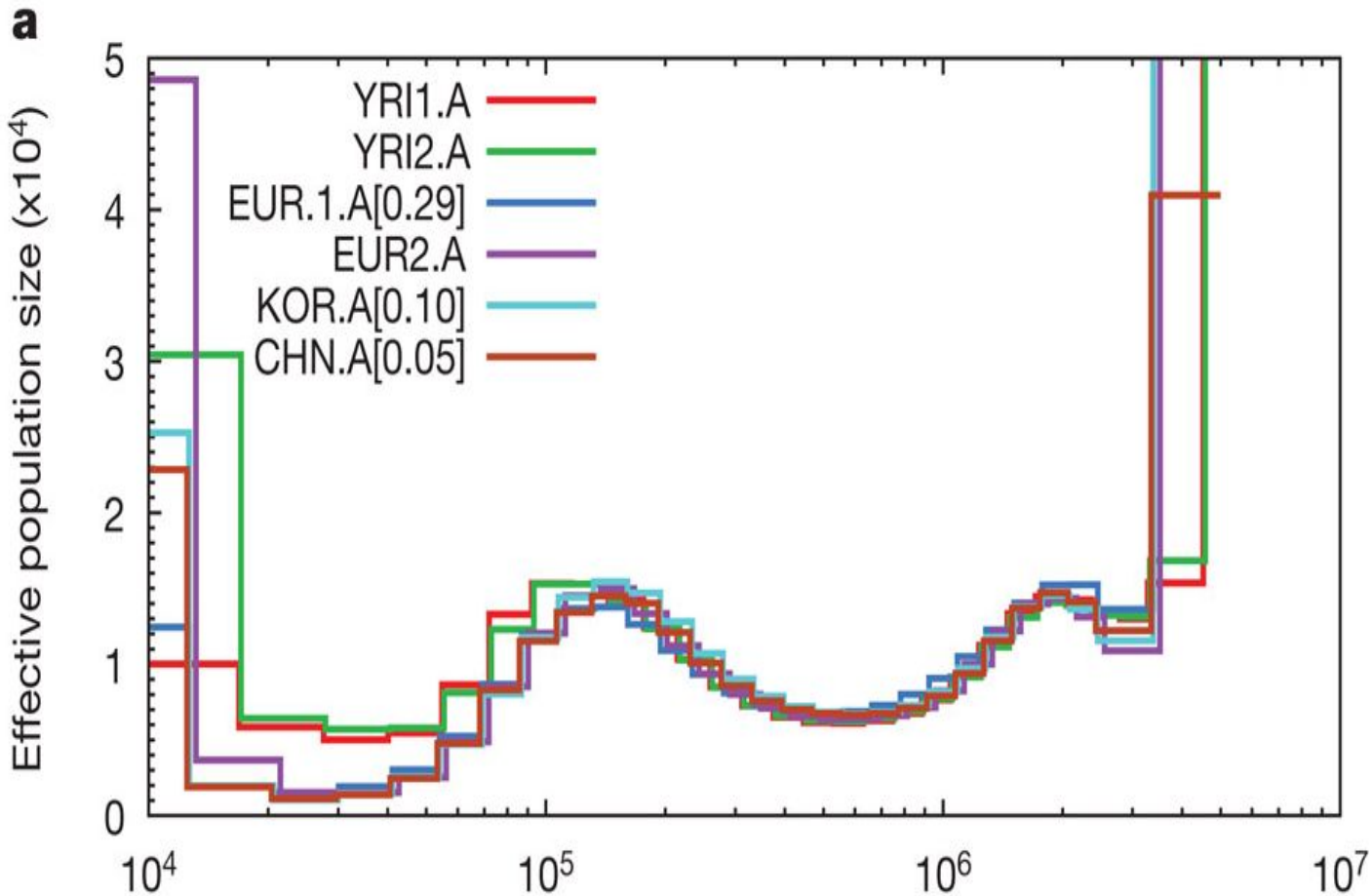
- **Intérêt:**
  - Données disponibles (puces de génotypage, séquençage, RADseq).
  - Plus de locus $\rightarrow$ estimation plus précise.
- **Obstacles:**
  - Généalogies G_i et G_j pour deux locus proches différentes mais corrélées.
  - Corrélation difficile à modéliser, prise en compte de la recombinaison.

Plan de l'exposé

1. Estimation à partir de locus indépendants
   - Généalogie à un locus
   - Méthodes d'estimation
2. Estimation à partir de données génomiques haut débit
   - Données complètes: les modèles SMC
   - Données résumées
   - L'approche ABC (Approximate Bayesian Computation)
3. Conclusions

PSMC (Li et Durbin, 2011)

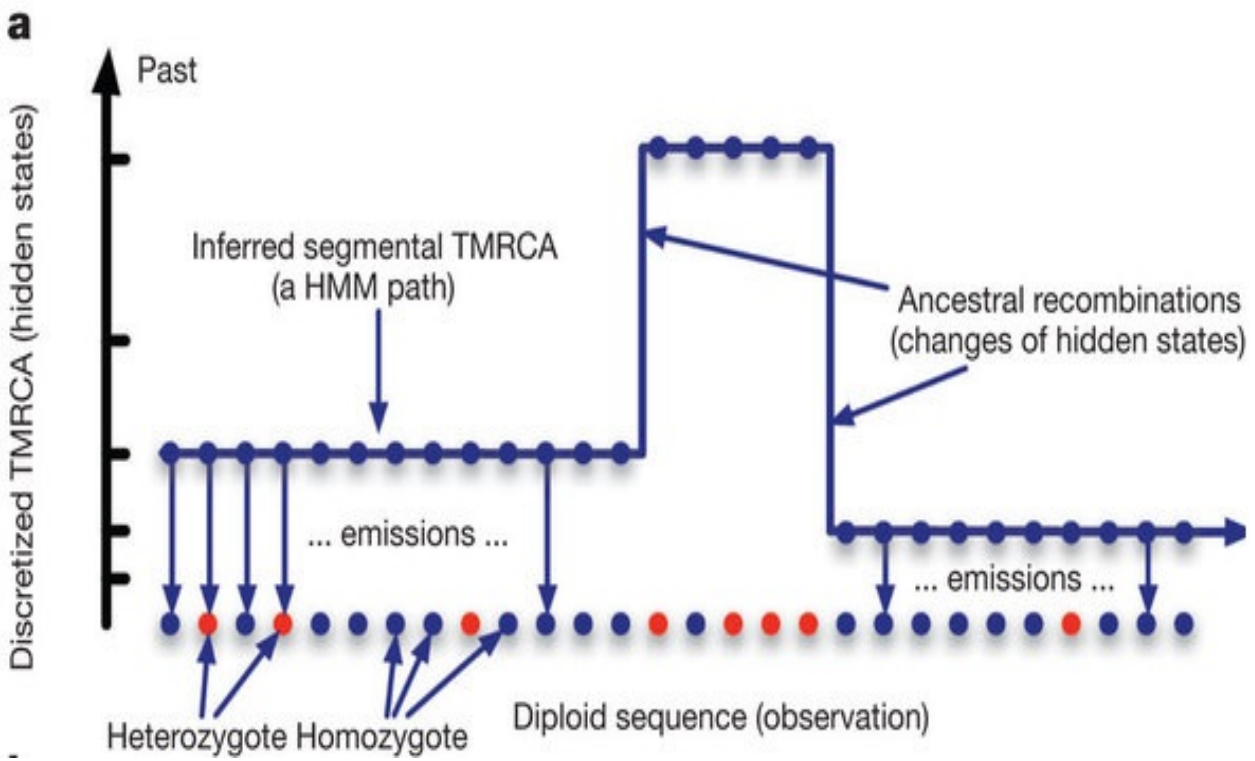
Estimation basée sur le génome entier d'un individu diploïde.



Principe

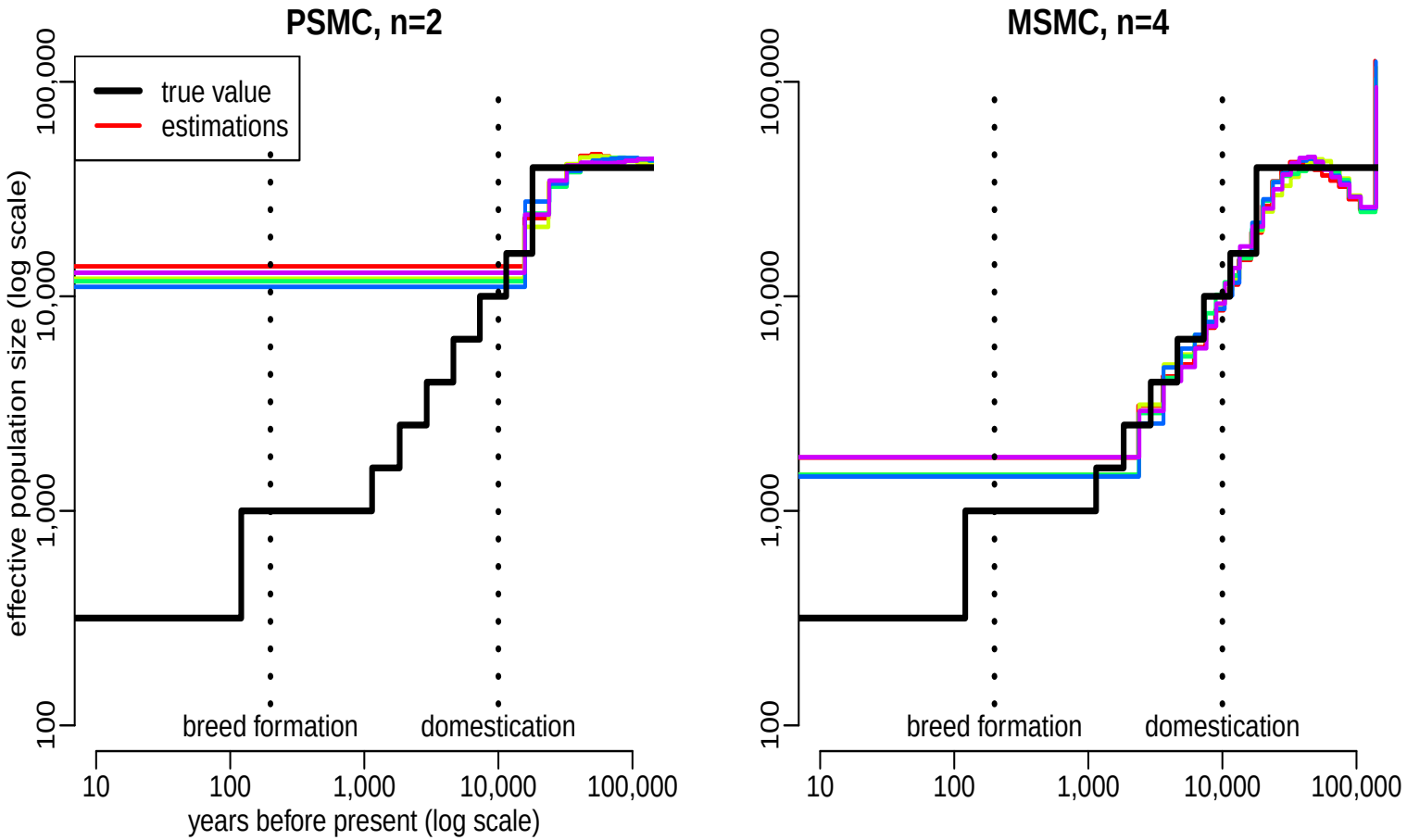
PSMC = Pairwise Sequentially Markovian Coalescent

- Généalogie simplifiée: $G = T_2$.
- Approximation de Markov: $G_{i+1} = f(G_i)$
  $\rightarrow$ Chaîne de Markov cachée.



Autres méthodes SMC

- dical (Sheehan *et al*, 2013), MSMC (Schiffels et Durbin, 2014).
- Petit nombre d'individus ($\approx$ 5 diploïdes max).
  $\rightarrow$ Faible précison pour la démographie récente.



Plan de l'exposé

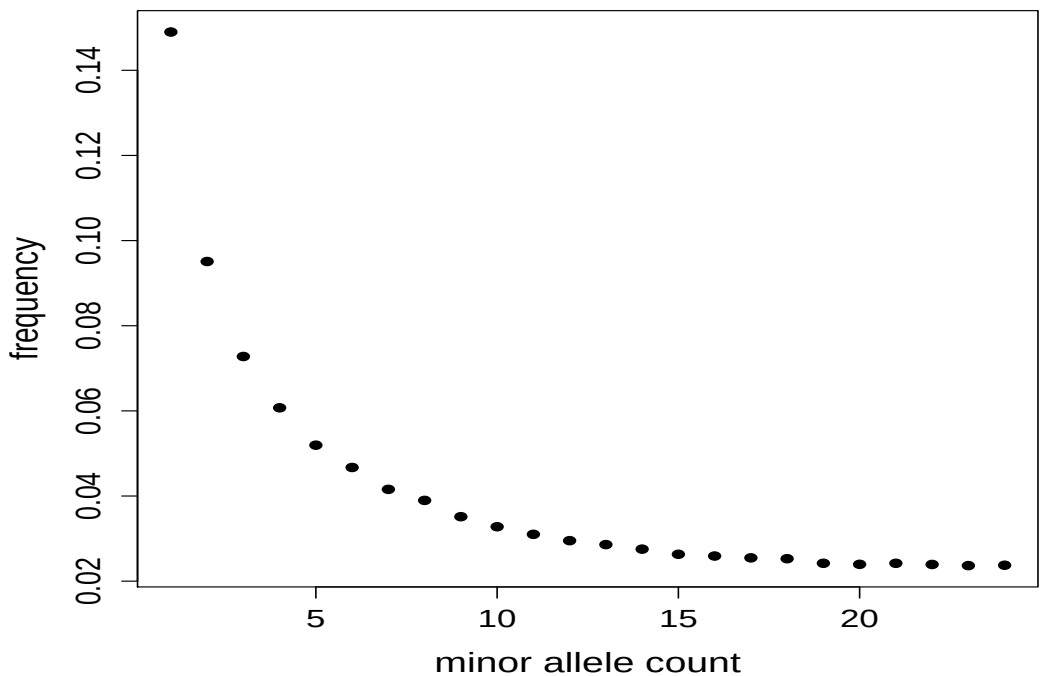
1. Estimation à partir de locus indépendants
   - Généalogie à un locus
   - Méthodes d'estimation
2. Estimation à partir de données génomiques haut débit
   - Données complètes: les modèles SMC
   - Données résumées
   - L'approche ABC (Approximate Bayesian Computation)
3. Conclusions

Principe

- Remplacer les données complètes $\mathcal{D}$ par un ensemble de statistiques $\mathcal{S}$ résumant ces données.
- **Avantage:** $\mathbb{P}(\mathcal{S} \mid N())$ plus facile à calculer que $\mathbb{P}(\mathcal{D} \mid N())$
- **Inconvénient:** Estimation un peu moins précise.

Spectre des fréquences alléliques (AFS)

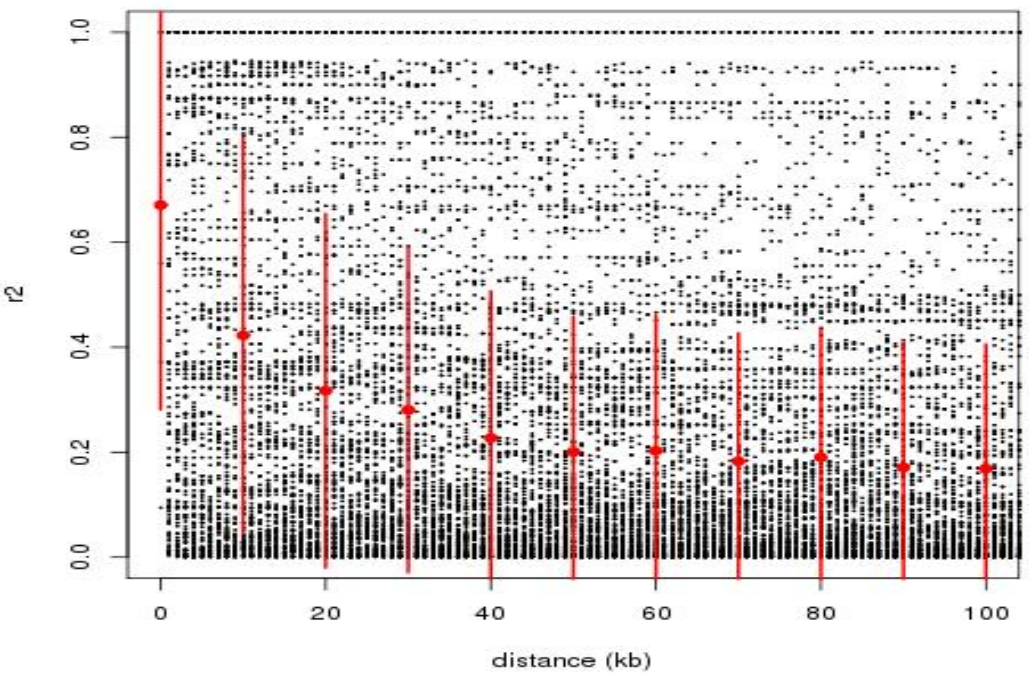
- Proportion de SNPs sur le génome.
- Parmi ces SNPs, proportion de ceux ayant i copies de l'allèle le moins fréquent, pour i de 1 à $n/2$ (n taille de l'échantillon).
- Formule analytique approchée pour $\mathbb{P}(\mathcal{S} \mid N())$ (Bhaskar *et al*, 2015; Liu *et al*, 2015).



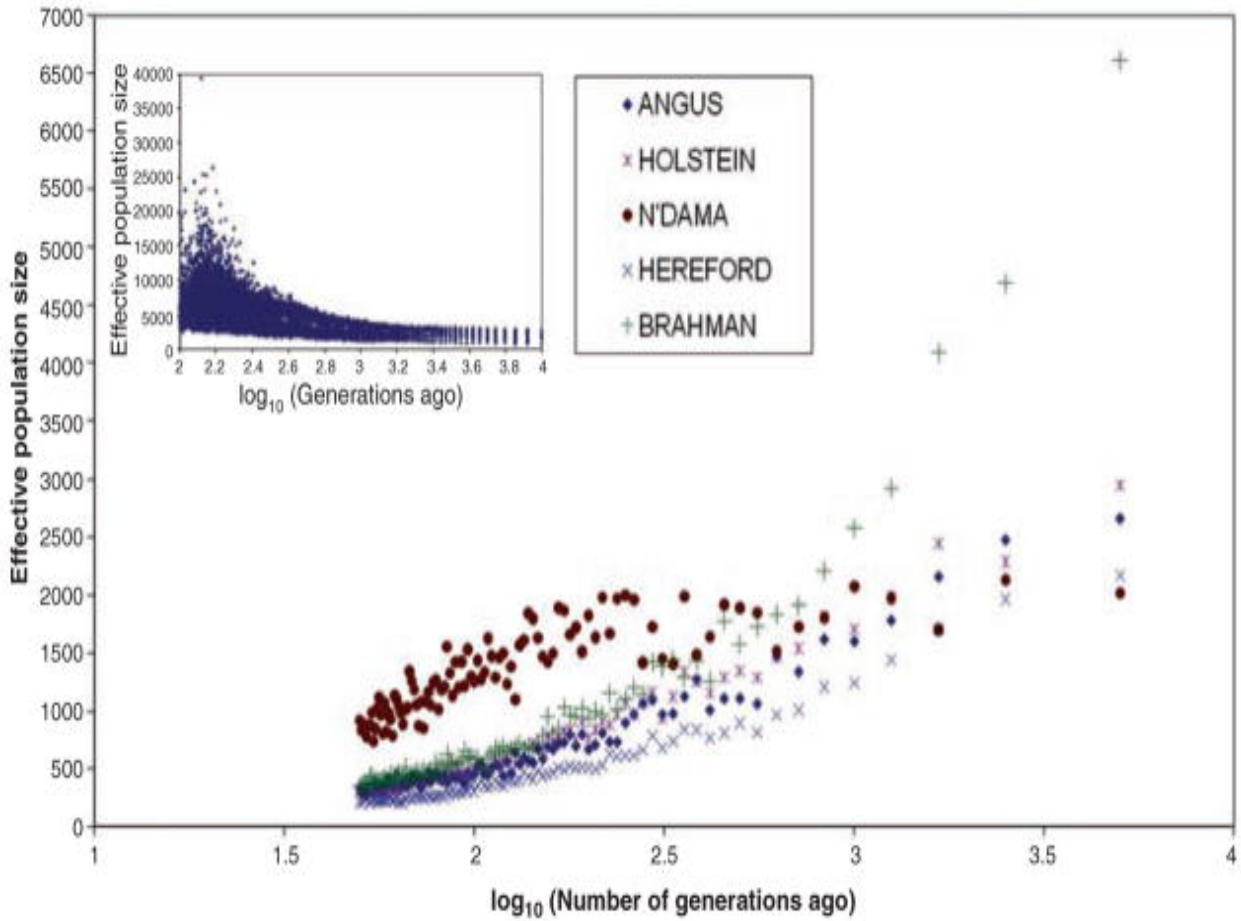
Déséquilibre de liaison (LD)

- r^2 = corrélation des génotypes observés entre deux SNPs.
- r^2 diminue quand le taux de recombinaison c augmente.
- Approximation de Hayes *et al* (2003):

$$\hat{r^2}(c) \approx \frac{1}{1 + 2N(1/2c)c}$$



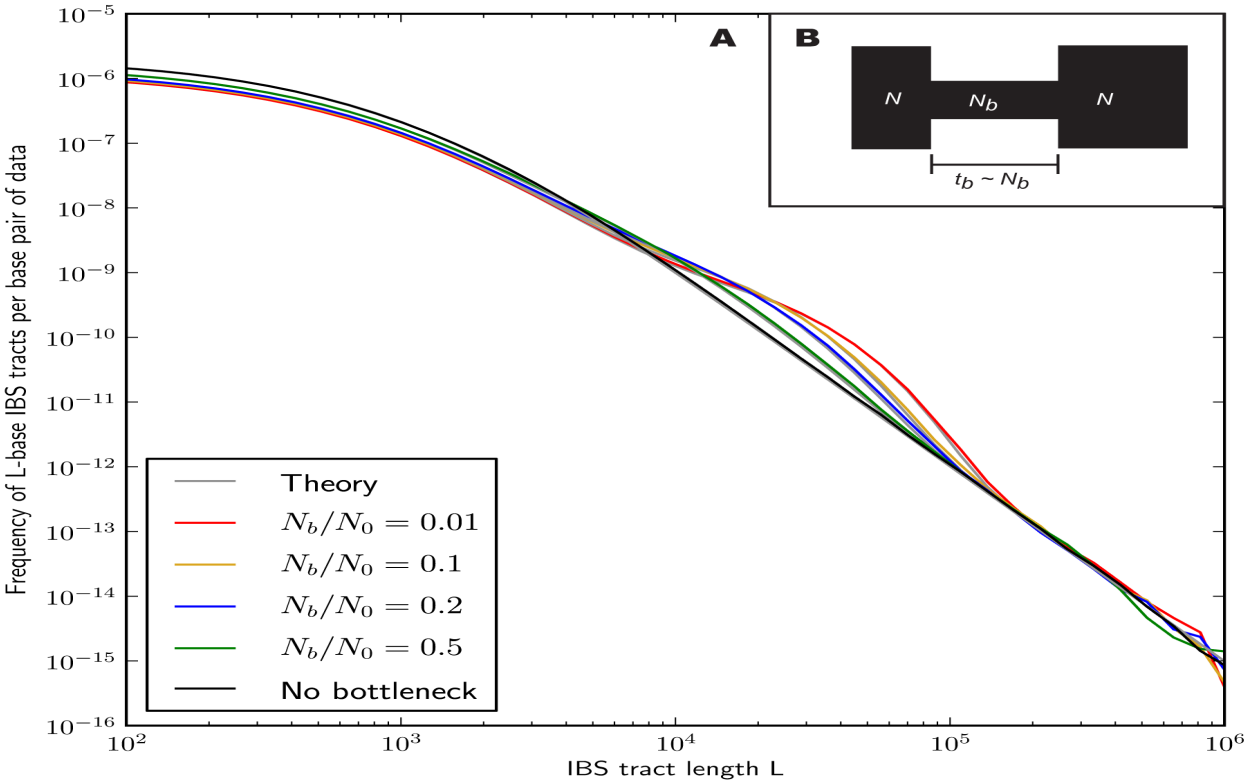
Exemple



The Bovine HapMap consortium (2009)

Segments IBS (Identical By State)

- Zone du génome sans polymorphisme.
- McLeod et al (2013), Harris et Nielsen (2013): formules approchées pour $\mathbb{P}(\mathcal{S} \mid N())$, $\mathcal{S}$ distribution de la longueur des segments IBS chez un individu.



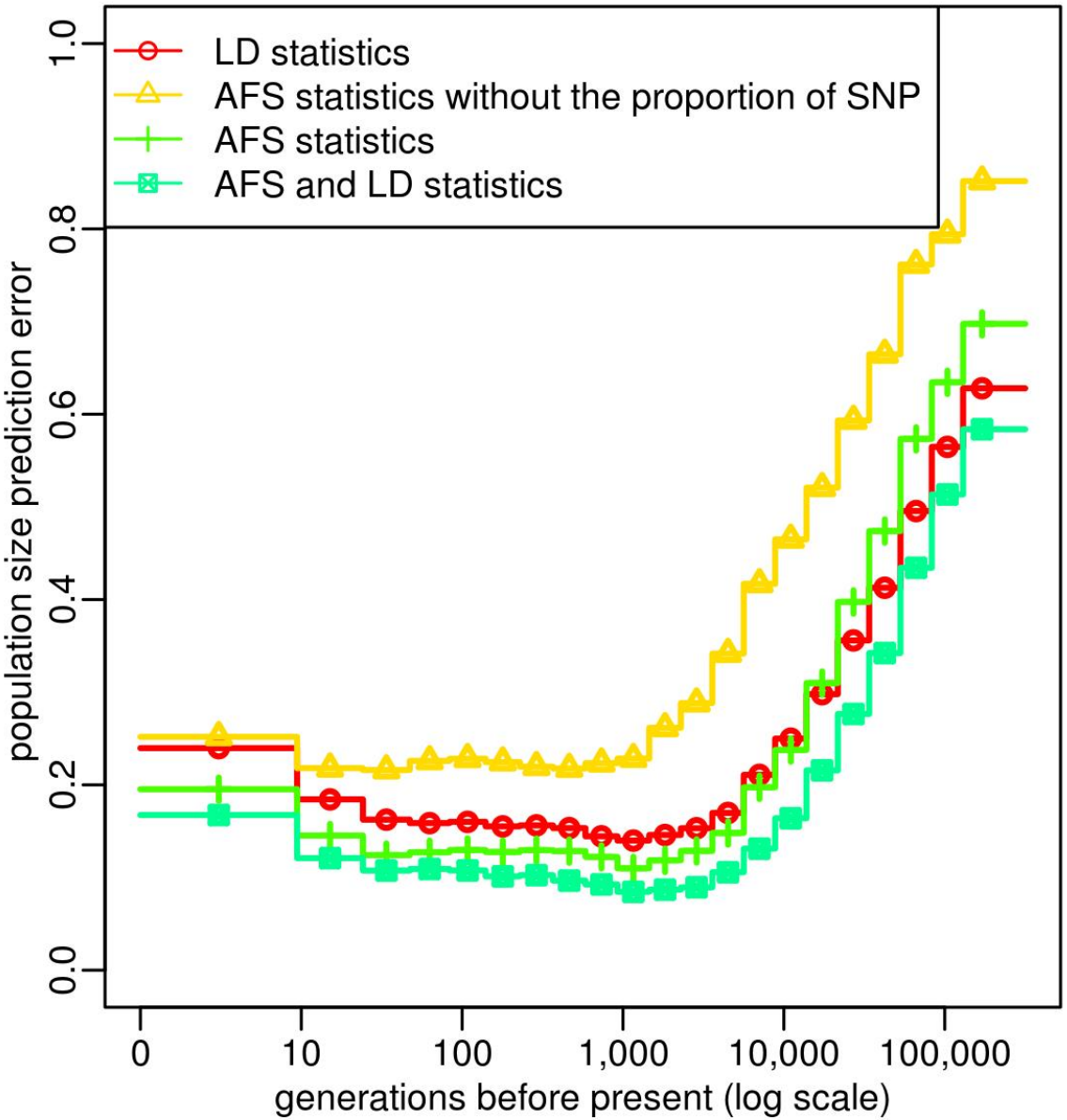
Plan de l'exposé

1. Estimation à partir de locus indépendants
   - Généalogie à un locus
   - Méthodes d'estimation
2. Estimation à partir de données génomiques haut débit
   - Données complètes: les modèles SMC
   - Données résumées
   - L'approche ABC (Approximate Bayesian Computation)
3. Conclusions

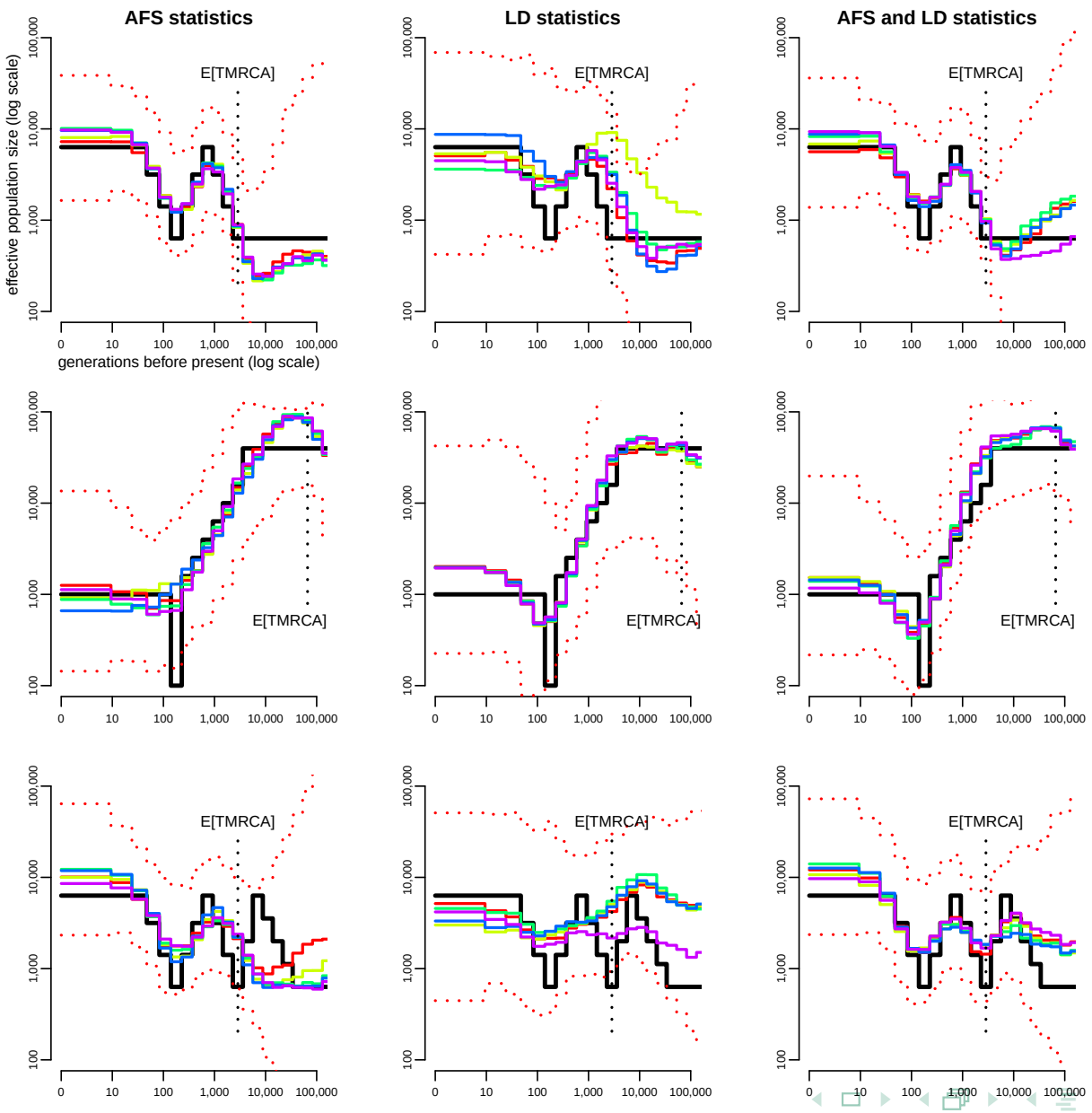
Principe (Beaumont *et al*, 2002)

- **Approche par simulation:**
  1. Tirer une histoire démographique selon une loi a priori.
  2. Simuler un échantillon de génomes selon cette histoire.
  3. Caculer les statistiques résumantes.
  4. Conserver les paramètres de l'histoire si les statistiques simulées ressemblent à celles observées.
- **Avantages:**
  - Combiner différentes catégories de statistiques.
  - Pas d'approximations du modèle.
  - Mesure de l'incertitude du résultat (approche Bayesienne).
- **Inconvénients:** temps de calcul important!
- Boitard *et al* (2016): combiner AFS et LD.

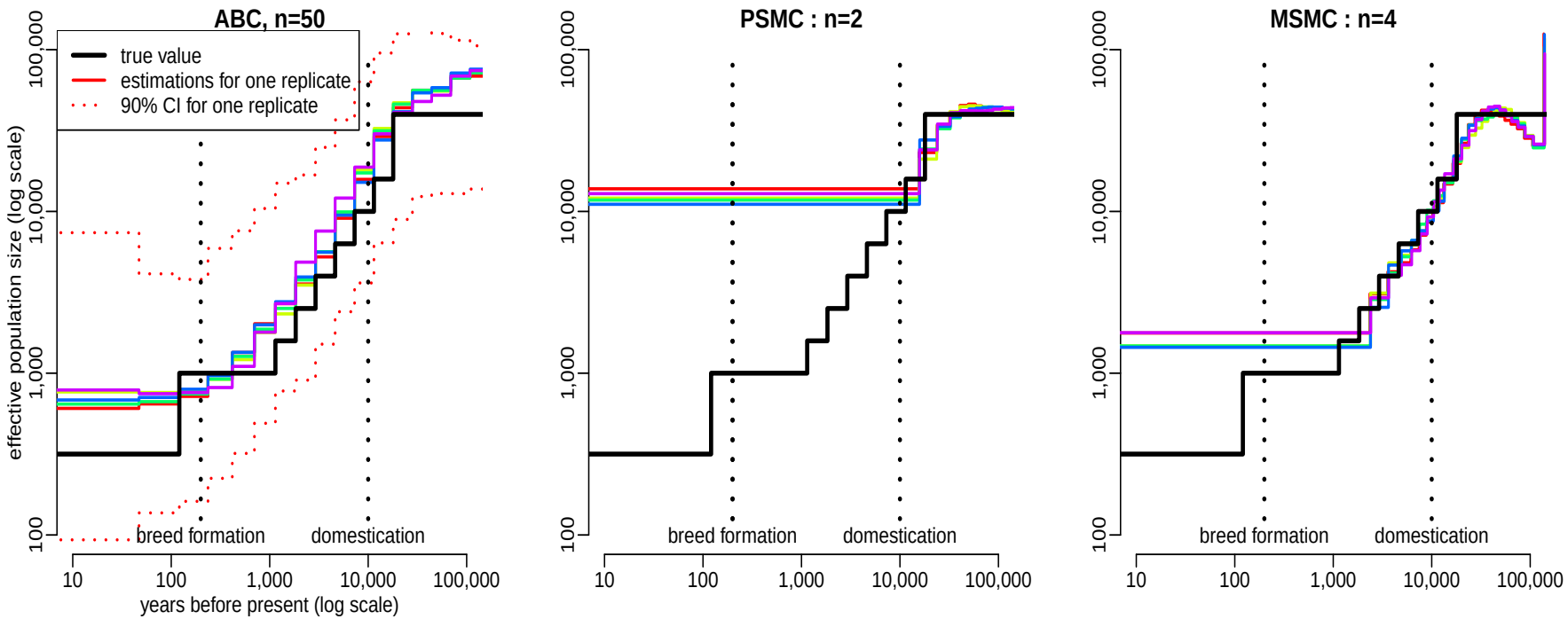
Erreur de prédiction moyenne ($n = 50$)



AFS et LD complémentaires

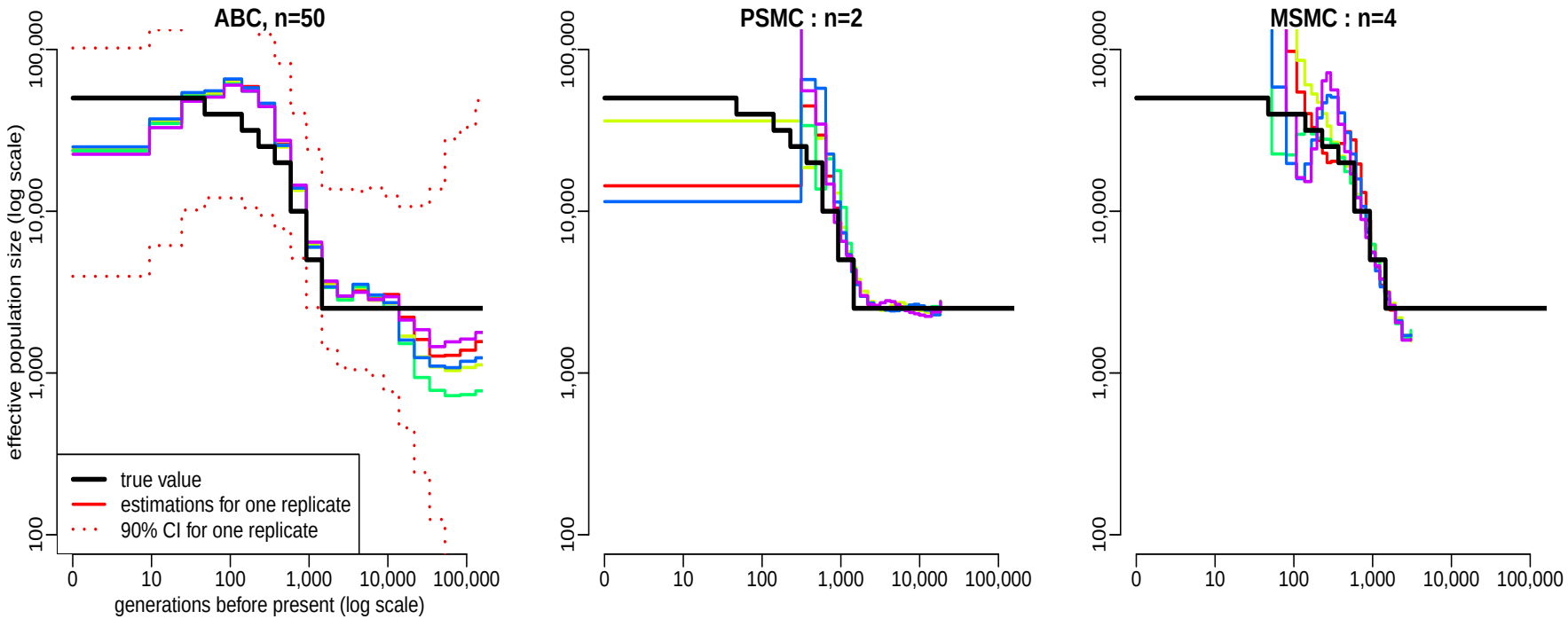


Comparaison avec les approches SMC

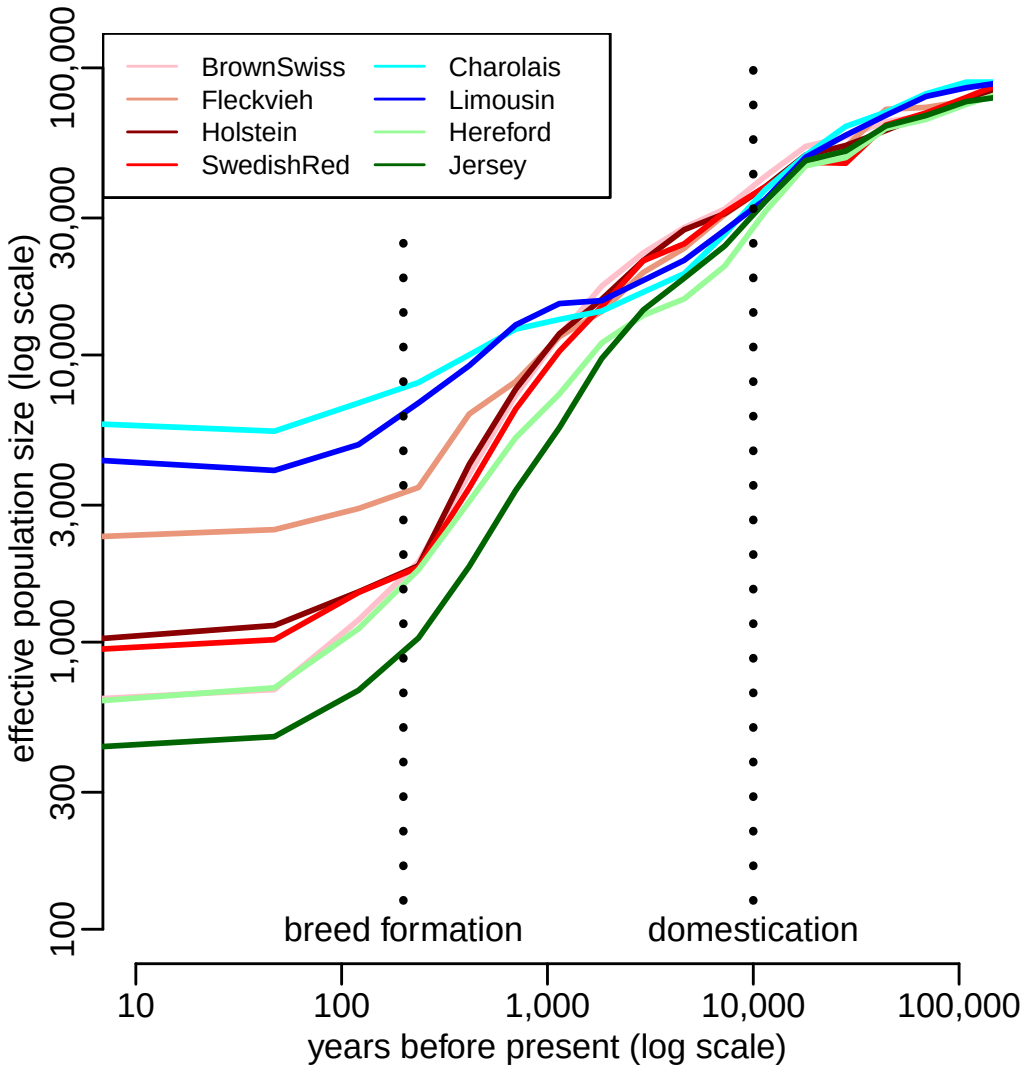


Utiliser des grands échantillons améliore l'estimation de l'histoire récente.

Comparaison avec les approches SMC



Application chez la vache (projet 1000 génomes)



- Les histoires dans chaque race **divergent à partir de la domestication**.
- Déclin continu **antérieur à la domestication**, similaire MacLeod *et al* (2013).
- Classement des races d'après leur taille récente cohérent.

Plan de l'exposé

1 Estimation à partir de locus indépendants
- Généalogie à un locus
- Méthodes d'estimation

2 Estimation à partir de données génomiques haut débit
- Données complètes: les modèles SMC
- Données résumées
- L'approche ABC (Approximate Bayesian Computation)

3 Conclusions

Conclusions

- Diversité génomique présente contient beaucoup d'information sur variations passées de la taille efficace.
- Résolution accrue par le haut débit.
- Approche SMC très prometteuse mais:
  - Méthodes actuelles peu précises pour l'histoire récente.
  - Séquences continues nécessaires.
- Approches par statistiques résumantes: bonne alternative, adaptées à données plus variées.
- ABC: grande flexibilité mais mise en ouevre plus longue.