Variations de la taille d'une population : estimation à partir de génomes entiers.

Simon Boitard, Willy Rodríguez, Flora Jay, Stefano Mona, Frédéric Austerlitz

Variations de la taille d'une population: estimation à partir de génomes entiers.

Simon Boitard[1], Willy Rodriguez[2], Flora Jay[3], Stefano Mona[4], Frédéric Austerlitz[3]

1 : Génétique, Physiologie et Systèmes d'Elevage (INRA & INP), Toulouse
2 : Institut de Mathématiques de Toulouse (Univ. Toulouse & CNRS)
3 : Eco-anthropologie et Ethnobiologie (MNHN & CNRS & Univ. Paris Diderot)
4 : Institut de Systématique, Evolution et Biodiversité (EPHE & MNHN & CNRS & UPMC), Paris

GDR AIEM, 30 nov - 1 dec 2016

La taille d'une population

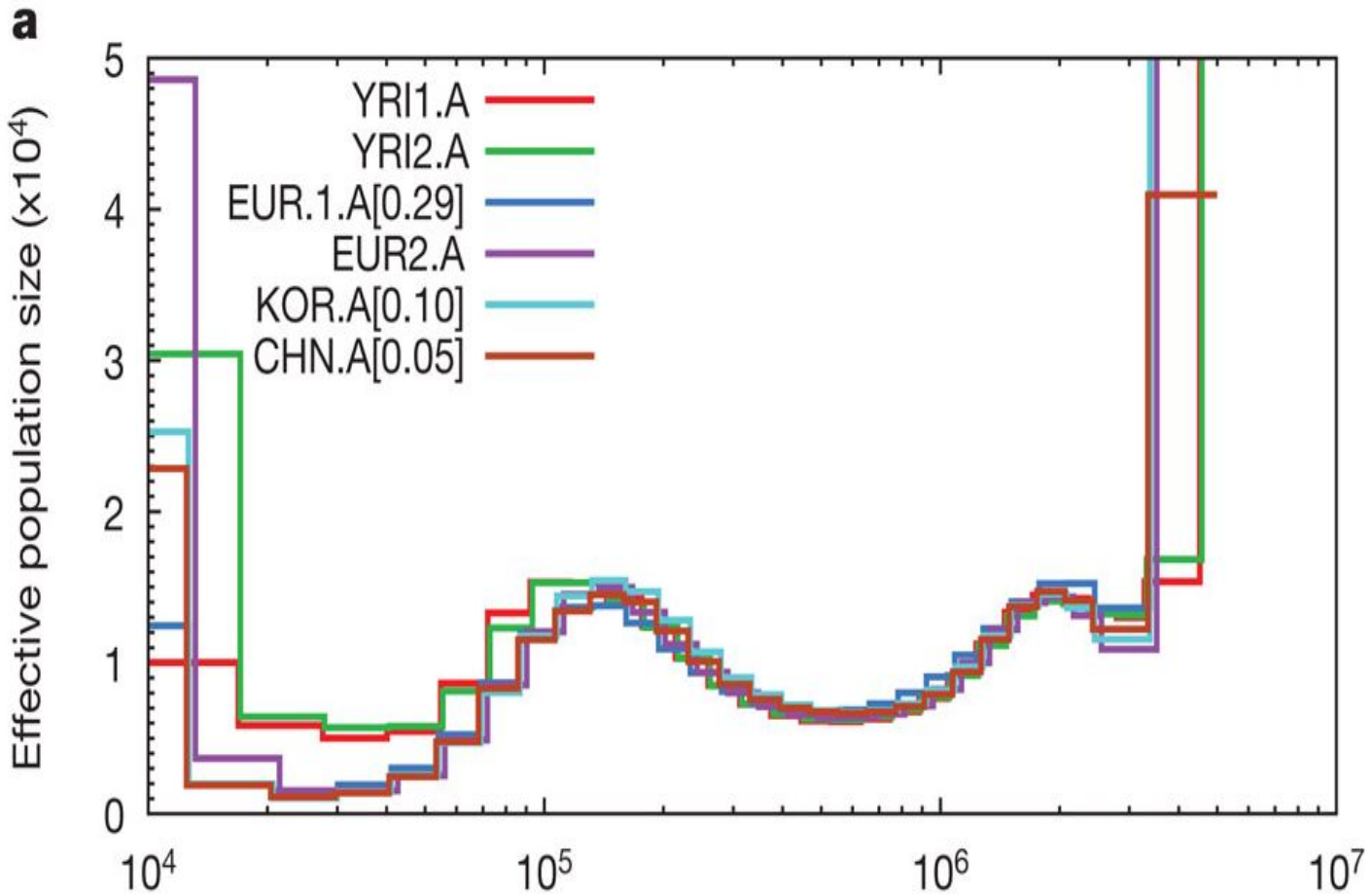
- **Modèle de Wright-Fisher:** population isolée de taille constante, générations non chevauchantes, reproduction aléatoire.
  $\rightarrow$ taille $=$ nombre d'individus dans cette population idéale.
- **Intérêt?**
  - Comparer des populations, identifier des populations en danger.
  - Hypothèse nulle pour détecter des locus sous sélection.
- **Estimation** à partir de données moléculaires actuelles:
  - Proportion de sites polymorphes (Watterson, 1975).
  - Nombre de différences entre deux séquences homologues (Tajima, 1983).

Modèles à taille variable

- **Motivation:**
  - Vision dynamique, identifier les populations en déclin.
  - Interprétation des variations de taille: évènements géologiques, climatiques, anthropomorphiques . . .
  - Meilleure détection des locus sous sélection.
- **Estimation:**
  - Plus d'allèles rares (resp. fréquents) dans les populations en expansion (resp. en déclin).
  - Nombreuses méthodes: Bottleneck (Cornuet et Luikart, 1996), Msvar (Beaumont, 1999), Beast (Drummond and Rambaut, 2007), VarEff (Nikolic and Chevalet, 2014).
  - Petit nombre de locus **indépendants sans recombinaison**

PSMC (Li and Durbin, 2011)

Estimation basée sur le génome entier d'un individu diploïde, grâce à une approximation du coalescent avec recombinaison.

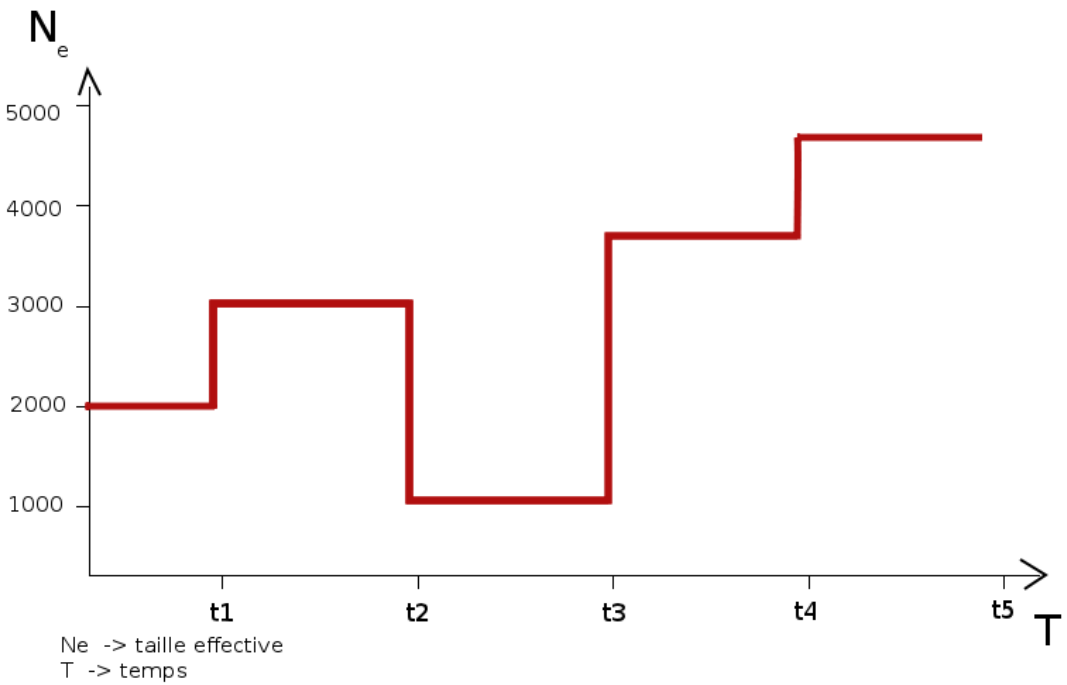


Motivation

- Plusieurs méthodes récentes exploitant des génomes entiers : dical (Sheehan *et al*, 2013), MacLeod *et al* (2013), Harris and Nielsen (2013), MSMC (Schiffels and Durbin, 2014).
- Limitées à un petit nombre d'individus ($\approx$ 5 diploïdes max).
  $\rightarrow$ Faible précison pour l'histoire récente.
- **Objectif:** Développer une méthode exploitant de **grands échantillons** de **génomes entiers**.

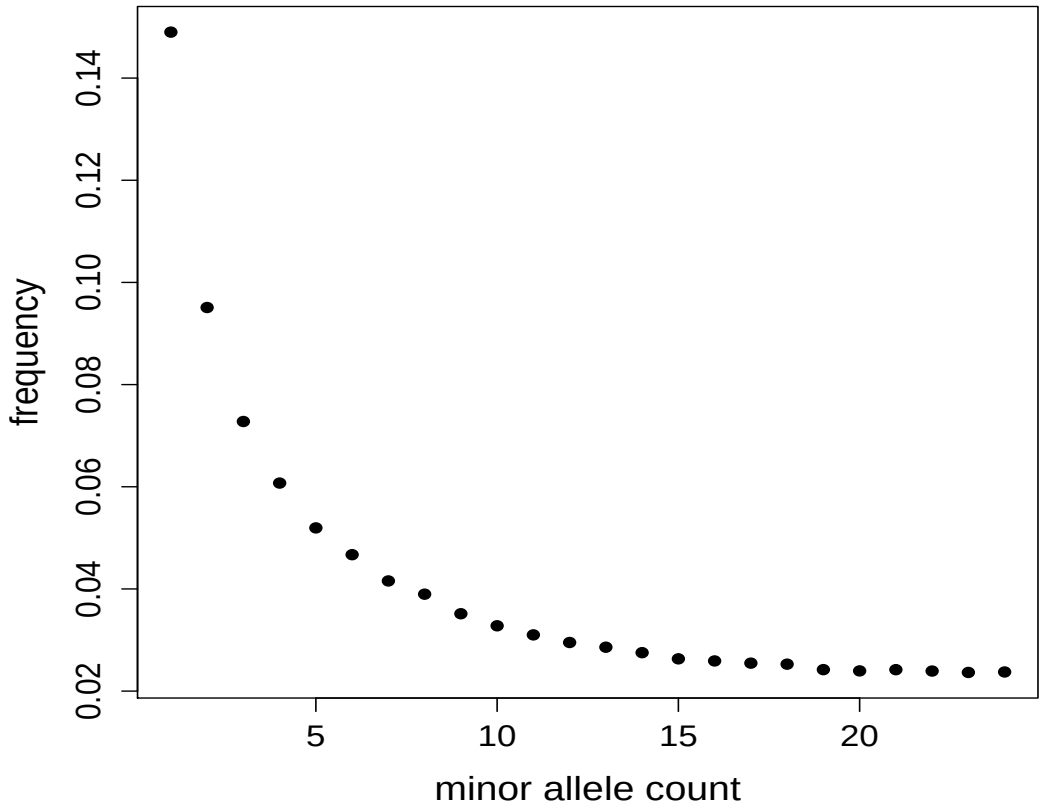
Approche ABC (Approximate Bayesian Computation)

- **Approche par simulation:**
  1. Tirer une histoire démographique selon une loi a priori.
  2. Simuler un échantillon de génomes selon cette histoire.
  3. Conserver les paramètres de l'histoire si l'échantillon simulé ressemble à l'échantillon observé.
- **Astuce fondamentale:** A l'étape 2, remplacer les données brutes par des statistiques résumant ces données.



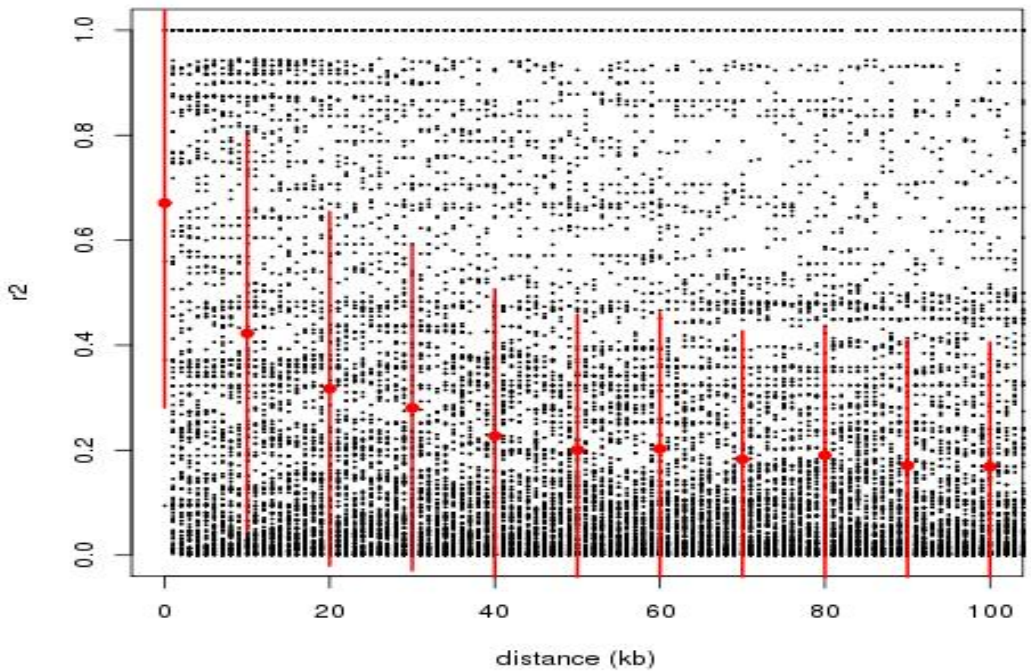
Statistiques résumantes : fréquences alléliques (AFS)

- 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 (taille de l'échantillon).



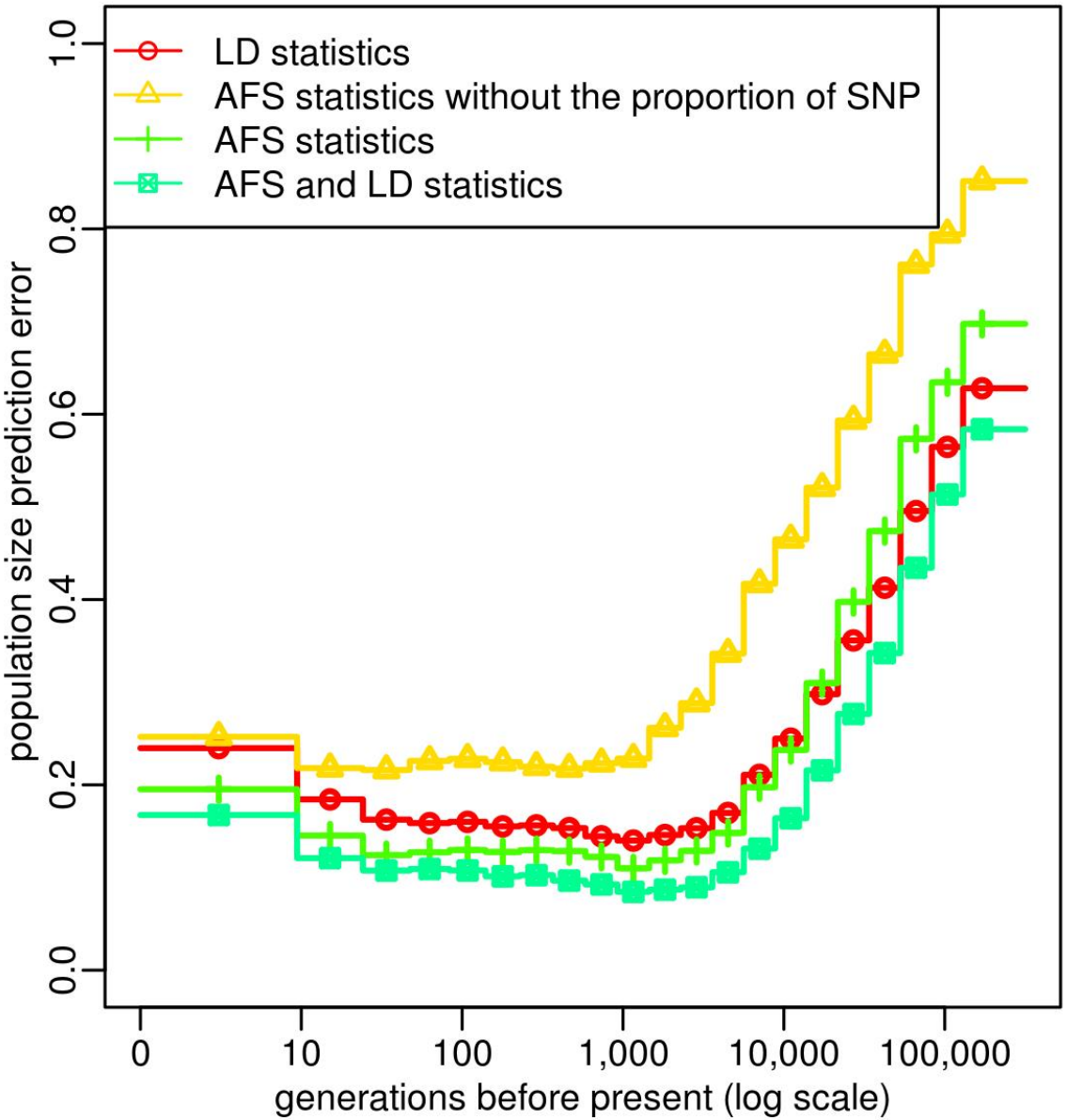
Statistiques résumantes : déséquilire de liaison (LD)

- r^2 : pour une paire de SNPs, corrélation des génotypes.

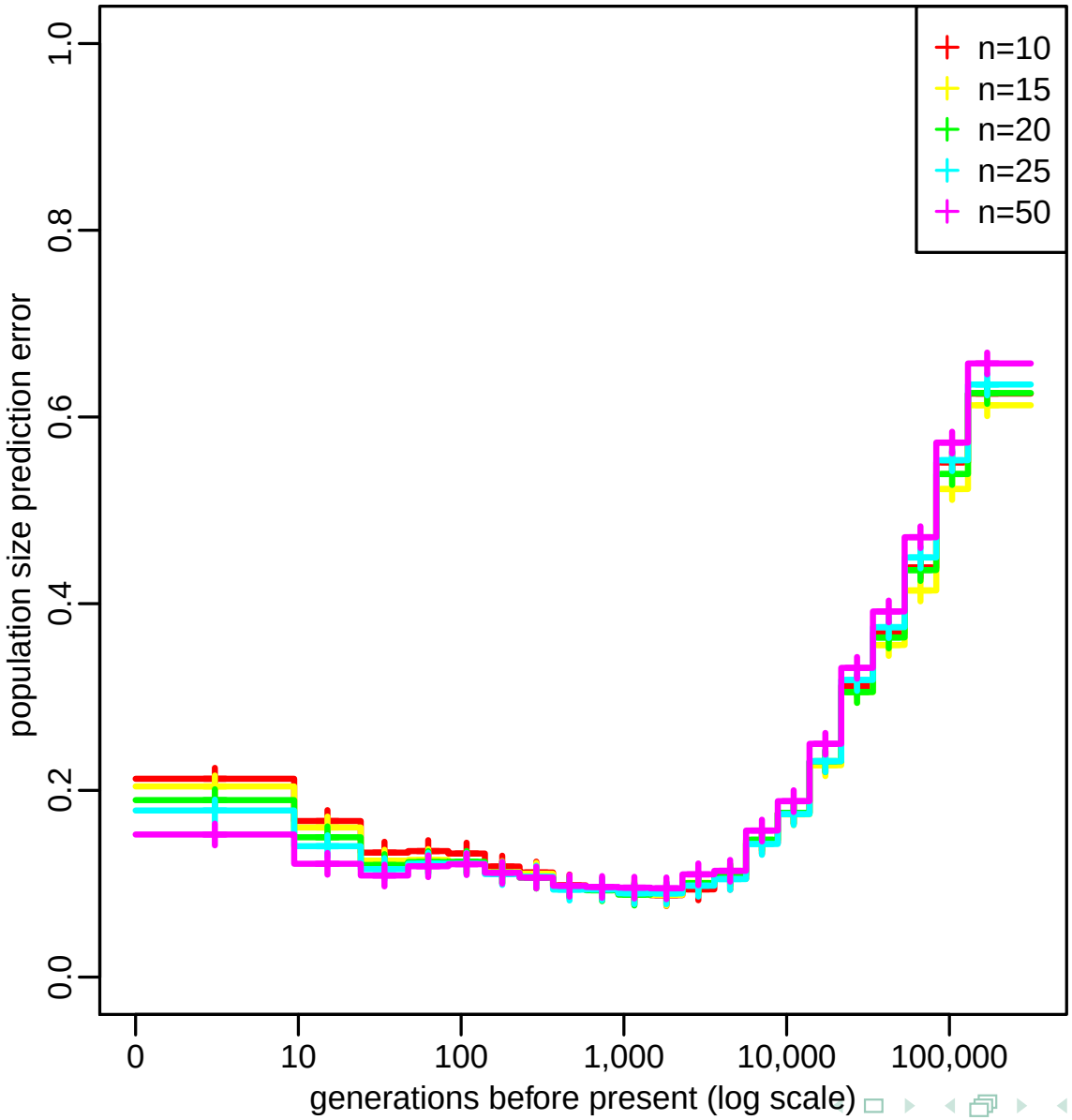


- Statistiques : r^2 moyen pour m catégories de distance entre SNPs, de 100bp à 2Mb.

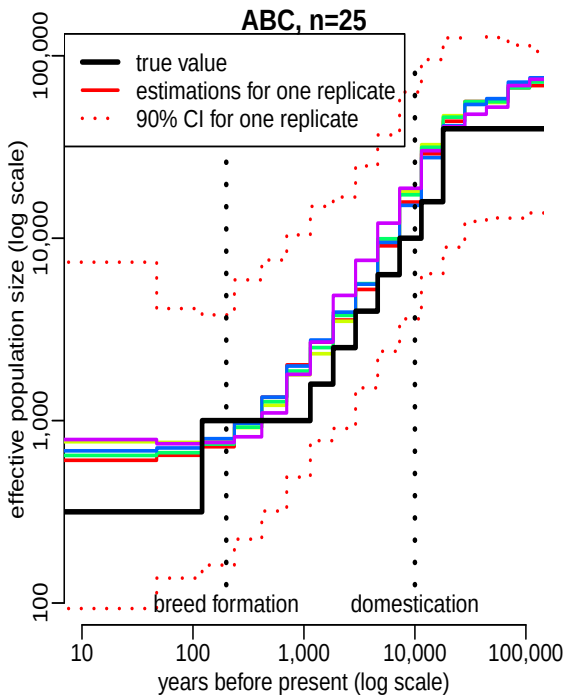
Erreur de prédiction moyenne ($n = 25$)

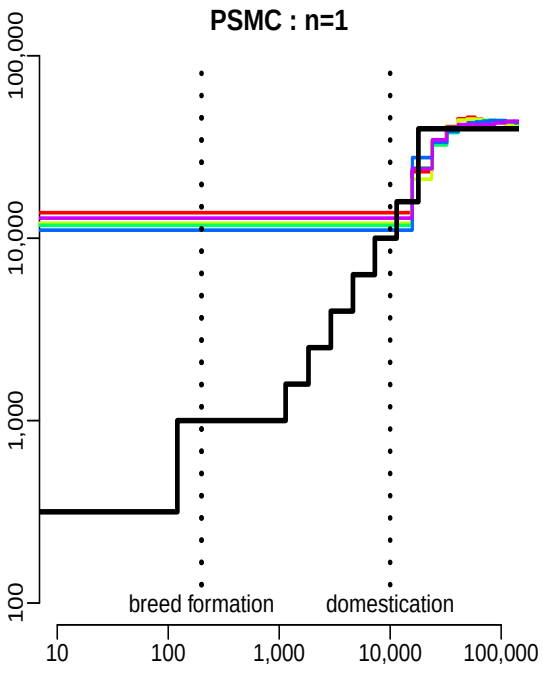


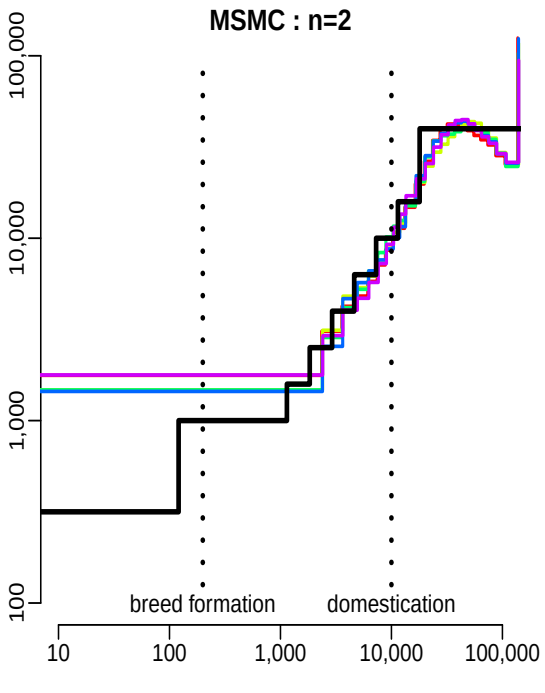
Influence de la taille d'échantillon



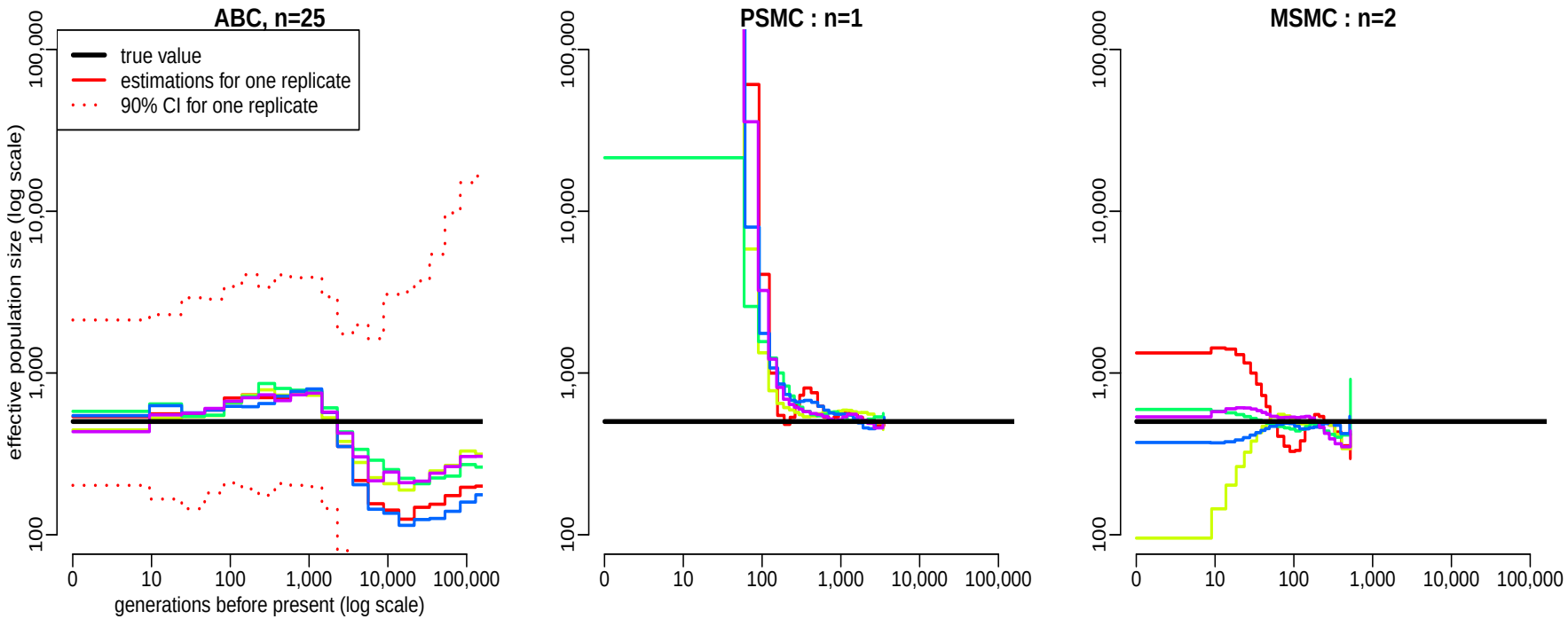
Exemple: population en déclin



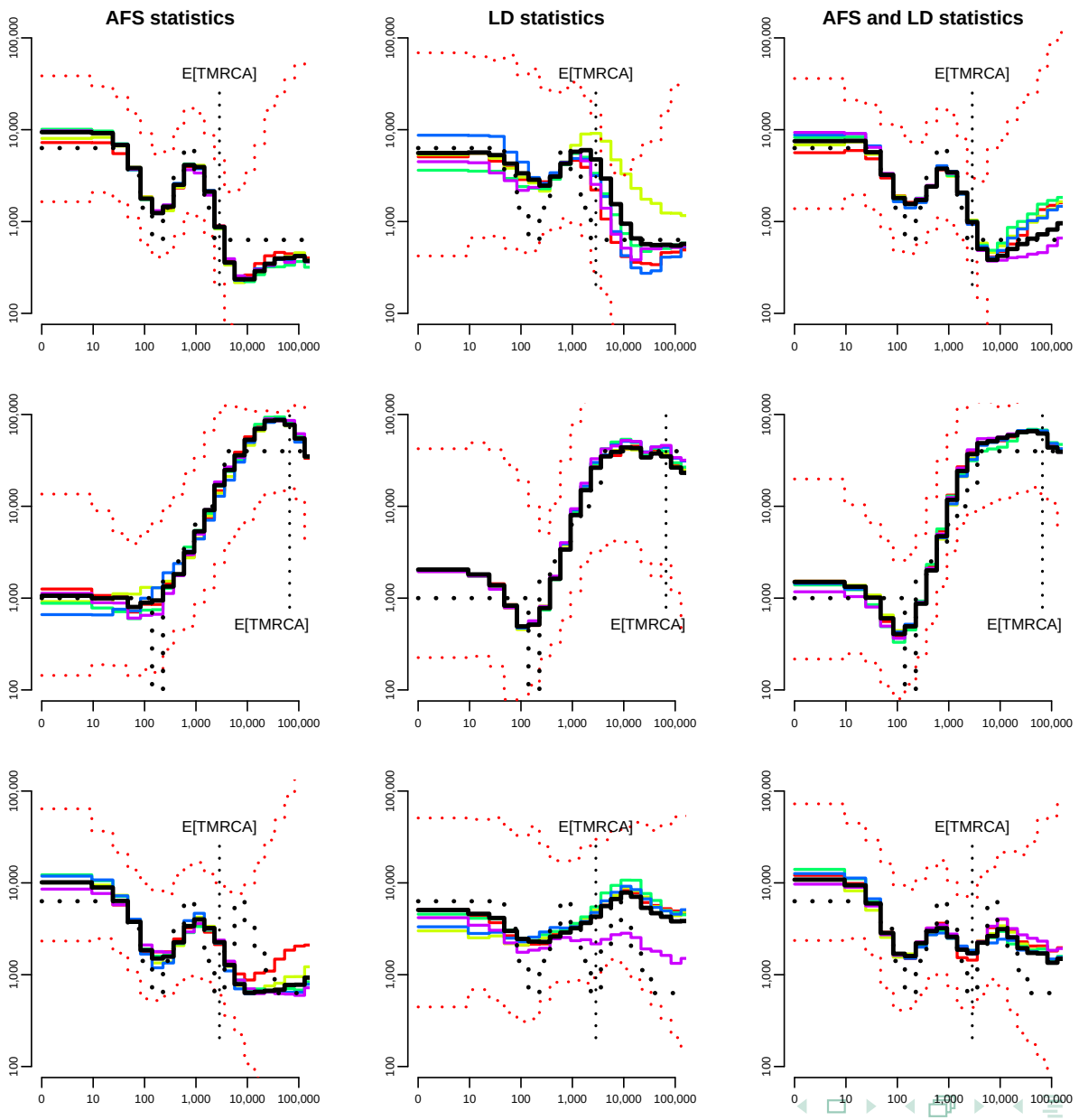




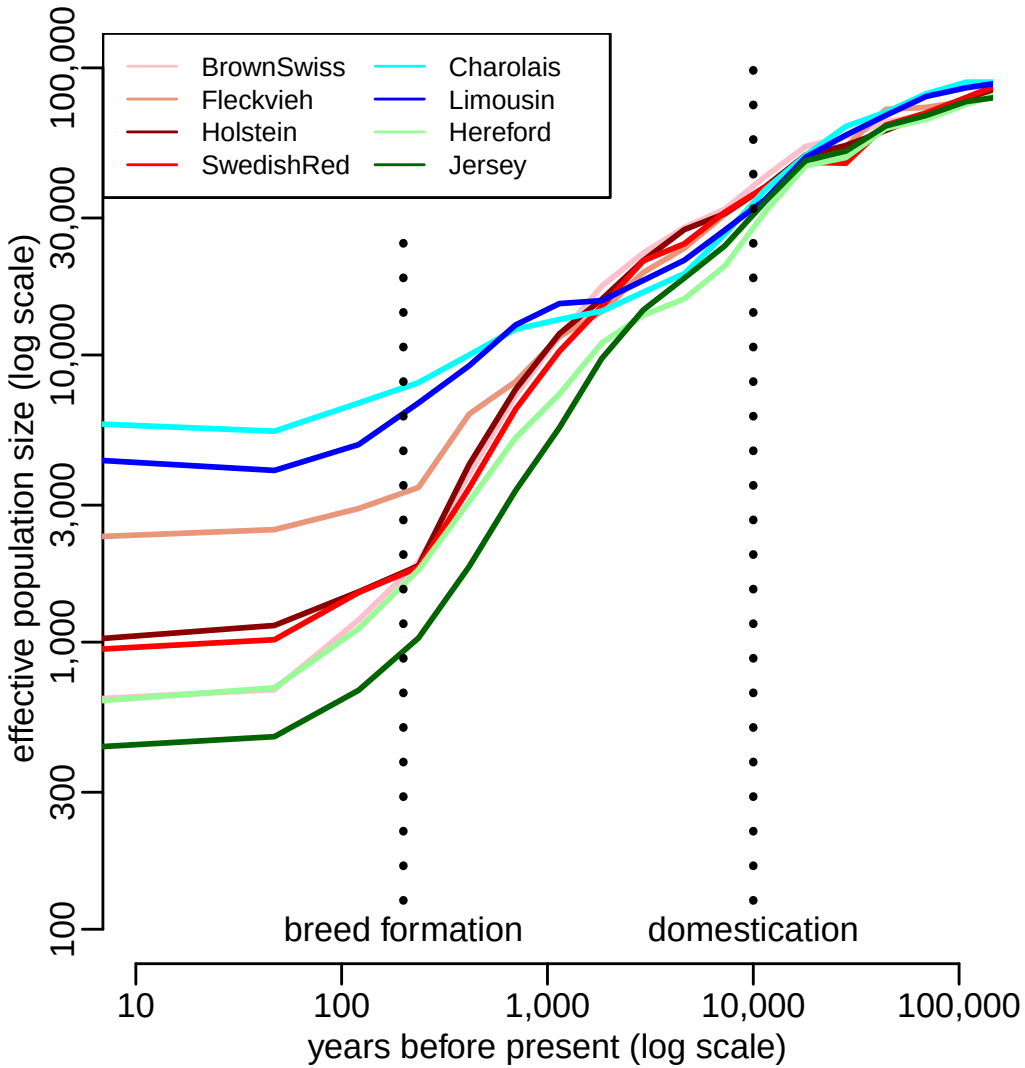
Exemple: population constante



LD et AFS complémentaires

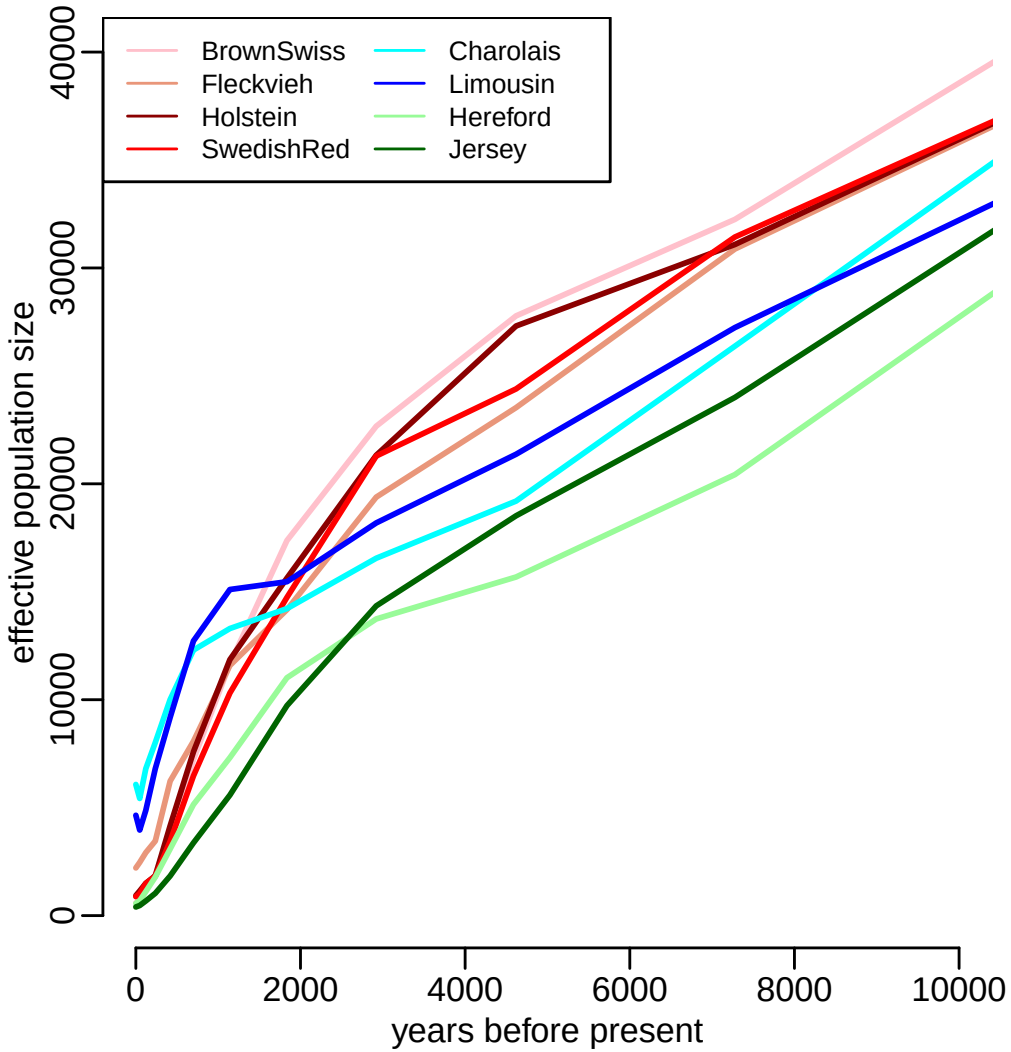


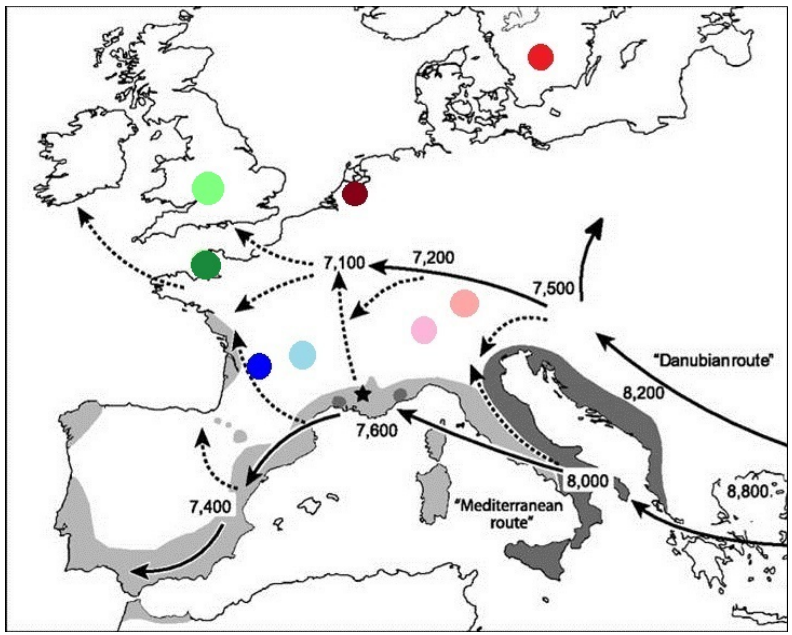
Application : projet 1000 génomes bovins



- 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.

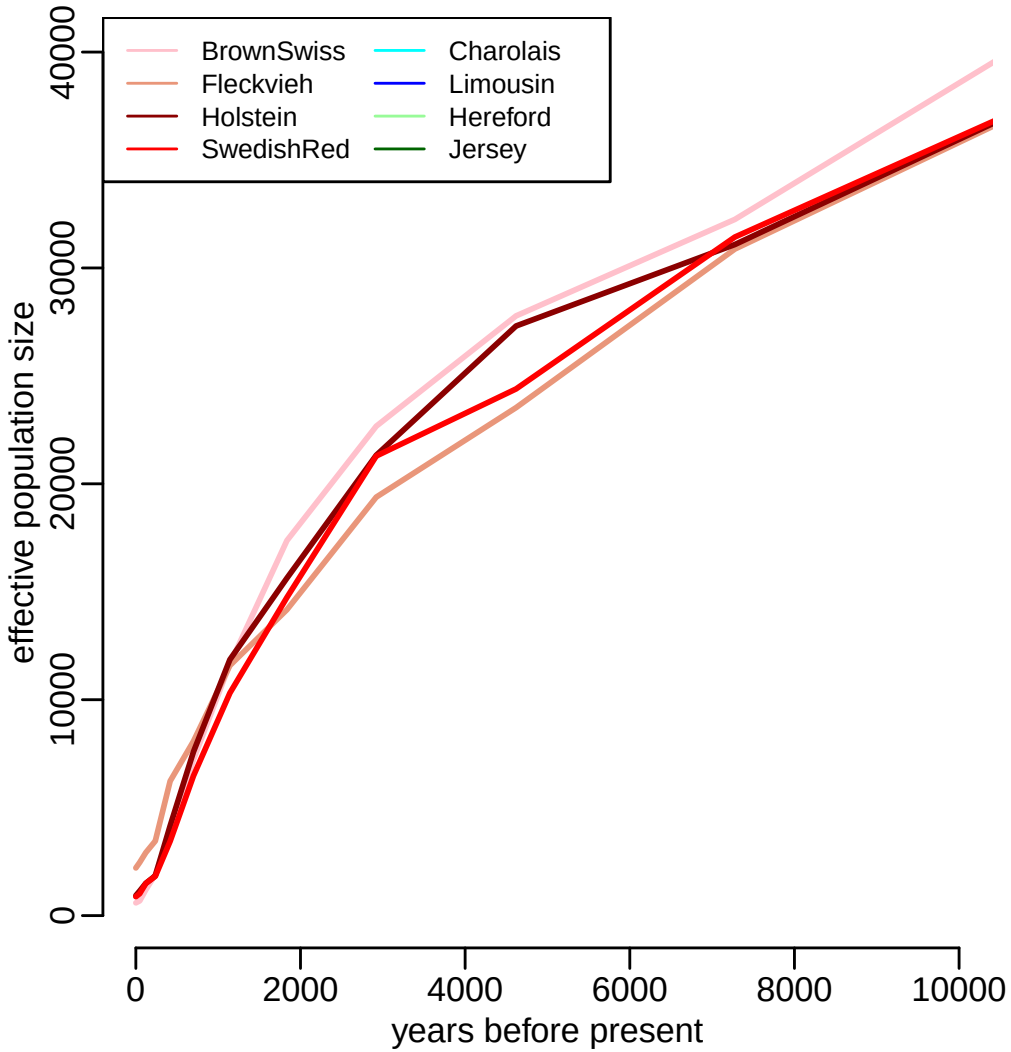
Différences entre races : 3 groupes géographiques

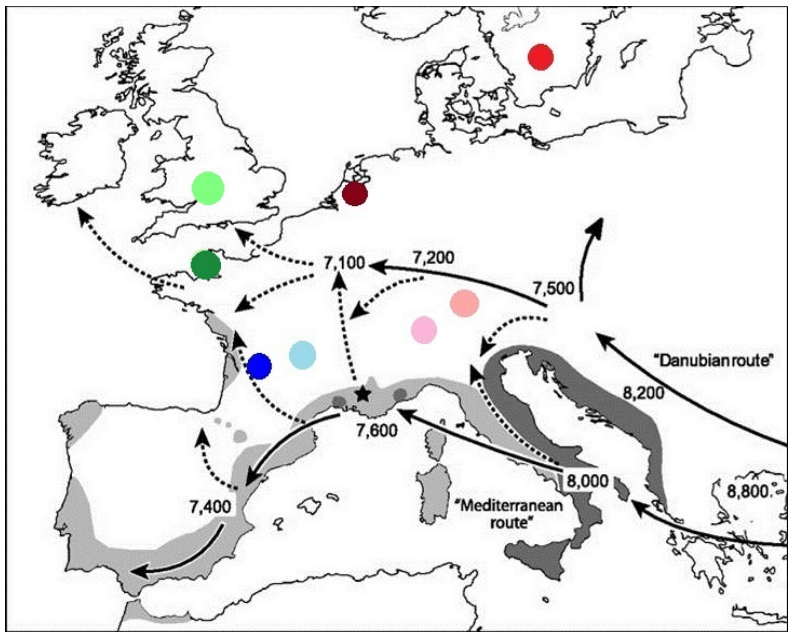




- origine “Danubienne”.
- origine “Méditerranéenne”.
- origine plus complexe.

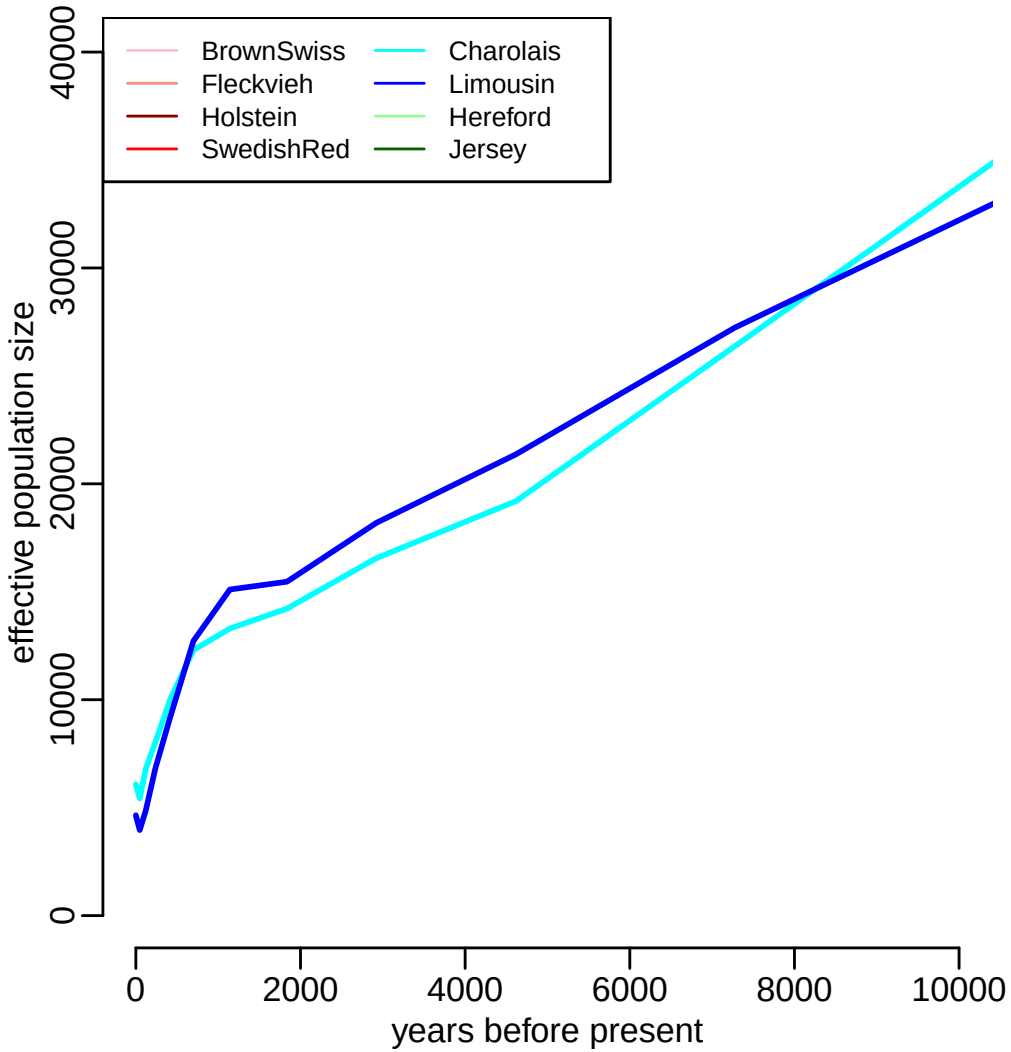
Différences entre races : 3 groupes géographiques

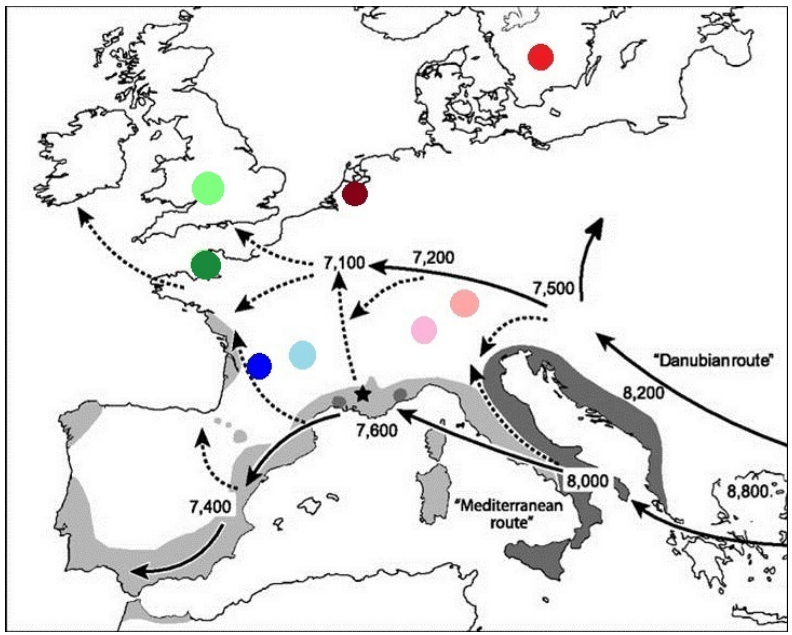




- origine “Danubienne”.
- origine “Méditerranéenne”.
- origine plus complexe.

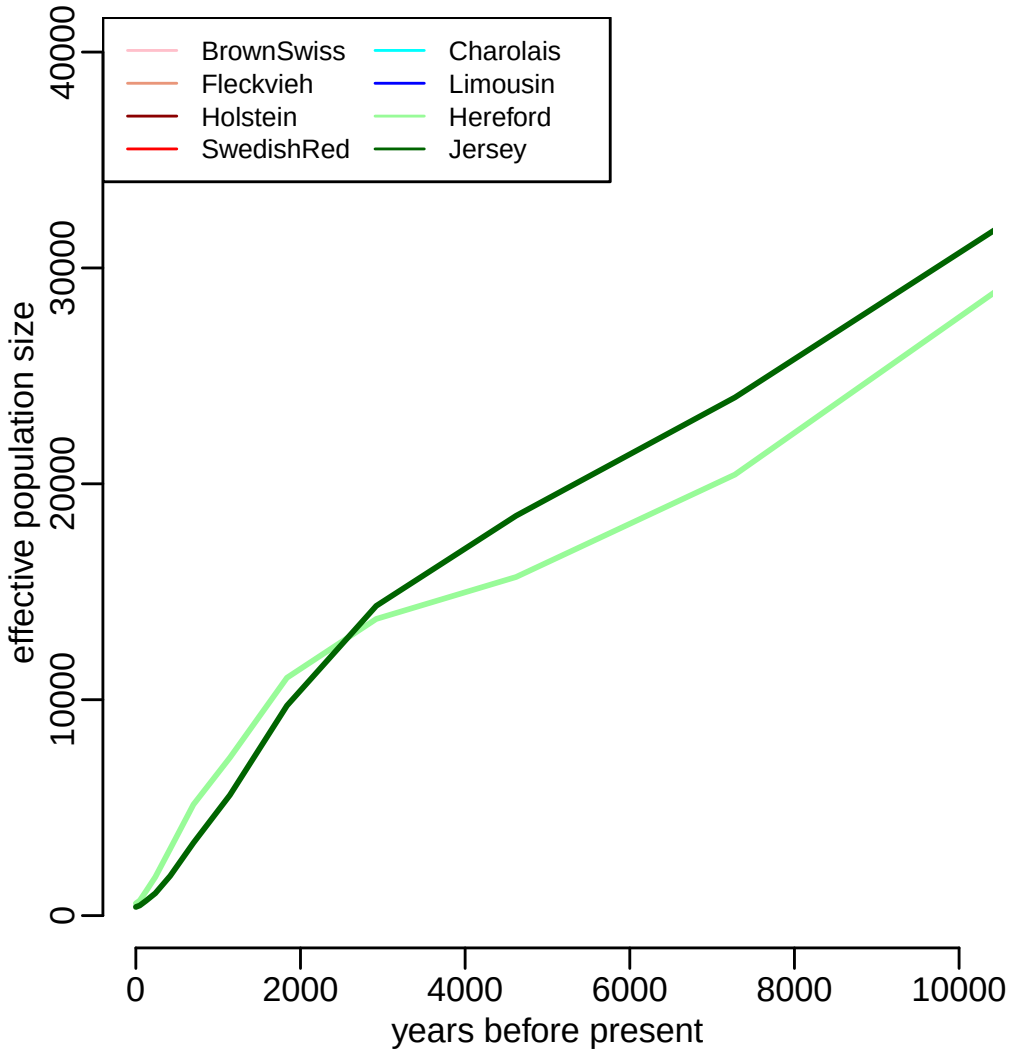
Différences entre races : 3 groupes géographiques

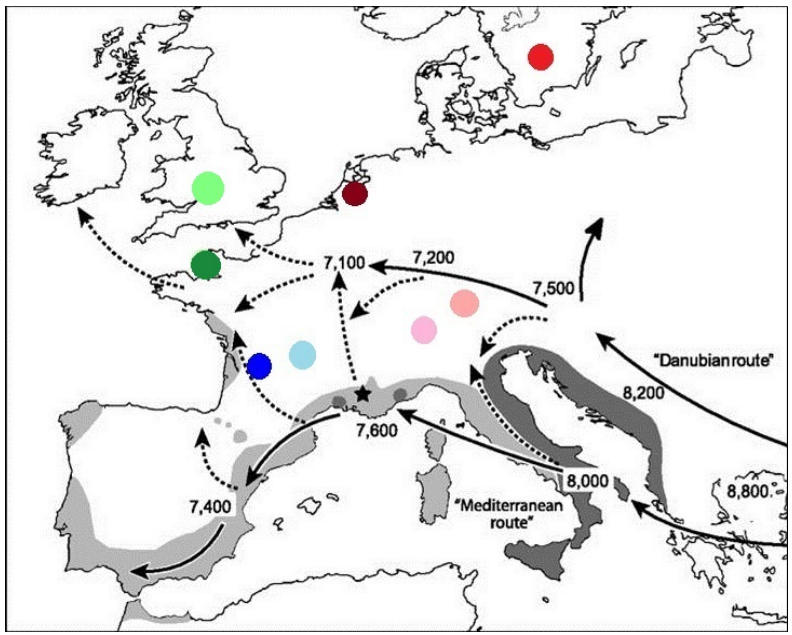




- origine “Danubienne”.
- origine “Méditerranéenne”.
- origine plus complexe.

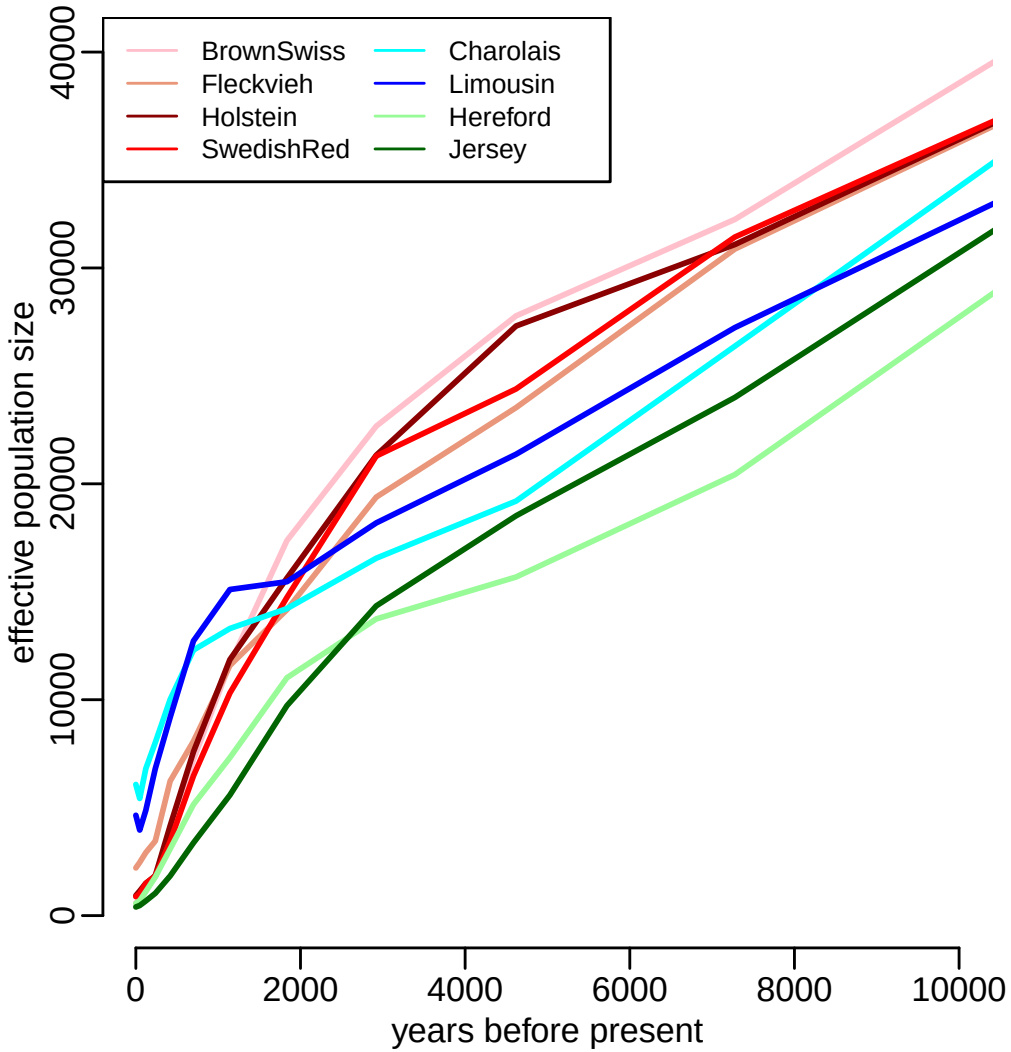
Différences entre races : 3 groupes géographiques

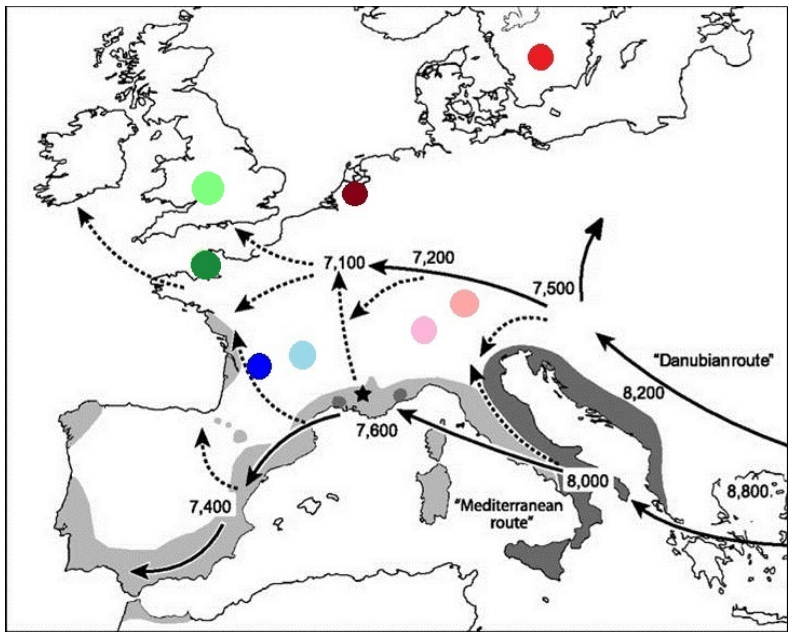




- origine “Danubienne”.
- origine “Méditerranéenne”.
- origine plus complexe.

Différences entre races : 3 groupes géographiques





- origine “Danubienne”.
- origine “Méditerranéenne”.
- origine plus complexe.

Conclusions

- ABC permet d'exploiter des grands échantillons de génomes entiers.
- Bonne estimation de l'histoire démographique dans son ensemble, y compris partie récente.
- Combiner LD et AFS améliore la prédiction.
- Insensible aux erreurs de phasage et peu affecté par erreurs de séquençage (MAF 20%).
- Ref: Boitard *et al* (2016), *PLoS Genet* 12(3): e1005877
- url: https://forge-dga.jouy.inra.fr/projects/popsizeabc/

Perspectives

- Tenir compte de la **structure des populations**, qui peut générer de faux changements de tailles (cf Olivier Mazet).
- Utiliser des **données temporelles**: ADN ancien, cryobanques.

Remerciements

- Plateforme bioinfo Genotoul.
- Projet 1000 génomes bovin (Ben Hayes, Didier Boichard).
- Lounes Chikhi (Université Toulouse III & CNRS), Olivier Mazet, Simona Grusea (INSA Toulouse).
- Membres de l'ANR Demochips.
- Bertrand Servin.