Détection de locus sous sélection positive

Simon Boitard, Bertrand Servin

INRA, GenPhySE, Toulouse

M2 BBS, Atelier Génétique Statistique, 2017-2018

Contexte

- La plupart des régions du génome évoluent sous neutralité, mais certaines ont évolué sous sélection positive / adaptative, naturelle ou artificielle.
- En étudiant la diversité génétique à l'échelle du génome entier (puces de génotypage, séquençage), on peut chercher à identifier ces régions : scans de sélection.
- Sélection positive modifie les fréquences alléliques du locus sous sélection, mais aussi des locus voisins.
- Enjeu théorique (mécanismes évolutifs) et appliqué (gènes liés à des fonctions stratégiques en médecine, agronomie . . .).

Plan de l'exposé

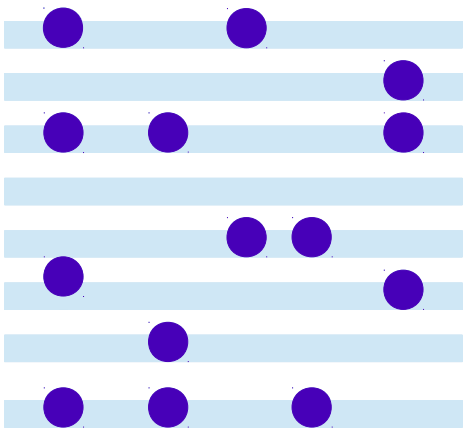
1. Effets de la sélection positive sur le génome
2. Tests basés sur la différenciation entre populations
   - Approche SNP par SNP
   - Approche haplotypique
3. Résultats
   - Simulations
   - Application chez le mouton
   - Application chez la vache
4. Conclusions et Perspectives
5. TP

Plan de l'exposé

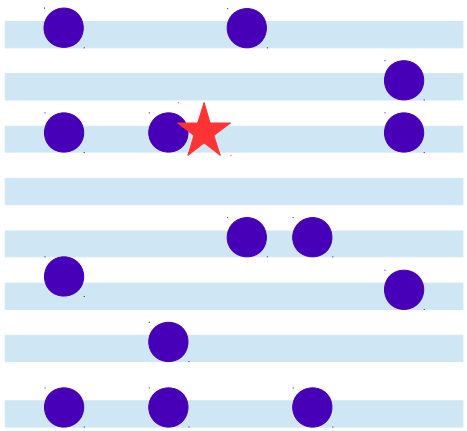
1. Effets de la sélection positive sur le génome
2. Tests basés sur la différenciation entre populations
   - Approche SNP par SNP
   - Approche haplotypique
3. Résultats
   - Simulations
   - Application chez le mouton
   - Application chez la vache
4. Conclusions et Perspectives
5. TP

Diversité génétique autour d'un locus sous sélection positive

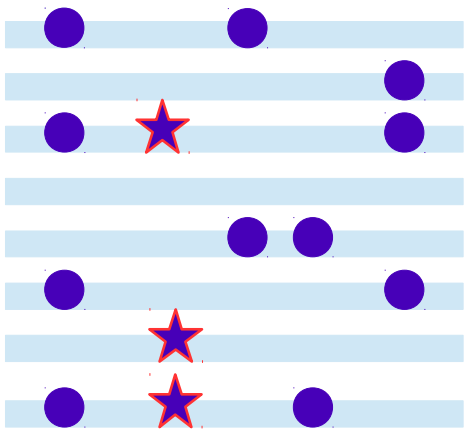
Pas de sélection



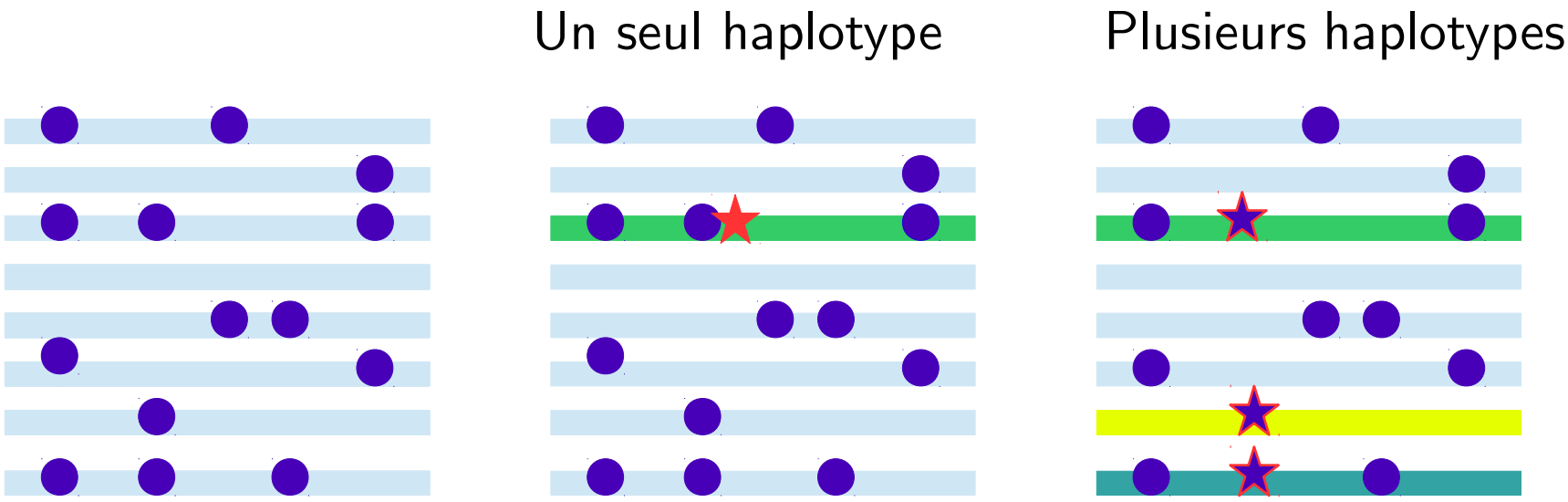
Sélection sur un nouvel alléle



Sélection sur un allèle existant



Diversité génétique autour d'un locus sous sélection positive

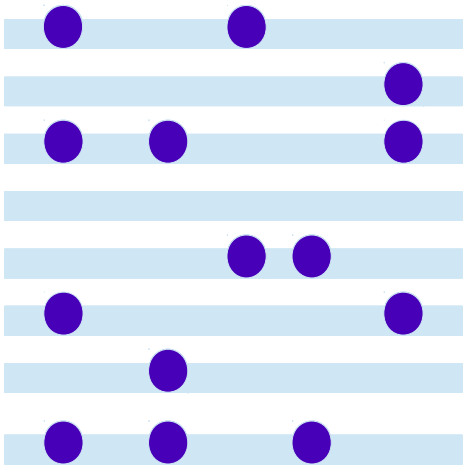


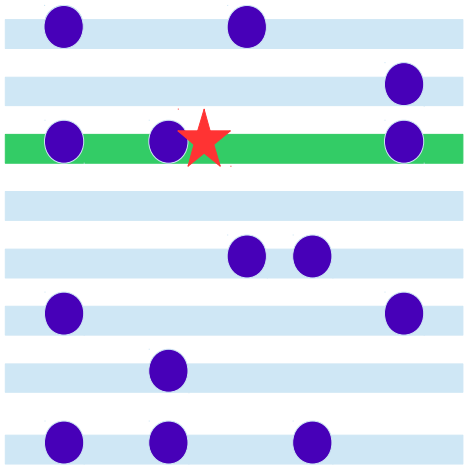
Diversité génétique autour d'un locus sous sélection positive

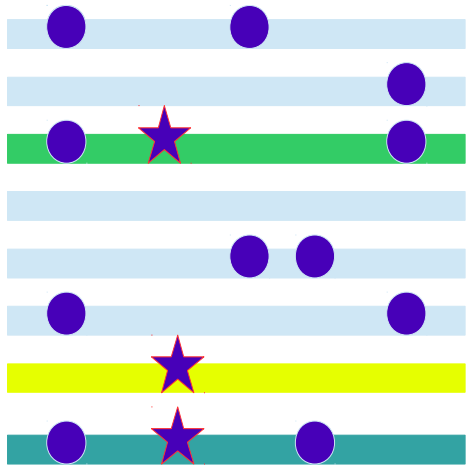
Evolution neutre

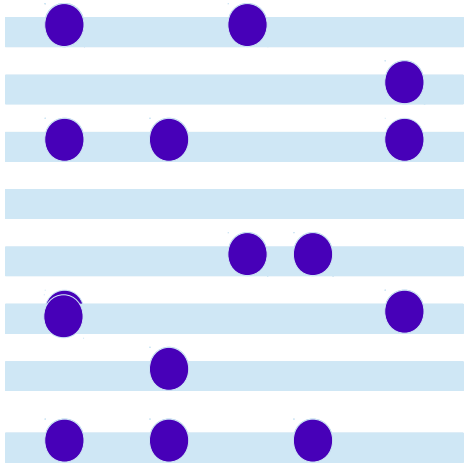
Un haplotype augmente en fréquence

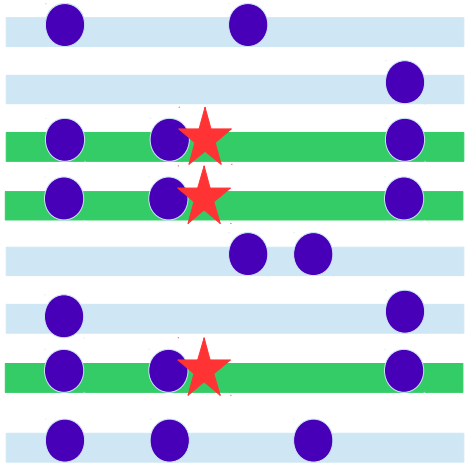
Plusieurs haplotypes augmentent en fréquence

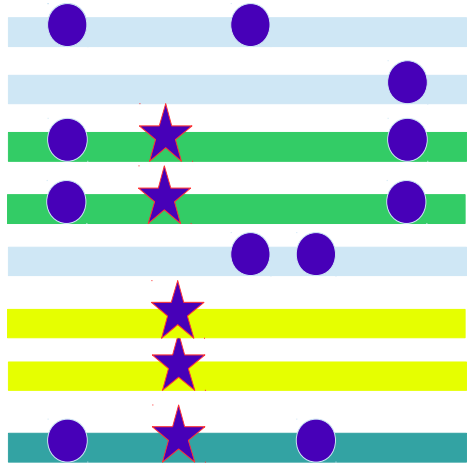




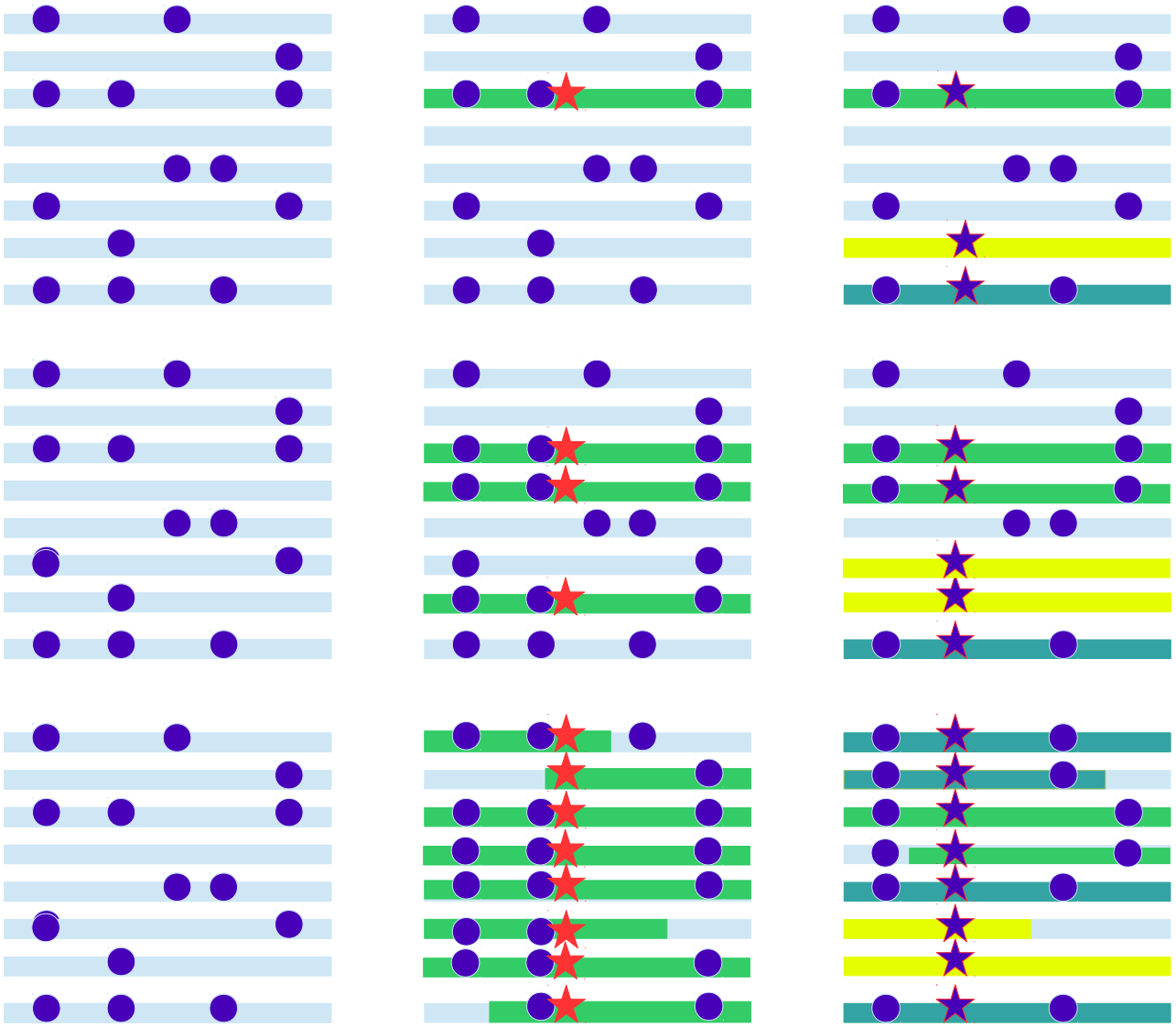




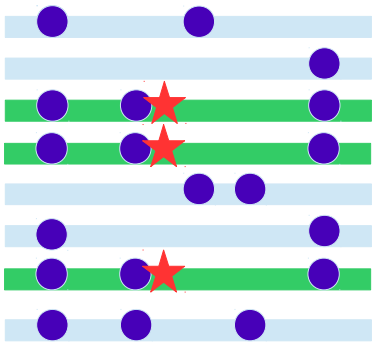


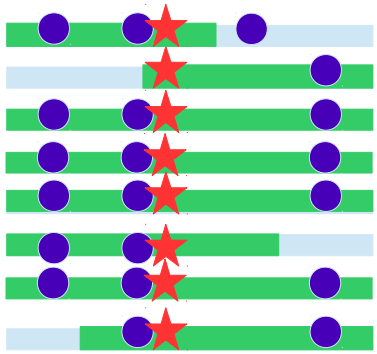


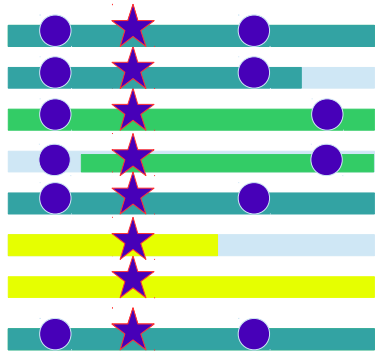
Diversité génétique autour d'un locus sous sélection positive



Effets de la sélection



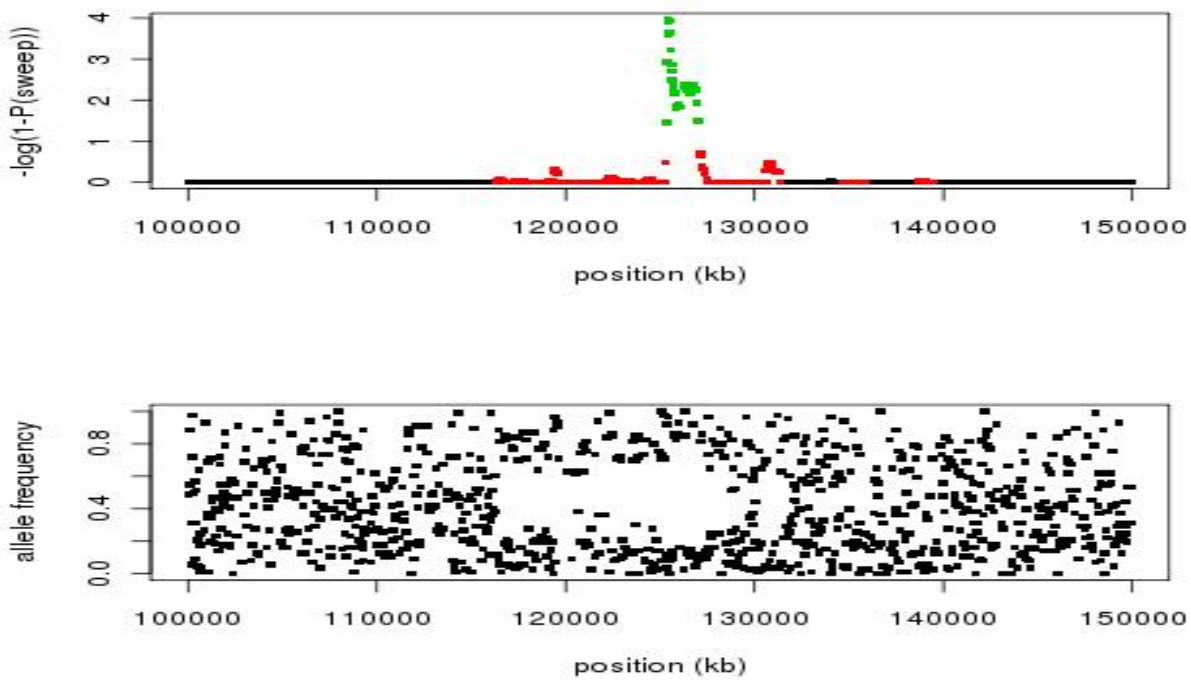




Scénario de sélection	partiel	“hard“	”soft”
Fréquences alléliques	élevées	extrêmes	intermédiaires ou élevées
Fréquences haplotypiques	élevée pour un haplotype	un haplotype fixé	élevées pour plusieurs haplotypes

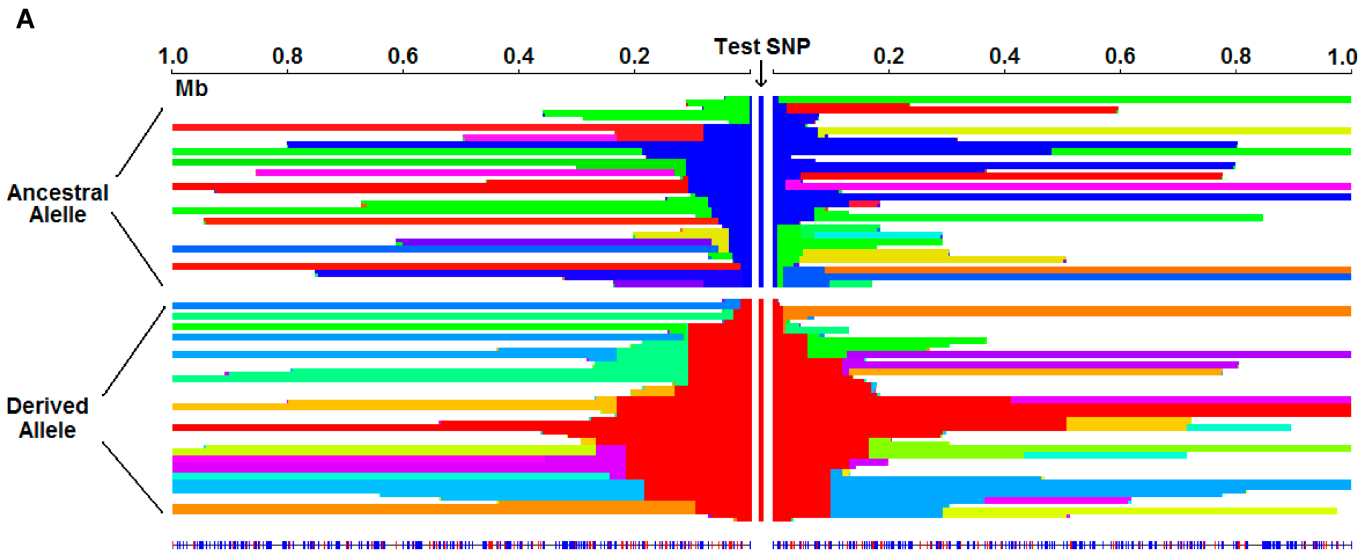
Méthodes de détection

- Baisse de la diversité génétique / modification des fréquences alléliques dans **une population**.
- CLR (Nielsen *et al* 2005), Freq-HMM (Boitard *el al*, 2009).
- **Hard sweeps**.



Méthodes de détection

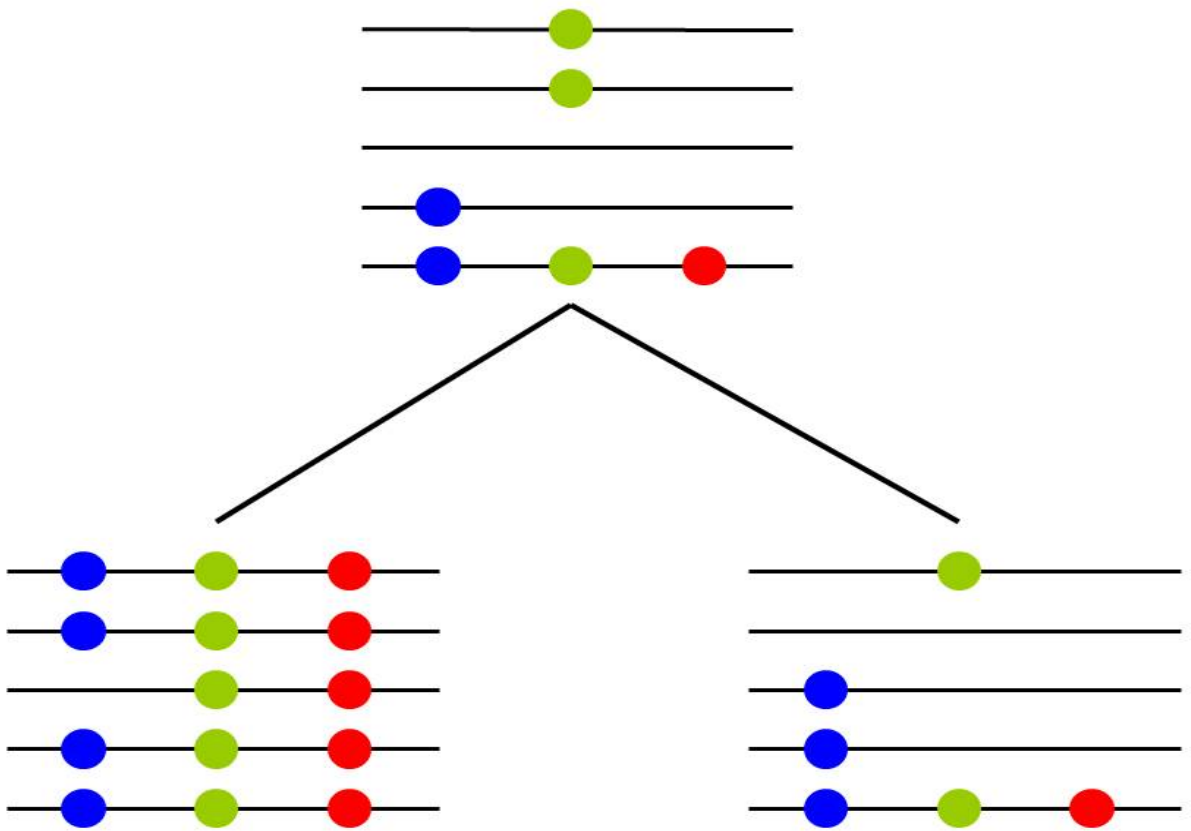
- Recherche d'haplotypes longs à forte fréquence dans **une population**.
- EHH (Sabeti *et al* 2002), iHS (Voight *et al* 2006).
- **Sweeps partiels**.



- Extension à **deux populations** : XP-EHH (Sabeti *et al*, 2007).

Méthodes de détection

- Différenciation génétique élevée **entre plusieurs populations**.
- **Scénarios de sélection plus variés** mais postérieurs à la divergence des population $\rightarrow$ **récents**.



Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

5 TP

Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

5 TP

Utilisation des F_{ST}

Pour un SNP :

- $p = (p_1, \ldots, p_i, \ldots, p_n)$ fréquences de l'allèle 1 dans n populations,

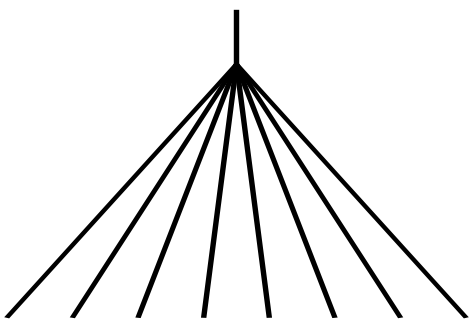
$$F_{ST} = \frac{s_p^2}{\bar{p}(1-\bar{p})}$$

  $\bar{p}$ et s_p^2 moyenne et variance de p.
- Tester H_1 : "sélection dans (au moins) une population" contre H_0 : "évolution neutre".
- H_0 **rejeté si** F_{ST} **grand.**

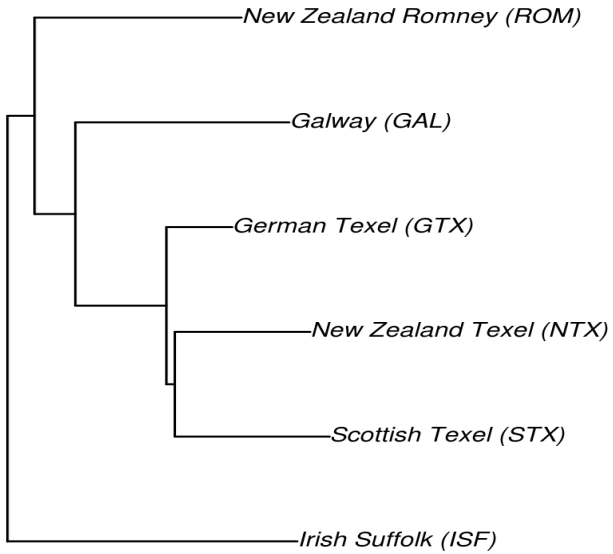
Test de Lewontin et Krakauer (1973)

$$LK = \frac{n-1}{\bar{F}_{ST}} F_{ST}$$

- Sous H_0, LK suit un χ^2 à $n-1$ degrés de liberté . . .
- . . . si les populations ont une **phylogénie en étoile**, **sans migration**, et des **tailles efficaces égales**.



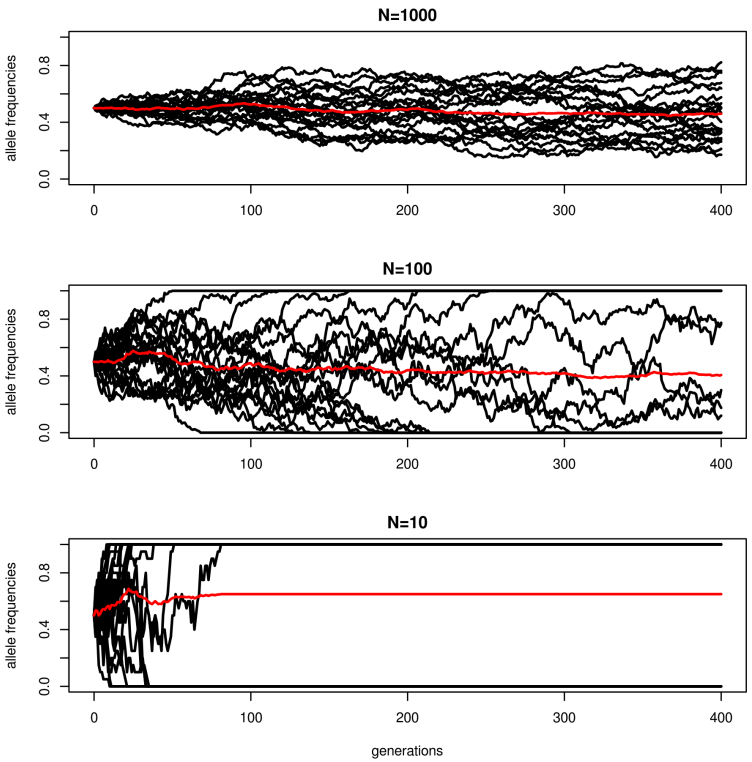
Test FLK (Bonhomme *et al*, 2010



- **Estime la phylogénie** moyenne sur génome.
  $\rightarrow$ **Loi de** p **sous** H_0.
- Pour chaque SNP, mesure l'**écart** des fréquences observées **à ce modèle nul** : FLK, extention de LK.

Dérive génétique dans une population

Population de taille constante N, un locus bi-allélique, alléle dérivé de fréquence p_0.



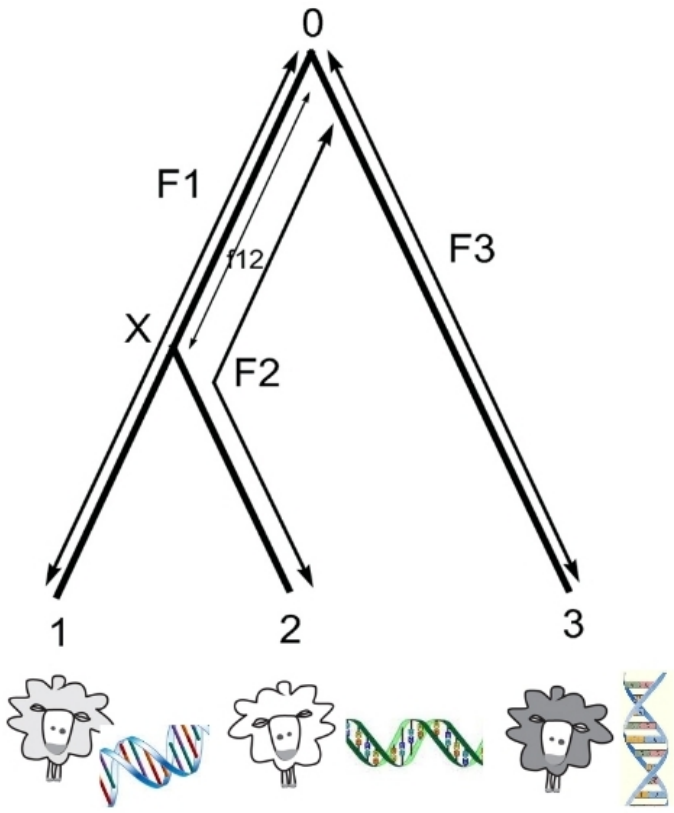
$$\mathbb{E}(p(t)) = p_0 \quad (1)$$

$$Var(p(t)) = F_t p_0(1-p_0) \quad (2)$$

Indice de fixation de Wright-Fisher

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t \approx \frac{t}{2N}$$

Extension à plusieurs populations sans migration



$$Var(p_i) = F_i p_0(1-p_0)$$
$$Cov(p_i, p_j) = f_{ij} p_0(1-p_0)$$

$$F_3 = 1 - \left(1 - \frac{1}{2N_3}\right)^t \approx \frac{t}{2N_3}$$
$$f_{12} = 1 - \left(1 - \frac{1}{2N_{12}}\right)^{t_{12}} \approx \frac{t_{12}}{2N_{12}}$$

matrice de kinship

$$F = \begin{pmatrix} F_1 & f_{12} & 0 \\ f_{12} & F_2 & 0 \\ 0 & 0 & F_3 \end{pmatrix}$$

$$\rightarrow Var(p) = F p_0(1-p_0)$$

Estimation de la matrice de kinship

- En utilisant un grand nombre de SNPs sur le génome, on calcule la **distance génétique de Reynolds** $\mathcal{D}_{ij}$ (Reynolds, Weir and Cockerham, 1983) entre toutes les paires de populations i and j.
- On construit un arbre de phylogénie en accord avec toutes ces distances (neighbour joining).
- On sait que

$$E(\mathcal{D}_{ij}) = \frac{F_i + F_j}{2}$$

Test FLK

Pour un SNP donné

- $\hat{p_0} = f(p, F)$
- $FLK = g(p - \hat{p_0}, F)$
- Sous H_0, FLK suit un χ^2 à $n - 1$ degrés de liberté $\rightarrow$ p-valeur.

Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

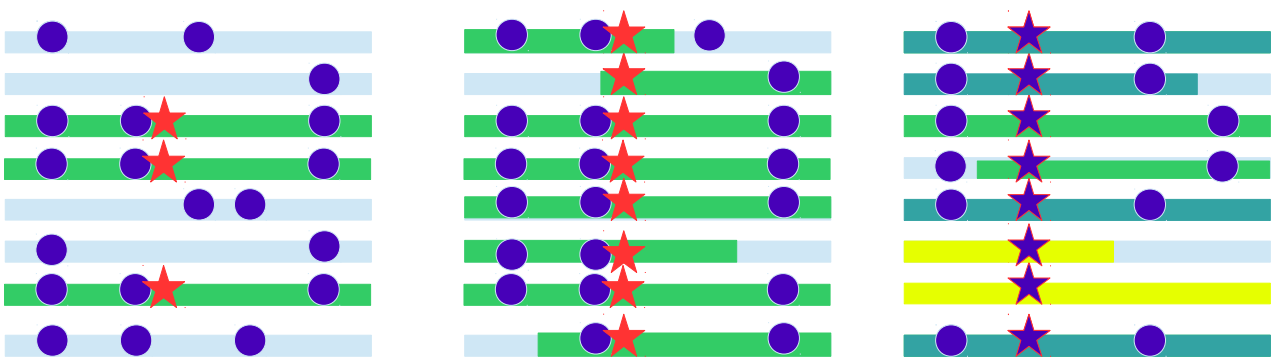
3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

5 TP

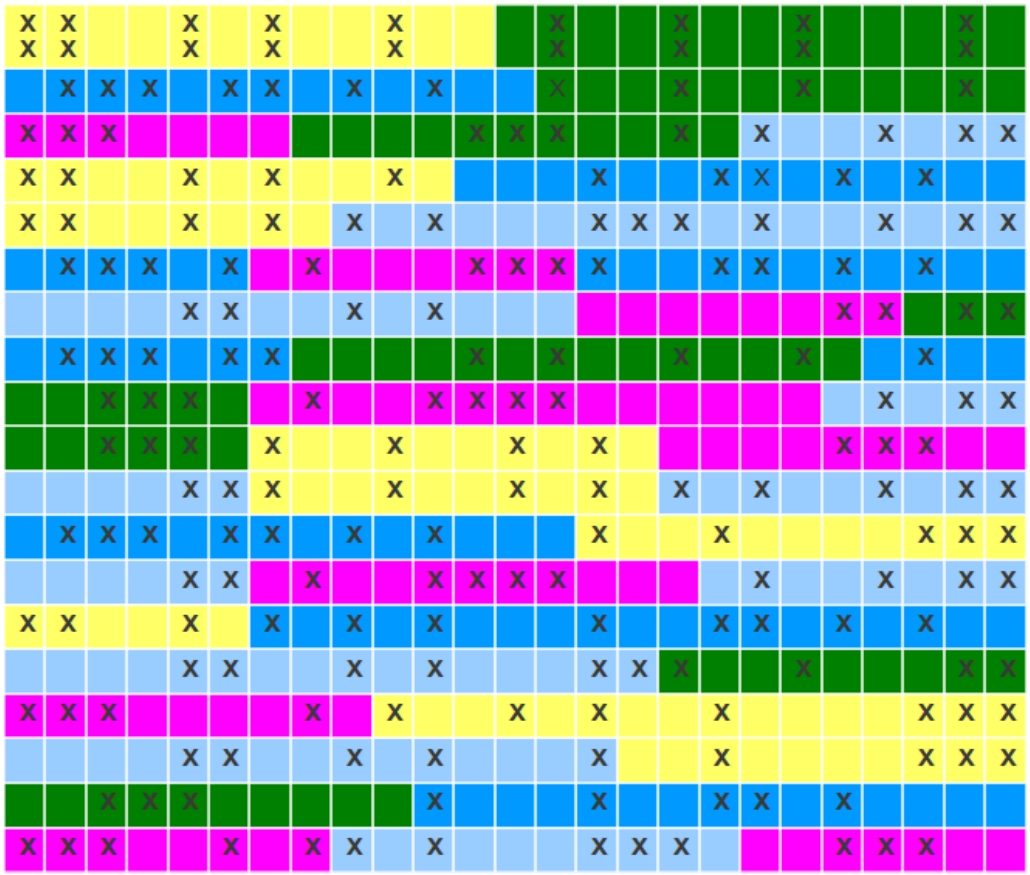
Principe

- Fréquences alléliques corrélées, fréquences haplotypiques affectées par la sélection.
- Appliquer FLK sur des haplotypes locaux.



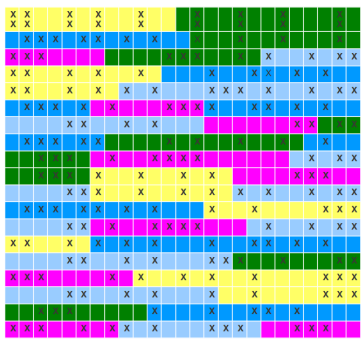
Modèle de Scheet and Stephens (2006)

Estime des haplotypes locaux autour de chaque SNP (Scheet and Stephens, 2006).



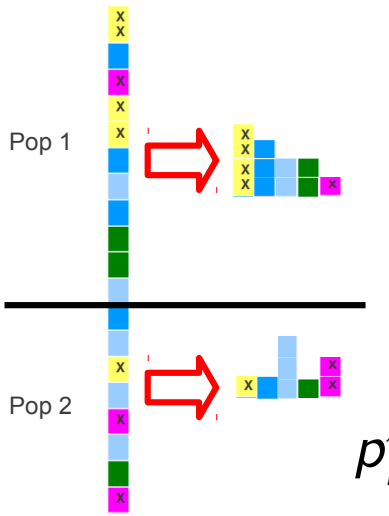
Test hapFLK (Fariello *et al*, 2013)

Estimation des clusters



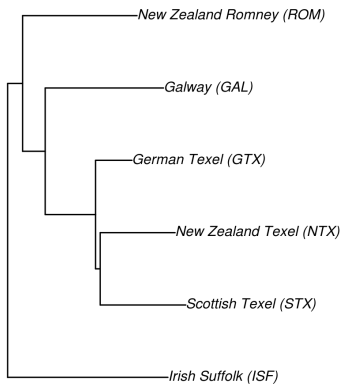
Algorithme EM

Fréquence des clusters pour chaque SNP ℓ et population j:



$$p_{kj}^{\ell} = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbb{P}(z_{ik}^{\ell} | \Theta)$$

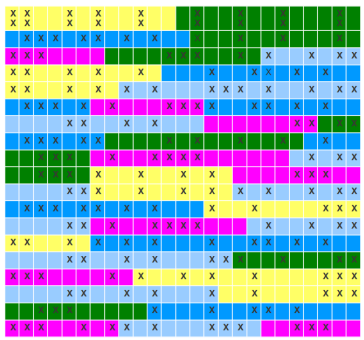
Calcul de FLK, en considérant les clusters comme des allèles.



Moyenne de FLK sur les différents runs $\rightarrow$ hapFLK

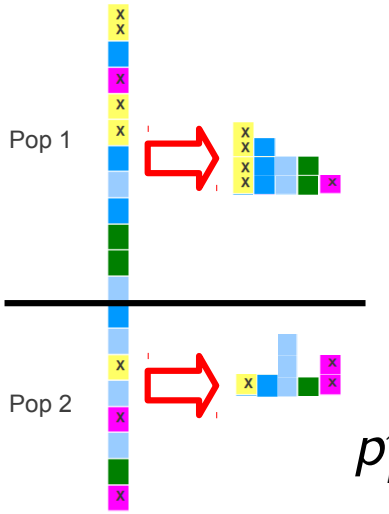
Test hapFLK (Fariello *et al*, 2013)

Estimation des clusters



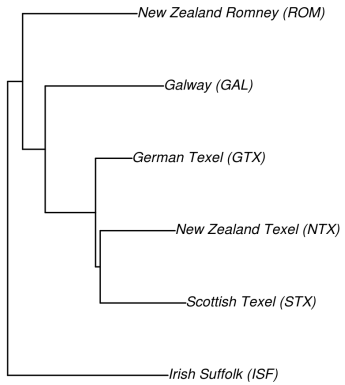
Algorithme EM

Fréquence des clusters pour chaque SNP ℓ et population j:



$$p_{kj}^{\ell} = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbb{P}(z_{ik}^{\ell} | \Theta)$$

Calcul de FLK, en considérant les clusters comme des allèles.



Moyenne de FLK sur les différents runs → hapFLK

Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

5 TP

Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

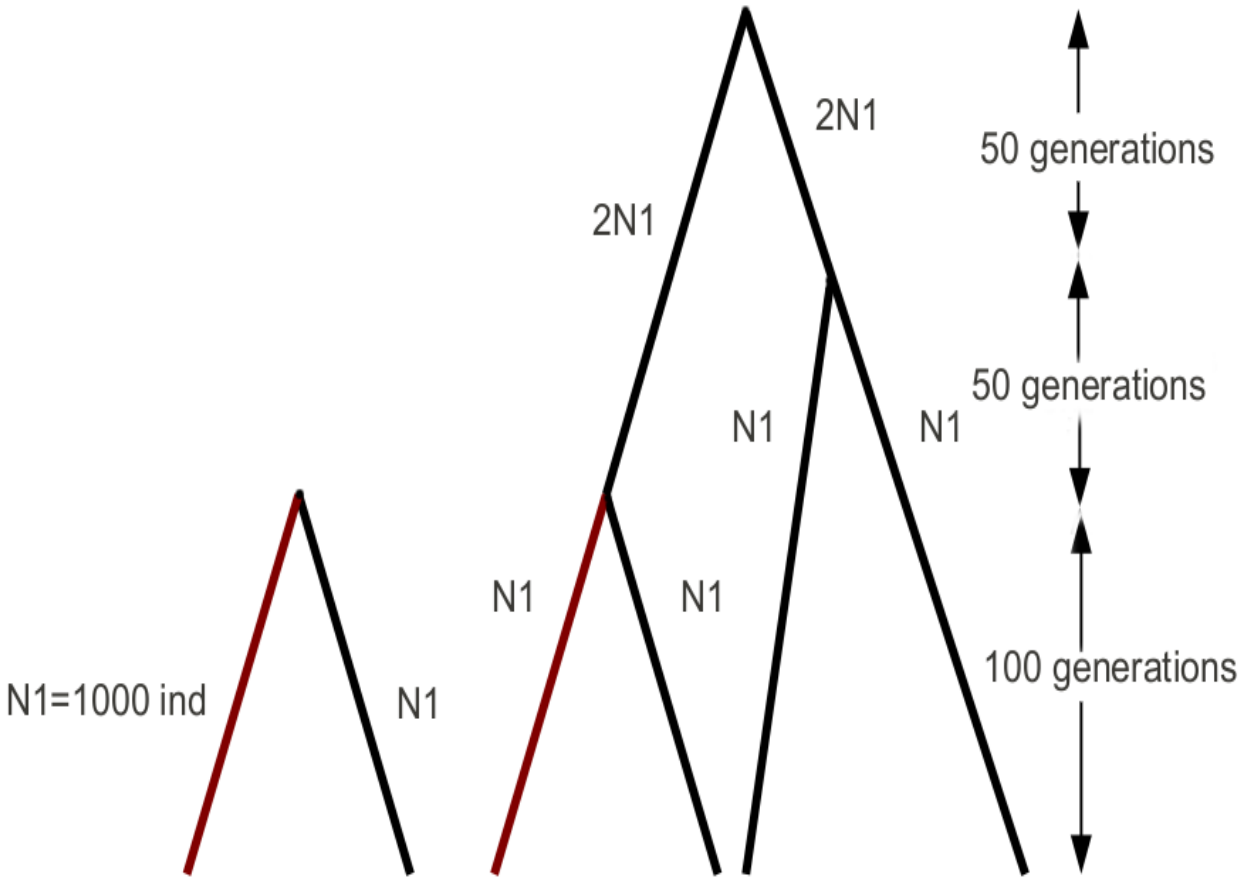
3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

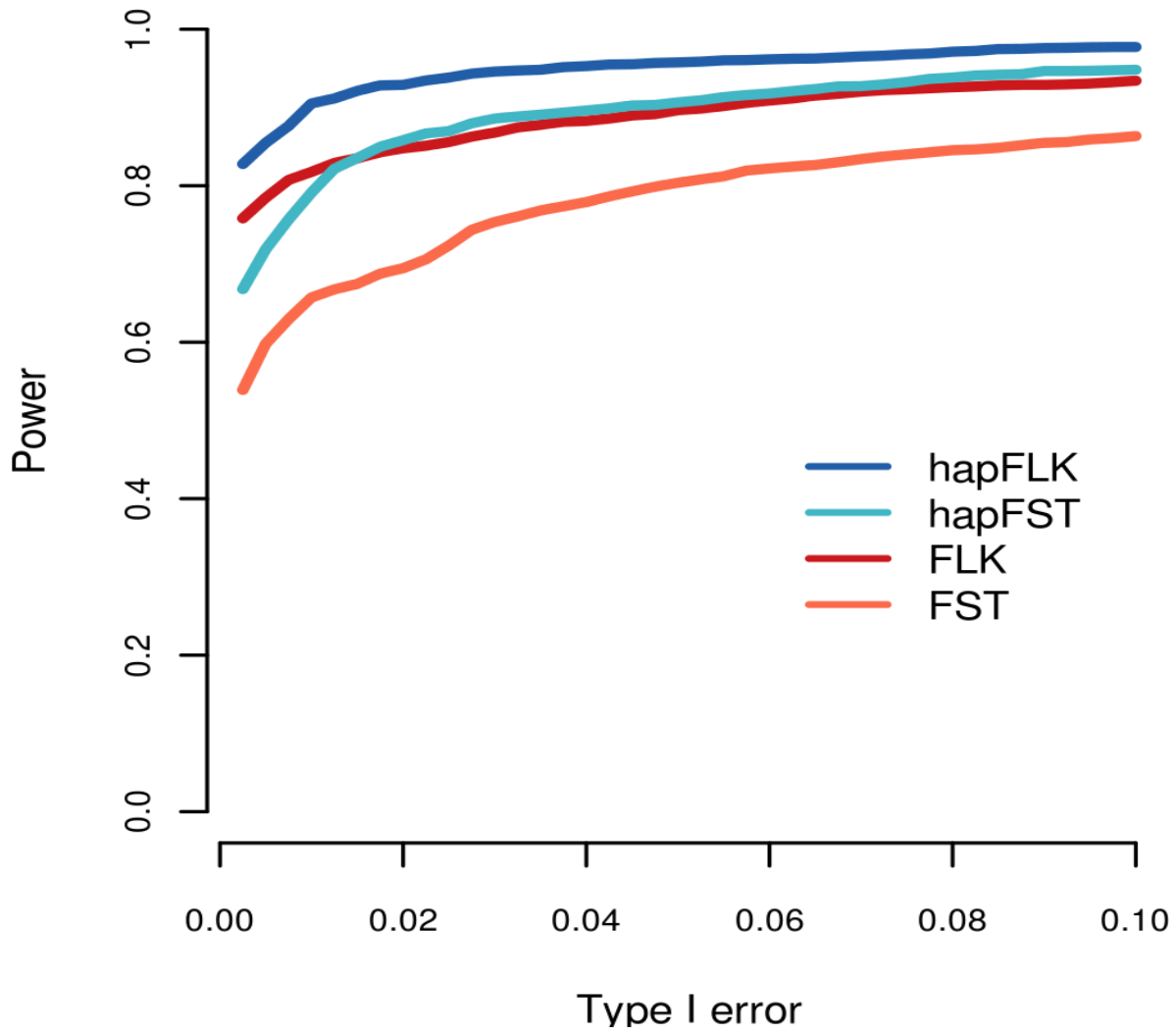
5 TP

Simulations

Locus de 5Mb avec 100 SNPs (génotypage haut débit).



Puissance de détection



Hard sweep, 4 populations.

Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

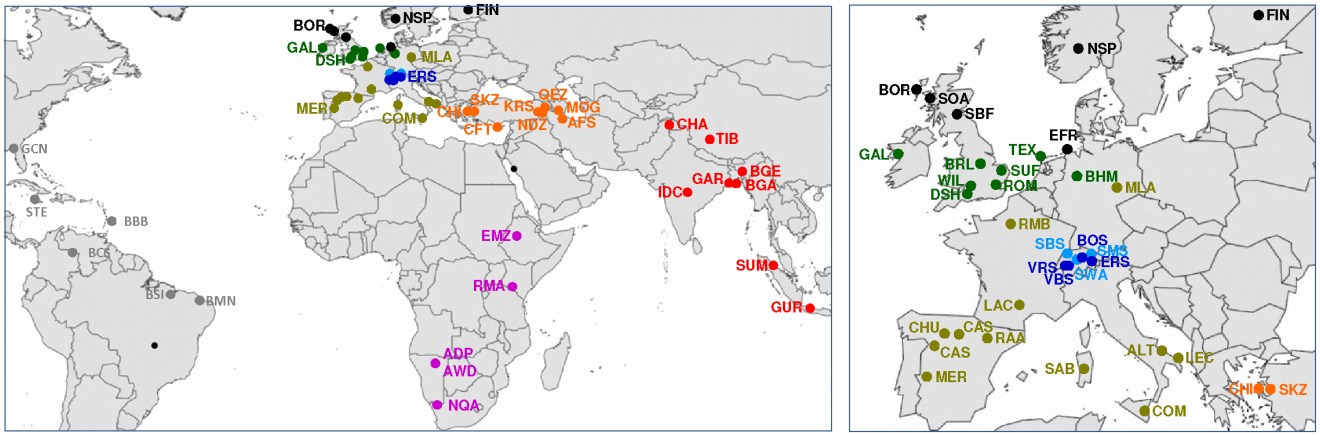
3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

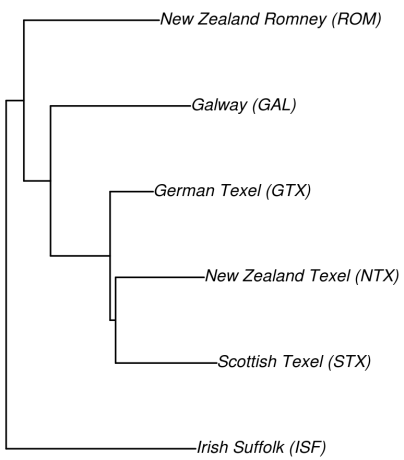
5 TP

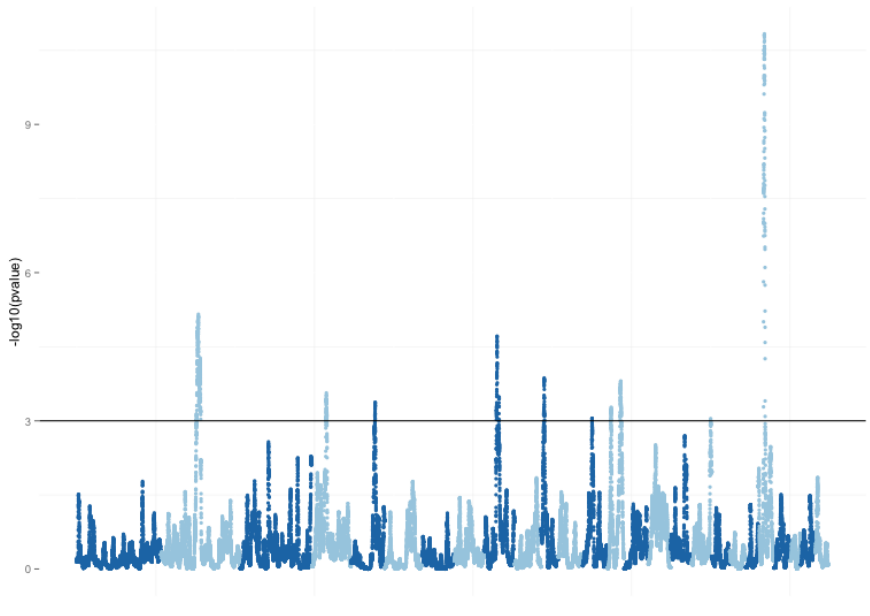
Projet Sheep HapMap

74 populations, 50K SNPs (Kijas *et al*, 2012)



Focus sur les populations d'Europe du Nord

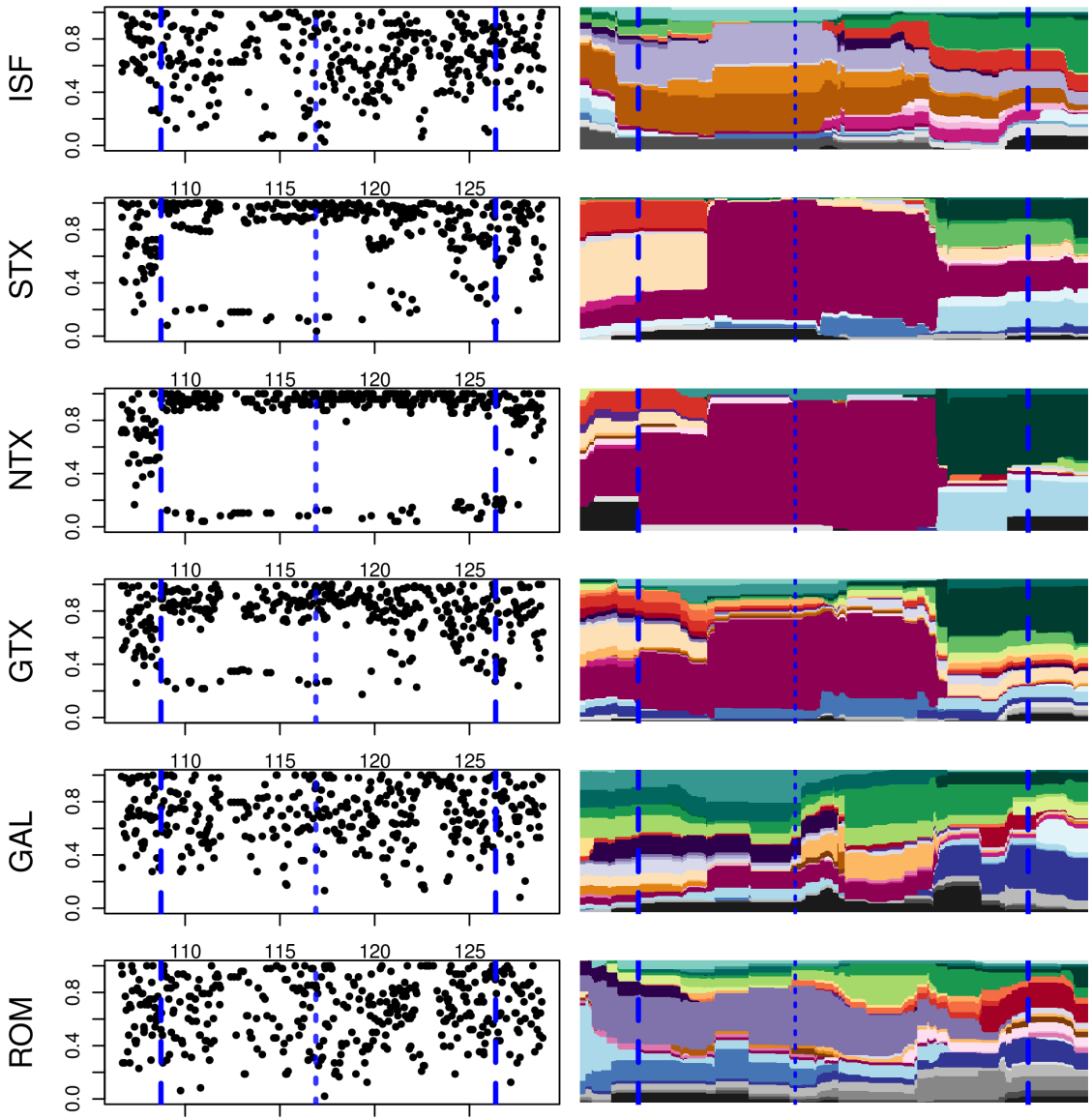


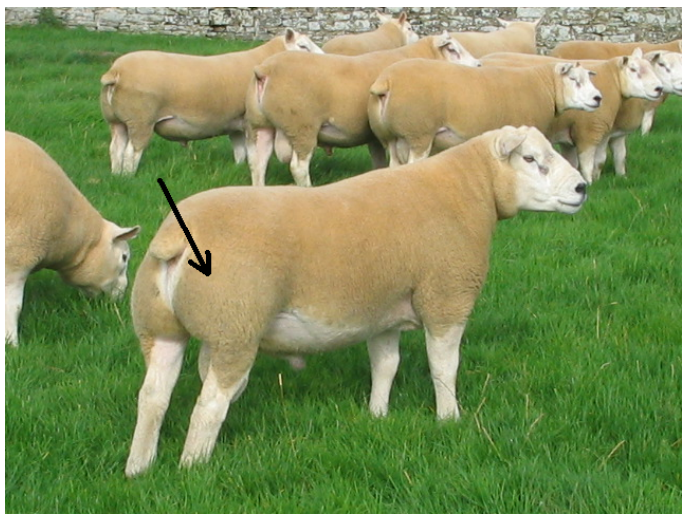


Hard sweep dans les races Texel

freq. allèliques

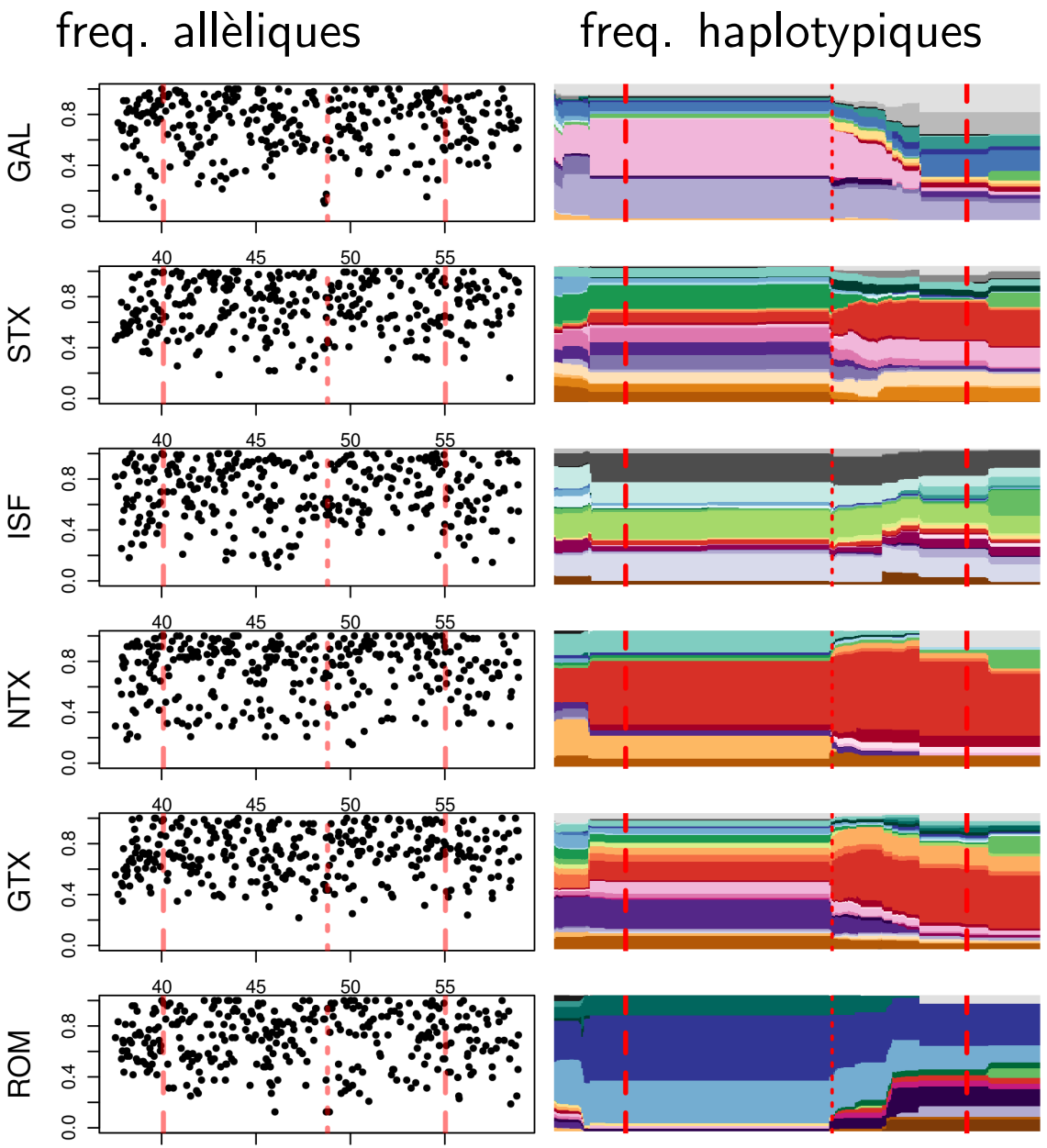
freq. haplotypiques





Mutation candidate dans MSTN

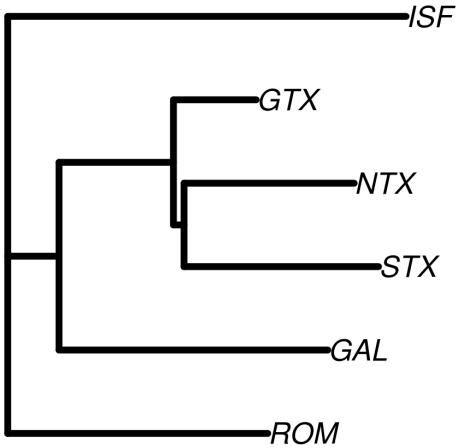
Soft sweep partiel en Nouvelle Zélande



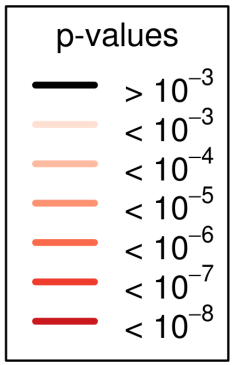
- Un cluster assez fréquent en New Zealand Texel (NTX)
- Deux clusters fixés en New Zealand Rommey (ROM).

Les races de Nouvelle Zélande sont sous selection

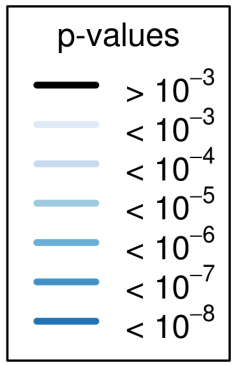
Whole genome tree



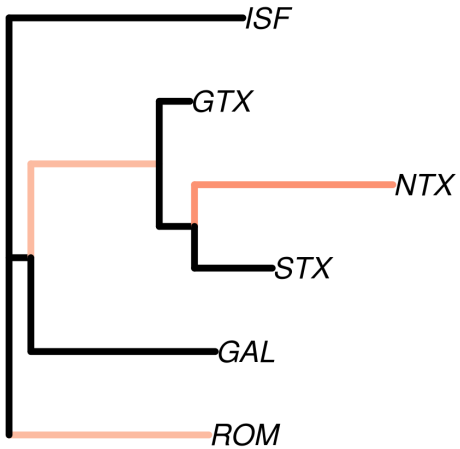
SNP trees

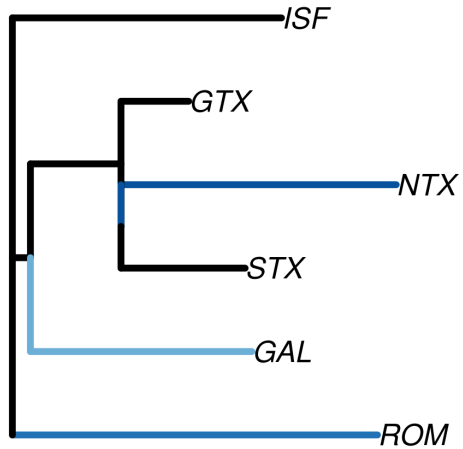


Haplotype trees



OAR 14





Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

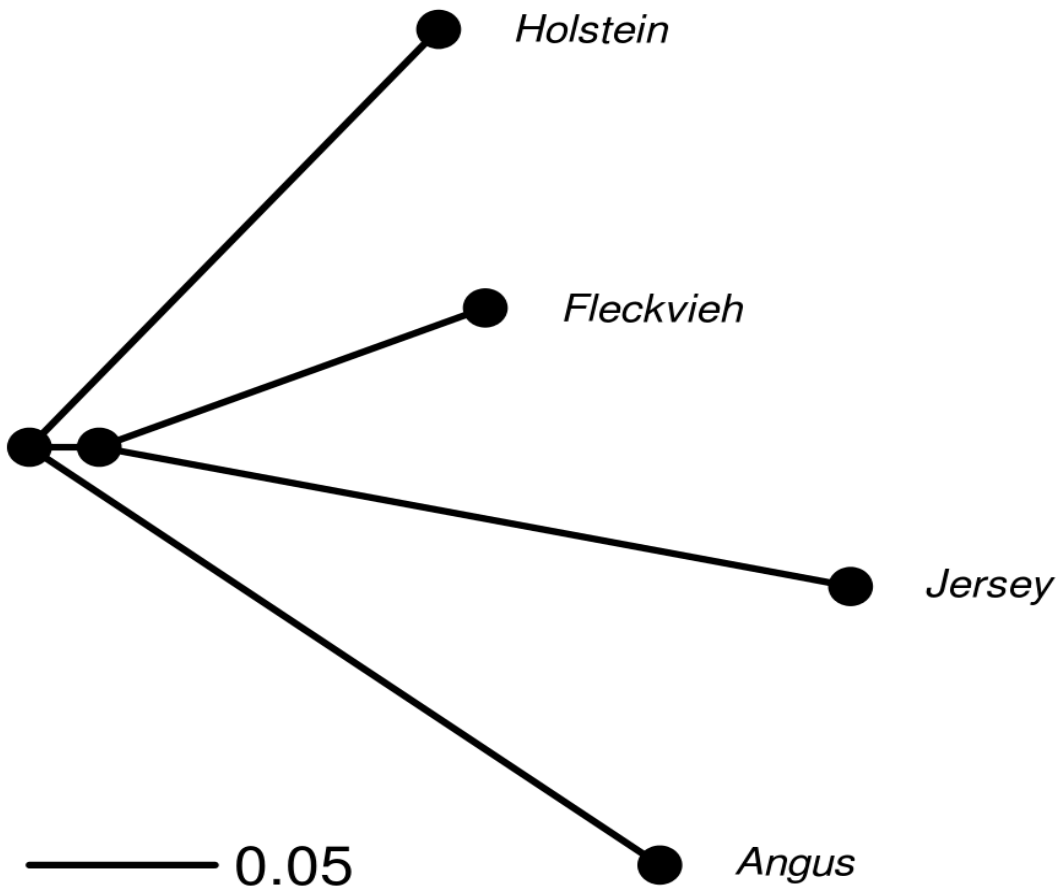
3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

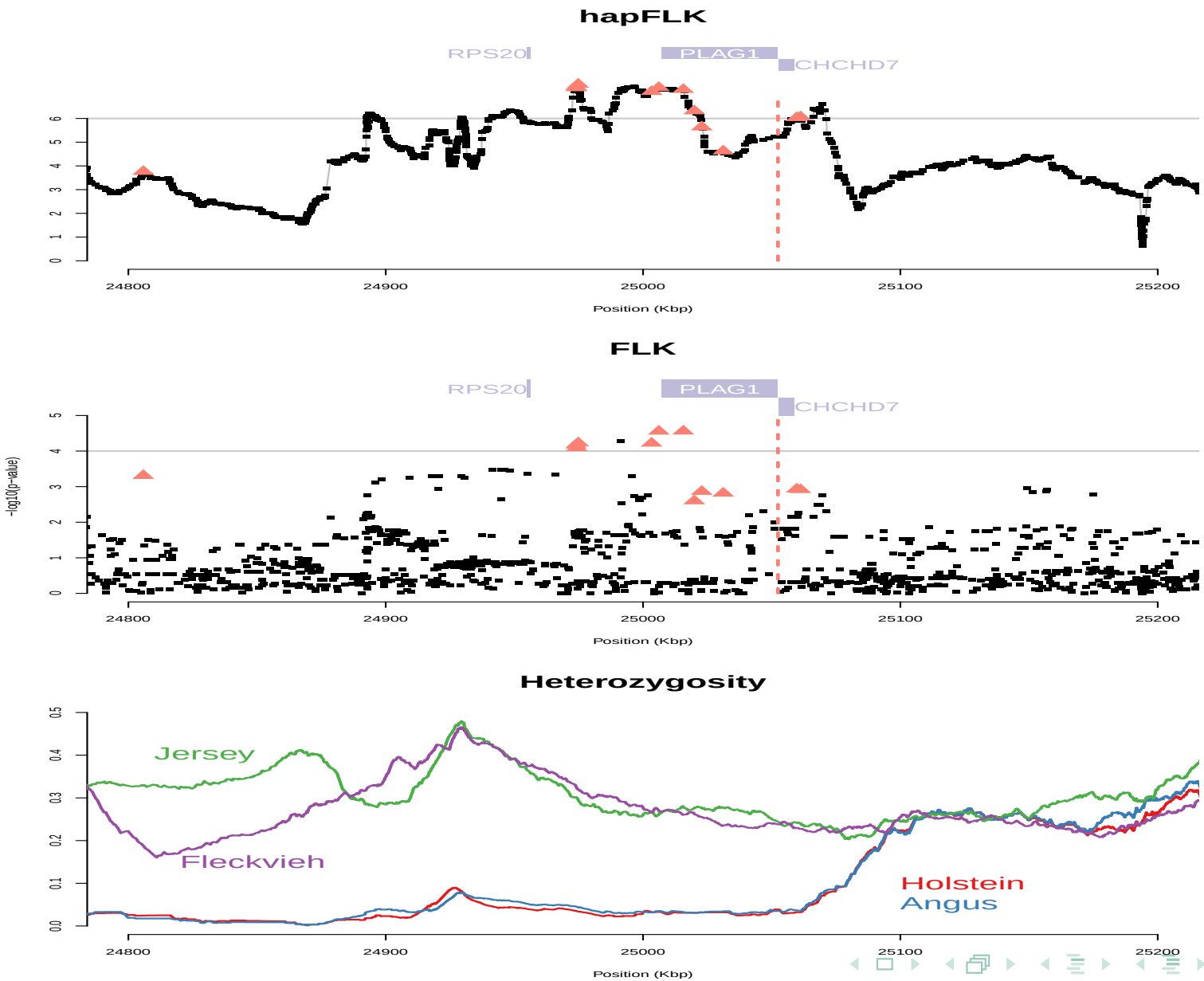
5 TP

Projet 1000 genomes bovins, run2

- 234 séquences issues de 4 races (90 utilisées).
- 29 millions de variants bi-allèliques (SNPs et indels)



Sélection autour du gène PLAG1 - variant candidat?



Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

5 TP

Conclusion

- L'approche hapFLK:
  - Détection de locus sous sélection positive par comparaison d'échantillons de plusieurs populations.
  - Tient compte des différences de tailles des populations et de leur structure hiérarchique.
  - Tient compte de la corrélation entre marqueurs (haplotypes).

  $\rightarrow$ Puissance de détection accrue.
- Autres avantages:
  - Nombre de populations arbitraire.
  - Données non phasées et éventuellement manquantes.
  - Plusieurs types de sélection détectés.
- Limitations:
  - Modèle de dérive pure (pas de migration).

Références

- Méthodes:
  - M. Bonhomme, C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, M. San Cristobal (2010). Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186: 241-26.
  - M. I. Fariello, S. Boitard, H. Naya, M. SanCristobal, B. Servin (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193: 929-941.
- Applications:
  - M.I. Fariello, B. Servin, G. Tosser-Klopp, R. Rupp, C. Moreno, International Sheep Genome Consortium, M. San Cristobal and S. Boitard (2014). Selection Signatures in Worldwide Sheep Populations. *PLoS ONE* 9(8), e103813.
  - P.F. Roux, S. Boitard, Y. Blum et al (2015). Combined QTL and Selective Sweep Mappings with Coding SNP Annotation and Cis-eQTL Analysis Revealed PARK2 and JAG2 as New Candidate Genes for Adiposity Regulation. *G3* 5(4) 517-529.
  - S. Boitard, M. Boussaha, A. Capitan, D. Rocha and B. Servin (2016). Uncovering Adaptation from Sequence Data: Lessons from Genome Resequencing of Four Cattle Breeds. *Genetics* 203: 433-450.

Plan de l'exposé

1 Effets de la sélection positive sur le génome

2 Tests basés sur la différenciation entre populations
- Approche SNP par SNP
- Approche haplotypique

3 Résultats
- Simulations
- Application chez le mouton
- Application chez la vache

4 Conclusions et Perspectives

5 TP

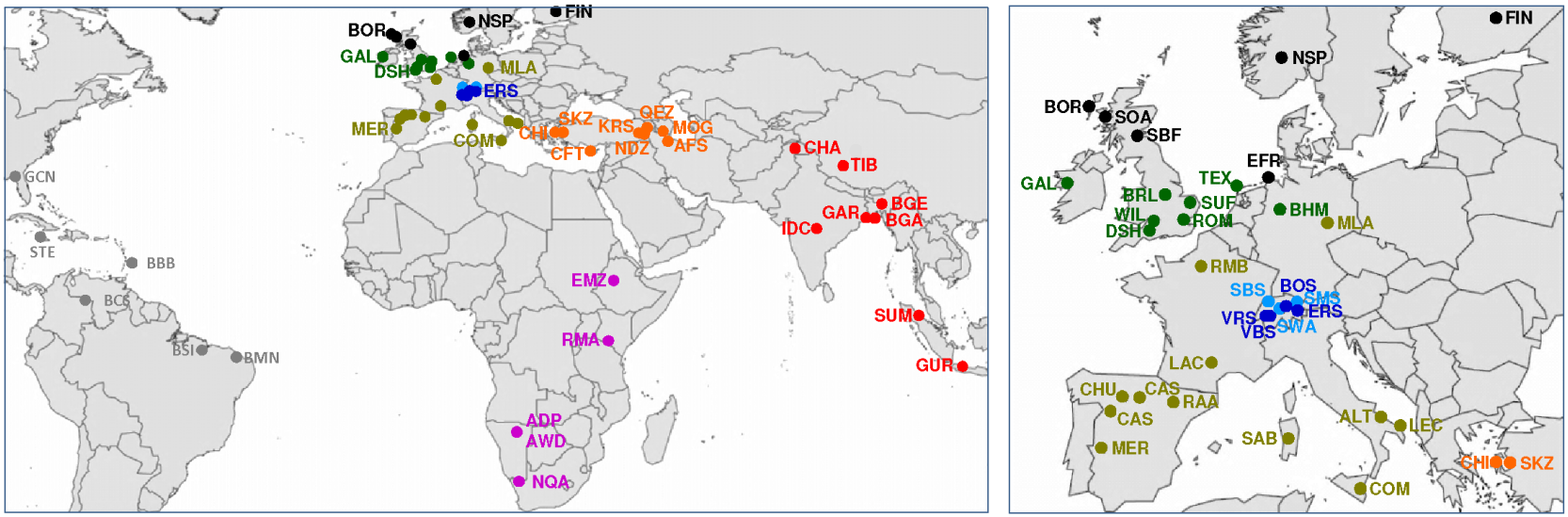
Logiciel `hapflk` software

- Construit une phylogénie des populations (libraries `R`)
- Calcule les statistiques FLK and hapFLK.
- Calcule les p-valeurs associées à ces statistiques.
- Trace des courbes aidant l'interprétation : fréquences des clusters, arbres locaux.

Information technique

- Adresse: `https://forge-dga.jouy.inra.fr/projects/hapflk`
- Code source `python`, besoin de `numpy`, `scipy` et d'un compilateur `C`.
- Scripts additionels en `R`, besoin de `ape`, `phangorn` et `ggplot2`.

Jeu de données: projet Sheep HapMap, Europe du Nord



Kijas *et al.* (2012) PLoS Biology

6 Populations + Outgroup (Soay), 388 animaux, 50K SNPs.

Disponible ici : `http://www.sheephapmap.org`

Avant de commencer...

- Se rappeler des hypothèses du modèle :
  - Dérive pure (pas de mutations depuis divergence, pas de migration).
  - F_i petit (< 0.2), pour le calcul des p-valeurs.
- Il faut donc
  - Enlever les populations admixées ou très consanguines.
  - Enlever les variants rares (à l'échelle de la méta-population), susceptibles d'être apparus récemment.
- Réaliser d'abord une étude de diversité :
  - PCA, clustering (LEA) pour enlever les individus outliers.
  - Enlever les individus trop apparentés.

Les étapes de l'analyse

1. Calculer la matrice de kinship
   Calculer FLK pour tous les SNPs.
2. Calculer hapFLK pour tous les SNPs.
3. Calculer les p-valeurs de hapFLK pour tous les SNPs.
4. Tracer les p-valeurs et détecter les régions significatives (FLK et hapFLK).
5. Etudier plus en détail les régions significatives :
   - Fréquence des clusters.
   - Arbres locaux.