Detecting selective sweeps from pooled next generation sequencing (NGS) samples

Simon S. Boitard, Christian Schloetterer, Viola Nolte, Ram Vinay Pandey, Andreas Futschik

▶ **To cite this version:**

Simon S. Boitard, Christian Schloetterer, Viola Nolte, Ram Vinay Pandey, Andreas Futschik. Detecting selective sweeps from pooled next generation sequencing (NGS) samples. theoretical and empirical advances in evolutionary genomics, Mar 2012, Roscoff, France. hal-02807026

HAL Id: hal-02807026
https://hal.inrae.fr/hal-02807026

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détection de gènes sous sélection dans une population à partir d'individus séquencés en pool

Simon Boitard, Andreas Futschik[†], Viola Nolte[*], Ram Vinay Pandey[*], Christian Schlötterer[*]

INRA, LGC
[†] Institut für Statistik und Operations Research, Universität Wien, Austria
[*] Institut für Populationsgenetik, Veterinärmedizinische Universität Wien, Austria

Introduction

- Une majorité de locus neutres dans le génome, quelques uns sous sélection (naturelle ou artificielle).
- Détection des loci sous sélection est un enjeu théorique (comprendre évolution, adaptation des espèces . . .) et appliqué (zones du génome d'intérêt médical, agronomique . . .).
- Scans génomiques possibles à partir de données de type puces SNP haute densité ou NGS.
- Une approche possible : recherche des zones de faible diversité génétique intra population.

Plan de l'exposé

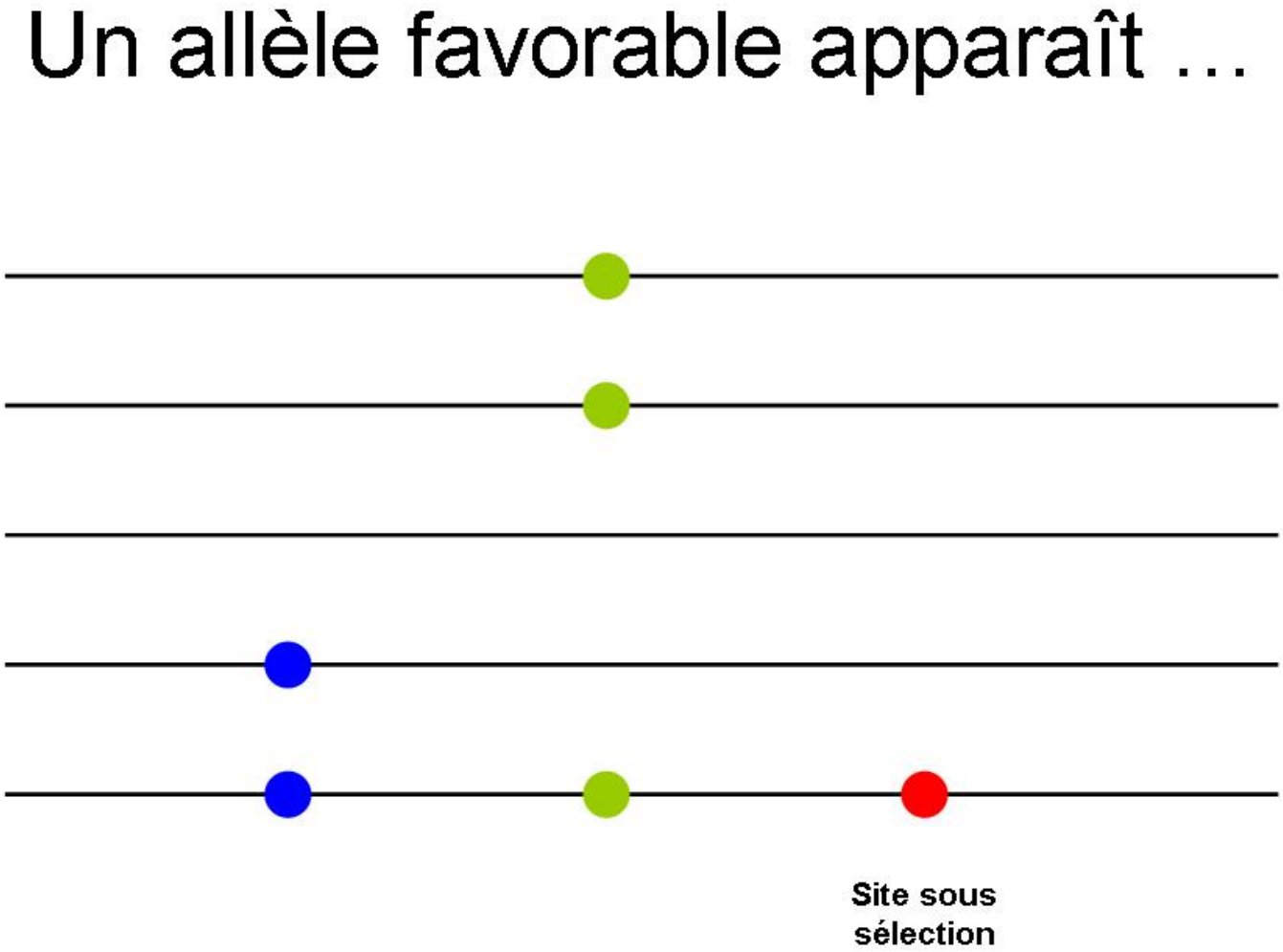
1. Détection de zones de faible diversité génétique intra population

2. Utilisation d'individus séquencés en pool
   - Méthode
   - Résultats de simulation
   - Résultats sur données réelles

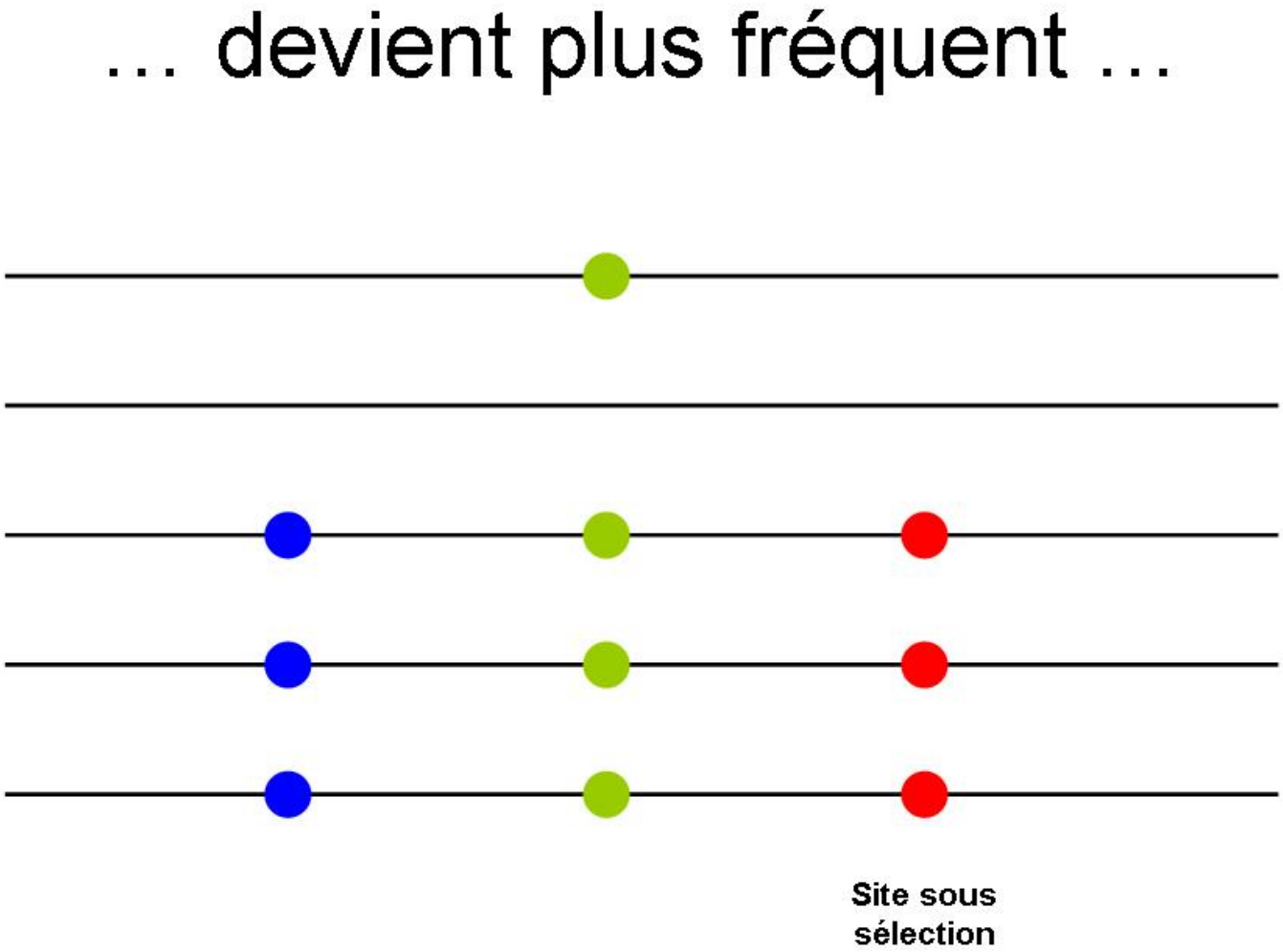
1. Détection de zones de faible diversité génétique intra population

2. Utilisation d'individus séquencés en pool
   - Méthode
   - Résultats de simulation
   - Résultats sur données réelles

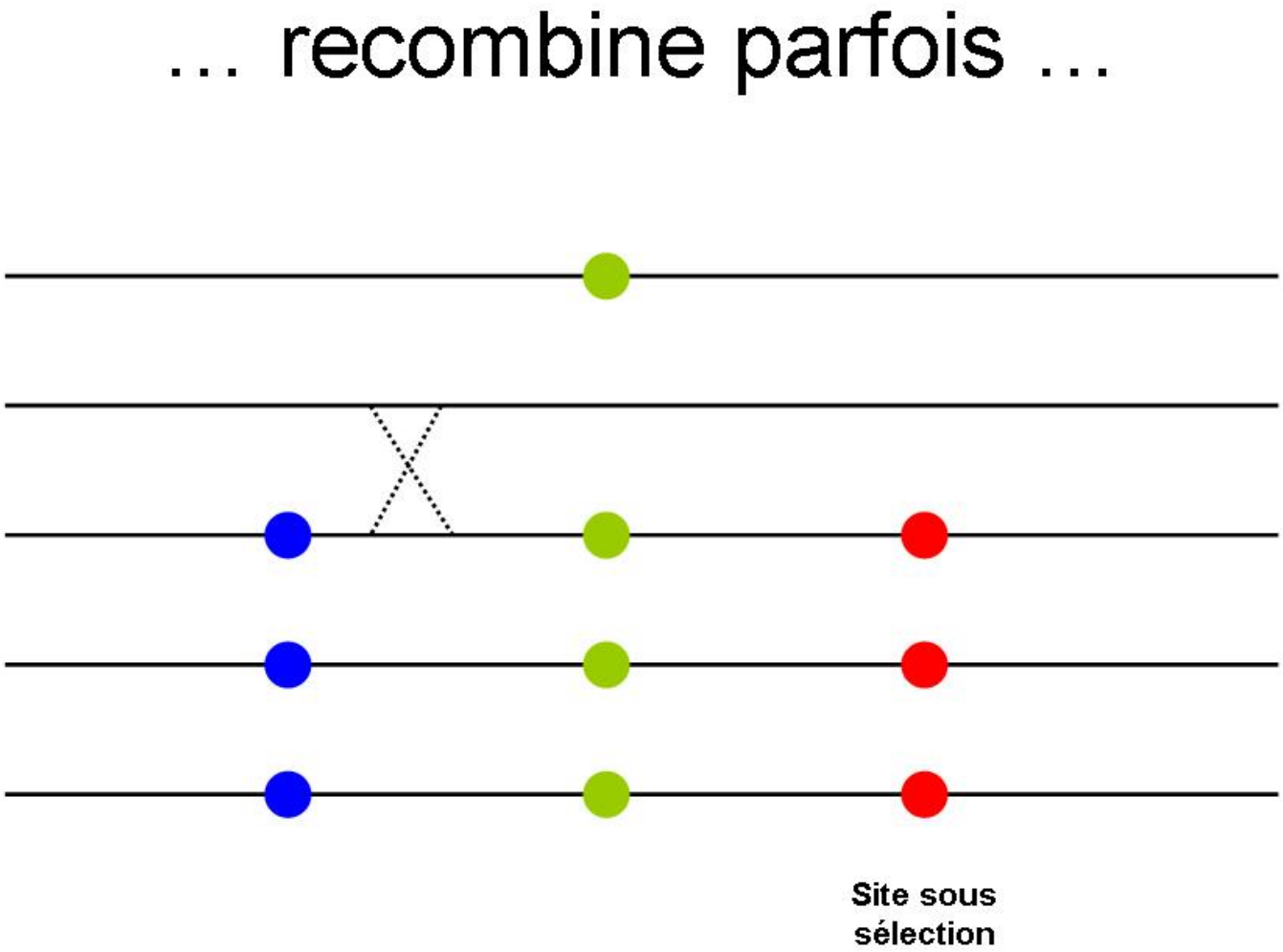
Balayage sélectif



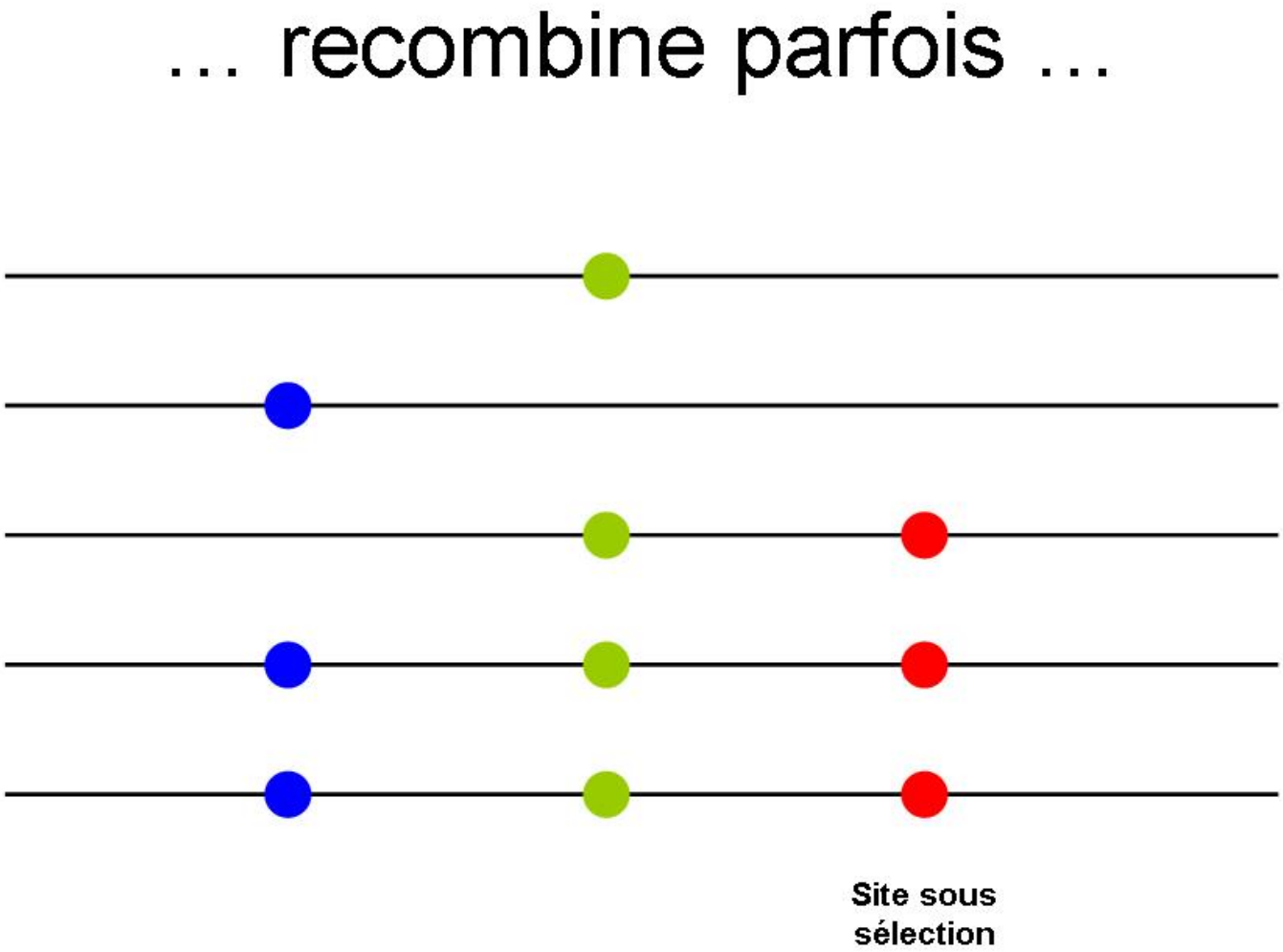
Balayage sélectif



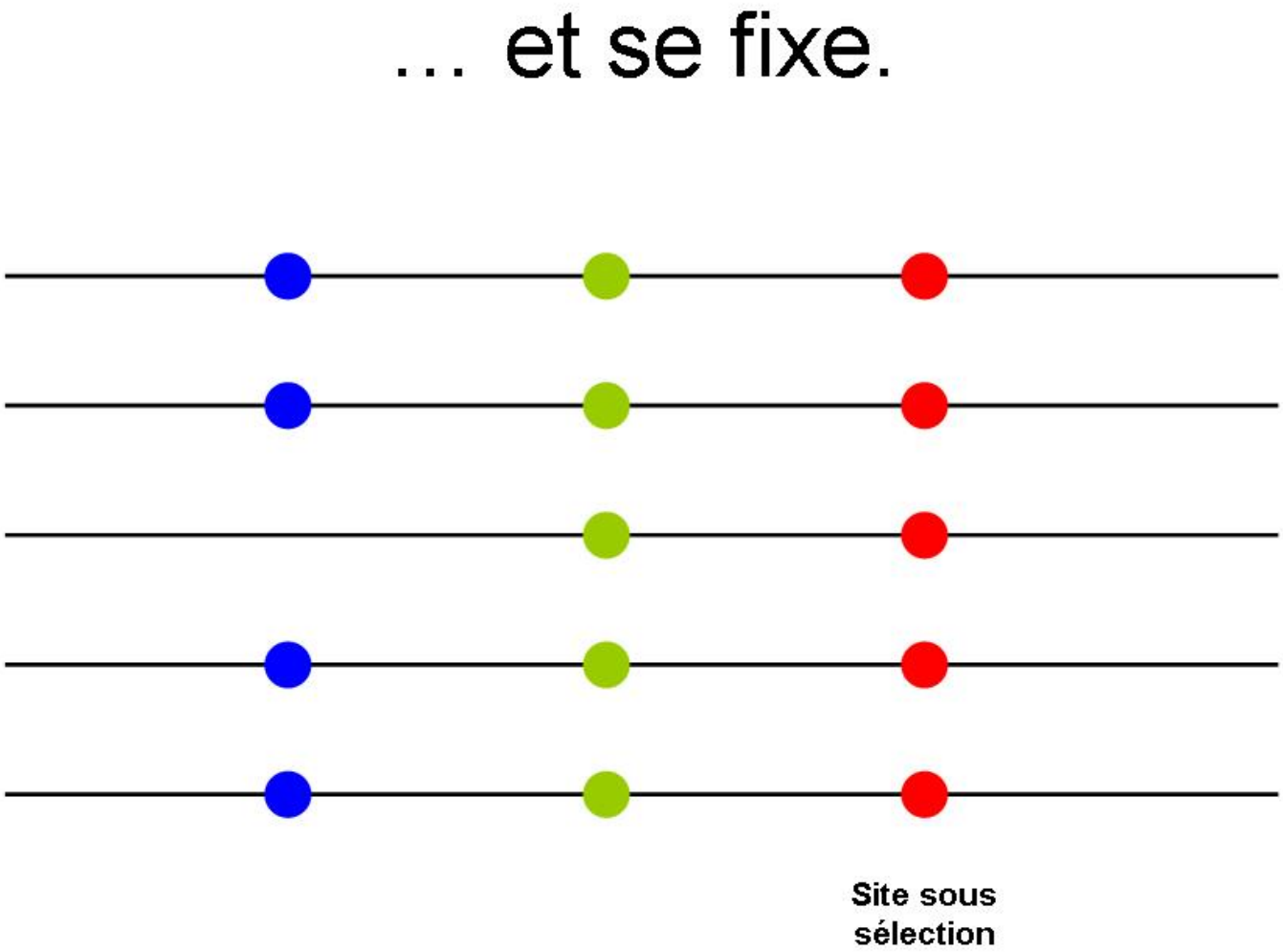
Balayage sélectif



Balayage sélectif



Balayage sélectif



Données SNP

- Echantillon de n haplotypes. 0 = ancestral, 1 = muté.

$$0-0-0-0-1-0-0-0-0-0-0-\ldots$$
$$0-0-0-0-1-0-0-0-1-0-0-\ldots$$
$$0-0-0-0-0-0-0-0-0-0-0-\ldots$$

- y_i nombre d'allèles 1 au site i (valeurs entre 0 et n).

$$0-0-0-0-2-0-0-0-1-0-0-\ldots$$

- Distribution de Y_i dépend du modèle d'évolution
  $\rightarrow$ un moyen de détecter des évènements de sélection.

Effet d'un balayage sélectif sur la distribution de Y_i

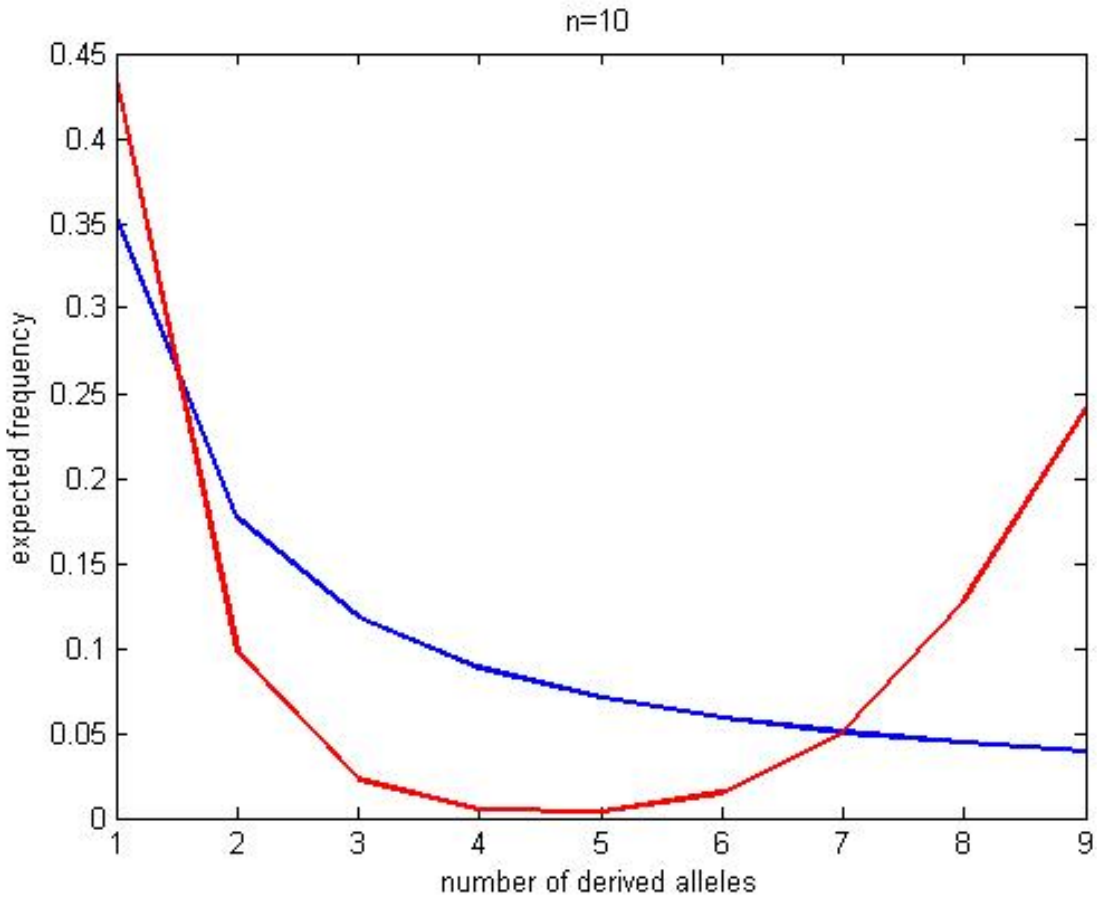


Figure: Distribution du nombre d'allèles mutés (Y_i) à un site neutre (i) loin d'un site sélectionné (courbe bleue) (ii) proche d'un site sélectionné (courbe rouge)

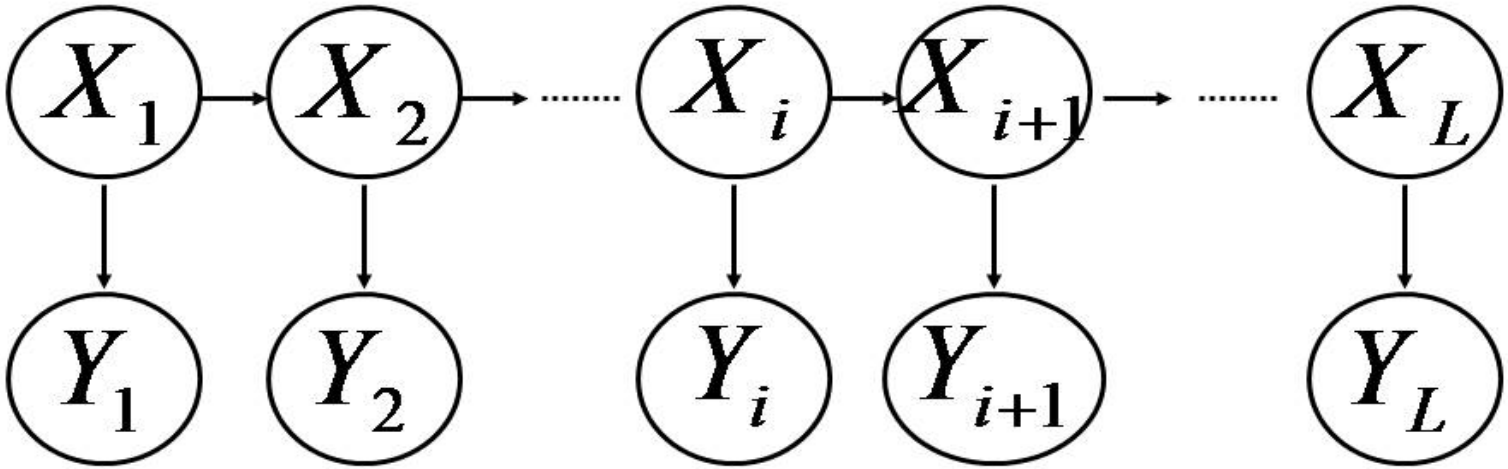
Influence de la démographie

L'histoire démographique de la population a aussi un effet sur la distribution des fréquence alléliques.
Pour compenser cet effet :

1. Estimer la distribution de Y_i à l'aide de données tout génome.
   $\approx$ distribution sous neutralité.
2. Calculer la distribution sous sélection à partir de la distribution neutre (Nielsen *et al* 2005).
3. Identifier les régions où la distribution locale des fréquences ressemble à la distribution sous sélection.

Modèle de Chaîne de Markov cachée (Boitard *et al*, 2009)

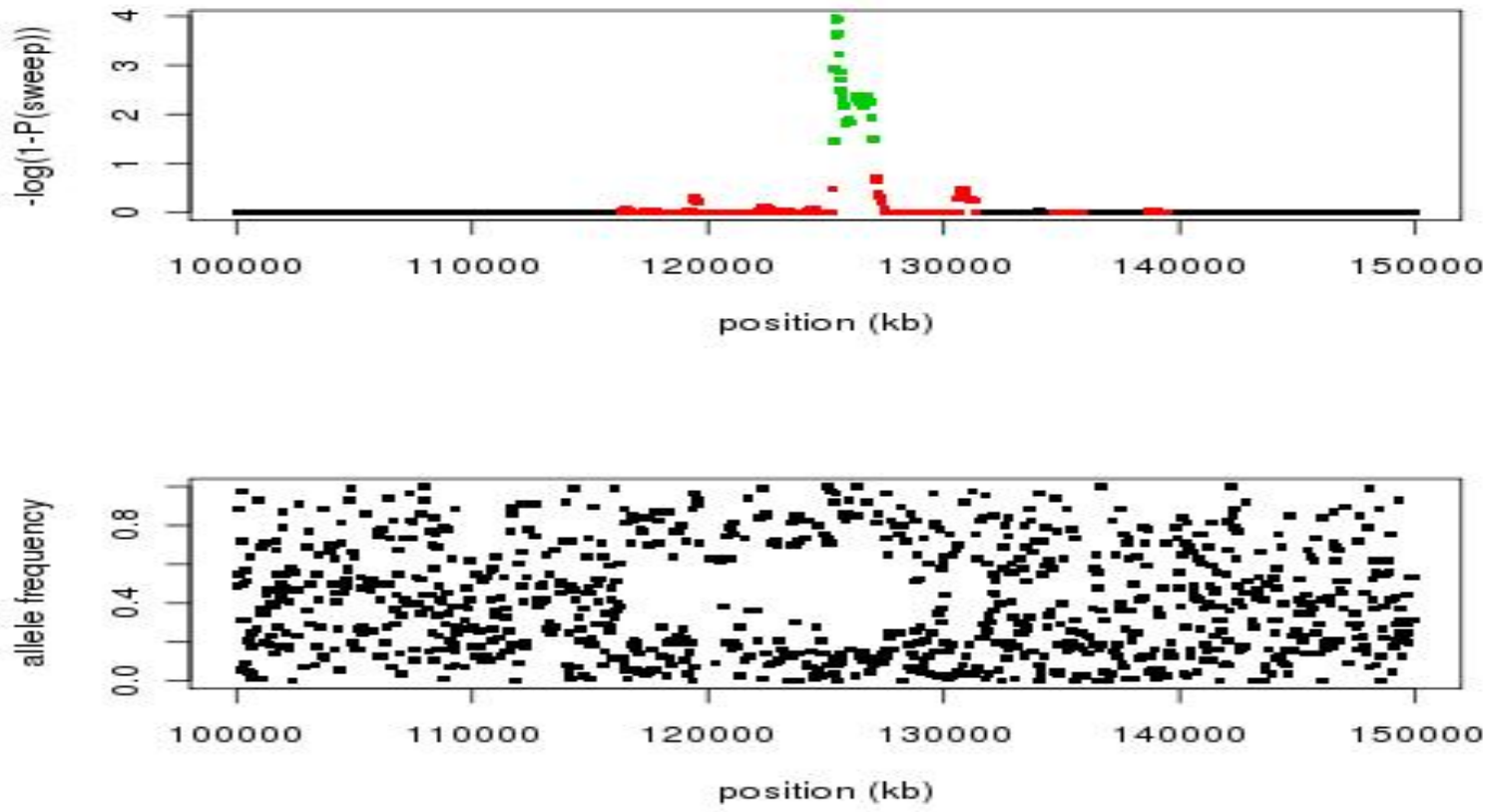
Hidden states : "Neutral", "Intermediate" or "Sweep"



Observed states : number of copies of allele 1

Objective : predict the most likely hidden states *X given the observed states Y*

Exemple : sweep mouton Texel OAR2



- Sweep autour de GDF8 (myostatin), régulateur de la masse musculaire. Mutation connue.

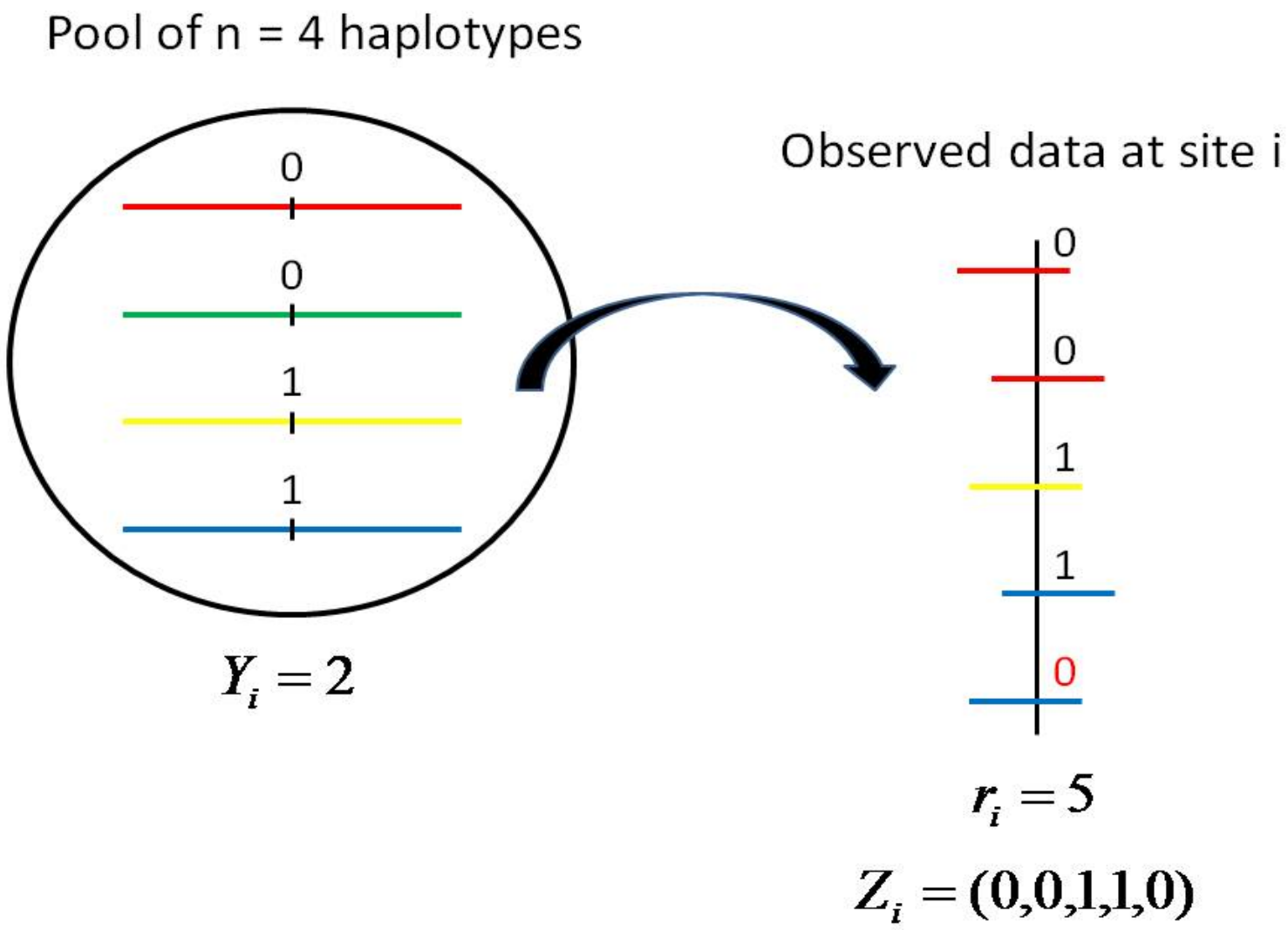
1. Détection de zones de faible diversité génétique intra population

2. Utilisation d'individus séquencés en pool
   - Méthode
   - Résultats de simulation
   - Résultats sur données réelles

Principe

- La méthode de détection est basée seulement sur les fréquences alléliques (pas les haplotype), donc pooler les individus ne pose pas de problème a priori.
- Pour l'estimation des fréquences alléliques, séquençage en pool est plus efficace (moins cher pour même précision) que séquençage individuel (Futschik et Schlötterer, 2010).

Protocole expérimental



Méthode

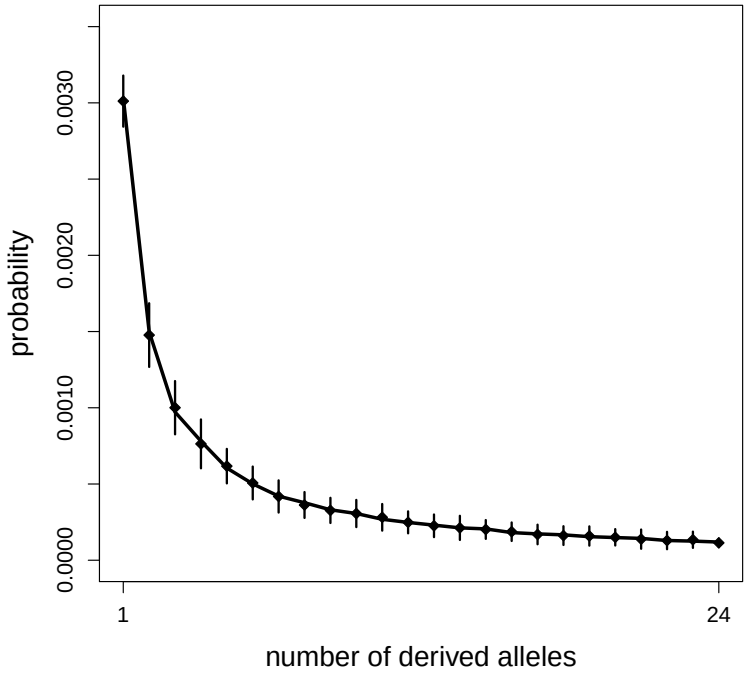
- On n'observe plus la fréquence allélique Y_i, mais des données Z_i découlant (de manière aléatoire) de Y_i.
- On commence par calculer $\mathbb{P}(Z_i \mid Y_i)$ pour $0 \leq Y_i \leq n$.
- On intègre cette information dans calculs précédents (estimation de la distribution des Y_i et recherche des traces de sélection).
- Ainsi, on ne jette aucune donnée mais les positions i avec une couverture plus grande, ou les allèles lus avec précision supérieure, ont automatiquement plus de poids dans les analyses.

Vraisemblance

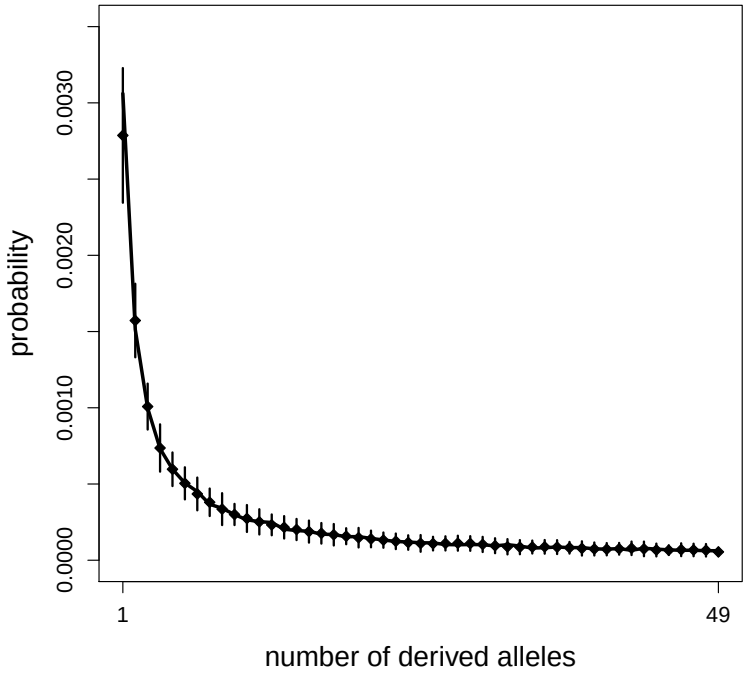
$$\begin{aligned}\mathbb{P}(Z_i \mid Y_i) &= \prod_{j:Z_{i,j}=1} ((1-p_{i,j})\frac{Y_i}{n} + p_{i,j}(1-\frac{Y_i}{n})) \\ &* \prod_{j:Z_{i,j}=0} ((1-p_{i,j})(1-\frac{Y_i}{n}) + p_{i,j}\frac{Y_i}{n})\end{aligned}$$

- Y_i nombre de copies de l'allèle 1.
- Z_i vecteur des allèles observés.
- $p_{i,j}$ probabilité d'erreur de séquençage (déduite du PHRED score).

Estimation de la distribution de Y_i



$n = 25$



$n = 50$

100 échantillons simulés de taille $L = 100kb$, couverture globale 100X.

Puissance de détection de la sélection

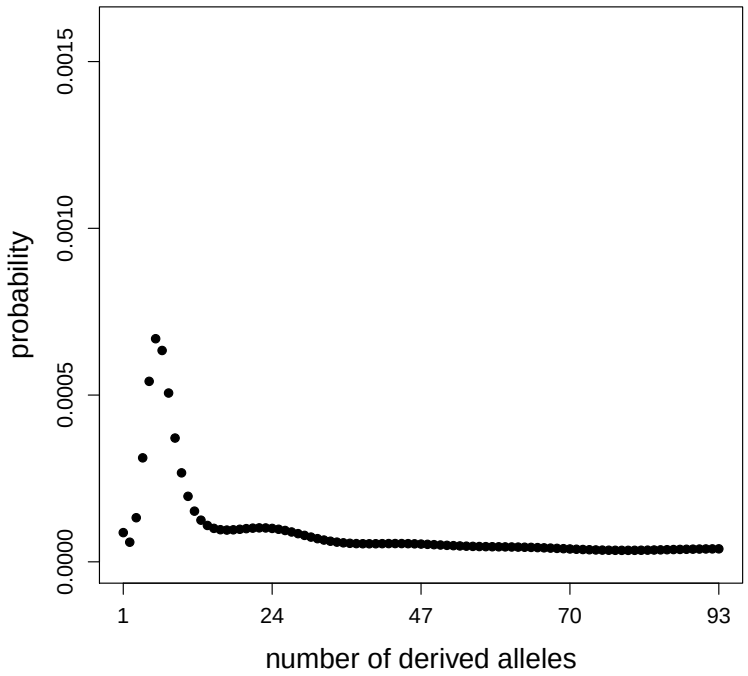
Taille d'échantillon	$n = 25$	$n = 50$	$n = 100$	$n = 200$
NGS en pool	0.91	0.90	0.91	0.90
Séquence complète	0.89	0.88	0.87	0.87

500 échantillons simulés avec un gène sous sélection, $L = 100kb$, couverture globale 100X.

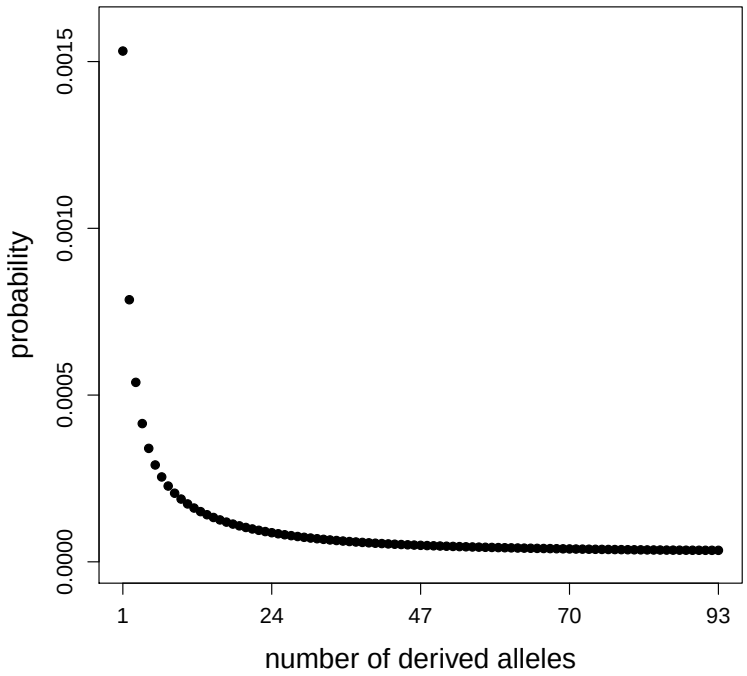
Application chez Drosophile (melanogaster)

- 2 échantillons de 97 femelles ($n = 194$), issues d'une même population naturelle prélevée près de Vienne.
- Séquençage en pool, couverture 100X et 87X.
- Etude du chromosome X.

Distribution des fréquences



Toutes les bases

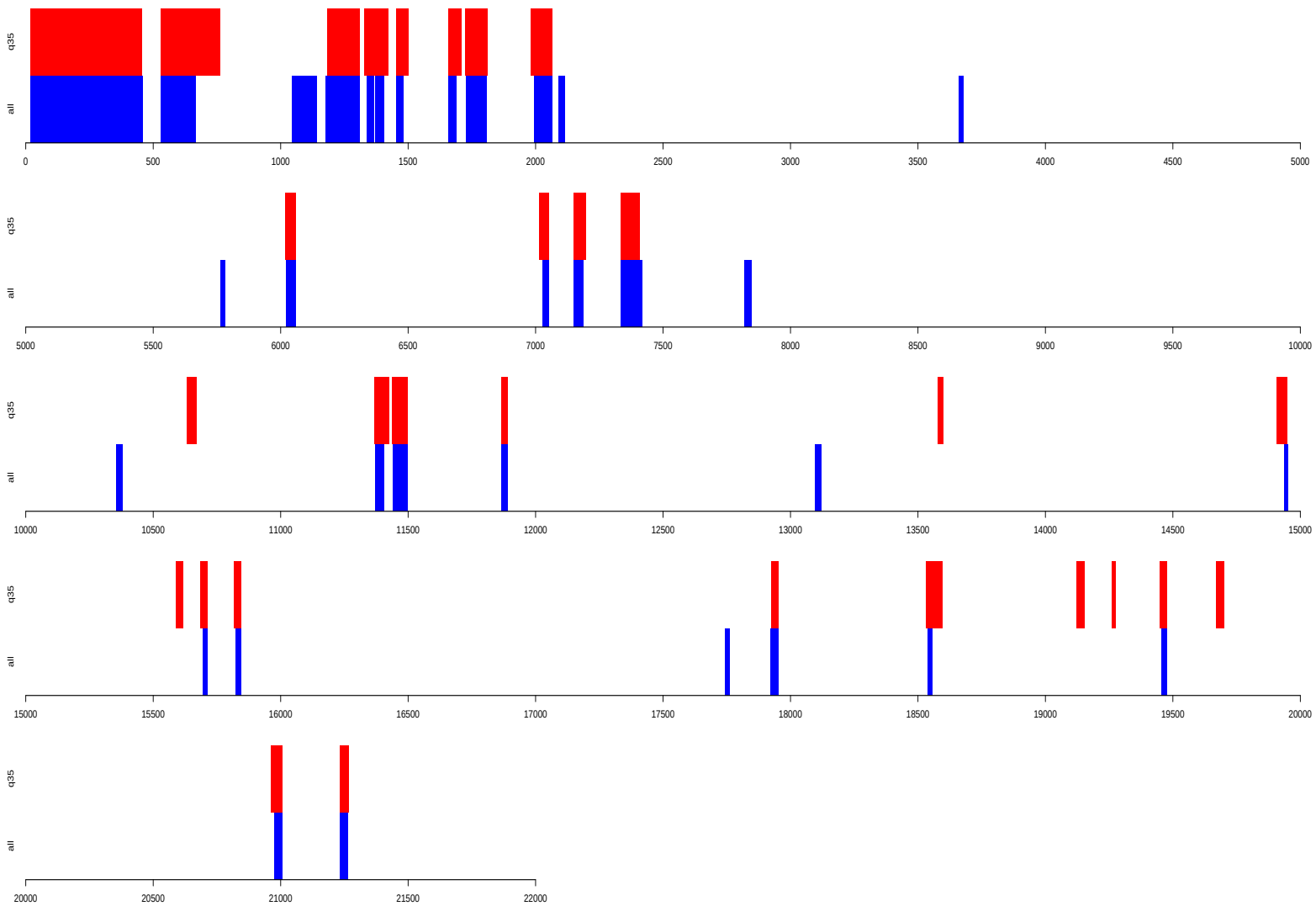


PHRED score $>$ 35

Distribution des fréquences

- Le biais d'estimation provient du fait que les PHRED scores ne correspondent pas à des vraies probabilités d'erreurs de séquençage.
- PHRED sur-estimés $\rightarrow$ probabilité des fréquences faibles sous-estimée.
- Calibrage des PHRED scores possible (GATK), mais nécessite une jeu de données pour lequel vraie séquence connue.

Signatures de sélection



Conclusions et perspectives

- Séquençage en pool de nombreux animaux à faible profondeur (1X par chromosome) est une approche prometteuse pour génétique des populations.
- Très bons résultats sur simulation.
- Résultats concluants sur données réelles pour la recherche de signatures de sélection.
- Réfléchir à la prise en compte des erreurs de séquençage.