Méthodes d'apprentissage statistique pour les tests d'association écologique

Kevin Caye

Co-encadrants : Olivier François (TIMC-IMAG), Olivier Michel (GIPSA-lab), Jean-Luc Bosson (TIMC-IMAG)

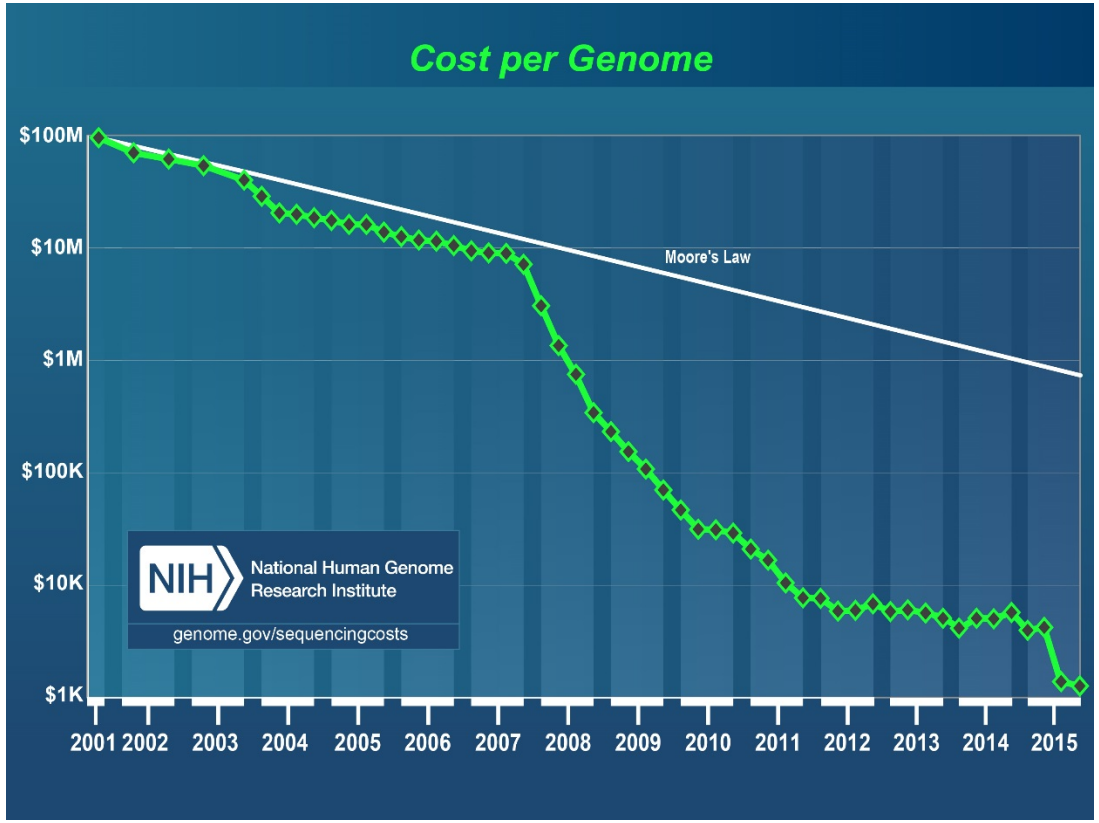
I-Contexte
II-Problématique et objectifs
III-Résultats préliminaires

I-Contexte

Données massives en génétique

Besoin de méthodes adaptées :

- aux problèmes d'association génotype/phénotype
- aux problèmes dus à la grande dimension des données



Approche statistique pour l'association génétique

Trouver des gènes en lien statistique avec des variables écologiques

$$\begin{pmatrix} \text{ACTGA} \ldots \text{TGTG} \\ \text{ACGGT} \ldots \text{ATTG} \\ \vdots \\ \text{TCAGA} \ldots \text{CCCC} \end{pmatrix} \sim \text{variables écologiques}$$

Exemple :

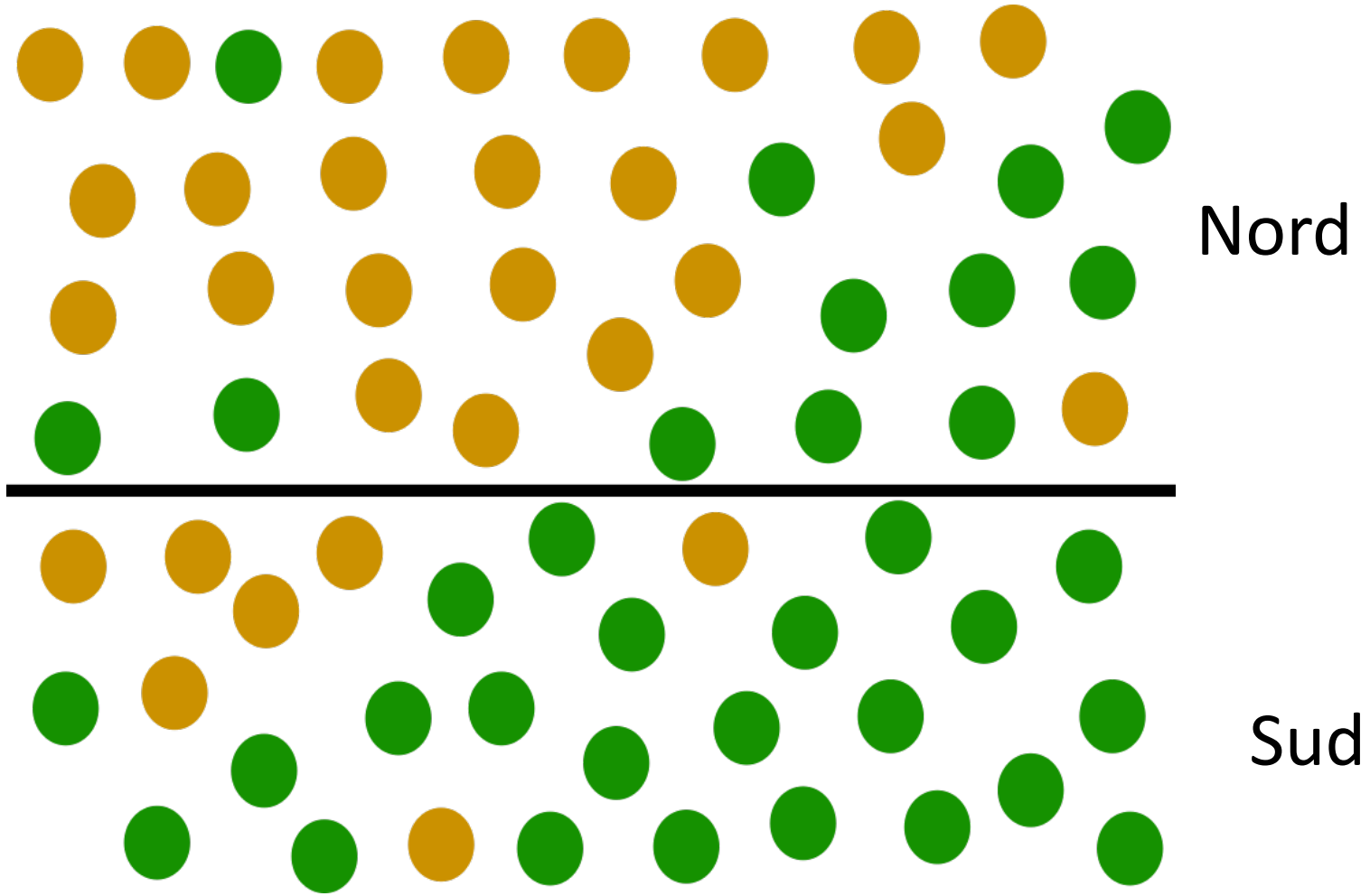
- gènes *EPAS1* et *EGLN1* liés à la tolérance à l'hypoxie chez l'humain.

II-Problématique et objectifs

Les facteurs de confusion : structure de population

Allèle A

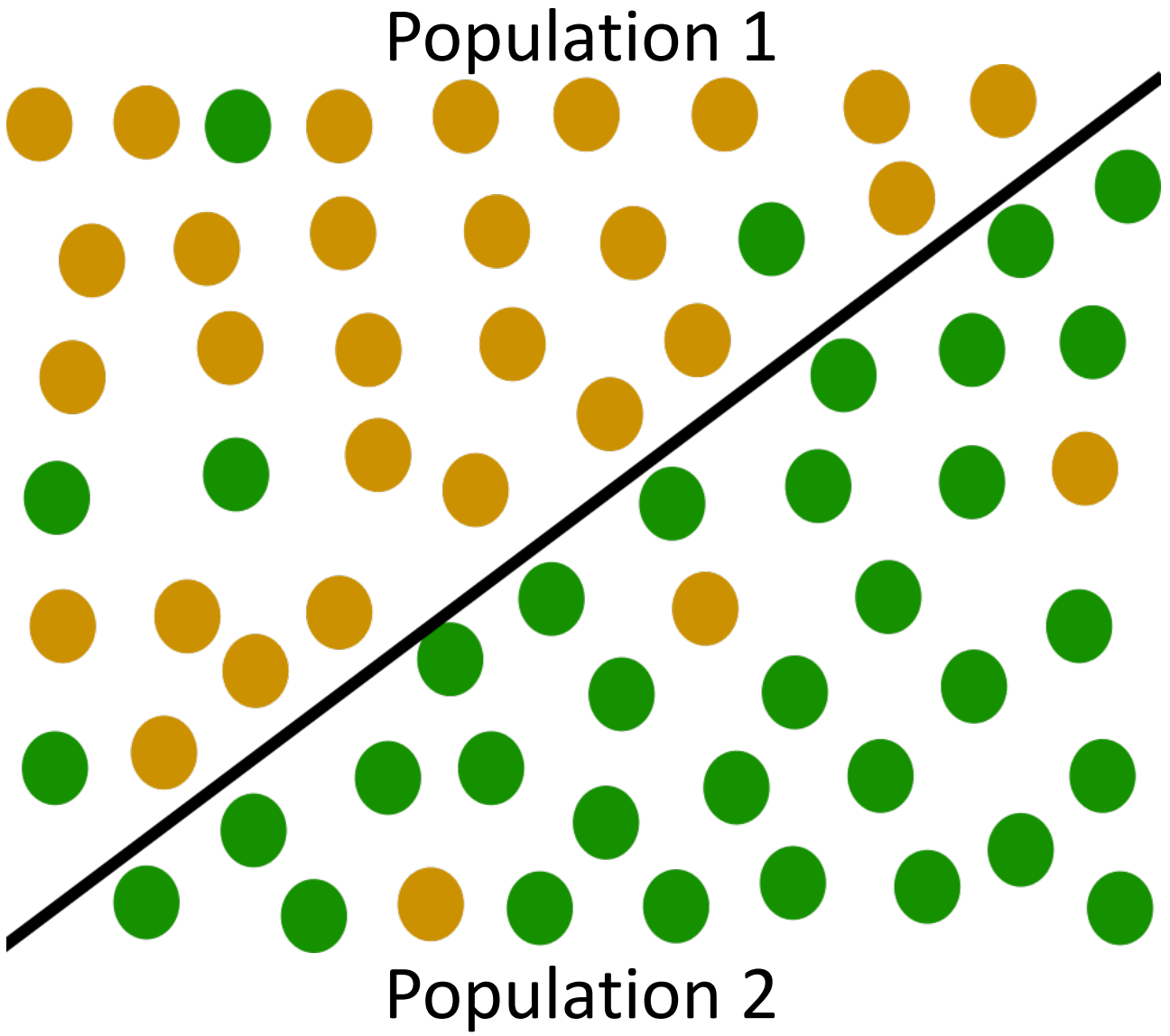
Allèle B



Les facteurs de confusion : structure de population

Allèle A

Allèle B



Contrôle des fausses découvertes

On veut fournir une liste de gènes candidats lié à l'adaptation des populations :

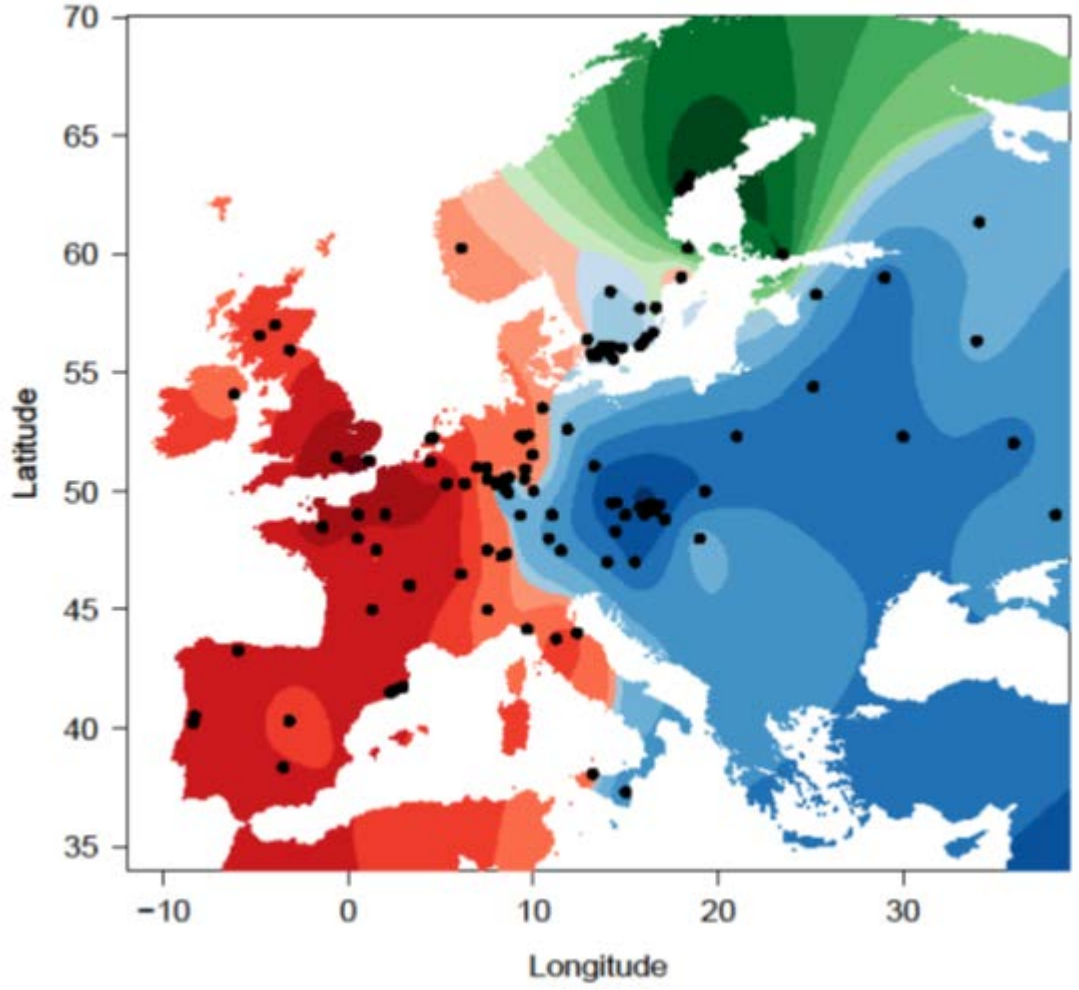
locus 560, locus 2 362, locus 693, locus 10 002, locus 563, locus 98 000, locus 89 652, locus 789 623, locus 78,…

Objectif : proposer des méthodes de contrôle des faux positifs dus aux facteurs de confusion et à l'incertitude statistique (échantillonnage, biais de mesure,…).

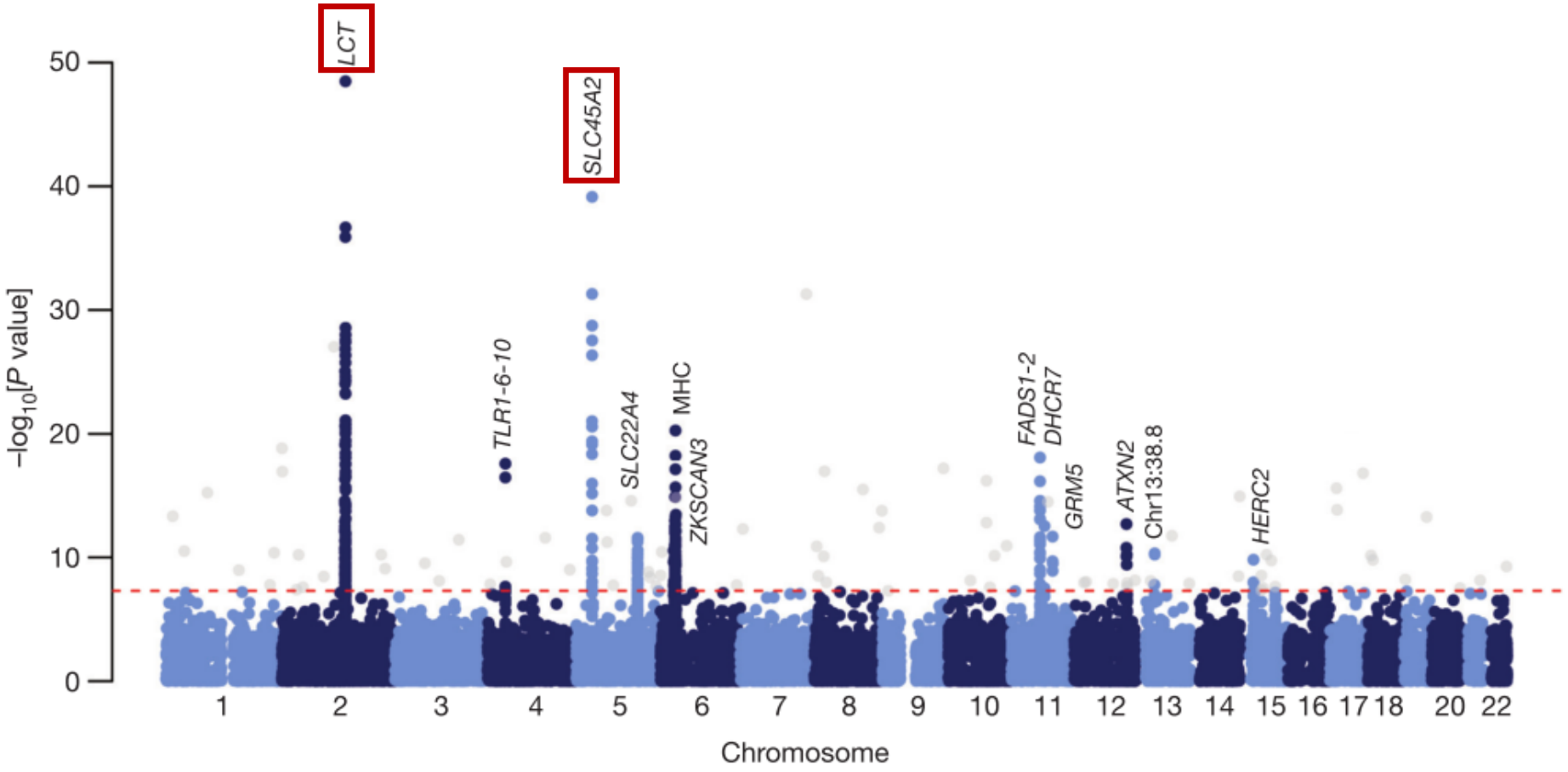
III-Résultats préliminaires

Travail publié : *TESS3: fast inference of spatial population structure and genome scans for selection*

- Estimation des coefficients individuels d'ascendance génétique.
- Méthode fondée sur un problème de minimisation des moindres carrés.
- Scan à la sélection : trouver des gènes liés à l'adaptation à l'environnement.



Résultat : Population Reference Sample



Mathieson, Iain, et al. "Genome-wide patterns of selection in 230 ancient Eurasians." *Nature* 528.7583 (2015): 499-503.

Travail à venir

Tester de nouveaux algorithmes de complexité plus faible.

Améliorer l’estimation des coefficients d’ascendance individuel et augmenter la puissance des tests de neutralité adaptative des locus.

Application à des études d’association génotype / phénotype (Welcome Trust Case Control Consortium)

Merci de votre attention