LEA: Un paquetage R pour la génomique des populations

Olivier François
Université Grenoble-Alpes

Grenoble, Juin 2015

LEA : **Landscape and Ecological** genomewide **Association** studies

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015

doi: 10.1111/2041-210X.12382

APPLICATION

LEA: An R package for landscape and ecological association studies

Eric Frichot[1] and Olivier François[1]*

LEA: Un programme R pour l'écologie moléculaire (à haut-débit)

- Objectifs généraux: Evaluer les effets des interactions de longue durée entre les organismes d'une population et leurs environnements (climat, régimes alimentaires, exposition aux pathogènes).

LEA: Un programme R pour l'écologie moléculaire (à haut-débit)

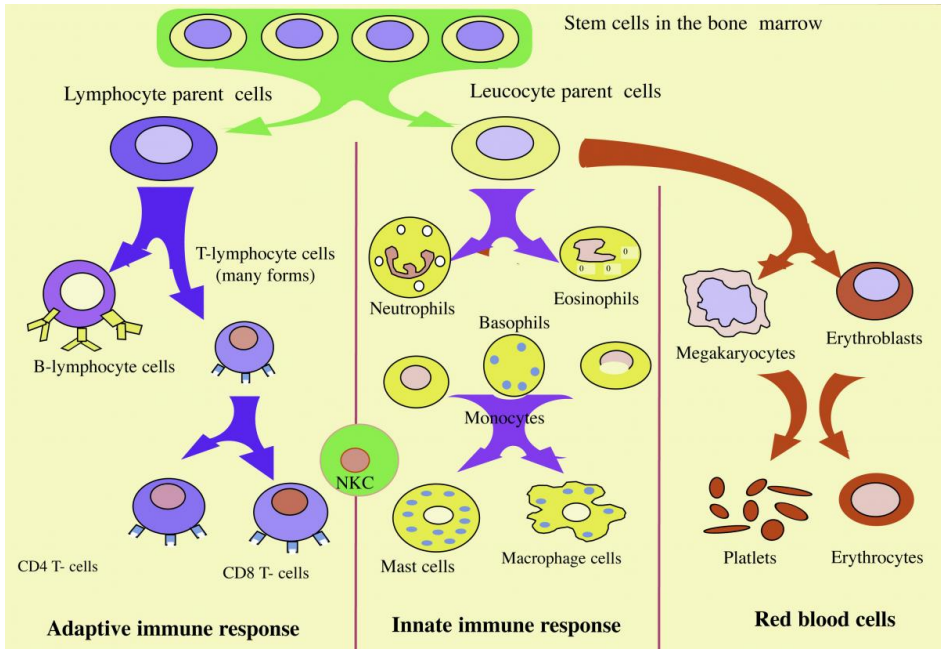
- Analyser la structure génétique des populations.
- Effectuer des cribles génomiques pour détecter les régions du génome répondant à la sélection naturelle.

Exemples chez l'homme

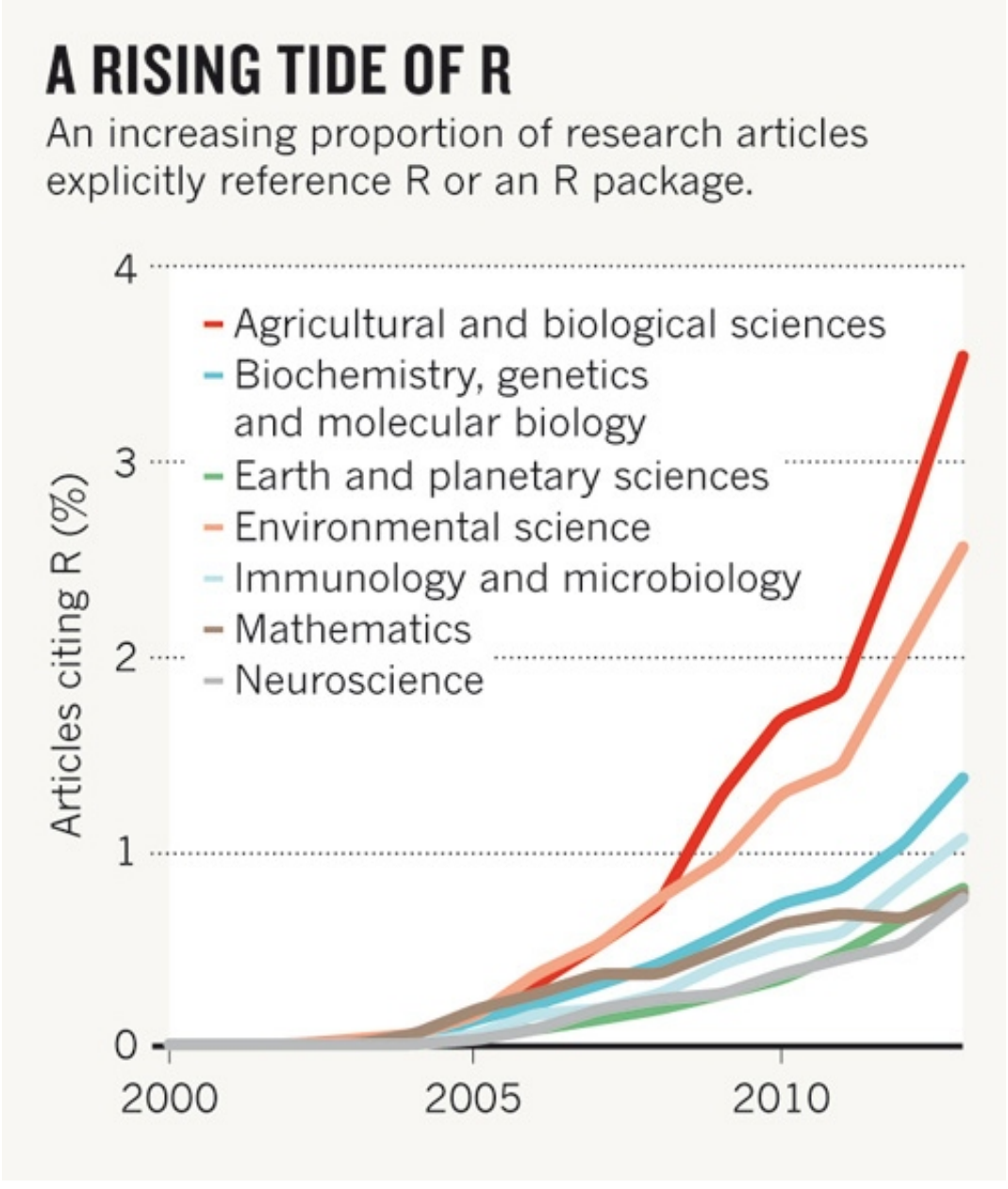
- Des mutations des gènes *EGLN1* and *PPARA* confèrent une **tolérance à l'hypoxie** et permettent l'adaptation à la **haute altitude** chez les populations tibétaines (Simonsen et al. Science 2011).

Exemples chez l'homme

- Excès de gènes associés à des **maladies auto-immunes** en réponse à des pressions sélectives exercées par les pathogènes (Fumagalli et al. Plos Genetics 2012).



Pourquoi choisir R?



Nature, 2015

Bases de données

- Projets de genomique humaine : **HGDP** (Li et al. Science 2008), **The 1000 Genome project** (Nature 2012)
- Matrices de génotypes de SNPs[1] de **très grandes** tailles : $(Y_{i\ell}) \sim$ **1–5 Giga** entrées ($Y_{i\ell} = 0,1,2$).

[1] SNP = Single Nucleotide Polymorphism (locus de l'ADN montrant une variation dans les populations)

Bases de données

- Bases de données environnementales: **Worldclim**, **WHO**
- Bases de données bioinformatiques: dbSNP, Genome browsers, NHI-EBI GWAS catalog, etc

Modèles statistiques implantés dans LEA

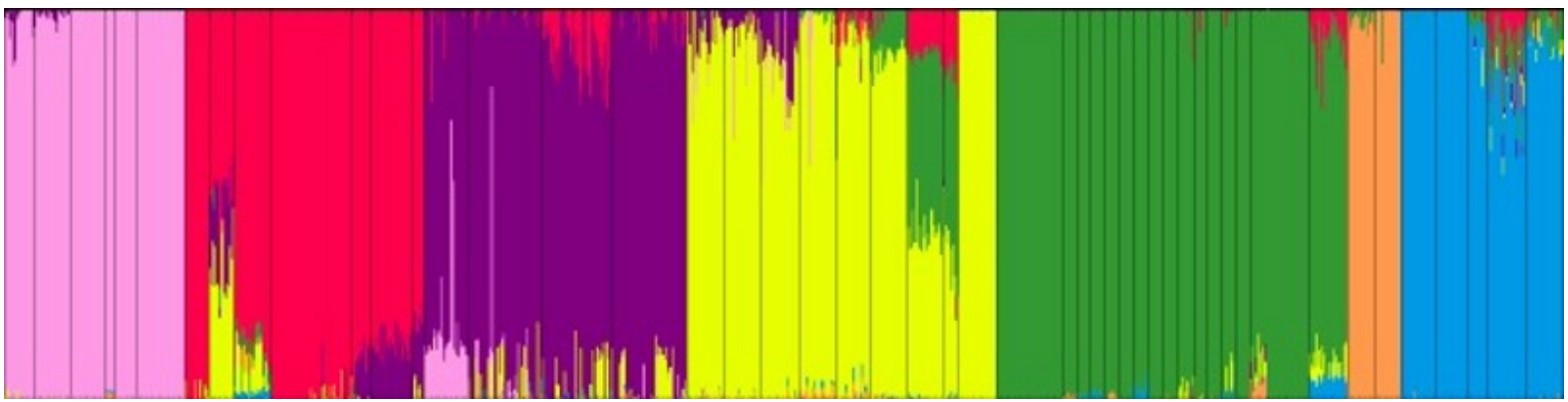
- **snmf** : Factorisation de matrice non-négative (avec contraintes de convexité)
  - Estimation de la structure génétique des populations et des coefficients individuels d'ascendance génétique.
  - Détection de régions génomiques répondant à la sélection naturelle (adaptation).

Modèles statistiques implantés dans LEA

- **lfmm** : Modèles mixtes à facteurs latents
  - Détection de régions génomiques associées à des variables écologiques
  - Correction des facteurs de confusion induits par la structure génétique des populations.

Coefficients d'ascendance génétique

- Les modèles de métissage génétique supposent que les **individus ont hérité leurs gènes de K populations ancestrales (K n'est pas connu)**.
- **Coefficient d'ascendance génétique** : la **fraction de genome de l'individual i provenant de la population ancestrale k, Q_{ik}**.



Méthode **snmf**

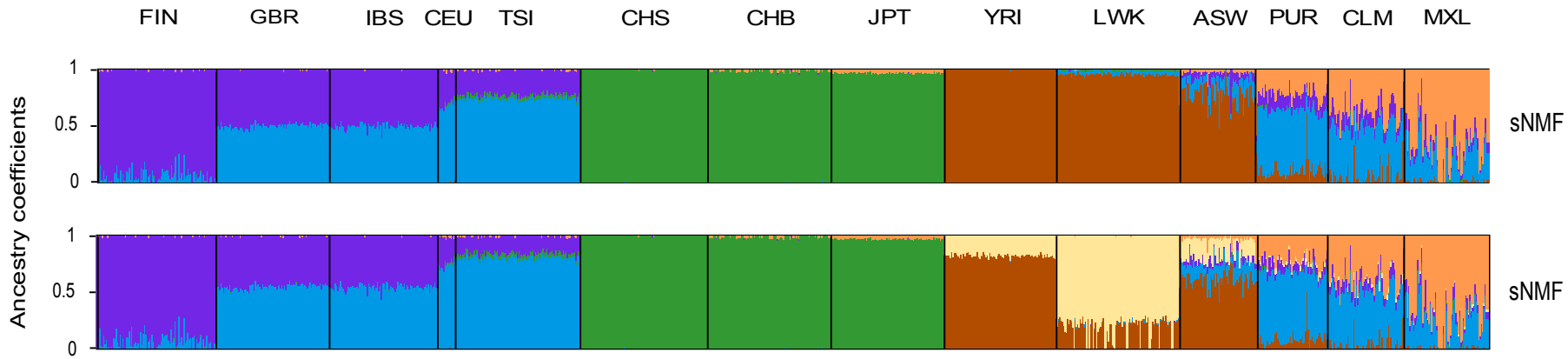
- La fonction **snmf** calcule des estimateurs de moindres carrés de la matrice $\mathbf{Q}$ (de rang K)

$$\mathrm{LS}(\mathbf{Q},\mathbf{F}) = \|\mathbf{Y} - \mathbf{Q}\mathbf{F}\|_{\mathrm{F}}^2 + \sqrt{\alpha}\sum_{i=1}^{n}\|\mathbf{Q}_{i\cdot}\|_1^2\,, \quad \mathbf{Q},\mathbf{F} \geq 0\,,$$

- sous les contraintes

$$\sum_{k=1}^{K}\mathbf{Q}_{ik} = 1\,, \quad \sum_{j=0}^{2}\mathbf{F}_{k\ell}(j) = 1\,, \quad j = 0,1,2.$$

Résultat pour les données 1000 Genomes

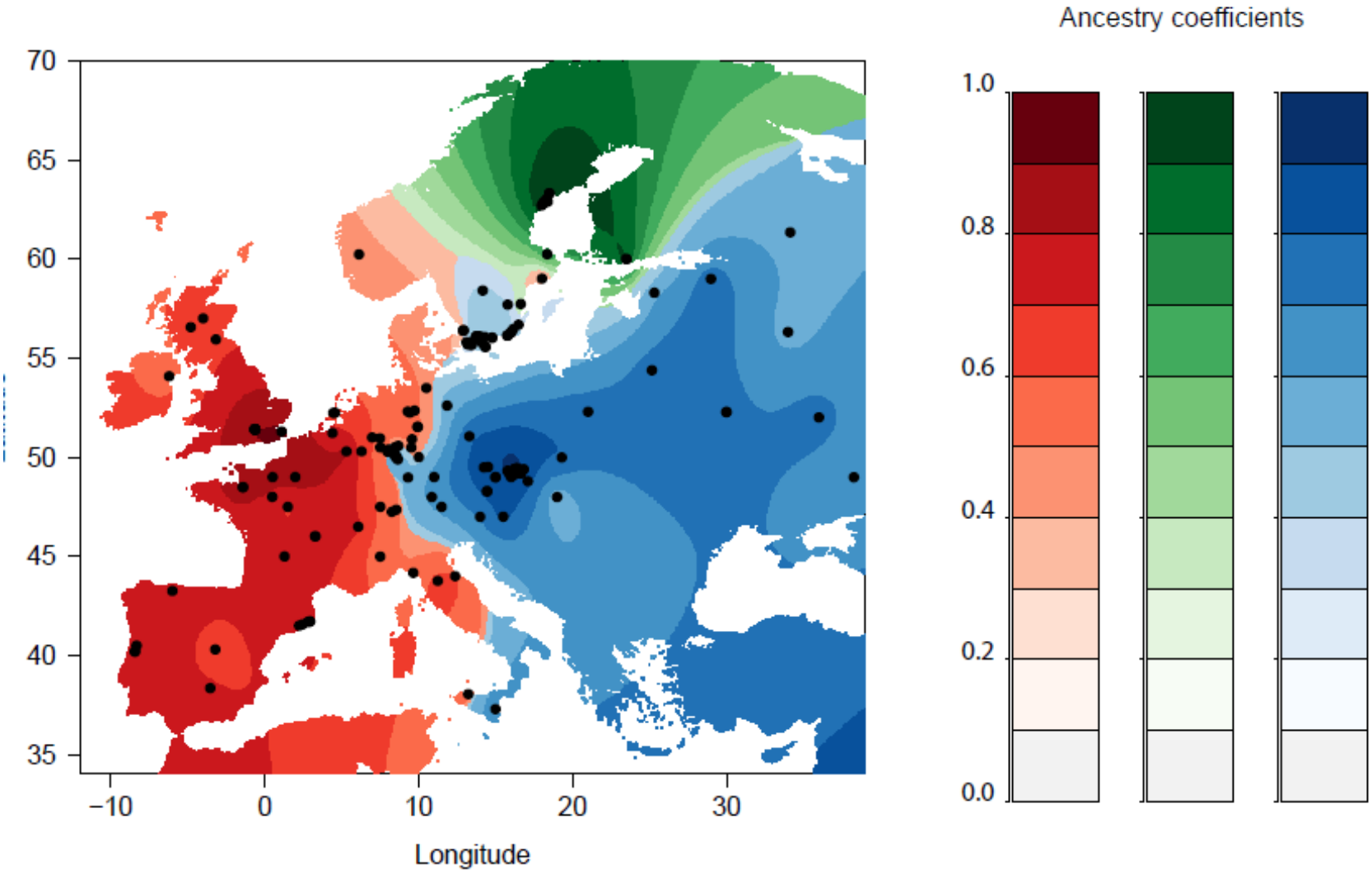


Dimension de la matrice : $n = 2\text{k}$, $p = 3\text{M}$,
Temps de calcul CPU $\approx$ 30 min
(Frichot et al. Genetics 2014)

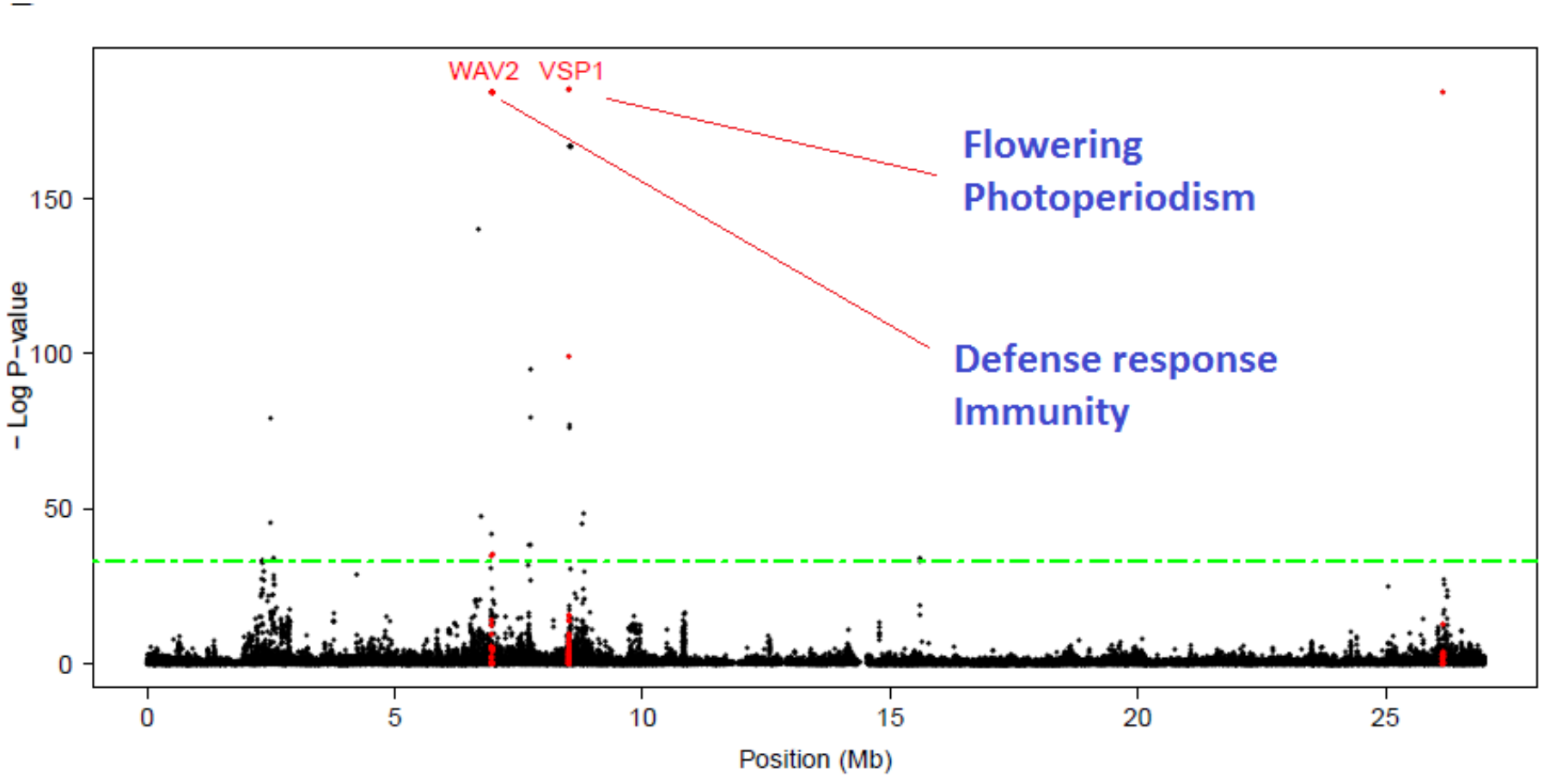
Résultats pour la plante *A. thaliana*

- Lignées européennes de la plante modèle *A. thaliana* (216K SNPs, 170 individus).
- La valeur du nombre de populations, K, est choisie selon un critère d'entropie croisée $\rightarrow K = 3$.

Résultat pour *A. thaliana*



Crible génomique pour *A. thaliana*



lfmm: Méthode d'association génomique

- **lfmm: latent factor mixed model**
- Estimer **la corrélation entre la fréquence d'allèle** pour un SNP donné et une **variable d'intérêt écologique**.
- Les cribles génomiques cherchent à identifier les SNPs montrant des différences significatives en comparaison avec le fond génomique.
- **Caveat**: La structure de population et d'autres facteurs de confusion, créent de nombreux faux positifs.

lfmm: Méthode d'association génomique

- **Latent factor mixed models**

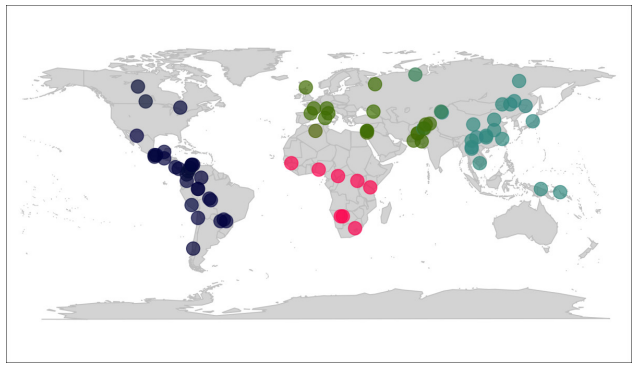
$$Y_{i\ell} = \mu_\ell + \beta_\ell^T X_i + \mathbf{U_i^T V}_\ell + \epsilon_{i\ell},$$

où β_ℓ est un vecteur de coefficients de régression, U_i est un facteur latent, et la matrice V_ℓ contient les poids correspondants (*loadings*) (Frichot et al. Mol. Biol. Evol. 2013).

- K facteurs latents.

Human Genome Diversity Project (660k SNP arrays)

- Echantillons d'ADN de 1,043 individus répartis mondialement dans 52 populations
- Données climatiques pour chacune des 52 populations (WorldClim database, résolution de $1km^2$)



- 11 variables bioclimatiques interpolées, collectées sur une période de 50 ans (1950-2000)
- Nombre de facteurs latents choisi par un test de Tracy-Widom.

GWAS-SNPs associated with environmental predictors.

Gene	Trait association	$-\log_{10}$ P-value
OCA2/HERC2	Eye and hair color, pigmentation	9.15
DHCR7	Vitamin D levels	7.78
SLC45A2	Hair color	6.90
Intergenic *MUC7*	Alcoholism	8.91
ZMIZ1	Crohn's disease	8.77
KLK3	Prostate Cancer	8.61
ICOSLG	Celiac disease	7.02
HLA-DRA	Systemic sclerosis	6.97
NCAPG-LCORL	Height	9.43
BOK	Brain structure and development	9.43

Genic SNPs associated with environmental predictors.

Gene	Annotation (dbSNPs)	$-\log_{10} P$-value
EPHB4	Heart morphogenesis and angiogenesis	16.54
NRG1	Nervous system development, cell proliferation	16.21
RBM19	Regulation of embryonic development	15.98
EYA2	Eye development and DNA repair	15.9
POLA1	Mitotic cell cycle and cell proliferation	15.87

Résumé

- Des algorithmes rapides de factorisation matricielle (modèles à facteurs latents)
- Permettant d'étudier la structure génétique des populations
- De lancer des cribles génomiques tout en contrôlant les fausses découvertes.

LEA

LEA: an R package for Landscape and Ecological Association Studies

Bioconductor version: Release (3.1)

LEA is an R package dedicated to landscape genomics and ecological association tests. LEA can run analyses of population structure and genome scans for local adaptation. It includes statistical methods for estimating ancestry coefficients from large genotypic matrices and evaluating the number of ancestral populations (snmf, pca); and identifying genetic polymorphisms that exhibit high correlation with some environmental gradient or with the variables used as proxies for ecological pressures (lfmm), and controlling the false discovery rate. LEA is mainly based on optimized C programs that can scale with the dimension of very large data sets.

Author: Eric Frichot <eric.frichot at gmail.com>, Olivier Francois <olivier.francois at imag.fr>