

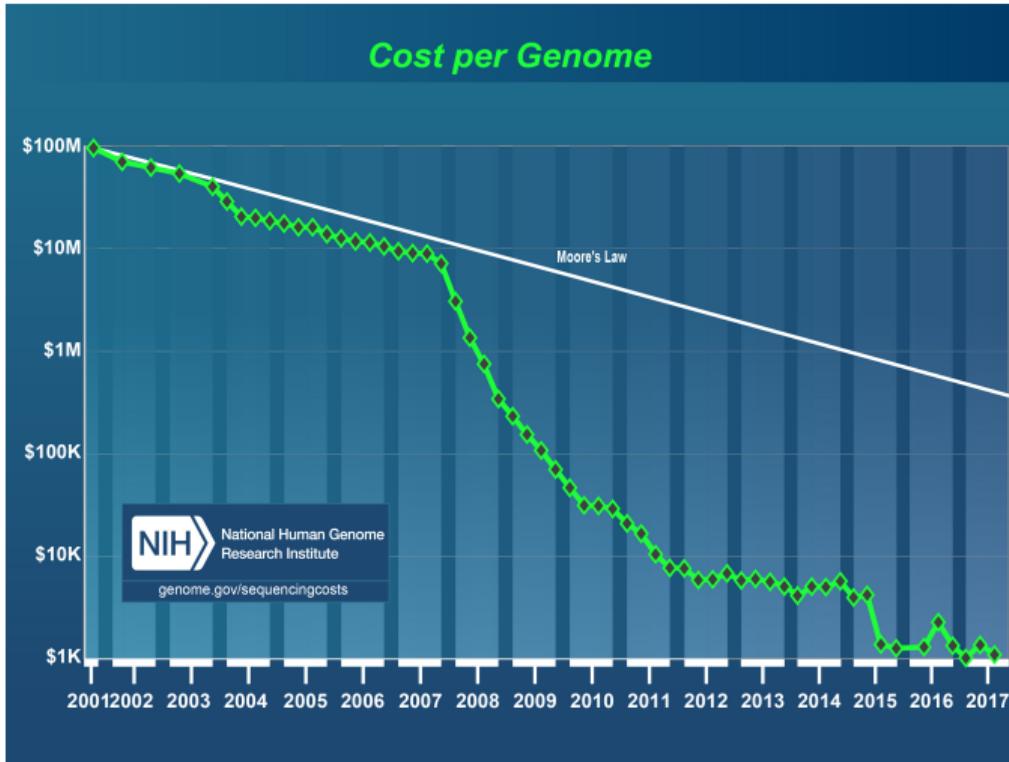
Méthodes de factorisation matricielle pour la génomique des populations et les tests d'association

Kévin CAYE

Thèse dirigée par Olivier FRANÇOIS
et co-encadrée par Olivier MICHEL et Jean-Luc BOSSON



Volume des données en génétique



Les données génétiques

- ▶ L'ADN est une longue séquence de nucléotides : A, C, T, G (3 milliards chez l'humain)
- ▶ Les données sont constituées de polymorphismes nucléotidiques appelés SNPs (10 millions chez l'humain)
- ▶ Il y a deux allèles possibles par locus

ADNs	...	G	A	T	C	C	...
	...	G	A	A	C	C	...
	...	G	A	A	C	C	...
	...	G	A	T	C	C	...
	...	G	A	T	C	C	...
	...	G	A	T	C	C	...

Illustration d'un SNP. Le nucléotide différent entre les séquences d'ADN est un SNP.

Matrice de génotype

Les données sont rangées dans une matrice de taille $n \times p$:

$$\mathbf{Y} = \begin{bmatrix} 0 & 1 & 2 & 2 & \cdots & \cdots & \cdots \\ 1 & 1 & 0 & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ 0 & 0 & 2 & 0 & \cdots & \cdots & \cdots \end{bmatrix}$$

Illustration d'une matrice de génotype. Chaque élément de la matrice est le nombre de fois que l'allèle muté est observé pour un individu donné à un locus donné

Problématiques

Inférence des coefficients de métissage à l'aide de données géographiques

- ▶ Kévin Caye, Timo Deist, Helena Martins, Olivier Michel, Olivier François (2016) TESS3 : fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources* 16 (2), 540-548. DOI : [10.1111/1755-0998.12471](https://doi.org/10.1111/1755-0998.12471).
- ▶ Kévin Caye, Flora Jay, Olivier Michel, Olivier Francois (2017). Fast Inference of Individual Admixture Coefficients Using Geographic Data. *The Annals of Applied Statistics*. DOI : [10.1101/080291](https://doi.org/10.1101/080291).
- ▶ Kévin Caye, Olivier François, Flora Jay. tess3r : An R package for estimating and visualizing spatial population structure based on geographically constrained non-negative matrix factorization and population genetics.

Problématiques

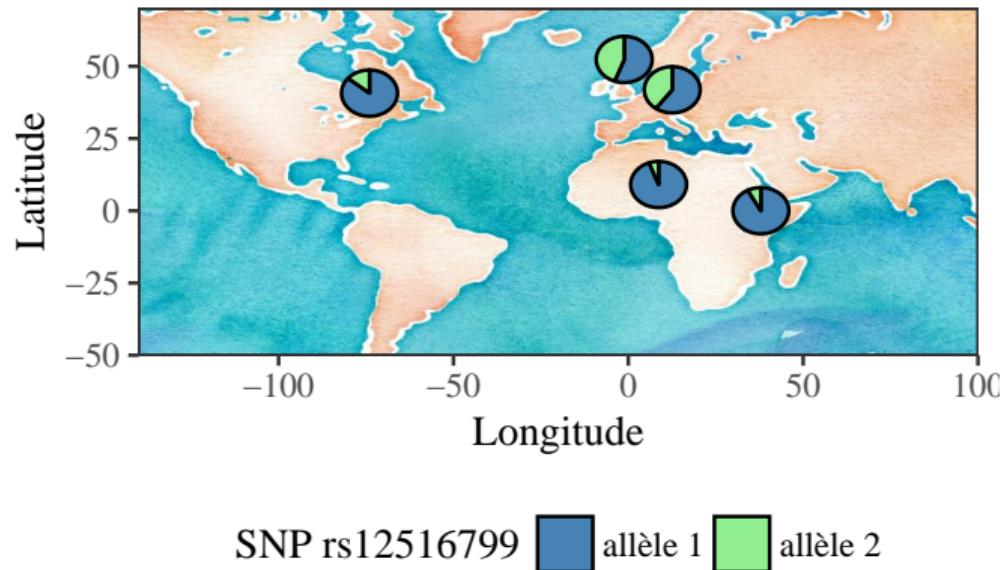
Correction des facteurs de confusion pour les études d'association

- ▶ Olivier François, Kevin Caye. Genome-wide association studies in geographically structured populations. *Molecular Ecology Resources*, papier invité. En révision.
- ▶ Kevin Caye, Olivier François. Optimal confounder adjustment in genome and epigenome-wide association studies. En préparation.
- ▶ Kévin Caye, Olivier François. `lfmm` : An R package for correcting association studies for confounding factors.

Section 2

Inférence des coefficients de métissage à l'aide de données géographiques

La structure génétique des populations



Différenciation allélique entre des populations. Distribution des allèles du SNP rs17066888 dans des populations européenne, africaine et afro-américaine (données 1000Genome Consortium, 2015).

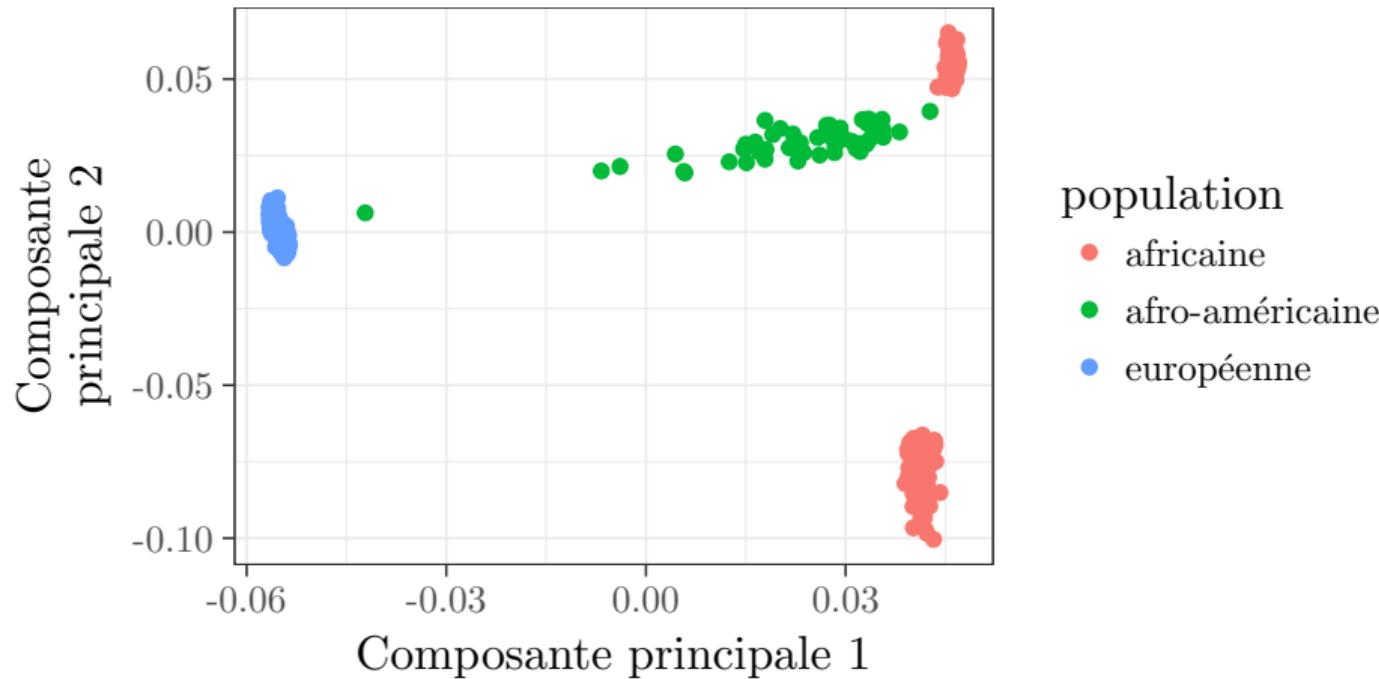
Pressions évolutives :

- ▶ la mutation
- ▶ la sélection
- ▶ la dérive génétique
- ▶ la migration

Pourquoi étudier la structure génétique des populations ?

- ▶ Représentation synthétique de données multivariées
- ▶ Étude de l'histoire démographique des populations (Li et al., 2008)
- ▶ Étude d'association de gènes avec une maladie (Marchini et al., 2004)

Visualisation de la structure génétique des populations avec l'ACP



Scores des deux premières composantes principales calculés sur des données de SNPs d'invidus humains de populations européenne, africaine et afro-américaine (données 1000Genome Consortium, 2015).

Le modèle de métissage (Pritchard et al., 2000)

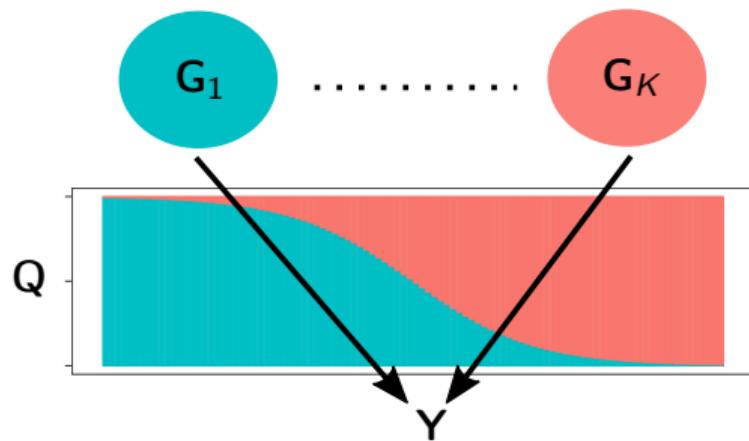


Illustration du modèle de structure génétique de population.

$$\Pr(\mathbf{Y}_{i,\ell} = j) = \sum_{k=1}^K \mathbf{G}_{(d+1)\ell+j,k} \mathbf{Q}_{i,k}$$

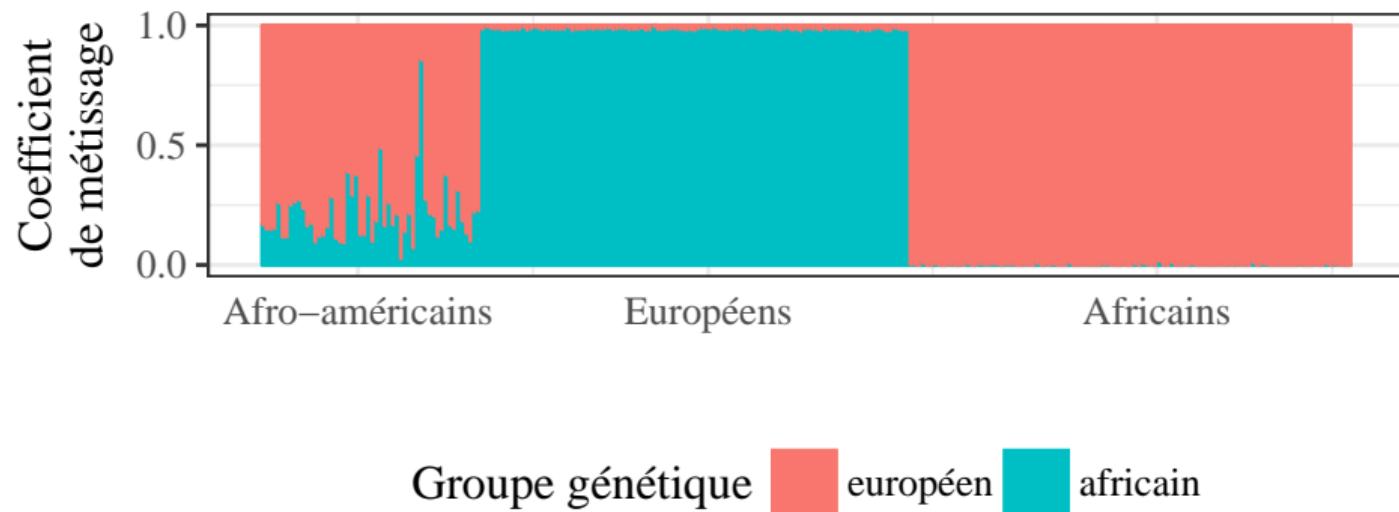
où

- ▶ d est la ploïdie ($d = 2$ pour une espèce diploïde)
- ▶ $\Pr(\mathbf{Y}_{i,\ell} = j)$ est la probabilité d'observer l'allèle j au locus ℓ chez l'individu i
- ▶ $\mathbf{G}_{(d+1)\ell+j,k}$ est la fréquence d'apparition de l'allèle j au locus ℓ dans le groupe génétique k .
- ▶ $\mathbf{Q}_{i,k}$ est la proportion de gènes de l'individu i provenant du groupe k .

Méthodes d'estimation des coefficients de métissage

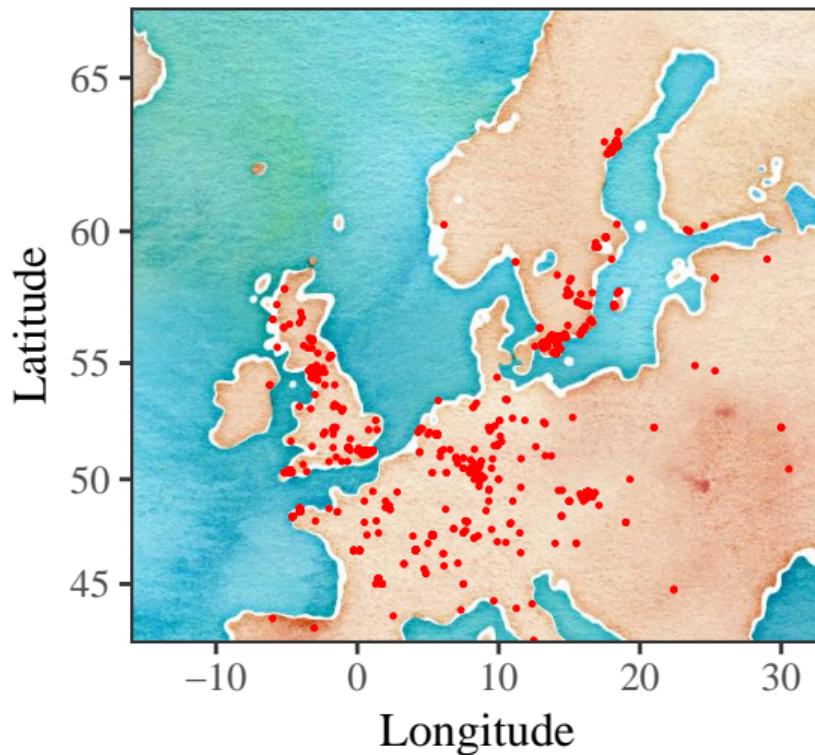
Méthode	Modèle	Algorithme	Référence
STRUCTURE	bayésien	MCMC	Pritchard et al. (2000)
FRAPPE	vraisemblance	EM	Tang et al. (2005)
ADMIXTURE	vraisemblance	optimisation quasi-Newton alternée	Alexander et Lange (2011)
fastStructure	bayésien	inférence variationnelle bayésienne	Raj et al. (2014)
PSIKO	ACP	SVD	Popescu et al. (2014)
sNMF	factorisation parcimonieuse matricielle	optimisation quadratique alternée avec projection	Frichot, Mathieu et al. (2014)

Visualisation des coefficients de métissage (Q)



Estimation par le logiciel `snmf` (Frichot, Mathieu et al., 2014) des coefficients de métissage pour un jeu de données composé d'individus humains provenant de populations européenne, africaine et afro-américaine.

Données géographiques



Méthodes d'estimation des coefficients de métissage à l'aide de données géographiques

Méthode	Modèle	Algorithme	Référence
TESS	bayésien	MCMC	Chen et al. (2007)
GENELAND	bayésien	MCMC	Guedj et Guillot (2011)
BAPS	bayésien	optimisation stochastique	Corander et al. (2008)
TESS3	factorisation matricielle régularisée sur graphe	moindres carrés alternés projetés	Caye, Jay et al. (2017)
conStruct	bayésien	MCMC	Bradburd et al. (2017)

Estimation des matrices d'ascendance génétique

Fritchot, Mathieu et al. (2014) cherchent à décomposer la matrice de génotype :

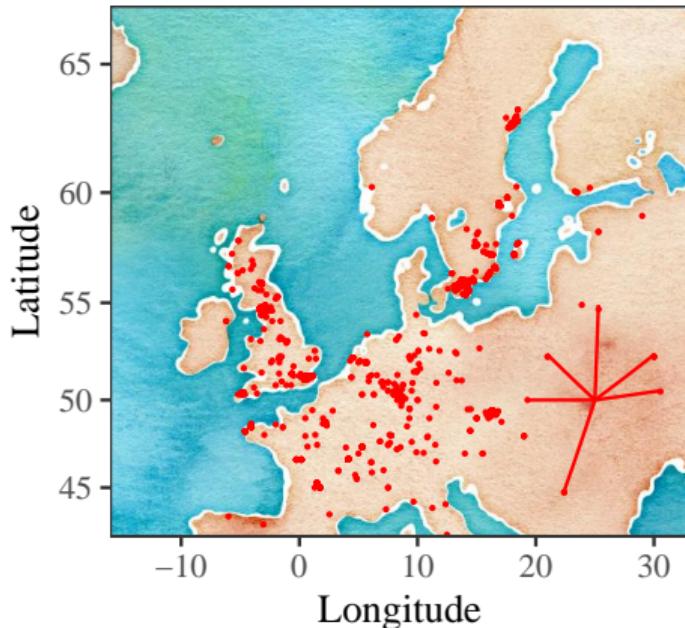
$$\begin{bmatrix} Q \\ \hline \end{bmatrix} \times \begin{bmatrix} G \\ \hline \end{bmatrix} \approx \begin{bmatrix} Y \\ \hline \end{bmatrix}$$

où

$$Q \geq 0, \quad \sum_{k=1}^K Q_{i,k} = 1, \quad i = 1 \dots n$$

$$G \geq 0, \quad \sum_{j=0}^d G_{(d+1)\ell+j,k} = 1, \quad \ell = 1 \dots p.$$

Information géographique



Entre chaque individu i et j , nous avons le poids de graphe

$$W_{i,j} = \exp(-\text{dist}(x_i, x_j)^2 / \sigma^2)$$

où

- ▶ $\text{dist}(x_i, x_j)$ est la fonction de distance entre les coordonnées géographiques x_i et x_j .
- ▶ σ est le paramètre d'échelle géographique

Problème des moindres carrés

Pour estimer les matrices d'ascendance on cherche à minimiser la fonction

$$\mathcal{L}(\mathbf{Q}, \mathbf{G}) = \|\mathbf{Y} - \mathbf{Q}\mathbf{G}^T\|_{\text{F}}^2 + \frac{\alpha}{2} \sum_{i,j=1}^n \mathbf{W}_{i,j} \|\mathbf{Q}_{i,.} - \mathbf{Q}_{j,.}\|^2$$

avec les contraintes

$$\mathbf{Q} \geq 0, \quad \sum_{k=1}^K \mathbf{Q}_{i,k} = 1, \quad i = 1 \dots n$$

$$\mathbf{G} \geq 0, \quad \sum_{j=0}^d \mathbf{G}_{(d+1)\ell+j,k} = 1, \quad \ell = 1 \dots p.$$

Algorithme de descente par blocs de coordonnées

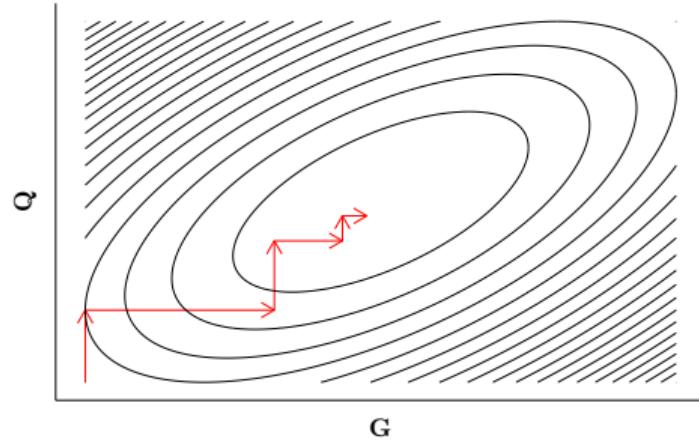


Illustration de l'algorithme de descente par blocs de coordonnées.

On alterne deux étapes jusqu'à la convergence vers un point critique :

- ▶ optimisation de \mathcal{L} selon \mathbf{Q} avec \mathbf{G} fixé
- ▶ optimisation de \mathcal{L} selon \mathbf{G} avec \mathbf{Q} fixé

Nous présentons deux algorithmes utilisant ce principe.

Algorithme de descente par blocs de coordonnées

Algorithme d'optimisation quadratique alternée (AQP)

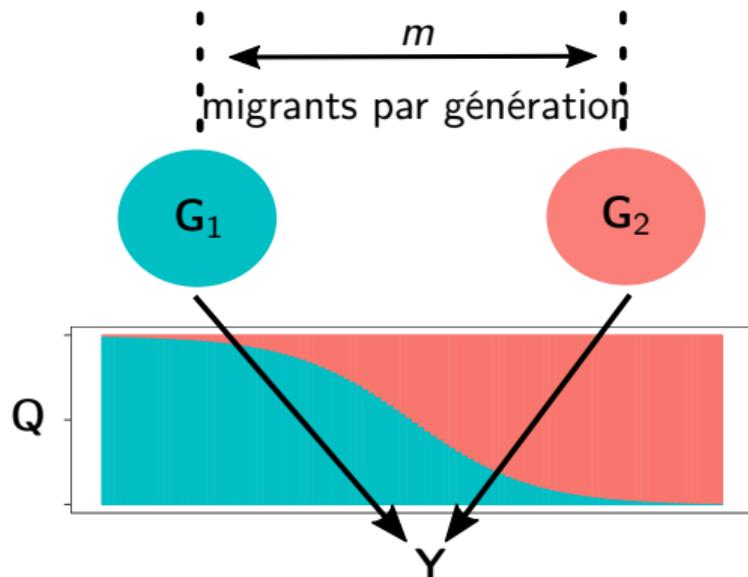
- ▶ D'après Grippo et Sciandrone (2000), AQP converge vers un minimum local de la fonction objectif \mathcal{L}
- ▶ L'étape de calcul de \mathbf{Q} implique de résoudre un problème d'optimisation quadratique de taille $n \times K$

Algorithme des moindres carrés alternés projetés (APLS)

- ▶ L'étape de calcul de \mathbf{Q} peut être séparée en n moindres carrés régularisés en norme L_2
- ▶ Dans nos comparaisons, APLS fournit de bonnes approximations de AQP tout en étant plus rapide

Nous utilisons APLS dans la suite

Simulation de génotypes métissés spatialement

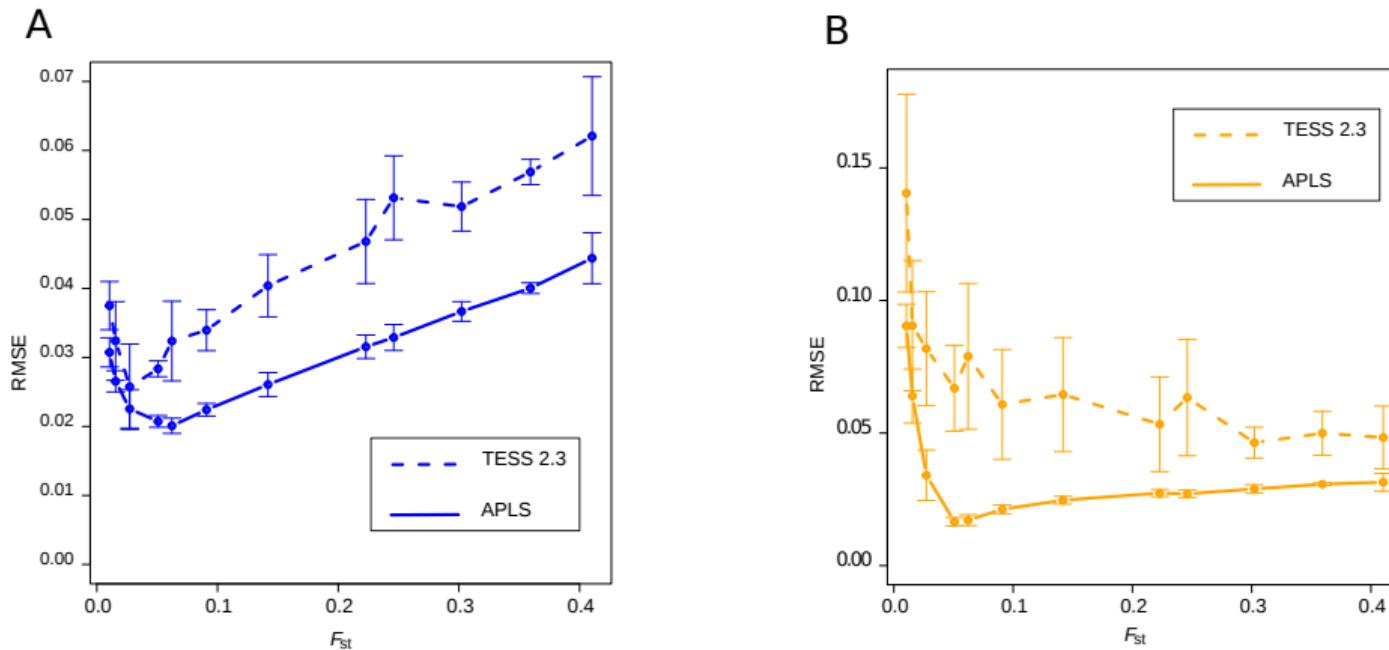


- ▶ La matrice **G** est simulée par un modèle de Wright à deux îles
- ▶ La matrice **Q** est simulée selon un gradient longitudinal
- ▶ La matrice **Y** est générée en tirant des gènes des deux populations sources avec des probabilités données par les coefficients de métissage

On simule plusieurs génotypes pour avoir plusieurs valeurs de différenciation mesurées par

$$F_{ST} = \frac{1}{1 + 4N_0m}$$

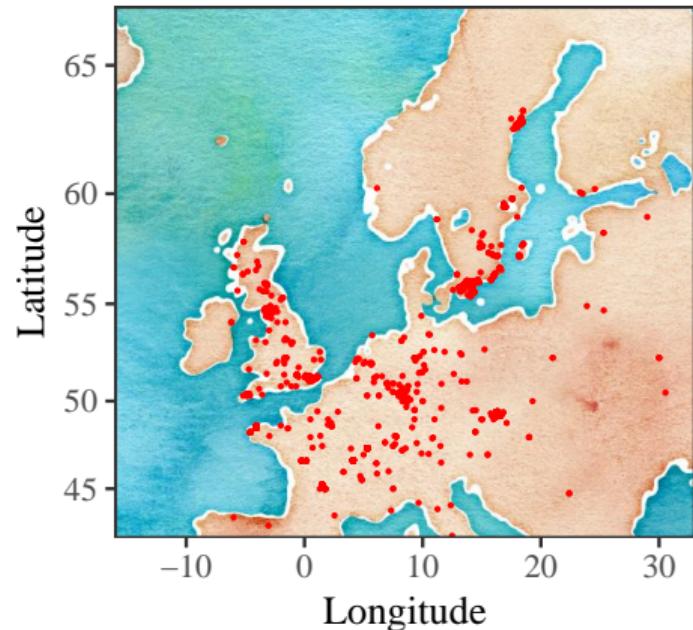
Comparaison avec une méthode bayésienne TESS 2.3 (Caye, Deist et al., 2015)



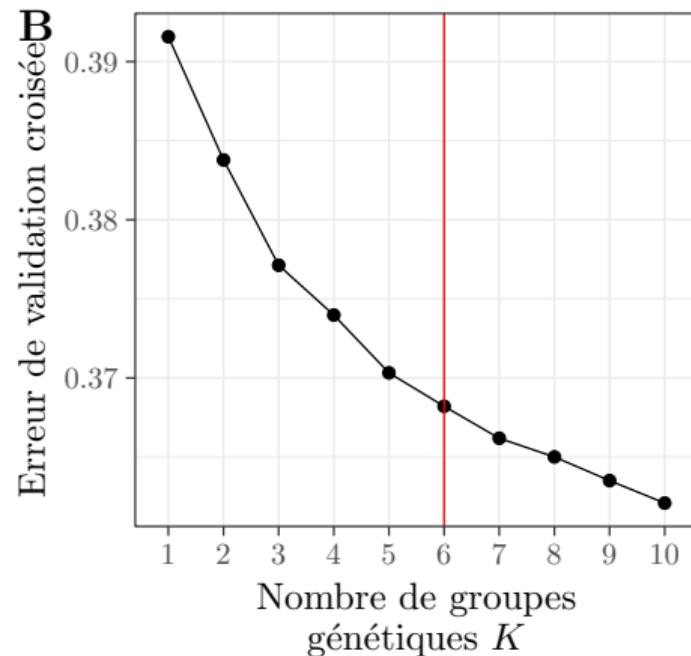
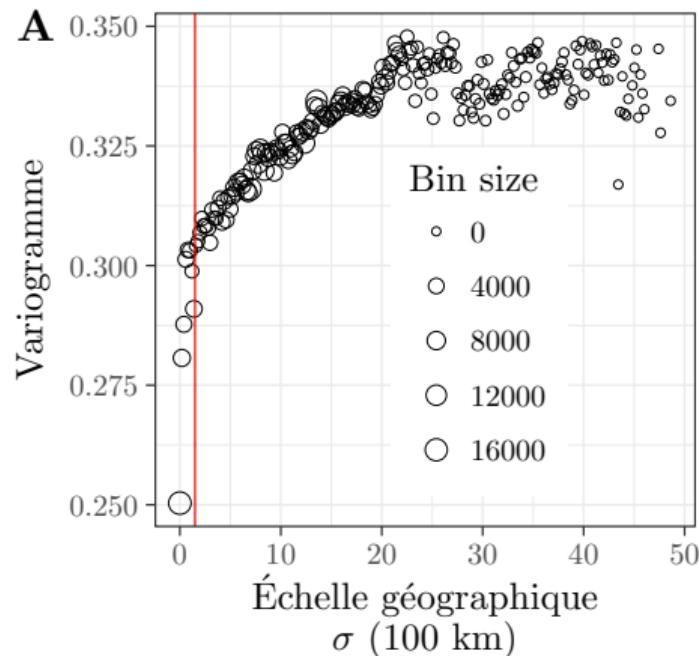
Racine de l'erreur quadratique moyenne (RMSE) pour l'estimation de Q (figure A) et G (figure B). APLS était 30 fois plus rapide que TESS 2.3.

Application à des données *Arabidopsis thaliana*

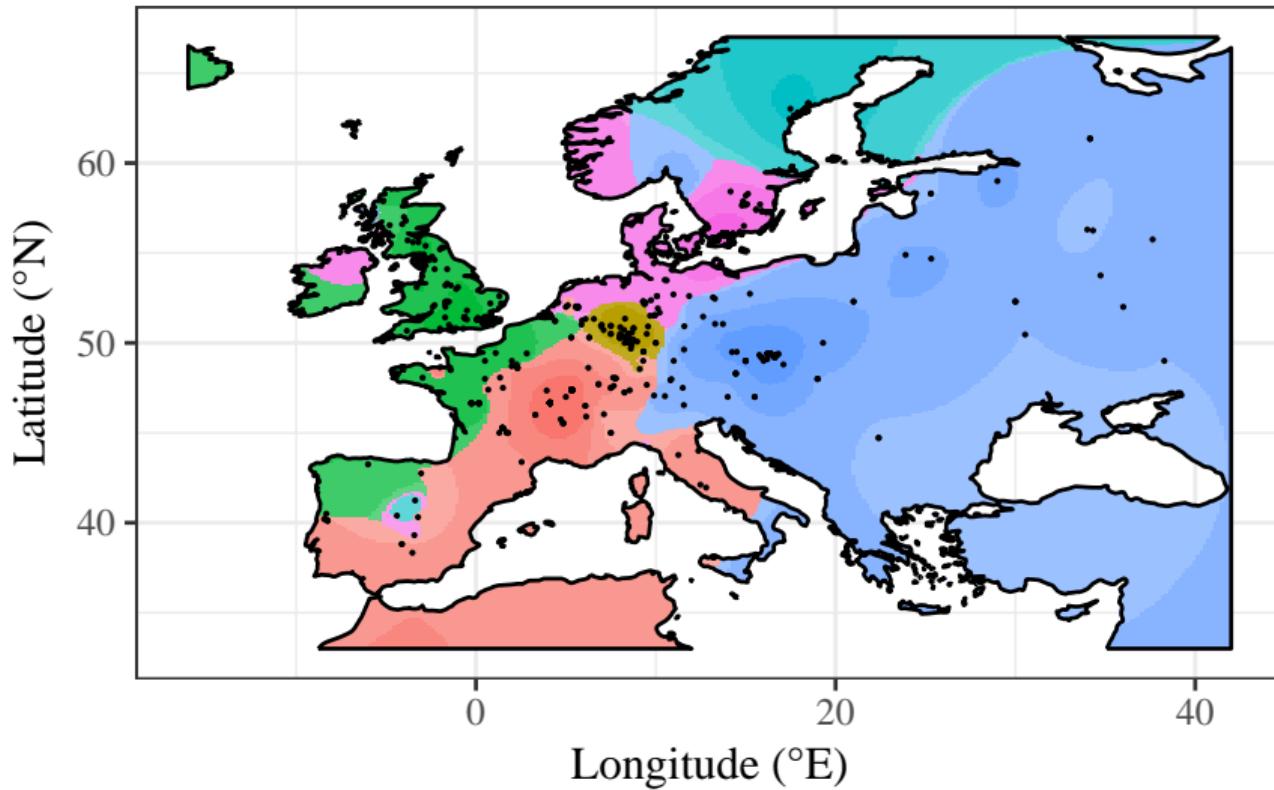
On étudie 214k SNPs pour 1 095 écotypes européens des espèces végétales *A.thaliana* (Horton et al., 2012).



Choix des paramètres d'échelle géographique (σ) et du nombre de groupes génétiques (K)



Carte des coefficients de métissage



Section 3

Correction des facteurs de confusion pour les études d'association

Test d'association

Objectif

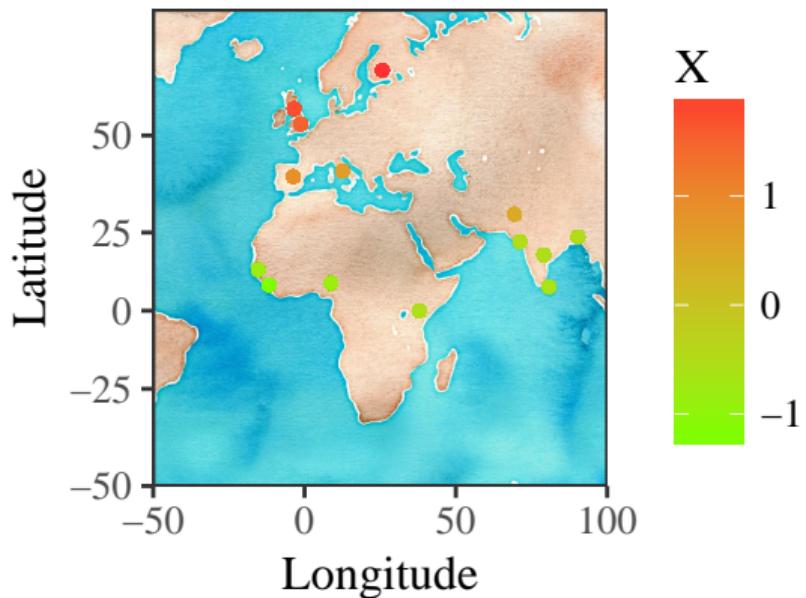
Déetecter les SNPs dont les fréquences sont corrélées à une variable d'intérêt

$$\begin{bmatrix} 0 & 1 & 2 & 2 & \dots & \dots & 0 & \dots \\ 1 & 1 & 0 & \dots & \dots & \dots & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots & \dots & \vdots & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots & \dots & \vdots & \dots \\ 0 & 0 & 2 & 0 & \dots & \dots & 1 & \dots \end{bmatrix} \sim \begin{bmatrix} 0.2 \\ 1.5 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

Exemple de variable d'intérêt

- ▶ Maladie : diabète, maladie cœliaque (Dubois et al., 2010)
- ▶ Phénotype : la taille (Wood et al., 2014)
- ▶ Environnement : la température (Frichot, Schoville et al., 2013)

Étude d'association entre des données génétiques et un gradient climatique



Données génétiques du projet
1000Genome (Consortium, 2015)

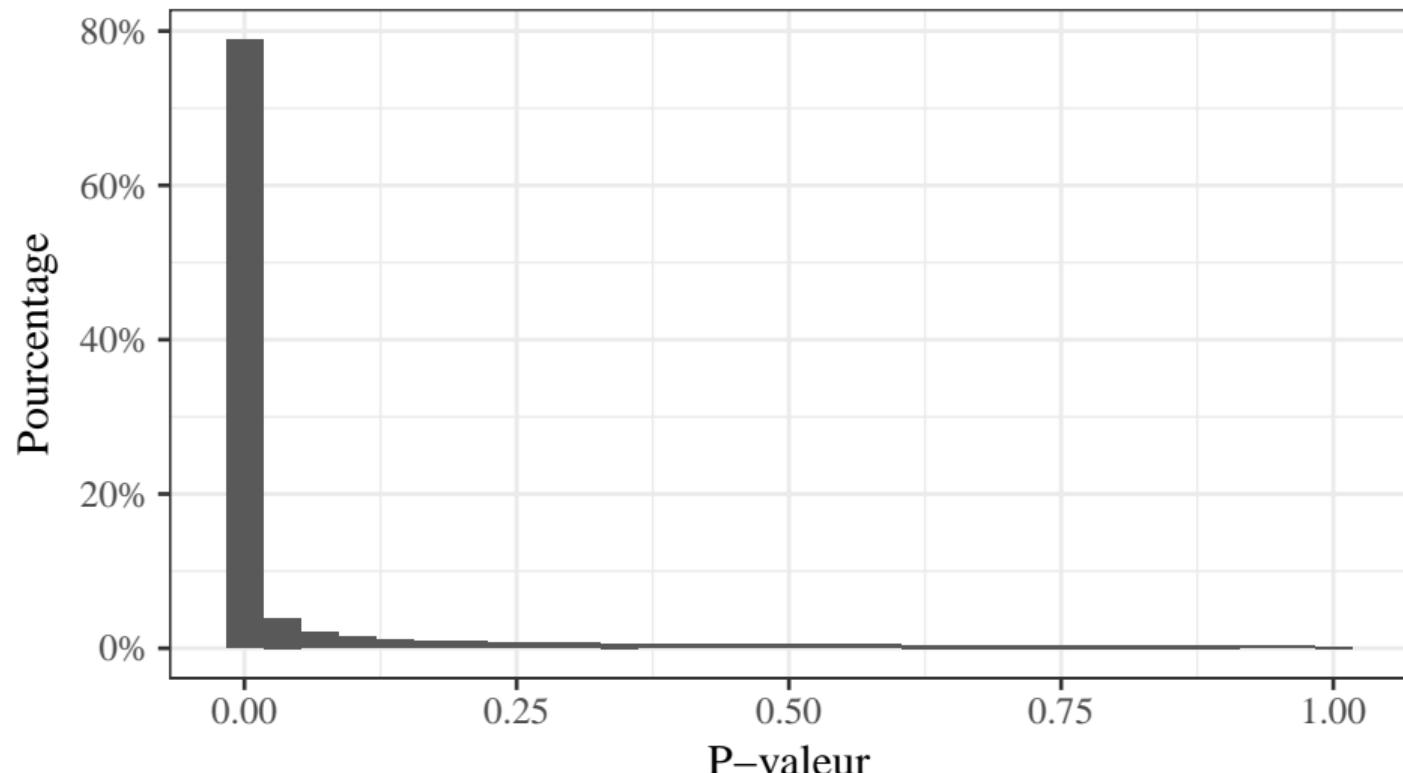
- ▶ 1409 individus de 14 populations
- ▶ 5397214 SNPs

Variable d'exposition

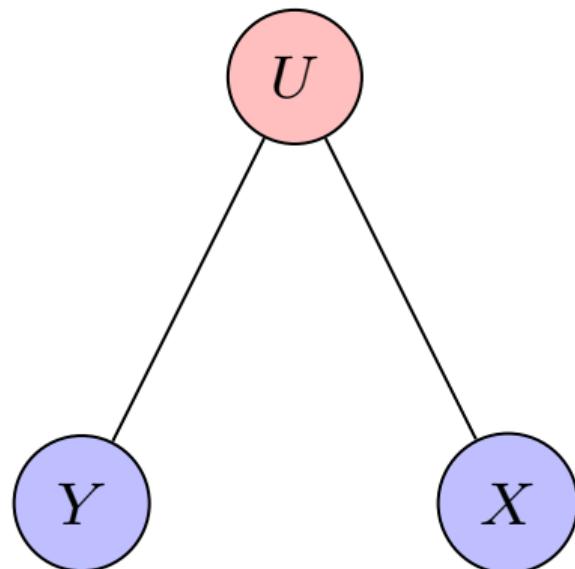
- ▶ données climatiques de la base WordClim
- ▶ première composante principale

On veut identifier les SNPs associés au climat

Histogramme des p -valeurs de l'étude d'association entre des données génétiques et un gradient climatique



Modèles mixtes à facteurs latents (LFMM)



$$\mathbf{Y} = \mathbf{XB}^T + \mathbf{UV}^T + \mathbf{E}$$

où

- ▶ \mathbf{U} est la matrice des variables latentes de taille $n \times K$
- ▶ \mathbf{V} est la matrice des effets des variables latentes $p \times K$
- ▶ \mathbf{B} est l'effet de la variable \mathbf{X} sur \mathbf{Y} de taille $p \times 1$
- ▶ \mathbf{E} est la matrice de bruit résiduel de taille $n \times p$

Méthodes d'estimation pour les modèles de régression à facteurs latents

Méthode	Modèle	Algorithme	Référence
sva-twostep	ACP et régression linéaire	moindres carrés ordinaire et SVD	Leek et J. D. Storey (2007)
sva-irw	<i>weighted</i> -ACP et régression linéaire	moindres carrés ordinaire et <i>weighted</i> -SVD	Leek et J. D. Storey (2008)
cate	analyse factorielle et régression linéaire	EM ou SVD et estimation des moindres carrés généralisée	Wang et al. (2017)
ridgeLFMM	factorisation matricielle avec régularisation L_2	SVD et estimation des moindres carrés régularisée en norme L_2	
lassoLFMM	factorisation matricielle avec régularisation L_1	<i>soft-thresholded</i> SVD et estimation des moindres carrés régularisée en norme L_1	

Estimateur des moindres carrés régularisé en norme L2

Fonction objectif

$$\mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \mathbf{B}) = \frac{1}{2} \left\| \mathbf{Y} - \mathbf{U}\mathbf{V}^T - \mathbf{X}\mathbf{B}^T \right\|_F^2 + \frac{\lambda}{2} \|\mathbf{B}\|_2^2$$

Estimateurs

1. On calcule

$$\hat{\mathbf{U}}\hat{\mathbf{V}}^T = \sqrt{\mathbf{P}_\lambda}^{-1} \text{svd}_K(\sqrt{\mathbf{P}_\lambda} \mathbf{Y})$$

où

$$\mathbf{P}_\lambda = \mathbf{Id}_n - (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{Id}_n)^{-1} \mathbf{X}^T \mathbf{X}$$

2. On calcule

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{Id}_d)^{-1} \mathbf{X}^T (\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T),$$

Estimateur des moindres carrés régularisé en norme L2

Si $\lambda \rightarrow 0$

- ▶ $P_\lambda = \text{Id}_n - (X^T X)^{-1} X^T X$
- ▶ P_λ n'est plus inversible
- ▶ **U** et **V** sont calculées sur le résidu de la régression linéaire de **Y** par **X**

Si $\lambda \rightarrow \infty$

- ▶ $P_\lambda = \text{Id}_n$
- ▶ **U** et **V** sont données par la SVD de rang K

Estimateur des moindres carrés régularisé en norme L2

Théorème 1

Pour λ strictement supérieur à zéro, les estimateurs des paramètres de LFMM régularisés en norme L_2 définissent un minimum global de la fonction objectif $\mathcal{L}_{\text{ridge}}$.

Idée de la preuve

$$\begin{aligned}\mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, \mathbf{B}) &\geq \mathcal{L}_{\text{ridge}}(\mathbf{U}, \mathbf{V}, (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{Id}_d)^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{UV}^T)) \\ &= \frac{1}{2} \left\| \sqrt{\mathbf{P}_\lambda} (\mathbf{Y} - \mathbf{UV}^T) \right\|_F^2\end{aligned}$$

Estimateur des moindres carrés régularisé en norme L1

Fonction objectif

$$\mathcal{L}_{\text{lasso}}(\mathbf{W}, \mathbf{B}) = \frac{1}{2} \left\| \mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T \right\|_F^2 + \mu \|\mathbf{B}\|_1 + \gamma \|\mathbf{W}\|_*$$

où

- ▶ \mathbf{W} est la matrice latente telle que

$$\mathbf{W} = \mathbf{U}\mathbf{V}^T$$

- ▶ $\|\mathbf{W}\|_*$ est la norme nucléaire

Estimateur des moindres carrés régularisé en norme L1

Algorithme de descente par blocs de coordonnées

On initialise

$$\hat{\mathbf{W}}_{t=0} = 0$$

$$\hat{\mathbf{B}}_{t=0} = 0$$

On alterne les étapes :

1. Calculer $\hat{\mathbf{B}}_t$ le point minimum de

$$\mathcal{L}'_{\text{lasso}}(\mathbf{B}) = \frac{1}{2} \|(\mathbf{Y} - \hat{\mathbf{W}}_{t-1}) - \mathbf{X}\mathbf{B}^T\|_F^2 + \mu \|\mathbf{B}\|_1 \quad (1)$$

2. Calculer $\hat{\mathbf{W}}_t$ le point minimum de

$$\mathcal{L}''_{\text{lasso}}(\mathbf{W}) = \frac{1}{2} \|(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_t^T) - \mathbf{W}\|_F^2 + \gamma \|\mathbf{W}\|_* \quad (2)$$

Estimateur des moindres carrés régularisé en norme L1

Théorème 2

L'algorithme d'estimation des moindres carré régularisé en norme L_1 converge vers un minimum global de la fonction objectif $\mathcal{L}_{\text{lasso}}$.

La preuve s'appuie sur les travaux de Tseng (2001)

Les principales hypothèses sont :

- ▶ $\mathbf{W}, \mathbf{B} \mapsto \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2$ est convexe et différentiable (terme d'attache aux données dans $\mathcal{L}_{\text{lasso}}$)
- ▶ $\mathbf{B} \mapsto \|\mathbf{B}\|_1$ est continue et convexe (terme de régularisation dans $\mathcal{L}_{\text{lasso}}$)
- ▶ $\mathbf{W} \mapsto \|\mathbf{W}\|_*$ est continue et convexe (terme de régularisation dans $\mathcal{L}_{\text{lasso}}$)

Tests d'hypothèse corrigés pour les facteurs de confusion (Price et al., 2006)

Pour chaque variable expliquée \mathbf{Y}_j

$$\mathbf{Y}_j = \hat{\mathbf{U}}\boldsymbol{\gamma}_j^T + \mathbf{X}\boldsymbol{\beta}_j + \mathbf{E}_j,$$

où $\hat{\mathbf{U}}$ est l'estimation de la matrice des variables latentes.

On teste l'hypothèse (test de Student)

$$H_0 : \beta_j = 0$$

On cherche une liste de candidats $\Gamma = \{1, \dots, J\}$ telle que

$$p(\beta_j = 0 | j \in \Gamma) = T$$

où T est le taux de fausse découverte souhaité.

On utilise la q -valeur John D Storey et Tibshirani, 2003

Comparaison des méthodes sur des données simulées à partir des données 1000Genomes

Données simulées

- ▶ On calcule les K premières composantes principales

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^T + \mathbf{\epsilon}$$

- ▶ On simule \mathbf{U}' et \mathbf{X}' en contrôlant leur corrélation.
- ▶ On calcule une nouvelle matrice telle que

$$\mathbf{Y}' = \mathbf{U}'\mathbf{V}^T + \mathbf{X}'\mathbf{B}'^T + \mathbf{\epsilon}$$

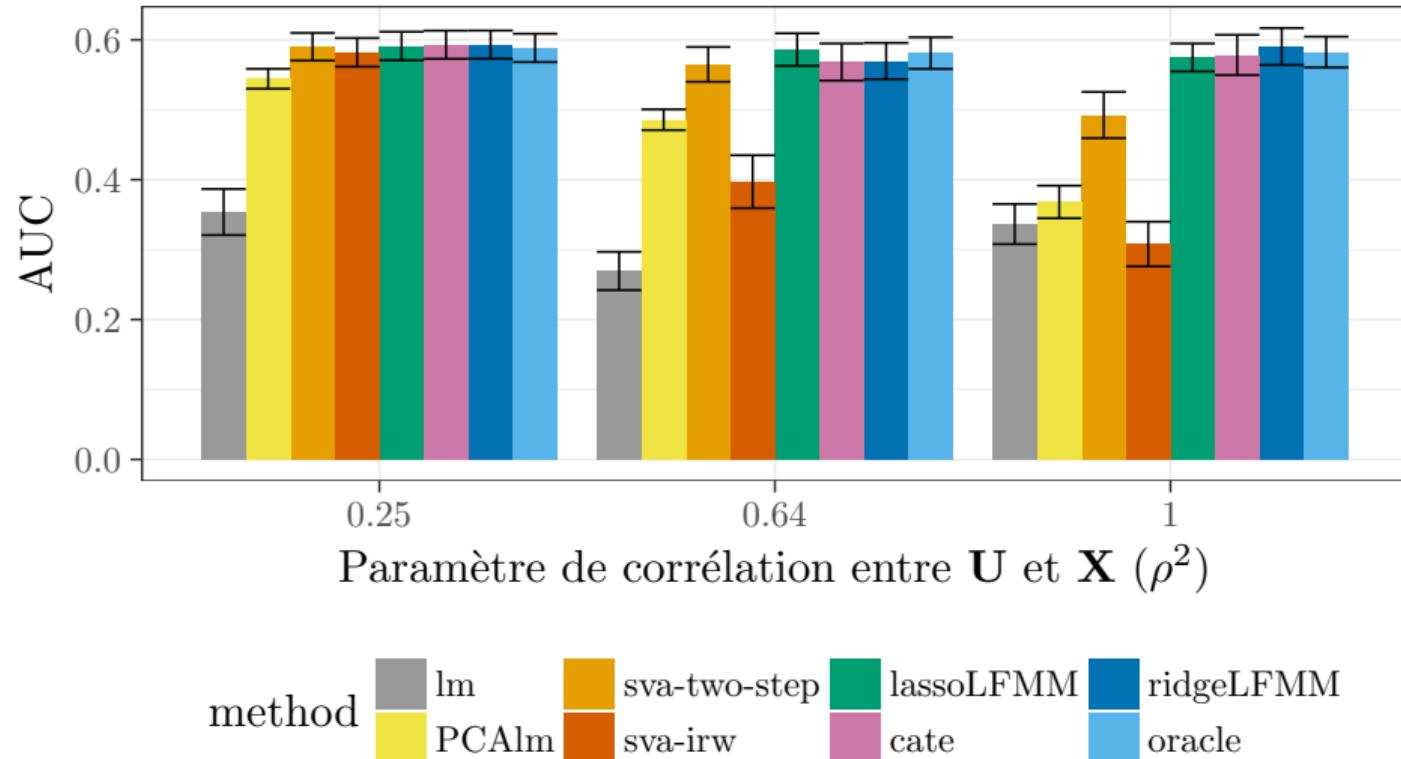
Critère

- ▶ AUC : aire sous la courbe précision (1 - FDR) \times rappel (puissance)

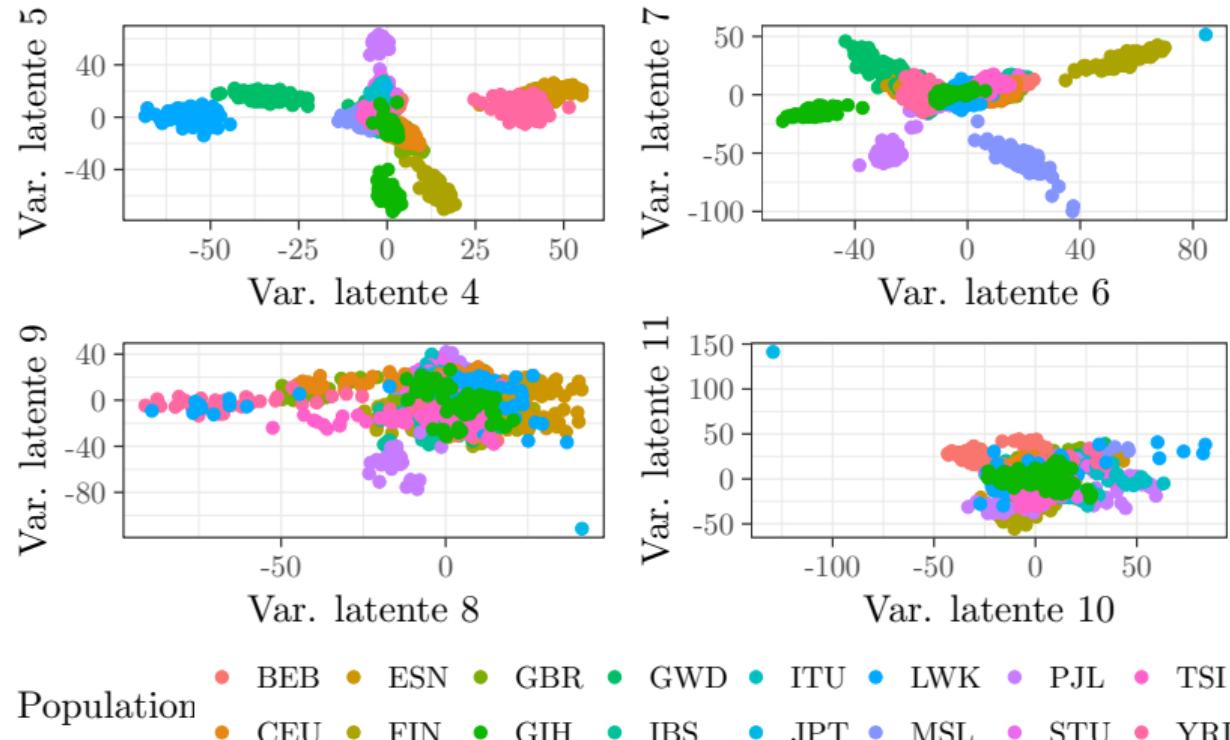
On compare les méthodes

- ▶ lm
- ▶ lmPCA
- ▶ sva-twostep
- ▶ sva-irw
- ▶ cate
- ▶ oracle
- ▶ ridgeLFMM
- ▶ lassoLFMM

Résultat de la comparaison des méthodes sur des données simulées



Choix de K pour l'étude d'association entre des données génétiques et un gradient climatique



Étude d'association entre des données génétiques et un gradient climatique

SNPs	DéTECTÉ par les méthodes	Description du phénotype
rs10908907	ridgeLFMM, cate	Alcoholism (heaviness of drinking)
rs10496731	lassoLFMM	Body Height
rs2472297	ridgeLFMM, cate, lassoLFMM	Caffeine metabolism
rs2256175	ridgeLFMM, cate, lassoLFMM	Cholesterol total
rs2472297	ridgeLFMM, cate, lassoLFMM	Coffee consumption (cups per day)
rs2278544, rs2322659	lassoLFMM	Congenital lactase deficiency
rs4954218	ridgeLFMM, cate, lassoLFMM	Corneal structure
rs882300	ridgeLFMM, cate, lassoLFMM	Electrocardiographic traits
rs882300	ridgeLFMM, cate, lassoLFMM	Electrocardiography
rs2256175	ridgeLFMM, cate, lassoLFMM	Giant cell arteritis
rs2256175, rs2104012, rs2853977	rs6085576, rs1983716,	Height
rs6430549	ridgeLFMM, cate, lassoLFMM	Hematocrit
rs2278544, rs2322659	lassoLFMM	Lactose intolerance
rs882300	ridgeLFMM, cate, lassoLFMM	Multiple sclerosis
rs1123848	ridgeLFMM, cate, lassoLFMM	Neuroblastoma
rs17158483	lassoLFMM	Obesity-related traits

Étude d'association entre des niveaux de méthylation de l'ADN et la polyarthrite rhumatoïde (EWAS)

Données (Liu et al., 2013)

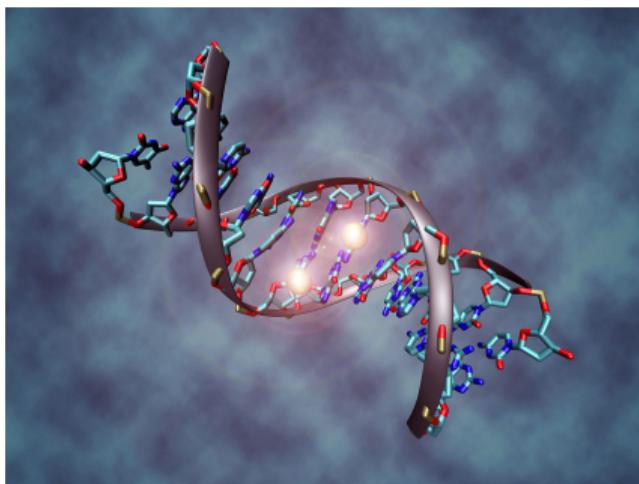
- ▶ \mathbf{Y} contient le niveau de méthylation de 485577 sites de l'ADN chez 699 individus (354 cas et 335 contrôles)
- ▶ \mathbf{X} représente la maladie polyarthrite rhumatoïde

Facteurs de confusion

- ▶ composition cellulaire
- ▶ âge
- ▶ genre
- ▶ consommation de tabac

Objectif

Trouver les sites de méthylation associés à la maladie polyarthrite rhumatoïde.



(Wikipedia, 2017)

Diagramme de Venn de l'étude d'association entre des sites de méthylation et la polyarthrite rhumatoïde (EWAS)

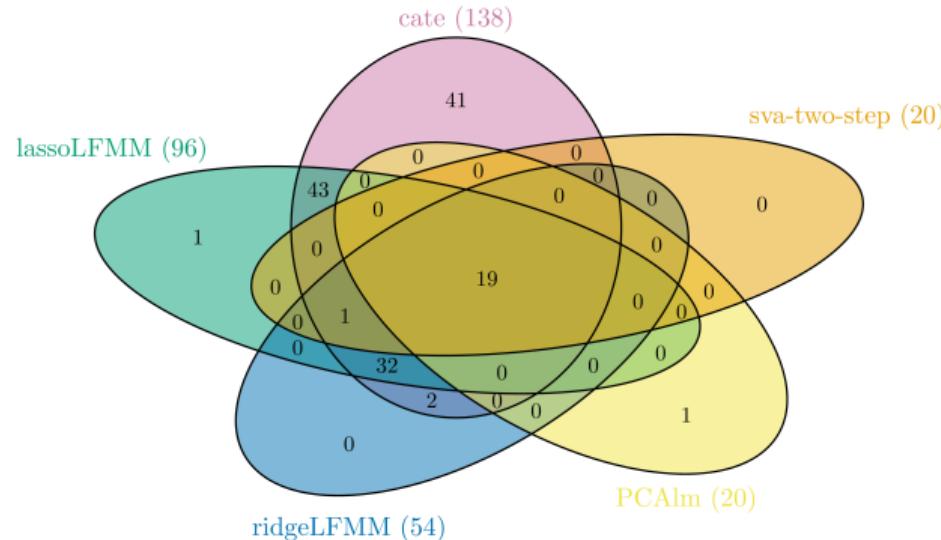


Diagramme de Venn des listes contrôlées à un taux de fausses découvertes de 1 %.

Sites de méthylation trouvés dans d'autres études (Rahmani et al., 2016 ; Zou et al., 2014) (EWAS)

ID	Chr	Position	Gene	PCAlm	lassoLFMM	cate	ridgeLFMM
cg16411857	16	57023191	NLRC5	9.2e-13	2.4e-12	6.6e-12	5.3e-12
cg07839457	16	57023022	NLRC5	1.9e-11	4.5e-11	1.1e-10	9.7e-11
cg05428452	6	32712979	HLA-DQA2	5.4e-11	4.6e-11	8.5e-11	8.8e-11
cg02508743	8	56903623	LYN	2.9e-08	2.7e-08	2.7e-08	2.8e-08
cg20821042	6	32709158	HLA-DQA2	6.5e-08	6.1e-08	9.6e-08	1.0e-07
cg13081526	6	32449961		1.5e-07	1.2e-07	2.0e-07	2.2e-07
cg18052547	6	32552547	HLA-DRB1	1.8e-07	1.8e-07	3.0e-07	3.1e-07
cg25372449	6	32490350	HLA-DRB5	2.5e-07	2.6e-07	4.5e-07	4.6e-07
cg02030958	13	110386267		4.0e-07	7.8e-08	6.0e-08	1.1e-07
cg16171858	3	58472734		4.6e-07	1.6e-07	2.7e-08	3.8e-08
cg03280622	8	145023013	PLEC1	4.7e-07	5.0e-09	5.8e-09	3.8e-08
cg24150157	19	51891210	LIM2	6.2e-07	3.1e-07	1.6e-07	2.1e-07
cg26244575	12	76354015		6.9e-07	2.7e-09	5.0e-10	4.2e-09
cg05370853	6	32606634	HLA-DQA1	7.1e-07	3.0e-07	3.3e-07	4.4e-07
cg14989316	10	80757927	LOC283050	7.3e-07	6.1e-08	7.8e-08	2.1e-07
cg17360552	6	32725332	HLA-DQB2	8.1e-07	6.1e-07	1.1e-06	1.2e-06
cg01373248	3	18480297	SATB1	8.1e-07	1.4e-07	1.1e-07	2.5e-07
cg26164488	2	64440295		9.3e-07	3.5e-09	1.6e-09	1.4e-08
cg05874806	2	102350276	MAP4K4	1.1e-06	1.1e-06	4.7e-07	5.6e-07

Section 4

Conclusions

tess3r

- ▶ Nouveau modèle pour l'estimation de la structure génétique des populations à partir de données génétique et géographique
- ▶ Deux algorithmes pour l'inférence des paramètres
- ▶ Même précision statistique que le logiciel bayésien TESS 2.3
- ▶ Algorithme 30 fois plus rapide que TESS 2.3
- ▶ Visualisation de la structure génétique de population dans l'espace
- ▶ Package R

- ▶ Deux nouvelles méthodes d'estimation des facteurs de confusion pour corriger les études d'association
- ▶ Résultats théoriques sur la convergence des algorithmes
- ▶ Sur les simulations même puissance que l'oracle
- ▶ Sur des données réelles les méthodes reposant sur le modèle mixte à facteurs latents découvrent plus d'association
- ▶ Sur les données réelles les associations découvertes peuvent varier largement entre les méthodes
- ▶ Package R

Perspectives

- ▶ Perspective de maintenance des logiciels
- ▶ données manquantes
- ▶ Construction des tests d'hypothèse (glm, modèle à effets fixes et aléatoires)
- ▶ Convergence statistique des estimateurs pour LFMM
- ▶ Utilisation de méthodes basées sur la factorisation matricielle à d'autres études : RNA-Seq, données méthylation au débit

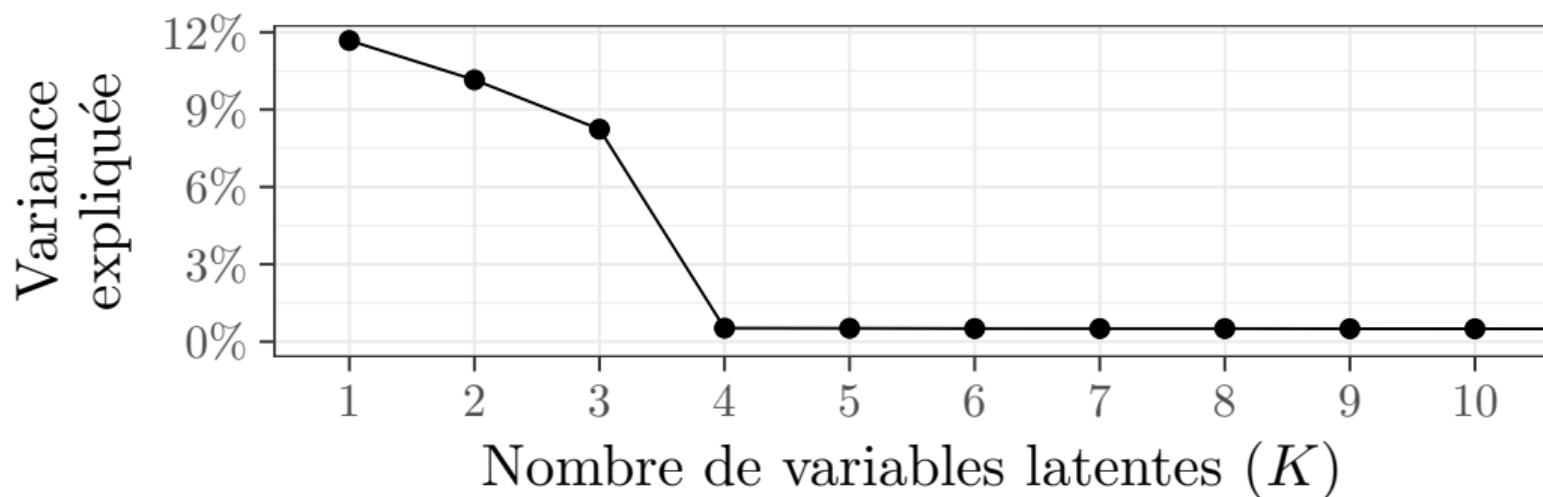
Merci de votre attention !

Choix du nombre de variables latentes

On projette \mathbf{Y} sur l'espace orthogonal à \mathbf{X} en prenant $\lambda = 0$

$$\mathbf{D}_0 \mathbf{Q}^T \mathbf{Y} = \mathbf{D}_0 \mathbf{Q}^T \mathbf{U} \mathbf{V}^T + \mathbf{D}_0 \mathbf{Q}^T \mathbf{E}.$$

On calcule les valeurs singulières pour visualiser la variance expliquée par chaque variable latente (scree plot).



GEAS : gènes candidats

