

Introduction aux études d'associations pangénomiques (GWAS)

Vincent Segura
INRA, UMR0588 BioForA

*Master Agrosciences, Environnement, Territoire, Paysage, Forêt
Parcours BICG*



- 1 Introduction
- 2 Le modèle linéaire mixte
- 3 Avancées méthodologiques

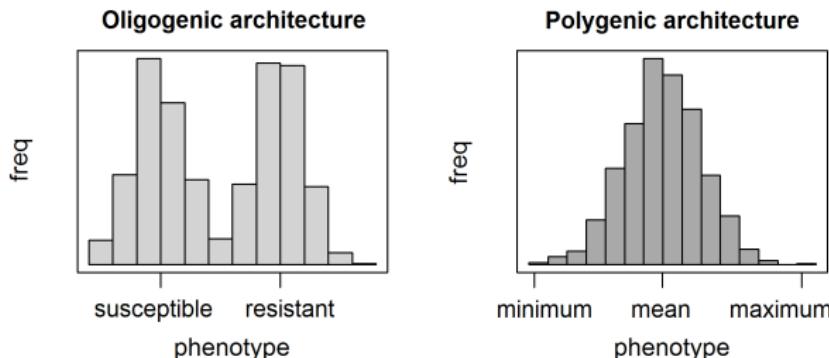
1 Introduction

2 Le modèle linéaire mixte

3 Avancées méthodologiques

Contexte

- Questions de recherche :
 - Quelle est l'**architecture génétique** de caractères d'intérêt agronomique, écologique ?



- Quelles **régions génomiques** contrôlent leur variabilité ?
- Application : Sélection assistée par marqueurs et/ou génomique diagnostic medical

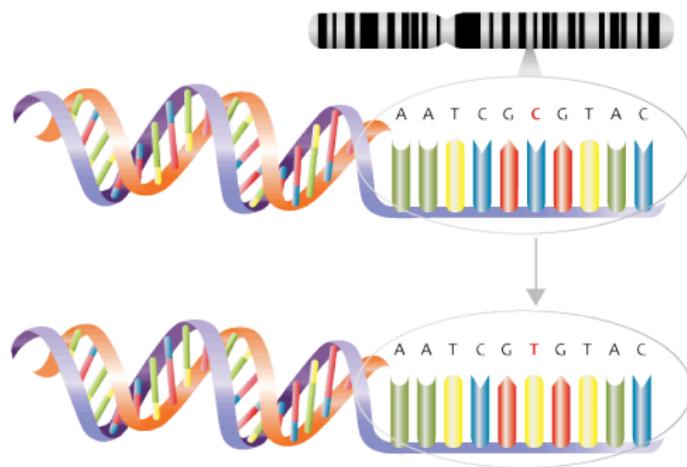
Définition

- Génétique d'association : Association entre **phénotype** (caractère d'intérêt) et **génotype** à des marqueurs moléculaires (SNP) au sein d'une **population**

Population		Genotype			Phenotype	
Individual		SNP1	SNP2	...	SNPi	Trait
1		CC	AA	...	CC	-1.6598
2		CT	AG	...	GC	0.4594
3		CC	AA	...	GG	-1.3315
...	
N		TT	GG	...	CC	0.9483

Rappel sur la notion de SNP

- Un **SNP** (single-nucleotide polymorphism) ou polymorphisme d'un seul nucléotide est la variation (polymorphisme) observée pour **une paire de bases le long du génome au sein d'une population**



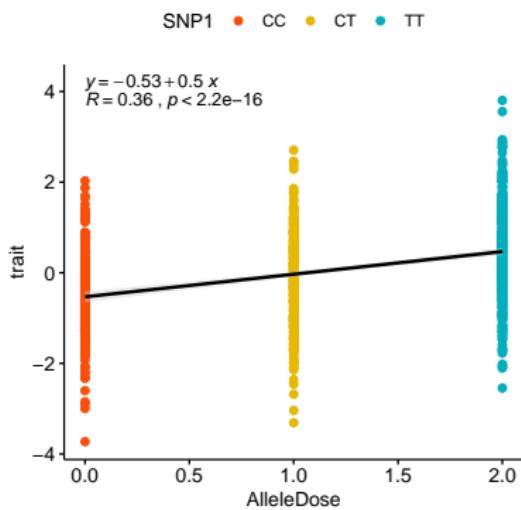
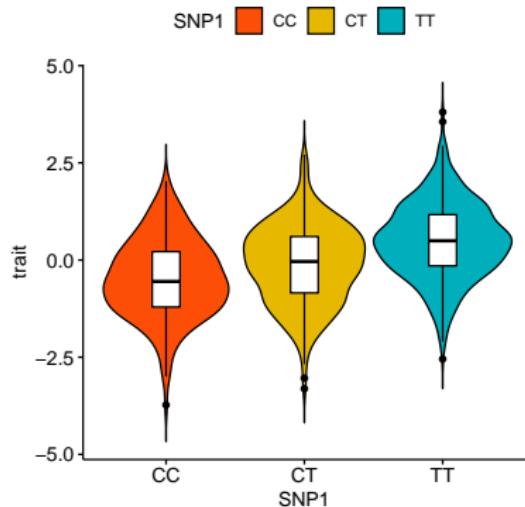
Définition

- Génétique d'association : Association entre **phénotype** (caractère d'intérêt) et **génotype** à des marqueurs moléculaires (SNP) au sein d'une **population**

Population		Genotype			Phenotype	
Individual		SNP1	SNP2	...	SNPi	Trait
1	CC	AA	...	CC	-1.6598	
2	CT	AG	...	GC	0.4594	
3	CC	AA	...	GG	-1.3315	
...
N	TT	GG	...	CC	0.9483	

Définition

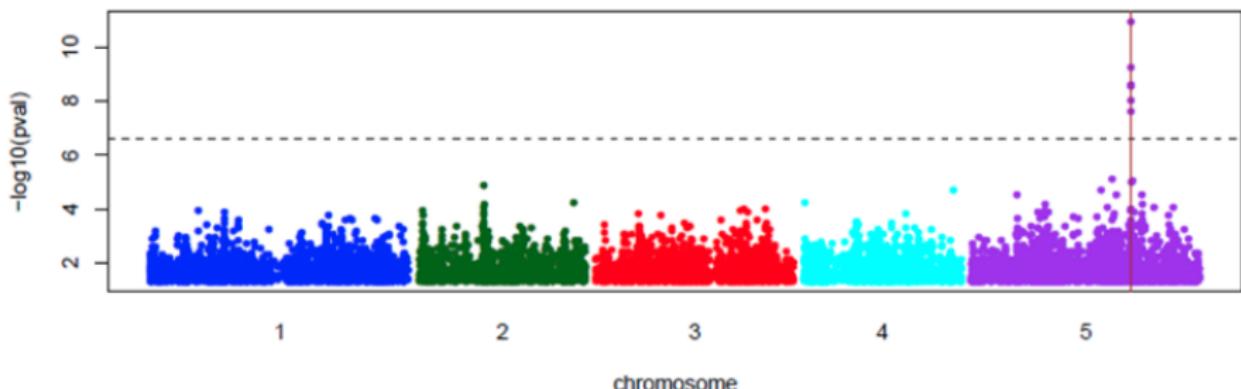
- Génétique d'association : Association entre **phénotype** (caractère d'intérêt) et **génotype** à des marqueurs moléculaires au sein d'une population



Mise en œuvre pratique

- **Phénotypage** : mesure d'un caractère d'intérêt
- **Génotypage** de (très) nombreux SNP (puce de génotypage ou séquençage)
- Modèle de **regression** linéaire simple répété pour chaque SNP entre son génotype et le caractère d'intérêt ⇒ **p-valeurs**
- Correction pour **tests multiples** (e.g. Bonferroni)
- Identification de marqueurs dont l'effet sur le caractère est **significatif**

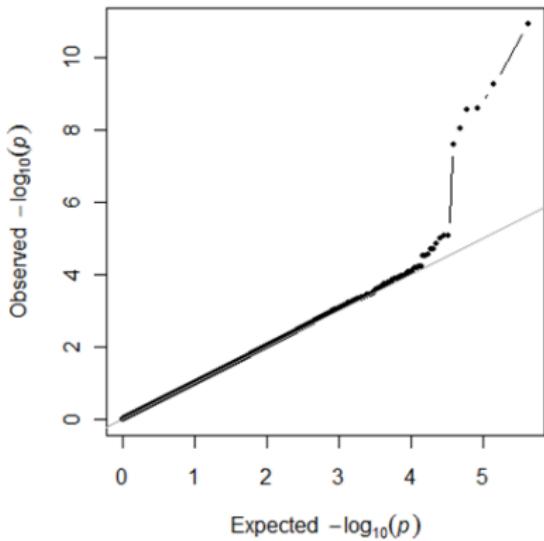
Le Manhattan plot



- Chaque **point** correspond à un **test** entre génotype et phénotype
- Les points sont ordonnés sur l'axe des x selon la **localisation génomique** des SNP
- L'axe des y représente la **p-valeur** en $-\log_{10}$
- La ligne horizontale en pointillés représente le **seuil de significativité** après correction pour tests multiples

Le QQ-plot

quantile-quantile p-value plot



- Distribution des **p-valeurs observées** (axe des y) contre celles **attendues** (axe des x) selon une loi uniforme
- Généralement en échelle logarithmique ($-\log_{10}$)

Le concept de déséquilibre de liaison (DL)

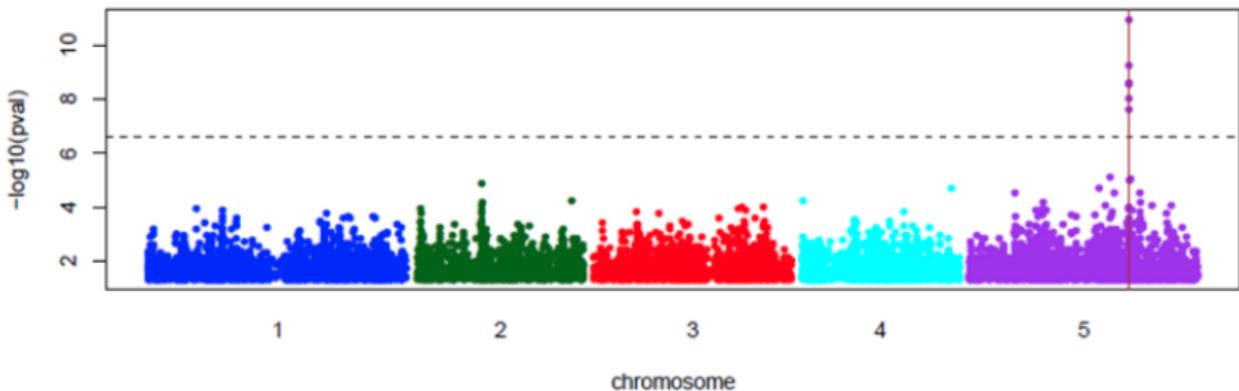
- Association non aléatoire entre allèles à des locus différents

	Ind1		Ind2		Ind3		Ind4		Ind5		Ind6		Ind7		Ind8		Ind9		Ind10	
SNP1	C	G	G	C	G	G	C	C	C	G	G	C	C	C	C	G	G	C	G	G
SNP2	A	A	T	A	T	A	A	T	T	T	A	A	T	T	T	A	T	T	A	A
SNP3	C	C	G	C	G	C	C	G	G	G	C	C	G	G	G	C	G	C	C	C

⇒ SNP1 et SNP2 sont à l'équilibre, tandis que SNP2 et SNP3 sont en DL

- Généralement le long du génome les **marqueurs voisins** présentent du déséquilibre de liaison

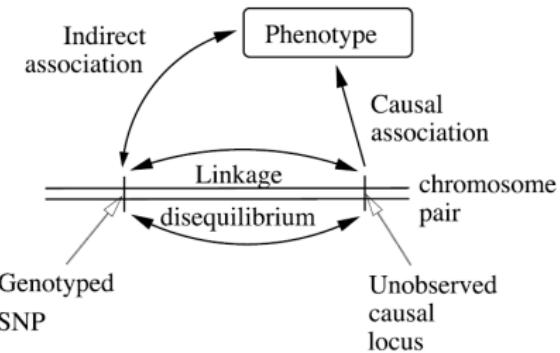
DL et Manhattan plot



- Le DL entre marqueurs voisins est généralement visible sur le Manhattan plot ⇒ Notion de **pic d'association**

La notion d'association indirecte

- Le concept de DL est central en génétique d'association

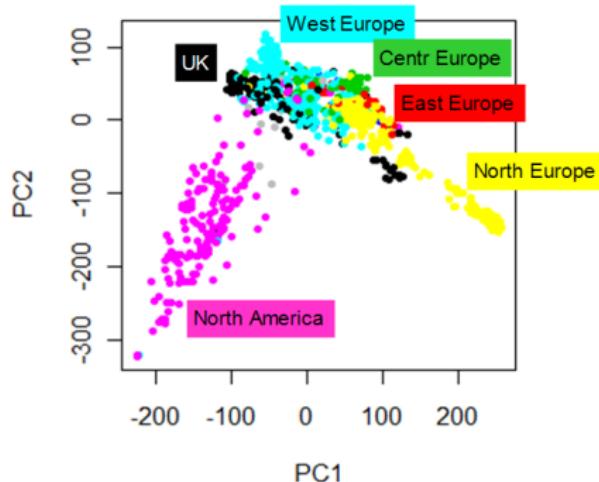


(Astle et Balding, Stat Sci 2009)

- Lorsqu'une association est détectée, le polymorphisme n'est pas nécessairement causal... c'est la notion d'association indirecte
- Le DL et la densité de génotypage conditionnent le succès d'une étude d'association

Le problème de la structure des populations

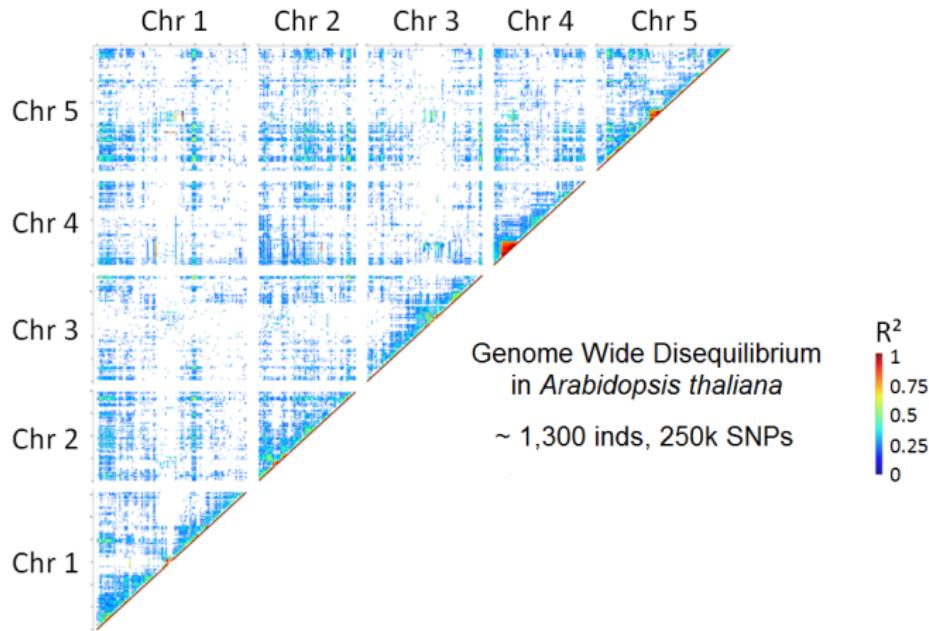
- Les individus ne sont pas indépendants !



Arabidopsis thaliana
1 179 inds
250k SNP
⇒ Structuration en sous-populations
(Horton et al., Nat Gen 2012)

- Les fréquences alléliques varient entre les sous-populations et génèrent du DL à longue distance

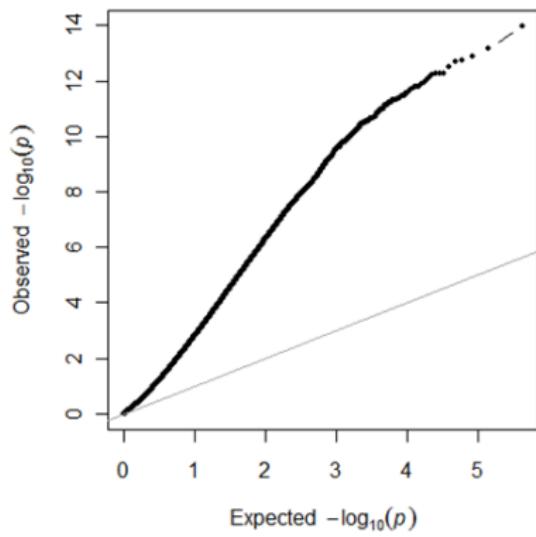
DL à l'échelle du génome chez *A. thaliana*



Conséquences pour les études d'association

- Le test statistique d'association est **biaisé**

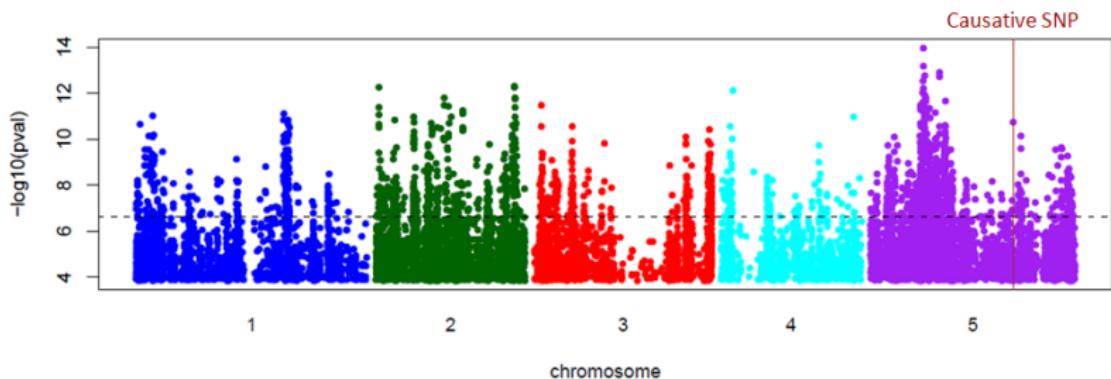
QQ-Plot
GWAS pour un phénotype simulé



Conséquences pour les études d'association

- Le test statistique est **biaisé**
- Le taux de **faux-positifs** est élevé

Manhattan-Plot
GWAS pour un phénotype simulé



1 Introduction

2 Le modèle linéaire mixte

3 Avancées méthodologiques

Modélisation statistique de l'effet des gènes

- Le modèle "marqueur par marqueur" n'est pas approprié pour les caractères quantitatifs
- Modèle infinitésimal de R.A. Fisher (1918) : les caractères quantitatifs sont contrôlés par un nombre infini de gènes à effet faible

$$Y_i = \mu + \sum_{j=1}^m Z_{ij} u_j + \epsilon_i,$$

où Z_{ij} est le génotype du i^{eme} individu pour le j^{eme} parmi les m polymorphismes causaux, $u_j \sim N(0, \sigma_g^2/m)$ et $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

- Le modèle infinitésimal peut s'écrire sous la forme d'un **modèle linéaire mixte** :

$$Y_i = \mu + g_i + \epsilon_i, \text{ où } \mathbf{g} \sim N\left(0, \sigma_g^2 \frac{\mathbf{Z}\mathbf{Z}^\top}{m}\right)$$

Le modèle linéaire mixte pour la génétique d'association

- Il semble ainsi "naturel" d'utiliser le modèle linéaire mixte en génétique d'association pour des caractères quantitatifs
- Cela consiste à intégrer un terme **polygénique** au modèle de détection d'association (Yu *et al.*, 2002)

$$Y_i = \beta_0 + x_{ij}\beta_j + g_i + \epsilon_i, \text{ où } \mathbf{g} \sim N(0, \sigma_g^2 \mathbf{K}) \text{ & } \epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I})$$

- En pratique, les polymorphismes causaux sont inconnus, et $\mathbf{Z}\mathbf{Z}^\top/m$ est approximé par une **matrice d'apparentement** typiquement estimée à partir de tous les marqueurs disponibles, e.g. $\mathbf{K} = \mathbf{W}\mathbf{W}^\top/m'$, où \mathbf{W} la matrice normalisée de génotypes composée de m' marqueurs
- Problème: **Temps de calcul** long, infaisable avec les outils classiques lorsque m' est très grand

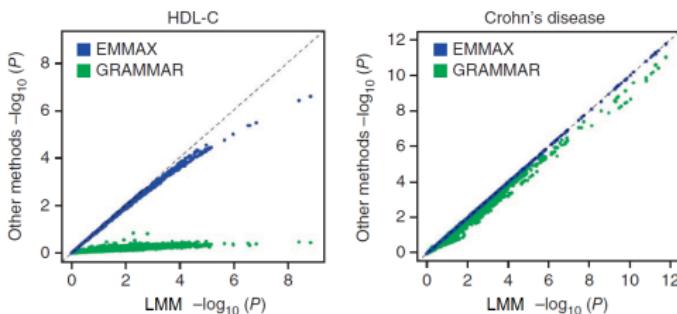
Modèle linéaire mixte pour les GWAS

- Méthodes approximatives :

- GRAMMAR (Aulchenko *et al.*, 2007): régression sur résidus d'un modèle linéaire mixte
- P3D (Zhang *et al.*, 2010) & EMMAX (Kang *et al.*, 2010):

Hypothèse : $\delta = \sigma_\epsilon^2 / \sigma_g^2$ varie peu au cours d'une GWAS

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \Rightarrow \mathbf{y} = \mathbf{X}\beta + \boldsymbol{\eta}, \text{Var}(\boldsymbol{\eta}) \propto \sigma_g^2 \mathbf{K} + \sigma_\epsilon^2 \mathbf{I}$$



(Zhou & Stephens, 2012)

Modèle linéaire mixte pour les GWAS

- Méthodes **approximatives** :

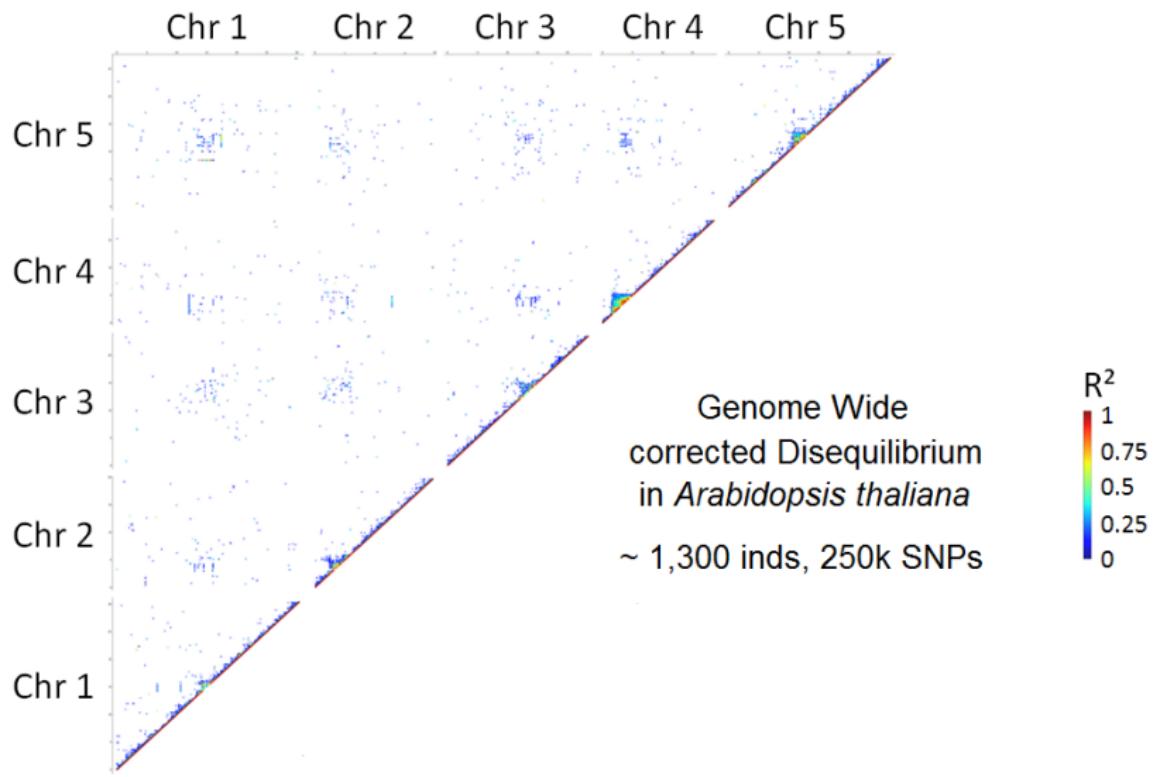
- GRAMMAR (Aulchenko *et al.*, 2007): régression sur résidus d'un modèle linéaire mixte
- P3D (Zhang *et al.*, 2010) & EMMA (Kang *et al.*, 2010):

- Méthodes **exactes** :

- FaST-LMM (Lippert *et al.*, 2011)
- GEMMA (Zhou & Stephens, 2012)

Méthodes	Temps de calcul	
	HDL-C	Crohn's disease
Exactes	EMMA	~ 9 d
	GEMMA	33 min
Approximatives	FaST-LMM	6.8 h
	EMMAX	44 min
	GRAMMAR	1.6 min

Le modèle mixte diminue le DL à longue distance

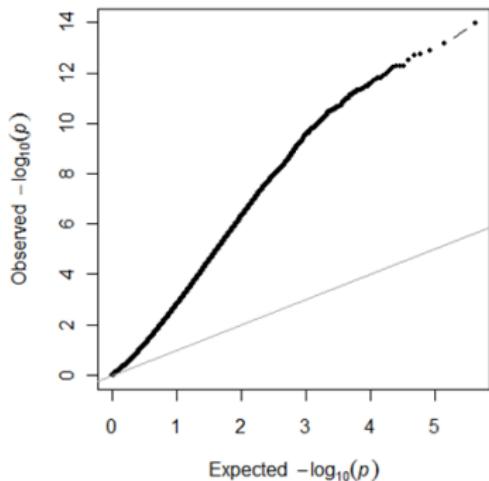


Le modèle mixte diminue le biais statistique

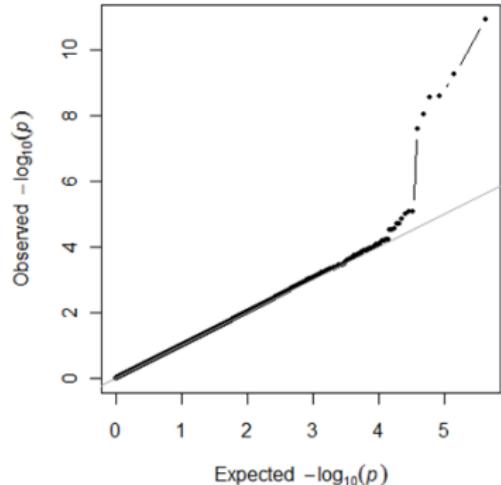
QQ-Plots

GWAS pour un phénotype simulé

Régression linéaire



EMMAX

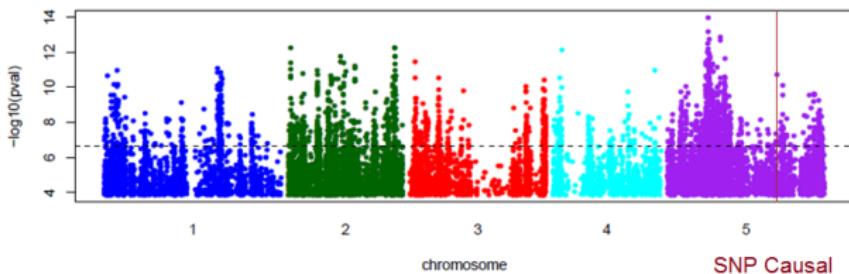


Le modèle mixte diminue le taux de faux-positifs

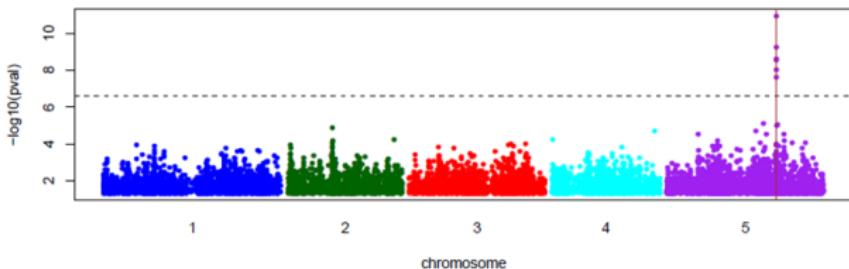
Manhattan Plots

GWAS pour un phénotype simulé

Régression
Linéaire



Modèle
Linéaire
Mixte

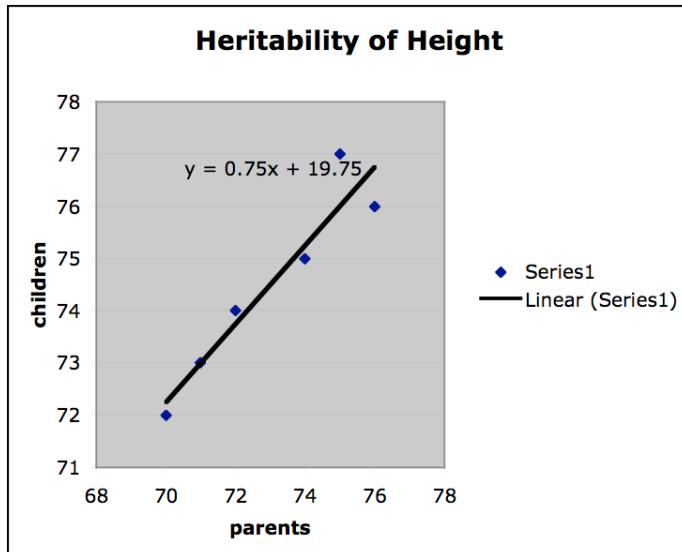


En résumé, la GWAS :

- Permet de tester l'association entre **génotype** et **phénotype** pour un (très) grand nombre de SNP répartis le long du génome
- Peut être réalisé dans des divers types de **populations**, notamment des populations naturelles dans lesquelles les individus ont pu accumuler un certain nombre de **recombinaisons**
- Tire partie du **déséquilibre de liaison** entre marqueurs et allèles causaux
- Est sensible aux variations de **fréquence alléliques** et en conséquence au **DL à longue distance** typiquement causé par la **structure** sous-jacente de la population étudiée
- Ce problème peut-être pris en compte en modélisant l'**effet de tous les gènes** sur le caractère dans le modèle d'association

- 1 Introduction
- 2 Le modèle linéaire mixte
- 3 Avancées méthodologiques

O



La taille est héritable à 80-90%

U

•

•

j

U " h " u o " M V) k U h = # U V 8 U 456
" K " " U - " h '8 o = " M V) k U h = # U V 8 U 8 ‡

Références manquantes et additionnelles

- 'Œ' 'o' 'U' '8
 'V' '8 y

h '° O '— 'V '° k ') ' h ' V ' V
 'V ' k '8 ' 11

") K ' V ' k '8 ' 7

Avancées méthodologiques

- Le modèle mixte fonctionne bien mais le pouvoir de détection des GWAS reste **limité** et pourrait être amélioré
- Quelques pistes d'amélioration :
 - Considération de **plusieurs loci** dans le modèle de détection via l'utilisation de cofacteurs
 - Modification de la **matrice d'apparentement**
 - Prise en compte explicite des **variants rares**
 - Combinaison de **plusieurs phenotypes** dans un modèle mixte multivarié

1 Introduction

2 Le modèle linéaire mixte

3 Avancées méthodologiques

- Approche multi-locus
 - Modifications de la matrice d'apparentement
 - FaST-LMM-Select
 - MLM-LOCO
 - Prise en compte du DL dans l'estimation de K
 - Le modèle linéaire mixte compressé
- Variants rares
- Approche multi-caractères

Approche multi-locus : motivations

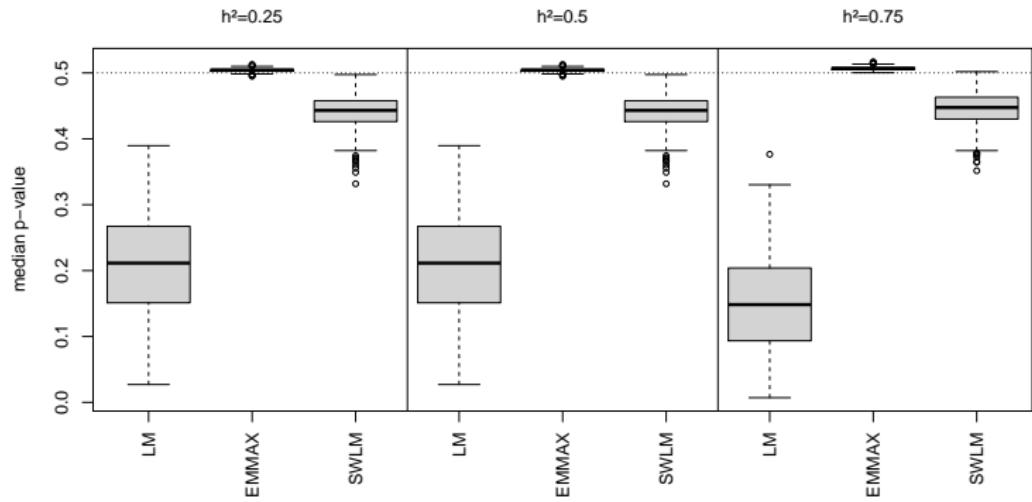
- Le modèle mixte fonctionne bien lorsque les caractères sont contrôlés par de **nombreux loci** à "faible" effet
- De plus, il est désormais **implémenté** de façon **efficace** et donc réellement applicable aux études d'association **pangénomiques**
- Mais, le modèle mixte ne permet pas de prendre en compte les **loci** dont l'**effet** est "relativement fort"
- Il devrait donc être possible d'améliorer les **performances** du modèle linéaire mixte dans les études d'associations, notamment lorsque l'**architecture** des caractères étudiés s'éloigne du modèle infinitésimal

Le modèle causal comme point de départ...

- Si les **locus causaux** étaient **tous connus** et génotypés, nous n'aurions pas besoin du modèle mixte. A la place, une **régression linéaire multiple** fonctionnerait très bien
- Le **problème** c'est que les locus causaux sont en grande majorité inconnus, puisque c'est eux que l'on recherche dans les études d'association...
- Ainsi, plusieurs approches **multi-locus** ont été proposées dans le contexte des études d'association, comme la **régression "stepwise"** ou la **régression pénalisée** (Hoggart *et al.*, 2008; Malo *et al.*, 2008; Croiseau & Cordell, 2009; Cho *et al.*, 2010; Ayers & Cordell; 2010)

Intérêt des approches multi-locus

- Il est d'ailleurs intéressant de noter que ces approches **multi-locus** permettent de prendre en compte une partie du **biais statistique** généralement attribué à la **structure des populations** (Setakis *et al.*, 2006; Pikkukookana & Sillanpää, 2009)

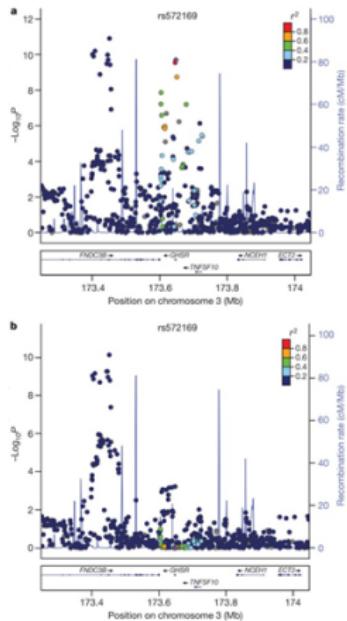


Intérêt des approches multi-locus

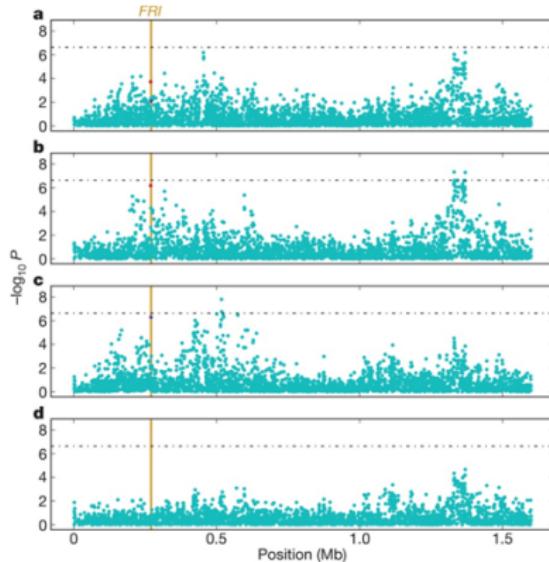
- Il est d'ailleurs intéressant de noter que ces approches **multi-locus** permettent de prendre en compte une partie du **biais statistique** généralement attribué à la **structure des populations** (Setakis *et al.*, 2006; Pikkukookana & Sillanpää, 2009)
- Dans les GWAS, utiliser des marqueurs "candidats" comme cofacteurs permet d'accroître **le pouvoir de détection** (Ma *et al.*, 2010)
- Cela peut aussi permettre de clarifier les **associations synthétiques** qui apparaissent typiquement dans les situations d'**hétérogénéité allélique** (Dickson *et al.*, 2010; Platt *et al.*, 2010)

Exemples d'utilisation de cofacteurs

Human Height
(Allen *et al.*, 2010)



FRI expression in *A. thaliana*
(Atwell *et al.*, 2010)



Approche multi-locus : comment ?

MLMM: Multi-Locus Mixed-Model (Segura et al., 2012)

- Problème : Impossible de tester **toutes les combinaisons**...
- Solution : Combiner la **régression multiple** avec le modèle mixte

① **Modèle null (étape 0)** : $\mathbf{y} = \beta_0 + \mathbf{Z}\mathbf{u} + \epsilon$

② **Inclusion forward (étapes forward 1 to i)** :

- GWAS scan avec EMMAX¹, en utilisant les $\hat{\sigma}_g^2$ & $\hat{\sigma}_\epsilon^2$ estimées dans le modèle mixte à l'étape forward $i - 1$
- Sélection de x_i , le SNP le plus significatif, pour l'inclure comme cofacteur dans le modèle mixte à l'étape i : $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$

③ **Élimination Backward (étapes backward 1 to i)** : Élimination itérative du cofacteur le moins significatif dans le dernier modèle mixte jusqu'au modèle null

¹ou n'importe qu'elle implémentation efficace du modèle mixte

Multi-Locus GWAS: comment?

Critère d'arrêt de l'inclusion forward et identification du meilleur modèle

- Critère d'arrêt de l'inclusion forward : $\hat{h}^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_\epsilon^2) \approx 0$
- Identification du meilleur modèle :
 - *BIC* (Schwarz, 1978)
 - *Extended BIC* : $EBIC = BIC + 2\log C_m^k$, où C_m^k représente l'étendue des possibles (Chen & Chen, 2010)
 - *mBonf* : Sélectionne le modèle le moins parcimonieux dont tous les cofacteurs sont significatifs (seuil 5% après correction de Bonferroni)

Évaluation des performances de MLMM avec des simulations

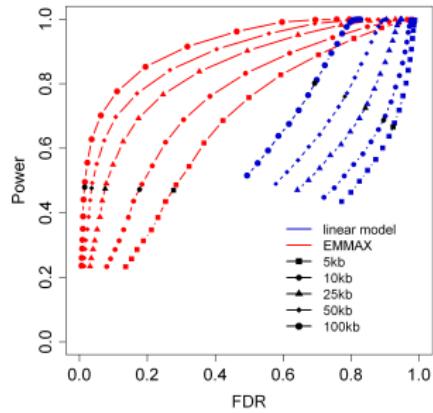
- Simulations effectuées à partir d'un jeu de données réel d'*A. thaliana* comprenant 1 326 accessions génotypées pour 214 051 SNP (Horton *et al.*, 2012)
- Modèle additif avec 100 SNP causaux échantillonnés au hasard et ayant des effets β tirés d'une distribution exponentielle
- Trois valeurs d'héritabilité au sens large : 0.25, 0.5, and 0.75
- Comparaison de 4 méthodes de détection:

	Simple-locus	Multi-locus
Régression Linéaire	LM	SWLM
Modèle mixte	EMMAX	MLMM

- SNP causaux conservés ou non dans le jeu de données

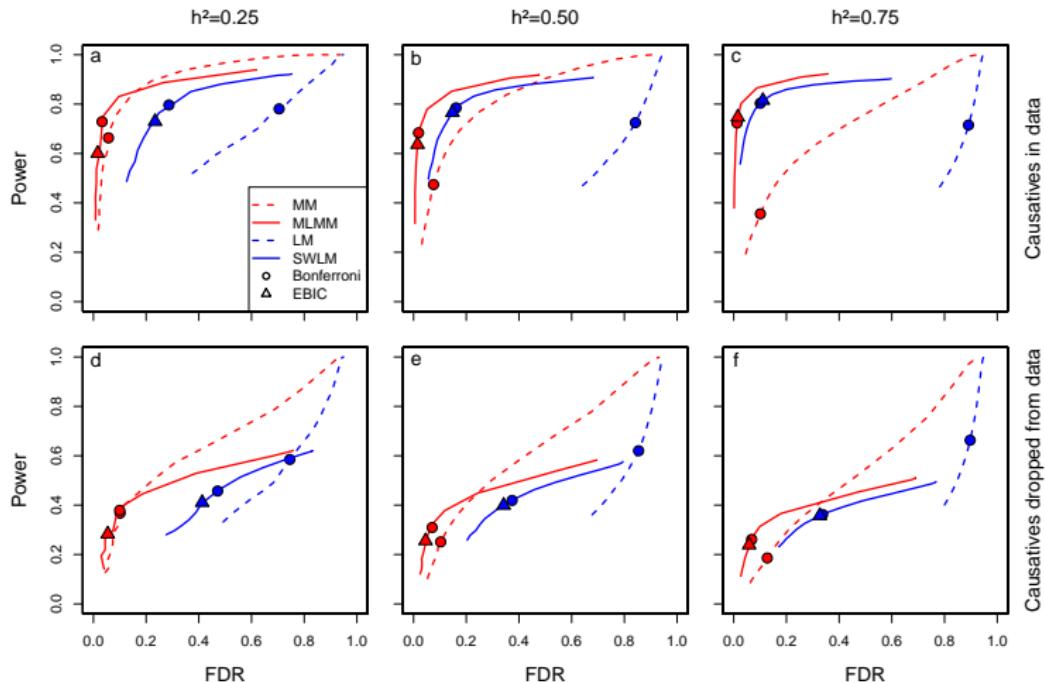
Évaluation des performances de MLMM avec des simulations

- Pouvoir de détection vs. FDR pour différentes p-values et fenêtres autour des SNP causaux



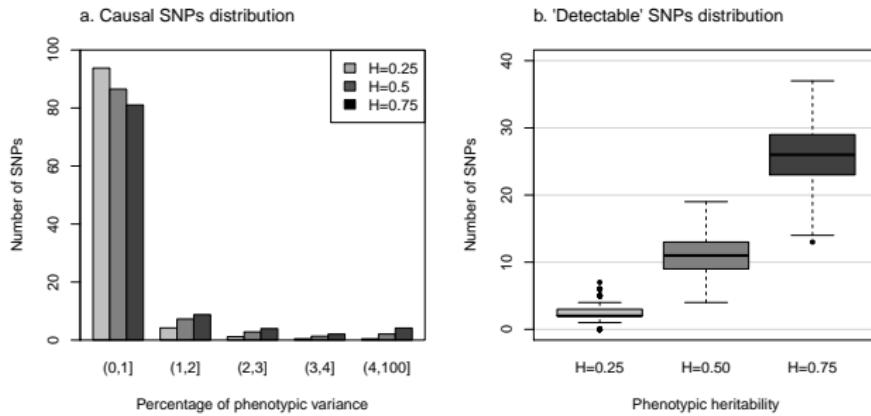
- Pouvoir de détection calculé par rapport aux SNP "détectables" car significatifs au seuil de 5% après correction de Bonferroni dans le modèle linéaire causal

Évaluation des performances de MLMM avec des simulations



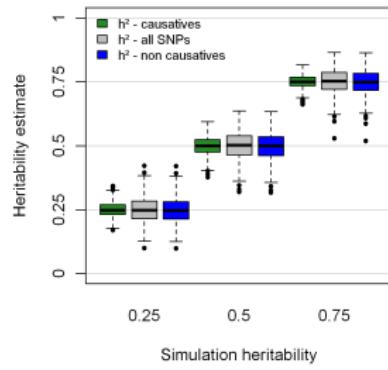
Évaluation des performances de MLMM avec des simulations

- Pour des seuils de significativité relativement stringents, MLMM présente des **performances similaires ou meilleures** que les autres méthodes testées
- Les performances de MLMM sont d'autant meilleures que l'héritabilité du caractère... ou plutôt que le **nombre de locus à effet "fort"** est élevé



Évaluation des performances de MLMM avec des simulations

- Pour des seuils de significativité relativement stringents, MLMM présente des **performances similaires ou meilleures** que les autres méthodes testées
- Les performances de MLMM sont d'autant meilleures que l'héritabilité du caractère... ou plutôt que le **nombre de locus à effet "fort"** est élevé
- Les critères de qualité des modèles (EBIC et mBonf) permettent d'obtenir un **faible FDR**
- Lorsque les SNP causaux ne sont pas dans le jeu de données les **performances de toutes les méthodes** sont **affectées** alors que les valeurs d'héritabilité estimées restent correctes



Application à des données réelles : [Na⁺] dans les feuilles d'*A. thaliana*

OPEN  ACCESS Freely available online

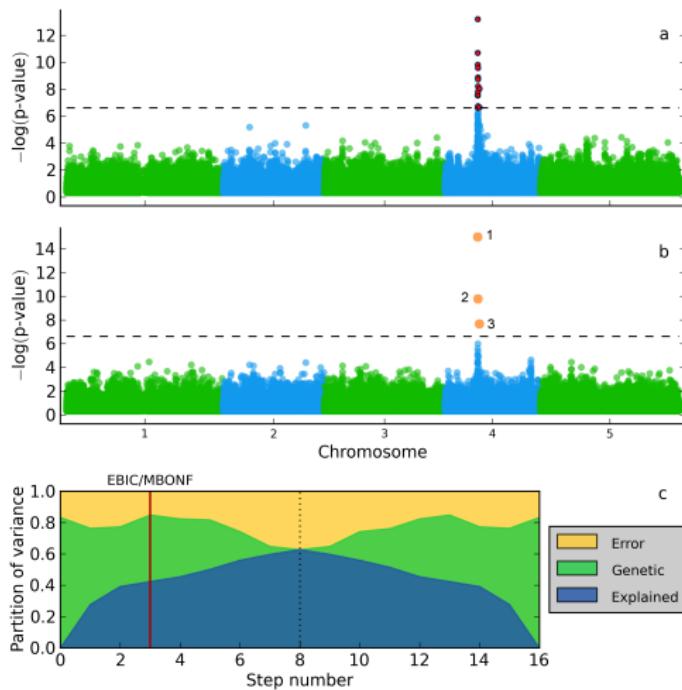
PLOS GENETICS

A Coastal Cline in Sodium Accumulation in *Arabidopsis thaliana* Is Driven by Natural Variation of the Sodium Transporter AtHKT1;1

Ivan Baxter^{1*}, Jessica N. Brazelton^{2*}, Danni Yu³, Yu S. Huang⁴, Brett Lahner², Elena Yakubova², Yan Li⁵, Joy Bergelson⁵, Justin O. Borevitz⁵, Magnus Nordborg⁶, Olga Vitek³, David E. Salt²

- 6 reps de 342 accessions en chambre de culture
- Le SNP le **plus significatif** est situé dans le premier exon du gène AtHKT1;1 qui code pour un **transporteur de sodium**
- L'**allèle (-)** de ce SNP est **sur-représenté** dans les **accessions** originaires de **régions côtières** ou dont les **sols sont salins**

Application à des données réelles : [Na²⁺] dans les feuilles d'*A. thaliana*

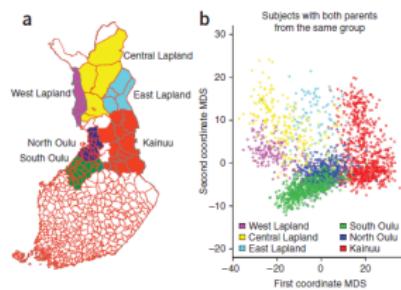


Application à des données réelles : Jeu de données NFBC (Humain)

Genome-wide association analysis of metabolic traits in a birth cohort from a founder population

Chiara Sabatti^{1,2,28}, Susan K Service^{3,28}, Anna-Liisa Hartikainen⁴, Anneli Pouta⁵, Samuli Ripatti⁶, Jae Brodsky², Chris G Jones^{3,7}, Noah A Zaitlen⁷, Teppo Varilo^{8,9}, Marika Kaakinen¹⁰, Ulla Sovio¹¹, Aimo Ruokonen¹², Jaana Laitinen¹³, Eveliina Jakkula⁶, Lachlan Coin¹¹, Clive Hoggart¹¹, Andrew Collins¹⁴, Hannu Turunen⁶, Stacey Gabriel¹⁵, Paul Elliott¹¹, Mark I McCarthy¹⁶⁻¹⁸, Mark J Daly^{15,19-23}, Marjo-Riitta Järvelin^{5,11,24}, Nelson B Freimer^{3,25,26} & Leena Peltonen^{6,15,27}

- North Finland Birth Cohort
- 5 000 inds; 360k SNP
- 9 caractères métaboliques
- Data ré-analysées avec EMMAX
(Kang *et al.*, 2010)



Application à des données réelles :

Jeu de données NFBC (Humain)

SNP	Chr.	Position	Gene	p-value		Previously identified by	
				EBIC	mBonf	Sabatti et al.	Kang et al.
Associated with mmol/l TG							
rs673548	2	21091049	APOB		5.1×10^{-8}	Y	Y
rs1260326	2	27584444	GCKR	1.5×10^{-10}	7.9×10^{-11}	Y	Y
rs10096633	8	19875201	LPL	1.6×10^{-8}	2.4×10^{-8}	Y	Y
Associated with mmol/l HDL							
rs1532085	15	56470658	LIPC	9.2×10^{-12}	8.0×10^{-12}	Y	Y
rs3764261	16	55550825	CETP	2.7×10^{-32}	3.7×10^{-23}	Y	Y
rs7499892	16	55564091	CETP		9.5×10^{-8}	N	N
rs255049	16	66570972	LCAT	1.3×10^{-8}	4.8×10^{-8}	Y	Y
rs1800961	20	42475778	HNF4A		1.5×10^{-7}	N	N
Associated with mmol/l LDL							
rs646776	1	109620053	CELSR2	4.2×10^{-16}	4.2×10^{-16}	Y	Y
rs693	2	21085700	APOB	7.1×10^{-12}	7.1×10^{-12}	Y	Y
rs11668477	19	11056030	LDLR	1.0×10^{-9}	1.0×10^{-9}	Y	Y
rs157580	19	50087106	TOMM40-APOE	2.2×10^{-17}	2.2×10^{-17}	Y	N
rs405509	19	50100676	TOMM40-APOE	1.3×10^{-12}	1.3×10^{-12}	N	N
Associated with mg/l CRP							
rs2369146	1	157934819	CRP	4.5×10^{-9}	2.8×10^{-9}	N	N
rs2794520	1	157945440	CRP	1.1×10^{-29}	6.6×10^{-30}	Y	Y
rs2650000	12	119873345	HNF1A	1.3×10^{-12}	1.0×10^{-12}	Y	Y
rs8106922	19	50093506	TOMM40-APOE		1.6×10^{-12}	N	N
rs439401	19	50106291	TOMM40-APOE		2.2×10^{-9}	N	N

MLMM: Bilan

- **Pouvoir de détection** potentiellement **plus élevé** que celui des approches simple-locus existantes
 - Détection de **nouvelles régions génomiques**
 - Mise en évidence de possibles cas d'**hétérogénéité allélique**
- Évaluation des pourcentages de variance expliquée et restant à expliquer ⇒ vers une meilleure compréhension de l'**architecture génétique** des caractères
- **Implémentation** dans 2 langages :
 - **Python** <https://github.com/bvilhjal/mixmogam/>
 - **R** <https://github.com/Gregor-Mendel-Institute/MultLocMixMod>

1 Introduction

2 Le modèle linéaire mixte

3 Avancées méthodologiques

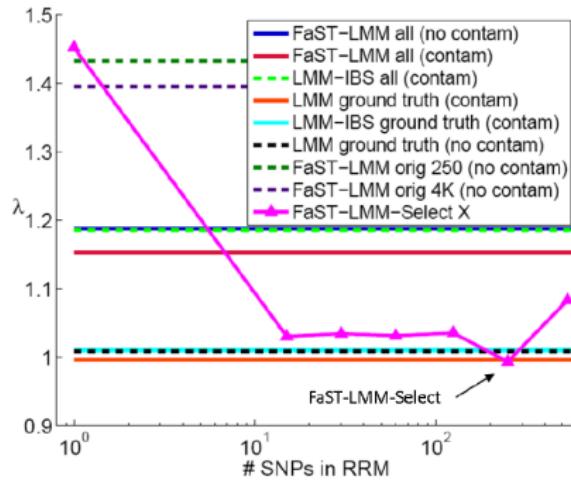
- Approche multi-locus
- Modifications de la matrice d'apparentement
 - FaST-LMM-Select
 - MLM-LOCO
 - Prise en compte du DL dans l'estimation de K
 - Le modèle linéaire mixte compressé
- Variants rares
- Approche multi-caractères

FaST-LMM-Select (Listgarten *et al.*, 2012)

- Motivés par la **ressemblance** entre le **modèle mixte** et la **régression multiple**, Listgarten *et al.* ont identifié **2 problèmes majeurs** dans l'utilisation du modèle mixte pour les études d'association :
 - ➊ **Dilution:** La **matrice de covariance (K)** de l'effet aléatoire polygénique ne devraient pas être construit à partir de tous les SNP, parceque la plupart d'entre-eux ne sont vraisemblablement pas en DL avec un locus causal. A la place, il faudrait utiliser un **sous-échantillon de SNP "pertinents"** (ideallement les loci causaux)
 - ➋ **Contamination proximale:** le **SNP testé** comme effet fixe dans le modèle devrait être **exclu de la matrice K**

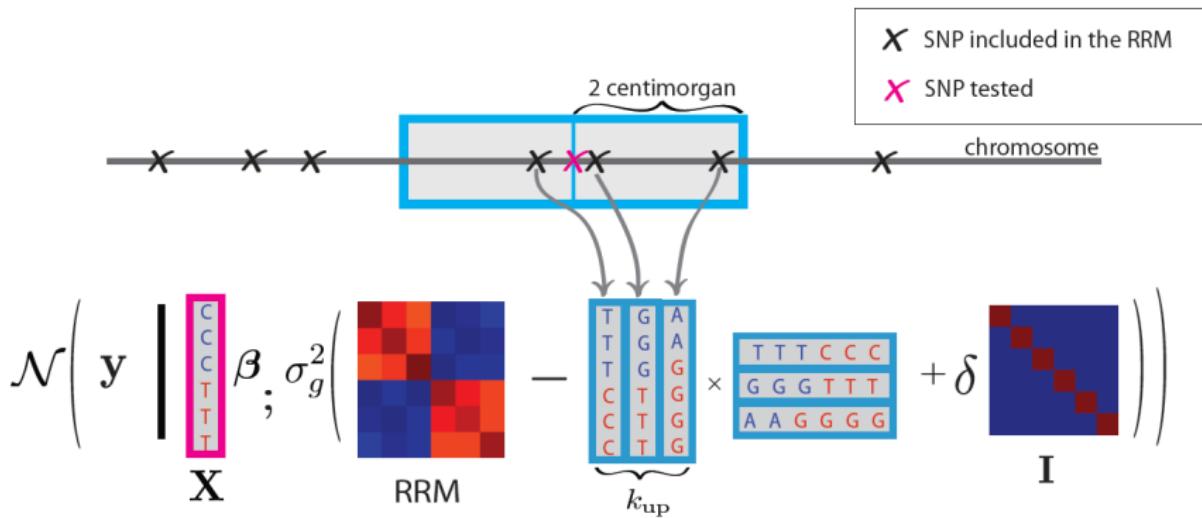
FaST-LMM-Select pour résoudre le problème de dilution

- Trier les SNP par ordre croissant de p-valeur en régression linéaire
- Construire des matrices K avec un nombre croissant de SNP (triés) jusqu'à atteindre le premier minimum de λ_{GC}



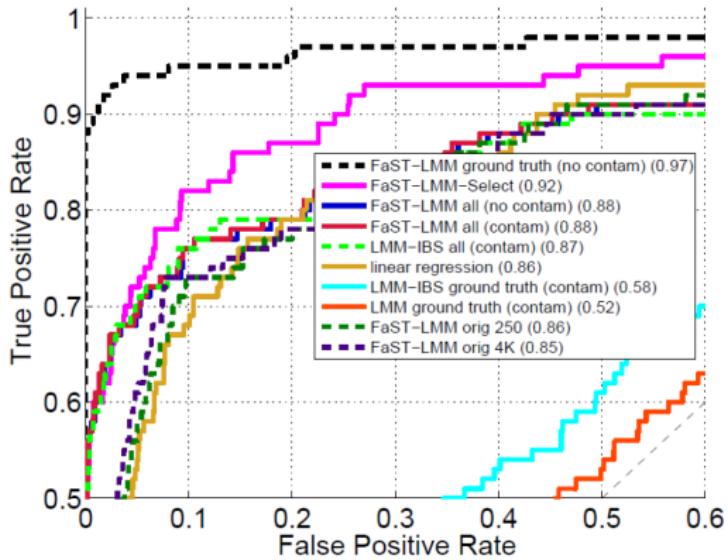
FaST-LMM-Select pour résoudre le problème de contamination proximale

- Retirer de la matrice K précédemment sélectionnée le SNP en cours de test ainsi que ceux qui lui sont proches (e.g. 2 cM or 2 Mb)



Performances de FaST-LMM-Select sur des données simulées

- 100,000 SNP simulés selon le modèle de Balding-Nichols; 3,000 inds; 2 pops; 100 SNP causaux sélectionnés de façon aléatoire, $H = 0.5$



Performances of FaST-LMM-Select sur des données réelles

- Données sur la maladie de Crohn (WTCCC, 2007); 14,925 inds (dont $\approx 2,000$ cas); 356,441 SNP

Table 1 | Comparison of calibration, power and computational costs on a GWAS of Crohn's disease

Algorithm	Algorithm parameters			Algorithm performance					
	SNP selection method	No. SNPs for matrix	Proximal contamination avoided?	λ_{GC}	No. false positives	No. false negatives	Runtime without speedup (min)	Runtime with speedup (min)	Memory usage (GB)
FaST-LMM-Select	Select	310	Yes	1.08	0	1	1.3×10^3	45	<1
FaST-LMM (all)	All	All	Yes	1.09	2	2	4.0×10^6	4,567	86
FaST-LMM (orig 310)	Equally spaced	310	Yes	1.26	9	1	1.1×10^3	6	<1
FaST-LMM (orig 4,000)	Equally spaced	4,000	Yes	1.17	5	1	2.1×10^5	30	2
Traditional	All	All	No	0.97	2	6	4.2×10^1	NA	45

The original version of FaST-LMM, which used equally spaced SNPs to estimate genetic similarity, was evaluated using 310 SNPs (the same number used by FaST-LMM-Select) and 4,000 SNPs (as used in the original version of FaST-LMM (ref.2)). The five algorithms yielded substantially different P values (Supplementary Fig. 1), which in turn led to different SNPs being deemed significant (using the P value threshold of 5×10^{-7} (ref. 6)). Previous studies were used to determine the gold standard in order to label the false positive and false negative loci (Supplementary Table 1). Details of the analysis are described in the Supplementary Methods.

FaST-LMM-Select implémentation et disponibilité

- Les options de **FaST-LMM-Select** sont inclus dans le **logiciel FaST-LMM**
- FaST-LMM est implémenté en **C++ & Python**
- Exécutables pour **Windows** et **Linux** et **code source** :
<http://research.microsoft.com/en-us/um/redmond/projects/mscompbio/fastlmm/>
- Version **Python** : <https://github.com/microsoftgenomics/fast-lmm>

MLM-LOCO : "Leave One Chromosome Out"

- Extrait du manuel de FaST-LMM :

"Excluding the SNP you are testing from the genetic similarity matrix and also those SNPs in close proximity to it in a naïve way is **extremely computationally expensive**.

A **computationally efficient** approach for performing the **exclusion** is to use a **similarity matrix** computed from **all but chromosome i** when testing SNPs on **chromosome i**."

- Proposé par Rinten et al., 2014 & Yang et al., 2014
- Implementé dans **GCTA** <http://cnsgenomics.com/software/gcta/>

Prise en compte du DL dans l'estimation de K

- Rincent *et al.*, 2014 ont analysé la **puissance** de détection en GWAS le long du génome dans des panels de diversité chez le maïs et ont identifié des **chutes de puissance** dans les **régions présentant un fort DL**
- Pour prendre en compte ce problème, ils ont proposé de **prendre en compte le DL** entre les marqueurs en les **pondérant** par leur **contribution** à l'apparentement
- Une approche similaire avait déjà été proposée par Speed *et al.* (2012) pour améliorer les estimation d'héritabilité génomique $h^2 \Rightarrow$ implementé dans LDAK <http://www.ldak.org/>

Table 3 Number of QTL detected with the three statistical models in each panel at different thresholds assuming different genetic models (50 or 100 QTL)

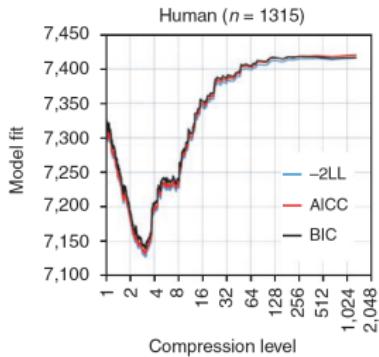
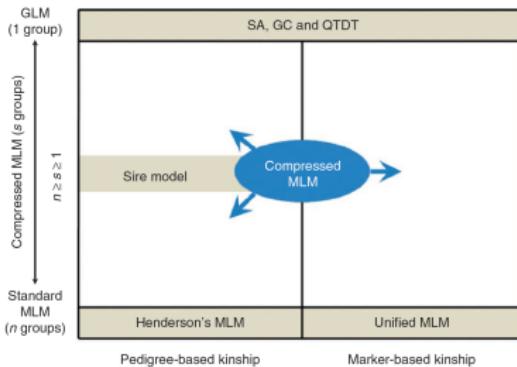
Panel	Nb QTL	Approach	T^a		$10 \times T$		$100 \times T$	
			Average	SD	Average	SD	Average	SD
C-K	50	M_K_{Freq}	1.4	1.0	2.5	1.2	4.2	1.6
C-K	50	M_K_{Chr}	1.7	1.1	3.2	1.5	4.9	1.7
C-K	50	M_K_{LD}	1.6	1.1	2.6	1.3	4.3	1.7
C-K	100	M_K_{Freq}	0.3	0.5	0.9	0.8	2.1	1.2
C-K	100	M_K_{Chr}	0.5	0.7	1.3	1.0	2.8	1.5
C-K	100	M_K_{LD}	0.4	0.6	1.1	0.9	2.3	1.4
CF-Dent	50	M_K_{Freq}	1.2	1.0	2.2	1.3	3.6	1.3
CF-Dent	50	M_K_{Chr}	2.1	1.4	3.4	1.5	5.3	1.6
CF-Dent	50	M_K_{LD}	1.3	1.1	2.5	1.3	4.1	1.4
CF-Dent	100	M_K_{Freq}	0.3	0.6	0.9	0.9	2.0	1.4
CF-Dent	100	M_K_{Chr}	0.8	1.0	1.7	1.3	3.4	1.7
CF-Dent	100	M_K_{LD}	0.5	0.7	1.0	1.1	2.4	1.4
CF-Flint	50	M_K_{Freq}	1.4	1.0	2.4	1.1	3.7	1.2
CF-Flint	50	M_K_{Chr}	1.8	1.2	3.0	1.0	4.5	1.3
CF-Flint	50	M_K_{LD}	1.4	0.9	2.4	1.1	4.0	1.3
CF-Flint	100	M_K_{Freq}	0.3	0.6	0.8	0.9	1.9	1.1
CF-Flint	100	M_K_{Chr}	0.6	0.8	1.4	1.2	2.8	1.4
CF-Flint	100	M_K_{LD}	0.4	0.7	1.0	1.1	2.1	1.3

We computed the average and the standard deviation of the number of QTL detected in the 100 runs of simulation.

^a Significance threshold T was set considering a type I risk of 5% with a Bonferroni correction assuming 40 000 tests.

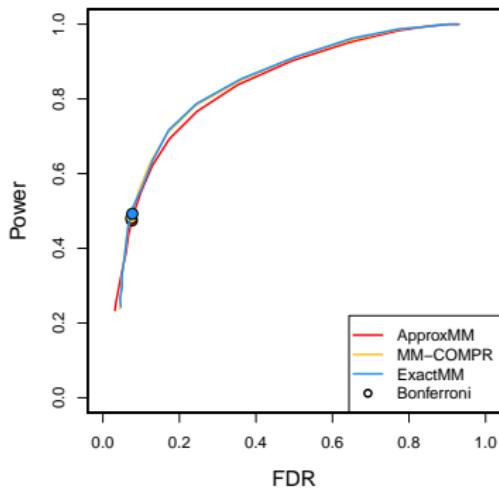
Le modèle linéaire mixte compressé (Zhang et al., 2010)

- Simplification ou **compression de la matrice K** par une approche de classification ascendante hiérarchique en faisant varier le nombre de groupes de n à 1 (n étant le nombre d'individus dans la population)
- Parmi l'ensemble des K matrices, **selection de celle qui permet le meilleur ajustement** dans le modèle nul



Performances du modèle linéaire mixte compressé

- Récupère la puissance du modèle mixte exact en utilisant une approche approximative (EMMAX/P3D).



- Données simulées à partir de données génotypiques réelles *A. thaliana* (*Horton et al., 2012*)
- 1,326 inds; 250k SNP
- 100 loci causaux, effets tirés d'une distribution exponentielle
- $H = 0.5$

- Implémenté dans le package R GAPIT (*Lipka et al., 2012*) :
<http://zzlab.net/GAPIT/>

1 Introduction

2 Le modèle linéaire mixte

3 Avancées méthodologiques

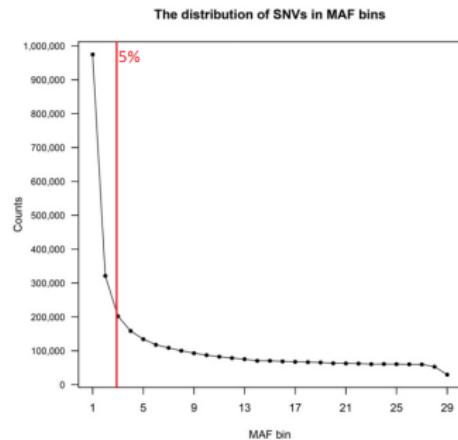
- Approche multi-locus
- Modifications de la matrice d'apparentement
 - FaST-LMM-Select
 - MLM-LOCO
 - Prise en compte du DL dans l'estimation de K
 - Le modèle linéaire mixte compressé

• Variants rares

- Approche multi-caractères

Variants rares

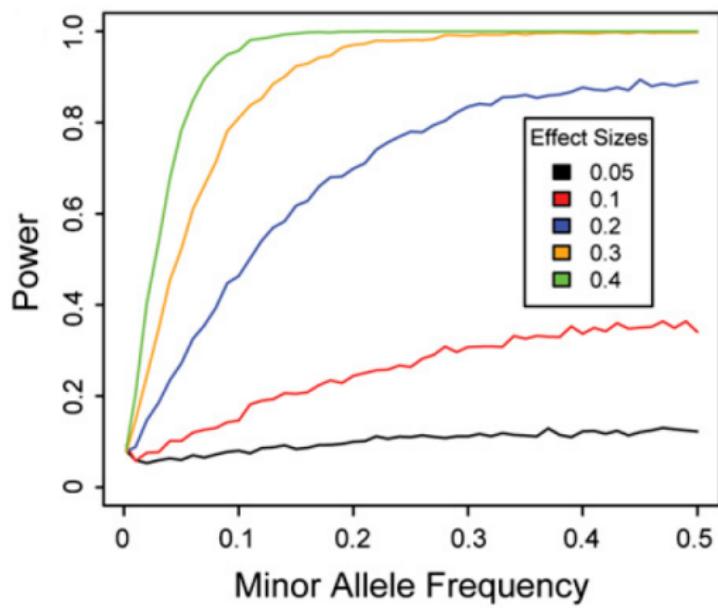
- Les variants rares ($MAF < 5\%$) contribuent potentiellement à l'architecture génétique des caractères complexes
- Parmi d'autres hypothèses, ils ont notamment été proposés pour expliquer le problème de l'héritabilité manquante (Manolio *et al.*, 2009)
- Les variants rares sont en fait très nombreux et des études ont montré qu'ils ont plus de chance d'être fonctionnellement importants que les variants à fréquence intermédiaire (Zhu *et al.*, 2011)
e.g. Distribution de la MAF chez 29 génomes humains



Le problème de la détection des variants rares:

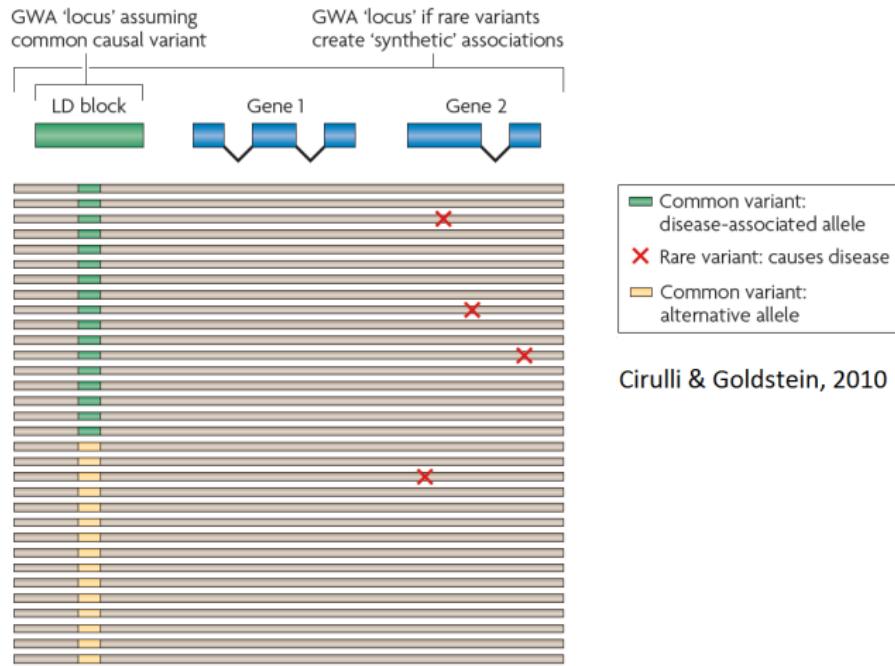
1. Faible puissance de détection

Myles *et al.*, 2009



Le problème de la détection des variants rares:

2. Ils génèrent des associations synthétiques



Approches pour la detection des variants rares

- **Combinaison des variants rares dans un groupe pour une analyse globale de leur effet conjoint**
- Les groupes peuvent être des **genes, voies de biosynthèse, regions génomiques** selon une fenêtre glissante le long du génome, etc...
- De nombreuses méthodes ont été proposées pour effectuer l'**analyse de groupes de SNP**:
 - Méthodes basées sur des **résumés/synthèses des statistiques** après regroupement (e.g. CAST)
 - Méthodes basées sur une analyse de **similarité entre les séquences** individuelles (e.g. TreeLD, Margarita)
 - Approches de **régression multiple**, dont les approches pénalisées
 - **Modèle linéaire mixte**

Detection de variants rares avec le modèle mixte : Sequence Kernel Association Test (SKAT)

- Le modèle linéaire mixte réconcilie les approches basées sur la similarité et celles basées sur la régression
- Il a donc été proposé dans un contexte de détection des variants rares par Wu *et al.* (2011)
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{s} + \boldsymbol{\epsilon}, \mathbf{s} \sim N(0, \sigma_s^2 \mathbf{K}_s) \text{ et } \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 \mathbf{I}),$$
 où, \mathbf{K}_s est une matrice d'apparentement estimée à partir du **groupe de SNP** que l'on souhaite **conjointement tester**
- La significativité est testée avec un **test du rapport de vraisemblance** (LRT): $D = -2 \ln(L_{H_0}/L_{H_1}) \sim \chi^2$ à 1 ddl
- Wu *et al.* ont proposé d'inclure dans le modèle linéaire mixte des axes d'ACP en **effets fixes** pour controller l'**effet confondant de la structure des populations**

Detection de variants rares avec le modèle mixte : FaST-LMM-Set (Listgarten *et al.*, 2013)

- Une autre approche évidente pour prendre en compte l'effet de la structure : le **modèle linéaire mixte à 2 effets aléatoires**

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{s} + \mathbf{c} + \epsilon,$$

$$\mathbf{s} \sim N(0, \sigma_s^2 \mathbf{K}_s), \mathbf{c} \sim N(0, \sigma_c^2 \mathbf{K}_c) \text{ & } \epsilon \sim N(0, \sigma_\epsilon^2 \mathbf{I}),$$

où \mathbf{K}_s est une matrice d'apparentement estimée à partir du **groupe de SNP** que l'on souhaite **conjointement tester**, et \mathbf{K}_c est une matrice d'apparentement estimée à partir des **autres SNP** pour prendre en compte la **structure**

- Listgarten *et al.* ont proposé de sélectionner les SNP pour estimer \mathbf{K}_c de la même manière que dans FaST-LMM-Select (pour prendre en compte les effets de dilution et de contamination proximale)
- Une approche similaire appelée **regional heritability mapping** a aussi été proposée par Nagamine *et al.* (2012) et Uemoto *et al.* (2013)

Performances de FaST-LMM-Set avec des données réelles

- Données sur la maladie de Crohn et l'hypertension (WTCCC, 2007); 14,925 inds (dont \approx 2,000 cases); 356,441 SNP
- SNP groupés par gène : 12,058 sets
- *Maladie de Crohn*: 41 groupes de SNP significatifs parmi lesquels 2 n'avaient pas été détectés par l'approche classique
Hypertension: 10 groupes de SNP significatifs parmi lesquels 3 n'avaient pas été détectés par l'approche classique
- Dans les 2 cas, les genes qui n'avaient pas été détectés par l'approche classique sont de bons candidats, dont RFWD2 pour la maladie de Crohn qui a par ailleurs été validé par une meta-analyse (Franke *et al.*, 2010)

Implementation et disponibilité

- FaST-LMM-Set est implémenté dans la version Python de FaST-LMM: <https://github.com/microsoftgenomics/fast-lmm>
- Regional heritability mapping est implementé dans le logiciel REACTA: <http://www.epcc.ed.ac.uk/projects-portfolio/reacta>
- Ce type d'approche est aussi faisable avec le logiciel ASReml (version stand alone ou version R), mais ce programme est sous licence par VSN International
- Bibliothèques R pour faire des modèles linéaires mixtes à deux effets aléatoires avec structure de covariance spécifiée par l'utilisateur: kinship, varComp, breedR, sommer...

1 Introduction

2 Le modèle linéaire mixte

3 Avancées méthodologiques

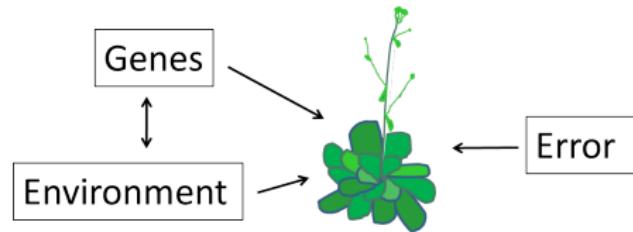
- Approche multi-locus
- Modifications de la matrice d'apparentement
 - FaST-LMM-Select
 - MLM-LOCO
 - Prise en compte du DL dans l'estimation de K
 - Le modèle linéaire mixte compressé
- Variants rares
- Approche multi-caractères

Approche multi-caractères

- Lorsque des **mesures multiples** sont prises sur des individus, les **phénotypes** qui en résultent sont souvent **corrélés** car :
 - ils sont contrôlés par les même locus : **pleïotropie**
 - ils sont contrôlés par des locus **liés**, DL inclus
 - ils ont été collectés dans les **mêmes conditions environnementales**
- La prise en compte des **corrélations** phénotypiques dans les modèles d'association devrait donc permettre de détecter plus facilement les **locus à effet pleïotropique**
- Cela devrait aussi permettre d'améliorer la **puissance de détection** par rapport aux analyses caractère par caractère

Approche multi-caractères

- Dans le cas particulier où les phénotypes consistent en **un caractère unique** mesuré dans **plusieurs environnements**, leur analyse conjointe devrait permettre:
 - de décomposer les effets **génétiques**, **environnementaux** ainsi que leur **interaction**
 - d'identifier des **allèles** qui **intéragissent** avec **l'environnement** et ainsi d'étudier les bases génétiques de la **plasticité phénotypique**



Comment analyser conjointement plusieurs phénotypes en GWAS

- Plusieurs approches ont été proposées pour faire des **GWAS sur plusieurs phenotypes**:
 - En **combinant des statistiques de test** issues des analyses univariées (O'Brien, 1984; Yang *et al.*, 2010), bibliothèque R 'CUMP'.
 - En analysant une **combinaison linéaire** des caractères issue d'une ACP (Klei *et al.*, 2008) ou d'une analyse canonique des corrélations (CCA) (Ferreira and Purcell, 2009), logiciels 'PCHAT' et 'Plink'.
 - O'Reilly *et al.* (2012) ont aussi proposé d'**inverser le modèle de regression** pour identifier la combinaison linéaire de phénotypes qui est le plus associée avec chaque SNP, bibliothèque R 'MultiPhen'.
- Mais, ces approches ne permettent pas de **prendre en compte l'effet confondant de la structure des populations**

Le modèle linéaire mixte pour l'analyse multivariée

- En génétique animale, les modèles mixtes multivariés sont utilisés de façon routinière pour l'analyse multi-caractères dans des pédigrées complexes et ainsi décomposer les correlations phénotypiques en composantes génétiques and environnementales (Lynch and Walsh, 1998)
- Les modèles mixtes multivariés ont récemment suscité de l'intérêt en génétique humaine pour notamment estimer des heritabilité de caractères et des correlations génétiques entre caractères (Yang *et al.*, 2011; Deary *et al.*, 2012; Lee *et al.*, 2012; Vattikuti *et al.*, 2012)

Le modèle linéaire mixte multivarié en GWAS

MTMM: Multi-Trait Mixed-Model (Korte *et al.*, 2012)

- Modèle mixte ***multivarié***

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{x}\beta + (\mathbf{x} \times \mathbf{s}_1)\alpha + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

$$var(\mathbf{u}) = \begin{bmatrix} \sigma_{g11}^2 \sigma_{g12}^2 \\ \sigma_{g12} \sigma_{g22}^2 \end{bmatrix} \otimes \mathbf{K}, var(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma_{\epsilon11}^2 \sigma_{\epsilon12}^2 \\ \sigma_{\epsilon12} \sigma_{\epsilon22}^2 \end{bmatrix} \otimes \mathbf{I}$$

- Paramètres génétiques

- Heritabilités : $h_i^2 = \sigma_{gii}^2 / (\sigma_{gii}^2 + \sigma_{\epsilon ii}^2)$
- Correlations génétiques : $\rho_{g12} = \sigma_{g12} / \sqrt{\sigma_{g11}^2 \sigma_{g22}^2}$

Approche multi-caractères : comment ?

MTMM: Multi-Trait Mixed-Model

- Les **variances** sont estimées avec le **modèle mixte multivarié** au moyen d'un outil dédié comme par exemple **ASReml**, ou les bibliothèques R **breedR** (Muñoz & Sanchez) ou **sommer** (Covarrubias-Pazaran, 2016).
- Les **associations** sont testées au moyen de **GLS F-tests** (philosophie EMMAX) entre les modèles imbriqués suivants :
 - *Modèle complet* : $\mathbf{y} = \mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{x}\beta + (\mathbf{x} \times \mathbf{s}_1)\alpha + \mathbf{Z}\mathbf{u} + \epsilon$
 - *Modèle réduit* : $\mathbf{y} = \mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{x}\beta + \mathbf{Z}\mathbf{u} + \epsilon \Leftrightarrow \alpha = 0$
 - *Modèle nul* : $\mathbf{y} = \mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{Z}\mathbf{u} + \epsilon \Leftrightarrow \alpha = 0 \text{ and } \beta = 0$

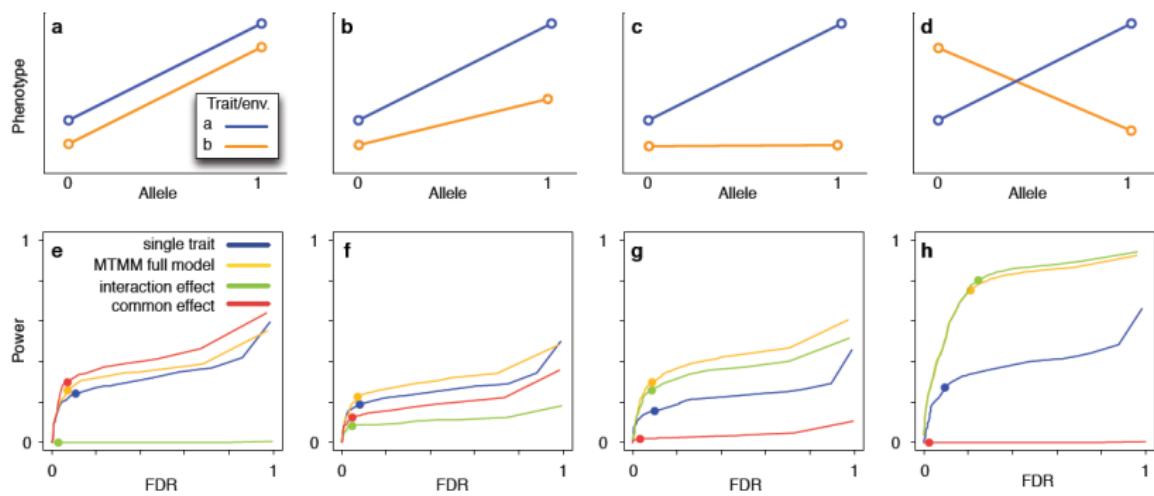
Évaluation des performances de MTMM avec des simulations

2 000 paires de caractères corrélés

Jeu de données génotypiques réel d'*A. thaliana* (Horton *et al.*, 2012)

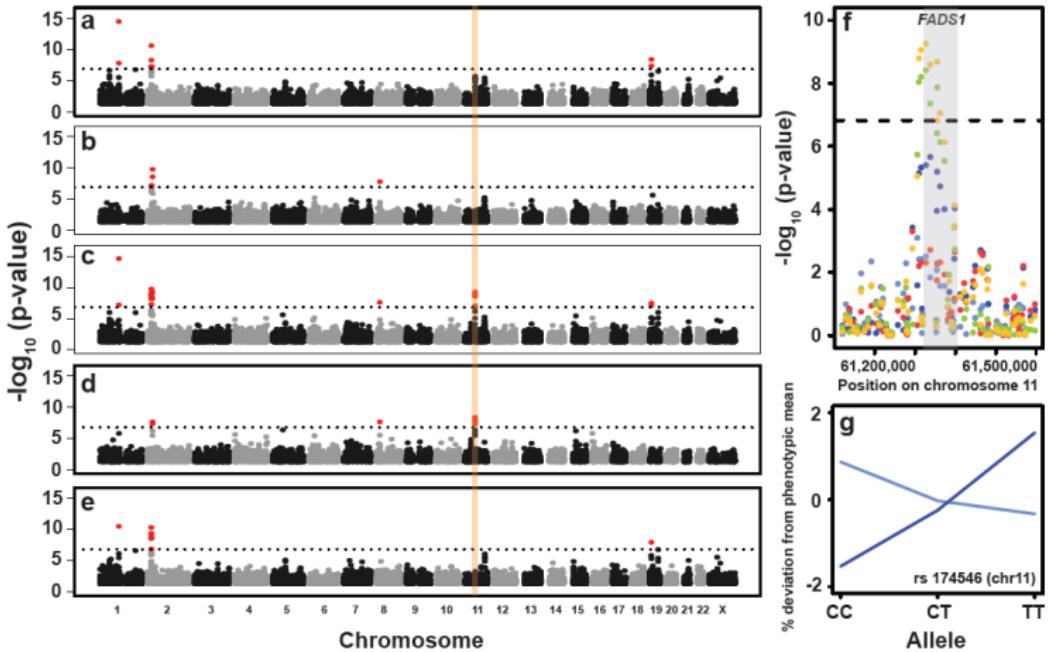
1 SNP à effet relativement fort (jusqu'à 2%)

10 000 SNP à faible effet (background)



Utilisation de MTMM pour détecter des locus à effet pleiotropique chez l'humain

Exemple de la paire LDL-TG du jeu de données NFBC



Utilisation de MTMM pour détecter des interactions GxE chez *A. thaliana*

- Date de floraison chez *A. thaliana* (Li et al., 2010)
- 459 accessions, 214 051 SNP
- Climats simulés en chambre de culture
- Plan factoriel: 2 saisons × 2 lieux

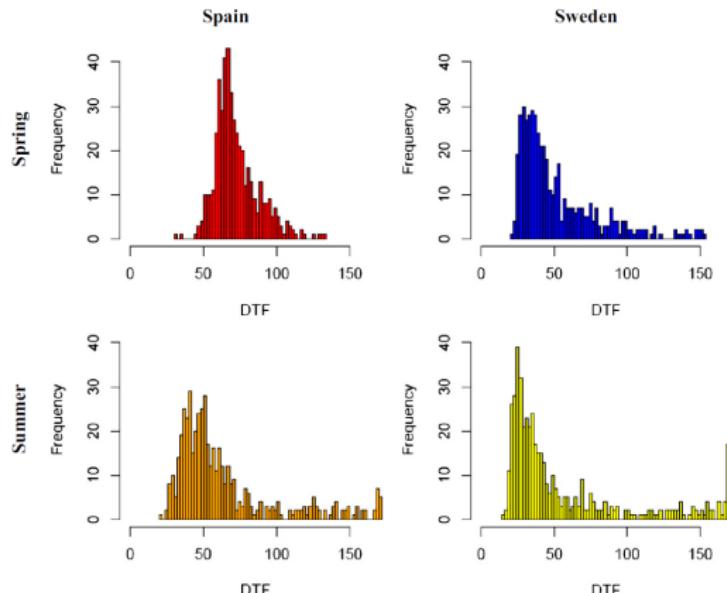
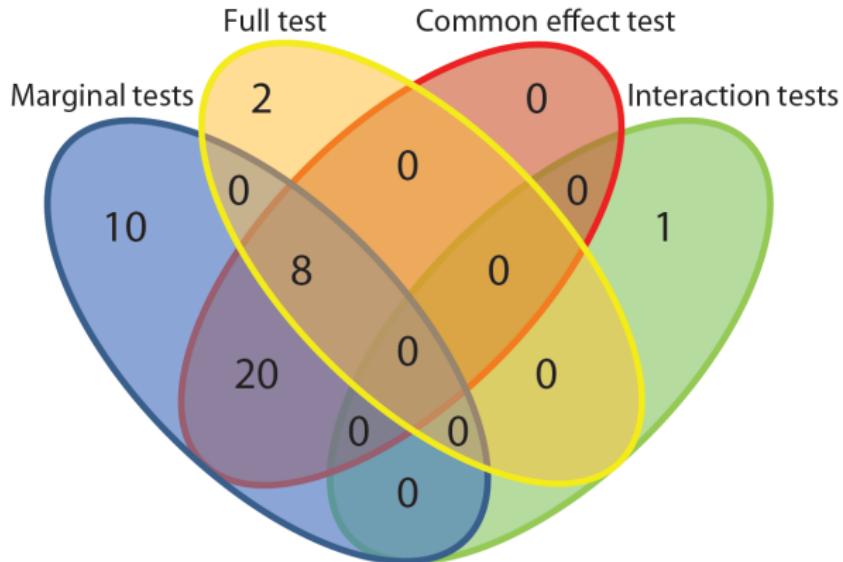


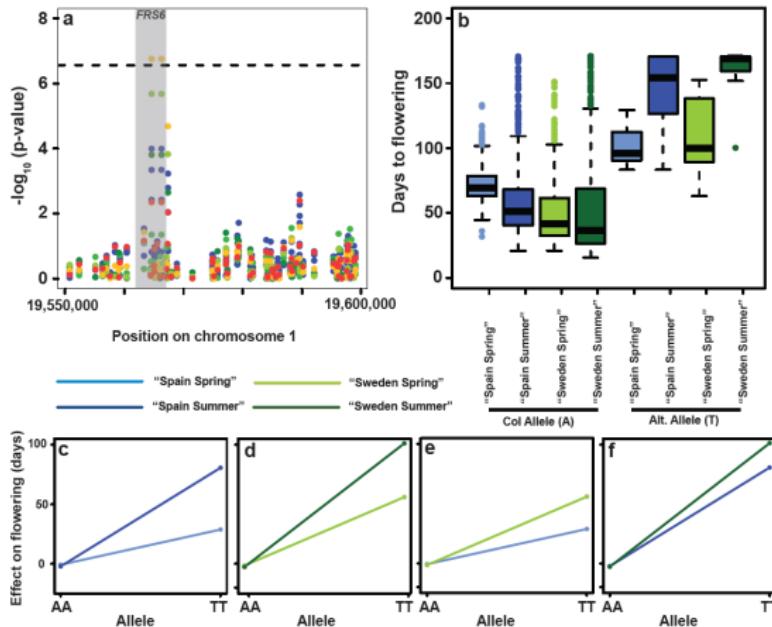
Fig. 1. Histogram of FT in four environments. Shown is the average of four replicates for each accession. DTF, days to flower.

Utilisation de MTMM pour détecter des interactions GxE chez *A. thaliana*

Associations détectées avec MTMM



Utilisation de MTMM pour détecter des interactions GxE chez *A. thaliana*



MTMM : Bilan

- L'approche modèle mixte multivarié est "**classique**" en **génétique animale** pour estimer les paramètres génétiques...
La philosophie EMMAX permet d'adapter cette approche à la **détection d'associations**
- Le **pouvoir de détection** de MTMM dépend du niveau de **corrélation génétique** entre les caractères analysés
- MTMM peut ainsi permettre de détecter de **nouvelles régions génomiques**, mais ne remplace pas les analyses univariées
- Chez les **plantes**, MTMM présente un intérêt pour la détection **d'interactions GxE** dans le cas de dispositifs multi-environnementaux
- **Implémentation** sous R qui dépend du package **ASReml-R** (Butler *et al*, 2009) : <https://cynin.gmi.oeaw.ac.at/home/resources/mtmm/>

Références bibliographiques

- Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-838 (2010).
- Astle, W. & Balding D.J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statist. Sci.* 24, 451-471 (2009).
- Aulchenko, Y.S. et al. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177, 577-585 (2007).
- Atwell, S. et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627-631 (2010).
- Ayers, K.L. & Cordell, H.J. SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression. *Genetic Epidemiology* 34, 879-891 (2010).
- Baxter, I. et al. A Coastal Cline in Sodium Accumulation in *Arabidopsis thaliana* Is Driven by Natural Variation of the Sodium Transporter AtHKT1;1. *Plos Genetics* 6, 8 (2010).
- Butler, D.G. et al. ASReml-R reference manual (2009).
- Chen, J.H. & Chen, Z.H. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95, 759-771 (2008).
- Cho, S. et al. Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis. *Annals of Human Genetics* 74, 416-428 (2010).

Références bibliographiques

- Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* 11, 415-25 (2010).
- Covarrubias-Pazaran G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLoS One.* 11, e0156744 (2016).
- Croiseau, P. & Cordell, H.J. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. *BMC Proceedings* 3(2009).
- Deary, I.J. et al. Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* 482, 212-5 (2012).
- Dickson, S.P. et al. Rare Variants Create Synthetic Genome-Wide Associations. *Plos Biology* 8, 12 (2010).
- Ferreira, M.A. & Purcell, S.M. A multivariate test of association. *Bioinformatics* 25, 132-3 (2009).
- Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399-433 (1918).
- Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* 42, 1118-1125 (2010).
- Hoggart, C.J. et al. Design and analysis of admixture mapping studies. *American Journal of Human Genetics* 74, 965-978 (2004).

Références bibliographiques

- Horton, M.W. et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics* 44, 212-216 (2012).
- Kang, H.M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42, 348-354 (2010).
- Klei, L. et al. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol* 32, 9-19 (2008).
- Korte, A. et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* 44, 1066-1071 (2012).
- Lee, S.H. et al. Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* (2012).
- Li, Y. et al. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* 107, 21199-204 (2010).
- Lipka, A.E. et al. GAPIT: Genome Association and Prediction Integrated Tool. *Bioinformatics* 28, 2397-2399 (2012).
- Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nature Methods* 8, 833-837 (2011).
- Listgarten, J. et al. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* 29, 1526-1533 (2013)
- Listgarten, J. et al. Improved linear mixed models for genome-wide association studies. *Nature Methods* 9, 525-526 (2012).

Références bibliographiques

- Lynch, M. & Walsh, B. Genetics and Analysis of Quantitative Traits, (Sinauer, 1998).
- Ma, L. et al. Multi-locus Test Conditional on Confirmed Effects Leads to Increased Power in Genome-wide Association Studies. Plos One 5, e15006 (2010).
- Malo, N. et al. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. American Journal of Human Genetics 82, 375-85 (2008).
- Manolio, T.A. et al. Finding the missing heritability of complex diseases. Nature 461, 747-753 (2009).
- Myles, S. et al. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. The Plant Cell 21, 2194-2202 (2009).
- Muñoz, F. & Sanchez L. breedR: Statistical Methods for Forest Genetic Resources Analysts. R package version 0.12-2.
- Nagamine, Y. et al. Localising Loci underlying Complex Trait Variation Using Regional Genomic Relationship Mapping. PLoSOne, 7(10): e46501 (2012).
- O'Brien, P.C. Procedures for comparing samples with multiple endpoints. Biometrics 40, 1079-87 (1984).
- O'Reilly, P.F. et al. MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. Plos One 7, e34861 (2012).
- Pikkukookana, P. & Sillanpaa, M.J. Correcting for relatedness in Bayesian models for genomic data association analysis. Heredity 103, 223-237 (2009).

Références bibliographiques

- Platt, A. et al. Conditions Under Which Genome-Wide Association Studies Will be Positively Misleading. *Genetics* 186, 1045-1052 (2010).
- Rincent, R. et al. Recovering Power in Association Mapping Panels with Variable Levels of Linkage Disequilibrium. *Genetics* 197, 375–387 (2014).
- Sabatti, C. et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* 41, 35-46 (2009).
- Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* 6, 461-464 (1978).
- Segura, V. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44, 825-830 (2012).
- Setakis, E. et al. Logistic regression protects against population structure in genetic association studies. *Genome Research* 16, 290-296 (2006).
- Speed, D. Improved Heritability Estimation from Genome-wide SNPs. *Am J Hum Genet* 91, 1011–1021 (2012).
- Uemoto, Y. et al. The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. *Frontiers in genetics* 4, 232 (2013)
- Vattikuti, S. et al. Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. *PLoS Genet* 8, e1002637 (2012).

Références bibliographiques

- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-78 (2007).
- Wu, M.C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics* 89, 82-93 (2011).
- Yang, J. et al. Mixed model association methods: advantages and pitfalls. *Nat Genet* 46, 100-106 (2014).
- Yang, J. et al. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88, 76-82 (2011).
- Yang, Q. et al. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet Epidemiol* 34, 444-54 (2010).
- Yu, J.M. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38, 203-208 (2006).
- Zhang, Z.W. et al. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42, 355-360 (2010).
- Zhu, Q. et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *American Journal of Human Genetics* 88, 458-68 (2011).
- Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Genetics* 11, 407-409 (2014).