Génomique Intégrative et Modélisation des Maladies Métaboliques
UMR 8199 (CNRS / Université de Lille 2 / Institut Pasteur de Lille)

Développement et Application de Méthodologies Statistiques pour Etudes Longitudinales d'Association Génétique

Comité de suivi de thèse: deuxième année

Mickaël Canouil
mickael.canouil@cnrs.fr

Direction de thèse

Dr. Ghislain Rocheleau & Pr. Philippe Froguel

26 Septembre 2016

Sommaire

Introduction

En 2014, la prévalence de diabète de type 2 (DT2) a été estimée à près de 9% chez l'adulte de 18 ans et plus.

Sur la dernière décennie, l'essor des études d'association pangénomiques (GWAS) a permis l'identification de :

- 65 variants associés à la susceptibilité au DT2 ;
- 36 variants associés à la glycémie à jeun (FG) chez les normoglycémiques.

Introduction

La grande majorité des GWAS a utilisé un design transversal, quand un design longitudinal offre la possibilité :

- de décrire la trajectoire temporelle d'une variable ;
- d'accroître la puissance pour détecter des variants génétiques associés à la trajectoire.

La modélisation de ces trajectoires temporelles optimiserait les tests d'association et l'exploitation des phenotypes disponibles.

Objectifs

Cette thèse s'organise sur deux principaux objectifs :

1. Développement et implémentation des approches basées notamment sur les modèles joints ;
2. Application à un jeu de données (p.ex. cohorte D.E.S.I.R., FRAMINGHAM, etc) ;
3. Optimisation du temps de calcul avec R (p.ex. lme4, portage Julia, etc).

Matériels

Le laboratoire (UMR CNRS 8199) dispose de l'accès à la cohorte prospective D.E.S.I.R. (Données Epidémiologiques sur le Syndrome d'Insulino-Résistance), comptant 5 212 individus suivis pendant 9 ans, tous les 3 ans (0, 3, 6 et 9 ans).

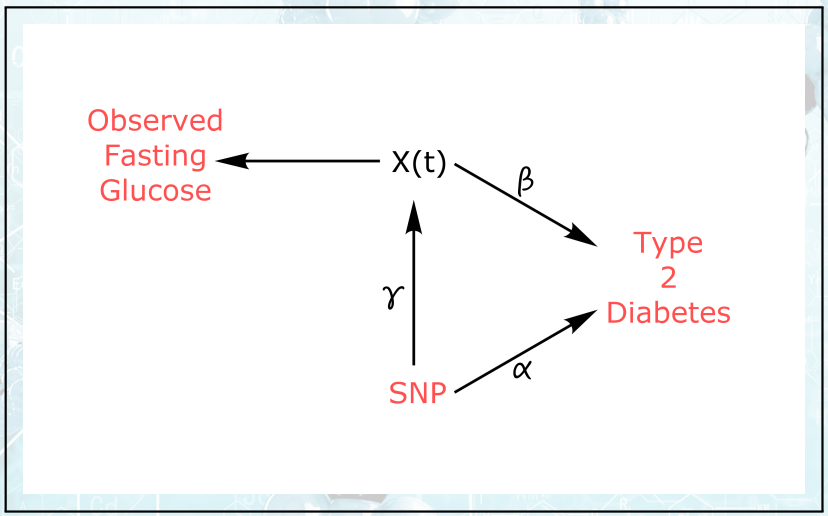
En plus de données phénotypiques (p.ex. FG, hba1c, etc), des données génotypiques sont également disponibles pour une grande partie de ces individus (4 364) : 124 095 SNPs (fréquence allèlique $> 1\%$ et Hardy-Weinberg $p < 1 \times 10^{-3}$).

Cette cohorte comporte 179 cas incidents de DT2, définis à partir d'une glycémie supérieure à 7 mmol/L ou par la prise d'un traitement anti-diabétique.

Méthodes : Modèle Joint

L'approche par modèle joint a éré décrite par Tsiatis and Davidian [2004] et Ibrahim et al. [2010],

avec une implémentation dans l'extension JM [Rizopoulos, 2010] du logiciel R (version 3.2.3)[R Core Team, 2015].



$X(t)$: trajectoire de FG inférée des données longitudinales observées ;

α : effet du SNP sur le DT2 ;

γ : effet du SNP sur la trajectoire de FG ;

β : effet de la trajectoire de FG sur le DT2.

Méthodes : Modèle Joint

Le modèle joint se décompose en deux parties :

- Composante longitudinale (Modèle linéaire mixte)

$$Y_{ij} = X_{ij} + \epsilon_{ij} \tag{1}$$

$$Y_{ij} = \theta_{0i} + \theta_{1i} \times t_{ij} + \gamma \times Z_i + (\delta \times W_i) + \epsilon_{ij} \tag{2}$$

$$\boldsymbol{\theta} \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}); \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

- Composante de survie (Modèle de Cox)

$$h_i(t) = h_0(t) exp(\beta X_i(t) + \alpha Z_i) \tag{3}$$

Méthodes : Simulation

Simulation des données selon les Equations 1 à 3, avec la fonction de risque de base fixée : $h_0(t) = \lambda$.

Les temps d'événements ont été générés selon une distribution exponentielle (Cox à risque proportionnel) [Austin, 2012].

$$H(T) = \int_0^T \lambda exp(\beta \times X(t) + \alpha \times Z) dt \tag{4}$$

$$T = \frac{1}{\beta\theta_1} log\left(-\frac{\beta\theta_1 \times log(1-u)}{\lambda exp(\beta\theta_0 + (\beta\gamma + \alpha)Z)} + 1\right) \tag{5}$$

Méthodes : Simulation

Paramètres initiaux pour la simulation des données basés sur le SNP de TCF7L2 (SNP le plus fortement associé au DT2).

Paramètres	Valeurs
Effectif (N)	5000
Temps de mesures (en années)	$0, 3, 6, 9$
Incidence à neuf ans (I)	5%
LMM : Trajectoire $\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$	$\mathcal{N}_2\left(\begin{bmatrix} 4.50 \\ 0.013 \end{bmatrix}, \begin{bmatrix} 0.16 & 0 \\ 0 & 1 \times 10^{-3} \end{bmatrix}\right)$
LMM : Effet du SNP (γ)	0.025
Cox : Effet du SNP (α)	0.2
JM : Effet de la trajectoire (β)	3.50

Simulation

α : effet du SNP sur le diabète

Paramètre	Estimée	Simulée	Fréquence allélique
α	0.215 [-0.0762, 0.551]	0.23	0.05
	0.219 [0.00346, 0.459]		0.10
	0.22 [0.0741, 0.373]		0.25
	0.219 [0.101, 0.343]		0.50
	0.219 [0.0825, 0.349]		0.75
	0.218 [0.0284, 0.398]		0.90
	0.218 [-0.0538, 0.461]		0.95

Simulation

β : effet du SNP sur la trajectoire de la glycémie à jeun

Paramètre	Estimée	Simulée	Fréquence allélique
	3.56 [3.29, 3.85]		0.05
	3.57 [3.3, 3.85]		0.10
	3.57 [3.3, 3.86]		0.25
β	3.56 [3.29, 3.85]	3.60	0.50
	3.57 [3.29, 3.85]		0.75
	3.57 [3.29, 3.85]		0.90
	3.57 [3.29, 3.85]		0.95

Simulation

γ : effet de la trajectoire de la glycémie à jeun sur le diabète

Paramètre	Estimée	Simulée	Fréquence allélique
	0.0196 [-0.0164, 0.0558]		0.05
	0.0195 [-0.00712, 0.0456]		0.10
	0.0194 [0.00111, 0.038]		0.25
γ	0.0197 [0.00322, 0.0353]	0.02	0.50
	0.0196 [0.00115, 0.0385]		0.75
	0.0196 [-0.00678, 0.0457]		0.90
	0.0195 [-0.0165, 0.0558]		0.95

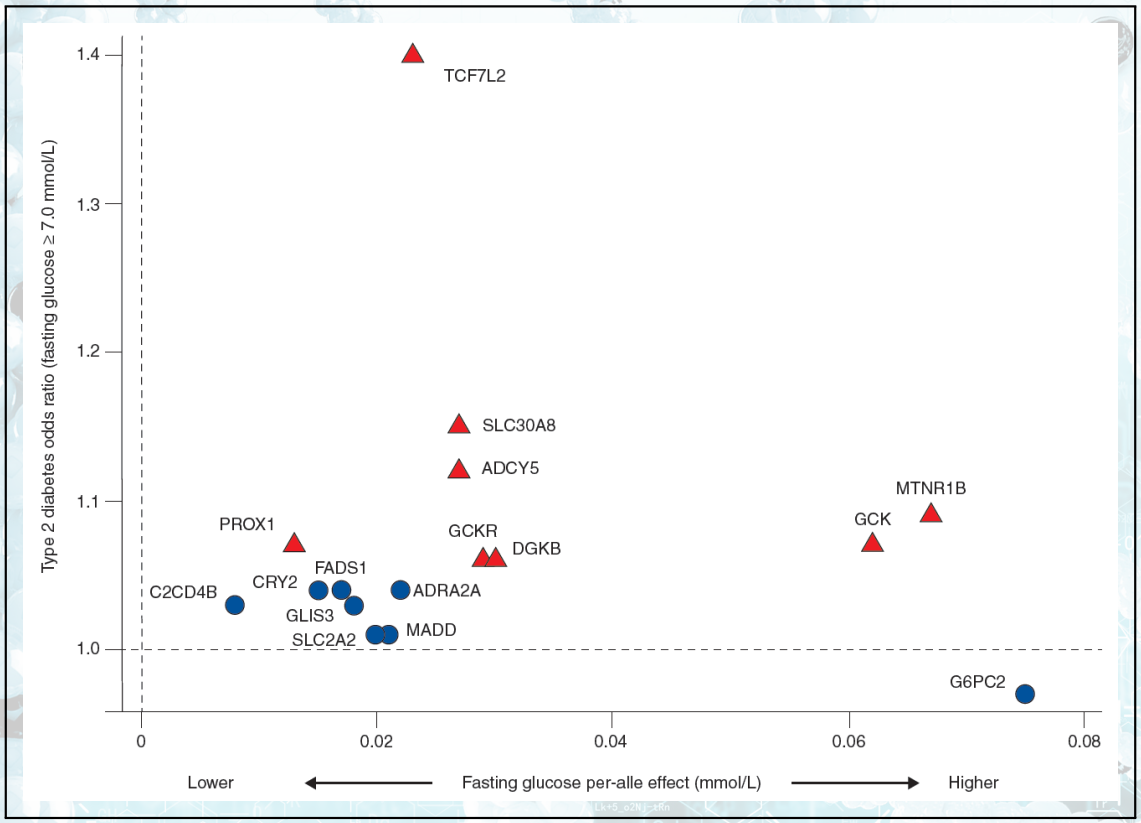
Simulation

À partir des résultats de simulations, nous pouvons recommandés les conditions suivantes (pour une incidence de 5%) :

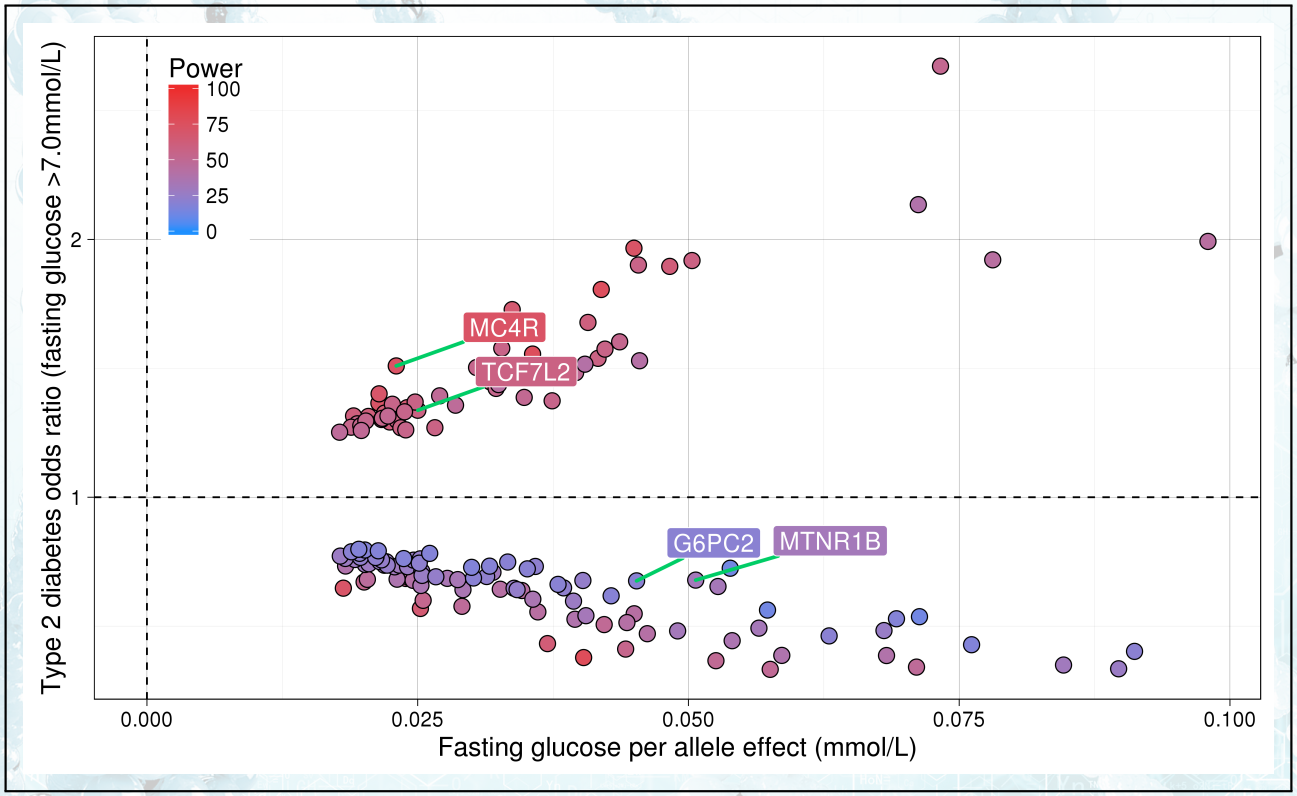
- Fréquence allélique $> 5\%$
- Nombre de mesures > 4
- Taille de population 1 000
- Données manquantes (MCAR/MAR) $< 25\%$

Historique

Identification de 80 et 100 loci associés au DT2 et au FG, notamment via GWAS et méta-analyses [Dupuis et al., 2010; Yaghootkar and Frayling, 2013; Vaxillaire et al., 2014].



Application D.E.S.I.R.



- 124 095 SNPs ont été analysés dans la cohorte D.E.S.I.R. avec la puce MetaboChip [Voight et al., 2012].
- 145 loci trouvés comme associés à la fois à FG et à la survenue du DT2 (au seuil de 5%)

Application D.E.S.I.R.

La puissance statistique a été calculée :

1. au niveau du modèle joint, à l'aide de la formule de Chen et al. [2011] :

$$z_{\tilde{\beta}} = \pm\sqrt{Df(1-f)(\beta\gamma+\alpha)^2} + z_{1-\tilde{\alpha}/2}$$

avec

- D le nombre de DT2 incidents,
- f la fréquence de l'allèle à risque,

2. au niveau des paramètres γ et α respectivement pour l'effet du SNP sur la trajectoire de FG et sur le risque de DT2.

Application D.E.S.I.R.

SNP	α	γ	β	Power
rs1942873_G (MC4R)	0.412	0.023	3.15	72.70
rs55899248_C (TCF7L2)	0.291	0.025	3.49	57.60
rs10830962_G (MTNR1B)	-0.388	0.0507	3.24	31.50
rs12475693_C (G6PC2)	-0.392	0.0452	3.17	20.30

(En bleu : p-value < 0.05, en rouge : p-value $< 5 \times 10^{-8}$)

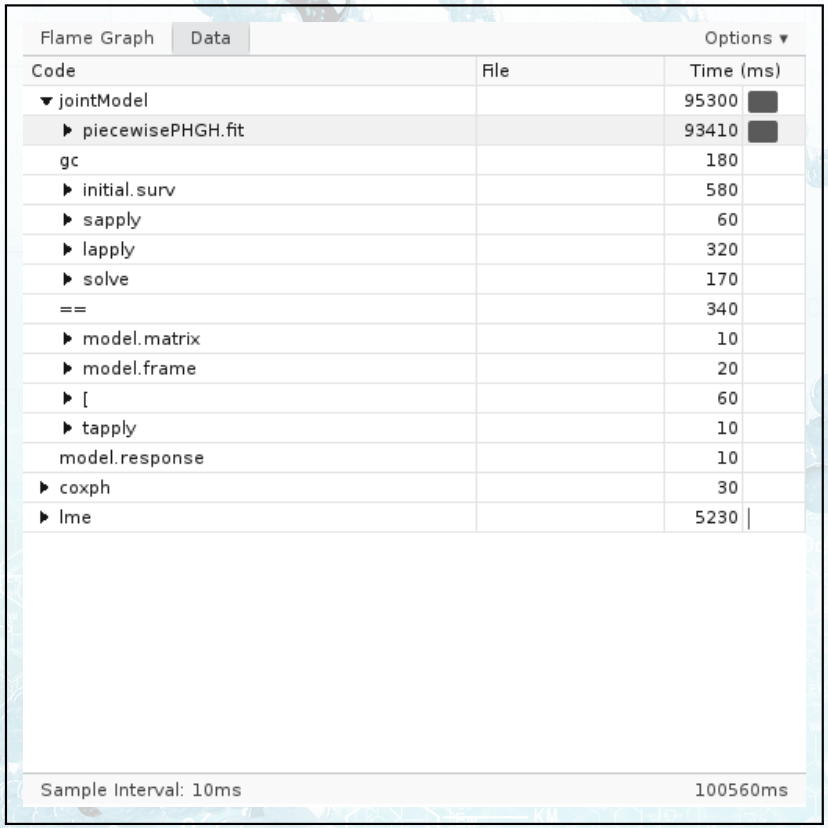
Application D.E.S.I.R.

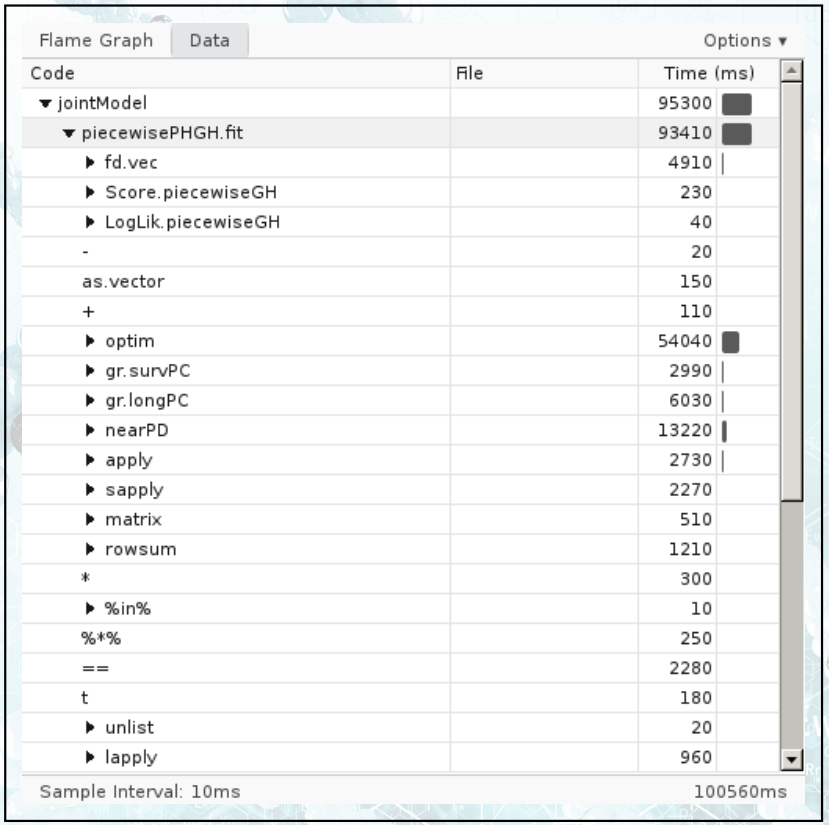
En bleu, lorsque le modèle joint (JM) présente une puissance statistique supérieure à une approche transversale (CM) : régression linéaire/logistique (JM / CM).

SNP	α	γ
rs1942873_G (MC4R)	47.7 / 71.6	64.9 / 52.8
rs55899248_C (TCF7L2)	66.6 / 35.3	91.6 / 81.0
rs10830962_G (MTNR1B)	55.1 / 32.1	60.8 / 47.1
rs12475693_C (G6PC2)	76.2 / 56.5	56.7 / 44.5

(Erreur de type 1 : 5.81% $\pm$ 0.83 / 5.47% $\pm$ 0.64 pour JM et CM)

Performance





Performance

Malgré les optimisations apportées au niveau des paramètres de convergence de l'étape EM du modèle joint (extension JM [Rizopoulos, 2010]), qui représente 50% du temps total, il est possible d'apporter d'autres optimisations de façon simple : comme l'utilisation des fonctions de bas niveau de R :

```
Unit: nanoseconds
                expr  min     lq    mean median   uq   max neval cld
    {     seq(10) } 4197 4756.0 6286.90 4979.5 5447 39636   100   b
 {     seq_len(10) }  146  156.5  531.88  177.0  206 23635   100  a
```

Perspectives

1. Ecriture d'un rapport scientifique (Bourse SFD-Lilly : 21 Mai 2017) et d'un article sur l'application du modèle joint à la cohorte D.E.S.I.R.
2. Validation des SNPs/gènes mis en évidence par le modèle joint (p.ex. cohorte de réplication)
3. Etude d'autre trait tel que l'HbA1c (hémoglobine glyquée)
4. Inclusion des individus incident pour l'IFG (Impaired Fasting Glucose ; $FG > 6.1 mM/L$) en plus des individus DT2 ($FG > 7 mM/L$)
5. Optimisation du code (algorithme de JM) pour une exécution sur des puces GWAS et/ou imputées

Congrès

- *SMPGD 2016* (Statistical Methods for Post Genomic Data) :

  Présentation orale

  "Longitudinal Genetic Modelling : Revisiting Associations of SNPs Associated with Blood Fasting Glucose in Normoglycemic Individuals"

- *IGES 2016* (International Genetic Epidemiology Society) :

  Présentation poster

  "Single Nucleotide Polymorphisms Associated With Fasting Blood Glucose Trajectory And Type 2 Diabetes Incidence : A Joint Modelling Approach"

Congrès

- *SFD 2017* (Société Francophone du Diabète)
- *SFdS 2017* (Société Française de Statistique)
- *Rencontres R 2017*
- *useR 2017*