Integrated Genomics and Metabolic Diseases Modeling
UMR 8199 (CNRS / Université de Lille 2 / Institut Pasteur de Lille)

Longitudinal Genetic Modelling

Revisiting Associations of SNPs Associated with Blood Fasting Glucose in Normoglycemic Individuals

Mickaël Canouil[1], Ghislain Rocheleau[1], Loïc Yengo[1], Philippe Froguel[1,2]

[1] Univ. Lille, CNRS, Institut Pasteur de Lille, UMR 8199 - EGID, F-59000 Lille, France

[2] Department of Genomics of Common Disease, Imperial College London, London, United Kingdom

mickael.canouil@cnrs.fr
(ghislain.rocheleau@cnrs.fr)

Statistical **M**ethods for **P**ost **G**enomic **D**ata

February 11-12, 2016

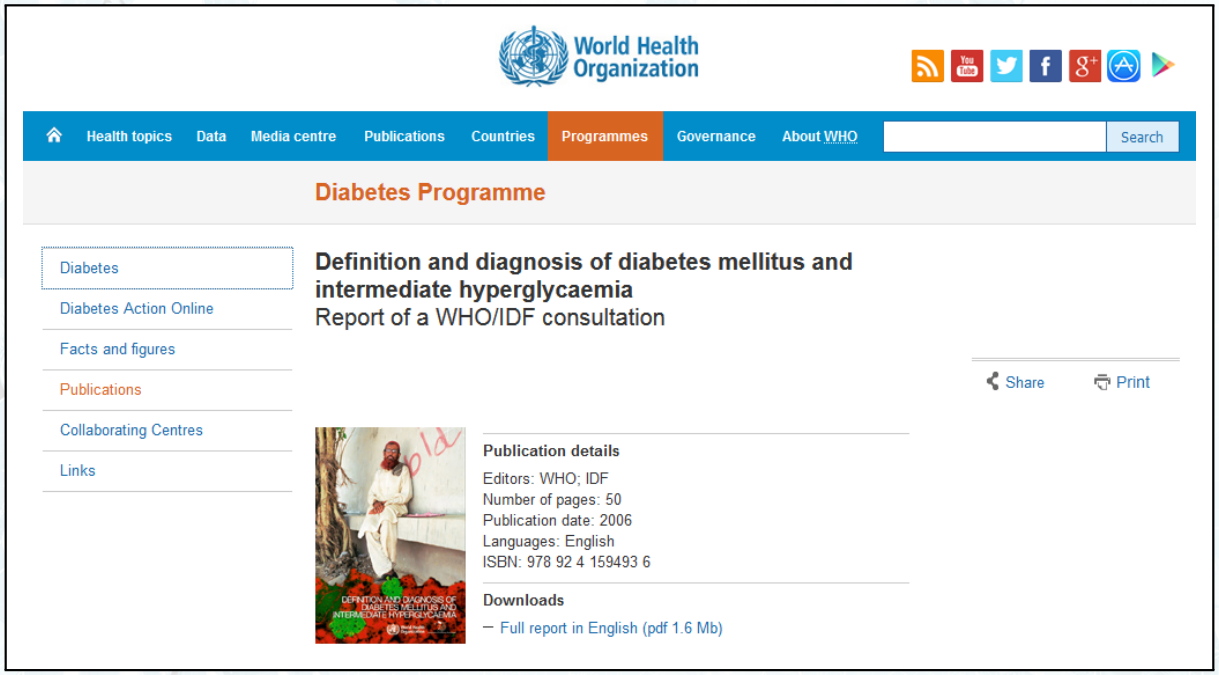
Background

- Genome-wide association studies (GWAS) have been extensively used to identify genetic markers.
- Typically, GWAS are based on cross-sectional data.
- The recent use of approaches to account for longitudinal data is driven among other things, by the need to increase statistical power.
- One possible solution is to increase sample size by meta-analysing multiple cohorts [Scott et al., 2012].

Data

- Metabochip DNA arrays (Illumina) [Voight et al., 2012] assayed in individuals recruited in the French cohort D.E.S.I.R. (**D**onnées **É**pidémiologiques sur le **S**yndrome d'**I**nsulino-**R**ésistance).
- D.E.S.I.R. is a prospective cohort (5,212 individuals) followed up for 9 years (measured at baseline, 3, 6 and 9 years) for many biological traits and several pathologies (e.g. type 2 diabetes, cardiovascular diseases, etc.).

Data



The focus of this study is on normoglycemic individuals, defined for fasting plasma glucose (FPG) < 7.0 mmol/l [World Health Organization, 2006].

Data

Table : Clinical characteristics of the 5,212 individuals in D.E.S.I.R.

	Men			Women		
	N	mean	sd	N	mean	sd
SEX	2,576	-	-	2,636	-	-
AGE	2,576	46.64	10	2,636	46.9	10.04
BMI	2,531	25.46	3.359	2,584	24.04	4.116
FPG (year = 0)	2,572	5.532	0.8921	2,633	5.188	0.743
FPG (year = 3)	2,234	5.594	0.972	2,264	5.23	0.736
FPG (year = 6)	2,004	5.657	1.033	2,059	5.279	0.7603
FPG (year = 9)	1,953	5.697	1.036	2,024	5.304	0.8011

Data

In this table, we focus on fasting plasma glucose measured at baseline, 3, 6 and 9 years (4 measures).

Table : Missing data rates for the 5,212 individuals in D.E.S.I.R.

	Men	Women
SEX	0.00%	0.00%
AGE	0.00%	0.00%
BMI	1.75%	1.97%
FPG (year = 0)	0.155%	0.114%
FPG (year = 3)	13.3%	14.1%
FPG (year = 6)	22.2%	21.9%
FPG (year = 9)	24.2%	23.2%

Methods

Y_i is the measured fasting plasma glucose and G_i denotes the genotypes coded 0, 1 or 2.

Linear Model (Baseline)

$$Y_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Methods

Y_i is the measured fasting plasma glucose and G_i denotes the genotypes coded 0, 1 or 2.

Linear Model (Baseline)

$$Y_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Linear Model (Average of m measures)

$$\bar{Y}_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \frac{\sigma^2}{m})$$

Methods

Y_i is the measured fasting plasma glucose and G_i denotes the genotypes coded 0, 1 or 2.

Linear Model (Baseline)

$$Y_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Linear Model (Average of m measures)

$$\bar{Y}_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \frac{\sigma^2}{m})$$

Two-Step

1. $Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \sim \mathcal{N}_m(0, V^\dagger{}_i \equiv Z^\dagger{}_i D^\dagger Z^\dagger{}'_i + \sigma^{\dagger 2} I_m)$
2. $\hat{b}_{0i} = \beta_0^* + \beta_g^* G_i + \epsilon_i^*$, where $\epsilon_i^* \sim \mathcal{N}(0, \sigma^{*2})$

Methods

Y_i is the measured fasting plasma glucose and G_i denotes the genotypes coded 0, 1 or 2.

Linear Model (Baseline)

$$Y_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Linear Model (Average of m measures)

$$\bar{Y}_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \frac{\sigma^2}{m})$$

Two-Step

1. $Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \sim \mathcal{N}_m(0, V^{\dagger}{}_i \equiv Z^{\dagger}{}_i D^{\dagger} Z^{\dagger}{}_i' + \sigma^{\dagger 2} I_m)$
2. $\hat{b}_{0i} = \beta_0^* + \beta_g^* G_i + \epsilon_i^*$, where $\epsilon_i^* \sim \mathcal{N}(0, \sigma^{*2})$

Linear Mixed Model (LMM)

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \beta_g G_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim \mathcal{N}_m(0, V_i \equiv Z_i D Z_i' + \sigma^2 I_m)$$

Methods

Y_i is the measured fasting plasma glucose and G_i denotes the genotypes coded 0, 1 or 2.

Linear Model (Baseline)

$$Y_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Linear Model (Average of m measures)

$$\bar{Y}_i = \beta_0 + \beta_g G_i + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \frac{\sigma^2}{m})$$

Two-Step

1. $Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} \sim \mathcal{N}_m(0, V^\dagger{}_i \equiv Z^\dagger{}_i D^\dagger Z^\dagger{}_i' + \sigma^{\dagger 2} I_m)$
2. $\hat{b}_{0i} = \beta_0^* + \beta_g^* G_i + \epsilon_i^*$, where $\epsilon_i^* \sim \mathcal{N}(0, \sigma^{*2})$

Linear Mixed Model (LMM)

$$Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \beta_g G_i + \epsilon_{ij}, \text{ where } \epsilon_{ij} \sim \mathcal{N}_m(0, V_i \equiv Z_i D Z_i' + \sigma^2 I_m)$$

Generalised Estimating Equations (GEE)

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 t_{ij} + \beta_g G_i \text{ and } \mathbb{V}(Y_i) = V_i \text{ (Compound Symmetry).}$$

Top associated SNPs

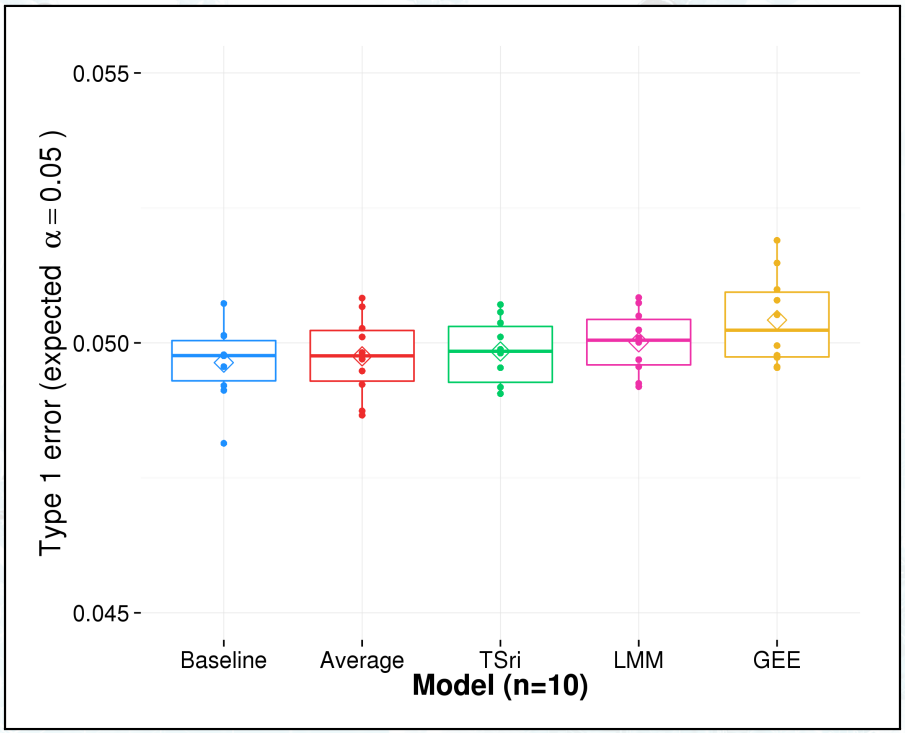
SNP	Chromosome	Position (hg18)	Gene	Minor Allele	MAF	P value
rs560887	2	169,471,394	G6PC2	T	0.30	1.5e-15
rs2908289	7	44,190,467	GCK	A	0.19	4.4e-05
rs16913693	9	110,720,180	IKBKAP	G	0.02	4.1e-04
rs2191349	7	15,030,834	DGKB/TMTM195	G	0.44	1.1e-03
rs6072275	20	39,177,319	TOP1	A	0.13	4.7e-03
rs340874	1	212,225,879	PROX1	T	0.45	7.8e-03
rs3783347	14	99,909,014	WARS	T	0.20	1.3e-02
rs3829109	9	138,376,587	DNLZ	A	0.26	1.4e-02
rs11607883	11	45,796,285	CRY2	G	0.46	3.0e-02
rs2302593	19	50,888,474	GIPR	G	0.48	3.7e-02

SNPs selected based on Yaghootkar and Frayling [2013] and Vaxillaire et al. [2014].

Results: Type 1 error

Type 1 error was computed for 10 SNPs previously shown to be significantly associated with fasting plasma glucose in Vaxillaire et al. [2014].

Testing H0: $\beta_g = 0$ versus H1: $\beta_g = d \neq 0$



100,000 genotypes permutations were performed for each selected SNP.

Results: Statistical power (post-hoc)

SNP	Chromosome	Position (hg18)	Gene	Minor Allele	MAF	P value
rs560887	2	169,471,394	G6PC2	T	0.30	1.5e-15
rs2908289	7	44,190,467	GCK	A	0.19	4.4e-05
rs16913693	9	110,720,180	IKBKAP	G	0.02	4.1e-04
rs2191349	7	15,030,834	DGKB/TMTM195	G	0.44	1.1e-03
rs6072275	20	39,177,319	TOP1	A	0.13	4.7e-03
rs340874	1	212,225,879	PROX1	T	0.45	7.8e-03
rs3783347	14	99,909,014	WARS	T	0.20	1.3e-02
rs3829109	9	138,376,587	DNLZ	A	0.26	1.4e-02
rs11607883	11	45,796,285	CRY2	G	0.46	3.0e-02
rs2302593	19	50,888,474	GIPR	G	0.48	3.7e-02

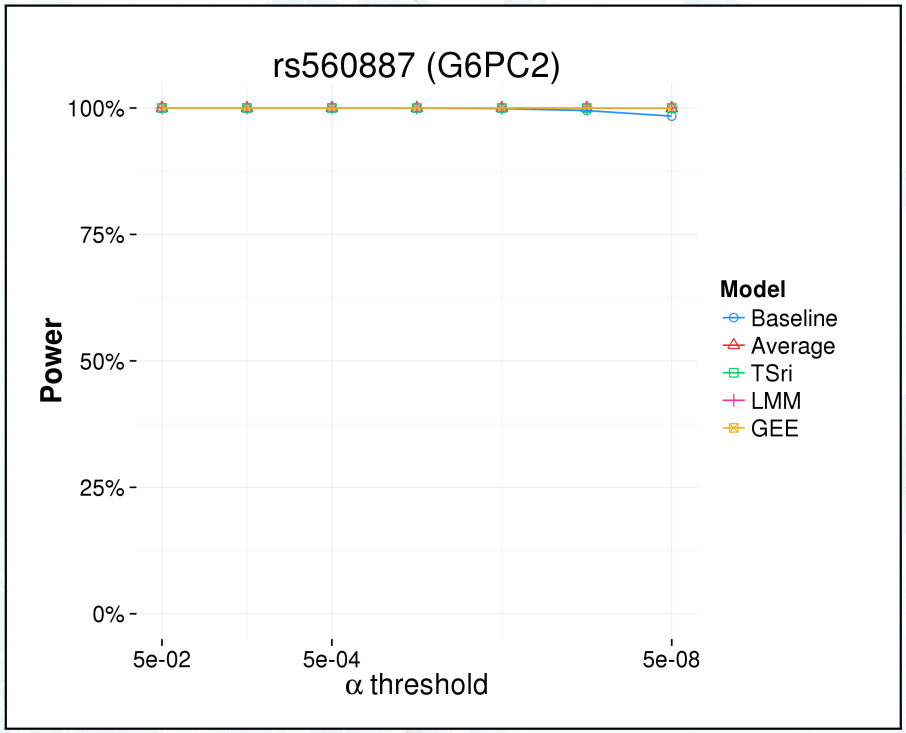
SNPs selected based on Yaghootkar and Frayling [2013] and Vaxillaire et al. [2014].

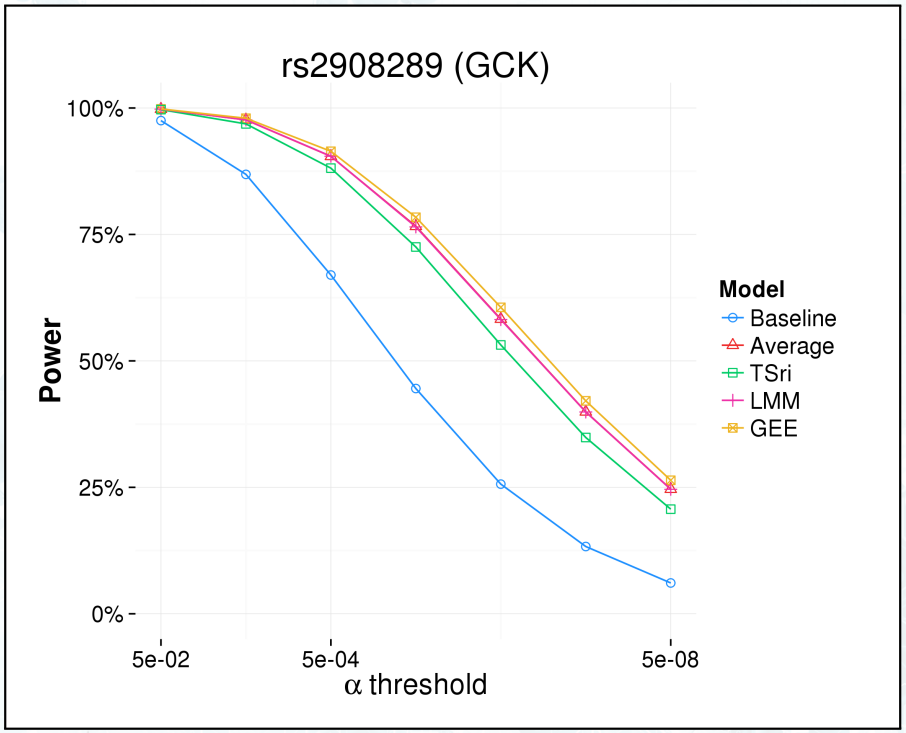
Results: Statistical power (post-hoc)

SNP	Chromosome	Position (hg18)	Gene	Minor Allele	MAF	P value
rs560887	**2**	**169,471,394**	**G6PC2**	**T**	**0.30**	**1.5e-15**
rs2908289	**7**	**44,190,467**	**GCK**	**A**	**0.19**	**4.4e-05**
rs16913693	**9**	**110,720,180**	**IKBKAP**	**G**	**0.02**	**4.1e-04**
rs2191349	7	15,030,834	DGKB/TMTM195	G	0.44	1.1e-03
rs6072275	**20**	**39,177,319**	**TOP1**	**A**	**0.13**	**4.7e-03**
rs340874	1	212,225,879	PROX1	T	0.45	7.8e-03
rs3783347	14	99,909,014	WARS	T	0.20	1.3e-02
rs3829109	9	138,376,587	DNLZ	A	0.26	1.4e-02
rs11607883	11	45,796,285	CRY2	G	0.46	3.0e-02
rs2302593	19	50,888,474	GIPR	G	0.48	3.7e-02

SNPs selected based on Yaghootkar and Frayling [2013] and Vaxillaire et al. [2014].

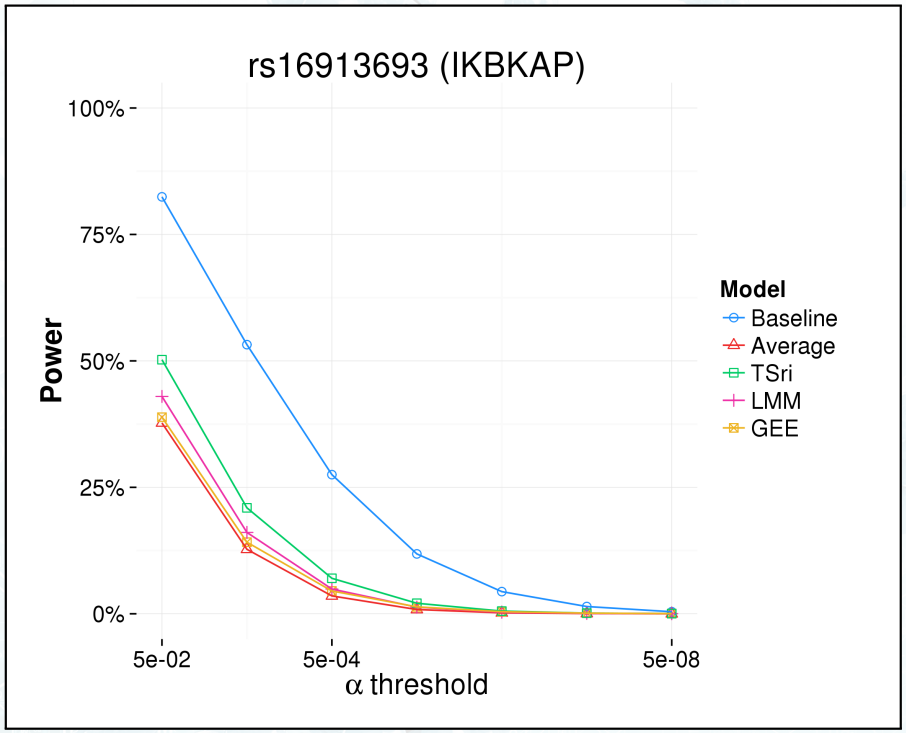
Results: Statistical power (post-hoc)

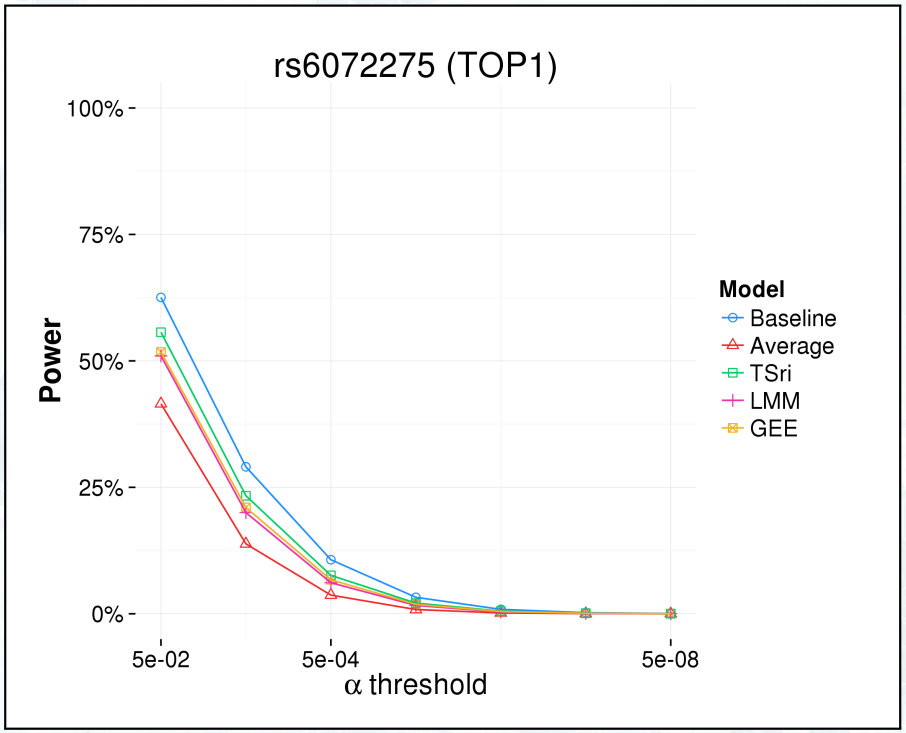




100,000 resamplings were performed for each selected SNP.

Results: Statistical power (post-hoc)





100,000 resamplings were performed for each selected SNP.

Non-Centrality Parameter & statistical power

Closed-form formulas for testing SNP association under the Cross-Sectional (CS), the Random Intercept (RI) and the Random Intercept and Slope model (RIS)

$$NCP_{CS} = nd^2 \left(\frac{2p(1-p)}{\sigma^2} \right) \quad (1)$$

$$NCP_{RI} = NCP_{CS} \left(\frac{m\sigma^2}{\sigma^2 + m\sigma^2_{b0}} \right) \quad (2)$$

$$NCP_{RIS} = NCP_{RI} U \quad (3)$$

$$\text{with} \quad U = \frac{(\sigma^2 + \sigma^2_{b1} \sum_{j=1}^{m} (t_j - \bar{t})^2)(\sigma^2 + m\sigma^2_{b0})}{(\sigma^2 + \sigma^2_{b1} \sum_{j=1}^{m} (t_j - \bar{t})^2)(\sigma^2 + m\sigma^2_{b0}) - m\rho^2 \sigma^2_{b0} \sigma^2_{b1} \sum_{j=1}^{m} (t_j - \bar{t})^2} \geq 1 \quad (4)$$

Non-Centrality Parameter & statistical power

This implies that $NCP_{RIS} \geq NCP_{RI}$ but no guarantee that $NCP_{RIS} > NCP_{CS}$

SNP	*Gene*	NCP_{CS}	NCP_{RIS} (NCP_{RI})
rs560887	G6PC2	63.33	93.08 (92.69)
rs2908289	GCK	17.02	26.37 (26.26)
rs16913693	IKBKAP	12.75	6.55 (6.53)
rs6072275	TOP1	7.78	6.78 (6.76)

Non-Centrality Parameter & statistical power

This implies that $NCP_{RIS} \geq NCP_{RI}$ but **no guarantee** that $NCP_{RIS} > NCP_{CS}$

SNP	*Gene*	NCP_{CS}		NCP_{RIS} (NCP_{RI})
rs560887	G6PC2	63.33	<	93.08 (92.69)
rs2908289	GCK	17.02	<	26.37 (26.26)
rs16913693	IKBKAP	12.75	>	6.55 (6.53)
rs6072275	TOP1	7.78	>	6.78 (6.76)

$\mathcal{F}$uture research

- Characterise parameter space for which NCP_{RIS} is indeed higher than NCP_{CS};
- Derive formula for the non-centrality parameter when testing SNP*time interaction effect;
- Check results consistency according to missing data distribution (MCAR, MAR and MNAR).

Acknowledgements

Thank you for your attention!

CNRS UMR 8199

Ghislain Rocheleau

Loïc Yengo

Philippe Froguel

Fundings:

This work was supported by grants from "European Genomic Institute for Diabetes" (E.G.I.D., ANR-10-LABX-46), "European Commission", "Société Francophone du Diabète" and "Lilly".