Single Nucleotide Polymorphisms Associated with Fasting Blood Glucose Trajectory and Type 2 Diabetes Incidence: A Joint Modelling Approach

Mickaël Canouil[1,2,3] Philippe Froguel[1,2,3,4] and Ghislain Rocheleau[1,2,3]

[1]University of Lille, UMR 8199 - EGID, F-59000 Lille, France.

[2]CNRS, UMR 8199, F-59000 Lille, France.

[3]Institut Pasteur de Lille, F-59000 Lille, France.

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

Corresponding authors

Mickaël Canouil, EGID - UMR 8199, Pôle Recherche - 1er étage Aile Ouest, 1 place de Verdun, 59045 Lille CEDEX, France. E-mail: mickael.canouil@cnrs.fr

Ghislain Rocheleau, Maelstrom Research - The Research Institute of the McGill University Health Centre (RI MUHC), 2155 Guy, 4th Floor, Office 458, Montreal, Quebec, H3H 2R9, Canada. E-mail: grocheleau@maelstrom-research.org

Abstract

In observational cohorts, longitudinal data are collected with repeated measurements at predetermined time points for many biomarkers, along with other covariates measured at baseline. In these cohorts, time to a certain event of interest occurring is commonly reported and very often, a relationship will be observed between a biomarker repeatedly measured over time and that event. Joint models were designed to efficiently estimate statistical parameters by combining a mixed model for the longitudinal biomarker trajectory and a survival model for the event risk, using a set of random effects to account for the link between the two types of data.

First, we checked model consistency based on different simulation scenarios, varying sample size, minor allele frequency, number of repeated measurements and missing data patterns.

Second, using genotypes assayed with the Metabochip DNA arrays (Illumina) from close to 4,500 subjects recruited in the French cohort D.E.S.I.R. (Données Épidémiologiques sur le Syndrome d'Insulino-Résistance), we assessed the feasibility of implementing the joint modelling approach in a real high-throughput genomic dataset.

In our study, the event of interest was onset of type 2 diabetes (T2D), and the longitudinal biomarker repeatedly measured over time was fasting plasma glucose level.

To the best of our knowledge, joint models have never been applied into a genetic epidemiology context and could help identify novel loci sharing effects on both glycaemic traits and T2D.

Key words

genetic association; joint modelling; longitudinal studies

Introduction

With the increased availability of longitudinal and survival data within prospective cohorts, joint models have emerged to account for both types of data, particularly when dealing with the informative/non-informative dropouts which occur in such cohorts. Joint models have been studied and overviewed in the literature (Chen, Ibrahim, & Chu, 2011; Elashoff, li, & Li, 2016; Tsiatis & Davidian, 2004; Wulfsohn & Tsiatis, 1997) and software implementation has been proposed within different software and platforms (Diggle & Kenward, 1994; Elashoff, Li, & Li, 2008; Proust-Lima, Joly, Dartigues, & Jacqmin-Gadda, 2009; Rizopoulos, 2010; Rizopoulos & Ghosh, 2011; Sun, Sun, & Liu, 2007). The main idea behind the joint modelling is: 1) to model efficiently the survival process with a time-varying covariate, accounting for missing data and measurements errors, and 2) to account for informative dropouts in the longitudinal data. To model the two components of a joint model, a linear mixed effects (LME) model and a Cox proportional hazards model (CoxPH), are classically used to, respectively, fit the longitudinal component, and the survival component. Unlike the CoxPH model, in which the time-varying covariate is assumed to be exogenous, i.e. not modified by the occurrence of a previous event (Kalbfleisch & Prentice, 2002), the joint modelling framework allows to account for an endogenous time-varying covariate. An example of an endogenous covariate is given by the relationship between fasting glucose is irremediably modified due to medication.

Two approaches can be used for the estimation and inference of the model parameters: a "naive" Two-Step (TS) method or a joint likelihood method (JM). The first method consists in estimating the random effects of the trajectory, as provided by a LME model, and including them as a time-varying covariate into a CoxPH model, then using partial likelihood of the CoxPH model for the parameter estimation (Therneau & Grambsch, 2000). The second method is based on a joint likelihood of the two components (longitudinal and survival) at the same time. Comparison of these two approaches showed that the latter offers more consistent and efficient estimators than the former (Albert & Shih, 2010a, 2010b). But JM could be challenging to compute, especially achieving convergence at the Expectation-Maximisation (EM) step. Moreover, depending on the number of time points and/or the sample size, the overall computation time can substantially increase.

In this paper, we conducted a comprehensive simulation study to compare two joint model approaches, JM and TS, for joint modelling of the longitudinal and survival components. Our main goal is to show that joint modelling approach, when compared to separate modelling, might improve statistical power to detect an effect on either, or both, longitudinal and survival processes, while resulting in a bias reduction in parameter estimation. We also compared JM with TS approach and show that in the context where highly demanding computation and convergence issues might arise in JM computation, the TS offers a good alternative to JM in a reasonable time span, especially when applied at the genome-scale level. We also investigated and decomposed the computational time required by the R package "JM" (Rizopoulos, 2010, 2016), on the one

hand, and by the TS approach combining the R packages: "survival" (Therneau, 2017) and "nlme" (Pinheiro, Bates, & R-core, 2017).

Finally, we applied these approaches to a real dataset, the prospective D.E.S.I.R. cohort (*Données Épidémiologiques sur le Syndrome d'Insulino-Résistance*), which includes 5,212 individuals with extensive phenotypic measures recorded at 4 different occasions spanning a 9-year follow up (data collected every 3 years). These individuals were genotyped using the Illumina Metabochip DNA array which interrogates nearly 200,000 SNPs) (Voight et al., 2012). Relying on cross-sectional genome-wide association study design, the D.E.S.I.R. cohort was instrumental in identifying novel loci associated to prevalent type 2 diabetes (T2D) and to blood fasting glucose (FG) level in normoglycemic subjects (Bouatia-Naji et al., 2008; Rung et al., 2009; Sladek et al., 2007). We specifically focused on prediabetes conditions, such as IFG (Impaired Fasting Glucose), which is part of the diagnostic definition of T2D (FG > 7.0 mmol/L), and on time-to-onset of T2D, in order to possibly identify loci, novel or published, which simultaneously associate with the risk of developing T2D and with increasing blood FG. Our results were then compared to the genetic variants as reported in the literature (Vaxillaire et al., 2014; Welter et al., 2014), and to the meta-analyses results published by large consortia, such as, DIAGRAM (Morris et al., 2012) or the MAGIC (Dupuis et al., 2010) consortium.

Methods

Models Formulation

Joint Likelihood Model (JM)

The standard formulation of the joint model involves two components: a longitudinal component and a time-to-event component. Let n denote the sample size, and Y_{ij} the longitudinal measurements collected for each subject at time points $t_{ij}, i = 1, \cdots, n, j = 1, \cdots, m_i$, where m_i is the number of measurements of subject i. The longitudinal component (measurements) typically consists of a (generalised) linear mixed effect (LME) model, whose within-subject correlation matrix is modelled using random-effect parameter vector $b_i = \begin{pmatrix} \theta_{0i} \\ \theta_{1i} \end{pmatrix}$.

Under the joint likelihood framework as implemented in "JM" (Rizopoulos, 2010, 2016), within the class of "shared parameter models" (Elashoff et al., 2016; Rizopoulos, 2012), we define

$$Y_{ij} = X_{ij} + \epsilon_{ij} \qquad (1)$$

where Y_{ij} is the observed value and X_{ij} is the true (unobserved) value of the longitudinal measurement at time t_{ij}. The quantity ϵ_{ij} is a random error term usually assumed to be normally distributed:

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \qquad (2)$$

The quantity X_{ij} is typically called the trajectory function and is usually specified as a linear (or quadratic) function of time t_{ij}. We also define Z_i, a variable denoting the genotype of subject i, and W_i, a set of adjusting covariates:

$$Y_{ij} = X_{ij} + \epsilon_{ij} = \theta_{0i} + \theta_{1i}t_{ij} + \gamma Z_i + \delta W_i + \epsilon_{ij} \qquad (3)$$

For representation purposes, the term δW_i will be omitted henceforth. Random effects θ_{0i} (intercept) and θ_{1i} (slope) are assumed bivariate Normal: $\theta \sim \mathcal{N}_2(\mu, \Sigma)$, and supposed independently distributed from ϵ_{ij}. The coefficient γ assesses the genotypic (additive) effect of variable Z_i in the trajectory function. To account for possible varying slopes, an interaction term between Z_i and time t_{ij} could be added into the trajectory function. The interaction term was not considered in our study.

The time-to-event (survival) component usually consists of a parametric (e.g. exponential or Weibull distribution) or semi-parametric (e.g. Cox proportional hazards) model. T_i denotes the event time for subject i, and C_i the right censoring time (e.g. end of the follow-up). Let Δ_i be the event indicator: $\Delta_i = 0$, if $T_i > C_i$, and $\Delta_i = 1$, if $T_i <= C_i$. Under the Cox proportional hazards model, variable T_i is specified using the following equation:

$$\lambda_i(t) = \lambda_0(t)\exp(\beta X_i(t) + \alpha Z_i) \qquad (4)$$

where $\lambda_i(t)$ is the hazard function at time t_i and $\lambda_0(t)$ is the unspecified baseline hazard function, which we assume piecewise constant with two knots placed at intermediate time points in the follow-up. The coefficient α measures the effect of Z_i on the hazard function, while β measures the association between the trajectory function and the hazard function. In this formulation, we suppose that the subject-specific parameters $b_i = \begin{pmatrix}\theta_{0i}\\ \theta_{1i}\end{pmatrix}$ included in the trajectory $X_i(t)$ could modify the hazard function, which implies that β is the parameter linking the longitudinal and survival components.

Two-Step Model (TS)

As an alternative to JM, and based on the work of Tsiatis, DeGruttola, & Wulfsohn (1995), the two-step model estimates parameters of the joint model by first, estimating parameters of the trajectory function $X_i(t)$ in Equation (3), and second, by substituting this estimated trajectory, say $X_i^*(t)$, into (4) before fitting the Cox survival model.

Simulation Study

Simulation studies were carried out to further examine the sensitivity of the JM estimations under several scenarios. Parameters were set based on values estimated from the strongest SNPS associated with T2D, that is rs17747324 in gene *TCF7L2* (Table 1) (C; $OR = 1.43$; $p = 8.5\times10^{-55}$ (Morris et al., 2012); FG C; $\beta = 0.025$; $p = 6.47\times10^{-08}$ (Dupuis et al., 2010)).

Longitudinal data were simulated according to Equation (3), while event times were generated from an exponential distribution for the CoxPH model (Austin, 2012)

$$\lambda_0(t) = \lambda \qquad (5)$$

$$H_i(T_i) = \int_0^{T_i} \lambda \exp(\beta X_i(t) + \alpha Z_i) dt \qquad (6)$$

$$T_i = \frac{1}{\beta\theta_{1i}} \log\left(1 - \frac{\beta\theta_{1i} \times \log(1-u)}{\lambda \exp(\beta\theta_{0i} + (\beta\gamma + \alpha)Z_i)}\right) \qquad (7)$$

where λ was set to achieve the targeted incidence rate in the simulated dataset.

Datasets were simulated by varying the number of longitudinal measurements $m \in \{2; 3; 4; 5\}$, the number of subjects $n \in \{500; 1{,}000; 2{,}500; 5{,}000; 10{,}000\}$, the allele frequency $f \in \{0.05; 0.1; 0.25; 0.5\}$ and the incidence rate $d \in \{0.025; 0.05; 0.1\}$, thereby leading to 240 different scenarios. Each scenario was simulated 500 times.

The Root-Mean Square Error (RMSE)

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)} \qquad (8)$$

was used to assess precision for estimation of β, γ and α, when testing association between Y_{ij}, and T_i, Z_i effect on Y_{ij} and Z_i effect on T_i, respectively. We compared JM and TS approaches with the linear mixed effect model and the Cox regression model with time-varying covariate. Power and Type I error were computed for each model. The computational burden of each approach was also investigated as our goal is to implement all of these at a genome-wide scale.

Computational times

Based on our simulations, we provide approximate computational times for four sample sizes with parameters as listed in Table 1, when using an UNIX system with Intel® Xeon® CPU E7- 4870 @ 2.40GHz (80 such CPUs available computing in parallel). Table 3 shows computational time for one model, and when extrapolating the total computational time for 100,000 SNPs, which is the approximate number of SNPs on the Metabochip, after we applied data cleaning and the quality-control over common SNPs (minor allele frequency > 0.05).

Real Data

SNP genotyping was performed with Metabochip DNA arrays (Voight et al., 2012) using Illumina HiScan technology and GenomeStudio software (Illumina, San Diego, USA) in 5,212 subjects from the French cohort D.E.S.I.R. (Balkau, 1996). These subjects have been followed up for 9 years, and extensive phenotypic data has been recorded at 4 different times during that follow-up. Quality control was performed using PLINK

1.90 beta version (Chang et al., 2015; Purcell & Chang, 2015). SNPs with call rate greater or equal to 95%, with no significant deviation from Hardy-Weinberg equilibrium at $p > 1\times10^{-5}$, and with minor allele frequency (MAF) over 5% were kept for analysis, resulting in 101,305 SNPs. Due to missing phenotypes which did not allow to confirm T2D status, 232 subjects were removed. An additional 554 subjects were excluded due to individual call rate lower than 95%, leaving 4,426 subjects for analysis after these quality control steps (Figure 1).

Principal component analysis was performed in a combined dataset comprised of the 4,426 subjects, and of the subjects from the publicly available 1,000 Genomes database (The 1000 Genomes Project Consortium, 2015). SNPS retained for analysis were restricted to those common in both samples. The first two components were sufficient to discriminate ethnic origin. Non-Caucasian subjects (62) were excluded from the analysis. A further 12 prevalent T2D cases at baseline were also removed.

The final dataset included 4,352 subjects, of which 167 were diagnosed as T2D incident cases.

Using the joint modelling approach implemented in the package JM (Rizopoulos, 2010, 2016) within R software version 3.3.3 (R Core Team, 2015), all 101,305 SNPs were tested for joint association with blood fasting glucose and T2D. Based on the joint modelling formulation (see in Equations (3) and (4)), let Y_{ij} denote the observed values of blood fasting glucose (FG), and let Z_i represent the genotype of individual i at each SNP, along with W_i covariates such as age, sex and BMI. Finally, let T_i is the time at which a subject is diagnosed with T2D.

As illustrated in Figure 2, the association between each SNP and the FG longitudinal values is captured through the parameter (γ); association between each SNP and the time at onset of T2D is captured through the parameter α; the association between the longitudinal values of FG and the onset of T2D is assessed using the parameter β. By convention, a subject is diagnosed diabetic if his/her measured value of FG is above 7.0 mmol/L and/or is under lowering glycaemia treatment. In the joint modelling framework, the trajectory of FG is viewed as a dropout process, since all FG values become missing after T2D diagnosis, as a result of diabetic subjects being placed under treatment to lower and regulate the glucose level in their blood. In this case, FG is considered as an endogenous covariate, because the dropout process is not independent from the measured glucose values prior to T2D diagnosis.

Results

Comparison of estimation accuracy

We explored the influence of several factors on the estimator accuracy when using a linear mixed effect model (LME) to estimate γ, and when using a Cox regression model with time-varying covariate (CoxPH) to estimate β and α, and compared it with the accuracy as obtained under the joint modelling approach (JM) and its Two-Step approximation (TS). For each simulated scenario, we measured the root-mean-square error (RMSE) for all three parameters of interest (γ, β and α).

Due to the complexity of the estimating algorithm within JM, convergence could not be obtained ($4.53 \pm 5.81\%$ of convergence issues in average per scenario) for the whole set of 500 simulations (i.e. algorithm "piecewise-PH-aGH" for a time-dependent relative risk model with a piecewise constant baseline risk function, using the adaptive Gauss-Hermite quadrature rule to approximate integrals within the Expectation-Maximisation (EM) step).

RMSE for parameter γ (Figure 3) showed performance quite similar between JM and TS, which was expected given the formulation of the joint model within the "Shared Parameter Models" framework, in which Y_i (mean of Y_{ij} modelled within LME according to Equation (3)) links the longitudinal data to the time of event.

RMSE for parameter β (Figure 4) and for parameter α (Figure 5) was smaller within the joint modelling framework (either JM or TS) than within the more classical CoxPH model. While RMSE for β was uniformly the same in the CoxPH model across all scenarios, under JM or TS it decreased whenever any of the sample size, incidence rate or allele frequency increases.

Differences in RMSE for parameter α were less important than for parameter β, where TS performed as equally well as CoxPH with time-dependent covariate, probably because partial likelihood inferences were used in both approaches. JM estimations were less biased in almost all scenarios when the sample size was greater than 2,500.

Overall, our simulations revealed that JM is less biased than when separate approaches are used to model the effect of Z_i on the longitudinal Y_i, and on the time-to-event T_i. While separate approaches performed well for parameters γ and α, the bias for β was the greatest observed across all scenarios.

In addition, statistical power and type I error were also studied (Table 2), for the default simulation settings (Table 1), and showed similar results between JM and TS. Nevertheless, the latter simulations highlighted convergence issues that might occur within the joint likelihood approach (19.4% of the power simulation study).

Computational time

Computational times are reported in Table 3. We observed that the time required to complete JM or TS algorithms increases linearly with respect to sample size in our simulations. However, these figures are very optimistic since our simulations did not include any covariate or more complex random parameter.

To investigate further computational time issues, we profiled the execution of the main function "*jointmodel*" from the R package "JM", which implements the joint likelihood modelling approach as described in this paper. In the "JM" package, the linear mixed effect sub-model is handled by the function "*lme*" from the "nlme" package. One may argue that using a faster approach, e.g. as implemented in the R package "lme4", the computational time might be decreased. As shown in Figure 6, the main issue is within the "*jointmodel*" function which took over 95 % of the global computation time. After examination of the call tree diagram, we can see that the more time-consuming task within the "*jointmodel*" function is the optimisation of the EM algorithm (described in Rizopoulos (2012), Appendix B), despite the use of a calculation tricks (i.e. adaptive Gauss-Hermite quadrature for numerical integration).

Application in real data

Applying R package JM to our D.E.S.I.R. cleaned dataset lead to 265 SNPs (Figure 7) which were globally associated (with p-value<0.05) with FG and T2D event through their respective parameters γ and α.

Amongst these 265 SNPs (163 unique genes), we identified 17 genes (Table 4) which were already reported to be associated with FG and/or T2D risk.

In Figure 8, we specifically focused on parameters γ and α. After Bonferroni correction (nominal p-value $\simeq 5\times10^{-7}$), no genetic variants showed a highly significant association with both parameters γ and α simultaneously; only SNPs in the following genes (or within a 100 kb window) remained significant when testing for γ: *G6PC2/ABCB11*, *GCK/YKT6*, *GCKR* and *MTNR1B*, with effect per risk allele of increasing FG from 0.10 mmol/L to 0.047 mmol/L. Zooming in on simultaneous associations with the longitudinal and survival processes revealed well known genes, such as *TCF7L2*, which was shown in many meta-analyses to be associated with elevated FG and increased risk of T2D (Table 5). *MTNR1B* was also found to be associated (34 SNPs within 30kb) with $\alpha = -0.44$ ($\mathrm{p-value} = 9.37\times10^{-04}$) and $\gamma = 0.099$ ($\mathrm{p-value} = 1.33\times10^{-23}$) for SNP rs10830963, the SNP usually reported.

While rs17747324 showed consistent results, with the DIAGRAM meta-analysis for both α and γ (Table 5), rs10830963 showed an opposite effect on T2D compared to the effect reported in MAGIC for FG ($\alpha = 0.104$, $\mathrm{p-value} = 7.3\times10^{-07}$).

To better compare JM and TS, we repeated the analysis on the whole dataset using TS. As shown in Figure 9, approximation of p-values can be inaccurate, especially for parameter α; for parameter γ, approximations were quite close to the p-values provided via the joint likelihood framework.

Discussion

With the ever-increasing availability of genomic data generated by genotyping arrays and next generation sequencing, the need to develop and implement efficient models is important to ensure that statistical analysis will be achieved in a reasonable time frame. In this paper, we proposed a comparison of two approaches, namely the joint model (JM) and the two-step model (TS), to infer parameters accounting for a simultaneous SNP effect on longitudinal and survival processes without omitting information about value dropouts or status of the longitudinal variable of interest. In our real data application, FG is the longitudinal trait, whereas T2D diagnostic defines survival time of interest, both being linked together by the fact that an upper threshold on FG actually defines T2D onset (currently, FG > 7 mmol/L). Through simulations over different scenarios, we showed that joint models are less biased than classical separate approaches, could provide more insight regarding the event of interest, and could assess the potential impact of a SNP on incident cases of T2D.

By looking at different statistical measures, such as RMSE for bias in the model estimators, and by estimating computational time using the available R implementation of joint models, our study revealed that the use of an approximate method, such as TS, at a genome-wide scale might represent a good trade-off between bias and computational time. TS could be used to overcome the computational burden of current joint likelihood methods by exploiting available softwares performing the two steps, LME and CoxPH, and could help filter out SNPs with low or undetectable association during a first preliminary scan. However, depending on the dataset parameters (sample size, incidence rate, number of measures), a joint likelihood method is highly preferred to obtain accurate estimation of parameters γ and α, describing the SNP effect on the trajectory of FG and time-to-onset of T2D. Finally, using parallel and grid computing approaches will reduce the computational time to a more suitable time frame when applied at a genome-wide level (i.e. with millions of SNPs).

In our real data application, results observed for *MTNR1B* in the French cohort D.E.S.I.R., even if they seemed inconsistent with previous studies, may uncover some interesting peculiarities pertaining to T2D incident cases in this population. In the literature, SNPs in *MTNR1B* were reported for being associated with increased blood FG and elevated T2D risk, but meta-analyses were performed on populations with different genetic backgrounds, and the two traits were never co-analysed jointly. However, we recognize that *MTNR1B* associations identified in our study need to be confirmed and replicated in other longitudinal cohorts, as they might represent cohort-specific associations. In addition, a major limitation of our study is the low number of incident T2D cases in the D.E.S.I.R. cohort (167 incident T2D cases over 5,212 subjects followed up over 9 years).

Acknowledgments

This study was supported by grants for funding of scientific research conducted in France and within the European Union: "Centre National de la Recherche Scientifique", "Université de Lille 2", "Institut Pasteur de Lille", "Société Francophone du Diabète", "Lilly", "Contrat de Plan Etat-Région", "Agence Nationale de la Recherche", ANR-10-LABX-46, ANR EQUIPEX Ligan MP: ANR-10-EQPX-07-01, European Research Council GEPIDIAB - 294785.

Conflict of interest disclosure

The authors declare that they have no conflict of interest.

List of tables

Table 1. Parameters and numerical values used for sensitivity analysis and simulations (based on results from rs17747324 within gene *TCF7L2*.

Parameters	Values
Number of subjects (n)	4,352
Number of measures (m)	4
Incidence rate (d)	0.0384
Minor allele frequency (f)	0.244
Random effects (θ)	$\sim \mathcal{N}_2\left(\begin{bmatrix} 4.55 \\ 0.0108 \end{bmatrix}, \begin{bmatrix} 0.143 & -0.00109 \\ -0.00109 & 6.8\times10^{-04} \end{bmatrix}\right)$
SNP effect on Y_{ij} (γ)	0.0229
SNP effect on T_i (α)	0.265
Association between Y_{ij} and T_i (β)	3.17
Error term (ϵ)	$\sim \mathcal{N}(0, 0.305^2)$

Table 2. RMSE, Type 1 Error and Power from default simulation settings rs17747324 (*TCF7L2*).

	Joint Model			Two-Step Model		
	RMSE	T1E	POWER	RMSE	T1E	POWER
α	0.137	0.036	45.4%	0.139	0.051	48.5%
γ	0.01	0.051	61.8%	0.01	0.05	58.8%

Table 3. Approximate computational times (in seconds) using function *system.time* of R software. System time is computed ten times per sample size (number of subjects). Extrapolation are displayed (in days) for 100,000 tests.

	Joint Model		Two-Step Model	
Sample Size	mean (sd) per test	100K test	mean (sd) per test	100K test
500	51 s (3.4)	59 d	0.71 s (0.066)	0.82 d
2,500	100 s (11)	120 d	3.1 s (0.092)	3.6 d
5,000	180 s (25)	210 d	6.3 s (0.17)	7.3 d
10,000	340 s (34)	400 d	9 s (0.22)	10 d

Table 4. List of loci found to be associated within the joint modelling framework with both FG and T2D, previously shown as associated with FG and/or T2D in the NHGRI GWAS Catalogue (Welter et al., 2014).

SNP (gene)	α (p-value)	γ (p-value)	β (p-value)	Power ($\beta\gamma + \alpha$)
rs6945660_G (ETV1)	0.55 (3.7×10^{-02})	0.0352 (2.5×10^{-02})	3.48 (9.6×10^{-45})	69.7%
rs1942873_C (MC4R)	0.41 (1.3×10^{-02})	0.0234 (3.7×10^{-02})	3.14 (1.9×10^{-41})	69.6%
rs55899248_G (TCF7L2)	0.292 (2.7×10^{-02})	0.0253 (1.7×10^{-02})	3.49 (1.7×10^{-44})	55.3%
rs17301514_A (ST6GAL1)	-0.657 (4.4×10^{-03})	0.0451 (3.4×10^{-03})	3.65 (2.9×10^{-45})	45.8%
rs833425_C (PTPRD)	0.321 (5.0×10^{-02})	0.0432 (4.2×10^{-03})	3.51 (1.3×10^{-43})	44.2%
rs7072870_A (C10orf35)	-0.404 (7.5×10^{-03})	0.0248 (2.2×10^{-02})	3.58 (1.7×10^{-45})	39.6%
rs61871514_A (KCNQ1)	0.425 (4.7×10^{-02})	0.0457 (2.0×10^{-02})	3.18 (8.5×10^{-42})	39.4%
rs9883865_A (ADAMTS9)	-0.598 (7.5×10^{-04})	0.0426 (1.2×10^{-02})	3.2 (5.9×10^{-42})	34.9%
rs114508985_C (HLA)	-0.294 (2.1×10^{-02})	0.0209 (3.0×10^{-02})	3.22 (8.2×10^{-43})	27.1%
rs10814856_T (GLIS3)	-0.265 (4.0×10^{-02})	0.0248 (1.5×10^{-02})	3.2 (1.5×10^{-42})	18.5%
rs73025532_C (SLC22A1)	-0.377 (4.8×10^{-02})	0.0317 (3.6×10^{-02})	3.58 (1.3×10^{-45})	17.3%
rs11769484_C (JAZF1)	-0.254 (4.8×10^{-02})	0.0221 (3.6×10^{-02})	3.21 (2.1×10^{-42})	16.9%
rs6450176_G (ARL15)	-0.291 (1.8×10^{-02})	0.0365 (3.0×10^{-04})	3.54 (2.2×10^{-45})	15.2%
rs4712580_C (CDKAL1)	-0.289 (4.2×10^{-02})	0.0313 (7.4×10^{-03})	3.57 (1.2×10^{-45})	14.0%
rs10830963_G (MTNR1B)	-0.44 (9.4×10^{-04})	0.0991 (1.3×10^{-23})	3.25 (3.6×10^{-42})	10.2%
rs853787_T (ABCB11)	-0.247 (4.3×10^{-02})	0.0831 (9.3×10^{-19})	3.21 (1.7×10^{-42})	3.3%
rs560887_C (G6PC2)	-0.315 (1.2×10^{-02})	0.0992 (9.6×10^{-25})	3.21 (1.3×10^{-42})	2.6%

Table 5. Effect sizes on FG and T2D risk estimated using JM. Comparison is shown with effect sizes as reported by consortia meta-analyses in genes MTNR1B and TCF7L2.

	α (p-value)		γ (p-value)		β (p-value)
SNP (gene)	JM (D.E.S.I.R.)	DIAGRAM	JM (D.E.S.I.R.)	MAGIC	JM (D.E.S.I.R.)
rs10830963_G (MTNR1B)	-0.44 (9.4×10^{-04})	0.104 (7.3×10^{-07})	0.0991 (1.3×10^{-23})	0.079 (1.3×10^{-68})	3.25 (3.6×10^{-42})
rs17747324_C (TCF7L2)	0.265 (4.1×10^{-02})	0.358 (8.5×10^{-55})	0.0229 (3.0×10^{-02})	0.025 (6.5×10^{-08})	3.17 (8.9×10^{-42})

List of figures

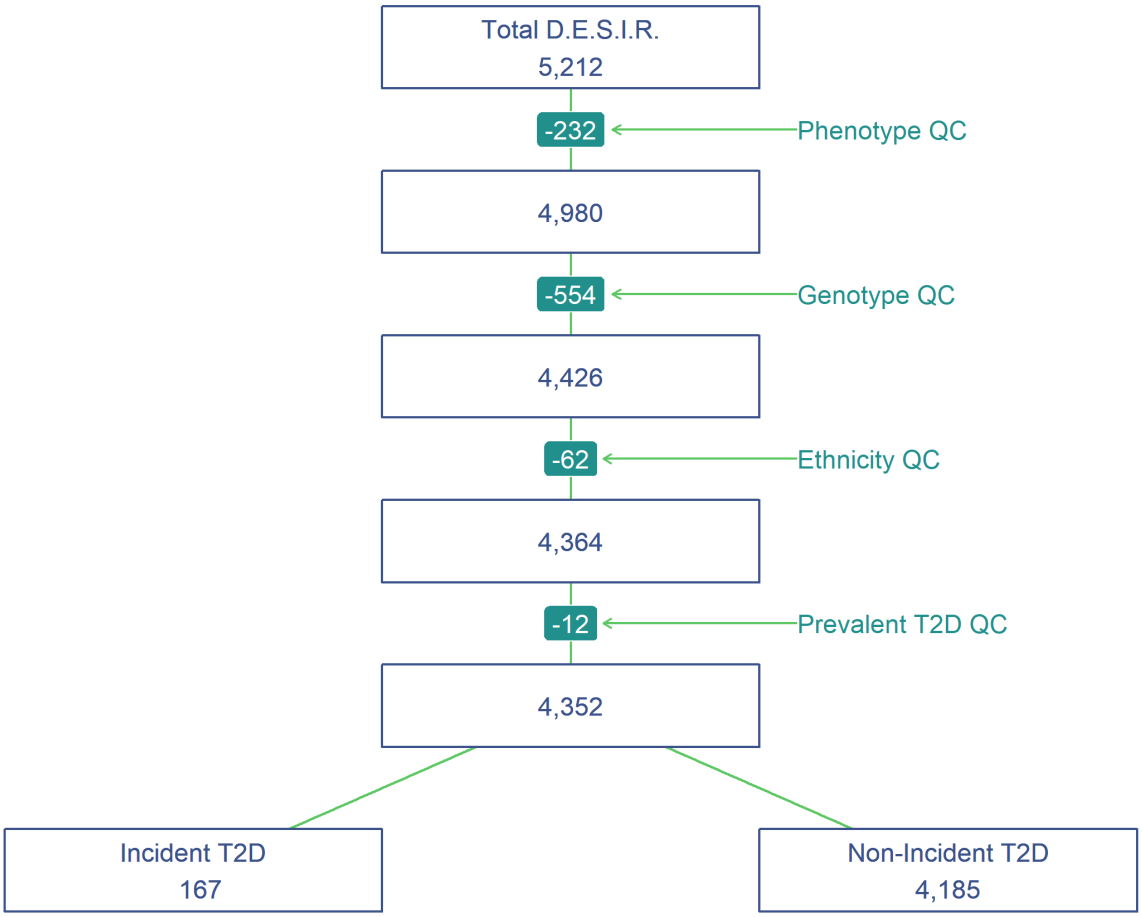

Figure 1. Flowchart quality control on subjects from the French cohort D.E.S.I.R.

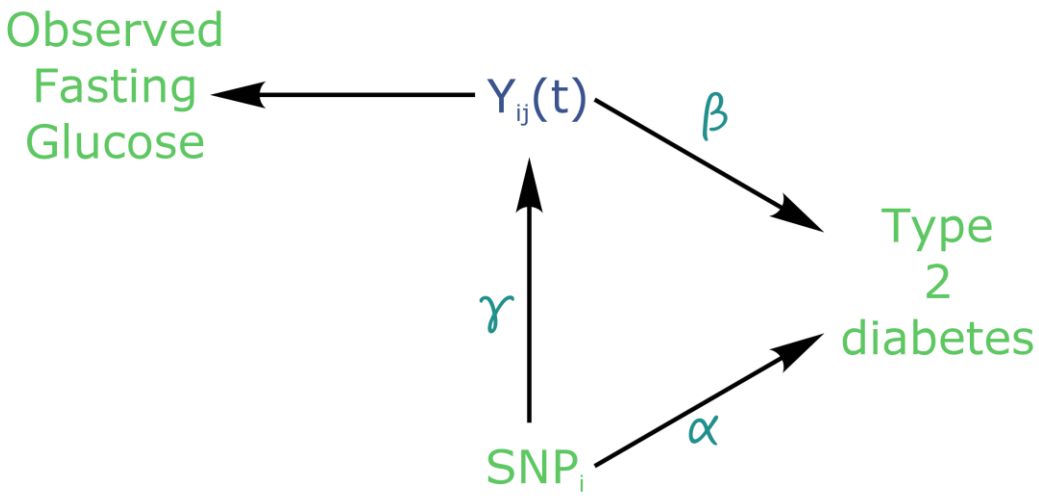

Figure 2 Causal diagram for joint modelling applied to fasting glucose (FG) and type 2 diabetes (T2D) (adapted from Ibrahim, Chu, & Chen (2010)). SNP: Single Nucleotide Polymorphism.

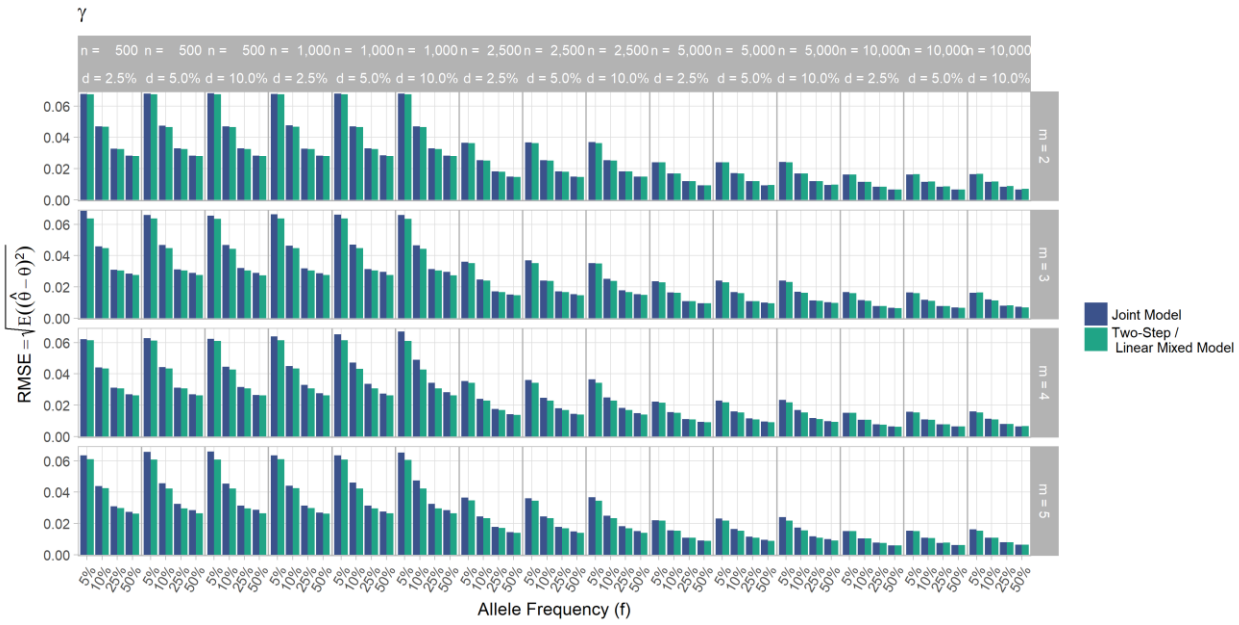

Figure 3 Simulation study for accuracy of estimator $\hat{\gamma}$ provided by the joint model (JM package) and by the linear mixed effect model (nlme package). m: number of measures; n: number of subjects; d: incidence rate.

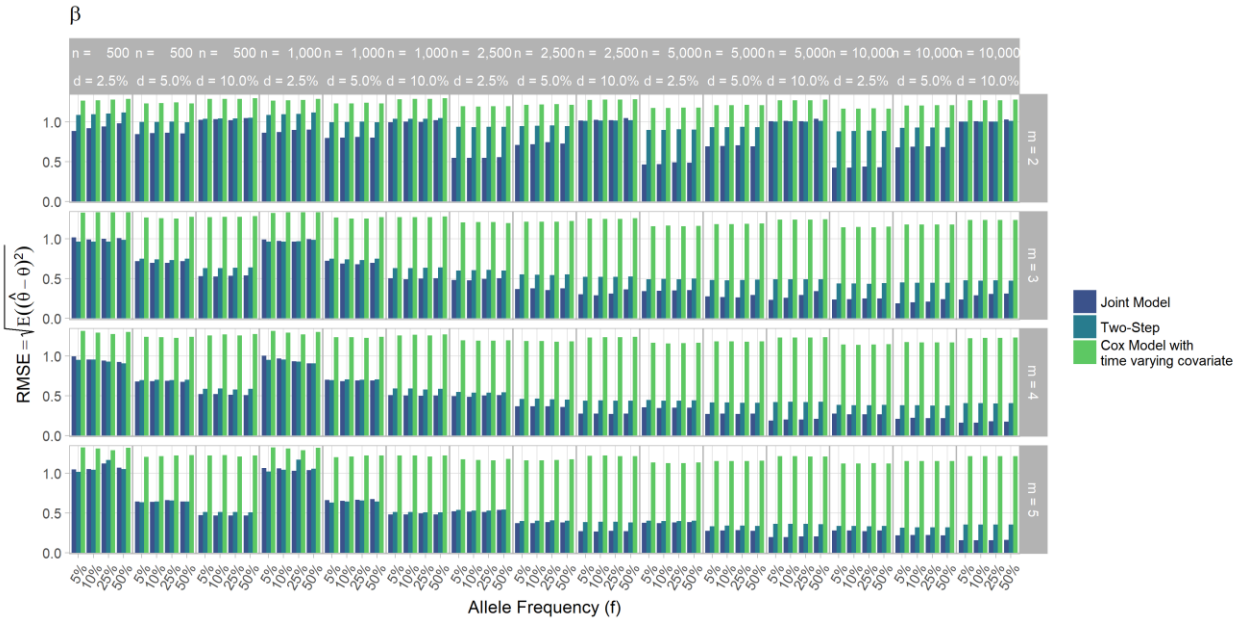

Figure 4 Simulation study for accuracy of estimator $\hat{\beta}$ provided by the joint model (JM package) and by the linear mixed effect model (nlme package). m: number of measures; n: number of subjects; d: incidence rate.

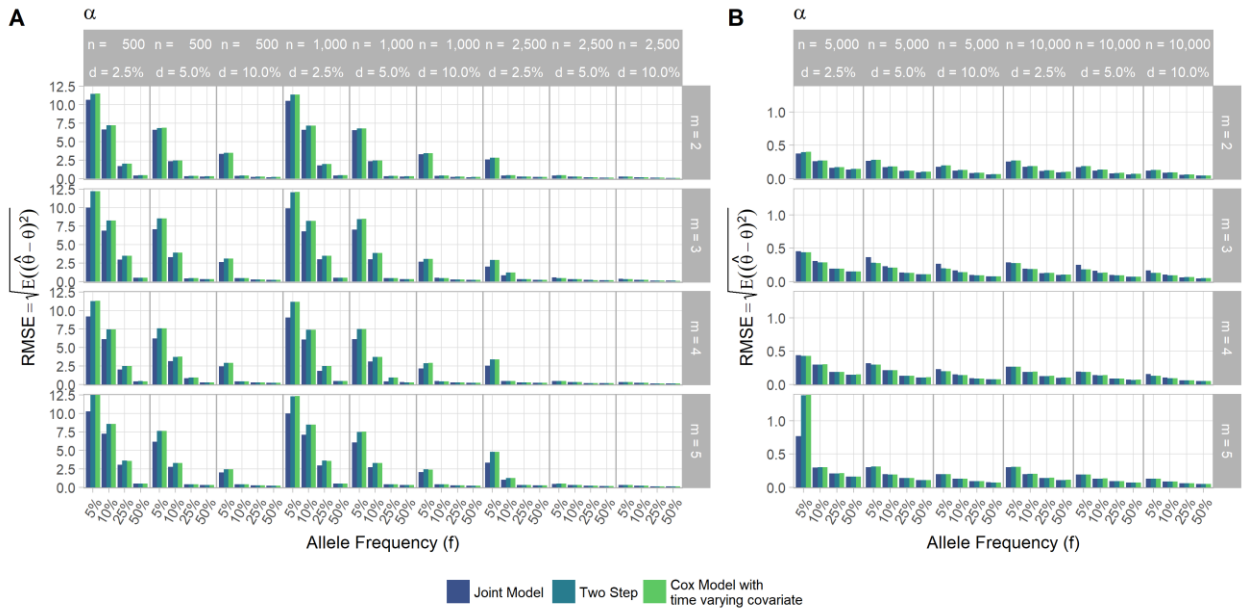

Figure 5 Simulation study for accuracy of estimator $\hat{\alpha}$ provided by the joint model (JM package) and by the linear mixed effect model (nlme package). m: number of measures; n: number of subjects; d: incidence rate.

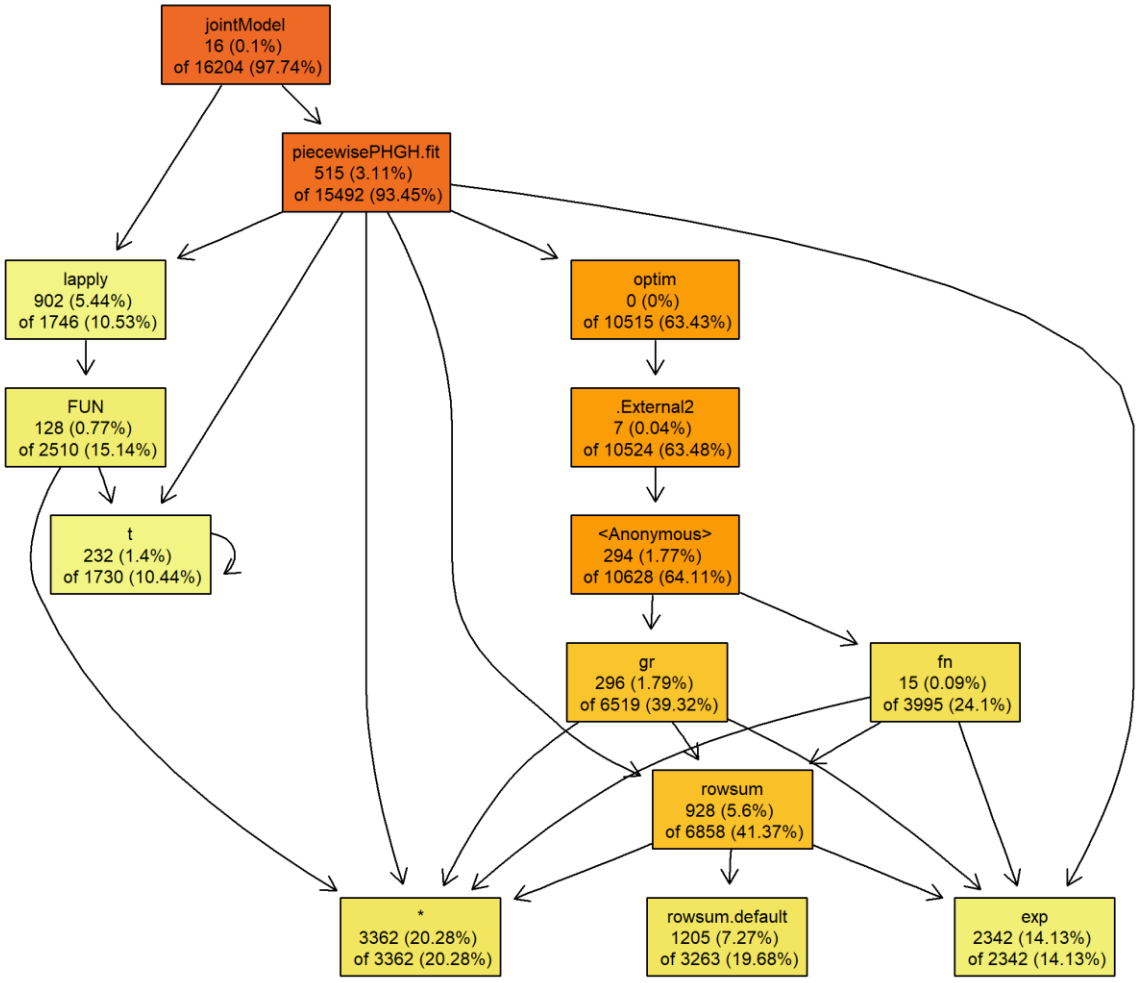

Figure 6 Call tree diagram of the main function *jointmodel* in the R package JM. Call based on a simulated dataset with three longitudinal measures and 5,000 subjects (other parameter values set as in Table 1).

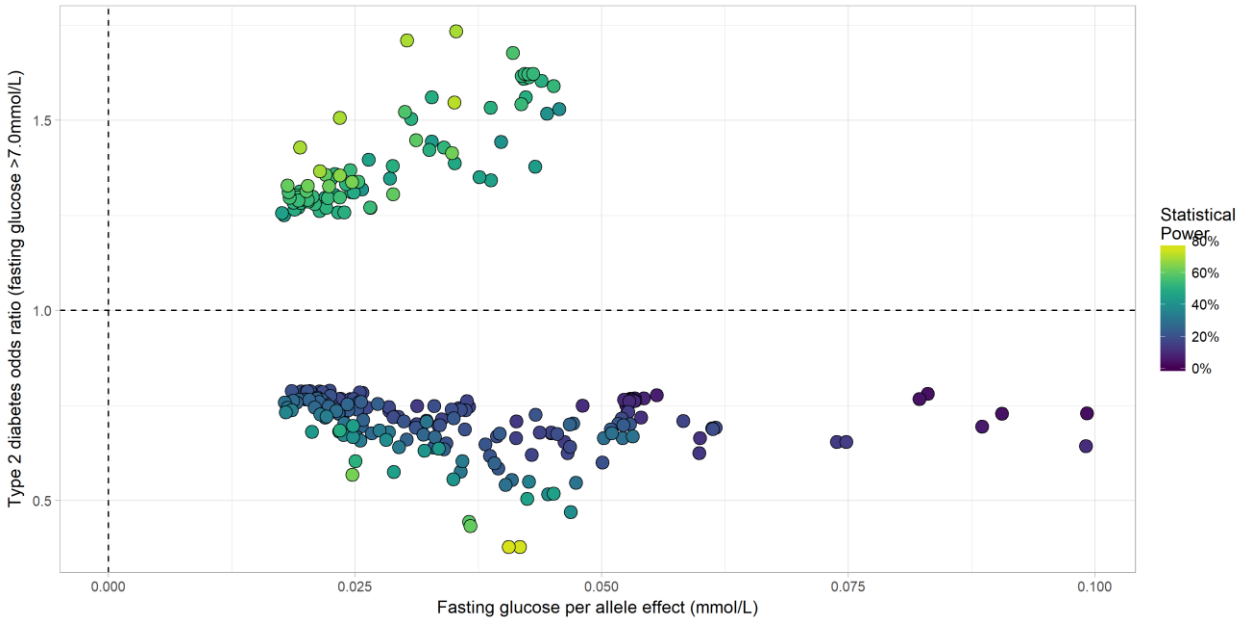

Figure 7 Results from statistical analysis using JM (Rizopoulos, 2010, 2016). Estimated effects of γ are displayed on the x-axis, with corresponding estimated odds ratio $\exp(\alpha)$ on the y-axis. Statistical power reported is the theoretical (retrospective) power to detect a joint effect $\beta\gamma + \alpha$ based on estimated model parameters (Chen et al., 2011).

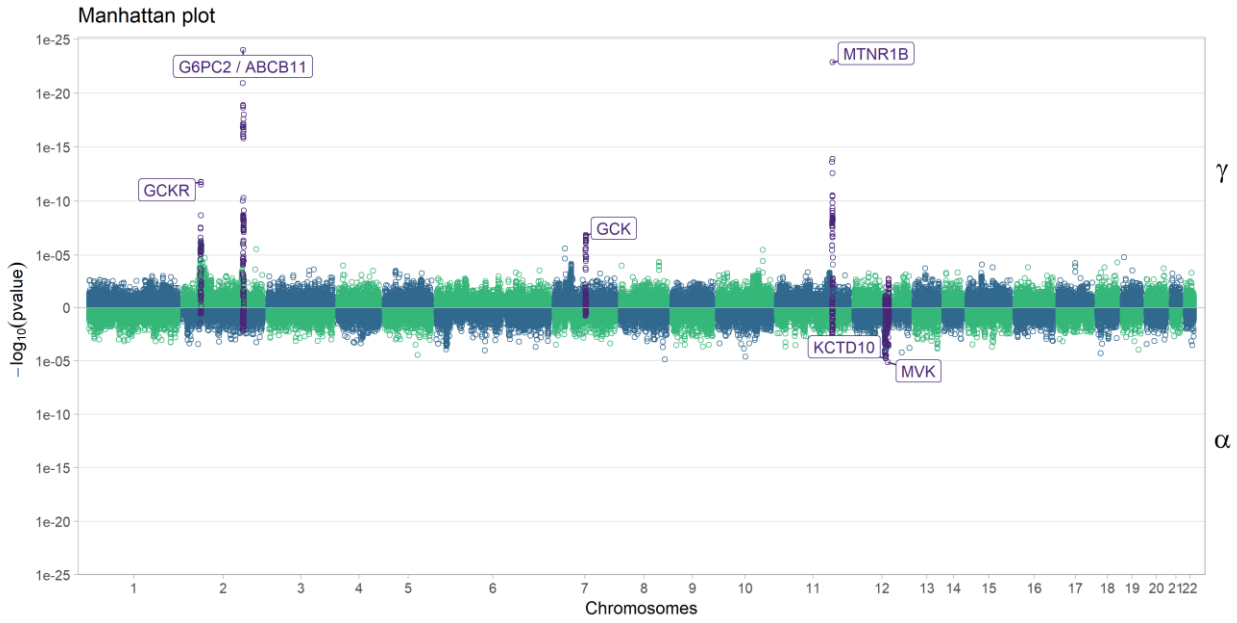

Figure 8 Manhattan plot for estimated effects of γ and α using JM. Results are presented for the cleaned set of 101,305 SNPs.

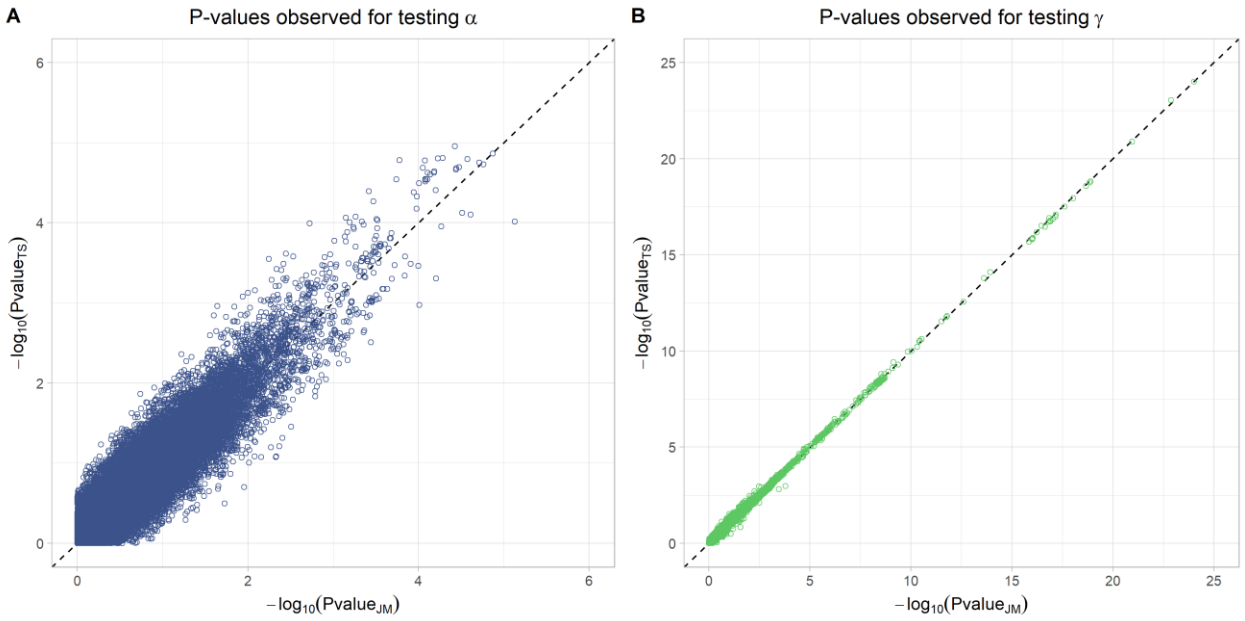

Figure 9 Testing for α (SNP effect on onset of T2D) and γ (SNP effect on the trajectory of FG) using Two-Step approach compared to Joint Model approach. On the x-axis, $-\log_{10}(p)$ from the Joint Model and on the y-axis the corresponding $-\log_{10}(p)$ from the approximate Two-Step approach.

References

Albert, P. S., & Shih, J. H. (2010a). An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *The Annals of Applied Statistics*, *4*(3), 1517–1532. https://doi.org/10.1214/10-AOAS339

Albert, P. S., & Shih, J. H. (2010b). On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. *Biometrics*, *66*(3), 983–987. https://doi.org/10.1111/j.1541-0420.2009.01324_1.x

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, *31*(29), 3946–3958. https://doi.org/10.1002/sim.5452

Balkau, B. (1996). An epidemiologic survey from a network of French Health Examination Centres, (D.E.S.I.R.): epidemiologic data on the insulin resistance syndrome. *Revue D'épidémiologie et de Santé Publique*, *44*(4), 373–375.

Bouatia-Naji, N., Rocheleau, G., Van Lommel, L., Lemaire, K., Schuit, F., Cavalcanti-Proença, C., … Froguel, P. (2008). A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science (New York, N.Y.)*, *320*(5879), 1085–1088. https://doi.org/10.1126/science.1156849

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, L. M., Ibrahim, J. G., & Chu, H. (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in Medicine*, *30*(18), 2295–2309. https://doi.org/10.1002/sim.4263

Diggle, P., & Kenward, M. G. (1994). Informative Drop-Out in Longitudinal Data Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *43*(1), 49–93. https://doi.org/10.2307/2986113

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., … Barroso, I. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, *42*(2), 105–116. https://doi.org/10.1038/ng.520

Elashoff, R. M., Li, G., & Li, N. (2008). A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics*, *64*(3), 762–771. https://doi.org/10.1111/j.1541-0420.2007.00952.x

Elashoff, R., li, G., & Li, N. (2016). *Joint Modeling of Longitudinal and Time-to-Event Data* (1st ed.). Chapman and Hall/CRC.

Ibrahim, J. G., Chu, H., & Chen, L. M. (2010). Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal of Clinical Oncology*, *28*(16), 2796. https://doi.org/10.1200/JCO.2009.25.0654

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118032985

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., … McCarthy, M. I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990. https://doi.org/10.1038/ng.2383

Pinheiro, J., Bates, D., & R-core. (2017). *Nlme: Linear and nonlinear mixed effects models*. Retrieved from https://CRAN.R-project.org/package=nlme

Proust-Lima, C., Joly, P., Dartigues, J.-F., & Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis*, *53*(4), 1142–1154. https://doi.org/10.1016/j.csda.2008.10.017

Purcell, S., & Chang, C. (2015). PLINK v1.90b3.36.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, *35*(9), 1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Rizopoulos, D. (2016). *JM: Joint modeling of longitudinal and survival data*. Retrieved from https://CRAN.R-project.org/package=JM

Rizopoulos, D., & Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, *30*(12), 1366–1380. https://doi.org/10.1002/sim.4205

Rung, J., Cauchi, S., Albrechtsen, A., Shen, L., Rocheleau, G., Cavalcanti-Proença, C., … Sladek, R. (2009). Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature Genetics*, *41*(10), 1110–1115. https://doi.org/10.1038/ng.443

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., … Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, *445*(7130), 881–885. https://doi.org/10.1038/nature05616

Sun, J., Sun, L., & Liu, D. (2007). Regression Analysis of Longitudinal Data in the Presence of Informative Observation and Censoring Times. *Journal of the American Statistical Association*, *102*(480), 1397–1406. https://doi.org/10.1198/016214507000000851

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Therneau, T. M. (2017). *Survival: Survival analysis*. Retrieved from https://CRAN.R-project.org/package=survival

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. (K. Dietz, M. Gail, K. Krickeberg, J. Samet, & A. Tsiatis, Eds.). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4757-3294-8

Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica*, *14*, 809–834.

Tsiatis, A. A., DeGruttola, V., & Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, *90*(429), 27–37. https://doi.org/10.2307/2291126

Vaxillaire, M., Yengo, L., Lobbens, S., Rocheleau, G., Eury, E., Lantieri, O., … Froguel, P. (2014). Type 2 diabetes-related genetic risk scores associated with variations in fasting plasma glucose and development of impaired glucose homeostasis in the prospective DESIR study. *Diabetologia*, *57*(8), 1601–1610. https://doi.org/10.1007/s00125-014-3277-x

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., … Boehnke, M. (2012). The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genetics*, *8*(8), e1002793. https://doi.org/10.1371/journal.pgen.1002793

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(D1), D1001–D1006. https://doi.org/10.1093/nar/gkt1229

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, *53*(1), 330. https://doi.org/10.2307/2533118