Published in final edited form as:

Genet Epidemiol. 2017 November ; 41(7): 620–635. doi:10.1002/gepi.22058.

A genetic stochastic process model for genome-wide joint analysis of biomarker dynamics and disease susceptibility with longitudinal data

Liang He^{1,*}, Ilya Zhbannikov¹, Konstantin G. Arbeev¹, Anatoliy I. Yashin¹, and Alexander M. Kulminski^{1,*}

¹Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, NC 27708

Abstract

Unraveling the underlying biological mechanisms or pathways behind the identified effects of genetic variations on complex diseases remains one of the major challenges in the post-GWAS era. To further explore the relationship between genetic variation, biomarkers, and diseases for elucidating underlying pathological mechanism, a huge effort has been placed on examining pleiotropic and gene-environmental interaction effects. We propose a novel genetic stochastic process model (GSPM) that can be applied to GWAS and jointly investigate the genetic effects on longitudinally-measured biomarkers and risks of diseases. This model is characterized by more profound biological interpretation and takes into account the dynamics of biomarkers during follow-up when investigating the hazards of a disease. We illustrate the rationale and evaluate the performance of the proposed model through two genome-wide association studies. One is to detect single nucleotide polymorphisms (SNPs) having interaction effects on type 2 diabetes (T2D) with body mass index (BMI) and the other is to detect SNPs affecting the optimal BMI level for protecting from T2D. We identified multiple SNPs that showed interaction effects with BMI on T2D, including a novel SNP rs11757677 in the *CDKALI* gene ($p=5.77e-07$). We also found a SNP rs1551133 located on 2q14.2 that reversed the effect of BMI on T2D ($p=6.70e-07$). In conclusion, the proposed GSPM provides a promising and useful alternative in GWAS of longitudinal data for interrogating pleiotropic and interaction effects to gain more insights into the relationship between genes, quantitative biomarkers and risks of complex diseases.

Keywords

genetic stochastic process model; joint association model; type 2 diabetes; body mass index; gene-environment interaction

*Corresponding authors: Alexander M. Kulminski, Liang He, kulminsk@duke.edu, lianghe@mit.edu, Address: Alexander M. Kulminski, Center for Population Health and Aging, Duke University, Street Address: 2024 W. Main Street, Room A102F, Durham, NC, 27705, USA; Liang He, MIT Computer Science and Artificial Intelligence Laboratory The Stata Center, Building 32, Street Address: 32 Vassar Street, Cambridge, MA, 02139, USA.

Competing Interests

The authors declare that they have no competing interests.

Accession numbers

This manuscript was prepared using a limited access dataset obtained through dbGaP (accession number phs000280.v2.p1)

Introduction

Genome wide association studies (GWAS) have so far identified thousands of genetic variations associated with a variety of complex traits and diseases; the potential biological or pathological role of the vast majority of these statistically significant signals is still elusive (Price et al., 2015). Unravelling the underlying biological mechanisms, or pathways behind the identified genetic variations on complex diseases, remains one of the major challenges in the post-GWAS era. General genetic analysis using regression models is successful in identifying significant association, but provides little information on the biological interpretation. The influence of an identified genetic variation on a complex disease may be mediated by its effect on the level of certain causal intermediate phenotypes, which is called a mediated effect (Hardy et al., 2014; Sritara et al., 2014). A genetic variation may affect both biomarkers and diseases through independent pathways, which results in biological pleiotropic effects (Solovieff et al., 2013). A genetic variation can also modulate the effect of some causal biomarkers, or a biomarker can affect the effect of a causal variant, both of which lead to the presence of gene-environment interaction (GEI) effects (Belsky et al., 2014). On the other hand, genetic effects may result from influence on the frailty threshold of some key causal biomarkers. Thus, to further explore the relationship between genetic variations, biomarkers, and diseases for elucidating underlying pathological mechanism, recently a huge effort has been placed on examining pleiotropic effects and GEI effects.

Pleiotropic effects can often be interrogated using multivariate statistical models, which estimate the genetic effects on multiple phenotypes jointly (Shriner, 2012). Such strategy has several drawbacks, including difficulties in jointly modeling different types of phenotypes and in handling missing data. For example, in longitudinal data, diseases are often presented as time-to-event outcomes while physiological indices are continuous traits. Statistical methods based on combining summary statistics can overcome these issues (Xu et al., 2003; Zhu et al., 2015) and have been applied to combine the evidence for a large number of biomarkers and age-related diseases (He et al., 2016a). Numerous joint models of longitudinal and time-to-event data based on common latent random effects (Guo and Carlin, 2004; Tsiatis and Davidian, 2004) have been proposed, but the application of these models to GWAS is still limited. On the other hand, estimation of the contributions of genes alone without consideration of their interactions with environmental factors, which is typical situation in GWAS, would incorrectly estimate the proportion of the disease that is explained by genes, the environment, and their joint effect (Hunter, 2005). Investigation of GEI improves the understanding of the underlying mechanism of identified genetic variations. GEI is also a potential candidate for the cause of missing heritability for many complex traits and diseases. For example, a previous study on dynamic heritability has revealed that the genetic contribution to body mass index (BMI) changes throughout the lifespan, which indicates the involvement of GEI on the trajectory of BMI (He et al., 2016b). Recently, several models including a two-step filtering approach have been proposed to increase the efficiency and statistical power for genome-wide environmental interaction (GWEI) analysis (Winham and Biernacka, 2013).

A model based on stochastic process has been proposed to jointly estimate the dynamics of physiological indices or biomarkers with continuous measurements and their relationship

with mortality across human lifespan (Woodbury and Manton, 1977; Yashin et al., 1985). This stochastic process model (SPM) has an advantage of encapsulating multiple parameters that characterize important concepts in aging research such as allostatic load, diffusion coefficient, and biomarker norms, and thus can improve our understanding of genetic influence on aging process (Yashin et al., 2007). The use of SPM with continuously changing biomarkers in statistical analyses of longitudinal data has been described in (Yashin and Manton, 1997; Yashin et al., 1986). Further development of this model includes incorporation of hidden heterogeneity (Arbeev et al., 2014; Yashin et al., 2008) and further incorporation of genetic markers (in the version of the model termed as GenSPM) (Arbeev et al., 2009). Although the SPM together with its discrete time version has been adopted to study how the levels of various biomarkers affect aging process (Yashin et al., 2009, 2013), this model has not been broadly applied in the context of GWAS partly due to its computational complexity in the estimation and inference procedures. The practical computational issue particularly manifests in GWAS where thousands and millions of single nucleotide polymorphisms (SNPs) need to be evaluated within reasonable time.

Exploiting the intrinsic virtue of joint modeling, the SPM should be an ideal candidate to directly detect pleiotropic effects of genetic variations on biomarkers and a disease and gene-biomarker interaction effects. In this work, we leveraged the joint model developed in (Yashin et al., 2007) to propose a novel modification of the genetic stochastic process model (termed as GSPM) that has an analytical solution and can be applied to GWAS to investigate genetic effects on biomarkers and diseases over a portion of the life course. We overcome the computational issue in the estimation procedure by solving the likelihood function analytically with some reasonable assumptions on the parameters. We incorporate genetic information into the SPM and illustrate how the inference on GEI and pleiotropic effects can be drawn using the proposed model. We evaluate the type 1 error rate with simulation data. We then apply the GSPM to an NIH-funded large scale longitudinal dataset to investigate GEI effects on type 2 diabetes (T2D). GEI may play a more important role in age-related diseases such as T2D, cardiovascular diseases (CVD) and neurodegenerative diseases (ND) because the genetic variations associated with age-related diseases are typically not subject to natural selection due to the late onset of these diseases after reproduction. Thus, the effects of these genetic variations tend to be modulated by various endophenotypes that can serve as a proxy for environmental exposures and other exogenous factors such as social influences. For example, interactions between genes and physical activity, diet, BMI in T2D have been reported (Cornelis and Hu, 2012; Franks, 2011). Given the recognized evidence of causal effect of BMI on T2D (Holmes et al., 2014), we focus on detecting SNPs through a genome-wide search that modulate the causal effect of BMI on T2D. We further perform a GWAS using the GSPM to examine whether there exist any SNPs that can affect the threshold of BMI on T2D.

Method

Model specification

The SPM, originally proposed in (Woodbury and Manton, 1977), is described by jointly characterizing the dynamics of a continuous physiological index $Y(t)$ (e.g., a biomarker or

risk factor of interest such as BMI; in this work $Y(t)$ is assumed to be a one-dimensional function.) and a hazard function $\mu(t)$ of a disease or mortality. More recently, an extended model with more explicitly biologically-interpretable parameters and its application to longitudinal data on aging study are provided (Yashin et al., 2007). More specifically, the model assumes that the dynamics of the biomarker over age t (or other time scales) obeys the stochastic differential equation

$$dY(t) = a(t)(Y(t) - f_1(t))dt + b(t)dW_t, \quad \text{given the initial value } Y(t_0) = Y_0 \text{ at } t_0, \quad (1)$$

where $f_1(t) (>0)$ describes the asymptotic equilibrium level or mean allostatic state that the biomarker should be adaptively subject to over t , and $a(t) (<0)$ is the rate of the adaptive capacity, indicating how rapidly the physiological state measured by the biomarker approaches its asymptotic level. Thus, the drift term $a(t)(Y(t) - f_1(t))$ in the equation (1) characterizes the allostasis process (McEwen and Wingfield, 2003) of the physiological state using a negative feedback mechanism. The parameter $b(t) (>0)$ measures the magnitude of random errors comprising the influence of all other random factors at age t , and W_t is a Wiener process. We assume that $b(t)$ is a deterministic function of age and adopt the Ito definition for the stochastic integral $\int_0^T b(t)dW_t$. Based on this modeling, $Y(t)$ is a one-dimensional stochastic process of diffusion type, and follows a Gaussian distribution at any time t if the initial distribution $Y(t_0)$ is Gaussian.

We assume that the process $Y(t)$ terminates at the occurrence of some event of interest such as death (in the case of a disease, we can always assume that this process stops at the onset of the disease.) with a positive probability, and the conditional hazard of this event is modeled by

$$\mu(t, Y(t)) = \mu_0(t) + q(t)(Y(t) - f(t))^2 \quad (2)$$

where $\mu_0(t)$ is the baseline hazard corresponding to the physiological norms, $f(t) (>0)$ is the optimal value of the biomarker conferring the lowest risk of the event, and $q(t)$ describes the extent to which a deviation from the optimal value of this biomarker affects the incidence of the event. The parameter $f(t)$ can be different from $f_1(t)$. For example, if we consider BMI and diabetes, the allostatic adaptation process of BMI might not be optimized for diabetes. Average BMI follows a steady increasing trend after adolescence, but the optimal level of BMI for avoiding diabetes stays low.

Suppose that the initial distribution of the biomarker $Y(t_0)$ is Gaussian. Then the conditional distribution $P(Y(t) | T > t)$ given the event $\{T > t\}$ is also Gaussian (Yashin, 1984). It has been shown (Yashin et al., 2007) that the marginal hazard can be expressed by

$$\mu(t) = \mu_0(t) + q(t)(m(t) - f(t))^2 + q(t)\gamma(t), \quad (3)$$

where $m(t)$ and $\gamma(t)$ are the conditional expectation and conditional variance of $Y(t)$ that conform to the system of ordinary differential equations

$$\begin{cases} \frac{dm(t)}{dt} = a(t)(m(t) - f_1(t)) - 2q(t)\gamma(t)(m(t) - f(t)), \\ \frac{d\gamma(t)}{dt} = 2a(t)\gamma(t) + b^2(t) - 2q(t)\gamma^2(t) \end{cases}, \quad (4)$$

given the initial values $m(t_0) = e_0$ and $\gamma(t_0) = v_0$. The parameters can be estimated from longitudinal data that have the format shown in Table 1. Suppose that we have N individuals and observe n_i measurements $y_1^i, \dots, y_{n_i}^i$ from the individual i ($i \in \{1, \dots, N\}$) at the time points $t_1^i, \dots, t_{n_i}^i$ and a variable $\delta_i \in \{1, 0\}$ indicating whether failure or censoring occurs at the end of the follow-up τ_i for the individual i . Previous study (Yashin et al., 2007) has shown that the joint log-likelihood of $\mu(t)$ and $Y(t)$ can be expressed by

$$l \propto \sum_{i=1}^N \left(\sum_{j=1}^{n_i} -\frac{1}{2} (\log(\gamma^i(t_j^i | t_{j-1}^i)) - \frac{(y^i(t_j^i) - m^i(t_j^i | t_{j-1}^i))^2}{\gamma^i(t_j^i | t_{j-1}^i)}) \right) + \delta_i \log(\mu^i(\tau_i)) - \int_{t_0^i}^{\tau_i} \mu^i(t) dt,$$

where $Y^i(t_j^i | t_{j-1}^i)$ represents $Y^i(t_j^i)$ given the initial value $Y^i(t_{j-1}^i)$.

Parameter estimation based on the likelihood

Maximum likelihood estimation (MLE) is not straightforward because it is hard to solve the equation system (4) explicitly in general. Numerical methods (e.g., the Runge-Kutta methods (Ascher and Petzold, 1998)) are required to evaluate the values of $m(t)$ and $\gamma(t)$ at each time point. Furthermore, the lack of explicit expression of $m(t)$ and $\gamma(t)$ would result in the difficulty in expressing the likelihood function analytically. Consequently, the MLE resorts to some heuristic simplex methods (Nelder and Mead, 1965) which do not require a gradient function. Although this strategy has been shown to work properly (Zhbannikov et al., 2017), it has certain major drawbacks when applied in the context of genetic association analysis. First, in general these numerical methods are too time-consuming to be efficiently applied in a large-scale analysis, such as GWAS given currently available resources. Second, the simplex methods adopted in the R implementation (Zhbannikov et al., 2017) do not provide a standard error for the estimation per se. Thus, the evaluation of significance, which is important in GWAS, resorts to resampling methods which can substantially increase the computational burden due to multiple testing. Finally, the results from the simplex methods are more sensitive to the choice of initial values.

To overcome these issues, we propose the GSPM dedicated to genetic association analysis that solves the Riccati equation in (4) analytically by assuming that the parameters

(a, b, q, f, f_1) are independent of t . We show that under such assumptions, the analytical solution to the equation (4) is

$$\gamma(t) = \frac{2a+u}{4q} + \frac{u}{c_1 e^{ut} - 2q'}$$

$$m(t) = (2q - c_1 e^{ut})^{-1} \left[-\frac{2w}{u} - c_1 e^{ut} \left(\frac{2a(f-f_1)}{u} + f \right) + c_2 e^{\frac{ut}{2}} \right],$$

where we have $u = 2\sqrt{a^2 + 2qb^2}$ and $w = q(2a(f-f_1) - fu)$. The constants c_1 and c_2 are determined by the initial mean and variance of $Y(t)$ at t_0 . The detailed derivation of the solution can be found in the supplemental material Text S1A. We further assume that the baseline hazard has the exponential form $\mu_0(t) = \mu_0 e^{\theta t}$ with $\mu_0, \theta > 0$, and thus the minus log-likelihood turns out to be

$$\begin{aligned} -l \propto \sum_{i=1}^N \left(\sum_{j=1}^{n_i} \frac{1}{2} \log \left(\frac{2a+u}{4q} + \frac{u}{c_1^{i,j} e^{ut_j^i} - 2q} \right) + \frac{1}{2} \frac{\left(y^i(t_j^i) - (2q - c_1^{i,j} e^{ut_j^i})^{-1} \left(-\frac{2w}{u} - c_1^{i,j} e^{ut_j^i} \left(\frac{2a(f-f_1)}{u} + f \right) + c_2^{i,j} e^{\frac{ut_j^i}{2}} \right) \right)^2}{\frac{2a+u}{4q} + \frac{u}{c_1^{i,j} e^{ut_j^i} - 2q}} \right. \\ \left. - \delta_i \log(\mu^i(\tau_i)) \right. \\ \left. + \int_{t_0}^{\tau_i} \mu^i(t) dt, \right. \end{aligned}$$

where $c_1^{i,j}, c_2^{i,j}$ are determined by the value at the previous time point t_{j-1}^i , and we show within each interval

$$\begin{aligned} \int_{t_{j-1}^i}^{t_j^i} \mu^i(t) dt &= \left(qr^* + \frac{4qa^2}{u^2} (f-f_1)^2 - \frac{u}{2} \right) (t_j^i - t_{j-1}^i) \\ &+ \frac{1}{2} \log \frac{2q - \left(\frac{u}{r_0 - r^*} + 2q \right) e^{u(t_j^i - t_{j-1}^i)}}{\frac{-u}{r_0 - r^*}} \\ &+ \frac{\mu_0}{\theta} (e^{\theta t_j^i} - e^{\theta t_{j-1}^i}) \\ &+ q \left(\left(\frac{c_2^2}{c_1 u} + 32 \frac{qa^2}{u^3} (f-f_1)^2 \right) \left(\frac{1}{2q - \left(\frac{u}{r_0 - r^*} + 2q \right) e^{u(t_j^i - t_{j-1}^i)}} - \frac{1}{r_0 - r^*} \right) - 16 \frac{ac_2(f-f_1)}{c_1 u^2} \left(\frac{q}{2q - \left(\frac{u}{r_0 - r^*} + 2q \right) e^{u(t_j^i - t_{j-1}^i)}} \right) \right) \end{aligned}$$

in which $r^* = \frac{a + \sqrt{a^2 + 2qb^2}}{2q}$ and r_0 is the initial variance at t_{j-1}^i that should be 0. As only elementary functions are involved in the log-likelihood, the analytical expressions of the gradient and the Hessian matrix can also be easily obtained using mathematic software (e.g., the R package 'Deriv' (<https://cran.r-project.org/web/packages/Deriv/index.html>)) albeit

tedious. The detailed derivation of the analytical likelihood can be found in the supplemental material Text S1B. Given the analytical expressions, MLE can be conducted using gradient-based numerical optimization algorithms (e.g., the Newton-Raphson method). In the current implementation, we choose the L-BFGS algorithm (Byrd et al., 1995) with box constraints for MLE as it does not require computing the Hessian matrix explicitly within each iteration, and therefore has a computational advantage in the context of analysis of large-scale genotype datasets.

Incorporating genetic models

In order for the GSPM to inspect the effects of genetic variations on the parameters, we connect the parameters of interest to genotypes through linear functions. Suppose that we observe an additively coded genotype G_i (i.e., $G_i \in \{0, 1, 2\}$; the genotype can be imputed.) at a specific locus for an individual i . We extend the SPM model by adopting the following linear relationship between a parameter (without loss of generality, we take the parameter a as an example.) and the genotype

$$a^i = a_{G_0} + \beta_a \cdot G_i,$$

where a_{G_0} is the rate of the adaptive regulation for those individuals with $G_i=0$ and β_a is the effect size (slope). This link function can be used for multiple parameters if more than one parameter is of interest. Further extension to the inclusion of other age-independent covariates such as sex in this function is straightforward. Time-dependent parameters are only allowed in the general model (Arbeev et al., 2009), and the estimation is implemented in the stpm R package (Zhbannikov et al., 2017). Note that $a^i < 0$ leads to

$a_{G_0} + \beta_a \cdot G_i < 0, \forall G_i$, so the MLE based on this parameterization would involve constraints of a linear combination of the two parameters. To apply the L-BFGS algorithm with box

constraints, we re-parameterize $\beta_a = \frac{a_{G_2} - a_{G_0}}{2}$ to have simpler constraints on the parameters. The re-parameterization leads to

$$a^i = a_{G_0} + \frac{a_{G_2} - a_{G_0}}{2} \cdot G_i = a_{G_2} \frac{G_i}{2} + a_{G_0} \left(1 - \frac{G_i}{2}\right) < 0, \forall G_i \Rightarrow a_{G_0} < 0, a_{G_2} < 0,$$

in which a_{G_2} is interpreted as the rate of the adaptive regulation for those individuals with $G_i=2$. In the case of rare variants where a dominant model is almost equivalent to an additive model, we can choose the dominant model ($G_i=2$ if there is at least one copy of a minor allele; otherwise $G_i=0$).

If there is no genetic effect on a , we would have $\beta_a = \frac{a_{G_2} - a_{G_0}}{2} = 0$. Thus, the inference of the existence of any genetic effect on a can be drawn by testing the null hypothesis $\beta_a = 0$. Denote by H the Hessian matrix of the minus log-likelihood which can be obtained either from the numerical approximation methods (e.g., L-BFGS) or analytically, and thus $(H^{-1})_a$

is the covariance matrix for the estimates $(\hat{a}_{G_2}, \hat{a}_{G_0})$. Under the null hypothesis, it follows for a large sample size

$$\hat{\beta}_a = \frac{\hat{a}_{G_2} - \hat{a}_{G_0}}{2} \sim \mathcal{N}(0, v(H^{-1})_a v'), \quad (5)$$

where $v = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \end{pmatrix}$ and $(\cdot)'$ represents the transpose, from which a p-value can be calculated. Minor extension of this strategy can also be used to test a null hypothesis involving multiple parameters such as $\beta_a = \beta_q = 0$. In this case, under the null hypothesis, we have

$$\hat{\beta} = (\hat{\beta}_a, \hat{\beta}_q)' = \left(\frac{\hat{a}_{G_2} - \hat{a}_{G_0}}{2}, \frac{\hat{q}_{G_2} - \hat{q}_{G_0}}{2} \right)' = W(\hat{a}_{G_0} \hat{a}_{G_2} \hat{q}_{G_0} \hat{q}_{G_2})' \sim \mathcal{N}(0, W H_{a,q}^{-1} W'),$$

where $W = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix}$. Thus, testing $\beta = 0$ can be based on the statistic

$$\hat{\beta}' (W H_{a,q}^{-1} W')^{-1} \hat{\beta} \sim \chi_2^2.$$

This kind of tests is useful for detecting pleiotropic effects on the biomarker and the disease.

Implementation

The proposed GSPM along with its estimation are implemented as a function (the major algorithms are implemented in C++) integrated in the *stpm* R package (Zhbannikov et al., 2017), which is available on CRAN (<https://cran.r-project.org/web/packages/stpm/index.html>). In the current implementation, five parameters (a, b, q, f, μ_0) can be linked to genetic information separately or collectively. The L-BFGS algorithm with bound constraints provided in the *nloptr* R package (Ypma, 2014) is employed for the optimization of the log-likelihood. We also provide a global optimization by using the “multi-level single-linkage” (MLSL) algorithm (Kan and Timmer, 1987) for the situations where computational burden is not a major concern. Both an analytical Hessian and a Hessian approximated by the L-BFGS algorithm at the stationary point are reported if the algorithm converges.

Results

Evaluation of the type I error rate with simulation data

We inspected the type I error rate of the tests for the genetic effect on the adaptive capacity a , the diffusion coefficient b , the vulnerability index q , the optimal state f and the baseline hazard μ_0 by using simulation datasets. We investigated the five genetic models separately:

- (1) $a^i = a_{G_0} + \beta_a \cdot G_i$, (2) $b^i = b_{G_0} + \beta_b \cdot G_i$, (3) $q^i = q_{G_0} + \beta_q \cdot G_i$, (4) $f^i = f_{G_0} + \beta_f \cdot G_i$ and (5) $\mu_0^i = \mu_{0G_0} + \beta_{\mu_0} \cdot G_i$. We randomly generated 1,000 datasets under the null hypothesis with

the parameters $a_{G_0} = -0.03$, $b_{G_0} = 0.1$, $q_{G_0} = 0.005$, $f_{G_0} = 2$, $f_1 = 3$, $\mu_{0G_0} = 0.001$, $\theta = 10^{-5}$ and $\beta_a = \beta_b = \beta_q = \beta_f = \beta_{\mu_0} = 0$, each of which contained 1,000 individuals. The genotype G_i for each individual was generated by sampling under an assumption of the Hardy-Weinberg equilibrium (HWE) from some allele frequency that was randomly generated from a uniform distribution $Unif(5\%, 95\%)$. P-values were calculated based on (5) with the Hessian matrix approximated by the L-BFGS algorithm. The Q-Q plots (Figure 1a–1d) of the p-values showed that the distribution of p-values under the null hypothesis for a , b , q and f conformed to a uniform distribution, indicating no inflation of the overall type I error rate. The Q-Q plot for μ_0 (Figure 1e) suggested that the type I error rate was lower than expected, indicating that the test for μ_0 is underpowered. We suspected that an overly small μ_0 compared to the quadratic function of q might reduce the power, so we performed another simulation study with $\mu_{0G_0} = 0.01$. The Q-Q plot (Figure 1f) showed that the deviation from the expected p-values disappeared after μ_0 was increased.

An application to BMI and T2D with an ARIC dataset

In this analysis, we used the GSPM to gain more insights into the relationship between SNPs, BMI and T2D. We performed genome-wide analyses by applying the GSPM to the Atherosclerosis Risk in Communities study (ARIC) (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000090.v1.p1) which is one of the large-scale NIH-funded longitudinal studies and has been widely used for research on the relationship between T2D and related risk factors (Raynor et al., 2013). Details of the study design, collection of samples, measurement of phenotypes and genotyping have been described in previous publications (Sharrett, 1992; The ARIC investigators, 1989). We included all available observations of BMI (at most 4 observations for each individual) throughout the follow-up period. We excluded any individual who was diagnosed with T2D prior to the entry of the study. Totally 22,101 observations from 8,284 independent Caucasians were included in the analyses. Genotyping of the ARIC participants was conducted using an Affymetrix 6.0 array (1,000K SNPs). All genotyped SNPs were further filtered for the analyses according to the exclusion criteria of the HWE test ($p < 1e-05$) and allele frequency ($MAF < 5\%$).

We first estimated all parameters related to BMI and T2D (including the adaptive capacity to approach allostasis a , the diffusion coefficient of BMI b , the asymptotic BMI level subject to allostasis f_1 , the optimal state of BMI against T2D f , the baseline hazard $\mu_0(t) = \mu_0 e^{\theta t}$ and the multiplier in the quadratic part of hazard q) by fitting the GSPM without genetic components. The assumption of a constant diffusion coefficient b for BMI could be reasonable at the population level as a previous twin study has shown that the environmental variance component of BMI increases almost linearly over age and the genetic variance component levels off after middle age (He et al., 2016b). Due to the limited precision of the numerical procedure, we rescaled the BMI levels by a factor of 100 (i.e., BMI/100).

We found that the estimated parameters characterizing the trajectory of BMI were $a = 0.003$, $b = 0.001$ and $f_1 = 0.73$, suggesting that the level of BMI kept almost constant and there was little adaptive capacity during the age interval between 40 and 80. These estimates were in

line with the Spaghetti plot of BMI for the ARIC cohort (Figure 2), showing that BMI leveled off and the mean value remained flat throughout the follow-up. The estimated optimal level (i.e., the parameter f) of BMI for the lowest risk of T2D is ~ 17 . As only 32 observations of BMI were lower than 17 among the 8,284 individuals under investigation, this estimate of f suggested a J-shape hazard function of T2D with respect to BMI. The estimated θ was 4.5×10^{-3} , suggesting that the baseline hazard increased very mildly.

It would be desirable to detect any SNP with pleiotropic effects on adaptive capacity for BMI and the hazard of T2D (i.e., a SNP affecting both a and μ_0). But as little adaptive capacity for BMI was observed from the cohort, we turned to test SNP-BMI interaction effects on T2D and genetic effects on the optimal BMI level against T2D. In all of the following analyses, we assumed a nearly constant baseline hazard by fixing $\theta=10^{-5}$ as the estimated θ was very small. We compared the performance of the proposed GSPM in these real data analyses with general survival models including the Cox proportional hazards (PH) model (Cox, 1972) and the Aalen's additive hazards model (Aalen, 1978). We employed the *coxph* function in the *survival* R package and the *aalen* function in the *timereg* R package (Scheike and Zhang, 2011) for the Cox model and the Aalen's additive model, respectively. In both models, we treated BMI as a time-varying variable based on the longitudinal observations and the effects of both BMI and genotypes were assumed to be constants. To estimate the interaction effects, we added an interaction term (genotype*BMI) in the models. Haploreg v4.1 (Ward and Kellis, 2012), the *haploR* R package (Zhbannikov et al., 2016) and GRASP (Leslie et al., 2014) were employed for functional annotation and search for previous relevant reports for identified SNPs.

Identifying SNP-BMI interaction effects on T2D

We performed a genome-wide association analysis using the GSPM to identify any SNPs (total 607,861 SNPs) that modulate the J-shape effect of BMI on T2D (The hazard functions are illustrated in Figure 3A). As we were interested in the hazard of T2D, we only estimated q and μ_0 and substituted the other parameters with their estimated values, which increased the statistical power and saved computational time. It should be noted that μ_0 also needs to be modeled as a function of genotypes, although q is the parameter characterizing the gene-interaction effects. Otherwise any direct effect of a SNP on μ_0

would be reflected in the estimate of q . More specifically, we assumed that q and μ_0 are

linear functions of a genotype (i.e., $q^i = q_{G_2} \frac{G_i}{2} + q_{G_0} \left(1 - \frac{G_i}{2}\right)$ and

$\mu_0^i = \mu_{0G_2} \frac{G_i}{2} + \mu_{0G_0} \left(1 - \frac{G_i}{2}\right)$) and we were interested in testing the null hypothesis $q_{G_0} = q_{G_2}$. In this analyses, we observed that the estimate of μ_0 would be zero for some SNPs, which led to problematic interpretation. We found that this often indicated the misspecification of the coding model for genotypes. We shifted to a dominant model for these SNPs and this solved the issue. The Q-Q plot (Figure S1) confirmed no sign of inflated type I error rate (the genomic inflation factor is 1.02) after such manipulation.

Table 2 shows a list of the SNPs with p-values exceeding the suggestive significance level ($<1 \times 10^{-5}$) identified from the GWAS. We found that no SNPs passed the genome-wide

significance ($<5 \times 10^{-8}$) (Figure 4) and five SNPs (rs9874828, rs9877871, rs11757677, rs4977708, rs7185166) had a p-value below 1×10^{-6} . The p-values of the interaction effects from the Aalen's additive model for three of these five SNPs were comparable to those from the GSPM while the p-values from the Cox model for the other two SNPs were much larger (Table 2). The top two SNPs rs9874828 and rs9877871 located on 3q22.2 are in high LD with each other as shown in the regional plot (Figure S3). Rs9874828 is an intron variant of *EPHB1* and rs9877871 is a downstream variant associated with *EPHB1*. Haploreg indicated that rs9874828 was an eQTL of *STK24* in peripheral blood monocytes and GRASP showed that it was nominally associated with multiple phenotypes including high-density lipoprotein (HDL) cholesterol change with statins ($p=2.3e-03$) (Barber et al., 2010) and birth weight ($p=3.2e-02$) (Horikoshi et al., 2013). Rs11757677 is an intron variant of *CDKAL1* located on 6p22.3 (Figure S4), and is an eQTL of *MLLT3* and *SNX16* in peripheral blood monocytes. GRASP results for this SNP showed that it was nominally associated with multiple biomarkers including waist hip ratio ($p=7.8e-03$) (Heid et al., 2010), diastolic blood pressure ($p=1.4e-02$) (International Consortium for Blood Pressure Genome-Wide Association Studies et al., 2011), LDL and HDL cholesterol levels ($p=2.7e-02$ and $4.6e-02$) (Barber et al., 2010) and serum creatinine ($p=3.7e-02$) (Köttgen et al., 2010). Rs4977708 is an intergenic variant located on 9p21.2 (Figure S5), and is an eQTL of *KIF11* in peripheral blood monocytes (Zeller et al., 2010). This SNP has been found nominally associated with obesity with early age of onset ($p=1.8e-03$) (Bradfield et al., 2012), adiponectin levels ($p=2.3e-03$) (Dastani et al., 2012) and hypertension ($p=9.5e-03$) (Yang et al., 2012). Rs7185166 is an intron variant of *ZFHX3* located on 16q22.3 (Figure S6), and it has been found to be associated with T2D ($p=8.1e-04$) (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research et al., 2007).

For most of the identified SNPs (Table 2), the minor allele increased its main effect on the risk of T2D adjusted for BMI ($\beta_{\mu_0} < 0$) while reduced the modulating effect of BMI on the risk of T2D ($\beta_{q_0} < 0$). It seems that the SNP-BMI interaction effect is negatively correlated with the main effect for these SNPs. We further searched for the top SNP with only significant interaction effects but no main effects (i.e., $P_q < 0.05$, $P_{\mu_0} \geq 0.05$) and identified rs17650892 with $P_q = 1.16e-05$, located on 4q32.3 in the gene *PALLD*.

Identifying SNPs associated with optimal BMI level for avoiding T2D

In this analysis, we examined which SNPs (total 607,861 SNPs) affected the optimal BMI level for attaining the lowest risk of T2D (the hazard functions are illustrated in Figure 3B), which can be considered as the genetic predisposition to the frailty threshold of BMI. Similar to the previous analysis, we plugged in the estimates of a , b and f_1 for characterizing the dynamics of BMI. We assumed that f was a function of a genotype

($f^i = f_{G_2} \frac{G_i}{2} + f_{G_0} \cdot \left(1 - \frac{G_i}{2}\right)$), and q and μ_0 were free parameters shared by the groups with different genotypes. Note that under this assumption, any genetic effect on μ_0 would also be reflected in the effect on f . Therefore, we further compared different models using Akaike information criterion (AIC) (Akaike, 1998) to choose the optimal model once significant SNPs were identified.

As shown in Figure 5, no genome-wide significant SNPs were identified. Three SNPs (rs1551133, rs2613310 and rs965892) yielded p-values ($p=6.70e-07$, $1.41e-06$ and $p=1.47e-06$) exceeding the suggestive significance level ($<1 \times 10^{-5}$) (Table 3). Interestingly, the estimated optimal levels of BMI for the homozygous genotypes of rs1551133 were 34.5 and 19.6, suggesting that the minor allele of this SNP reversed the effect of BMI on the risk of T2D. We found that the AA genotype of this SNP increased the incidence of T2D but lowered the BMI levels at onset (Table 4). Then, we fitted a model in which both f and μ_0 were treated as functions of the genotype, and the p-value for f became genome-wide significant ($P_f=2.90e-09$). Rs1551133 is a regulatory region variant located on 2q14.2, and Ensembl database showed that it was correlated with expression of many genes. GRASP showed that rs1551133 was associated with multiple phenotypes including schizophrenia ($p=1.7e-05$) and chronic obstructive pulmonary disease ($p=2.4e-05$). Rs2613310 and rs965892, located in an intron of the *SYK* gene on 9q22.2, were in complete LD with each other. Haploreg results indicated that these two SNPs were eQTLs of *SYK* and they were significantly associated with a metabolite C5-DC(C6-OH) / C16:2. The Cox model and the additive hazards model yielded p-values of $5.1e-06$ and $2.55e-05$ for rs2613310, respectively, and similar results were observed for rs965892. To further investigate whether these SNPs affected the optimal value of BMI against T2D or they had main effects on the risk of T2D independent of BMI, we fitted another model in which μ_0 instead of f was modelled as a function of the genotype, and compared the AICs. The result showed that the model with genetic effect on f was better.

Discussion

The SPM is a powerful and sophisticated tool that can simultaneously model the dynamics of biomarkers and their joint influence on mortality or diseases. By incorporating the parameters that characterize complex physiological processes and thus, have better biological interpretation, this mathematical model is beyond a simple statistical model in terms of biological interpretation and has the potential to gain more insights into the underlying biological mechanisms. In this article, we propose a fast stochastic process model that can be utilized specifically in large-scale analyses for investigating the relationship between a genetic variation, dynamics of a biomarker over a portion of individuals' life course, and the onset of a disease. We solve an analytical solution to the differential equations for the one-dimensional SPM under an assumption that the model coefficients are time-invariant. Our likelihood-based strategy dramatically shortens the computational burden of the estimation procedure, so that the proposed GSPM together with the proposed estimation and inference methods makes it possible for its broad application in the context of large-scale genetic analysis. In the analysis of the ARIC dataset which contained more than 8,000 individuals and 20,000 observations, it took ~3–5s to analyze each SNP on an Intel-i5 PC while, in contrast, the general estimation method employing the Runge-Kutta methods (Ascher and Petzold, 1998) and the heuristic simplex methods (Nelder and Mead, 1965) would require ~10 minutes for the same dataset. In addition, it substantially reduces the cost of drawing inference by avoiding resampling methods as the covariance matrix can be easily obtained. All of these improvements pave the way for its application in GWAS to investigate interaction effects, pleiotropic effects and so forth. To

pay the price for achieving this, we have to assume simple functions for most parameters. But these assumptions are reasonable in many cases as we have shown in the analysis of T2D. Such assumption might not be realistic in other situations (e.g., f and q tend to be age-dependent if the baseline hazard increases dramatically with respect to age). In some cases the piecewise constancy of model parameters may be assumed. This strategy might be relevant when one has to detect persistent effects of genetic variants on mortality (or disease) risks at relatively long age intervals in which the values of selected biomarkers and baseline hazard experience substantial changes. A more rigorous strategy is to check the age-independent assumption with the general model (Arbeev et al., 2009) allowing all parameters to depend on age (implemented in the *stpm* R package (Zhbannikov et al., 2017)) before running the GSPM for GWAS. Then, the GWAS follows given some evidence of the assumption being reasonable, and more refined analyses can be performed for GWAS top hits with the general model (Arbeev et al., 2009) as well.

In the ARIC data we analyzed, the influence of accounting for the dynamics of BMI on the results should be minor as the trajectory of BMI is almost flat, suggesting that the assumptions and models related to the hazards are responsible for the difference and can dramatically affect the signals to be identified. It seems that the proposed GSPM can produce very different results from the Cox PH model. The results from the identified SNPs suggested that the estimates from the GSPM are more comparable to those from the Aalen's additive model, which is quite reasonable as they both assume an additive form of the hazards while the Cox model assumes proportional hazards. The major difference from the Aalen's additive hazard model is that in the GSPM we assume that the impact of the biomarker on the hazard follows a quadratic function that can be U-shape or J-shape, while in the Aalen's model the hazard functions are linear. Therefore, the GSPM is more suitable for the situation where the biomarker conforms to an approximate U- or J- shape effect on the hazard of the diseases or mortality. In some situations, the quadratic form is a more realistic and flexible assumption and model comparison can also be used to select a better model. In the case of T2D, a previous study has shown that a linear relationship of BMI led to inflated type I error rate and the inclusion of quadratic and cubic terms came to a rescue (Cornelis et al., 2012).

The analysis of SNP-BMI interaction on T2D identifies 23 SNPs with suggestive significance. Three out of the four independent SNPs (rs9874828 and rs9877871 are in high LD) that had a p-value lower than $1e-06$ are eQTLs, indicating their potential involvement in the gene expression and regulatory activity. The results from GRASP show that rs9874828 (in *EPHBI*), rs11757677 (in *CDKALI*) and rs4977708 are associated with a wide spectrum of phenotypes. *CDKALI* is well known to be implicated in T2D and numerous associated genetic variants have been reported (Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research et al., 2007; Scott et al., 2007; Steinthorsdottir et al., 2007; Zeggini et al., 2007). Here we found that the variant rs11757677 in *CDKALI* exhibits significant interaction effect with BMI. Rs4977708 is associated with gene expression of *KIF11* in peripheral blood monocytes ($p=3.88e-06$) (Zeller et al., 2010). Multiple SNPs in the *IDE-KIF11-HHEX* region have been linked to *T2D* (Furukawa et al., 2008; Pascoe et al., 2007; Qian et al., 2012), but it is still unclear which gene is the culprit. It has been shown that the expression of these genes are

impacted during adipogenesis (Zhao et al., 2012). The finding of rs4977708 may suggest that *KIF11* is the gene implicated in the development of T2D. A SNP rs823968 in *EPHB1* in modest LD ($r^2=0.2$) with rs9874828 is associated with T2D in a Finnish population (Scott et al., 2007). In addition, rs7185166 is located in *ZFHX3*, variants in which are reportedly associated with a diverse range of diseases including coronary heart disease (Sun et al., 2015), atrial fibrillation (Gudbjartsson et al., 2009; Liu et al., 2014) and stroke (Liu et al., 2014). A recent study has found that a mutation in this gene disrupts circadian locomotor rhythms in mice (Parsons et al., 2015), which may explain its pleiotropic effects on these seemingly unrelated diseases. The interaction effects of rs7185166 with BMI on T2D may be related to its pleiotropic effects, suggesting that SNPs may be tied to some fundamental biological functions that can affect multiple biomarkers and work collectively with BMI to boost the risk of T2D.

Previous studies suggest that some populations such as Indians may have considerably lower frailty threshold of BMI for T2D (Chan et al., 2009; He et al., 2015), which means that the optimal level of BMI for avoiding T2D is much lower for these populations. Limited storage capacity of peripheral adipose tissue which has been shown to be involved in insulin resistance (Lotta et al., 2017) can be one of the causes. No genome-wide significant SNPs that affect the threshold value of BMI for T2D risk were identified from our analysis. It should be pointed out that any identified genetic effect on f may result from its direct effect on the baseline hazard of T2D that is independent of BMI (i.e., the genetic effect on μ_0) if we assume the homogeneous baseline hazard between the genotype groups. Once a significant SNP is detected, model comparison can be employed to select a better model for the data and further distinguish these two different cases. In this analysis, we found two intron SNPs (rs2613310 and rs965892) in the *SYK* gene that showed suggestive significance. Although the Cox and Aalen models showed consistent results for these SNPs, our model yielded smaller p-values, which may suggest that the GSPM is a more realistic model for these SNPs. Our results further indicate that these SNPs are more likely associated with the threshold value of BMI for the risk of T2D. The two SNPs have been found to be associated with bone mineral density (rs2613310, $p=2.6e-02$) (Styrkarsdottir et al., 2008), coronary artery disease (rs2613310, $p=2.6E-04$) and type 1 diabetes (T1D) (rs965892, $p=9.7e-04$) (Wellcome Trust Case Control Consortium, 2007). Blockade of *SYK* is reported a promising therapeutic target to prevent T1D (Colonna et al., 2008; Geahlen, 2014).

We have demonstrated the performance of the proposed GSPM in the application to a genome-wide investigation of the relationship between genetic variations, BMI and T2D. However, there are some limitations. As previously mentioned, it is difficult to solve the Riccati equation analytically in general case. The choice of the functional form for the parameters in the GSPM is restricted. This restriction may impede its application, particularly in some situations where the baseline hazard function substantially deviates from a constant. Further research needs to be carried out in order to relax the assumption for at least some of the parameters or allow piecewise constancy. In addition, the assumption of the symmetric quadratic form of the hazard function with respect to the biomarker might not be realistic for some diseases. For example, people with overly low BMI are not subject to increased risk of T2D. Further generalization to non-symmetric functional forms would remarkably broaden the flexibility and increase the scope of the model for more biomarkers

and diseases (Yashin et al., 2009). Unlike a linear regression model or the Cox model, the GSPM is built on a model involving more specific assumptions, which requires extra attention. These assumptions need to be checked when the GSPM is applied in GWAS, particularly for identified SNPs because any violation of the assumptions may lead to incorrect estimates and misleading interpretation. This has been illustrated in the real data analysis, in which adoption of a wrong genetic coding model results in unrealistic estimates of the parameter.

In conclusion, we proposed a novel model for joint analyses of longitudinal and time-to-event data that can be used to investigate complex relationship between genetic variations, biomarkers and diseases in GWAS. The proposed GSPM can simultaneously estimate genetic effects on the trajectory of a biomarker and the hazard function from the longitudinal data, and takes into account the stochastic process of the biomarker between observations and outcomes. Compared to other statistical models, this mathematical model based on stochastic process is featured with more detailed characterization and modeling of physiological processes, and thus may provide better biological interpretation. This model is particularly suitable for investigating pleiotropic effects on a biomarker, a disease, and gene-environment interaction effects. Our genome-wide analyses for BMI and T2D have demonstrated that the type I error rate is well controlled and it can detect novel associated variants with interaction effects and offer superior statistical power in some situations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

The research reported in this paper was supported by Grants No P01 AG043352, R01AG047310 and P30AG034424 from the National Institute on Aging (NIA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health (NIH).

The Atherosclerosis Risk in Communities Study (ARIC) is carried out as a collaborative study supported by the NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions. Funding for GENEVA was provided by the National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle). The authors wish to thank Ms. Shouse who assisted in the proof-reading of the manuscript.

References

- Aalen O. Nonparametric Inference for a Family of Counting Processes. *Ann Stat.* 1978; 6:701–726.
- Akaike, H. Selected Papers of Hirotugu Akaike, (Springer). 1998. Information theory and an extension of the maximum likelihood principle; p. 199-213.
- Arbeev KG, Akushevich I, Kulminski AM, Arbeeva LS, Akushevich L, Ukraintseva SV, Culminskaya IV, Yashin AI. GENETIC MODEL FOR LONGITUDINAL STUDIES OF AGING, HEALTH, AND LONGEVITY AND ITS POTENTIAL APPLICATION TO INCOMPLETE DATA. *J Theor Biol.* 2009; 258:103–111. [PubMed: 19490866]

- Arbeev KG, Akushevich I, Kulminski AM, Ukraintseva SV, Yashin AI. Joint Analyses of Longitudinal and Time-to-Event Data in Research on Aging: Implications for Predicting Health and Survival. *Front Public Health*. 2014; 2
- Ascher UM, Petzold LR. Computer methods for ordinary differential equations and differential-algebraic equations (Siam). 1998
- Barber MJ, Mangravite LM, Hyde CL, Chasman DI, Smith JD, McCarty CA, Li X, Wilke RA, Rieder MJ, Williams PT, et al. Genome-wide association of lipid-lowering response to statins in combined study populations. *PloS One*. 2010; 5:e9763. [PubMed: 20339536]
- Belsky DW, Suppli NP, Israel S. Gene-environment interaction research in psychiatric epidemiology: a framework and implications for study design. *Soc Psychiatry Psychiatr Epidemiol*. 2014; 49:1525–1529. [PubMed: 25216778]
- Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, Warrington NM, Hypponen E, Holst C, Valcarcel B, Thiering E, et al. A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat Genet*. 2012; 44:526–531. [PubMed: 22484627]
- Byrd R, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*. 1995; 16:1190–1208.
- Chan JCN, Malik V, Jia W, Kadowaki T, Yajnik CS, Yoon KH, Hu FB. Diabetes in Asia: epidemiology, risk factors, and pathophysiology. *JAMA*. 2009; 301:2129–2140. [PubMed: 19470990]
- Colonna L, Ogryzlo-Harbers SN, Catalano G, Flores M, Chew C, Masuda E, Schmitz J, Thomas J, Clynes R. Targeting Syk in autoimmune diabetes. *FASEB Journal*. 2008; 22:1074.27.
- Cornelis MC, Hu FB. Gene-environment interactions in the development of type 2 diabetes: recent progress and continuing challenges. *Annu Rev Nutr*. 2012; 32:245–259. [PubMed: 22540253]
- Cornelis MC, Tchetgen EJT, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol*. 2012; 175:191–202. [PubMed: 22199026]
- Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B Methodol*. 1972; 34:187–220.
- Dastani Z, Hivert MF, Timpson N, Perry JRB, Yuan X, Scott RA, Henneman P, Heid IM, Kizer JR, Lyytikäinen L-P, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet*. 2012; 8:e1002607. [PubMed: 22479202]
- Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research. Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PIW, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007; 316:1331–1336. [PubMed: 17463246]
- Franks PW. Gene $\times$ environment interactions in type 2 diabetes. *Curr Diab Rep*. 2011; 11:552–561. [PubMed: 21887612]
- Furukawa Y, Shimada T, Furuta H, Matsuno S, Kusuyama A, Doi A, Nishi M, Sasaki H, Sanke T, Nanjo K. Polymorphisms in the IDE-KIF11-HHEX gene locus are reproducibly associated with type 2 diabetes in a Japanese population. *J Clin Endocrinol Metab*. 2008; 93:310–314. [PubMed: 17971426]
- Geahlen RL. Getting Syk: spleen tyrosine kinase as a therapeutic target. *Trends Pharmacol Sci*. 2014; 35:414–422. [PubMed: 24975478]
- Gudbjartsson DF, Holm H, Gretarsdottir S, Thorleifsson G, Walters GB, Thorgeirsson G, Gulcher J, Mathiesen EB, Njølstad I, Nyrnes A, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat Genet*. 2009; 41:876–878. [PubMed: 19597491]
- Guo X, Carlin BP. Separate and joint modeling of longitudinal and event time data using standard computer packages. *Am Stat*. 2004; 58:16–24.
- Hardy DS, Racette SB, Hoelscher DM. Macronutrient Intake as a Mediator with FTO to Increase Body Mass Index. *J Am Coll Nutr*. 2014; 33:256–266. [PubMed: 25144299]
- He L, Tuomilehto J, Qiao Q, Söderberg S, Daimon M, Chambers J, Pitkaniemi J, DECODA study group. Impact of classical risk factors of type 2 diabetes among Asian Indian, Chinese and Japanese populations. *Diabetes Metab*. 2015; 41:401–409. [PubMed: 26381573]

- He L, Kernogitski Y, Kulminskaya I, Loika Y, Arbeev KG, Loiko E, Bagley O, Duan M, Yashkin A, Ukraintseva SV, et al. Pleiotropic Meta-Analyses of Longitudinal Studies Discover Novel Genetic Variants Associated with Age-Related Diseases. *Front Genet.* 2016a; 7
- He L, Sillanpää MJ, Silventoinen K, Kaprio J, Pitkäniemi J. Estimating modifying effect of age on genetic and environmental variance components in twin models. *Genetics.* 2016b; 202:1313–1328. [PubMed: 26868768]
- Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G, Zillikens MC, Speliotes EK, Mägi R, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet.* 2010; 42:949–960. [PubMed: 20935629]
- Holmes MV, Lange LA, Palmer T, Lanktree MB, North KE, Almoguera B, Buxbaum S, Chandrupatla HR, Elbers CC, Guo Y, et al. Causal Effects of Body Mass Index on Cardiometabolic Traits and Events: A Mendelian Randomization Analysis. *Am J Hum Genet.* 2014; 94:198–208. [PubMed: 24462370]
- Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Taal HR, Hennig BJ, Bradfield JP, St Pourcain B, Evans DM, Charoen P, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat Genet.* 2013; 45:76–82. [PubMed: 23202124]
- Hunter DJ. Gene–environment interactions in human diseases. *Nat Rev Genet.* 2005; 6:287–298. [PubMed: 15803198]
- International Consortium for Blood Pressure Genome-Wide Association Studies. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Smith AV, Tobin MD, Verwoert GC, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature.* 2011; 478:103–109. [PubMed: 21909115]
- Kan AHGR, Timmer GT. Stochastic global optimization methods part II: Multi level methods. *Math Program.* 1987; 39:57–78.
- Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, Glazer NL, Parsa A, Gao X, Yang Q, Smith AV, et al. New loci associated with kidney function and chronic kidney disease. *Nat Genet.* 2010; 42:376–384. [PubMed: 20383146]
- Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinforma Oxf Engl.* 2014; 30:i185–194.
- Liu Y, Ni B, Lin Y, Chen X, Fang Z, Zhao L, Hu Z, Zhang F. Genetic Polymorphisms in ZFHX3 Are Associated with Atrial Fibrillation in a Chinese Han Population. *PLOS ONE.* 2014; 9:e101318. [PubMed: 24983873]
- Lotta LA, Gulati P, Day FR, Payne F, Ongen H, van de Bunt M, Gaulton KJ, Eicher JD, Sharp SJ, Luan J'an, et al. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat Genet.* 2017; 49:17–26. [PubMed: 27841877]
- McEwen BS, Wingfield JC. The concept of allostasis in biology and biomedicine. *Horm Behav.* 2003; 43:2–15. [PubMed: 12614627]
- Nelder JA, Mead R. A simplex method for function minimization. *Comput J.* 1965; 7:308–313.
- Parsons MJ, Brancaccio M, Sethi S, Maywood ES, Satija R, Edwards JK, Jagannath A, Couch Y, Finelli MJ, Smyllie NJ, et al. The Regulatory Factor ZFHX3 Modifies Circadian Function in SCN via an AT Motif-Driven Axis. *Cell.* 2015; 162:607–621. [PubMed: 26232227]
- Pascoe L, Tura A, Patel SK, Ibrahim IM, Ferrannini E, Zeggini E, Weedon MN, Mari A, Hattersley AT, McCarthy MI, et al. Common Variants of the Novel Type 2 Diabetes Genes CDKAL1 and HHEX/IDE Are Associated With Decreased Pancreatic β -Cell Function. *Diabetes.* 2007; 56:3101–3104. [PubMed: 17804762]
- Price AL, Spencer CCA, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. *Proc R Soc B.* 2015; 282:20151684.
- Qian Y, Lu F, Dong M, Lin Y, Li H, Chen J, Shen C, Jin G, Hu Z, Shen H. Genetic Variants of IDE-KIF11-HHEX at 10q23.33 Associated with Type 2 Diabetes Risk: A Fine-Mapping Study in Chinese Population. *PLOS ONE.* 2012; 7:e35060. [PubMed: 22506066]

- Raynor LA, Pankow JS, Duncan BB, Schmidt MI, Hoogveen RC, Pereira MA, Young JH, Ballantyne CM. Novel Risk Factors and the Prediction of Type 2 Diabetes in the Atherosclerosis Risk in Communities (ARIC) Study. *Diabetes Care*. 2013; 36:70–76. [PubMed: 22933437]
- Scheike TH, Zhang MJ. Analyzing competing risk data using the R *timereg* package. *J Stat Softw*. 2011; 38
- Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, et al. A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science*. 2007; 316:1341–1345. [PubMed: 17463248]
- Sharrett AR. The Atherosclerosis Risk in Communities (ARIC) Study. Introduction and objectives of the hemostasis component. *Ann Epidemiol*. 1992; 2:467–469. [PubMed: 1342297]
- Shriner D. Moving toward System Genetics through Multiple Trait Analysis in Genome-Wide Association Studies. *Front Genet*. 2012; 3:1. [PubMed: 22303408]
- Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet*. 2013; 14:483–495. [PubMed: 23752797]
- Sritara C, Thakkestian A, Ongphiphadhanakul B, Chailurkit L, Chanprasertyothin S, Ratanachaiwong W, Vathesatogkit P, Sritara P. Causal relationship between the AHSG gene and BMD through fetuin-A and BMI: multiple mediation analysis. *Osteoporos. Int J Establ Result Coop Eur Found Osteoporos Natl Osteoporos Found USA*. 2014; 25:1555–1562.
- Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet*. 2007; 39:770–775. [PubMed: 17460697]
- Styrkarsdottir U, Halldorsson BV, Gretarsdottir S, Gudbjartsson DF, Walters GB, Ingvarsson T, Jonsdottir T, Saemundsdottir J, Center JR, Nguyen TV, et al. Multiple genetic loci for bone mineral density and fractures. *N Engl J Med*. 2008; 358:2355–2365. [PubMed: 18445777]
- Sun S, Zhang W, Chen X, Song H. The CAA repeat polymorphism in the ZFH3 gene is associated with risk of coronary heart disease in a Chinese population. *Tohoku J Exp Med*. 2015; 235:261–266. [PubMed: 25797214]
- The ARIC investigators. The atherosclerosis risk in communities (ARIC) study: design and objectives. *Am J Epidemiol*. 1989; 129:687–702. [PubMed: 2646917]
- Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sin*. 2004:809–834.
- Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012; 40:D930–D934. [PubMed: 22064851]
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447:661–678. [PubMed: 17554300]
- Winham SJ, Biernacka JM. Gene-environment interactions in genome-wide association studies: current approaches and new directions. *J Child Psychol Psychiatry*. 2013; 54:1120–1134. [PubMed: 23808649]
- Woodbury MA, Manton KG. A random-walk model of human mortality and aging. *Theor Popul Biol*. 1977; 11:37–48. [PubMed: 854860]
- Xu X, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics*. 2003; 4:223–229. [PubMed: 12925518]
- Yang HC, Liang YJ, Chen JW, Chiang KM, Chung CM, Ho HY, Ting CT, Lin TH, Sheu SH, Tsai WC, et al. Identification of IGF1, SLC4A4, WWOX, and SFMBT1 as hypertension susceptibility genes in Han Chinese with a genome-wide gene-based association study. *PloS One*. 2012; 7:e32907. [PubMed: 22479346]
- Yashin AI. Dynamics in survival analysis: conditional Gaussian property versus Cameron-Martin formula. 1984
- Yashin AI, Manton KG. Effects of Unobserved and Partially Observed Covariate Processes on System Failure: A Review of Models and Estimation Strategies. *Stat Sci*. 1997; 12:20–34.
- Yashin AI, Manton KG, Vaupel JW. Mortality and aging in a heterogeneous population: A stochastic process model with observed and unobserved variables. *Theor Popul Biol*. 1985; 27:154–175. [PubMed: 4023952]

- Yashin AI, Manton KG, Stallard E. Evaluating the effects of observed and unobserved diffusion processes in survival analysis of longitudinal data. *Math Model.* 1986; 7:1353–1363.
- Yashin AI, Arbeev KG, Akushevich I, Kulminski A, Akushevich L, Ukraintseva SV. Stochastic model for analysis of longitudinal data on aging and mortality. *Math Biosci.* 2007; 208:538–551. [PubMed: 17300818]
- Yashin AI, Arbeev KG, Akushevich I, Kulminski A, Akushevich L, Ukraintseva SV. Model of hidden heterogeneity in longitudinal data. *Theor Popul Biol.* 2008; 73:1–10. [PubMed: 17977568]
- Yashin AI, Ukraintseva SV, Arbeev KG, Akushevich I, Arbeeve LS, Kulminski AM. Maintaining physiological state for exceptional survival: what is the normal level of blood glucose and does it change with age? *Mech Ageing Dev.* 2009; 130:611–618. [PubMed: 19635493]
- Yashin AI, Arbeev KG, Wu D, Arbeeve LS, Kulminski A, Akushevich I, Culminskaya I, Stallard E, Ukraintseva SV. How lifespan associated genes modulate aging changes: lessons from analysis of longitudinal data. *Front Genet.* 2013; 4:3. [PubMed: 23346098]
- Ypma J. Introduction to nloptr: an R interface to NLOpt. 2014
- Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JRB, Rayner NW, Freathy RM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science.* 2007; 316:1336–1341. [PubMed: 17463249]
- Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, Maouche S, Germain M, Lackner K, Rossmann H, et al. Genetics and Beyond – The Transcriptome of Human Monocytes and Disease Susceptibility. *PLoS ONE.* 2010; 5
- Zhao J, Deliard S, Aziz AR, Grant SF. Expression analyses of the genes harbored by the type 2 diabetes and pediatric BMI associated locus on 10q23. *BMC Med Genet.* 2012; 13:89. [PubMed: 22998375]
- Zhbannikov I, Arbeev KG, Yashin AI. haploR: Query Haploreg and RegulomeDB. 2016
- Zhbannikov IY, Arbeev K, Akushevich I, Stallard E, Yashin AI. stpm: an R package for stochastic process model. *BMC Bioinformatics.* 2017; 18:125. [PubMed: 28231764]
- Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, Smith JA, Yanek LR, Sun YV, Edwards TL, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet.* 2015; 96:21–36. [PubMed: 25500260]

Figure 1. Q-Q plots of the p-values under the null hypotheses to test genetic effect on the parameters

a , b , q , f and μ_0 . A) the Q-Q plot for a , B) the Q-Q plot for b , C) the Q-Q plot for q , D) the Q-Q plot for f , E) the Q-Q plot for μ_0 with $\mu_{0G0}=0.001$, F) the Q-Q plot for μ_0 with $\mu_{0G0}=0.01$.

Figure 2. A spaghetti plot of the longitudinal observations of BMI from the ARIC cohort
 The figure illustrates all individual-specific observations (maximum 4 observations for each individual) of BMI between the age of 40 and 80 used in the analysis of the ARIC dataset. The dash curve with shaded areas is estimated from a lowess smoother with standard error bounds.

Figure 3. An illustration of the tests for genetic effects on the effect of BMI on T2D and the optimal BMI level against T2D

A) Comparison of the hazard functions in the two genotype groups with different μ_0 and q , but the same $f=0.1769$, which is estimated from the ARIC dataset. This is the model used in the GWAS for detecting SNP-BMI interaction effects. This is performed by testing whether q is different across genotype groups. B) Comparison of the hazard functions in the two genotype groups with different f , but the same μ_0 and q . This is the model used in the GWAS for detecting the BMI frailty threshold on T2D. This is performed by testing whether f is different across genotype groups with μ_0 and q shared by the genotype groups.

Figure 4. A Manhattan plot for the GWAS using the GSPM model to detect SNPs modulating the effects of BMI on T2D

The purple horizontal line indicates the genome-wide significance (5×10^{-8}). The p-values (607,861 SNPs included in this analysis) were calculated from the test for the heterogeneity of q .

Figure 5. A Manhattan plot for the GWAS using the GSPM model to detect SNPs affecting the frailty threshold of BMI on T2D

The purple horizontal line indicates the genome-wide significance (5×10^{-8}). The p-values (607,861 SNPs included in this analysis) were calculated from the test for the heterogeneity of f .

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

An example of the format of a longitudinal dataset used in the proposed model. There are i individuals, each of which has n_i observations of the biomarker at age $t_{n_i}^i$ and an event status δ_i at τ_i . The genotype G_i for individual i is an integer in $\{0,1,2\}$.

Table 1

Subject	Visit	Age	Biomarker	Disease status	Genotype
1	1	t_1^1	y_1^1	0	G_1
...	...	...	...	0	G_1
1	n_1	$t_{n_1}^1$	$y_{n_1}^1$	0	G_1
1		τ_1		δ^1	G_1
...	...	...	...	...	...
i	1	t_1^i	y_1^i	0	G_i
...	...	...	...	0	G_i
				0	G_i
i	n_i	$t_{n_i}^i$	$y_{n_i}^i$	0	G_i
i		τ_i		δ^i	G_i

Table 2

The results from the GWAS of gene-BMI interaction effects on the risk of T2D using the GSPM. The rows in boldface are the SNPs with a p-value P_q of suggestive significance ($p < 1e-06$).

CHR	Rs ID	Pos	A1	A2	AF (coding)	P (HWE)	β_q	P_q	β_{μ_0}	P (Aalen)	P (Cox)
2	rs11684487	167094349	T	C	0.696	0.057	0.260	2.39E-06	-59e-03	1.45e-03	
2	rs4667515	167107167	T	C	0.699	0.134	0.247	9.41E-06	-54e-03	2.41e-03	
2	rs13031728	167107813	G	A	0.697	0.577	0.252	5.11E-06	-1.53e-03	2.55e-03	
2	rs9917308	167129851	A	G	0.665	0.292	0.247	4.63E-06	-1.4e-03	3.93e-03	
2	rs12467601	167131779	C	T	0.659	0.046	0.247	4.98E-06	-	3.31e-03	
3	rs12491147	55544001	C	T	0.795	0.671	0.283	2.18E-06	-	2.21e-03	
3	rs9874828	136443519	A	T	0.660	0.946	0.264	2.89E-07	-	1.71e-02	2.60e-07
3	rs9877871	136465447	G	A	0.671	0.939	0.264	1.53E-07	-	2.52e-02	2.60e-07
6	rs4145451	12372611	C	A	0.750	0.159	0.274	5.74E-06	-	4.24e-03	
6	rs11757677	21181008	A	G	0.628	0.267	0.257	5.77E-07	-	8.52e-03	5.03e-08
6	rs9295492	21184246	A	T	0.622	0.428	0.230	8.32E-06	-	2.10e-02	
6	rs4401656	21203121	G	T	0.621	0.541	0.226	7.29E-06	-	2.74e-02	
6	rs12210919	37753490	C	T	0.907	0.727	0.282	3.33E-06	-	3.42e-02	
9	rs4977708	26590785	T	G	0.479	0.091	0.196	6.12E-07	-	5.94e-04	1.28e-03
9	rs7873088	111232842	T	C	0.745	0.001	0.247	5.60E-06	-	2.64e-02	
11	rs658845	56050851	A	G	0.784	0.324	0.313	4.06E-06	-	5.94e-02	
11	rs2187511	56063429	C	G	0.784	0.210	0.316	3.00E-06	-	5.36e-02	
11	rs10893739	127142613	A	G	0.876	0.660	0.266	5.02E-06	-	6.04e-02	
12	rs7964876	122124658	T	C	0.929	0.126	0.292	7.00E-06	-	3.47e-02	
15	rs12902949	39954777	A	G	0.700	0.378	0.238	2.74E-06	-	6.53e-03	
15	rs17269348	56611933	C	T	0.875	0.342	0.298	3.26E-06	-	1.74e-02	
16	rs7185166	71444684	C	T	0.584	0.188	-0.193	2.83E-07	1.01e-03	6.26e-04	5.61e-02

CHR: chromosome number; Rs ID: SNP marker name; Pos: SNP position on the chromosome; A1: the other allele; A2: the coding allele; AF: the allele frequency of the coding allele; P (HWE): the p-value from the HWE test; P (Aalen): the p-value from the Aalen's additive model; P (Cox): the p-value from the Cox PH model.

The results from the GWAS of the optimal state of BMI for protecting from T2D using the GSPM. The row in boldface is the SNP with a p-value of suggestive significance ($p<1e-06$).

Table 3

CHR	Rs ID	Pos	A1	A2	AF (coding)	P (HWE)	f_{G0}	f_{G2}	β_f	P_f	P (Aalen)	P (Cox)
2	rs1551133	119253354	A	C	0.924	0.508	0.345	0.196	-0.074	6.70E-07	1.19e-01	7.74e-02
9	rs2613310	92670266	C	T	0.785	0.782	0.125	0.181	0.028	1.41E-06	3.47e-05	3.04e-06
9	rs965892	92669610	G	C	0.785	0.672	0.125	0.181	0.028	1.47E-06	2.55e-05	5.1e-06

CHR: chromosome number; Rs ID: SNP marker name; Pos: SNP position on the chromosome; A1: the other allele; A2: the coding allele; AF: the allele frequency of the coding allele; P (HWE): the p-value from the HWE test; P (Aalen): the p-value from the Aalen's additive model; P (Cox): the p-value from the Cox PH model.

The T2D incidence and the average BMI levels (with standard errors) at the onset of three genotype groups of rs1551133.

Table 4

Genotype of rs1551133	Sample size	T2D events	Incidence	Average BMI at onset	SD of average BMI
AA	55	7	12.7%	25.6	3.26
AC	1138	45	3.95%	29.7	5.81
CC	7091	418	5.89%	30.3	5.27