Meta-analysis of Complex Diseases at Gene Level with Generalized Functional Linear Models

Ruzong Fan,*[,1] Yifan Wang,* Chi-yang Chiu,* Wei Chen,[†] Haobo Ren,[‡] Yun Li,[§] Michael Boehnke, Christopher I. Amos,[††] Jason H Moore,[‡‡] and Momiao Xiong****

*Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, [†]Division of Pulmonary Medicine, Allergy and Immunology, University of Pittsburgh, Medical Center, Pittsburgh, Pennsylvania 15224, [‡]Regeneron Pharmaceuticals, Inc., Basking Ridge, New Jersey 07920, [§]Departments of Genetics and Biostatistics, University of North Carolina, Chapel Hill, North Carolina, 27599, **Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109, [††]Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, New Hampshire 03756, [‡‡]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, and [§§]Human Genetics Center, University of Texas, Houston, Texas 77225

ABSTRACT We developed generalized functional linear models (GFLMs) to perform a meta-analysis of multiple case-control studies to evaluate the relationship of genetic data to dichotomous traits adjusting for covariates. Unlike the previously developed meta-analysis for sequence kernel association tests (MetaSKATs), which are based on mixed-effect models to make the contributions of major gene loci random, GFLMs are fixed models; *i.e.*, genetic effects of multiple genetic variants are fixed. Based on GFLMs, we developed chi-squared-distributed Rao's efficient score test and likelihood-ratio test (LRT) statistics to test for an association between a complex dichotomous trait and multiple genetic variants. We then performed extensive simulations to evaluate the empirical type I error rates and power performance of the proposed tests. The Rao's efficient score test statistics of GFLMs are very conservative and have higher power than MetaSKATs when some causal variants are rare and some are common. When the causal variants are all rare [*i.e.*, minor allele frequencies (MAF) < 0.03], the Rao's efficient score test statistics have similar or slightly lower power than MetaSKATs. The LRT statistics generate accurate type I error rates for homogeneous genetic-effect models and may inflate type I error rates for heterogeneous genetic-effect models owing to the large numbers of degrees of freedom and have similar or slightly higher power than the Rao's efficient score test statistics. GFLMs were applied to analyze genetic data of 22 gene regions of type 2 diabetes data from a meta-analysis of eight European studies and detected significant association for 18 genes ($P < 3.10 \times 10^{-6}$), tentative association for 2 genes (*HHEX* and *HMGA2*; $P \approx 10^{-5}$), and no association for 2 genes, while MetaSKATs detected none. In addition, the traditional additive-effect model detects association at gene *HHEX*. GFLMs and related tests can analyze rare or common variants or a combination of the two and can be useful in whole-genome and whole-exome association studies.

KEYWORDS meta-analysis; rare variants; common variants; association mapping; complex traits; functional data analysis

FOR association studies of many complex traits, multiple studies may have been conducted that have collected the same phenotypic traits. For example, a large number of studies of type 2 diabetes (T2D) have been conducted to evaluate the relationship between single-nucleotide polymorphisms (SNPs) and T2D (Morris *et al.* 2012; Scott *et al.* 2012; Li *et al.* 2014). The sample size of an individual study can be small or moderate and may not always lead to a significant association signal at a genome-wide requirement. It is desirable to combine multiple studies for a unified meta-analysis in order to reach rigorous significant threshold levels (Zeggini and Ioannidis 2009; Evangelou and Ioannidis 2013; Liu *et al.* 2014). By combining multiple studies together, one can get a sample with a large sample size, and it is more likely to produce significant results. However, different studies may contain different genetic data or covariates, which make analysis

doi: 10.1534/genetics.115.180869
Manuscript received July 18, 2015; accepted for publication December 9, 2015; published Early Online December 29, 2015.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180869/-/DC1
[1]Corresponding author: Biostatistics and Bioinformatics Branch, Eunice Kennedy Shriver National Institute of Child Health and Human Development, 6100 Executive Blvd., MSC 7510, Bethesda, MD 20892. E-mail: fanr@mail.nih.gov

of the combined data difficult. It is important to develop statistical methods that analyze the combined data of multiple studies.

To perform an association meta-analysis for complex traits, one may take two strategies: (1) single-genetic-variant-based approaches and (2) gene-based-variant-analysis approaches. The single-genetic-variant approaches use only one genetic variant at a time and are useful to analyze common variants (Zeggini *et al.* 2008; Hindorff *et al.* 2009; Stahl *et al.* 2010). Gene-based association analysis uses multiple genetic variants to detect an association. In recent years, there has been a great deal of interest in developing statistical methods and tests for gene-based association analysis of complex traits (Hu *et al.* 2013; Liu *et al.* 2014). Gene-based analysis can lead to higher power and improve multiple-comparison problems compared to single-marker analysis because fewer tests are required. More important, gene-based analysis can be the only way to analyze rare variants that have minor allele frequencies (MAFs) < 0.01–0.05 because it could be powerless to use a single rare variant in an analysis.

Burden tests and kernel-based test methods are popular approaches to performing rare variant-gene-based association analyses. Burden tests collapse rare variants into a single variable to test for an association with a complex trait and to reduce the high dimensionality of genetic data (Li and Leal 2008; Madsen and Browning 2009; Han and Pan 2010; Morris and Zeggini 2010; Price *et al.* 2010; Neale *et al.* 2011). Kernel-based test methods are based on mixed-effect models in which the regression coefficients of multiple genetic variants are random with means of zero and constant variance. The association is tested by testing a null hypothesis of zero variance by a sequence kernel association test (SKAT). The SKAT and its optimal unified test (SKAT-O) were found to have higher power than burden tests (Wu *et al.* 2011; Lee *et al.* 2012). By extending SKAT and SKAT-O to perform meta-analyses, Lee *et al.* (2013) developed the meta-analysis for sequence kernel association test (MetaSKAT) and its optimal unified test (MetaSKAT-O) to carry out meta-analyseis.

The regression coefficients of genetic terms in the models of SKAT and MetaSKAT were assumed to be random because the number of genetic variants is usually large for modern genetic data. In population genetics, however, the genetic effects of major gene loci are usually assumed to be fixed, while the contributions of polygenic loci are modeled as a random term (Fisher 1918). The high dimensionality of modern genetic data does not necessarily imply that traditional population genetics theory is not correct because the number of causal variants may not be large. A fixed model should be fine to analyze the major gene locus data in most cases if the dimension of the genetic data can be properly reduced.

By viewing genetic variant data as realizations of an underlying stochastic process, functional regression models were proposed to reduce the dimensionality and to perform a gene-based association analysis of quantitative, qualitative, and survival traits (Luo *et al.* 2011, 2012, 2013; Fan *et al.* 2013, 2014, 2015, 2016; Vsevolozhskaya *et al.* 2014; Zhang *et al.* 2014; Wang *et al.* 2015). For quantitative traits, functional linear models lead to both *F*- and chi-squared-distributed test statistics that are almost always more powerful than SKAT and SKAT-O (Luo *et al.* 2012; Fan *et al.* 2013, 2015; Wang *et al.* 2015). For dichotomous and survival traits, functional regression models lead to test statistics that are more powerful than SKAT and SKAT-O except in some cases where the causal variants are all rare (Luo *et al.* 2011, 2013; Fan *et al.* 2014, 2016; Vsevolozhskaya *et al.* 2014). Therefore, functional regression models are found to outperform other methods and potentially to be useful in gene-based association analysis of complex traits.

In our functional regression models, the genetic effects are treated as a function of the physical position, and the genetic-variant data are viewed as stochastic functions of the physical position, so any orders of linkage disequilibrium (LD) are taken care of in the models (Ross 1996). The regression coefficients of genetic terms in the SKAT and MetaSKAT models do no depend on the physical position, while our genetic-effect function depends on the physical position and is actually a function of physical position. Hence, the functional regression models can fully use LD and physical position information. The functional regression models are a natural extension of traditional population genetics because we model the genetic effects of major gene loci as fixed functions.

In this paper, generalized functional linear models (GFLMs) are developed for a meta-analysis of multiple studies. GFLMs can analyze rare or common variants or a combination of the two. Both chi-squared-distributed Rao's efficient score test statistics and likelihood-ratio test (LRT) statistics are introduced to test for an association between disease traits and multiple genetic variants. Extensive simulations are performed to evaluate the type I error rates and power performance of the GFLMs and tests. The proposed methods were applied to analyze T2D data from a meta-analysis of eight European studies.

Materials and Methods

Consider a meta-analysis with L case-control studies in a genomic region. For the ℓth study, we assume that there are n_ℓ individuals who are sequenced in the genomic region at m_ℓ variants. We assume that the m_ℓ variants are located with ordered physical positions $0 \le t_{\ell 1} < \cdots < t_{\ell m_\ell}$. To make the notation simpler, we normalized the region $[t_{\ell 1}, t_{\ell m_\ell}]$ to be [0, 1]. For the ith individual in the ℓth study, let $y_{\ell i}$ denote his or her dichotomous trait (here $y_{\ell i} = 1$ indicates that the individual is an affected case of the disease of interest, $y_{\ell i} = 0$ indicates that the individual is a normal control individual), $G_{\ell i} = [X_{\ell i}(t_{\ell 1}), \ldots, X_{\ell i}(t_{\ell m_\ell})]'$ denotes his or her genotypes of the m_ℓ variants, and $Z_{\ell i} = (z_{\ell i 1}, \ldots, z_{\ell i c_\ell})'$ denotes his or her c_ℓ covariates. Hereafter in this paper, a prime denotes the transpose of a vector or matrix. For the genotypes, we assume that $X_{\ell i}(t_{\ell j})$ (= 1, 2, 3) is the number of minor alleles of the individual at the jth variant located at position $t_{\ell j}$.

Traditional additive-effect models

By using logistic regression, an additive-effect model (AEM) can be used to analyze the relation between the disease trait $y_{\ell i}$ and the m_ℓ variants in the ℓth study as (Cordell and Clayton 2002)

$$\text{logit}\,(\pi_{\ell i}) = \alpha_{\ell 0} + Z_{\ell i}^{'}\alpha_\ell + \sum_{j=1}^{m_\ell} X_{\ell i}(t_{\ell j})\beta_{\ell j},$$
$$\ell = 1, 2, \ldots, L; i = 1, 2, \ldots, n_\ell \qquad (1)$$

where $\pi_{\ell i} = P(y_{\ell i} = 1)$ is the disease probability, $\alpha_{\ell 0}$ is the regression intercept, $\alpha_\ell = (\alpha_{\ell 1}, \ldots, \alpha_{\ell c_\ell})'$ is a $c_\ell \times 1$ column vector of regression coefficients of covariates, and $\beta_{\ell j}$ is the additive genetic effect of variant j for the ℓth study. The number of the parameters of the model (1) can be large, so it may not be powerful. Despite the potential drawbacks, the model (1) can be easily implemented by standard statistical software such as R. If the number of genetic variants is large, one may decompose the genotype matrix into the product of an orthogonal matrix **Q** and a triangular matrix **R** via Gram-Schmidt process to remove the redundancy to facilitate computation in applications, *i.e.*, the **QR** decomposition.

β-Smooth-only GFLMs

To model the relation between the disease trait $y_{\ell i}$ and the m_ℓ variants, we propose the following functional logistic regression model:

$$\text{logit}\,(\pi_{\ell i}) = \alpha_{\ell 0} + Z_{\ell i}^{'}\alpha_\ell + \sum_{j=1}^{m_\ell} X_{\ell i}(t_{\ell j})\beta_\ell(t_{\ell j}),$$
$$\ell = 1, 2, \ldots, L; i = 1, 2, \ldots, n_\ell \qquad (2)$$

where $\beta_\ell(t_{\ell j})$ is the genetic effect of the variant at position $t_{\ell j}$, and the other terms are similar to those in the AEM (1). Note that we have L studies, so the effect of a common covariate can be either the same or different across the studies: (1) heterogeneous: we treat α_ℓ as different for different studies, *i.e.*, $\alpha_\ell, \ell = 1, \ldots, L$, are all different; and (2) homogeneous: if a covariate is present in different studies, we model its regression coefficient by one common coefficient.

In the model (2), $\beta_\ell(t_{\ell j})$ is introduced as the genetic effect of the variant at position $t_{\ell j}$. We assume that $\beta_\ell(t)$ is a continuous function of the physical position t. One may expand it by B-spline or Fourier or linear spline basis functions. Formally, let us expand the genetic-effect function $\beta_\ell(t)$ by a series of K_β basis functions $\psi(t) = [\psi_1(t), \ldots, \psi_{K_\beta}(t)]'$ as $\beta_\ell(t) = [\psi_1(t), \ldots, \psi_{K_\beta}(t)](\beta_{\ell 1}, \ldots, \beta_{\ell K_\beta})' = \psi(t)'\beta_\ell$, where $\beta_\ell = (\beta_{\ell 1}, \ldots, \beta_{\ell K_\beta})'$ is a vector of coefficients $\beta_{\ell 1}, \ldots, \beta_{\ell K_\beta}$. We consider two types of basis functions: (1) the B-spline basis $\psi_k(t) = B_k(t), k = 1, \ldots, K_\beta$ and (2) the Fourier basis $\psi_1(t) = 1$, $\psi_{2r+1}(t) = \sin(2\pi r t)$, and $\psi_{2r}(t) = \cos(2\pi r t)$, $r = 1, \ldots, (K_\beta - 1)/2$. Here, for Fourier basis, K_β is taken as a positive odd integer (de Boor 2001; Ramsay and Silverman 2005; Ramsay *et al.* 2009; Ferraty and Romain 2010; Horváth and Kokoszka 2012). Replacing $\beta_\ell(t_{\ell j})$ by the expansion, the model (2) can be revised as

$$\text{logit}\,(\pi_{\ell i}) = \alpha_{\ell 0} + Z_{\ell i}^{'}\alpha_\ell + \left\{\sum_{j=1}^{m_\ell} X_{\ell i}(t_{\ell j})\left[\psi_1(t_{\ell j}), \ldots, \psi_{K_\beta}(t_{\ell j})\right]\right\}$$
$$\times\left(\beta_{\ell 1}, \ldots, \beta_{\ell K_\beta}\right)' = \alpha_{\ell 0} + Z_{\ell i}^{'}\alpha_\ell + W_{\ell i}^{'}\beta_\ell \qquad (3)$$

where $W_{\ell i}^{'} = \sum_{j=1}^{m_\ell} X_{\ell i}(t_{\ell j})[\psi_1(t_{\ell j}), \ldots, \psi_{K_\beta}(t_{\ell j})]$. In the model (2) and its revised version (3), we use the raw genotype data $G_{\ell i} = [X_{\ell i}(t_{\ell 1}), \ldots, X_{\ell i}(t_{\ell m_\ell})]'$ directly in the analysis.

General GFLM

In this subsection we view the ith individual's genotype data as a genetic variant function (GVF) $X_{\ell i}(t),\ t \in [0, 1]$ in addition to treating the genetic effects as functions $\beta_\ell(t)$. To relate the GVF to the phenotypic traits adjusting for covariates, we consider the following functional logistic regression model:

$$\text{logit}\,(\pi_{\ell i}) = \alpha_{\ell 0} + Z_{\ell i}^{'}\alpha_\ell + \int_0^1 X_{\ell i}(t)\beta_\ell(t)dt,$$
$$\ell = 1, 2, \ldots, L; i = 1, 2, \ldots, n_\ell \qquad (4)$$

where $\beta_\ell(t)$ is the genetic effect of GVF $X_{\ell i}(t)$ at position t, and the other terms are similar to those in the β-smooth-only model (2). In this model, the integration term $\int_0^1 X_{\ell i}(t)\beta_\ell(t)dt$ is used to replace the summation term $\sum_{j=1}^{m_\ell} X_{\ell i}(t_{\ell j})\beta_\ell(t_{\ell j})$ in the β-smooth-only model (2).

Estimation of GVF: Let $\phi_k(t),\ k = 1, \ldots, K$, be a series of K basis functions, such as the B-spline basis and Fourier basis functions. Let Φ denote the $m_\ell \times K$ matrix containing the values $\phi_k(t_{\ell j})$, where $j \in 1, \ldots, m_\ell$. Denote $\phi(t) = [\phi_1(t), \ldots, \phi_K(t)]'$. Using the discrete realizations $G_{\ell i} = [X_{\ell i}(t_{\ell 1}), \ldots, X_{\ell i}(t_{\ell m_\ell})]'$, we may estimate the GVF $X_{\ell i}(t)$ using an ordinary linear square smoother as follows (Ramsay and Silverman 2005, Chapter 4):

$$\hat{X}_{\ell i}(t) = \left[X_{\ell i}(t_{\ell 1}), \ldots, X_{\ell i}(t_{\ell m_\ell})\right]\Phi\left[\Phi'\Phi\right]^{-1}\phi(t) \qquad (5)$$

Revised GFLM: As in the β-smooth-only case, the genetic effect $\beta_\ell(t)$ is expanded by a series of basis functions $\beta_\ell(t) = \psi(t)'\beta_\ell$. Replacing $X_{\ell i}(t)$ in (4) by $\hat{X}_{\ell i}(t)$ in (5) and $\beta_\ell(t)$ by the expansion, we have the following revised logistic regression model:

$$\text{logit}\,(\pi_{\ell i}) = \alpha_{\ell 0} + Z_{\ell i}^{'}\alpha_\ell + \left\{\left[X_{\ell i}(t_{\ell 1}), \ldots, X_{\ell i}(t_{\ell m_\ell})\right]\Phi\left[\Phi'\Phi\right]^{-1}\right.$$
$$\left.\times\int_0^1 \phi(t)\psi'(t)dt\right\}\beta_\ell$$
$$= \alpha_{\ell 0} + Z_{\ell i}^{'}\alpha_\ell + W_{\ell i}^{'}\beta_\ell \qquad (6)$$

where $W_{\ell i}^{'} = [X_{\ell i}(t_{\ell 1}), \ldots, X_{\ell i}(t_{\ell m_\ell})]\Phi[\Phi'\Phi]^{-1}\int_0^1 \phi(t)\psi'(t)dt$. In this revised regression model, one needs to calculate $\Phi[\Phi'\Phi]^{-1}$ and $\int_0^1 \phi(t)\psi'(t)dt$ in order to get $W_{\ell i}$. In the statistical package R, there are readily available codes to calculate them (Ramsay *et al.* 2009).

Table 1 Association analysis of T2D status in eight European cohorts by heterogeneous Rao's efficient score test statistics (Het-Rao), Het-MetaSKAT-O, and Het-MetaSKAT

	P-values of Het-Rao						
	Basis of both GVF and $\beta_\ell(t)$		Basis of β-smooth-only		Additive effect	*P*-values of Het-Meta	
Gene	B-spline	Fourier	B-spline	Fourier	Model (1)	SKAT	SKAT-O
PCSK9	$\mathbf{3.23 \times 10^{-11}}^a$	$\mathbf{4.60 \times 10^{-11}}^a$	$\mathbf{3.23 \times 10^{-11}}^a$	$\mathbf{4.60 \times 10^{-11}}^a$	10^{-5}	0.792	0.059
APOB	$\mathbf{1.13 \times 10^{-22}}^a$	$\mathbf{2.52 \times 10^{-20}}^a$	$\mathbf{1.13 \times 10^{-22}}^a$	$\mathbf{2.52 \times 10^{-20}}^a$	$\mathbf{6.49 \times 10^{-15}}^a$	0.499	0.517
IGF2BP2	$\mathbf{7.06 \times 10^{-9}}^a$	$\mathbf{3.10 \times 10^{-11}}^a$	$\mathbf{7.06 \times 10^{-9}}^a$	$\mathbf{3.10 \times 10^{-11}}^a$	$\mathbf{9.29 \times 10^{-17}}^a$	0.531	0.503
CDKAL1	$\mathbf{9.07 \times 10^{-20}}^a$	$\mathbf{9.01 \times 10^{-22}}^a$	$\mathbf{9.07 \times 10^{-20}}^a$	$\mathbf{9.01 \times 10^{-22}}^a$	$\mathbf{2.11 \times 10^{-9}}^a$	0.961	0.800
JAZF1	$\mathbf{8.03 \times 10^{-29}}^a$	$\mathbf{2.61 \times 10^{-27}}^a$	$\mathbf{8.03 \times 10^{-29}}^a$	$\mathbf{2.61 \times 10^{-27}}^a$	$\mathbf{1.91 \times 10^{-12}}^a$	0.032	0.046
LPL	4.92×10^{-5}	$\mathbf{5.09 \times 10^{-8}}^a$	4.92×10^{-5}	$\mathbf{5.09 \times 10^{-8}}^a$	$\mathbf{7.34 \times 10^{-12}}^a$	0.590	0.795
CDKN2B	$\mathbf{2.94 \times 10^{-35}}^a$	$\mathbf{9.98 \times 10^{-28}}^a$	$\mathbf{2.94 \times 10^{-35}}^a$	$\mathbf{9.98 \times 10^{-28}}^a$	$\mathbf{6.17 \times 10^{-25}}^a$	0.554	0.410
CDC123	$\mathbf{1.66 \times 10^{-18}}^a$	$\mathbf{6.98 \times 10^{-18}}^a$	$\mathbf{1.66 \times 10^{-18}}^a$	$\mathbf{6.98 \times 10^{-18}}^a$	$\mathbf{1.31 \times 10^{-14}}^a$	0.039	0.072
IDE	$\mathbf{1.47 \times 10^{-21}}^a$	$\mathbf{6.62 \times 10^{-23}}^a$	$\mathbf{1.47 \times 10^{-21}}^a$	$\mathbf{6.62 \times 10^{-23}}^a$	$\mathbf{3.66 \times 10^{-16}}^a$	0.414	0.630
KIF11	$\mathbf{1.57 \times 10^{-23}}^a$	$\mathbf{1.91 \times 10^{-23}}^a$	$\mathbf{1.57 \times 10^{-23}}^a$	$\mathbf{1.91 \times 10^{-23}}^a$	$\mathbf{1.68 \times 10^{-21}}^a$	0.768	0.913
HHEX	3.48×10^{-5}	2.97×10^{-5}	5.10×10^{-6}	2.97×10^{-5}	$\mathbf{2.95 \times 10^{-6}}^a$	0.480	0.691
TCF7L2	$\mathbf{7.51 \times 10^{-11}}^a$	$\mathbf{6.06 \times 10^{-10}}^a$	$\mathbf{7.51 \times 10^{-11}}^a$	$\mathbf{6.06 \times 10^{-10}}^a$	1.02×10^{-4}	0.021	0.042
KCNQ1	$\mathbf{3.67 \times 10^{-31}}^a$	$\mathbf{4.94 \times 10^{-29}}^a$	$\mathbf{3.67 \times 10^{-31}}^a$	$\mathbf{4.94 \times 10^{-29}}^a$	$\mathbf{2.64 \times 10^{-8}}^a$	0.572	0.797
MTNR1B	$\mathbf{2.09 \times 10^{-17}}^a$	$\mathbf{2.27 \times 10^{-15}}^a$	$\mathbf{2.09 \times 10^{-17}}^a$	$\mathbf{2.27 \times 10^{-15}}^a$	$\mathbf{8.54 \times 10^{-14}}^a$	0.295	0.456
HMGA2	1.68×10^{-5}	1.99×10^{-4}	1.68×10^{-5}	1.99×10^{-4}	6.18×10^{-2}	0.699	0.887
TSPAN8	$\mathbf{4.78 \times 10^{-38}}^a$	$\mathbf{9.39 \times 10^{-38}}^a$	$\mathbf{5.89 \times 10^{-38}}^a$	$\mathbf{1.48 \times 10^{-37}}^a$	$\mathbf{1.01 \times 10^{-36}}^a$	0.747	0.923
HNF1A	$\mathbf{1.71 \times 10^{-16}}^a$	$\mathbf{1.10 \times 10^{-15}}^a$	$\mathbf{1.71 \times 10^{-16}}^a$	$\mathbf{1.10 \times 10^{-15}}^a$	$\mathbf{3.56 \times 10^{-26}}^a$	0.272	0.441
OASL	$\mathbf{6.01 \times 10^{-35}}^a$	$\mathbf{1.06 \times 10^{-28}}^a$	$\mathbf{6.01 \times 10^{-35}}^a$	$\mathbf{1.06 \times 10^{-28}}^a$	$\mathbf{8.85 \times 10^{-24}}^a$	0.530	0.416
FTO	$\mathbf{1.26 \times 10^{-25}}^a$	$\mathbf{1.14 \times 10^{-26}}^a$	$\mathbf{1.26 \times 10^{-25}}^a$	$\mathbf{1.14 \times 10^{-26}}^a$	$\mathbf{1.17 \times 10^{-21}}^a$	0.048	0.090
LDLR	0.373	0.477	0.373	0.477	0.427	0.233	0.400
APOE	$\mathbf{2.07 \times 10^{-31}}^a$	$\mathbf{2.19 \times 10^{-27}}^a$	$\mathbf{2.07 \times 10^{-31}}^a$	$\mathbf{2.19 \times 10^{-27}}^a$	$\mathbf{7.19 \times 10^{-30}}^a$	0.042	0.082
GIPR	5.99×10^{-3}	9.56×10^{-3}	5.99×10^{-3}	9.56×10^{-3}	0.013	0.808	0.303

The results of "Basis of both GVF and $\beta_\ell(t)$" were based on smoothing both GVF and genetic-effect functions $\beta_\ell(t)$ of model 6, and the results of "Basis of β-smooth-only" were based on the smoothing $\beta_\ell(t)$ only approach of model 3, and the *P*-values of Het-MetaSKAT and Het-MetaSKAT-O were based of the R package MetaSKAT. GVF, genetic variant function.

[a] Associations that attain a threshold significance of $P < 3.1 \times 10^{-6}$.

Test statistics of association

We consider the revised regression models [(3) and (6)] as usual logistic regressions that model the genetic effect of GVFs adjusted for covariates. First, assume that the genetic effects among the L studies are heterogeneous. To test for an association between the genetic variants and the disease trait, the null hypothesis is H_0: $\beta_\ell = (\beta_{\ell 1}, \ldots, \beta_{\ell K_\beta})' = 0$, $\ell = 1, \ldots, L$. We may test the null hypothesis by a chi-squared-distributed Rao's efficient score statistic with a degree of freedom of LK_β. The Rao's efficient score statistic is denoted by GFLM Het-Rao. An alternative approach is to use a LRT statistic to test for association, which is also chi-squared distributed with LK_β degrees of freedom and is denoted by GFLM Het-LRT.

If the genetic effects are homogeneous, *i.e.*, $\beta_\ell = (\beta_{\ell 1}, \ldots, \beta_{\ell K_\beta})' = \beta = (\beta_1, \ldots, \beta_{K_\beta})'$, $\ell = 1, \ldots, L$, we may test for association by testing a simplified null hypothesis H_0: $\beta = (\beta_1, \ldots, \beta_{K_\beta})' = 0$. Again, one may use a chi-squared-distributed Rao's efficient score statistic and a chi-squared-distributed LRT statistic to test the null hypothesis. Both the chi-squared-distributed Rao's efficient score statistic and the LRT statistic have a degree of freedom of K_β and are denoted by GFLM Hom-Rao and GFLM Hom-LRT, respectively.

For the AEM (1), the null hypothesis is H_0: $\beta_\ell = (\beta_{\ell 1}, \ldots, \beta_{\ell m_\ell})' = 0$, $\ell = 1, \ldots, L$, under an assumption of heterogeneous genetic effect. The corresponding chi-squared-distributed Rao's efficient score test and LRT statistics are chi-squared distributed with $\sum_{\ell=1}^{L} m_\ell$ degrees of freedom. The tests are denoted as AEM Het-Rao and AEM Het-LRT, respecyively. Assume that each individual of the L studies is sequenced at the same variants at $0 \le t_1 < \cdots < t_m$ and so $m_1 = \cdots = m_\ell = m$. In addition, assume that the genetic effects are homogeneous, *i.e.*, $\beta_\ell = (\beta_{\ell 1}, \ldots, \beta_{\ell m_\ell})' = \beta = (\beta_1, \cdots, \beta_m)'$. Then the AEM (1) is simplified as

$$\text{logit}\,(\pi_{\ell i}) = \alpha_{\ell 0} + Z'_{\ell i}\alpha_\ell + \sum_{j=1}^{m} X_{\ell i}(t_j)\beta_j,$$
$$\ell = 1, 2, \ldots, L; i = 1, 2, \ldots, n_\ell \qquad (7)$$

The null hypothesis of no association between the genetic variants and the disease trait is H_0: $\beta = (\beta_1, \ldots, \beta_m)' = 0$. The corresponding Rao and LRT statistics are chi-squared distributed with m degrees of freedom. The tests are denoted as AEM Hom-Rao and AEM Hom-LRT, respectively.

Parameters of functional data analysis

In the data analysis and simulations, we used a functional data analysis procedure in the statistical package R. We use two functions in library fda of R package as follows to create basis:
basis = create.bspline.basis(norder = order, nbasis = bbasis)
basis = create.fourier.basis(c(0,1), nbasis = fbasis)

Table 2 Association analysis of T2D status in eight European cohorts by homogeneous Rao's efficient score test statistics (Hom-Rao), Hom-MetaSKAT-O, and Hom-MetaSKAT

	P-values of Hom-Rao						
	Basis of both GVF and $\beta_t(t)$		Basis of β-smooth only		Additive effect	*P*-values of Hom-Meta	
Gene	B-spline	Fourier	B-spline	Fourier	Model (1)	SKAT	SKAT-O
PCSK9	0.079	0.034	0.181	0.229	0.780	0.063	0.025
APOB	0.035	0.081	0.012	0.021	0.873	0.807	0.623
IGF2BP2	0.017	4.48×10^{-3}	1.13×10^{-3}	2.30×10^{-4}	0.041	0.417	0.368
CDKAL1	0.190	0.214	0.056	0.081	0.416	0.473	0.646
JAZF1	0.446	0.422	0.199	0.302	0.476	0.352	0.094
LPL	0.075	0.013	0.080	0.011	0.148	0.416	0.559
CDKN2B	0.001	5.63×10^{-5}	6.04×10^{-3}	0.015	0.147	0.325	0.430
CDC123	0.039	0.027	0.076	0.071	0.040	0.129	0.210
IDE	0.241	0.138	0.155	0.308	0.368	0.252	0.389
KIF11	0.040	0.036	0.065	0.187	0.864	0.667	0.802
HHEX	0.020	0.004	0.030	0.034	0.378	0.684	0.711
TCF7L2	$\mathbf{2.67 \times 10^{-14}}$[a]	$\mathbf{4.36 \times 10^{-14}}$[a]	$\mathbf{1.94 \times 10^{-16}}$[a]	$\mathbf{1.07 \times 10^{-15}}$[a]	$\mathbf{8.11 \times 10^{-7}}$[a]	1.37×10^{-4}	3.03×10^{-4}
KCNQ1	0.061	0.143	0.106	0.142	0.103	0.420	0.601
MTNR1B	6.70×10^{-4}	4.55×10^{-4}	0.012	$\mathbf{4.90 \times 10^{-9}}$[a]	0.357	0.523	0.641
HMGA2	0.757	0.911	0.671	0.903	0.598	0.880	1
TSPAN8	0.448	1.47×10^{-5}	2.40×10^{-3}	3.84×10^{-4}	0.910	0.991	0.836
HNF1A	0.135	0.046	0.087	7.83×10^{-3}	0.194	0.661	0.363
OASL	0.075	0.026	0.032	0.030	0.371	0.477	0.305
FTO	7.02×10^{-4}	1.83×10^{-4}	3.70×10^{-6}	$\mathbf{2.07 \times 10^{-7}}$[a]	0.283	0.291	0.428
LDLR	0.951	0.916	0.933	0.933	0.907	0.876	0.727
APOE	0.449	0.155	0.024	4.77×10^{-3}	0.045	0.038	0.070
GIPR	0.058	0.038	0.037	0.128	0.034	0.306	0.250

The results of "Basis of both GVF and $\beta_t(t)$" were based on smoothing both GVF and genetic-effect functions $\beta_t(t)$ of model 6, and the results of "Basis of β-smooth-only" were based on the smoothing $\beta_t(t)$ only approach of model 3, and the *P*-values of Hom-MetaSKAT and Hom-MetaSKAT-O were based of the R package MetaSKAT. GVF, genetic variant function.

[a] Associations that attain a threshold significance of $P < 3.1 \times 10^{-6}$.

The three parameters were taken as *order = 4*, *bbasis = 10*, *fbasis = 11* for the heterogeneous genetic-effect model and as *order = 4*, *bbasis = 12*, *fbasis = 13* for the homogeneous genetic-effect model in all data analyses and simulations. To make sure that the results are valid and stable, we tried a wide range of parameters: (1) $8 \le K = K_\beta \le 13$ for the heterogeneous genetic-effect model and (2) $10 \le K = K_\beta \le 21$ for the homogeneous genetic-effect model. The results are similar to each other (data not shown).

Data availability

Computer Program: The methods proposed in this paper are implemented by using the procedure of functional data analysis (fda) in the statistical package R. The R codes for data analysis and simulations are available from http://www.nichd.nih.gov/about/org/diphr/bbb/software/fan/Pages/default.aspx.

Results

Meta-analysis of T2D in eight European cohorts

The proposed methods were applied to analyze a set of studies investigating T2D that includes eight European cohorts: the FIN-D2D 2007 study (D2D2007), the Diabetes Genetic study (DIAGEN), the Finnish Diabetes Prevention Study (DPS), the Finland–United States Investigation of NIDDM Genetics study (FUSION Stage 2), the Nord-Trøndelag Health Study 2 (HUNT), the Metabolic Syndrome in Men study (METSIM), and the Tromsø study (TROMSO). The sample sizes of cases and controls for each study are provided in Supporting Information, Table S1. For each cohort, 54,741 genetic variants are genotyped and are located in 97 genetic regions across the 22 autosomes. For our analysis, we used the literature on T2D as a reference for gene selection and found that 22 gene regions were fine mapped (Zeggini *et al.* 2008; Voight *et al.* 2010; Morris *et al.* 2012; Scott *et al.* 2012; Li *et al.* 2014; Liu *et al.* 2014). We used Builder Mar. 2006 (NCBI36/hg18) to determine gene positions, and 5 kb was used to extend the gene region on each side of a gene. A summary of the 22 genes and the number of genetic variants in each gene region are given in Table S2.

Association analysis between T2D status and each of the 22 genes was performed by the proposed methods and MetaSKAT. Except for METSIM, age and sex were used as covariates. For METSIM, age was used as a covariate because no females were included in the study. A significance threshold of $P < 3.1 \times 10^{-6}$ was taken from Liu *et al.* (2014). If a *P*-value is around 10^{-5} but larger than 3.1×10^{-6}, we call it a "tentative significant association signal."

Table 1 reports the results of association analysis of the eight European cohorts by heterogeneous Rao's efficient score test (Het-Rao), Het-MetaSKAT-O, and Het-MetaSKAT, and Table 2 reports the results by homogeneous Rao's efficient score test (Hom-Rao), Hom-MetaSKAT-O, and Hom-MetaSKAT. The results of Het-LRT and Hom-LRT are reported

Table 3 Simulation study settings

Scenario	Population	Sample sizes			Covariates		
		Study 1	Study 2	Study 3	Study 1	Study 2	Study 3
1	EUR	1600	2200	3200	(z_1, z_2)	(z_1, z_2)	(z_1, z_2)
2	EUR	1600	2200	3200	(z_1, z_2, z_3)	(z_1, z_2, z_3)	(z_1, z_2, z_3)
3	EUR + AA	1600	2200	3200	(z_1, z_2, z_3)	(z_1, z_2, z_3)	(z_1, z_2, z_3)

Sample sizes are total sample sizes in each study, in which half are cases and the rest half are control individuals. Covariates represent covariates in each study. EUR refers to the scenario where all three studies had EUR samples. EUR + AA refers to the scenario where studies 1 and 2 had EUR samples and study 3 had AA samples. z_1 is a binary covariate taking values 0 and 1 each with probability 0.5, and z_2 and z_3 are continuous covariates and distributed as standard normal.

in Table S3 and Table S4, respectively. At the significance threshold of $P < 3.1 \times 10^{-6}$, we observe associations for 17 genes, *PCSK9*, *APOB*, *IGF2BP2*, *CDKAL1*, *JAZF1I CDKN2B*, *CDC123*, *IDE*, *KIF11*, *TCF7L2*, *KCNQ1*, *MTNR1B*, *TSPAN8*, *HNF1A*, *OASL*, *FTO*, and *APOE*, by both Het-Rao and Het-LRT in both the revised β-smooth-only GFLM (3) and the revised general GFLM (6) for both B-spline and Fourier basis functions in Table 1 and Table S3. For the *LPL* gene, a significant association signal is observed by both Het-Rao and Het-LRT in both the β-smooth-only GFLM (3) and the revised general GFLM (6) for Fourier basis functions, while B-spline basis functions lead to tentative association signals. Tentative association signals are observed for two genes, *HHEX* and *HMGA2*, in Table 1 and Table S3, respectively. Only two genes, *LDLR* and *GIPR*, show no association signal.

By both Hom-Rao and Hom-LRT in both the revised β-smooth-only GFLM (3) and the revised general GFLM (6) for both B-spline and Fourier basis functions in Table 2 and Table S4, association is observed for gene *TCF7L2* at the significance threshold of $P < 3.1 \times 10^{-6}$. By both Hom-Rao and Hom-LRT in the β-smooth-only GFLM (3) for Fourier basis functions, significant association signals are observed for two genes, *MTNR1B* and *FTO*, in Table 2 and Table S4. Tentative association signals are observed for two genes, *CDKN2B* and *TSPAN8*, by both Hom-Rao and Hom-LRT in Table 2 and Table S4 for Fourier basis functions in the revised general GFLM (6), respectively.

The *P*-values of Hom-LRT in Table S4 are very similar to those of Hom-Rao in Table 2, and the *P*-values of Het-LRT in Table S3 are slightly smaller than those of Het-Rao in Table 1. Hence, the LRT statistics can be slightly more powerful than the Rao's efficient score test statistics. It is noteworthy that most association signals are detected by Het-LRT and Het-Rao, but Hom-LRT and Hom-Rao only detect association signals for three genes, *TCF7L2*, *MTNR1B*, and *FTO*, reflecting the presence of heterogeneity of the genetic effects.

In addition to the results of GFLMs 3 and 6, MetaSKAT, and MetaSKAT-O, Table 1, Table 2, Table S3, and Table S4 report the results of traditional additive-effect models 1 and 7. Additive-effect models 1 and 7 detect most association signals of GFLMs 3 and 6 in Table 1 and Table 2. In particular, the Het-Rao and Het-LRT of the AEM (1) detect association for *HHEX* in Table 1 and Table S3.

It is noteworthy that Het-MetaSKAT-O, Het-MetaSKAT, Hom-MetaSKAT-O, and Hom-MetaSKAT do not detect any significant signals in any of the 22 genes. The 22 genes are from the literature on T2D, and each of them contains SNPs that are associated with T2D. Thus, significant association signals for T2D are expected for most of the 22 genes if a gene-based method is sensitive. However, MetaSKAT detected no associations for the 22 genes, although our GFLMs and AEM detect associations for 19 genes. Therefore, MetaSKAT is less sensitive than the proposed LRT and Rao's efficient score test statistics for the T2D data in the European cohorts. In Table S5, Table S6, Table S7, and Table S8, we report the results of Rao's efficient score tests by dividing the data between rare and common variants based on a cutoff of 0.03. It is worth noting that the 22 genes contain both rare and common variants and that the associations are mainly from common variants. SKAT and MetaSKAT are designed to analyze rare variants, while the GFLMs and the AEM can analyze rare or common variants or a combination of the two.

When we analyze the data sets separately for each study by the method proposed in Fan *et al.* (2014), significant association is only detected at *TCF7L2* in the study of Norway by Rao's efficient score test and the LRT (data not shown). Thus, it is advantageous to perform a meta-analysis of multiple studies.

A simulation study

Simulations were performed to evaluate the performance of the proposed methods for two cases: (1) all causal variants are rare, and (2) some causal variants are rare and some are common. Three scenarios listed in Table 3 were considered for the simulations. Scenarios 1 and 2 used the European-like (EUR) sequence data, which are the same as those in Lee *et al.* (2012). Scenario 3 used both the EUR and African-American-like (AA) sequence data. The EUR sequence data were generated using COSI's calibrated best-fit models, and the generated European haplotypes mimick CEPH Utah individuals with ancestry from northern and western Europe in terms of site-frequency spectrum and LD pattern (Schaffner *et al.* 2005, Figure 4; International HapMap Consortium 2007). Similarly, the AA sequence data mimick the Yoruba from Ibadan (YRI) (Nigeria in Africa) individuals with a 20:80 mixture of Europeans and Africans, together with parameters calibrated to model realistic demographic history (including bottleneck, population expansion, and migration events). The EUR data included 10,000 chromosomes covering 1-Mb regions, and the AA data included 45,000 chromosomes covering 0.1-Mb regions.

Type I error simulations: To evaluate the type I error rates of the proposed models and tests, we generated phenotype data sets by using the model

$$\text{logit}(\pi_{\ell i}) = \alpha_0 + 0.5z_{\ell i1} + 0.5z_{\ell i2}, \qquad \ell = 1,2,3 \tag{8}$$

for scenario 1 in Table 3 and

$$\text{logit}(\pi_{\ell i}) = \alpha_0 + 0.5z_{\ell i1} + 0.5z_{\ell i2} + 0.5z_{\ell i3}, \qquad \ell = 1,2,3 \tag{9}$$

for scenarios 2 and 3 in Table 3, where $z_{\ell i1}$ is a dichotomous covariate taking values 0 and 1 with a probability of 0.5, $z_{\ell i2}$ and $z_{\ell i3}$ are continuous covariates from standard normal distributions $N(0,1)$, and $\alpha_0 = -4.60 = \log[0.01/(1-0.01)]$ was chosen to provide a disease prevalence of 0.01 under a null hypothesis $z_{\ell ij} = 0$. To obtain genotype data, 3-kb subregions were randomly selected in the 1-Mb regions of EUR and AA data. The ordered genotypes were these variants in the 3-kb subregions. Note that the trait values are not related to the genotypes, so the null hypothesis holds. We calculated empirical type I error rates for both Rao's efficient score test and LRT statistics.

The sample sizes of the data sets were taken as 1600 (study 1), 2200 (study 2), and 3200 (study 3), respectively. For each study, half the sample consists of cases, and the remaining half consists of control individuals. The simulation settings are summarized in Table 3. For each sample-size combination, 10^6 phenotype-genotype data sets were generated to fit the proposed models and to calculate the test statistics and related *P*-values. Then an empirical type I error rate was calculated as the proportion of 10^6 *P*-values that were smaller than a given α level (*i.e.*, 0.05, 0.01 and 0.001, 0.0001, respectively).

Empirical power simulations: To evaluate the power performance of the proposed models and tests, we simulated data sets under the alternative hypothesis by randomly selecting 3-kb subregions to obtain causal variants for the disease traits as follows: once a 3-kb subregion was selected, a subset of *p* causal variants located in the 3-kb subregion was then randomly selected to obtain ordered genotypes $[g(t_1), \ldots, g(t_p)]$. Then we generated the disease traits by

$$\text{logit}(\pi_{\ell i}) = \alpha_0 + 0.5z_{\ell i1} + 0.5z_{\ell i2} + \beta_{\ell i1}g(t_1) + \cdots + \beta_{\ell ip}g(t_p), \qquad \ell = 1,2,3 \tag{10}$$

for scenario 1 in Table 3 and by

$$\text{logit}(\pi_{\ell i}) = \alpha_0 + 0.5z_{\ell i1} + 0.5z_{\ell i2} + 0.5z_{\ell i3} + \beta_{\ell i1}g(t_1) + \cdots + \beta_{\ell ip}g(t_p), \qquad \ell = 1,2,3 \tag{11}$$

for scenarios 2 and 3, where α_0 and $z_{\ell ij}$ are the same as in models 8 and 9, and β is as follows: we used $|\beta_{\ell ij}| = c_\ell |\log_{10}(\text{MAF}_j)|/2$, where MAF_j is the MAF of the *j*th variant. Three different settings were considered: 5, 10, and 20% of variants in the 3-kb subregion are chosen as causal variants. When 5, 10, and 20% of the variants were causal, two parameter settings of genetic effects were considered for c_ℓ: (1) homogeneous and (2) heterogeneous (Table 4). In the homogeneous case, the genetic effects are the same for the three studies, *i.e.*, $c_1 = c_2 = c_3$. In the heterogeneous case, the genetic effects are different for the three studies, *i.e.*, $c_2 = c_1 + 0.15$ and $c_3 = c_1 - 0.15$. For each setting, 1000 data sets were simulated to calculate the empirical power as the proportion of *P*-values that are smaller than an $\alpha = 0.0001$ level.

Table 4 Simulation parameter settings

		Percent of causal variants		
Genetic effect	**Study (c_ℓ)**	**5**	**10**	**20**
Homogeneous	1 (c_1)			
	2 (c_2)	0.60	0.46	0.35
	3 (c_3)			
Heterogeneous	1 (c_1)	0.60	0.46	0.35
	2 (c_2)	0.60 + 0.15	0.46 + 0.15	0.35 + 0.15
	3 (c_3)	0.60 − 0.15	0.46 − 0.15	0.35 − 0.15

The constants c_ℓ in $\beta_\ell = c_\ell/\log_{10}(\text{MAF})$ of power simulations, $\ell = 1,2,3$, are given in this table for two cases: (1) homogeneous genetic effect and (2) heterogeneous genetic effect.

Type I error simulation results: The empirical type I error rates are reported in Table 5 and Table 6. In Table 5, only rare variants were used to generate genotype data, but none of them relates to the trait. In Table 6, all variants were used to generate genotype data. For the GFLMs Hom-Rao and Het-Rao, all empirical type I error rates are below the nominal α levels for both B-spline and Fourier basis functions (columns 4–7 of Table 5 and Table 6). Therefore, the chi-squared-distributed Rao's efficient score statistics are very conservative and can be useful in whole-genome and whole-exome association studies.

For the GFLM Hom-LRT, all empirical type I error rates are around the nominal α levels for both B-spline and Fourier basis functions when all variants were used to generate genotype data (bottom parts of columns 8–11 of Table 6). For the GFLM Het-LRT, the empirical type I error rates are slightly higher than the nominal α levels when all variants were used to generate genotype data (top parts of columns 8–11 of Table 6), and the GFLM Het-LRT statistics can inflate type I error rates.

When only rare variants were used to generate genotype data, the empirical type I error rates are much higher than the nominal α levels for both B-spline and Fourier basis functions for GFLM Het-LRT statistics (top parts of columns 8–11 of Table 5). Relatively, the empirical type I error rates of GFLM Hom-LRT statistics are only slightly higher than the nominal α levels for both B-spline and Fourier basis functions (bottom parts of columns 8–11 of Table 5).

In Fan *et al.* (2014), it was found that the Rao's efficient score test statistics are very conservative when the sample is small or moderate from a single study (*i.e.*, the sample ranges

Table 5 Empirical type I error rates of Rao's efficient score test statistics and LRT statistics at different α levels based on 10^6 simulated data sets when only rare variants were used to generate genotype data

			Rao's efficient score test statistics of GFLMs				LRT statistics of GFLMs			
			Basis of both GVF and $\beta_t(t)$		Basis of β-smooth-only		Basis of both GVF and $\beta_t(t)$		Basis of β-smooth-only	
Type of test	Scenario	α Level	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier
GFLM Het-Rao	Scenario 1	0.05	0.040308	0.041429	0.040307	0.041429	0.107753	0.094461	0.107752	0.094468
or		0.01	0.006502	0.006787	0.006502	0.006787	0.029244	0.024788	0.029243	0.024790
GFLM Het-LRT		0.001	0.000442	0.000489	0.000442	0.000489	0.004305	0.003603	0.004305	0.003603
		0.0001	0.000024	0.000032	0.000024	0.000032	0.000606	0.000548	0.000606	0.000548
	Scenario 2	0.05	0.041607	0.042350	0.041607	0.042349	0.104186	0.090845	0.104186	0.090849
		0.01	0.006992	0.007279	0.006992	0.007279	0.027457	0.023561	0.027457	0.023563
		0.001	0.000543	0.000574	0.000543	0.000574	0.003972	0.003349	0.003972	0.003349
		0.0001	0.000032	0.000041	0.000032	0.000041	0.000570	0.000503	0.000570	0.000503
	Scenario 3	0.05	0.042585	0.043306	0.042585	0.043306	0.095901	0.085546	0.095895	0.085547
		0.01	0.007270	0.007443	0.007270	0.007443	0.024544	0.021358	0.024543	0.021359
		0.001	0.000568	0.000585	0.000568	0.000585	0.003431	0.002897	0.003431	0.002897
		0.0001	0.000041	0.000058	0.000041	0.000058	0.000443	0.000395	0.000443	0.000395
GFLM Hom-Rao	Scenario 1	0.05	0.046652	0.047087	0.047218	0.047167	0.058747	0.056513	0.058836	0.057511
or		0.01	0.008560	0.008799	0.008629	0.008713	0.012600	0.012024	0.012507	0.012261
GFLM Hom-LRT		0.001	0.000737	0.000828	0.000775	0.000794	0.001454	0.001364	0.001428	0.001366
		0.0001	0.000051	0.000070	0.000066	0.000064	0.000161	0.000144	0.000163	0.000166
	Scenario 2	0.05	0.047177	0.047423	0.047544	0.047543	0.058428	0.056090	0.058055	0.057137
		0.01	0.008855	0.008809	0.008910	0.008838	0.012636	0.011741	0.012536	0.012229
		0.001	0.000760	0.000761	0.000841	0.000783	0.001355	0.001198	0.001440	0.001289
		0.0001	0.000071	0.000079	0.000077	0.000069	0.000151	0.000160	0.000174	0.000146
	Scenario 3	0.05	0.048264	0.048039	0.048500	0.048655	0.056015	0.054683	0.053643	0.052776
		0.01	0.008940	0.008962	0.009228	0.009348	0.011540	0.011083	0.010987	0.010708
		0.001	0.000759	0.000769	0.000846	0.000839	0.001173	0.001099	0.001130	0.001081
		0.0001	0.000060	0.000064	0.000059	0.000084	0.000122	0.000111	0.000105	0.000116

The results of "Basis of both GVF and $\beta_t(t)$" were based on smoothing both the GVF and genetic-effect functions $\beta_t(t)$ of model 6, and the results of "Basis of β-smooth-only" were based on the smoothing $\beta_t(t)$ only approach of model 3. GVF, genetic variant function.

from 200 to 2000). Hence, the results of this paper are consistent with those of Fan *et al.* (2014) for the Rao's efficient score test statistics. When the sample is smaller than or equal to 2000, Fan *et al.* (2014) found that the LRT statistics inflate the type I error rates. In this paper, we have a very big sample size of 7000 by combining three studies for a unified analysis, and the GFLM Hom-LRT controls the type I error rates correctly, but the GFLM Het-LRT still may inflate the type I error rates.

In short, the chi-squared-distributed Rao's efficient score test statistics of GFLMs Hom-Rao and Het-Rao are very conservative. If the sample size is large, GFLM Hom-LRT statistics control the type I error rates well when all variants were used to generate genotype data and can slightly inflate the type I error rates when only rare variants were used to generate genotype data. The GFLM Het-LRT statistics may inflate the type I error rates, which may be due to the large degrees of freedom.

Statistical power results: We compared the power performance of the proposed tests with MetaSKAT based on the simulated COSI sequence data. The empirical power levels at the $\alpha = 0.0001$ level are plotted in Figure 1, Figure 2, Figure 3, Figure 4, Figure S1, Figure S2, Figure S3, and Figure S4. In all theses figures, "GVF&Beta, B-sp" (or "GVF&Beta, F-sp") means that both GVF and the genetic-effect function $\beta(t)$ were smoothed by B-spline (or Fourier) basis functions, and "Beta, B-sp" (or "Beta, F-sp") means that only the genetic-effect function $\beta(t)$ was smoothed by B-spline (or Fourier) basis functions (*i.e.*, β-smooth-only). Moreover, the results of Het-MetaSKAT, Het-MetaSKAT-O, Hom-MetaSKAT, and Hom-MetaSKAT-O using the R package MetaSKAT are reported for power comparison (Lee *et al.* 2013).

In Figure 1, Figure 2, Figure 3, and Figure 4, the results of GFLM Hom-Rao are reported, and the Rao's efficient score test statistics are constructed using the homogeneous effect model. In Figure S1, Figure S2, Figure S3, and Figure S4, the results of GFLM Het-Rao are reported, and the Rao's efficient score test statistics are constructed using the heterogeneous effect model. Moreover, the results of AEM Het-Rao for the additive-effect model (1) are reported in each figure for a comparison. In Figure 1, Figure 2, Figure S1, and Figure S2, the simulated data are generated under the assumption of homogeneous genetic effect, and in Figure 3, Figure 4, Figure S3, and Figure S4, the simulation data are generated under the assumption of heterogeneous genetic effect (Table 4).

When some causal variants are rare and some are common, the GFLM Hom-Rao has higher power than MetaSKAT and MetaSKAT-O for scenarios 1 and 2 and has similar power as MetaSKAT and MetaSKAT-O for scenario 3 in Figure 1 and Figure 3. The GFLM Het-Rao also has higher or similar power

Table 6 Empirical type I error rates of Rao's efficient score test statistics and LRT statistics at different α levels based on 10^6 simulated data sets when all variants were used to generate genotype data

			Rao's efficient score test statistics of GFLMs				LRT statistics of GFLMs			
			Basis of both GVF and $\beta_t(t)$		Basis of β-smooth-only		Basis of both GVF and $\beta_t(t)$		Basis of β-smooth-only	
Type of test	Scenario	α Level	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier	B-spline	Fourier
GFLM Het-Rao	Scenario 1	0.05	0.038922	0.040414	0.038922	0.040414	0.070972	0.061845	0.070972	0.061845
or		0.01	0.006645	0.007022	0.006645	0.007022	0.016140	0.013552	0.016140	0.013552
GFLM Het-LRT		0.001	0.000518	0.000573	0.000518	0.000573	0.001946	0.001558	0.001946	0.001558
		0.0001	0.000041	0.000056	0.000041	0.000056	0.000227	0.000177	0.000227	0.000177
	Scenario 2	0.05	0.040400	0.041604	0.040400	0.041604	0.069546	0.061199	0.069546	0.061199
		0.01	0.007057	0.007462	0.007057	0.007462	0.015841	0.013613	0.015841	0.013613
		0.001	0.000550	0.000665	0.000550	0.000665	0.001863	0.001560	0.001863	0.001560
		0.0001	0.000044	0.000041	0.000044	0.000041	0.000208	0.000193	0.000208	0.000193
	Scenario 3	0.05	0.040580	0.041450	0.040580	0.041450	0.064665	0.058161	0.064666	0.058161
		0.01	0.007031	0.007417	0.007031	0.007417	0.014029	0.012416	0.014029	0.012416
		0.001	0.000550	0.000637	0.000550	0.000637	0.001528	0.001390	0.001528	0.001390
		0.0001	0.000036	0.000050	0.000036	0.000050	0.000172	0.000156	0.000172	0.000156
GFLM Hom-Rao	Scenario 1	0.05	0.045807	0.045358	0.047004	0.046821	0.051164	0.048432	0.052932	0.051544
or		0.01	0.008605	0.008461	0.008918	0.008894	0.010409	0.009512	0.010960	0.010461
GFLM Hom-LRT		0.001	0.000842	0.000779	0.000852	0.000898	0.001107	0.000938	0.001167	0.001111
		0.0001	0.000078	0.000086	0.000069	0.000078	0.000119	0.000110	0.000111	0.000126
	Scenario 2	0.05	0.045847	0.045606	0.046779	0.046988	0.050780	0.048540	0.052439	0.051368
		0.01	0.008598	0.008638	0.008938	0.009000	0.010240	0.009649	0.010825	0.010428
		0.001	0.000768	0.000752	0.000799	0.000821	0.001023	0.000922	0.001080	0.001043
		0.0001	0.000074	0.000056	0.000077	0.000072	0.000103	0.000081	0.000132	0.000116
	Scenario 3	0.05	0.043291	0.043601	0.043397	0.043609	0.045600	0.045334	0.045182	0.044837
		0.01	0.008050	0.008195	0.008154	0.008286	0.008801	0.008745	0.008737	0.008715
		0.001	0.000742	0.000790	0.000709	0.000746	0.000873	0.000877	0.000797	0.000805
		0.0001	0.000072	0.000061	0.000070	0.000070	0.000084	0.000079	0.000079	0.000078

The results of "Basis of both GVF and $\beta_t(t)$" were based on smoothing both the GVF and genetic-effect functions $\beta_t(t)$ of model 6, and the results of "Basis of β-smooth-only" were based on the smoothing $\beta_t(t)$ only approach of model 3. GVF, genetic variant function.

as MetaSKAT and MetaSKAT-O in Figure S1 and Figure S3. Therefore, the proposed Rao's efficient score test statistics have good power performance when some causal variants are rare and some are common. By a comparison of power levels in Figure 1 *vs.* Figure S1 and Figure 3 *vs.* Figure S3, the power levels of the GFLM Hom-Rao are generally higher than those of GFLM Het-Rao, which may be due to the large degrees of freedom of the GFLM Het-Rao. The AEM Het-Rao has slightly lower power levels than GFLM Hom-Rao and GFLM Het-Rao but performs well.

When the causal variants are all rare, the GFLM Hom-Rao has slightly lower or similar power levels as MetaSKAT and MetaSKAT-O in Figure 2, Figure 4, Figure S2, and Figure S4. Again, the power levels of the GFLM Hom-Rao in Figure 2 and Figure 4 are generally higher than the corresponding power levels of GFLM Het-Rao in Figure S2 and Figure S4. The AEM Het-Rao has low power levels.

In each graph, we compared five Rao's efficient score test statistics: one is based on the additive-effect model (1), two are based on B-spline basis functions, and two are based on Fourier basis functions. In the two Rao's efficient score test statistics to use B-spline (or Fourier) basis functions, one is to smooth both the GVFs and the genetic-effect function $\beta(t)$, and the other is only to smooth the genetic-effect function $\beta(t)$ (*i.e.*, β-smooth-only). The four Rao's efficient score test statistics of the GFLMs have similar power. The power levels of β-smooth-only are almost identical to those of smoothing both the GVFs and genetic-effect function $\beta(t)$ by B-spline basis (or Fourier basis). Thus, the tests do not strongly depend on whether the genotype data are smoothed or not. In addition, the Rao's efficient score test statistics do not strongly depend on which basis functions are used. We also calculated the empirical power levels of the LRT statistics, which provide very similar empirical power levels as the Rao's efficient score test statistics (data not shown).

Discussion

In this paper, GFLMs are developed to perform a meta-analysis of multiple case-control studies to connect genetic data to dichotomous traits adjusting for covariates. Based on the GFLMs, chi-squared-distributed Rao's efficient score test and LRT statistics are introduced to test for an association between a complex trait and multiple genetic variants. Extensive simulations are performed to evaluate empirical type I error rates and the power performance of the proposed GFLMs and tests. We show that the proposed Rao's efficient score test statistics are very conservative. The Rao's efficient score test statistics have higher power than MetaSKAT when some causal variants are rare and some

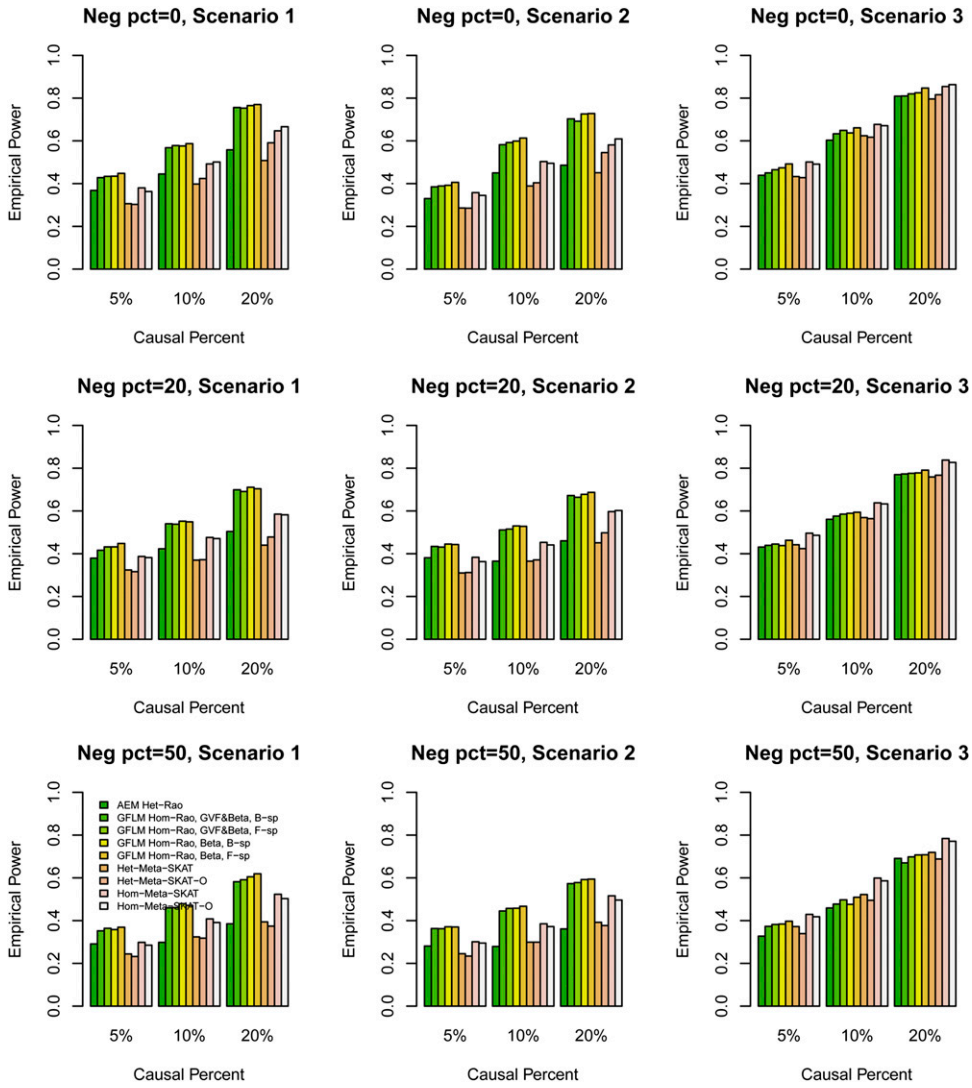

Figure 1 The empirical power of the GFLM Hom-Rao of models (3) and (6) as well as the AEM Het-Rao of the additive-effect model (1) and MetaSKAT at $\alpha = 0.0001$ when some causal variants are rare and some are common and the genetic effect is simulated as homogeneous. When Neg pct = 0, all causal variants had positive effects; when Neg pct = 20, 20/80% of causal variants had negative/positive effects; when Neg pct = 50, 50/50% of causal variants had negative/positive effects.

are common. When the causal variants are all rare (*i.e.*, MAF < 0.03), the Rao's efficient score test statistics have similar or slightly lower power than MetaSKAT. For homogeneous genetic effect models, the GFLM Hom-LRT generates accurate type I error rates. For heterogeneous genetic models, the GFLM Het-LRT may inflate type I error rates owing to large degrees of freedom. The GFLMs and related test statistics can be useful in whole-genome and whole-exome association studies.

The GFLMs and AEM were applied to analyze the genetic data of 22 gene regions of T2D data from a meta-analysis of eight European studies and detected significant associations for 19 genes ($P < 3.1 \times 10^{-6}$), tentative association for 1 gene ($P \approx 10^{-5}$), and no association for 2 genes, while MetaSKAT detected none. Because the 22 genes are from the literature on T2D showing that each of them contains SNPs that are associated with T2D, the association is confirmed by our fixed models and related tests for the 19 genes, although MetaSKAT failed to confirm any of the associations. One may note that the European cohorts were analyzed by MetaSKAT in Lee *et al.* (2013), but no results were reported for the dichotomous traits of T2D.

Unlike other methods such as SKAT or MetaSKAT, which are based on mixed-effect models, GFLMs are fixed-effect models, and the genetic effects of multiple genetic variants are assumed to be fixed. The formulation of the β-smooth-only model (2) is similar to that of SKAT and MetaSKAT. However, the assumptions are totally different. Specifically, the regression coefficients β_ℓ of genetic variant terms in the models of SKAT and MetaSKAT are random, while the genetic effects $\beta_\ell(t_{\ell j})$ in model 2 are fixed unknown functions. Our GFLMs are a natural extension of traditional population genetics without a polygenic term because we consider the population data. By using functional data-analysis techniques, we develop procedures to estimate the genetic-effect functions $\beta_\ell(t)$ and introduce test statistics to test for an association.

If the causal genetic variants are all rare, the number of causal rare variants is large, and each contributes a small amount to the trait, it would be reasonable to assume the genetic

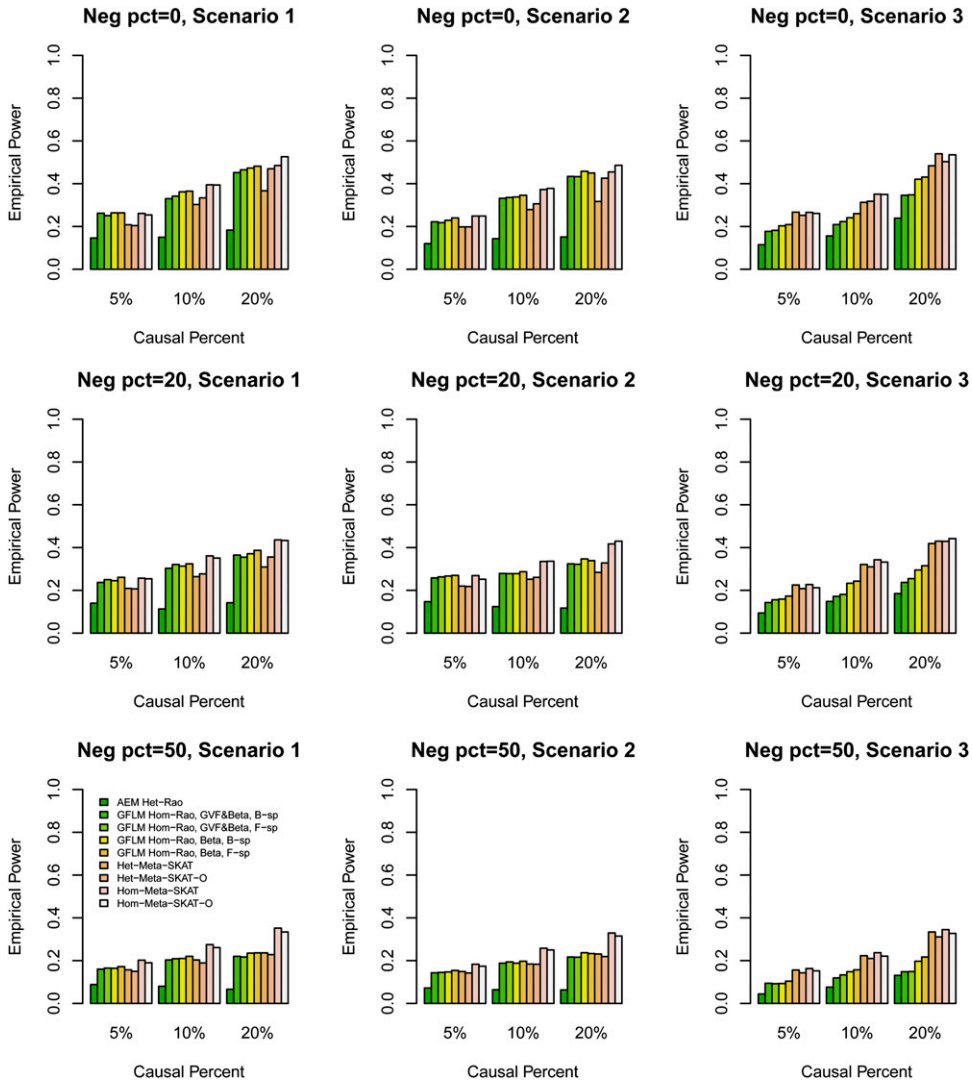

Figure 2 The empirical power of the GFLM Hom-Rao of models (3) and (6) as well as the AEM Het-Rao of the additive-effect model (1) and MetaSKAT at $\alpha = 0.0001$ when all causal variants are rare and the genetic effect is simulated as homogeneous. When Neg pct = 0, all causal variants had positive effects; when Neg pct = 20, 20/80% of causal variants had negative/positive effects; when Neg pct = 50, 50/50% of causal variants had negative/positive effects.

contribution of major gene loci to be random, and then the mixed models of SKAT and MetaSKAT can be valid. In our power comparison, we found that the proposed Rao's efficient score test statistics have similar or slightly lower power than MetaSKAT when the causal variants are all rare. However, the proposed Rao's efficient score test statistics have higher power than MetaSKAT when some causal variants are rare and some are common (in this case, it is likely that the effects of a few genetic variants of the major gene locus are large, so fixed-effect models perform well). It is noteworthy that this paper deals with dichotomous traits. For quantitative traits, it was found that functional linear models lead to both *F*- and chi-squared-distributed score test statistics that are more powerful than SKAT and MetaSKAT (Luo *et al.* 2012; Fan *et al.* 2013).

In the proposed models and tests, we do not make any assumptions about whether the genetic variants are rare or common variants or a combination of the two. The proposed models are very flexible and can analyze rare or common variants or a combination of the two. We do assume that the number of genetic variants in a genetic region is large, which is true for modern genetic data. When a large number of genetic variants are available in a genetic region, estimation of the GVF is accurate, which makes our GFLMs very reliable. In our simulation and data analysis, models 2 and 4 perform very close to each other.

In Fan *et al.* (2013, 2014), we investigated the performance of the mixed models by making the regression coefficients β of genetic-effect function random in the frameworks of our functional regression models. It was found that the mixed models perform well only when the causal genetic variants are all rare and the traits are dichotomous (for rare variants, we used an artificial cutoff of 0.03). For most diseases, the causal variants can be both rare and common. Because the proposed models are very flexible in analyzing rare or common variants, we focus on fixed-effect models in this paper. In our simulations, we treat the regression effect of covariates as heterogeneous. We also investigate the performance of the proposed models by treating the regression effect of covariates as homogeneous, and we find that the results are similar in terms of empirical power performance and type I error rates.

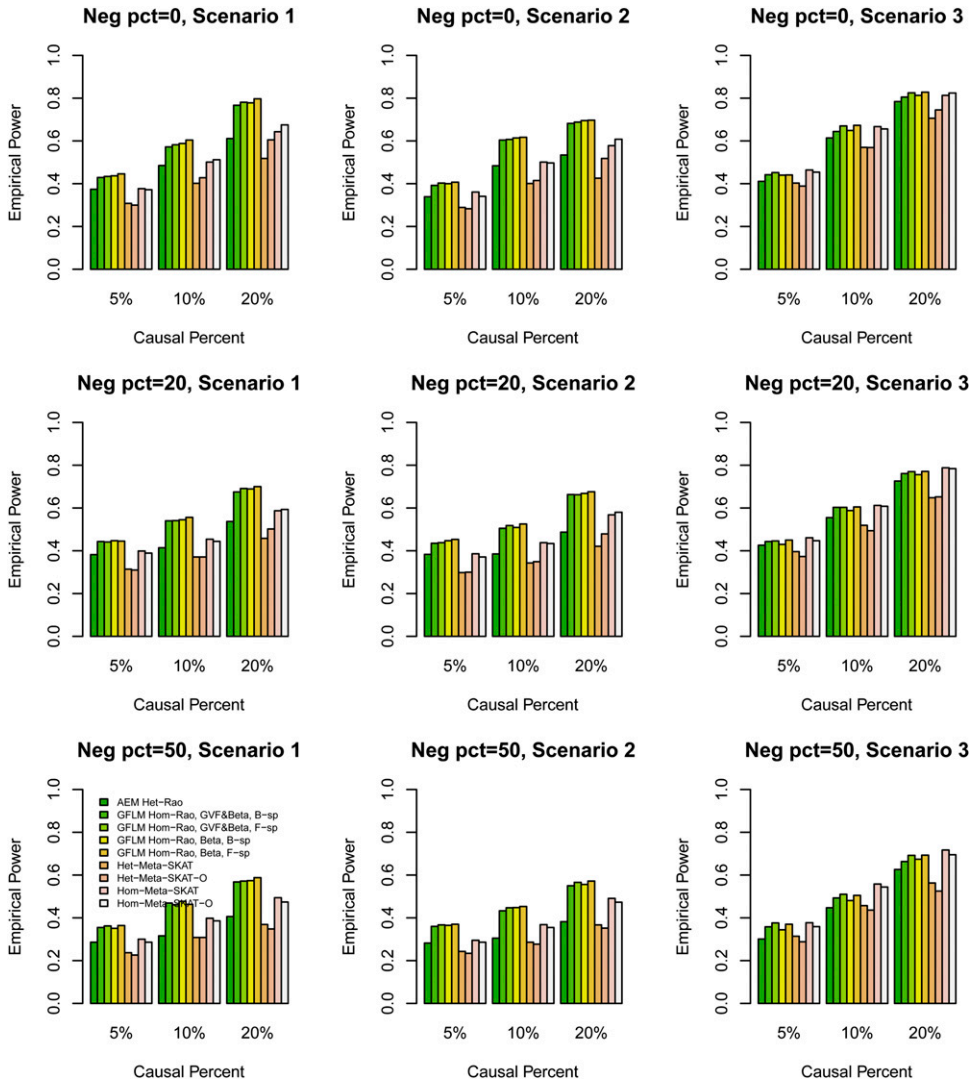

Figure 3 The empirical power of the GFLM Hom-Rao of models (3) and (6) as well as the AEM Het-Rao of the additive-effect model (1) and MetaSKAT at $\alpha = 0.0001$ when some causal variants are rare and some are common and the genetic effect is simulated as heterogeneous. When Neg pct = 0, all causal variants had positive effects; when Neg pct = 20, 20/80% of causal variants had negative/positive effects; when Neg pct = 50, 50/50% of causal variants had negative/positive effects.

For small- and moderate-sample-size single studies when the sample sizes are smaller than or equal to 2000, the LRT statistics of GFLMs were found to inflate the type I error rates, while chi-squared-distributed Rao's efficient score test statistics control type I error rates correctly (Fan *et al.* 2014). Hence, Rao's efficient score test statistics were recommended for small- and moderate-sample-size single studies. In this paper, we show that Rao's efficient score test statistics control the type I error rates correctly when the sample sizes of combined multiple studies are large. For homogeneous genetic-effect models, the LRT statistics were found to have correct type I error rates; for heterogeneous genetic-effect models, the LRT statistics inflated the type I error rates. Therefore, one needs to be cautious about using LRT statistics for dichotomous traits. For quantitative traits, both the LRT and *F*-distributed statistics have correct type I error rates and good power performance for a sample with a sample size $\geq$ 1500 (Fan *et al.* 2013).

The proposed method requires individual genotype data and is more powerful than MetaSKAT and MetaSKAT-O when genotype data are available from all studies. However, the proposed method cannot meta-analyze summary statistics. If only summary statistics of GFLMs are available from different studies using Fan *et al.* (2014), it is still an open question as to how to use them for a meta-analysis.

Acknowledgments

Two anonymous reviewers and the editors, Dr. Chiara Sabatti and Dr. Gary Churchill, provided very good and insightful comments for us to improve the manuscript. We greatly thank the European cohorts groups for letting us analyze the data and use them as examples. Dr. Heather M. Stringham and Dr. Tanya M. Teslovich kindly sent us the data of the European cohorts and patiently answered many questions about the cohorts, and we greatly appreciated them. This study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (http://biowulf.nih.gov). No GWAS data were generated in this paper. This study was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and

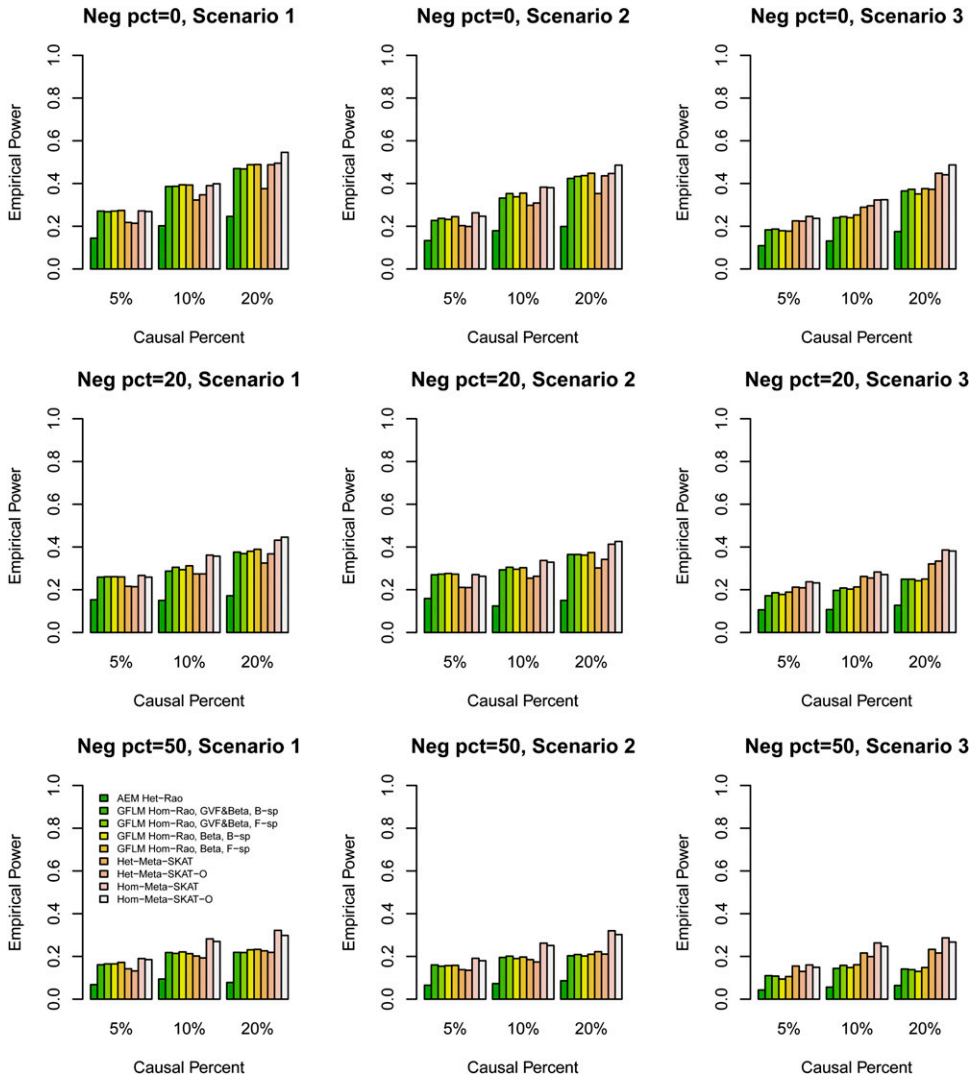

Figure 4 The empirical power of the GFLM Hom-Rao of models (3) and (6) as well as the AEM Het-Rao of the additive-effect model (1) and MetaSKAT at $\alpha = 0.0001$ when all causal variants are rare and the genetic effect is simulated as heterogeneous. When Neg pct = 0, all causal variants had positive effects; when Neg pct = 20, 20/80% of causal variants had negative/positive effects; when Neg pct = 50, 50/50% of causal variants had negative/positive effects.

Human Development, National Institutes of Health (R.F., Y.W., and C.-y.C.), by NIH grants R01-EY024226 and R01-HG007358 (to W.C.), by the University of Pittsburgh (R.F. is an unpaid collaborator on grant R01-EY024226), by NIH grants R01-HG006292 and R01-HG006703 (to Y.L.), and by NIH grants LM-009012 and LM-010098 (to J.H.M).

Literature Cited

Cordell, H. J., and D. G. Clayton, 2002 A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am. J. Hum. Genet. 70: 124–141.

de Boor, C., 2001 *Applied Mathematical Sciences 27: A Practical Guide to Splines*, Rev. Ed. Springer, New York.

Evangelou, E., and J. P. A. Ioannidis, 2013 Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. 14: 379–389.

Fan, R. Z., Y. F. Wang, J. L. Mills, A. F. Wilson, J. E. Bailey-Wilson *et al.*, 2013 Functional linear models for association analysis of quantitative traits. Genet. Epidemiol. 37: 726–742.

Fan, R. Z., Y. F. Wang, J. L. Mills, T. C. Carter, I. Lobach *et al.*, 2014 Generalized functional linear models for case-control association studies. Genet. Epidemiol. 38: 622–637.

Fan, R. Z., Y. F. Wang, M. Boehnke, W. Chen, Y. Li *et al.*, 2015 Gene level meta-analysis of quantitative traits by functional linear models. Genetics 200: 1089–1104.

Fan, R. Z., Y. F. Wang, Y. Qi, Y. Ding, D. E. Weeks *et al.*, 2016 Gene-based association analysis for censored traits via functional regressions. Genet. Epidemiol. (in press).

Ferraty, F., and Y. Romain, 2010 *Oxford Handbook of Functional Data Analysis*. Oxford University Press, New York.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Philos. Trans. R. Soc. Edinb. 52: 399–433.

Han, F., and W. A. Pan, 2010 Data-adaptive sum test for disease association with multiple common or rare variants. Hum. Hered. 70: 42–54.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106: 9362–9367.

Horváth, L., and P. Kokoszka, 2012 Inference for Functional Data *w*ith Applications. Springer, New York.

Hu, Y. J., S. I. Berndt, S. Gustafsson, A. Ganna Genetic Investigation of Anthropometric Traits (GIANT) Consortium*et al.*, 2013 Meta-analysis of gene-level associations for rare variants based on single-variant statistics. Am. J. Hum. Genet. 93: 42–53.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder *et al.*, 2012 Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. 91: 224–237.

Lee, S., T. M. Teslovich, M. Boehnke, and X. Lin, 2013 General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet. 93: 42–53.

Li, B., and S. M. Leal, 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83: 311–321.

Li, S., B. Mukherjee, J. M. G. Taylor, K. M. Rice, X. Wen *et al.*, 2014 The role of environmental heterogeneity in meta-analysis of gene-environment interactions with quantitative traits. Genet. Epidemiol. 38: 416–429.

Liu, D. J., G. M. Peloso, X. Zhan, O. L. Holmen, M. Zawistowski *et al.*, 2014 Meta-analysis of gene-level tests for rare variant association. Nat. Genet. 46: 200–204.

Luo, L., E. Boerwinkle, and M. Xiong, 2011 Association studies for next-generation sequencing. Genome Res. 21: 1099–1108.

Luo, L., Y. Zhu, and M. Xiong, 2012 Quantitative trait locus analysis for next-generation sequencing with the functional linear models. J. Med. Genet. 49: 513–524.

Luo, L., Y. Zhu, and M. Xiong, 2013 Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. Eur. J. Hum. Genet. 21: 217–224.

Madsen, B. E., and S. R. Browning, 2009 A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5: e1000384.

Morris, A. P., and E. Zeggini, 2010 An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 34: 188–193.

Morris, A. P., B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segre *et al.*, 2012 Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44: 981–990.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin *et al.*, 2011 Testing for an unusual distribution of rare variants. PLoS Genet. 7: e1001322.

Price, A. L., G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. 86: 832–838.

Ramsay, J. O., and B. W. Silverman, 2005 *Functional Data Analysis*, Ed. 2. Springer, New York.

Ramsay, J. O., G. Hooker, and S. Graves, 2009 *Functional Data Analysis with R and Matlab*. Springer, New York.

Ross, S. M., 1996 *Stochastic Processes*, Ed. 2. Wiley, New York.

Schaffner, S. F., C. Foo, S. Gabriel, D. Reich, M. J. Daly *et al.*, 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15: 1576–1583.

Scott, R. A., V. Lagou, R. P. Welch, E. Wheeler, M. E. Montasser *et al.*, 2012 Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat. Genet. 44: 991–1005.

Stahl, E. A., S. Raychaudhuri, E. F. Remmers, G. Xie, S. Eyre *et al.*, 2010 Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42: 508–514.

Voight, B. F., L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina *et al.*, 2010 Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. 42: 579–589.

Vsevolozhskaya, O. A., D. V. Zaykin, M. C. Greenwood, C. Wei, and Q. Lu, 2014 Functional analysis of variance for association studies. PLoS One 9: e105074.

Wang, Y. F., A. Y. Liu, J. L. Mills, M. Boehnke, A. F. Wilson *et al.*, 2015 Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genet. Epidemiol. 39: 259–275.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89: 82–93.

Zeggini, E., and J. P. A. Ioannidis, 2009 Meta-analysis in genome-wide association studies. Pharmacogenomics 10: 191–201.

Zeggini, E., L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini *et al.*, 2008 Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet. 40: 638–645.

Zhang, F., E. Boerwinkle, and M. Xiong, 2014 Epistasis analysis for quantitative traits by functional regression models. Genome Res. 24: 989–998.

Communicating editor: C. Sabatti

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.180869/-/DC1

Meta-analysis of Complex Diseases at Gene Level with Generalized Functional Linear Models

Ruzong Fan, Yifan Wang, Chi-yang Chiu, Wei Chen, Haobo Ren, Yun Li, Michael Boehnke, Christopher I. Amos, Jason H Moore, and Momiao Xiong

DOI: 10.1534/genetics.115.180869

Supporting Information: "Meta-analysis of Complex Diseases at Gene Level by Generalized Functional Linear Models"

Information Of the Eight European Cohorts

For the eight European cohorts, we performed association analysis between T2D and 22 genes. The sample sizes of each study are presented in Table S.1. The information of the 22 genes is given in Table S.2. The results of association analysis by Het-LRT and Hom-LRT are reported in Tables S.3 and S.4.

Table S.1: **Sample sizes of the cases and controls for each of the seven studies.**

Study	# of Cases	# of Controls	Total
D2d-2007	281	1794	2075
DIAGEN	429	1042	1471
DPS	193	219	412
DRs EXTRA	108	1049	1157
FUSION Stage 2	806	1694	2500
METSIM	572	774	1346
Norway	1143	1347	2490
Total	3532	7919	11451

Table S.2: **Summary of 22 genes and the number of genetic variants in each gene region by Mar. 2006 (NCBI36/hg18).** The number of variants is the number of genetic variants in a region of **Start (-5Kb) - End (+5Kb) Positions**. * The gene region of *PCSK9* is (55277737, 55303114), and (55271537, 55286109) is the region in the database. # The Length is the length of the region in bp.

Gene	**Chromosome Region**	**Gene Positions (bp)**	**Start (-5Kb) - End (+5Kb) Positions (Length#)**	**Number of Variants**
*PCSK9**	1	55277737 - 55303114	55271537 - 55286109 (14572)	74
APOB	2	21077806 - 21120450	21072806 - 21125450 (52644)	223
IGF2BP2	3	186844221 - 187025521	186839221 - 187030521 (191300)	231
CDKAL1	6	20642667 - 21340613	20637667 - 21345613 (707946)	560
JAZF1	7	27836718 - 28186962	27831718 - 28191962 (360244)	384
LPL	8	19840862 - 19869050	19835862 - 19874050 (38188)	212
CDKN2B	9	21992902 - 21999312	21987902 - 22004312 (16410)	64
CDC123	10	12277971 - 12332593	12272971 - 12337593 (64622)	265
IDE	10	94201421 - 94323832	94196421 - 94328832 (132411)	327
KIF11	10	94342805 - 94405132	94337805 - 94410132 (72327)	216
HHEX	10	94439661 - 94445388	94434661 - 94450388 (15727)	30
TCF7L2	10	114699999 - 114917426	114694999 - 114922426 (227427)	258
KCNQ1	11	2422797 - 2826916	2417797 - 2831916 (414119)	660
MTNR1B	11	92342437 - 92355596	92337437 - 92360596 (23159)	106
HMGA2	12	64504507 - 64646338	64499507 - 64651338 (151831)	214
TSPAN8	12	69805144 - 69838046	69800144 - 69843046 (42902)	54
HNF1A	12	119900932 - 119924697	119895932 - 119929697 (33765)	71
OASL	12	119942478 - 119961428	119937478 - 119966428 (28950)	108
FTO	16	52295376 - 52705882	52290376 - 52710882 (420506)	191
LDLR	19	11061038 - 11105505	11056038 - 11110505 (54467)	43
APOE	19	50100879 - 50104490	50095879 - 50109490 (13611)	35
GIPR	19	50863342 - 50877557	50858342 - 50882557 (24215)	37

Table S.3: **Association analysis of type 2 diabetes status in eight European cohorts by heterogeneous likelihood ratio test statistics (Het-LRT), Het-Meta-SKAT-O, and Het-Meta-SKAT.** The associations that attain a threshold significance of p-value $< 3.1 \times 10^{-6}$ are marked by red. The results of "Basis of both GVF and $\beta_\ell(t)$" were based on smoothing both GVF and genetic effect functions $\beta_\ell(t)$ of model (6), and the results of "Basis of β-smooth only" were based on smoothing $\beta_\ell(t)$ only approach of model (3), and the p-values of Het-Meta-SKAT and Het-Meta-SKAT-O were based of R package MetaSKAT. Abbreviation: GVF = genetic variant function.

Gene	p-values of the Het-LRT					p-values of Het-Meta	
	Basis of both GVF and $\beta_\ell(t)$		Basis of β-smooth only		Additive Effect Model (1)		
	B-spline	Fourier	B-spline	Fourier		SKAT	SKAT-O
PCSK9	2.20×10^{-11}	3.23×10^{-11}	2.20×10^{-11}	3.23×10^{-11}	6.50×10^{-7}	0.792	0.059
APOB	2.18×10^{-24}	1.98×10^{-22}	2.18×10^{-24}	1.98×10^{-22}	5.40×10^{-23}	0.499	0.517
IGF2BP2	4.11×10^{-9}	1.26×10^{-11}	4.11×10^{-9}	1.26×10^{-11}	2.81×10^{-24}	0.531	0.503
CDKAL1	1.11×10^{-20}	1.14×10^{-22}	1.11×10^{-20}	1.14×10^{-22}	2.06×10^{-20}	0.961	0.800
JAZF1	7.44×10^{-31}	6.98×10^{-29}	7.44×10^{-31}	6.98×10^{-29}	2.24×10^{-21}	0.032	0.046
LPL	3.11×10^{-5}	1.47×10^{-8}	3.11×10^{-5}	1.47×10^{-8}	3.45×10^{-19}	0.590	0.795
CDKN2B	8.94×10^{-41}	1.73×10^{-31}	8.94×10^{-41}	1.73×10^{-31}	3.31×10^{-31}	0.554	0.410
CDC123	3.37×10^{-20}	3.33×10^{-19}	3.37×10^{-20}	3.33×10^{-19}	3.12×10^{-23}	0.039	0.072
IDE	1.10×10^{-22}	3.18×10^{-24}	1.10×10^{-22}	3.18×10^{-24}	3.92×10^{-24}	0.414	0.630
KIF11	2.74×10^{-24}	1.03×10^{-24}	2.74×10^{-24}	1.03×10^{-24}	6.07×10^{-29}	0.768	0.913
HHEX	1	1.38×10^{-5}	1	1.38×10^{-5}	1.11×10^{-6}	0.480	0.691
TCF7L2	7.39×10^{-11}	5.86×10^{-10}	7.39×10^{-11}	5.86×10^{-10}	3.51×10^{-8}	0.021	0.042
KCNQ1	5.15×10^{-33}	8.89×10^{-31}	5.15×10^{-33}	8.89×10^{-31}	8.44×10^{-31}	0.572	0.797
MTNR1B	3.93×10^{-18}	5.82×10^{-16}	3.93×10^{-18}	5.82×10^{-16}	5.56×10^{-17}	0.295	0.456
HMGA2	9.19×10^{-6}	1.14×10^{-4}	9.19×10^{-6}	1.14×10^{-4}	2.36×10^{-3}	0.699	0.887
TSPAN8	3.50×10^{-43}	3.38×10^{-41}	4.60×10^{-43}	5.08×10^{-41}	1.17×10^{-40}	0.747	0.923
HNF1A	2.28×10^{-17}	2.04×10^{-16}	2.28×10^{-17}	2.04×10^{-16}	2.42×10^{-30}	0.272	0.441
OASL	1.76×10^{-37}	2.03×10^{-30}	1.76×10^{-37}	2.03×10^{-30}	7.35×10^{-30}	0.530	0.416
FTO	6.79×10^{-27}	9.53×10^{-28}	6.79×10^{-27}	9.53×10^{-28}	8.53×10^{-29}	0.048	0.090
LDLR	0.340	0.327	0.340	0.327	0.196	0.233	0.400
APOE	7.76×10^{-34}	9.91×10^{-30}	7.76×10^{-34}	9.91×10^{-30}	3.45×10^{-32}	0.042	0.082
GIPR	3.43×10^{-3}	1.81×10^{-3}	3.43×10^{-3}	1.81×10^{-3}	0.004	0.808	0.303

Table S.4: **Association analysis of type 2 diabetes status in eight European cohorts by homogeneous likelihood ratio test statistics (Hom-LRT), Hom-Meta-SKAT-O, and Hom-Meta-SKAT.** The associations that attain a threshold significance of p-value $< 3.1 \times 10^{-6}$ are marked by red. The results of "Basis of both GVF and $\beta_\ell(t)$" were based on smoothing both GVF and genetic effect functions $\beta_\ell(t)$ of model (6), and the results of "Basis of β-smooth only" were based on smoothing $\beta_\ell(t)$ only approach of model (3), and the p-values of Hom-Meta-SKAT and Hom-Meta-SKAT-O were based of R package MetaSKAT. Abbreviation: GVF = genetic variant function.

Gene	p-values of the Hom-LRT					p-values of Hom-Meta	
	Basis of both GVF and $\beta_\ell(t)$		Basis of β-smooth only		Additive Effect Model (1)		
	B-spline	Fourier	B-spline	Fourier		SKAT	SKAT-O
PCSK9	0.082	0.034	0.185	0.233	0.681	0.063	0.025
APOB	0.035	0.081	0.012	0.020	0.724	0.807	0.623
IGF2BP2	0.017	4.69×10^{-3}	1.22×10^{-3}	2.54×10^{-4}	0.013	0.417	0.368
CDKAL1	0.192	0.216	0.058	0.083	1	0.473	0.646
JAZF1	0.447	0.425	0.201	0.305	0.042	0.352	0.094
LPL	0.076	0.013	0.081	0.011	0.018	0.416	0.559
CDKN2B	1.50×10^{-3}	5.11×10^{-5}	6.18×10^{-3}	0.015	0.070	0.325	0.430
CDC123	0.038	0.026	0.076	0.071	0.008	0.129	0.210
IDE	0.242	0.142	0.155	0.310	0.055	0.252	0.389
KIF11	0.041	0.037	0.065	0.184	0.364	0.667	0.802
HHEX	0.014	7.33×10^{-4}	0.016	0.026	0.243	0.684	0.711
TCF7L2	3.54×10^{-14}	6.11×10^{-14}	2.69×10^{-16}	1.47×10^{-15}	5.70×10^{-8}	1.37×10^{-4}	3.03×10^{-4}
KCNQ1	0.060	0.141	0.105	0.141	0.008	0.420	0.601
MTNR1B	6.76×10^{-4}	4.69×10^{-4}	0.012	4.90×10^{-9}	0.198	0.523	0.641
HMGA2	0.759	0.912	0.671	0.904	0.269	0.880	1
TSPAN8	0.450	1.26×10^{-5}	2.42×10^{-3}	3.91×10^{-4}	0.785	0.991	0.836
HNF1A	0.133	0.046	0.085	7.79×10^{-3}	0.087	0.661	0.363
OASL	0.074	0.026	0.031	0.030	0.264	0.477	0.305
FTO	7.08×10^{-4}	1.84×10^{-4}	3.84×10^{-6}	2.17×10^{-7}	0.048	0.291	0.428
LDLR	0.950	0.916	0.932	0.932	0.849	0.876	0.727
APOE	0.450	0.152	0.024	4.34×10^{-3}	0.048	0.038	0.070
GIPR	0.060	0.038	0.011	0.088	0.020	0.306	0.250

Table S.5: **Association analysis of type 2 diabetes status in eight European cohorts by heterogeneous Rao's efficient score test statistics (Het-Rao), Het-Meta-SKAT-O, and Het-Meta-SKAT, Using Rare Variants (MAFs≤0.03).** The associations that attain a threshold significance of p-value $< 3.1 \times 10^{-6}$ are marked by red. The results of "Basis of both GVF and $\beta_\ell(t)$" were based on smoothing both GVF and genetic effect functions $\beta_\ell(t)$ of model (6), and the results of "Basis of β-smooth only" were based on smoothing $\beta_\ell(t)$ only approach of model (3), and the p-values of Het-Meta-SKAT and Het-Meta-SKAT-O were based of R package MetaSKAT. Abbreviation: GVF = genetic variant function.

Gene	Number of Rare Variants (MAFs≤0.03)	p-values of the Het-Rao					p-values of Het-Meta	
		Basis of both GVF and $\beta_\ell(t)$		Basis of β-smooth only		Additive Effect Model		
		B-spline	Fourier	B-spline	Fourier		SKAT	SKAT-O
PCSK9	42	0.669	0.356	0.999	0.283	0.448	0.590	0.198
APOB	175	0.198	0.200	0.198	0.200	0.516	0.412	0.542
IGF2BP2	157	0.240	0.393	0.240	0.393	0.270	0.530	0.468
CDKAL1	323	0.092	0.310	0.092	0.310	0.164	0.933	0.549
JAZF1	252	0.490	0.134	0.490	0.134	0.525	0.034	0.062
LPL	140	0.963	0.946	0.966	0.946	0.758	0.612	0.509
CDKN2B	40	0.692	0.545	0.676	0.648	0.726	0.592	0.805
CDC123	178	0.726	0.857	0.726	0.886	0.874	0.149	0.246
IDE	208	0.402	0.387	0.402	0.387	0.614	0.548	0.703
KIF11	134	0.596	0.598	0.579	0.578	0.622	0.772	0.688
HHEX	25	0.811	0.477	0.731	0.775	0.731	0.421	0.613
TCF7L2	170	1.02×10^{-3}	1.60×10^{-3}	1.02×10^{-3}	1.60×10^{-3}	0.036	0.122	0.217
KCNQ1	356	0.267	0.251	0.267	0.340	0.037	0.248	0.419
MTNR1B	83	0.286	0.376	0.320	0.482	0.324	0.485	0.684
HMGA2	164	0.330	0.399	0.330	0.399	0.385	0.440	0.657
TSPAN8	46	0.700	0.660	0.853	0.660	0.737	0.688	0.885
HNF1A	24	0.485	0.698	0.371	0.390	0.371	0.662	0.796
OASL	67	0.794	0.794	0.767	0.817	0.841	0.765	0.411
FTO	75	0.571	0.520	0.605	0.541	0.593	0.298	0.464
LDLR	31	0.714	0.763	0.714	0.763	0.714	0.177	0.309
APOE	21	0.188	0.151	0.188	0.151	0.188	0.018	0.022
GIPR	25	0.671	0.357	0.671	0.417	0.671	0.884	1

Table S.6: **Association analysis of type 2 diabetes status in eight European cohorts by heterogeneous Rao's efficient score test statistics (Het-Rao), Het-Meta-SKAT-O, and Het-Meta-SKAT, using common variants (MAFs>0.03).** The associations that attain a threshold significance of p-value $< 3.1 \times 10^{-6}$ are marked by red. The results of "Basis of both GVF and $\beta_\ell(t)$" were based on smoothing both GVF and genetic effect functions $\beta_\ell(t)$ of model (6), and the results of "Basis of β-smooth only" were based on smoothing $\beta_\ell(t)$ only approach of model (3), and the p-values of Het-Meta-SKAT and Het-Meta-SKAT-O were based of R package MetaSKAT. Abbreviation: GVF = genetic variant function.

Gene	Number of Common Variants (MAFs>0.03)	p-values of the Het-Rao					p-values of Het-Meta	
		Basis of both GVF and $\beta_\ell(t)$		Basis of β-smooth only		Additive Effect Model		
		B-spline	Fourier	B-spline	Fourier		SKAT	SKAT-O
PCSK9	32	1.38×10^{-14}	8.83×10^{-15}	1.38×10^{-14}	8.83×10^{-15}	3.38×10^{-8}	0.821	0.115
APOB	48	2.09×10^{-28}	2.52×10^{-26}	2.09×10^{-28}	2.52×10^{-26}	4.24×10^{-21}	0.803	0.783
IGF2BP2	74	4.21×10^{-6}	8.42×10^{-10}	4.21×10^{-6}	8.42×10^{-10}	1.14×10^{-20}	0.484	0.716
CDKAL1	237	4.86×10^{-26}	1.67×10^{-27}	4.86×10^{-26}	1.67×10^{-27}	5.05×10^{-11}	0.845	0.424
JAZF1	132	1.89×10^{-27}	6.18×10^{-27}	1.89×10^{-27}	6.18×10^{-27}	2.25×10^{-14}	0.266	0.132
LPL	72	2.75×10^{-6}	4.36×10^{-8}	2.75×10^{-6}	4.36×10^{-8}	3.22×10^{-15}	0.320	0.144
CDKN2B	24	5.09×10^{-30}	1.06×10^{-36}	5.09×10^{-30}	1.06×10^{-36}	8.83×10^{-31}	0.378	0.290
CDC123	87	2.11×10^{-18}	2.33×10^{-19}	2.11×10^{-18}	2.33×10^{-19}	4.99×10^{-18}	0.047	0.085
IDE	119	3.11×10^{-23}	5.67×10^{-19}	3.11×10^{-23}	5.67×10^{-19}	7.34×10^{-22}	0.214	0.142
KIF11	82	5.29×10^{-31}	1.17×10^{-33}	5.29×10^{-31}	1.17×10^{-33}	9.60×10^{-27}	0.472	0.579
HHEX	5	1.62×10^{-4}	5.36×10^{-8}	1.68×10^{-6}	5.36×10^{-8}	8.60×10^{-7}	0.545	0.740
TCF7L2	88	1.09×10^{-7}	3.40×10^{-10}	1.09×10^{-7}	3.40×10^{-10}	1.08×10^{-5}	9.16×10^{-3}	2.31×10^{-4}
KCNQ1	304	1.75×10^{-26}	1.61×10^{-25}	5.67×10^{-27}	1.63×10^{-27}	9.27×10^{-9}	0.680	0.870
MTNR1B	23	8.08×10^{-13}	4.47×10^{-15}	8.08×10^{-13}	4.47×10^{-15}	3.72×10^{-17}	1.67×10^{-3}	3.27×10^{-3}
HMGA2	50	2.85×10^{-3}	4.55×10^{-3}	2.85×10^{-3}	4.55×10^{-3}	0.109	0.785	1.000
TSPAN8	8	4.84×10^{-7}	2.56×10^{-11}	1.48×10^{-11}	2.56×10^{-11}	1.58×10^{-11}	0.633	0.758
HNF1A	47	2.80×10^{-24}	7.38×10^{-22}	2.80×10^{-24}	7.38×10^{-22}	6.14×10^{-28}	0.159	0.262
OASL	41	1.21×10^{-34}	9.41×10^{-35}	1.21×10^{-34}	9.41×10^{-35}	1.98×10^{-27}	0.194	0.323
FTO	116	2.39×10^{-23}	1.28×10^{-30}	2.39×10^{-23}	1.28×10^{-30}	5.75×10^{-24}	0.028	0.052
LDLR	12	0.610	0.233	0.610	0.282	0.236	0.433	0.656
APOE	14	6.30×10^{-31}	3.38×10^{-31}	6.30×10^{-31}	3.38×10^{-31}	1.37×10^{-31}	0.563	0.736
GIPR	12	1.45×10^{-3}	3.33×10^{-3}	1.45×10^{-3}	8.26×10^{-3}	2.19×10^{-3}	0.311	0.085

Table S.7: **Association Analysis of Type 2 Diabetes Status in Eight European Cohorts by Homogeneous Rao's Efficient Score Test Statistics (Hom-Rao), Hom-Meta-SKAT-O, and Hom-Meta-SKAT, Using Rare Variants (MAFs≤0.03).** The associations that attain a threshold significance of p-value $< 3.1 \times 10^{-6}$ are marked by red. The results of "Basis of both GVF and $\beta_\ell(t)$" were based on smoothing both GVF and genetic effect functions $\beta_\ell(t)$ of model (6), and the results of "Basis of β-smooth only" were based on smoothing $\beta_\ell(t)$ only approach of model (3), and the p-values of Hom-Meta-SKAT and Hom-Meta-SKAT-O were based of R package MetaSKAT. Abbreviation: GVF = genetic variant function.

Gene	**Number of Rare Variants (MAFs≤0.03)**	**p-values of the Het-Rao**					**p-values of Het-Meta**	
		Basis of both GVF and $\beta_\ell(t)$		**Basis of β-smooth only**		**Additive Effect Model**		
		B-spline	**Fourier**	**B-spline**	**Fourier**		**SKAT**	**SKAT-O**
PCSK9	42	0.214	0.381	0.415	0.460	0.374	0.125	0.130
APOB	175	0.906	0.383	0.399	0.364	0.768	0.757	0.663
IGF2BP2	157	0.062	0.503	0.324	0.162	0.327	0.432	0.322
CDKAL1	323	0.613	0.508	0.755	0.717	0.086	0.384	0.513
JAZF1	252	0.783	0.311	0.475	0.295	0.365	0.363	0.239
LPL	140	0.738	0.712	0.552	0.399	0.614	0.426	0.482
CDKN2B	40	0.770	0.212	0.620	0.423	0.657	0.577	0.747
CDC123	178	0.600	0.468	0.735	0.372	0.776	0.145	0.199
IDE	208	0.496	0.417	0.475	0.450	0.392	0.355	0.477
KIF11	134	0.488	0.536	0.503	0.601	0.846	0.629	0.564
HHEX	25	0.820	0.516	0.759	0.625	0.369	0.575	0.691
TCF7L2	170	7.87×10^{-3}	6.29×10^{-4}	2.22×10^{-4}	1.56×10^{-4}	0.055	0.022	0.038
KCNQ1	356	0.354	0.570	0.588	0.728	0.101	0.544	0.761
MTNR1B	83	0.317	0.245	0.352	0.390	0.602	0.586	0.542
HMGA2	164	0.745	0.548	0.295	0.455	0.792	0.906	0.887
TSPAN8	46	0.349	0.377	0.679	0.414	0.905	0.984	0.863
HNF1A	24	0.532	0.795	0.496	0.529	0.223	0.473	0.606
OASL	67	0.892	0.928	0.424	0.564	0.610	0.420	0.337
FTO	75	0.304	0.580	0.093	0.043	0.354	0.657	0.488
LDLR	31	0.257	0.442	0.380	0.485	0.943	0.628	0.601
APOE	21	0.056	0.096	0.044	0.070	0.036	0.146	0.141
GIPR	25	0.453	0.317	0.453	0.322	0.544	0.760	0.880

Table S.8: **Association Analysis of Type 2 Diabetes Status in Eight European Cohorts by Homogeneous Rao's Efficient Score Test Statistics (Hom-Rao), Hom-Meta-SKAT-O, and Hom-Meta-SKAT, using common variants (MAFs>0.03).** The associations that attain a threshold significance of p-value $< 3.1 \times 10^{-6}$ are marked by red. The results of "Basis of both GVF and $\beta_\ell(t)$" were based on smoothing both GVF and genetic effect functions $\beta_\ell(t)$ of model (6), and the results of "Basis of β-smooth only" were based on smoothing $\beta_\ell(t)$ only approach of model (3), and the p-values of Hom-Meta-SKAT and Hom-Meta-SKAT-O were based of R package MetaSKAT. Abbreviation: GVF = genetic variant function.

Gene	Number of Common Variants (MAFs>0.03)	p-values of the Het-Rao					p-values of Het-Meta	
		Basis of both GVF and $\beta_\ell(t)$		Basis of β-smooth only		Additive Effect Model		
		B-spline	Fourier	B-spline	Fourier		SKAT	SKAT-O
PCSK9	32	0.280	0.348	0.241	0.101	0.746	0.070	0.041
APOB	48	0.478	0.544	0.114	0.091	0.756	0.822	0.770
IGF2BP2	74	3.52×10^{-3}	3.97×10^{-3}	4.97×10^{-5}	$4.92\times^{-5}$	0.025	0.258	0.414
CDKAL1	237	0.220	0.280	0.095	0.024	0.583	0.625	0.776
JAZF1	132	0.224	0.360	0.082	0.126	0.731	0.266	0.148
LPL	72	0.069	0.091	0.033	0.042	0.062	0.303	0.139
CDKN2B	24	0.084	0.007	0.002	0.003	0.040	0.098	0.121
CDC123	87	0.023	0.014	0.064	0.044	0.017	0.168	0.249
IDE	119	0.296	0.260	0.036	0.052	0.407	0.132	0.140
KIF11	82	0.430	0.090	0.183	0.018	0.685	0.480	0.559
HHEX	5	0.690	0.390	0.548	0.425	0.495	0.660	0.660
TCF7L2	88	3.15×10^{-14}	1.03×10^{-14}	1.44×10^{-15}	2.99×10^{-15}	1.07×10^{-09}	6.97×10^{-5}	3.38×10^{-5}
KCNQ1	304	0.088	0.083	0.040	0.201	0.128	0.349	0.482
MTNR1B	23	0.057	0.015	0.085	0.058	0.065	0.181	0.195
HMGA2	50	0.304	0.845	0.270	0.580	0.463	0.597	0.702
TSPAN8	8	0.871	4.88×10^{-10}	0.087	2.77×10^{-3}	0.593	0.803	1
HNF1A	47	0.049	0.104	0.010	0.018	0.219	0.631	0.401
OASL	41	0.267	0.140	3.61×10^{-3}	7.09×10^{-3}	0.205	0.542	0.432
FTO	116	4.32×10^{-3}	5.87×10^{-4}	1.16×10^{-5}	3.86×10^{-6}	0.380	0.154	0.222
LDLR	12	0.775	0.616	0.816	0.635	0.667	0.914	1
APOE	14	0.184	0.163	5.09×10^{-3}	1.59×10^{-3}	0.123	0.032	0.053
GIPR	12	0.223	0.075	0.157	0.106	0.040	0.043	0.041

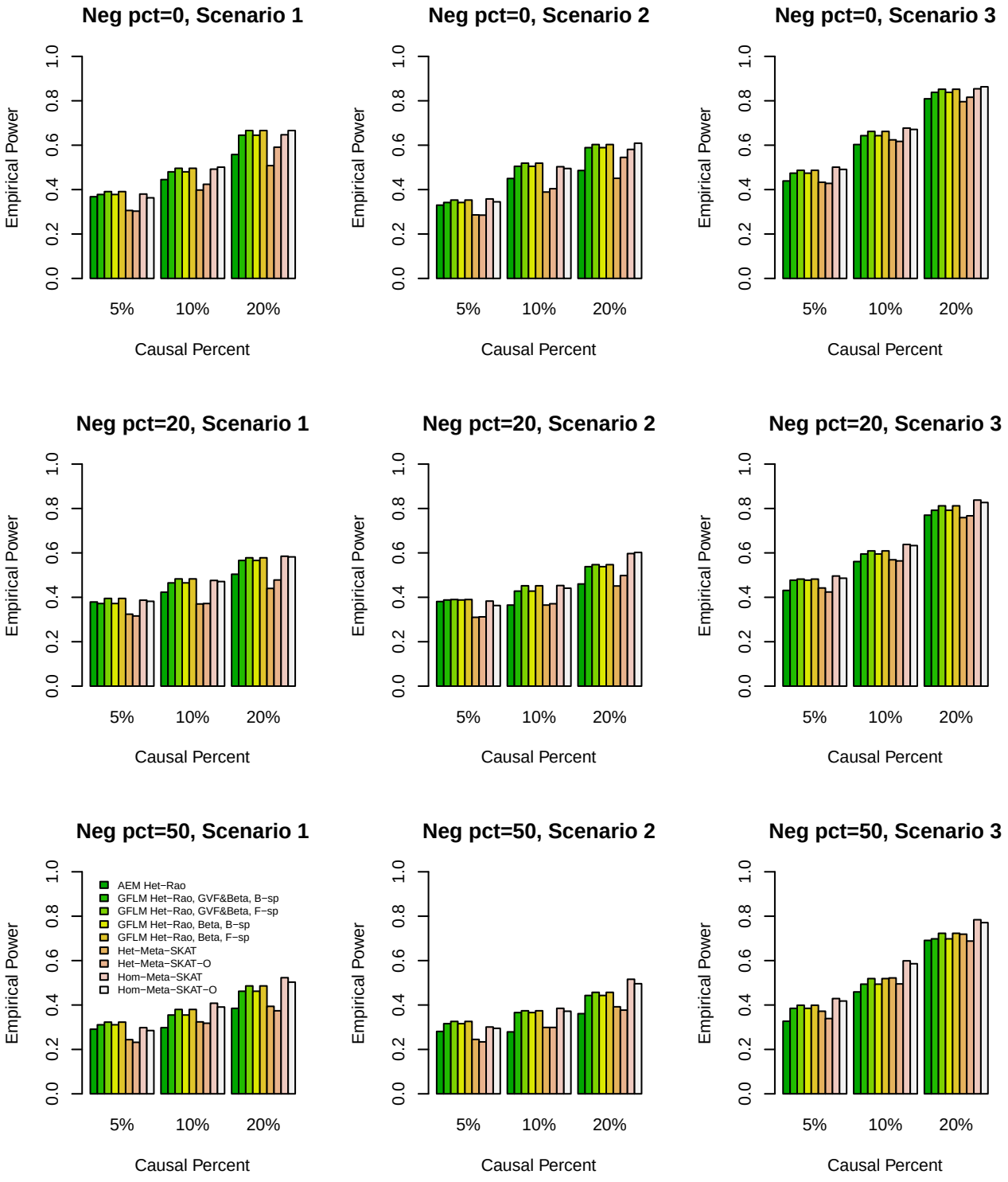

Figure S.1: **The empirical power of the heterogeneous Rao's efficient score test statistics (Het-Rao) of the models (1), (3), and (6) and MetaSKAT at $\alpha = 0.0001$, when some causal variants are rare and some are common and the genetic effect is simulated as homogeneous.** When Neg pct $= 0$, all causal variants had positive effects; when Neg pct $= 20$, 20%/80% causal variants had negative/positive effects; when Neg pct $= 50$, 50%/50% causal variants had negative/positive effects.

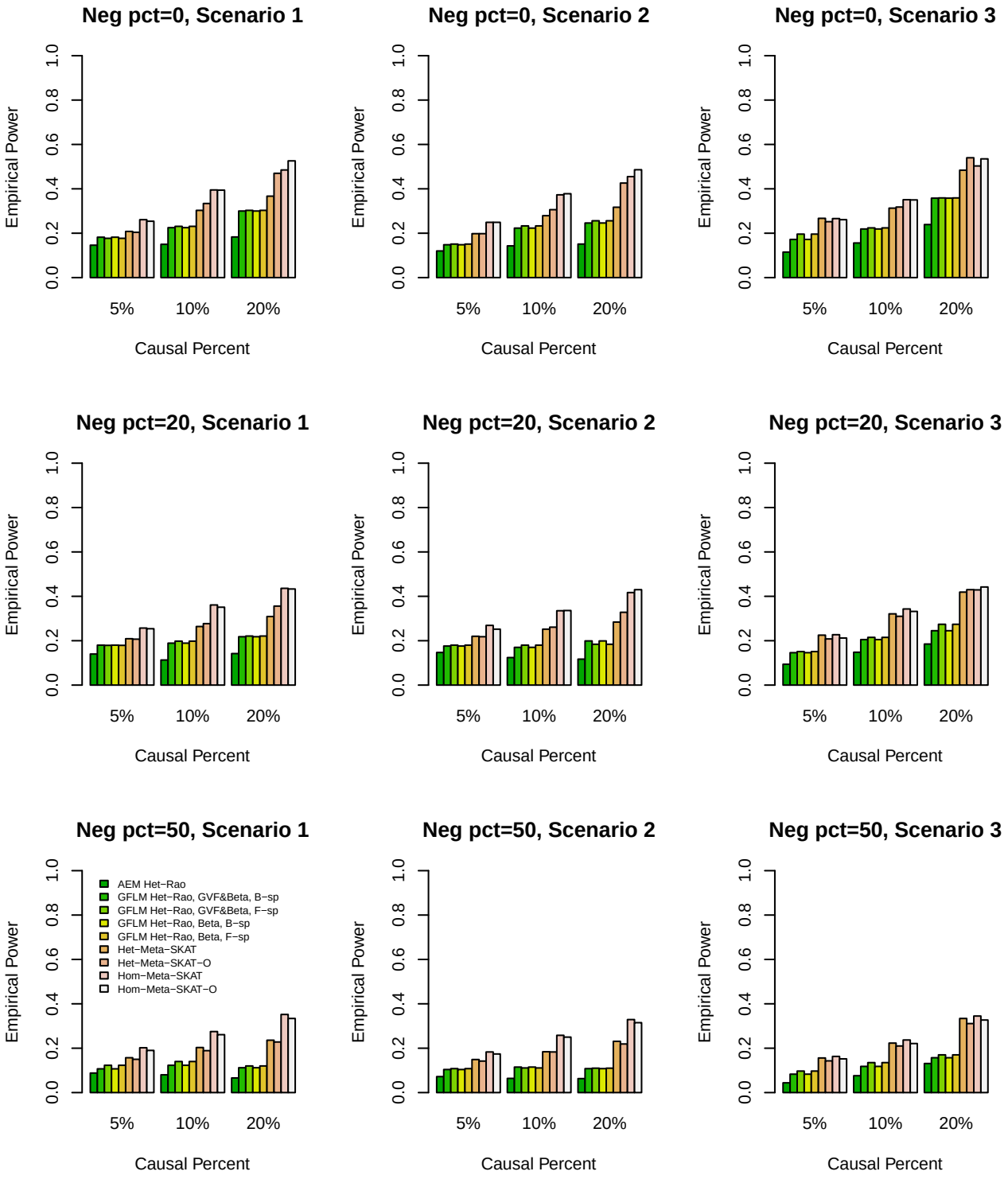

Figure S.2: **The empirical power of the heterogeneous Rao's efficient score test statistics (Het-Rao) of the models (1), (3), and (6) and MetaSKAT at $\alpha = 0.0001$, when all causal variants are rare and the genetic effect is simulated as homogeneous.** When Neg pct = 0, all causal variants had positive effects; when Neg pct = 20, 20%/80% causal variants had negative/positive effects; when Neg pct = 50, 50%/50% causal variants had negative/positive effects.

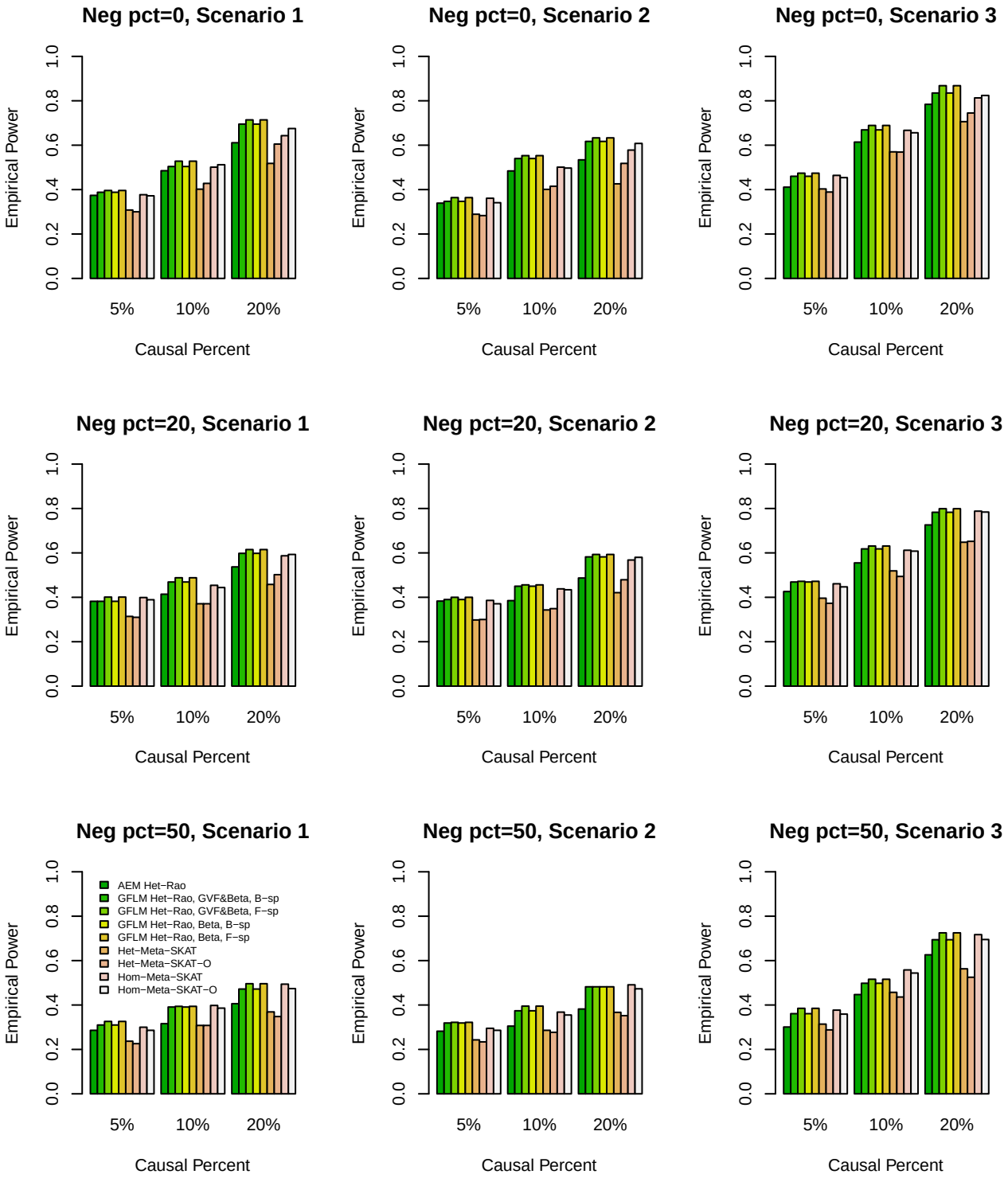

Figure S.3: **The empirical power of the heterogeneous Rao's efficient score test statistics (Het-Rao) of the models (1), (3), and (6) and MetaSKAT at $\alpha = 0.0001$, when some causal variants are rare and some are common and the genetic effect is simulated as heterogeneous.** When Neg pct = 0, all causal variants had positive effects; when Neg pct = 20, 20%/80% causal variants had negative/positive effects; when Neg pct = 50, 50%/50% causal variants had negative/positive effects.

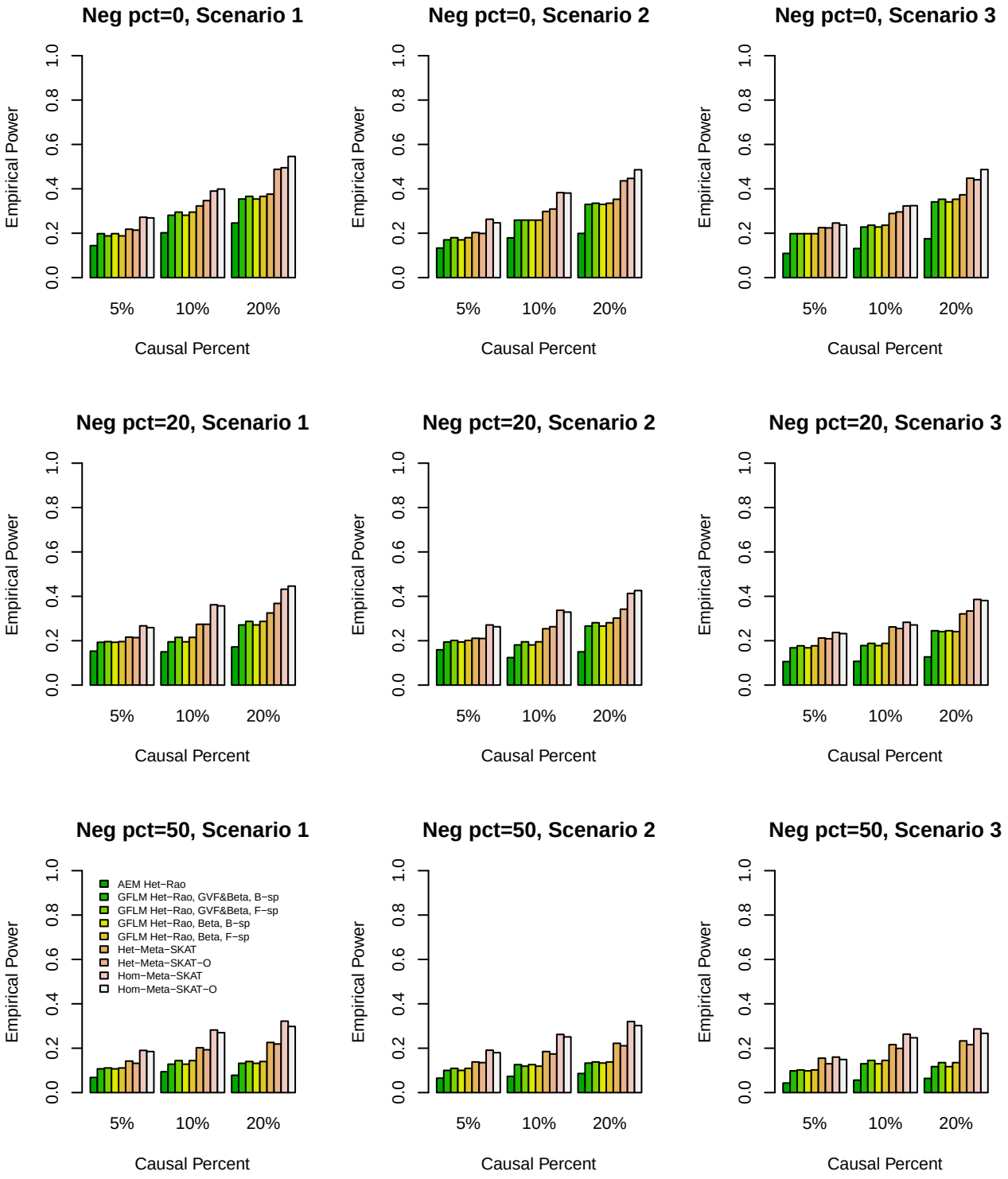

Figure S.4: **The empirical power of the heterogeneous Rao's efficient score test statistics (Het-Rao) of the models (1), (3), and (6) and MetaSKAT at $\alpha = 0.0001$, when all causal variants are rare and the genetic effect is simulated as heterogeneous.** When Neg pct = 0, all causal variants had positive effects; when Neg pct = 20, 20%/80% causal variants had negative/positive effects; when Neg pct = 50, 50%/50% causal variants had negative/positive effects.