HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

J Am Stat Assoc. 2015 June 1; 110(510): 560–572. doi:10.1080/01621459.2015.1008099.

Analysis of Sequence Data Under Multivariate Trait-Dependent Sampling

Ran Tao[1], **Donglin Zeng**[1], **Nora Franceschini**[2], **Kari E. North**[2], **Eric Boerwinkle**[3], and **Dan-Yu Lin**[1]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

[2]Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599

[3]Human Genetics Center, University of Texas Health Science Center, Houston, TX 77030

Abstract

High-throughput DNA sequencing allows for the genotyping of common and rare variants for genetic association studies. At the present time and for the foreseeable future, it is not economically feasible to sequence all individuals in a large cohort. A cost-effective strategy is to sequence those individuals with extreme values of a quantitative trait. We consider the design under which the sampling depends on multiple quantitative traits. Under such trait-dependent sampling, standard linear regression analysis can result in bias of parameter estimation, inflation of type I error, and loss of power. We construct a likelihood function that properly reflects the sampling mechanism and utilizes all available data. We implement a computationally efficient EM algorithm and establish the theoretical properties of the resulting maximum likelihood estimators. Our methods can be used to perform separate inference on each trait or simultaneous inference on multiple traits. We pay special attention to gene-level association tests for rare variants. We demonstrate the superiority of the proposed methods over standard linear regression through extensive simulation studies. We provide applications to the Cohorts for Heart and Aging Research in Genomic Epidemiology Targeted Sequencing Study and the National Heart, Lung, and Blood Institute Exome Sequencing Project.

Keywords

Association studies; Gene-level tests; Linear regression; Quantitative traits; Rare variants; Sequencing studies

1. INTRODUCTION

The past few years have seen progressive advances in high-throughput sequencing technologies that allow the sequencing of genomic regions for association studies. However,

Ran Tao is Doctoral Student (taor@live.unc.edu), and Donglin Zeng is Professor (dzeng@bios.unc.edu), Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Nora Franceschini is Research Assistant Professor (no-raf@unc.edu), and Kari E. North is Associate Professor (kari north@unc.edu), Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599. Eric Boerwinkle (eric.boerwinkle@uth.tmc.edu) is Professor and Director, Human Genetics Center, University of Texas Health Science Center, Houston, TX 77030. Dan-Yu Lin is Dennis Gillings Distinguished Professor (lin@bios.unc.edu), Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599.

the cost of performing high-throughput sequencing on a large number of individuals is still high and will likely remain so in the near future. If a quantitative trait is of primary interest, then a cost-effective strategy is to sequence individuals with the extreme trait values. This trait-dependent sampling (TDS) strategy can substantially increase statistical power when compared to a random sample of the same size (Allison 1997; Page and Amos 1999; Slatkin 1999; Chen et al. 2005; Huang and Lin 2007; Lin et al. 2013).

Many sequencing studies are derived from large, population-based cohorts, such as the Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators 1989), Cardiovascular Health Study (CHS) (Fried et al. 1991), and Framingham Heart Study (FHS) (Dawber et al. 1951). In these cohorts, hundreds of traits are measured at baseline and follow-up visits. Investigators are often interested in multiple (potentially correlated) quantitative traits. One may select an equal number of individuals from the upper and lower tails of each trait distribution or select individuals from one tail of each trait distribution and use a random sample as a common comparison group. The former design was adopted by the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) (Lin et al. 2013). The latter design was recently used in the Cohorts for Heart and Aging Research in Genomic Epidemiology Targeted Sequencing Study (CHARGE-TSS) (Lin et al. 2014).

The NHLBI ESP European American (EA) sample consists of 2538 individuals who were selected for sequencing from six cohorts: ARIC, CHS, FHS, Coronary Artery Risk Development in Young Adults (CARDIA) study (Friedman et al. 1988), Multi-Ethnic Study of Atherosclerosis (MESA) (Bild et al. 2002), and Women's Health Initiative (WHI) (The Women's Health Initiative Study Group 1998). The project contains several studies, each of which was focused on a particular trait and some of which selected individuals with extreme values of quantitative traits, including low-density lipoprotein (LDL) and blood pressure (BP). The CHARGE-TSS involves three cohorts, ARIC, CHS and FHS, in which ~200 individuals with extreme values from each of 14 traits, as well as a random sample of ~2000 individuals, were selected for sequencing at a total of 77 genomic loci that had been identified by genome-wide association studies (GWAS) to be associated with one or more traits (Lin et al. 2014).

Standard linear regression analysis based on least squares (LS) estimation only uses the sequenced individuals and treats them as if they were randomly selected from the whole cohorts. Thus, the multivariate TDS design is ignored with this approach. If the genetic variant of interest is independent of all the traits used in the sampling, then the LS method has correct type I error. If the genetic variant affects certain traits used in the sampling, however, then the LS method yields biased estimates of the genetic effects. The type I error for testing the genetic effect on one trait may also be inflated if other traits that are used in sampling are affected by the genetic variant.

Analysis methods for the univariate TDS design, such as that of Lin et al. (2013), may be applied to the multivariate TDS design. Lin et al. (2013) analyzed the LDL data in the NHLBI ESP by performing separate analysis in each study and combining the summary statistics. This approach may not preserve the type I error because it cannot properly handle

sequenced individuals with extreme values in multiple traits, as elaborated in Section 2. In the CHARGE-TSS, the selection of individuals with the extreme values of the pulmonary function was based on both the forced expiratory volume in the first second (FEV_1) and the ratio of FEV_1 to forced vital capacity (FEV_1/FVC) (Lin et al. 2014). The univariate approach is not applicable to this case because it does not allow the selection of an individual to depend on multiple traits. Another limitation of the univariate approach is that it cannot perform simultaneous inference on multiple traits.

In this paper, we develop a valid and efficient likelihood-based approach to making inferences about genetic effects under multivariate TDS. In our formulation, the sampling can depend on multiple quantitative traits in any manner. Quantitative traits are related to genetic variants and covariates through a multivariate linear regression model while the distributions of genetic variants and covariates are unspecified. We derive the likelihood that accounts for the TDS and utilizes all available data. The computation is challenging due to the presence of missing trait values with arbitrary patterns, the multivariate nature of the model, and a potentially infinite-dimensional covariate distribution. We develop a novel expectation-maximization (EM) algorithm (Dempster et al. 1977) to maximize the likelihood. We establish the consistency, asymptotic normality, and asymptotic efficiency of the resulting estimators by using novel arguments to deal with the challenging issue of partially missing trait values. We construct single-variant and gene-level association tests (Li and Leal 2008; Madsen and Browning 2009; Price et al. 2010; Lin and Tang 2011; Wu et al. 2011) for assessing the marginal genetic effects on each trait or the joint effects on any subset of traits. We demonstrate the superiority of the proposed methods over the univariate approach and standard linear regression through extensive simulation studies. Finally, we provide applications to the CHARGE-TSS and NHLBI ESP data.

2. METHODS

Let $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_K)^{\mathrm{T}}$ be a $K \times 1$ vector of quantitative traits, $\boldsymbol{G}$ be a $d \times 1$ vector of genetic variables, and $\boldsymbol{Z}$ be a $p \times 1$ vector of covariates (including the unit component). We relate $\boldsymbol{Y}$ to $\boldsymbol{G}$ and $\boldsymbol{Z}$ through the multivariate linear model:

$$\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{G} + \boldsymbol{\gamma}\boldsymbol{Z} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\boldsymbol{\beta}$ is a $K \times d$ matrix of regression parameters for the genetic effects, $\boldsymbol{\gamma}$ is a $K \times p$ matrix of regression parameters for the covariate effects, and $\boldsymbol{\varepsilon}$ is a K-variate normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. In single-variant analysis, $d = 1$, and G is a scalar that codes the number of minor alleles the individual carries at the variant site under the additive model or indicates whether the individual carries any minor allele (or two minor alleles) at that site under the dominant (or recessive) model. In gene-level analysis for rare variants, $\boldsymbol{G}$ is a (weighted) sum of the numbers of mutations across multiple variant sites within a gene or the vector of genotypes for individual variants.

Under the multivariate TDS design, $\boldsymbol{Y}$ is measured on all the N individuals in the cohort (with potential missing values), and $\boldsymbol{G}$ is only collected for a sub-sample of size n. The selection may depend on observed $\boldsymbol{Y}$ in an arbitrary manner. Under the “one-tail” design

used in the CHARGE-TSS, the sequenced individuals include those with extreme values of each quantitative trait of interest plus a random sample. Under the “two-tail” design used in the NHLBI ESP, the sequenced individuals have the largest or smallest trait values. If $\boldsymbol{Z}$ contains demographic/environmental variables and ancestry information, such as the percentage of African ancestry or the principal components (PCs) for ancestry, which is estimated from the GWAS marker data, then $\boldsymbol{Z}$ may potentially be available for all N individuals. If the ancestry information is obtained from the sequence data, then $\boldsymbol{Z}$ is available only for the n sequenced individuals. Because it is often difficult to retrieve covariate information for nonsequenced individuals, especially when multiple cohorts are involved, we require $\boldsymbol{Z}$ to be available only for the n sequenced individuals.

We arrange the records such that the first n individuals are the sequenced ones and the remaining $(N - n)$ are the nonsequenced ones. Then the data consist of $(\boldsymbol{Y}_i^{obs}, \boldsymbol{Z}_i, \boldsymbol{G}_i)$ for $i = 1, \ldots, n$ and $\boldsymbol{Y}_i^{obs}$ for $i = n + 1, \ldots, N$, where $\boldsymbol{Y}_i^{obs}$ is the observed part of $\boldsymbol{Y}_i$. We include all the individuals with at least one nonmissing trait — the largest possible sample — in the analysis. We assume that the observations on $\boldsymbol{Y}$ are missing at random. We require $\boldsymbol{Z}$ to be completely observed for all sequenced individuals, which is the case in both the CHARGE-TSS and NHLBI ESP.

We represent β, γ, and Σ by θ. We show in Appendix A.1 that the observed-data likelihood takes the form

$$\prod_{i=1}^{n}[f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{Z}_i,\boldsymbol{G}_i)f(\boldsymbol{Z}_i,\boldsymbol{G}_i)]\prod_{i=n+1}^{N}\int_{\boldsymbol{z},\boldsymbol{g}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z},\boldsymbol{g})dF(\boldsymbol{z},\boldsymbol{g}), \quad (2)$$

where $f_{\boldsymbol{\theta}}(\cdot|\boldsymbol{z},\boldsymbol{g})$ is the joint density of $\boldsymbol{Y}^{obs}$ conditional on $(\boldsymbol{Z},\boldsymbol{G}) = (\boldsymbol{z},\boldsymbol{g})$, $f(\cdot,\cdot)$ is the joint density of $(\boldsymbol{Z},\boldsymbol{G})$, and $F(\cdot,\cdot)$ is the distribution function of $f(\cdot,\cdot)$. Note that we do not assume a specific form for $f(\cdot,\cdot)$ in (2). Thus, $f(\cdot,\cdot)$ is infinite-dimensional when $\boldsymbol{Z}$ contains continuous covariates. We estimate $f(\cdot,\cdot)$ by the discrete probabilities at the observed distinct values of $(\boldsymbol{Z}_i,\boldsymbol{G}_i)$, $i = 1, \ldots, n$, denoted by $(\boldsymbol{z}_1,\boldsymbol{g}_1), \ldots, (\boldsymbol{z}_m,\boldsymbol{g}_m)$, $m \le n$, and maximize the above function over other parameters. Denote the point mass at $(\boldsymbol{z}_j,\boldsymbol{g}_j)$ as q_j, $j = 1, \ldots, m$. The objective function to be maximized is equivalent to

$$\sum_{i=1}^{n}\left[\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{Z}_i,\boldsymbol{G}_i)+\log\sum_{j=1}^{m}I\{(\boldsymbol{Z}_i,\boldsymbol{G}_i)=(\boldsymbol{z}_j,\boldsymbol{g}_j)\}\,q_j\right]+\sum_{i=n+1}^{N}\log\sum_{j=1}^{m}f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z}_j,\boldsymbol{g}_j)q_j, \quad (3)$$

where $I(\cdot)$ is the indicator function.

We present in Appendix A.2 a novel EM algorithm for maximizing (3) that is computationally efficient and numerically stable. In addition, we prove in Appendix A.3 that the resulting maximum likelihood estimators (MLEs) are consistent, asymptotically normal, and asymptotically efficient. Thus, the corresponding association tests have correct type I error and are the most powerful of all valid tests.

Inferences about the genetic effects on the traits of interest are flexible under our likelihood framework, as detailed in Appendix A.4. For single-variant analysis, G is a scalar, and β reduces to a $K \times 1$ vector. We can use the Wald, score, or likelihood ratio statistics to test any subset of β. The Wald tests are the most efficient computationally because we only need to fit the model once no matter how many and what kind of hypotheses we are interested in; to perform the score or likelihood ratio tests, we need to obtain the restricted MLEs under each null hypothesis. For variants with moderate minor allele frequencies (MAFs), the three types of tests give similar results.

To perform a burden test for rare variants, we define G as the total number of mutations among variants whose MAFs are below a pre-specified threshold, such as 1% or 5%, with the corresponding tests denoted by T1 and T5, respectively; alternatively, we define G as a weighted sum of the mutation counts, using weights such as those defined by Madsen and Browning (2009) to reflect each variant's MAF, with the corresponding test denoted by MB. For detecting variants with opposite effects on the traits, we extend the sequence kernel association test (SKAT) (Wu et al. 2011) to the multivariate TDS setting. We can test the null hypothesis that there is no genetic effect on a particular trait or the "global" null hypothesis that there is no genetic effect on any trait. All our gene-level tests are based on the score statistics, which are statistically more accurate and numerically more stable than the Wald statistics for rare variants (Lin and Tang 2011).

Lin et al. (2013) proposed a likelihood-based approach for the univariate TDS design. They derived efficient estimators for both the primary trait, which is used for sampling, and the secondary trait, which is not used for sampling. Suppose that we wish to make inference on the first trait under a multivariate TDS design with K traits. We can analyze the first trait as the primary trait by treating the individuals with extreme values of the first trait as sequenced individuals and all others as nonsequenced individuals. We can also analyze the first trait as a secondary trait with each of the remaining $(K-1)$ traits as the primary trait. We can then combine the summary statistics of the K analyses. This meta-analysis is not valid because it does not account for the correlations of the K statistics caused by overlapping individuals. To avoid overlaps of sequenced individuals, we let each individual be considered "sequenced" in only one analysis. This strategy, however, will introduce bias into the univariate analysis because the "selection" for one trait depends on other traits. We label these two methods as (a) and (b), respectively.

For the design that contains a random sample, such as the one-tail design adopted by the CHARGE-TSS, each individual in the cohort has a positive probability of being selected. Then the inverse probability weighting (IPW) method commonly used in survey sampling can be adopted. The IPW method avoids the joint modeling of the traits and thus can handle quantitative, binary, and censored traits simultaneously. It yields unbiased effect estimation and correct type I error. Such weighting methods, however, are substantially less efficient than the LS method (T. Lumley, personal communication, April 19, 2012). Efficiency is a major concern in association studies since many genetic effects are small and the correction for multiple comparisons is extremely severe for tens of thousands of variants. In addition, IPW is not applicable to the design that does not contain a random sample.

3. RESULTS

3.1 Simulation Studies

We evaluated the performance of the MLE and LS methods in extensive simulation studies. The ARIC data in the CHARGE-TSS are more complex than the NHLBI ESP data because the former contain more sampling traits and more sequenced individuals with extreme trait values than the latter. Thus, we designed our simulation studies to mimic the ARIC data in the CHARGE-TSS.

We generated 11 traits from the multivariate linear model given in (1) in which G is the number of minor alleles for a SNP with MAF of 0.1, Z is a normally distributed confounder (representing a PC for ancestry or some other genetically related variable) with mean G and unit variance, and the error terms are multivariate normal with mean 0, variances 1, and correlations r under compound symmetry. (The Pearson correlation between G and Z is ~0.17.) We generated a cohort of 9000 individuals and selected individuals for sequencing as follows: we first selected a random sample of 1000 individuals; we then selected 100 individuals with the largest values of Y_1 from the remaining 8000 individuals; and we continued to select 100 individuals with the largest values of Y_2 from the remaining 7900 individuals, and so on, until we reached a "sequenced" sample of 2100 individuals. We set $\beta_1 = 0$ and considered two cases of non-zero effects for the other 10 traits: Case 1. five traits with the same effect, i.e., $\beta_2 = \cdots = \beta_6 = 0.2$, $\beta_7 = \cdots = \beta_{11} = 0$; and Case 2. six traits with opposite effects, i.e., $\beta_2 = \beta_4 = \beta_6 = 0.2$, $\beta_3 = \beta_5 = \beta_7 = -0.2$, $\beta_8 = \cdots = \beta_{11} = 0$. The value of 0.2 for β corresponds to R^2 of 0.7% and 4.0% under $\gamma = 0$ and 0.3, respectively; the value of −0.2 corresponds to R^2 of 0.7% and 0.2% under $\gamma = 0$ and 0.3, respectively. We assessed the bias, type I error, and power of the MLE and LS methods. The nominal significance level α was set to 0.001. All results are based on 100,000 replicates.

Table 1 shows the results for trait 1 (null effect) and trait 2 (positive effect) in Case 1. The MLE method provides unbiased estimation of genetic effects and correct type I error. The LS method is approximately unbiased for β_1 when the confounder has no effect and the traits are strongly correlated, and it has a negative bias for β_1 when there is confounding or the traits are weakly correlated or independent. When the confounder has no effect, the LS method substantially overestimates β_2. The bias is larger when the correlations are lower. When there is confounding, the bias decreases as the correlations increase. When the traits are weakly correlated or independent, the LS method yields highly inflated type I error, whether or not the confounder has an effect. The type I error is also inflated when the traits are strongly correlated and the confounder has an effect. The MLE method is more powerful than the LS method because its standardized test statistic tends to be larger. The largest power difference is 0.188 under $\gamma = 0.3$ and $r = 0.5$. The MLE method always yields smaller root mean squared error (RMSE) than the LS method (see Table S1 of the Supplementary Material).

Table 2 shows the results for trait 1 (null effect), trait 2 (positive effect), and trait 3 (negative effect) in Case 2. The MLE method continues to provide unbiased estimation of genetic effects and correct type I error. The LS method tends to overestimate the effect on trait 2 and underestimate the effect on trait 3, and the bias can be as high as 26%, which is higher than

in Case 1. The LS method also has inflated type I error (as high as 80%) when there is confounding. When the confounder has no effect, the LS method generally has correct type I error, although it is not as powerful as the MLE method; the power differences are larger when the correlations are higher, which is opposite to what we find in Case 1. The MLE method always yields smaller root mean squared error (RMSE) than the LS method (see Table S1). For both Case 1 and Case 2, we conducted other simulations with larger genetic effects and lower MAFs or with 10% random missingness in all traits. The results are similar to those of Tables 1 and 2 and thus not shown.

Due to the presence of a random sample, it was possible to evaluate the IPW method. We set the weights for individuals with extreme trait values at 1 and set the weights for individuals in the random sample at 9. These weights are not exactly equal to the inverse selection probabilities, which are difficult to calculate under the sequential selection mechanism, but the approximations are good enough for our illustration. The results for Case 1 and Case 2 are summarized in Table S2. Comparing Table S2 with Tables 1 and 2, we observe that although the IPW method preserves the type I error, it is substantially less powerful than the MLE and LS methods.

We also conducted simulation studies under the two-tail design. Specifically, we generated the cohort in the same manner as in the previous simulation studies but sequentially selected 95 individuals from the upper and lower tails of each trait distribution to reach a "sequenced" sample of 2090 individuals. The results that are analogous to those shown in Tables 1 and 2 are summarized in Tables S3 and S4. The MLE method continues to perform well. Because the two-tail sampling is more extreme than the one-tail sampling used in the previous simulation studies, the LS method tends to yield more bias. The loss of power by the LS method compared to the MLE method tends to be more severe under the two-tail design than under the one-tail design (with maximal differences of 0.583 vs. 0.188). In addition, the MLE method is generally more powerful under the two-tail design than under the one-tail design (with the power difference being as high as 0.184).

We conducted additional simulation studies under simple random sampling. We generated the cohort in the same manner as before but selected a simple random sample of 2100 individuals. The LS method is valid in this setting. The power is approximately 0.61 for all traits with non-zero effects (positive or negative) in both Case 1 and Case 2 with any combination of γ and r. When comparing with the power estimates for trait 2 in Tables 1 and S3 and traits 2 and 3 in Tables 2 and S4, we see that the two multivariate TDS designs are much more efficient than simple random sampling.

To assess the robustness to the normality assumption, we simulated data in the setup of Case 1 under the one-tail design but let $\boldsymbol{\varepsilon}$ follow a multivariate t distribution $t_\nu(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the scale matrix, and ν is the degrees of freedom. We set $\gamma = 0.3$ and $r = 0.05$. We added a variation of the MLE method that applies the inverse normal transformation to the trait values, which is referred to as MLE-INV. The results are summarized in Table S5. The MLE method has appreciable bias and inflated type I error for trait 1 (null effect) when ν is small but performs reasonably well when ν is moderate or large. The MLE-INV method has

better control of the type I error than the MLE method when ν is small. The LS method is biased and its performance worsens as ν increases.

To compare our multivariate approach with the univariate approach of Lin et al. (2013), we simulated a cohort of 10,000 individuals with two traits. We set the genetic variable to be the number of minor alleles for a SNP with MAF of 0.1, the effect sizes at 0.2 and 0 for the two traits; we did not include any confounder in the model. We adopted the two-tail design by sequentially selecting 250 individuals from the upper and lower tails of the two trait distributions. We used score tests for both approaches. We set the nominal significance level at 0.001 and varied the correlation between the two traits and the proportion of random missingness for each trait. As shown in Table S6, the univariate approach has inflated type I error, which is caused by the underestimation of the variance in method (a) and the bias in method (b). The inflation increases as the correlation between the two traits becomes stronger. There is power loss in (b) as compared to the multivariate approach, which is caused by the larger variances of the test statistics. The power difference is larger when the correlation is higher and is not affected much by the level of missingness.

3.2 CHARGE-TSS ARIC Data

We considered the ARIC data in the CHARGE-TSS. As described, a random sample plus individuals with extreme values for 11 traits were selected from ~9000 ARIC whites who provided informed consent for use of their genetic data and had sufficient DNA for sequencing. The selected individuals were sequenced for 77 genomic loci that had previously been found to be associated with one or more of 14 traits. (Three traits were not used for sampling in the ARIC data.) After quality control (QC), the genotype data included 31,813 SNPs and 2003 individuals. Details for the design, sample selection criteria, genotype QC, and annotation can be found in Lin et al. (2014).

We removed individuals without PCs (calculated from GWAS data) and obtained 9103 individuals, among whom 1927 were sequenced. Table 3 shows the number of individuals with nonmissing trait values in the cohort, the specific sampling strategy, and the achieved number of extreme cases for sequencing, as well as that number after QC for each of the 11 traits. (Note that the numbers of extreme cases for all traits may add up to be greater than n since some individuals may have extreme values for multiple traits.) Of the 11 traits used for sampling, stroke is an age-at-onset trait that cannot be incorporated into our model. We treated the 60 individuals who were selected solely due to stroke as nonsequenced individuals. As noted before, the pulmonary function trait comprised two traits — FEV_1 and FEV_1/FVC — such that the total number of traits entering into the analysis remained at 11. C-reactive protein (CRP) and retinal venule diameter have about 20% missingness in the whole cohort, while all the other traits have less than 5% missingness.

In the CHARGE-TSS, the selections for certain traits were based on the residuals of the original values adjusted for various covariates. For those traits, we used the residuals in the analysis. Most of the traits are positively correlated, and there is no pairwise correlation less than −0.15. The correlations are 0.56 between fast insulin and body mass index (BMI), 0.49 between the two pulmonary function traits, 0.30 between BMI and CRP, and 0.22 between fast insulin and hematocrit, as well as between fast insulin and CRP. All the other positive

correlations are well below 0.2, and many of them are essentially 0 (see Table S7). We included age, gender, study centers, and the top five PCs as covariates.

We focused on BMI. We restricted the single-variant analysis to SNPs with MAFs larger than 5% and ended up with 2971 SNPs. We chose the additive genetic model. Table 4 shows the top 10 SNPs for the MLE method and the corresponding LS results. The LS method consistently yields larger effect estimates for SNPs with positive effects and smaller effect estimates for SNPs with negative effects. This is similar to what we find in most scenarios under Case 2 in the simulation studies. As shown in Figure S1 of the Supplemental Material, the *p*-values for the MLE and LS methods are similar.

In gene-level analysis of rare variants, we considered "functional coding" variants, i.e., non-synonymous, splicing, and stop-gain variants, and ended up with a total of 2360 variants. We removed any targeted region with minor allele count (MAC) — the number of individuals with at least one mutation — less than five. For MB and SKAT tests, we only included variants with MAFs less than 5%. Table S8 shows the results for the top five targeted regions in each of the four types of tests based on the MLE method. We also performed gene-level tests of the global null hypothesis that there is no genetic effect on any trait. Table S9 shows the results for the top five targeted regions in each of the four types of tests. It would be worthwhile to follow up the regions identified in Tables S8 and S9 in larger samples.

3.3 NHLBI ESP EA Data

The NHLBI ESP EA data consist of the six cohorts mentioned previously and include four types of study designs. The first study is a TDS study consisting of 872 individuals who were selected from the upper and lower tails of the LDL and BP distributions. The second study is a random sample of 721 individuals with measurements on a common set of phenotypes; this study is referred to as the deeply phenotyped reference (DPR). The third study is a case-control study of early myocardial infarction (MI) consisting of 220 cases and 390 controls. The fourth study is a case-only study of stroke consisting of 335 individuals with ischemic stroke. Exome sequencing was performed on the selected individuals at the University of Washington and the Broad Institute. We implemented the genotype QC steps described by Lin et al. (2013) and obtained 1,281,645 variants.

In the TDS study, we excluded individuals (either sequenced or nonsequenced) who were not eligible for either the LDL or BP selection. In the FHS, which contains related individuals, we removed one individual from each pair of first- or second-degree relatives. The actual sample selections for LDL and BP were based on the residuals rather than the original values. We used the LDL residuals (log-transformed LDL values adjusted for age, age-squared, gender, and lipid medication) and BP residuals (mean of the residuals for diastolic and systolic BPs adjusted for age, gender, BMI, and anti-hypertensive medication) as the trait values in the analysis. We considered LDL as the trait of interest and removed individuals with missing LDL values in the DPR, MI, and stroke studies. Note that individuals with missing LDL or BP values (but not both) were still included in the analysis of the TDS study. Table 5 summarizes the sample sizes of the four studies in each cohort after QC.

In the TDS study, we used both the MLE and LS methods to analyze LDL. For case-control and case-only studies with rare diseases, standard linear regression analysis of secondary quantitative traits conditional on the disease status yields approximately correct results (Lin and Zeng 2009). Because early MI and ischemic stroke are relatively rare, we performed standard linear regression in the MI (adjusted for the MI status), stroke, and DPR studies. We included cohorts and sequencing centers/targets as covariates. We performed meta-analysis of the four studies using software MASS (Tang and Lin 2013).

We restricted the single-variant analysis to SNPs with MACs ≥5 and ended up with 109,607 SNPs. We chose the additive model and used score statistics to ensure numerical accuracy for SNPs with low MACs. Figure 1 shows the quantile-quantile plots using the MLE and LS methods in the TDS study only and in all four studies. Although the trends in the quantile-quantile plots of the TDS study appear to be similar between the MLE and LS methods, the MLE method clearly produces more significant results than the LS method in the meta-analysis. Table 6 lists the top 10 SNPs for the MLE method in the meta-analysis. For the MLE method, the top SNP (chr19:45397229) in the meta-analysis is also the top SNP in the TDS study, with the *p*-value in the meta-analysis being much more significant (2.08×10^{-10} vs. 2.64×10^{-7}). For the LS method, although the top SNP remains the same, its *p*-value in the meta-analysis is less significant than that in the TDS study (1.17×10^{-6} vs. 4.29×10^{-7}).

The forest plots shown in Figure S2 help to explain the results in Figure 1 and Table 6. The MLE estimates in the TDS study are very similar to the estimates in the DPR, MI, and stroke studies. (The estimates in the stroke study tend to have large standard errors due to its small sample size.) Thus, the MLE estimates from the meta-analysis are similar to the MLE estimates in the TDS study but with smaller standard errors. Because of its bias, the LS method yields larger effect estimates as well as (proportionately) larger standard errors than the MLE method in the TDS study, such that the two methods have similar standardized test statistics in the TDS study. Because the LS estimates in the TDS study are much larger than the LS estimates in the other three studies, meta-analysis of the LS estimates from the four studies yields less significant results than the MLE meta-analysis.

We also performed single-variant analysis in the TDS study using the univariate approach of Lin et al. (2013). Figure S3 compares the *p*-values for the multivariate and univariate methods. The two methods yield similar results for most SNPs. This is because the correlation between LDL and BP among individuals in the TDS study is only 0.01. Note that the multivariate approach produces a more significant *p*-value for the top SNP (chr19:45397229) than the univariate approach does (2.64×10^{-7} vs. 1.24×10^{-5}).

In gene-level analysis for rare variants, we considered variants that are nonsynonymous, stop-gain, stop-loss, or splicing mutations. Other steps were the same as in the analysis of the CHARGE-TSS ARIC data. The results are displayed in Figures S4–S7 and in Tables S10–S13. The conclusions regarding the performance of the MLE and LS methods are similar to those of the single-variant analysis. Again, the MLE method yields more significant results than the LS method. We also performed gene-level tests of the global null hypothesis. The results are displayed in Figure S8 and in Tables S14–S16. The strongest signals appear in the T1 tests.

4. DISCUSSION

Multivariate TDS is a useful and cost-effective design when investigators are interested in multiple quantitative traits but cannot afford to sequence all cohort members. The CHARGE-TSS and NHLBI ESP are two recent examples of this design. It is not hard to envision that many large-scale whole-exome and whole-genome sequencing projects will adopt similar multivariate TDS designs. As demonstrated in the simulation studies and in the two real examples, standard linear regression without regard to the sampling design can result in estimation bias, type I error inflation, and power loss, and the existing methods for univariate TDS have important limitations.

In this paper, we propose for the first time a valid and efficient likelihood-based approach to making inferences under multivariate TDS, paying special attention to gene-level tests for rare variants. The methodology is very general and can be applied to both genetic and non-genetic studies. The proposed EM algorithm is stable and the software is available on our website.

Our approach is scalable to whole-exome and whole-genome sequencing studies. In our single-variant analysis of the NHLBI ESP EA data, it took ~5 seconds on an IBM HS21 machine to perform one association analysis. The computation time increases as the number of traits or the percentage of missing data increases. When there are no covariates or covariates are categorical (i.e. when m is small), the computation is fast. When there are continuous covariates, we recommend splitting the genome and using multiple CPUs.

As shown in the simulation studies, the MLE method has appreciable bias and inflated type I error when the normality assumption on $\boldsymbol{\varepsilon}$ is severely violated. In practice, one should inspect the trait distributions and explore parametric transformations, such as the log transformation, or the rank-based inverse normal transformation. In genome-wide studies, a well-behaved quantile-quantile plot for the association tests would imply that non-normality has no undue influence on the type I error.

For single-variant analysis, we compared the MLE method with the univariate LS method. It is also possible to consider the multivariate LS method. If one is only interested in the marginal genetic effects on each trait and the traits are completely observed for all sequenced individuals, then univariate and multivariate LS methods yield the same results. If there is a small proportion of missingness, then the two methods should still yield similar results. If one is interested in the joint genetic effects on multiple traits, then a multivariate model is necessary. We adopt a multivariate model in our MLE approach primarily because the sampling scheme involves multiple traits. Our model is more elaborate than a univariate model, but it is the only approach that provides valid and efficient inferences for the multivariate TDS design.

In both the simulation studies and the real examples, all traits in the model are used in the sampling process. In practice, investigators may be interested in secondary quantitative traits which are not directly used for sampling but are correlated with the primary traits. (Note that standard linear regression is valid only when a secondary trait is independent of all primary traits, which is an unlikely scenario.) It is straightforward to analyze secondary traits with

our MLE method. Using a multivariate normal distribution for the primary and secondary traits, one can include each secondary trait of interest as an additional "primary" trait and use our MLE method with these (K + 1) traits.

Our approach does not require $\boldsymbol{Z}$ for nonsequenced individuals. In the NHLBI ESP, part of $\boldsymbol{Z}$ (sequencing centers/targets) is not available for nonsequencd individuals. In the CHARGE-TSS, $\boldsymbol{Z}$ is available for all individuals. Incorporating $\boldsymbol{Z}$ of nonsequenced individuals into the analysis has two advantages. First, it allows the selection of individuals for sequencing to depend on $\boldsymbol{Z}$. Second, it improves the efficiency of estimation. Then the likelihood involves the conditional distribution of $\boldsymbol{G}$ given $\boldsymbol{Z}^{(1)}$, which is the part of $\boldsymbol{Z}$ that is correlated with $\boldsymbol{G}$. We plan to incorporate kernel smoothing into the likelihood to handle continuous components in $\boldsymbol{Z}^{(1)}$. Table S17 shows the estimated distribution of ($\boldsymbol{Z}$, $\boldsymbol{G}$) in the analysis of the second most significant SNP in the NHLBI ESP EA sample; there is no strong evidence of correlation between $\boldsymbol{Z}$ and $\boldsymbol{G}$. A similar issue arises when some part of $\boldsymbol{Z}$ is subject to missingness. We denote that part of $\boldsymbol{Z}$ and $\boldsymbol{G}$ as $\tilde{\boldsymbol{G}}$ and denote the rest of $\boldsymbol{Z}$ as $\tilde{\boldsymbol{Z}}$. We plan to formulate the conditional distribution of $\tilde{\boldsymbol{G}}$ given $\tilde{\boldsymbol{Z}}$ through general odds ratio functions (Hu et al. 2010).

We have focused on the inference procedures rather than the design aspects. Although our simulation studies indicate that the two-tail design can be more efficient than the one-tail design, the optimal design remains unknown. It is unclear what the best sampling strategy is when multiple quantitative traits are of equal interest. Because our likelihood framework applies to any multivariate TDS, our variance formulas can be used to compare the efficiencies of different designs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the National Institute of Health grants R01CA082659, P01CA142538, R37GM047845.

References

Allison DB. Transmission-Disequilibrium Tests for Quantitative Traits. American Journal of Human Genetics. 1997; 60:676–690. [PubMed: 9042929]

Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacobs DR Jr, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-Ethnic Study of Atherosclerosis: Objectives and Design. American Journal of Epidemiology. 2002; 156:871–881. [PubMed: 12397006]

Chen Z, Zheng G, Ghosh K, Li Z. Linkage Disequilibrium Mapping of Quantitative-Trait Loci by Selective Genotyping. American Journal of Human Genetics. 2005; 77:661–669. [PubMed: 16175512]

Dawber TR, Meadors GF, Moore FE Jr. Epidemiological Approaches to Heart Disease: the Framingham Study. American Journal of Public Health and the Nations Health. 1951; 41:279–286.

Dempster AP, Laird NM, Rubin DB. Maximum Likelihood From Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39:1–38.

Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, O'Leary DH, Psaty B, Rautaharju P, Tracy RP, Weiler PG. The Cardiovascular Health Study: Design and Rationale. Annals of Epidemiology. 1991; 1:263–276. [PubMed: 1669507]

Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB Jr, DRJ, Liu K, Savage PJ. CARDIA: Study Design, Recruitment, and Some Characteristics of the Examined Subjects. Journal of Clinical Epidemiology. 1988; 41:1105–1116. [PubMed: 3204420]

Hu YJ, Lin DY, Zeng D. A General Framework for Studying Genetic Effects and Gene-Environment Interactions with Missing Data. Biostatistics. 2010; 11:583–598. [PubMed: 20348396]

Huang BE, Lin DY. Efficient Association Mapping of Quantitative Trait Loci With Selective Genotyping. American Journal of Human Genetics. 2007; 80:567–576. [PubMed: 17273979]

Li B, Leal SM. Methods for Detecting Associations With Rare Variants for Common Diseases: Application to Analysis of Sequence Data. American Journal of Human Genetics. 2008; 83:311–321. [PubMed: 18691683]

Lin DY, Tang ZZ. A General Framework for Detecting Disease Associations With Rare Variants in Sequencing Studies. American Journal of Human Genetics. 2011; 89:354–367. [PubMed: 21885029]

Lin DY, Zeng D. Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies. Journal of the American Statistical Association. 2006; 101:89–104.

Lin DY, Zeng D. Proper Analysis of Secondary Phenotype Data in Case-Control Association Studies. Genetic Epidemiology. 2009; 33:256–265. [PubMed: 19051285]

Lin DY, Zeng D, Tang ZZ. Quantitative Trait Analysis in Sequencing Studies Under Trait-Dependent Sampling. Proceedings of the National Academy of Sciences. 2013; 110:12247–12252.

Lin H, Wang M, Brody JA, Bis JC, Dupuis J, Lumley T, McKnight B, Rice KM, Sitlani CM, Reid JG, Bressler J, Liu X, Davis BC, Johnson AD, O'Donnell CJ, Kovar CL, Dinh H, Wu Y, Newsham I, Chen H, Broka A, DeStefano AL, Gupta M, Lunetta KL, Liu CT, White CC, Xing C, Zhou Y, Benjamin EJ, Schnabel RB, Heckbert SR, Psaty BM, Muzny DM, Cupples LA, Morrison AC, Boerwinkle E. Strategies to Design and Analyze Targeted Sequencing Data: Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study. Circulation: Cardiovascular Genetics. 2014; 7:335–343. [PubMed: 24951659]

Louis TA. Finding the Observed Information Matrix When Using the EM Algorithm. Journal of the Royal Statistical Society, Series B. 1982; 44:226–233.

Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genetics [online]. 2009; 5:e1000384. Available at http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000384.

Page GP, Amos CI. Comparison of Linkage-Disequilibrium Methods for Localization of Genes Influencing Quantitative Traits in Humans. American Journal of Human Genetics. 1999; 64:1194–1205. [PubMed: 10090905]

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei L-J, Sunyaev SR. Pooled Association Tests for Rare Variants in Exon-resequencing Studies. The American Journal of Human Genetics. 2010; 86:832–838. [PubMed: 20471002]

Slatkin M. Disequilibrium Mapping of a Quantitative-Trait Locus in an Expanding Population. American Journal of Human Genetics. 1999; 64:1765–1773.

Tang ZZ, Lin DY. MASS: Meta-Analysis of Score Statistics for Sequencing Studies. Bioinformatics. 2013; 29:1803–1805. [PubMed: 23698861]

The ARIC Investigators . The Atherosclerosis Risk in Communites (ARIC) Study: Design and Objectives. American Journal of Epidemiology. 1989; 129:687–702. [PubMed: 2646917]

The Women's Health Initiative Study Group . Design of the Women's Health Initiative Clinical Trial and Observational Study-Examples from the Women's Health Initiative. Controlled Clinical Trials. 1998; 19:61–109. [PubMed: 9492970]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data With the Sequence Kernel Association Test. American Journal of Human Genetics. 2011; 89:82–93. [PubMed: 21737059]

APPENDIX: TECHNICAL DETAILS

A.1 Derivation of the Observed-Data Likelihood

Let $\boldsymbol{V}_i \equiv (V_{i1}, \ldots, V_{iK})^{\mathrm{T}}$ be a $K\times 1$ vector of ones and zeros indicating which components of $\boldsymbol{Y}_i$ are observed or missing for the ith individual. Let R_i indicate, by the values 1 versus 0, whether the ith individual is selected for sequencing. We make the following assumptions:

Assumption 1

The conditional distribution of $\boldsymbol{V}_i$ given $(\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)$ is a function of $(\boldsymbol{Y}_i^{obs}, \boldsymbol{Z}_i, \boldsymbol{G}_i)$ for sequenced individuals and a function of $\boldsymbol{Y}_i^{obs}$ for nonsequenced individuals.

Assumption 2

The distribution of $\boldsymbol{R} \equiv (R_1, \ldots, R_N)$ depends on $(\boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{G}) \equiv \{(\boldsymbol{V}_1, \boldsymbol{Y}_1, \boldsymbol{Z}_1, \boldsymbol{G}_1), \ldots, (\boldsymbol{V}_N, \boldsymbol{Y}_N, \boldsymbol{Z}_N, \boldsymbol{G}_N)\}$ only through $\boldsymbol{V}\circ\boldsymbol{Y} \equiv (\boldsymbol{V}_1\circ\boldsymbol{Y}_1, \ldots, \boldsymbol{V}_N\circ\boldsymbol{Y}_N)$, where "$\circ$" denotes component-wise product.

Assumption 3

$f(\boldsymbol{R}|\boldsymbol{V}\circ\boldsymbol{Y})\prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)\prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i)$ does not contain parameters $\boldsymbol{\theta}$ and F.

Under Assumptions 1–2, the complete-data density for the underlying variables $(R_i, \boldsymbol{V}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)$, $i = 1, \ldots, N$, is

$$f(\boldsymbol{R}, \boldsymbol{V}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{G}) = f(\boldsymbol{R}|\boldsymbol{V}\circ\boldsymbol{Y})\prod_{i=1}^{N} f(\boldsymbol{V}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)$$
$$= f(\boldsymbol{R}|\boldsymbol{V}\circ\boldsymbol{Y})\prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i) f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{Z}_i, \boldsymbol{G}_i) f(\boldsymbol{Z}_i, \boldsymbol{G}_i) \times \prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i) f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{Z}_i, \boldsymbol{G}_i) f(\boldsymbol{Z}_i, \boldsymbol{G}_i).$$

The observed data are $(R_i, \boldsymbol{V}_i, \boldsymbol{V}_i\circ\boldsymbol{Y}_i, R_i\boldsymbol{Z}_i, R_i\boldsymbol{G}_i)$, $i = 1, \ldots, N$, whose density is obtained by integrating over the unobserved variables in the complete-data density, i.e.,

$$f(\boldsymbol{R}, \boldsymbol{V}, \boldsymbol{V}\circ\boldsymbol{Y}, \boldsymbol{R}\circ\boldsymbol{Z}, \boldsymbol{R}\circ\boldsymbol{G})$$
$$= f(\boldsymbol{R}|\boldsymbol{V}\circ\boldsymbol{Y})\prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)\left\{\int_{\boldsymbol{Y}^{mis}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{Z}_i, \boldsymbol{G}_i) d\boldsymbol{Y}^{mis}\right\} f(\boldsymbol{Z}_i, \boldsymbol{G}_i)$$
$$\times \prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i)\int_{\boldsymbol{z},\boldsymbol{g}}\left\{\int_{\boldsymbol{Y}^{mis}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{z}, \boldsymbol{g}) d\boldsymbol{Y}^{mis}\right\} dF(\boldsymbol{z}, \boldsymbol{g})$$
$$= f(\boldsymbol{R}|\boldsymbol{V}\circ\boldsymbol{Y})\prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)\prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i\circ\boldsymbol{Y}_i)$$
$$\times \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{Z}_i, \boldsymbol{G}_i) f(\boldsymbol{Z}_i, \boldsymbol{G}_i)\prod_{i=n+1}^{N}\int_{\boldsymbol{z},\boldsymbol{g}} f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z}, \boldsymbol{g}) dF(\boldsymbol{z}, \boldsymbol{g}),$$

where $\boldsymbol{R} \circ \boldsymbol{Z} = (R_1\boldsymbol{Z}_1, \ldots, R_N\boldsymbol{Z}_N)$, $\boldsymbol{R} \circ \boldsymbol{G} = (R_1\boldsymbol{G}_1, \ldots, R_N\boldsymbol{G}_N)$, and $\boldsymbol{Y}^{mis}$ is the missing part of $\boldsymbol{Y}$. We can ignore $f(\boldsymbol{R}|\boldsymbol{V} \circ \boldsymbol{Y})\prod_{i=1}^{n} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i, \boldsymbol{Z}_i, \boldsymbol{G}_i)\prod_{i=n+1}^{N} f(\boldsymbol{V}_i|\boldsymbol{V}_i \circ \boldsymbol{Y}_i)$ because of Assumption 3. The remaining part of the above density is exactly the observed-data likelihood given in (2).

A.2 Estimation

To calculate the MLEs for (3), we use the EM algorithm in which missing data contain the partially missing $\boldsymbol{Y}_i$'s and the missing observations on $(\boldsymbol{Z}, \boldsymbol{G})$ for individuals not selected for sequencing. The complete-data log-likelihood function is

$$\sum_{i=1}^{N}\left[\sum_{j=1}^{m} I\{(\boldsymbol{Z}_i, \boldsymbol{G}_i)=(\boldsymbol{z}_j;\boldsymbol{g}_j)\}\,\{\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{z}_j, \boldsymbol{g}_j)+\log q_j\}\right].$$

At the tth iteration, the M-step maximizes

$$\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{\psi}_{ij}^{(t)}\left[E\{\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}_i|\boldsymbol{z}_j, \boldsymbol{g}_j)|\,\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j;\hat{\boldsymbol{\theta}}^{(t)}\}+\log q_j\right],$$

where $E(\cdot|\,\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j;\hat{\boldsymbol{\theta}}^{(t)})$ is the conditional expectation given $\boldsymbol{Y}_i^{obs}$, $(\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j)$, evaluated at $\hat{\boldsymbol{\theta}^{(t)}}$, and $\hat{\psi}_{ij}^{(t)}$ is the conditional probability of $I\{(\boldsymbol{Z}_i, \boldsymbol{G}_i) = (\boldsymbol{z}_j, \boldsymbol{g}_j)\} = 1$ given $\boldsymbol{Y}_i^{obs}$, $(\boldsymbol{z}_1, \boldsymbol{g}_1), \ldots, (\boldsymbol{z}_m, \boldsymbol{g}_m)$, evaluated at $\hat{\boldsymbol{\theta}^{(t)}}$, $\hat{q}_1^{(t)}, \ldots, \hat{q}_m^{(t)}$. That is,

$$\hat{\psi}_{ij}^{(t)} = \begin{cases} I\{(\boldsymbol{Z}_i, \boldsymbol{G}_i)=(\boldsymbol{z}_j, \boldsymbol{g}_j)\} & i=1, \ldots, n; \\ \dfrac{f_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z}_j, \boldsymbol{g}_j)\hat{q}_j^{(t)}}{\sum_{l=1}^{m} f_{\hat{\boldsymbol{\theta}}^{(t)}}(\boldsymbol{Y}_i^{obs}|\boldsymbol{z}_l, \boldsymbol{g}_l)\hat{q}_l^{(t)}} & i=n+1, \ldots, N. \end{cases}$$

Write $\boldsymbol{W}_j=(\boldsymbol{g}_j^{\mathrm{T}}, \boldsymbol{z}_j^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. The M-step involves the following calculations:

$$(\hat{\boldsymbol{\eta}}_k^{(t+1)})^{\mathrm{T}}=\left(\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{\psi}_{ij}^{(t)}\boldsymbol{W}_j^{\otimes 2}\right)^{-1}\left[\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{\psi}_{ij}^{(t)}E\{Y_{ki}|\,\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j;\hat{\boldsymbol{\theta}}^{(t)}\}\boldsymbol{W}_j\right],\ 1 \le k \le K,$$

$$\hat{\Sigma}^{(t+1)}=N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{\psi}_{ij}^{(t)}E\left\{(\boldsymbol{Y}_i-\hat{\boldsymbol{\eta}}^{(t+1)}\boldsymbol{W}_j)^{\otimes 2}|\,\boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j;\hat{\boldsymbol{\theta}}^{(t)}\right\},$$

$$\hat{q}_j^{(t+1)}=N^{-1}\sum_{i=1}^{N}\hat{\psi}_{ij}^{(t)},$$

where $\boldsymbol{\eta}_k$ is the kth row of $\boldsymbol{\eta}$, and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$. We start with initial values $\hat{\boldsymbol{\eta}^{(0)}} = \boldsymbol{0}$, $\hat{\boldsymbol{\Sigma}^{(0)}}$ being the sample covariance matrix based on those $\boldsymbol{Y}_i$'s with complete observations, and

$\hat{q}_j^{(0)}=n^{-1}\sum_{i=1}^{n} I\{(\boldsymbol{Z}_i,\boldsymbol{G}_i)=(\boldsymbol{z}_j,\boldsymbol{g}_j)\}$, j = 1, …, m, and iterate until convergence to obtain the MLEs $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\Sigma}},\hat{q}_1,\cdots,\hat{q}_m)$. In the above expressions, the conditional expectations can be evaluated by using the fact that the missing part of $\boldsymbol{Y}_i$, denoted by $\boldsymbol{Y}_i^{mis}$, given $\boldsymbol{Y}_i^{obs}$ and ($\boldsymbol{z}_j$, $\boldsymbol{g}_j$), follows a normal distribution with mean $\boldsymbol{\beta}_i^{mis}\boldsymbol{g}_j+\boldsymbol{\gamma}_i^{mis}\boldsymbol{z}_j+\sum_i^{mo}\{\sum_i^{oo}\}^{-1}(\boldsymbol{Y}_i^{obs}-\boldsymbol{\beta}_i^{obs}\boldsymbol{g}_j-\boldsymbol{\gamma}_i^{obs}\boldsymbol{z}_j)$ and variance $\sum_i^{mm}-\sum_i^{mo}\{\sum_i^{oo}\}^{-1}\{\sum_i^{mo}\}^{\mathrm{T}}$, where $\boldsymbol{\beta}_i^{mis}$ and $\boldsymbol{\beta}_i^{obs}$ are the corresponding parts for $\boldsymbol{Y}_i^{mis}$ and $\boldsymbol{Y}_i^{obs}$ in $\boldsymbol{\beta}$, and the same partitions apply to $\boldsymbol{\gamma}$ to yield $\boldsymbol{\gamma}_i^{mis}$ and $\boldsymbol{\gamma}_i^{obs}$ and to $\boldsymbol{\Sigma}$ to yield $\sum_i^{mm}$, $\sum_i^{mo}$, and $\sum_i^{oo}$.

We estimate the asymptotic covariance matrix of the MLEs by the Louis formula (Louis 1982). We use A_{kl} to denote the (k, l)th element of any matrix $\boldsymbol{A}$. For i = 1, …, N and j = 1, …, m, we calculate the derivatives of log $f(\boldsymbol{Y}_i|\boldsymbol{z}_j,\boldsymbol{g}_j)$ + log q_j to obtain the $\{K(p+d)+K(K+1)/2+m\}\times 1$ complete-data score vector

$$\boldsymbol{l}_{1ij}=[\boldsymbol{S}_{1ij}^{\mathrm{T}},\ldots,\boldsymbol{S}_{Kij}^{\mathrm{T}},T_{11ij},T_{12ij},\ldots,T_{KKij},\boldsymbol{P}_{ij}^{\mathrm{T}}]^{\mathrm{T}},$$

where $\boldsymbol{S}_{kij}=\boldsymbol{W}_j\boldsymbol{e}_k^{\mathrm{T}}\hat{\sum}^{-1}(\boldsymbol{Y}_i-\hat{\boldsymbol{\eta}}\boldsymbol{W}_j)$, with $\boldsymbol{e}_k$ being the kth canonical vector of length K, i.e. with 1 in the kth position and 0 in all the other positions,

$$T_{klij}=-\frac{1}{2}\{1+I(k\neq l)\}(\hat{\sum}^{-1})_{kl}+\frac{1}{4}\{1+I(k\neq l)\}(\boldsymbol{Y}_i-\hat{\boldsymbol{\eta}}\boldsymbol{W}_j)^{\mathrm{T}}\hat{\sum}^{-1}(\boldsymbol{e}_{kl}+\boldsymbol{e}_{lk})\hat{\sum}^{-1}(\boldsymbol{Y}_i-\hat{\boldsymbol{\eta}}\boldsymbol{W}_j),\quad k\leq l,$$

with $\boldsymbol{e}_{kl}=\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}$ and $\boldsymbol{P}_{ij}$ = (0, …, 0, $1/\hat{q}_j$, 0, …, 0)$^{\mathrm{T}}$. We also calculate the second derivatives as a $\{K(p+d)+K(K+1)/2+m\}\times\{K(p+d)+K(K+1)/2+m\}$ matrix, which is the block diagonal matrix

$$\boldsymbol{l}_{2ij}=\begin{bmatrix}\boldsymbol{l}_{11ij} & \boldsymbol{0}_{\{K(p+d)+K(K+1)/2\}\times m}\\ \boldsymbol{0}_{m\times\{K(p+d)+K(K+1)/2\}} & \boldsymbol{l}_{22ij}\end{bmatrix},$$

where

$$\boldsymbol{l}_{11ij}=\begin{bmatrix} \frac{\partial \boldsymbol{s}_{1ij}}{\partial \boldsymbol{\eta}_1} & \cdots & \frac{\partial \boldsymbol{s}_{1ij}}{\partial \boldsymbol{\eta}_K} & \frac{\partial \boldsymbol{s}_{1ij}}{\partial \sum_{11}} & \cdots & \frac{\partial \boldsymbol{s}_{1ij}}{\partial \sum_{KK}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{s}_{Kij}}{\partial \boldsymbol{\eta}_1} & \cdots & \frac{\partial \boldsymbol{s}_{Kij}}{\partial \boldsymbol{\eta}_K} & \frac{\partial \boldsymbol{s}_{Kij}}{\partial \sum_{11}} & \cdots & \frac{\partial \boldsymbol{s}_{Kij}}{\partial \sum_{KK}} \\ \frac{\partial \boldsymbol{s}_{1ij}}{\partial \sum_{11}}^{\mathrm{T}} & \cdots & \frac{\partial \boldsymbol{s}_{Kij}}{\partial \sum_{11}}^{\mathrm{T}} & \frac{\partial T_{11ij}}{\partial \sum_{11}} & \cdots & \frac{\partial T_{11ij}}{\partial \sum_{KK}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \boldsymbol{s}_{1ij}}{\partial \sum_{KK}}^{\mathrm{T}} & \cdots & \frac{\partial \boldsymbol{s}_{Kij}}{\partial \sum_{KK}}^{\mathrm{T}} & \frac{\partial T_{KKij}}{\partial \sum_{11}} & \cdots & \frac{\partial T_{KKij}}{\partial \sum_{KK}} \end{bmatrix},$$

and $\boldsymbol{l}_{22ij}$ is a diagonal matrix with diagonal elements $\{0, \ldots, 0, -1/\hat{q}_j^2, 0, \ldots, 0\}$. In the above matrix,

$$\begin{aligned} \frac{\partial \boldsymbol{S}_{kij}}{\partial \boldsymbol{\eta}_l} &= -\boldsymbol{W}_j \boldsymbol{W}_j^{\mathrm{T}} \boldsymbol{e}_k{}^{\mathrm{T}} \hat{\sum}^{-1} \boldsymbol{e}_l, \\ \frac{\partial \boldsymbol{S}_{kij}}{\partial \sum_{k'l'}} &= -\tfrac{1}{2}\{1+I(k' \neq l')\} \boldsymbol{W}_j \boldsymbol{e}_k^{\mathrm{T}} \hat{\sum}^{-1} (\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'}) \hat{\sum}^{-1} (\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}} \boldsymbol{W}_j), \\ \frac{\partial T_{klij}}{\partial \sum_{k'l'ij}} &= \tfrac{1}{4}\{1+I(k \neq l)\}\{1+I(k' \neq l')\} \left\{ \hat{\sum}^{-1} (\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'}) \hat{\sum}^{-1} \right\}_{kl} \\ & \quad -\tfrac{1}{8}\{1+I(k \neq l)\}\{1+I(k' \neq l')\} (\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}} \boldsymbol{W}_j)^{\mathrm{T}} \\ & \quad \left\{ \hat{\sum}^{-1} (\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'}) \hat{\sum}^{-1} (\boldsymbol{e}_{kl}+\boldsymbol{e}_{lk}) \hat{\sum}^{-1} \right\} (\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}} \boldsymbol{W}_j) \\ & \quad -\tfrac{1}{8}\{1+I(k \neq l)\}\{1+I(k' \neq l')\} (\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}} \boldsymbol{W}_j)^{\mathrm{T}} \\ & \quad \left\{ \hat{\sum}^{-1} (\boldsymbol{e}_{kl}+\boldsymbol{e}_{lk}) \hat{\sum}^{-1} (\boldsymbol{e}_{k'l'}+\boldsymbol{e}_{l'k'}) \hat{\sum}^{-1} \right\} (\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}} \boldsymbol{W}_j). \end{aligned}$$

We then calculate the information matrix as

$$\begin{aligned} \boldsymbol{Q} = -\sum_{i=1}^{N} \sum_{j=1}^{m} \hat{\psi_{ij}} E\{\boldsymbol{l}_{2ij} | \boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j\} - \sum_{i=1}^{N} [\sum_{j=1}^{m} \hat{\psi_{ij}} E\{\boldsymbol{l}_{ij}^{\otimes 2} | \boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j\} \\ - \left(\sum_{j=1}^{m} \hat{\psi_{ij}} E\{\boldsymbol{l}_{1ij} | \boldsymbol{Y}_i^{obs}, \boldsymbol{z}_j, \boldsymbol{g}_j\} \right)^{\otimes 2}]. \end{aligned}$$

To account for the constraint that $\sum_{j=1}^{m} q_j = 1$, we define $\boldsymbol{D}$ to be the derivative matrix of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, q_1, \cdots, q_m)$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, q_1, \cdots, q_{m-1})$. Then, the covariance matrix for $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\Sigma}}, \hat{q_1}, \cdots, \hat{q_{m-1}})$ is estimated by $\boldsymbol{\Omega} = \boldsymbol{F}^{-1}$, where $\boldsymbol{F} = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{D}$.

A.3 Asymptotic Properties

Let Θ denote the parameter space of $\boldsymbol{\theta}$, which is a bounded open set in the interior of the domain of $\boldsymbol{\theta}$, and $\mathcal{F}$ denote the space of the joint distributions of $(\boldsymbol{Z}, \boldsymbol{G})$. Let $\boldsymbol{\theta}_0 \in \Theta$ and $F_0 \in \mathcal{F}$ denote the true values of $\boldsymbol{\theta}$ and F. We impose the following regularity conditions and state the asymptotic results in Theorem 1.

Assumption 4

With probability one, $\Pr(R = 1, V_k = V_l = 1 | \boldsymbol{V} \circ \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{G})$ is bounded away from zero, for each pair of k and $l \in \{1, \ldots, K\}$.

Assumption 5

For any nonzero $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, $\Pr(\boldsymbol{\beta}\boldsymbol{G} + \boldsymbol{\gamma}\boldsymbol{Z} = \boldsymbol{0}) < 1$.

Assumption 6

The density function of F_0 is positive in its support and continuously differentiable with respect to a suitable measure.

Theorem 1

Under Assumptions 1–6, $\hat{\boldsymbol{\theta}}$ and $\hat{F}(\cdot, \cdot)$ are consistent in that $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| + \sup_{\boldsymbol{z},\boldsymbol{g}} |\hat{F}(\boldsymbol{z}, \boldsymbol{g}) - F_0(\boldsymbol{z}, \boldsymbol{g})| \to 0$ almost surely. In addition, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semi-parametric efficiency bound.

Proof—The observed-data likelihood given in (2) is similar to the likelihood given in (6) of Lin and Zeng (2006), which pertains to haplotype rather than genotype effects. In (2), $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs} | \boldsymbol{Z}, \boldsymbol{G})$ is the density of a multivariate linear regression model with partial missingness in $\boldsymbol{Y}$, whereas in (6) of Lin and Zeng (2006), $m_g(Y, \boldsymbol{X}; \boldsymbol{\theta})$, which reduces to $P_{a,\boldsymbol{\beta},\xi}(Y | \boldsymbol{X})$ when haplotypes are replaced by genotypes, is the density of a univariate generalized linear model with Y being always observed. If we can verify that Conditions 1–3 for $P_{a,\boldsymbol{\beta},\xi}(\boldsymbol{Y} | \boldsymbol{X})$ in Lin and Zeng (2006) are satisfied by $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs} | \boldsymbol{Z}, \boldsymbol{G})$, we can use Theorem 1 of Lin and Zeng (2006) to show the consistency, asymptotic normality, and asymptotic efficiency of our estimators.

Before verifying Conditions 1–3 in Lin and Zeng (2006), we need some additional notation. Suppose that there are s distinct missing patterns in $\boldsymbol{Y}$, each with a positive probability of being observed. Let δ_t be the indicator of the tth missing pattern. Let $\boldsymbol{Y}^{obs(t)}$ and $\boldsymbol{Y}^{mis(t)}$ denote the observed and missing parts of $\boldsymbol{Y}$ for the tth missing pattern, $t = 1, \ldots, s$. Then $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs} | \boldsymbol{Z}, \boldsymbol{G})$ can be rewritten as $\prod_{t=1}^{s} \{f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)} | \boldsymbol{Z}, \boldsymbol{G})\}^{\delta_t}$.

Condition 1 in Lin and Zeng (2006) pertains to the identifiability of the regression model. Suppose that two sets of parameters $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ yield the same likelihood value. Then $\prod_{t=1}^{s} \{f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)} | \boldsymbol{Z}, \boldsymbol{G})\}^{\delta_t} = \prod_{t=1}^{s} \{f_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{Y}^{obs(t)} | \boldsymbol{Z}, \boldsymbol{G})\}^{\delta_t}$ for sequenced individuals. By Assumption 4, we can find, for each pair of k and $l \in \{1, \ldots, K\}$, some $t_0 \in \{1, \ldots, s\}$, such that Y_k and Y_l are observed in the t_0th missing pattern. Setting $\delta_{t_0} = 1$, $\delta_t = 0$, and $t \neq t_0$, we have $f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t_0)} | \boldsymbol{Z}, \boldsymbol{G}) = f_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{Y}^{obs(t_0)} | \boldsymbol{Z}, \boldsymbol{G})$, where both sides are multivariate normal densities. Because Y_k and Y_l are components of $\boldsymbol{Y}^{obs(t_0)}$, we have $\boldsymbol{\eta}^k = \tilde{\boldsymbol{\eta}}^k$, $\boldsymbol{\eta}^l = \tilde{\boldsymbol{\eta}}^l$, $\Sigma_{kk} = \tilde{\Sigma}_{kk}$, $\Sigma_{ll} = \tilde{\Sigma}_{ll}$, and $\Sigma_{kl} = \tilde{\Sigma}_{kl}$. Condition 1 in Lin and Zeng (2006) is verified.

Conditions 2 and 3 in Lin and Zeng (2006) are the same if we replace haplotypes by genotypes. Thus, it remains to show that the information operator for $\boldsymbol{\theta}$ and F is

continuously invertible at the true parameter values. This is tantamount to showing that the score function at any non-trivial submodel is non-zero because the information operator is the sum of an invertible operator and a compact operator mapping the score space of (θ_0, F_0) to itself. To this end, suppose that there exists a constant vector $\boldsymbol{u}$, such that

$$\boldsymbol{u}^{\mathrm{T}}\left\{\sum_{t=1}^{s}\delta_t\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)}|\boldsymbol{Z},\boldsymbol{G})\right\}=0. \quad \text{(A.1)}$$

Let $\boldsymbol{b}^{(t)}\equiv(b_1^{(t)},\ldots,b_K^{(t)})^{\mathrm{T}}=\{D(\boldsymbol{V}^{(t)})\sum D(\boldsymbol{V}^{(t)})\}^{+}\{\boldsymbol{V}^{(t)}\circ(\boldsymbol{Y}-\boldsymbol{\eta}\boldsymbol{W})\}$, where $\boldsymbol{V}^{(t)}$ represents $\boldsymbol{V}$ in the tth missing pattern, $D(\boldsymbol{V}^{(t)})$ represents the diagonal matrix with the diagonal vector being $\boldsymbol{V}^{(t)}$, and $\boldsymbol{A}^{+}$ represents the Moore-Penrose generalized inverse of any square matrix $\boldsymbol{A}$. Then

$$\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\boldsymbol{Y}^{obs(t)}|\boldsymbol{Z},\boldsymbol{G})=\left[(\boldsymbol{S}_1^{(t)})^{\mathrm{T}},\ldots,(\boldsymbol{S}_K^{(t)})^{\mathrm{T}},T_{11}^{(t)},T_{12}^{(t)},\ldots,T_{KK}^{(t)}\right]^{\mathrm{T}},$$

where $\boldsymbol{S}_k^{(t)}=\boldsymbol{W}b_k^{(t)}$, and

$$T_{kl}^{(t)}=-\frac{1}{2}\{1+I(k\neq l)\}[\{D(\boldsymbol{V}^{(t)})\sum D(\boldsymbol{V}^{(t)})\}^{+}]_{kl}+\frac{1}{2}\{1+I(k\neq l)\}b_k^{(t)}b_l^{(t)},\quad k\leq l.$$

By Assumption 4, we can find, for each pair of k and $l\in\{1,\ldots,K\}$, $k\leq l$, some $t_0\in\{1,\ldots,s\}$, such that $V_k^{(t_0)}=V_l^{(t_0)}=1$. Set $\delta_{t_0}=1$, $\delta_t=0$, and $t\neq t_0$. Since Y_k and Y_l can take arbitrary values and $b_k^{(t_0)}$ and $b_l^{(t_0)}$ are non-degenerate linear functions of Y_k and Y_l, we see that $b_k^{(t_0)}$ and $b_l^{(t_0)}$ can take arbitrary values. By examining the linear and quadratic terms of $b_k^{(t_0)}$ and $b_l^{(t_0)}$ in equation (A.1), we conclude that their corresponding coefficients must be zero. That is, $\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{W}=0$, $\boldsymbol{u}_l^{\mathrm{T}}\boldsymbol{W}=0$, and $u_{kl}=0$, where $\boldsymbol{u}_k$, $\boldsymbol{u}_l$, and u_{kl} are the components of $\boldsymbol{u}$ associated with $\boldsymbol{S}_k^{(t)}$, $\boldsymbol{S}_l^{(t)}$, and $T_{kl}^{(t)}$, respectively. By Assumption 5, $\boldsymbol{u}_k=\boldsymbol{0}$ and $\boldsymbol{u}_l=\boldsymbol{0}$. It follows that $\boldsymbol{u}=\boldsymbol{0}$. Thus, the score function is non-zero at any non-trivial submodel, and Conditions 2 and 3 in Lin and Zeng (2006) hold.

Remark

Condition 1 suggests that we need to observe with positive probability each pair of components of $\boldsymbol{Y}$ in some individuals selected for sequencing in order for the MLE method to be applicable. We do not require a fully-observed $\boldsymbol{Y}$ for any individual. On the other hand, both the CHARGE-TSS ARIC data and NHLBI ESP EA data contain a large proportion of sequenced individuals with fully-observed $\boldsymbol{Y}$. Thus, Condition 1 is not an issue but mainly serves theoretical purposes.

A.4 Association Tests

For Wald tests employed in single-variant analysis, we estimate all parameters under the alternative hypothesis. Suppose that we decompose $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}_a^{\mathrm{T}}, \boldsymbol{\beta}_b^{\mathrm{T}})^{\mathrm{T}}$ and wish to test the null hypothesis H_0^a:$\boldsymbol{\beta}_a$=**0**. The Wald test statistic is $T_a \equiv \hat{\boldsymbol{\beta}}_a^{\mathrm{T}} \boldsymbol{\Omega}_{aa}^{-1} \hat{\boldsymbol{\beta}}_a$, where $\hat{\boldsymbol{\beta}_a}$ is the MLE of $\boldsymbol{\beta}_a$, and $\boldsymbol{\Omega}_{aa}$ is the covariance matrix of $\hat{\boldsymbol{\beta}_a}$, which is the submatrix of $\boldsymbol{\Omega}$ corresponding to $\boldsymbol{\beta}_a$. We refer T_a to the $\chi^2_{d_a}$ distribution, with the degree of freedom d_a being the dimension of $\boldsymbol{\beta}_a$. In particular, to test the genetic effect on each trait, we consider the null hypothesis $H_0^{(k)}$:β_k=0 for k = 1, …, K. The test statistic is $T_k \equiv \hat{\beta}_k^2 / \Omega_{kk}$, where Ω_{kk} is the variance estimate of $\hat{\beta_k}$. We refer T_k to the χ^2_1 distribution.

Gene-level tests for rare variants rely on score statistics. To test the global null hypothesis that there is no genetic effect on any trait, i.e. H_0: $\boldsymbol{\beta}$ = **0**, we calculate the restricted MLE of ($\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}$, q_1, ⋯, q_{m-1}) under H_0. This is accomplished through the above EM algorithm in which $\boldsymbol{\beta}$ is set to **0** and only ($\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}$, q_1, ⋯, q_{m-1}) is estimated. The score statistic for testing H_0: $\boldsymbol{\beta}$ = **0** is $\boldsymbol{U}_1 \equiv \sum_{i=1}^{N}\sum_{j=1}^{m} \hat{\psi}_{ij} \boldsymbol{l}_{1ij}^{(1)}$, where $\boldsymbol{l}_{1ij}^{(1)}$ is the subvector of $\boldsymbol{l}_{1ij}$ corresponding to $\boldsymbol{\beta}$. It can be shown that $\boldsymbol{U}_1$ is asymptotically normal with mean **0** and covariance matrix $\boldsymbol{V}_1 = \boldsymbol{F}_{11} - \boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}$, where $\begin{bmatrix} \boldsymbol{F}_{11} & \boldsymbol{F}_{12} \\ \boldsymbol{F}_{21} & \boldsymbol{F}_{22} \end{bmatrix}$ is the partition of $\boldsymbol{F}$ with respect to $\boldsymbol{\beta}$ and the other parameters.

For T1 and T5 tests, G is the total number of mutations among variants whose MAFs are below 1% and 5%, respectively. For the MB test, G is the weighted sum of mutations with weights defined as $\{\mathrm{MAF}(1-\mathrm{MAF})\}^{-1/2}$ for each variant (Madsen and Browning 2009). For the above three tests, G is a scalar, and d = 1. The test statistic for testing H_0: $\boldsymbol{\beta}$ = **0** is $T_{(1)} \equiv \boldsymbol{U}_1^T \boldsymbol{V}_1^{-1} \boldsymbol{U}_1$. We refer $T_{(1)}$ to the χ^2_K distribution.

For SKAT, $\boldsymbol{G}$ is a vector of the genotypes of individual variants within a gene. A SKAT-type statistic can be defined as $\boldsymbol{Q}_2 \equiv \boldsymbol{U}_1^{\mathrm{T}} \boldsymbol{B} \boldsymbol{U}_1$, where $\boldsymbol{B}$ is a diagonal matrix of weights that depend on the MAFs through a beta function. The null distribution of $\boldsymbol{Q}_2$ is approximated by $\sum_{j=1}^{Kd} \lambda_j \chi^2_{1,j}$, where ($\lambda_1$, …, λ_{Kd}) are the eigenvalues of $\boldsymbol{V}_1^{1/2} \boldsymbol{B} \boldsymbol{V}_1^{1/2}$, and ($\chi^2_{1,1}, \ldots, \chi^2_{1,Kd}$) are independent χ^2_1 random variables (Wu et al. 2011).

To test the genetic effect on a particular trait, say, the k_0th trait, i.e. H_0: $\boldsymbol{\beta}_{k_0}$ = **0**, where $\boldsymbol{\beta}_{k_0}$ is the k_0th row of $\boldsymbol{\beta}$ reflecting the genetic effect on the k_0th trait, we estimate ($\{\boldsymbol{\eta}_k\}_{k=1,\ldots,K,k\neq k_0}$, $\boldsymbol{\gamma}_{k_0}$, $\boldsymbol{\Sigma}$, q_1, ⋯, q_{m-1}) under H_0. This is accomplished through the above EM algorithm (with a modified M-step) in which $\boldsymbol{\beta}_{k_0}$ is set to **0** and only ($\{\boldsymbol{\eta}_k\}_{k=1,\ldots,K,k\neq k_0}$, $\boldsymbol{\gamma}_{k_0}$, $\boldsymbol{\Sigma}$, q_1, ⋯, q_{m-1}) is estimated. The M-step for estimating $\boldsymbol{\eta}$ is

$$\left[\hat{\boldsymbol{\eta}}_1^{(t+1)},\ldots,\hat{\boldsymbol{\gamma}}_{k_0}^{(t+1)},\ldots,\hat{\boldsymbol{\eta}}_K^{(t+1)}\right]^{\mathrm{T}}=\left[\boldsymbol{A}^{\mathrm{T}}\left\{\left(\hat{\sum}^{(t)}\right)^{-1}\otimes\left(\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{\psi}_{ij}^{(t)}\boldsymbol{W}_j^{\otimes 2}\right)\right\}\boldsymbol{A}\right]^{-1}$$
$$\boldsymbol{A}^{\mathrm{T}}\left[\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{\psi}_{ij}^{(t)}\left\{\left(\hat{\sum}^{(t)}\right)^{-1}\otimes\boldsymbol{W}_j\right\}E\{\boldsymbol{Y}_i|\,\boldsymbol{Y}_i^{obs},\boldsymbol{z}_j,\boldsymbol{g}_j;\hat{\boldsymbol{\theta}}^{(t)}\}\right],$$

where $\boldsymbol{A}$ is a $pK\times(pK-1)$ matrix constructed by deleting the $\{p(k_0-1)+1\}$th column of the $pK\times pK$ identity matrix $\boldsymbol{I}_{pK}$, and $\boldsymbol{A}\otimes\boldsymbol{B}$ denotes the Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. The score statistic for testing H_0: $\beta_{k0}=\mathbf{0}$ is $\boldsymbol{U}_2\equiv\sum_{i=1}^{N}\sum_{j=1}^{m}\hat{\psi}_{ij}\boldsymbol{l}_{1ij}^{(21)}$, where $\begin{bmatrix}\boldsymbol{l}_{1ij}^{(21)}\\ \boldsymbol{l}_{1ij}^{(22)}\end{bmatrix}$ and $\begin{bmatrix}\boldsymbol{F}_{11}^{(2)} & \boldsymbol{F}_{12}^{(2)}\\ \boldsymbol{F}_{21}^{(2)} & \boldsymbol{F}_{22}^{(2)}\end{bmatrix}$ are the partitions of $\boldsymbol{l}_{1ij}$ and $\boldsymbol{F}$ with respect to $\boldsymbol{\beta}_{k0}$ and the other parameters. It can be shown that $\boldsymbol{U}_2$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{V}_2\equiv\boldsymbol{F}_{11}^{(2)}-\boldsymbol{F}_{12}^{(2)}\left(\boldsymbol{F}_{22}^{(2)}\right)^{-1}\boldsymbol{F}_{21}^{(2)}$. All tests of H_0: $\boldsymbol{\beta}_{k0}=\mathbf{0}$ can be constructed in a similar manner. For SKAT tests, we use the vector of genotypes of individual variants as the genetic variables for the k_0th trait and use the burden scores for other traits to ensure numerical stability.

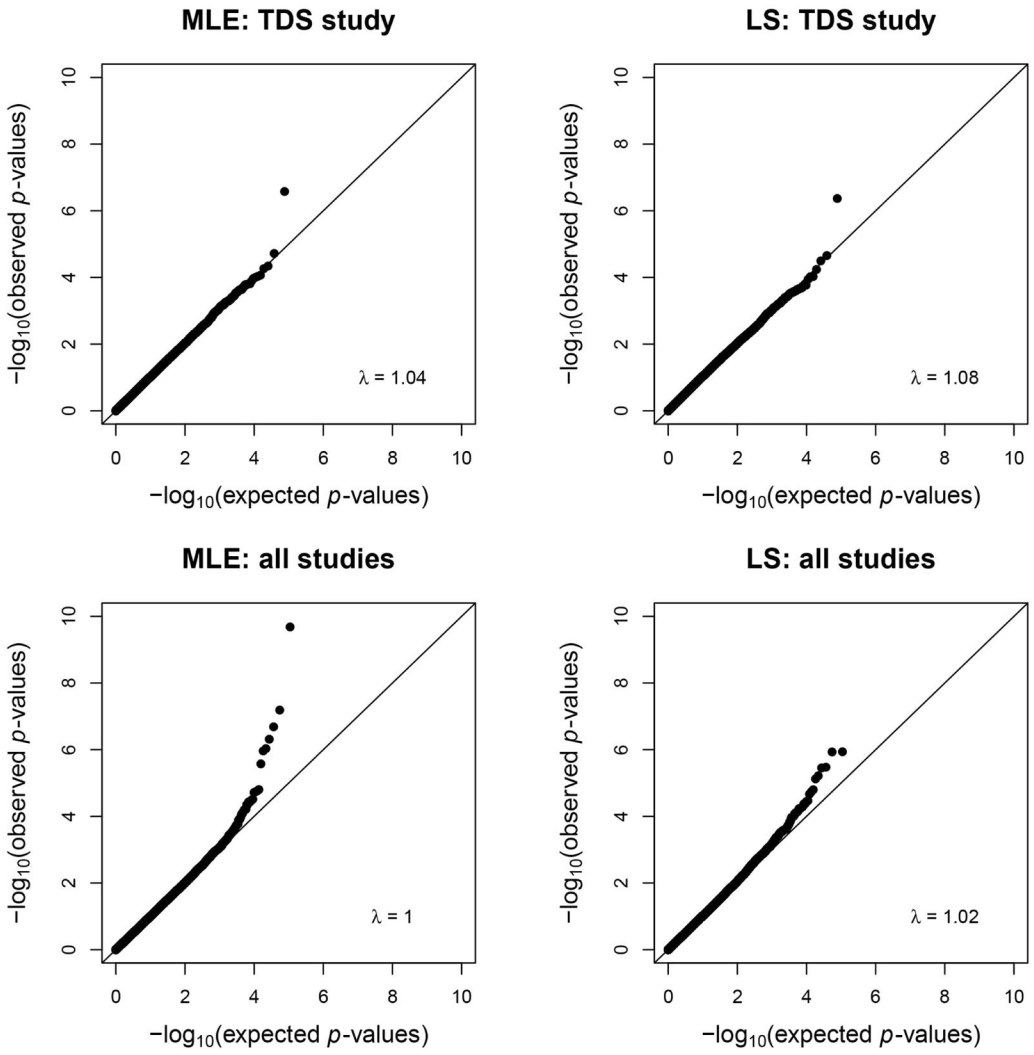

Figure 1.

Quantile-quantile plots for the single-variant analysis of the LDL data using the MLE and LS methods in the TDS study only and in all four studies included in the NHLBI ESP EA sample. The values of the genomic control λ, defined as the ratio between the observed median of the test statistics and the median of the χ_1^2 distribution, are also shown.

Table 1

Simulation Results for Estimating the Genetic Effects on Trait 1 (Null Effect) and Trait 2 (Positive Effect) in Case 1, Five Traits with the Same Effect

			MLE				LS			
Trait	***γ***	***r***	**Bias**	**SE**	**SEE**	**Power**	**Bias**	**SE**	**SEE**	**Power**
1	0.0	0.00	0.000	0.048	0.048	0.0010	−0.018	0.059	0.060	0.0014
		0.05	0.000	0.049	0.049	0.0010	−0.015	0.059	0.059	0.0012
		0.10	0.000	0.050	0.049	0.0011	−0.010	0.058	0.059	0.0010
		0.20	0.000	0.050	0.050	0.0010	−0.007	0.058	0.059	0.0010
		0.50	0.001	0.050	0.050	0.0010	0.002	0.058	0.059	0.0008
	0.3	0.00	0.000	0.044	0.044	0.0010	−0.026	0.053	0.053	0.0023
		0.05	0.000	0.044	0.044	0.0008	−0.028	0.053	0.053	0.0026
		0.10	0.000	0.045	0.045	0.0010	−0.032	0.052	0.053	0.0032
		0.20	0.000	0.046	0.046	0.0010	−0.032	0.052	0.053	0.0031
		0.50	0.000	0.048	0.048	0.0009	−0.034	0.051	0.053	0.0030
2	0.0	0.00	0.000	0.048	0.048	0.817	0.033	0.060	0.059	0.732
		0.05	0.000	0.048	0.048	0.805	0.033	0.059	0.059	0.743
		0.10	0.000	0.049	0.049	0.793	0.033	0.059	0.059	0.749
		0.20	0.001	0.049	0.049	0.775	0.031	0.058	0.058	0.753
		0.50	0.001	0.051	0.051	0.744	0.024	0.057	0.058	0.722
	0.3	0.00	0.000	0.044	0.043	0.902	0.018	0.053	0.053	0.799
		0.05	0.000	0.044	0.044	0.888	0.013	0.053	0.052	0.780
		0.10	0.000	0.045	0.045	0.876	0.009	0.052	0.052	0.761
		0.20	0.000	0.046	0.046	0.854	0.002	0.052	0.052	0.723
		0.50	0.000	0.049	0.049	0.799	−0.013	0.051	0.052	0.611

NOTE: SE and SEE stand for standard error and standard error estimate, respectively.

Table 2

Simulation Results for Estimating the Genetic Effects on Trait 1 (Null Effect), Trait 2 (Positive Effect), and Trait 3 (Negative Effect) in Case 2, Six Traits with Opposite Effects

			MLE				LS			
Trait	γ	r	**Bias**	**SE**	**SEE**	**Power**	**Bias**	**SE**	**SEE**	**Power**
1	0.0	0.00	0.000	0.050	0.050	0.0011	−0.003	0.061	0.061	0.0010
		0.05	0.000	0.050	0.050	0.0011	−0.003	0.061	0.061	0.0011
		0.10	0.000	0.050	0.050	0.0010	−0.003	0.061	0.061	0.0010
		0.20	0.000	0.051	0.051	0.0010	−0.002	0.060	0.060	0.0009
		0.50	0.000	0.050	0.050	0.0010	−0.001	0.060	0.060	0.0009
	0.3	0.00	0.000	0.045	0.045	0.0011	−0.008	0.054	0.054	0.0012
		0.05	0.000	0.045	0.045	0.0010	−0.011	0.054	0.054	0.0011
		0.10	0.000	0.046	0.046	0.0011	−0.014	0.053	0.054	0.0013
		0.20	0.000	0.047	0.046	0.0010	−0.020	0.054	0.054	0.0018
		0.50	0.000	0.049	0.049	0.0010	−0.025	0.052	0.053	0.0018
2	0.0	0.00	0.000	0.049	0.049	0.792	0.052	0.062	0.061	0.787
		0.05	0.000	0.049	0.049	0.782	0.048	0.062	0.061	0.780
		0.10	0.001	0.050	0.049	0.773	0.045	0.061	0.060	0.772
		0.20	0.001	0.050	0.050	0.762	0.038	0.060	0.060	0.751
		0.50	0.001	0.051	0.050	0.753	0.020	0.059	0.059	0.668
	0.3	0.00	0.000	0.044	0.044	0.888	0.037	0.055	0.054	0.859
		0.05	0.000	0.045	0.045	0.875	0.031	0.054	0.054	0.840
		0.10	0.000	0.046	0.046	0.862	0.025	0.054	0.054	0.818
		0.20	0.000	0.047	0.047	0.842	0.015	0.053	0.053	0.771
		0.50	0.000	0.049	0.049	0.798	−0.008	0.052	0.053	0.622
3	0	0.00	0.000	0.050	0.050	0.754	−0.044	0.060	0.061	0.759
		0.05	0.000	0.051	0.051	0.746	−0.041	0.060	0.061	0.750
		0.10	0.000	0.051	0.051	0.742	−0.038	0.059	0.061	0.741
		0.20	0.000	0.051	0.051	0.734	−0.031	0.059	0.060	0.718
		0.50	0.000	0.050	0.050	0.747	−0.015	0.058	0.059	0.631
	0.3	0.00	0.000	0.045	0.045	0.873	−0.047	0.053	0.054	0.895

J Am Stat Assoc. Author manuscript; available in PMC 2016 June 01.

			MLE				LS			
Trait	γ	r	**Bias**	**SE**	**SEE**	**Power**	**Bias**	**SE**	**SEE**	**Power**
		0.05	0.000	0.046	0.046	0.861	−0.047	0.053	0.054	0.900
		0.10	0.000	0.046	0.046	0.850	−0.047	0.053	0.054	0.904
		0.20	0.000	0.047	0.047	0.831	−0.046	0.052	0.054	0.907
		0.50	0.000	0.048	0.048	0.798	−0.039	0.051	0.054	0.888

NOTE: SE and SEE stand for standard error and standard error estimate, respectively.

Table 3

Summary of the ARIC Data in the CHARGE-TSS

Trait	No. (%) of non-missing values	Sampling strategy	No. sequenced (No. after QC)
ECG PR interval	8996 (98.82)	high residual	94 (92)
ECG QRS interval	9053 (99.45)	high residual	90 (89)
Blood pressure	9091 (99.87)	high/low residual	93 (89)
Body mass index	9095 (99.91)	high	90 (79)
Fasting insulin	8896 (97.73)	high	94 (94)
C-reactive protein	7211 (79.22)	high residual	93 (90)
Hematocrit	9071 (99.65)	low residual	97 (85)
Retinal venule diameter	7099 (77.99)	high	156 (154)
Carotid wall thickness	8725 (95.85)	high	91 (87)
Pulmonary: FEV_1	8958 (98.41)	low	186 (185)
Pulmonary: FEV_1/FVC	8956 (98.39)		
Stroke		early onset	74 (70)
Random sample			946 (913)
Total	9103 (100.00)		2003 (1927)

NOTE: For the sampling strategy, “high” (“low”) means sampling from the upper (lower) tail of the trait distribution; “residual” indicates that the sampling is based on the residuals of the original values adjusted for covariates.

Table 4

Top 10 SNPs in the Single-Variant Analysis of the BMI Data in the CHARGE-TSS ARIC Sample

			MLE			LS		
Variant ID	**Gene**	**MAF**	**Est**	**SE**	***p*-value**	**Est**	**SE**	***p*-value**
chr02:000649384	*TMEM18*	2.87E-01	1.12E-01	3.21E-02	4.89E-04	1.34E-01	4.04E-02	9.07E-04
chr02:000669959	*TMEM18*	2.98E-01	−1.06E-01	3.19E-02	8.79E-04	−1.34E-01	4.11E-02	1.09E-03
chr12:000547464	*NINJ2*	6.43E-02	−1.96E-01	5.92E-02	8.98E-04	−2.47E-01	7.41E-02	8.38E-04
chr01:068340029	*WLS*	4.94E-01	−9.41E-02	2.86E-02	9.93E-04	−1.17E-01	3.65E-02	1.36E-03
chr02:000648937	*TMEM18*	2.95E-01	1.01E-01	3.23E-02	1.72E-03	1.19E-01	4.07E-02	3.31E-03
chr02:000648595	*TMEM18*	3.00E-01	9.75E-02	3.19E-02	2.27E-03	1.15E-01	4.04E-02	4.43E-03
chr02:000645222	*TMEM18*	1.12E-01	−1.44E-01	4.74E-02	2.47E-03	−1.71E-01	5.96E-02	4.09E-03
chr02:000649218	*TMEM18*	2.60E-01	1.01E-01	3.36E-02	2.59E-03	1.23E-01	4.24E-02	3.76E-03
chr02:000647954	*TMEM18*	2.95E-01	9.83E-02	3.27E-02	2.61E-03	1.15E-01	4.10E-02	4.97E-03
chr02:000648157	*TMEM18*	2.99E-01	9.34E-02	3.20E-02	3.53E-03	1.10E-01	4.04E-02	6.35E-03

NOTE: Variant ID is in “chromosome:position” format, where the positions are based on the reference human genome (NCBI Genome Build 36, 2006). Est and SE stand for the genetic effect estimate and standard error, respectively.

Table 5

Sample Size Summary of the NHLBI ESP EA Data

	With nonmissing LDL					
	LDL	**BP**	**DPR**	**MI**	**Stroke**	**Nonsequenced**
ARIC	172	93	84	136	6	9553
CARDIA	14	66	32	0	0	1530
CHS	15	3	77	43	1	1186
FHS	12	52	34	147	15	2245
MESA	60	19	159	0	7	1310
WHI	46	8	286	156	49	5115
Total	319	241	672	482	78	20939

Table 6

Top 10 SNPs in the Single-Variant Analysis of the LDL Data in the NHLBI ESP EA Sample

			MLE		LS	
Variant ID	**Gene**	**MAC**	**All studies**	**TDS study**	**All studies**	**TDS study**
chr19:045397229	*TOMM40*	132	2.08E-10	2.64E-07	1.17E-06	4.29E-07
chr01:109814880	*CELSR2*	538	6.48E-08	8.57E-05	3.51E-06	9.42E-05
chr12:101685691	*UTP20*	546	2.06E-07	2.45E-04	6.08E-06	2.10E-04
chr12:101685852	*UTP20*	548	4.85E-07	5.25E-04	7.53E-06	4.61E-04
chr12:101693534	*UTP20*	614	9.28E-07	1.62E-03	3.35E-06	1.44E-03
chr12:101776996	*UTP20*	554	1.09E-06	6.76E-04	1.85E-05	6.15E-04
chr19:002039746	*MKNK2*	9	2.66E-06	1.91E-06	1.20E-02	9.17E-06
chr07:121513561	*PTPRZ1*	492	1.57E-05	3.89E-03	5.87E-05	3.84E-03
chr01:186089112	*HMCN1*	916	1.73E-05	1.08E-04	4.28E-03	1.14E-04
chr12:101705477	*UTP20*	560	1.83E-05	3.67E-03	1.05E-04	3.55E-03

NOTE: Variant ID is in “chromosome:position” format, where the positions are based on the human reference sequence (UCSC Genome Browser, hg19).