Title

Simultaneous Modeling of Disease Status and Clinical Phenotypes To Increase Power in Genome-Wide Association Studies

Journal

Genetics, 205(3)

ISSN

0016-6731

Authors

Bilow, Michael
Crespo, Fernando
Pan, Zhicheng
et al.

Publication Date

2017-03-01

DOI

10.1534/genetics.116.198473

Peer reviewed

Simultaneous Modeling of Disease Status and Clinical Phenotypes To Increase Power in Genome-Wide Association Studies

Michael Bilow,* Fernando Crespo,†,‡ Zhicheng Pan,§ Eleazar Eskin,*, and Susana Eyheramendy†,1**
*Department of Computer Science, §Bioinformatics Program, and **Department of Human Genetics, University of California, Los Angeles, California 90095, †Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile 8320000, and ‡El Centro de Desarrollo y Transferencia Tecnológica, Facultad de Ingeniería, Ciencia y Tecnología, Universidad Bernardo O'Higgins, Santiago, Chile 8320000

ABSTRACT Genome-wide association studies have identified thousands of variants implicated in dozens of complex diseases. Most studies collect individuals with and without disease and search for variants with different frequencies between the groups. For many of these studies, additional disease traits are also collected. Jointly modeling clinical phenotype and disease status is a promising way to increase power to detect true associations between genetics and disease. In particular, this approach increases the potential for discovering genetic variants that are associated with both a clinical phenotype and a disease. Standard multivariate techniques fail to effectively solve this problem, because their case–control status is discrete and not continuous. Standard approaches to estimate model parameters are biased due to the ascertainment in case–control studies. We present a novel method that resolves both of these issues for simultaneous association testing of genetic variants that have both case status and a clinical covariate. We demonstrate the utility of our method using both simulated data and the Northern Finland Birth Cohort data.

KEYWORDS multivariate analysis; covariates

GENETIC case–control association studies have found thousands of associations between genetic variants and disease (Spencer *et al.* 2009; Welter *et al.* 2014; Zhou and Stephens 2014). These studies can be designed to include cases and controls in two different ways. First, they can include an equal number of cases and controls. Second, they can include cases and controls obtained randomly from a population cohort. The true prevalence of the disease in the population can be inferred from a population cohort, but not from the design in which there are an equal number of cases and controls. The latter case is strongly influenced by a selection bias, which can affect the estimation of the true genetic effects on the phenotype (Zaitlen *et al.* 2012).

Many case–control association studies consider the genetic association with a single phenotype at a time even if they have collected additional clinical phenotypes, such as body mass index, high-density lipoproteins (HDL) cholesterol, low-density lipoproteins (LDL) cholesterol, smoking habits, *etc*. (Kuo and Feingold 2010). In cases where it is well known that a clinical phenotype affects disease status, such as body mass index on diabetes, most commonly the clinical phenotype is incorporated in the analysis as a covariate. Studies have shown that when the clinical phenotype is correlated with disease status, this approach can lose its power to detect genetic associations (Pirinen *et al.* 2012; Zaitlen *et al.* 2012). Zaitlen *et al.* (2012) propose a model that is based on the liability threshold model, which performs informed conditioning on the clinical phenotypes in order to remedy the deleterious effect that a clinical phenotype has when it is incorporated into the analysis as a covariate. Also, in Zaitlen *et al.* (2012), the model parameters are estimated using external epidemiology data to avoid the problem of selection bias in the design of equally sampled cases and controls.

We propose an alternative approach to incorporate the clinical phenotype into the model. Specifically, we propose a model that jointly assesses the effect of the genetic variant on the clinical phenotype and the disease. Previously,

doi: 10.1534/genetics.116.198473
Manuscript received November 17, 2016; accepted for publication December 19, 2016; published Early Online January 27, 2017.
[1]Corresponding author: Department of Statistics, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago 6904411, Chile. E-mail: susana@mat.puc.cl

several methods have been proposed to identify genetic variants associated with multiple phenotypes (Korte *et al.* 2012; Zhou and Stephens 2014; Furlotte and Eskin 2015). These methods increase power when phenotypes are correlated (Neuhaus 1998; Mefford and Witte 2012) and contribute to the understanding and discovery of pleiotropic effects (Chanock *et al.* 2007; Frayling 2007; Amos *et al.* 2008; Hung *et al.* 2008; Thorgeirsson *et al.* 2008).

However, these previous methods assume that both phenotypes are continuous. Unfortunately, the case–control disease status is coded as variables with discrete values. Often a disease is coded with a one, and a control without the disease is coded with a zero. Performing association for both the case–control status and the clinical covariate is challenging, and only a few methods have been proposed for this scenario (Liu *et al.* 2009; Prerau *et al.* 2009). It is harder to model the correlation structure and perform inference in such a model when one variable is discrete and one is continuous as opposed to when both are continuous. A second challenge is that diseases are typically rare and therefore the design of a case–control study involves substantially oversampling cases compared to a representative population cohort (Kuo and Feingold 2010). This creates a complex distribution of the clinical covariate among the individuals in the study (Bays *et al.* 2007).

In this paper, we present a novel method for simultaneous association of a genetic variant that has both the case–control status and the clinical covariate. Our method combines the liability threshold framework with multivariate methods. Specifically, our method explicitly handles the issue of ascertainment/selection bias by developing an expectation-maximization algorithm for estimating the model parameters, which reweighs the individuals to correct for the ascertainment bias. We demonstrate the increase in statistical power of our approach utilizing both simulated and real datasets. Using simulations, we show that modeling a correlated environmental effect, which impacts both the case–control status and the clinical covariate, significantly increases power compared to traditional approaches. We also show the utility of our results by analyzing multiple phenotypes from the Northern Finland Birth Cohort.

Methods

Liability threshold model

Consider the liability threshold framework for modeling disease status. Denote Y^d a discrete random variable that takes a one for cases and a zero for controls. Assume that the distribution of Y^d is Bernoulli with $\Pr(Y^d = 1) = I(Y > t)$, where $I(x) = 1$ if x is true and $I(x) = 0$ if x is false and Y is a latent (unobserved) random variable representing disease liability. Given a vector of covariates X (including genetic factors), a vector β of the covariate effect sizes, error $e \sim \mathcal{N}(0,1)$, and a mean liability μ, we assume $Y = \mu + X^T\beta + e$. Therefore, the liability Y is normally distributed with a mean equal to $\mu + X^T\beta$ and variance equal to one. Even though $X^T\beta$ can include covariates other than genotypes, for simplicity, we assume that there is only one covariate, the genotype.

Following Zaitlen *et al.* (2012), we assume that the disease prevalence in the population f is known, for example from a prior epidemiological study of the disease. We determine the liability threshold t using $t = -\mathbf{\Phi}^{-1}(1-f)$, where $\mathbf{\Phi}(x)$ is the standard normal cumulative density function.

We can now write the log-likelihood of the observed and latent variables, the so-called complete log-likelihood, as:

$$\log \mathcal{L}(\beta|\mathbf{Y},X) = \log\left(\prod_{j=1}^{n} I(Y_j>0)^{Y_j^d}(1-I(Y_j>0))^{1-Y_j^d} \times 1/(\sqrt{2\pi})e^{-\frac{1}{2}(Y_j-\mu-X_j\beta)^2}\right)$$
$$= \log\left(\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(Y_j-\mu-X_j\beta)^2}\right).$$

To assess the association of the genetic variant with the disease, we perform a likelihood ratio test (LRT) in which the null hypothesis, $H_0 : \beta = 0$ assumes no association while the alternative hypothesis $H_1 : \beta \neq 0$ assumes that the association is different from zero, where β is the genetic effect in the threshold liability model.

The LRT under the null hypothesis has chi-squared distribution with one degree of freedom and is calculated as follows:

$$2\left(\log\mathcal{L}\left(\beta=\hat{\beta}|\hat{\mathbf{Y}},X\right)-\log\mathcal{L}\left(\beta=0|\hat{\mathbf{Y}},X\right)\right) \sim \chi_1^2. \quad (1)$$

In this model the expected liability and the parameters of the model are estimated using an expectation-maximization (EM) algorithm.

Multiple phenotype model for two continuous phenotypes

Let Y_1, Y_2 be two phenotypes represented as $N \times 1$ continuous-valued column vectors. The values of Y_1 and Y_2 for each individual j are denoted Y_{1j} and Y_{2j}, respectively. Let the true effect size of $\mathbf{X}$ on Y_1 be β_1 and the true effect size on Y_2 be β_2. Let μ_1 and μ_2 be the true means of Y_1 and Y_2, respectively. Then we can represent the two phenotypes as follows:

$$\begin{aligned} Y_1 &= \mu_1 + \sum_{i=1}^{m}\mathbf{X}\beta_{1i} + \mathbf{e}_1 \\ Y_2 &= \mu_2 + \sum_{i=1}^{m}\mathbf{X}\beta_{2i} + \mathbf{e}_2 \end{aligned} \quad (2)$$

The distribution of the errors, $\mathbf{e}_1$ and $\mathbf{e}_2$, is a bivariate normal distribution with mean zero, variance equal to σ_1^2 and σ_2^2, respectively, and covariance equal to $\rho\sigma_1\sigma_2$.

$$\begin{pmatrix} \mathbf{e}_{1j} \\ \mathbf{e}_{2j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}\right) \tag{3}$$

We can express the log-likelihood of the data in the following way:

$$\begin{aligned} &\log\mathcal{L}(\beta_1,\beta_2,\mu_1,\mu_2,\rho|Y_1,Y_2,X) \\ &\quad = -\frac{m}{2}\log\Big((2\pi)^2\sigma_1^2\sigma_2^2(1-\rho^2)\Big) \\ &\quad - \sum_{j=1}^{m}\left[\frac{1}{2}\begin{pmatrix} Y_{1j}-X_j\beta_1-\mu_1 \\ Y_{2j}-X_j\beta_2-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}^{-1}\right. \\ &\quad \left.\times \begin{pmatrix} Y_{1j}-X_j\beta_1-\mu_1 \\ Y_{2j}-X_j\beta_2-\mu_2 \end{pmatrix}\right] \end{aligned} \tag{4}$$

Consider again an LRT to assess the association of genetic variants with the continuous phenotypes. The likelihood of the model under the null hypothesis $H_0 : \beta_1 = 0, \beta_2 = 0$ is compared with the likelihood of the model under the alternative hypothesis $H_1 : \beta_1 = \hat{\beta}_1, \beta_2 = \hat{\beta}_2$, by calculating:

$$\begin{aligned} &2(\log\mathcal{L}(\beta_1 = \hat{\beta}_1, \mu_1 = \hat{\mu}_1, \beta_2 = \hat{\beta}_2, \mu_2 \\ &\quad = \hat{\mu}_2, \rho = \hat{\rho}|Y_1,Y_2,X) - \log\mathcal{L}(\beta_1 = 0, \mu_1 \\ &\quad = \hat{\mu}_1, \beta_2 = 0, \mu_2 = \hat{\mu}_2, \rho = \hat{\rho}|Y_1,Y_2,X)]) \sim \chi_2^2 \end{aligned} \tag{5}$$

which asymptotically distributes as χ_2^2 under the null hypothesis.

Extending the liability threshold framework to model discrete and continuous phenotypes simultaneously: the BinCont model

In this study, we assume Y_1 is an observed continuous phenotype and Y_2 is a latent (unobserved) disease liability. For each individual (indexed by j), instead of observing the continuous-valued Y_{2j}, we observe case–control status Y_{2j}^d. As before, we denote cases as $Y_{2j}^d = 1$ and controls as $Y_{2j}^d = 0$. Each individual's case–control status depends on their liability Y_{2j} and a liability threshold value t:

$$Y_{2j}^d = \begin{cases} 1 & \text{if } Y_{2j} > t \\ 0 & \text{if } Y_{2j} \le t \end{cases} \tag{6}$$

In other words, Y_{2j}^d has Bernoulli distribution with parameters equal to $Pr(Y_{2j}^d = 1) = I(Y_{2j} > t)$ where $I()$ is an indicator function.

For each individual j, the joint distribution of Y_{2j} and Y_{1j} is a bivariate normal defined by:

$$\begin{pmatrix} Y_{1j} \\ Y_{2j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 + X_j\beta_1 \\ \mu_2 + X_j\beta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{7}$$

The log-likelihood of this model is given by

$$\begin{aligned} &\log\mathcal{L}(\beta_1,\beta_2,\mu_1,\mu_2,\rho|Y_1,Y_2,X) \\ &\quad = \log\Bigg(\prod_{j=1}^{n} I(Y_{2j}>0)^{Y_{2j}^d}(1-I(Y_{2j}>0))^{1-Y_{2j}^d} \\ &\qquad \times f(Y_{1j},Y_{2j}|\beta_1,\beta_2,\mu_1,\mu_2,\rho,X_j)\Bigg) \\ &\quad = \log\Bigg(\prod_{j=1}^{n} f(Y_{1j},Y_{2j}|\beta_1,\beta_2,\mu_1,\mu_2,\rho,X_j)\Bigg) \end{aligned}$$

where $f(Y_{1j},Y_{2j}|X\beta_1,X\beta_2,\mu_1,\mu_2,\rho)$ is the same bivariate normal distribution of Equation 4. An LRT is considered again to assess the null hypothesis $H_0 : \beta_{1\ \text{SNP}} = 0, \beta_{2\ \text{SNP}} = 0$ of no association between any of the phenotypes and the genetic variant *vs.* the alternative hypothesis $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$. The LRT is expressed as:

$$\begin{aligned} &2(\log\mathcal{L}(\beta_1 = \hat{\beta}_1, \mu_1 = \hat{\mu}_1, \beta_2 = \hat{\beta}_2, \mu_2 \\ &\quad = \hat{\mu}_2, \rho = \hat{\rho}|Y_1,Y_2,X) - \log\mathcal{L}(\beta_1 = 0, \mu_1 \\ &\quad = \hat{\mu}_1, \beta_2 = 0, \mu_2 = \hat{\mu}_2, \rho = \hat{\rho}|Y_1,Y_2,X)]) \sim \chi_2^2 \end{aligned} \tag{8}$$

which asymptotically distributes as χ_2^2 under the null hypothesis.

Extending the BinCont model to overcome selection bias: the BinContSelection model

In order to correct for selection bias, we reweigh the individuals. The weight of each control individual is set to $2(1-f)$, and the weight of each case individual is set to $2f$. Note that since the controls are undersampled, their weights are bigger than 1 while the opposite is true of the cases, and the sum of the weights of all the $n/2$ controls plus all the $n/2$ cases is n, the total sample size. We then use the weights when estimating the parameters in the model. A graphical description of this model is shown in Figure 1.

Inferring model parameters using expectation maximization

We estimate the parameters of our BinContSelection model using the EM algorithm. The EM algorithm is an iterative algorithm for finding maximum likelihood estimators in the presence of missing data. The algorithm alternates between two steps until convergence: the expectation step (or E-step) and the maximization step (or M-step). In the E-step we compute the conditional expectation of the complete log-likelihood of the model given the observed data. In the M-step the parameters are estimated using maximum likelihood.

Initial conditions: We initialize the parameter of the model to be equal to zero, $\beta_1 = \beta_2 = 0$, and the expected liability $\hat{Y}_2^{(0)}$ is initialized with the conditional expectation given the observed disease status (*i.e.*, $\mathrm{E}(Y_{2j}|Y_{2j}^d)$, obtained from the univariate liability threshold model.

E-step: The E-step consists of computing the expected value of the complete log-likelihood (*i.e.*, the log-likelihood of the

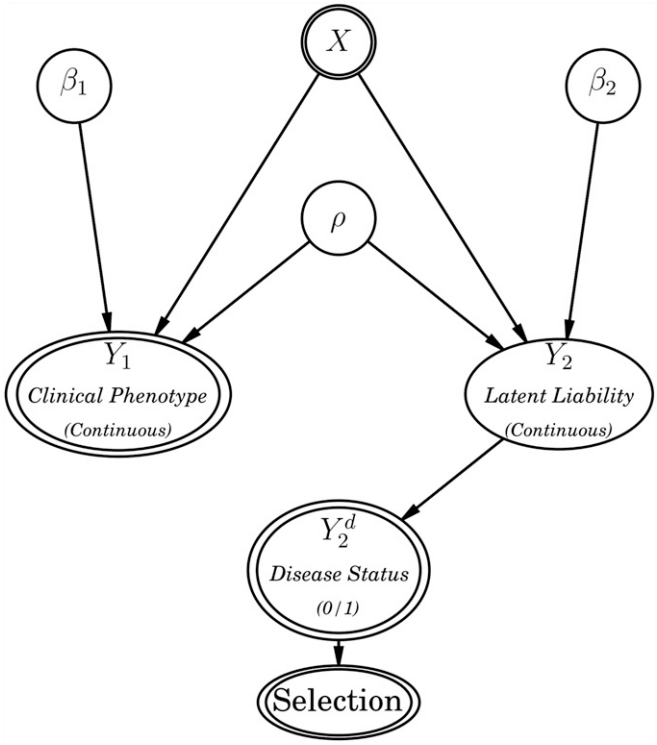

Figure 1 Graphical representation of our model.

observed and latent random variables) of the model given the observed data and the current estimates of the parameters. From Equations 4 and 8, and a bit of algebra, all that remains to be estimated in this step is: $\mathbb{E}[Y_2|Y_1, Y_2^d, \beta_1^{(t)}, \beta_2^{(t)}, \rho^{(t)}] = \hat{Y}_2^{(t+1)}$. Therefore, in the $t+1$ iteration we estimate:

$$\hat{Y}_{2j}^{(t+1)} = X\hat{\beta}_2^{(t)} + \begin{cases} \frac{1}{\boldsymbol{\Phi}_2\left(-\infty, Y_1 - X\hat{\beta}_1, -X\hat{\beta}_2^{(t)}, \infty\right)} \left[\phi\left(-X\hat{\beta}_2^{(t)}\right)\boldsymbol{\Phi}\left(\frac{\mu_{1*}^{(t)}}{\sigma_*^{(t)}}\right) - \right. \\ \left. \hat{\rho}^{(t)}\phi(Y_1 - X\hat{\beta}_1)\left(1 - \boldsymbol{\Phi}\left(-\mu_{2*}^{(t)}/\sigma_*^{(t)}\right)\right)\right] \text{ if } Y_{2j}^d = 1 \\ \frac{-1}{\boldsymbol{\Phi}_2\left(-\infty, Y_1 - X\hat{\beta}_1, -\infty, -X\hat{\beta}_2^{(t)}\right)} \left[\phi\left(-X\hat{\beta}_2^{(t)}\right)\boldsymbol{\Phi}\left(\mu_{1*}^{(t)}/\sigma_*^{(t)}\right) + \right. \\ \left. \hat{\rho}^{(t)}\phi(Y_1 - X\hat{\beta}_1)\boldsymbol{\Phi}\left(-\mu_{2*}^{(t)}/\sigma_*^{(t)}\right)\right] \text{ if } Y_{2j}^d = 0 \end{cases} \quad (9)$$

Here, ϕ is the standard normal probability density function, $\boldsymbol{\Phi}_2$ is the standard bivariate normal cumulative distribution function $\mu_{1*}^{(t)} = (Y_1 - X\hat{\beta}_1) + \hat{\rho}^{(t)}X\hat{\beta}_2^{(t)}$: $\mu_{2*}^{(t)} = X\hat{\beta}_2^{(t)} + \hat{\rho}^{(t)}(Y_1 - X\hat{\beta}_1)$ and $\sigma_*^{(t)} = \sqrt{1 - (\hat{\rho}^{(t)})^2}$.

M-Step: In the $t+1$-th iteration of the M-step, we compute the maximum likelihood estimator of $\beta_1^{(t+1)}$ $\hat{\beta}_2^{(t+1)}$ and $\hat{\rho}^{(t+1)}$, using $\hat{Y}_2^{(t+1)}$, and the known weights for cases and controls.

Convergence: We alternate the **E-** and **M-**steps until the estimates for β_1, β_2, and ρ converge.

We consider that the estimates converged if: $\frac{\hat{\beta}_1^{(t)} - \hat{\beta}_1^{(t+1)}}{\hat{\beta}_1^{(t)}} < 10^{-3}, \frac{\hat{\beta}_2^{(t)} - \hat{\beta}_2^{(t+1)}}{\hat{\beta}_2^{(t)}} < 10^{-3}, \frac{\hat{\rho}^{(t)} - \hat{\rho}^{(t+1)}}{\hat{\rho}^{(t)}} < 10^{-3}$.

National Finland Birth Cohort data

The National Finland Birth Cohort 1966 enrolled almost everyone born in 1966 in Finland's two most northern provinces. The North Finland Birth Cohort (NFBC) dataset consists of 10 phenotypes and genotypes at 331,476 genetic variants measured in 5327 individuals. The phenotypes include LDL cholesterol and triglyceride levels (TG), which are used in the analysis.

Data availability

The software implementing the methods described in this paper is available at http://genetics.cs.ucla.edu/multipheno/ and https://github.com/facrespo/BivariateProbitContinueEM.

Results

Overview of the method

We propose a novel approach for incorporating clinical phenotypes into case–control studies where we assume that the genetic variant (X) can affect both the clinical phenotype (Y_1) and the disease status (Y_2^d). We assume the liability threshold model. Each individual has an underlying liability (Y_2) and the disease status is a deterministic function of this liability (*i.e.*, if the liability is greater than a threshold (t) depending on the disease prevalence then the individual has the disease ($Y_2^d = 1$)). Specifically, we assume the underlying model of Figure 1. In our approach, we assume that the genetic variant can have an effect on both the clinical phenotype (β_1) and the disease liability (β_2). We also assume that the clinical phenotype and disease liability have a correlation of ρ.

Given this model and the observed data for a set of individuals, we apply a maximum likelihood approach to estimate the parameters and perform a statistical test of the hypothesis $\beta_1 = \beta_2 = 0$. If we reject this hypothesis, then we declare the genetic variant associated with the clinical phenotype and/or the disease status.

Multivariate analysis significantly improves power in genome-wide association in simulation studies

We compare the performance of our method to traditional approaches through simulations. In our approach, we simulate the case where an SNP affects both the case–control status and a covariate. We simulated studies of 5000 individuals evenly split between cases and controls. We assumed two different prevalences of disease status, 40 and 0.1%, and generated the individuals by sampling from a liability threshold model (see *Methods*). We simulated a single SNP with minor allele frequency 0.2. The effect size of the SNP on the liability (referred to as β_2) was chosen such that a univariate association between the SNP and case–control status would have 50% power. Along with disease status, we simulated a clinical phenotype correlated to the liability as

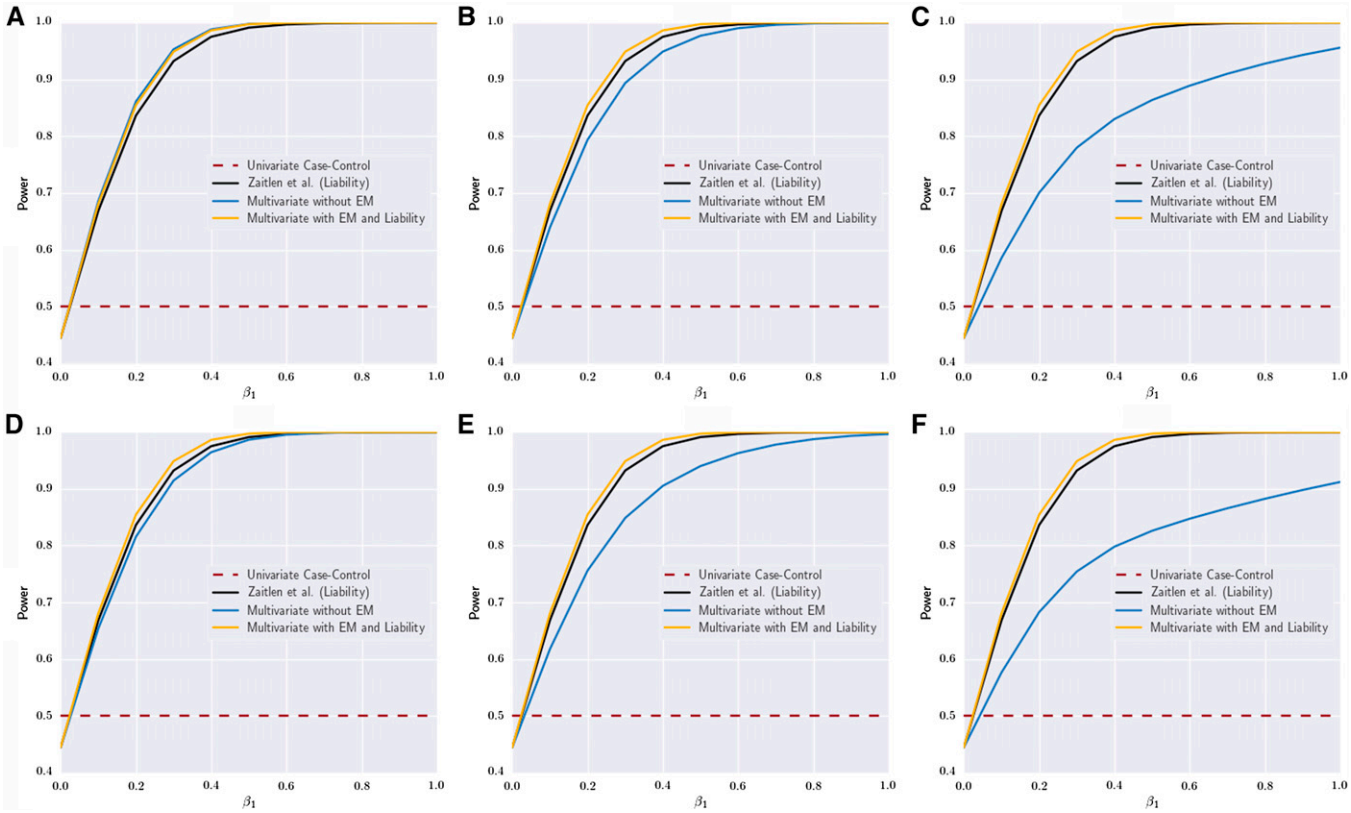

Figure 2 (A–C) Low selection bias ($f = 40\%$); (D–F) high selection bias ($f = 0.1\%$). In A and D, $\rho = .2$; $\rho = 0.5$ in B and E; $\rho = 0.8$ in C and F. The power of each test at each value of β_1 is shown as a line.

described in the *Methods*. The goal of our simulation is to measure the effect of analyzing both the case–control status and the covariate. Therefore, in each simulation we fixed the effect size of the variant on the case–control status (β_2) so that the univariate test of the case–control status will have a power of 0.5. We then vary the effect of the genetic variant on the clinical covariate (β_1).

We evaluate the performance of our method and compare it with three other approaches. The first method is the single univariate test applied to the disease status. The second is a multivariate approach applied to the disease status and clinical phenotype modeled as a multivariate normal distribution. We note that the data clearly violate the assumptions of the multivariate model, because the disease status is discrete and does not follow the normal distribution. The third method is the Zaitlen *et al.* (2012) liability threshold model treating the clinical phenotype as a covariate. We show the results of the simulated data in Figure 2. In each plot, the x axis corresponds to the effect size of the clinical covariant and the y-axis corresponds to the estimated statistical power, as measured by the fraction of 10,000 simulations that achieve a statistically significant association.

The power of a single univariate test on the case–control status is shown in red in Figure 2. The power of a multivariate test using the case–control status and the clinical trait is shown in blue. The Zaitlen *et al.* (2012) liability threshold-based model implemented in the software package LTSOFT is shown as a black line. The power of our method is shown as a gold line. It is also important to note the role of the correlation between the clinical phenotype and the SNP (ρ), which varies among the columns of the figure.

As expected, when the effect of the genetic variant on the clinical covariate is very low, the univariate method outperforms all of the other methods. However, even for modest levels of genetic effects on the clinical covariate, all of the multivariate methods increase the overall power. Our method has the highest power in all scenarios. The advantage of our method is greater when there are substantial amounts of selection bias (D, E, F) compared to lower amounts of selection bias (A, B, C). Similarly, the advantage of our method is more substantial when the correlation between the clinical covariate and the disease liability is lower (A and D *vs.* C and F). This is because the advantage of our method is that we explicitly estimate the underlying liability using all of the data. However, when the correlation is high, the covariate itself is a good approximation for the underlying liability. In this scenario, the multivariate method and our method provide very similar results. We further perform simulations to measure the false-positive rate of the methods by performing simulations where the SNP has no effect on the trait ($\beta_1 = \beta_2 = 0$). In this simulation, since we are exactly simulating the null distribution, we verify that our likelihood ratio statistics

Table 1 P-value comparison of methods in Sabatti *et al.* (2009) and this paper on LDL

	Sabatti *et al.*	Univariate	Multivariate	EM
rs646776	2.19×10^{-12}	8.08×10^{-8}	1.17×10^{-8}	1.11×10^{-9}
rs4844614	2.38×10^{-7}	8.89×10^{-7}	1.61×10^{-7}	1.02×10^{-8}
rs673548	-	6.50×10^{-5}	5.07×10^{-5}	7.60×10^{-7}

For significantly associated SNPs, EM adds power. rs673548 was not reported for LDL in Sabatti *et al.* (2009) but reported a nearby SNP (rs693) which is in high LD and has a *P*-value of 2.99^{-11}. rs673548 was reported for TG.

follow the χ^2 distribution as expected by theory and the false-positive rate is controlled.

Analysis of the NFBC dataset

We demonstrate the utility of our method using data from the NFBC dataset by showing that multiple phenotype analysis can increase power compared to single phenotype analysis in some cases. The idea behind our evaluation strategy is that we will analyze a subset of the NFBC data using both univariate and multivariate strategies and compare what we discover to what was discovered in the full NFBC dataset, which we treat as the gold standard. We evaluate our performance by assessing if we can recover what was discovered in the full dataset by only analyzing a subset of the samples from the NFBC using multiple phenotypes.

In particular, we transformed LDL measurements to case–control data by dichotomizing the top 10% as LDL cases and sampling at random an equal number of individuals from the bottom 90% of LDL as LDL controls. This scheme is in line with the liability threshold model. We used TG as the clinical phenotype. In our subset, we are only considering 20% of the individuals of the full dataset and also have dichotomized the data, both of which reduce the statistical power to discover loci associated with LDL. We first perform univariate association analysis only on the dichotomized LDL data and report the *P*-value. As expected, the *P*-value is much less significant than the same locus in the full NFBC dataset. We then incorporate the TG phenotype into our analysis using both the standard multivariate analysis and our liability threshold modeling approach.

Table 1 reports the three SNPs that are associated with LDL and also have a signal for TG. The univariate test reports the *P*-value when associating the SNP with the dichotomized LDL phenotype. As expected, the *P*-values are much weaker than what was observed in the full dataset. However, running the multivariate approaches incorporating the TG phenotype increased power as evidenced with more significant *P*-values to those SNPs.

Discussion

In this paper we presented a method for incorporating a clinical phenotype into case–control studies under the assumption that the genetic variant can affect both. We apply our method in tandem with a method that incorporates the clinical phenotype as a covariate, such as the method of Zaitlen *et al.* (2012). Intuitively, our method will have higher power to detect genetic variants that affect both phenotypes, while a traditional method would have higher power to detect genetic variants that only affect the disease status.

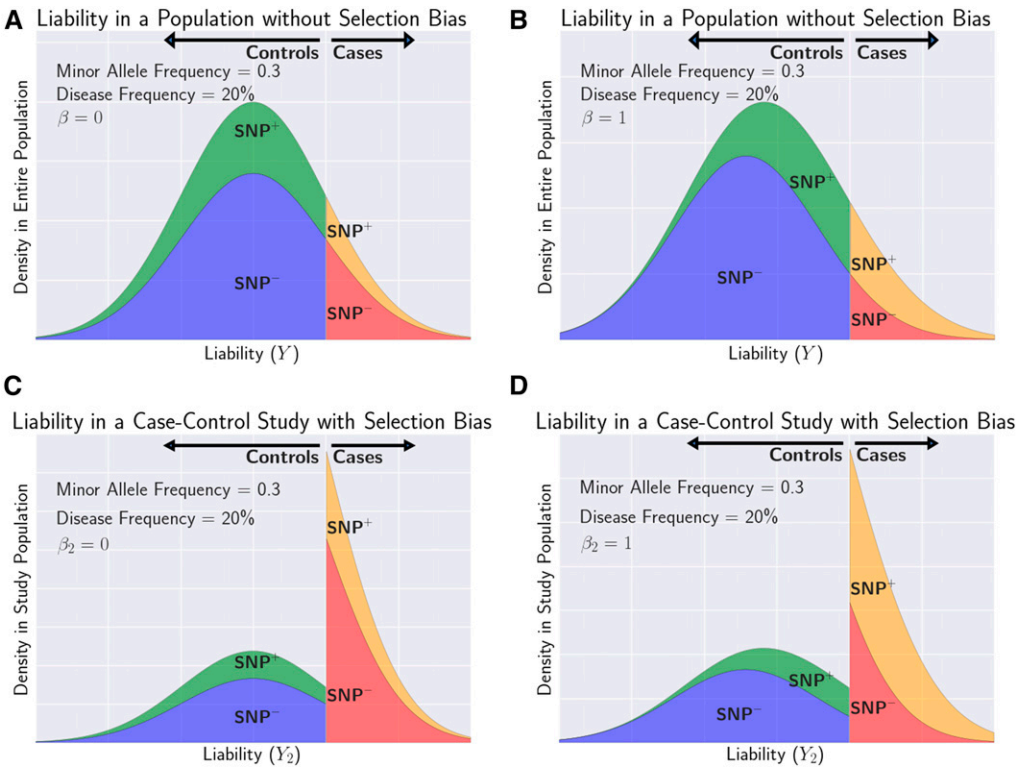

Figure 3 An illustration of the distribution of liability in a case–control study under selection bias. In the figure, the disease has an incidence of 10% in the population. (A and B) Twenty percent of the sample contains individuals with the disease. (C and D) Fifty percent of the sample contains individuals with the disease resulting in a large oversampling of individuals with liability values just over the threshold. In A and C, the SNP does not have an effect on the disease and the frequency of the SNP is the same in each group. In B and D, the SNP affects the disease and a change in frequencies of the SNPs between the cases and controls is observed.

When we model the correlation between the phenotypes, it is critically important to remove the effect of ascertainment or selection bias. The oversampling in case–control studies greatly distorts the distribution of the underlying liability as shown in Figure 3.

We compare our method to a standard multivariate regression approach, which treats the clinical phenotype and the disease status as following the multivariate normal distribution. In the genetics literature, methods that assume continuous traits are often applied to case–control data ignoring this assumption (Price *et al.* 2006; Kang *et al.* 2010). Therefore, we include this comparison method even though the multivariate normal assumptions are clearly violated by the discrete disease status.

Acknowledgments

S.E. and F.C. acknowledge support from Conicyt/Fondap 15110006, and S.E. acknowledges support from Fondecyt 1120813. M.B. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589, and 1331176, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782, and R01-ES022282. E.E. is supported in part by the National Institutes of Health BD2K award, U54EB020403. We acknowledge the support of the National Institute Of Neurological Disorders (NINDS) Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691). No competing financial interests exist.

Literature Cited

Amos, C. I., X. Wu, P. Broderick, I. P. Gorlov, J. Gu *et al.*, 2008 Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet. 40: 616–622.

Bays, H. E., R. H. Chapman, S. Grandy, and S. I. Group, 2007 The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. Int. J. Clin. Pract. 61: 737–747.

Chanock, S. J., T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter *et al.*, 2007 Replicating genotype-phenotype associations. Nature 447: 655–660.

Frayling, T. M., 2007 Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat. Rev. Genet. 8: 657–662.

Furlotte, N. A., and E. Eskin, 2015 Efficient multiple trait association and estimation of genetic correlation using the matrix-variate linear mixed-model. Genetics 200: 59–68.

Hung, R. J., J. D. McKay, V. Gaborieau, P. Boffetta, M. Hashibe *et al.*, 2008 A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452: 633–637.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42: 348–354.

Korte, A., B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long *et al.*, 2012 A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat. Genet. 44: 1066–1071.

Kuo, C.-L. L., and E. Feingold, 2010 What's the best statistic for a simple test of genetic association in a case-control study? Genet. Epidemiol. 34: 246–253.

Liu, J., Y. Pei, C. J. Papasian, and H. W. Deng, 2009 Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet. Epidemiol. 33: 217–227.

Mefford, J., and J. S. Witte, 2012 The covariate's dilemma. PLoS Genet. 8: e1003096.

Neuhaus, J. M., 1998 Theory and methods. J. Am. Stat. Assoc. 93: 1124–1129.

Pirinen, M., P. Donnelly, and C. C. Spencer, 2012 Including known covariates can reduce power to detect genetic effects in case-control studies. Nat. Genet. 44: 848–851.

Prerau, M. J., A. C. Smith, U. T. Eden, Y. Kubota, M. Yanike *et al.*, 2009 Characterizing learning by simultaneous analysis of continuous and binary measures of performance. J. Neurophysiol. 102: 3060–3072.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909.

Sabatti, C., S. K. Service, A.-L. L. Hartikainen, A. Pouta, S. Ripatti *et al.*, 2009 Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat. Genet. 41: 35–46.

Spencer, C. C. A., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5: e1000477.

Thorgeirsson, T. E., F. Geller, P. Sulem, T. Rafnar, A. Wiste *et al.*, 2008 A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 452: 638–642.

Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall *et al.*, 2014 The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42: D1001–D1006.

Zaitlen, N., S. Lindström, B. Pasaniuc, M. Cornelis, G. Genovese *et al.*, 2012 Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet. 8: e1003032.

Zhou, X., and M. Stephens, 2014 Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods 11: 407–409.

Communicating editor: G. A. Churchill