Modeling Hybrid Traits for Comorbidity and Genetic Studies of Alcohol and Nicotine Co-Dependence

Heping Zhang, Dungang Liu, Jiwei Zhao, and Xuan Bi*

January 23, 2018

*Heping Zhang is Susan Dwight Bliss Professor (Email: heping.zhang@yale.edu), Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520; Dungang Liu is Assistant Professor (Email: dungang.liu@uc.edu), Department of Operations, Business Analytics and Information Systems, University of Cincinnati Lindner College of Business, Cincinnati, OH 45221; Jiwei Zhao is Assistant Professor (Email: zhaoj@buffalo.edu), Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY 14214; and Xuan Bi is Postdoctoral Associate (Email: xuan.bi@yale.edu), Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520. The work is supported by the grant R01 DA016750-09 from the National Institute on Drug Abuse, and was completed when DL and JZ were postdoctoral associates at Yale University. The real data used in this manuscript were obtained from dbGaP at `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1` (accession number phs000092.v1.p1). The data collection was funded by NIH grants U01 HG004422, U01 HG004446, U10 AA008401, P01 CA089392, R01 DA013423, U01 HG004438, and HHSN268200782096C. The authors wish to thank Dr. Chintan Mehta and Ms. Donna DelBasso for their careful proofreading. We are especially indebted to the Editor, anonymous Associate Editor and referees for their constructive comments and guidance which led to significant improvement of our work.

Modeling Hybrid Traits for Comorbidity and Genetic Studies of Alcohol and Nicotine Co-Dependence

Abstract

We propose a novel multivariate model for analyzing hybrid traits and identifying genetic factors for comorbid conditions. Comorbidity is a common phenomenon in mental health in which an individual suffers from multiple disorders simultaneously. For example, in the Study of Addiction: Genetics and Environment (SAGE), alcohol and nicotine addiction were recorded through multiple assessments that we refer to as hybrid traits. Statistical inference for studying the genetic basis of hybrid traits has not been well-developed. Recent rank-based methods have been utilized for conducting association analyses of hybrid traits but do not inform the strength or direction of effects. To overcome this limitation, a parametric modeling framework is imperative. Although such parametric frameworks have been proposed in theory, they are neither well-developed nor extensively used in practice due to their reliance on complicated likelihood functions that have high computational complexity. Many existing parametric frameworks tend to instead use pseudo-likelihoods to reduce computational burdens. Here, we develop a model fitting algorithm for the full likelihood. Our extensive simulation studies demonstrate that inference based on the full likelihood can control the type-I error rate, and gains power and improves the effect size estimation when compared with several existing methods for hybrid models. These advantages remain even if the distribution of the latent variables is misspecified. After analyzing the SAGE data, we identify three genetic variants (rs7672861, rs958331, rs879330) that are significantly associated with the comorbidity of alcohol and nicotine addiction at the chromosome-wide level. Moreover, our approach has greater power in this analysis than several existing methods for hybrid traits.Although the analysis of the SAGE data motivated us to develop the model, it can be broadly applied to analyze any hybrid responses.

Key words: comorbidity, association, EM algorithm, latent variable, ordinal outcome.

1 Introduction

Identifying genetic variants that contribute to diseases is critically important for understanding their biological etiologies and, in turn, determining optimal treatment programs. Genetic studies in recent years have identified numerous genetic variants associated with diseases. A comprehensive catalog of findings from Genome-wide Association Studies (GWAS) can be found at `http://www.genome.gov/gwastudies`, where millions of genetic variants, namely single nucleotide polymorphisms (SNPs), are studied. This catalog makes it clear that genetic studies focus primarily on a single disease or trait.

Genetic studies of single phenotypes are likely to be inadequate for mental illnesses and behavioral disorders, which are characterized by variations in several traits. Genetic studies of such disorders warrant assessments across multiple phenotypes. For example, the Study of Addiction: Genetics and Environment (SAGE) recorded varying degrees of alcohol addition through multiple assessments, such as the maximum alcohol consumption in 24 hours (a continuous trait), whether or not feeling bad when controlling alcohol use (a binary trait), and the severity of alcohol symptoms (an ordinal trait). Alcohol addition was thus characterized by multiple variables of different types, which are referred to as hybrid traits. Furthermore, in mental health research and behavioral science, comorbidity of multiple disorders is common. An individual who is addicted to alcohol is more likely to suffer from nicotine addiction and mood disorders (Li and Burmeister, 2009). To examine the benefit of analyzing multivariate traits from a statistics perspective, Zhu and Zhang (2009) conducted extensive simulation studies and found that jointly testing correlated traits improves power over testing single traits one at a time. Similar findings have been reported recently; e.g., Lange et al. (2003); Yang et al. (2010); Zhang et al. (2010); Zhu et al. (2012); He et al. (2012, 2013); Galesloot et al. (2014); Jiang et al. (2014).

In most genetic studies of multivariate traits, traits are assumed to be exclusively quantitative, binary, or ordinal traits but not a mix of them, as reviewed by Zhang (2011); Galesloot et al. (2014). This assumption is however overly restrictive in mental health studies. For example, in SAGE, addiction to alcohol and nicotine is assessed by

the number of drinks, the number of cigarettes smoked, and the severity of alcohol or nicotine symptoms (no, mild, moderate, and severe). As a consequence, the methods designed for quantitative scales such as MV-PLINK (Ferreira and Purcell, 2009) and the commonly used principal component approach (Klei et al., 2008) are not appropriate.

Several existing methods developed for analyzing ordinal traits can be utilized for testing associations between hybrid traits and a genetic marker (Zhang et al., 2010; Zhu et al., 2012; Jiang et al., 2014; O'Reilly et al., 2012; He et al., 2013)). In particular, MultiPhen (O'Reilly et al., 2012) is designed to address arbitrary types of traits. However, these methods do not estimate the size and direction of the effects nor the dependence among the phenotypes. We aim to resolve this major deficiency in the genetic analysis of hybrid traits, particularly mental disorders by proposing a parametric multivariate hybrid (MH) model for jointly modeling hybrid traits with ordinal components. Tests of appropriately fitted parametric models may have considerably greater power than those relying on non-parametric approaches. More importantly, the MH model provides useful information that improves our understanding of comorbidity.

The key assumption behind the MH model is that the observed ordinal traits originate from some latent continuous variables, and this assumption makes it easier for us to form a joint distribution of the hybrid traits. Using latent variables for this purpose has previously been explored by many authors. Anderson and Pemberton (1985) and Poon and Lee (1987) termed it the conditional grouped continuous model, which de Leon and Carrièrgre (2007) and de Leon and Carrièrgre (2013) extended to a more general mixed data model. Boscardin et al. (2008) proposed a generalized multivariate probit model to study a repeated measures setting in a Bayesian framework. However, to the best of our knowledge, existing methods avoid directly computing the maximum likelihood estimate in this MH model. Several composite likelihood methods, which are less computationally demanding, have also been studied. For example, de Leon (2005) studied the pairwise likelihood approach, which is, however, less efficient than the maximum likelihood estimate. Due to the scarcity of computationally tractable algorithms for computing the maximum likelihood estimate, the general framework of MH modeling has not been utilized in practice, especially for applications that require performing a large number of tests for associating hybrid traits with risk factors, such as

in a GWAS of comorbid mental health traits. Genetic studies of comorbidity critically need statistically efficient and computationally practical methods, which we attempt to introduce in this work.

To obtain the maximum likelihood estimate from the full likelihood, we propose a Parameter-Expanded Expectation Conditional Maximization (PX-ECM) procedure. It extends the conditional version of the EM algorithm by transforming the latent variables and expanding the parameter space (Ruud, 1991; Meng and Rubin, 1993; Liu et al., 1998; Kawakatsu and Largey, 2009). Like the EM algorithm, the PX-ECM possesses an advantageous property that the likelihood is monotonically increasing in subsequent iterations, which guarantees the solution to be a local maximizer and, moreover, yields fully efficient parameter estimates by directly maximizing the full likelihood. Our numerical studies confirm that our estimation procedure improves both power in hypothesis testing and precision in estimating effect sizes, as well as provides directions.

In Section 2, we state our aim in analyzing the SAGE data, present the MH model for analyzing hybrid traits of comorbidity, and present results from the data analysis. For comparison, we also use the traditional univariate analysis, a reverse regression method, a multivariate nonparametric test based on the generalized Kendall's Tau, and the Fisher's combination. Our analysis reveals novel genetic markers that have not previously been reported as being associated with alcohol or nicotine addiction, which provides evidence that our parametric inference method can be valuable for genetic association studies of comorbid traits more broadly. To assess the performance of our method, in Section 3, we conduct simulation studies to compare it with four competing methods. The simulation results indicate that our parametric inference improves power in most scenarios. In particular, it compares favorably with the MultiPhen method (O'Reilly et al., 2012). Our method also improves, to a great extent, effect size estimation, compared to univariate analysis. These advantages remain even when the distribution of latent variables is misspecified. We conclude this work with some remarks in Section 4. We defer technical details to the Appendix, including the model fitting algorithm and its properties.

2 A Study of Comorbidity

2.1 The Study of Addiction: Genetics and Environment

The Study of Addiction: Genetics and Environment (SAGE) was a major undertaking to identify novel genetic risk factors for alcohol, smoking, and drug addiction through a large-scale genome-wide association study. The SAGE data include 4121 European and African Americans from three data sets: the Collaborative Study on the Genetics of Alcoholism (COGA), the Family Study of Cocaine Dependence (FSCD), and the Collaborative Genetic Study of Nicotine Dependence (COGEND). Each subject was diagnosed based on DSM-IV symptoms for alcohol, nicotine and other illicit drugs. Most studies of these data have been limited to single-trait based analyses with a few exceptions (Chen et al., 2011; Zhao and Zhang, 2016).

In our analysis, we study alcohol and nicotine dependence simultaneously as previous research showed that patients' dependence on these two substances are closely related (Li and Burmeister, 2009). We used the following four measures of addiction as the MH traits: 1) the continuous trait "max-drinks" that measures the largest number of alcoholic drinks consumed in 24 hours; 2) the ordinal trait "alc-sx" that measures the severity of alcohol symptoms (no, mild, moderate, and severe); 3) the binary trait "cig-daily" that reflects daily smoking or not; and 4) the ordinal trait "cig-sx" that measures the severity of nicotine symptoms (no, mild, moderate, and severe).

2.2 Multivariate Hybrid Model

We consider M continuous traits $Y = (Y^{(1)}, \ldots, Y^{(M)})^T$ and L ordinal traits $W = (W^{(1)}, \ldots, W^{(L)})^T$. For the SAGE data, $M = 1$ and $L = 3$, if we treat a binary trait as a special case of an ordinal trait. We use $Y^{(1)}, W^{(1)}, W^{(2)}$, and $W^{(3)}$ to represent the four traits introduced above, respectively.

For each ordinal trait $W^{(l)}$, we assume that there exists a latent continuous variable

$Z^{(l)}$ such that

$$W^{(l)} = \begin{cases} 1 & \text{if } Z^{(l)} \le c_0^{(l)}, \\ 2 & \text{if } c_0^{(l)} < Z^{(l)} \le c_1^{(l)}, \\ \vdots & \vdots \\ k_l - 1 & \text{if } c_{k_l-3}^{(l)} < Z^{(l)} \le c_{k_l-2}^{(l)}, \\ k_l & \text{if } c_{k_l-2}^{(l)} < Z^{(l)}, \end{cases} \tag{1}$$

where $k_l \ge 2$, $-\infty < c_0^{(l)} < c_1^{(l)} < \ldots < c_{k_l-2}^{(l)} < \infty$. The values of $W^{(l)}$, $1, 2, \ldots, k_l$, are merely symbols, which represent the order and should not be interpreted as numerical quantities. The role of the latent variables $Z = (Z^{(1)}, \ldots, Z^{(L)})^T$ is to facilitate modeling the joint distribution of Y and W.

In addition to the traits, the observed variables include the genotype G and p-dimensional covariate X. Conditional on G and X, each continuous trait follows

$$Y^{(m)} = \alpha_m + \beta_m G + \gamma_m^T X + \epsilon_m$$

where ϵ_m follows the normal distribution with mean zero and variance σ_m^2. Therefore,

$$\Pr(Y^{(m)} \le y^{(m)}) = \Phi(\sigma_m^{-1}(y^{(m)} - \alpha_m - \beta_m g - \gamma_m^T x)).$$

We use the similar technique to define the marginal distribution of $Z^{(l)}$:

$$Z^{(l)} = \mu_l + \theta_l G + \eta_l^T X + \varepsilon_l.$$

We have

$$\Pr(W^{(l)} \le i) = \Pr(Z^{(l)} \le c_{i-1}^{(l)}) = \Phi\left(\frac{c_{i-1}^{(l)} - (\mu_l + \theta_l G + \eta_l^T X)}{\sigma_{M+l}}\right). \tag{2}$$

Furthermore, for jointly modeling hybrid traits, we assume that the $(M+L)$-dimensional random vector $(\sigma_1^{-1}\epsilon_1, \ldots, \sigma_M^{-1}\epsilon_M, \sigma_{M+1}^{-1}\varepsilon_{M+1}, \ldots, \sigma_{M+L}^{-1}\varepsilon_{M+L})$ follows the multivariate normal distribution with mean zero and correlation matrix Λ.

In genetic association studies, the parameters β_m's and θ_l's represent the genetic effects, γ_m's and η_l's represent the environmental effects, the off-diagonal elements in the correlation matrix Λ reflect the trait-trait correlation (conditional on the known risk factors). In the Appendix, we will discuss the technical aspects of fitting this model and making statistical inference.

Here we also briefly introduce testing procedures to study the association between multivariate traits (Y, W) and the genotype G; that is,

$$H_0 : \beta = 0, \theta = 0 \text{ vs } H_1 : \beta \neq 0, \text{ or } \theta \neq 0.$$

Computational burdens of variance estimation raise challenges for efficiently using the classical Wald's test. Our proposed PX-ECM algorithm is feasible for computing the likelihood ratio test. Under H_0, the PX-ECM algorithm can also be implemented by removing biomarker G. Suppose the maximum likelihood estimate is $\hat{\Theta}$ in the whole parameter space and $\tilde{\Theta}$ when restricting to H_0. For simplicity, we denote the p.d.f. of $[Y, W|G, X]$ for each individual by $f(\Theta; Y_i, W_i)$, that is, $\text{Lik}(\Theta) = \Pi_{i=1}^{n} f(\Theta; Y_i, W_i)$. Following the classic likelihood theory, we can show that, under H_0, as sample size $n \to \infty$, we have

$$-2 \log \frac{\text{Lik}(\tilde{\Theta})}{\text{Lik}(\hat{\Theta})} \xrightarrow{d} \chi^2_{M+L}.$$

Therefore, the null hypothesis H_0 is rejected if $-2 \log \frac{\text{Lik}(\tilde{\Theta})}{\text{Lik}(\hat{\Theta})} > \chi^2_{M+L,\alpha}$, where $\chi^2_{M+L,\alpha}$ is the $(1-\alpha)$-th quantile of the chi-square distribution χ^2_{M+L}, and α is a pre-specified nominal level of significance.

2.3 Data Analysis Results

We performed a genome-wide association study to identify SNPs associated with comorbid addiction in the SAGE sample. After following data quality control steps in Jiang et al. (2014), we used data from 3,564 subjects with 950,705 SNPs in our analysis.

To use the MH model, we code the genotype G of each SNP as the observed number of minor alleles. Covariates X include gender, race (European or African), study (COGA, FSCD, or COGEND), and the first two principal components of the genetic relatedness matrix to adjust for population stratification (Figure 1 reveals two major principal components for population stratification based on the genotype data).

[Figure 1 approximately here]

For comparison, we consider four competitors including univariate analysis with Bonferroni correction (Univariate-BC), the MultiPhen method, the generalized Kendall's Tau method (G-Kendall), and Fisher's combination of p-values method with a bootstrap correction (Fisher-boot). The Univariate-BC method fits a (linear or probit) regression model to each trait, and the p-value from Wald test is then adjusted by Bonferroni correction (Laird and Lange, 2011). The MultiPhen method (O'Reilly et al., 2012) is based on reverse regression, i.e., treating the genotype as an ordinal response and regressing it on phenotypes including the disease traits and covariates. For implementation, we use the R package "MultiPhen". The G-Kendall method is a nonparametric method based on Kendall's Tau while adjusting for covariates (Zhu et al., 2012). The Fisher-boot method uses Fisher's method to combine p-values from univariate trait analysis. The correlation between the p-values is corrected by bootstrap (Kwak et al., 2013).

Because we are particularly interested in p-values that are extremely small, we first screened the entire genomewide with simple tests. Specifically, we impose a screening criterion of requiring the p-value from the univariate analysis with Bonferroni correction less than 0.05, as well as the p-value from the Kendall's Tau less than 1×10^{-4}, which led to 86 candidate SNPs for intensive computation of their p-values. An R code is available upon request.

We find that the SNP rs958331, located on the gene CARD11 in Chromosome 7, has a p-value of 4.51×10^{-7}, reaching the chromosome wide significance level ($\alpha = 0.05/50138 = 9.97 \times 10^{-7}$). CARD11 is an oncogene in human diffuse large B cell lymphoma (Lenz et al., 2008), and has not been previously associated with addiction. For the same SNP, as reported in Table 1, the p-values are 3.01×10^{-6}, 3.33×10^{-2}, 1.61×10^{-2} and 9.51×10^{-5}, using the MultiPhen (O'Reilly et al., 2012), the Fisher's approach, the univariate approach with the Bonferroni correction (an extra multiplier of 4 must be factored in the Bonferroni correction due to the four traits) and the nonparametric multivariate approach based on generalized Kendall's Tau, respectively. Tests for this SNP using the competing approaches were not significant at the chromosome wide level.

Meanwhile, we find that SNP rs7672861, located in the intergenic region on Chromosome 4, has a p-value of 5.22×10^{-7}, reaching the chromosome wide significance level

($\alpha = 0.05/55634 = 8.99 \times 10^{-7}$). Although SNP rs7672861 was not previously identified as a risk factor of comorbid addiction, it is in linkage disequilibrium with SNPs in gene RNF150, which is well known as being associated with chronic obstructive pulmonary disease (Kim et al., 2012). A haploview for the LD blocks in the proximity of SNP rs7672861 is provided in Figure 2, which demonstrates the relative proximity of the SNP 7672861 and the gene RNF150. We should note that all competing methods except the univariate approach also detect SNP rs7672861, as presented in Table 1. In addition, SNP rs879330 located in gene COL18A1 on chromosome 21 passes the chromosome wide significance from the MultiPhen method, and our proposed method provides a p-value of similar magnitude to that from the MultiPhen.

[Figure 2 approximately here]

[Table 1 approximately here]

In the following, we further investigate the effects of three identified SNPs. For our joint model, Table 2 presents the estimates for parameters that are of interest in genetic studies. The rows corresponding to the three identified SNPs reveal that, if we examine the association with each single trait, only $Y^{(1)}$ (the continuous trait "max-drinks") yields a p-value (1.3×10^{-7}) below the chromosome-wide significance level. This observation underscores again the benefit of considering comorbidity of alcohol and nicotine dependence and modeling comorbid traits simultaneously. The gain of power is due to the strong correlation (conditional on the known risk factors) between the traits, as seen in Table 2. More specifically, the conditional correlations between alcohol-related and nicotine-related traits are 0.51 or greater. This magnitude implies that, in addition to those risk factors identified in the model, there remain other unknown but major genetic or environmental factors that account for the comorbidity of alcohol and nicotine dependence. Such information is also critical for genetic association studies, as it provides additional knowledge of comorbidity and calls for further search of other risk factors. We note that neither the nonparametric method nor combining univariate analyses method can unveil the conditional correlation among the traits as a result of unknown factors.

[Table 2 approximately here]

According to Table 2, rs958331 appears to be associated positively with $Y^{(1)}$, $W^{(2)}$, and $W^{(3)}$, but negatively with $W^{(1)}$, despite the fact that the four traits are positively associated with each other. Our analysis also reveals that a large proportion of the comorbidity of alcohol and nicotine dependence has not been explained by the known risk factors in the SAGE data, and hence further studies are warranted.

3 Simulation study

We use simulation studies to examine the performance of the jointly modeling approach with respect to (1) the power of detecting signals; (2) the bias and efficiency in parameter estimation; and (3) the robustness to model misspecifications. We consider again the four competing methods: the Univariate-BC, MultiPhen, G-Kendall, and Fisher-boot methods.

We first simulate six traits: two continuous, two binary and two ordinal traits. The genotype variable G takes values 0, 1, or 2 with the minor allele frequency (MAF) being 0.3 or 0.1. Two covariates X_1 and X_2 are generated from the normal distributions $N(0, 1)$ and $N(0, 4)$, independent of G. We set the regression coefficients such that the genetic effects are much smaller than the covariate effects (e.g., 0.1 versus 1.0), which is often the case in genetic studies. The correlation of the latent error components is set to be 0.7. Discretizing the latent continuous variables yields the binary and ordinal traits. The two ordinal traits have 4 and 3 levels, respectively. This simulation setting involves 44 parameters for us to estimate in the MH model. We consider three choices of sample size, $n = 600, 1200$ and 2000.

Tables 3-4 reveal that all methods control the type I error reasonably well when all the genetic effects $\beta_m = \theta_l = 0$, with the Univariate-BC method being relatively conservative. Tables 5-6 present the power analysis results when the genetic effects $\beta_m = \theta_l = 0.1$. We observe that the likelihood ratio test (LRT) from our MH model gains considerable power relative to the Univariate-BC and MultiPhen methods, the two most popular methods

for genetic association studies. Our method is expected to be more powerful than the Univariate-BC because we explicitly model the between-trait correlations. A striking observation is that the MultiPhen method, which also accommodates the multivariate traits, is sometimes less powerful than the Univariate-BC method. It is slightly better than the nonparametric G-Kendall method. One explanation is that the MultiPhen method ignores the relationship among the traits. We also observe that the power of our MH-LRT method is comparable to the Fisher-boot method, which adjusts the correlation using bootstrap. The latter method, however, is even more computationally intensive. For example, in the current simulation setting, it needs at least 3000 bootstrap replicates to ensure a reasonably controlled type I error rate. Moreover, unlike our method, it does not improve the precision of the effect size estimation. We also examine the type I error and power while mimicking our real data analysis in Section 2. Table 7 shows the results when we simulate data using the effect sizes from the regression model for rs958331. In this scenario, we have two new observations: the MultiPhen method has severely inflated type I error rates; and the Fisher-boot method is much less powerful. Our method controls the type I error and has the highest power.

Unlike the four competitors, the MH model explicitly models the between-trait association. Therefore, our inference on the regression slope parameters, which represent genetic and environmental effects, is expected to be more efficient than that based on the univariate-trait analysis. The PX-ECM algorithm helps us achieve the efficiency by maximizing the full likelihood, rather than marginal or pseudo likelihoods. Table 8 exemplifies our gain of efficiency over the univariate-trait (marginal) analysis ($n = 2000$, MAF=0.3). For the regression slope and correlation parameters, it tabulates their true values, the mean and standard deviation of their estimates from multivariate and univariate analyses. To assess the relative efficiency, we calculate the ratio of the mean squared error (MSE) of the two types of estimates. Table 8 suggests: for the slope parameters associated with continuous traits, the univariate trait analysis is fully efficient, and for the binary and ordinal traits, our inference can be 10 to 40 times more efficient than the univariate method. It is important to note that unlike our proposed method, the univariate trait analysis does not estimate the strength of the association, marked as "NA"s in the table.

Similar to the Univariate-BC and Fisher-boot methods, the MH model requires some parametric assumptions. In particular, we require that the latent variables underlying the categorical traits follow normal distributions. In what follows, we numerically examine the robustness of our inference to misspecification of latent distributions. Specifically, we simulate the latent errors from χ^2 distributions with 4 degrees of freedom, which is a highly skewed distribution with a skewness value of $\sqrt{2}$. We examine the type I error, power, and parameter estimation using the same simulation setting as above. The results are reported in Tables 9-11. Table 9 reassures the robustness of all methods in terms of the type I error rates. Table 10 indicates that our method is much more powerful than the MultiPhen and Univariate-BC methods. As for the effect size estimation, recall that the distribution assumed for the latent variable in the model was different from the one from which the data were simulated. As a result from this misspecification, we can observe from Table 11 that both methods are biased. However, our joint modeling approach is less biased than the method based on the marginal likelihood (see the 3rd and 5th columns of Table 11). Moreover, the last column confirms that our inference is still much more efficient than marginal inference.

To summarize, our method has the following advantages. First, it gains substantial power relative to several existing methods such as the MultiPhen Method and univariate-trait analysis. Second, it improves, to a great extent, effect size (regression slope) estimation over the univariate-trait analysis. Third, our method is the only one to estimate the between-trait correlation while adjusting for covariates. The partial correlation estimates can provide deeper insight into comorbidity. These advantages remain even when the distribution of the latent variables of our MH model is misspecified.

4 Discussion

Studying comorbidity is important for mental health research where psychiatric conditions are characterized by multiple assessments. Because these assessments typically consist of a hybrid of continuous and ordinal variables, comorbidity raises major challenges for statistical inference. This problem is evident from the analysis of the SAGE

data. To meet the challenges of analyzing such data, we proposed to use the multivariate hybrid model and develop sound methods for computation and inference. Although parametric models have previously been proposed to analyze hybrid responses, they have been underdeveloped and have limited applications. Our work presents the first major attempt to tackle this challenging problem. After applying our model to the SAGE, we unraveled a novel gene that is commonly known as an oncogene in human diffuse large B cell lymphoma, but was not previously associated with addiction. Our finding underscores the potential of our model to improve power for genetic association studies of comorbidity. Not only did we provide statistically significant evidence for our finding, but also we assessed the validity of our model and offered insights into how we were able to discover this SNP that was otherwise not detectable by the existing methods.

While our own primary motivation was to evaluate the genetic contribution to complex traits, such data are common in many areas, especially in behavioral sciences. As we noted above, there are a limited number of methods to analyze mixed types of outcomes. Nonparametric tests are among one of the main approaches, which yield p-values, but offer no explicit information on the association direction and strength. Thus, a parametric framework is critical to fill in this need. Moreover, parametric models are far more convenient for the purpose of incorporating covariates in the analysis which allows, in turn, an examination of confounding and interactive effects. These are critical aspects of regression analysis involving a range of known and unknown attributing factors.

Given the complexity involved in modeling the mixed outcomes, it is not surprising that we must resolve some major technical challenges. In fact, model fitting for such outcomes is particularly difficult, and has rarely been discussed in the literature. We developed a stable and feasible PX-ECM algorithm to compute the maximum likelihood estimates. Our PX-ECM algorithm is designed to accommodate the mixture of continuous and ordinal traits simultaneously. It still retains the key feature of EM-type algorithms such as monotonicity of the observed likelihood.

In the analysis of multiple traits, the proposed parametric model has clear advantages over the rank-based nonparametric methods proposed in Zhang et al. (2010) and Zhu et al. (2012). First, since our parametric approach explicitly models the correlation

structure of continuous and categorical traits, it can achieve remarkable gain of power in detecting plausible association of genetic variants with the complex disease of interest, as we demonstrated in the simulation and real data studies. Importantly, the parametric framework can yield other important inference outcomes such as the direction and strength of the association (captured by the regression coefficients), which play a crucial role in understanding the genetic and environmental effect in the development of a complex disease.

The normality (or other distributional) assumption for continuous data is commonly used for convenience in parametric models. When the response variable is binary or ordinal and if an ordered probit/logistic model is used, we can assume a certain density function of the latent variable for the categorical response. Our simulation studies demonstrate that our model is robust to the misspecification of the distribution of the latent variables. Nonetheless, if such an assumption is of clear concern, we may consider rank-based methods such as the generalized Kendall's Tau method. But the rank-based methods are generally not as powerful as parametric methods as we also demonstrated in our simulation studies. Thus, parametric and nonparametric methods complement each other and both are useful.

References

Anderson, J. and Pemberton, J. (1985), "The grouped continuous model for multivariate ordered categorical variables and covariate adjustment." *Biometrics*, 41, 875–885.

Boscardin, W. J., Zhang, X., and Belin, T. R. (2008), "Modeling a mixture of ordinal and continuous repeated measures," *Journal of Statistical Computation and Simulation*, 78, 873–886.

Chen, X., Cho, K., Singer, B., and Zhang, H. (2011), "The nuclear transcription factor PKNOX2 is a candidate gene for substance dependence in European-origin women," *PLoS One*, 6, e16002.

de Leon, A. (2005), "Pairwise likelihood approach to grouped continuous model and its extension," *Statistics & Probability Letters*, 75, 49–57.

de Leon, A. and Carrièrgre, K. (2007), "General mixed-data model: Extension of general location and grouped continuous models," *Canadian Journal of Statistics*, 35, 533–548.

— (2013), *Analysis of Mixed Data: Methods & Applications*, Chapman and Hall/CRC.

Ferreira, M. A. and Purcell, S. M. (2009), "A multivariate test of association," *Bioinformatics*, 25, 132–133.

Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L., and Vermeulen, S. H. (2014), "A comparison of multivariate genome-wide association methods," *PloS one*, 9, e95923.

He, J., Li, H., Edmondson, A. C., Rader, D. J., and Li, M. (2012), "A Gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies," *Biostatistics*, 13, 497–508.

He, Q., Avery, C. L., and Lin, D.-Y. (2013), "A General Framework for Association Tests With Multivariate Traits in Large-Scale Genomics Studies," *Genetic epidemiology*, 37, 759–767.

Jiang, Y., Li, N., and Zhang, H. (2014), "Identifying Genetic Variants for Addiction via Propensity Score Adjusted Generalized Kendall's Tau," *Journal of the American Statistical Association*, 109, 905–930.

Kawakatsu, H. and Largey, A. G. (2009), "EM algorithms for ordered probit models with endogenous regressors," *The Econometrics Journal*, 12, 164–186.

Kim, D. K., Cho, M. H., Hersh, C. P., Lomas, D. A., Miller, B. E., Kong, X., Bakke, P., Gulsvik, A., Agustí, A., Wouters, E., et al. (2012), "Genome-wide association analysis of blood biomarkers in chronic obstructive pulmonary disease," *American Journal of Respiratory and Critical Care Medicine*, 186, 1238–1247.

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008), "Pleiotropy and principal components of heritability combine to increase power for association analysis," *Genetic Epidemiology*, 32, 9–19.

Kwak, M., Zheng, G., and Wu, C. O. (2013), "Joint tests for mixed traits in genetic association studies," in *Analysis of Mixed Data: Methods & Applications*, Chapman and Hall/CRC, pp. 31–41.

Laird, N. M. and Lange, C. (2011), *The Fundamentals of Modern Statistical Genetics*, Springer.

Lange, C., Silverman, E. K., Xu, X., Weiss, S. T., and Laird, N. M. (2003), "A multivariate family-based association test using generalized estimating equations: FBAT-GEE," *Biostatistics*, 4, 195–206.

Lenz, G., Davis, R., Ngo, V., Lam, L., George, T., Wright, G., Dave, S., Zhao, H., Xu, W., Rosenwald, A., Ott, G., Muller-Hermelink, H., Gascoyne, R., Connors, J., Rimsza, L., Campo, E., Jaffe, E., Delabie, J., Smeland, E., Fisher, R., Chan, W., and LM, S. (2008), "Oncogenic CARD11 mutations in human diffuse large B cell lymphoma," *Science*, 319, 1676–1679.

Li, M. D. and Burmeister, M. (2009), "New insights into the genetics of addiction," *Nature Reviews Genetics*, 10, 225–231.

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter expansion to accelerate EM: the PX-EM algorithm," *Biometrika*, 85, 755–770.

Meng, X.-L. and Rubin, D. B. (1993), "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, 80, 267–278.

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M.-R., and Coin, L. J. (2012), "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS," *PloS One*, 7, e34861.

Poon, W.-Y. and Lee, S.-Y. (1987), "Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients," *Psychometrika*, 52, 409–430.

Ruud, P. A. (1991), "Extensions of estimation methods using the EM algorithm," *Journal of Econometrics*, 49, 305–341.

Yang, Q., Wu, H., Guo, C.-Y., and Fox, C. S. (2010), "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests," *Genetic epidemiology*, 34, 444–454.

Zhang, H. (2011), "Statistical Analysis in Genetic Studies of Mental Illnesses," *Statistical Science*, 26, 116–129.

Zhang, H., Liu, C.-T., and Wang, X. (2010), "An association test for multiple traits based on the generalized Kendalls tau," *Journal of the American Statistical Association*, 105, 473–481.

Zhao, J. and Zhang, H. (2016), "Modeling multiple responses via bootstrapping margins with an application to genetic association testing," *Statistics and Its Interface*, 9, 47–56.

Zhu, W., Jiang, Y., and Zhang, H. (2012), "Nonparametric Covariate-Adjusted Association Tests Based on the Generalized Kendall's Tau," *Journal of the American Statistical Association*, 107, 1–11.

Zhu, W. and Zhang, H. (2009), "Why do we test multiple traits in genetic association studies? (with discussion)," *Journal of the Korean Statistical Society*, 38, 1–10.

Table 1: A summary of significant results by MultiPhen, Fisher's approach, univariate analysis with Bonferroni correction, Kendall's Tau and the proposed method PX-ECM. Chromosome-wide significance level is provided as threshold. Fisher's approach is based on a permutation test of 3000 replications. More replications are expected but are not conducted due to computational restriction. Chromosome-wide significant results are highlighted in bold.

SNP	Gene	Chromosome
rs7672861	intergenic	4
rs958331	CARD11	7
rs879330	COL18A1	21

SNP	Threshold	MultiPhen	Fisher	Univariate-BC	Kendall's Tau	PX-ECM
rs7672861	8.99×10^{-7}	$\mathbf{7.57 \times 10^{-7}}$	**0**	1.36×10^{-5}	$\mathbf{9.04 \times 10^{-8}}$	$\mathbf{5.22 \times 10^{-7}}$
rs958331	9.97×10^{-7}	3.01×10^{-6}	3.33×10^{-2}	1.61×10^{-2}	9.51×10^{-5}	$\mathbf{4.51 \times 10^{-7}}$
rs879330	3.99×10^{-6}	$\mathbf{2.91 \times 10^{-6}}$	1.67×10^{-3}	1.26×10^{-5}	2.10×10^{-3}	1.01×10^{-5}

Table 2: Parameter estimation of the detected SNPs rs7672861, rs958331 and rs879330 based on the MH model for the SAGE data analysis

	Alcohol-related				Nicotine-related			
	Max-drinks		Alc-sx		Cig-daily		Cig-sx	
rs7672861	0.095	(0.018)	0.059	(0.025)	0.156	(0.037)	0.082	(0.028)
rs958331	0.021	(0.025)	-0.084	(0.035)	0.030	(0.041)	0.059	(0.031)
rs879330	0.029	(0.032)	0.068	(0.043)	0.073	(0.053)	0.195	(0.047)
Gender	-0.693	(0.028)	-0.556	(0.040)	-0.166	(0.051)	-0.209	(0.043)
Race	0.045	(0.038)	-0.228	(0.055)	-0.344	(0.062)	0.075	(0.052)
Max-drinks	1		0.766	(0.009)	0.513	(0.020)	0.539	(0.016)
Alc-sx			1		0.585	(0.023)	0.663	(0.014)
Cig-daily					1		0.939	(0.012)
Cig-sx							1	

Alc-sx and Cig-sx represent the severity of alcohol and nicotine symptoms, respectively. Within the parentheses is the estimated standard error of the corresponding parameter estimate. The upper section is the coefficient estimates and the lower section is the dependence among the four traits.

Table 3: Empirical type I error rate for association test when MAF=0.3

Sample size	Method	α=0.1	α=0.05	α=0.01
$n = 600$	Univariate-BC	0.068	0.032	0.006
	MultiPhen	0.100	0.051	0.009
	G-Kendall	0.090	0.045	0.008
	Fisher-boot	0.100	0.048	0.009
	MH-LRT	0.107	0.054	0.011
$n = 1200$	Univariate-BC	0.070	0.034	0.006
	MultiPhen	0.103	0.050	0.010
	G-Kendall's Tau	0.096	0.050	0.010
	Fisher-boot	0.102	0.051	0.010
	MH-LRT	0.111	0.058	0.011
$n = 2000$	Univariate-BC	0.062	0.031	0.006
	MultiPhen	0.104	0.054	0.010
	G-Kendall's Tau	0.098	0.048	0.009
	Fisher-boot	0.104	0.051	0.011
	MH-LRT	0.111	0.059	0.012

Results are obtained based on 5000 simulation replicates.

Table 4: Empirical type I error rate for association test when MAF=0.1

Sample size	Method	α=0.1	α=0.05	α=0.01
$n = 600$	Univariate-BC	0.068	0.030	0.007
	MultiPhen	0.106	0.054	0.012
	G-Kendall's Tau	0.098	0.047	0.009
	Fisher-boot	0.094	0.046	0.010
	MH-LRT	0.112	0.059	0.014
$n = 1200$	Univariate-BC	0.064	0.031	0.007
	MultiPhen	0.095	0.049	0.012
	G-Kendall's Tau	0.092	0.041	0.010
	Fisher-boot	0.100	0.046	0.010
	MH-LRT	0.109	0.057	0.010
$n = 2000$	Univariate-BC	0.061	0.030	0.006
	MultiPhen	0.109	0.058	0.011
	G-Kendall's Tau	0.101	0.053	0.009
	Fisher-boot	0.101	0.050	0.011
	MH-LRT	0.113	0.054	0.011

Results are obtained based on 5000 simulation replicates.

Table 5: Empirical power for association test when MAF=0.3

Sample size	Method	α=0.1	α=0.05	α=0.01
$n = 600$	Univariate-BC	0.285	0.198	0.065
	MultiPhen	0.226	0.137	0.030
	G-Kendall	0.200	0.121	0.024
	Fisher-boot	0.337	0.234	0.091
	MH-LRT	0.326	0.209	0.075
$n = 1200$	Univariate-BC	0.394	0.290	0.122
	MultiPhen	0.365	0.249	0.099
	G-Kendall's Tau	0.337	0.227	0.080
	Fisher-boot	0.517	0.410	0.204
	MH-LRT	0.513	0.382	0.188
$n = 2000$	Univariate-BC	0.574	0.431	0.223
	MultiPhen	0.557	0.441	0.205
	G-Kendall's Tau	0.537	0.400	0.169
	Fisher-boot	0.703	0.592	0.349
	MH-LRT	0.731	0.633	0.399

Results are obtained based on 1000 simulation replicates.

Table 6: Empirical power for association test when MAF=0.1

Sample size	Method	α=0.1	α=0.05	α=0.01
$n = 1200$	Univariate-BC	0.200	0.124	0.043
	MultiPhen	0.198	0.129	0.047
	G-Kendall's Tau	0.193	0.116	0.033
	Fisher-boot	0.330	0.223	0.070
	MH-LRT	0.290	0.183	0.060
$n = 2000$	Univariate-BC	0.295	0.198	0.070
	MultiPhen	0.277	0.186	0.064
	G-Kendall's Tau	0.267	0.164	0.052
	Fisher-boot	0.401	0.309	0.149
	MH-LRT	0.418	0.289	0.131
$n = 3000$	Univariate-BC	0.353	0.235	0.090
	MultiPhen	0.347	0.241	0.095
	G-Kendall's Tau	0.348	0.226	0.082
	Fisher-boot	0.544	0.397	0.178
	MH-LRT	0.550	0.408	0.202

Results are obtained based on 1000 simulation replicates.

Table 7: Type I error and power when mimicking the effect size in analysis for rs958331

	Method	α=0.1	α=0.05	α=0.01
Type I error	Univariate-BC	0.078	0.039	0.007
	MultiPhen	0.186	0.109	0.035
	G-Kendall's Tau	0.073	0.032	0.006
	Fisher-boot	0.102	0.047	0.009
	MH-LRT	0.091	0.044	0.008
Power	Univariate-BC	0.882	0.765	0.462
	MultiPhen	0.996	0.994	0.975
	G-Kendall's Tau	0.998	0.992	0.965
	Fisher-boot	0.841	0.607	0.153
	MH-LRT	0.999	0.996	0.974

Results are obtained based on 5000 simulation replicates.

Table 8: Parameter estimation using full and marginal likelihoods ($n = 2000$)

		Joint estimation		Marginal estimation		
	True value	Mean	SE	Mean	SE	ratio.MSE
Regression coefficients						
Trait 1	1.00	1.00	0.10	1.00	0.10	1.00
	0.10	0.10	0.07	0.10	0.07	1.00
	1.00	1.00	0.04	1.00	0.04	1.00
	1.00	1.00	0.02	1.00	0.02	1.00
Trait 2	1.00	1.01	0.15	1.01	0.15	1.00
	0.10	0.10	0.10	0.10	0.10	1.00
	1.00	1.00	0.07	1.00	0.07	1.00
	1.00	1.00	0.03	1.00	0.03	1.00
Trait 3	-1.40	-1.41	0.11	-1.30	0.14	2.53
	0.10	0.10	0.06	0.09	0.07	1.11
	1.00	1.00	0.05	0.92	0.08	4.21
	1.00	1.01	0.04	0.93	0.07	6.10
Trait 4	-2.20	-2.21	0.13	-1.90	0.17	6.54
	0.10	0.10	0.07	0.09	0.07	1.04
	1.00	1.01	0.06	0.86	0.07	6.75
	1.00	1.01	0.05	0.85	0.07	11.65
Trait 5	1.90	1.91	0.08	1.62	0.14	14.20
	0.10	0.10	0.04	0.08	0.04	0.95
	1.00	1.00	0.04	0.85	0.07	19.63
	1.00	1.00	0.03	0.85	0.06	37.11
Trait 6	1.10	1.11	0.09	1.01	0.11	2.50
	0.10	0.10	0.05	0.09	0.05	1.00
	1.00	1.00	0.04	0.92	0.07	7.04
	1.00	1.00	0.03	0.92	0.07	11.53
Correlation coefficients						
	0.70	0.70	0.01	NA	NA	
	0.70	0.70	0.03	NA	NA	
	0.70	0.70	0.03	NA	NA	
	0.70	0.70	0.02	NA	NA	
	0.70	0.70	0.02	NA	NA	
	0.70	0.70	0.03	NA	NA	
	0.70	0.70	0.03	NA	NA	
	0.70	0.70	0.02	NA	NA	
	0.70	0.70	0.02	NA	NA	
	0.70	0.70	0.04	NA	NA	
	0.70	0.70	0.03	NA	NA	
	0.70	0.70	0.03	NA	NA	
	0.70	0.70	0.04	NA	NA	
	0.70	0.70	0.04	NA	NA	
	0.70	0.70	0.03	NA	NA	
Cutoff points						
	2.10	2.10	0.07	1.79	0.14	23.75
	2.10	2.10	0.07	1.78	0.14	27.41
	2.60	2.61	0.08	2.39	0.18	11.60

Results are obtained based on 1000 simulation replicates.

Table 9: Empirical type I error rate for association test when the latent distributions are misspecified as χ^2 (MAF=0.3)

Sample size	Method	α=0.1	α=0.05	α=0.01
$n = 600$	Univariate-BC	0.083	0.042	0.007
	MultiPhen	0.109	0.047	0.008
	G-Kendall's Tau	0.088	0.044	0.009
	Fisher-boot	0.102	0.050	0.013
	MH-LRT	0.102	0.060	0.008
$n = 1200$	Univariate-BC	0.069	0.034	0.007
	MultiPhen	0.101	0.049	0.010
	G-Kendall's Tau	0.094	0.048	0.010
	Fisher-boot	0.100	0.051	0.010
	MH-LRT	0.108	0.056	0.011
$n = 2000$	Univariate-BC	0.061	0.033	0.009
	MultiPhen	0.109	0.061	0.009
	G-Kendall's Tau	0.092	0.053	0.006
	Fisher-boot	0.082	0.046	0.009
	MH-LRT	0.107	0.057	0.016

Results are obtained based on 1000 simulation replicates.

Table 10: Empirical power for association test when the latent distributions are misspecified as χ^2 (MAF=0.3)

Sample size	Method	α=0.1	α=0.05	α=0.01
$n = 600$	Univariate-BC	0.207	0.143	0.050
	MultiPhen	0.244	0.155	0.056
	G-Kendall's Tau	0.220	0.121	0.032
	Fisher-boot	0.307	0.219	0.088
	MH-LRT	0.310	0.197	0.072
$n = 1200$	Univariate-BC	0.391	0.254	0.094
	MultiPhen	0.383	0.259	0.099
	G-Kendall's Tau	0.349	0.247	0.090
	Fisher-boot	0.516	0.380	0.169
	MH-LRT	0.506	0.376	0.171
$n = 2000$	Univariate-BC	0.568	0.439	0.242
	MultiPhen	0.519	0.409	0.209
	G-Kendall's Tau	0.541	0.405	0.208
	Fisher-boot	0.709	0.600	0.360
	MH-LRT	0.686	0.580	0.347

Results are obtained based on 1000 simulation replicates.

Table 11: Parameter estimation when the latent distributions are misspecified as χ^2 ($n = 2000$)

		Joint estimation		Marginal estimation		
	True value	Mean	SE	Mean	SE	ratio.MSE
Regression coefficients						
Trait 1	1.00	0.99	0.11	0.99	0.11	1.00
	0.10	0.10	0.07	0.10	0.07	1.00
	1.00	1.00	0.04	1.00	0.04	1.00
	1.00	1.00	0.02	1.00	0.02	1.00
Trait 2	1.00	1.00	0.16	1.00	0.16	1.00
	0.10	0.10	0.11	0.10	0.11	1.00
	1.00	1.00	0.07	1.00	0.07	1.00
	1.00	1.00	0.03	1.00	0.03	1.00
Trait 3	-1.40	-1.37	0.11	-1.27	0.12	2.18
	0.10	0.10	0.06	0.09	0.07	1.18
	1.00	0.96	0.06	0.89	0.07	2.96
	1.00	0.96	0.05	0.89	0.05	3.55
Trait 4	-2.20	-1.98	0.13	-1.87	0.15	2.02
	0.10	0.09	0.06	0.08	0.07	1.22
	1.00	0.88	0.06	0.82	0.06	1.86
	1.00	0.88	0.05	0.82	0.05	1.91
Trait 5	1.90	1.95	0.08	1.63	0.13	10.76
	0.10	0.10	0.04	0.08	0.04	0.89
	1.00	1.02	0.04	0.86	0.07	12.37
	1.00	1.02	0.03	0.86	0.06	16.53
Trait 6	1.20	1.21	0.08	1.10	0.12	3.47
	0.10	0.10	0.05	0.09	0.05	0.96
	1.00	1.03	0.05	0.93	0.08	3.33
	1.00	1.03	0.04	0.93	0.07	4.15
Correlation coefficients						
	0.70	0.70	0.01	NA	NA	
	0.70	0.67	0.03	NA	NA	
	0.70	0.68	0.03	NA	NA	
	0.70	0.67	0.02	NA	NA	
	0.70	0.67	0.02	NA	NA	
	0.70	0.67	0.03	NA	NA	
	0.70	0.68	0.03	NA	NA	
	0.70	0.67	0.02	NA	NA	
	0.70	0.67	0.02	NA	NA	
	0.70	0.69	0.04	NA	NA	
	0.70	0.66	0.03	NA	NA	
	0.70	0.71	0.04	NA	NA	
	0.70	0.70	0.03	NA	NA	
	0.70	0.69	0.04	NA	NA	
	0.70	0.66	0.03	NA	NA	
Cutoff points						
	2.00	2.09	0.08	1.75	0.14	6.18
	2.10	2.09	0.08	1.77	0.14	22.23
	2.60	2.63	0.10	2.39	0.19	7.17

Results are obtained based on 1000 simulation replicates.

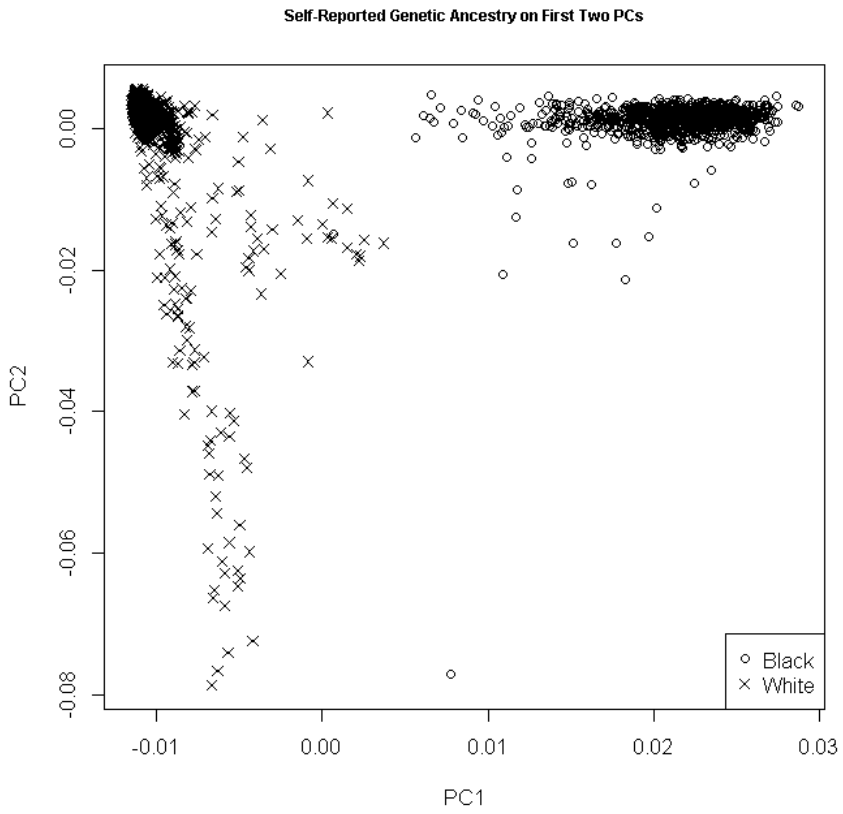

Figure 1: Principal component (PC) analysis of the quality-controlled SAGE genotype data to demonstrate population stratification.

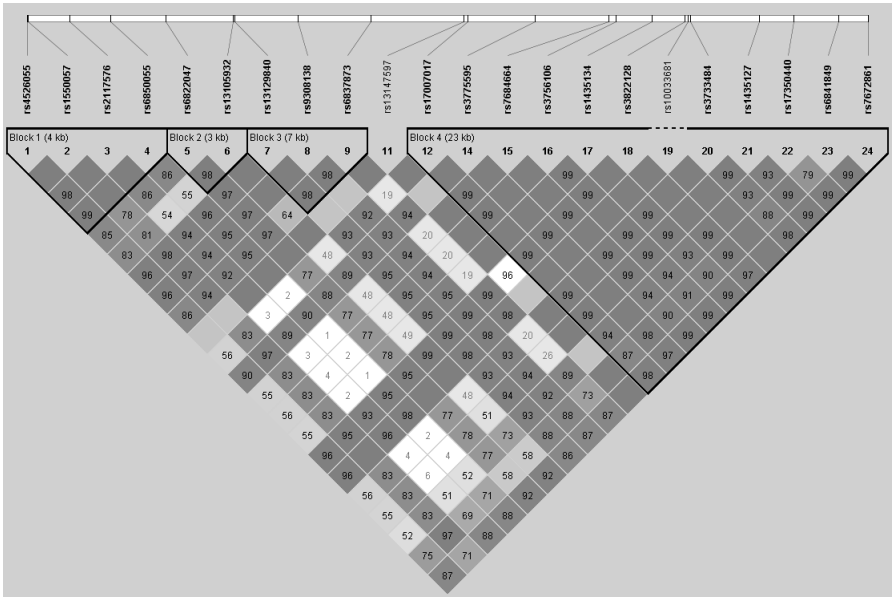

Figure 2: Haploview of the proximity of SNP 7672861, which is at the right end of the plot. The gene RNF150 covers from the left end to the right of SNP 9308138, which is about 33kb from the SNP 7672861.