

## Mixed modelling to characterize genotype–phenotype associations

A. S. Foulkes<sup>1,\*</sup>,<sup>†</sup>,<sup>‡</sup>, M. Reilly<sup>3</sup>, L. Zhou<sup>2</sup>, M. Wolfe<sup>3,4</sup> and D. J. Rader<sup>3,4</sup>

<sup>1</sup>*Department of Biostatistics, University of Massachusetts, School of Public Health, 404 Arnold House, 715N. Pleasant Street, Amherst, MA 01003-9304, U.S.A*

<sup>2</sup>*Department of Biostatistics, 423 Guardian Drive, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, U.S.A.*

<sup>3</sup>*Cardiovascular Division, University of Pennsylvania School of Medicine, 421 Curie Blvd, Philadelphia, PA 19104, U.S.A.*

<sup>4</sup>*Center for Experimental Therapeutics, 421 Curie Blvd, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, U.S.A.*

### SUMMARY

We propose using mixed effects models to characterize the association between multiple gene polymorphisms, environmental factors and measures of disease progression. Characterizing high-order gene–gene and gene–environment interactions presents an analytic challenge due to the large number of candidate genes and the complex, undescribed interactions among them. Several approaches have been proposed recently to reduce the number of candidate genes and *post hoc* approaches to identify gene–gene interactions are described. However, these approaches may be inadequate for identifying high-order interactions in the absence of main effects and generally do not permit us to control for potential confounders. We describe how mixed effects models and related testing procedures overcome these limitations and apply this approach to data from a cohort of subjects at risk for cardiovascular disease. Four (4) genetic polymorphisms in three genes of the same gene family are considered. The proposed modelling approach allows us first to test whether there is a significant genetic contribution to the variability observed in our disease outcome. This contribution may be through main effects of multi-locus genotypes or through an interaction between genotype and environmental factors. This approach also enables us to identify specific multi-locus genotypes that interact with environmental factors in predicting the outcome. Mixed effects models provide a flexible statistical framework for controlling for potential confounders and identifying interactions among multiple genes and environmental factors that explain the variability in measures of disease progression. Copyright © 2005 John Wiley & Sons, Ltd.

**KEY WORDS:** multi-locus genotype; phenotype; mixed models; high-order interactions; cardiovascular disease

\*Correspondence to: Andrea S. Foulkes, Department of Biostatistics, University of Massachusetts, School of Public Health, 404 Arnold House, 715N. Pleasant Street, Amherst, MA 01003-9304, U.S.A

<sup>†</sup>E-mail: foulkes@schoolph.umass.edu

<sup>‡</sup>Assistant Professor of Biostatistics.

Contract/grant sponsor: NCCR/NIH; contract/grant number: M01-RR00040

Contract/grant sponsor: National Center for Research Resources; contract/grant number: NIH-RR15532-02

Contract/grant sponsor: NIH; contract/grant number: RO1 HL73278-01

Contract/grant sponsor: W.W. Smith Charitable Trust; contract/grant number: H0204

Contract/grant sponsor: NHLBI

Contract/grant sponsor: HIDDK

*Received November 2003*

*Accepted June 2004*

## 1. INTRODUCTION

There has been a marked increase in the prevalence of obesity and related metabolic disorders in our society over the past decade [1,2]. Obesity, a major risk factor for cardiovascular morbidity and mortality, has a combination of environmental and genetic influences [3–6]. The major environmental contributors of overeating and physical inactivity are well described whereas the genetic influences are poorly understood [7]. Heritability studies suggest a strong familial component to obesity [8] and linkage analyses have identified a number of novel obesity-related loci in different populations [9–11]. In addition, case-control studies have found associations between a number of candidate gene single nucleotide polymorphisms (SNPs) and obesity measures, although results have not been consistent across studies [7]. These inconsistencies may relate to study design, differences in study populations or to complex higher-order interactions between candidate genes or genes and environmental factors that may not be detected when looking at the main effect of individual gene SNPs [12].

The strong genetic predisposition to obesity, at a time of increasing dietary excess and sedentary lifestyles, underscores the need for a greater understanding of the complex interaction of genetic and environmental influences on the development of obesity and its resultant complications. Because of the potential co-ordinated effects of distinct lipase gene family members on plasma triglyceride, free fatty acid bioavailability and adipocyte fat storage [13–16], we examine combinations of four SNPs in three lipase genes as an example to test for higher-order interactions between genes of candidate gene pathways that may act synergistically to influence obesity.

Characterizing the association between multiple SNPs and disease outcomes will offer new insight into disease aetiology while providing tools for making individualized treatment decisions. However, this presents an analytic challenge due to the large number of SNPs and the complex, uncharacterized relationships among them. Several methods have been proposed recently for the analysis of SNP data. Standard regression techniques can provide insight into main effects of single-locus SNPs while controlling for potential confounders by clinical and environmental factors; however, these techniques may be inadequate for identifying gene–gene and high-order interactions given sample size limitations. Recursive partitioning (RP) is a natural approach to identifying structure in complex data [17] and its usefulness for this setting is described in Reference [18]. RP can provide insight into high-order interactions that cannot be obtained with traditional modelling techniques. In addition, it is non-parametric and does not impose the constraints of ordinary linear regression. However, the structure imposed by the tree can be limiting in certain data settings. For example, consider the setting in which the presence of two SNPs leads to a highly unfavourable outcome, while each individual polymorphism is relatively unimportant (compared to other polymorphisms under consideration.) In this case, the tree may not identify the two-way interaction since a split on either of the individual SNPs is unlikely.

Combinatorial partitioning (CPM) uses a sums of squares criterion similar to the criterion proposed in the context of RP but has the additional benefit of considering multiple combinations of genotypes [19]. Unfortunately, CPM is computationally difficult when examining interactions among three or more polymorphisms and methods for incorporating covariates with CPM have not been described. Using a combination of dimension reduction techniques (described below) and recursive partitioning may also lend insight into high-order gene–gene

or gene-environment interactions as described in References [20, 21] but has similar limitations in terms of incorporating covariates.

Permutation procedures and related methods have been proposed recently for high-dimensional data settings including complex disease-gene association studies [22] and microarray analyses [23]. The usefulness of these tools for reducing the number of candidate SNPs dramatically has been demonstrated [22]. Combining dimension reduction techniques and permutation testing is described for microarray data [24] and HIV genetics data [25]. The former utilizes a permutation test procedure to test for association while the latter is based on a confidence band for a normal probability plot. While these approaches can provide insight into high-order gene-gene interactions, extensions that allow us to control for covariates or assess gene-environment interactions have not been described.

Dimension reduction techniques such as clustering have also been described. In the microarray setting, cluster analysis is used to group genes that have similar levels of expression across all observations. For a given individual, the average expression level over all genes within a cluster is then used as a summary measure and testing is performed based on this measure [24]. Cluster analysis has also been used in the context of HIV to group patients with genotypically similar viral populations [20, 26]. An alternative dimension reduction technique termed *patterning* is described in the HIV literature [21, 25] and involves assigning observations to the same group when the corresponding multi-locus genotypes are *identical*. In the context of analysing SNP data, the associations between disease status and groups based on patterning and hierarchical and *K*-means clustering is described in Reference [27].

Finally, a few recent manuscripts describe methods for using a combination of clustering and mixed effects models in high-dimensional data settings [26, 28]. For example, Reference [28] involves constructing prediction intervals around cluster specific latent factors where clusters are defined as groups of *genes* while Reference [26] uses the model fitting procedure to estimate clinically relevant parameters relating to clusters of *patients*. We propose extending these approaches to identify significant multi-locus genotypes and interactions between genotypes and environmental or host specific factors. This method may offer new insight into the effects of combinations of genetic polymorphisms on disease outcome measures. It provides a flexible statistical framework for controlling for potential confounders and allows for identification of interactions between multi-locus genotypes and environmental and host-specific factors.

The following section describes the proposed method. Section 3 illustrates the approach using data from a cohort of 914 asymptomatic subjects at risk for cardiovascular disease and four lipase gene family polymorphisms. We also provide an example of a *post hoc* approach to identifying interactions between SNPs and environmental factors based on a permutation testing procedure and standard regression modelling. Finally, a discussion of the relative merits of the mixed modelling approach is provided in Section 4.

## 2. METHODS

In this section, we begin by defining *genotype groups* and then describe a general framework for modelling the relationships among these groups, phenotype and environmental factors. Finally, we describe approaches to testing the significance of genotype effects and their interactions with environmental factors in predicting phenotype.

### 2.1. Creating genotype groups

Genotype groups are defined simply as groups of individuals with similar or identical multi-locus genotypes. As described above, organizing people into groups based on their multi-locus genotype is described in analyses of human genetic data (see for example Reference [27]). Our motivation for this first stage dimension reduction includes the following: First, it is a computationally useful dimension reduction step. That is, we are able to reduce the number of variables from multiple SNPs to a single genotype group indicator. Secondly, in the context of understanding risk for complex diseases, such as cardiovascular disease, lung disease or cancer, it is intuitively (and potentially clinically) appealing to think of an individual's multi-locus genotype as a single contributor. Finally, modelling the relationship between genotype groups and phenotype will potentially capture information on how genes are interacting with each other, even in the absence of main SNP effects.

Suppose we have  $N$  observations and  $S$  single-locus genotypes for each observation. We begin by grouping these observations into  $L$  groups so that observations with similar or identical genotypes across the  $S$  loci are in the same group while observations with different genotypes are in different groups. Two approaches to arriving at genotype groups can be used. The first, which we term *patterning* is simply to group observations with identical genotypes into the same genotype group. This technique is useful if any of the following criteria are met: (1) Polymorphisms are highly correlated so that the presence of a polymorphism at one locus tends to occur with a particular polymorphism at another locus. For example, this is observed when genes are in linkage disequilibrium and is referred to as a haplotype effect. (2) The number of loci under consideration is relatively small or (3) The population prevalence of a single locus polymorphism is low. In general, if there are three possible genotypes (homozygous wildtype, heterozygous, or homozygous variant) at each of  $S$  loci, then there are  $3^S$  possible genotype patterns. For example, if  $S=4$ , there are 81 possible patterns. If (1) or (3) holds then the number of observed patterns may be substantially less than the number of possible patterns as we see in the example provided in Section 3.

Alternatively, genotype groups can be defined as groups of people with similar (though not necessarily identical) genotypes across the  $S$  loci. Standard cluster analysis techniques, such as hierarchical or  $k$ -means clustering [29] as well as phylogenetic analyses such as neighbour joining [30] can be used to arrive at these groups. Cluster analysis generally requires a measure of similarity between pairs of observations. Several approaches to defining distance between genetic sequences have been proposed [31–33]. In the context of analysing SNP data, distance has been defined as the squared Euclidean distance between sequences where genetic sequences are coded as sequences of numbers each representing homozygous wildtype, heterozygous, and homozygous variant for the corresponding locus [27]. While patterning has the advantage of creating well-defined groups (i.e. assignment of new individuals to these groups is straightforward), clustering is useful when the number of patterns approaches the number of observations in the data set. Clustering also provides a means of grouping patients with rare multi-locus genotypes and has the advantage that prior information about the importance of particular polymorphisms (e.g. recessive vs dominant effect) can be used via a weight matrix in the distance calculations. Hartigan's criterion [29] and the Gap statistic [34] are useful approaches to determining the appropriate number of clusters.

## 2.2. A permutation test procedure

Here, we describe a permutation test procedure that represents a straightforward extension of the method described in Reference [22]. This method is illustrated in Section 3 and compared to the proposed mixed modelling approach. Let  $g_1, \dots, g_L$  represent the genotype groups resulting from patterning or clustering observations. We begin by defining one of these groups to be a reference group. The choice of reference group will depend on the goal of testing. In our case, we assign the group with the largest number of observations to be the reference group. In other words, the most prevalent genotype group is treated as the reference group. We refer to this most prevalent genotype group as the *wildtype* group and without loss of generality we denote it  $g_L$ . We now calculate a summary statistic for each of the remaining groups which we denote,  $W_1, \dots, W_{L-1}$ . For example, a continuously distributed outcome,  $W_l$  may be the  $T$ -statistic or Wilcoxon rank sum statistic for testing the null hypothesis  $H_0 : \mu_l = \mu_L$  where  $\mu_l$  is the true population mean of the response  $Y$  for group  $g_l$ ,  $l = 1, \dots, L - 1$ . Alternatively, for a binary response,  $W_l$  can be defined as the chi-squared statistic for a test of difference between the proportion of responders in group  $l$  versus the reference group.

Let the order statistics corresponding to these test statistics be denoted  $W_{(1)}, \dots, W_{(L-1)}$  where  $W_{(1)} < W_{(2)} < \dots < W_{(L-1)}$ . We aim to determine whether the observed order statistics are different from what we expect to see under the null hypothesis of no association. To achieve this we consider the following permutation test procedure:

1. Randomly permute the values of  $Y$ .
2. Calculate a summary statistic for each genotype group:  $w_{(1)}, \dots, w_{(L-1)}$ .
3. Repeat steps 1 and 2,  $P$  times. Let  $w_{p(1)}, \dots, w_{p(L-1)}$  be the order statistics for the  $p$ th iteration,  $p = 1, \dots, P$ .
4. Compare the observed order statistics to the distribution of these expected order statistics.

In the final step, we use the distribution of expected order statistics to determine the probability that what we observe arises from the null distribution. Suppose for example that the first observed order statistic is  $W_{(1)}$ . Further suppose that  $W_{(1)} < w_{p(1)}$  for  $k$  values of  $p$ . The  $p$ -value associated with this observed order statistic is then given by  $k/P$ . If  $P = 1000$ , and  $W_{(1)} < w_{p(1)}$  for 25 of these permutations, then the probability that we observe something as extreme as we do is  $25/1000 = 0.025$ .

Adjustments for covariates can be achieved by first fitting a model for the response as a function of these predictor variables. The residuals from the final model can then be used as the response,  $Y$  in the procedure described above. This approach allows for identification of multi-loci genotypes that provide information above that of the selected covariates. Alternatively, subjects can be stratified according to the value of a covariate and the analysis can be repeated within each strata. Finally, genotype groups that are significant based on the permutation test using the raw data (i.e. without fitting a model) can be included in a regression model and their significance can be assessed with additional variables in the model. Interactions between this subset of significant genotype groups and other factors can also be modeled.

## 2.3. The mixed effects modelling framework

This section describes a mixed modelling framework that will allow us to: (1) assess whether there is a significant genetic contribution to the variability observed in our disease

outcome. This contribution may be through main effects of multi-locus genotypes or through an interaction between genotype and environmental factors and (2) identify specific multi-locus genotypes that interact with environmental factors in predicting the outcome. The first set of hypotheses are tested using standard model fitting procedures based on likelihood ratio tests. Posterior means of random pattern effects and corresponding variance estimates are used to assess significance of specific multi-locus genotypes and genotype–environment interactions. These are described in more detail below. First, the mixed effects model is provided.

Again let  $g_1, \dots, g_L$  represent the genotype groups resulting from patterning or clustering observations. These observed groups can be thought of as a random sample from the general population of genotypes. It is therefore natural to treat them as random effects in a linear model. The general form of this model is given in Reference [35] and equation (1) where  $Y$  is a vector of responses,  $X$  and  $Z$  are matrices of covariates and  $\beta$  is the vector of parameters corresponding to  $X$ . We assume  $b \sim N(0, D)$  is a vector of random genotype group effects,  $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$  is measurement error and  $b$  is independent of  $\varepsilon$ . This model provides a flexible framework for adjusting for observed potential confounders (through the fixed effects) and allowing genotype effects to vary across covariate values (through the random effects.)

$$Y = X\beta + Zb + \varepsilon \quad (1)$$

Consider for example a setting in which we observe  $L$  multi-locus genotypes and interest lies in assessing whether there is an interaction between any of these genotypes and a covariate (for example, gender) on an outcome measure,  $Y$ . Further suppose that we want to control for a potentially confounding variable,  $x$ . In this case equation (1) reduces to equation (2) where  $i = 1, \dots, L$  represents genotype and  $j$  represents individual so that  $Y_{ij}$  is the outcome for the  $j$ th observation with genotype  $i$  and  $x_{ij}$  is this person's corresponding covariate value. We refer to  $b_{0i}$  as the intercept random genotype effect and  $b_{1i}$  as the slope or gender random genotype effect. In this example,  $b = [\mathbf{b}_0, \mathbf{b}_1]$  where each  $\mathbf{b}_k = (b_{k1}, b_{k2}, \dots, b_{kL})$ ,  $k = 0, 1$ , is a vector of length  $L$ .

$$Y_{ij} = \beta_0 + \text{gender}_{ij} \beta_1 + x_{ij} \beta_2 + b_{0i} + \text{gender}_{ij} b_{1i} + \varepsilon_{ij} \quad (2)$$

As mentioned above, there are two stages of testing that may be of clinical interest. First, we want to know whether there is an overall effect of genotype and/or an interaction between genotype and a specific environmental factor on the outcome. Second, if there is such an effect, which multi-locus genotypes differ from the others. A likelihood ratio test can be employed to test for departures from the null hypothesis that the random genotype and genotype by environmental factor random effects have zero variance. Rejection of this null would suggest that these factors interact in predicting our outcome,  $Y$ .

In the presence of a significant interaction, interest lies in identifying the particular genotype patterns whose interaction with gender is significantly greater (or less than) zero. We now describe how to estimate the random effect and construct pointwise prediction intervals around them in order to test that they are equal to zero. As described in Reference [36], it is straightforward to show that the conditional expectation of  $b$  given the observed data  $y$  is given by  $E(b|y) = DZ'V^{-1}(y - X\beta)$ . Here  $V$  is the variance of  $Y$  and  $D$  is the variance of  $b$ . The best linear unbiased predictor (BLUP) of  $b$ , denoted  $\hat{b}$  is given by replacing  $\beta$  by its weighted least squares estimate  $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$  and described by equation (3) [36]. Furthermore,

it can be shown that  $\text{var}(\tilde{b}) = W = DZ'PZD$  where  $P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$  and  $\text{var}(\tilde{b} - b) = D - DZ'PZD$  [36].

$$\tilde{b} = E(b|y) = DZ'V^{-1}(y - X\hat{\beta}) \quad (3)$$

Pointwise prediction intervals for the random effects are given by equation (4). Under the null hypothesis that a specific random effect equals 0 we expect this prediction interval to contain 0. If the prediction interval does not contain 0 then we will reject this null hypothesis in favour of the alternative that the random effect is not equal to 0. Returning to our example, we can construct pointwise prediction intervals around each of the random genotype pattern effects and random gender by pattern effects using the formula in equation (4). If the prediction interval for a given random effect does not contain 0 then we conclude that with  $(1 - \alpha)100$  per cent certainty the true random effect is not equal to 0; i.e. the effect of a genotype pattern is different from 0 or a genotype pattern effect on the outcome varies by gender. Note that in choosing  $\alpha$ , consideration needs to be given to the number of tests.

$$\text{PI}(b) = \tilde{b} \pm z_{\alpha/2}(D - DZ'PZD)^{1/2} \quad (4)$$

### 3. EXAMPLE: PATIENTS AT RISK FOR CARDIOVASCULAR DISEASE

The Study of Inherited Risk of Coronary Atherosclerosis (SIRCA) is a cross-sectional study of genetic factors associated with cardiovascular risk factors and coronary atherosclerosis in healthy volunteer subjects enriched for family history of premature coronary artery disease (CAD). Study design and preliminary findings have been published [37, 38]. Entry criteria included (1) men aged 30–65 and women aged 35–70 and (2) family history of premature CAD, defined as having at least one first degree relative with premature CAD (males prior to age 60 and females prior to age 70). Exclusion criteria included any established clinical CAD and were designed to exclude extremes of major established CAD risk factors (known diabetes, total cholesterol >300 mg/dL; cigarette smoking >1 pack per day, blood pressure >160/100 mmHg, and serum creatinine >3.0) in order to facilitate the discovery of novel genetic influences on CAC.

Obesity, a complex trait with a strong genetic component [3, 4, 7], is a major risk factor for cardiovascular disease [5, 6, 39]. A family of lipase enzymes (hepatic lipase, lipoprotein lipase and endothelial lipase) plays important and co-ordinated roles in the metabolism of plasma triglyceride and regulation of plasma free fatty acid bioavailability for uptake and storage in adipocyte [13–16]. Thus, combinations of common genetic variations in this pathway of genes involved in lipoprotein remodelling may act synergistically to influence adipose tissue fat stores and obesity. For the purpose of these analyses, we examine the association of four SNPs in three lipase genes (endothelial lipase—EL 2237 3'UTR (EL3'UTR') and EL-Thr111Ile (ELexon3), hepatic lipase—HL-514C/T (HL-514) and lipoprotein lipase—LPL hindIII/intron (LPLhindIII)) with body mass index (BMI), a measure of obesity used commonly in epidemiological studies and clinical practice, in the SIRCA cohort.

Genotyping of lipase single nucleotide polymorphisms (SNPs) was performed on Orchid's 25K SNPstreamTM platform (Orchid Bioscience, Princeton, NJ) using a high throughput single-base primer extension method. A set of three primers was chosen for each SNP; two PCR primers were selected to amplify a 100–200 base pair product and a 25 bp single-base

Table I. Observed genotype patterns with at least 10 observations.

Pattern	EL3'UTR'			ELexon3			HL-514			LPLhindIII			N
	AA	GA	GG	CC	CT	TT	CC	CT	TT	TT	GT	GG	
1		×		×			×					×	14
2	×				×		×					×	14
3	×			×			×				×		22
4		×		×			×				×		42
5			×	×			×				×		14
6	×				×		×				×		39
7		×			×		×				×		40
8	×					×	×				×		26
9			×	×				×		×	×		13
10		×		×				×		×	×		32
11	×			×				×		×	×		13
12	×				×			×		×	×		23
13		×			×			×			×		30
14	×			×			×			×			30
15		×		×			×			×	×		57
16			×	×			×			×	×		32
17	×				×		×			×	×		50
18		×			×		×			×	×		38
19	×					×	×			×	×		13
20	×			×				×		×	×		24
21		×		×				×		×	×		29
22	×				×			×		×	×		29
23		×			×			×		×	×		18
												Total	642

primer extension (SBE) primer. (EL3'UTR: AGTGCAACCCAWGAGAWCCCCAACAGC, GTGTTCAATAGACATTTGCTCAATTA, GTACTCTGCCTGACGAGGAAC; ELexon3: GGGGAGCCAGTCAACCAC, AACTACATTGGCGTCTTTCTCTCTT, TGCAGATGAGCGGTATCTTTG; LPLhindIII: AACATTACCCAGZTTGATCATGTA, AAAATGGATGTGAATATGCCATG, ATTCTGATGTGGCCTGAGTGT; HL-514: GTGTGZTGCAGAAAA-CCCTTZACCCCC, CAAATTTCTGTTGGGTTTCAGTGA, GTCACCTGGCAAGGGCATC). SNP-IT primers were extended by one base at the polymorphic site of interest using DNA polymerase I. The extension mixes contained two labelled terminating nucleotides (one fluorescein, one biotin). An ELISA-based technique was utilized for detection of the extension product; antiluorescein-alkaline phosphatase (Boehringer Mannheim, Indianapolis, IN) for fluoresceinated nucleotides (allele 1) and antibiotin-horseradish peroxidase conjugate (Zymed, San Francisco, CA) for biotinylated nucleotides (allele 2). Raw OD data were analysed by GetGenos software program. Genotype calls were corroborated by visual inspection.

We begin by creating genotype groups using the patterning approach. Of the 914 subjects in our cohort, 738 (80.7 per cent) are non-Hispanic Caucasian with complete genotype and BMI information and are considered in subsequent analyses. A total of 55 (of a possible  $3^4=81$ ) patterns are observed. The 23 patterns with at least 10 observations (representing 76.4 per cent ( $N=642$ ) of complete observations) are described in Table I. We refer to the most prevalent pattern (pattern 15 in this case) as the *wildtype* pattern. This pattern is treated as the reference group for subsequent analyses.

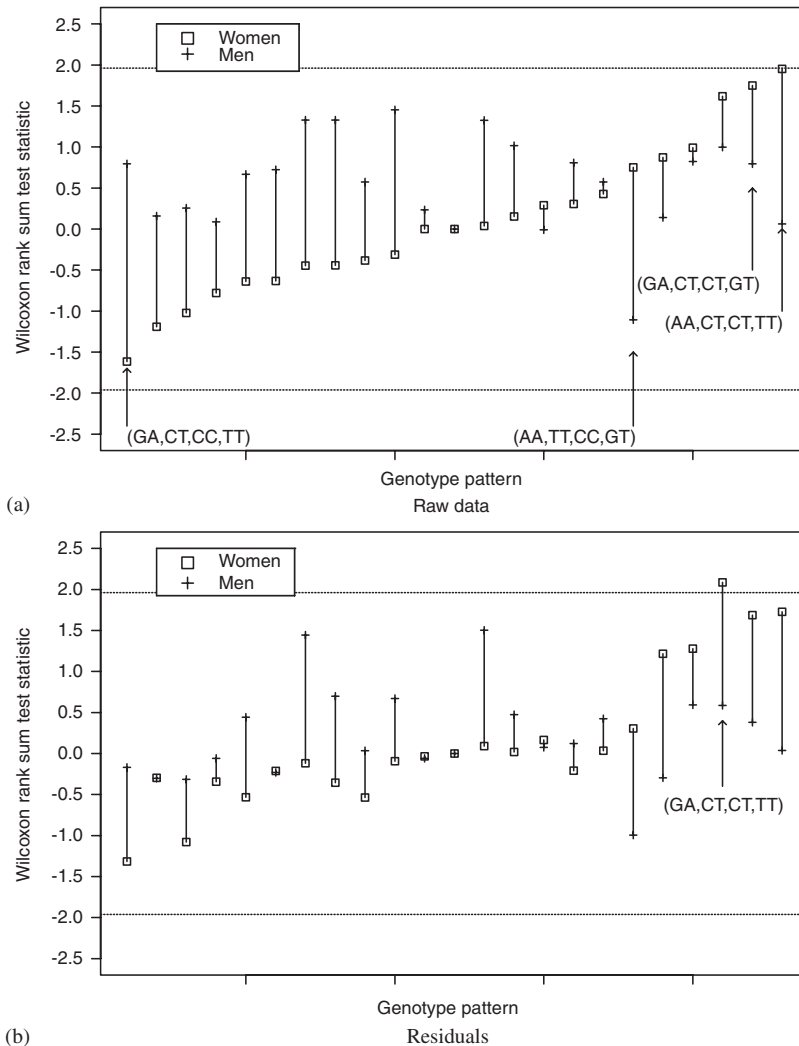


Figure 1. Wilcoxon rank sum statistics by gender: (a) Raw data; and (b) Residuals.

Because of known gender-dependent differences in measures of obesity [40], Wilcoxon rank sum statistics were generated for the tests comparing each genotype pattern to the wildtype pattern within each gender group based on the raw BMI data. This is illustrated in Figure 1(a). The genotypes reported in parentheses in this figure refer to the genotypes for EL3'UTR', ELexon3, HL-514 and LPLhindIII, respectively. None of the patterns are significant (combining genders and within gender groups) based on the permutation test procedure. Also provided in Figure 1(b) are the Wilcoxon test statistics based on the residuals from a model for BMI that includes main effects for gender, age, age<sup>2</sup>, exercise and systolic blood pressure (BP). The test statistics can be thought of as corresponding to tests of genotype effects after adjusting for

Table II. Summary of model fitting procedure.

	Fixed effects	Random pattern effects	LRT	Comparison
Model I	gender, age, age <sup>2</sup> , exercise, BP, smoke, alcohol	intercept, gender		
Model II	gender, age, age <sup>2</sup> , exercise, BP, smoke	intercept, gender	0.0576	(II vs I)
Model III*	gender, age, age <sup>2</sup> , exercise, BP	intercept, gender	3.67	(III vs II)
Model IV	gender, age, age <sup>2</sup> , exercise	intercept, gender	97.24**	(IV vs III)
Model V	gender, age, age <sup>2</sup> , BP	intercept, gender	23.71**	(V vs III)
Model VI	gender, age, exercise, BP	intercept, gender	9.01**	(VI vs III)
Model VII	gender, age, age <sup>2</sup> , exercise, BP	intercept	9.74**	(VII vs III)

\*Final model.

\*\*Indicates the test statistic is significant at the 0.05 level.

these factors. In general, this adjustment tends to result in more similar test statistics between males and females. Again none of the tests (combining gender and within gender groups) are significant using the permutation procedure.

Interestingly, using either the residuals or the raw data, the test statistics vary between males and females. The largest positive statistic in women is genotype (*AA, CT, CT, TT*) while in men it is (*AA, CT, CC, GT*). On the other hand the largest negative test statistic is (*GA, CT, CC, TT*) in women and (*AA, TT, CC, GT*) in men. While the differences in the value of the test statistics between genders suggest a possible interaction between genotype and gender, this approach does not permit quantification of the significance of this interaction. We now demonstrate how a mixed modelling procedure will provide the tools to achieve this objective while similarly allowing us to identify multi-locus genotype effects and control for observed potential confounders.

We now consider a mixed model for log(BMI) as the outcome and random pattern specific effects. A series of models are fit to the data, as described in Table II. Likelihood ratio tests are used to arrive at the best model (in this case Model III). This model includes random genotype pattern effects as well as random genotype by gender effects. Restricted maximum likelihood mean parameter estimates for the final model are given by  $\hat{\beta} = (2.4, 0.029, 0.018, -0.00019, -0.060, 0.0040)$  for the intercept, gender, age, age<sup>2</sup>, exercise and systolic blood pressure (BP), respectively. The estimated variance parameters are  $\hat{\sigma}_e^2 = 0.023$ ,  $\hat{D}[1, 1] = 0.0029$ ,  $\hat{D}[1, 2] = -0.0026$  and  $\hat{D}[2, 2] = 0.0023$  where  $D$  is block diagonal. A quantile–quantile plot of residuals from the final model is provided in Figure 2 and suggests a reasonable fit with a slightly heavy right tail.

Interestingly, the model fitting procedure results in random pattern and pattern by gender effects. In other words, there is significant variability (greater than zero) across genotype groups in the effect of gender on BMI. Empirical Bayes estimates and corresponding 95 per cent prediction intervals for the random pattern effects and pattern by gender effects are illustrated in Figure 3. Here patterns with at least 10 observations are illustrated though all patterns are included in the model fitting procedure. The analysis of the pattern by gender random effects suggests that genotypes (*GA, CT, CT, TT*) and (*GA, CT, CT, GT*) are associated with a lower BMI in men compared to women while genotypes (*GA, CT, CC, TT*) and (*AA, CT, CC, TT*) are associated with a higher BMI in men than women. This results are consistent with the findings based on the stratified analysis described by Figure 1. After adjusting for multiple testing, we cannot identify a specific pattern or pattern by gender effects is

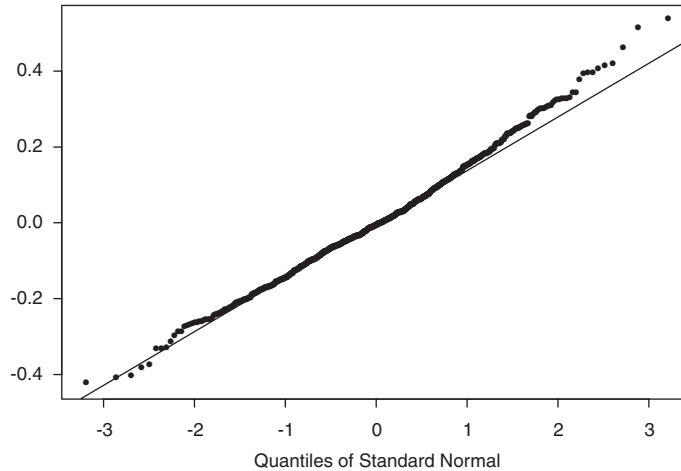


Figure 2. Quantile–quantile plots of residuals from final model.

different from 0. In summary, while we are unable to detect a specific multi-locus genotype for which the effect of gender on BMI differs from zero, our results suggest there is significant variability in the effect of gender on BMI across genotypes.

#### 4. DISCUSSION

This paper demonstrates the usefulness of combining dimension reduction and mixed modelling for the analysis of multiple SNP–phenotype association data. Mixed effects modelling is a powerful parametric tool for drawing from the totality of the data to make inferences about individual groups. Importantly, it provides a flexible framework for adjusting for observed potential confounders in association studies while providing the tools for examining high-order gene–gene and gene–environment interactions. In the example provided, we consider the relationships among four SNPs in three lipid genes, multiple environmental variables and BMI as phenotype.

The analyses described in Section 3 resulted in several interesting findings that are not observable in univariate analyses. First, based on the results of the Wilcoxon tests (Figure 1) there is some suggestion that women with genotype ( $GA, CT, CT, TT$ ) have a greater BMI than women with the most prevalent genotype, ( $GA, CC, CC, TT$ ) while this is not true for men. The results of the mixed modelling (Figure 3) similarly draw attention to genotype ( $GA, CT, CT, TT$ ), in this case suggesting that women with this genotype have a greater BMI than men with the same genotype. In both cases, adjustments for the potential confounding effects of age, exercise, systolic blood pressure, smoking and alcohol use were made. Based on the mixed modelling, we found genotype ( $GA, CT, CT, GT$ ) had a similar effect across gender while genotypes ( $AA, CT, CC, TT$ ) and ( $GA, CT, CC, TT$ ) resulted in a greater BMI in men than women. While suggestive of a possible relationship, the results were not statistically significant after adjusting for multiple testing.

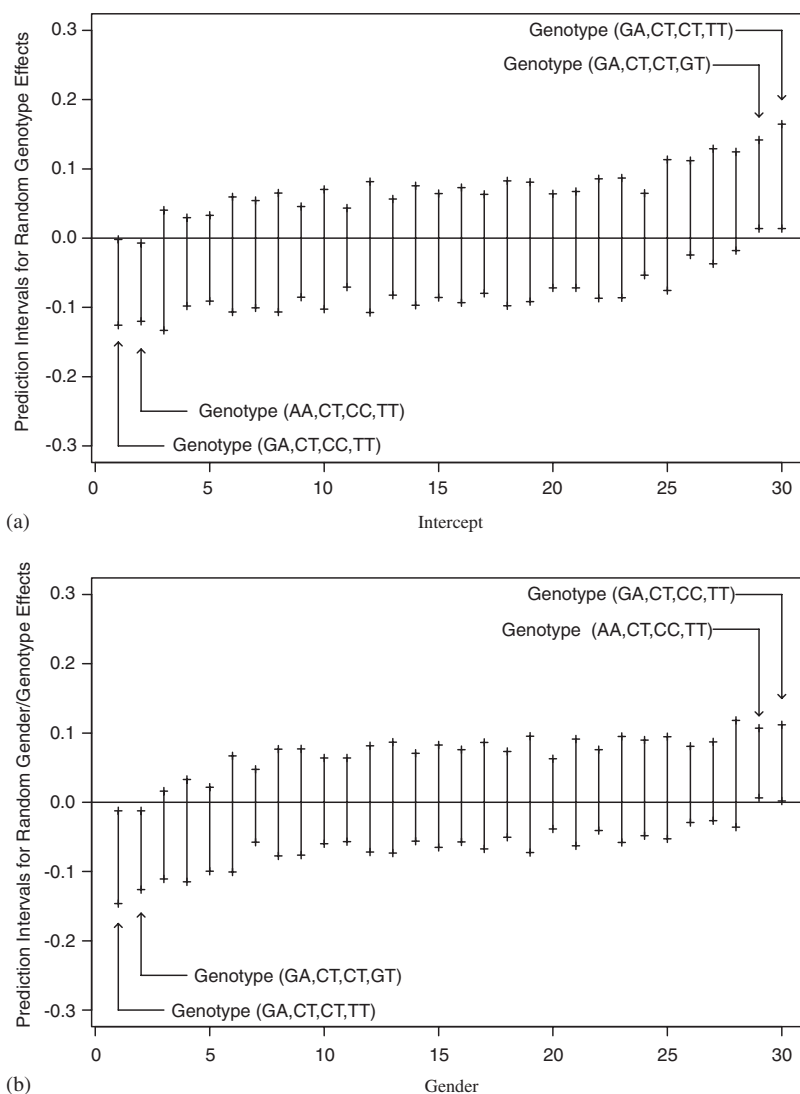


Figure 3. Ninety-five per cent prediction intervals for random pattern effects: (a) Intercept; and (b) Gender.

The mixed model fitting procedure did, however, identify a significant random pattern by gender interaction effect. In other words, the distribution of random genotype by gender effects has a non-zero variance, indicating that not all genotype by gender effects are the same. This suggests that accounting for this interaction improves prediction though further research is required to assess the clinical significance of this contribution. It also suggests that there is a significant genetic contribution to the variability in BMI explained by the three genes described in this example.

The primary advantage of first defining genotype groups is that it allows for assessing the effect of multiple polymorphisms within and across many gene loci. Furthermore, discovery of interactions is not precluded in the absence of main effects. However, one limitation of this approach is the potentially large number of patterns. As described in Section 2, one alternative is to define genotype groups using a clustering algorithm that groups patients with similar, though not necessarily identical genotypes. However, depending on the patient population and genes under consideration, the data may not result in well defined clusters. Another alternative is to create genotype patterns based on subsets of the genes under consideration. For example, if we are examining 10 genes, we can define patterns based on combinations of three of these genes. These subsets can be chosen exhaustively or based on prior biological information about the relationships among the genes and disease aetiology. This is similar to examining three-way SNP interactions. Using mixed models, we could then assess simultaneously the contributions of multiple pattern variables, each constructed using a different subset of genes. Emerging information on haplotype block structure will facilitate the use of smaller numbers of SNPs and therefore genotype groups to address issues of multi-locus interactions in association with clinical phenotypes.

Finally, in the example provided we considered a continuous outcome (BMI) and a potential interaction between genotype and a categorical variable (gender). Application of these methods to settings in which the response is binary and/or the environmental variable is continuous is straightforward. In the case of a categorical outcome, a non-linear mixed effects model may be appropriate [36]. Methods for characterizing the relationships among genotypes and environmental factors and their impact on measures of disease progression have broad implications for biological research across multiple disciplines. Furthermore, understanding how environmental factors interact with genotype is a key component in making more informed, individualized treatment decisions. While this manuscript focuses on one clinical data setting, the methods can be applied broadly to settings in which more polymorphisms are available or there is interest in characterizing additional gene–environment interactions.

#### ACKNOWLEDGEMENTS

This study was funded in part by a Grant M01-RR00040 from the NCRR/NIH supporting the University of Pennsylvania General Clinical Research Center (GCRC). M.P.R. is supported by a Mentored Patient-Oriented Research Career Development Award from the National Center for Research Resources (NIH-RR15532-02), by (RO1 HL73278-01) from the NIH, and by the W.W. Smith Charitable Trust (H0204). D.J.R. is supported by grants from the NHLBI, HHDK and NCRR and is a recipient of the Burroughs Wellcome Fund Clinical Scientist Award in Translational Research and a recipient of a Doris Duke Distinguished Clinical Investigator Award.

#### REFERENCES

1. Ford ES, Giles WH, Dietz WH. Prevalence of the metabolic syndrome among us adults: findings from the third national health and nutrition examination survey. *Journal of the American Medical Association* 2002; **287**: 356–359.
2. Flegal KM, Carroll MD, Ogden CL, Johnson CL. Prevalence and trends in obesity among us adults. *Journal of the American Medical Association* 2002; **288**:1723–1727.
3. Hill JO. Genetic and environmental contributions to obesity. *American Journal of Clinical Nutrition* 1998; **68**:991.
4. Aitman TJ. Genetic medicine and obesity. *New England Journal of Medicine* 2003; **348**:2138–2139.
5. Kannel WB, Cupples LS, Ramaswami R, Stokes 3rd J, Kreger BE, Higgins M. Regional obesity and risk of cardiovascular disease: the Framingham study. *Journal of Clinical Epidemiology* 1991; **44**:183–190.

6. Rashid MN, Fuentes F, Touchon RC, Wehner PS. Obesity and the risk for cardiovascular disease. *Preventive Cardiology* 2003; **6**:42–47.
7. Swarbrick MM, Vaisse C. Emerging trends in the search for genetic variants predisposing to human obesity. *Current Opinion in Clinical Nutrition* 2003; **6**(4):6369–6375.
8. Pausova Z, Gossard F, Gaudet D, Tremblay J, Kotchen RA, Cowley AW, Hamet P. Heritability estimates of obesity measures in siblings with and without hypertension. *Hypertension* 2001; **38**:41–47.
9. Li WD, Li S, Wang S, Zhang S, Zhao H, Price RA. Linkage and linkage disequilibrium mapping of genes influencing human obesity in chromosome region 7q22.1–7q35. *Diabetes* 2003; **52**:1557–1561.
10. Deng HW, Deng H, Liu YJ, Liu YZ, Xu FH, Shen H, Conway T, Li JL, Huang QY, Davies KM, Recker RR. A genome-wide linkage scan for quantitative-trait loci for obesity phenotypes. *American Journal of Human Genetics* 2002; **70**:1138–1151.
11. Bray MS, Boerwinkle D, Hanis CL. Linkage analysis of candidate obesity genes among the Mexican-American population of Starr County, Texas. *Genetic Epidemiology* 1999; **16**:397–411.
12. Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. *Science* 2002; **298**(5602):2345–2349.
13. Clement K, Dina C, Basdevant A, Chastang N, Pelloux X, Lahlou N, Berlan M, Langin D, Guy-Grand B, Froguel P. A sib-pair analysis study of 15 candidate genes in french families with morbid obesity: indication for linkage with islet 1 locus on chromosome 5q. *Diabetes* 1999; **48**:398–402.
14. Arner P. Genetic variance and lipolysis regulation: implications for obesity. *Annals of Medicine* 2001; **33**:542–546.
15. Jin W, Marchadier D, Rader DJ. Lipases and hdl metabolism. *Trends in Endocrinology and Metabolism* 2002; **13**:174–178.
16. Deeb SS, Zambon A, Carr MC, Ayyobi AF, Brunzell JD. Hepatic lipase and dyslipidemia: interactions between genetic variants, obesity, gender and diet. *Journal of Lipid Research* 2003; **44**(7):1279–1286.
17. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Wadsworth Inc.: Belmont, CA, 1984.
18. Zhang H, Bonney G. Use of classification trees for association studies. *Genetic Epidemiology* 2000; **19**:323–332.
19. Nelson M, Kardia S, Ferrell R, Sing C. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research* 2001; **11**:458–470.
20. Sevin A, De Gruttola V, Nijhuis M, Schapiro JM, Foulkes AS, Para MF, Boucher CAB. Evaluating the relationship between drug susceptibility phenotype and genotype among HIV from patients treated with protease inhibitors. *The Journal of Infectious Diseases* 2000; **182**:59–67.
21. Foulkes AS, DeGruttola V, Hertogs K. Combining genotype groups and recursive partitioning: an application to HIV-1 genetics data. *Journal of the Royal Statistical Society C, Part 2* 2004; **53**:311–323.
22. Hoh J, Wille A, Zee R, Cheng S, Reynolds R, Lindpaintner K, Ott J. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Annals of Human Genetics* 2000; **64**:413–417.
23. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001; **98**(9):5116–5121.
24. Tibshirani R, Hastie T, Balasubramanian N, Eisen M, Sherlock G, Brown P, Botstein D. Exploratory screening of genes and clusters from microarray experiments. *Statistica Sinica* 2002; **12**(1):47–60.
25. DiRienzo G, DeGruttola V. Nonparametric methods to predict HIV drug susceptibility phenotype from genotype. *Statistics in Medicine* 2003; **22**(17):2785–2798.
26. Foulkes AS, DeGruttola V. Characterizing the relationship between HIV-1 genotype and phenotype: prediction based classification. *Biometrics* 2002; **58**:145–156.
27. Hoehe M, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd K, Berrettini W, Church G. Sequence variability and candidate gene analysis in complex disease: association of  $\mu$  opioid receptor gene variation with substance dependence. *Human Molecular Genetics* 2000; **9**(19):2895–2908.
28. Li H, Hong F. Cluster-rasch models for microarray gene expression data. *Genome Biology* 2001; **2**(8):research0031.1–0031.13.
29. Hartigan JA. *Clustering Algorithms*. Wiley: New York, 1975.
30. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 1987; **4**:406–425.
31. Jukes TH, Cantor CR. *Mammalian Protein Metabolism III*. Academic Press: New York, 1969.
32. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 1980; **16**:111–120.
33. Nei M, Gojobori T. Simple methods of estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 1986; **3**(5):418–426.
34. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistical Society, Series B* 2001; **32**:411–423.
35. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.

36. McCulloch CE, Searle SR. *Generalized, Linear, and Mixed Models*. Wiley: New York, 2001.
37. Valdes ML, Wolkfe AM, Tate HC, Geftere W, Rut A, Rader DJ. Association of traditional risk factors with coronary calcification in persons with a family history of premature coronary heart disease: the study of the inherited risk of coronary atherosclerosis. *Journal of Investigative Medicine* 2001; **49**:353–361.
38. Valdes AM, Wolfe ML, O'Brien EJ, Spurr NK, Geftere W, Rut A, Groot DJ, Rader PH. Val64ile polymorphism in the c-c chemokine receptor 2 is associated with reduced coronary artery calcification. *Arterio Thrombosis Vascular Biology* 2002; **22**:1924–1928.
39. Hubert HB, Feinleib M, McNamara PM, Castelli WP. Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the framingham heart study. *Circulation* 1983; **67**:968–977.
40. Jackson AS, Stanforth PR, Gagnon J, Rankinen T, Leon AS, Rao DC, Skinner JS, Bouchard C, Wilmore JH. The effect of sex, age and race on estimating percentage body fat from body mass index: the Heritage Family Study. *International Journal of Obesity Related Metabolic Disorder* 2002; **26**(6):789–796.