A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy

Shaoyu Li[1], Qing Lu[2], Wenjiang Fu[2], Roberto Romero[3] and Yuehua Cui[1‡]

[1]*Department of Statistics & Probability,* [2]*Department of Epidemiology, Michigan State University, East Lansing, Michigan 48824*
[3]*The Perinatology Research Branch, NICHD, NIH, DHHS, Bethesda, MD, and Detroit, 48201*

Abstract: Human diseases developed during pregnancy could be caused by the *direct* effects of both maternal and fetal genes, and/or by the *indirect* effects caused by genetic conflicts. Genetic conflicts exist when the effects of fetal genes are opposed by the effects of maternal genes, or when there is a conflict between the maternal and paternal genes within the fetal genome. The two types of genetic conflicts involve the functions of different genes in different genome and are genetically distinct. Differentiating and further dissecting the two sets of genetic conflict effects that increase disease risk during pregnancy present statistical challenges, and have been traditionally pursued as two separate endeavors. In this article, we develop a unified framework to model and test the two sets of genetic conflicts via a regularized regression approach. Our model is developed considering real situations in which the paternal information is often completely missing; an assumption that fails most of the current family-based studies. A mixture model-based penalized logistic regression is proposed for data sampled from a natural population. We develop a variable selection procedure to select significant genetic features. Simulation studies show that the model has high power and good false positive control under reasonable sample sizes and disease allele frequency. A case study of small for gestational age (SGA) is provided to show the utility of the proposed approach. Our model provides a powerful tool for dissecting genetic conflicts that increase disease risk during pregnancy, and offers a testable framework for the genetic conflict hypothesis previously proposed.

Key words: Complex disease, Genetic conflicts, Genomic imprinting, Maternal-fetal genotype incompatibility, Penalized mixture logistic regression

‡ To whom correspondence should be addressed.

1 Introduction

The pregnancy process is unique in mammals in which the fetus resides inside of its mother's womb and there are significant interactions between the mother and the fetus through the placenta. A fetus carries one copy of its mother's genes, hence there is an underlying harmony of interest between a mother and her fetus, as most biologists assumed. On the other hand, a fetus also carries one copy of the genome from its father. The maternal and paternal copies inherited by a fetus may not be identical and the underlying harmony may be broken. The paternal copy in the fetus' genome that favors fetal growth may not always benefit its mother. Moreover, there are intensive chemical exchanges, including the fetus obtaining nutrition and disposing of wastes through its mother's blood (Haig 2004). What is beneficial to the fetus may not always be optimal for its mother which thus affects development of both mother and offspring. For example, the placental hormones may impose a negative effect on maternal receptors and, consequently complications may develop. According to the genetic conflict theory proposed by Haig (1993), when the balance of the mutual benefits is broken — especially under conditions of extreme conflicts — disease may occur.

Any diseases developed during pregnancy may impose great risks to the mother or the fetus or to both. For example, mothers could develop diseases such as preterm delivery or, Pre-eclampsia (PE) with a syndrome of hypertension and proteinuria. PE is a leading cause of maternal mortality, and affects at least 5-7% of pregnancies (Kaunitz et al. 1985). On the other hand, the fetus may be affected by disease and may show growth retardation or overgrowth. Such diseases, developed during pregnancy, have unique genetic bases and are generally complex due to the different interests of the three sets of genes: 1) genes in the mother; 2) maternally derived genes in the fetus; and 3) paternally derived genes in the fetus. At a particular locus, the maternal-fetal pair contains three distinct alleles: the non-inherited maternal alleles (NIMA) in the mother's genome, the maternally derived fetal allele (MDFA) shared by the mother and the fetus, and the paternally derived fetal alleles (PDFA) (Haig 2004). Quite often, the NIMAs do not discriminate among offspring, and are called non-discriminating maternal alleles (Haig 2004). Occasionally, a subset of discriminating maternal alleles may trigger negative effects in fetal growth, and in extreme cases could cause the early demise of an embryo (Haig 2004).

During the gestation period, at least two interrelated sources of conflicts may exist: 1) a conflict between genes expressed in the mother and in the fetus; and 2) a conflict between the MDFA and the PDFA in the fetus' genome (Haig 1993). The first conflict is called either the parent-offspring conflict or

the maternal-fetal (MF) conflict to distinguish it from the second conflict, which can be termed as fetal-fetal (FF) conflict. The MF conflict is due to the genetic differences between the maternal and fetal genes, which is also termed maternal-fetal genotype incompatibility (MFGI). For some disorders, the maternal genotype influences fetal development through the mediation of an altered uterine environment, and may increase the risk of birth defects for the fetus. Meanwhile, the fetus' genotype may produce chemicals that are harmful to its mother, or the fetus may demand a great blood supply more than its mother can provide, resulting in hypertension in the mother — a common symptom of PE (Odent 2001). In both cases, the MF genetic conflict can trigger negative effects in either the mother or the fetus. One well-known example of this is the RhD MFGI which increases susceptibility to schizophrenia (Palmer et al. 2002). When the mother is Rh negative (d/d) and the fetus is Rh positive (D/d) at the RhD locus, harmful effects may be imposed on the fetus due to Rh incompatibility.

The FF genetic conflict can be explained by different investments of the PDFA and the MDFA in the fetus' genome, which is also termed genomic imprinting. When a gene is maternally (or paternally) imprinted, the PDFA (or MDFA) may partially or completely dominate the expression of the other allele, resulting in maternal (or paternal) imprinting (Pfeifer 2000). In the presence of genomic imprinting, the PDFA (or MDFA) may favor greater levels of investment in the fetus than does the MDFA (or PDFA). Examples of genomic imprinting have been increasingly documented in the literature, including both the well-studied maternally imprinted IGF2 genes and the paternally imprinted IGF2R gene.

There have been developed statistical methods for testing MFGI, such as, the MFG test (Sinshermer et al. 2003), the exact MFG test (Minassian et al. 2005), and the v-MFG test (Hsieh et al. 2006a, 2006b). The MFG test is a generalization of the log-linear model developed by Weinberg and her colleagues (Weinberg et al. 1998). Most of these methods have been developed for the case-parents design, which requires genotyping trios that consist of a mother, a father and an affected child. Some of the methods were developed to allow partial missing parental genetic information (e.g., Hsieh et al. 2006b). In a genetic association study related to pregnancy, it often happens that no genetic information is available for the father at all. When this happens, a key step in their method for the inference of identity-by-state, would fail, hence the direct application of their method would not apply.

Statistical methods for mapping genomic imprinting have also been developed. For a quantitative trait, Hanson et al. (2001) proposed a variance components framework that partitions the additive genetic effects into sex-specific components. The method requires the estimation of allele-specific

sharing probability among sibpairs in a family, hence, it does not allow for complete parental information to be missed. Shete and Zhou (2005) recently proposed a parametric approach that incorporates the information about sex-specific recombinations. By modifying the transmission/disequilibrium test (TDT), Hu et al. (2007) proposed a method that allows for testing imprinting for case-control data. However, the TDT approach requires genotyping both affected individuals and their parents, and is greatly limited when paternal genotype information is missing. Moreover, for large family-based association studies, the method is practically unfeasible, due to the costs associated with genotyping parental genomes.

So far, the dissection of the two aforementioned genetic conflicts that increase disease risk has been largely pursued as two separate endeavors. Cordell et al. (2004) proposed a method for modelling the parent-of-origin effect and the maternal-fetal interaction effect simultaneously. But the authors did not consider the genetic conflict effect between the maternal and fetal genome. No study has been proposed to date to model the two sets of genetic conflict effects simultaneously, in a unified framework. Moreover, most of the current approaches assume a case-parent-trio design in which the parental genotype information is allowed for partially missing, but not completely missing. Their applications are limited in practice when the paternal data are completely missing. According to the genetic conflict theory proposed by Haig (2004), both types of conflicts could trigger great negative effects resulting in great risk to either the mother or the fetus or both during pregnancy. Methods that address the conflict effects separately are greatly limited.

Addressing the limitations of the current methods, we propose a unified framework to model the two sets of genetic conflicts that increase disease risk during pregnancy after adjusting for the maternal and fetal main genetic effects and the effects of clinical risk factors. The identified conflict effects can help scientists explain phenomenon that can not be interpreted with traditional approaches. The results can also be incorporated into a predictive genetic test model to predict future disease risk (e.g., Lu and Elston 2008). Our main objective in this paper is to search for genetic factors conferring disease risk during pregnancy, which does not rule out the possibility of large effects from environmental factors, such as maternal diet. Thus, environmental effects are adjusted when searching for genetic factors. We assume a case-control, population-based study design with paired mother-fetus genotype data. Assuming a quadratic penalty function, a mixture model-based penalized logistic regression is developed to account for the missing information for double heterozygote maternal-fetal paired data when inferring genomic imprinting. A detailed estimation procedure in an expectation-maximization (EM) framework is developed. Details about tuning parameter selection using generalized

cross validation and variable selection are presented. Extensive simulations are conducted to evaluate the performance of the proposed method. To show the model's utility, we apply the method to a real data set, which is generated for the purpose of understanding the disease etiology of small for gestational age (SGA) during pregnancy. Genes that function in different conflict forms are detected in association with SGA.

2 Statistical methods

2.1 The genetic model

Consider a sample of n mother-fetus pairs collected from a population. By n mother-fetus pairs we mean there are total n mothers and n fetus with one to one relationship. A number of candidate genes are selected, based on prior knowledge of the disease, and a number of SNPs are then genotyped for each candidate gene. Genetic conflicts can cause disease to either the mother or the fetus. We first consider the case in which fetuses are the affected individuals, due to a genetic conflict. We further assume no paternal genotype information is available. Let $y_i = 1$ if the ith child has the disease, $y_i = 0$ otherwise. Note that the paired mother-fetus data only produce one binary response variable, and we assume a child is the affected one. The model can be extended to the case whereby a mother is the affected one. Let $\mathbf{y}$ represent an $n \times 1$ vector for the observed disease status. Assume that there are two alleles, B and b, at a marker locus with a population frequency of q and $1 - q$, respectively. Without loss of generality, we assume B is the minor disease allele. Let us denote the maternal genotype by $\mathcal{G}_{\mathrm{M}}$ and the offspring genotype by $\mathcal{G}_{\mathrm{O}}$. We use the subscript letters M and F to denote the maternal and paternal origins of an offspring allele, respectively. For example, an offspring's genotype $B_{\mathrm{M}}b_{\mathrm{F}}$ means that allele B is inherited from the mother and allele b is inherited from the father. For convenience, we denote the marker genotypes BB, Bb and bb by the numbers 2, 1 and 0 (i.e. the number of copies of the marker allele B carried by an individual). Table 1 lists all of the possible combinations of mother-fetus genotypes.

Note that the parental origin of an offspring allele can be explicitly identified in most cases, except for a case in which both the mother and the fetus carry heterozygote genotype Bb. Column 4 in Table 1 denotes the relative frequencies of all possible linkage phases in offspring, when considering the allelic parental origin. When the parent-of-origin linkage phase is explicit, the relative frequency is 1, as it is in most cases. For the double heterozygous maternal-fetal genotype pair, the chance that an offspring inherits the B allele from the mother is denoted by π_1 and, similarly, $\pi_2(= 1 - \pi_1)$ for an offspring

Table 1: The maternal-fetal genotype pairs and their mean genotypic values

$\mathcal{G}_{\mathrm{M}}$	$\mathcal{G}_{\mathrm{O}}$	$\mathcal{G}_{\mathrm{MO}}$	Frequency	Mean genotypic values
BB	$B_{\mathrm{M}}B_{\mathrm{F}}$	$(2,2)$	1	$\mu + a_m + a_o$
	$B_{\mathrm{M}}b_{\mathrm{F}}$	$(2,1)$	1	$\mu + a_m + d_o + i_c$
Bb	$B_{\mathrm{M}}B_{\mathrm{F}}$	(1,2)	1	$\mu + d_m + a_o + i_c$
	$B_{\mathrm{M}}b_{\mathrm{F}}$	$(1,1)$	π_1	$\mu + d_m + d_o$
	$b_{\mathrm{M}}B_{\mathrm{F}}$	$(1,1)$	π_2	$\mu + d_m + d_o + i_m$
	$b_{\mathrm{M}}b_{\mathrm{F}}$	(1,0)	1	$\mu + d_m - a_o + i_c$
bb	$b_{\mathrm{M}}B_{\mathrm{F}}$	$(0,1)$	1	$\mu - a_m + d_o + i_m + i_c$
	$b_{\mathrm{M}}b_{\mathrm{M}}$	$(0,0)$	1	$\mu - a_m - a_o$

inheriting the b allele from the mother. The probability of π_1 is generally unknown and needs to be estimated from the observed data.

Due to the unique environment that exists during pregnancy, the three sets of genes may have different levels of investments in fetal growth and fitness. The effect of the maternal genotype on its fetus, termed maternal effect, represents one contributing source. Following the traditional quantitative genetic theory (Lynch and Walsh 1998), the maternal effect is modelled by two genetic parameters, the additive effect (a_m) and the dominance effect (d_m). Similarly, the main effect of an offspring genotype on its own risk is also modelled by the additive effect (a_o) and the dominance effect (d_o). When a gene in the fetus' genome is imprinted due to different levels of self-interest from the maternal and paternal genes, the expressions for the two reciprocal genotypes $B_{\mathrm{M}}b_{\mathrm{F}}$ and $b_{\mathrm{M}}B_{\mathrm{F}}$ will be different depending on the underlying imprinting mechanism. Following Shete and Amos (2002), the genetic conflict effect caused by imprinting is modelled by the imprinting parameter i_m, which distinguishes the expression of two reciprocal heterozygotes $B_{\mathrm{M}}b_{\mathrm{F}}$ and $b_{\mathrm{M}}B_{\mathrm{F}}$. When $i_m = 0$, the expressions for the two reciprocal genotypes are the same and there is no imprinting.

Depending on the underlying mechanism of incompatibility, three sets of incompatible effects could exist due to: 1) an extra copy of a disease allele presented in the mother genome; 2) an extra copy of a disease allele presented in the fetus genome; or 3) combination of the two cases. For example, when a fetus is homozygous and its mother carries an allele that codes for an antigen, the fetus may produce an allogeneic response to the mother's antigen that is detrimental to the mother. The mother may experience a preterm premature rupture of membranes. In general, the first two sets of incompatibility effect are due to an extra allele copy in either the mother or the fetus' genome, and can

thus be termed as the allelic incompatibility effect. The third incompatibility occurs when any mismatch between the maternal and fetal genotype exists; this can be termed as the genotype incompatibility. Parimi et al. (2008) divided the MF conflict into six categories, and compared the performance of the different incompatibility models. They found that an incompatibility model similar to the aforementioned third model is the most powerful one. Following Parimi et al. (2008), we model the MF conflict attributable to the MFGI by the parameter i_c. An MFGI effect exists whenever there is a mismatch between the mother and the fetus's genotype, following the third incompatibility model in Parimi et al. (2008). A detailed list of the genotypic means for all possible maternal-fetal pairs is given in the last column in Table 1, with the overall mean denoted by μ.

2.2 Logistic regression and its limitations

Assuming unrelated case-control data sampled from a population, the logistic regression model is a natural choice for fitting the binary data with the form given by

$$\begin{aligned} \log \frac{p_i}{1-p_i} &= \mu + X_{i1}^* a_m + X_{i2}^* d_m + X_{i3}^* a_o + X_{i4}^* d_o + X_{i5}^* i_m + X_{i6}^* i_c + U_i' \boldsymbol{\gamma} \\ &= X_i^{*\prime} \boldsymbol{\beta} + U_i' \boldsymbol{\gamma} = X_i' \boldsymbol{\theta} \end{aligned} \tag{1}$$

where $p_i = \Pr(y_i = 1 | X_i^*, U_i)$; U_i is a $p \times 1$ vector of covariates, including clinical risk factors such as mother's age or smoking status; $\boldsymbol{\gamma}$ is a $p \times 1$ vector of the covariates effect; $X_i^* = (1, X_{i1}^*, X_{i2}^*, X_{i3}^*, X_{i4}^*, X_{i5}^*, X_{i6}^*)'$ is a 7×1 dummy vector; $\boldsymbol{\beta} = (\mu, a_m, d_m, a_o, d_o, i_m, i_c)'$ is the coefficients vector; and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$. Here we use the symbol $\prime$ to denote the matrix transpose. Specifically, the indicator variables X_{i1}^* and X_{i2}^* define the mother genotype status with

$$X_{i1}^* = \begin{cases} +1 & \text{for } BB \\ 0 & \text{for } Bb \\ -1 & \text{for } bb \end{cases}$$

and

$$X_{i2}^* = \begin{cases} 1 & \text{for } Bb \\ 0 & \text{for } BB \text{ or } bb \end{cases}$$

X_{i3}^* and X_{i4}^* are defined in a similar way for the fetus' genotype, and

$$X_{i5}^* = \begin{cases} 1 & \text{for } b_{\mathrm{M}} B_{\mathrm{F}} \\ 0 & \text{otherwise} \end{cases}$$

$$X^*_{i6} = \begin{cases} 1 & \text{for } BB - B_{\mathrm{M}}b_{\mathrm{F}},\ Bb - B_{\mathrm{M}}B_{\mathrm{F}}/b_{\mathrm{M}}b_{\mathrm{F}} \text{ or } bb - b_{\mathrm{M}}B_{\mathrm{F}} \text{ pair} \\ 0 & \text{otherwise} \end{cases}$$

Assuming independence and a known parent-specific linkage phase for the double heterozygous maternal-fetal genotype pair, the log-likelihood function for the joint maternal-fetal pairs can be expressed as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(y_i|X_i) = \sum_{i=1}^{n} \{y_i \log p_i + (1-y_i)\log(1-p_i)\} \qquad (2)$$

where $f_i(y_i|X_i) = \frac{exp(X_i'\boldsymbol{\theta} y_i)}{1+exp(-X_i'\boldsymbol{\theta})}$; $p_i = \Pr(y_i = 1|X_i) = [1 + exp(-X_i'\boldsymbol{\theta})]^{-1}$. The logistic regression coefficients preserve typical nice MLE properties when all the predictors are independent. The Wald or the likelihood ratio test can be applied to test the significance of an individual variable or sets of variables.

However, there are certain issues that arise when fitting the regular logistic regression in the current setting: 1) When the Hardy-Weinberg equilibrium (HWE) holds, the genotype frequency remains constant at each generation. For a polymorphic locus, we expect that the number of individuals from each generation who carry certain genotype will be proportional to that genotype frequency. This is true with large samples. With a finite sample size, however, certain genotype(s) may often dominate the other genotype(s) resulting in extremely unbalanced distributions among the three genotypes at a given marker locus. This may also happen when HWE does not hold for a tested polymorphic locus. When this does happen, multicollinearity may appear among the predictors, due to the dummy coding scheme of the variables. Consequently, the ordinary maximum likelihood method may lead to estimates with highly inflated variances, which make the ordinary inference procedure fails; and 2) In certain cases, a column of the design matrix $\mathbf{X}$ could contain many zeros or even all zeros, due to the dummy coding scheme (see Park and Hastie 2008 for a similar situation in modelling SNP interactions). For example, when few $B_{\mathrm{M}}b_{\mathrm{F}}$ genotypes are observed in the offspring, the column coding for the imprinting effect contains many zeros, resulting in an unreliable imprinting estimate i_m. This actually happens quite often in real data with a finite sample size so the regular logistic regression model may not fit well.

To avoid these problems, we introduce a penalized logistic regression (PLR), with L_2 regularization to penalize the regression coefficients. The L_2 regularization on the sum of the squares of the regression coefficients is known as ridge regression (Hoerl and Kennard 1970), and has been applied to logistic regression analysis (le Cessie and Houwelingen 1992). In the next section, we will introduce the PLR and estimation procedures.

2.3 Penalized logistic regression

Instead of maximizing the log-likelihood function in (2) directly, the PLR maximizes the log-likelihood function subject to a L_2-norm penalty on the regression coefficients (excluding the intercept) (le Cessie and Houwelingen 1992; Lee et al. 1988). This is equivalent to maximizing the following penalized log-likelihood function

$$\ell(\boldsymbol{\theta}, \lambda) = \ell(\boldsymbol{\theta}) - \frac{\lambda}{2} \sum_{j=1}^{p+6} \theta_j^2 \tag{3}$$

where $\ell(\boldsymbol{\theta})$ denotes the unrestricted log-likelihood function given by (2); θ_j $(j = 1, \cdots, p+6)$ are the regression coefficients, excluding the intercept μ; and λ is the tuning parameter, which is determined based on the data. When $\lambda \to 0$, (3) yields to the unrestricted maximum likelihood estimator, whereas if $\lambda \to \infty$, θ_j $(j = 1, \cdots, p+6)$ shrinks toward zeros. The logistic regression with L_2 regularization can handle multicollinearity problems efficiently and has attractive properties, as discussed in Park and Hastie (2008).

The estimation of the parameters can be done in an iterative way by modifying the Newton-Raphson algorithm. Following the iteratively re-weighted ridge regressions (IRRR) algorithm proposed by Park and Hastie (2008), at the $(t+1)$th step, we update the parameters $\boldsymbol{\theta}$ by

$$\begin{aligned} \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda})^{-\mathbf{1}} \mathbf{X}'\mathbf{W} \left\{ \mathbf{X}\boldsymbol{\theta}^{(\mathbf{t})} + \mathbf{W}^{-\mathbf{1}}(\mathbf{y} - \mathbf{p}) \right\} \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda})^{-\mathbf{1}} \mathbf{X}'\mathbf{W}\mathbf{z}. \end{aligned} \tag{4}$$

Where $\mathbf{X}$ is the $n \times (p+7)$ matrix of the predictors; $\mathbf{p}$ is the vector of probability estimates with $p_i = [1 + exp(-X_i'\boldsymbol{\theta})]^{-1}$; $\mathbf{W}$ is the diagonal matrix with the diagonal elements $p_i(1-p_i)$; $\boldsymbol{\Lambda}$ is the diagonal matrix with the diagonal elements $(0, \lambda, \cdots, \lambda)$; and $\mathbf{z} = \mathbf{X}\boldsymbol{\theta}^{(t)} + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p})$ is the "working" response in the IRRR algorithm (Park and Hastie 2008).

2.4 Penalized logistic regression with missing data

As shown in Table 1, when double heterozygous maternal-fetal pairs present, the allelic parental origin for fetus genotype can not be explicitly distinguished. The parental origin linkage phase in the fetus is missing. Consider the maternal-fetal genotype pair (Bb, Bb). A fetus can inherit a B allele from its mother or father with probability π_1 or π_2 $(= 1 - \pi_1)$, respectively. The missing probability can be estimated from the observed data. In this section,

we give details on how to fit the PLR model with missing data and derive an estimation procedure using the EM algorithm.

Let $\mathcal{G}_i$ be the observed genotype information for the ith maternal-fetal pair, and g_i be the genotype information with a known allelic parental origin in the fetus genotype corresponding to $\mathcal{G}_i$. Considering the fetus' allelic parent-of-origin information, the likelihood function can be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{g_i \in \mathcal{G}_i} \Pr(g_i|\mathcal{G}_i) f_i(y_i|X_i) \tag{5}$$

where $\Pr(g_i|\mathcal{G}_i)$ is the probability of a fetus carrying an allelic specific parent-of-origin genotype g_i that corresponds to the observed genotype $\mathcal{G}_i$, and $f_i(y_i|X_i)$ is the logistic regression function. As shown in Table 1, in most cases, g_i and $\mathcal{G}_i$ are one-to-one correspondence, hence, $\Pr(g_i|\mathcal{G}_i) = 1$. For the double heterozygous maternal-fetal pair (Bb, Bb), the fetus parent-of-origin genotype phase is missing, with $\Pr(g_i|\mathcal{G}_i) = \pi_1$ if a fetus carries genotype $B_{\mathrm{M}}b_{\mathrm{F}}$ and π_2 if $b_{\mathrm{M}}B_{\mathrm{F}}$.

Rewriting the likelihood function in (5), we obtain the observed likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n_1} f_i(y_i|X_i) \prod_{i=n_1+1}^{n} \left(\sum_{k=1}^{2} \pi_k f_{ik}(y_i|X_i) \right) \tag{6}$$

where n_1 is the number of total cell counts with a known parent-of-origin genotype phase in the fetus, and $n - n_1$ is the number of observed double heterozygous maternal-fetal pairs. Let $Z = (Z_{n_1+1}, Z_{n_1+2}, \cdots, Z_n)$ be a vector of random classification variables, with $Z_i = 1$ if the inherited B allele for the ith fetus is from mother and $Z_i = 2$ otherwise. Then, the complete likelihood can be written as

$$L^c(\boldsymbol{\theta}) = \prod_{i=1}^{n_1} f_i(y_i|X_i) \prod_{i=n_1+1}^{n} \prod_{k=1}^{2} \pi_k^{I(Z_i=k)} f_{ik}(y_i|X_i)^{I(Z_i=k)}$$

and

$$\begin{aligned} \frac{L^c(\boldsymbol{\theta})}{L(\boldsymbol{\theta})} &= \frac{\prod_{i=n_1+1}^{n} \prod_{k=1}^{2} \pi_k^{I(Z_i=k)} f_{ik}(y_i|X_i)^{I(Z_i=k)}}{\prod_{i=n_1+1}^{n} (\sum_{k=1}^{2} \pi_k f_{ik}(y_i|X_i))} \\ &= \prod_{i=n_1+1}^{n} \prod_{k=1}^{2} \left(\frac{\pi_k f_{ik}(y_i|X_i)}{\sum_{k=1}^{2} \pi_k f_{ik}(y_i|X_i)} \right)^{I(Z_i=k)} \end{aligned}$$

This leads to the posterior distribution of Z_i as

$$\Pi_{ik} = \Pr(Z_i = k|\mathbf{y}, \boldsymbol{\theta}) = \frac{\pi_k f_{ik}(y_i|X_i)}{\sum_{k=1}^{2} \pi_k f_{ik}(y_i|X_i)}, i = n_1 + 1, n_1 + 2, \cdots, n; k = 1, 2. \tag{7}$$

We can then get the expectation of the complete log-likelihood function with respect to the posterior distribution of Z as

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}) &= E_{(Z|\boldsymbol{\theta})}\left[\log L^c(\boldsymbol{\theta})\right] \\ &= E_{(Z|\boldsymbol{\theta})}\left[\sum_{i=1}^{n_1}\log f_i(y_i|X_i) + \sum_{i=n_1+1}^{n}\sum_{k=1}^{2} I(Z_i=k)\left(\log\pi_k + \log f_{ik}(y_i|X_i)\right)\right] \\ &= \sum_{i=1}^{n_1} y_i X_i'\boldsymbol{\theta} - \log(1+e^{X_i'\boldsymbol{\theta}}) + \sum_{i=n_1+1}^{n}\sum_{k=1}^{2}\Pi_{ik}(\log\pi_k + y_i X_{ik}'\boldsymbol{\theta} - \log(1+e^{X_{ik}'\boldsymbol{\theta}}))] \end{aligned} \tag{8}$$

In the M-step, we maximize function (8) under the L_2 constraint for the coefficients, except the intercept μ. That is

$$\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \quad \text{s.t.} \sum_{j=1}^{p+6}\theta_j^2 \le t.$$

which is equivalent to maximizing

$$\ell^*(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) - \frac{\lambda}{2}\sum_{j=1}^{p+6}\theta_j^2 \tag{9}$$

By a simple mathematical calculation, the first and second derivative of function $\ell^*(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ can be derived

$$\begin{aligned} \frac{\partial\ell^*(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= \sum_{i=1}^{n_1}(y_i-p_i)X_i' + \sum_{i=n_1+1}^{n}\sum_{k=1}^{2}\Pi_{ik}(y_i-p_{ik})X_{ik}' - \lambda\boldsymbol{\theta} \\ &= \mathbf{X}'\mathbf{V}(\mathbf{y}-\mathbf{p}) - \lambda\boldsymbol{\theta} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2\ell^*(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= -\sum_{i=1}^{n_1} p_i(1-p_i)X_i'X_i - \sum_{i=n_1+1}^{n}\sum_{k=1}^{2}\Pi_{ik}p_{ik}(1-p_{ik})X_{ik}'X_{ik} - \lambda I \\ &= -(\mathbf{X}'\mathbf{V}\mathbf{W}\mathbf{X} + \boldsymbol{\Lambda}) \end{aligned}$$

where $\mathbf{V}$ is an $N\times N$ ($N = n_1 + 2(n-n_1)$) diagonal matrix with diagonal elements $\{\mathbf{1}_{1\times n_1}, \boldsymbol{\Pi}_{1\times(n-n_1),1}, \boldsymbol{\Pi}_{1\times(n-n_1),2}\}$ where $\Pi_{ik} = P(Z_i = k|\mathbf{y},\boldsymbol{\theta})$, $\lambda\boldsymbol{\theta} = \{0, \lambda\boldsymbol{\theta}_1, \cdots, \lambda\boldsymbol{\theta}_{p+6}\}$ and $\lambda I = \text{diag}\{0,\lambda,\cdots,\lambda\}$. Applying the Newton-Raphson algorithm at the M-step, the iteration formula for $\boldsymbol{\theta}$ at the (t+1)th

step can be updated as

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} + (\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-\mathbf{1}}(\mathbf{X'V}(\mathbf{y}-\mathbf{p}) - \boldsymbol{\Lambda}\boldsymbol{\theta}^{(t)}) \\ &= \boldsymbol{\theta}^{(t)} + (\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-\mathbf{1}}\mathbf{X'VW}[\mathbf{X}\boldsymbol{\theta}^{(t)} + \mathbf{W}^{-\mathbf{1}}(\mathbf{y}-\mathbf{p})] \\ &= \boldsymbol{\theta}^{(t)} + (\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-\mathbf{1}}\mathbf{X'VWz}\end{aligned} \tag{10}$$

where $\mathbf{X}$ is an $N\times(p+7)$ matrix, and $\mathbf{y} = (y_1, \ldots, y_{n_1}, y_{n_1+1}, \ldots, y_n, y_{n_1+1}, \ldots, y_n)$ is an $N \times 1$ vector. $\mathbf{W}$ is an $N \times N$ diagonal matrix with diagonal elements $\{p_1(1-p_1), \ldots, p_{n_1}(1-p_{n_1}), p_{n_1+1,1}(1-p_{n_1+1,1}), \ldots, p_{n,1}(1-p_{n,1}), p_{n_1+1,2}(1-p_{n_1+1,2}), \ldots, p_{n,2}(1-p_{n,2})\}$. And $\boldsymbol{\Lambda}$ is a $(p+7)\times(p+7)$ diagonal matrix with the diagonal elements $\{0, \lambda, \ldots, \lambda\}$.

The EM algorithm with PLR is summarized as follows:

E-step: For given values of the initial parameters $\boldsymbol{\theta}$ and $\pi_1^{(0)} = 0.5$, and the tuning parameter λ, we calculate Π_{ik} given in (7).

M-step: At step $t+1$, we calculate $\pi_k^{(t+1)} = \frac{\sum_{i=n_1+1}^{n}\Pi_{ik}^{(t+1)}}{\sum_{i=n_1+1}^{n}\sum_{k=1}^{2}\Pi_{ik}^{(t+1)}}$, $\mathbf{V}^{(t+1)}$, $\mathbf{W}^{(t+1)}$, and $\mathbf{z}^{(t+1)}$, updating $\boldsymbol{\theta}$ by (10). Repeat the EM steps until convergence.

The variance of the estimator $\hat{\boldsymbol{\theta}}$ can be calculated by

$$\begin{aligned}Var(\hat{\boldsymbol{\theta}}) &= Var[(\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-\mathbf{1}}\mathbf{X'VWz}] \\ &= (\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-\mathbf{1}}\mathbf{X'V}Var[\mathbf{y}-\mathbf{p}](\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-\mathbf{1}}\mathbf{X'V} \\ &= (\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-1}\mathbf{X'VWVX'}(\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-1}\end{aligned}$$

which is the sandwich estimate defined by Gray (1992). Following Hastie and Tibshirani (1990) and Park and Hastie (2008), the effective degree of freedom for the PLR regression with missing data can be approximated by

$$df(\lambda) = tr((\mathbf{X'VWX} + \boldsymbol{\Lambda})^{-1}\mathbf{X'VWX}) \tag{11}$$

2.5 Choosing tuning parameter λ

The quadratic regularization offers the advantages emphasized in previous sections. In addition, it can be used to smooth a model and control the effective degree of freedom (Park and Hastie 2008). The algorithm described above is for a fixed λ value. In a real analysis, the tuning parameter needs to be selected. Ordinary cross validation (OCV) such as leave-one-out cross validation or k-fold cross validation can be used for selecting λ. However, OCV is computationally intensive, since for each value of λ, we have to fit the model and estimate the parameters using the EM algorithm. A computationally efficient alternative to choose λ is to use the generalized cross validation (GCV), as it only requires fitting the model once and can greatly reduce the computational

burden. Here, we adopt a non-linear GCV approach proposed by Fu (2005). The non-linear GCV requires fitting the model once, and it takes the form

$$\text{GCV} = \frac{\text{Dev}}{n(1 - vs/n)^2} \tag{12}$$

where Dev$= 2\ell(\mathbf{y}, \mathbf{y}) - 2\ell(\mathbf{y}, \mathbf{p})$ is the model deviance (McCullagh and Nelder 1989); $v(= p + 6)$ is the number of covariates in the model excluding the intercept; and s is the shrinkage rate, which is defined as

$$s = \frac{\|\hat{\boldsymbol{\theta}}(\lambda)\|^2}{\|\hat{\boldsymbol{\theta}}\|^2} \tag{13}$$

where $\hat{\boldsymbol{\theta}}(\lambda)$ are the coefficient estimates with the L_2 regularization, and $\hat{\boldsymbol{\theta}}$ are the ordinary coefficient estimates without constraint. The shrinkage rate s defines the degree of shrinkage, with 0 for complete shrinkage and 1 for no shrinkage (Tibshirani 1996). A full model fitted with all of the parameters is used to choose the tuning parameter λ, which is then fixed for subsequent analysis.

2.6 Variable selection

Unlike LASSO (Tibshirani 1996), or more recently, the adaptive LASSO (Zou 2006), which do shrinkage and variable selection simultaneously, the L_2 penalty only does shrinkage but not variable selection. In fact, none of the coefficients shrink to zero unless the distribution of the variable is extremely sparse (Park and Hastie 2008). Regular hypothesis testing procedures, such as the likelihood ratio test or the score test, do not work under the penalized framework due to the difficulty of deriving the null distribution of the test statistic. For ease of interpretability, we need a method for variable selection. The classic forward selection procedure is used to accomplish this.

The forward selection starts with a null model, in which only the intercept term is included. Then, for each forward step, factors are added to the model one at a time. The choice of the factor to be added in each step is based on the AIC/BIC criteria, where AIC is defined by AIC=Dev+2df and BIC is defined by BIC=Dev+log(n)df. The model with the smallest AIC/BIC value is chosen as the final model. When at least one genetic parameter (excluding the intercept) is included in the final model, we declare that the SNP is associated with the disease under study.

Tuning parameter λ is selected by fitting the full model which includes all the factors. It is then fixed for subsequent variable selection. It is worth noting that this may not be the optimal regularization. To efficiently select variables,

we may need to search for the optimal regularization parameter λ at each selection step. For example, one may start with a tuning parameter λ chosen based on the null model, add one more variable to the model, then search for a better tuning parameter and continue. This, however, will incur extremely high computational costs. In the meantime, new problems may be introduced. For example, if we select λ at each step, it may vary at each selection step, resulting in difficulty in comparing the AIC/BIC values for model comparison. We leave this as an open issue for future investigation. A similar issue is also discussed in Zhu and Hastie (2004).

3 Simulation

Extensive simulation studies were conducted to evaluate the model performance and its statistical behavior under different sample sizes and different gene action modes. At an SNP locus, we assumed the same population allele frequency for both a mother and a father's minor allele. Assuming HWE, the genotype frequency can be calculated from the allele frequency for both parents. Using multinomial distribution, the genotype information for both parents can then be simulated. Assuming random mating in the simulated population, the offspring genotype can easily be simulated from the parents' genotype data. For example, if an individual's mother has genotype BB and father has genotype Bb, then the individual's genotype has a 50% chance of being BB and a 50% chance of being Bb. Individual disease status was then simulated from a bernoulli distribution, with the probability of being affected defined by the parameters given in Table 1. For simplicity, only genetic factors were simulated, even though non-genetic factors such as age or weight can also be simulated with specified parameter values. Once the phenotype data were simulated, paternal genotype data were discarded and only maternal-fetal pairs were recorded for further analysis.

Simulations were conducted assuming a case-control sample size (denoted by n) of 500 and 1000. The minor disease allele frequency (denoted by q) was assumed to be 0.1, 0.3, and 0.5. The given values for all of the genetic parameters were listed in Table 2. Scenario S0 assumes no genetic effect at all. Other scenarios assume different gene actions. For example, a disease phenotype may result from the function of only the main effects of maternal genes (S1) or fetal genes (S2), or may be due to the function of the conflict effects only (S5 and S6). The simulated samples contain roughly 50% cases and 50% controls. Data simulated with these configurations were subject to analysis with the proposed logistic regression with L_2 regularization and the ordinary logistic regression. Results from 1000 Monte Carlo repetitions were recorded.

Table 2: List of parameter values under different simulation designs

Scenario	a_m	d_m	a_o	d_o	i_m	i_c
S0	0	0	0	0	0	0
S1	0.8	0.8	0	0	0	0
S2	0	0	0.8	0.8	0	0
S3	0.8	0	0.8	0	0	0
S4	0	0	0.8	0	0.8	0
S5	0	0	0	0	0.8	0
S6	0	0	0	0	0	0.8

In all of the simulations, the AIC criterion performs consistently poorer than the BIC, hence the results from the AIC were not reported. Figure 1 shows the results for variable selection under different simulation scenarios. The top figure corresponds to scenario S0, in which the proportion of selection is equivalent to the false positive (or selection) rate. For each parameter, the red, green and blue bars correspond to the selection results with an allele frequency of 0.1, 0.3, and 0.5, respectively. The two bars with the same color under each allele frequency represent the selection results for a sample size of 500 and 1000. It is clear that the false selection rate for the parameters under different sample sizes and minor allele frequencies are all under the nominal level of 0.05, indicating a good false positive control of the proposed method. Figure 1 also shows the power (true positive) and the false selection rate for scenarios S1 to S6. As we expected, the selection power increases as the sample size and the minor allele frequency increase. The selection rates for true negatives are also under reasonable control. It is worth noting that the incompatibility effect (Fig. 1-S6) has a higher selection power, compared to the imprinting effect (Fig. 1-S5), given that both parameters have the same effect size. This might be due to the uncertainty of inferring the imprinting effect when applying the EM algorithm.

To compare the performance between the proposed PLR method with the ordinary logistic regression (OLR), we also analyzed the simulated data with the OLR method. Likelihood ratio test was applied to test significance for each parameter when the OLR was applied. Without loss of generality, comparisons were made only under the minor allele frequency of 0.3 and the sample size of 500. Figure 2 shows the power plot under scenarios S0-S6 with parameter values given in Table 2. It is clear that the OLR method has a relatively low power to detect true positives, especially for the maternal additive and dominance effects, the offspring dominance effect and the two genetic conflict effects. The OLR method gives constantly under-estimated false pos-

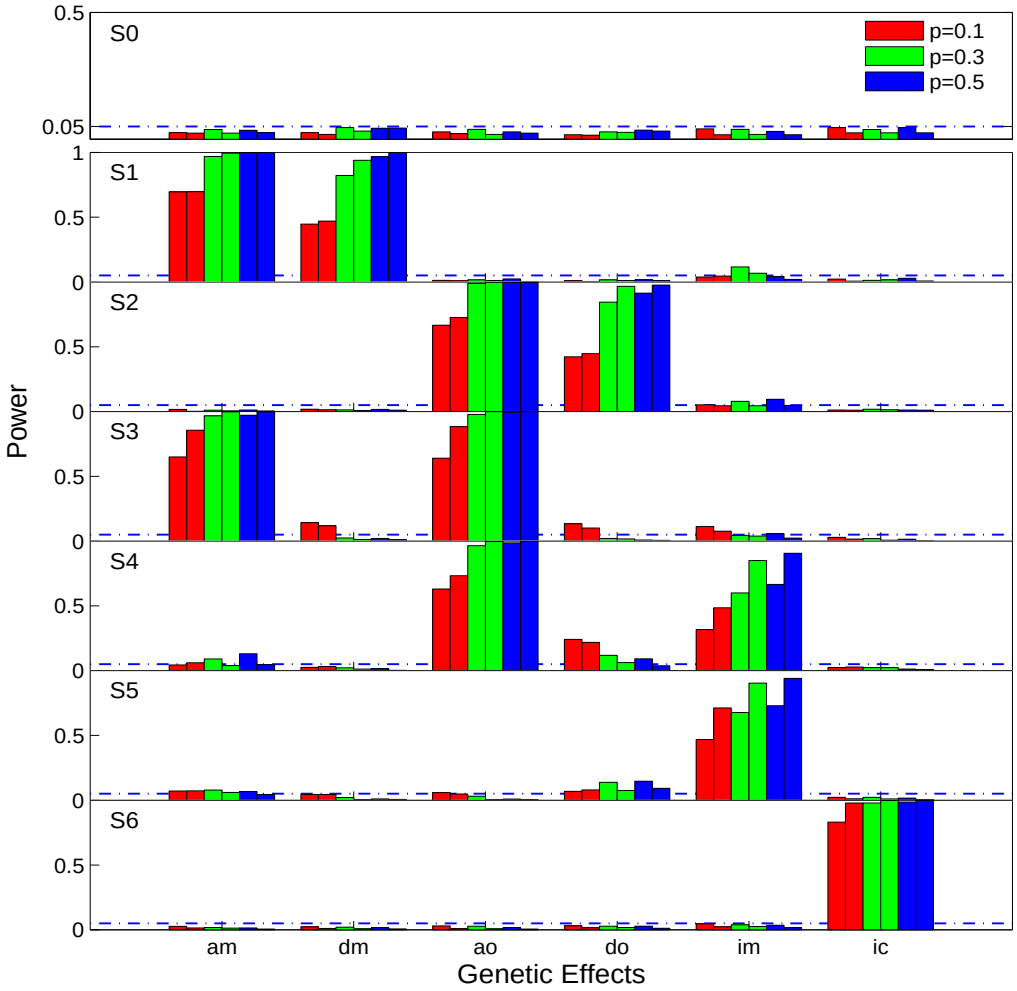

Figure 1: The plot of variable selection results under different simulation scenarios (Parameter values are listed in Table 2). There are three sets of colored bars associated with each parameter. The red, green and blue bars correspond to selection results with minor allele frequencies 0.1, 0.3 and 0.5, respectively. For each allele frequency, the left and right bars correspond to the selection results with the sample size 500 and 1000, respectively. The horizontal dash-dotted line indicates the nominal level of 0.05.

itive rates for parameters d_m and i_c for the seven simulation scenarios. The results indicate that the OLR method is less competitive than the proposed PLR method.

As suggested by one reviewer, we did additional simulations to check the model performance under population stratification and asymmetric mating. We followed the simulation design given in Sinsheimer et al. (2003).

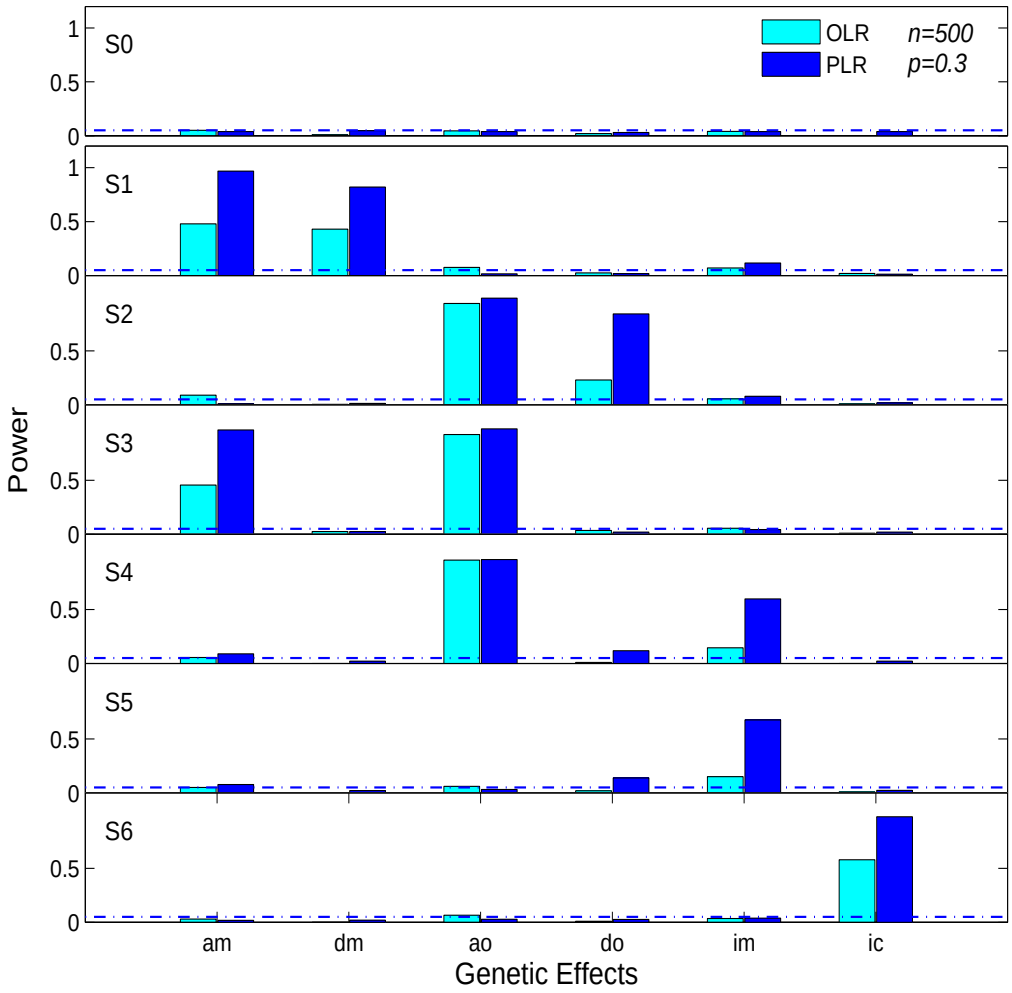

Figure 2: The comparison of power plot with the ordinary logistic regression (OLR) and the penalized logistic regression (PLR) under different simulation scenarios. Each plot corresponds to one simulation scenario listed in Table 2. The horizontal dash-dotted line indicates the nominal level of 0.05.

To evaluating the effect of population stratification, we simulated 1000 data sets of 400 child-parent triads, with parental genotype frequencies given by $P(BB) = 0.14$, $P(Bb) = 0.49$, and $P(bb) = 0.37$, and an additional 100 child-parent triads with parental genotype frequencies given by $P(BB) = 0.01$, $P(Bb) = 0.06$, and $P(bb) = 0.93$. We allowed the baselines to differ in these two populations. The cases and controls were sampled with a roughly 1:1 ratio in each subpopulation. The genotype frequency indicates deviation from Hardy-Winberg equilibrium (HWE) at the simulated locus. We assumed an

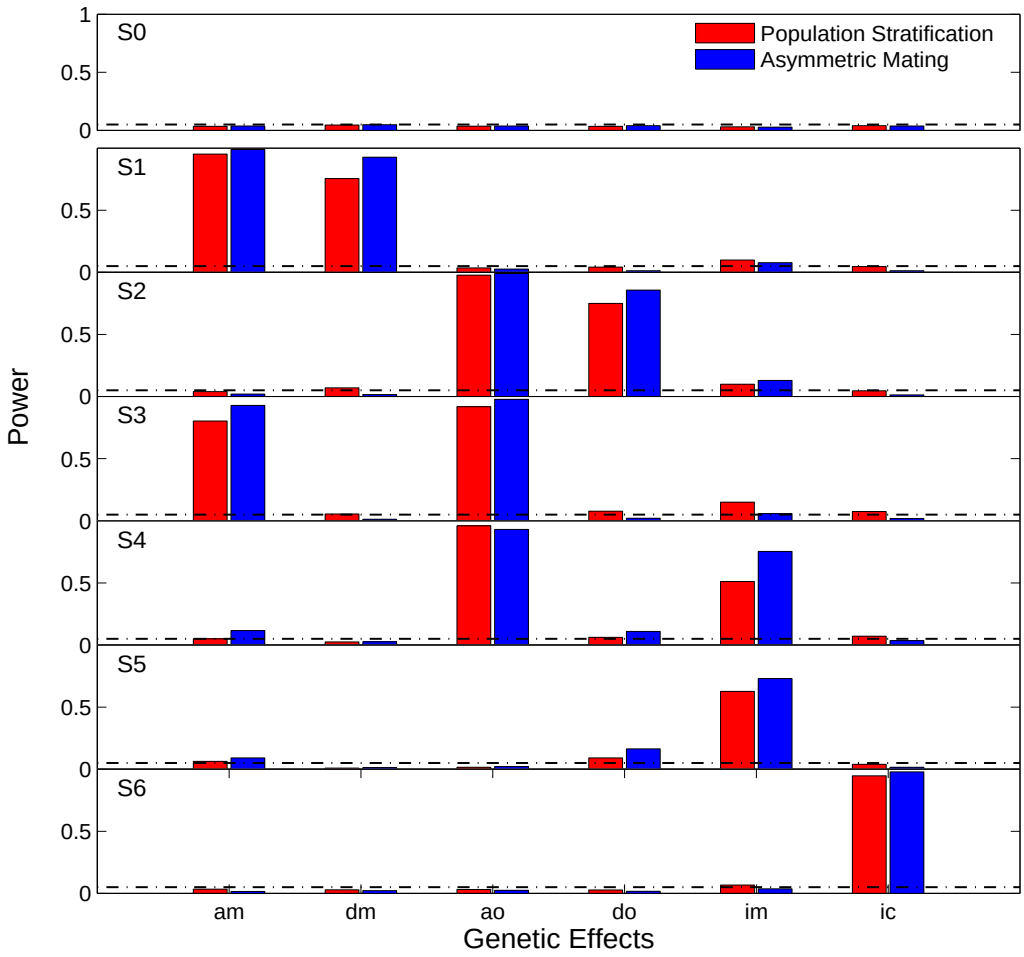

Figure 3: The power plot for variable selection under population stratification and asymmetric mating for different simulation scenarios. Each plot corresponds to one simulation scenario listed in Table 2. The horizontal dash-dotted line indicates the nominal level of 0.05.

equal effect size for each parameter in the two populations. For simplicity, the simulation was done with sample size $n = 500$. A list of parameter values used for the simulation study is given in Table 2. The red bars in Figure 3 show the power plots under population stratification for different simulation scenarios. It is observed that the true and false positives are under reasonable control with the proposed method under different simulation scenarios, indicating the method robustness to population stratification.

To evaluating the effect of asymmetric mating on variable selection, we simulated 1000 data sets — each composed of two populations of child-parent

triads — allowing for asymmetric mating between the males and females in the two populations, following the simulation design given in Sinsheimer et al. (2003). For each data set, 400 child-parent triads were simulated with both parental genotype frequencies given by $P(BB) = 0.14$, $P(Bb) = 0.49$, and $P(bb) = 0.37$. An additional sample of 100 child-parent triads were simulated with paternal genotype frequencies given by $P(BB) = 0.01$, $P(B = b) = 0.06$, $P(bb) = 0.93$, and maternal genotype frequencies given by $P(BB) = 0.14$, $P(Bb) = 0.49$, and $P(bb) = 0.37$. Again, the cases and controls were sampled with a roughly 1:1 ratio in each subpopulation. Simulation results are shown in Figure 3 (indicated by blue bars). The results indicate that the method performs reasonably good, even though there were slightly inflated false positives observed for the maternal additive (S4 and S5), offspring dominance effect (S4 and S5) and imprinting effect (S1 and S2), under the asymmetric mating design. Sinsheimer et al. (2003) previously showed that asymmetry mating leads to bias in the maternal allelic effects in their approach. Since ridge regression gives biased parameter estimations, it is meaningless to report bias estimation in this study. Thus it is hard to compare the two methods in this regard.

In summary, the proposed method performs reasonably well under different scenarios. When there is no gene effect, the false positive rate for each parameter is reasonably controlled. We also observed a reasonable selection power for those significant parameters, in which the power varies depending on the size of sample and minor allele frequency. In addition, the proposed method is reasonably robust to population stratification and asymmetric mating, even though it appears that population stratification has a relatively larger negative effect than the asymmetric mating does, on the selection power of the imprinting effect.

4 A case study

Our method is applied to a genetic association study of small for gestational age (SGA) neonates. Infants whose weight falls below the 10th percentile for gestational age are classified as SGA (Cardosi 2006). Experimental evidence has shown that SGA infants have an elevated risk of developing metabolic disease — particularly obesity, insulin resistance, carbohydrate intolerance and dyslipidemia — and confer a substantial risk of morbidity and mortality both in the perinatal period and later in life (reviewed in Saenger et al. 2007). For example, SGA may lead to complications for newborns, including respiratory complications, hypotension, hypoglycemia, necrotizing enterocolitis, and neonatal death (Bernstein et al. 2000; Villar et al. 1990). Consequently, children born with SGA are prone to neurological impairment and delayed

cognitive development (Paz et al. 1995; Taylor and Howie 1989).

An increasing number of SGA infants has been reported in recent years, but the etiology of SGA remains largely unknown. It has been commonly recognized that a complication of pregnancy and/or delivery is a complex trait determined by multiple environmental and genetic factors (Hao et al. 2004). The importance of genetic factors in the regulation of fetal growth has been recognized and increasing evidence has been documented in the literature (Reviewed in Saenger et al. 2007). The genetic conflict theory proposed by Haig (2004) offers an alternative explanation of the disease risk associated with SGA. However, no genetic association study has been reported regarding the genetic conflict effects that increase SGA risk.

To understand the genetic basis of SGA, particularly any genetic conflict effects that increase the infant SGA rate, a number of candidate genes were selected from pathways, including immune response and angiogenesis. Subjects were recruited through the Department of Obstetrics and Gynecology at Sotero del Rio Hospital in Puente Alto, Chile. SNPs were selected for genotyping in order to capture at least 90% of the haplotypic diversity of each gene. A total of 488 SNPs from 164 genes were analyzed in 340 SGA mother-offspring pairs and in 585 control mother-offspring pairs after eliminating SNPs that show deviation from HWE in the controls, SNPs with a minor allele frequency of less than 0.05, and families with obvious relationship errors and genotyping errors in the SNPs. No evidence of population stratification was revealed using the genomic control method (Devlin and Roeder 1999), since the estimated inflation factor was near one (i.e., no inflation). A number of clinical risk factors potentially associated with SGA risk were also measured. These include maternal age, maternal weight and height at birth, maternal hypertension, allergy and asthma condition, maternal BMI index, number of preterm deliveries, neonatal height, head size and gender. A logistic regression was first conducted for all of the clinical covariates, to select the one(s) that should be included when estimating the effects of the genetic parameters. Among the group of analyzed covariates, only maternal weight was significant at the 5% level. Considering that neonatal gender might be associated with the risk of SGA, it was also included in the analysis. Thus, two covariates were included in the model together with other genetic factors with the aim of assessing which genetic risk factors contribute to the risk of SGA, after adjusting for the effects of the two environmental factors.

A full list of selected SNPs with their genetic effects is shown in the supplemental table. For each SNP, the first row corresponds to the selected features with estimated effects, where "-" sign indicates the feature(s) is(are) not significant. In addition to the estimated values, there are additional two rows, with values given in parenthesis. Details about these two rows are explained

Table 3: Partial list of genes and SNPs showing significant association with SGA. Non-significant effects are indicated by "-" sign. The permutation and bootstrap scores (given in the parenthesis) are listed as italic and regular fonts, respectively.

Gene	rs Number	Location	Significant effects							
symbol			a_m	d_m	a_o	d_o	i_m	i_c	BS	MW
FGF4	634043245	Exon 3	-	-	-	0.361	-	-	-	-
			(0.5)	*(1.5)*	*(1)*	*(0.5)*	*(3)*	*(1)*	*(0.5)*	*(1.5)*
			(6)	(0)	(7.5)	(49.5)	(8)	(2.5)	(1)	(0)
	634043464	Down-	0.748	-	-	-	-	-	-	-
		stream	*(1.5)*	*(0.5)*	*(2)*	*(0.5)*	*(2)*	*(1)*	*(2)*	*(2.5)*
			(66)	(9)	(1)	(4)	(11.5)	(16.5)	(0)	(0)
LPL	22220155	Intron 6a	-	-	-	-	-	0.363	-	-
			(0.5)	*(4.5)*	*(0.5)*	*(3)*	*(0.5)*	*(2.5)*	*(2.5)*	*(2.5)*
			(6.5)	(16.5)	(2)	(3)	(3)	(68.5)	(3.5)	(3)
	612980414	Intron 8	-	-	-	-	-	0.456	-	-
			(2)	*(0.5)*	*(1)*	*(0.5)*	*(0.5)*	*(3.5)*	*(0.5)*	*(0.5)*
			(2.5)	(3)	(10)	(6.5)	(1)	(42)	(1.5)	(3)
PON1	20209376	intron 5	-	-	-	-	-1.223	-	-	-
			(0)	*(1)*	*(1)*	*(1)*	*(5)*	*(1.5)*	*(1)*	*(1)*
			(0)	(0.5)	(5.5)	(21.5)	(83.5)	(2.5)	(1)	(0.5)
IL2RA	23884895	Intron 5	-	-	-	-	-	0.519	-	-
			(1.5)	*(0.5)*	*(1)*	*(0.5)*	*(3)*	*(1)*	*(0)*	*(0)*
			(3.5)	(4.5)	(5.5)	(1.5)	(4)	(76)	(1.5)	(2)

BS refers to baby sex; MW refers to maternal weight.

at the end of this section. Note that the ridge estimator is a biased estimate. The selected feature only indicates its relative importance in association with a disease trait. An actual inference about the feature can be done by simply fitting those selected features into an ordinary logistic regression model. The estimated values listed in the Tables are refitted values.

A partial list of significant SNPs is given in Table 3 to demonstrate the model implementation. Gene FGF4 (fibroblast growth factor 4), located on chromosome 11, is one of the genes being selected. The gene is essential for mammalian embryogenesis and fetal growth (Lamb and Rizzino 1998). We detected a fetal dominance effect at Exon 3 and an additive maternal effect at downstream position. This gene is regulated by a powerful downstream enhancer (Lamb and Rizzino 1998). Thus, the significant maternal SNP located

at the downstream position might play a regulation role in LD with an SNP that plays a regulation role. No conflict effects were detected for this gene.

Two SNPs in gene LPL (lipoprotein lipase) showed a pure incompatibility effect. Studies have shown that LPL is associated with intrauterine growth restriction (IUGR), which reduces the placental supply of nutrients to the fetus and prevents the fetus from achieving its growth potential (Saenger et al. 2007; Tabano et al. 2006). The definition of IUGR has a great deal of overlap with SGA (Wollmann 1998), and both share similar genetic risk factors. Gene IL2RA also shows an incompatibility effect. This gene was previously reported in Parimi et al. (2008) to have a significant incompatibility effect increasing the risk of PE in mother. Thus the two diseases might share a common risk factor, and the result might indicate the heterogeneity of the gene function. One SNP located at gene PON1 (Paraoxonase-1) showed an imprinting effect. PON1 belongs to a family of at least three genes, including PON2 and PON3, all of which map to human chromosome 7q21.3 and are imprinted (Morison et al. 2005).

In the real data analysis, the λ value was chosen as a gradient of 2^{ω}, $\omega = -10 : 10$. To show the performance of GCV criterion in selecting tuning parameter λ, we used gene IL2RA as an example. Fig. 4 plots the selection results for that gene. A minimum GCV value was reached when λ takes value 8. We adopted an adaptive selection procedure to choose λ — that is, when increasing the value of λ caused an increase in GCV value, the selection was automatically terminated.

Note that the variable selection procedure does not give a p-value for a selected feature. If the data were perturbed, a different set of features might be selected. To assess whether the selected feature(s) was(were) false identification(s), we applied a permutation analysis similar to the one used by Wang et al. (Wang et al. 2007). We permuted the data by randomly reshuffling the relationship between the disease status and the genetic markers (keep the maternal-fetal paired relationship), and applied the proposed method to the permuted data sets. 200 permutation runs were repeated. Each permutated data set represents a random sample generated from the null distribution. The results for each SNP are tabulated in Table 3. The values given in parenthesis and shown in italic font correspond to the percentage being selected. Among the 200 runs, the selection rate for all of the factors is less than or close to 5%, indicating that our method indeed selects the relevant genetic effects, with few false positives.

Alternatively, one can apply a bootstrap assessment (Efron and Tibshirani 1993). The bootstrap analysis provides a measure of how likely the features are to be selected (Park and Hastie 2008). To do so, we randomly bootstrapped 200 samples and ran the variable selection with a fixed λ value selected based

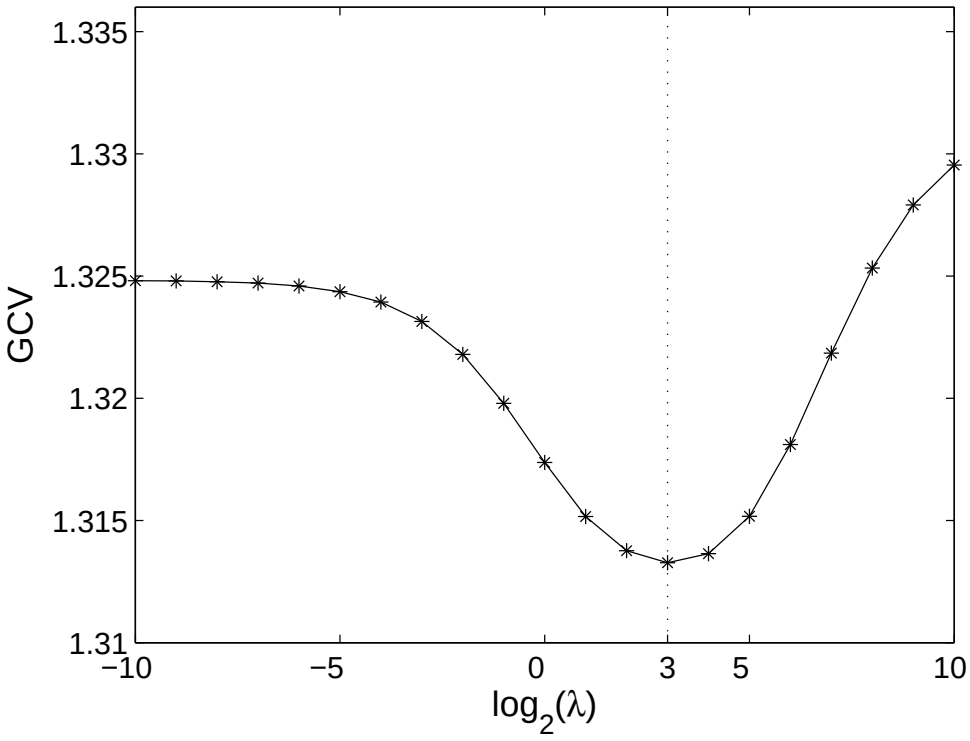

Figure 4: The GCV values versus the tuning parameter λ for gene IL2RA.

on the original data. The frequency for each feature included in the model was counted with the percentage listed in parenthesis and shown in a regular font (Table 3). Clearly, those features selected based on the original data had a relatively high chance of being selected in the bootstrap analysis, indicating their importance. For example, the incompatibility effect for gene IL2RA had a 79.5% chance being selected in the bootstrapped samples, whereas the chance of selecting other factors was smaller than this number. However, some features showed a relatively high selection frequency in the bootstrap analysis, but were not selected based on the original data. For example, the fetal dominance effect in gene PON1 had a 21.5% chance of being selected in the bootstrap analysis, indicating less important for this feature. It is also possible that the dominance effect is highly correlated with the imprinting effect, such an occurrence can be detected by calculating the co-occurrence matrix, as proposed by Park and Hastie (2008).

5 Discussion

Pregnancy has traditionally been viewed as a harmonious cooperation between the mother and the fetus (Haig 2004). This view, however, fails to recognize aspects of potential genetic conflicts among the three sets of genes — the maternal gene, the PDFA and the MDFA in the fetus' genome. Due to the unique environment that exists during pregnancy, the maternal genotype unlike its paternal counterpart, has the opportunity to influence fetal development, through the mediation of the altered uterine environment. The fetus' genes also play pivotal roles in regulating and controlling its own growth. Thus, the normal function of fetal growth relies on the coordinated function of these three sets of genes. Any unusual function caused by different levels of investment among these three sets of genes may result in a functional abnormality in fetal growth, and may subsequently lead to disease, such as the small sized or under-weight infants defined as SGA neonates. Therefore, a "disease" developed during pregnancy could have unique genetic bases.

Quite often the two genetic conflict effects, the genomic imprinting and the maternal-fetal genotype incompatibility, are ignored when searching for disease factors in an association study. The existence of such effects, however, may lead to incorrect interpretations of the (marginal) effects of particular genes when performing a human genetic association study at the individual level. In this article, we make an attempt to model and test those genetic conflicts that can arise during pregnancy to increase disease risk. Our model considers both maternal and fetal main effects, as well as genetic conflict effects, when dissecting disease gene association, and is a unified framework. Simulation studies under different scenarios show that the model has reasonable power and good false positive control.

We applied our model to a real data set for SGA neonates. The results indicate that there are a relatively large proportion of genetic risk factors that confer conflict effects (see the supplemental table). The selected genes listed in Table 3 are further evaluated with both permutation and bootstrap analyses. The results confirm that it is highly possible that the selected factors may trigger effects in increasing SGA risk. Increasing evidence has shown that the majority of imprinted genes in mammals play important roles in controlling embryonic growth and development, and some are involved in post-natal development (Isles and Holland 2005; Tycko and Morison 2002). Imprinted genes also play a role in the regulation of placental blood vessel development and in controlling nutrient transport. Thus, they may also indirectly control fetal growth and development (Constancia et al. 2004). This explains in part why a substantial number of imprinting effects are identified in this study. Among the SNPs showing imprinting effects, the majority show maternal imprinting

(or paternal expression). Some of this phenomenon can be explained by Haig's genetic conflict theory, in which the paternal copy always favors fetal growth (Haig 2004). Empirical study also indicates that fetal genes — especially those expressed by the father — may have a substantial influence on fetal growth when the maternal copy is restrained (Dunger et al. 2006).

In this study, the MF genetic conflict is modelled in a composite way whereby a genetic conflict exists whenever there is a mismatch between the maternal and the fetal genotype. Parimi et al. (2008) previously broke down the MF conflict into six different categories, each one corresponding to a unique conflict mechanism. Their simulation study indicates that the one that was coded in a composite way, as we did in the current work, is the most powerful model. However, when an MF conflict is detected, it is worth distinguishing which model is the optimal one. For example, what if the MF conflict is due to gestational drive (the mother has an allele that the fetus does not) or symmetric incompatibility (the fetus has any allele that the mother does not) (Parimi et al. 2008). This can be determined using a statistical model selection method — fitting different conflict models assuming different MF conflict mechanisms. The AIC or BIC type selection method can be applied for this purpose.

Our model is developed under the L_2 regularized regression considering binary disease phenotype. The advantage of this ridge-type regression is discussed in section 2, particularly for the multicolinearity problem. There has been a great interest shown in the statistical literature in developing efficient variable selection method. Most methods are developed under the L_1 regularization framework — for example, the LASSO (Tibshirani 1996) and the adaptive LASSO (Zou 2006). The advantage of the L_1 regularization method is that it can handle a large number of variables, the so-called large p small n problem. Moreover, these algorithms can do variable estimation and selection simultaneously, and are computationally favorable when p is large. However, the LASSO-type algorithm randomly picks just one variable in a cluster of highly correlated variables and disregards the remaining ones (Zou and Hastie 2005). Thus, the results might not be biologically justifiable. For example, when there is imprinting or incompatibility effect which might be correlated with the maternal or the fetal main effect, the LASSO algorithm may end up choosing the main effect and lead to wrong inference. Moreover, Tibshirani (1996) pointed out that for regular $n > p$ situations, the prediction performance of LASSO is dominated by the ridge regression if there are high correlations between the predictors.

The model developed in this article allows one to test the effects of maternal and fetal main effects, as well as the genetic conflicts that increase fetal disease risk. The incorporation of the maternal effect in searching for

an association signal in a case-control design can avoid misinterpretation of an association signal, as shown by Buyske (2008). It is worth noting that the symmetric design also allows one to test these effects on maternal behavior or any pregnancy-related diseases in the mother, such as PE. As shown in the simulation studies, the model is quite robust to population stratification and asymmetric mating (Fig. 3). Simulations also indicate the relative merit of the proposed penalized method against the regular logistic regression (Fig. 2). With increasing computing power, the method is feasible for large-scale candidate gene or genomewide association studies, especially as the genotyping cost is rapidly decreasing. A computer program written in R can be downloaded at the author's website at http://www.stt.msu.edu/∼cui.

Acknowledgement

This work was supported in part by NSF grant DMS-0707031 and by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, NIH, DHHS.

References

Bernstein IM, Horbar JD, Badger GJ, Ohlsson A, Golan A. (2000) Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. The Vermont Oxford Network. *American Journal of Obstetrics and Gynecology* **182**: 198-206.

Buyske, S. (2008) Maternal genotype effects can alias case genotype effects in casecontrol studies. *European Journal of Human Genetics* **16**: 783-785.

Constancia M, Kelsey G, Reik W. (2004) Resourceful imprinting. *Nature* **432**: 53-57.

Devlin B, Roeder K. (1999) Genomic control for association studies. *Biometrics* **55**: 997-1004.

Dunger DB, Petry CJ, Ong KK. (2006) Genetic variations and normal fetal growth. *Hormone Research* **65**: 34-40.

Efron B, Tibshirani R. (1993) *An Introduction to the Bootstrap*. Chapman and Hall: Boca Raton.

Fu WJ. (2005) Nonlinear GCV and quasi-GCV for shrinkage models. *J. Stat. Plan. Infer.* **131**: 333-347.

Gardosi J. (2006) New definition of small for gestational age based on fetal growth potential. *Hormone Research* **65**: 15-18.

Gray R. (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of American Statistical Association* **87**: 942-951.

Haig D. (1993) Genetic conflicts in human pregnancy. *Quarterly Review of Biology* **68**: 495-532.

Haig D. (2004) Evolutionary conflicts in pregnancy and calcium metabolism - A review. *Placenta* **25** Supplement A; Trophoblast Research, Vol. 18: S10-S15.

Hanson RL, Kobes S, Lindsay RS, Kmowler WC. (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *American Journal of Human Genetics* **68**: 951-962.

Hao K, Wang X, Niu T, Xu X, et al. (2004) A candidate gene association study on preterm delivery: application of high-throughput genotyping technology and advanced statistical methods. *Human Molecular Genetics* **13**: 683-691.

Hastie T, Tibshirani R. (1990) *Generalized Additive Models* Chapman and Hall: London.

Hoerl AE, Kennard RW. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**: 55-67.

Hsieh H-J, Palmer CGS, Harney S, Newton JL, Wordsworth P, Brown MA, Sinsheimer JS. (2006a) The v-MFG test: Investigating maternal, offspring, and maternal -fetal genetic incompatibility effects on disease and viability. *Genetic Epidemiology* **30**: 333-347.

Hsieh H-J, Palmer CGS, Sinsheimer JS. (2006b) Allowing for missing data at highly polymorphic genes when testing for maternal, offspring and maternal-fetal genotype incompatibility effects. *Human Heredity* **62**: 165-174.

Hu YQ, Zhou JY, Fung WK. (2007) An extension of the transmission disequilibrium test incorporating imprinting. *Genetics* **175**: 1489-1504.

Isles AR, Holland AJ. (2005) Imprinted genes and mother-offspring interactions. *Early Human Development* **81**: 73-77.

Kaunitz AM, Hughes JM, Grimes DA, Smith JC, Rochat RW, Kafrissen ME. (1985) Causes of maternal mortality in the United States. *Obstetrics & Gynecology* **65**: 605-612.

Lamb K, Rizzino A. (1998) Effects of differentiation on the transcriptional regulation of the FGF-4 gene: Critical roles played by a distal enhancer. *Molecular Reproduction and Development* **51**: 218-224.

le Cessie S, van Houwelingen JC. (1992) Ridge estimators in logistic regression. *Applied Statistics* **41**: 191-201.

Lee A, Silvapulle M. (1988) Ridge estimation in logistic regression. *Communications in Statistics, Simulation and Computation* **17**: 1231-1257.

Lu Q, Elston RC. (2008) Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. *American Journal of Human Genetics* **82**: 641-651.

Lynch M, Walsh B. (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

McCullagh P, Nelder JA. (1989) *Generalized Linear Models* Chapman and Hall: London.

Minassian SL, Palmer CG, Sinsheimer JS. (2005) An exact maternal-fetal genotype incompatibility (MFG) test. *Genetic Epidemiology* **28**: 83-95.

Morison IM, Ramsay JP, Spencer HG. (2005) A census of mammalian imprinting. *Trends in Genetics* **21**: 457-465.

Odent M. (2001) Hypothesis: Preeclampsia as a maternal-fetal conflict. *Medscape General Medicine* **3**: 2.

Parimi N, Tromp G, Kuivaniemi H, Nien JK, Gomez R, Romero R, Goddard KA. (2008) Analytical approaches to detect maternal/fetal genotype incompatibilities that increase risk of pre-eclampsia. *BMC Medical Genetics* **9**: 60.

Park MY, Hastie T. (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**: 30-50.

Paz I, Gale R, Laor A, Danon YL, Stevenson DK, Seidman DS. (1995) The cognitive outcome of full-term small for gestational age infants at late adolescence. *Obstetrics & Gynecology* **85**: 452-456.

Pfeifer K. (2000) Mechanisms of genomic imprinting. *American Journal of Human Genetics* **67**: 777-787.

Saenger P, Czernichow P, Hughes I, Reiter EO. (2007) Small for gestational age: short stature and beyond. *Endocrine Reviews* **28**: 219-251.

Sinsheimer JS, Palmer CG, Woodward JA. (2003) Detecting genotype combinations that increase risk for disease: Maternal-fetal genotype incompatibility test. *Genetic Epidemiology* **24**: 1-13.

Shete S, Amos CI. (2002) Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *American Journal of Human Genetics* 70: 751-757.

Shete S, Zhou X. (2005) Parametric approach to genomic imprinting analysis with applications to angelmans syndrome. *Human Heredity* **59**: 26-33.

Tabano S, Alvino G, Antonazzo P, Grati FR, Miozzo M, Cetin I. (2006) Placental LPL gene expression is increased in severe intrauterine growth-restricted pregnancies. *Pediatric Research* **59**: 250-253.

Taylor DJ, Howie PW. (1989) Fetal growth achievement and neurodevelopmental disability. *British Journal of Obstetrics and Gynaecology* **96**: 789-794.

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* B **58**: 267-288.

Tycko B, Morison IM. (2002) Physiological functions of imprinted genes. *Journal of Cellular Physiology* **192**: 245-258.

Villar J, de Onis M, Kestler E, Bolanos F, Cerezo R, Bernedes H. (1990) The differential neonatal morbidity of the intrauterine growth retardation syndrome. *American Journal of Obstetrics and Gynecology* **163**: 151-157.

Wang L, Chen G, Li H. (2007) Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**: 1486-1494.

Weinberg CR, Wilcox AJ, Lie RT. (1998) A log-linear approach to case-parent triad data: Assessing effects of disease genes that act directly or through maternal effects, and may be subject to parental imprinting. *American Journal of Human Genetics* **62**: 969-978.

Wollmann HA. (1998) Intrauterine growth restriction: definition and etiology. *Hormone Research* **49**: 1-6.

Zhu J, Hastie T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**: 427-443.

Zou H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**: 1418-1429.

Zou H, Hastie T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* B **67**: 301-320.

Supplemental table: A full list of genes and SNPs showing significant effects associated with SGA

Gene symbol	rs Number	Location	Significant effects							
			a_m	d_m	a_o	d_o	i_m	i_c	MW	BS
APOC3	633938988	Exon 3	-	-	-	-0.372	-	-0.388	-	-
CCL2	633922561	Intron 1	-	-	-	-	-	0.644	-	-
CETP	8946997	Intron 10	-	-	-	0.423	-	-	-	-
COL1A2	28162145	Intron 4	-	-	0.424	-	-	-	-	-
	28139054	Intron 19	-	-	-	0.266	-	-	-0.420	-
	28171921	Intron 46	-	-	-0.321	-	-	-	-	-
	28698658	Intron 51	-	-	-	-	-	-0.515	-	-
COL4A1	633898711	Intron 1b	-	-	-	-	-0.836	-	-	-
	633901866	Intron 13	-	0.362	-	-	-	-	-	-0.363
COL4A3	634240165	Intron 2a	-	-	-	-	0.404	-	-	-
COL5A2	635134934	Exon 51	-	-	-	-	-1.330	-	-	-
CRHR2	617474044	Intron 5	-	-	-	-	0.489	-	-	-
	617474102	Intron 7	-	-	-	-	0.708	-	-	-
CSPG2	634208341	promoter	-	-	-	-	-	0.515	-	-
DAF	633860783	promoter	-0.349	-	-	-	-	-	-	-
F13A1	9370755	Exon 12	-	-	-	-	-	0.446	-	-
F2	9077693	Exon 6	0.280	-	-	-	-	-	-	-
	9084637	Intron 13	-0.195	-	-	-	-	-	-	-
FGF4	634043245	Exon 3	-	-	-	0.362	-	-	-	-
	634043464	Downstream	0.699	-	-	-	-	-	-	-
FLT4	22767327	Intron 7	-	-	-	-	0.569	-	-	-0.331
FN1	633938384	Exon 1	-	-	-	-	1.068	-	-	-
HSPG2	634092163	Intron 60	-	-	-	-	0.505	-	-	-
IFNGR2	5071132	Intron 5	-	-	-	-	-0.696	-	-	-
IGF1R	44530209	Intron 17a	-	-	-	-	0.643	-	-	-0.444
IL18BP	16402666	Intron 2	-	-0.388	-	-	-	0.505	-	-
IL2	634065022	Promoter	-	0.309	-	-	-	-	-	-
IL2RA	23884895	Intron 5	-	-	-	-	-	0.475	-	-
IL6	632284204	Promoter	-	-	-	-	-0.776	-	-	-
	3868962	Intron 2	-	-	-	-	-	-0.290	-	-
IL6R	24756885	Exon 2	0.484	-	-	-	-	-	-	-
LIPC	18683260	Promoter	-	-	-	-	-	-0.427	-	-0.331
LPL	22220155	Intron 6a	-	-	-	-	-	0.382	-	-
	612980414	Intron 8	-	-	-	-	-	0.456	-	-
LTF	633838634	Intron 13	-	-	-	-	-	0.432	-	-
MMP7	613913123	Exon 6	-	-	-	-	0.838	-	-	-
MMP9	17252653	Intron 4	-	-	-	-	-	0.810	-	-
NFKB1	659435702	Intron 22	-0.288	-	-	-	-	-	-	-0.273
NPY	3047643	Promoter	-	-	-	-	-	-0.272	-	-
PTGS1	628331732	Intron 7	-	0.385	-	-	-	-	-	-
REN	628862674	Intron 6	-	-	-	0.286	-	-	-	-
SPARC	1125290	Intron 5	-	-0.463	-	-	-	-	-	-
TIMP2	634841123	Exon 3	-	-	0.569	-	-	-	-	-
TLR9	4482153	Exon 2	0.263	-	-	-	-	-	-	-0.334
TNR	614058142	Intron 16a	-	-	-	-	0.374	-	-	–