Statistica Sinica **29** (2019), 1915-1937
doi:https://doi.org/10.5705/ss.202016.0114

IMPRINTING AND MATERNAL EFFECT DETECTION USING PARTIAL LIKELIHOOD BASED ON DISCORDANT SIBPAIR DATA

Fangyuan Zhang, Abbas Khalili and Shili Lin

Texas Tech University, McGill University and Ohio State University

Abstract: Numerous statistical methods have been developed to explore genomic imprinting and maternal effects, which are causes of parent-of-origin patterns in complex human diseases and are confounded. However, most of these methods have limitations: they may model only one of the two confounded epigenetic effects; they may make strong, yet unrealistic assumptions about the population to avoid over-parameterization; or they are applicable only to study designs that require the recruitment of difficult-to-obtain control families. In this study, we develop a partial likelihood method for detecting imprinting and maternal effects for a discordant sibpair design (LIME$_{DSP}$) utilizing all available sibship data without the need to recruit separate control families. By matching affected and unaffected probands and stratifying according to their familial genotypes, a partial likelihood component free of nuisance parameters can be extracted from the full likelihood. This alleviates the need to make assumptions about the population. Our theoretical analysis shows that the partial maximum likelihood estimators based on LIME$_{DSP}$ are consistent and asymptotically normally distributed. Using a closed-form formula, we compare a study design with more independent families and a design with larger families by keeping the total number of individuals that need to be genotyped fixed. We also conduct a simulation study to demonstrate the robust property of LIME$_{DSP}$ and show that it is a powerful approach that does not require recruiting control families. To illustrate its practical utility, LIME$_{DSP}$ is applied to a clubfoot disease data set and to the data from the Framingham Heart Study.

Key words and phrases: Ascertainment, association study, discordant sibpair design, imprinting effect, maternal effect, partial likelihood.

1. Introduction

Genome-wide association studies (GWAS) are used to identify common genetic variants associated with complex human traits and provide valuable insights into the genetic architecture of such traits. However, the variants identified thus far explain only a small proportion of the variability in most complex traits, leading to concerns about "missing heritability" (Manolio et al. (2009)). Efforts to

understand this missing heritability have revealed that, because gene expression is a dynamic process, DNA sequence polymorphism is not the only factor contributing to phenotypic variation. For example, other mechanisms that may be involved include epigenetic modification and transcriptional/translational regulation (Hirschhorn (2009); Peters (2014)). As a result, researchers are increasingly focusing on epigenetic factors, including imprinting and maternal genotype effects (Kohda (2013)).

Genomic imprinting is an epigenetic factor involving methylation and histone modifications that completely or partially silence the expression of a gene inherited from a particular parent, without altering the genetic sequence (Patten et al. (2014)). As such, genomic imprinting can lead to a parent-of-origin pattern in gene expressions, that is, an unequal expression of a heterozygous genotype, depending on whether the imprinted variant is inherited from the mother (maternal imprinting) or from the father (paternal imprinting). The imprinting effect has been hailed as a key factor in understanding the interplay between the epigenome and genome (Ferguson-Smith (2011)). On the other hand, the maternal genotype effect, another epigenetic effect, can also lead to a parent-of-origin pattern. This effect refers to the phenomenon in which the genotype of a mother is expressed in the phenotype of her offspring. This is usually attributed to the mother passing extra mRNAs and proteins to her offspring during pregnancy, which may change the expression level of certain genes.

Normal genetic imprinting contributes to a wide range of human growth and development (Wilkinson, Davies and Isles (2002); Peters (2014)). However, the deregulation of imprinted genes has been found to contribute to a number of complex human diseases, such as Beckwith–Wiedemann syndrome, Silver–Russell syndrome, Angelman syndrome, and Prader–Willi syndrome (Lim and Maher (2009)). At the same time, studies have shown that maternal effects play an important role in a variety of diseases, especially those related to pregnancy outcomes, such as childhood cancers and birth defects (Haig (2004)), certain psychiatric illnesses (Palmer et al. (2008)), and pregnancy complications (Svensson et al. (2009)). However, limited data availability and the insufficient power of current methods means that very few genes have been identified as having genomic imprinting or maternal effects.

Because both imprinting and maternal effects exhibit parent-of-origin patterns, family data are needed to trace inheritance paths. Here a common study design is that of case-parent triads, which may also include control-parent triads. Based on this design, numerous methods have been proposed to model

imprinting and maternal effects simultaneously in order to avoid potential confounding, because methods that attempt to detect only one of these effects may have inflated false positive or false negative rates when the other effect exists as well (see Lin (2013) and the references therein). However, almost all of these methods rely on strong, yet unrealistic assumptions about the population (e.g., mating symmetry) to avoid over-parameterization. The likelihood ratio test is a classic example (Weinberg, Wilcox and Lie (1998)). An exception is the recently proposed partial likelihood method for detecting imprinting and maternal effects (LIME), which alleviates the need to make unrealistic assumptions (Yang and Lin (2013)). However, the study design for the LIME method requires the recruitment of both case families and control families (Yang and Lin (2013)), with information from additional siblings accounted for in an extension to this method (Han, Hu and Lin (2013)). Thus, the price paid for avoiding assumptions that are difficult to satisfy is the need for separate control families, which are typically difficult to recruit. Recently, a mixture modeling approach was proposed for detecting imprinting. However, the data employed in this approach are gene expressions from a population sample (Li et al. (2015)), which differs from our family-based design.

To enjoy the benefits of the LIME method without needing control families, here we propose a LIME method based on a discordant sibpair design (LIME$_{DSP}$). The proposed method borrows from the work of Yang and Lin (2013) and Han, Hu and Lin (2013), but considers an alternative study design in which a nuclear family is recruited if there is a discordant sibpair; that is, one sibling is affected and the other is unaffected. Data from additional siblings (whether affected or not) may also be incorporated to further increase the method's power. The idea of LIME$_{DSP}$ is to match affected proband-parent triads with unaffected proband-parent triads, and then to factor out common terms involving mating-type probabilities, the nuisance parameters. By doing so, LIME$_{DSP}$ circumvents the problem of over-parameterization, unrealistic assumptions, and the need for control families in the original LIME design. When control families are available, they can be utilized to further increase the statistical power of the method. Finally, note that the discordant sibpair design is popular in linkage and association studies (Horvath and Laird (1998)), which provide practical applications for LIME$_{DSP}$.

2. Partial Likelihood Method (LIME$_{DSP}$)

2.1. Notation and genetic model

Consider a candidate genetic marker with two alleles A and B, where A is the allele of interest, the variant allele, which may represent disease susceptibility or an epigenetic effect. In a nuclear family, let F and M be the random variables denoting the number of A alleles carried by the father and mother respectively, which can take values 0, 1, or 2, corresponding to genotypes BB, AB, or AA, respectively. Similarly, let C_i be a random variable denoting the number of A alleles (i.e. the genotype) of child i, for $i = 1, 2, \ldots$. Specifically, C_1 and C_2 denote the affected and unaffected probands, respectively, through which the family is recruited, whereas C_i, for $i = 3, \ldots,$ denote the additional siblings, if any. Then D_i, for $i = 1, 2, \ldots,$ denotes the disease status of a child (affected = 1; normal = 0). Thus, $D_1 = 1$ and $D_2 = 0$. The development of LIME$_{DSP}$ is based on a multiplicativ risk model for disease prevalence for a triad family:

$$P(D = 1|M = m, F = f, C = c) = \delta r_1^{I(c=1)} r_2^{I(c=2)} r_{im}^{I(c=1_m)} s_1^{I(m=1)} s_2^{I(m=2)}, \quad (2.1)$$

where r_1 and r_2 denote the effects of one or two copies of an individual's own variant allele, r_{im} denotes the imprinting effect, s_1 and s_2 denote the effects of one or two copies of the mother's variant allele, and δ is the phenocopy rate. The notation $c = 1_m$ indicates that the child's genotype is AB, where the variant allele A is from the mother. We need to estimate the model parameters, collectively denoted as $\boldsymbol{\theta} = (\delta, r_1, r_2, r_{im}, s_1, s_2)^T$, although the phenocopy rate δ may also be regarded as a nuisance parameter. Note that all parameters are positive, and a parameter is identifiable and estimable only if the required data are available. Furthermore, $r_{im} > 1, < 1, = 1$ signify paternal, maternal, or no imprinting effects, respectively. Although no restriction is placed on s_1 and s_2, they are typically ≥ 1, with the equality denoting no maternal effect. A further constraint placed on the parameters is that $P(D|M = m, F = f, C = c) \leq 1$.

2.2. Ascertainment and probability formulation

Because families are ascertained through discordant sibpairs, the probability of the observed data from a family will be conditional on the affection status of the two probands only (i.e., not on any additional siblings):

$$\begin{aligned} &P(M = m, F = f, C_1 = c_1, C_2 = c_2, C_i = c_i, D_i = d_i, i = 3, \cdots \mid D_1 = 1, D_2 = 0) \\ &= P(M = m, F = f, C_1 = c_1 | D_1 = 1, D_2 = 0) \\ &\quad P(M = m, F = f, C_2 = c_2 | D_1 = 1, D_2 = 0) \end{aligned} \quad (2.2)$$

$$\times \prod_{i\geq 3} P(C_i = c_i|M = m, F = f)P(D_i = d_i|M = m, F = f, C_i = c_i) \tag{2.3}$$

$$\times \frac{P(D_1 = 1, D_2 = 0)}{P(M = m, F = f)P(D_1 = 1|M = m, F = f)P(D_2 = 0|M = m, F = f)}. \tag{2.4}$$

A detailed derivation of this formula can be found in Supplementary Material S1. On the right-hand side of the above formula, the probability of the observed data is expressed as the product of three components: the proband-parents triad probability (mother, father, and child) conditional on the proband disease status (2.2), the joint probability of the genotypes and phenotypes of any additional siblings given the parents' genotypes (2.3), and the remaining part (2.4). The component expressed in 2.2 regarding the contribution from the probands can be thought of as being obtained from a "retrospective" design, which can be turned into a "prospective" design using stratification, as discussed in detail below. The second component, given in 2.3, accounts for information from additional siblings and is formulated using a "prospective" design and free of any nuisance parameters. The last component shown in 2.4 is the remaining term that contains the nuisance parameters. Whereas the prospective part is straightforward, involving parameters of interest only, as can be seen from disease risk model (2.1), the retrospective part is more intricate and is examined in detail in the following subsection.

We first note that, in (2.2),

$$\begin{aligned} &P(M = m, F = f, C_1 = c_1|D_1 = 1, D_2 = 0) \\ &= \frac{P(M = m, F = f, C_1 = c_1, D_1 = 1, D_2 = 0)}{P(D_1 = 1, D_2 = 0)}. \end{aligned} \tag{2.5}$$

There are 15 possible combinations of genotypes for the parents (M, F) and a child (C); these, together with their labeling (types), are listed in Table 1, with the corresponding probability for the numerator in (2.5) given in the last column of the top segment. Similarly, the probability $P(M = m, F = f, C_2 = c_2, D_1 = 1, D_2 = 0)$ is given in the last column of the bottom segment of the table. Derivations of the probabilities for a few of the cases are provided in Supplementary Material S2. In the expressions in Table 1, μ_{mf} ($m = 0, 1, 2$, $f = 0, 1, 2$) denotes the mating-type probabilities, that is, $\mu_{mf} = P(M = m, F = f)$. Note that we do not make any assumptions about the mating-type probabilities, such as Hardy-Weinberg equilibrium (HWE) or even mating symmetry; thus, μ_{mf} is not necessarily equal to μ_{fm}. As shown in the table, these nuisance parameters can be factored out completely from the six model parameters. This

Table 1. Joint probability of mother-father-child triad genotypes and proband disease status.

(a). Triad genotype with affected child				
Type	m	f	c	$P(M=m,F=f,C_1=c,D_1=1,D_2=0)^a$
1	0	0	0	$\mu_{00}\delta(1-\delta)$ [b]
2	0	1	0	$\mu_{01}(1/2)\delta(1/2)(2-\delta-\delta r_1)$
3	0	1	1	$\mu_{01}(1/2)\delta r_1(1/2)(2-\delta-\delta r_1)$
4	0	2	1	$\mu_{02}\delta r_1(1-\delta r_1)$
5	1	0	0	$\mu_{10}(1/2)s_1\delta(1/2)(2-\delta s_1-\delta r_1 r_{im}s_1)$
6	1	0	1	$\mu_{10}(1/2)\delta r_1 r_{im}s_1(1/2)(2-\delta s_1-\delta r_1 r_{im}s_1)$
7	1	1	0	$\mu_{11}(1/4)\delta s_1(1/4)(4-\delta s_1-\delta s_1 r_1-\delta s_1 r_1 r_{im}-\delta r_2 s_1)$
8	1	1	1	$\mu_{11}(1/4)\delta s_1 r_1(1+r_{im})(1/4)(4-\delta s_1-\delta s_1 r_1-\delta s_1 r_1 r_{im}-\delta r_2 s_1)$
9	1	1	2	$\mu_{11}(1/4)\delta s_1 r_2(1/4)(4-\delta s_1-\delta s_1 r_1-\delta s_1 r_1 r_{im}-\delta r_2 s_1)$
10	1	2	1	$\mu_{12}(1/2)\delta r_1 s_1(1/2)(2-\delta r_1 s_1-\delta r_2 s_1)$
11	1	2	2	$\mu_{12}(1/2)\delta r_2 s_1(1/2)(2-\delta r_1 s_1-\delta r_2 s_1)$
12	2	0	1	$\mu_{20}\delta r_1 s_2 r_{im}(1-\delta r_1 s_2 r_{im})$
13	2	1	1	$\mu_{21}(1/2)\delta r_1 s_2 r_{im}(1/2)(2-\delta r_1 s_2 r_{im}-\delta r_2 s_2)$
14	2	1	2	$\mu_{21}(1/2)\delta r_2 s_2(1/2)(2-\delta r_1 s_2 r_{im}-\delta r_2 s_2)$
15	2	2	2	$\mu_{22}\delta r_2 s_2(1-\delta r_2 s_2)$
(b). Triad genotype with unaffacted child				
Type	m	f	c	$P(M=m,F=f,C_2=c,D_1=1,D_2=0)^a$
1	0	0	0	$\mu_{00}\delta(1-\delta)$
2	0	1	0	$\mu_{01}(1/2)(1-\delta)(1/2)\delta(1+r_1)$
3	0	1	1	$\mu_{01}(1/2)(1-\delta r_1)(1/2)\delta(1+r_1)$
4	0	2	1	$\mu_{02}\delta r_1(1-\delta r_1)$
5	1	0	0	$\mu_{10}(1/2)(1-\delta s_1)(1/2)\delta s_1(1+r_1 r_{im})$
6	1	0	1	$\mu_{10}(1/2)(1-\delta r_1 r_{im}s_1)(1/2)s_1\delta(1+r_1 r_{im})$
7	1	1	0	$\mu_{11}(1/4)(1-\delta s_1)(1/4)\delta s_1(1+r_1+r_1 r_{im}+r_2)$
8	1	1	1	$\mu_{11}(1/4)(2-\delta s_1 r_1(1+r_{im}))(1/4)\delta s_1(1+r_1+r_1 r_{im}+r_2)$
9	1	1	2	$\mu_{11}(1/4)(1-\delta s_1 r_2)(1/4)\delta s_1(1+r_1+r_1 r_{im}+r_2)$
10	1	2	1	$\mu_{12}(1/2)(1-\delta r_1 s_1)(1/2)\delta s_1(r_1+r_2)$
11	1	2	2	$\mu_{12}(1/2)(1-\delta r_2 s_1)(1/2)\delta s_1(r_1+r_2)$
12	2	0	1	$\mu_{20}\delta r_1 s_2 r_{im}(1-\delta r_1 s_2 r_{im})$
13	2	1	1	$\mu_{21}(1/2)(1-\delta r_1 s_2 r_{im})(1/2)\delta s_2(r_1 r_{im}+r_2)$
14	2	1	2	$\mu_{21}(1/2)(1-\delta r_2 s_2)(1/2)\delta s_2(r_1 r_{im}+r_2)$
15	2	2	2	$\mu_{22}\delta r_2 s_2(1-\delta r_2 s_2)$

Note: [a]M, F, and C are the number of variant alleles carried by the mother, father, and child in a triad, and take values of 0, 1, or 2; the mating type probability for $(M,F)=(m,f)$ is denoted by μ_{mf}; $D_1=1$ ($D_2=0$) indicates that the child is affected (unaffected). [b]Notation for model parameters, δ: the phenocopy rate; r_1: relative risk of carrying one variant allele; r_2: relative risk of carrying two variant alleles; r_{im}: imprinting effect parameter with a single variant allele from mother; s_1: maternal effect with mother carrying one variant allele; s_2: maternal effect with mother carrying two copies of the variant allele.

observation forms the basis of the partial likelihood formulation.

2.3. Organization of data

Table 1 shows that, conditional on each possible triad genotype vector (m, f, c), the counts of the affected and unaffected proband-parent triads share the same nuisance parameter components μ_{mf}. Thus, the proportion of affected proband-parents triads among all triads with that genotype vector is free of nuisance parameters. For example, among all proband-parent triads with the genotype combination (m, f, c), the probability of observing an affected proband-parent triad is

$$\begin{aligned} p_{mfc} &= \frac{NP(m, f, C_1 = c|D_1 = 1, D_2 = 0)}{NP(m, f, C_1 = c|D_1 = 1, D_2 = 0) + NP(m, f, C_2 = c|D_1 = 1, D_2 = 0)} \\ &= \frac{P(m, f, C_1 = c, D_1 = 1, D_2 = 0)}{P(m, f, C_1 = c, D_1 = 1, D_2 = 0) + P(m, f, C_2 = c, D_1 = 1, D_2 = 0)} \\ &= \frac{P(D = 1|m, f, c)P(D = 0|m, f)}{P(D = 1|m, f, c)P(D = 0|m, f) + P(D = 0|m, f, c)P(D = 1|m, f)}, \end{aligned} \tag{2.6}$$

which includes only those parameters in (1). This manipulation turns the data from a retrospective design into a "prospective" design using stratifation according to each triad genotype combination. We denote the denominator of (2.6) as S_{mfc}. Thus, $p_{mfc} = P(D = 1|m, f, c)P(D = 0|m, f)/S_{mfc}$.

By applying this idea to the overall likelihood, we can extract a partial likelihood component that only involves the parameters of interest. Let n^1_{mfc} and n^0_{mfc} denote the count of affected proband-parent triads and unaffected proband-parent triads, respetively, with genotype $M = m$, $F = f$, and $C = c$. Note that $N = \sum_{m,f,c} n^1_{mfc} + \sum_{m,f,c} n^0_{mfc}$ is the number of independent families. Similarly, let sn^1_{mfc} and sn^0_{mfc} denote the counts of affected additional sibling-parent triads and unaffected additional sibling-parent triads, respectively, with genotype combination $M = m$, $F = f$, and $C = c$. Recall that we denote the vector of the parameters of interest by $\boldsymbol{\theta} = (\delta, r_1, r_2, r_{im}, s_1, s_2)^\top$. We further denote the vector of nuisance parameters (including the mating-type probabilities) by $\boldsymbol{\phi}$. Then, according to the three component factorization,

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}) =$$
$$\prod_{m,f,c} [P(m, f, C_1 = c|D_1 = 1, D_2 = 0)]^{n^1_{mfc}} [P(m, f, C_2 = c|D_1 = 1, D_2 = 0)]^{n^0_{mfc}}$$

$$\times \prod_{m,f,c} [P(c|m,f)]^{sn^1_{mfc}+sn^0_{mfc}} [P(D=1|m,f,c)]^{sn^1_{mfc}} [P(D=0|m,f,c)]^{sn^0_{mfc}}$$

$$\times \prod_{m,f,c} \left[\frac{P(D_1=1,D_2=0)}{P(m,f)P(D_2=0|m,f)P(D_1=1|m,f)}\right]^{n^1_{mfc}}$$

$$\propto \prod_{m,f,c} p^{n^1_{mfc}}_{mfc}(1-p_{mfc})^{n^0_{mfc}} \prod_{m,f,c} q^{sn^1_{mfc}}_{mfc}(1-q_{mfc})^{sn^0_{mfc}} \tag{2.7}$$

$$\times \prod_{m,f,c} S^{n^1_{mfc}+n^0_{mfc}}_{mfc} \left[\frac{P(D_1=1,D_2=0)}{P(m,f)P(D_2=0|m,f)P(D_1=1|m,f)}\right]^{n^1_{mfc}}, \tag{2.8}$$

where p_{mfc} and S_{mfc} are defined as above, and $q_{mfc} = P(D=1|M=m,F=f,C=c)$.

Note that all of the nuisance parameters in $\boldsymbol{\phi}$ are present only in (2.8), whereas the factors in (2.7) contain only the parameters in $\boldsymbol{\theta}$, which is therefore taken as our partial likelihood. The parameters in $\boldsymbol{\theta}$ can be inferred by maximizing the partial likelihood instead of the full likelihood to avoid estimating the nuisance parameters (Cox (1975)). In fact, the first factor of the partial likelihood component can be regarded as the likelihood of the reorganized data, conditional on each possible triad (m,f,c) type. Within each type, the counts of the affected-proband triads follow a "renormalized" binomial distribution with the conditional probability p_{mfc}. The second factor, on the other hand, represents the contributions from the additional siblings. Because the affection statuses of the additional siblings are obtained prospectively, the probability of observing affected sibling-parent triads in a particular familial genotype combination (m,f,c) is simply the penetrance probability. Furthermore, by design, p_{mfc} does not involve population disease prevalence information $P(D=1)$, which is another nuisance parameter.

2.4. Partial likelihood and asymptotic properties

From the above organization of the data, it is clear that the log partial likelihood $l_{par}(\boldsymbol{\theta})$ is as follows:

$$l_{par}(\boldsymbol{\theta}) = \sum_{m,f,c} \left\{ n^1_{mfc} \times \log[p_{mfc}] + n^0_{mfc} \times \log[1-p_{mfc}] \right\}$$
$$+ \sum_{m,f,c} \left\{ sn^1_{mfc} \times \log[q_{mfc}] + sn^0_{mfc} \times \log[1-q_{mfc}] \right\}.$$

By solving the score-type equation

$$\frac{\partial l_{par}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = l'_{par}(\boldsymbol{\theta}) = \mathbf{0}, \tag{2.9}$$

the maximum partial likelihood estimator (MPLE) of $\boldsymbol{\theta}$ can be obtained following the work of Zhang, Khalili and Lin (2016).

We use n to represent the total number of the four types of triads inferred from the families in the partial log-likelihood $l_{par}(\boldsymbol{\theta})$: affected proband-parent triads, unaffected proband-parent triads, affected additional sibling-parent triads, and unaffected additional sibling-parent triads. That is,

$$n = \sum_{m,f,c} n^0_{mfc} + \sum_{m,f,c} n^1_{mfc} + \sum_{m,f,c} sn^0_{mfc} + \sum_{m,f,c} sn^1_{mfc}.$$

As we can see from the partial likelihood, these four types of triads contribute independent information, conditional on the genotype of the parents. Thus, n is regarded as the effective sample size. We study the asymptotic properties of the MPLE of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}_n$, as the effective sample size n tends to infinity.

Let $\boldsymbol{\theta}_0$ denote the true value of the parameter vector $\boldsymbol{\theta} = (\delta, r_1, r_2, r_{im}, s_1, s_2)^\top$. We assume that $\boldsymbol{\theta}_0$ is an interior point of the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^6$.

Theorem 1. *Under the regularity conditions provided in Supplementary Material S3, we have the following:*

(i) The likelihood equation has a unique consistent solution $\widehat{\boldsymbol{\theta}}_n$, i.e. $\widehat{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}_0$ with probability tending to one.

(ii) Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \longrightarrow N(0, I^{-1}(\boldsymbol{\theta}_0))$, where $I(\boldsymbol{\theta}_0)$ is the information matrix given by

$$I(\boldsymbol{\theta}_0) = \sum_{m,f,c} \frac{[p'_{mfc}(\boldsymbol{\theta}_0)][p'_{mfc}(\boldsymbol{\theta}_0)]^\top \times B_{mfc}}{p_{mfc}(\boldsymbol{\theta}_0)(1 - p_{mfc}(\boldsymbol{\theta}_0))} + \sum_{m,f,c} \frac{[q'_{mfc}(\boldsymbol{\theta}_0)][q'_{mfc}(\boldsymbol{\theta}_0)]^\top \times C_{mfc}}{q_{mfc}(\boldsymbol{\theta}_0)(1 - q_{mfc}(\boldsymbol{\theta}_0))},$$

where $0 \le B_{mfc} < 1$ and $0 \le C_{mfc} < 1$ are the limits in probability of $\{(n^1_{mfc} + n^0_{mfc})/n\}$ and $\{(sn^1_{mfc} + sn^0_{mfc})/n\}$, respectively, when $n \to \infty$.

The proof of the theorem can be found in Supplementary Material S3. Note that although the consistent solution of partial likelihood score equation (2.9) is unique (Chanda (1954); Lindsay (1980)), there may exist inconsistent roots.

2.5. Combining data from the two study designs

In a real data analysis, both case-control family data and discordant sibpair

data may exist. Therefore, it is important to combine all information to make full use of the data, leading to the proposal of LIME$_{D+}$. Suppose data set A is obtained from a case-control family design. Then, the LIME method of Yang and Lin (2013) is applied to extract the partial likelihood $pL_A(\boldsymbol{\theta})$. On the other hand, if data set B is the consequence of a discordant sibpair study design, then we use the currently proposed LIME$_{DSP}$ approach to obtain the partial likelihood component $pL_B(\boldsymbol{\theta})$. The total partial likelihood for all available data is then $pL(\boldsymbol{\theta}) = pL_A(\boldsymbol{\theta}) * pL_B(\boldsymbol{\theta})$, given that the data in sets A and B are independent. Note that if both studies focus on the the same underlying disease model, then the parameters of interest are identical. The model parameters in $\boldsymbol{\theta}$ are estimated by maximizing the partial likelihood $pL(\boldsymbol{\theta})$. The MPLE of LIME$_{D+}$ has the same asymptotic properties as those of LIME$_{DSP}$.

3. Evaluation of Information Content

In practical applications, resources are finite. As such, it is important to have a good understanding of the information contained in commonly used study designs. Questions of interest include the roles of additional siblings in the DSP design, and in particular, whether it is better to recruit additional siblings (if available) or additional independent families by considering "per individual" information. To facilitate this investigation, we consider eight disease models (Table 2). The first three models have no imprinting nor maternal effects. Models 4 has maternal effects only, models 5 and 6 have imprinting effects only, and models 7 and 8 have both types of parent-of-origin effects. For each of these eight models, we consider eight scenarios, which are combinations of two levels of minor allele frequency (MAF) {0.1, 0.3}, two levels of population disease prevalence $P(D = 1)$ (PREV) {0.05, 0.15}, and two levels of HWE (not hold = 0, hold = 1). Suppose p is the MAF, then the probabilities of a genotype taking the values 0, 1, and 2 are $(1-p)^2(1-\zeta)+(1-p)\zeta$, $2p(1-p)(1-\zeta)$, and $p^2(1-\zeta)+p\zeta$, respectively, where ζ is the inbreeding parameter (Weir (1996)). When HWE holds, $\zeta = 0$. When HWE does not hold, ζ is set to 0.1 and 0.3 for males and females, respectively. Note that with the specification of each scenario and a disease model, the penetrance probability (2.1) is fully specified. Because the summation over the 15 joint probabilities $P(D = 1, M, F, C)$ is equal to the disease prevalence $P(D = 1)$, the phenocopy rate can be solved from the equation.

Intuitively, including additional siblings in a DSP design will typically in-

Table 2. Eight disease models and eight scenarios comprised of three factors.

	Model Parameters[a]					Scenario Factors[b]		
model/scenario	r_1	r_2	r_{im}	s_1	s_2	MAF	PREV	HWE
1	1	1	1	1	1	0.1	0.05	0
2	2	3	1	1	1	0.1	0.05	1
3	1	3	1	1	1	0.1	0.15	0
4	1	3	1	2	2	0.1	0.15	1
5	1	3	3	1	1	0.3	0.05	0
6	3	3	1/3	1	1	0.3	0.05	1
7	1	3	3	2	2	0.3	0.15	0
8	3	3	1/3	2	2	0.3	0.15	1

Note: [a]The notation for the model parameters is the same as that in Table 1. [b]MAF: minor allele frequency; PREV: prevalence (rare = 0.05; common = 0.15); HWE: Hardy-Weinberg equilibrium (Yes = 1; No = 0); a specification of a disease model and a scenario completely determines the penetrance model specified in equation (2.1).

crease the information available for estimating the model parameters, and, hence the detection power for a fixed sample of N families. In fact, this is demonstrated using a theoretical calculation of "per family" information content (Supplementary Fig. S1). However, including additional siblings leads to a larger number of total individuals, and hence greater genotyping and phenotyping costs, even if the number of families N remains fixed. As such, whether it is beneficial to recruit additional siblings is no longer clear from the perspective of "per individual" information content, which is the average information contributed by a single family member. We take up this investigation by considering three study designs, D, $D+1$, and $D+2$, denoting a DSP design with 0, 1, and 2 additional siblings, respectively, leading to a total of 4, 5, and 6 individuals, respectively, per family. Figure 1 shows the information content per individual for the three study designs when HWE holds and MAF is 0.3 (scenarios 6 and 8 in Table 2) for all eight disease models. Plots for the other scenarios are given in the Supplementary Material, Fig. S2-4. Unsurprisingly, the figures show that there is essentially no information for inferences on the maternal effect parameters s_1, s_2 when only discordant sibpairs are recruited. This is because the two siblings in a discordant sibpair share the same mother, which provides a very limited contrast for the maternal effect. A theoretical explanation is provided in Supplementary Material S4. Fortunately, when additional siblings are available, maternal effects can be estimated. For the other parameters r_1, r_2, and r_{im}, the efficiency depends on the disease prevalence. When the disease prevalence is high (0.15), recruiting

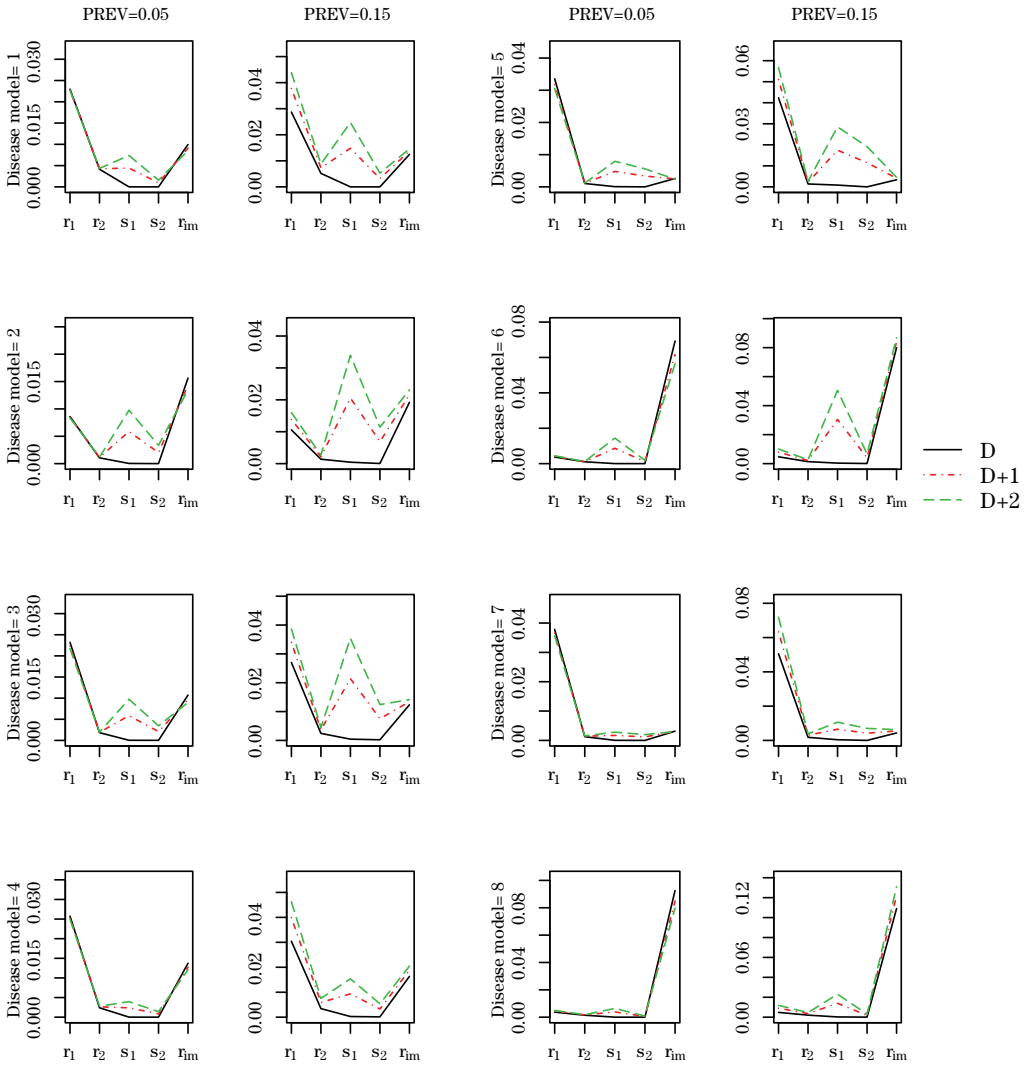

Figure 1. Information content per individual for eight disease models and two PREVs when HWE holds and MAF is 0.3. Each curve depicts the information for estimating one of the five parameters for data types D, $D+1$, and $D+2$.

additional siblings, which are likely to include affected cases given the common disease, will increase the efficiency. On the other hand, when the disease prevalence is low (0.05), recruiting additional independent families or siblings leads to fairly similar results (apart from estimating the maternal effects), although having a larger number of independent families is slightly better for estimating the other parameters. Thus, depending on the disease prevalence and the which

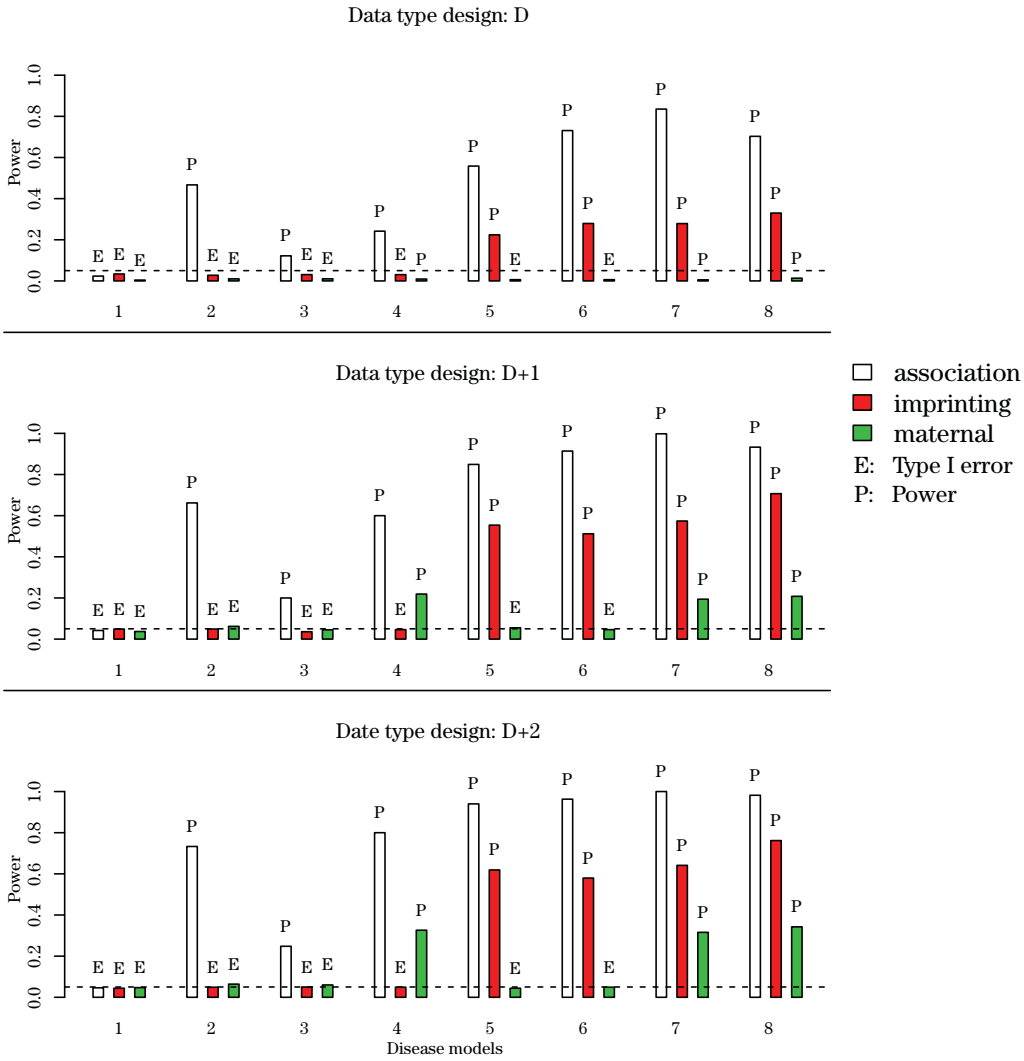

Figure 2. Type I error rate and power of LIME$_{DSP}$ under eight disease models and scenario 1, as given in Table 2. The three rows represent three data types: D, $D+1$, and $D+2$. The three bars refer to association, imprinting, and maternal effects, respectively. The horizontal line marks the nominal level of 0.05.

parameters are of greater interest, the most efficient design may vary.

4. Simulation

Given our understanding of LIME$_{DSP}$ from the theoretical analysis, in this section, we demonstrate its empirical performance with finite samples by studying

its size and power in a simulation for a typical sample size in genetic epidemiology. We consider D, $D+1$, and $D+2$ designs, each with 300 families. All combinations of the eight disease models and eight population scenarios are included, leading to 192 ($3 \times 8 \times 8$) simulation settings, with 1,000 simulated data sets under each setting.

Figure 2 shows the empirical type I error rates and the power of LIME$_{DSP}$ under all eight disease models and scenario 1. The three rows represent the three designs considered. The three bars refer to association, imprinting, and maternal effects, respectively. The results show that the type I error rates are close to the nominal value of 0.05, marked by a horizontal dashed line for association under model 1, the imprinting effect under models 1, 2, 3, and 4, and the maternal effect under models 1, 2, 3, 5, and 6, across all three designs. Note that when there are no additional siblings (i.e. the D design), the type I error rate for the maternal effect is rather low. This is not surprising because, as we discussed earlier, such data provide no information for inferring the maternal effect. Comparing the three designs, we can see that the power increases as additional siblings are recruited, especially when detecting the maternal effect. Note that LIME$_{DSP}$ is incapable of detecting the maternal effect when there are only discordant sibpairs, but that the power increases when additional siblings are available. The results for the other seven scenarios are similar and are shown in the Supplementary Material, Fig. S5-11.

5. Real Data Analysis

To illustrate the application of LIME$_{DSP}$ and LIME$_{D+}$ to real human genetic studies, we consider two complex diseases with established genetic bases, namely, clubfoot and the Framingham Heart Study (FHS). Both studies are family based and have extended pedigrees. In the clubfoot data, we extract nuclear families with discordant sibpairs and additional siblings, if available. Thus, LIME$_{DSP}$ is applicable to these data. For the FHS, we extract nuclear families that have discordant sibpairs or are case-parent or control-parent triads, all potentially involving additional siblings, and analyze these data using LIME$_{D+}$.

5.1. Analysis of the clubfoot data

Clubfoot is a congenital deformity in which the affected foot appears to have been rotated internally at the ankle. With treatment, most patients recover completely during early childhood and are able to walk and participate in athletics. Thus, understanding the underlying causal mechanism is important for

Table 3. Top SNPs for association, imprinting, and maternal effects for the clubfoot data using LIME_{DSP}.

Effect	SNP	Chr	Position (BP)*	Gene	$-\log_{10}$ (P-value)
Association	rs1568717	15	61362446	RORA	3.52
Imprinting	rs2145214	20	42237066	IFT52	11.99
	rs11048527	12	26604100	ITPR2	11.10
	rs6785520	3	170991646	TNIK	10.97
Maternal	rs9446305	6	71598570	B3GAT2	4.55
	rs11766624	7	69887084	AUTS2	4.50
	rs585157	13	99045319	FARP1	4.47

*Position (BP) is the genomic position of the SNP relative to the start of the chromosome (Chr) in terms of the base pair (BP).

the development of effective treatment strategies. Our LIME_{DSP} analysis uses 87 discordant sibpairs with 33 additional siblings. These range from discordant sibpairs without additional siblings to discordant sibpairs with six siblings. The data are obtained from dbGaP (`www.ncbi.nlm.nih.gov/gap/`).

Among the top (i.e. the smallest p-values) single nucleotide polymorphisms (SNPs) identified by LIME_{DSP} (Table 3), some reside within genes that have been identified in the literature, either for symptoms directly related to clubfoot or for other congenital diseases. For example, two SNPs (rs11048527 and rs6785520) with very small p-values for imprinting effects are in genes that have recently been found to be associated with clubfoot. Specifically, a duplication in a region of the gene ITPR2 was found in a patient presenting symptoms that include clubfoot (Al-Qattan (2013)). The most direct evidence of the involvement of the gene TNIK comes from the study of Zhang et al. (2014), in which the authors showed that the p-value for the association between the gene and clubfoot is less than 0.001. As another example, one of the top SNPs (rs9446305) with some evidence of a maternal effect is in gene B3GAT2, the association of which with the clubfoot syndrome has been discussed (`http://biograph.be/concept/graph/C1866294/C1412717`). In addition, SNP rs11766624, residing in the AUTS2 gene, also has a relatively small p-value for detecting the maternal effect. It has been found that deletion of exon 6 of the AUTS2 gene can cause congenital disorders, including eversion of the feet. Note that multiple studies have identified rare mutations in the AUTS2 gene with autism, another congenital disease (Oksenberg et al. (2013)). In fact, autism has been found to be related to maternal effects (Zandi et al. (2006)), consistent with our finding.

Table 4. Top SNPs for association, imprinting and maternal effects for the Framingham Heart Study data using LIME_{D+}.

Effect	SNP	Chr	Position (BP)*	Gene	$-\log_{10}$ (P-value)
Association	rs16892095	4	15518356	CC2D2A	15.65
	rs2229188	7	92134309	CYP51A1	15.11
Imprinting	rs2290201	8	82394701	FABP4	5.32
	rs2213162	12	48390721	COL2A1	4.46
	rs1562705	2	142796062	LRP1B	4.36
	rs6471053	8	133310740	KCNQ3	4.10
Maternal	rs2272487	3	126451936	CHCHD6	8.44
	rs9852584	3	126445456	CHCHD6	6.26
	rs13230531	7	6114558	CHCHD6	5.52
	rs7741727	6	132069916	ENPP3	5.19
	rs1370656	2	178607997	PDE11A	5.18
	rs7133914	12	40702910	LRRK2	5.16

*The Position(BP) is the genomic position of the SNP relative to the start of the chromosome (Chr) in terms of base pair (BP).

LIME_{DSP} also identifies some other genes that have been reported to be associated with complex developmental traits in the literature. For example, RORA is related to autism (Nguyen et al. (2010)), and TNIK and FARP1 are related to fetal brain outgrowth and development (Coba et al. (2012)). In a recent study, gene IFT52 was linked to skeletal ciliopathy, manifestations of which include congenital diseases (Girisha et al. (2016)). A list of the top-20 SNPs (with the smallest p-values) identified by LIME_{DSP} for each of the association, imprinting, and maternal effects can be found in the Supplementary Material, Tables S1-3. Given the large number of SNPs investigated, some of those identified may not be genome-wide significant. A complete set of results for all of the SNPs analyzed are provided in the Supplementary Material, Fig. S12-14.

5.2. Analysis of the FHS data

The FHS is a long-term, ongoing cardiovascular risk study on cohorts of residents in Framingham, Massachusetts. We focus on hypertension, a multifactorial complex trait, which can increase the risk of coronary heart disease. A person is classified as hypertensive if his/her systolic blood pressure is $\geq$ 140mmHg, or diastolic blood pressure is $\geq$ 90mmHg, or if he/she takes medication to control blood pressure. In this analysis, we focus on 263 DSP families (with 229 additional siblings), 436 case-parent triads, and 281 control-parent triads (with 230 additional siblings in total). Because the data comprise not only DSP families,

but also case-control families, we use the LIME$_{D+}$ procedure, which is applicable to a mixture of these two types of families.

Many top SNPs identified as associated with the hypertensive trait by LIME$_{D+}$ (top segment of Table 4) have been identified in the literature as related to hypertension, cardiovascular-related disorders, or other complex diseases. Specifically, SNP rs16892095, residing in the intron region of gene CC2D2A on chromosome 4, is found to be associated with the Meckel and Joubert syndromes, conditions that may be related to atrial septal defects (Elmali et al. (2014)). In addition, rs2229188 is an SNP associated with hypertension. It is in the intron region of gene CYP51A1 on chromosome 7. There are a number of haplotypes involving rs2229188 that are inferred to be strongly associated with hypertension (Wang and Lin (2014)).

Several of the genes found to potentially exert an imprinting effect on hypertension (middle segment of Table 4) are worth discussing. Previous research suggests that the FABP4 level, related to adiposity and metabolic disorders, is a novel predictor of cardiovascular mortality in end-stage renal disease (Furuhashi et al. (2011)). In addition, FABP4 has been found to contribute to blood pressure elevation and the atherogenic metabolic phenotype, and an elevated FABP4 level is predisposed by a family history of hypertension (Ota et al. (2012)). Gene COL2A1 in chromosome 12 is highly expressed in endocardial cushions and is very important in heart valve function (Peacock et al. (2008)). Furthermore, LRP1B is important in the development of atherosclerosis, a disease that affects arterial blood vessels (`www.scbt.com/datasheet-49230-lrp1b-n-19-antibody.html`). On the other hand, gene KCNQ3 in chromosome 8, together with other KCNQ channels, is believed to play a functional role in pulmonary artery smooth muscle (Joshi, Balan and Gurney (2006)).

Finally, four of the top genes for maternal effects that harbor multiple SNPs (last segment of Table 4) have been discussed in the literature. In particular, gene CHCHD6 has been identified as having a hypertension risk effect in a linkage analysis on chromosome 3 (Chiu et al. (2014)). On the other hand, gene ENPP3 in chromosome 6 is a member of the ENPP family. Rucker et al. (2007) demonstrated the presence of this family in the cardiac system, which suggests that these enzymes could contribute to the fine-tuning control of the nucleotide levels at the nerve terminal endings of left ventricles involved in several cardiac pathologies. As another example, gene PDE11A is associated with the development of adrenocortical hyperplasia, which leads to Cushing syndrome (Horvath et al. (2006)), and Cushing syndrome has clinical manifestations of arterial hyper-

tension. Finally, LRRK2 mutant mice was found to have caused blood pressure changes (Herzig et al. (2011)). A list of the top-20 SNPs (with the smallest p-values) identified by LIME$_{D+}$ for associatoon, imprinting, and maternal effects can be found in the Supplementary Material, Tables S4-6. As with the clubfoot study, some of the SNPs identified may not reach genome-wide significance. A complete set of results for all of the SNPs analyzed is provided in the Supplementary Material, Fig. S15-17.

6. Discussion

Imprinting and maternal effects are two confounding epigenetic factors that are increasingly being explored for their roles in complex traits. The partial likelihood method proposed in this paper, LIME$_{DSP}$, provides a robust approach for detecting these two effects without needing to make unrealistic assumptions or requiring the collection of separate control families. Based on the asymptotic property of LIME and a closed-form formula for calculating information, we provide a tool for comparing the relative efficiency of various study designs for a specific underlying disease model. We carried out a simulation study with finite samples to demonstrate the robustness of LIME$_{DSP}$ without sacrificing power.

We further applied LIME$_{DSP}$ and LIME$_{D+}$ to two data sets to illustrate their utility in analyses of real data. The results show that many of our findings are consistent with those in the literature, but potential novel genes also emerged. Interestingly, for the FHS data, even though 2,332 of the 48,071 SNPs investigated (about 5%) failed the HWE test at the 0.1% level, none needed to be removed in our analysis, because LIME$_{D+}$ is robust to departures from HWE. In fact, four of the SNPs among the top-20 presented in the Supplementary Material, Table S4 (including one with a small p-value of 3×10^{-7}), failed the HWE test, which would not have been studied using traditional methods for detecting an association. We also checked for the familial consistency of geneotypes and did not uncover any problems. For the clubfoot data, a large proportion of the SNPs (over 60%) failed the HWE tests. This is not surprising because the sample is composed of roughly 50% Hispanic and 50% non-Hispanic subjects. Further HWE testing within each of the two subsamples showed that less than 5% of the SNPs failed the test, which is similar to the result from the FHS data. As investigated and discussed in Yang and Lin (2013), the LIME methodology is robust to this type of population stratification, that is, when the sample is a mixture from two subpopulations in which HWE may or may not hold. Therefore, the

results presented in this paper remain valid.

Because proband information is required in our analysis, we investigated the sensitivity of LIME$_{D+}$ against the designations by studying the variability of the outcomes with multiple sets of proband labeling. We considered SNP rs1562705 as an example, using 100 replications to test for imprinting effects. In each replication, a discordant sibpair was chosen randomly as probands from every DSP family and a child was chosen randomly as the proband for each case or control family. From the plot of the $-\log_{10}$ (p-value) versus the replication index (Supplementary Material, Fig. S18), we can see that although there is variation across the 100 replications, the results remain qualitatively the same because the p-values are small (less than 10^{-3}). Thus, the proposed method is robust to the somewhat arbitrary designations of probands, echoing the results from an earlier study (Han, Hu and Lin (2013)), which included only case and control families.

Despite its advantages, LIME$_{DSP}$ has several limitations. A disadvantage of LIME$_{DSP}$ when compared to LIME is that it cannot be applied directly to families when the father's genotype is missing. This is because after we match the affected proband-mother pair with the unaffected proband-mother pair using the child-mother genotype combination, nuisance parameters can no longer be separated from the parameters of interest. Details are provided in Supplementary Material S5. A potential solution is to infer the haplotype frequencies first using information from nearby loci, and then applying LIME$_{DSP}$ based on the imputed data from compatible haplotypes. By weighting the likelihood according to the probabilities of the compatible haplotypes, a preliminary simulation shows that the empirical type I error is close to the nominal value, whereas the power is close to that when using the complete family data (results not shown). However, the HWE assumption is generally needed to infer haplotypes, which leads to bias if the assumption is violated, such as when population stratification exists. Therefore, further study is needed to find a satisfactory solution.

The DSP design addresses a practical difficulty in recruiting control families. As such, design efficiency is not the foremost criterion. Nevertheless, it is important to understand the relative efficiency of these two designs, namesly, DSP versus family case-control, to quantify the information loss in a more practicable design. To this end, we compared the "per individual" information for these two study designs (Supplementary Material S6). The results (Supplementary Material, Fig. S19-S26) show that the family case-control design is typically more powerful, especially in detecting maternal effects. Nevertheless, LIME$_{DSP}$ can be more informative than LIME for estimating some of the parameters, espe-

cially when there is a severe imbalance between the number of case families and the number of control families. This is illustrated by a simulation study; details are provided in Supplementary Material S6. Because control families are more difficult to recruit, LIME$_{DSP}$ is a useful addition to the statistical toolbox for genetic analyses. Most importantly, if data from both types of study designs are available, they should be utilized fully, as demonstrated in our FHS analysis.

Supplementary Materials

The Supplementary Material contains detailed derivations of the probability for a DSP with an arbitrary number of siblings, additional information on the calculation of the probabilities in Table 1, the regularity conditions and proof of Theorem 1, estimations of maternal effects for a DSP design without additional siblings and a DSP design with missing father genotypes, the relative efficiency of LIME$_{DSP}$ vs. LIME, and supplementary tables and figures.

Acknowledgment

The authors would like to thank the two anonymous reviewers for their constructive comments and suggestions. Support from the NSF grant DMS-1208968 and allocations of computing resources from the Ohio Supercomputer Center are also gratefully acknowledged. The clubfoot and the Framingham Heart Study data are downloaded from dbGaP (`https://www.ncbi.nlm.nih.gov/gap`).

References

Al-Qattan, M. M. (2013). Central and ulnar cleft hands: a review of concurrent deformities in a series of 47 patients and their pathogenesis. *The Journal of Hand Surgery, European Volume* **39**, 510–519.

Chanda, K. C. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika* **41**, 56–61.

Chiu, Y., Chung, R., Lee, C., Kao, H., Hou, L. and Hsu, F. (2014). Identification of rare variants for hypertension with incorporation of linkage information. *BMC Proceedings* **8**, S109.

Coba, M. P., Komiyama, N. H., Nithianantharajah, J., Kopanitsa, M. V., Indersmitten, T., Skene, N. G., Tuck, E. J., Fricker, D. G., Elsegood, K. A., Stanford, L. E., Afinowi, N. O., Saksida, L. M., Bussey, T. J., O'Dell, T. J. and Grant, S. G. (2012). TNiK is required for postsynaptic and nuclear signaling pathways and cognitive function. *The Journal of Neuroscience* **32**, 13987–13999.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Elmali, M., Ozmen, Z., Ceyhun, M., Tokatlioglu, O., Incesu, L. and Diren, B. (2014). Joubert

syndrome with atrial septal defect and persistent left superior vena cava. *Diagnostic and Interventional Radiology* **13**, 94–96.

Ferguson-Smith, A. C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. *Nature Reviews Genetics* **12**, 565–575.

Furuhashi, M., Ishimura, S., Ota, H., Hayashi, M., Nishitani, T., Tanaka, M., Yoshida, H., Shimamoto, K., Hotamisligil, G. S. and Miura, T. (2011). Serum fatty acid-binding protein 4 is a predictor of cardiovascular events in end-stage renal disease. *PLoS One* **6**, e27356.

Girisha, K. M., Shukla, A., Trujillano, D., Bhavani, G. S., Kadavigere, R. and Rolfs, A. (2016). A homozygous nonsense variant in IFT52 is associated with a human skeletal ciliopathy. *Clinical Genetics* **90**, 536–539.

Haig, D. (2004). Evolutionary conflicts in pregnancy and calcium metabolism - a review. *Placenta* **25 Suppl A**, S10–S15.

Han, M., Hu, Y.-Q. and Lin, S. (2013). Joint detection of association, imprinting and maternal effects using all children and their parents. *European Journal of Human Genetics* **21**, 1449–1456.

Herzig, M. C., Kolly, C., Persohn, E., Theil, D., Schweizer, T., Hafner, T., Stemmelen, C., Troxler, T. J., Schmid, P., Danner, S., Schnell, C. R., Mueller, M., Kinzel, B., Grevot, A., Bolognani, F., Stirn, M., Kuhn, R. R., Kaupmann, K., van der Putten, P. H., Rovelli, G. and Shimshek, D. R. (2011). Lrrk2 protein levels are determined by kinase function and are crucial for kidney and lung homeostasis in mice. *Human Molecular Genetics* **20**, 4209–4223.

Hirschhorn, J. N. (2009). Genomewide association studies - illuminating biologic pathways. *New England Journal of Medicine* **360**, 1699–1701.

Horvath, A., Boikos, S., Giatzakis, C., Robinson-White, A., Groussin, L., Griffin, K. J., Stein, E., Levine, E., Delimpasi, G., Hsiao, H. P., Keil, M., Heyerdahl, S., Matyakhina, L., Libe, R., Fratticci, A., Kirschner, L. S., Cramer, K., Gaillard, R. C., Bertagna, X., Carney, J. A., Bertherat, J., Bossis, I. and Stratakis, C. A. (2006). A genome-wide scan identifies mutations in the gene encoding phosphodiesterase 11a4 (pde11a) in individuals with adrenocortical hyperplasia. *Nature Genetics* **38**, 794–800.

Horvath, S. and Laird, N. M. (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data. *The American Journal of Human Genetics* **63**, 1886–1897.

Joshi, S., Balan, P. and Gurney, A. M. (2006). Pulmonary vasoconstrictor action of kcnq potassium channel blockers. *Respiratory Research* **7**, 31.

Kohda, T. (2013). Effects of embryonic manipulation and epigenetics. *Journal of Human Genetics* **58**, 416–420.

Li, S., Chen, J., Guo, J., Jing, B.-Y., Tsang, S.-Y. and Xue, H. (2015). Likelihood ratio test for multi-sample mixture model and its application to genetic imprinting. *Journal of the American Statistical Association* **110**, 867–877.

Lim, D. H. and Maher, E. R. (2009). Human imprinting syndromes. *Epigenomics* **1**, 347–369.

Lin, S. (2013). Assessing the effects of imprinting and maternal genotypes on complex genetic traits. In *Lecture Notes in Statistics* 210 (Edited by M. L. T. Lee, M. Gail, R. Pfeier, G. Satten, T. Cai and A. Gandy), 285-300. New York: Springer, Ch. Risk Assessment and Evaluation of Predictions.

Lindsay, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philosophical Transactions of the Royal Society of London A* **296**, 639–

662.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461**, 747–753.

Nguyen, A., Rauch, T. A., Pfeifer, G. P. and Hu, V. W. (2010). Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. *The FASEB Journal* **24**, 3036–3051.

Oksenberg, N., Stevison, L., Wall, J. and Ahituv, N. (2013). Function and regulation of auts2, a gene implicated in autism and human evolution. *PLoS Genetics* **9**, e1003221.

Ota, H., Furuhashi, M., Ishimura, S., Koyama, M., Okazaki, Y., Mita, T., Fuseya, T., Yamashita, T., Tanaka, M., Yoshida, H., Shimamoto, K. and Miura, T. (2012). Elevation of fatty acid-binding protein 4 is predisposed by family history of hypertension and contributes to blood pressure elevation. *American Journal of Hypertension* **25**, 1124–1130.

Palmer, C. G., Mallery, E., Turunen, J. A., Hsieh, H. J., Peltonen, L., Lonnqvist, J., Woodward, J. A. and Sinsheimer, J. S. (2008). Effect of Rhesus D incompatibility on schizophrenia depends on offspring sex. *Schizophrenia Research* **104**, 135–145.

Patten, M. M., Ross, L., Curley, J. P., Queller, D. C., Bonduriansky, R. and Wolf, J. B. (2014). The evolution of genomic imprinting: theories, predictions and empirical tests. *Heredity (Edinb)* **113**, 119–128.

Peacock, J. D., Lu, Y., Koch, M., Kadler, K. E. and Lincoln, J. (2008). Temporal and spatial expression of collagens during murine atrioventricular heart valve development and maintenance. *Developmental Dynamics* **237**, 3051–3058.

Peters, J. (2014). The role of genomic imprinting in biology and disease: an expanding view. *Nature Publishing Group* **15**, 517–530.

Rucker, B., Almeida, M. E., Libermann, T. A., Zerbini, L. F., Wink, M. R. and Sarkis, J. J. (2007). Biochemical characterization of ecto-nucleotide pyrophosphatase/ phosphodiesterase (e-npp, e.c. 3.1.4.1) from rat heart left ventricle. *Molecular and Cellular Biochemistry* **306**, 247–254.

Svensson, A. C., Sandin, S., Cnattingius, S., Reilly, M., Pawitan, Y., Hultman, C. M. and Lichtenstein, P. (2009). Maternal effects for preterm birth: a genetic epidemiologic study of 630,000 families. *American Journal of Epidemiology* **170**, 1365–1372.

Wang, M. and Lin, S. (2014). Famlbl: detecting rare haplotype disease association based on common snps using case-parent triads. *Bioinformatics* **30**, 2611–2618.

Weinberg, C. R., Wilcox, A. J. and Lie, R. T. (1998). A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *The American Journal of Human Genetics* **62**, 969–978.

Weir, B. (1996), Genetic Data Analysis II. Methods for Discrete Population Genetic Data, Sinauer Associates, Inc. Publishers.

Wilkinson, L. S., Davies, W. and Isles, A. R. (2002). Genomic imprinting effects on brain development and function. *Nature Publishing Group* **8**, 832.

Yang, J. and Lin, S. (2013). Robust partial likelihood approach for detecting imprinting and maternal effects using case-control families. *The Annals of Applied Statistics* **7**, 249–268.

Zandi, P. P., Kalaydjian, A., Avramopoulos, D., Shao, H., Fallin, M. D. and Newschaffer, C. J. (2006). Rh and ABO maternal - fetal incompatibility and risk of autism. *American Journal of Medical Genetics B* **141**, 643–647.

Zhang, F., Khalili, A. and Lin, S. (2016). Optimum study design for detecting imprinting and maternal effects based on partial likelihood. *Biometrics* **72**, 95–105.

Zhang, T.-X., Haller, G., Lin, P., Alvarado, D. M., Hecht, J. T., Blanton, S. H., Stephens Richards, B., Rice, J. P., Dobbs, M. B. and Gurnett, C. A. (2014). Genome-wide association study identifies new disease loci for isolated clubfoot. *Journal of Medical Genetics* **51**, 334–339.

Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA.

E-mail: fangyuan.zhang@ttu.edu

Department of Mathematics and Statistics, McGill University, Montreal, QC H3A 0B9, Canada.

E-mail: abbas.khalili@mcgill.ca

Department of Statistics, Ohio State University, Columbus, OH 43210, USA.

E-mail: shili@stat.osu.edu

(Received March 2016; accepted January 2018)