Statistica Sinica

Imprinting and Maternal Effect Detection Using Partial Likelihood Based on Discordant Sibship Data

Journal:	*Statistica Sinica*
Manuscript ID	SS-2016-0114
Manuscript Type:	Original Article
Date Submitted by the Author:	08-Mar-2016
Complete List of Authors:	Lin, Shili; Ohio State University, Statistics Zhang, Fangyuan; Texas Tech University
Keywords:	Ascertainment, Partial likelihood, Discordant Sib-Pair design

Imprinting and Maternal Effect Detection Using Partial Likelihood Based on Discordant Sibship Data

Fangyuan Zhang[1] and Shili Lin[2*]

[1]Department of Mathematics and Statistics, Texas Tech University

[2]Department of Statistics, Ohio State University, Columbus, OH 43210, USA

[*]Corresponding author email: shili@stat.osu.edu

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. However, most of them either only model one of these two confounded epigenetic effects, make strong yet unrealistic assumptions about the population to avoid over-parameterization, or are only applicable to study designs that require recruitment of difficult-to-obtain control families. In this paper, we develop a partial Likelihood method for detecting Imprinting and Maternal Effects for a Discordant Sib-Pair 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. Theoretical analysis shows that the partial maximum likelihood estimators based on LIME$_{DSP}$ are consistent and asymptotically normally distributed. Based on the close-form formula for computing information, we compared a study design with more independent families versus one with larger families by keeping the total number of individuals needed to be genotyped fixed. We further carried out a simulation study, which demonstrates the robust property of LIME$_{DSP}$ and shows that it is a powerful approach without resolving to collecting control families. To illustrate its practical utility, LIME$_{DSP}$ was applied to a clubfoot dataset and the Framingham Heart Study.

Keywords: Ascertainment; Association study; Discordant Sib-Pair design; Imprinting effect; Maternal effect; Partial likelihood

1 INTRODUCTION

Genome-wide association studies (GWAS) represent a powerful tool in identifying common genetic variants that are associated with complex human traits, and have provided valuable insights into the genetic architecture of such traits. However, the variants identified have explained only a small proportion of the variability in most complex traits, leading to concerns about "missing heritability" (Manolio *et al.* 2009). In an effort to understand this missing heritability, it is realized that, since gene expression is a dynamic process, DNA sequence polymorphism is not the only contributing factor to phenotypic variation. Other mechanisms may also be involved, such as epigenetic modification and transcriptional/translational regulation (Hirschhorn 2009; Peters 2014). Therefore, epigenetic factors, including imprinting and maternal genotype effects, which were largely ignored before, have been brought to the attention and become a research focus in the hunt for missing heritability (Kohda 2013).

Genomic imprinting is an epigenetic factor involving methylation and histone modifications that completely or partially silences the expression of a gene inherited from a particular parent without altering the genetic sequence (Patten *et al.* 2014). It can lead to a parent-of-origin pattern in gene expressions, i.e. 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). Imprinting effect is hailed as a key factor in understanding the interplay between the epigenome and genome (Ferguson-Smith 2011). On the other hand, maternal genotype effect, as another epigenetic effect, can also lead to parent-of-origin pattern. Maternal genotype effect refers to the phenomenon that 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 the 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 *et al.* 2002; Peters 2014). However, deregulation of imprinted genes have been found to contribute to a number of complex human diseases. The most well-known examples are Beckwith-Wiedemann Syndrome, Silver-Russell Syndrome, Angelman Syndrome, and Prader-Willi Syndrome (Lim and Maher 2009). Meanwhile, It has been well established that for a variety of diseases, especially those that are related to pregnancy outcomes, such as childhood cancers and birth defects (Haig 2004), certain psychiatric illness (Palmer *et al.* 2008), and pregnancy complications (Svensson *et al.* 2009), maternal effects play an important role. However, to date, due to limited data availability and insufficient power of methods, only very few genes have been detected to have genomic imprinting or maternal effects.

As both imprinting and maternal effects exhibit parent-of-origin patterns, family data are needed to trace the inheritance path. One common study design is case-parent triads, which may also include control-parent triads. Based on such a design, numerous methods have been proposed to model imprinting and maternal effect simultaneously to avoid potential confounding (see (Lin 2013) and references therein). However, almost all of them rely on strong yet unrealistic assumptions about the population, such as mating symmetry, to avoid over-parameterization, with the Logarithm Likelihood Ratio Test (LL-LRT) as a classic example (Weinberg *et al.* 1998). The exception is the recently proposed partial Likelihood method for detecting Imprinting and Maternal Effects (LIME), which alleviates the need to make the unrealistic assumptions (Yang and Lin 2013). However, the price to pay is the need for separate control families, which are typically much more difficult to recruit compared to recruitment of individual controls in a population design. Most recently, a mixture modeling approach was proposed for detecting imprinting, but we note that the data type was gene expression from a population sample (Li *et al.* 2015), which differs from our family-based design.

To reap the benefit of LIME but without the requirement of control families, in this paper, we propose a LIME method based on a Discordant Sib-Pair design (LIME$_{DSP}$). It borrows the idea from Yang and Lin (2013), but considers an alternative study design in which a nuclear family is recruited if there is a discordant sibpair, i.e., one sibling is affected and the other is unaffected. Data from additional siblings (affected or not) may also be incorporated to further increase power. The idea of LIME$_{DSP}$ is to match affected proband-parent triads with unaffected proband-parent triads and factor out common terms involving mating type probabilities, the nuisance parameters. As such LIME$_{DSP}$ circumvents the problem of over parameterization, unrealistic assumptions, and the need for control families. However, when control families are available, they can be utilized as well to further increase statistical power. Finally, we note that, discordant sibpair design is popular in linkage and association studies (Horvath and Laird 1998), which provide an outlet 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 code for disease susceptibility or epigenetic effect. In a nuclear family, let F and M be the random variables denoting the number of A alleles carried by father and mother respectively, which can take values 0, 1, or 2, corresponding to genotype BB, AB or AA, respectively. Similarly, let C_i be the random variable denoting the number of A alleles, that is, the genotype of child i, $i = 1, 2, \cdots$. Specifically, C_1 and C_2 are designated for the affected and unaffected probands, respectively, through which the family is recruited, whereas $C_i, i = 3, \cdots$, are for the additional siblings, if any. D_i, $i = 1, 2, \cdots$, denote disease status of children (1 - affected; 0 - normal). Thus, $D_1 = 1$ and $D_2 = 0$. The development of LIME$_{DSP}$ is based on a multiplicative relative risk model for disease prevalence for a triad family:

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

where r_1 and r_2 denote the effect of one or two copies of an individual's own variant allele, r_{im} denotes imprinting effect, s_1 and s_2 denote the effect of one or two copies of the mother's variant allele, and δ is the phenocopy rate. The notation $c = 1_m$ denotes that the child's genotype is AB, where variant allele A is from mother. We are interested in the estimation of 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 the parameters are positive. Further, $r_{im} > 1, < 1, = 1$ signify paternal, maternal, or no imprinting effect, 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

As the ascertainment criterion is discordant sibpair, 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):

$$P(M = m, F = f, C_1 = c_1, C_2 = c_2, C_i = c_i, D_i = d_i, i = 3, \cdots \mid D_1 = 1, D_2 = 0)$$

$$= P(M = m, F = f, C_1 = c_1|D_1 = 1, D_2 = 0)P(M = m, F = f, C_2 = c_2|D_1 = 1, D_2 = 0) \quad (2)$$

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

$$\times \frac{P(D_1 = 1, D_2 = 0)}{P(M = m, F = f)P(D_1 = 1|M = m, F = f)P(D_2 = 0|M = m, F = f)}. \quad (4)$$

On the right hand side of the above formula, we write the probability of the observed data as the product of three parts: triad probability (mother, father and child) conditioned on proband disease status (2), additional sibling genotype and phenotype joint probability given parents' genotypes (3), and the remaining part (4). The first part (2) containing the probands can be thought of as obtained from a "retrospective" design, whereas the second part (3) for the additional siblings are treated as from a "prospective" design. While the prospective part is straight-forward, involving parameters of interest only, as can be seen from model (1), the retrospective part is much more intricate and will be closely examined in the following subsection.

We first note that, in (2),

$$P(M=m,F=f,C_1=c_1|D_1=1,D_2=0)=\frac{P(M=m,F=f,C_1=c_1,D_1=1,D_2=0)}{P(D_1=1,D_2=0)}. \quad (5)$$

There are 15 possible combinations of genotypes for parents (M, F) and a child (C) in total; their enumeration and labeling (type) are listed in Table 1, with the corresponding probability for the numerator in (5) listed in the 5^{th} column. Similarly, the probability $P(M=m,F=f,C_2=c_2,D_1=1,D_2=0)$ are given in the last column of the table. Derivations of the probabilities for a few of the cases are given in the Supplementary Material S1 as examples. In the expressions in Table 1, the μ_{mf}'s $(m=0,1,2,\ f=0,1,2)$ are the mating type probabilities, that is, $\mu_{mf}=P(M=m,F=f)$. Note that we do not make any assumption about the mating type probabilities such as Hardy-Weinberg Equilibrium (HWE) or even mating symmetry, and thus μ_{mf} is not necessarily equal to μ_{fm}. As can be seen from the table, these nuisance parameters can be factored out completely from the 6 model parameters. This observation forms the basis of the partial likelihood formulation.

2.3 Organization of Data

It can be seen from Table 1 that conditional on each possible triad genotype vector (m,f,c), the count of the affected proband-parent triads and that of 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 will be free of nuisance parameters. For example, among all proband-parent triads with genotype type combination being (m,f,c), the probability of observing an affected proband-parent triad is

$$\begin{aligned}
p_{mfc} &= \frac{NP(m,f,C_1=c|D_1=1,D_2=0)}{NP(m,f,C_1=c|D_1=1,D_2=0)+NP(m,f,C_2=c|D_1=1,D_2=0)}\\
&= \frac{P(m,f,C_1=c,D_1=1,D_2=0)}{P(m,f,C_1=c,D_1=1,D_2=0)+P(m,f,C_2=c,D_1=1,D_2=0)}\\
&= \frac{P(D=1|m,f,c)P(D=0|m,f)}{P(D=1|m,f,c)P(D=0|m,f)+P(D=0|m,f,c)P(D=1|m,f)}, \quad (6)
\end{aligned}$$

where only parameters in (1) are involved. This manipulation turns data from a retrospective design into a “prospective” one through stratifying according to each triad genotype combination. We denote the denominator of (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 whole likelihood, we can extract out 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 with genotype $M = m$, $F = f$, and $C = c$, respectively. 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 with genotype combination $M = m$, $F = f$ and $C = c$, respectively. Recall that we denote the vector of 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 mating type probabilities) by $\boldsymbol{\phi}$. Then according to the three component factorization,

$$\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\phi}) \quad = \quad & \prod_{m,f,c} [P(m, f, C_1 = c|D_1 = 1, D_2 = 0)]^{n^1_{mfc}} [P(m, f, C_2 = c|D_1 = 1, D_2 = 0)]^{n^0_{mfc}} \\
& \times \prod_{m,f,c} [P(c|m, f)]^{sn^1_{mfc}+sn^0_{mfc}} [P(D = 1|m, f, c)]^{sn^1_{mfc}} [P(D = 0|m, f, c)]^{sn^0_{mfc}} \\
& \times \prod_{m,f,c} \left[\frac{P(D_1 = 1, D_2 = 0)}{P(m, f)P(D_2 = 0|m, f)P(D_1 = 1|m, f)} \right]^{n^1_{mfc}} \\
\propto \quad & \prod_{m,f,c} p^{n^1_{mfc}}_{mfc} (1 - p_{mfc})^{n^0_{mfc}} \prod_{m,f,c} q^{sn^1_{mfc}}_{mfc} (1 - q_{mfc})^{sn^0_{mfc}} \qquad (7) \\
\times \quad & \prod_{m,f,c} S^{n^1_{mfc}+n^0_{mfc}}_{mfc} \left[\frac{P(D_1 = 1, D_2 = 0)}{P(m, f)P(D_2 = 0|m, f)P(D_1 = 1|m, f)} \right]^{n^1_{mfc}}, \qquad (8)
\end{aligned}$$

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

We note that, all the nuisance parameters in $\boldsymbol{\phi}$ are only present in (8), while the factors in (7) contain only parameters in $\boldsymbol{\theta}$ and is therefore taken as our partial likelihood. The parameters in $\boldsymbol{\theta}$ can be inferred through maximizing the partial likelihood instead of the full likelihood to avoid estimating the nuisance parameters (Cox 1975). In fact, the first factor of 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, 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. As the affection statuses of the additional siblings are obtained prospectively, the probability of observing affected sibling-parent triads with certain 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:

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

By solving the score-type equation

$$\frac{\partial l_{par}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = l'_{par}(\boldsymbol{\theta}) = \mathbf{0}, \tag{9}$$

the *maximum partial likelihood estimator* (MPLE) of $\boldsymbol{\theta}$ can be obtained.

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,

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

As one can see from the partial likelihood, these four types of trios contribute independent information conditioned on the genotype of the parents. Thus, n is regarded as the effective sample size. We study the asymptotic properties of the *maximum partial likelihood estimator* (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 S2, we have:*

(i) The likelihood equation has an 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 and is 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 $0 \leq B_{mfc} < 1$ *and* $0 \leq C_{mfc} < 1$ *are the limits in probability of* $\{\frac{n^1_{mfc}+n^0_{mfc}}{n}\}$, $\{\frac{sn^1_{mfc}+sn^0_{mfc}}{n}\}$*, respectively, when* $n \to \infty$.

The proof of the theorem can be found in Supplementary Material S2. Note that although the consistent solution of partial likelihood score equation (9) is unique (Chanda 1954; Lindsay 1980), there may exist inconsistent roots.

2.5 Combining Data From the Two Study Designs

In 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, we assume that 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 the available data is then $pL(\boldsymbol{\theta}) = pL_A(\boldsymbol{\theta}) * pL_B(\boldsymbol{\theta})$ as data in sets A and B are independent. Note that if both studies are concerned about the the same underlying disease model, then the parameter of interests are identical, as assumed in the LIME$_{D+}$ procedure. The model parameters in $\boldsymbol{\theta}$ are estimated by maximizing the partial likelihood $pL(\boldsymbol{\theta})$. The MPLE of LIME$_{D+}$ enjoys the same asymptotic properties as LIME$_{DSP}$.

3 EVALUATION of INFORMATION CONTENTS

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 a total of 8 disease models (Table 2). The first three models portrait no imprinting nor maternal effects. Models 4 has maternal effect only, models 5 and 6 have imprinting effect only, and model 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 Hardy-Weinberg equilibrium (HWE) {not hold = 0, hold = 1}. Suppose p is the MAF, then the probabilities of a genotype taking values of 0, 1 and 2 are $(1-p)^2(1-\zeta) + (1-p)\zeta$, $2p(1-p)(1-\zeta)$, and $p^2(1-\zeta) + p\zeta$, where ζ is the inbreeding parameter (Weir, 1996). When HWE holds, $\zeta = 0$. When HWE does not hold, ζ is set to be 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 (1) is fully specified. As the summation over the 15 joint probabilities $P(D = 1, M, F, C)$ equals the disease prevalence $P(D = 1)$, the phenocopy rate can be solved from the equation.

It is intuitive to understand that including additional siblings to a DSP design will typically increase the information for estimating model parameters and hence detection power for a fixed sample of N families, which in fact is demonstrated through theoretical calculation of “per family” information content (Supplementary Fig. S1). However, additional siblings will lead to a larger number of total individuals, hence greater genotyping and phenotyping cost, 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 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 8 disease models. Plots for other scenarios are given in the Supplementary Fig. S2-4. It is not surprising to see from the figures that there is essentially no information for inference on 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 very limited contrast for maternal effect. The theoretical explanation can be found in Supplementary Material S3. Fortunately, when additional siblings are available, maternal effects can be estimated. For the other parameters r_1, r_2 and r_{im}, which design is more efficient depends on the disease prevalence. When disease prevalence is high (0.15), recruiting additional siblings, which are likely to include affected ones given the common disease, will increase the efficiency. On the other hand, when disease prevalence is low (0.05), recruiting more independent families or more siblings leads to fairly similar results (apart from for estimating the maternal effects), although larger number of independent families has a slight edge for estimating the other parameters. Thus, depending on the disease prevalence and the parameters one is more interested in, the most efficient design may be different.

4 SIMULATION

With a good 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 through simulation for a typical sample size in genetic epidemiology. We consider the D, $D+1$ and $D+2$ designs, each with 300 families. All combinations of the eight disease models and 8 population scenarios are entertained, leading to 192 $(3 \times 8 \times 8)$ simulation settings, with 1000 simulated data sets under each setting.

Figure 2 shows empirical type I error rate and power of LIME$_{DSP}$ under all 8 disease models and scenario 1. The three rows represent the three designs considered. The three bars refer to association, imprinting effect, and maternal effect, respectively, in that order. The results show that the type I error rates are close to the nominal value 0.05, marked by a horizontal dashed line for association under model 1, imprinting effect under models 1, 2, 3, 4, and maternal effect under models 1, 2, 3, 5, 6, across all three designs. Note that when there are no additional siblings, the D design, the type I error rate for maternal effect is rather low, not surprisingly as we discussed earlier since such data provide no information on inferring maternal effect. Comparing across the three designs, we can see that power increases as more additional siblings are recruited, especially for detecting maternal effect. LIME$_{DSP}$ is incapable of detecting maternal effect when there are only discordant sibpairs, but the power increases when additional siblings are also available. The results for the other seven scenarios are similar and are shown in the Supplementary 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, whose genetic bases have been established, club foot and Framingham Heart Study (FHS). Both studies are family based, involving extended pedigrees. In the club foot data, we extracted out nuclear families with discordant sibpairs and additional siblings, if available. Thus LIME$_{DSP}$ is applicable to the data. For the FHS, we extracted out nuclear families that have discordant sibpairs or are case-parent, control-parent triads, all potentially involving additional siblings, and are analyzed using LIME$_{D+}$.

5.1 Analysis of the Club Foot Data

Club foot is a congenital deformity in which the affected foot appears to have been rotated internally at the ankle. With treatment, the vast majority of patients recover completely during early childhood and are able to walk and participate in athletics. Thus, understanding the underlying causal mechanism is important in aiding the development of effective treatment strategies. Our LIME$_{DSP}$ analysis makes use of 87 discordant sibpairs with 33 additional siblings. They range from discordant sibpairs without additional siblings to with 6 siblings. The data are obtained from dbGaP (www.ncbi.nlm.nih.gov/gap/).

Among the top SNPs (with the smallest p-values) identified by LIME$_{DSP}$ (Table 3), some reside within genes that have been implicated in the literature, either for symptoms directly related to clubfoot or for other congenital diseases. For example, two SNPs (rs11048527 and rs6785520) that are found to have 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 include club foot (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 association between the gene and clubfoot is less than 0.001. As another example, one of the top SNPs (rs9446305) with some evidence for maternal effect is in gene B3GAT2, whose association 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 relatively small p-value for detecting maternal effect. It has been found that deletion of exon 6 of the AUTS2 gene can cause congenital disorders, including eversion of the feet. It is interesting to point out 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 effect (Zandi *et al.*

2006), consistent with our finding.

LIME$_{DSP}$ also identified some other genes that have been reported to be associated with other complex developmental traits in the literature. For example, RORA is related to autism (Nguyen *et al.* 2010), whereas TNIK and FARP1 are related to fetal brain outgrowth and development (Coba *et al.* 2012). In a most recent study, gene IFT52 is linked to skeletal ciliopathy, whose manifestations 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 association, imprinting, and maternal effect can be found in Supplementary Tables S1-3. Given the large number of SNPs investigated, some of the SNPs identified may not be genome-wide significant. A complete results of all the SNPs analyzed are provided as Supplementary Fig. S12-14.

5.2 Analysis of the Framingham Heart Study Data

Framingham Heart Study (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 has taken medication to control blood pressure. In this analysis, we focus on 263 DSP families (with 229 additional siblings) and 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 to be associated with the hypertensive trait by LIME$_{D+}$ (top segment of Table 4) have been previously implicated in the literature to be 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 Meckel and Joubert syndromes, conditions that may be related to atrial septal defect (Elmali *et al.* 2014). Also, rs2229188 is another SNP identified to be 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 also worth discussing. Previous research suggests that FABP4 level, being 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 atherogenic metabolic phenotype, and the elevation of 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). It is also found that another gene, LRP1B, is important in the development of atherosclerosis, a disease that affects the arterial blood vessel (www.scbt.com/datasheet-49230-lrp1b-n-19-antibody.html). On the other hand, gene KCNQ3 in Chromosome 8, together with other KCNQ channels, are believed to play a functional role in pulmonary artery smooth muscle (Joshi *et al.* 2006).

Finally, four genes harboring multiple SNPs that are among the top ones for maternal effect (last segment of Table 4) have also been discussed in the literature previously. In particular, Gene CHCHD6 has been identified to have 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 cardiac system, which suggests that these enzymes could contribute with the fine-tuning control of the nucleotide levels at the nerve terminal endings of left ventricles that are involved in several cardiac pathologies. As another example, gene PDE11A is associated with the development of adrenocortical hyperplasia leading to Cushing syndrome (Horvath *et al.* 2006), while Cushing syndrome has clinical manifestations of arterial hypertension. Finally, gene LRRK2 is also implicated in a previous study, as LRRK2 mutant mice can cause blood pressure changes (Herzig *et al.* 2011). A list of the top-20 SNPs (with the smallest p-values) identified by LIME$_{D+}$ for each of associatoon, imprinting, and maternal effect can be found in Supplementary Tables S4-6. As with the clubfoot study, some of the SNPs identified may not reach genome-wide significance. A complete results of all the SNPs analyzed are provided as Supplementary Fig. S15-17.

6 DISCUSSION

Imprinting and maternal effects are two confounding epigenetic factors that have been increasingly 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 the need to make unrealistic assumptions or to require the collection of separate control families. Based on the asymptotic property of LIME and the close-form formula for calculating information, our work provides 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 analysis of real data. The results from these analyses show that many of our findings are consistent with those in the literature, but potential novel genes also emerged. It is interesting to note that, for the FHS data, even though 2332 of the 48071 SNPs investigated (about 5%) failed the HWE test at the 0.1% level, none need to be removed for our analysis as LIME$_{D+}$ is robust to departure from HWE. In fact, four of the SNPs among the top-20 presented in Supplementary Table S4 (including one with a small p-value of 3×10^{-7}) failed the HWE test, which would not have been studied using a traditonal methods for detecting association. We have also checked for familial consistency of geneotypes and did not find any problem. For the club foot data, a very large proportion of the SNPs (over 60%) failed the HWE tests. This is not surprising as 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 SNPs failed the test, similar to the result from the FHS data. As investigated and discussed in Yang and Lin (2013), the LIME methodology is in fact 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 within each. Therefore, the results presented in this paper remain valid.

Despite the advantages, LIME$_{DSP}$ has its own limitations. One disadvantage of LIME$_{DSP}$ when compared to LIME, is that it cannot be directly applied to families with father's genotype missing. This is because after we match affected proband-mother pair with unaffected proband-mother pair by the child-mother genotype combination, nuisance parameters can no longer be separated from the parameters of interest. Details are provided in Supplementary Material S4. One potential solution is to infer haplotype frequencies first by utilizing information from nearby loci, and then apply LIME$_{DSP}$ based on imputed data from compatible haplotypes. By weighting the likelihood according to the probabilities of the compatible haplotypes, preliminary simulation shows that the empirical type I error is close to the nominal ones, while the power is close to using complete family data (results not shown). However, HWE assumption is generally needed to infer haplotype, which will lead to bias, if such an assumption is violated, such as when there is population stratification. Further study is therefore needed to find a satisfactory solution.

The DSP design is to address 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, DSP versus family case-control, to quantify information loss with the more practicable design. To this end, we compared the "per individual" information for these two study designs (Supplementary Material S5). Indeed, the results (Supplementary Fig. S18- S25) show that the family case-control design is typically more powerful, especially in detecting maternal effect, not surprisingly as discussed earlier. Nevertheless, LIME$_{DSP}$ can in fact be more informative than LIME for estimating some of the parameters, especially when there is a severe imbalance between the numbers of case families and the number of control families. Regardless, since control families are much harder to recruit, LIME$_{DSP}$ is an useful addition to the statistical toolbox for genetic analysis. Most importantly, if data from both types of study designs are available, they should be utilized fully as we demonstrated in the FHS analysis.

SUPPLEMENTARY MATERIALS

This supplementary document contains additional information on calculation of probabilities in Table 1, regularity conditions and proof of Theorem 1, estimation of maternal effect with DSP design without additional siblings, DSP design with missing father genotypes, relative efficiency of LIME$_{DSP}$ vs. LIME, and supplementary tables and figures.

Acknowledgements

The authors gratefully acknowledge the NSF grant DMS-1208968 and allocations of computing resources from the Ohio Supercomputer Center.

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, 13987–99.

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, 663–663.

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”, , .

Haig, D. (2004), “Evolutionary Conflicts in Pregnancy and Calcium Metabolism - A Review”, *Placenta*, 25 Suppl A, S10–5.

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, 1886–897.

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, 416–20.

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, 347–69.

Lin, S. (2013), "Assessing the Effects of Imprinting and Maternal Genotypes on Complex Genetic Traits.", in *Lecture Notes in Statistics*, edited by M.-L. T. Lee, M. Gail, R. Pfeiffer, G. Satten, T. Cai, and A. Gandy, volume 210, chapter Risk Assessment and Evaluation of Predictions, 285–300, Springer: New York.

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, 747–53.

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, 3036–51.

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, 135–45.

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”, *Molec- ular 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, 1365–72.

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-ParentCTriad 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, 969–78.

Wilkinson, L. S., Davies, W., and Isles, A. R. (2002), “Gnomic 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, 643–7.

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, 334–9.

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	$\mu_{01}\frac{1}{2}\delta\frac{1}{2}(2-\delta-\delta r_1)$
3	0	1	1	$\mu_{01}\frac{1}{2}\delta r_1\frac{1}{2}(2-\delta-\delta r_1)$
4	0	2	1	$\mu_{02}\delta r_1(1-\delta r_1)$
5	1	0	0	$\mu_{10}\frac{1}{2}s_1\delta\frac{1}{2}(2-\delta s_1-\delta r_1 r_{im}s_1)$
6	1	0	1	$\mu_{10}\frac{1}{2}\delta r_1 r_{im}s_1\frac{1}{2}(2-\delta s_1-\delta r_1 r_{im}s_1)$
7	1	1	0	$\mu_{11}\frac{1}{4}\delta s_1\frac{1}{4}(4-\delta s_1-\delta s_1 r_1-\delta s_1 r_1 r_{im}-\delta r_2 s_1)$
8	1	1	1	$\mu_{11}\frac{1}{4}\delta s_1 r_1(1+r_{im})\frac{1}{4}(4-\delta s_1-\delta s_1 r_1-\delta s_1 r_1 r_{im}-\delta r_2 s_1)$
9	1	1	2	$\mu_{11}\frac{1}{4}\delta s_1 r_2\frac{1}{4}(4-\delta s_1-\delta s_1 r_1-\delta s_1 r_1 r_{im}-\delta r_2 s_1)$
10	1	2	1	$\mu_{12}\frac{1}{2}\delta r_1 s_1\frac{1}{2}(2-\delta r_1 s_1-\delta r_2 s_1)$
11	1	2	2	$\mu_{12}\frac{1}{2}\delta r_2 s_1\frac{1}{2}(2-\delta r_1 s_1-\delta r_2 s_1)$
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	$\mu_{21}\frac{1}{2}\delta r_1 s_2 r_{im}\frac{1}{2}(2-\delta r_1 s_2 r_{im}-\delta r_2 s_2)$
14	2	1	2	$\mu_{21}\frac{1}{2}\delta r_2 s_2\frac{1}{2}(2-\delta r_1 s_2 r_{im}-\delta r_2 s_2)$
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	$\mu_{01}\frac{1}{2}(1-\delta)\frac{1}{2}\delta(1+r_1)$
3	0	1	1	$\mu_{01}\frac{1}{2}(1-\delta r_1)\frac{1}{2}\delta(1+r_1)$
4	0	2	1	$\mu_{02}\delta r_1(1-\delta r_1)$
5	1	0	0	$\mu_{10}\frac{1}{2}(1-\delta s_1)\frac{1}{2}\delta s_1(1+r_1 r_{im})$
6	1	0	1	$\mu_{10}\frac{1}{2}(1-\delta r_1 r_{im}s_1)\frac{1}{2}s_1\delta(1+r_1 r_{im})$
7	1	1	0	$\mu_{11}\frac{1}{4}(1-\delta s_1)\frac{1}{4}\delta s_1(1+r_1+r_1 r_{im}+r_2)$
8	1	1	1	$\mu_{11}\frac{1}{4}(2-\delta s_1 r_1(1+r_{im}))\frac{1}{4}\delta s_1(1+r_1+r_1 r_{im}+r_2)$
9	1	1	2	$\mu_{11}\frac{1}{4}(1-\delta s_1 r_2)\frac{1}{4}\delta s_1(1+r_1+r_1 r_{im}+r_2)$
10	1	2	1	$\mu_{12}\frac{1}{2}(1-\delta r_1 s_1)\frac{1}{2}\delta s_1(r_1+r_2)$
11	1	2	2	$\mu_{12}\frac{1}{2}(1-\delta r_2 s_1)\frac{1}{2}\delta s_1(r_1+r_2)$
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	$\mu_{21}\frac{1}{2}(1-\delta r_1 s_2 r_{im})\frac{1}{2}\delta s_2(r_1 r_{im}+r_2)$
14	2	1	2	$\mu_{21}\frac{1}{2}(1-\delta r_2 s_2)\frac{1}{2}\delta s_2(r_1 r_{im}+r_2)$
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 mother, father and child in a triad, which 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). 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 variant allele.

Table 2. Eight disease models represented by relative risks 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]Notations for the model parameters are the same as 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 (1).

Table 3. Top SNPs for associtition, imprinting, and maternal effects for the club foot date 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

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

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).

Figure Legends

Figure 1: Information content per individual for 8 disease models and two PREVs when HWE holds and MAF is 0.3. Each curve depicts the information for estimating one of the 5 parameters, for data types D, $D+1$ and $D+2$.

Figure 2: Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 1 as given in Table 2. Three rows represent three data types: D, $D+1$ and $D+2$. The three bars refer to association, imprinting effect and maternal effect, respectively, in that order. The horizontal line marks the nominal a level of 0.05.

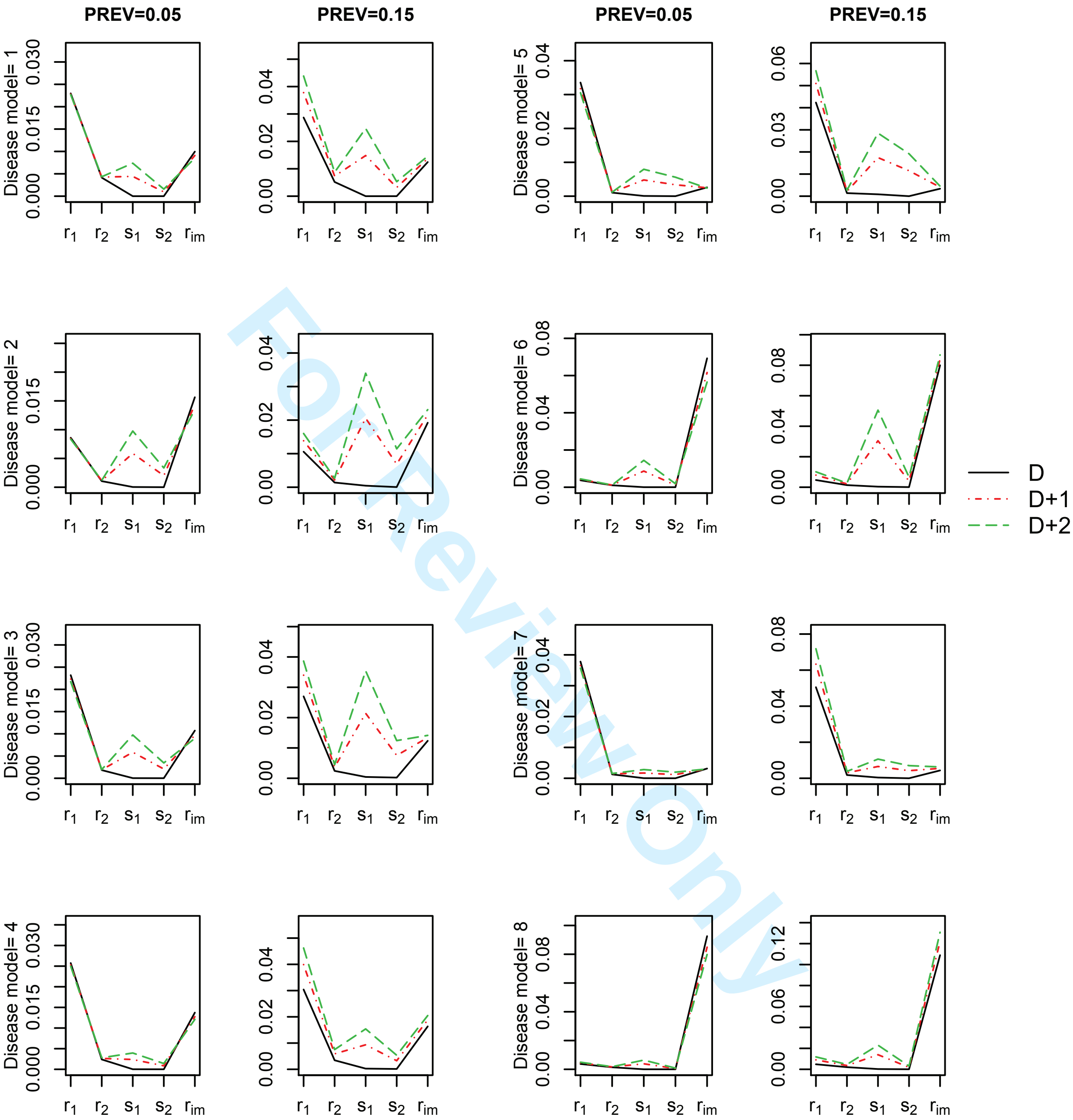


Figure 1. Information content per individual for 8 disease models and two PREVs when HWE holds and MAF is 0.3. Each curve depicts the information for estimating one of the 5 parameters, for data types D, $D + 1$ and $D + 2$.

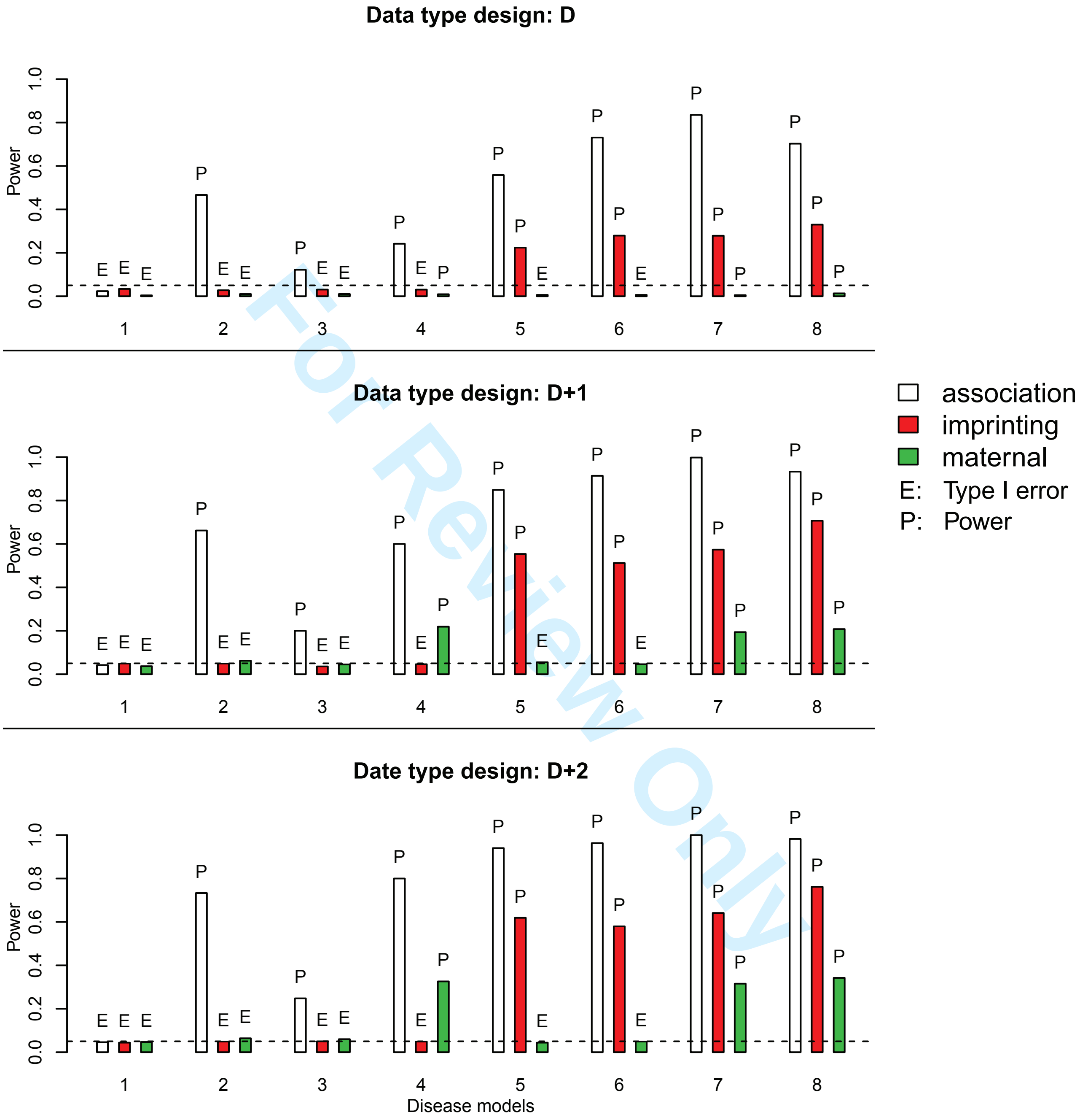


Figure 2. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 1 as given in Table 2. Three rows represent three data types: D, $D+1$ and $D+2$. The three bars refer to association, imprinting effect and maternal effect, respectively, in that order. The horizontal line marks the nominal a level of 0.05.

Supplementary Material for“Imprinting and Maternal Effect Detection Using Partial Likelihood Based on Discordant Sibship Data”

Fangyuan Zhang and Shili Lin

March 7, 2016

S1. Calculation of Probabilities in Table 1.

Consider a candidate genetic marker with two alleles A and B, where A is the allele of interest, the variant allele, which may code for disease susceptibility or epigenetic effect. In a nuclear family, let F and M be the random variables denoting the number of A alleles carried by father and mother respectively, which can take values 0, 1, or 2, corresponding to genotype BB, AB or AA, respectively. Similarly, let C_i be the random variable denoting the number of A alleles, that is, the genotype of child i, $i = 1, 2, \cdots$. Specifically, C_1 and C_2 are designated for the affected and unaffected probands, respectively, through which the family is recruited, whereas $C_i, i = 3, \cdots$, are for the additional siblings, if any. $D_i,\ i = 1, 2, \cdots$, denote disease status of children (1 - affected; 0 - normal). Thus, $D_1 = 1$ and $D_2 = 0$.

In table 1, the formulas to calculate the joint probabilities are as follows:

$$\begin{aligned}
&P(M=m, F=f, C_1=c, D_1=1, D_2=0)\\
&= P(M=m, F=f)P(C_1=c|M=m, F=f)\\
&\times P(D_1=1|M=m, F=f, C_1=c)P(D_2=0|M=m, F=f), and
\end{aligned}$$

$$\begin{aligned}
&P(M=m, F=f, C_2=c, D_1=1, D_2=0)\\
&= P(M=m, F=f)P(C_2=c|M=m, F=f)\\
&\times P(D_2=0|M=m, F=f, C_2=c)P(D_1=1|M=m, F=f).
\end{aligned}$$

For all types other than type 8 (Table 1), if a child has one copy of the variant allele, the parental origin can be unambiguously identified, and hence the joint probability can be easily obtained by extracting the relevant factors from the relative risk model for disease prevalence.

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

where the parameters: r_1 and r_2 denote the effect of one or two copies of an individual's own variant allele, r_{im} denotes imprinting effect, s_1 and s_2 denote the effect of one or two copies of the mother's variant allele, and δ is the phenocopy rate. The notation $c = 1_m$ denotes that the child's genotype is heterozygous, where the variant allele is from mother. The indicator variable D denotes the disease status of a child (1 - affected; 0 - normal).We use μ_{mf}'s ($m = 0, 1, 2,\ f = 0, 1, 2$) to denote the mating type probabilities.

For example, in the familial genotype combination $(m, f, c) = (2, 0, 1)$,

$$\begin{aligned}
&P(M = 2, F = 0, C_1 = 1, D_1 = 1, D_2 = 0)\\
&= P(M = 2, F = 0)P(C_1 = 1|M = 2, F = 0)\\
&\times P(D_1 = 1|M = 2, F = 0, C_1 = 1)P(D_2 = 0|M = 2, F = 0)\\
&= \mu_{20}\delta r_1 s_2 r_{im}(1 - \delta r_1 s_2 r_{im}),
\end{aligned}$$

and

$$\begin{aligned}
&P(M = 2, F = 0, C_2 = 1, D_1 = 1, D_2 = 0)\\
&= P(M = 2, F = 0)P(C_2 = 1|M = 2, F = 0)\\
&\times P(D_2 = 0|M = 2, F = 0, C_2 = 1)P(D_1 = 1|M = 2, F = 0)\\
&= \mu_{20}(1 - \delta r_1 s_2 r_{im})\delta r_1 s_2 r_{im}.
\end{aligned}$$

For type 8, in which $(m, f, c) = (1, 1, 1)$, the variant allele carried by the child can be inherited either from the mother or the father with equal probabilities and, as such, the joint probability ends up being the summation of two probabilities weighted equally. We show the calculation of $P(M = 1, F = 1, C_1 = 1, D_1 = 1, D_2 = 0)$ as an example:

$$\begin{aligned}
&P(M = 1, F = 1, C_1 = 1, D_1 = 1, D_2 = 0)\\
&= P(M = 1, F = 1)P(C_1 = 1_m|M = 1, F = 1)\\
&\times P(D_1 = 1|M = 1, F = 1, C_1 = 1_m)P(D_2 = 0|M = 1, F = 1)\\
&+ P(M = 1, F = 1)P(C_1 = 1_f|M = 1, F = 1)\\
&\times P(D_1 = 1|M = 1, F = 1, C_1 = 1_f)P(D_2 = 0|M = 1, F = 1)\\
&= 1/4\mu_{11}\delta r_1 s_1(1 + r_{im})1/4(4 - \delta s_1 - \delta r_1 s_1 - \delta r_1 s_1 r_{im} - \delta r_2 s_1).
\end{aligned}$$

S2. Regularity Conditions and Proof of Theorem 1

The LIME$_{DSP}$ uses a multiplicative relative risk model for the disease prevalence are as given in (1) above. The vector of parameters of interest is denoted by

$$\boldsymbol{\theta} = (\delta, r_1, r_2, r_{im}, s_1, s_2).$$

Let n^1_{mfc} and n^0_{mfc} denote the count of affected proband-parent triads and unaffected proband-parent triads with genotype $M = m$, $F = f$, and $C = c$, respectively. Similarly, let sn^1_{mfc} and sn^0_{mfc} denote the counts of affected additional sibling-parent triads and unaffected additional sibling-parent triads with genotype combination $M = m$, $F = f$ and $C = c$, respectively.

To make inference about $\boldsymbol{\theta}$, we use the partial log-likelihood

$$\begin{aligned} l_{par}(\boldsymbol{\theta}) &= \sum_{m,f,c} \left\{ n^1_{mfc} \times \log[p_{mfc}(\boldsymbol{\theta})] + n^0_{mfc} \times \log[1 - p_{mfc}(\boldsymbol{\theta})] \right\} \\ &+ \sum_{m,f,c} \left\{ sn^1_{mfc} \times \log[q_{mfc}(\boldsymbol{\theta})] + sn^0_{mfc} \times \log[1 - q_{mfc}(\boldsymbol{\theta})] \right\} \\ &= l_{t1}(\boldsymbol{\theta}) + l_{t2}(\boldsymbol{\theta}). \end{aligned}$$

The effective total sample size, called n, in the partial log-likelihood $l_{par}(\boldsymbol{\theta})$, is computed as

$$\begin{aligned} n &= \sum_{m,f,c} [n^0_{mfc} + n^1_{mfc}] + \sum_{m,f,c} [sn^0_{mfc} + sn^1_{mfc}] \\ &= (N + N) + (sN^0_t + sN^1_t) \\ &= n_t + sn_t \end{aligned}$$

where N denotes the total number of independent families, and (sN^0_t, sN^1_t) are the total number of unaffected and affected siblings in all complete families, respectively. Hence n_t is the total number of probands children, and sn_t is the total number of additional siblings besides discordant sibpair.

The *maximum partial likelihood estimator* (MPLE) of $\boldsymbol{\theta}$ is denoted by

$$\widehat{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} \;\; l_{par}(\boldsymbol{\theta})$$

which is assumed to be obtained by solving the score-type equation

$$\frac{\partial l_{par}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = l'_{par}(\boldsymbol{\theta}) = l'_{t1}(\boldsymbol{\theta}) + l'_{t2}(\boldsymbol{\theta}) = \mathbf{0}.$$

We study the theoretical properties of $\widehat{\boldsymbol{\theta}}_n$, as the effective sample size $n = n_t + sn_t$ tends to infinity. We should note that here when $n \to \infty$, each of the sample sizes (n_t, sn_t) also tend to infinity, at the same rate, such that

$$\frac{n_t}{n} \longrightarrow 1 \;, \; \frac{sn_t}{n} \longrightarrow 1.$$

Clearly, this is under the assumption that both sums $\sum$ are present in the partial log-likelihood $l_{par}(\boldsymbol{\theta})$ defined above. If, however, there are no additional siblings, the theorem still holds and the proof is analogous.

Regularity Conditions

Let $\boldsymbol{\theta}_0$ be the true value of the parameter of interest. In what follows we denote

$$C_{r_n}(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^6 : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le r_n\}$$

as some neighborhood of $\boldsymbol{\theta}_0$, with radius r_n, where $r_n \to 0$, as n tends to infinity. Later on, we will see that this rate is $n^{-1/2}$. The regularity conditions are:

R1. The true value $\boldsymbol{\theta}_0$ of the parameter vector $\boldsymbol{\theta}$ is an interior point of the compact parameter space $\boldsymbol{\Theta}$.

R2. The cell probabilities $p_{mfc}(\boldsymbol{\theta})$ and $q_{mfc}(\boldsymbol{\theta})$ admit up to their third-order partial derivatives with respect to the elements of the parameter vector $\boldsymbol{\theta} = (\delta, r_1, r_2, r_{im}, s_1, s_2)$, for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

R3. The cell probabilities $p_{mfc}(\boldsymbol{\theta})$ and $q_{mfc}(\boldsymbol{\theta})$ are bounded away from the boundaries zero and one, at least for those $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$. Further, the partial derivatives of the cell probabilities, up to third order, are bounded by some constants, for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

R4. Identifiability: for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, $p_{mfc}(\boldsymbol{\theta}_1) = p_{mfc}(\boldsymbol{\theta}_2)$, $q_{mfc}(\boldsymbol{\theta}_1) = q_{mfc}(\boldsymbol{\theta}_2)$, for all (m, f, c) combinations, imply that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

R5. The information matrix

$$I(\boldsymbol{\theta}) = -E\{l''_{par}(\theta)\} = -E\{\frac{\partial^2 l_{par}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\}$$

is positive definite for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

We adopt the line of proof provided in Chanda (1954) and Lindsay (1980) to our partial likelihood context.

Proof of Theorem 1

Proof of Part (i) of Theorem 1. For simplicity in notation, we denote the vector of parameters of interest as $\boldsymbol{\theta} = (\delta, r_1, r_2, r_{im}, s_1, s_2) = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$. By the regularity Condition R2, for the first part of the partial log-likelihood, $l_{t1}(\boldsymbol{\theta})$, representing proband triads, we have that

$$\frac{\partial l_{t1}(\boldsymbol{\theta})}{\partial \theta_j} = l'_{t1,j}(\boldsymbol{\theta}) = l'_{t1,j}(\boldsymbol{\theta}_0) + \sum_{k=1}^{6} l''_{t1,jk}(\boldsymbol{\theta}_0)(\theta_k - \theta_k^0) + \frac{1}{2}\sum_{l,k}^{6} l'''_{t1,jkl}(\widetilde{\boldsymbol{\theta}})(\theta_k - \theta_k^0)(\theta_l - \theta_l^0) \quad (2)$$

for $j = 1, 2, \ldots, 6$, where $\widetilde{\boldsymbol{\theta}}$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$; $l''_{t1,jk}(\cdot)$ and $l'''_{t1,jkl}(\cdot)$ are the second and third-order partial derivatives of the function $l_{t1}(\cdot)$, respectively. For $j, k, l = 1, 2, 3, 4, 5, 6$, we have

$$
\begin{aligned}
l'_{t1,j}(\boldsymbol{\theta}) &= \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \left\{ \frac{n^1_{mfc}}{p_{mfc}(\boldsymbol{\theta})} - \frac{n_{mfc} - n^1_{mfc}}{1 - p_{mfc}(\boldsymbol{\theta})} \right\} \\
l''_{t1,jk}(\boldsymbol{\theta}) &= \sum_{m,f,c} \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \times \left\{ \frac{n^1_{mfc}}{p_{mfc}(\boldsymbol{\theta})} - \frac{n_{mfc} - n^1_{mfc}}{1 - p_{mfc}(\boldsymbol{\theta})} \right\} \\
&- \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \left\{ \frac{n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^2} + \frac{n_{mfc} - n^1_{mfc}}{[1 - p_{mfc}(\boldsymbol{\theta})]^2} \right\} \\
l'''_{t1,jkl}(\boldsymbol{\theta}) &= \sum_{m,f,c} \frac{\partial^3 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_l} \times \left\{ \frac{n^1_{mfc}}{p_{mfc}(\boldsymbol{\theta})} - \frac{n_{mfc} - n^1_{mfc}}{1 - p_{mfc}(\boldsymbol{\theta})} \right\} \\
&- \sum_{m,f,c} \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_l} \times \left\{ \frac{n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^2} + \frac{n_{mfc} - n^1_{mfc}}{[1 - p_{mfc}(\boldsymbol{\theta})]^2} \right\} \\
&- \sum_{m,f,c} \left[\frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} + \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} \right] \times \left\{ \frac{n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^2} + \frac{n_{mfc} - n^1_{mfc}}{[1 - p_{mfc}(\boldsymbol{\theta})]^2} \right\} \\
&- \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_l} \left\{ \frac{-2n^1_{mfc}}{[p_{mfc}(\boldsymbol{\theta})]^3} + \frac{2(n_{mfc} - n^1_{mfc})}{[1 - p_{mfc}(\boldsymbol{\theta})]^3} \right\}
\end{aligned}
$$

for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

For every triad type (m, f, c), denote the ratio

$$
r^1_{mfc} = \frac{n^1_{mfc}}{n_{mfc}}
$$

where $n_{mfc} = n^0_{mfc} + n^1_{mfc}$. The form of the partial log-likelihood $l_{par}(\boldsymbol{\theta})$ suggests that, for each triad type (m, f, c) and conditional on n_{mfc}, we have $n^1_{mfc} | n_{mfc} \sim Binomial(n_{mfc}, p_{mfc}(\boldsymbol{\theta}))$. By using a double conditional expectation technique, it is thus easy to see that $E(r^1_{mfc}) = p_{mfc}(\boldsymbol{\theta})$. Now, we have that

$$
\begin{aligned}
n^{-1} E\{l'_{t1,j}(\boldsymbol{\theta})\} &= 0 \\
-n^{-1} E\{l''_{t1,jk}(\boldsymbol{\theta})\} &= \sum_{m,f,c} \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \left\{ \frac{E(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})][1 - p_{mfc}(\boldsymbol{\theta})]} \right\} = I_{t1,jk}(\boldsymbol{\theta})
\end{aligned}
$$

for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$, where $E(\cdot)$ is the expected value under the model with the parameter $\boldsymbol{\theta}$.

Further, by the regularity condition R3, for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$,

$$
\begin{aligned}
n^{-1} |l'''_{t1,jkl}(\boldsymbol{\theta})| &\leq \sum_{m,f,c} 2 \left| \frac{\partial^3 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_l} \right| + \sum_{m,f,c} \left| \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_l} \right| \times \left\{ \frac{(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})][1 - p_{mfc}(\boldsymbol{\theta})]} \right\} \\
&+ \sum_{m,f,c} \left| \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} + \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial^2 p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_l} \right| \times \left\{ \frac{(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})][1 - p_{mfc}(\boldsymbol{\theta})]} \right\} \\
&+ 2 \sum_{m,f,c} \left| \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \frac{\partial p_{mfc}(\boldsymbol{\theta})}{\partial \theta_l} \right| \left\{ \frac{(n_{mfc}/n)}{[p_{mfc}(\boldsymbol{\theta})]^2} + \frac{(n_{mfc}/n)}{[1 - p_{mfc}(\boldsymbol{\theta})]^2} \right\} \\
&= O_p(1),
\end{aligned}
$$

which implies that $l'''_{t1,jkl}(\boldsymbol{\theta}) = O_p(n)$, for any $\boldsymbol{\theta} \in C_{r_n}(\boldsymbol{\theta}_0)$.

On the other hand, by the law of large numbers, we have that

$$r^1_{mfc} = \frac{n^1_{mfc}}{n_{mfc}} \xrightarrow{w.p.o} p_{mfc}(\boldsymbol{\theta}_0) \quad , \quad \frac{n_{mfc}}{n} \xrightarrow{w.p.o} E\left(\frac{n_{mfc}}{n}\right) = B_{mfc} \tag{3}$$

for some constant $0 < B_{mfc} < 1$, as $n \to \infty$, where w.p.o stands for with probability tending to one. Thus, using (3), as $n \to \infty$, we have

$$l'_{t1,j}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0 \ , \ l''_{t1,jk}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} I_{t1,jk}(\boldsymbol{\theta}_0) \ , \ l'''_{t1,jkl}(\boldsymbol{\theta}_0)/n = O_p(1). \tag{4}$$

for $j, k, l = 1, 2, \ldots, 6$.

By similar argument,s and under the regularity conditions R1-R5, for the remaining three terms of the partial log-likelihood, we have that

$$\begin{aligned} n^{-1}E\{l'_{t2,j}(\boldsymbol{\theta})\} &= 0 \\ -n^{-1}E\{l''_{t2,jk}(\boldsymbol{\theta})\} &= \sum_{(m,f,c)} \frac{\partial q_{mfc}(\boldsymbol{\theta})}{\partial \theta_j} \times \frac{\partial q_{mfc}(\boldsymbol{\theta})}{\partial \theta_k} \times \left\{ \frac{E(sn_{mfc}/n)}{[q_{mfc}(\boldsymbol{\theta})][1 - q_{mfc}(\boldsymbol{\theta})]} \right\} = I_{t2,jk}(\boldsymbol{\theta}) \\ n^{-1}\{l'''_{t2,jkl}(\boldsymbol{\theta})\} &= O_p(1) \quad \text{as} \quad n \to \infty. \end{aligned}$$

Thus, similar to (4), as $n \to \infty$, we have that

$$l'_{t2,j}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0 \ , \ l''_{t2,jk}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} I_{t2jk}(\boldsymbol{\theta}_0) \ , \ l'''_{t2,jkl}(\boldsymbol{\theta}_0)/n = O_p(1),$$

for $j, k, l = 1, 2, \ldots, 6$.

Using the above results, we have that

$$l'_{par}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} 0 \ , \ l''_{par}(\boldsymbol{\theta}_0)/n \xrightarrow{w.p.o} \boldsymbol{I}(\boldsymbol{\theta}_0) \ , \ l'''_{par}(\boldsymbol{\theta}_0)/n = O_p(1) \tag{5}$$

as $n \to \infty$. Here $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is a 6×6 information matrix constructed based on the $\{I_{t1,jk}(\boldsymbol{\theta}), I_{t2,jk}(\boldsymbol{\theta})\}$, for $j, k = 1, 2, \ldots, 6$.

Thus consider the score-type equation divided by the total sample size n, which leads to the equations

$$n^{-1} \sum_{k=1}^{6} l''_{par,jk}(\boldsymbol{\theta}_0)(\theta_k - \theta^0_k) = -n^{-1} l'_{par,j}(\boldsymbol{\theta}_0) - \frac{1}{2} n^{-1} \sum_{l,k=1}^{6} l'''_{par,jkl}(\widetilde{\boldsymbol{\theta}})(\theta_k - \theta^0_k)(\theta_l - \theta^0_l)$$

for $j = 1, \ldots, 6$. By expanding the summation on the left hand side and re-writing with respect to each $\theta_k - \theta^0_k$, we have that

$$\theta_k - \theta^0_k = \sum_{j=1}^{6} [\frac{-1}{n} l'_{par,j}(\boldsymbol{\theta}_0)] \times l^*_{par,jk}(\boldsymbol{\theta}_0) - \frac{1}{2} \sum_{l,r=1}^{6} \left[(\theta_r - \theta^0_r)(\theta_l - \theta^0_l) \left(\sum_{j=1}^{6} [\frac{1}{n} l'''_{par,jrl}(\widetilde{\boldsymbol{\theta}})] \times l^*_{par,jk}(\boldsymbol{\theta}_0) \right) \right] \tag{6}$$

for $k = 1, \ldots, 6$, where $l^*_{par,jk}(\boldsymbol{\theta}_0)$ are the elements of the inverse matrix $\Big(l''_{par,jk}(\boldsymbol{\theta}_0)/n; j,k = 1,\ldots,6\Big)^{-1}$. By (5), the first term on the right hand side of the above equations tends to zero, as $n \to \infty$. This implies that the equations in (6) have at least one solution, in terms of $\theta_k - \theta_k^0$, that satisfies

$$\hat{\theta}_k - \theta_k^0 \longrightarrow^p 0 \quad ; \quad k = 1, \ldots, 6,$$

as $n \to \infty$. Thus, there exists a solution, say, $\widehat{\boldsymbol{\theta}}_n$ of the score-type equation $l'_{par}(\boldsymbol{\theta}) = \mathbf{0}$ such that $\widehat{\boldsymbol{\theta}}_n \longrightarrow^p \boldsymbol{\theta}_0$, as $n \to \infty$.

Now we prove the uniqueness of such consistent estimator. Under the regularity conditions R1-R5, and consistency of $\hat{\boldsymbol{\theta}}_n$, we have that

$$\frac{1}{n} l''_{par}(\hat{\boldsymbol{\theta}}_n) + I(\boldsymbol{\theta}_0) = o_p(1) \tag{7}$$

as n tends to ∞, where $I(\boldsymbol{\theta}_0)$ is the positive definite information matrix. Let us assume that there exist two such consistent estimators, say, $\hat{\boldsymbol{\theta}}_{1n}$ and $\hat{\boldsymbol{\theta}}_{2n}$ of $\boldsymbol{\theta}_0$ that are the solutions of the score-type equation

$$l'_{par}(\boldsymbol{\theta}) = 0.$$

By the extension of Rolle's theorem to multivariate case, there exists a point $\tilde{\boldsymbol{\theta}}_n$ laying inside a hyper-cell with the vector $\hat{\boldsymbol{\theta}}_{1n} - \hat{\boldsymbol{\theta}}_{2n}$ as its diagonal, such that

$$l''_{par}(\tilde{\boldsymbol{\theta}}_n) = 0. \tag{8}$$

On the other hand, since $\hat{\boldsymbol{\theta}}_{1n}$ and $\hat{\boldsymbol{\theta}}_{2n}$ are consistent estimators, so is $\tilde{\boldsymbol{\theta}}_n$ and it must satisfy (7). But clearly (7) and (8) contradict. This implies that the consistent estimator $\hat{\boldsymbol{\theta}}_n$ is unique. This completes the proof of Part(i). ♠

The result of Lemma 1 below is used for proving Part (ii) of Theorem 1.

Lemma 1 *Under the regularity conditions R1-R5, we have that*

$$\frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0))$$

as $n \to \infty$.

Proof of Lemma 1. Consider the partial-score function

$$\begin{aligned} \frac{\partial l_{par}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = l'_{par}(\boldsymbol{\theta}_0) &= l'_{t1}(\boldsymbol{\theta}_0) + l'_{t2}(\boldsymbol{\theta}_0) \\ &= \sum_{m,f,c} \frac{n_{mfc} \times p'_{mfc}(\boldsymbol{\theta}_0)}{p_{mfc}(\boldsymbol{\theta}_0)[1 - p_{mfc}(\boldsymbol{\theta}_0)]} \times [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)] \\ &+ \sum_{m,f,c} \frac{sn_{mfc} \times q'_{mfc}(\boldsymbol{\theta}_0)}{q_{mfc}(\boldsymbol{\theta}_0)[1 - q_{mfc}(\boldsymbol{\theta}_0)]} \times [s^1_{mfc} - q_{mfc}(\boldsymbol{\theta}_0)], \end{aligned}$$

where $p'_{mfc}(\boldsymbol{\theta}_0)$ and $q'_{mfc}(\boldsymbol{\theta}_0)$ are the 6-dimensional vectors of the partial derivatives of the cell probabilities $p_{mfc}(\boldsymbol{\theta})$ and $q_{mfc}(\boldsymbol{\theta})$, with respect to $\boldsymbol{\theta}$, which are evaluated at the true $\boldsymbol{\theta}_0$. Also,

$$r^1_{mfc} = \frac{n^1_{mfc}}{n_{mfc}} \quad , \quad s^1_{mfc} = \frac{sn^1_{mfc}}{sn_{mfc}},$$

are the ratios of the number of cases among: proband (m, f, c) triads and additional (m, f, c) sibling triads respectively.

We first try to find the limiting distribution of $l'_{t1}(\boldsymbol{\theta}_0)/\sqrt{n}$, as $n \to \infty$. We have that

$$\frac{l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} = \sum_{m,f,c} \frac{p'_{mfc}(\boldsymbol{\theta}_0)}{p_{mfc}(\boldsymbol{\theta}_0)[1 - p_{mfc}(\boldsymbol{\theta}_0)]} \times \sqrt{\frac{n_{mfc}}{n}} \times \sqrt{n_{mfc}}\, [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)]$$

In what follows we use the Wald device. For any non-zero vector $\boldsymbol{v} \in \mathbb{R}^6$,

$$w_n(\boldsymbol{\theta}_0) = \frac{\boldsymbol{v}^\top l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} = \sum_{m,f,c} \frac{u_{mfc}(\boldsymbol{\theta}_0)}{p_{mfc}(\boldsymbol{\theta}_0)[1 - p_{mfc}(\boldsymbol{\theta}_0)]} \times \sqrt{\frac{n_{mfc}}{n}} \times \sqrt{n_{mfc}}\, [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)]$$

where $u_{mfc}(\boldsymbol{\theta}_0) = \boldsymbol{v}^\top p'_{mfc}(\boldsymbol{\theta}_0)$ is a scalar. Note that conditional on the n_{mfc}'s, the ratios r^1_{mfc}'s are independent, each having the conditional asymptotic distribution

$$\sqrt{n_{mfc}}\, [r^1_{mfc} - p_{mfc}(\boldsymbol{\theta}_0)] \longrightarrow^d N(0, p_{mfc}(\boldsymbol{\theta}_0)(1 - p_{mfc}(\boldsymbol{\theta}_0))$$

as $n \to \infty$. Note that since n_{mfc}'s are following a multinomial distribution, say, with the joint probability mass function $g(n_{mfc}; m, f, c)$, then

$$F_n(w) = P(w_n(\boldsymbol{\theta}_0) \le w) = \sum_{\{m,f,c:n_{mfc}=0\}}^{n_t} P(w_n(\boldsymbol{\theta}_0) \le w | n_{mfc}, m, f, c)\; g(n_{mfc}; m, f, c).$$

On the other hand, as $n \to \infty$, since $n_{mfc}/n \overset{p}{\to} E(n_{mfc}/n) = B_{mfc}$, for some constant $0 < B_{mfc} < 1$, then

$$(w_n(\boldsymbol{\theta}_0) | n_{mfc}, m, f, c) \longrightarrow^d N(0, \sigma^2(\boldsymbol{\theta}_0))$$

where

$$\sigma^2(\boldsymbol{\theta}_0) = \sum_{m,f,c} \frac{u^2_{mfc}(\boldsymbol{\theta}_0) \times B_{mfc}}{p_{mfc}(\boldsymbol{\theta}_0)(1 - p_{mfc}(\boldsymbol{\theta}_0))}.$$

Therefore, for $w \in \mathbb{R}$, as $n \to \infty$,

$$F_n(w) \longrightarrow \frac{1}{\sigma(\boldsymbol{\theta}_0)} \Phi\left(\frac{w}{\sigma(\boldsymbol{\theta}_0)}\right)$$

where $\Phi(\cdot)$ is the distribution function of the standard normal. This implies that

$$w_n(\boldsymbol{\theta}_0) \longrightarrow^d N(0, \sigma^2(\boldsymbol{\theta}_0))$$

as $n \to \infty$. Hence,

$$\frac{l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N\left(\mathbf{0}, \sum_{m,f,c} \frac{[p'_{mfc}(\boldsymbol{\theta}_0)][p'_{mfc}(\boldsymbol{\theta}_0)]^\top \times B_{mfc}}{p_{mfc}(\boldsymbol{\theta}_0)(1-p_{mfc}(\boldsymbol{\theta}_0))}\right), \quad n \to \infty.$$

Similarly, we have

$$\frac{l'_{t2}(\boldsymbol{\theta}_0)}{\sqrt{n}} \quad \longrightarrow^d \quad N\left(\mathbf{0}, \sum_{m,f,c} \frac{[q'_{mfc}(\boldsymbol{\theta}_0)][q'_{mfc}(\boldsymbol{\theta}_0)]^\top \times C_{mfc}}{q_{mfc}(\boldsymbol{\theta}_0)(1-q_{mfc}(\boldsymbol{\theta}_0))}\right),$$

for some constants $0 < C_{mfc} < 1$, such that, as $n \to \infty$,

$$\frac{sn_{mfc}}{n} \longrightarrow^p C_{mfc}.$$

Thus, by the independence of the ratios r^1_{mfc} and s^1_{mfc}, as the effective sample size $n = n_t + sn_t$ tends to infinity, we have

$$\frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} = \frac{l'_{t1}(\boldsymbol{\theta}_0)}{\sqrt{n}} + \frac{l'_{t2}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N\left(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0)\right)$$

where $\boldsymbol{I}(\boldsymbol{\theta}_0) = \boldsymbol{I}_{t1}(\boldsymbol{\theta}_0) + \boldsymbol{I}_{t2}(\boldsymbol{\theta}_0)$, and

$$\begin{aligned}
\boldsymbol{I}_{t1}(\boldsymbol{\theta}_0) &= \sum_{m,f,c} \frac{[p'_{mfc}(\boldsymbol{\theta}_0)][p'_{mfc}(\boldsymbol{\theta}_0)]^\top \times B_{mfc}}{p_{mfc}(\boldsymbol{\theta}_0)(1-p_{mfc}(\boldsymbol{\theta}_0)},\\
\boldsymbol{I}_{t2}(\boldsymbol{\theta}_0) &= \sum_{m,f,c} \frac{[q'_{mfc}(\boldsymbol{\theta}_0)][q'_{mfc}(\boldsymbol{\theta}_0)]^\top \times C_{mfc}}{q_{mfc}(\boldsymbol{\theta}_0)(1-q_{mfc}(\boldsymbol{\theta}_0)},
\end{aligned}$$

are 6×6-dimensional positive definite information matrices.

Hence, as $n \to \infty$, we have that

$$\frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0)). \tag{9}$$

This completes the proof of Lemma 1. ♠

Proof of Part (ii) of Theorem 1. Let $\widehat{\boldsymbol{\theta}}_n$ be the MPLE, which satisfies the score-type equation

$$l'_{par}(\widehat{\boldsymbol{\theta}}_n) = 0.$$

By the regularity conditions R1-R5, we have that

$$\begin{aligned}
\mathbf{0} &= \frac{1}{n}\, l'_{par}(\boldsymbol{\theta}_0) + \frac{1}{n}\, l''_{par}(\boldsymbol{\theta}_0)(1+o_p(1)) \times (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\\
&= \frac{1}{n}\, l'_{par}(\boldsymbol{\theta}_0) + \left[\frac{1}{n}\, l''_{par}(\boldsymbol{\theta}_0) + \boldsymbol{I}(\boldsymbol{\theta}_0) - \boldsymbol{I}(\boldsymbol{\theta}_0)\right](1+o_p(1)) \times (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)
\end{aligned}$$

where by (5) $l''_{par}(\boldsymbol{\theta}_0)/n + \boldsymbol{I}(\boldsymbol{\theta}_0) = o_p(1)$. Therefore, by the result of Lemma 1,

$$\sqrt{n}\ (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \times \frac{l'_{par}(\boldsymbol{\theta}_0)}{\sqrt{n}} \longrightarrow^d N(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)),$$

as $n \to \infty$. This completes the proof of Part(ii) of Theorem 1. ♠

S3. Estimation of Maternal Effect with the DSP Design without Additional Siblings

To analyze the information for detecting parent-of-origin effects, especially maternal effect, we take a closer look at p_{mfc} in the partial likelihood:

$$\begin{aligned} p_{mfc} &= \frac{P(D=1|m,f,c)P(D=0|m,f)}{P(D=1|m,f,c)P(D=0|m,f)+P(D=0|m,f,c)P(D=1|m,f)} \\ &= 1/\left(1+\frac{P(D=0|m,f,c)}{P(D=0|m,f)}/\frac{P(D=1|m,f,c)}{P(D=1|m,f)}\right). \end{aligned}$$

$$\begin{aligned} \frac{P(D=1|m,f,c)}{P(D=1|m,f)} &= \frac{\delta r_1^{I(C=1)} r_2^{I(C=2)} r_{im}^{I(C=1_m)} s_1^{I(M=1)} s_2^{I(M=2)}}{\sum_{c*} p(c*|m,f)\delta r_1^{I(C*=1)} r_2^{I(C*=2)} r_{im}^{I(C*=1_m)} s_1^{I(M=1)} s_2^{I(M=2)}} \\ &= \frac{r_1^{I(C=1)} r_2^{I(C=2)} r_{im}^{I(C=1_m)}}{\sum_{c*} p(c*|m,f) r_1^{I(C*=1)} r_2^{I(C*=2)} r_{im}^{I(C*=1_m)}}. \end{aligned} \tag{10}$$

$$\frac{P(D=0|m,f,c)}{P(D=0|m,f)} = \frac{1-\delta r_1^{I(c=1)} r_2^{I(c=2)} r_{im}^{I(c=1_m)} s_1^{I(m=1)} s_2^{I(m=2)}}{1-\sum_{c*} p(c*|m,f)\delta r_1^{I(c*=1)} r_2^{I(c*=2)} r_{im}^{I(c*=1_m)} s_1^{I(m=1)} s_2^{I(m=2)}}. \tag{11}$$

We can see that for maternal effect, (10) is totally independent of parameters s_1 and s_2. Though (11) includes maternal effect parameters, when there is only maternal effect, i.e. $r_1 = r_2 = r_{im} = 1$, maternal effect parameters will be canceled out again. Furthermore, when there are other effects besides maternal effect, only (F, M) belonging to $\{(1,2),(2,1),(1,0),(0,1),(1,1)\}$ is informative for (11), and if disease penetrance for these combinations with different offspring genotype are similar, for example, $P(D = 1|M = 1, F = 2, C = 1)$ is similar as $P(D = 1|M = 1, F = 2, C = 2)$, then the combination is again almost non-informative. On the other hand, most of child-parent genotype combinations are informative for detecting imprinting effect for both (10) and (11). This is consistent with the result from the simulation that the power to detect maternal effect is very low when only such discordant sibpairs without additional siblings are recruited, whereas when additional siblings are also recruited, the power will increase, as no term can be canceled.

S4. DSP design with missing father genotypes

In LIME proposed by Yang and Lin (2013), nuclear families with father's genotype missing can still contribute to the estimation of the parameters. However, as we elaborate in the following, LIME$_{DSP}$ cannot be generalized to the discordant sibpairs design with father's genotype missing. Following the same idea as in complete data, denote n_{mc}^1 as the count of affected proband-mother pairs with genotype $M = m$ and $C_1 = c$, and n_{mc}^0 as the count of unaffected proband-mother pairs with genotype $M = m$ and $C_2 = c$. Let n_p denote the count of independent families. To keep it focused, we assume there are no additional siblings. Thus, the likelihood can be written as follows, where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ denote the parameters of interest and the nuisance parameters, respectively. That is,

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\phi})_p &= \prod_{m,c} [p_{mc}^{n^1_{mc}} (1 - p_{mc})^{n^0_{mc}}] \prod_{m,c} S_{mc}^{n^1_{mc}+n^0_{mc}} \qquad (12) \\ &\times \prod_{j=1}^{n_p} \frac{P(M_j = m_j, C_{j1} = c_{j1}, C_{j2} = c_{j2})}{P(M_j = m_j, C_{j1} = c_{j1}) P(M_j = m_j, C_{j2} = c_{j2})} \frac{P(D_1 = 1, D_2 = 0)}{P(D_1 = 1|m_j, c_{j2}) P(D_2 = 0|m_j, c_{j1})}, \end{aligned}$$

where the j represents the j^{th} DSP in the data, and

$$p_{mc} = \frac{P(M = m, C_1 = c|D_1 = 1, D_2 = 0)}{P(M = m, C_1 = c|D_1 = 1, D_2 = 0) + P(M = m, C_2 = c|D_1 = 1, D_2 = 0)},$$

and the denominator is denoted as S_{mc}. However, we can rewrite the probability as

$$p_{mc} = \frac{1}{1 + \frac{P(M=m,C_1=c,D_1=1,D_2=0)}{P(M=m,C_2=c,D_1=1,D_2=0)}}.$$

Then, as we can see from Supplementary Table S7, p_{mc} still involves nuisance parameters, thus we cannot extract out a partial likelihood component to estimate parameters.

S5. Relative Efficiency of LIME$_{DSP}$ vs. LIME

To compare the relative efficiency of the LIME and LIME$_{DSP}$ study designs, we compare the "per individual" information when LIME$_{DSP}$ is applied to a D+2 design, with LIME to a T+3 study design, where a T+3 design refers to a case-parent/control-parent study design in which each family (either a case family or a control family) has 3 additional siblings. We chose to compare these two designs as the total number of individuals per family is equal to 6 in both designs. We vary the proportion of case families for the T+3 design from 0.025 to 0.975 by 0.025. Figures S12-19 are for disease model 1-8 under scenario 8, where the horizontal line is the information per individual for the D+2 design, while the circles represent that for the T+3 data. We can see that, as expected, a balanced setting, the proportion of case families being 0.5, is generally the most informative, in which case the D+2 design is not as efficient as the T+3 design. However, when such a balanced setting is not available, the D+2 design can be more efficient. This is especially true for making inference about association and imprinting effects. However, the T+3 design typically has more power than D+2 for inference about maternal effect, as we discussed earlier.

References

Chanda, K. C. (1954), "A Note on the Consistency and Maxima of the Roots of Likelihood Equations", *Biometrika*, 41, 56–61.

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.

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.

Table S1 Top-20 SNPs having the smallest p-values for association with club foot using LIME$_{DSP}$

Rank	SNP	Chr	Position(BP)	Gene	-log$_{10}$(P-value)
1	rs1023913	9	23003004	TOX3	4.7633
2	rs6040798	20	11602357		4.7631
3	rs1870488	6	63933078	WDR55	4.2773
4	rs292202	5	73582314	FAM53A	4.137
5	rs12523740	6	32897704		3.8777
6	rs10484209	4	37074039		3.8774
7	rs2953299	2	51852092		3.8746
8	rs1327992	6	4310124	CTB-32H22.1	3.7614
9	rs11594622	10	72580602		3.6976
10	rs17712426	10	83563646		3.6968
11	rs17035675	4	106457953		3.6754
12	rs6933121	6	79856243		3.6512
13	rs17141297	10	17580107		3.6244
14	rs12512863	4	24134430		3.6105
15	rs2650703	10	63236710	LOC101928781	3.5965
16	rs3115763	2	138763552	HNMT	3.5646
17	rs11980754	7	4408130		3.5394
18	rs1568717	15	61362446	RORA	3.5223
19	rs915895	6	32190216	KCND3	3.5093
20	rs2384549	12	115349867		4.9359

Table S2 Top-20 SNPs having the smallest p-values for imprinting effect on club foot using LIME$_{DSP}$

Rank	SNP	Chr	Position(BP)	Gene	-log$_{10}$(P-value)
1	rs1079295	5	5165951	MT1A	13.4218
2	rs2405941	18	73740843		13.2871
3	rs2320214	18	4420249	DLGAP1	12.4824
4	rs13384546	2	185616127	ZNF804A	12.2454
5	rs2145214	20	42237066	IFT52	11.9946
6	rs213134	17	32823258		11.7425
7	rs7162435	15	56121333	NEDD4	11.5518
8	rs6151826	5	80080680	MSH3	11.4768
9	rs2520121	16	26577301		11.4644
10	rs1224524	6	67250007		11.3491
11	rs10413941	19	49347707	PLEKHA4	11.1828
12	rs11610123	12	47500730	PCED1B	11.1069
13	rs11048527	12	26604100	ITPR2	11.1035
14	rs6785520	3	170991646	TNIK	10.9721
15	rs17117977	11	115130709		10.7654
16	rs13228877	7	34199973		10.6878
17	rs3743308	15	69563185	DRAIC	10.6850
18	rs11789529	9	130164412		10.5804
19	rs908296	2	9814639		10.4491
20	rs12223323	11	26298810	ANO3	10.3638

Table S3 Top-20 SNPs having the smallest p-values for maternal effect on club foot using method LIME$_{DSP}$

Rank	SNP	Chr	Position(BP)	Gene	-log$_{10}$(P-value)
1	rs2384549	12	115349867		4.9359
2	rs3781503	10	121571506	INPP5F	4.9039
3	rs9446305	6	71598570	B3GAT2	4.5466
4	rs10224932	7	31035681		4.515
5	rs11766624	7	69887084	AUTS2	4.4982
6	rs585157	13	99045319	FARP1	4.467
7	rs9540648	13	34951551		4.3431
8	rs10499527	7	21243187		4.3245
9	rs1005391	4	16386448		4.2718
10	rs6711382	2	152531076	NEB	4.2556
11	rs7801891	7	17133513		4.2536
12	rs9818949	3	197683750	IQCG	4.2419
13	rs723636	6	160580493	SLC22A1	4.2334
14	rs2018193	1	153079071		4.215
15	rs10066164	5	13945188	DNAH5	4.2147
16	rs7546648	1	152931206		4.2143
17	rs17559561	4	132367852		4.1886
18	rs1529557	2	37898991		4.1799
19	rs12550249	8	13140608	DLC1	4.1429
20	rs17712426	10	83563646		3.6968

Table S4 Top-20 SNPs having the smallest p-values for association with hypertension using LIME$_{D+}$

Rank	SNP	Chr	Position(BP)	Gene	-log$_{10}$(P-value)
1	rs16892095	4	15518356	CC2D2A	15.65
2	rs11128437	3	75447270		15.48
3	rs4125931	4	49489497		15.35
4	rs2405219	18	731439945	SMIM21	15.26
5	rs2229188	7	92134309	CYP51A1	15.11
6	rs4702048	5	14750799	ANKH	14.44
7	rs12626631	21	45001813	HSF2BP	14.22
8	rs3734815	6	29694680	HLA-F	14.08
9	rs13202088	6	163174689	PACRG	13.64
10	rs52828135	15	unknown		13.50
11	rs6485742	11	12454075	PARVA	12.82
12	rs11843435	13	69479766		11.17
13	rs4707557	6	90362782	MDN1	11.16
14	rs7032988	9	91837409		9.93
15	rs2013347	17	22171189		8.73
16	rs11672918	19	8943393	ZNF558	8.62
17	rs13255458	8	41636070	ANK1	8.61
18	rs2272487	3	126733094	CHCHD6	8.41
19	rs2947658	3	125607009		8.07
20	rs12256916	10	38344894	ZNF33A	7.99

Table S5 Top-20 SNPs having the smallest p-values for imprinting effect on hypertension using LIME$_{D+}$

Rank	SNP	Chr	Position(BP)	Gene	-log$_{10}$(P-value)
1	rs16892095	4	15518356	CC2D2A	15.65
2	rs11128437	3	75447270		15.48
3	rs4125931	4	49489497		15.35
4	rs2405219	18	731439945	SMIM21	15.26
5	rs2229188	7	92134309	CYP51A1	15.11
6	rs4702048	5	14750799	ANKH	14.44
7	rs12626631	21	45001813	HSF2BP	14.22
8	rs3734815	6	29694680	HLA-F	14.08
9	rs13202088	6	163174689	PACRG	13.64
10	rs52828135	15	unknown		13.50
11	rs6485742	11	12454075	PARVA	12.82
12	rs11843435	13	69479766		11.17
13	rs4707557	6	90362782	MDN1	11.16
14	rs7032988	9	91837409		9.93
15	rs2013347	17	22171189		8.73
16	rs11672918	19	8943393	ZNF558	8.62
17	rs13255458	8	41636070	ANK1	8.61
18	rs2272487	3	126733094	CHCHD6	8.41
19	rs2947658	3	125607009		8.07
20	rs12256916	10	38344894	ZNF33A	7.99

Table S6 Top-20 SNPs having the smallest p-values for maternal effect on hypertension using LIME$_{D+}$

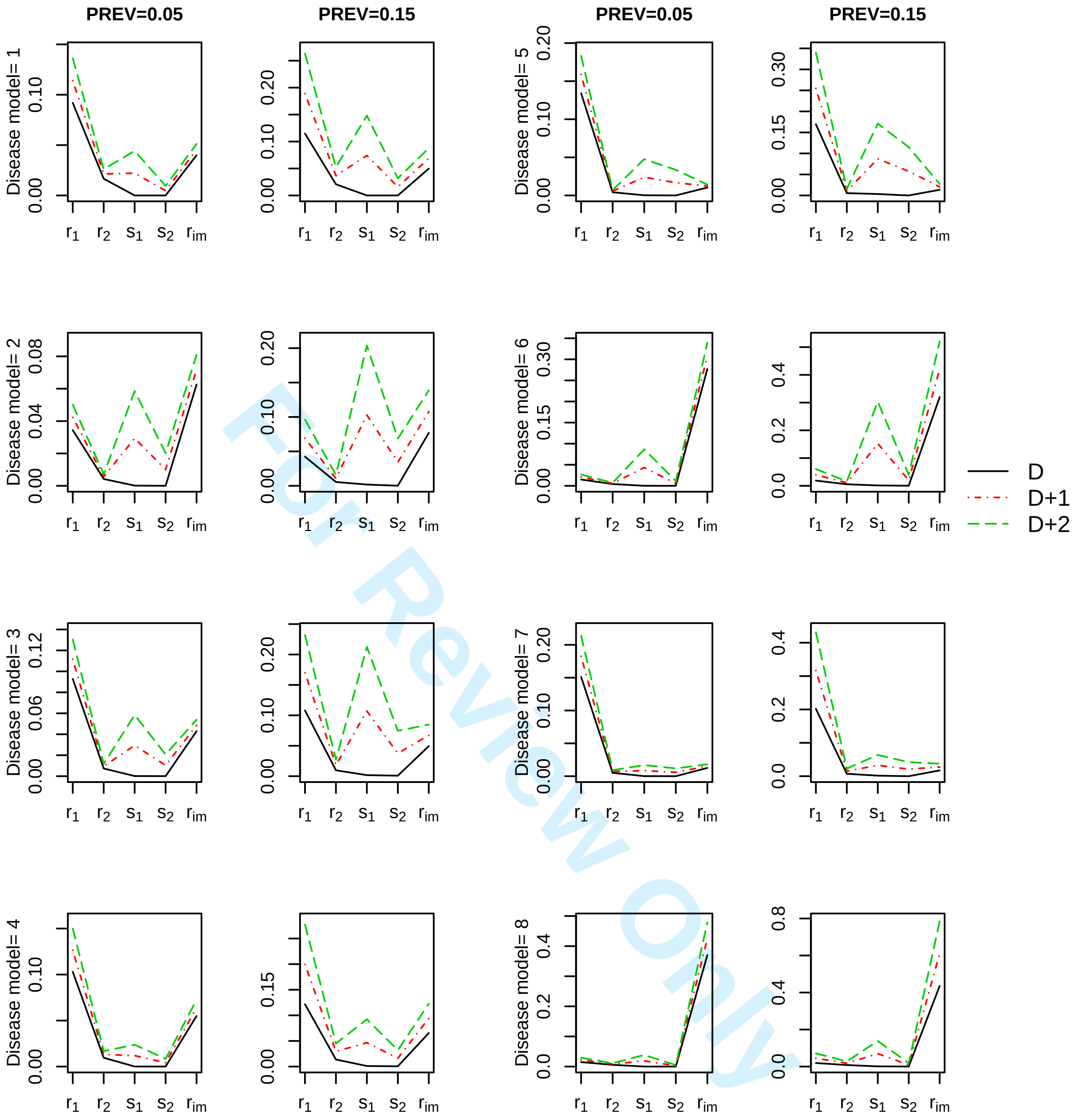
Rank	SNP	Chr	Position(BP)	Gene	-log$_{10}$(P-value)
1	rs2272487	3	126451936	CHCHD6	8.44
2	rs9852584	3	126445456	CHCHD6	6.26
3	rs13230531	7	6114558	CHCHD6	5.52
4	rs17631957	14	81755544	STON2	5.49
5	rs820866	5	73978700		5.43
6	rs6086342	20	8096104		5.23
7	rs7741727	6	132069916	ENPP3	5.19
8	rs1370656	2	178607997	PDE11A	5.18
9	rs7133914	12	40702910	LRRK2	5.16
10	rs17601580	6	132061419	ENPP3	5.07
11	rs3856154	1	225565014	DNAH14	5.03
12	rs2165661	11	100142833	CNTN5	4.99
13	rs12368599	12	12908793	GPRC5A	4.92
14	rs17158657	15	84405464	ADAMTSL3	4.90
15	rs16832191	3	120944943	STXBP5L	4.88
16	rs3205144	3	172349215	NCEH1	4.82
17	rs4813864	20	8515840	PLCB1	4.78
18	rs17460330	4	36338943	DTHD1	4.76
19	rs10209069	2	153384254	FMNL2	4.71
20	rs390878	4	103213241	SLC39A8	4.67

Table S7. Joint probabilities of $P(M = m, C_1 = c, D_1 = 1, D_2 = 0)$ and $P(M = m, C_2 = c, D_1 = 1, D_2 = 0)$

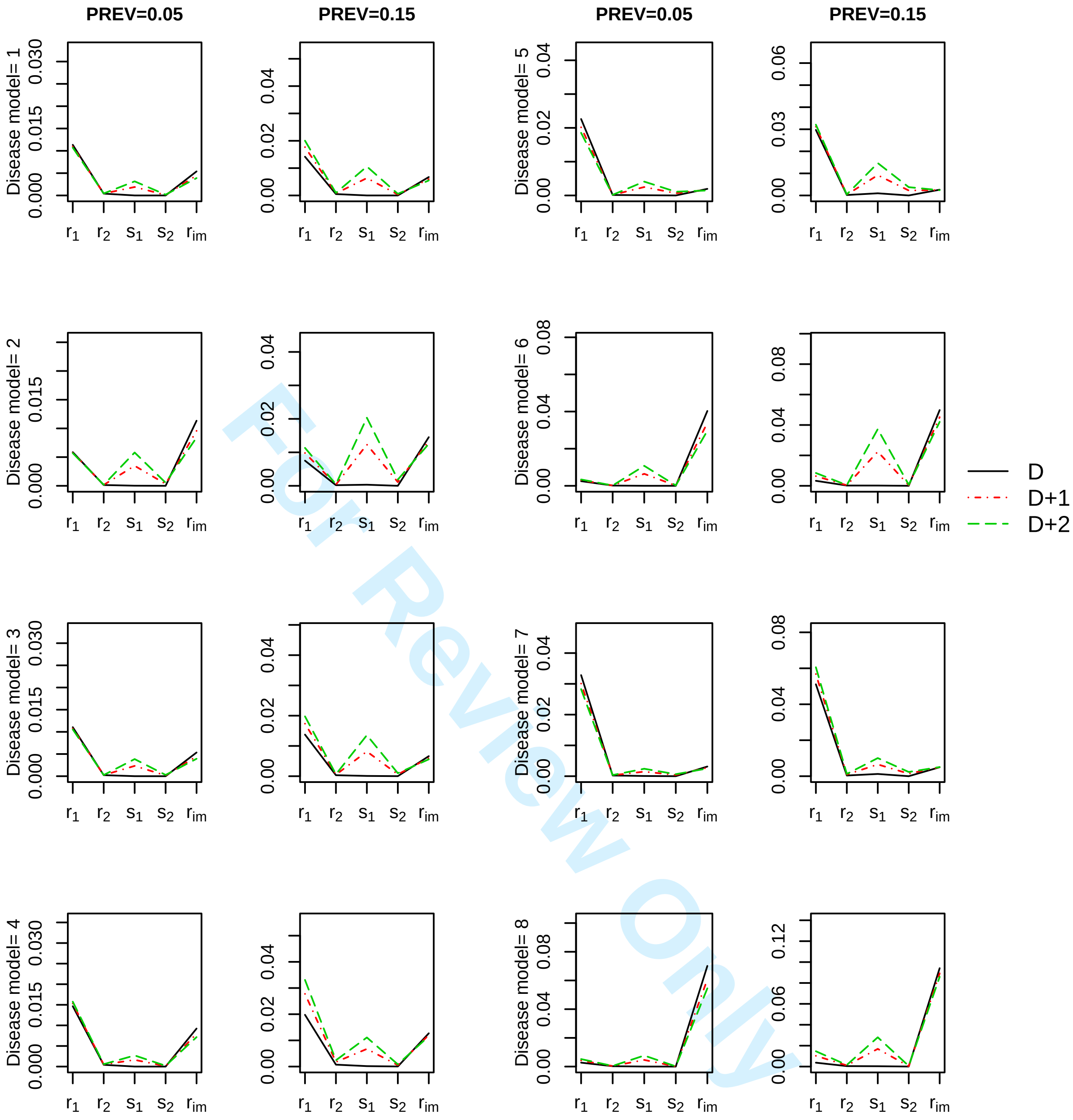
Type	m	c	$P(M = m, C_1 = c, D_1 = 1, D_2 = 0)$
1	0	0	$\mu_{00}(1-\delta)\delta + \frac{1}{4}\mu_{01}\delta(2-\delta-\delta r_1)^a$
2	0	1	$\frac{1}{4}\mu_{01}\delta r_1(2-\delta r_1-\delta) + \mu_{02}(1-\delta r_1)\delta r_1$
3	1	0	$\frac{1}{4}\mu_{10}\delta s_1(2-\delta s_1-\delta s_1 r_1 r_{im})$ $+\frac{1}{16}\mu_{11}\delta s_1(4-\delta s_1-\delta s_1 r_1(1+r_{im})-\delta s_1 r_2)$
4	1	1	$\frac{1}{4}\mu_{10}\delta s_1 r_1 r_{im}(2-\delta s_1-\delta s_1 r_1 r_{im})$ $+\frac{1}{16}\mu_{11}\delta s_1 r_1(1+r_{im})(4-\delta s_1-\delta s_1 r_1-\delta s_1 r_1 r_i m-\delta r_2 s_1)$ $+\frac{1}{4}\mu_{12}\delta r_1 s_1(2-\delta r_1 s_1-\delta r_2 s_1)$
5	1	2	$\frac{1}{16}\mu_{11}\delta s_1 r_2(4-\delta s_1-\delta s_1 r_1(1+r_{im})-\delta s_1 r_2)$ $+\frac{1}{4}\mu_{12}\delta s_1 r_2(2-\delta s_1 r_1-\delta s_1 r_2)$
6	2	1	$\mu_{20}(1-\delta s_2 r_1 r_{im})\delta s_2 r_1 r_{im}$ $+\frac{1}{4}\mu_{21}\delta s_2 r_1 r_{im}(2-\delta s_2 r_1 r_{im}-\delta s_2 r_2)$
7	2	2	$\frac{1}{4}\mu_{21}\delta s_2 r_2(2-\delta s_2 r_1 r_{im}-\delta s_2 r_2) + \mu_{22}(1-\delta s_2 r_2)\delta r_2 s_2$

Type	m	c	$P(M = m, C_2 = c, D_1 = 1, D_2 = 0)$
1	0	0	$\mu_{00}(1-\delta)\delta + \frac{1}{4}\mu_{01}(1-\delta)\delta(1+r_1)$
2	0	1	$\frac{1}{4}\mu_{01}(1-\delta r_1)\delta(1+r_1) + \mu_{02}(1-\delta r_1)\delta r_1$
3	1	0	$\frac{1}{4}\mu_{10}(1-\delta s_1)\delta s_1(1+r_1 r_{im})$ $+\frac{1}{16}\mu_{11}(1-\delta s_1)\delta s_1(1+r_2+r_1(1+r_{im}))$
4	1	1	$\frac{1}{4}\mu_{10}(1-\delta s_1 r_1 r_{im})\delta s_1(1+r_1 r_{im})$ $+\frac{1}{16}\mu_{11}[2-\delta r_1 s_1(1-r_{im})]\delta s_1(1+r_1(1+r_{im})+r_2)$ $+\frac{1}{4}\mu_{12}(1-\delta r_1 s_1)\delta s_1(r_1+r_2)$
5	1	2	$\frac{1}{16}\mu_{11}(1-\delta s_1 r_2)\delta s_1(1+r_2+r_1(1+r_{im}))$ $+\frac{1}{4}\mu_{12}(1-\delta s_1 r_2)\delta s_1(r_1+r_2)$
6	2	1	$\mu_{20}(1-\delta s_2 r_1 r_{im})\delta s_2 r_1 r_{im}$ $+\frac{1}{4}\mu_{21}(1-\delta s_2 r_1 r_{im})\delta s_2(r_2+r_1 r_{im})$
7	2	2	$\frac{1}{4}\mu_{21}(1-\delta s_2 r_2)\delta s_2(r_1 r_{im}+r_2) + \mu_{22}(1-\delta s_2 r_2)\delta r_2 s_2$

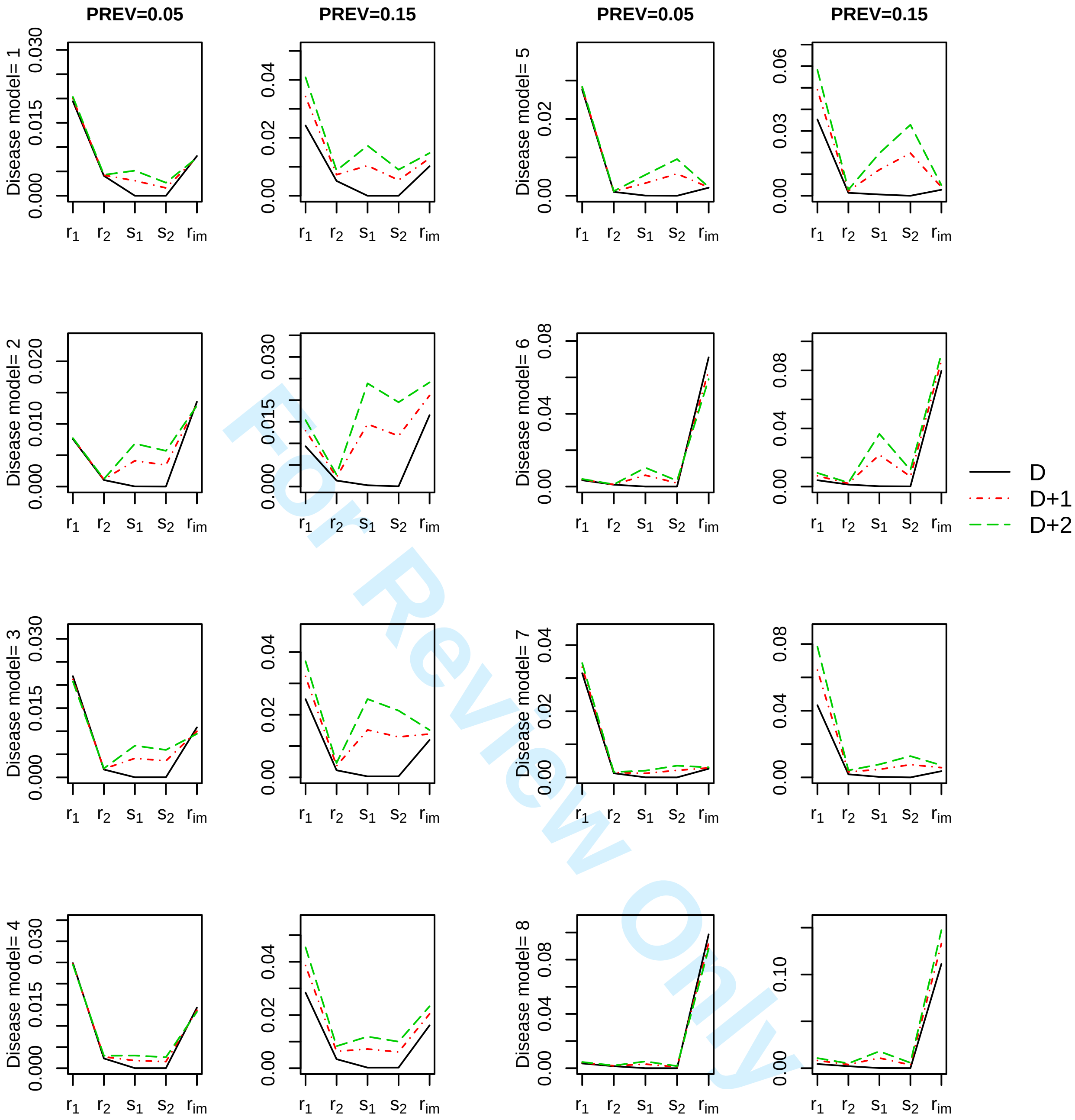
Note: ${}^a r_1$: relative risk of carrying one variant allele; r_2: relative risk of carry ing two variant alleles; r_{im}: imprinting effect parameter with a single variant allele from mother; s_1: maternal effect with mother carrying one variant allele; s_2: maternal effect with mother carrying two variant allele. In addition, mating type probability of $(M, F) = (m, f)$ is denoted by μ_{ij}.



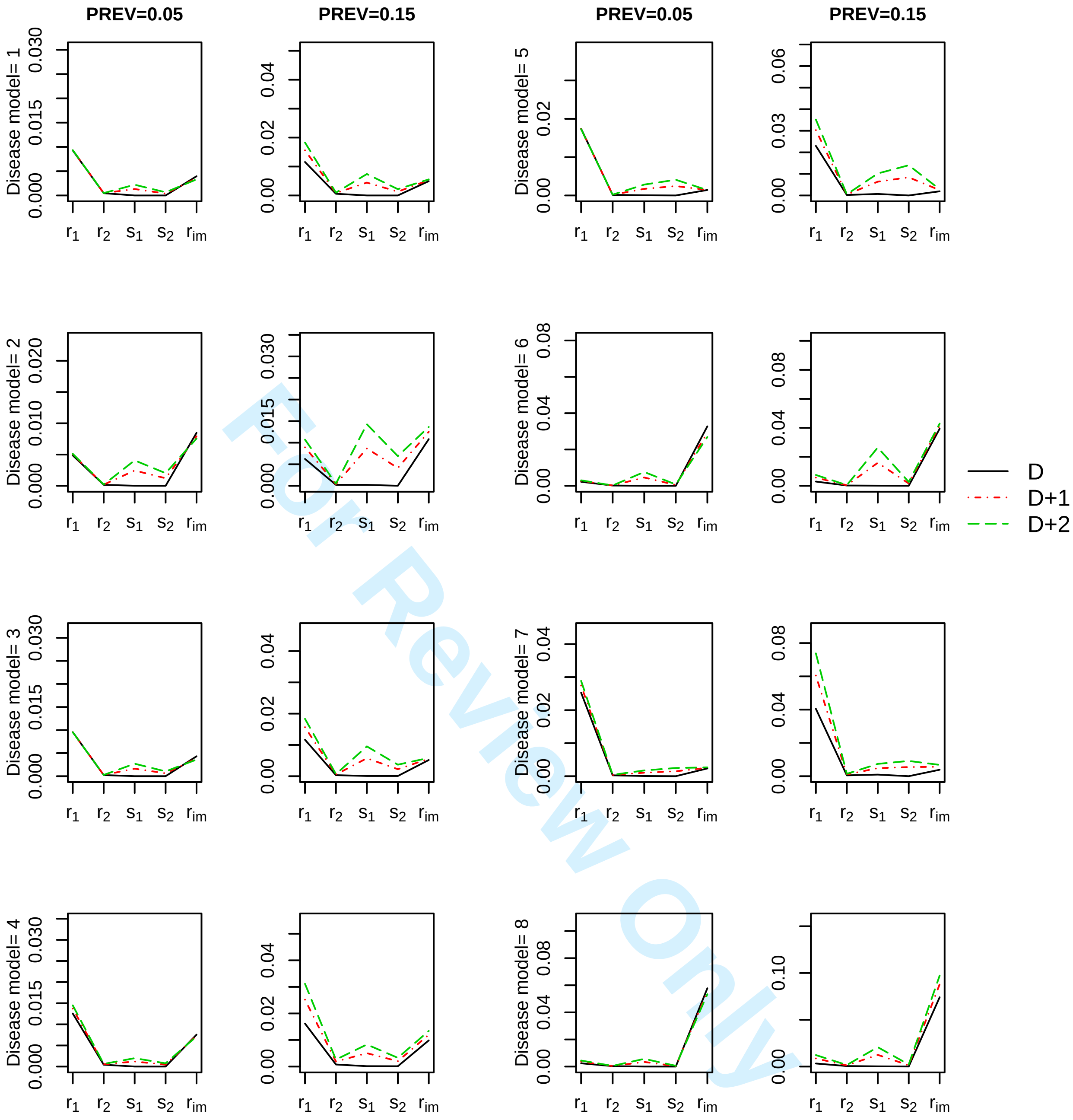
Supplementary Figure 1. Information content per family for 8 disease models and two PREVs when HWE holds and MAF is 0.3. Each curve provides the information for estimating one of the 5 parameters, for data types D, $D+1$ and $D+2$.



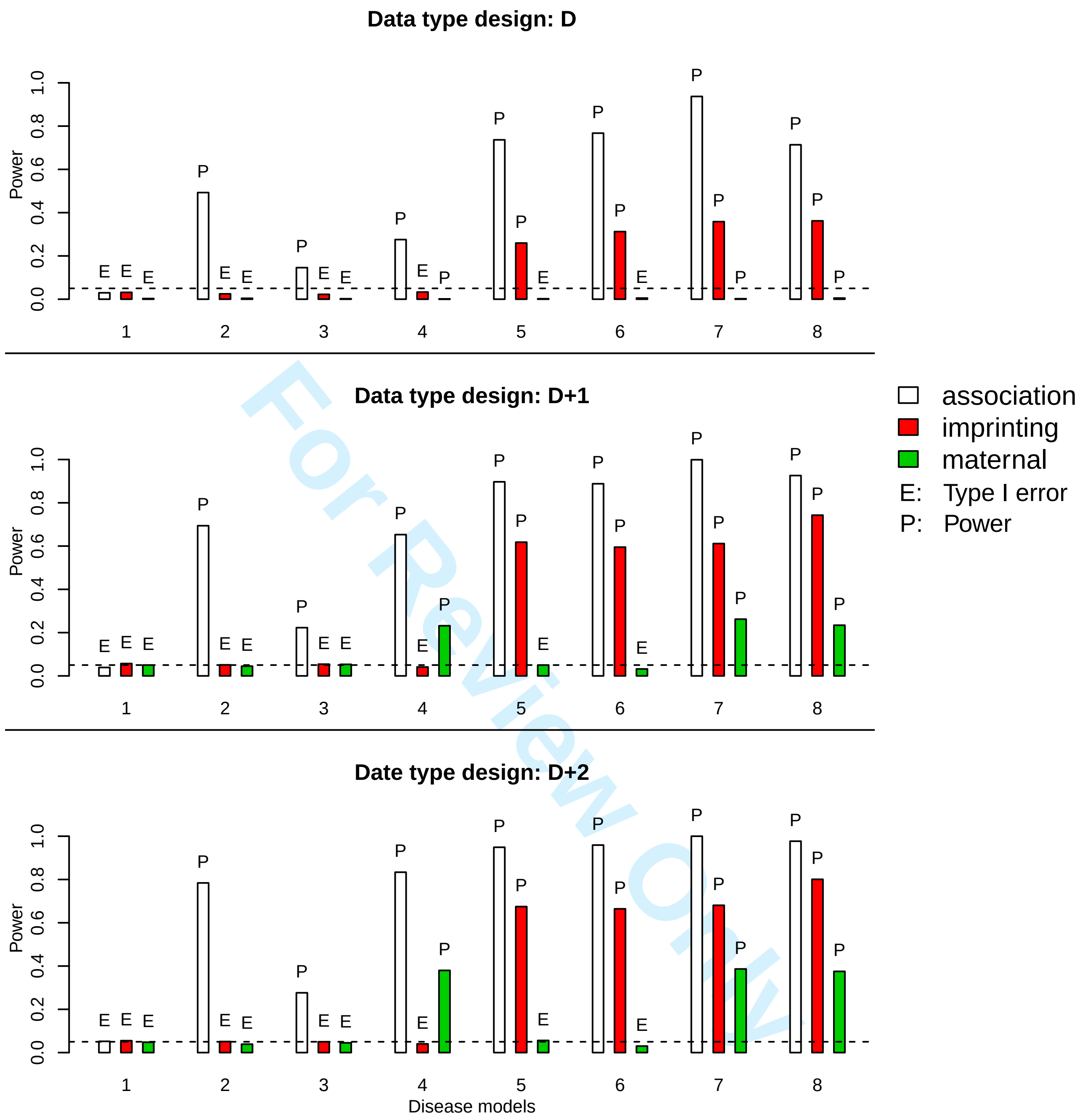
Supplementary Figure 2. Information content per individual for 8 disease models and two PREVs when HWE holds and MAF is 0.1. Each curve provides the information for estimating one of the 5 parameters, for data types D, $D+1$ and $D+2$.



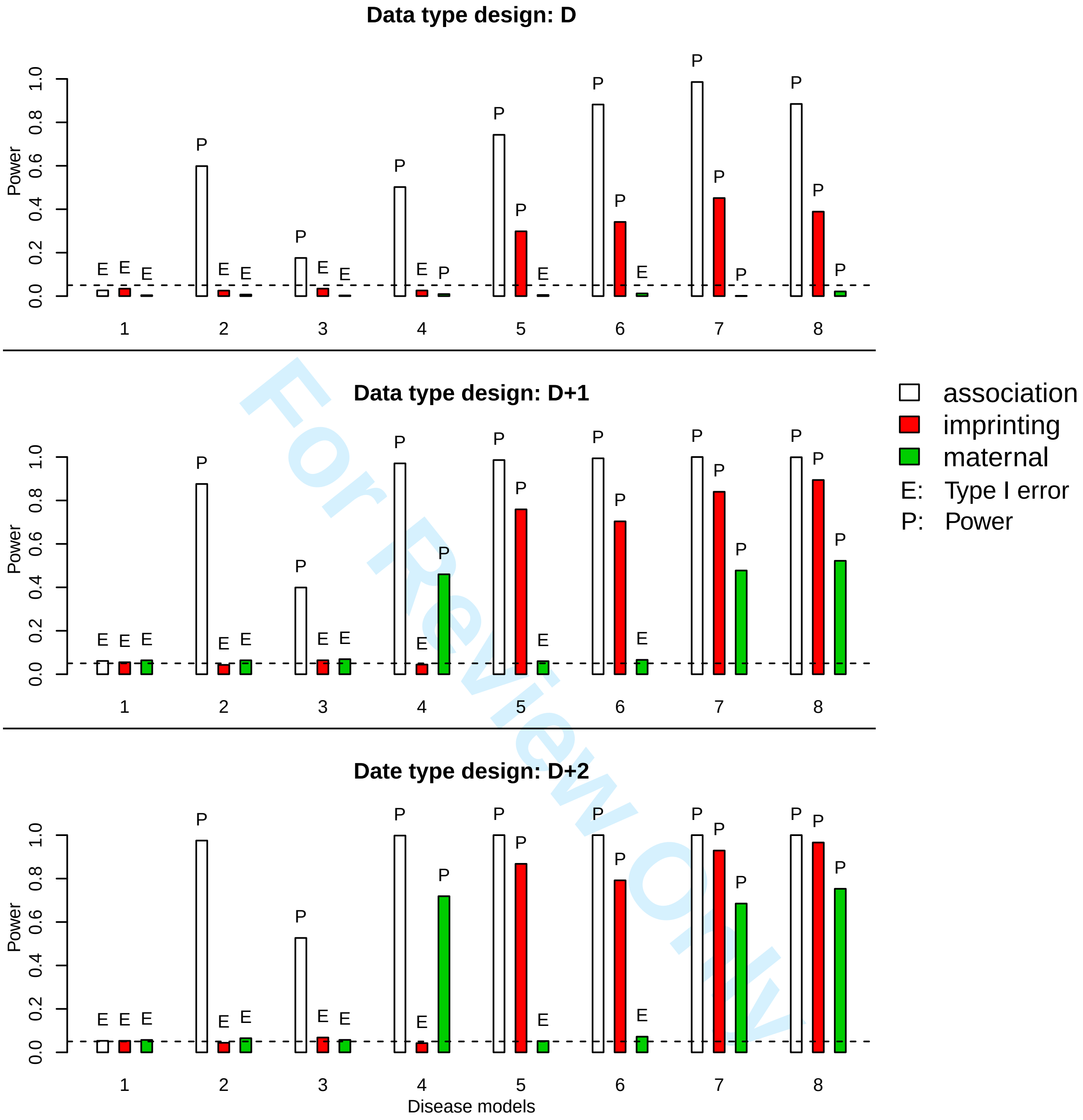
Supplementary Figure 3. Information content per individual for 8 disease models and two PREVs when HWE does not hold and MAF is 0.3. Each curve provides the information for estimating one of the 5 parameters, for data types D, $D+1$ and $D+2$.



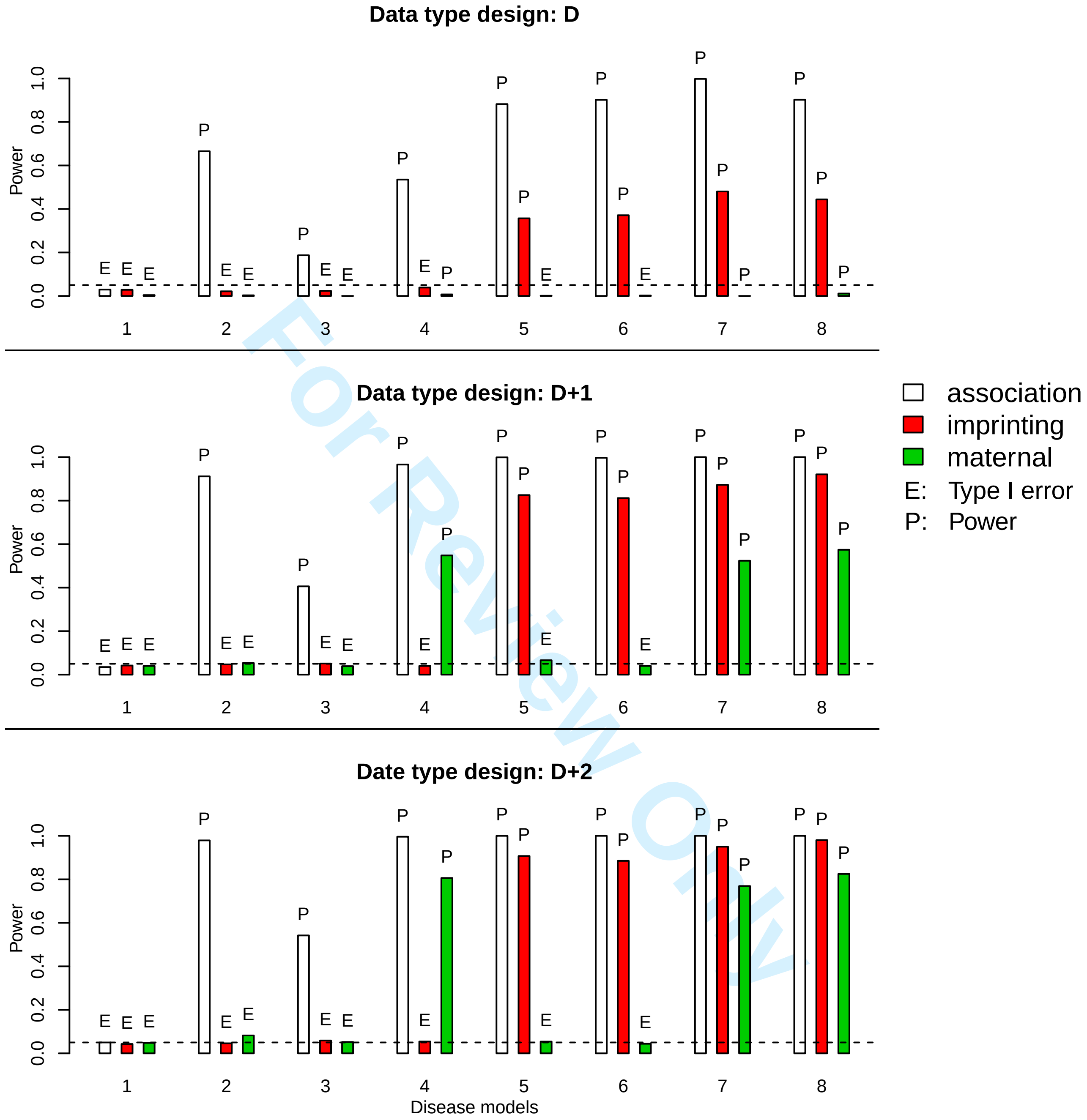
Supplementary Figure 4. Information content per individual for 8 disease models and two PREVs when HWE does not hold and MAF is 0.1. Each curve provides the information for estimating one of the 5 parameters, for data types D, $D+1$ and $D+2$.



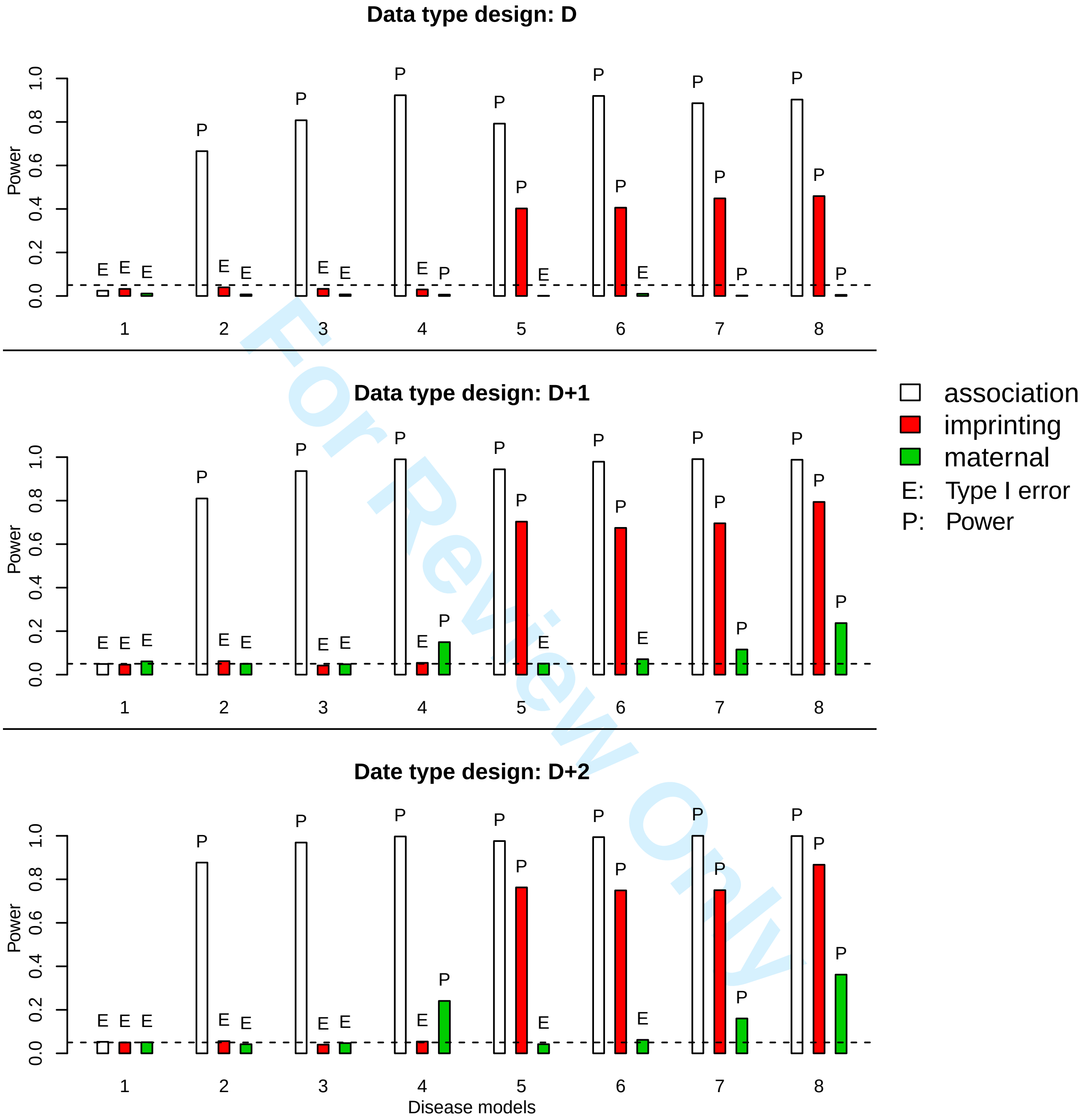
Supplementary Figure 5. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 2 as given in Table 2. Three rows represent three data types: D, $D+1$ and $D+2$. The bars of color white, red and green refer to association, imprinting effect and maternal effect. The horizontal line marks the nominal a level of 0.05.



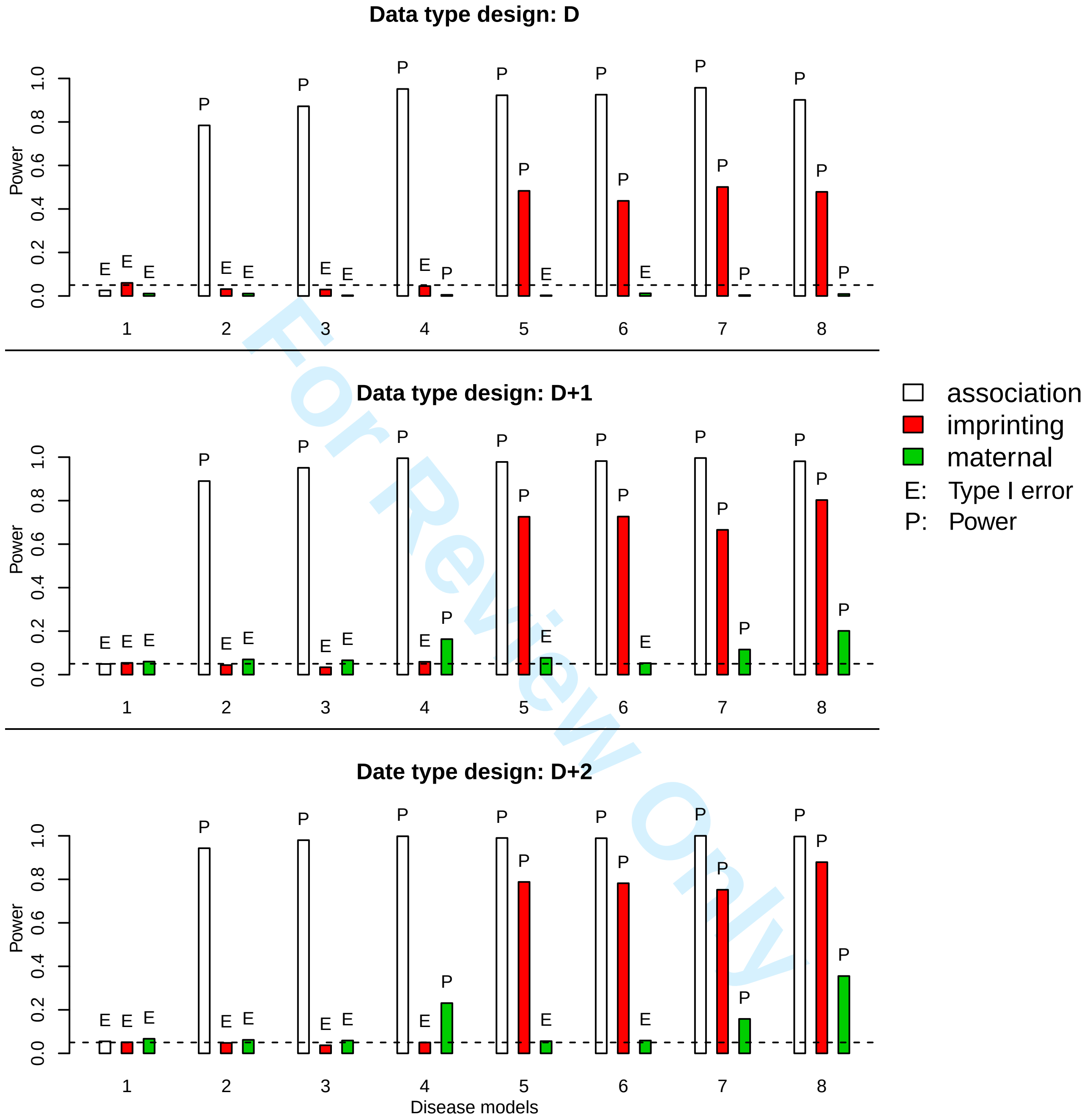
Supplementary Figure 6. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 3 as given in Table 2. Three rows represent three data types: D, $D+1$ and $D+2$. The bars of color white, red and green refer to association, imprinting effect and maternal effect. The horizontal line marks the nominal a level of 0.05.



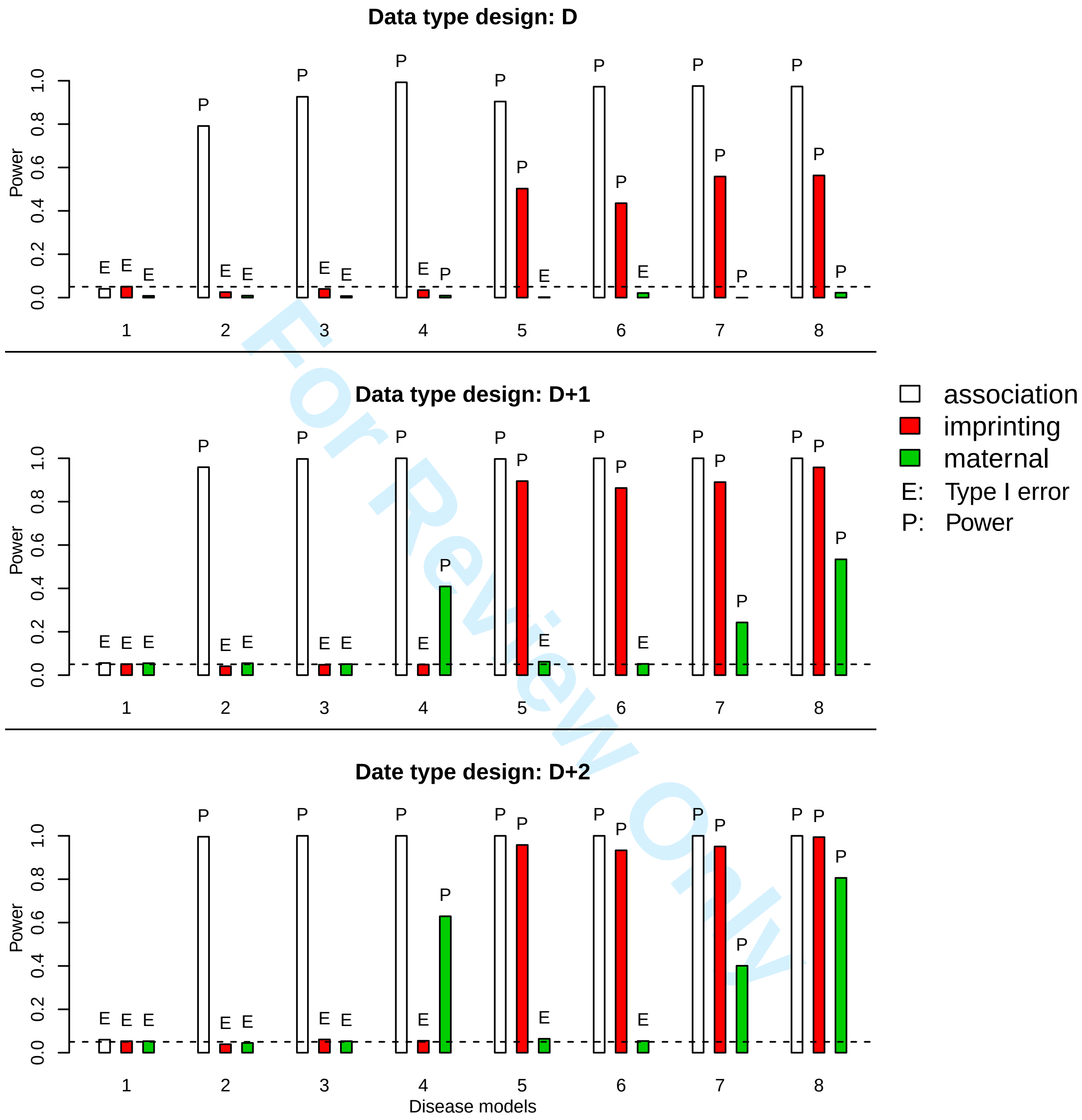
Supplementary Figure 7. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 4 as given in Table 2. Three rows represent three data types: D, $D+1$ and $D+2$. The bars of color white, red and green refer to association, imprinting effect and maternal effect. The horizontal line marks the nominal a level of 0.05.



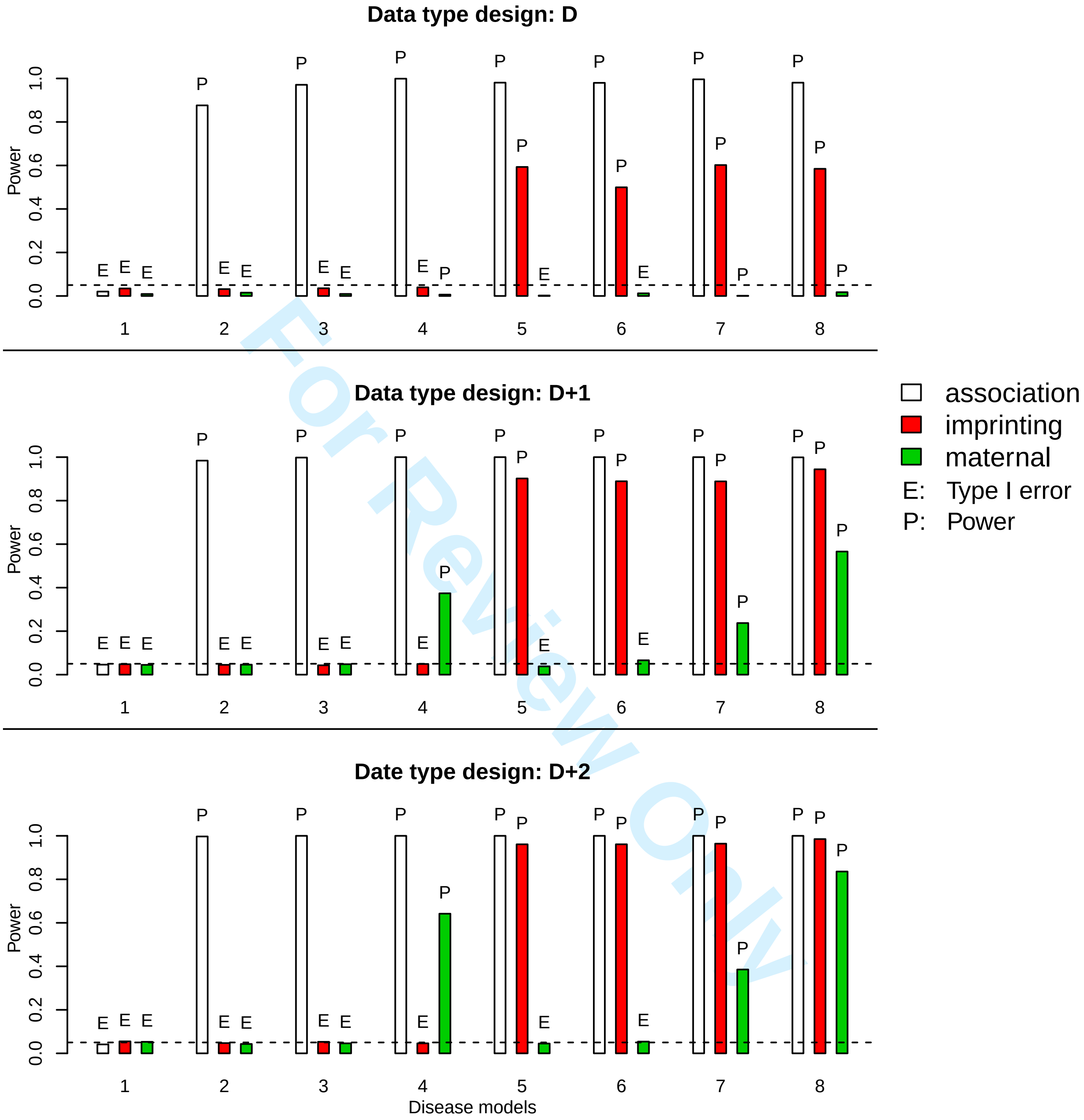
Supplementary Figure 8. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 5 as given in Table 2. Three rows represent three data types: D, $D + 1$ and $D + 2$. The bars of color white, red and green refer to association, imprinting effect and maternal effect. The horizontal line marks the nominal a level of 0.05.



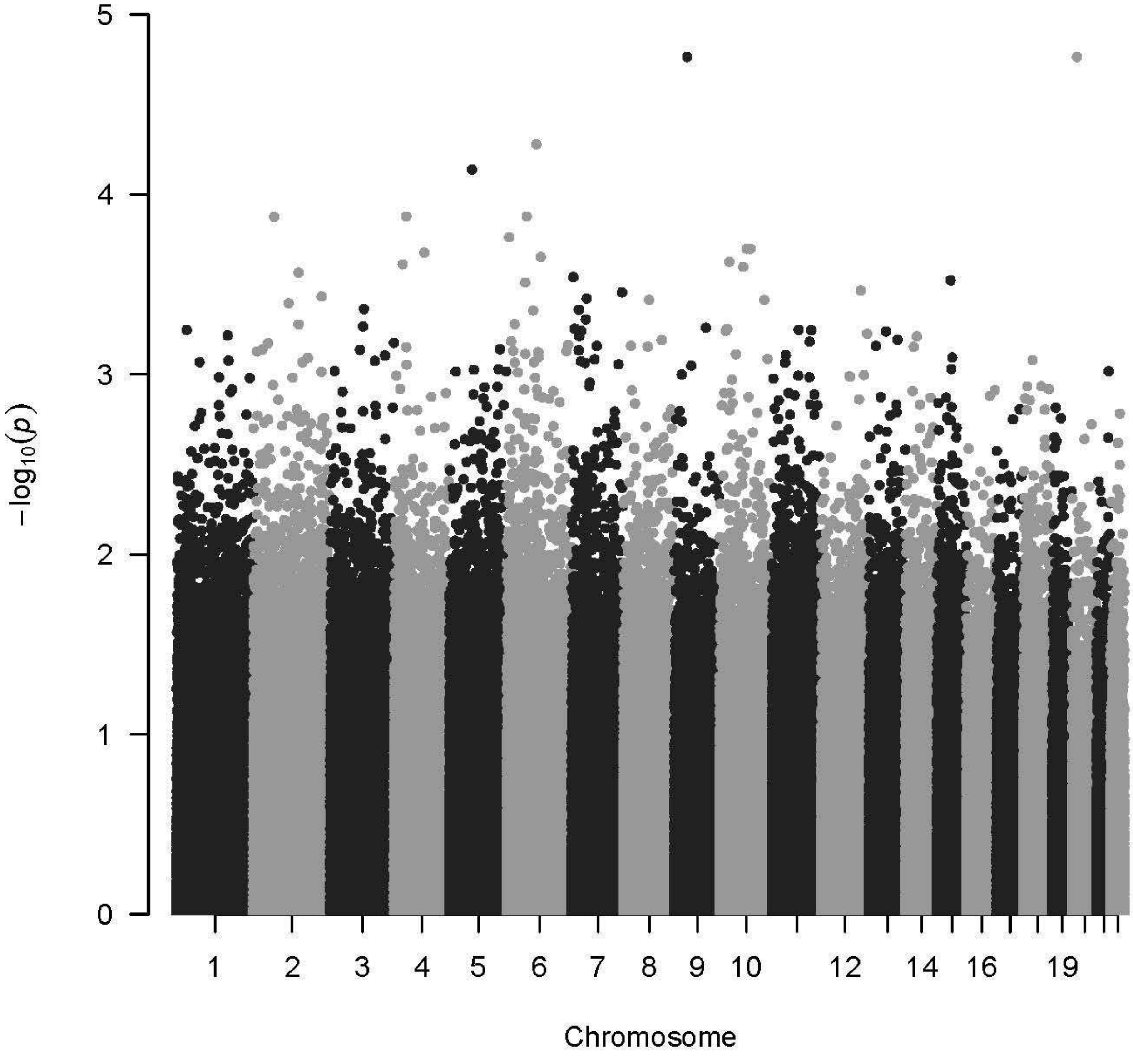
Supplementary Figure 9. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 6 as given in Table 2. Three rows represent three data types: D, $D+1$ and $D+2$. The bars of color white, red and green refer to association, imprinting effect and maternal effect. The horizontal line marks the nominal a level of 0.05.



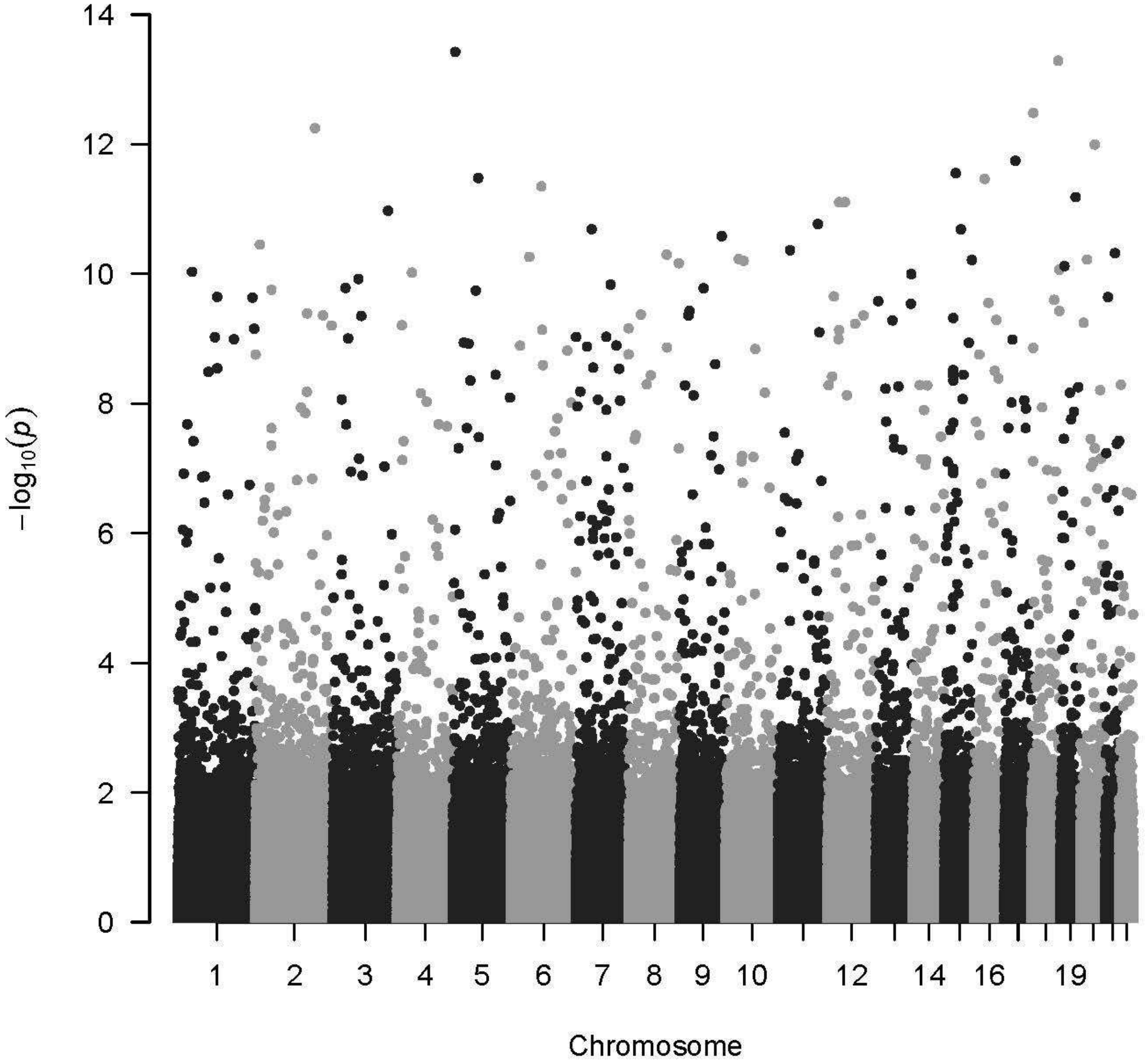
Supplementary Figure 10. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 7 as given in Table 2. Three rows represent three data types: D, $D + 1$ and $D + 2$. The bars of color white, red and green refer to association, imprinting effect and maternal effect. The horizontal line marks the nominal a level of 0.05.



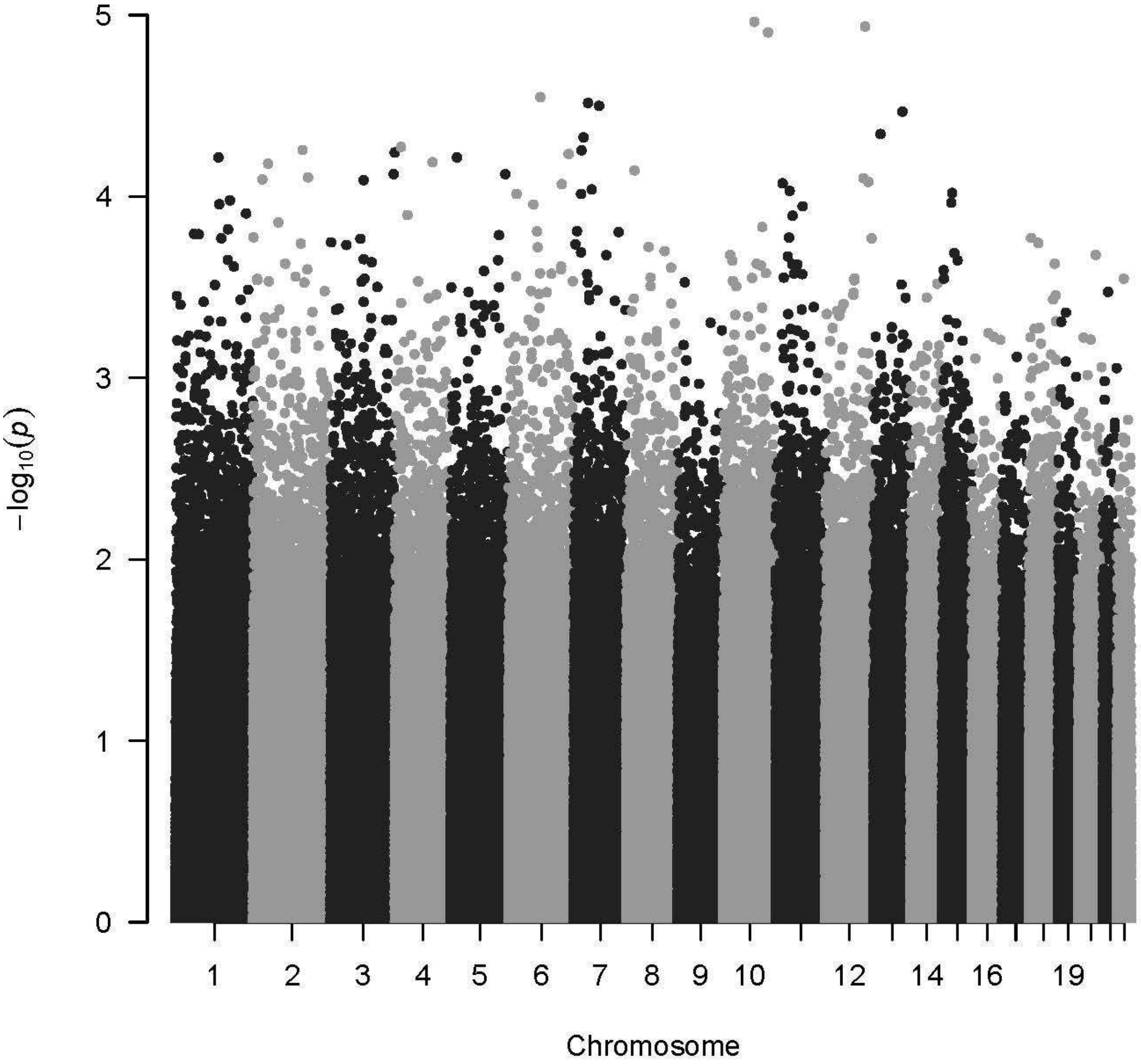
Supplementary Figure 11. Type I error rate and power of LIME$_{DSP}$ under 8 disease models and scenario 8 as given in Table 2. Three rows represent three data types: D, $D+1$ and $D+2$. The bars of color white, red and green refer to association, imprinting effect and maternal effect. The horizontal line marks the nominal a level of 0.05.



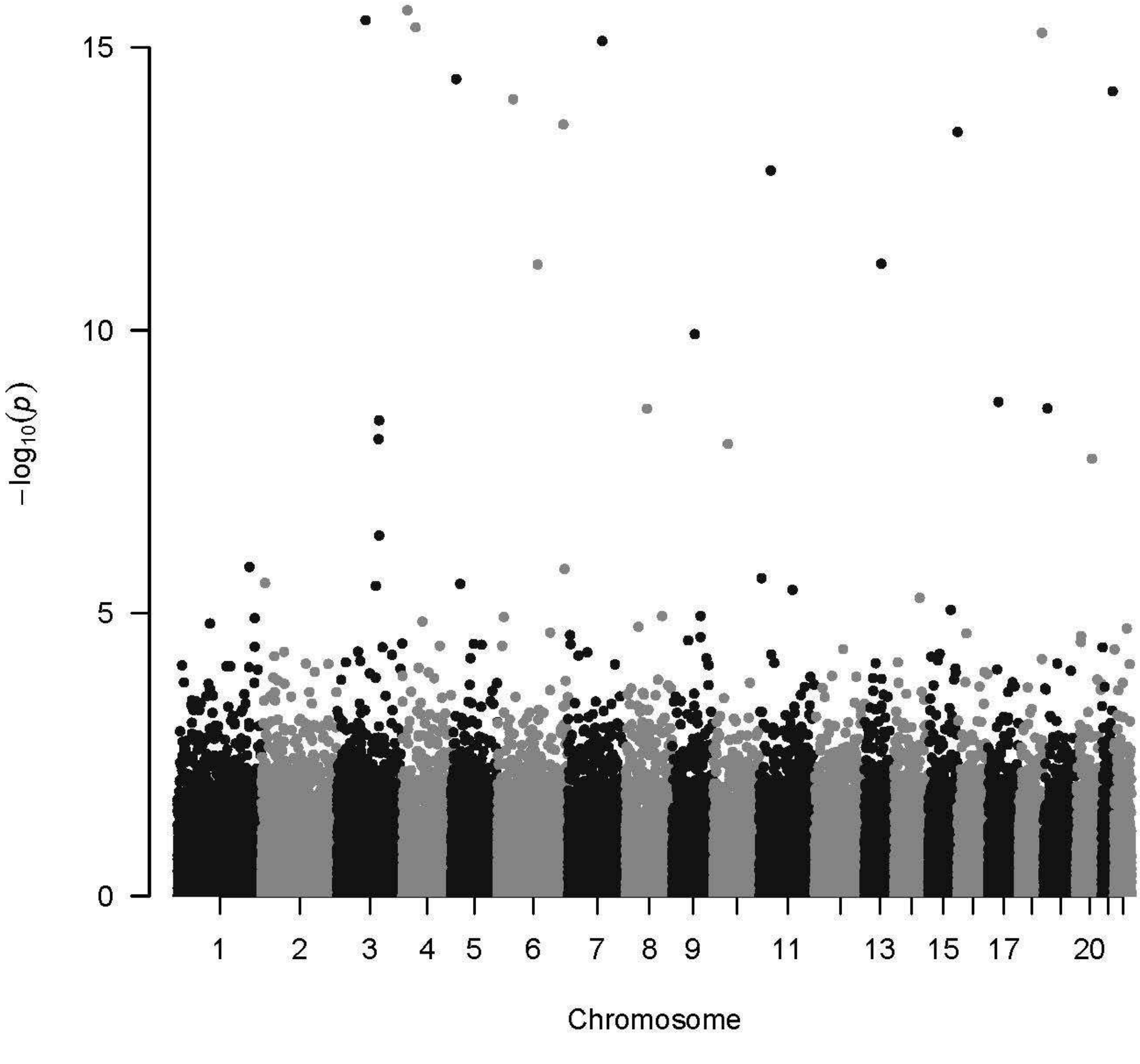
Supplementary Figure 12. Manhattan plot of -$\log_{10}$(p-value) for tests of association effect on club foot.



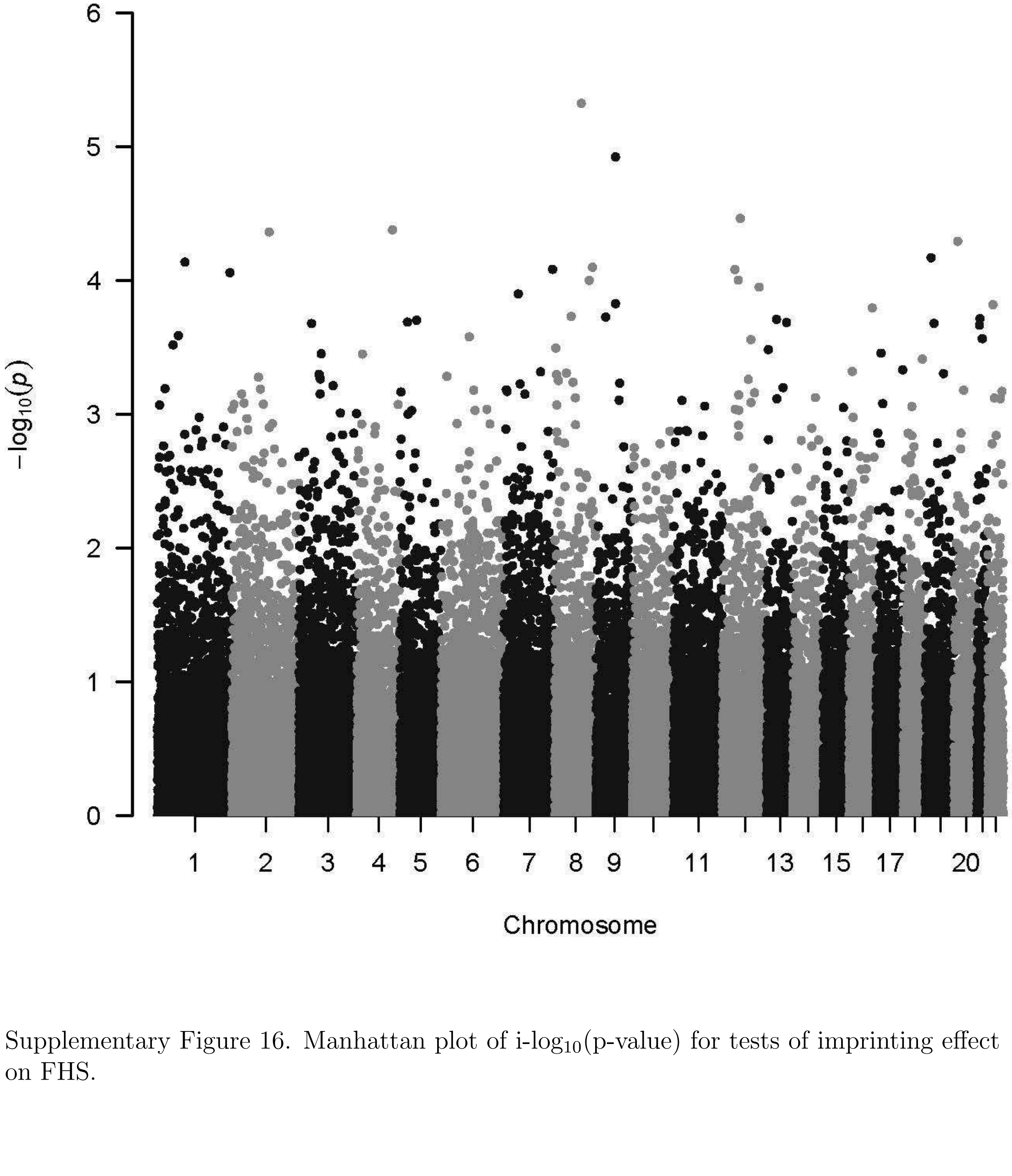
Supplementary Figure 13. Manhattan plot of -$\log_{10}$(p-value) for tests of imprinting effect on club foot.



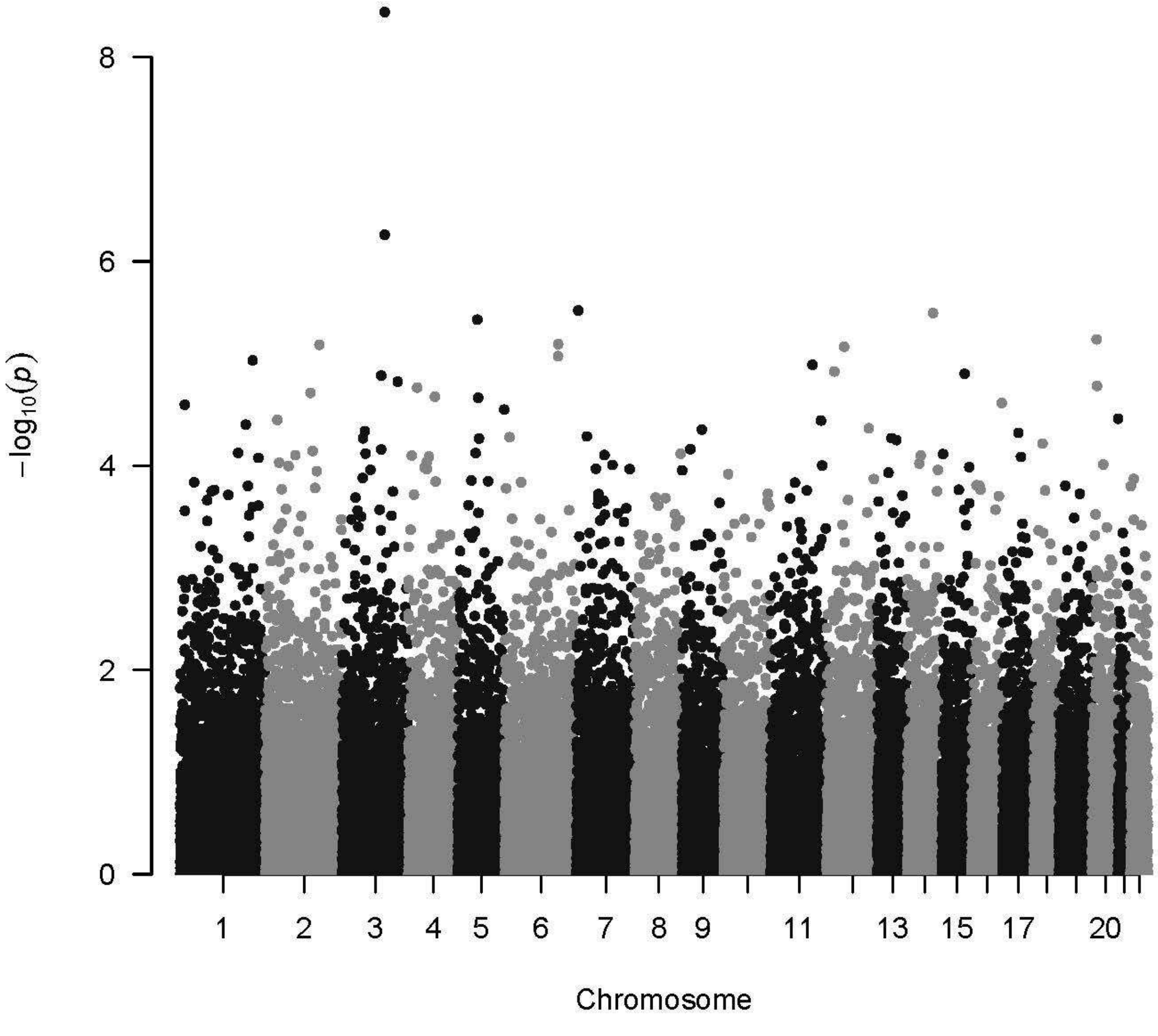
Supplementary Figure 14. Manhattan plot of -$\log_{10}$(p-value) for tests of maternal effect on club foot.



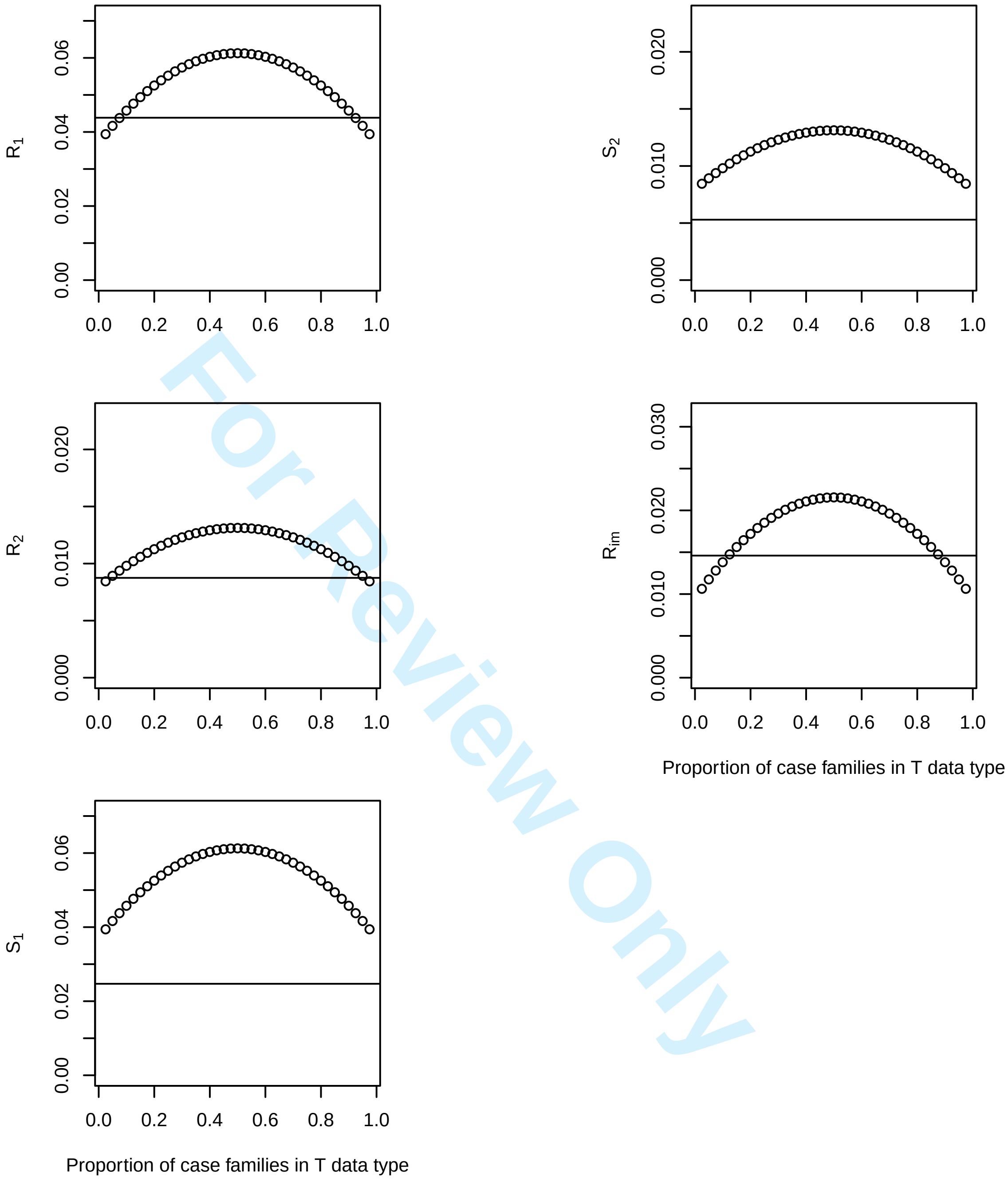
Supplementary Figure 15. Manhattan plot of -$\log_{10}$(p-value) for tests of association effect on FHS.



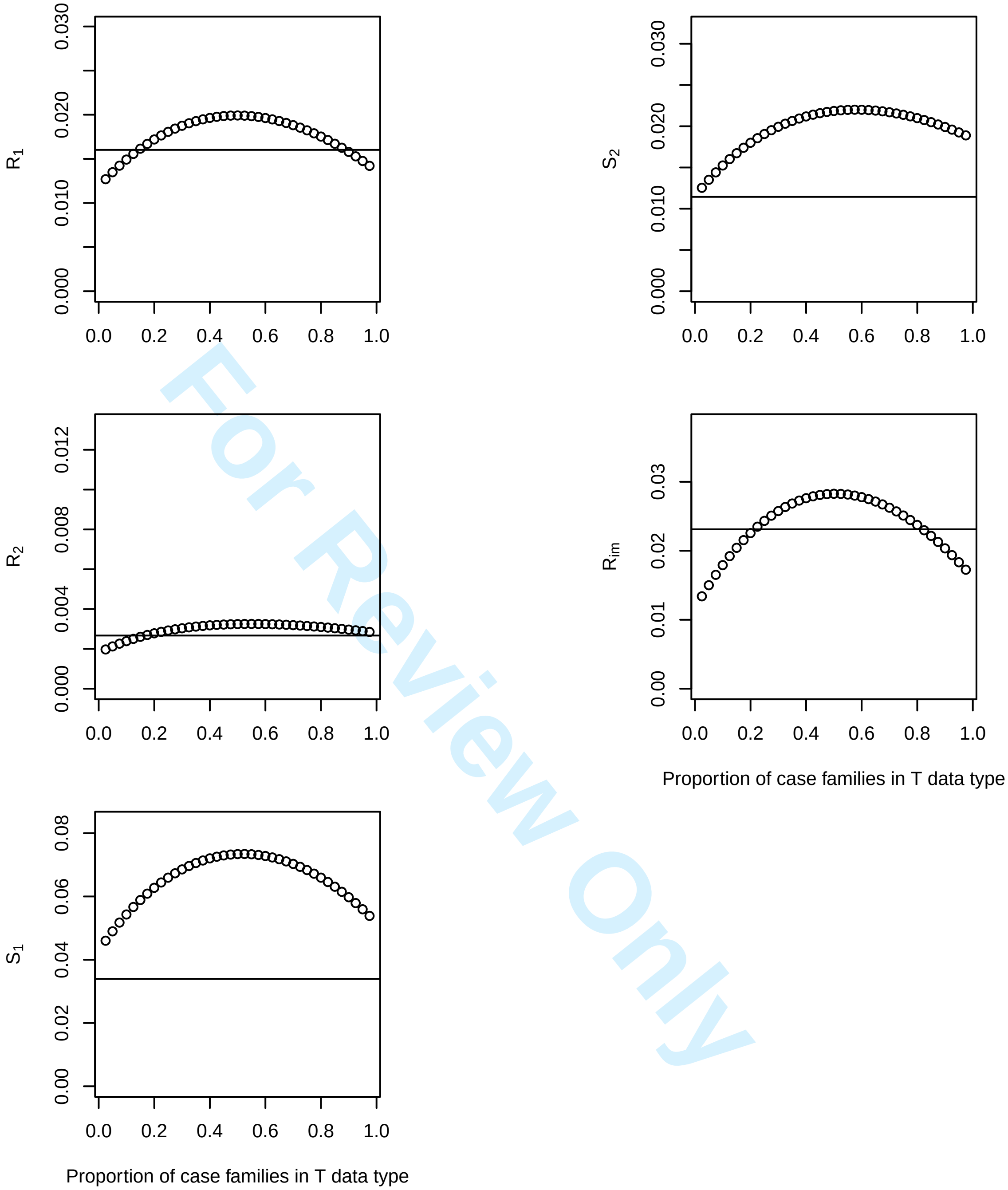
Supplementary Figure 16. Manhattan plot of i-$\log_{10}$(p-value) for tests of imprinting effect on FHS.



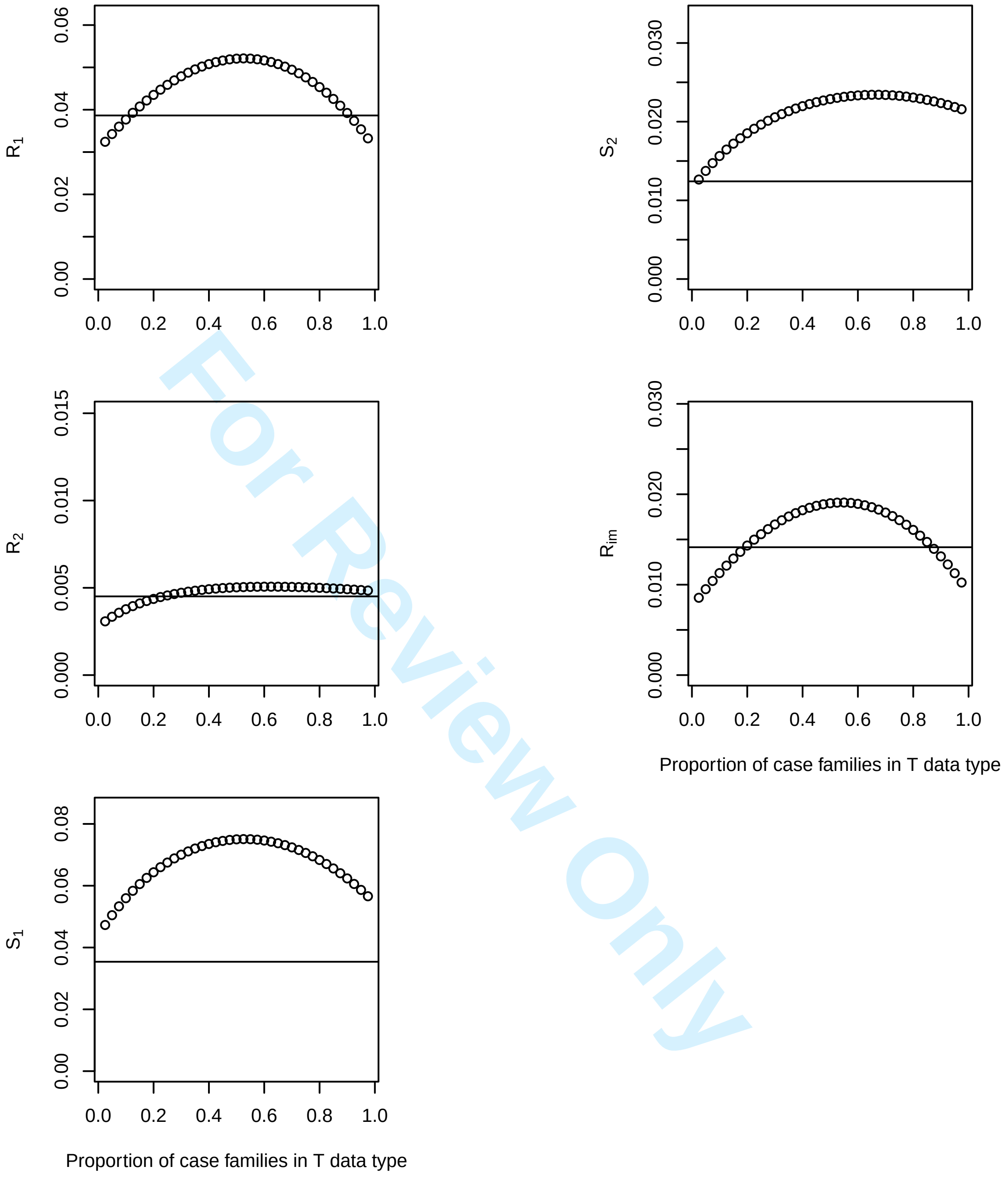
Supplementary Figure 17. Manhattan plot of -$\log_{10}$(p-value) for tests of maternal effect on FHS.



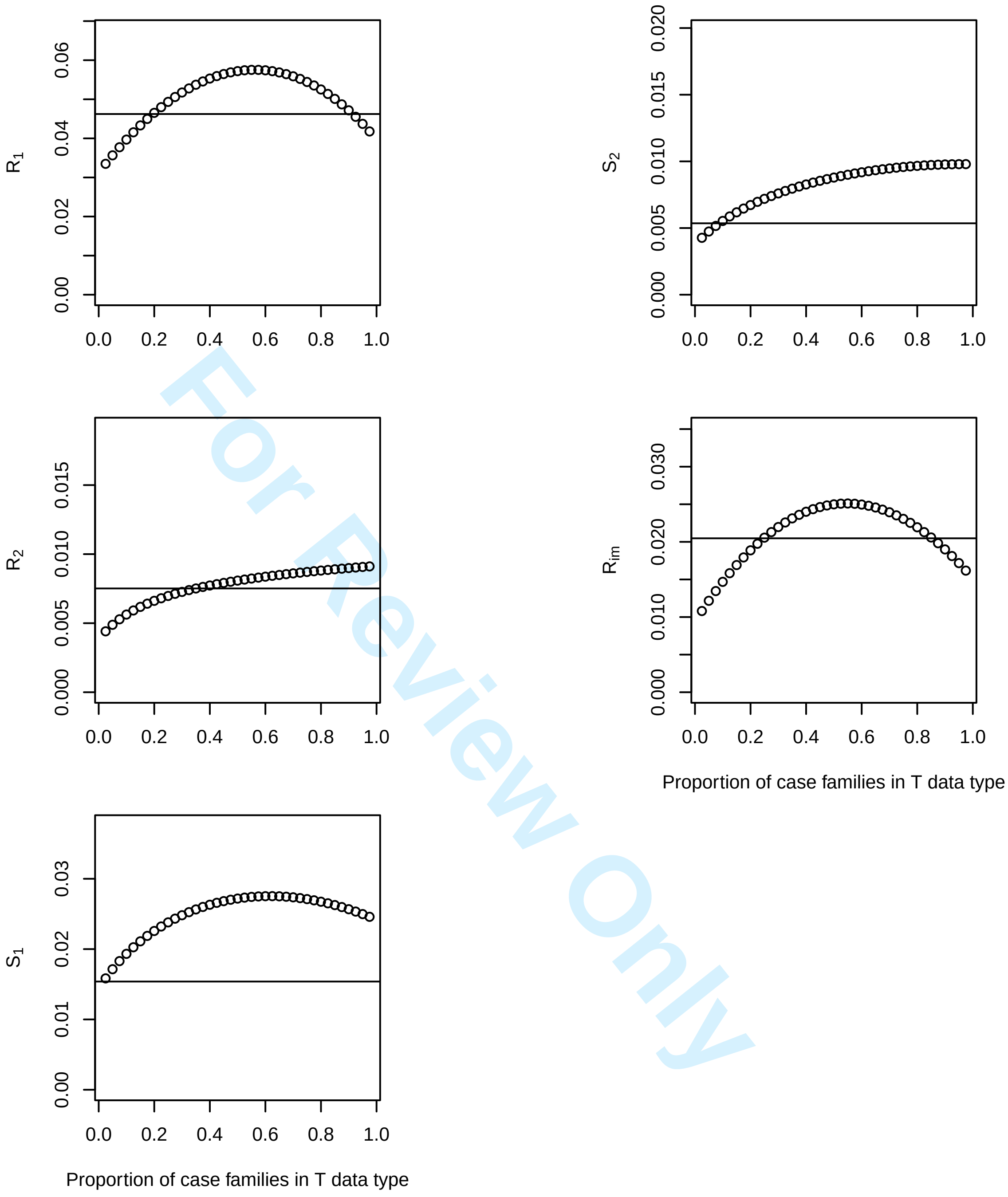
Supplementary Figure 18. Information content per individual for inference of parameters under disease model 1 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.



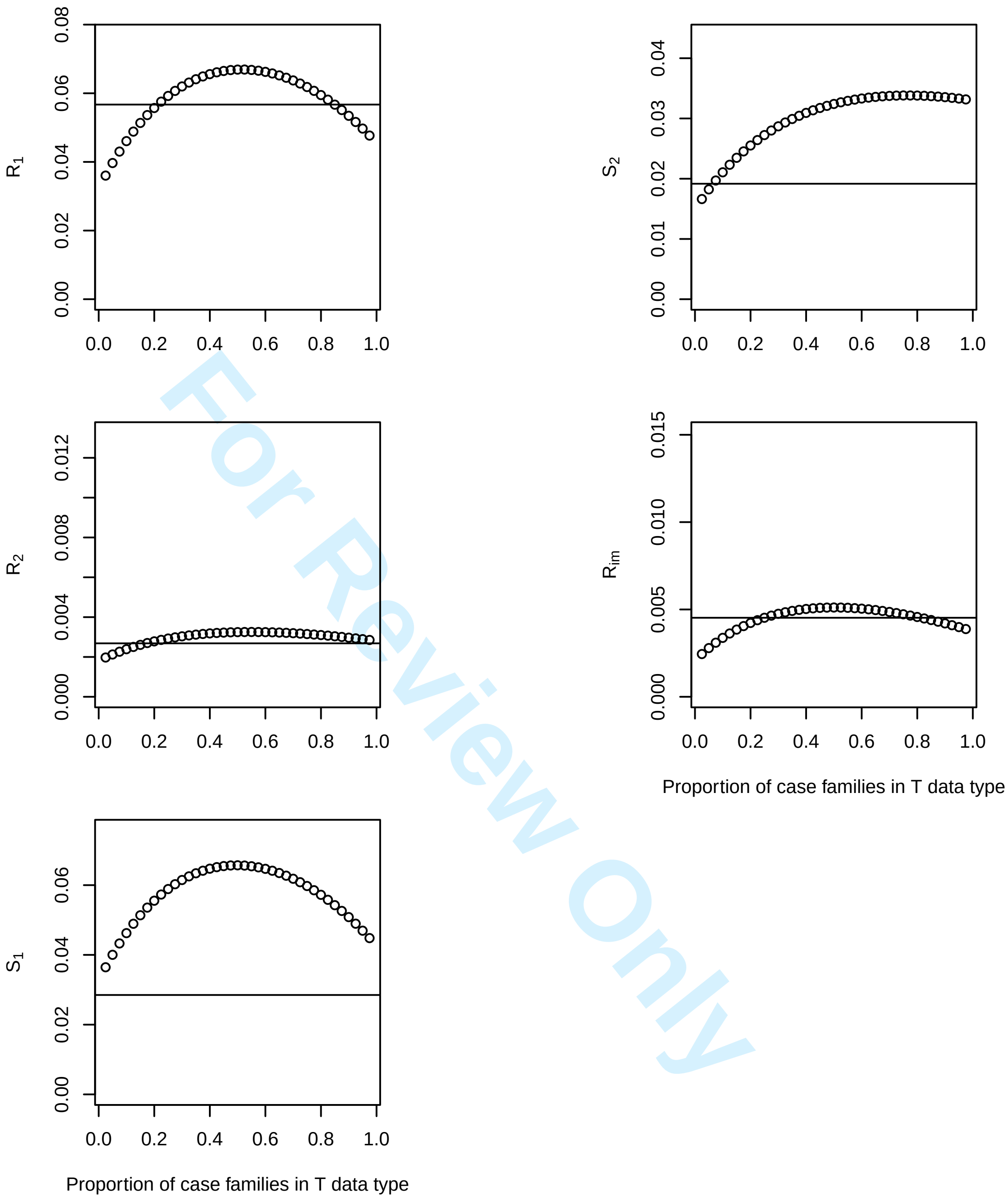
Supplementary Figure 19. Information content per individual for inference of parameters under disease model 2 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.



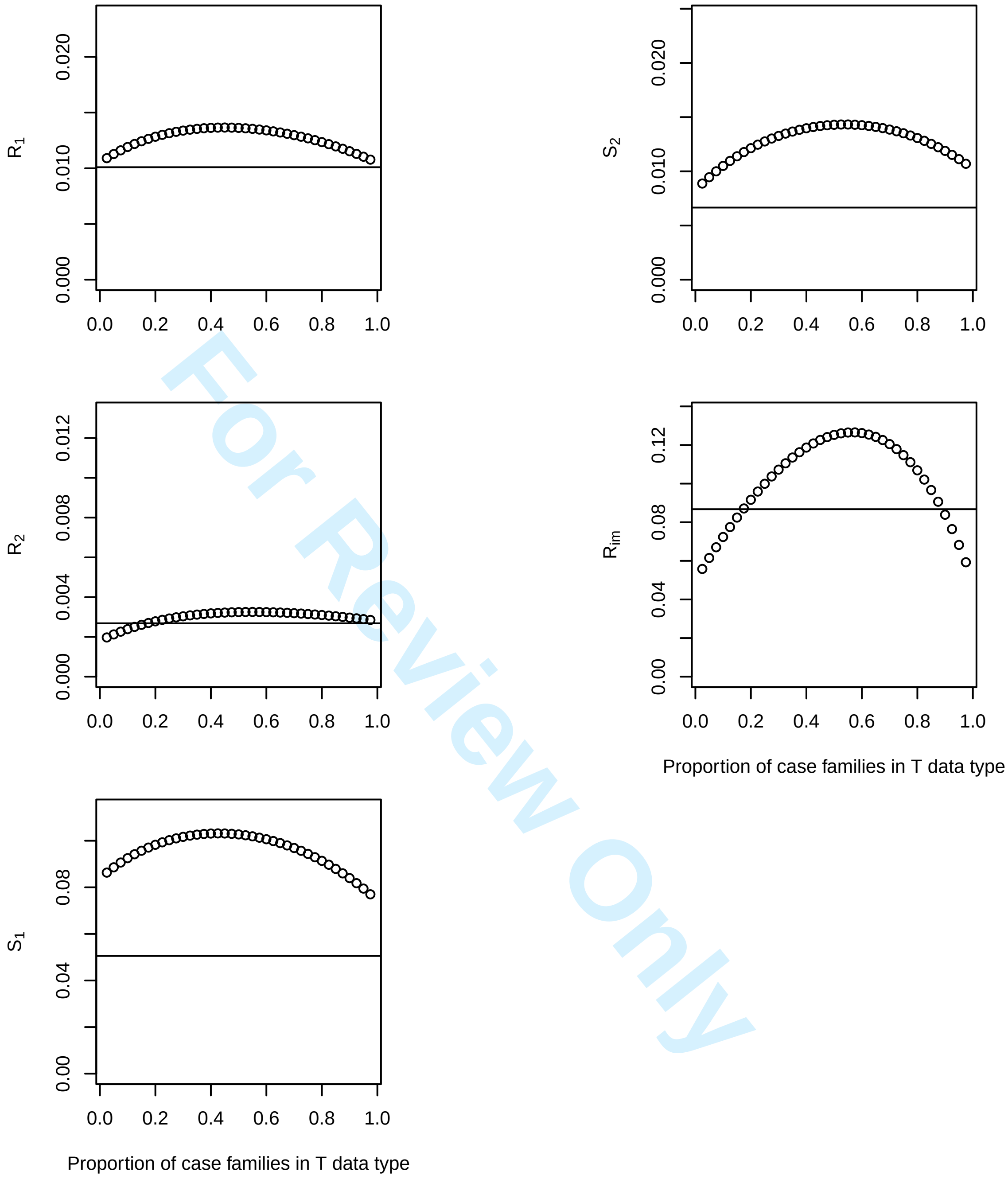
Supplementary Figure 20. Information content per individual for inference of parameters under disease model 3 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.



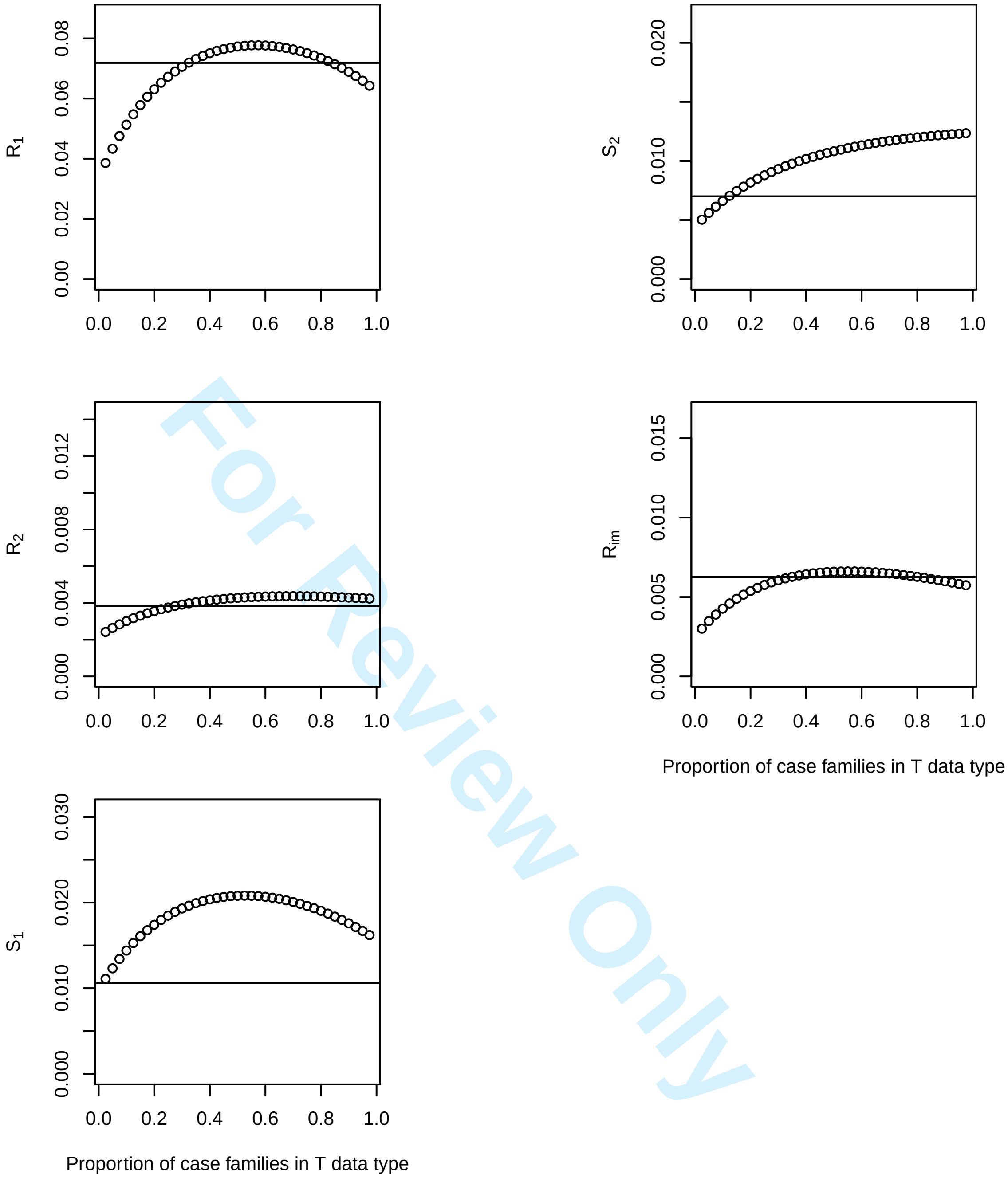
Supplementary Figure 21. Information content per individual for inference of parameters under disease model 4 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.



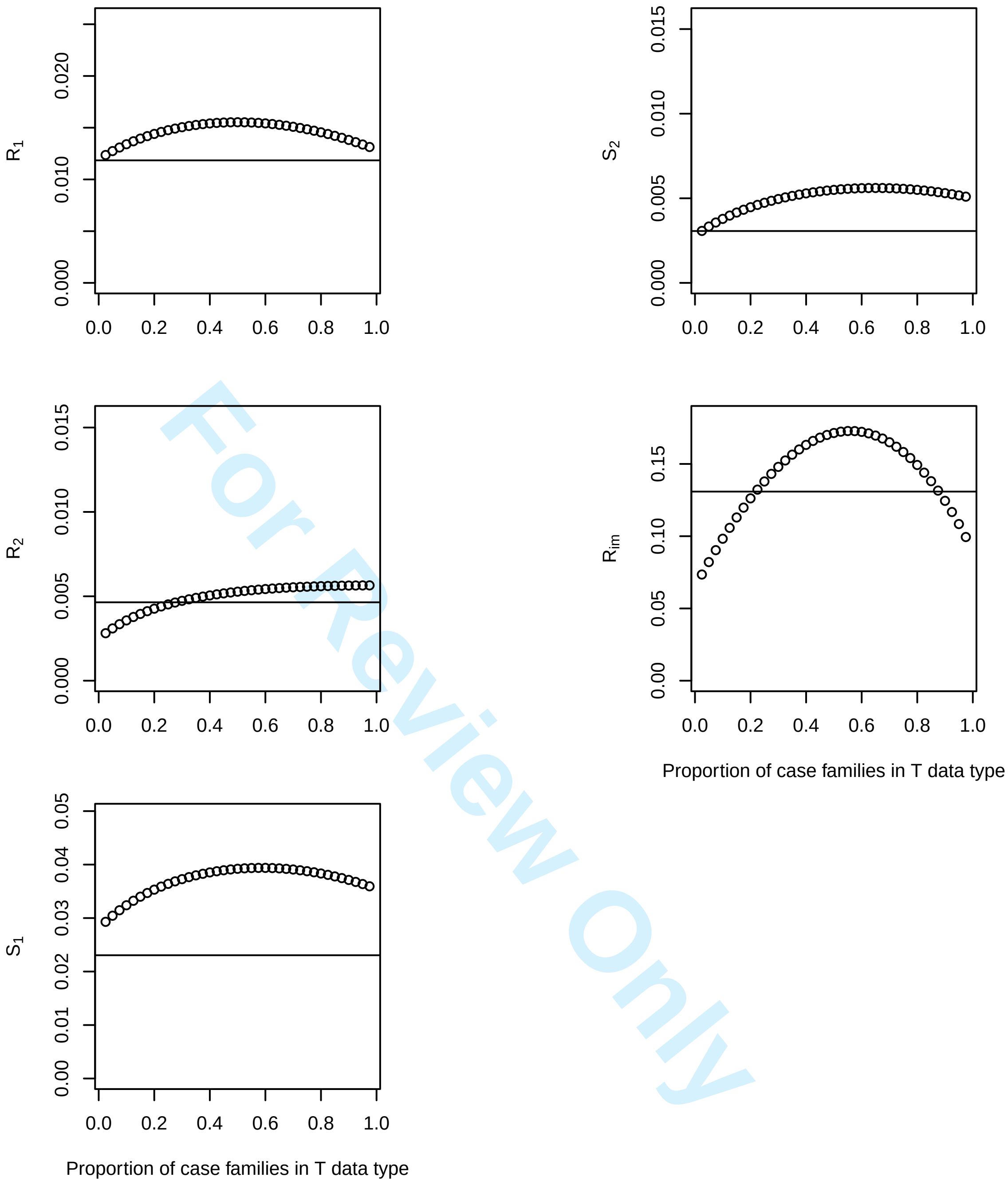
Supplementary Figure 22. Information content per individual for inference of parameters under disease model 5 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.



Supplementary Figure 23. Information content per individual for inference of parameters under disease model 6 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.



Supplementary Figure 24. Information content per individual for inference of parameters under disease model 7 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.



Supplementary Figure 25. Information content per individual for inference of parameters under disease model 8 and scenario 8. The horizontal line refers the information content per individual for LIME$_{DSP}$ applying to the D+2 design. The small circles represent information content per individual for LIME when applied to the T+3 design, with the proportion of case families varying from 0.025 to 0.975 by 0.025.