



Incorporating information from markers in LD with test locus for detecting imprinting and maternal effects

Fangyuan Zhang¹ · Shili Lin²

Received: 8 July 2019 / Revised: 26 November 2019 / Accepted: 4 February 2020 / Published online: 20 February 2020
© The Author(s), under exclusive licence to European Society of Human Genetics 2020

Abstract

Numerous statistical methods have been developed to explore genomic imprinting and maternal effects by identifying parent-of-origin patterns in complex human diseases. However, because most of these methods only use available locus-specific genotype data, it is sometimes impossible for them to infer the distribution of parental origin of a variant allele, especially when some genotypes are missing. In this article, we propose a two-step approach, LIMEhap, to improve upon a recent partial likelihood inference method. In the first step, the distribution of the missing genotypes is inferred through the construction of haplotypes by using information from nearby loci. In the second step, a partial likelihood method is applied to the inferred data. To substantiate the validity of the proposed procedures, we simulated data in a genomic region of gene *GPXI*. The results show that, by borrowing genetic information from nearby loci, the power of the proposed method can be close to that with complete genotype data at the locus of interest. Since the inference on the genotype distribution is made under the assumption of Hardy–Weinberg Equilibrium (HWE), we further studied the robustness of LIMEhap to violation of HWE. Finally, we demonstrate the utility of LIMEhap by applying it to an autism dataset.

Introduction

Genomic imprinting and maternal effects are both important epigenetic factors that have been explored as potential sources of heritability unexplained by genome-wide association studies. Genomic imprinting and maternal effects are involved in many complex human diseases, including Prader–Willi, Beckwith–Weidemann, and Angelman syndromes [1], and childhood cancers [2]. In this paper, we focus on developing an extension to a statistical method so that it is more powerful for detecting imprinting and maternal effects; therefore, we only describe these two epigenetic factors briefly that is sufficient for the purpose of

explaining our work. Complete and detailed descriptions of their complex mechanisms can be found in the literature [3–5]. Genomic imprinting refers to the process of differential epigenetic DNA modifications of the parental alleles, which leads to unequal expression of a heterozygous genotype depending on whether the variant allele is inherited from the mother or from the father [3]. Genomic imprinting may vary among individuals (a phenomenon termed polymorphic imprinting) depending on the stages of development, tissues, genetic background, and environment [4]. A maternal genotype effect, on the other hand, is a phenomenon wherein the phenotype of an individual is influenced by the genotype of the mother, not merely the allelic copy inherited. Maternal effects usually occur due to the additional mRNAs or proteins passed from the mother to the fetus during pregnancy and remained in the child after birth. Although imprinting and maternal effects arise from two different biological processes, their effects can mask one another [5]. Thus, it is important that these two confounding effects are considered together using statistical methods to avoid false positives/negatives ([6] and references therein).

A number of statistical methods have been proposed to study imprinting and maternal effects jointly, dated back to the work by Weinberg et al. [7]. Nevertheless, this and most

Supplementary information The online version of this article (<https://doi.org/10.1038/s41431-020-0590-3>) contains supplementary material, which is available to authorized users.

✉ Shili Lin
shili@stat.osu.edu

¹ Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX, USA

² Department of Statistics, The Ohio State University, Columbus, OH, USA

of the work that followed make the assumption of mating symmetry to avoid overparametrization so that hypothesis testing on the existence of imprinting and/or maternal genotype effects can be carried out [6]. The most frequently used design for studying maternal and imprinting effects is that of complete triads, where the genotypes of mother, father, and child are all required to be observed [8] to facilitate the identification of the parental origin of the allele of interest. In practice, incomplete triads are common, as one parent may be unavailable for genetic study: parents may be deceased or unavailable; they may refuse to participate in the study; or the father may need to be excluded post hoc due to nonpaternity. Although the need for complete triad data has been relaxed in more recent methods so that mother–child pairs may also be included in a study, data are nonetheless not fully utilized [9–11].

Among the methods that allow for missing father genotypes, LIME, a partial Likelihood inference method for detecting both imprinting and maternal effects, stands out as it does not require assumptions about mating type probabilities [11]. Furthermore, nonproband siblings may also be included in a study in addition to the child proband in triads and pairs. However, when a father's genotype is missing, LIME has to ignore pairs in which both the mother and the child's genotypes are heterozygous, as parental origin of the child's allele cannot be determined and thus the parameters of interest and the nuisance parameters are not separable.

While Yang and Lin [11] demonstrated that ignoring mother–child pair both with heterozygous genotypes may not lead to a significant loss of information in some scenarios for LIME, incorporating genotype information for markers that are in linkage disequilibrium (LD) with the test locus can generally help infer parental origin [12–14]. As such, we propose LIMEhap, an approach that adopts the methodology of LIME, but utilizes additional information from nearby markers that are in LD with the test locus to help infer missing genotypes. Therefore, LIMEhap can detect both imprinting and maternal effects and distinguish between them by using all the complete triad data and incomplete mother–child pair data when available. To substantiate the validity of the proposed procedure and to quantify the potential gain in power, we carried out an extensive simulation study by considering 32 settings. Our results show that, by inferring the parental origin of the minor allele at the test locus through borrowing genetic information from nearby loci, the power of the proposed method can be close to that using complete genotype data at the test locus with well-controlled type I error rate. This illustrates that the use of nearby marker in LD with the test locus helps resolve parental origin ambiguity to a great extent, consistent with observations in the literature. We further study the robustness of LIMEhap to

violation of Hardy–Weinberg Equilibrium (HWE), as inference on the genotype distribution is made under the assumption of HWE. Finally, we applied LIMEhap to an autism spectrum disorder dataset to demonstrate its practical utility.

Materials and methods

Enrichment of test locus information

Suppose a disease locus has two alleles A and a , where a (typically the minor allele, that is, with a smaller frequency than A) is the variant allele. In a nuclear family, we use F , M , and C to represent the genotypes of father, mother, and a child, respectively. The genotypes AA , Aa , and aa are coded as 0, 1, or 2, respectively, representing the number of the variant allele. We also use 1_m to indicate a child with genotype Aa whose minor allele a is inherited from the mother.

We use a multiplicative risk model for the disease penetrance:

$$P(D = 1|M, F, C) = \delta R_1^{I(C=1)} R_2^{I(C=2)} R_{im}^{I(C=1_m)} S_1^{I(M=1)} S_2^{I(M=2)}, \quad (1)$$

where δ is the phenocopy rate; R_1 and R_2 denote the effects of one or two copies of an individual's own variant allele, respectively; R_{im} denotes the imprinting effect; S_1 and S_2 denote the effects of one or two copies of the mother's variant allele, respectively; and finally, D is an indicator variable denoting the disease status of the child, with 1 being affected and 0 being unaffected. As shown in Yang and Lin [11], all the parameters are identifiable and estimable; therefore, we can distinguish maternal effects from imprinting effects using the model as specified in (1). We denote the vector of parameters of interest by $\theta = (\delta, R_1, R_2, R_{im}, S_1, S_2)^\top$ and the vector of nuisance parameters (including mating type probabilities, to be elaborated in the following) by ϕ . We note that this model is assumed to apply to all individuals; therefore, the methodology is expected to suffer from power loss in the presence of polymorphic imprinting.

When the father's genotype is missing, there are seven genotype combinations of the (M, C) pair, listed in Table 1. Therein, the μ 's are mating type probabilities; that is, $\mu_{ij} = P(M = i, F = j)$, $i, j = 0, 1, 2$. As we can see from the table, for the first six categories, the mating type probabilities (the nuisance parameters) and the risk parameters can be completely separated as multiplicative factors in the joint probabilities of mother–child genotypes and child's disease status (either $D = 1$ or $D = 0$). However, for the 7th category, $(M, C) = (1, 1)$, separation cannot be achieved. This is

Table 1 Joint probabilities of disease status and genotypes, with further expansion of the $(M,C) = (1, 1)$ category and sample size^a for each category.

Type	M	F	C	$P(D = 1, M, C)$	$P(D = 0, M, C)$	#cases	#controls
1	0	.	0	$(\mu_{00} + \frac{1}{2}\mu_{01})\delta$	$(\mu_{00} + \frac{1}{2}\mu_{01})(1 - \delta)$	n_{00}^1	n_{00}^0
2	0	.	1	$(\mu_{02} + \frac{1}{2}\mu_{01})\delta R_1$	$(\mu_{02} + \frac{1}{2}\mu_{01})(1 - \delta R_1)$	n_{01}^1	n_{01}^0
3	1	.	0	$(\frac{1}{2}\mu_{10} + \frac{1}{4}\mu_{11})\delta S_1$	$(\frac{1}{2}\mu_{10} + \frac{1}{4}\mu_{11})(1 - \delta S_1)$	n_{10}^1	n_{10}^0
4	1	.	2	$(\frac{1}{4}\mu_{11} + \frac{1}{2}\mu_{12})\delta R_2 S_1$	$(\frac{1}{4}\mu_{11} + \frac{1}{2}\mu_{12})(1 - \delta R_2 S_1)$	n_{12}^1	n_{12}^0
5	2	.	1	$(\mu_{20} + \frac{1}{2}\mu_{21})\delta R_1 S_2 R_{im}$	$(\mu_{20} + \frac{1}{2}\mu_{21})(1 - \delta R_1 S_2 R_{im})$	n_{21}^1	n_{21}^0
6	2	.	2	$(\frac{1}{2}\mu_{21} + \mu_{22})\delta R_2 S_2$	$(\frac{1}{2}\mu_{21} + \mu_{22})(1 - \delta R_2 S_2)$	n_{22}^1	n_{22}^0
7	1	.	1	$\mu_{11}\frac{1}{4}\delta R_1 S_1(1 + R_{im}) + \mu_{10}\frac{1}{2}\delta R_1 S_1 R_{im} + \mu_{12}\frac{1}{4}\delta R_1 S_1$	$\mu_{11}\frac{1}{4}[1 - \delta R_1 S_1(1 + R_{im})] + \mu_{10}\frac{1}{2}[1 - \delta R_1 S_1 R_{im}] + \mu_{12}\frac{1}{4}[1 - \delta R_1 S_1]$	$n_{11h}^1 + n_{101h}^1 + n_{121h}^1$	$n_{11h}^0 + n_{101h}^0 + n_{121h}^0$
(7-1)	1	0	1	$\mu_{10}\frac{1}{2}\delta R_1 S_1 R_{im}$	$\mu_{10}\frac{1}{2}[1 - \delta R_1 S_1 R_{im}]$	n_{101h}^1	n_{101h}^0
(7-2)	1	2	1	$\mu_{12}\frac{1}{2}\delta R_1 S_1$	$\mu_{12}\frac{1}{2}[1 - \delta R_1 S_1]$	n_{121h}^1	n_{121h}^0
(7-3)	1	1	1	$\mu_{11}\frac{1}{4}\delta R_1 S_1(1 + R_{im})$	$\mu_{11}\frac{1}{4}[1 - \delta R_1 S_1(1 + R_{im})]$	n_{111h}^1	n_{111h}^0

^a $n_{f1/h}^0$ is the rounded sum of the probabilities of all case-mother pairs with $(M, F, C) = (1, f, 1), f = 0, 1, 2$. $n_{f1/h}^0$ is defined similarly for control-mother pairs. $\mu_{mf} = P(M = m, F = f), m, f = 0, 1, 2$, are the mating type probabilities and are treated as nuisance parameters.

because when both mother and child are heterozygous, the parental origin of the child’s variant allele cannot be inferred when father’s genotype is missing. As such, LIME ignores data from this category, with a justification given in Yang and Lin [11], where they demonstrated that the loss of information with the ignorance of this category is not substantial in some situations. Nevertheless, it is important to recover as much information as possible, especially given the scarcity of family data compared with population case-control samples. For brevity in the main text, a detailed description of the LIME procedure is delegated to Supplementary Material S1. In the remaining of this subsection, we describe a method that uses haplotype information in the genomic region spanned by the test locus and nearby markers in LD with it to impute the missing father’s genotype so that such mother–child pairs can also be used.

For ease of exposition, we explain the idea in a simple scenario with only two loci. The method for multiple loci proceeds analogously. Suppose the test locus has two alleles, A and a , and a nearby marker also has two alleles, B and b . We assume that these two loci are in high LD so that there is no recombination between these two loci when transmitting the chromosomal segment from parents to offspring. There are four possible haplotypes formed by these two loci: $AB, Ab, aB,$ and ab . Suppose in a triad, the mother’s unphased genotype is (Aa, Bb) , where the first entry in the parentheses denotes the genotype at the test locus and the second denotes the genotype at the marker locus. For the child, the genotype is (Aa, BB) . Then at the test locus, $(M, C) = (1, 1)$. Suppose further that haplotype ab occurs with a negligible frequency, then one can safely assume that the mother’s phased two-locus genotype is $Ab|aB$, where the haplotype before and after the “|” make up the haplotype pair of the mother. Since no recombination is allowed between these two loci, she must have passed aB to the child since the child has genotype BB at the additional marker locus. Therefore, the parental origin of the a allele in the child must be from the mother.

In a more complex scenario, we assume that all four haplotypes have frequencies that are too high to be ignored. In this situation, we still cannot determine the parental origin of the a allele even with additional information from the marker locus. For such cases, we will infer all compatible haplotype configurations for the family and compute their corresponding probabilities, as shown schematically in Table S1 and explained in the caption. By doing so, methods that cannot make use of the pair data with $(M, C) = (1, 1)$ at the test locus, such as LIME, can now borrow information from nearby markers to estimate the distribution of the familial genotype configurations, which provides a probabilistic approach to incorporate the partial information into these methods.

LIMEhap

In this section, we introduce LIMEhap, an extension of LIME, to make use of the $(M,C) = (1,1)$ category based on inferred haplotype distributions. Pairs with $(M,C) = (1,1)$ are processed first. Utilizing information from loci in LD with the test locus and assuming that there are no recombinations in the region spanned by these loci, we infer compatible haplotype configurations and their corresponding probabilities for the nuclear family (which may include additional siblings) using HAPLORE [15]. Since HWE is a necessary assumption in HAPLORE, we will explore the robustness of LIMEhap when the HWE assumption is violated.

Define n_{mc}^1 and n_{mc}^0 as the numbers of case-mother and control-mother pairs, respectively, with $M = m$ and $C = c$. The child in each of these pairs is a proband (either affected with $D = 1$ or unaffected with $D = 0$). Define sn_{mc}^1 and sn_{mc}^0 similarly for sibling-mother pairs, where these siblings are not probands. We assume that the total number of families derived from case-mother pairs, $n_p^1 = \sum_{m,c} n_{mc}^1$, and the total number of families derived from control-mother pairs, $n_p^0 = \sum_{m,c} n_{mc}^0$, are fixed. Then the likelihood from the observed data can be written, up to a proportionality, as

$$L(\theta, \phi) \propto \left\{ \prod_{m,c} P(m, c | D = 1)^{n_{mc}^1} P(m, c | D = 0)^{n_{mc}^0} P(D = 1 | m, c)^{sn_{mc}^1} P(D = 0 | m, c)^{sn_{mc}^0} \right\}$$

Note that in the above likelihood, for each case-mother or control-mother pair, the contribution is through a conditional probability conditioning on their affection status (first two factors in the likelihood), whereas for the additional siblings, their contributions are factored into the likelihood prospectively since they are not probands (last two factors in the likelihood).

Following the argument of LIME, one can extract the following partial likelihood that is free of the nuisance parameters:

$$L_p(\theta) = \left\{ \prod_{m,c} (p_{mc})^{n_{mc}^1} (1 - p_{mc})^{n_{mc}^0} \prod_{m,c} (q_{mc})^{sn_{mc}^1} (1 - q_{mc})^{sn_{mc}^0} \right\}, \tag{2}$$

where p_{mc} and q_{mc} for $(m,c) \neq (1,1)$ are as described in Supplementary Material S1, and we focus on defining these two quantities for $(m,c) = (1,1)$ in the following. We first note that in the original LIME method [11], the likelihood and partial likelihood do not include the $(m,c) = (1,1)$ category. Nevertheless, their argument for extracting the partial likelihood, now including all pairs, remains valid, although the corresponding p_{11} and q_{11} are defined differently and explained in the following.

We let $n_{11} = n_{11}^1 + n_{11}^0$ be the total number of nuclear families with $(m,c) = (1,1)$ and n_{1f1h}^1 and n_{1f1h}^0 be, respectively, the summation of the inferred probabilities of case-mother and control-mother pairs with triad genotype configuration $(m,f,c) = (1,f,1)$ at the test locus derived from HAPLORE as explained in subsection of enrichment of test locus information, where f may take the value of 0, 1, or 2. Following the original idea of LIME [11], we use $p_{1f1}(\theta)$ to represent the probability of being a case-parent triad among all families (including all case- and control-families) with $(m,f,c) = (1,f,1)$, $f = (0,1,2)$, and is defined as

$$p_{1f1}(\theta) = \frac{1}{s_{1f1}} \frac{n_p^1 P(D = 1 | M = 1, F = f, C = 1)}{P(D = 1)}, \quad f = 0, 1, 2,$$

where

$$s_{1f1}(\theta) = \frac{n_p^1 P(D = 1 | M = 1, F = f, C = 1)}{P(D = 1)} + \frac{n_p^0 P(D = 0 | M = 1, F = f, C = 1)}{P(D = 0)}$$

is also free of the nuisance parameters. Then, the probability of being a case-mother pair among families with $(m,c) = (1,1)$, $p_{11}(\theta)$, is the weighted average of the p_{1f1} 's with the weight for each $f(=0,1,2)$ being the proportion of families with $(m,f,c) = (1,f,1)$: $(n_{1f1h}^1 + n_{1f1h}^0)/n_{11}$. That is,

$$p_{11}(\theta) = p_{101} \frac{n_{101h}^1 + n_{101h}^0}{n_{11}} + p_{111} \frac{n_{111h}^1 + n_{111h}^0}{n_{11}} + p_{121} \frac{n_{121h}^1 + n_{121h}^0}{n_{11}}. \tag{3}$$

Similarly, we define sn_{1f1h}^1 and sn_{1f1h}^0 as, respectively, the summation of the inferred probabilities of affected or unaffected sibling-mother pairs with triad genotype configuration $(m,f,c) = (1,f,1)$ at the test locus derived from HAPLORE. Then the penetrance probability for $(m,c) = (1,1)$ is calculated as the weighted average of disease penetrance $q_{1f1}(\theta) = P(D = 1 | M = 1, F = f, C = 1)$ with the weight for each $f(=0,1,2)$ being the proportion of the additional siblings with $(m, f, c) = (1,f,1)$: $(sn_{1f1h}^1 + sn_{1f1h}^0)/sn_{11}$. That is,

$$q_{11}(\theta) = q_{101} \frac{sn_{101h}^1 + sn_{101h}^0}{sn_{11}} + q_{111} \frac{sn_{111h}^1 + sn_{111h}^0}{sn_{11}} + q_{121} \frac{sn_{121h}^1 + sn_{121h}^0}{sn_{11}}.$$

It is noteworthy to point out that the above idea for taking care of the $(M,C) = (1,1)$ category is in fact parallel to the way that LIME handles the $(M,F,C) = (1,1,1)$ category, where the uncertain origin of the a allele is being

attributed to father or mother, each with probability 0.5. Thus, the disease penetrance is a weighted average of the two possibilities with the weights being their respective probabilities (50% for each). For LIMEhap here, we have three possibilities, and we use the inferred probabilities from HAPLORE as the weights.

Simulation study

We carry out a simulation study to evaluate the performance of LIMEhap and compare its information gain over LIME. To generate data that reflect LD structure in real situations, we consider 5 SNPs in the *GPXI* gene (shown in Supplementary Table S2), where the underlying haplotypes have been previously constructed and used in other studies, including those that investigated the properties of methods for detecting parent-of-origin effects [13, 16, 17]. We consider four scenarios, which are combinations of two levels of population disease prevalence $P(D = 1)$ (PREV) {0.05, 0.15}, and two levels of HWE {not hold = 0, hold = 1}. 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 [18]. 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.

To facilitate this investigation, we consider a total of eight disease models (Table 2). The first three models portrait no imprinting nor maternal effects. Model 4 has maternal effect only, models 5 and 6 have imprinting effect only, and model 7 and 8 have both parent-of-origin effects. With the specification of each scenario and a disease model,

Table 2 Parameters for eight disease models used in the simulation study.

Model	Parameters				
	R_1	R_2	R_{im}	S_1	S_2
1	1	1	1	1	1
2	2	3	1	1	1
3	1	3	1	1	1
4	1	3	1	2	2
5	1	3	3	1	1
6	3	3	1/3	1	1
7	1	3	3	2	2
8	3	3	1/3	2	2

R_1 : relative risk of carrying one copy of the variant allele.

R_2 : relative risk of carrying two copies of the variant allele.

R_{im} : imprinting effect parameter with a single variant allele from mother; $R_{im} > 1 (< 1, \text{ or } = 1)$ represents paternal imprinting (maternal imprinting, or no imprinting effect).

S_1 : maternal effect with mother carrying one copy of the variant allele.

S_2 : maternal effect with mother carrying two copies of the variant allele.

the penetrance probability in (1) is fully specified. In all, we consider a total of 32 settings (four scenarios and eight disease models).

We generate 1000 replications under each of these 32 settings. Each replicate consists of 150 case families and 150 control families. Firstly, parental haplotypes are generated. Then, the haplotypes of their proband children are created according to the transmission probability assuming no recombination. Affection status D of the probands are determined by a Bernoulli trial, with the success probability calculated based on (1) with M, F, C as mother, father, and child's genotypes at the test locus, which are deduced completely from the generated haplotypes. A family with an affected child is recruited as a case family, whereas a family with an unaffected child is recruited as a control family. The process of generating M, F, C , and D is repeated until we have collected sufficient numbers of case and control families to meet the preset sample size. This process returns triad data with affected or unaffected probands. To generate pair data, we remove all the fathers' genotype information from the triad data. In addition, we also consider the scenario with one additional nonproband sibling, whose haplotype is generated based on the parents' haplotypes and transmission probability as for the probands, and whose disease status is assigned prospectively based on (1). We denote P and T as pair data without the $(M, C) = (1, 1)$ category and triad data, respectively. We further denote $P + 1$ and $T + 1$ as pair and triad families, each with one additional sibling. On the other hand, we use $P + hap$, $P + 1 + hap$ to represent the corresponding data that include the $(M, C) = (1, 1)$ pairs.

Real data analysis

Autism spectrum disorder (ASD) is a serious neuron-developmental disorder that impairs the ability to communicate and interact, typically starting in childhood [19, 20]. Parents usually notice signs in the first two years of their child's life. The disease is highly heritable, but the underlying genetic determinants are yet fully understood [21, 22]. Some studies have amassed evidence that suggests the involvement of parent-of-origin effects, including both imprinting and maternal effects [23–25]. To uncover the genetic architecture of ASD, the Autism Genome Project (AGP) Consortium investigated 2611 nuclear families. We obtained the genotype and phenotype data of 4222 individuals with pedigree information from dbGAP (Accession: phs000267.v1.p1) [26]. In our application of LIMEhap, we focus on the 41,940 SNPs on chromosome 7, as this chromosome contains more than 50 genes that have been implicated to be associated with ASD in the literature. One flanking marker on each side were used for each test locus. In addition, as ignoring structure due to differential ancestry

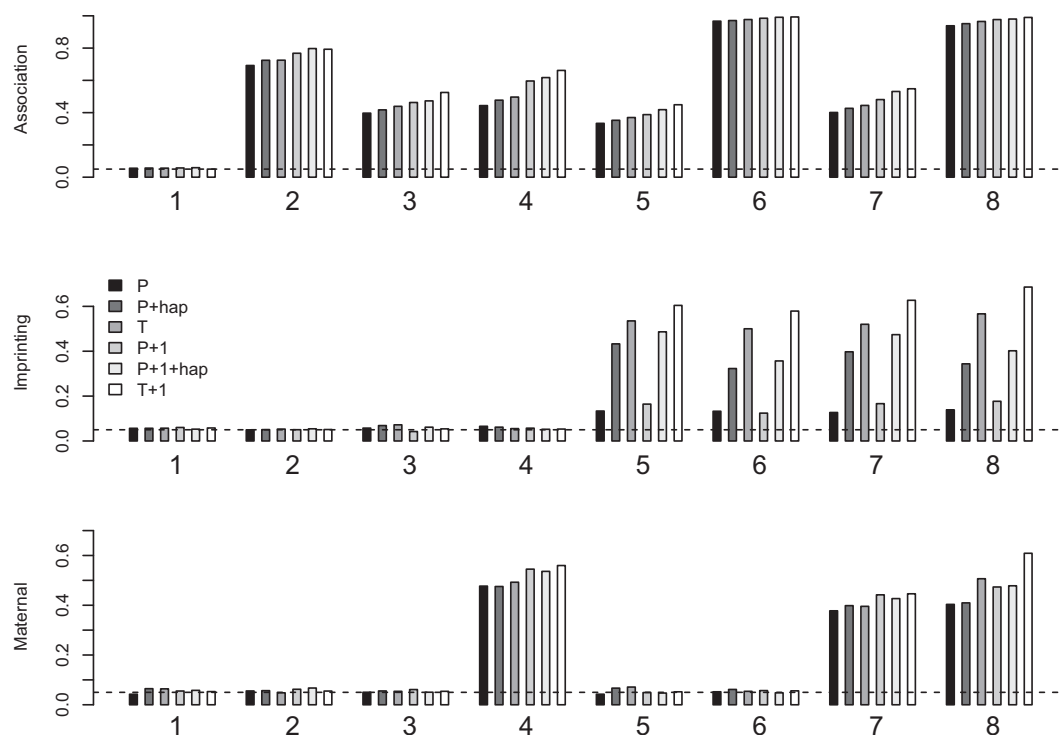


Fig. 1 Type I error and power for eight disease models (labeled 1–8 corresponding to the numbering in Table 2) for the scenario in which HWE does not hold and $PREV = 0.05$. The top segment gives the outcome for association, in which the results for model 1 represent type I errors while the rest are power. The middle segment provides the outcome for imprinting, in which the results for models 1–4 represent type I errors while the rest are power. The bottom

segment presents the outcome for maternal effect, in which the results for models 1–3, 5–6 represent type I errors while the rest are power. For each model, the results for six types of data utilization are given in the order provided in the caption, and note that $P + hap$ and $P + 1 + hap$ give the analysis results using LIMEhap and the rest using the original LIME.

can lead to an excess of spurious findings and reduced power, we only utilize families with European ancestry (the indicator variable for European ancestry is available). About 88% of the individuals in the data belong to this category, which come from 1193 nuclear families including 1177 affected and 32 unaffected offspring.

Results

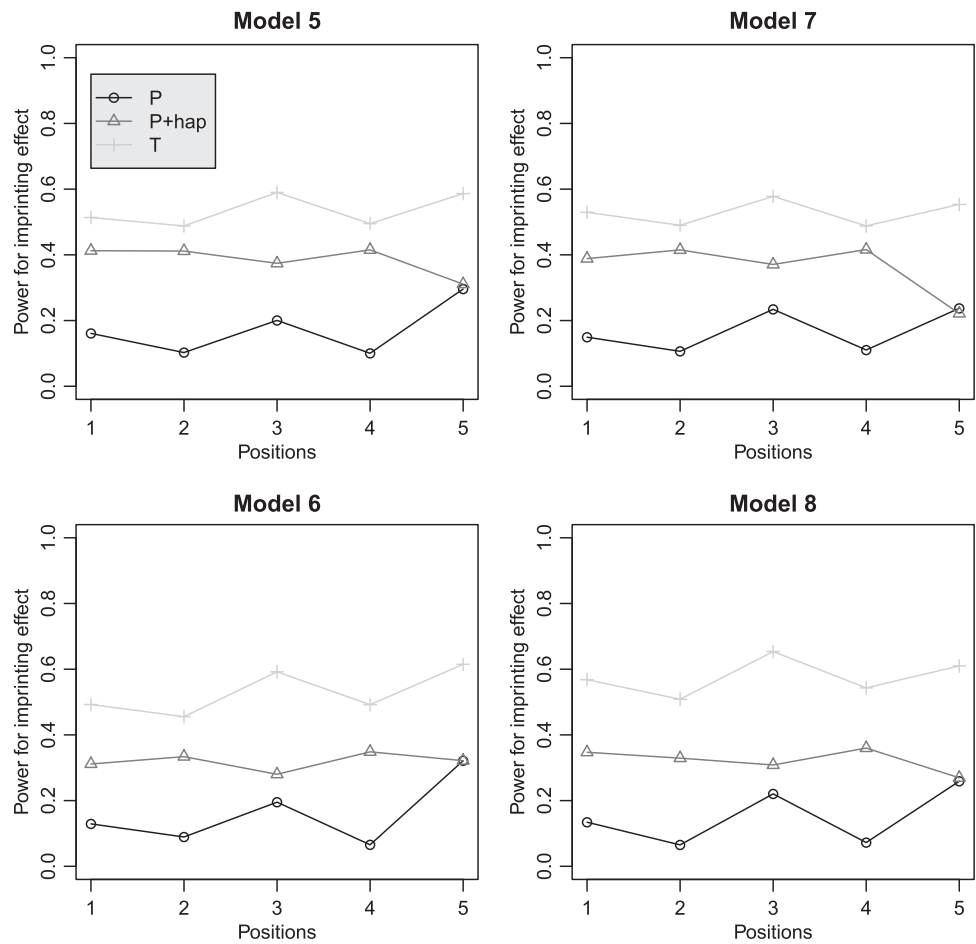
Simulation study

Type I error and power

We applied LIMEhap to data types $P + hap$ and $P + 1 + hap$. For comparison, we also applied LIME to the P , $P + 1$, T , and $T + 1$ data types. Although it is expected that the results from LIMEhap is somewhere between the corresponding P and T data, we are interested in assessing how much information can be gained using $P + hap$ over P , or $P + 1 + hap$ over $P + 1$, and how close $P + hap$ or $P + 1 + hap$ can get to T or $T + 1$ respectively. For our first set of

analyses, we use SNP1 as the test locus and the rest of the SNPs to infer haplotype. Figure 1 shows the type I error rate and power when HWE does not hold and $PREV = 0.05$. We can see that all the type I error rates are close to the nominal values of 0.05, even when HWE does not hold. This shows that even though HAPLORE infers haplotype based on the HWE assumption, LIMEhap is still robust to violation of such an assumption, inheriting the property from the original LIME. Compared with LIME, LIMEhap obtains much greater power for detecting imprinting effect for all four non-null models and comparable power for detecting association and maternal effects. The numerical values for the type I error and power are given in Supplementary Tables S4–9. In all cases, for detecting imprinting effects, the inclusion of the additional information from the $(M, C) = (1, 1)$ category cuts down the information loss (from not having complete triad data) by at least half. The reason that there is a drastic increase in power for detecting imprinting effect is that using markers in LD with the test locus helps resolve, even if not completely, the parental origin of the variant allele in the child to a great extent. The information increase is nontrivial since, typically, more than

Fig. 2 Power for imprinting effect at different test locus when HWE holds and $PREV = 0.05$. For each locus, the genotype information from the rest of the other four loci is used to help resolve the familial haplotype ambiguity. Only results from Models 5–8 are presented here since there are no imprinting effects for models 1–4.



10% of the families fall into the $(M,C) = (1,1)$ category. Similar results for type I error and power are obtained for the other three scenarios of HWE and $PREV$ combinations and are delegated to Supplementary Figures S1–3.

In reality, there might be a mixture of triad data and pair data. To simulate this scenario, we also randomly set a father’s genotype to be missing with probability 0.5 in each case-parent families and 0.7 in each control-parent families. We then apply the two-step LIMEhap procedure to these resulting mix datasets. Results in Supplementary Figures S4–7 and Tables S10–13 show that LIMEhap also controls type I error well and gains considerable power for detecting imprinting effect.

Position of test locus

The results presented above all use the 1st SNP in the 5-SNP block as the test locus. To explore whether detection power can be substantially influenced by the position of the test locus relative to the additional loci, we also position the test locus to be at the 2nd, 3rd, 4th and 5th SNP, and we use the rest of the SNPs as additional loci for inferring

fathers’ genotypes. The power for detecting imprinting effect, presented in Fig. 2, is for the scenario where HWE holds and $PREV = 0.05$. From the figure, we first see that the power for using the additional markers is increased regardless of the test locus position, except for when the test locus is at the last position. To understand the result when the last SNP (SNP 5) is the test locus, we first note that the SNP is the most informative, or nearly the most informative; therefore, there is limited additional information that can be gained by utilizing information from the other SNPs. Further, the last SNP is in fact in low LD with the other SNPs (Supplementary Table S3). Perhaps the most likely reason for lack of increase in power in this instance is due to its minor allele frequency (MAF) being almost equal to 0.5 (Supplementary Table S2), rendering less certainty about the familial haplotype configurations even with additional information from nearby loci. This argument is consistent with that made in another study [13]. Similar results for the other three scenarios of HWE and $PREV$ combinations can be found in Supplementary Figures S8–10.

Although one might have expected the greatest information gain to be when the test locus is in the middle of the

LD block (i.e., when SNP 3 is the test locus), this is however, not the case. Even though this may be counter-intuitive, the results in fact is not surprising upon careful inspection of Fig. 2 and Supplementary Figures S8–10, Tables S2 and S3, and using the same line of arguments as for SNP 5. It turns out that SNP 3 is either the most informative or nearly the most informative, in par with SNP 5; hence, as discussed before, utilizing the flanking SNPs will only gain limited additional information. Further, like SNP 5, it has low LD with the rest of the four SNPs (Supplementary Table S3). However, compared with SNP 5, the MAF of SNP 3 is 0.28 (much less than 0.5); thus, using the flanking markers did lead to some, albeit limited, power gain over using pair data only. On the contrary, the most information gain is in the situation where the test SNP itself is less informative and has a relatively small MAF, such as the 2nd or the 4th SNP. Using SNP 2 as an example, which has a MAF of 0.18, we examined whether additional markers, beyond the flanking ones, will necessarily lead to greater information for inferring parental origins of the minor allele. The results in Supplementary Figure S23 show that there was very substantial power gain for detecting imprinting effect when the flanking markers SNP 1 and SNP 3 were used (note that the LD between SNP 1 and 2 was 0.78). However, inclusion of additional loci (adding SNPs 4 and 5 one at a time) in this situation did not lead to further power gain. Taken together, we have demonstrated that the position of the test locus relative to the additional loci is not the most important factor for determining information gain. Rather, whether substantial information will be gained by using additional markers depends not only on its LD profile with the other markers, but also on its own informativeness and its MAF. Further, it appears that the use of the two immediate flanking markers is sufficient for helping to resolve the parental origin of the minor allele in a child. Including additional markers does not seem to be necessary.

The two-step procedure of LIMEhap is practical, but note that the disease model and affection status of probands are ignored in the first step, which may lead to bias in estimating the effect size in the second step. To quantify this concern, we checked the mean value of the estimates for each parameter under different settings, and found that the empirical means are all very close to the corresponding true values, alleviate the concern regarding bias, although outliers exit (Supplementary Figures S11–18). However, note that this is not a unique problem of LIMEhap; rather, it is a general phenomenon in LIME approaches when the sample size is limited [27].

Our above results are all based on a balanced design, where the number of case families is the same as the number of control families. To explore the impact of an unbalanced design, we also performed a small simulation

study with 210 case families and 90 control families. We can see that LIMEhap still has type I error close to the nominal value and much higher power for detecting imprinting effects than that of LIME (Supplementary Figures S19–22). Nevertheless, the absolute power is lower compared with that with the same total number of 300 families of a balanced design.

Real data analysis

Using LIMEhap, we identified a number of SNPs that have potential association, imprinting, or maternal effects at the 0.05 significance level after the Bonferroni correction. We then checked for violation of the HWE assumption at the 5% nominal level, and none of these significant SNPs failed the test. Although LIME is not susceptible to deviation from HWE, and the second-step of LIMEhap has inherited this property, we nonetheless still test HWE for all markers since such is assumed in HAPLORE. Our results show that some of the estimated effects are very large, which could very well be due to the small number of control families [27]. Thus these results need to be further scrutinized before they can be reported confidently, and we chose only to report SNPs with reasonable effect sizes. This issue is further elaborated in the Discussion Section. In the following, we focus on discussing SNPs that are found to be significant by LIMEhap (at the Bonferroni-corrected 5% level) and have also been implicated in the literature previously.

The top segment of Table 3 presents the SNPs that have been found to have potential imprinting effects by LIMEhap and have met the above general criterion. Specifically, SNPs rs1608628 and rs10247167 both fall in the *SLI4* region, which is related to specific language impairment according to OMIM (<http://www.omim.org/>). It has been shown in Ruser et al. [28] that impaired communication is part of the broader autism phenotype, especially among male family members. SNPs rs4729043 and rs917325 are within gene *HDAC9*, which has been identified to be associated with ASDs [29]. SNPs rs1978201, rs4718958, and rs2299456 are within the autism susceptibility loci *AUTS1*, *AUTS2*, and *AUTS9*, respectively (OMIM). SNPs rs1567277 and rs854721 are within gene *PPP1R9A*, which has been examined as a candidate gene for ASDs [29]. Gene *PCLO*, which includes SNP rs7807790, is identified to be associated with ASDs in the Chinese Han population [30].

For imprinting effects, three SNPs have been found to meet the criterion; that is, their Bonferroni-corrected p value are smaller than 0.05 and they have been previously discussed as having an effect on ASD by other investigators (middle segment of Table 3). In fact, these SNPs all fall within the *SLI4* gene region. The SNPs that were found to have potential maternal effects by LIMEhap and have also been implicated in the literature are provided in the bottom

Table 3 SNPs inferred from LIMEhap with potential association, imprinting, or maternal effects on ASD and have been implicated previously in the literature.

Effect type	SNP	GRCh38 ref. seq. ^a	Gene	−log ₁₀ (<i>P</i> value)
Association	rs1608628	NC_000007.14:g.147010952T>C	<i>SLI4</i> (<i>CNTNAP2</i>)	8.049***
	rs4729043	NC_000007.14:g.92448402G>A	<i>HDAC9</i>	7.496**
	rs10247167	NC_000007.14:g.145581582C>T	<i>SLI4</i>	7.255**
	rs2299456	NC_000007.14:g.126528712T>C	<i>AUTS9</i> (<i>GRM8</i>)	6.962**
	rs1527677	NC_000007.14:g.94958854C>T	<i>PPP1R9A</i>	6.756**
	rs917325	NC_000007.14:g.18248970A>G	<i>HDAC9</i>	6.553*
	rs7807790	NC_000007.14:g.83154530T>C	<i>PCLO</i>	6.379*
	rs854721	NC_000007.14:g.94971522G>A	<i>PPP1R9A</i>	6.156*
	rs4718958	NC_000007.14:g.70548605G>A	<i>AUTS2</i> (<i>KIAA0442</i>)	6.008*
	rs1978201	NC_000007.14:g.105200029T>C	<i>AUTS1</i>	5.980*
Imprinting	rs842454	NC_000007.14:g.158784449G>A	<i>SLI4</i> (<i>ESYT2</i>)	6.978**
	rs7787974	NC_000007.14:g.158735479G>A	<i>SLI4</i> (<i>ESYT2</i>)	6.936**
	rs2657323	NC_000007.14:g.158851177T>C	<i>SLI4</i>	6.267*
Maternal	rs3807843	NC_000007.14:g.8140025A>G	<i>ICA1</i>	8.071***
	rs17347159	NC_000007.14:g.18354475C>A	<i>HDAC9</i>	7.951***
	rs3734989	NC_000007.14:g.132179832T>C	<i>PLXNA4</i>	7.417**
	rs41567	NC_000007.14:g.82990312T>C	<i>PCLO</i>	7.405**
	rs11977905	NC_000007.14:g.15435562A>C	<i>AGMO</i>	7.394**
	rs2107829	NC_000007.14:g.82984933A>T	<i>PCLO</i>	7.028**
	rs10225065	NC_000007.14:g.106650074A>G	<i>AUTS1</i>	6.897**
	rs6466064	NC_000007.14:g.105383770T>C	<i>AUTS1</i> (<i>SRPK2</i>)	6.612*
	rs10488060	NC_000007.14:g.156755951T>C	<i>SLI4</i> (<i>LMBR1</i>)	6.450*
	rs6958145	NC_000007.14:g.8148337C>G	<i>ICA1</i>	6.325*
	rs11981093	NC_000007.14:g.156937382C>T	<i>SLI4</i>	6.151*
	rs6956114	NC_000007.14:g.111857472G>A	<i>AUTS9</i> (<i>DOCK4</i>)	6.093*
	rs13228314	NC_000007.14:g.152048953C>T	<i>SLI4</i> (<i>GALNT11</i>)	6.018*
	rs12154389	NC_000007.14:g.111855897G>A	<i>AUTS9</i> (<i>DOCK4</i>)	5.928*

*0.05 significant; **0.01 significant; and ***0.001 significant, all with Bonferroni correction.

^aThese SNPs are all in Chromosome 7.

segment of Table 3. To elaborate, SNPs rs3807843 and rs6958145 are both in gene *ICA1*, while Salyakina et al. [31] provides evidence for the involvement of *ICA1* on 7p21.3 in ASDs. SNP rs17347159 falls in the gene *HDAC9*. SNP rs3734989 is within *PLXNA4*. Suda et al. [32] found decreased expression of axon-guidance proteins such as *PLXNA4* in the brains of people with autism. Both SNPs rs41567 and rs2107829 are within gene *PCLO*. The Study by Fenster and Garner [33] suggests that the alterations in the expression of Piccolo or the *PCLO* gene could contribute to developmental disabilities and mental retardation. SNP rs11977905 is within the gene *AGMO*, where a rare CNV in the *AGMO* (*TMEM195*) gene has been identified with autism [34]. SNPs rs10225065 and rs105024216 are both in autism susceptibility loci *AUTS1*. SNPs rs6956114 and rs12154389 are both in autism susceptibility loci *AUTS9*. SNPs rs10488060, rs11981093, and rs13228314 are all in *SLI4* gene region.

Discussion

In this article, we propose a two-step approach, LIMEhap, to improve upon a recent partial likelihood inference method, LIME, for detecting imprinting and maternal effects. The improvement is to make fuller usage of mother–child pair data by utilizing all pairs regardless of their genotypes instead of discarding certain pairs with ambiguous parental origin of the variant allele. Our simulation shows that LIMEhap has empirical type I error rates close to the nominal value and achieves higher power than the original LIME. Further, the position of the test locus relative to the additional loci in the LD block does not appear to be important. On the other hand, the MAF of the test locus and its informativeness do influence the extent of power gain. Moreover, there may be little additional information gain for delineating the parental origin beyond the usage of the immediate flanking markers of a test locus.

Although the first step of LIMEhap requires the HWE assumption, our simulation results show that it is robust to HWE violation, at least for the 32 settings considered. Note that the second step of LIMEhap does not assume HWE, as no nuisance parameters are in the partial likelihood. Nevertheless, it would be advantageous for the haplotype frequency inference procedure in the first step to be free of the HWE assumption as well to ensure that the entire LIMEhap procedure is robust to such violation regardless of the underlying setting. To this end, the algorithm described in Kong et al. [35] is a potential choice, as it does not require HWE, although extended pedigrees are needed in order to infer phase information.

From the results of our analysis of the AGP data using LIMEhap, we see that the gene region *SLI4* was repeatedly implicated to have potential association, imprinting, and maternal effects with ASD. Although there was an abundance of findings of association between *SLI4* and ASD in the literature, little was said about imprinting nor maternal effects in previous investigations. However, although not directly supporting the evidence of epigenetic effects, according to OMIM (<http://WWW.omim.org/>), the *SLI4* region is related to specific language impairment, which is part of the broader autism phenotype especially for male family members. Notwithstanding the possibilities of false positives, the considerable number of novel findings of potential maternal and imprinting effects are likely the consequence of increased power by making use of all available data and by considering the joint effects of both factors to diminish the impacts of potential confounding. These are advantages of LIMEhap compared with methods used in previous studies.

Despite advantages seen in both simulation and real data analyses, LIMEhap has its limitations. Due to the concern about the impact of an unbalanced design on the type I error and power, we carried out a small simulation, which showed that type I errors remain well maintained yet there was a power loss compared with a balanced design. It is worth pointing out, though, that lower power with an unbalanced design is a common problem in statistical methods for case-control (families) studies. In real data analyses, deviations from a balanced design can be more extreme than what we have explored, which was the case with the autism data analyzed. This situation with the AGP data is by no means unique, though, as it is a fact that control families are harder to recruit than case families; therefore, it is warranted to explore whether LIMEhap can be extended to the setting where only discordant sib-pair families are available without the need to recruit separate control families [36].

Another limitation, which is not unique to LIMEhap either, is the expected limited power for detecting imprinting effects when imprinting varies among

individuals. This could potentially be an issue in our ASD analysis. Although the AGP genotypes were obtained mainly with DNA samples from blood, a small proportion of the samples were obtained from cell lines, buccal, and other sources. To the best of our knowledge, there is a dearth of statistical methods for analyzing data exhibiting polymorphic imprinting. Thus, future research is warranted to develop methods with adequate power in such a situation. However, it is out of the scope of the current research.

Acknowledgements The authors would like to thank the Section Editor and the anonymous reviewers for their constructive comments and suggestions, which, in our view, have led to improved presentation and greater clarity. The autism spectrum disorder family data, made available by the AGP, were downloaded from dbGaP (Accession: phs000267.v1.p1).

Funding This work was supported in part by the National Science Foundation grant DMS-1208968.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Lawson HA, Cheverud JM, Wolf JB. Genomic imprinting and parent-of-origin effects on complex traits. *Nat Rev Genet.* 2013;14:609–17.
2. Noursome D, Lupo PJ, Okcu MF, Scheurer ME. Maternal and offspring xenobiotic metabolism haplotypes and the risk of childhood acute lymphoblastic leukemia. *Leuk Res.* 2013;37: 531–5.
3. Ferguson-Smith AC. Genome imprinting: the emergence of an epigenetic paradigm. *Nat Rev.* 2011;12:565–75.
4. Naumova AK, Croteau S. Mechanisms of epigenetic variation: polymorphic imprinting. *Curr Genomics.* 2004;5:417–29.
5. Hager R, Cheverud JM, Wolf JB. Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics.* 2008;178:1755–62.
6. Lin S. Assessing the effects of imprinting and maternal genotypes on complex genetic traits. In: Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A, editors. *Risk assessment and evaluation of predictions.* New York, NY, USA: Springer; 2013. p. 285–300.
7. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet.* 1998;62:969–78.
8. Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet.* 1999;65:229–35.
9. Ainsworth HF, Unwin J, Jamison DL, Cordell HJ. Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet Epidemiol.* 2011;35:19–45.

10. Shi M, Umbach DM, Vermeulen SH, Weinberg CR. Making the most of case-mother/control-mother studies. *Am J Epidemiol*. 2008;168:541–7.
11. Yang J, Lin S. Robust partial likelihood approach for detecting imprinting and maternal effects using case-control families. *Ann Stat*. 2013;7:249–68.
12. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, et al. Parental origin of sequence variants associated with complex diseases. *Nature*. 2009;462:868–74.
13. Lin D, Weinberg CR, Feng R, Hochner H, Chen J. A multi-locus likelihood method for assessing parent-of-origin effects using case-control mother–child pairs. *Genet Epidemiol*. 2013;37:152–62.
14. Howey R, Mamasoula C, Töpf A, Nudel R, Goodship JA, Keavney BD, et al. Increased power for detection of parent-of-origin effects via the use of haplotype estimation. *Am J Hum Genet*. 2015;97:419–34.
15. Zhang K, Sun F, Zhao H. HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics*. 2005;21:90–103.
16. Chen J, Peters U, Foster C, Chatterjee N. A haplotype based test of association using data from cohort and nested case-control epidemiologic studies. *Hum Hered*. 2004;58:18–29.
17. Chen J, Chatterjee N. Haplotype based association analysis in cohort and nested case-control studies. *Biometrics*. 2006;62:28–35.
18. Weir BS. Genetic data analysis II. Sunderland, MA, USA: Sinauer; 1996.
19. Lahiri DK, Sokol DK, Erickson C, Ray B, Ho CY, Maloney B. Autism as early neurodevelopmental disorder: evidence for an sAPP α -mediated anabolic pathway. *Front Cell Neurosci*. 2013;7:94.
20. Xiao Z, Qiu T, Ke X, Xiao X, Xiao T, Liang F, et al. Autism spectrum disorder as early neurodevelopmental disorder: evidence from the brain imaging abnormalities in 2-3 years old toddlers. *J Autism Dev Disord*. 2014;44:1633–40.
21. Sandin S, Lichtenstein P, Kuja-Halkola R, Hultman C, Larsson H, Reichenberg A. The heritability of autism spectrum disorder. *JAMA*. 2017;318:1182–4.
22. Tick B, Bolton P, Happé S, Rutter M, Rijdsdijk F. Heritability of autism spectrum disorders: a meta-analysis of twin studies. *J Child Psychol Psychiatry*. 2016;57:585–95.
23. Fradin D, Cheslack-Postava K, Ladd-Acosta C, Newschaffer C, Chakravarti A, Arking DE, et al. Parent-of-origin effects in autism identified through genome-wide linkage analysis of 16,000 SNPs. *PLoS ONE*. 2010;5:e12513.
24. Loke YJ, Hannan AJ, Craig JM. The role of epigenetic change in autism spectrum disorders. *Front Neurol*. 2015;6:107.
25. Porokhovnik LN, Kostyuk SV, Ershova ES, Stukalov SM, Veiko NN, Korovina NY, et al. The maternal effect in infantile autism: elevated DNA damage degree in patients and their mothers. *Biomed Khim*. 2016;62:466–70.
26. Cheng Y, Quinn JF, Weiss LA. An eQTL mapping approach reveals that rare variants in the SEMA5A regulatory network impact autism risk. *Hum Mol Genet*. 2013;22:2960–72.
27. Zhang F, Khalili A, Lin S. Optimum study design for detecting imprinting and maternal effects based on partial likelihood. *Biometrics*. 2016;72:95–105.
28. Ruser TF, Arin D, Dowd M, Putnam S, Winklosky B, Rosen-Sheidley B, et al. Communicative competence in parents of children with autism and parents of children with specific language impairment. *J Autism Dev Disord*. 2007;37:1323–36.
29. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am J Hum Genet*. 2014;94:677–94.
30. Li H, Li Y, Shao J, Li R, Qin Y, Xie C, et al. The association analysis of RELN and GRM8 genes with autistic spectrum disorder in Chinese Han population. *Am J Med Genet*. 2008;B147:194–200.
31. Salyakina D, Cukier HN, Lee JM, Sacharow S, Nations LD, Ma D, et al. Copy number variants in extended autism spectrum disorder families reveal candidates potentially involved in autism risk. *PLoS One*. 2011;6:e26049–e26049.
32. Suda S, Iwata K, Shimmura C, Kamenoy Y, Anitha A, Thanseem I, et al. Decreased expression of axon-guidance receptors in the anterior cingulate cortex in autism. *Mol Autism*. 2011;2:14–14.
33. Fenster SD, Garner CC. Gene structure and genetic localization of the PCLO gene encoding the presynaptic active zone protein Piccolo. *Int J Neurosci*. 2002;20:161–71.
34. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. *Science*. 2007;316:445–9.
35. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet*. 2008;40:1068–75.
36. Zhang F, Khalili A, Lin S. Imprinting and maternal effect detection using partial likelihood based on discordant sibpair data. *Stat Sin*. 2019;29:1915–37.