Published in final edited form as:

Genet Epidemiol. 2017 November ; 41(7): 587–598. doi:10.1002/gepi.22060.

An efficient study design to test parent-of-origin effects in family trios

Xiaobo Yu^{1,2}, Gao Chen², and Rui Feng¹

¹Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

²Department of Neurosurgery, School of Medicine, The Second Affiliated Hospital of Zhejiang University, Hangzhou, China

Abstract

Increasing evidence has shown that genes may cause prenatal, neonatal, and pediatric diseases depending on their parental origins. Statistical models that incorporate parent-of-origin effects (POEs) can improve the power of detecting disease-associated genes and help explain the missing heritability of diseases. In many studies, children have been sequenced for genome-wide association testing. But it may become unaffordable to sequence their parents and evaluate POEs. Motivated by the reality, we proposed a budget-friendly study design of sequencing children and only genotyping their parents through single nucleotide polymorphism array. We developed a powerful likelihood-based method, which takes into account both sequence reads and linkage disequilibrium to infer the parental origins of children's alleles and estimate their POEs on the outcome. We evaluated the performance of our proposed method and compared it with an existing method using only genotypes, through extensive simulations. Our method showed higher power than the genotype-based method. When either the mean read depth or the pair-end length was reasonably large, our method achieved ideal power. When single parents' genotypes were unavailable or parental genotypes at the testing locus were not typed, both methods lost power compared with when complete data were available; but the power loss from our method was smaller than the genotype-based method. We also extended our method to accommodate mixed genotype, low-, and high-coverage sequence data from children and their parents. At presence of sequence errors, low-coverage parental sequence data may lead to lower power than parental genotype data.

Keywords

family study design; haplotype; imprinting; next generation sequence; pediatric disease; statistical method

---

Correspondence. Rui Feng, Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, 209 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, USA. ruifeng@mail.med.upenn.edu.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

1 | INTRODUCTION

Genome-wide association studies (GWAS) have been conducted to find causal genetic variants underlying many health outcomes. However, a large amount of the disease heritability remains unexplained and is partially accounted by undetected epigenetic and parent-of-origin effects (POEs) (Manolio et al., 2009). Increasing evidence suggests that certain genes increase the risk of diseases only when they are inherited from mothers but not from fathers, or vice versa. Some prenatal and neonatal disorders including Beckwith-Wiedemann syndrome (Weksberg et al., 1993), Prader-Willi/Angelman syndrome (Donlon, 1988), pediatric obesity (Wermter et al., 2008), and autism (Duyzend et al., 2016), are known to be affected by either paternal or maternal risk alleles.

The parent-of-origin coded alleles have been included in advanced statistical models. These models not only detect POEs, but also help recover missing heritability (Xiao, Ma, & Amos, 2013). The key in these models is to identify the maternal or paternal origins of children's alleles based on Mendelian inheritance, which has been mostly implemented in child-parent trios. For binary outcomes, Weinberg developed a log-linear model that stratified on both parental mating type and children's inherited alleles (Weinberg, 1999). For continuous traits, many used regression methods incorporating separate effects of maternal and paternal alleles (Hanson, Kobes, Lindsay, & Knowler, 2001; Shete & Amos, 2002; Shete, Zhou, & Amos, 2003; Whittaker, Gharani, Hindmarsh, & McCarthy, 2003). The methods above used only genotype data at a single testing locus and did not take advantage of the valuable intramarker linkage disequilibrium (LD) information. A likelihood-based method (Feng, Wu, Jang, Ordovas, & Arnett, 2011) borrowed the LD imbedded in the haplotypes around the testing locus and jointly modeled parent-of-origin and additive genetic effects, which noticeably improved the power of detecting POEs in families.

With the additional coverage for rare alleles and structure variants, the next-generation sequencing (NGS) (Metzker, 2010) technology has been applied more widely to detect disease-associated variants. In many studies, the DNA samples of children and their parents were available for further investigations. For example, a National Heart, Lung, and Blood Institute sponsored cohort study Prematurity and Respiratory Outcomes Program (PROP) collected clinical data and DNA samples from 765 preemies born between 23 and 28 gestational weeks (Poindexter et al., 2015). However, the cost of sequencing all family members was around 1 million dollars in 2016, which was unaffordable to the study team.

To fully utilize DNA samples and work within budgets at the same time, we propose a new hybrid design where we sequence children by more expensive NGS and genotype their parents by much cheaper single nucleotide polymorphism (SNP) array. For this design, we developed a powerful and efficient likelihood-based method for testing the POEs of children's alleles on their disease phenotypes. In combination with parents' genotypes, all sequence reads of children were incorporated to construct all members' haplotype phases and to identify the parent-of-origins of children's alleles. With the additional linkage information that can be extracted from sequence segments, our proposed method is expected to improve the power of POE testing regardless of the intermarker LD strength.

Missing genotypes or phenotypes often occur in family studies due to various reasons, such as, failed informed consents, geographical limits, single-parent families, and SNP genotyping array techniques (Lin, Hu, & Huang, 2008). For example, in PROP, DNA samples were collected from 95% of participating infants, 85% of their mothers, and only 65% of their fathers. If a parent's DNA is unavailable, the child's sequence reads and the other parent's genotypes remain usable to improve both the inference of haplotypes and parent-of-origins. In addition, some variants, especially rare variants, may not be typed in certain SNP arrays. If both parental genotypes at the testing locus are unavailable, neighboring linked markers can still be informative. Our method can handle both types of missing data.

To allow for diverse data collection variations, we also extended our method to mixed SNP, low-coverage sequence, and high-coverage sequence data. The benefits of our proposed design and approach were demonstrated through a series of simulations.

2 | MATERIAL AND METHODS

2.1 | Notation

Throughout the paper, we consider a parent-child trio design and assume that there are n independent children in a study. Suppose we are only interested in testing the POEs of children's alleles on the outcomes at a SNP locus, which is located within a genomic region with multiple strongly or weakly linked SNPs. The raw pair-end sequence reads of all children are available and have been aligned to the reference genome (i.e., alignment data from BAM files). These sequence reads are simplified to consecutive SNP codes. We denote the collection of such sequence codes for child i as S_c^i . At the testing locus, we denote the unobserved “sourced” genotype of child i as (g_{c-p}^i, g_{c-m}^i) where g_{c-p}^i and g_{c-m}^i are the paternal and maternal alleles inherited by this child, respectively. If the major and minor alleles are coded by 0 and 1, then the *sourced genotypes* have four possible values: (0, 0), (0, 1), (1, 0), and (1, 1). Different from typical 0-1-2 genotype coding, the notation of (g_{c-p}, g_{c-m}) separates the regular heterozygous genotypes into reciprocal heterozygotes. The haplotype pair of the child is called *diplotype*. The *sourced diplotype* of the child is denoted as (h_{c-p}^i, h_{c-m}^i) for the haplotype inherited from his/her father and mother, respectively. The *sourced genotypes or diplotypes* are not observed directly from the data and will be implicated in likelihood functions. For comparison with the existing method for SNP genotypes, we assume that child's genotypes are observed over the whole region and his/her genotype is denoted by G_c^i , a vector with single-locus genotypes as elements.

The genotypes of father i and mother i (from SNP array) within the region are denoted by G_p^i and G_m^i , respectively. The diplotype of the parents, with their own parent-of-origins unknown, is denoted as h_{p1}/h_{p2} and h_{m1}/h_{m2} , respectively. Haplotype h_p or h_m can take any value from 1 to t (t is the possible haplotypes within the region) and their population frequencies $\rho = (\rho_1, \rho_2, \dots, \rho_t)$ are either known or estimated from parental genotypes.

2.2 | Model

Let Y be a quantitative outcome and $\mathbf{x}$ be a vector of covariates (including intercept). Given each individual child's *sourced genotype* (g_{c-p}, g_{c-m}), the phenotypes of the children follow:

$$Y^i = \mathbf{x}^i \beta + g_{c-p}^i \gamma_p + g_{c-m}^i \gamma_m + \varepsilon^i, \text{ for } i=1, \dots, n. \quad (1)$$

where i is the individual child index (also family index), n is the total number of independent children, β is a vector of the covariate effects on the outcome, γ_p and γ_m are the genetic effects of paternally and maternally inherited alleles, respectively, and ε is the random error following a normal distribution with a mean of 0 and a variance of σ^2 . In order to test the equivalence of paternal and maternal effects (i.e., $\gamma_p = \gamma_m$), we reparameterize model (1) into an equivalent model:

$$Y^i = \mathbf{x}^i \beta + (g_{c-p}^i + g_{c-m}^i) \gamma_1 + (g_{c-p}^i - g_{c-m}^i) \gamma_2 + \varepsilon^i, \\ \text{for } i=1, \dots, n. \quad (2)$$

where $\gamma_1 = \frac{\gamma_p + \gamma_m}{2}$ stands for the total genetic effect and $\gamma_2 = \frac{\gamma_p - \gamma_m}{2}$ indicates the POE. The null hypothesis of interest is that the effects from the paternal and maternal alleles are equal. So under the null hypothesis, $\gamma_2 = 0$ and the model becomes a general additive model $Y^i = \mathbf{x}^i \beta + (g_{c-p}^i + g_{c-m}^i) \gamma_1 + \varepsilon^i$. The sign of γ_2 indicates the direction of POEs in favor of maternal or paternal minor allele's risk. In addition, the magnitude of γ_1 and γ_2 reveals the relationship between maternal and paternal allelic effects. In particular, when $\gamma_1 = \gamma_2 = 0$, maternal allele has no effect on child's outcome (complete imprinting).

2.3 | Log-likelihood with observed parental genotypes and children's sequences

Prior to formulating likelihood function, we want to introduce three key probabilities in family data.

- I. Penetrance probability: It is the probability of children's phenotype given their sourced genotype, denoted as $\Phi(Y^i | X^i, g_{c-p}^i, g_{c-m}^i; \mu^i, \sigma^2)$, a normal probability density function with mean $\mu^i = \mathbf{x}^i \beta + (g_{c-p}^i + g_{c-m}^i) \gamma_1 + (g_{c-p}^i - g_{c-m}^i) \gamma_2$ and variance σ^2 .
- II. (Conditional) Inheritance probability: It is the probability of a child receiving his haplotypes given the parents' diplotypes and his own sequence reads, i.e., $p((h_{c-p}^i, h_{c-m}^i) | h_{1p}^i / h_{2p}^i, h_{1m}^i / h_{2m}^i, S_c^i)$. Because maternal and paternal inheritance are independent,

$$p\left((h_{c-p}^i, h_{c-m}^i) | h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i, S_c^i\right) = \frac{p\left(h_{c-p}^i | h_{1p}^i/h_{2p}^i\right) \cdot \left(h_{c-m}^i | h_{1m}^i/h_{2m}^i\right) \cdot I\left((h_{c-p}^i, h_{c-m}^i) \sim S_c^i\right)}{\sum_{\text{all possible } (h_{c-p}^i, h_{c-m}^i) \sim S_c^i} p\left(h_{c-p}^i | h_{1p}^i/h_{2p}^i\right) \cdot \left(h_{c-m}^i | h_{1m}^i/h_{2m}^i\right)}$$

where $I(A)$ is a binary indicator and equals 1 when statement A is true.

“(h_{c-p}^i, h_{c-m}^i) $\sim S_c^i$ ” means that both haplotypes of the child h_{c-p}^i and h_{c-m}^i match with his/her sequence reads S_c^i . $p(h_{c-p}^i | h_{1p}^i/h_{2p}^i)$ is 1/2 for a heterozygous father ($h_{1p}^i \neq h_{2p}^i$) and 1 for a homozygous father ($h_{1p}^i = h_{2p}^i$), similarly ($h_{c-m}^i | h_{1m}^i/h_{2m}^i$) is 1/2 or 1 depending on the mother’s heterozygosity. Given the sequences of a child, only a subset of this child’s possible haplotypes (h_{c-p}^i, h_{c-m}^i) is compatible with his/her sequence reads. The total number of possible (h_{c-p}^i, h_{c-m}^i) can be reduced, and thus the conditional probability becomes 1, 1/2, 1/3, and 1/4. For example, if parental haplotypes are $h_1/h_2, h_3/h_4$ and the child’s sequence data are only compatible with $h_1/h_3, h_2/h_3$, and h_1/h_4 , this child has an equal chance of getting (h_1, h_3), (h_2, h_3), or (h_1, h_4).

III. Parental phasing probability: It is the probability of parental diplotypes, given observed parental genotypes and child’s sequences and can be written as $p(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i | G_p^i, G_m^i, S_c^i)$. It defines a prior how likely parental diplotypes exist in population given the observed data. As parental haplotypes must be compatible with their observed genotypes or child’s sequence reads, the conditional probability of an incompatible haplotype combination is zero. Because the diplotype of an individual fully determines his genotypes within the region, joint probability of h_{1p}^i/h_{2p}^i and G_p^i is the same as the probability of h_{1p}^i/h_{2p}^i . Thus the probability can be calculated according to

$$\begin{aligned} p(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i | G_p^i, G_m^i, S_c^i) &= \frac{p(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i, G_p^i, G_m^i | S_c^i)}{p(G_p^i, G_m^i | S_c^i)} \\ &= \frac{p(h_{1p}^i/h_{2p}^i) \cdot p(h_{1m}^i/h_{2m}^i) \cdot I(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i \sim G_p^i, G_m^i, S_c^i)}{\sum_{h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i \sim G_p^i, G_m^i, S_c^i} p(h_{1p}^i/h_{2p}^i) \cdot p(h_{1m}^i/h_{2m}^i)}. \end{aligned}$$

If haplotypes follow the Hardy-Weinberg Equilibrium (HWE), the diplotype probability is based on the haplotype population frequencies, i.e., $p(h_r/h_s) = 2\rho_r \cdot \rho_s$ if $r \neq s$ or ρ_r^2 if $r = s$; otherwise, we can use the diplotype population frequencies directly.

Given parental multilocus genotypes G_p and G_m and child’s sequence reads S_c , the log-likelihood of individual child i , as a function of $\beta, \gamma_1, \gamma_2, \sigma^2$, can be written as

$$l^i = \log p(Y^i | X^i, S_c^i, G_p^i, G_m^i)$$

$$= \log \left[\sum_{h_{1m}^i/h_{2m}^i, h_{1p}^i/h_{2p}^i \sim (G_p^i, G_m^i, S_c^i)} \left(\underbrace{\Phi(Y^i | X^i, g_{c-p}^i, g_{c-m}^i; \mu^i, \sigma^2)}_{I_p} \left((g_{c-p}^i, g_{c-m}^i) | (h_{c-p}^i, h_{c-m}^i) \right) \right) \underbrace{p((h_{c-p}^i, h_{c-m}^i) | h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i)}_{II} \right]$$

(3)

where I is the penetrance function, II is the inheritance probability, and III is the parental phasing probability. Because the *sourced genotype* of child i at the testing locus is a subset of his *sourced diplotype* (h_{c-p}^i, h_{c-m}^i) and can be uniquely determined by (h_{c-p}^i, h_{c-m}^i) , for that genotype (g_{c-p}^i, g_{c-m}^i) the probability $p((g_{c-p}^i, g_{c-m}^i) | (h_{c-p}^i, h_{c-m}^i))$ equals 1.

In theory, the summation in (3) should take over all possible diplotypes of parents, with each parent having $\ell(\ell+1)/2$ possible types. Because only the diplotypes that do not match the observed parents' genotypes or the sequences of their child have zero probabilities, the summation is taken over matched types only.

We use a trio example to illustrate how using sequence reads can help identify the origins of a child's alleles, as shown in Figure 1. There are five SNPs in a block and "SNP3" is the testing locus. Just looking at parental and child's genotypes at SNP3 (highlighted in yellow), we cannot tell the origins of this child's alleles A and T. But if we know that four possible haplotypes exist in the population (middle left) and observe the parental genotypes at all five loci (top left and right), we can infer that the father's haplotypes are AATTA and AAAAT, and the mother's haplotypes are TTTTA and TTAAT. Then the child's haplotype combination could be (1) one haplotype AATTA inherited from the father and the other TTAAT from the mother or (2) one haplotype AAAAT inherited from the father and the other TTTTA from the mother. Based on the population haplotype frequencies, we can calculate the conditional probability given the parental genotypes, which is 60% for (1) and 40% for (2). Therefore the child's paternal and maternal alleles at SNP3 can be (A, T) or (T, A) with 60% or 40% of chance. As A is the major allele, we denote (A, T) as (0, 1) and (T, A) as (1, 0). This is the basic principle underlying the haplotype-based method. The uncertainty/probability of *sourced genotypes* (A, T) or (T, A) is a function of haplotype frequencies and is incorporated into the likelihood. If we use "x" to indicate the SNP that is uncovered by sequence, all the sequence reads of this child can be denoted as TxAxx, xAxTx, xxTxA, TTxxx, xxAxT, AAxxx, AxTxx, and xTxAx. We can easily "stitch" pieces together to form only two haplotypes AATTA and TTAAT with the first one inherited from the father, i.e., the probability of the child's sourced genotypes being (1, 0) is 100%. In this case, our sequence-based method completely identifies the sources of the child's alleles. In a study, there can be a decent percentage of similar families and thus incorporating all sequence reads may improve the overall power of detecting POEs.

In our log-likelihood, both SNPs aligned along sequence reads and the LD information implied in haplotypes contribute to parts II and III and help infer the parent-of-origins of children's alleles.

The log-likelihood for all the data will be

$$l = \sum_{i=1}^N l^i, \quad (4)$$

and maximizing it yields the maximum likelihood estimation of the parameters of interests, including covariate effects, genetic effects, and phenotypic variance.

The log-likelihood under the null hypothesis is calculated similarly except that the penetrance probability is calculated under the null model:

$$Y^i = \mathbf{x}^i \beta + (g_{c-p}^i + g_{c-m}^i) \gamma_1 + \varepsilon^i, \text{ for } i=1, \dots, n. \quad (5)$$

2.4 | Log-likelihood with known parent-of-origins of children's genotypes (ideal and hypothetical)

When (g_{c-p}^i, g_{c-m}^i) is assumed to be known, then the log-likelihood ratio test (LRT) would be the doubled difference in the log-likelihood with least square estimates from linear regression of (5) and (2).

2.5 | Log-likelihood with only genotypes of children and parents, but without sequence data

The method utilizing the haplotypes can offer a better power than that using single-locus genotypes (Feng et al., 2011) and thus we will use it as the baseline comparison method. The log-likelihood function of individual i using haplotypes becomes:

$$l^i = \log p(Y^i | X^i, G_c^i, G_p^i, G_m^i)$$

$$= \log \left[\sum_{\substack{(g_{c-p}^i, g_{c-m}^i) \sim G_c^i \\ G_c^i = h_{1p}^i + h_{1m}^i \\ G_m^i = h_{1m}^i + h_{2m}^i \\ G_p^i = h_{1p}^i + h_{2p}^i}} \underbrace{\Phi(Y^i | X^i, (g_{c-p}^i, g_{c-m}^i); \mu^i, \sigma^2)}_I \underbrace{p((h_{1p}^i, h_{1m}^i) | h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i, G_c^i)}_{II} \underbrace{p(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i | G_p^i, G_m^i)}_{III} \right] \quad (6)$$

Log-likelihood (3) included the neighboring SNPs on one chromosome (identified through sequence reads), and thus parts II and III contain less possibilities than the counterparts in (6), which only use population haplotype frequencies.

2.6 | Log-likelihood with only single parents' genotypes available

Without loss of generalizability, we assume maternal genotypes are available and paternal genotypes are missing. The log-likelihood becomes

$$\begin{aligned}
 l^i(\beta, \gamma_1, \gamma_2, \sigma) &= \log p(Y^i | X^i, S_c^i, G_m^i) \\
 &= \log \left[\sum_{g_c^i \sim G_c^i = h_{c-p}^i + h_{c-m}^i} \left(\underbrace{\Phi(Y^i | X^i, (g_{c-p}^i, g_{c-m}^i); \mu^i, \sigma^2)}_I \right. \right. \\
 &\quad \left. \left. \underbrace{p((h_{c-p}^i, h_{c-m}^i) | h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i, S_c^i)}_{II} \underbrace{p(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i | G_m^i, S_c^i)}_{III} \right) \right]. \quad (7)
 \end{aligned}$$

When maternal genotypes are missing, function (7) can be easily revised. Part III will be updated with observed paternal genotypes, and the other parts of (7) remain the same.

There are more uncertainty in the inheritance and parental phases in log-likelihood (7) than in (3): when paternal genotypes are missing and two maternal haplotypes are inferable, we can know only one paternal haplotype that is inherited by the child and the possibilities of both homozygous and heterozygous paternal diplotypes need to be counted in parts II and III.

2.7 | Log-likelihood with sequence errors

Because mapping errors and base-calling errors are common in sequence data, we incorporate the error probability into the likelihood. If we assume the number of unmatched sequences follows a binomial distribution with the total number of sequences as the count and the error probability as the event rate, the last component in the log-likelihood can be written as

$$\begin{aligned}
 &p(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i | G_p^i, G_m^i, S_c^i) \\
 &= \frac{p(h_{1p}^i/h_{2p}^i) \cdot p(h_{1m}^i/h_{2m}^i) \cdot I(h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i \sim G_p^i, G_m^i, S_{c, \text{matched}}^i) \binom{m_i + u_i}{m_i} (1 - e)^{m_i} e^{u_i}}{\sum_{h_{1p}^i/h_{2p}^i, h_{1m}^i/h_{2m}^i \sim G_p^i, G_m^i, S_c^i} p(h_{1p}^i/h_{2p}^i) \cdot p(h_{1m}^i/h_{2m}^i) \binom{m_i + u_i}{m_i} (1 - e)^{m_i} e^{u_i}}, \quad (8)
 \end{aligned}$$

where m_i is the number of matched sequences, u_i is the number of unmatched sequences, $S_{c,matched}^i$ is the collection of sequence reads that match parental haplotypes, and e is the probability of locus-wise error rate. Because e is generally tiny and two or more unmatched reads can be removed at a screen stage, we included only cases at $u_i = 0$ or 1 in (8) as an approximation.

2.8 | LRT statistic

Our test statistic is a LRT statistic, i.e., twice of the difference between the maximum log-likelihood under the null and alternative hypotheses. The statistic follows the χ^2 distribution with one degree of freedom and thus $100(1-\alpha)$ percentile of χ_1^2 is the critical value for rejecting the null at the significance level of α .

2.9 | Extensions to other types of genetic data

Our method is extended to accommodate mixed SNP and sequences for children, both SNP and sequences for children, and all sequences (low or high coverage) for all family members. The last two parts in the log-likelihood functions change to conditional probabilities given observed sequence and/or genotype data. The penetrance function remains the same.

2.10 | Note about haplotype frequency inference

In our method, we estimated the haplotype frequencies ρ using all parental genotypes and a population-based expectation-maximization (EM) algorithm (Hawley & Kidd, 1995; Long, Williams, & Urbanek, 1995; Schaid, Rowland, Tines, Jacobson, & Poland, 2002). The estimates of the haplotype frequencies can be also obtained using additional sequence information. Specifically, given a set of initial values of the haplotype frequency estimates, the algorithm updates the estimates until convergence:

At step (k), the conditional probabilities of all haplotypes can be calculated by counting all possible parental diplotypes that match observed genotypes and sequences, and the estimates of haplotype frequencies can be updated by the following equation:

$$\hat{\rho}_{r,(k)} = \frac{1}{2n} \left[\sum_{(h_{m1}^i/h_{m2}^i, h_{p1}^i/h_{p2}^i) \sim G_p^i, G_s^i, S_c^i} [I(h_{m1}^i = h_r) + I(h_{m2}^i = h_r) + I(h_{p1}^i = h_r) + I(h_{p2}^i = h_r)] p(h_{m1}^i/h_{m2}^i, h_{p1}^i/h_{p2}^i | G_p^i, G_s^i, S_c^i) \right]$$

for $r = 1, \dots, t$.

Then the parental phasing probability $p(h_{m1}^i/h_{m2}^i, h_{p1}^i/h_{p2}^i | G_p^i, G_s^i, S_c^i)$, also part III in the likelihood, will be calculated using the final haplotype frequency estimates.

3 | SIMULATIONS

To evaluate the type-I error rate and power of our method, we conducted a series of simulations. Genotypes were generated based on a randomly picked 5-SNP region on chromosome 1 and many simulated 5-SNP blocks with different R^2 and minor allele frequencies (MAFs). Within the picked region, the actual haplotype frequencies in HapMap CEU samples were used. The adjacent locus distance was fixed at 120 bps initially. In each experiment (one dataset), 100 to 1,000 trio families were generated. In each family, four haplotypes of the parents were generated based on the haplotype frequencies and the child's haplotypes were inherited following the Mendelian law without recombination. For each individual, a pair-end sequence was randomly selected two-piece region with a fixed pair-end length and an insert size following a normal distribution with a mean of 1/5 of the pair-end length and a standard deviation (SD) of 1/20 of the pair-end length. The sequences had uniform coverage in the region and the mean coverage of the subject's sequences across loci had the same prespecified value. The middle SNP locus within the block was set to be the causal locus. A continuous covariate gestation age was generated from a normal distribution with a mean of 27 weeks and an SD of 2 weeks.

3.1 | Type-I error rate

Under the null hypothesis that there was no POE, the phenotype of the child was generated according to the model (1) using that causal SNP, with $\beta = \gamma_p = \gamma_m = 1$ and $\sigma^2 = 1$. We tested the POEs at the true causal locus using our sequence-based method (SM) given parental genotypes and child's sequence data. To compare with existing methods using only trio's genotypes, we also applied the previously developed haplotype-based method (HM) that demonstrated better power than a commonly used single-locus method. Moreover, we also evaluated our extended methods for augmented data types: (i) when both parents and children had sequences available, method for all sequences (ASM) was applied; (ii) when a half of the children had genotypes and the other half had sequences, method for mixed data (MM) was applied; and (iii) when children had both genotypes and sequences, the method applied was H + SM. Last, we fitted a linear regression model (LM) with the true parent-of-origin of children's alleles, which was supposed to have the best power possible given each sample and thus ideal for comparison.

We varied the pair-end read length from 100 to 250 bp and the mean read depth from 5× to 20×. Higher values in either the sequence read length or depth reflected better sequence coverage. For each set of parameters, we repeated the experiment 10,000 times.

3.2 | Power with complete data

Genotypes and sequence reads were generated as before. The phenotype of the child was generated from model (1) with different maternal and paternal effects, i.e., $\gamma_1 = 2\gamma_2$ and the specific value was determined so that the covariate, maternal, and paternal alleles explained 50%, 5%, and 1% of the phenotypic variance, respectively (a low heritability). Then HM, LM, SM, MM, and H + SM were applied. The same set of parameters including the family size, the mean read depth, and the pair-end read length were used as before. The experiment was repeated 1,000 times for each set of parameters.

We also evaluated the power of each method when maternal and paternal alleles explained 16% and 4% of the phenotype variance, which was likely for a gene or protein expression outcome.

When simulating sequence errors, we assumed that an error could occur with an equal fixed rate (e) ranging from 0.2% to 0.6% across base positions and then the original nucleoid code at the faulty position was mutated to any one of the other three. In the analysis, we first filtered out those that did not match either major or minor alleles and used the proportion of unmatched reads to estimate the error rate. Then we applied our models assuming the presence of sequence errors.

3.3 | Power with single parent data or untyped parental genotypes

To assess the information loss at the presence of missing genotypes, we further investigated the performance of both HM and our methods under two common missing scenarios. First, we considered the situation where the father's DNA was unavailable, so only the genotypes of the mother and children's sequence reads could be used to infer the parent of origin of children's haplotypes. Undoubtedly, there was more ambiguity in the origins of children's alleles, but more sequence reads with either more depth or longer reads might help recover the missing information. Second, all parental genotypes at the testing locus were assumed untyped or missing. Depending on the LD between the missing locus and neighbor loci and the children's sequence data, the missing alleles could be partially or completely inferred.

We also simulated haplotype blocks with various LD and MAFs. For a given R^2 , we generated multivariate continuous variables Z1–Z5 from a multivariate normal distribution

$$\begin{pmatrix} 1 & R^2 & R^4 & R^6 & R^8 \\ R^2 & 1 & R^2 & R^4 & R^6 \\ R^4 & R^2 & 1 & R^2 & R^4 \\ R^6 & R^4 & R^2 & 1 & R^2 \\ R^8 & R^6 & R^4 & R^2 & 1 \end{pmatrix}$$

with a mean of $\mathbf{0}$ and a variance-covariance matrix of $\begin{pmatrix} 1 & R^2 & R^4 & R^6 & R^8 \\ R^2 & 1 & R^2 & R^4 & R^6 \\ R^4 & R^2 & 1 & R^2 & R^4 \\ R^6 & R^4 & R^2 & 1 & R^2 \\ R^8 & R^6 & R^4 & R^2 & 1 \end{pmatrix}$ and then each variable of Z1–Z5 was dichotomized to a binary allele according to a fixed MAF. Each set of five alleles formed an independent parental haplotype. We varied R^2 from 0 to 1 and MAF from 0.05 to 0.5, fixed adjacent locus distance at 110 bps, read depth at 20 $\times$, and pair-end read length at 100 bps. For each fixed MAF and R^2 , experiments were repeated for 200 times.

Last, to understand the impact of various between-locus distances, we reran all simulations with inter-SNP distances artificially set at 50, 100, 200, and 300bps, keeping the genotype and phenotype data intact.

4 | RESULTS

Table 1 shows the empirical type I error rates at the nominal level of 0.05. The type I error rates of all our proposed methods and the HM using only genotypes were well reserved regardless of the number of trios, locus position, read length, and read depth. The empirical LRT followed the theoretical χ_1^2 distribution reasonably well except the tail probabilities.

Type I error rates were reserved for a sample size of 500 and more though can be inflated at a much higher significance level for a small sample size (Supporting Information Fig. S1).

Table 2 shows the power to detect a significant POE by SM and HM in contrast with the ideal LM when maternal and paternal heritability is 1% and 5% respectively. For a fixed sample and effect size, the power of detecting a POE increased as read length or the mean read depth increased. HM lost 4.2% (100 trios) to 12.2% (500 trios) of power compared with the optimal LM. Our proposed SM were always more powerful than HM and the power gains ranged from 3.4%, 5.8% to 10.0% for different mean read depths.

When a half of the children had genotype data but the other half had sequences, the power of MM was between HM and SM given the same depth and read length. When all children had both genotype and sequence data, the power of H + SM was higher than the comparable SM.

Figure 2 shows the power of detecting a significant POE at the same locus for HM, SM, and LM as in Table 2 under two different missing scenarios, compared with complete parental genotype case. The top and bottom panel is for low (1% and 5%) and high (4% and 16%) heritability, respectively.

When only single parents had genotype data (dotted lines), HM lost power up to 19% and our SM slightly less compared with both-parent scenario; the difference between single-parent and both-parent SMs seemed decreased when the mean read depth became larger. The comparison also demonstrated that deeper sequences on children could recover most of the missing contribution from the other parent and reach comparable power to that of HM using both parents. When parental genotypes at the testing locus were missing (dashed line), we observed a similar trend as in the missing parent scenario.

Figure 3 shows the power in relationship with the MAF at the testing locus and the average R^2 between the adjacent markers, at a low heritability. As MAFs increased, the power of both HM and SM increased. The power of SM was not sensitive to R^2 , while HM was the most powerful for blocks with medium R^2 (Feng, et al., 2011).

Figure 4 shows the relationship between the power of SM and the adjacent intermarker distance in 200 families for a low heritability and a pair-end length of 100 bp. Regardless of the mean read depth, the power of SM decreased with increasing adjacent SNP distances, especially in two missing scenarios. With complete parental data, the power reduced up to 14.6% when the distance changed from 50 to 300 bps, while in the locus missing scenario, the power reduced to 57.1%.

Table 3 shows the power of our proposed SM, MM, and H + SM at the presence of sequence errors, for a low heritability. At absence of sequence errors, deeper sequences increased the power. As errors occurred more often, the power of our methods reduced more though deeper sequences were more helpful to recover the lost power. The power reached the minimum when parents had low-pass sequences but no genotypes. With added genotypes to children's sequence data, our method had maintained the same power regardless of sequence error rate.

5 | DISCUSSION

In summary, we proposed a hybrid design and a powerful LRT to test the parent-of-origin of a SNP on an outcome. Our method is more powerful than the genotype-based method. The known factors influencing power the most were sample size and effect size.

Unique to sequence data, the read length and depth were also important and larger values of those lead to more power increase of our method over the genotype-based approach. It is ideal to have both parents' genotypes. However, if only single parent is available, it is desirable to have sufficient pair-end length and depth when sequencing children.

The other factors affecting power were specific to the characteristics of the testing locus, including its MAF and distances to other loci. When MAF is larger, SM has the best power. As SM does not require any LD within markers and most of the adjacent R^2 in human genome is lower than 0.01, it has broader applications than HM and sliding-window strategy can be fully adopted for genome-wide studies. The physical gaps between the testing locus and adjacent loci should be within the maximum pair-end sequence coverage and in general, shorter distances led to better power gain using our method. So for candidate gene studies, both LD and physical distances to others of the testing loci are informative to assess power.

Because sequence errors are common (Wall et al., 2014), we also considered a model accommodating errors. When parental sequences were low-pass, the power can be lower than that knowing only genotype data. So we suggest supplementing genotypes to sequences data to improve the phasing quality, if resources are available.

To analyze sequence data in families, Kong et al. (2009) used a combination of genealogy and long-range stitching approach to identify parent-of-origins of each genome sequence. Overlapping sequence tiles with each 6 cm in length were checked against the parental genotypes and then consecutive tiles were stitched together to form longer haplotypes and determine parental origins. At present, most laboratories cannot handle such long-range technology; because there are more variants in one long sequence, mapping algorithm needs to be tailored, which can be a daunting task itself. In addition, longer pair-end reads cost much more for the same amount of sequence reads. At University of Pennsylvania, increasing from 250 to 300 bp read length will cost five times more to generate the same number of sequence reads. From our result, we can see that with a read length of 250 bps and a mean read depth of 20 $\times$, our model can achieve almost the same power as that using the HM. Furthermore, although our current proposed method has focused on identifying POEs on a continuous outcome, the method can be revised to fit binary outcomes and case-control studies (Lin, Weinberg, Feng, Hochner, & Chen, 2013).

Our method is easily extended to nuclear families with multiple siblings. The inheritance probability of one child will be independent from other siblings so the overall inheritance will be the product of each child's inheritance probability. The complexity for nuclear families including more than one child is that a random effect term representing the common residual family effect needs to be added, which is beyond the scope of this article.

Our method is a one-step procedure. For whole-genome analysis, an alternative approach is to prescreen loci with only significant result for a joint association test of $\gamma_1 = \gamma_2 = 0$ and then to test $\gamma_2 = 0$ in a limited number of loci (Shete & Yu, 2005). The advantage of such method is that the number of tests performed in step 2 can be greatly reduced and thus each test can use a less stringent significance level to obtain better power. But the screening test has a degree of freedom of 2 and may miss loci with relatively large POEs but minor overall association effects.

Our study has limitations. First, our method depends on the population haplotype frequency estimates though it is not sensitive to inaccurate inference. Second, our likelihood does not incorporate the uncertainties in the population haplotype frequencies. For testing haplotype-disease association, others have averaged test statistics over many plausible estimates for the haplotypes to allow for such uncertainty (Stephens & Donnelly, 2003). It warrants further investigation whether we should adopt similar strategies. Third, our model allowing for sequencing errors is overly simplistic when the error rate was assumed to be constant across all loci and the mutation possibility was uniform for different nucleotides. Fourth, structural variation, including copy number variation (CNV) and short indels, can affect the mapping of genomic sequences. Sequence reads at these sites and inheritance probabilities need to be handled differently in more advanced approaches. Last, for the SNPs that are further away from adjacent markers and with little LD with other markers, neither haplotype-based model nor our model can help.

Our program will be available from <https://dbe.med.upenn.edu/biostat-research/RuiFeng>.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors want to thank Dr. Thomas J. Mariani at University of Rochester Medical Center and two Genetic Epidemiology reviewers for their meticulous review and valuable suggestions.

References

- Donlon TA. Similar molecular deletions on chromosome 15q11.2 are encountered in both the Prader-Willi and Angelman syndromes. *Human Genetics*. 1988; 80(4):322–328. [PubMed: 3198109]
- Duyzend MH, Nuttle X, Coe BP, Baker C, Nickerson DA, Bernier R, Eichler EE. Maternal modifiers and parent-of-origin bias of the autism-associated 16p11.2 CNV. *American Journal of Human Genetics*. 2016; 98(1):45–57. Retrieved from <https://doi.org/10.1016/j.ajhg.2015.11.017>. [PubMed: 26749307]
- Feng R, Wu Y, Jang GH, Ordovas JM, Arnett D. A powerful test of parent-of-origin effects for quantitative traits using haplotypes. *PLoS One*. 2011; 6(12):e28909. Retrieved from <https://doi.org/10.1371/journal.pone.0028909>. [PubMed: 22174922]
- Hanson RL, Kobes S, Lindsay RS, Knowler WC. Assessment of parent-of-origin effects in linkage analysis of quantitative traits. *American Journal of Human Genetics*. 2001; 68(4):951–962. Retrieved from <https://doi.org/10.1086/319508>. [PubMed: 11254452]
- Hawley ME, Kidd KK. Haplo—A program using the Em algorithm to estimate the frequencies of multisite haplotypes. *Journal of Heredity*. 1995; 86(5):409–411. [PubMed: 7560877]

- Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Stefansson K. Parental origin of sequence variants associated with complex diseases. *Nature*. 2009; 462(7275): 868–874. Retrieved from <https://doi.org/10.1038/nature08625>. [PubMed: 20016592]
- Lin DY, Hu Y, Huang BE. Simple and efficient analysis of disease association with missing genotype data. *American Journal of Human Genetics*. 2008; 82(2):444–452. Retrieved from <https://doi.org/10.1016/j.ajhg.2007.11.004>. [PubMed: 18252224]
- Lin DY, Weinberg CR, Feng R, Hochner H, Chen JB. A multi-locus likelihood method for assessing parent-of-origin effects using case-control mother-child pairs. *Genetic Epidemiology*. 2013; 37(2): 152–162. Retrieved from <https://doi.org/10.1002/gepi.21700>. [PubMed: 23184538]
- Long JC, Williams RC, Urbanek M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*. 1995; 56(3):799–810. [PubMed: 7887436]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009; 461(7265):747–753. Retrieved from <https://doi.org/10.1038/nature0849>. [PubMed: 19812666]
- Metzker ML. Sequencing technologies—The next generation. *Nature Reviews Genetics*. 2010; 11(1): 31–46. Retrieved from <https://doi.org/10.1038/nrg2626>.
- Poindexter BB, Feng R, Schmidt B, Aschner JL, Ballard RA, Hamvas A, Prematurity and Respiratory Outcomes Program. Comparisons and limitations of current definitions of bronchopulmonary dysplasia for the prematurity and respiratory outcomes program. *Annals of the American Thoracic Society*. 2015; 12(12):1822–1830. Retrieved from <https://doi.org/10.1513/AnnalsATS.201504-218OC>. [PubMed: 26397992]
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics*. 2002; 70(2):425–434. Retrieved from <https://doi.org/10.1086/338688>. [PubMed: 11791212]
- Shete S, Amos CI. Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *American Journal of Human Genetics*. 2002; 70(3):751–757. Retrieved from <https://doi.org/10.1086/338931>. [PubMed: 11836650]
- Shete S, Zhou XJ, Amos CI. Genomic imprinting and linkage test for quantitative-trait loci in extended pedigrees. *American Journal of Human Genetics*. 2003; 73(4):933–938. Retrieved from <https://doi.org/10.1086/378592>. [PubMed: 13680523]
- Shete S, Yu R. Genetic imprinting analysis for alcoholism genes using variance components approach. *BMC Genetics*. 2005; 6(Suppl 1):S161. [PubMed: 16451623]
- Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*. 2003; 73:1162–1169. [PubMed: 14574645]
- Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, Schaefer C, Risch N. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research*. 2014; 24(11):1734–1739. [PubMed: 25304867]
- Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *American Journal of Human Genetics*. 1999; 65(1):229–235. Retrieved from <https://doi.org/10.1086/302466>. [PubMed: 10364536]
- Weksberg R, Teshima I, Williams BR, Greenberg CR, Puschel SM, Chernos JE, Squire J. Molecular characterization of cytogenetic alterations associated with the Beckwith-Wiedemann syndrome (BWS) phenotype refines the localization and suggests the gene for BWS is imprinted. *Human Molecular Genetics*. 1993; 2(5):549–556. [PubMed: 8518793]
- Wermter AK, Scherag A, Meyre D, Reichwald K, Durand E, Nguyen TT, Brönnner G. Preferential reciprocal transfer of paternal/maternal DLK1 alleles to obese children: First evidence of polar overdominance in humans. *European Journal of Human Genetics*. 2008; 16(9):1126–1134. Retrieved from <https://doi.org/10.1038/ejhg.2008.64>. [PubMed: 18398438]
- Whittaker JC, Gharani N, Hindmarsh P, McCarthy MI. Estimation and testing of parent-of-origin effects for quantitative traits. *American Journal of Human Genetics*. 2003; 72(4):1035–1039. Retrieved from <https://doi.org/10.1086/374382>. [PubMed: 12644965]

Xiao FF, Ma JZ, Amos CI. A Unified Framework Integrating Parent-of-Origin Effects for Association Study. PLoS One. 2013; 8(8):e72208. Retrieved from <https://doi.org/ARTN e7220810.1371/journal.pone.0072208>. [PubMed: 23991061]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

FIGURE 1. Identify the parent-of-origins of child's alleles at SNP3 within a 5-SNP region using genotype and sequence data

FIGURE 2.

Power of detecting a significant parent-of-origin effect using three methods under two missing scenarios, compared with the complete data scenario, (A) for low heritability ($h^2 = 6\%$) and (B) for high heritability ($h^2 = 20\%$)

FIGURE 3.
Power for various minor allele frequencies and adjacent R^2

FIGURE 4.

Power for various adjacent SNP distances, when sample size is 500 and sequence end length is 100 bps

TABLE 1 Type I errors of test of parent-of-origin effect in trios with different sample sizes at $\alpha = 0.05$

Method	Depth	Number of trios		
		200	500	1000
HM (SNP genotype for all children)	–	0.054	0.051	0.053
MM (SNP genotype for 50% of children and sequence for 50% of children)	20×	0.053	0.051	0.053
SM (sequence for all children)	5×	0.053	0.051	0.052
	10×	0.054	0.052	0.053
	20×	0.054	0.050	0.054
H + SM (both SNP genotype and sequence for all children)	20×	0.053	0.050	0.054
LM		0.053	0.051	0.053

TABLE 2
Power of detecting a significant parent-of-origin effect at a testing locus within a random c-SNP block

Method	Depth	Read length (bp)	Number of trios		
			100	200	500
HM	–	–	0.195	0.358	0.698
MM (SNP genotype for 50% children and sequence for 50% of children)	20×	100	0.208	0.387	0.740
SM (sequence for all children)	5×	100	0.228	0.392	0.733
	10×	100	0.237	0.423	0.756
	20×	100	0.241	0.424	0.754
H + SM (both SNP genotype and sequence for all children)	20×	100	0.241	0.424	0.754
LM			0.249	0.465	0.815

TABLE 3

Power at presence of sequence error ($n = 500$)

Method	Parents data type	Children data type	Locus-wise error rate				
			0%	0.2%	0.4%	0.6%	
SM	Genotype	Sequence 5×	0.795	0.791	0.791	0.790	
		Sequence 10×	0.809	0.807	0.805	0.804	
		Sequence 20×	0.810	0.807	0.807	0.807	
H + SM	Genotype	Sequence 5× + genotype	0.806	0.803	0.801	0.802	
		Sequence 10× + genotype	0.812	0.811	0.810	0.810	
		Sequence 20× + genotype	0.812	0.811	0.811	0.812	
SM	Sequence 5×	Sequence 5×	0.821	0.816	0.759	0.747	
		Sequence 10×	0.825	0.820	0.783	0.768	
		Sequence 20×	0.826	0.823	0.798	0.788	