Published in final edited form as:

Genet Epidemiol. 2013 February ; 37(2): 152–162. doi:10.1002/gepi.21700.

A Multi-Locus Likelihood Method for Assessing Parent-of-Origin Effects Using Case-Control Mother-Child Pairs

Dongyu Lin¹, Clarice R. Weinberg², Rui Feng¹, Hagit Hochner³, and Jinbo Chen^{1,*}

¹Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

²Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangular Park, North Carolina

³Braun School of Public Health, Hebrew University-Hadassah Medical Center, Jerusalem, Israel

Abstract

Parent-of-origin effects have been pointed out to be one plausible source of the heritability that was unexplained by genome-wide association studies. Here, we consider a case-control mother-child pair design for studying parent-of-origin effects of offspring genes on neonatal/early-life disorders or pregnancy-related conditions. In contrast to the standard case-control design, the case-control mother-child pair design contains valuable parental information and therefore permits powerful assessment of parent-of-origin effects. Suppose the region under study is in Hardy-Weinberg equilibrium, inheritance is Mendelian at the diallelic locus under study, there is random mating in the source population, and the SNP under study is not related to risk for the phenotype under study because of linkage disequilibrium (LD) with other SNPs. Using a maximum likelihood method that simultaneously assesses likely parental sources and estimates effect sizes of the two offspring genotypes, we investigate the extent of power increase for testing parent-of-origin effects through the incorporation of genotype data for adjacent markers that are in LD with the test locus. Our method does not need to assume the outcome is rare because it exploits supplementary information on phenotype prevalence. Analysis with simulated SNP data indicates that incorporating genotype data for adjacent markers greatly help recover the parent-of-origin information. This recovery can sometimes substantially improve statistical power for detecting parent-of-origin effects. We demonstrate our method by examining parent-of-origin effects of the gene PPARGC1A on low birth weight using data from 636 mother-child pairs in the Jerusalem Perinatal Study.

Keywords

case-control study; imprinting; mother-child pair data; parent-of-origin effect

*Correspondence to: Jinbo Chen, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104. jinboche@mail.med.upenn.edu.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

Introduction

Genome-wide association studies (GWAS) have been successful in identifying common genetic variants that are associated with complex human traits. However, these newly identified variants generally have small effect sizes and thus can only explain a small proportion of genetic heritability [Eichler et al., 2010]. In the effort to understand this missing heritability, genetic effects related to parent-of-origin, such as imprinting effects or maternal genetic effects beyond the transmitted gene copy, have recently become a research focus [e.g., Kong et al., 2009]. Imprinting is an epigenetic phenomenon where the two gene copies at a locus have unequal expression and thus have different contributions to a phenotype. For autosomal genes, this asymmetry can arise through a mechanism called imprinting, where the effect of a gene copy depends on whether it was inherited from the father or from the mother. In other words, the parent of origin acts as an effect modifier for the allele effect [Dudbridge, 2008]. Imprinted genes are known to play an important role in certain neonatal disorders, such as Beckwith-Wiedemann, Prader-Willi, and Angelman syndromes [e.g., Falls et al., 1999]. Evidence has been emerging that fetal imprinted genes may also be involved in the development of maternal pregnancy-related conditions [See e.g., Kanayama et al., 2003; Petry et al., 2007; Saftlas et al., 2005; Wangler et al., 2005].

To date, most studies of parent-of-origin effects used traditional family-based designs, which provide most of the required parental-source information for children's alleles. But it can be relatively expensive to recruit families. The case-control design has been popularly used for GWAS because of its high cost-effectiveness [Risch and Merikangas, 1996]. Ordinary case-control data, however, does not provide parental information and thus does not permit the evaluation of parent-of-origin effects. Here, we consider an intermediate solution that recruits mother-child pairs in the case-control study, which we term as case-control mother-child design. This design calls for collecting phenotype data for pairs of cases and controls, where the case can either be the mother (e.g., for a pregnancy complication) or the child (e.g., for a birth defect) depending on the phenotype of interest, and one genotypes both the mother and child. Because maternal genotype data contains partial parental source information [Weinberg and Shi, 2009], this design allows one to assess parent-of-origin effects of children's genes. This design allows estimation and testing of separate and joint effects of both maternal and child genomes on either children's outcomes or pregnancy-related outcomes in mothers [Weinberg and Umbach, 2005]. Investigating parent-of-origin effects with mother-child pair data has been previously studied in the literature [See e.g., Ainsworth et al., 2011; Weinberg and Shi, 2009], with a focus on single SNP approaches that apply log-linear models for the estimation of relative risks.

In this work, we investigate the improvement in the power of the SNP-based method that can be achieved by incorporating genotype data for loci in linkage disequilibrium (LD) with the test locus in a maximum likelihood method. The genotype data for nearby loci can help identify parental sources of children's alleles [Browning and Browning, 2009; Kong et al., 2009], thereby improving the precision of estimation and the power of testing the parent-of-origin effect parameter. We develop an accompanying Expectation-Maximization (EM) algorithm to maximize the likelihood. We assume that information is available on the total

numbers of cases and controls in the cohort that would be available for case and control recruitment. This supplemented case-control design conforms to typical real settings for studying pregnancy-related disorders, owing to their relatively high incidence (e.g., the incidence of premature birth is around 12% in the United States). In our analysis of low birth weight in the Jerusalem Perinatal Study (JPS; Section “Analyzing Parent-of-Origin Effects of the Gene PPARGC1A on Low Birth Weight”), the total numbers of newborn babies who had normal or low birth weight were known for inclusion in the analysis.

The rest of this paper is organized as follows. We present the logistic regression model for analysis and a single SNP-based maximum likelihood method for the estimation and testing of parent-of-origin effects of children's genes in Section “Methods.” In Section “Maximum Likelihood Method that Exploits Genotype Data of Nearby Markers via Haplotypes,” we describe a maximum likelihood method that improves on the single-SNP method by exploiting genotype data for nearby markers in LD with the test locus. Results from extensive simulation studies are presented in Section “Simulation Studies” to assess the performance of our proposed method. In Section “Analyzing Parent-of-Origin Effects of the Gene PPARGC1A on Low Birth Weight,” we apply the proposed method to explore possible parent-of-origin effects of the gene PPARGC1A on low birth weight, using data from the JPS. We conclude with discussion in Section “Discussion.”

Methods

Without loss of generality, we assume a diallelic locus with a minor allele “ a ” and a wild-type allele “ A .” This locus may affect the phenotype status of an individual, which is denoted as Y and takes values “1” or “0”, depending on the presence or absence of the phenotype, respectively. Let p_a be the minor allele frequency (MAF). The child genotype g^c consists of two alleles, g_m^c and g_p^c , which were inherited from the mother and father, respectively. Next, we refer to the ordered pair (g_m^c, g_p^c) as the *sourced genotype*. For example, (A, a) refers to genotype Aa with A inherited from the mother and a from the father. Let g^m , g^p , and g^c denote the maternal, paternal, and child genotype, respectively. Numbers “0,” “1,” and “2” are used to represent genotypes AA , Aa , and aa in the mathematical equations below.

A Logistic Regression Model for Parent-of-Origin Effects

The logistic regression model has been commonly used for analyzing case-control genetic association studies and we use it to quantify a parent-of-origin effect of genes on the phenotype. For a rare outcome, this is equivalent to previous log-linear models [Ainsworth et al., 2011; Weinberg and Shi, 2009]. The following model allows maternal effects at the same locus, which is important since maternal effects could confound offspring genetic effects:

$$\text{logit } p(Y=1|g^m, g_m^c, g_p^c) = \beta_0 + \beta_1 f(g^m) + \gamma_1 I(g_m^c=a) + \gamma_2 I(g_p^c=a) + \gamma_3 I(g^c=2).$$

Here, indicator function “ $I(\text{statement})$ ” is 1 if the statement is true and 0 otherwise, β_1 is the log odds ratio (OR) parameter(s) that quantifies the maternal genetic effect, γ_1 and γ_2 are the

respective log OR parameters that quantify effects of single maternally and paternally inherited alleles, and γ_3 is the log OR parameter that quantifies the departure of the effects of the two alleles from an additive model. The numerical coding $f(g^m)$ for the maternal genotype can take different forms. For example, it can be “co-dominant,” with two indicator functions, $I(g^m = 1)$ and $I(g^m = 2)$, with β_1 becoming a vector of parameters. A reformulation of the above model can help clarify the nature of the parent-of-origin effect. The above model can be rewritten as follows:

$$\text{logit } p(Y=1|g^m, g_m^c, g_p^c) = \beta_0 + \beta_1 f(g^m) + \beta_2 g^c + \beta_3 \{I(g_m^c=a) - I(g_p^c=a)\} + \gamma_3 I(g^c=2) \quad (1)$$

Under this model, the parent-of-origin effect is quantified as β_3 , and the null hypothesis of no parent-of-origin effect is formulated as $\beta_3 = 0$. The parameter corresponding to the unsourced child genotype, β_2 , partials out the additive effect of the two alleles. The parent-of-origin parameter β_3 is one-half of the natural logarithm of the OR for (a, A) divided by the OR for (A, a) . Listed in Table 1 below are the possible sourced offspring genotypes inferred from all seven combinations of mother-child genotype pairs, their population distributions, and the corresponding log ORs under model (1). Model (1) can be modified to use a log link in place of the logistic link, in which case one is estimating risk ratios (and ratios of risk ratios for assessing imprinting).

Under model (1), one can test whether the offspring genotype g^c is associated with the phenotype Y , by formulating the null hypothesis as $\beta_2 = \beta_3 = \gamma_3 = 0$. But one would not test hypothesis $\beta_2 = 0$, for the same reason that one would not test the genetic main effect in the presence of gene-environment interactions. When association of individual maternal or paternal allele is of interest, it is easier to formulate the null hypothesis as $\gamma_1 = 0$ or $\gamma_2 = 0$ in the first model. The single parameter β_3 in model (1) allows more explicit evaluation of the parent-of-origin effect. Hereafter, we will denote $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \gamma_3)$ as the vector of log OR parameters of interest.

Likelihood Function for the Case-Control Mother-Child Pair Data at the Test Locus

We consider maximum likelihood methods for the estimation and testing of parent-of-origin effects with case-control mother-child pair data, assuming penetrance model (1). We assume that we know or can approximate N_1 and N_0 , the respective total numbers of mother/child pairs who are eligible for selection. Let n_1 and n_0 be the respective numbers of sampled case and control mother-child pairs. The obvious challenge for fitting the aforementioned model is that the parental origin of the two alleles is unknown for heterozygous children born to heterozygous mothers, as shown in Table 1. That is, if both g^m and g^c are equal to 1, one cannot infer whether the sourced offspring genotype is (A, a) or (a, A) . Simply ignoring these pairs may lead to bias and loss of efficiency [See, e.g., Weinberg, 1999] and thus is not desirable. We consider a maximum likelihood estimation (MLE) method for estimating the OR parameters of interest. To this end, we further assume random mating, Hardy-Weinberg Equilibrium (HWE), and Mendelian transmission in the parental population, such that the joint probability of maternal and child's genotypes $p(g^m, g^c)$ is a function of the MAF, p_a (Table 1). We use θ to denote the logit of p_a . Then, neglecting combinatorics, the retrospective likelihood function based on the observed case-control data can be written as

$$L_1(\mathbf{y}, g^m, g^c; \boldsymbol{\beta}, \theta) = \prod_{u=1}^{n_0+n_1} p(g_u^m, g_u^c | y_u) \prod_{y=0,1} p(Y=y)^{N_y}, \quad (2)$$

where

$$p(g^m, g^c | y) = \frac{p(g^m, g^c) p(y | g^m, g^c, \theta)}{\sum_{m', c'} p(g^{m'}, g^{c'}) p(y | g^{m'}, g^{c'}, \theta)}.$$

This likelihood has a similar form as that proposed in Dudbridge [2008] for analyzing SNP and haplotype effects with incomplete genotype data from nuclear families and unrelated subjects. Note that except when both the mother and child are heterozygous, the two alleles g_m^c and g_p^c are known (Table 1) and therefore $p(y | g^m, g^c) = p(y | g^m, g_m^c, g_p^c)$. If both the mother and child are heterozygous, then

$$p(y | g^m = g^c = 1) = p_a p(y | g^m = 1, g_m^c = A, g_p^c = a) + (1 - p_a) p(y | g^m = 1, g_m^c = a, g_p^c = A).$$

We obtain the MLE of parameters $\boldsymbol{\beta}$ by maximizing the likelihood function (2) with respect to $\boldsymbol{\beta}$ and θ . The variance-covariance matrix of the MLE for $(\boldsymbol{\beta}, \theta)$ can be estimated as the inverse of the observed information matrix. The parent-of-origin effect can be assessed using the standard Wald or a likelihood ratio test. Provided that the maternal effect is correctly modeled, the test statistic asymptotically follows a $\chi^2_{(1)}$ distribution under the null hypothesis. Careful attention must be paid to the modeling of the maternal genetic effect, i.e., the coding of $f(g^m)$, to ensure that the test for a parent-of-origin effect is valid. The safest approach is to allow a two-parameter, co-dominant modeling for the maternal effect.

Paternal genotype data, if available, can be exploited to increase the power of testing parent-of-origin effects. Paternal genotype data help resolve the parental origin of off-spring alleles in heterozygous mother-child pairs, and can also help improve the precision of estimating the parameter θ . Moreover, paternal genotypes can be incorporated into the likelihood-based analysis to improve the efficiency of estimating the association parameter $\boldsymbol{\beta}$. This can be achieved simply by replacing the probability $p_\theta(g^m, g^c)$ in L_1 with $p_\theta(g^m, g^c, g^p)$. Parental origins of most offspring genotypes in mother-father-child trios can be determined except when all three are heterozygous [Weinberg et al., 1998] (also see Supplementary Table S4).

In that case, g_m^c and g_p^c are equally likely to be the minor allele in the source population, and therefore, the conditional probability of the phenotype, $p(y | g^m = g^c = g^p = 1)$, is equal to the average of the two probabilities, $p(y | g^m = g^c = g^p = 1, g_p^c = a)$ and $p(y | g^m = g^c = g^p = 1, g_p^c = A)$.

Maximum Likelihood Method That Exploits Genotype Data of Nearby Markers via Haplotypes

In this section, we propose to improve power of the SNP-based likelihood method in Section “Methods” by incorporating genotype data for markers that are near to the test locus. When

both the mother and child are heterozygous at the test locus, haplotypes in the genomic region spanned by the test locus and nearby markers in LD with it can help identify the parental origin at the test locus [Browning and Browning, 2009; Kong et al., 2009]. For instance, consider a simple situation where genotype data are also available for a nearby diallelic marker with alleles B and b . Suppose the B/b locus is close enough that we can safely assume the rate of recombination between A/a and B/b in a single meiosis is zero. Assume that possible haplotypes in the population are known to be AB , Ab , and aB . Then, if the maternal bilocus genotype is $\{Aa, BB\}$ and the child's genotype is $\{Aa, bB\}$, the mother's haplotypes must be AB – aB , and the child inherited either (aB, Ab) or (AB, ab) from the parents. Since the latter pair does not occur in the population, the child's sourced genotype can be inferred as $(g_m^c, g_p^c) = (a, A)$. Parental source can be similarly inferred for maternal genotype $\{Aa, Bb\}$ and child genotype $\{Aa, BB\}$. This simple example demonstrates that genotype data for nearby marker loci can help identify parental origins of children's alleles via haplotypes. Readers are referred to Browning and Browning [2011] for more comprehensive discussions of this idea. Multiple markers in a genomic region, as haplotype tagging SNPs, are often genotyped in practice. Therefore, we propose to incorporate marker genotype data into the maximum likelihood method to increase the efficiency of parameter estimation.

Let $G^m = \{g_1^m, g_2^m, \dots, g_K^m\}$ and $G^c = \{g_1^c, g_2^c, \dots, g_K^c\}$ denote the maternal and offspring genotypes at the test locus and $K - 1$ nearby markers, where the first component refers to the test locus. We assume the K markers are close enough together that the probability of a recombination event in a single meiotic division is effectively 0. Let $\{h_1, h_2, \dots, h_S\}$ denote the S possible haplotypes in the genomic region spanned by the K loci, and let $\pi = (\pi_1,$

$\pi_2, \dots, \pi_S)$, with $\sum_{s=1}^S \pi_s = 1$, denote the corresponding haplotype frequencies, which typically are unknown. Next, we refer to the haplotype pair for a mother, $h_i^m h_j^m$, as the maternal diplotype, and the ordered haplotype pair for a child, (h_w^c, h_l^c) , as the sourced offspring diplotype. Here the subscripts i, j , and l refer to the i th, j th, and l th haplotypes in the haplotype set $\{h_1, h_2, \dots, h_S\}$, respectively, and h_w^c refers to the maternally inherited copy so that $w = i$ or $w = j$. To account for haplotype phase ambiguity, as has been done for haplotype-based association analysis [e.g., Chen et al., 2004; Chen and Chatterjee, 2006; Dudbridge, 2008; Epstein and Satten, 2003], we compute the likelihood by summing over individual likelihoods determined by possible diplotypes weighted by the diplotype probabilities. We assign probabilities to possible diplotypes under HWE, $p(h_i h_j) = 2^{I(i \neq j)} \pi_i \pi_j$.

We propose to jointly estimate the association parameters in model (1), haplotype frequencies π , and the parental origin of children's alleles using the likelihood function below based on all data (Y, G^m, G^c) . Let $\theta = (\theta_1, \theta_2, \dots, \theta_{S-1})$ be the polytomous logistic reparameterization of haplotype frequencies $\pi = (\pi_1, \pi_2, \dots, \pi_S)$, where

$\pi_s = \exp \theta_s / (1 + \sum_{s'=1}^{S-1} \exp \theta_{s'})$ for $1 \leq s \leq S - 1$, and $\pi_S = 1 / (1 + \sum_{s'=1}^{S-1} \exp \theta_{s'})$. Let $H_u^m = \{h_{iu}^m h_{ju}^m\} \sim G_u^m$ denote the set of diplotypes that are compatible with the genotype of the u th mother, $G_u^m = \{g_u^m, g_{2u}^m, \dots, g_{Ku}^m\}$, and let $H_u^c = \{(h_{wu}^c, h_{lu}^c)\} \sim G_u^c$ denote the set of sourced diplotypes that are compatible with the genotype of the u th child,

$G_u^c = \{g_u^c, g_{2u}^c, \dots, g_{K_u}^c\}$. Let $H(G^m, G^c) = \{h_i^m h_j^m, (h_w^c, h_l^c)\} \sim (G^m, G^c)$ be the collection of all distinct phased maternal and sourced offspring haplotype pairs that are simultaneously compatible with the mother-child pair with genotype (G^m, G^c) . We next make the assumption that the risk of the phenotype depends on the genotype at the test locus, and once we condition on the test locus, the risk does not depend on the genotype of any nearby loci. This assumption is equivalent to saying that the nearby SNPs serve only to provide information related to the parental source for the test SNP. The likelihood function can then be written as follows:

$$L_3(y; \beta, \theta) = \prod_{u=1}^{n_0+n_1} p(y_u, G_u^m, G_u^c) \prod_{y=0,1} p(Y=y)^{N_y-n_y}$$

$$= \prod_{u=1}^{n_0+n_1} \sum_{h_{iu}^m h_{ju}^m \in H_u^m, (h_{wu}^c, h_{lu}^c) \in H_u^c} p(y_u | g_u^m, g_{mu}^c, g_{pu}^c) p_\theta(h_{iu}^m h_{ju}^m, h_{wu}^c, h_{lu}^c) \times \prod_{y=0,1} \left\{ \sum_{(G^m, G^c) h_i^m h_j^m, (h_w^c, h_l^c) \in H(G^m, G^c)} p(Y=y) \right\}$$

Here, under HWE, random mating, and Mendelian inheritance, the joint probability $p_\theta(h_i^m, h_j^m, h_w^c, h_l^c)$ is equal to $\pi_i \pi_j \pi_l$. The parental origin of the two child's alleles at the test locus can be directly obtained from the sourced child's haplotype pair (h_w^c, h_l^c) . As mentioned above, the haplotype phase helps improve the resolution for identifying parental sources if the number of possible haplotypes in the population is small. The consequent statistical inference of the parent-of-origin effect is similar to that in Subsection “Likelihood Function for the Case-Control Mother-Child Pair Data at the Test Locus,” with the single SNP-based likelihood replaced by the above likelihood that incorporates the multilocus genotype data. Unknown parameters now include the reparameterized haplotype frequencies θ in addition to the association parameters β in model (1).

Direct maximization of L_3 involves a relatively high-dimensional optimization problem because the number of possible haplotypes in the population, S , could be large. We thus prefer to maximize L_3 via an iterative EM algorithm, in line with Excoffier and Slatkin [1995] and Epstein and Satten [2003]. It is implemented by iterating the following two steps. In step 1, we estimate the probability of each distinct haplotype. Let v_{ij}^s denote the number of copies of haplotype h_s in $\{h_i^m, h_j^m, h_l^c\}$, the two maternal haplotypes and the paternally inherited offspring haplotype. Number v_{ij}^s takes values 0, 1, 2, or 3. Define

$\hat{p}^{(r-1)}(y | g^m, g_m^c, g_p^c)$ as the estimated penetrance function by replacing β with its estimate at the $r-1$ st iteration, $\hat{\beta}^{(r-1)}$. Starting values of haplotype frequencies $\hat{\pi}^{(0)}$ can be obtained in two steps: first apply the EM algorithm [Excoffier and Slatkin, 1995] separately to the genotype data of case and control mothers; then combine the two sets of estimates using the inverse probability weighted method that accounts for case-control sampling (See Section “Analyzing Parent-of-Origin Effects of the Gene PPARGC1A on Low Birth Weight” for more details). In the E-step at the r th iteration, the expected number v_{ij}^s for a case or control with genotype data available is computed as

$$E \left\{ v_{ijl}^s \middle| y_u, G_u^m, G_u^c; \beta^{(r-1)}, \hat{\pi}^{(r-1)} \right\} = \frac{\sum_{h_{iu}^m h_{ju}^m \in H_u^m, (h_{wu}^c, h_{lu}^c) \in H_u^c} v_{ijl}^s \hat{p}^{(r-1)}(y_u | g_u^m, g_{mu}^c, g_{pu}^c) \hat{p}_{\hat{\pi}^{(r-1)}}(h_{iu}^m h_{ju}^m, h_{wu}^c, h_{lu}^c)}{\sum_{h_{iu}^m h_{ju}^m \in H_u^m, (h_{wu}^c, h_{lu}^c) \in H_u^c} \hat{p}^{(r-1)}(y_u | g_u^m, g_{mu}^c, g_{pu}^c) \hat{p}_{\hat{\pi}^{(r-1)}}(h_{iu}^m h_{ju}^m, h_{wu}^c, h_{lu}^c)},$$

and that for a subject who only has phenotype data available is computed as

$$E \left\{ v_{ijl}^s \middle| y; \beta^{(r-1)}, \hat{\pi}^{(r-1)} \right\} = \frac{\sum_{h_i^m h_j^m \in H, (h_w^c, h_l^c) \in H} v_{ijl}^s \hat{p}^{(r-1)}(Y=y | g^m, g_m^c, g_p^c) \hat{p}_{\hat{\pi}^{(r-1)}}(h_i^m h_j^m, h_w^c, h_l^c)}{\sum_{h_i^m h_j^m \in H, (h_w^c, h_l^c) \in H} \hat{p}^{(r-1)}(Y=y | g^m, g_m^c, g_p^c) \hat{p}_{\hat{\pi}^{(r-1)}}(h_i^m h_j^m, h_w^c, h_l^c)}.$$

Therefore, the estimated frequency for the s th haplotype is updated at the r th step using the following formula:

$$\hat{\pi}_s^{(r)} = \frac{1}{3(N_1 + N_0)} \left\{ \sum_{u=1}^{n_0+n_1} E \left\{ v_{ijl}^s \middle| y_u, G_u^m, G_u^c; \hat{\beta}^{(r-1)}, \hat{\pi}^{(r-1)} \right\} + \sum_{y=0,1} (N_y - n_y) E \left\{ v_{ijl}^s \middle| y; \hat{\beta}^{(r-1)}, \hat{\pi}^{(r-1)} \right\} \right\},$$

where $3(N_1 + N_0)$ is the total number of haplotype copies in $N_1 + N_0$ pairs. In the M-step, with π being fixed at $\hat{\pi}^{(r)}$, we maximize the likelihood function L_3 with respect to β to obtain the updated estimate $\hat{\beta}^{(r)}$. The final estimates are those obtained when the algorithm converges.

In general, genotype data for markers in LD with the test locus can help infer but cannot completely resolve the uncertain parental origin of the child's alleles at the test locus. For example, assuming both the mother and child are heterozygous at the test locus, if the diplotype configurations for both of them are $h_i h_j$ ($i \neq j$), the parental origin of the child's alleles will remain ambiguous. Suppose that the minor allele a resides on haplotype h_i but not on haplotype h_j . Then the probability that a was inherited from the mother is $\pi_j / (\pi_i + \pi_j)$. However, when the number of possible haplotypes S in the population is small compared to the theoretical number 2^K , the chance of having this source ambiguity becomes small, and we potentially gain power from the successfully resolved subjects. On the other hand, the increased number of nuisance haplotype frequency parameters can lead to a loss of power. We expect that the trade-off between the information loss due to the need to estimate haplotype frequencies balanced against the improvement in the resolution of parental source is relevant to the power of our haplotype-based likelihood method.

Simulation Studies

We designed simulation studies to evaluate the performance of the proposed likelihood method, with a focus on the power of likelihood ratio tests for detecting parent-of-origin effects. We first considered the simple situation when genotype data are available only for the test locus. We then assessed the extent of power improvement offered by incorporating

marker genotype data. Lastly, we investigated the robustness of the proposed method to violations of the assumptions of HWE and random mating.

Design of Simulation Studies

Summarized in Table 2 are scenarios we considered in our simulation studies. In each scenario, we first generated $N = 10,000$ maternal diplotypes, $h_i^m h_j^m$, and $N = 10,000$ paternal diplotypes under HWE based on random sampling using prespecified haplotype frequencies. When the simulation was conducted based on one single SNP, h_i^m and h_j^m refer to single alleles, which were randomly generated according to their respective frequencies. Each child's diplotype was generated by randomly sampling one maternal and one paternal haplotype. The phenotype Y was then generated based on model (1). We then randomly sampled $n_0 = 200$ control triads ($Y = 0$) and $n_1 = 200$ case triads ($Y = 1$).

We considered three likelihood methods: the likelihood method based on the mother-child pair genotype data for the test locus, as described above in Subsection “Likelihood Function for the Case-Control Mother-Child Pair Data at the Test Locus” ($P\text{-SNP}$); the likelihood method based on multi-locus mother-child pair genotype data, as described above in Section “Maximum Likelihood Method that Exploits Genotype Data of Nearby Markers via Haplotypes” ($P\text{-HAP}$); and the likelihood method based on the mother-father-child triad genotype data for the test locus ($T\text{-SNP}$). We use the $T\text{-SNP}$ method, which exploits both maternal and paternal genotype data, as a benchmark for the extent of additional power that $P\text{-HAP}$ can potentially achieve, which we believe can best help understand the merit of case-control mother-child pair design for assessing parent-of-origin effects. We note that $T\text{-SNP}$ uses genotype data at the test locus not only from case triads, but also from control triads as mentioned in Subsection “Likelihood Function for the Case-Control Mother-Child Pair Data at the Test Locus.” We retained genotype data for the test locus from all $n_0 + n_1$ pairs in all three methods. In $P\text{-HAP}$, we additionally retained the mother's and the child's nearby marker genotype data from the $n_0 + n_1$ pairs. In $T\text{-SNP}$, instead of the nearby marker genotype data, the paternal genotype data of the test locus were retained. Any other genetic information in the cohort was removed before the analyses. We used a likelihood ratio test to assess the null hypotheses of no parent-of-origin effect. In all scenarios, we adopted additive coding for both g^m and g^c , with respective OR parameters $\exp \beta_1 = 1.8$ and $\exp \beta_2 = 1.5$. The OR for the parent-of-origin effect $\exp \beta_3$ was set to be 1.0, 1.2, 1.5, or 2.0. We kept the phenotype prevalence in the range of 6~13% in all simulations by varying the intercept parameter β_0 . We used significance level 0.05 for testing parent-of-origin effects.

Performance of SNP-Based Methods $P\text{-SNP}$ and $T\text{-SNP}$

We first evaluated the performance of SNP-based tests, $P\text{-SNP}$ and $T\text{-SNP}$, as a function of the MAF, p_a . The nominal type I error rate 0.05 was well maintained in all the tests (data not shown). Figure 1 illustrates the power of testing a parent-of-origin effect of magnitude $\beta_3 = \log 1.5$ in model (1). Results for ORs 1.2 and 2.0 were largely similar (data not shown).

The power of the $P\text{-SNP}$ method appeared to be a concave function of the MAF in the range of (0, 0.5) with a peak at around $p_a = 0.20$. Incorporating paternal information ($T\text{-SNP}$) generally improved the power, and the improvement became more pronounced when the MAF p_a was

large. For example, for testing $\beta_3 = 0$ (when the truth was $\log 1.5$) at sample size $n_0 = n_1 = 200$, the test power increased by 0.02 (from 0.32 to 0.34) at $p_a = 0.05$; while at $p_a = 0.45$, the improvement became 0.27 (from 0.34 to 0.61). This can be explained as follows. Intuitively, parental origins of children's alleles are unidentifiable in the source population for the proportion $p_a(1 - p_a)$ of mother-child pairs (Table 1), but parental source ambiguity only happens to the proportion $2p_a^2(1 - p_a)^2$ of mother-father-child triads. Therefore, one can expect the gain in power due to the improved resolution of parental origins to be monotone in the range of $p_a \in (0, 0.5)$. Furthermore, the case-control mother-father-child triad data benefits from having more subjects than does the mother-child pair data, which helps improve the precision of estimating p_a .

Performance of $P_{\text{-HAP}}$ Based on Simulated Multi-Locus Genotype Data in Region GPX1

We investigated the improvement in power that can be achieved by incorporating nearby markers. We generated genotype data for five loci in the genomic region GPX1 based on published haplotype data [Chen et al., 2004], where seven haplotypes were reconstructed from the genotype data for five tagging SNPs. The configurations and frequencies of these haplotypes, the MAF of each SNP, and the LD structure of these five SNPs are shown in Supplementary Table S1. We designated each of the five SNPs in turn as the causative SNP and incorporated in $P_{\text{-HAP}}$ the genotype data for the remaining four marker loci. To evaluate the power improvement when various numbers of markers are incorporated, we also computed the power of $P_{\text{-HAP}}$ for testing the fifth locus S_5 , when genotype data for only the one, two, or three nearest marker SNPs were incorporated.

Results are summarized in Figure 2. The observed type I error rates were close to the nominal level 0.05 for all tests considered (data not shown). It is evident that incorporating genotype data for adjacent markers ($P_{\text{-HAP}}$) greatly improved the power for testing parent-of-origin effects, and the improvement was often comparable to that achieved by using only the one SNP but incorporating paternal genotype data ($T_{\text{-SNP}}$).

We observed that the magnitude of the power increase due to the incorporation of adjacent markers was influenced by both the MAF and the LD structure. For example, the power for testing locus S_1 increased by a factor of 1.36 when $\beta_3 = \log 1.2$, by 1.34 when $\beta_3 = \log 1.5$, and by 1.09 when $\beta_3 = \log 2.0$ (Panel A in Fig. 2). Unlike other SNPs, these gains were consistently slightly higher than those produced by incorporating no marker data but paternal genotype data. This might not only be because its MAF ($p_a = 0.18$) was closer to the information peak of mother-child pair data (Fig. 1), but also because it was in very strong LD with other loci (Supplementary Table S1). As another example, loci S_2 and S_4 had similar MAFs ($p_a = 0.15$), and the power was similar for testing their parent-of-origin effects with only the respective genotype data for S_2 and S_4 ($P_{\text{-SNP}}$). However, the power of $P_{\text{-HAP}}$ for testing S_2 was lower than that for S_4 (0.19 vs. 0.29 when $\beta_3 = \log 1.2$ and 0.19 vs. 0.24 when $\beta_3 = \log 1.5$). A closer look at the LD structure reveals that the LD between S_4 and its adjacent markers was stronger than that of S_2 . Simulation studies performed on locus S_5 where different numbers of markers were incorporated provided further support for the above observations. The power of $P_{\text{-SNP}}$ for S_5 was lower than for the other four SNPs since its MAF was large (Fig. 1), and incorporating several markers even in relatively weak LD led

to meaningful power improvement. Our results showed that the test power generally increased (Panel B in Fig. 2) with increasing numbers of linked markers.

Finally, our proposed maximum likelihood methods provided unbiased estimates of all model parameters. Supplementary Table S2 includes estimates of association parameters in model (1) with $n_1 = 200$ case pairs and $n_0 = 200$ control pairs. The true values for the parent-of-origin effect parameters are provided in the table, and we used multiplicative genetic models for both g^m and g^c with respective OR parameters 1.8 and 1.5. All of the averaged estimates appeared to be close to the true values. The estimation efficiency gained by incorporating paternal genotype data and by incorporating adjacent marker genotype data was in line with the power gain of the corresponding likelihood ratio tests discussed above. Our proposed method's estimated standard errors were close to the empirical standard errors.

Investigating the Influence of LD on the Power Gain of P_{HAP}

In simulation scenario C (Table 2), we further investigated the power gain of P_{HAP} as a function of LD between markers and the test locus. We considered a test locus with alleles A/a and a marker locus with alleles B/b , where the LD between the two loci was quantified by parameters D , r^2 , and D' [Foulkes, 2009]. We followed the simulation scheme described in Subsection “Design of Simulation Studies” with MAFs p_a and p_b fixed at 0.2 and β_3 at log 1.5. Results are summarized in Figure 3, where the power of the three methods for testing the parent-of-origin effect $\beta_3 = 0$ is displayed as a function of D , r^2 , and D' using the simulated bilocus data. P_{HAP} increased the power compared to P_{SNP} by up to a factor of 1.23. It is not surprising that when the two SNPs are in linkage equilibrium, i.e., $D = r^2 = D' = 0$, P_{HAP} and P_{SNP} had nearly identical power. This also held true when $r^2 = 1$ since the two SNPs became essentially identical. Therefore, in this two-SNP scenario, incorporating a marker SNP in moderate LD appeared to lead to higher power improvement than incorporating one with very strong LD. These results, together with those based on the GPX1 example where significant power gain was achieved by incorporating adjacent markers, suggest that the power improvement of P_{HAP} is a complex function of haplotype structures, haplotype frequencies, and MAFs.

Robustness to Violation of HWE

We investigated the robustness and bias of our likelihood methods when faced with violation of the assumption of HWE. To impose a violation, the parental genotype data were generated from the following distribution:

$$P(AA) = (1 - p_a)^2 + D_a, \quad p(Aa) = 2p_a(1 - p_a) - 2D_a, \quad \text{and} \quad p(aa) = p_a^2 + D_a,$$

where D_a characterizes the departure from HWE [Hernandez and Weir, 1989], with a positive D_a measuring an excess of homozygotes and a negative D_a measuring a deficiency of homozygotes. Further, let $\rho = D_a I(D_a < 0) / p_a^2 + D_a I(D_a > 0) / p_a(1 - p_a)$ be the inbreeding coefficient [Weir, 1996]. We fixed p_a at value 0.3 and changed the value and directions of D_a . The simulation setup corresponds to scenario D in Table 2. As expected, the bias in the estimation of the parent-of-origin effect parameter β_3 became greater when ρ became larger

(Fig. 4, upper panel). The bias was slightly more severe for larger MAFs (data not shown). The proposed likelihood ratio test for testing parent-of-origin effects had inflated type I error rates when $D_a < 0$ and deflated type I error when $D_a > 0$ (Fig. 4, lower panel). We note that the bias in method $T\text{-}SNP$ could be corrected by introducing parameters such as parental mating type [Weinberg and Shi, 2009]. Methods $P\text{-}SNP$ and $P\text{-}HAP$ may also be correctable by modifying the likelihood to incorporate parameters D_a or ρ . Similar extensions have been widely applied to haplotype-phenotype association studies with multi-locus genotype data [See e.g., Lin and Zeng 2006; Satten and Epstein, 2004].

Analyzing Parent-of-Origin Effects of the Gene PPARGC1A on Low Birth Weight

The JPS assembled a historical birth cohort with archival survey and medical record data of prenatal and perinatal characteristics for 17,003 offspring and their parents. The genotype data for around 1,500 SNPs in multiple candidate gene regions are available for 1,251 mother-child pairs, which were selected based on children's birth weight and mothers' prepregnancy body mass index (BMI). We applied $P\text{-}SNP$ and $P\text{-}HAP$ to examine parent-of-origin effects of nine loci in or near the gene PPARGC1A on intrauterine growth, as reflected by birth weight. We dichotomized birth weight using 2.5 kg as the cutoff point for low birth weight and focused on the subcohort of children whose mothers' prepregnancy BMI was less than 25. With this binary response variable, our analysis subcohort had $N_1 = 297$ underweight (< 2.5 kg) children at birth, among whom 117 pairs were genotyped, and $N_0 = 7,941$ children with normal birth weight, among whom 636 mother-child pairs had genotype data. We assume the genotype data were randomly available for normal-weight and for underweight deliveries in the cohort. The pairwise D' and r^2 of these nine SNPs are relatively large (Supplementary Fig. S1), indicating strong LD within this genomic region.

Our initial analysis of the data using $P\text{-}SNP$ yielded suggestive evidence that SNPs rs3736265 and rs8192678 had parent-of-origin effects on birth weight (p -values were 0.042 and 0.081, respectively), and that the main fetal effects of the two alleles appeared to be log-additive. Therefore, we focused on results for these two SNPs and assumed an additive association model (model (1) with $\gamma_3 = 0$). None of the nine maternal SNPs were alone significant in a standard logistic regression analysis. For illustrative purposes, we adopted an additive model for effects of maternal genotypes. To apply $P\text{-}HAP$, we first obtained unbiased initial estimates of the haplotype frequencies using the inverse probability weighting method [Horvitz and Thompson, 1952], i.e.,

$$\hat{\pi}_i = \left(N_0 \hat{\pi}_i^{(0)} / n_0 + N_1 \hat{\pi}_i^{(1)} / n_1 \right) / (N_0 / n_0 + N_1 / n_1),$$

where $\hat{\pi}_i^{(0)}$ and $\hat{\pi}_i^{(1)}$ are the respective haplotype frequency estimates in controls and in cases. Starting with these frequency estimates, final estimates were obtained by maximizing the proposed likelihood. Initial and maximum likelihood estimates of haplotype frequencies are summarized in Supplementary Table S3. None of the SNPs appeared to have a significant parent-of-origin effect on the risk of low birth weight after accounting for the multiplicity in testing nine SNPs. In line with the simulation results, incorporating the adjacent eight markers improved the estimation efficiency of the parent-of-origin effect in most instances.

Reported in Table 3 are the estimates of parameters in model (1) for SNPs rs3736265 and rs8192678, where the p -values were based on likelihood ratio tests. After adjusting for the maternal effect, the estimated parent-of-origin effect parameter β_3 for SNP rs3736265 was 0.920 (95% CI: [-0.015, 1.855]) with P_{-SNP} and was 0.736 (95% CI: [-0.156, 1.628]) with P_{-HAP} (likelihood ratio test P -values were 0.042 and 0.032, respectively). The estimate for SNP rs8192678 was 0.477 (95% CI: [-0.029, 0.983]) with P_{-SNP} and was 0.302 (95% CI: [-0.123, 0.727]) with P_{-HAP} (p -values were 0.081 and 0.050, respectively). These nonsignificant findings in our analyses could be due to the relatively small number of case subjects.

Examining separate effects of maternal and paternal alleles revealed an interesting phenomenon: the paternal copy tended to prevent the fetus from having low birth weight. For example, with P_{-HAP} , the log OR for the paternal copy was -1.540 (95% CI: [-2.897, -0.182]) for SNP rs3736265 and was -0.641 (95% CI: [-1.129, -0.153]) for SNP rs8192678. The corresponding log ORs for the maternal copy were -0.068 (95% CI: [-1.197, 1.061]) for SNP rs3736265 and -0.038 (95% CI: [-0.661, 0.586]) for SNP rs8192678. The paternal copy generally showed a stronger (significantly protective) effect against low birth weight than did the maternal copy. These data thus provide suggestive evidence for the “Mother-Fetus Competition” conjecture advanced by Haig [1993].

Discussion

Imprinted genes are of long interest as causal factors related to fetal development and early-life disorders. Pregnancy is a unique state in that there are two distinct but correlated genomes that can jointly influence fetal-maternal health. Recent studies suggest that imprinted fetal genes can influence maternal health during pregnancy. For example, several mutations in STOX1 were found in a recent study to co-segregate with preeclampsia, but the mother was at increased risk only when the fetus inherited the loss-of-function STOX1 variant from the mother [van Dijk et al., 2005]. It is thus of interest to consider parent of origin when analyzing effects of maternal and fetal genes on both maternal and fetal health.

In this work, we used maximum likelihood methods to study parent-of-origin effects of children's genes, where the phenotype of interest may concern either the mother or the child. The idea of including parent of origin effects in a likelihood framework was proposed by Weinberg et al. [1998] and further discussed by Ainsworth et al. [2011] and Gjessing and Lie [2006]. Our method was designed for the case-control mother-child pair design, and our simulation results supported the efficiency of this study design: under the required assumptions, one can investigate essentially all genetic effects that can be studied with family-based study designs. Nevertheless, investigators should also consider collecting DNA from fathers when possible, as triad designs offer additional opportunities to identify the parent of origin for both alleles and haplotypes.

We have developed an R (<http://www.r-project.org/>) program for the estimation of the parameters. With data similar to those generated in the simulation studies described in Section “Simulation Studies” (i.e., five markers, seven potential haplotypes, with 200 case pairs, and 200 control pairs), a Windows 7 PC with Intel i7 core CPUs required about 420 sec run time. The program can be requested from the first author. Our method assumes

HWE, random mating, and Mendelian inheritance, and is biased when the data do not meet these requirements. However, the assumption of HWE can be relaxed by introducing a fixation parameter as discussed above. We have also assumed that outcome status is known for all individuals in the cohort from which cases and controls are sampled. In hospital-based studies, this requirement may not be realistic. In the absence of known cohort totals, simulations (data not shown) suggested that our likelihood method performs well when the analyst has a reasonable phenotype prevalence estimate.

Currently, three GWAS imputation software packages, BEAGLE, SHAPE-IT, and Marchini et al. [2006], are capable of efficiently exploiting family relationships in haplotype phasing [Browning and Browning, 2011]. Simulation studies showed that the average rates of incorrect assignment of parental origin were very low with case-parent trio and/or parent-offspring duo data [Browning and Browning, 2009; Delaneau et al., 2012]. These methods were designed for GWAS data and not expected to be as accurate with a smaller number of markers.

Many case-control mother-child pair studies, such as the JPS study that we analyzed, are candidate-gene based, partly due to financial constraints for genotyping two genomes. GWASs of case-control mother-child pairs may work best if they only fully genotype one genome and then only some candidate genomic regions in the other genome [e.g., those that harbor promising signals]. For these less rich data, it remains an open question how well any of the existing software packages could incorporate family relationships in haplotype phasing/parent-of-origin inference. Our proposed method jointly assesses haplotype phasing and parent-of-origin effects, thereby efficiently exploiting family-relationship information. We assumed the loci under study are sufficiently close that there was no mother-to-child recombination between the test SNP and the nearby markers. If recombinations do occur, the multilocus method should still retain type I error rates, but power may be reduced. In a GWAS context with case-parent data, the phasing approaches that have been developed for dense SNPs could also be exploited to produce improvements in power for studying parent-of-origin effects.

Haplotype structure can also be exploited in a log-linear framework [e.g. Shi et al., 2007]. In fact, with case-control mother-child pair data and known cohort total numbers of cases and controls, one may employ a log linear model in constructing likelihoods (2) and (3) by using a log link. Parameters in the two base models have different interpretations (ORs vs. relative risks), and the use of the logarithm instead of the logit can introduce annoying problems, e.g., fitted probabilities that exceed one. However, under co-dominant coding of the maternal main effects (which is advisable to ensure validity of tests of imprinting), the two models should be close, and so are the corresponding likelihood ratio tests for imprinting effect, as we confirmed in unreported simulation studies. Therefore, the results we have shown related to power and estimation would also apply to a log-linear formulation. When the population prevalence is small (rare disease), the two models also yield parameter interpretations that coincide.

In a GWAS setting, one can never be sure that one has been lucky enough to have typed a causative SNP rather than a marker or that the OR (or parent-of-origin parameter) one is

estimating is not markedly attenuated compared to that for the nearby actual causative SNP. However, if in fact the SNP under study is a marker (and the investigator will not know that) then the parent-of-origin null hypothesis for the causative SNP implies the parent-of-origin null hypothesis for the nearby marker, so testing for parent-of-origin effects is still valid. Our method appeared to behave well for testing a noncausal marker that is close to the causal SNP. Despite decreased power, the type I error rates were close to the nominal level (data not shown).

The JPS example underscores the issue that ambiguity will remain about which of the SNPs that display apparent parent-of-origin effects is the real player vs. merely an informative marker. The problem of which (if any) is the “real” SNP, or whether the effect might actually be due to a haplotype but not a single SNP, arises in this context, just as it does in the context of main effects of SNPs in GWAS. The methods described in this paper will help the analyst identify linkage blocks worthy of future more detailed study via targeted sequencing. Only such careful follow-up analyses can definitively identify causative SNPs or causative haplotypes for parent-of-origin effects.

Our proposed methods can be extended to analyze combined data collected from family-based and population-based study designs. Retrospective likelihood based methods have been proposed for assessing associations with genotypes and haplotypes [e.g., Dudbridge, 2008], but it remains to evaluate the power increase for testing parent-of-origin effects through the incorporation of marker genotype data. For example, Berends et al. [2007] conducted both case-control and family studies to confirm the possible parent-of-origin effect of the gene STOX1 on preeclampsia. Using triads with heterozygous mothers only, they found that the fetal STOX1 gene appeared not to be related to the risk of preeclampsia. However, they only tested the Mendelian transmission of maternal genotypes, but not the parent-of-origin effect of the fetal gene. For the latter, homozygous mothers can be very informative in inferring parent-of-origin effects of children's alleles.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Drs. Jinbo Chen and Dongyu Lin were supported by NIH grant ES016626. Dr. Rui Feng was supported by NIH grant GM088566. Dr. Hagit Hochner was supported by NIH grant HL091244. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences, under project number ZO1 ES040007. We would also like to thank Dr. Min Shi at the Biostatistics Branch, National Institute of Environmental Health Sciences for her very helpful comments.

References

- Ainsworth HF, Unwin J, Jamison DL, Cordell HJ. Investigation of maternal effects, maternal-foetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet Epidemiol.* 2011; 35(1):19–45. [PubMed: 21181895]
- Berends AL, Bertoli-Avella AM, de Groot CJM, van Duijn CM, Oostra BA, Steegers EAP. STOX1 gene in pre-eclampsia and intrauterine growth restriction. *BJOG: Int J Obstet Gy.* 2007; 114(9): 1163–1167.

- Browning SR, Browning BL. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet.* 2009; 84:210–223. [PubMed: 19200528]
- Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev.* 2011; 12:703–714.
- Chen J, Chatterjee N. Haplotype-based association analysis in cohort and nested case-control studies. *Biometrics.* 2006; 62:28–35. [PubMed: 16542226]
- 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. [PubMed: 15604561]
- Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2012; 9:179–181. [PubMed: 22138821]
- Dudbridge F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered.* 2008; 66:87–98. [PubMed: 18382088]
- Eichler EE, Flint F, Gibson G, Kong A, Leal SM, Moore J, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* 2010; 11(6):446–450. [PubMed: 20479774]
- Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet.* 2003; 73:1316–1329. [PubMed: 14631556]
- Excoffier L, Slatkin GA. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol.* 1995; 12:921–927. [PubMed: 7476138]
- Falls JG, Pulford DJ, Wylie AA, Jirtle RL. Genomic imprinting: implications for human disease. *Am J Pathol.* 1999; 154(3):635–647. [PubMed: 10079240]
- Foulkes, AS. *Applied statistical genetics with R: for population-based association studies (use R).* Springer-Verlag; New York: 2009.
- Gjessing HK, Lie R. Case-parent triads: estimating single- and double-dose effects of fetal and maternal disease gene haplotypes. *Ann Hum Genet.* 2006; 70(3):382–396. [PubMed: 16674560]
- Haig D. Genetic conflicts in human pregnancy. *Q Rev Biol.* 1993; 68:495–532. [PubMed: 8115596]
- Hernandez JL, Weir BS. A disequilibrium approach to Hardy-Weinberg testing. *Biometrics.* 1989; 45:53–70. [PubMed: 2720060]
- Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc.* 1952; 47:663–685.
- Kanayama N, Takahashi K, Matsuura T, Sugimura M, Kobayashi T, Moniwa N, Tomita M, Nakayama K. Deficiency in p57Kip2 expression induces preeclampsia-like symptoms in mice. *Mol Hum Reprod.* 2003; 8:1129–1135. [PubMed: 12468647]
- Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Jonasdottir A, Sigurdsson A, Kristinsson KT, Jonasdottir A. Parental origin of sequence variants associated with complex diseases. *Nature.* 2009; 462(7275):868–874. others. [PubMed: 20016592]
- Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *J Am Stat Assoc.* 2006; 101:89–118.
- Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR. A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet.* 2006; 78:437–450. others. [PubMed: 16465620]
- Petry CJ, Ong KK, Dunger DB. Does the fetal genotype affect maternal physiology during pregnancy? *Trends Mol Med.* 2007; 13:414–421. [PubMed: 17900986]
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science.* 1996; 273:1516–1517. [PubMed: 8801636]
- Saftlas AF, Beydoun H, Triche E. Immunogenetic determinants of preeclampsia and related pregnancy disorders: a systematic review. *Obstet Gynecol.* 2005; 106:162–172. [PubMed: 15994633]
- Satten GA, Epstein MP. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol.* 2004; 27(3):192–201. [PubMed: 15372619]
- Shi M, Umbach DM, Weinberg CR. Identification of risk-related haplotypes using multiple SNPs from nuclear families. *Am J Hum Genet.* 2007; 81(1):53–66. [PubMed: 17564963]

- van Dijk M, Mulders J, Poutsma A, Könst AA, Lachmeijer AM, Dekker GA, Blankenstein MA, Oudejans CB. Maternal segregation of the Dutch preeclampsia locus at 10q22 with a new member of the winged helix gene family. *Nat Genet.* 2005; 37(5):514–519. [PubMed: 15806103]
- Wangler MF, Chang AS, Moley KH, Feinberg AP, Debaun MR. Factors associated with preterm delivery in mothers of children with Beckwith-Wiedemann syndrome: a case cohort study from the BWS registry. *Am J Med Genet.* 2005; 134:187–191. [PubMed: 15723285]
- Weir, BS. *Genetic data analysis II*. Sinauer; Sunderland, MA: 1996.
- Weinberg CR. Methods for detecting parent-of-origin effects in genetic studies of case-parent triads. *Am J Hum Genet.* 1999; 65:229–235. [PubMed: 10364536]
- Weinberg CR, Shi M. The genetics of preterm birth: using what we know to design better association studies. *Am J Epidemiol.* 2009; 170(11):1373–1381. [PubMed: 19854804]
- Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet.* 2005; 77:627–636. [PubMed: 16175508]
- Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent triad data: assessing effects of disease genes that act directly or through maternal effects, and may be subject to parental imprinting. *Am J Hum Genet.* 1998; 62(4):969–978. [PubMed: 9529360]

Figure 1.

The power of p-snp and t-snp for testing parent-of-origin effects. The null hypothesis is $\beta_3 = 0$, and the data were generated at $\beta_3 = \log 1.5$. Each legend indicates the method and the number of cases, which was equal to the number of controls.

Figure 2.

The power of p-snp, t-snp, and p-hap based on simulated genotype data for five diallelic loci in gene GPX1. Panel A: Each SNP was treated as the causal locus in turn. Each number beneath the x -axis is the MAF corresponding to the SNP above. Different values for the parent-of-origin effect parameter β_3 (0, log 1.2, log 1.5, and log 2.0) were considered and are provided on the right hand side of the figure. Method p-hap incorporated genotype data of all the other four SNPs. Panel B: The power of p-hap for testing SNP 5 when 1, 2, 3, or 4 adjacent markers were incorporated and SNP 5 is causal, and where the x -axis is the number of included nearby markers plus 1.

Figure 3.

The power increase of p-hap compared with p-snp and t-snp as a function of LD measured by D , r^2 , and D' , where genotype data for one marker were incorporated in p-hap. The MAFs of both SNPs were 0.2.

Figure 4. The bias and type I error rates of p-snp and t-snp when the distribution of the SNP genotype deviates from the Hardy-Weinberg equilibrium. The x -axis is the linkage disequilibrium parameter D_a and the inbreeding coefficient ρ .

Table 1

Joint distribution of mother-child genotypes and log ORs of offspring sourced genotype according to Model (1)

Observed		Inferred (g_m^c, g_p^c)	Probability $p(g^m, g_m^c, g_p^c)$	log OR of (g_m^c, g_p^c)
g^m	g^c			
AA *	AA	(A, A)	$(1 - p_a)^3$	0
	Aa	(A, a)	$p_a (1 - p_a)^2$	$\beta_2 - \beta_3$
Aa	AA	(A, A)	$p_a (1 - p_a)^2$	0
	Aa	(A,a)	$p_a^2 (1 - p_a)$	$\beta_2 - \beta_3$
		(a, A)	$p_a (1 - p_a)^2$	$\beta_2 + \beta_3$
	aa	(a, a)	$p_a^2 (1 - p_a)$	$2 \beta_2 + \gamma_3$
aa	Aa	(a, A)	$p_a^2 (1 - p_a)$	$\beta_2 + \beta_3$
	aa	(a, a)	p_a^3	$2 \beta_2 + \gamma_3$

* AA, Aa,and aa are coded as “0”, “1”, and “2”, respectively.

Table 2

Parameter settings used in simulation studies

Scenario	Range	Parameters					Assumptions	Number of replicates	Discussion section
		β_0	$\exp \beta_1$	$\exp \beta_2$	$\exp \beta_3$	p_a			
A	Single SNP	-4.0~-3.0 *	1.8	1.5	1.0, 1.2, 1.5, 2.0	0.05~0.45	HWE + RM **	5,000	Subsection "Performance of SNP-Based Methods <i>p-snp</i> and <i>t-snp</i> "
B	GPX1 ***	-3.0	1.8	1.5	1.0, 1.2, 1.5, 2.0	See Supplementary Table S1	HWE + RM	5,000/1,000 ****	Subsection "Performance of <i>p-hap</i> Based on Simulated Multi-Locus Genotype Data in Region GPX1"
C	Two markers	-3.0	1.8	1.5	1.5	0.20, 0.30	HWE + RM	1,000	Subsection "Investigating the Influence of LD on the Power Gain of <i>p-hap</i> "
D	Single SNP	-4.0~-3.0	1.8	1.5	1.0	0.10~0.40	RM	10,000	Subsection "Robustness to Violation of HWE"

* We varied β_0 so that the population prevalence of the phenotype was in the range of 6~7%.

** HWE: Hardy-Weinberg Equilibrium in the parental population. RM: Random Mating.

*** A five-locus genomic region based on published haplotype data, with seven reconstructed haplotypes from the genotype data (See Subsection "Performance of *p-hap* Based on Simulated Multi-Locus Genotype Data in Region GPX1").

**** 5,000 for estimating type I error rates and 1,000 for power.

Table 3

Estimated parent-of-origin effects of two SNPs in genomic region PPARGC1A on birth weight in the JPS

SNP	$f(g^m)$	Method	β_1		β_2		β_3		p -value LRT ***
			EST *	SB **	EST	SE	EST	SE	
rs3736265	g^m ****	<i>p-snp</i>	-0.498	0.437	-0.743	0.460	0.920	0.477	0.042
		<i>p-hap</i>	-0.263	0.406	-0.804	0.446	0.736	0.455	0.032
rs8192678	g^m	<i>p-snp</i>	-0.142	0.236	-0.257	0.186	0.477	0.258	0.081
		<i>p-hap</i>	0.024	0.218	-0.339	0.186	0.302	0.217	0.050

* MLE of the parameter.

** Estimated standard error of the MLE estimate.

*** P -value of testing $H_0: \beta_3 = 0$ using likelihood ratio test.**** The fitted model was $\text{logit}p(Y=1|g^m, g_m^c, g_p^c) = \beta_0 + \beta_1 g^m + \beta_2 g^c + \beta_3 \{I(g_m^c=a) - I(g_p^c=a)\}$.