Journal of Human Genetics (2014) 59, 541–548
© 2014 The Japan Society of Human Genetics All rights reserved 1434-5161/14
www.nature.com/jhg

ORIGINAL ARTICLE

Nonparametric method for detecting imprinting effect using all members of general pedigrees with missing data

Fangyuan Zhang and Shili Lin

Imprinting effects can lead to parent-of-origin patterns in complex human diseases. For a diallelic marker locus, Pedigree Parental-Asymmetry Test (PPAT) and its extension MCPPAT using pedigrees allowing for missing genotypes are simple and powerful for detecting imprinting effects. However, these approaches only take affected offspring into consideration, thus not making full use of the data available. In this paper, we propose Monte Carlo Pedigree Parental-Asymmetry Test using both affected and unaffected (MCPPATu) offsprings, which allows for missing genotypes through Monte Carlo sampling. Simulation studies demonstrate that MCPPATu controls the empirical type I error rate well under the null hypotheses of no parent-of-origin effects. It is also demonstrated that the use of additional information from unaffected offspring and partially observed genotypes in the analysis can greatly improve the statistical power. Indeed, for common diseases, MCPPATu is much more powerful than MCPPAT when all genotypes are observed and the power improvement is even greater when there is missing data. For rarer diseases, there are still substantial power gains with the inclusion of unaffected offspring, although the gains are less impressive compared with those for more common diseases.

Journal of Human Genetics (2014) **59,** 541–548; doi:10.1038/jhg.2014.67; published online 14 August 2014

INTRODUCTION

Genomic imprinting is an epigenetic factor that modulates the effects of genetic variants. It can lead to parent-of-origin patterns in gene expressions, and hence has been increasingly explored for its crucial role in the etiology of complex diseases.[1] More specifically, genomic imprinting is an effect of the epigenetic process involving methylation and histone modifications to silence the expression of a gene inherited from a particular parent (mother or father) without altering the genetic sequence. This process leads to unequal expression of a heterozygous genotype depending on whether the imprinted variant is inherited from the mother (maternal imprinting) or from the father (paternal imprinting), which has a key role in normal mammalian growth and development. Hence, genomic imprinting is hailed as a key factor in understanding the interplay between the epigenome and the genome.[2]

In addition to its involvement in growth and development, genomic imprinting also has an important role in a number of complex human diseases. Beckwith–Wiedemann Syndrome, Silver–Russell Syndrome, Angelman Syndrome (AS) and Prader–Willi Syndrome (PWS) are most well-known examples.[3–5] It is fascinating to note that AS and PWS are caused by the same genetic locus, albeit that one is due to the gene being maternally, and the other being paternally, imprinted. In fact, about 1% of all mammalian genes are estimated to be imprinted,[6] and thus it is expected that imprinting may have a role in many other complex human diseases, such as some cancers and type 2 diabetes.[1] However, there are still very few that have been identified thus far, partly because of insufficient amount of data and/or lack of sufficient power in existing statistical tests.

With the availability of the next-generation sequencing technology, scientists are now able to carry out direct studies of imprinting genomewide in the mouse efficiently.[7,8] Nevertheless, the controlled mating setup that was successful in mouse studies is not feasible in humans. Hence, powerful statistical methods for detecting and assessing imprinting effects on complex genetic traits are still indispensable.

Numerous statistical methods have been proposed to detect imprinting effect.[9–12] For a diallelic genetic marker locus, traditional tests like parental-asymmetry test (PAT) based on affected child–parent trio data[11] is simple and powerful. A series of generalizations of PAT widen its capability. CPAT was developed to extend PAT to nuclear families with an arbitrary number of affected children;[12] MCPPAT further extends the capability of PAT to use information on extended families that may have missing genotypes on some of the individuals;[13] a more recent extension, PATu, is for nuclear families taking unaffected children into account as well.[14] These methods are valid for testing for imprinting under the assumption of no maternal genotype effect, another kind of parent-of-origin effect.

Another set of tests that have been proposed for detecting imprinting effect also considers testing for maternal genotype effect, a simultaneous testing strategy.[15–18] In both papers by Yang and Lin,[17,18]

Department of Statistics, The Ohio State University, Columbus, OH, USA
Correspondence: Professor S Lin, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA.
E-mail: shili@stat.osu.edu

Received 27 March 2014; revised 5 June 2014; accepted 26 June 2014; published online 14 August 2014

unaffected children within the same family as the affected ones are also used in the analysis, which leads to substantial power increase in simulation studies, a fact motivated the work of Zhou *et al.*[14] It has been shown that simultaneous detection methods are robust to the presence of maternal genotype effect compared with PAT-type tests. However, as expected, they may be less powerful if there is *a priori* knowledge of no maternal effect.

The flurry of research in this area recently as summarized above all aims at increasing statistical power for detecting parents-of-origin effects by making fuller use of available data or by expanding the study design to include new data types. In this paper, we propose the Monte Carlo Pedigree Parental-Asymmetry Test using both affected and unaffected (MCPPATu) offspring from (extended) families of arbitrary sizes and structure to test for imprinting effect. When there is no missing data, the test is referred to as PPATu, which may be regarded as a generalization of PATu by using data from all individuals in a pedigree rather than just a nuclear family by ignoring other available information. A limited simulation study on PPATu was carried out previously.[19] On the other hand, when there is missing data, MCPPATu is a generalization of MCPPAT by also using unaffected offspring in addition to the affected ones. When the MC option is turned on to maximize its capability, MCPPATu uses all the observed data in a pedigree fully. Results from an extensive simulation study not only substantiate the validity of the proposed test with well controlled type I error but also show power gains with MCPPATu when compared with PPATu or MCPPAT for common diseases as well as for diseases that are relatively more rare.

MATERIALS AND METHODS

Notation

Suppose a disease susceptibility locus has two alleles, disease allele D and normal allele d. Denote the probability of an individual being affected with genotypes D/D, D/d, d/D and d/d by $\phi_{D/D}$, $\phi_{D/d}$, $\phi_{d/D}$ and $\phi_{d/d}$, respectively, where the allele before '/' is paternal and that after '/' is maternal. Consider a marker locus of interest with two alleles, M_1 and M_2. For convenience, let 0, 1 and 2 represent the marker genotypes M_2M_2, M_1M_2 and M_1M_1, respectively, which are basically counts of the number of M_1 allele, the allele of interest. We also use F, M and C to denote the marker genotypes of father, mother and child, respectively, which takes values in {0,1,2} depending on the individual's genotype. Throughout this study, we assume mating symmetry and no maternal effect. In the following, we describe the complete data version of the proposed test (PPATu) first before launching the full version with MC sampling to accommodate families with missing genotypes.

PPATu for complete pedigree data

Suppose that we have N pedigrees each with n_{1j} child–parents trios having an affected heterozygous child and n_{0j} trios having an unaffected heterozygous child, $j=1,\ldots,N$. Let $Q_{1j}=\sum_{i=1}^{n_{1j}} S_{1ji}$, where

$$\begin{aligned} S_{1ji} =& I\{F_{1ji} > M_{1ji}, C=1 \mid \text{child affected}\} \\ & - I\{F_{1ji} < M_{1ji}, C=1 \mid \text{child affected}\} \end{aligned}$$

$I\{\cdot\}$ is the indicator function taking value of 1 if the condition within the set of curly brackets is true and 0 otherwise, and i indexes a trio. Similarly, let $Q_{0j}=\sum_{i=1}^{n_{0j}} T_{0ji}$, where

$$\begin{aligned} T_{0ji} =& I\{F_{0ji} < M_{0ji}, C=1 \mid \text{child unaffected}\} \\ & - I\{F_{0ji} > M_{0ji}, C=1 \mid \text{child unaffected}\} \end{aligned}$$

In words, Q_{1j} looks for information pertaining to excess (deficiency) in transmission of the M_1 allele from the father (mother) to the affected child. In contrast, Q_{0j} looks for the opposite—deficiency (excess) in transmission of the M_1 allele from the father (mother) to the unaffected child. Then, the PPATu statistic for the whole data set is constructed as

$$\begin{aligned} T_{\text{PPATu}} &= \frac{\sum_{j=1}^{N}[(1-w)Q_{1j}+wQ_{0j}]}{\sqrt{\sum_{j=1}^{N}[(1-w)Q_{1j}+wQ_{0j}]^2}} \\ &= \frac{\sum_{j=1}^{N}\left[(1-w)\left(\sum_{i=1}^{n_{1j}} S_{1ji}\right)+w\left(\sum_{i=1}^{n_{0j}} T_{0ji}\right)\right]}{\sqrt{\sum_{j=1}^{N}\left[(1-w)\left(\sum_{i=1}^{n_{1j}} S_{1ji}\right)+w\left(\sum_{i=1}^{n_{0j}} T_{0ji}\right)\right]^2}}, \end{aligned} \tag{1}$$

which takes dependencies among the S's and the T's within the same pedigree into account, and also makes use of unaffected offspring. Further, w (a constant between 0 and 1) is a weight denoting the contribution from the unaffected offspring. When $w=0$, the PPATu statistic reduces to the PPAT statistic in Zhou *et al.*[13] In this article, we use $w=\rho$, the population prevalence of the disease, for the results presented in this paper; we also present results with other weights in the Supplementary Materials.

To study the asymptotic property of T_{PPATu}, we first consider the expectation of its numerator. In Table 1, the six informative trio genotype configurations for which the child is a heterozygous and the parents have different number of the M_1 allele are given in the first three columns. The joint probabilities of the genotypes and the child being affected, $\{s_1, s_2, \ldots, s_6\}$, are given in the next column. Their counter parts, the joint probabilities of the genotypes and the child being unaffected, $\{t_1, t_2, \ldots, t_6\}$, are given in the last column. Then,

$$\begin{aligned} & E\left(\sum_{i=1}^{n_{1j}} S_{1ji} + \sum_{i=1}^{n_{0j}} T_{0ji}\right) \\ &= n_{1j}(s_1+s_3+s_5-s_2-s_4-s_6)/\rho - n_{0j}(t_1+t_3+t_5-t_2-t_4-t_6)/(1-\rho), \end{aligned}$$

where ρ is the population prevalence of the disease. Under the null hypothesis of no-imprinting effect with the assumptions of mating symmetry and no maternal effect, we can see that $s_1=s_2$, $s_3=s_4$, $s_5=s_6$, $t_1=t_2$, $t_3=t_4$ and $t_5=t_6$. Let us consider s_1 and s_2 as an example. Under the assumption of mating symmetry, the probability of mating type $(M, F)=(1, 0)$ is the same as that of $(M, F)=(0, 1)$. Further, if there are no imprinting nor maternal effects, the probability that a child is affected only depends on his/her own genotype, but not those of the parents. Therefore, the penetrance probabilities are the same for phenotype configurations $(M, F, C)=(1,0,1)$ and $(M, F, C)=(0,1,1)$. This leads to $s_1=s_2$ as we can see from the detailed formulas provided in Table 1. Therefore, the expectation of the numerator is 0 under the null. To estimate the variance of the numerator under the null hypothesis, we note that $\text{Var}(Q_{1j}+Q_{0j})=E(Q_{1j}+Q_{0j})^2$ as a consequence of the expectation being 0 as shown above. Thus, $\sum_{j=1}^{N}(Q_{1j}+Q_{0j})^2$ provides an unbiased estimator of the variance of the numerator. By invoking the Central Limit Theorem, when the number of families N is sufficiently large, the standardized PPATu statistic in Equation (1) follows a standard normal distribution approximately.

To understand the potential gain in power from a theoretical perspective, we note that both Q_1 and Q_0 are of the same sign if there is indeed imprinting effect; that is, they are both positive (negative) if there is maternal (paternal) imprinting. Therefore, using information from unaffected offspring would make the test statistic to be farther away from zero under the alternative

Table 1 Six possible trio genotype configurations with heterozygous child genotype and their corresponding joint probabilities with child's affection status

F	M	C	$P\ (F, M, C, A=1)$	$P\ (F, M, C, A=0)$
1	0	1	$s_1=\mu_{10}\cdot\frac{1}{2}\cdot f_{101}$	$t_1=\mu_{10}\cdot\frac{1}{2}\cdot(1-f_{101})$
0	1	1	$s_2=\mu_{01}\cdot\frac{1}{2}\cdot f_{011}$	$t_2=\mu_{01}\cdot\frac{1}{2}\cdot(1-f_{011})$
2	0	1	$s_3=\mu_{20}\cdot 1\cdot f_{201}$	$t_3=\mu_{20}\cdot 1\cdot(1-f_{201})$
0	2	1	$s_4=\mu_{02}\cdot 1\cdot f_{021}$	$t_4=\mu_{02}\cdot 1\cdot(1-f_{021})$
2	1	1	$s_5=\mu_{21}\cdot\frac{1}{2}\cdot f_{211}$	$t_5=\mu_{21}\cdot\frac{1}{2}\cdot(1-f_{211})$
1	2	1	$s_6=\mu_{12}\cdot\frac{1}{2}\cdot f_{121}$	$t_6=\mu_{12}\cdot\frac{1}{2}\cdot(1-f_{121})$

M, F and C are the number of variant allele(s) carried by mother, father and child in a trio, which take values of 0, 1 or 2; the mating type probability for $(M, F)=(m,f)$ is denoted by μ_{mf}; $A=1$ ($A=0$) indicates that the child is affected (unaffected); f_{mfc} denotes the probability (penetrance) that a child is affected given the trio genotype configuration $(M, F, C)=(m,f,c)$.

hypothesis, and thereby can lead to an increase in power for detecting the effect if it indeed exists.

MCPPATu when there are missing genotypes

The PPATu as described above only uses trios that have complete genotype data. Hence, when there are missing genotypes in a trio, PPATu will discard it. To make full use of all data available, we propose the MCPPATu statistic that will include all trios in the analysis even when only partial genotypes are available. We first describe the statistic for a single pedigree. We define Q_1 and Q_0 as in PPATu, but with the subscript j suppressed for simpler notation without causing ambiguity. Further, we write Q_1 and Q_0 more fully as $Q_1(G_m, G_o, A)$ and $Q_0(G_m)$, respectively, to show explicitly that they not only depend on the observed genotypes (G_0) but also on the unobserved ones (G_m), and on the affection status (A) of all offspring. However, as Q_1 and Q_0 are no longer computable given the existence of missing data, we consider $Q_{1\mathrm{MC}}$ and $Q_{0\mathrm{MC}}$, the conditional expectations of Q_1 and Q_0 given the observed genotypes. That is,

$$Q_{1\mathrm{MC}} = E[Q_1 \,|\, G_o] = E[Q_1(G_m, G_o, A) \,|\, G_o],$$

$$Q_{0\mathrm{MC}} = E[Q_0 \,|\, G_o] = E[Q_0(G_m, G_o, A) \,|\, G_o].$$

Evaluation of the above expectations is usually not computationally feasible owing to the large number of summations over all sets of possible genotypes unless there are only a handful of individuals with missing genotypes. Hence, we propose to estimate $Q_{1\mathrm{MC}}$ and $Q_{0\mathrm{MC}}$ based on an MC simulation scheme following the work of Zhou *et al.*[13] Specifically, we draw independent samples G_{mk}, $k = 1, \ldots K$, from $P\ (G_m|G_o)$ using a peeling algorithm,[20] and the desired statistics are then estimated as

$$Q_{1\mathrm{MC}} \approx \frac{1}{K}\sum_{k=1}^{K} Q_1(G_{mk}, G_o, A) \quad \text{and}$$

$$Q_{0\mathrm{MC}} \approx \frac{1}{K}\sum_{k=1}^{K} Q_0(G_{mk}, G_o, A).$$

In practice, the number of simulated genotype sets needed to achieve good estimates depends on the degree of missingness. The more individuals missing in a pedigree, the larger the K should be. Further, K may be set to be different for different pedigrees if there is a wide range in the amount of missing data among the pedigrees. In our simulation study as well as in the real data analysis, we found that K up to 200 appears to work satisfactorily.

With data from N pedigrees, the MCPPATu statistic T_{MCPPATu} is formed analogous to T_{PPATu} as defined in Equation (1), but with Q_1 and Q_0 for each pedigree replaced by the corresponding $Q_{1\mathrm{MC}}$ and $Q_{0\mathrm{MC}}$. We can show (Supplementary Materials A.1) that the expectations of the $Q_{1\mathrm{MC}}$ and $Q_{0\mathrm{MC}}$ statistics are both zero under the null hypothesis of no imprinting effect. Therefore, T_{MCPPATu} also follows a standard normal distribution asymptotically, and can be used as a valid test statistic for imprinting.

RESULTS

Settings

An extensive simulation study was carried out to investigate the size and power of the proposed tests. As tests for imprinting effects are typically carried out in the presence of association, we assume no recombination between the disease susceptibility locus and the marker locus. The numbers in Table 2 shows 27 combinations of haplotype frequencies and penetrance probabilities of imprinting models. Specifically, we set nine combinations (referred to as settings) of three sets of haplotype frequencies and three sets of penetrances for homozygous genotypes, $\phi_{D/D}$ and $\phi_{d/d}$. For each combination, we further assign three imprinting effect models: no imprinting (NI; $\phi_{D/d} = \phi_{d/D}$), incomplete maternal imprinting (II; $\phi_{D/D} > \phi_{D/d} > \phi_{d/D} > \phi_{d/d}$) and complete maternal imprinting (CI; $\phi_{D/D} = \phi_{D/d} > \phi_{d/D} = \phi_{d/d}$). As there is complete symmetry between paternal and maternal imprinting using the proposed statistics, only models portraying maternal imprinting effect are considered without loss of generality. The prevalence is 29.9% for all 27 models (referred to as the 'A' models and the nine settings are referred to as the 'A' settings). Such a prevalence value represents the upper half of prevalence for common diseases and traits (Supplementary Table S1). To consider models that represent the lower end, we also consider another set of models whose penetrances are half of those shown in Table 2, leading to 9 'B' settings with 27 'B' models, all with a prevalence of 14.95%. We further consider 18 additional models ('C' models with 6 'C' settings) for which the prevalence is generally lower, ranging from 7% to 15% for all but four models (Supplementary Table S2). In all, we consider an extensive and thorough simulation study over a total of 72 scenarios. The 24 NI models (9 A's, 9 B's and 6 C's) are used to

Table 2 Simulation study scenarios through combinations of 3 sets of haplotype frequencies and 18 penetrance model

							Penetrance of imprinting model[b]				
	Hap frequencies										
							No	*Incomplete*		*Complete*	
Setting[a]	DM$_1$	dM$_1$	DM$_2$	dM$_2$	$\phi_{D/D}$	$\phi_{d/d}$	$\phi_{D/d}$	$\phi_{D/d}$	$\phi_{d/D}$	$\phi_{D/d}$	$\phi_{d/D}$
1A/1B	0.2	0.0	0.1	0.7	0.390	0.260	0.325	0.370	0.280	0.390	0.260
2A/2B	0.3	0.1	0.0	0.6	0.390	0.260	0.325	0.370	0.280	0.390	0.260
3A/3B	0.3	0.0	0.0	0.7	0.390	0.260	0.325	0.370	0.280	0.390	0.260
4A/4B	0.2	0.0	0.1	0.7	0.440	0.240	0.340	0.420	0.260	0.440	0.240
5A/5B	0.3	0.1	0.0	0.6	0.440	0.240	0.340	0.420	0.260	0.440	0.240
6A/6B	0.3	0.0	0.0	0.7	0.440	0.240	0.340	0.420	0.260	0.440	0.240
7A/7B	0.2	0.0	0.1	0.7	0.580	0.180	0.380	0.530	0.230	0.580	0.180
8A/8B	0.3	0.1	0.0	0.6	0.580	0.180	0.380	0.530	0.230	0.580	0.180
9A/9B	0.3	0.0	0.0	0.7	0.580	0.180	0.380	0.530	0.230	0.580	0.180

Abbreviations: CI, complete imprinting; II, incomplete maternal imprinting; NI, no imprinting.

[a] A penetrance model is specified by ($\phi_{D/D}$,$\phi_{D/d}$,$\phi_{d/D}$,$\phi_{d/d}$). When there is NI, $\phi_{D/d} = \phi_{d/D}$; when there is II, $\phi_{D/D} > \phi_{D/d} > \phi_{d/D} > \phi_{d/d}$; when there is CI, $\phi_{D/D} = \phi_{D/d} > \phi_{d/D} = \phi_{d/d}$. The first set of nine penetrance models (as shown) is made up by combinations of three sets of ($\phi_{D/D}$,$\phi_{d/d}$) and three sets of ($\phi_{D/d}$,$\phi_{d/D}$). The second set of nine penetrance models are obtained by halving all the penetrances in the first set.

[b] A setting is referred to the combination of a haplotype distribution (frequencies) and the penetrances of homozygous genotypes ($\phi_{D/D}$,$\phi_{d/d}$). The 27 models with penetrances as shown are referred to as the 'A' models and the 9 settings are referred to as the 'A' settings. The prevalence is 29.9% for all the 27 'A' models. By dividing all penetrance by 2, we obtain 9 'B' setting with 27 'B' models, all having a prevalence of 14.95%.

study type I error rates of the tests, while the remaining 48 (24 II models—9 A's, 9 B's and 6 C's and 24 CI models—9 A's, 9 B's and 6 C's) are for the study of power.

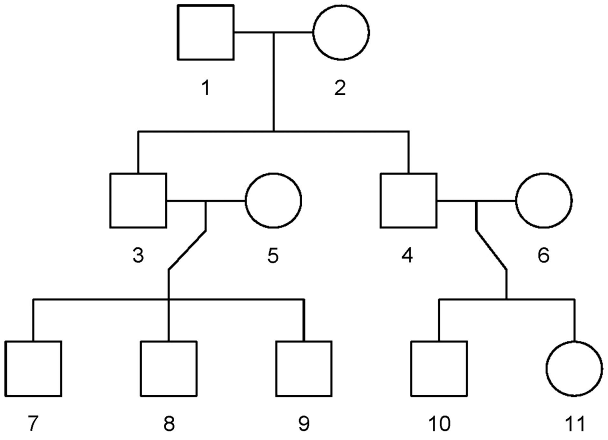

Figure 1 Pedigree structure used in the simulation study. Individuals 1, 3 and/or 6's genotypes may be missing in the analysis depending on the five incomplete data scenarios.

Each simulated data set contains 100 three-generation 11-member pedigrees with the structure shown in Figure 1. We obtain haplotype data by first generating the founders' haplotypes according to the specified haplotype frequencies and then generating the haplotype of the offspring without allowing for recombination. Next, we assign each individual's affection status according to the genotypes and the imprinting model. Genotypes of some individuals are removed in several ways as detailed below to assess the influence of incomplete data on the proposed statistics.

We simulate 1000 replicates under each of the 72 scenarios. We generate 100 MC samples of missing genotypes for each replicate. Estimated marker allele frequencies from the genotyped founders in each replicate are used in the MC sampling. All computations are based on the R environment, specifically the R package MC-PDT, which contains the PPAT and MCPPAT tests in its earlier version;[13] we have added to the package our implementations of PPATu and MCPPATu.

To assess the performance of PPATu and MCPPATu and compare with PPAT and MCPPAT, we considered five missing data scenarios: complete data without any missing genotype (MS_0), incomplete data with individual 1's genotype missing (MS_1), incomplete data with individual 3's genotype missing (MS_3), incomplete data with

Table 3 Empirical type I error (%) for four tests[a] based on 1000 replications with a nominal significance levels of 0.5% and five incomplete data scenarios for the 'A' settings

	MS$_0$[b]		MS$_1$				MS$_3$				MS$_6$				MS$_{1,6}$			
Setting	*T*	*Tu*	*T*	*Tu*	*MCT*	*MCTu*	*T*	*Tu*	*MCT*	*MCTu*	*T*	*Tu*	*MCT*	*MCTu*	*T*	*Tu*	*MCT*	*MCTu*
1A	0.5	0.5	0.3	0.7	0.6	0.4	0.2	0.6	0.6	0.5	0.6	0.5	0.5	0.5	0.4	0.3	0.5	0.6
2A	0.7	0.6	0.8	0.5	0.4	0.3	0.7	0.6	0.8	0.4	0.8	0.4	1.0	0.7	1.0	0.6	0.7	0.6
3A	0.3	0.5	0.4	0.3	0.3	0.4	0.2	0.2	0.6	0.5	0.4	0.4	0.2	0.5	0.3	0.2	0.2	0.4
4A	0.8	0.3	0.5	0.2	0.6	0.5	0.5	0.2	0.7	0.3	0.2	0.0	0.4	0.5	0.1	0.5	0.5	0.4
5A	0.5	0.5	0.7	0.8	0.4	0.3	0.1	0.3	0.6	0.3	0.8	0.4	0.7	0.6	0.5	0.2	0.6	0.3
6A	0.8	0.6	0.5	0.5	0.6	0.1	0.5	0.4	0.7	0.4	0.4	0.3	0.9	0.7	0.4	0.6	0.6	0.4
7A	0.7	0.6	0.8	0.5	0.7	0.5	0.8	0.6	0.9	0.7	0.5	0.6	0.7	0.5	0.3	0.0	0.9	0.6
8A	0.3	0.6	0.4	0.5	0.2	0.5	0.4	0.5	0.3	0.4	0.9	0.8	0.6	0.5	0.4	0.7	0.4	0.6
9A	0.6	0.7	0.4	0.3	0.6	0.4	0.5	0.2	0.5	0.6	0.5	0.5	0.6	0.4	0.6	0.3	0.7	0.4

Abbreviations: PPATu, Pedigree Parental-Asymmetry Test unaffected; MCPPATu, Monte Carlo Pedigree Parental-Asymmetry Test using both affected and unaffected.
[a]The abbreviations for the four tests are: T = PPAT, Tu = PPATu, MCT = MCPPAT and MCTu = MCPPATu.
[b]Under MS_0, there is no missing genotypes and therefore the MC versions are not applicable.

Table 4 Empirical type I error (%) of four tests[a] based on 1000 replications with a nominal significance level of 1% and five incomplete data scenarios for the 'B' settings

	MS$_0$[b]		MS$_1$				MS$_3$				MS$_6$				MS$_{1,6}$			
Setting	*T*	*Tu*	*T*	*Tu*	*MCT*	*MCTu*	*T*	*Tu*	*MCT*	*MCTu*	*T*	*Tu*	*MCT*	*MCTu*	*T*	*Tu*	*MCT*	*MCTu*
1B	0.7	1.0	0.6	0.7	0.8	1.3	0.8	0.5	1.3	0.8	1.3	0.9	0.8	0.9	0.4	0.9	0.7	1.0
2B	0.6	1.1	1.1	1.1	0.4	1.1	0.8	0.6	0.8	1.0	0.5	1.4	0.8	1.2	0.4	1.2	0.5	1.2
3B	0.8	1.1	1.1	0.9	1.3	0.8	0.6	0.9	0.9	1.0	0.8	0.7	0.9	1.2	0.8	1.4	1.3	1.1
4B	1.1	1.0	0.6	1.2	1.0	0.9	0.8	0.8	0.9	0.8	0.7	0.5	1.2	1.0	0.5	0.4	1.0	0.8
5B	0.6	1.2	0.6	0.6	0.6	1.1	0.8	0.9	0.6	1.2	0.9	1.3	0.8	0.9	0.5	0.6	0.6	0.9
6B	0.8	0.8	0.7	0.9	1.0	1.2	0.7	1.1	0.7	0.9	0.5	1.1	0.5	0.9	1.0	0.9	0.9	1.2
7B	0.6	0.9	0.3	1.0	0.4	1.1	0.4	0.7	0.6	0.8	0.7	1.0	0.8	0.7	0.7	1.1	0.7	0.9
8B	0.5	0.9	0.9	0.6	0.3	0.7	0.9	0.7	0.8	0.7	1.4	0.6	0.7	0.8	0.7	1.0	0.6	0.8
9B	0.8	1.2	0.8	1.7	0.9	1.4	0.5	0.6	0.8	1.1	0.7	1.1	0.6	1.0	0.7	0.8	0.8	1.0

Abbreviations: PPATu, Pedigree Parental-Asymmetry Test unaffected; MCPPATu, Monte Carlo Pedigree Parental-Asymmetry Test using both affected and unaffected.
[a]The abbreviations for the four tests are: T = PPAT; Tu = PPATu; MCT = MCPPAT; MCTu = MCPPATu.
[b]Under MS_0, there is no missing genotypes and therefore the MC versions are not applicable.

individual 6's genotype missing (MS_6) and incomplete data with both individual 1 and 6's genotypes missing ($MS_{1,6}$). The reason for choosing these missing schemes is to assess the effect of missing the data from the first generation (individual 1) or from the parental (second) generation who is either a nonfounder (individual 3) or a founder (individual 6). We also assess the effect of missing data on two individuals (individuals 1 and 6).

Size of tests

We investigated the empirical type I error rates for the 24 NI models with three nominal significance levels: 5, 1 and 0.5%. We did not choose to use even smaller nominal significance levels because imprinting effect tests are typically applied to only genetic markers that have been implicated for disease association, and hence the number of such markers are usually not very large. Thus, a test-wise level of 0.5% should be sufficiently small even when multiple testing is taken into account. The results for the 'A' settings at the 0.5% nominal level are given in Table 3, which show that the actual type I error rates are generally quite close to the corresponding nominal levels, substantiating the validity of MCPPATu and PPATu empirically. In addition, as an effort to evaluate whether a large proportion of cases with missing data would adversely affect the type I error rate, we split the pedigrees in which individual 3 has missing genotype into two subsets according to whether 3 is affected or not. The results

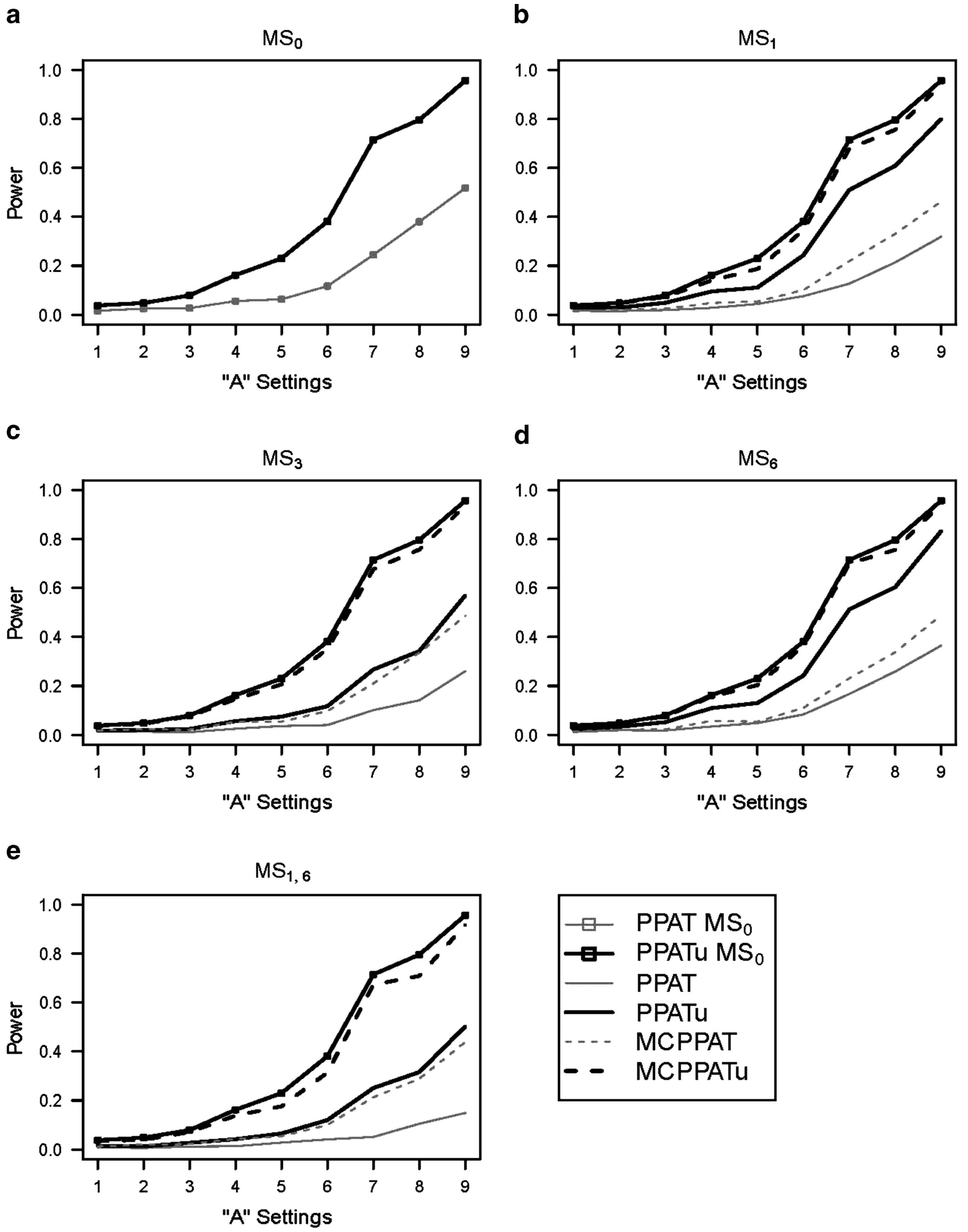

Figure 2 Power comparison among four tests (the proposed PPATu and MCPPATu and the existing PPAT and MCPPAT statistics) for the II models of the 'A' settings at the 0.5% significance level under five missing data scenarios: (**a**) MS_0, complete data without any missing genotypes; (**b**) MS_1, data missing the genotype of individual 1; (**c**) MS_3, data missing the genotype of individual 3; (**d**) MS_6, data missing the genotype of individual 6; and (**e**) $MS_{1,6}$, data missing the genotypes of both individuals 1 and 6.

(Supplementary Table S3) show that the empirical type I error rates still closely track those of the nominal values for three different levels. The results for the 'B' settings at the 1% nominal significance level are given in Table 4, which also shows that the type I error rates are well controlled for models with a smaller prevalence. The results for the other nominal type I error rates under the 'A' and 'B' settings and those under the 'C' settings (Supplementary Tables S4–S6) also show good control of type I error.

Power comparisons

The rest of the 48 scenarios (24 II and 24 CI) are used to study power. Figure 2 plots the estimated powers of the four statistics against the nine 'A' settings under the five missing data scenarios for the II models at the 0.5% significance level. When there is no missing data (under scenario MS_0), MCPPAT and MCPPATu are redundant as they give the same results as PPAT and PPATu, respectively, and thus only two power curves are given in Figure 2a. As MS_0 represents a scenario having maximally available data, the result from PPATu is treated as the 'gold standard' and plotted in the results for the other scenarios for ease of comparison. As expected, the gold standard achieves the highest power in all nine 'A' settings, indicating that using information from unaffected offspring can lead to gain in power with complete data without inflated type I error (Figure 2a). When missing data exist, PPATu generally have higher power than PPAT, and MCPPATu generally have higher power than MCPPAT, further showing the benefits of including unaffected offspring. It is also encouraging to see

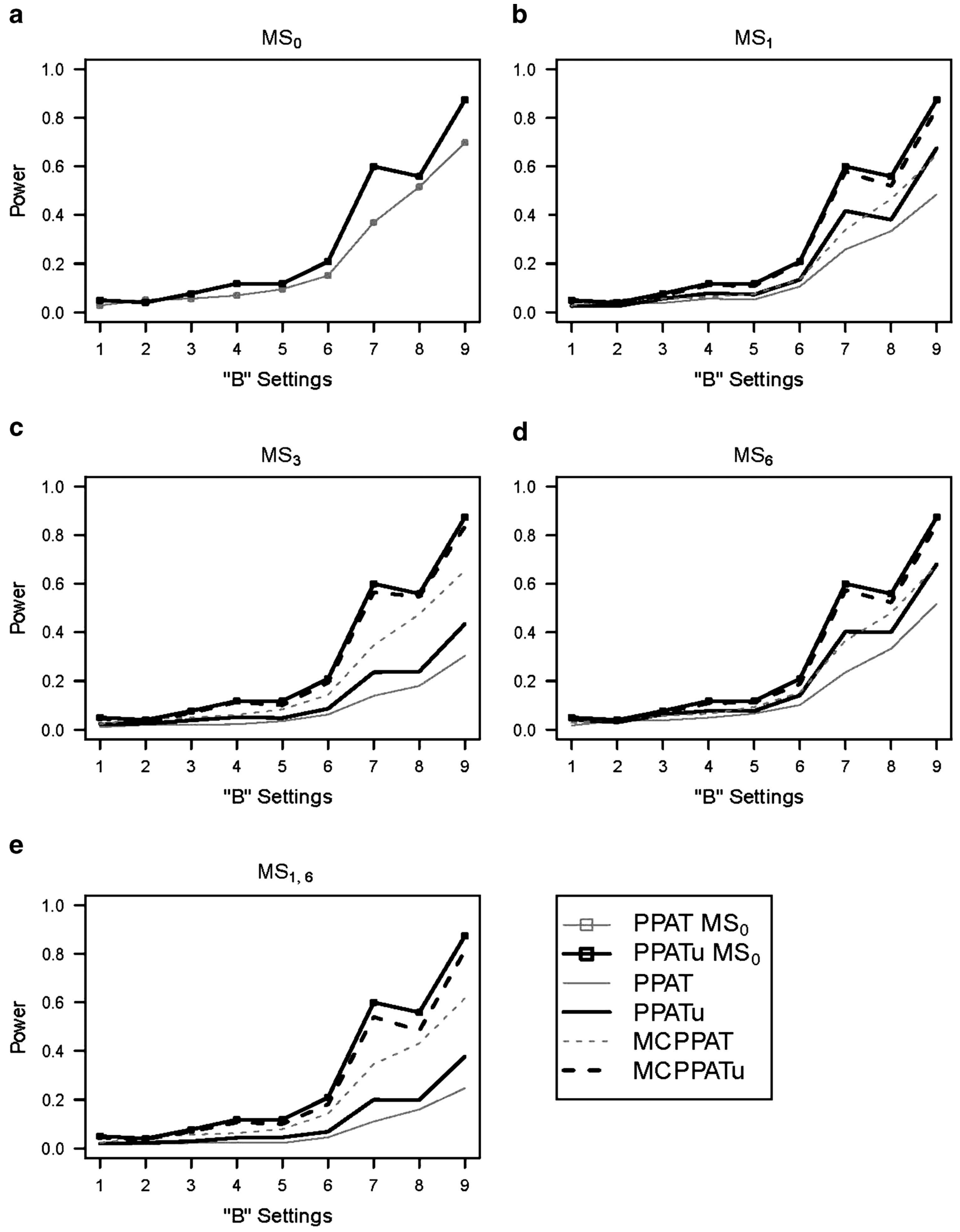

Figure 3 Power comparison among four tests (the proposed PPATu and MCPPATu and the existing PPAT and MCPPAT statistics) for the CI models of the 'B' settings at the 1% significance level under five missing data scenarios: (**a**) MS_0, complete data without any missing genotypes; (**b**) MS_1, data missing the genotype of individual 1; (**c**) MS_3, data missing the genotype of individual 3; (**d**) MS_6, data missing the genotype of individual 6; and (**e**) $MS_{1,6}$, data missing the genotypes of both individuals 1 and 6.

that, when there are missing genotypes, MCPPATu can recover the missing information well to achieve power almost reaching the gold standard (Figures 2b–e). Plots for the other combinations of imprinting models and significant levels for the 'A' settings can be found in the Supplementary Materials (Supplementary Figures S1–S5). The observation above for the II models at the 0.5% significance level applies to these figures as well.

To investigate whether the power gain with including unaffected individuals can still be realized for models with a much smaller prevalence, we carry out a power study under the 'B' settings. Results for the CI models at the 1% significance level are plotted in Figure 3. From these plots, we still see substantial power gains when unaffected individuals are included for all missing schemes, although the magnitudes of gains are all smaller compared to the corresponding 'A' settings. In general, the properties observed in the 'A' settings also apply to the 'B' settings. Most importantly, MCPPATu can recover almost all the missing information to achieve power close to that of complete data even when the prevalence is small. These observations also apply to the 'B' settings with the other combinations of imprinting models and significance levels (Supplementary Figures S6–S10) and the 'C' settings where the prevalence can be even lower (Supplementary Figures S11–S16).

DISCUSSION

Since epigenetic factors such as genomic imprinting may contribute to the explanation of missing heritability of complex traits, there is an increasing interest in factoring in such an effect in the study of disease-marker association. To contribute to this endeavor, in this paper, we propose a test for detecting an imprinting effect and show that it is generally more powerful than existing methods that are current state of the art. Our proposed MCPPATu test makes use of data from unaffected offspring in general pedigrees to enrich the information for testing for imprinting. Further, through MC sampling of unobserved genotypes conditioning on the observed ones, MCPPATu appears to be able to recover a great deal of the missing information, leading to even greater gain in power. Extensive simulation under 72 different scenarios with a wide range of prevalence shows that MCPPATu is more powerful than existing methods with well controlled type I error, even when genotypes are missing for a large proportion of cases. Under certain missing data scenarios, the gain in power through MC sampling can reach over 100% (Figure 2). To demonstrate practical feasibility of the method, we applied MCPPATu to a rheumatoid arthritis dataset. The results also point to potentially substantial increase in power given the large amount of missing genotype in the data (Supplementary Materials A.2).

To combine information from both affected and unaffected offspring, a weight needs to be specified. In our simulation results presented in this paper, we set the weight to be the population prevalence of the disease, which is typically available for common diseases from epidemiology studies (e.g. Supplementary Table S1). Further investigation with estimated population prevalence from the data for the 'B' settings show similar performances (Supplementary Table S7 and Supplementary Figures S17–S22). To show that utilizing unaffected offspring with other weights can also lead to increase in power without compromising type I error rate, we used the weight $w=1/2$ (the most extreme in some sense since the unaffected component is weighted as much as the affected component) for the 'A' settings. The results, presented in Supplementary Figures S23–S28 and Supplementary Table S8, show that including unaffected individuals in the analyses still lead to an increase of power without compromising the type I error even with this extreme setting of the weight. Nevertheless, using population prevalence as the weight is preferred as the results indicate greater power gains (Figure 2, Supplementary Figure S1–S5 versus S23–S28).

For performing MC sampling of missing genotypes, we need allele frequencies if there is missing founder genotypes. For the procedure to be valid, an important assumption is that the underlying population is homogeneous; that is, all pedigrees in the dataset are from the same population. Ding *et al.*[21] showed that when population stratification exists but the sub-populations are not very different in some key factors, the effect of ignoring the population structure may be minimal. However, if the sub-populations are very different in some key aspects, such as causing different missing patterns, then the impact could be large.

In this article, we focus on considering genetic information for each marker separately. However, it is possible to extend the method to consider haplotypes encompassing multiple markers as in previous studies.[22–24] Further, imprinted loci are believed to interact with one another owing to imprinted gene network;[20,25,26] therefore, it is of considerable interest to explore the potential of extending PPATu to that setting. For example, if two loci interact in a recessive manner to express epistatic imprinting, PPATu can be modified to capture this type of imprinting by revising the expression for excess/deficiency of transmission of the alleles by considering both loci together.

Web resources

The URLs for data presented herein are as follows:

MCPDT, http://www.stat.osu.edu/ ~ statgen/SOFTWARE/MC-PDT/;
R Project for Statistical Computing, http://www.r-project.org/.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank two anonymous referees for insightful and constructive comments, which, we believe, have led to an improved manuscript. This work was supported in part by a National Science Foundation Grant DMS-1208968.

1 Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462,** 868–874 (2009).
2 Ferguson-Smith, A. C. Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.* **12,** 565–575 (2011).
3 Falls, J. G., Pulford, D. J., Wylie, A. A. & Jirtle, R. L. Genomic imprinting: implications for human disease. *Am. J. Pathol.* **154,** 635–647 (1999).
4 Viljoen, D. & Ramesar, R. Evidence for paternal imprinting in familial Beckwith–Wiedemann syndrome. *J. Med. Genet.* **29,** 221–225 (1992).
5 Wakeling, E. L., Abu-Amero, S., Price, S. M., Stanier, P., Trembath, R. C., Moore, G. E. *et al.* Genetics of Silver–Russell syndrome. *Horm. Res.* **49,** 32–36 (1998).
6 Morison, I. M., Paton, C. J. & Cleverley, S. D. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.* **29,** 275–276 (2001).
7 Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G. P., Haig, D. *et al.* High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329,** 643–648 (2010).
8 Wang, X., Sun, Q., McGrath, S. D., Mardis, E. R., Soloway, P. D. & Clark, A. G. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS ONE* **3,** e3839 (2008).
9 Li, Y., Guo, Y., Wang, J., Hou, W., Chang, M. N., Liao, D. *et al.* A statistical design for testing transgenerational genomic imprinting in natural human populations. *PLoS ONE* **6,** e16858 (2011).
10 Li, X., Sui, Y., Liu, T., Wang, J., Li, Y., Lin, Z. *et al.* A model for family-based case-control studies of genetic imprinting and epistasis. *Brief. Bioinform.* (e-pub ahead of print 24 July 2013; doi:10.1093/bib/bbt050) .
11 Weinberg, C. R. Methods for detection of parent-of-origin effects in genetic studies of case–parents triads. *Am. J. Hum. Genet.* **65,** 229–235 (1999).

12 Zhou, J. Y., Hu, Y. Q., Lin, S. & Fung, W. K. Detection of parent-of-origin effects based on complete and incomplete nuclear families with multiple affected children. *Hum. Hered.* **67,** 1–12 (2009).
13 Zhou, J. Y., Ding, J., Fung, W. K. & Lin, S. Detection of parent-of-origin effects using general pedigree data. *Genet. Epidemiol.* **34,** 151–158 (2010).
14 Zhou, J. Y., Mao, W. G., Li, D. L., Hu, Y. Q., Xia, F. & Fung, W. K. A powerful parent-of-origin effects test for qualitative traits incorporating control children in nuclear families. *J. Hum. Genet.* **57,** 500–507 (2012).
15 Ainsworth, H. F., Unwin, J., Jamison, D. L. & Cordell, H. J. Investigation of maternal effects, maternal–fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genet. Epidemiol.* **35,** 19–45 (2011).
16 Weinberg, C. R., Wilcox, A. J. & Lie, R. T. A log-linear approach to case–parent–triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am. J. Hum. Genet.* **62,** 969–978 (1998).
17 Yang, J. & Lin, S. Detection of imprinting and heterogeneous maternal effects on high blood pressure using Framingham Heart Study data. *BMC. Proc.* **3,** S125 (2009).
18 Yang, J. & Lin, S. Likelihood approach for detecting imprinting and *in utero* maternal effects using general pedigrees from prospective family-vased association studies. *Biometrics* **68,** 477–485 (2012).
19 Zhang, F. & Lin, S. Detection of imprinting effects for hypertension based on general pedigrees utilizing all affected and unaffected individuals. *BMC Proc.* **8,** S52 (2014).
20 Cannings, C., Thompson, E. & Skolnick, M. Probability functions on complex pedigrees. *Adv. Appl. Prob.* **10,** 26–61 (1978).
21 Ding, J., Lin, S. & Liu, Y. Monte Carlo pedigree disequilibrium test for markers on the X chromosome. *Am. J. Hum. Genet.* **79,** 567–573 (2006).
22 Cheng, Y., Berg, A., Wu, S., Li, Y. & Wu, R. Computing genetic imprinting expressed by haplotypes. *Method Mol. Biol.* **573,** 189–212 (2009).
23 Wen, S., Wang, C. G., Berg, A., Li, Y., Chang, M. N., Fillingim, R. B. *et al.* Modeling genetic imprinting effects of DNA sequences with multilocus polymorphism data. *Algor. Mol. Biol.* **4,** 11 (2009).
24 Zhou, J. Y., Lin, S., Fung, W. K. & Hu, Y. Q. Detection of parent-of-origin effects in complete and incomplete nuclear families with multiple affected children using multiple tightly linked markers. *Hum. Hered.* **67,** 116–127 (2009).
25 Varrault, A., Gueydan, C., Delalbre, A., Bellmann, A., Houssami, S., Aknin, C. *et al.* Zac1 regulates an imprinted gene network critically involved in the control of embryonic growth. *Dev. Cell.* **11,** 711–722 (2006).
26 Wolf, J. B. Evolution of genomic imprinting as a coordinator of coadapted gene expression. *Proc. Natl Acad. Sci. USA* **110,** 5085–5090 (2013).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)