RESEARCH ARTICLE

Detection of parent-of-origin effects for quantitative traits using general pedigree data

HAI-QIANG HE[1], WEI-GAO MAO[1], DONGDONG PAN[2], JI-YUAN ZHOU[1]*, PING-YAN CHEN[1] and WING KAM FUNG[3]

[1]*Department of Biostatistics, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou 510515, People's Republic of China*
[2]*Department of Statistics, Yunnan University, Kunming 650091, People's Republic of China*
[3]*Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, People's Republic of China*

Abstract

Genomic imprinting is a genetic phenomenon in which certain alleles are differentially expressed in a parent-of-origin-specific manner, and plays an important role in the study of complex traits. For a diallelic marker locus in human, the parental-asymmetry tests Q-PAT(c) with any constant c were developed to detect parent-of-origin effects for quantitative traits. However, these methods can only be applied to deal with nuclear families and thus are not suitable for extended pedigrees. In this study, by making no assumption about the distribution of the quantitative trait, we first propose the pedigree parental-asymmetry tests Q-PPAT(c) with any constant c for quantitative traits to test for parent-of-origin effects based on nuclear families with complete information from general pedigree data, in the presence of association between marker alleles under study and quantitative traits. When there are any genotypes missing in pedigrees, we utilize Monte Carlo (MC) sampling and estimation and develop the Q-MCPPAT(c) statistics to test for parent-of-origin effects. Various simulation studies are conducted to assess the performance of the proposed methods, for different sample sizes, genotype missing rates, degrees of imprinting effects and population models. Simulation results show that the proposed methods control the size well under the null hypothesis of no parent-of-origin effects and Q-PPAT(c) are robust to population stratification. In addition, the power comparison demonstrates that Q-PPAT(c) and Q-MCPPAT(c) for pedigree data are much more powerful than Q-PAT(c) only using two-generation nuclear families selected from extended pedigrees.

[He H.-Q., Mao W.-G., Pan D., Zhou J.-Y., Chen P.-Y. and Fung W.-K. 2014 Detection of parent-of-origin effects for quantitative traits using general pedigree data. *J. Genet*. **93**, 339–347]

Introduction

Genomic imprinting is important in the study of complex traits, being a special genetic phenomenon in which certain alleles are differentially expressed in a parent-of-origin-specific manner. Increasingly, researchers have reported that genomic imprinting is related to human genetic diseases such as cat cry syndrome, Miller–Dieker syndrome, hereditary glomus tumours, oesophageal cancer, and Beckwith–Wiedemann, Prader–Willi and Angelman syndromes (vanTuinen *et al*. 1988; Overhauser *et al*. 1990; Hibi *et al*. 1996; Struycken *et al*. 1997; Falls *et al*. 1999; Ziegler and Konig 2006). Further, imprinting effects are suspected or hypothesized to play an important role in some other complex diseases such as autism, diabetes, hereditary paragangliomas, intrauterine growth retardation, neural tube defects, obesity and schizophrenia (Chatkupt *et al*. 1992; Temple *et al*. 1995; Abel 2004; Dong *et al*. 2005; Samaco *et al*. 2005).

Recently, there has been considerable interest in the detection of parent-of-origin effects for qualitative and quantitative traits in humans (Weinberg 1999; Shete and Amos 2002; Whittaker *et al*. 2003; Zhou *et al*. 2009; He *et al*. 2011). For qualitative traits and nuclear families, the parental-asymmetry test (PAT) is simple and powerful for the detection of parent-of-origin effects in the absence of the maternally mediated effect (Weinberg 1999). Zhou *et al*. (2012) extended PAT and proposed the C-PATu test by using both the control and case children in nuclear families with one or both parents. Zhou *et al*. (2010) further extended PAT and proposed a method to accommodate general pedigree data. On the other hand, the increasing interest of the influence

*For correspondence. E-mail: zhoujiyuan@gmail.com.

Keywords. general pedigree; missing data; Monte Carlo sample; parent-of-origin effects; quantitative trait.

of parental imprinting on heritability underscores the importance of incorporating imprinting effects into association analysis. Thus, several extensions of the transmission disequilibrium test (TDT) were developed to test for association based on nuclear family data by incorporating information on imprinting effects, which are much more powerful than the original TDT (Hu *et al.* 2007a, b; Xia *et al.* 2011).

For quantitative traits, the parental-asymmetry tests Q-PAT(c) with any constant c were developed to detect parent-of-origin effects at a diallelic marker locus (He *et al*. 2011). Their extensions Q-1-PAT(c) and Q-C-PAT(c) can serve as effective methods to tackle nuclear families with some individuals' genotypes missing (He *et al*. 2011). However, these methods for quantitative traits can only be applied to deal with two-generation nuclear families and thus are not suitable for extended pedigrees. On the other hand, it should be noted that there may be missing genotypes for some individuals in pedigrees. As such, to make full use of the available information from the pedigrees, Ding *et al*. (2006) utilized Monte Carlo (MC) sampling and simulated missing genotypes based on the observed ones from their relatives, and then developed a test for X-chromosomal markers to test for association in the presence of linkage.

In this study, by making no assumption about the distribution of the quantitative trait, we first propose the pedigree parental-asymmetry tests Q-PPAT(c) with any constant c for quantitative traits to test for parent-of-origin effects based on nuclear families with complete information from general pedigree data, in the presence of association between marker alleles under study and quantitative traits. When there are any genotypes missing in pedigrees, following Ding *et al*. (2006), we utilize MC sampling and estimation and further develop the Q-MCPPAT(c) statistics to test for parent-of-origin effects. Finally, various simulation studies are conducted to assess the performance of the proposed methods, for different sample sizes, genotype missing rates, degrees of imprinting effects and population models. Simulation results show that the proposed methods control the size well under the null hypothesis of no parent-of-origin effects, with the constant c being taken as the population mean of the quantitative trait or the mean quantitative trait value estimated based on all the nonfounders in the sample. Further, Q-PPAT(c) are robust to population stratification. In addition, the power comparison demonstrates that Q-PPAT(c) and Q-MCPPAT(c) for pedigree data are much more powerful than Q-PAT(c) only using two-generation nuclear families selected from extended pedigrees.

Materials and methods

Background and notations

Consider a candidate marker with two alleles M_1 and M_2. We use F, M and C_j to denote the genotype codes of the father, mother and the jth child in a nuclear family, respectively. Thus, F, M and C_j take values of 2, 1 and 0 for genotypes M_1M_1, M_1M_2 and M_2M_2, respectively. We also introduce four ordered genotypes to describe the parental origin of allele M_1 for a child: M_1/M_1, M_1/M_2, M_2/M_1 and M_2/M_2, where the left allele is taken to be paternal and the right one maternal. Let Q denote the quantitative trait value of an individual, and four mean trait values for genotypes M_1/M_1, M_1/M_2, M_2/M_1 and M_2/M_2 are respectively denoted by μ_{11},μ_{12}, $\mu_{21}$and μ_{22}. So, $\mu_{12} = \mu_{21}$ suggests no parent-of-origin effects; otherwise, there is a parent-of-origin effect (Weinberg *et al*. 1998). Also, $\mu_{12} < \mu_{21}(\mu_{12} > \mu_{21})$ is indicative of a paternal (maternal) imprinting effect when M_1 is positively associated with the trait, or vice versa when M_1 is negatively associated with the trait.

Throughout this study, missingness of a parental genotype is assumed to be independent of the parental underlying genotype. We also presume that mating symmetry is valid across parents within each mating type, i.e. $P(F = f, M = m) = P(F = m, M = f)$ for all f, $m = 0$, 1 and 2, and there is no maternally mediated genotype effect.

Existing Q-PAT(c) methods for two-generation nuclear families

We begin by describing the existing Q-PAT(c) here. Suppose that we collect n independent nuclear families and there are l_i children in the ith family. For any constant c, the following Q-PAT(c) tests were proposed to detect parent-of-origin effects for quantitative traits (He *et al*. 2011).

$$\text{Q-PAT}\,(c) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{l_i}\left(Q_{ij}-c\right)\left(I_{F_i>M_i,C_{ij}=1}-I_{F_i<M_iC_{ij}=1}\right)}{\sqrt{\sum_{i=1}^{n}\left[\sum_{j=1}^{l_i}\left(Q_{ij}-c\right)^2 I_{F_i\neq M_i,C_{ij}=1}+2\sum_{j<k}\left(Q_{ij}-c\right)\left(Q_{ik}-c\right)I_{F_i\neq M_i,C_{ij}=1,C_{ik}=1}\right]}} = \frac{s\,(c)}{\sqrt{\widehat{\sigma}^2\,(c)}},$$

where F_i, M_i and C_{ij} denote the genotypes of the father, mother and the jth child at the marker locus, and Q_{ij} denotes the value of quantitative trait for the jth child in the ith family, respectively, $i = 1, \ldots, n, j = 1, \ldots, l_i$. Specifically, $I_{F>M,C=1} = 1$ means that the copies of allele M_1 in father are more than mother and their child is heterozygous, which illustrates that the allele M_1 in their child is paternal, or vice versa for $I_{F<M,C=1} = 1$.

Q-PPAT(c) for general pedigree data

Note that Q-PAT(c) may suffer from power loss since they are not suitable for general pedigree data. As such, we propose the following Q-PPAT(c) statistics to test for parent-of-origin effects, which are the extension of Q-PAT(c) to accommodate pedigree data. Consider N pedigrees with the ith pedigree having l_i nonfounders, $i = 1, \ldots, N, j = 1, \ldots, l_i$. Let

$$s_p(c) = \sum_{i=1}^{N}\sum_{j=1}^{l_i}(Q_{ij}-c)\left(I_{F_{ij}>M_{ij},C_{ij}=1} - I_{F_{ij}<M_{ij},C_{ij}=1}\right),$$

where Q_{ij} and C_{ij} denote the value of the quantitative trait and the genotype of the jth nonfounder in the ith pedigree, F_{ij} and M_{ij} are the genotypes of the father and mother of the jth nonfounder, respectively. Under the null hypothesis of no parent-of-origin effects, $E(s_p(c)) = 0$, and the corresponding proof is similar to $E(s(c)) = 0$ (He *et al*. 2011). So,

$$\begin{aligned}
\mathrm{Var}\left(s_p(c)\right) &= \mathrm{Var}\left[\sum_{i=1}^{N}\sum_{j=1}^{l_i}(Q_{ij}-c)\left(I_{F_{ij}>M_{ij},C_{ij}=1} - I_{F_{ij}<M_{ij},C_{ij}=1}\right)\right] \\
&= \sum_{i=1}^{N}\left\{\mathrm{Var}\left[\sum_{j=1}^{l_i}(Q_{ij}-c)\left(I_{F_{ij}>M_{ij},C_{ij}=1} - I_{F_{ij}<M_{ij},C_{ij}=1}\right)\right]\right\} \\
&= \sum_{i=1}^{N} E\left[\sum_{j=1}^{l_i}(Q_{ij}-c)\left(I_{F_{ij}>M_{ij},C_{ij}=1} - I_{F_{ij}<M_{ij},C_{ij}=1}\right)\right]^2 \\
&= E\left\{\sum_{i=1}^{N}\left[\sum_{j=1}^{l_i}(Q_{ij}-c)^2 I_{F_{ij}\neq M_{ij},C_{ij}=1}\right.\right. \\
&\quad \left.\left. +2\sum_{j<k}(Q_{ij}-c)(Q_{ik}-c) I_{F_{ij}\neq M_{ij},C_{ij}=1,C_{ik}=1}\right]\right\}.
\end{aligned}$$

Let $\widehat{\sigma}_p^2(c) = \sum_{i=1}^{N}\left[\sum_{j=1}^{l_i}(Q_{ij}-c)^2 I_{F_{ij}\neq M_{ij},C_{ij}=1} + 2\sum_{j<k}(Q_{ij}-c)(Q_{ik}-c) I_{F_{ij}\neq M_{ij},C_{ij}=1,C_{ik}=1}\right]$. Then, $\widehat{\sigma}_p^2(c)$ is an unbiased estimate of the variance of $s_p(c)$. Therefore, the following Q-PPAT(c) statistics are proposed to test for parent-of-origin effects.

$$T(c) = \frac{\sum_{i=1}^{N} X_i(c)}{\sqrt{\sum_{i=1}^{N}[X_i(c)]^2}} = \frac{s_p(c)}{\sqrt{\widehat{\sigma}_p^2(c)}}, \qquad (1)$$

where,

$$X_i(c) = \sum_{j=1}^{l_i}(Q_{ij}-c)\left(I_{F_{ij}>M_{ij},C_{ij}=1} - I_{F_{ij}<M_{ij},C_{ij}=1}\right).$$

When the number of pedigrees is large enough, $T(c)$ approximates a standard normal distribution.

Q-MCPPAT(c) for general pedigree data with missing genotypes

When there are missing genotypes for some individuals in pedigree data, Q-PPAT(c) use only the subset of two-generation nuclear families with complete information and ignore those with incomplete information. On the other hand, MC sampling of the missing genotypes given the observed genotypes can incorporate the information on the observed genotypes in incomplete two-generation nuclear families into analysis, which may improve test power (Ding *et al*. 2006; Zhou *et al*. 2010). Therefore, we propose the following Q-MCPPAT(c) statistics. Let X be the contribution from a pedigree to $T(c)$ in equation (1). Here, we omit the subscript that indexes the pedigree and the constant c for convenient description later. Let G_m be the missing genotypes and G_o be the observed genotypes in a pedigree. Obviously, X is not computable in the presence of missing genotypes. To this end, we use X_MC, which is the conditional expectation of X given the observed genotypes G_o, to replace X, so as to estimate $T(c)$, that is,

$$X_\mathrm{MC} = E[X|G_\mathrm{o}] = E[X(G_\mathrm{m}, G_\mathrm{o})|G_\mathrm{o}],$$

where $X(G_\mathrm{m}, G_\mathrm{o})$ depends on G_m and G_o. Note that the calculation of X_MC may be very complicated or even impossible due to so many summations over all possible missing genotypes G_m given G_o. As such, following Ding *et al*. (2006), we use the following MC simulation scheme to estimate X_MC. Specifically, we first draw independent samples G_{mk}, $k = 1, \ldots, K$ from $P(G_\mathrm{m}|G_\mathrm{o})$ and then regard the average value of all the resulting $X(G_{mk}, G_\mathrm{o})$'s as the estimate of X_MC, i.e.,

$$X_\mathrm{MC} \approx \frac{1}{K}\sum_{k=1}^{K} X(G_{mk}, G_\mathrm{o}).$$

We use the SLINK software based on the peeling algorithm to implement this process efficiently (Weeks *et al*. 1990).

Therefore, equation (1) can still be used to compute the Q-MCPPAT(c) statistics with each X being replaced by X_MC. Suppose that all the pedigrees are drawn from a certain underlying population, then $E(X_\mathrm{MC}) = 0$ under the null hypothesis of no parent-of-origin effects (see appendix). On the other hand, it should be noted that the different c

values in Q-MCPPAT(c) may have different test powers (He *et al*. 2011). However, this will not affect the validity of Q-MCPPAT(c). In this study, c is taken as the population mean of the quantitative trait in the population under study or the mean trait value of all the nonfounders in the sample (He *et al*. 2011).

Results

Settings

We carry out simulation studies to assess the performance of the proposed methods. Consider a homogeneous population with the allele frequency of M_1 at the marker under study being fixed at 0.1. Assume that the quantitative trait value follows a normal distribution with mean 1.5 and variance 1, although the assumption is not necessary for the tests. We simulate the parent-of-origin effects of the marker by imposing a shift λ on the trait value for the person inheriting a maternal copy of M_1. The mean trait values for genotypes M_1/M_1, M_1/M_2, M_2/M_1 and M_2/M_2 are $1.5 + 0.9\lambda$, $1.5 - 0.1\lambda$, $1.5 + 0.9\lambda$ and $1.5 - 0.1\lambda$, respectively. λ takes values from 0 to 0.8 in increments of 0.2. Note that $\lambda = 0$ means no parent-of-origin effects for the quantitative trait which is used to simulate the type I error rates of the proposed tests.

Figure 1 shows three common pedigree structures used in our simulation studies: (a) nuclear family with five individuals, (b) three-generation family with 10 individuals, and (c) four-generation family with 12 individuals. For each simulation setting, 50 pedigrees under each of the three pedigree structures are simulated. Fifty MC samples of missing genotypes are generated for each replicate using the SLINK software (Weeks *et al*. 1990). Either the true marker allele frequencies or those estimated from the genotyped founders in each replicate are used in the MC sampling where applicable. For assessing the proposed tests (Q-PPAT(c) and Q-MCPPAT(c)) and comparing with the existing Q-PAT(c) tests, we consider the following five types of statistics: Q-PPAT$_{\text{full}}(c)$, Q-PPAT$_{\text{incom}}(c)$, Q-MCPPAT$_{\text{T}}(c)$, Q-MCPPAT$_{\text{E}}(c)$ and Q-PAT(c). The Q-PPAT$_{\text{full}}(c)$ are calculated based on complete pedigree data without any missing genotypes using Q-PPAT(c) tests, which are considered as the gold standard. The Q-PPAT$_{\text{incom}}(c)$ are on the basis of the incomplete data after removing the genotypes of individual 1 in nuclear family, individuals 1, 4 and 5 in three-generation family, and individuals 1 and 3 in four-generation family. As such, the Q-PPAT$_{\text{incom}}(c)$ only use the subset consisting of individuals 5, 6, 9, 10, 11 and 12 in four-generation family. Two Q-MCPPAT(c) versions (Q-MCPPAT$_{\text{T}}(c)$ and Q-MCPPAT$_{\text{E}}(c)$) are investigated. As in Q-PPAT$_{\text{incom}}(c)$, both versions of Q-MCPPAT(c) assume that the genotypes of the above-mentioned individuals are missing. Q-MCPPAT$_{\text{T}}(c)$ and Q-MCPPAT$_{\text{E}}(c)$ are obtained based on the true and estimated marker allele frequencies, respectively. Finally, Q-PAT(c) only use the subset of four-generation family, which contains individuals 9, 10, 11 and 12 in four-generation family; Q-PAT(c) can be regarded as the baseline.

In the simulation settings mentioned above, we assume that the genotypes of some fixed individuals in the families are missing. To further investigate the effect of genotype missing rate on the performance of the proposed methods, we consider the missing rate taken as 0.1, 0.2 and 0.3, which mean that the genotypes of 10%, 20% and 30% individuals in the families are randomly missing, respectively. We simulate 30 and 50 pedigrees under each of the three pedigree structures. Then, the corresponding sample sizes are 90 and 150, respectively. λ is fixed to be 0 and 0.6 to study the empirical type I error rate and power of the proposed methods, respectively. Besides, other simulation settings are the same as the above homogenous population. Further, five types of statistics are considered: Q-PPAT$_{\text{full}}(c)$, Q-MCPPAT$_{\text{T}}(c)$, Q-MCPPAT$_{\text{E}}(c)$, Q-PPAT$_{\text{incom}}(c)$ and Q-PAT(c). Q-PPAT$_{\text{full}}(c)$,

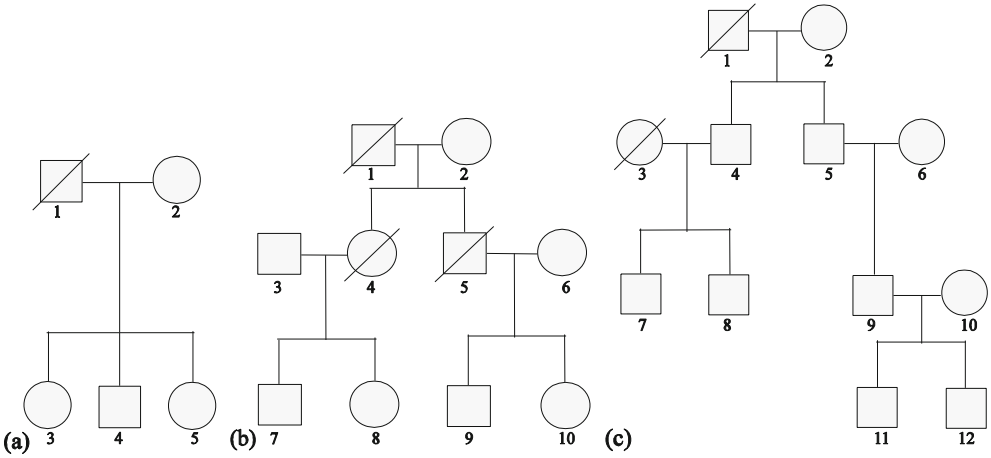

Figure 1. Pedigree structure used for the simulation studies. (a) nuclear family, (b) three-generation family, (c) four-generation family. Genotypes of individual 1 in nuclear family, individuals 1, 4 and 5 in three-generation family, and individuals 1 and 3 in four-generation family are assumed to be missing for the analysis based on incomplete data.

Table 1. Empirical type I error rates and powers of the proposed statistics based on 10000 replicates for a homogenous population.

	Complete data*		Incomplete data*							
λ	Q-PPAT$_{full}(c_1)$	Q-PPAT$_{full}(c_2)$	Q-MCPPAT$_T(c_1)$	Q-MCPPAT$_T(c_2)$	Q-MCPPAT$_E(c_1)$	Q-MCPPAT$_E(c_2)$	Q-PPAT$_{incom}(c_1)$	Q-PPAT$_{incom}(c_2)$	Q-PAT(c_1)	Q-PAT(c_2)
At nominal 5% level										
0	0.0494	0.0492	0.0493	0.0499	0.0498	0.0500	0.0425	0.0437	0.0417	0.0429
0.2	0.1970	0.1973	0.1874	0.1865	0.1866	0.1878	0.0692	0.0691	0.0567	0.0561
0.4	0.5930	0.5918	0.5663	0.5691	0.5685	0.5698	0.1318	0.1327	0.0944	0.0937
0.6	0.8947	0.8946	0.8744	0.8737	0.8747	0.8736	0.2596	0.2592	0.1634	0.1637
0.8	0.9854	0.9856	0.9780	0.9783	0.9788	0.9781	0.4049	0.4054	0.2505	0.2503
At nominal 1% level										
0	0.0071	0.0076	0.0093	0.0095	0.0089	0.0099	0.0037	0.0038	0.0036	0.0035
0.2	0.0609	0.0616	0.0585	0.0572	0.0593	0.0575	0.0077	0.0075	0.0050	0.0051
0.4	0.3291	0.3281	0.3062	0.3068	0.3053	0.3034	0.0219	0.0226	0.0103	0.0105
0.6	0.7132	0.7121	0.6782	0.6767	0.6784	0.6774	0.0652	0.0657	0.0236	0.0236
0.8	0.9264	0.9254	0.9069	0.9056	0.9074	0.9074	0.1250	0.1254	0.0457	0.0453

*c_1 is the population mean of the quantitative trait and c_2 is the mean trait value of all the nonfounders in the sample.

Table 2. Empirical type I error rates of the proposed statistics based on 10000 replicates for a homogenous population, having genotype missing rate being fixed at 0.1, 0.2 and 0.3, sample size being taken as 90 and 150, and λ being 0.

Sample size	Missing rate	Complete data*		Incomplete data*							
		Q-PPAT$_{full}(c_1)$	Q-PPAT$_{full}(c_2)$	Q-MCPPAT$_T(c_1)$	Q-MCPPAT$_T(c_2)$	Q-MCPPAT$_E(c_1)$	Q-MCPPAT$_E(c_2)$	Q-PPAT$_{incom}(c_1)$	Q-PPAT$_{incom}(c_2)$	Q-PAT(c_1)	Q-PAT(c_2)
At nominal 5% level											
	0.1	0.0492	0.0495	0.0501	0.0487	0.0501	0.0489	0.0487	0.0479	0.0434	0.0443
90	0.2	0.0477	0.0473	0.0495	0.0490	0.0476	0.0489	0.0466	0.0473	0.0452	0.0419
	0.3	0.0497	0.0492	0.0492	0.0489	0.0505	0.0489	0.0460	0.0461	0.0408	0.0397
	0.1	0.0495	0.0493	0.0506	0.0517	0.0503	0.0497	0.0475	0.0464	0.0444	0.0490
150	0.2	0.0417	0.0423	0.0480	0.0450	0.0449	0.0459	0.0446	0.0447	0.0428	0.0470
	0.3	0.0514	0.0507	0.0477	0.0494	0.0484	0.0485	0.0489	0.0493	0.0487	0.0431
At nominal 1% level											
	0.1	0.0076	0.0077	0.0077	0.0073	0.0079	0.0078	0.0071	0.0076	0.0052	0.0030
90	0.2	0.0074	0.0074	0.0083	0.0084	0.0083	0.0082	0.0061	0.0062	0.0036	0.0037
	0.3	0.0063	0.0061	0.0062	0.0064	0.0062	0.0063	0.0046	0.0047	0.0026	0.0023
	0.1	0.0099	0.0100	0.0093	0.0095	0.0100	0.0096	0.0083	0.0086	0.0066	0.0080
150	0.2	0.0069	0.0069	0.0067	0.0070	0.0066	0.0069	0.0058	0.0062	0.0041	0.0060
	0.3	0.0085	0.0088	0.0082	0.0076	0.0086	0.0089	0.0054	0.0053	0.0042	0.0053

*c_1 is the population mean of the quantitative trait and c_2 is the mean trait value of all the nonfounders in the sample.

Q-MCPPAT$_T(c)$ and Q-MCPPAT$_E(c)$ are similarly defined as above. However, Q-PPAT$_{incom}(c)$ are based on all the complete nuclear families from each pedigree. Q-PAT(c) are computed on the basis of the child–parent trios, where only one child–parent trio is randomly chosen from each pedigree.

Note that the marker allele frequencies are required in the MC sampling for Q-MCPPAT(c). To this end, Q-MCPPAT(c) need the assumption that the population under study is homogeneous (see Discussion). However, Q-PPAT(c) may be robust to population stratification, because marker allele frequencies are not needed in the calculation of Q-PPAT(c). To investigate the performance of Q-PPAT(c) in the presence of population stratification, we consider the same population stratification model as in He *et al*. (2011), which is composed of two subpopulations with equal proportions and Hardy–Weinberg equilibrium holds within each subpopulation. The allele frequency of M_1 is 0.1 (0.5) and the population mean quantitative trait value is 1.5 (0) in the first (second) subpopulation. We assume that the quantitative trait value follows a normal distribution with variance 1 in both subpopulations. The mean trait values of M_1/M_1, M_1/M_2, M_2/M_1 and M_2/M_2 in the first (second) subpopulation are $1.5 + 0.9\lambda$, $1.5 - 0.1\lambda$, $1.5 + 0.9\lambda$ and $1.5 - 0.1\lambda$ (0.5λ, -0.5λ, 0.5λ and -0.5λ), respectively. The λ values are taken as 0 to 0.8 in increments of 0.2. For each simulation setting, 50 pedigrees under each of the three types of pedigree structure in figure 1 are simulated, while without any missing genotypes. As such, we consider only the Q-PPAT(c) and Q-PAT(c) statistics here. The Q-PAT(c) are calculated based on the child–parent trios, where only one child–parent trio is randomly selected from each pedigree.

For each simulation setting mentioned above, we evaluated the empirical size and power of the proposed tests based on 10000 replicates at the significance levels of 5% and 1%.

Size and power comparison of Q-PPAT(c), Q-MCPPAT(c) and Q-PAT(c) for a homogeneous population

Table 1 gives the empirical type I error rates and simulated powers of all the statistics investigated at the significance levels of 5% and 1% based on the true and estimated c values (denoted by c_1 and c_2, respectively) and different λ values for a homogenous population. The empirical type I error rates of Q-PPAT$_{full}(c)$, Q-MCPPAT$_T(c)$ and Q-MCPPAT$_E(c)$ all stay close to the corresponding nominal levels ($\lambda = 0$), irrespective of the true or estimated c value, the true or estimated allele frequencies, which signifies the validity of the proposed Q-MCPPAT(c) and Q-PPAT(c). However, note that the sample sizes used in Q-PPAT$_{incom}(c)$ and Q-PAT(c) are both 50 (selected from 50 families of four-generation family type), which appears to be small. Thus, the simulated type I error rates of Q-PPAT$_{incom}(c)$ and Q-PAT(c) are a little conservative.

It is also shown in table 1 that for the true and estimated c values, the test powers of each type of statistics are almost the same (e.g. Q-PPAT$_{full}(c_1)$ vs Q-PPAT$_{full}(c_2)$).

Table 3. Powers of the proposed statistics based on 10000 replicates for a homogenous population, having genotype missing rate being fixed at 0.1, 0.2 and 0.3, sample size being taken as 90 and 150, and λ being 0.6.

Sample size	Missing rate	Complete data*				Incomplete data*					
		Q-PPA$_{full}(c_1)$	Q-PPAT$_{full}(c_2)$	Q-MCPPAT$_T(c_1)$	Q-MCPPAT$_T(c_2)$	Q-MCPPAT$_E(c_1)$	Q-MCPPAT$_E(c_2)$	Q-PPAT$_{incom}(c_1)$	Q-PPAT$_{incom}(c_2)$	Q-PAT(c_1)	Q-PAT(c_2)
At nominal 5% level											
90	0.1	0.6907	0.6877	0.6570	0.6566	0.6564	0.6570	0.5458	0.5457	0.1560	0.1546
	0.2	0.6837	0.6840	0.6250	0.6198	0.6294	0.6254	0.3998	0.3983	0.1405	0.1387
	0.3	0.6839	0.6844	0.5782	0.5770	0.5781	0.5773	0.2731	0.2722	0.1177	0.1203
150	0.1	0.8930	0.8936	0.8753	0.8756	0.8773	0.8753	0.7766	0.7772	0.2560	0.2580
	0.2	0.8952	0.8954	0.8473	0.8458	0.8461	0.8454	0.6259	0.6247	0.2324	0.2286
	0.3	0.8965	0.8947	0.8062	0.8050	0.8041	0.8061	0.4593	0.4572	0.1994	0.2044
At nominal 1% level											
90	0.1	0.4049	0.4034	0.3724	0.3713	0.3735	0.3717	0.2592	0.2589	0.0298	0.0288
	0.2	0.3992	0.4008	0.3295	0.3314	0.3319	0.3311	0.1405	0.1413	0.0214	0.0208
	0.3	0.4000	0.4009	0.2877	0.2864	0.2896	0.2880	0.0712	0.0696	0.0124	0.0151
150	0.1	0.7126	0.7118	0.6776	0.6780	0.6787	0.6772	0.5256	0.5244	0.0771	0.0753
	0.2	0.7107	0.7070	0.6328	0.6303	0.6314	0.6307	0.3321	0.3310	0.0656	0.0652
	0.3	0.7160	0.7153	0.5723	0.5738	0.5726	0.5699	0.1943	0.1930	0.0518	0.0503

*c_1 is the population mean of the quantitative trait and c_2 is the mean trait value of all the nonfounders in the sample.

Table 4. Empirical type I error rates and powers of the Q-PPAT(c) and Q-PAT(c) statistics based on 10000 replicates under the population stratification model.

λ	Complete data*			
	Q-PPAT(c_1)	Q-PPAT(c_2)	Q-PAT(c_1)	Q-PAT(c_2)
At nominal 5% level				
0	0.0490	0.0483	0.0501	0.0486
0.2	0.1761	0.1782	0.0764	0.0759
0.4	0.5295	0.5339	0.1641	0.1580
0.6	0.8447	0.8454	0.2935	0.2879
0.8	0.9724	0.9722	0.4570	0.4585
At nominal 1% level				
0	0.0082	0.0085	0.0073	0.0076
0.2	0.0532	0.0545	0.0178	0.0146
0.4	0.2831	0.2843	0.0481	0.0451
0.6	0.6342	0.6341	0.1053	0.1053
0.8	0.8893	0.8898	0.2075	0.2124

*c_1 is the population mean of the quantitative trait and c_2 is the mean trait value of all the nonfounders in the sample.

Further, Q-MCPPAT(c) have similar performance for the true and estimated allele frequencies needed in the MC sampling (Q-MCPPAT$_T$(c_1) vs Q-MCPPAT$_E$(c_1), Q-MCPPAT$_T$(c_2) vs Q-MCPPAT$_E$(c_2)). When the significance level and λ value are fixed, the Q-PPAT$_{full}$(c) (gold standard) indeed perform the best among all the statistics. All the other statistics based on Q-MCPPAT(c) or Q-PPAT(c) are more powerful than Q-PAT(c) (the baseline). Compared to Q-PAT(c), the power gain of Q-PPAT$_{incom}$(c) by only using individuals 5, 6, 9, 10, 11 and 12 of four-generation family type is moderate. However, the Q-MCPPAT(c), taking into account that the genotypes of several individuals are missing, lead to much more substantial power increase. They also have higher powers than Q-PPAT$_{incom}$(c), while they are just a little less powerful than the gold standard. These results demonstrate that, through MC sampling to infer the missing genotypes, much of the missing information can be recovered. On the other hand, with the λ value increasing (i.e., the degree of imprinting is increasing), the powers of all the statistics become large. Further, power decreases as the nominal level is reduced, and the effect is more profound for the Q-PPAT$_{incom}$(c) and Q-PAT(c).

Tables 2 and 3 list the empirical type I error rates and powers of all the examined statistics for two sample sizes (90 and 150) and three genotype missing rates (10, 20 and 30%) with λ being taken as 0 and 0.6 at the significance levels of 5% and 1% based on the true and estimated c values, respectively. Note that all the statistics, except for Q-PAT(c), control the type I error rates well (see table 2). The Q-PAT(c) statistics are conservative, which may be due to the small sample size of the child–parent trios. It can be seen from table 3 that for each genotype missing rate, Q-PPAT$_{full}$(c) have the highest power. Both Q-MCPPAT(c) and Q-PPAT(c) are much more powerful than Q-PAT(c). When the missing rate increases from 0.1 to 0.3 and other parameters are unchanged, the Q-MCPPAT(c) statistics show a little reduction in power, which appears to indicate that Q-MCPPAT(c) are not so sensitive to the genotype missing rate when the missing rate is small or moderate. However, the powers of Q-PPAT$_{incom}$(c) and Q-PAT(c) rapidly decline with the missing rate increasing.

Size and power comparison of Q-PPAT(c) and Q-PAT(c) under population stratification model

Table 4 shows the empirical type I error rates and powers of the Q-PPAT(c) and Q-PAT(c) statistics for the true and estimated c values and different λ values at the significance levels of 5% and 1%, under the population stratification model. It is seen from the table that all the statistics control the size well (λ = 0), irrespective of the true or estimated c value, which signifies that the Q-PPAT(c) and Q-PAT(c) statistics are robust to population stratification. As in tables 1 and 3, table 4 also displays that the power of Q-PPAT(c)/Q-PAT(c) based on the estimated c value is very close to that based on the true c value when the λ value and the significance level are unchanged. Further, Q-PPAT(c) are much more powerful than Q-PAT(c) using only trios from each pedigree. Finally, all the statistics have more power when the λ value increases, while the corresponding power becomes less when the nominal level is reduced.

Discussion

In this study, we extended the existing Q-PAT(c) tests to Q-PPAT(c), which accommodate general pedigree data for detecting parent-of-origin effects in the presence of association. When the genotypes of some individuals in the pedigrees were missing, we further proposed the Q-MCPPAT(c) based on MC sampling. Various simulation studies are conducted to assess the performance of the proposed methods, for different sample sizes, genotype missing rates, degrees of imprinting effects and population models. Our simulation results showed that both Q-PPAT(c) and

Q-MCPPAT(c) control the size well under the null hypothesis of no parent-of-origin effects. As expected, both of them are much more powerful than Q-PAT(c) when general pedigree data are available. Further, Q-MCPPAT(c) can have higher power than Q-PPAT(c) by recapturing most of the missing information. Our software, Q-MCPPAT, implemented in R (http://www.r-project.org/), is freely available at http://www.echobelt.org/web/UploadFiles/QMCPPAT.html.

Like any other method, Q-MCPPAT(c) have their own drawbacks. Note that the marker allele frequencies are needed in MC sampling based on the SLINK software (Weeks *et al*. 1990). If the number of genotyped founders in a data set is too small, the estimates of allele frequencies may not be so reliable. However, our limited simulation study appears to indicate that the results are not so sensitive to the minor perturbation of the allele frequencies. On the other hand, Q-MCPPAT(c) need the assumption that the population under study is homogeneous. If population stratification exists, but the subpopulations are similar except for the allele frequencies of some test markers, the overall effect of ignoring population structure may be minimal (Ding 2008). However, if the subpopulations are very different in some other aspects (e.g. level of missingness), the impact of assuming a single set of allele frequencies would be much larger, which needs to be further investigated in future. Finally, it is still worth emphasizing that Q-PPAT(c) do not require this assumption, and thus they are robust to population stratification and can be much more powerful than Q-PAT(c) using only nuclear family data.

Acknowledgements

We thank an anonymous reviewer for the helpful feedback and suggestions, which greatly improved our paper. This work was supported by the National Natural Science Foundation of China (81373098, 81072386 and 11301465) and the Hong Kong Research Grants Council Research Grant (766511M).

Appendix

Proof of $E(X_{MC}) = 0$ under the null hypothesis of no parent-of-origin effects

For a pedigree with n nonfounders, let S_{MC_i} denote the contribution from the ith nonfounder to X_{MC} and B_i be the event that the ith nonfounder is heterozygous. Suppose that Q_i is the value of the quantitative trait of the ith nonfounder and c is a constant. Further, let G_{o} be the collection of all the observed genotypes in this pedigree, $\mathbf{G}_{\mathrm{m}}$ be the set of all possible genotypes for individuals with missing genotypes given the observed genotypes G_{o}, and $\mathbf{G}_{\mathrm{o}}$ be the set of all possible genotypes for individuals with genotypes available in a pedigree of this type. Then, $G_{\mathrm{o}} \in \mathbf{G}_{\mathrm{o}}$,

$$X_{\mathrm{MC}} = \sum_{i=1}^{n} S_{\mathrm{MC}_i},$$

and

$$S_{\mathrm{MC}_i} = \sum_{G_{\mathrm{m}} \in \mathbf{G}_{\mathrm{m}}} I_{B_i}\,(Q_i - c)\, S_i\,(G_{\mathrm{m}}, G_{\mathrm{o}})\, P\,(G_{\mathrm{m}}\,|G_{\mathrm{o}}),$$

where $I_{B_i} = 1$ if the ith nonfounder is heterozygous and $I_{B_i} = 0$ otherwise; $S_i(G_{\mathrm{m}}, G_{\mathrm{o}}) = 1(-1)$ if the father of the ith nonfounder has more (less) copies of allele M_1 than his/her mother, else $S_i(G_{\mathrm{m}}, G_{\mathrm{o}}) = 0$, which is related to G_{m} and G_{o}. Therefore, we have

$$\begin{aligned}
E\left(S_{\mathrm{MC}_i}\right) &= E\left[\sum_{G_{\mathrm{m}} \in \mathbf{G}_{\mathrm{m}}} I_{B_i}\,(Q_i - c)\, S_i\,(G_{\mathrm{m}}, G_{\mathrm{o}})\, P\,(G_{\mathrm{m}}\,|G_{\mathrm{o}})\right] \\
&= E\left\{E\left[\sum_{G_{\mathrm{m}} \in \mathbf{G}_{\mathrm{m}}} I_{B_i}\,(Q_i - c)\, S_i\,(G_{\mathrm{m}}, G_{\mathrm{o}})\, P\,(G_{\mathrm{m}}\,|G_{\mathrm{o}})\,\middle|\, G_{\mathrm{o}}\right]\right\} \\
&= E\left\{(\mu - c)\left[\sum_{G_{\mathrm{m}} \in \mathbf{G}_{\mathrm{m}}} I_{B_i} S_i\,(G_{\mathrm{m}}, G_{\mathrm{o}})\, P\,(G_{\mathrm{m}}\,|G_{\mathrm{o}})\right]\right\} \\
&= (\mu - c) \sum_{G_{\mathrm{o}} \in \mathbf{G}_{\mathrm{o}}} \sum_{G_{\mathrm{m}} \in \mathbf{G}_{\mathrm{m}}} I_{B_i} S_i\,(G_{\mathrm{m}}, G_{\mathrm{o}})\, P\,(G_{\mathrm{m}}|G_{\mathrm{o}})\, P\,(G_{\mathrm{o}}).
\end{aligned}$$

If the ith nonfounder is heterozygous, then under the null hypothesis of no parent-of-origin effects,

$$E\left(S_{\mathrm{MC}_i}\right) = (\mu - c)\, E\,(S_i) = 0,$$

where $E\,(S_i) = 0$, because each heterozygous child has an equal chance of getting allele M_1 either from his/her father or mother under the null. So, we have

$$E\,(X_{\mathrm{MC}}) = \sum_{i=1}^{n} E\left(S_{\mathrm{MC}_i}\right) = 0.$$

References

Abel K. M. 2004 Fetal origins of schizophrenia: testable hypotheses of genetic and environmental influences. *Br. J. Psychiatry* **184**, 383–385.

Chatkupt S., Lucek P. R., Koenigsberger M. R. and Johnson W. G. 1992 Parental sex effect in spina bifida: a role for genomic imprinting. *Am. J. Med. Genet.* **44**, 508–512.

Ding J. 2008 *Monte Carlo pedigree disequilibrium test with missing data and population structure*. Ph.D. dissertation, The Ohio State University, Columbus, USA.

Ding J., Lin S. and Liu Y. 2006 Monte Carlo pedigree disequilibrium test for markers on the X chromosome. *Am. J. Hum. Genet.* **79**, 567–573.

Dong C. H., Li W. D., Geller F., Lei L., Li D., Gorlova O. Y. *et al.* 2005 Possible genomic imprinting of three human obesity-related genetic loci. *Am. J. Hum. Genet.* **76**, 421–437.

Falls J. G., Pulford D. J., Wylie A. A. and Jirtle R. L. 1999 Genomic imprinting: implications for human disease. *Am. J. Pathol.* **154**, 635–647.

He F., Zhou J.-Y., Hu Y.-Q., Sun F., Yang J., Lin S. *et al.* 2011 Detection of parent-of-origin effects for quantitative traits in complete and incomplete nuclear families with multiple children. *Am. J. Epidemiol.* **174**, 226–233.

Hibi K., Nakamura H., Hirai A., Fujikake Y., Kasai Y., Akiyama S. *et al.* 1996 Loss of H19 imprinting in esophageal cancer. *Cancer Res.* **56**, 480–482.

Hu Y.-Q., Zhou J.-Y. and Fung W. K. 2007a An extension of the transmission disequilibrium test incorporating imprinting. *Genetics* **175**, 1489–1504.

Hu Y.-Q., Zhou J.-Y., Sun F. and Fung W. K. 2007b The transmission disequilibrium test and imprinting effects test based on case-parent pairs. *Genet. Epidemiol.* **31**, 273–287.

Overhauser J., McMahon J., Oberlender S., Carlin M. E., Niebuhr E., Wasmuth J. J. *et al.* 1990 Parental origin of chromosome 5 deletions in the cri-du-chat syndrome. *Am. J. Med. Genet.* **37**, 83–86.

Samaco R. C., Hogart A. and LaSalle J. M. 2005 Epigenetic overlap in autism-spectrum neurodevelopmental disorders: MECP2 deficiency causes reduced expression of UBE3A and GABRB3. *Hum. Mol. Genet.* **14**, 483–492.

Shete S. and Amos C. I. 2002 Testing for genetic linkage in families by a variance-components approach in the presence of genomic imprinting. *Am. J. Hum. Genet.* **70**, 751–757.

Struycken P. M., Cremers C. W., Mariman E. C., Joosten F. B. and Bleker R. J. 1997 Glomus tumours and genomic imprinting: influence of inheritance along the paternal or maternal line. *Clin. Otolaryngol.* **22**, 71–76.

Temple I. K., James R. S., Crolla J. A., Sitch F. L., Jacobs P. A., Howell W. M. *et al.* 1995 An imprinted gene(s) for diabetes? *Nat. Genet.* **9**, 110–112.

vanTuinen P., Dobyns W. B., Rich D. G., Summers K. M., Robinson T. J., Nakamura Y. *et al.* 1988 Molecular detection of microscopic and submicroscopic deletions associated with Miller-Dieker syndrome. *Am. J. Hum. Genet.* **43**, 587–596.

Weeks D. E., Ott J. and Lathrop G. M. 1990 SLINK: a general simulation program for linkage analysis. *Am. J. Hum. Genet.* **47**, A204.

Weinberg C. R. 1999 Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am. J. Hum. Genet.* **65**, 229–235.

Weinberg C. R., Wilcox A. J. and Lie R. T. 1998 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.

Whittaker J. C., Gharani N., Hindmarsh P. and McCarthy M. I. 2003 Estimation and testing of parent-of-origin effects for quantitative traits. *Am. J. Hum. Genet.* **72**, 1035–1039.

Xia F., Zhou J.-Y. and Fung W. K. 2011 A powerful approach for association analysis incorporating imprinting effects. *Bioinformatics* **27**, 2571–2577.

Zhou J.-Y., Hu Y.-Q., Lin S. and Fung W. K. 2009 Detection of parent-of-origin effects based on complete and incomplete nuclear families with multiple affected children. *Hum. Hered.* **67**, 1–12.

Zhou J.-Y., Ding J., Fung W. K. and Lin S. 2010 Detection of parent-of-origin effects using general pedigree data. *Genet. Epidemiol.* **34**, 151–158.

Zhou J.-Y., Mao W.-G., Li D.-L., Hu Y.-Q., Xia F. and Fung W. K. 2012 A powerful parent-of-origin effects test for qualitative traits incorporating control children in nuclear families. *J. Hum. Genet.* **57**, 500–507.

Ziegler A. and Konig I. R. 2006 *A statistical approach to genetic epidemiology: Concepts and applications*. Wiley-VCh, New York, USA.

Received 9 July 2013, in revised form 21 August 2013; accepted 31 December 2013
Unedited version published online: 6 June 2014
Final version published online: 1 August 2014