Estimating genotype-specific call rates from offspring-parents trio data

Zhaoxia Yu

Department of Statistics
University of California, Irvine, CA 92697
yu.zhaoxia@ics.uci.edu

Abstract

Missing mechanism is an important issue in analyzing data. In genetic association studies, one important question is whether missingness of genotype data depends on the underlying true genotypes. Using the genotypic constraints imposed by the family structure of offspring-parents trios, we propose a method that estimates the three call rates at a single nucleotide polymorphism (SNP). We also propose a likelihood ratio test to test whether the call rates are all the same. We then apply the methods to a genome-wide association study and we show that rare genotypes are usually more difficult to call, which introduces systematic bias that tends to estimate the common allele to be over transmitted.

Running Title: Informative missingness in genotype calling

Introduction

The case-parents design is a simple yet efficient design that is robust against population stratification. In this design, parental genotypes are used as controls, and the transmission disequilibrium test (TDT) [1,2] and other alternative methods [3-5] have been widely used for testing genetic association using case-parents trio data. In families that only one parent's genotype is available, the allele transmitted from the parent to the affected offspring can be determined unambiguously when the genotype of the offspring is homozygous and the observed parental genotype is heterozygous. However, it has been shown that including such case-parent pairs in TDT can cause bias and these case-parent pairs should be excluded from TDT to ensure valid conclusions [6]. Many likelihood based methods have been proposed to handle missingness in parental genotypes [7-19]. Non-parametric approaches have also been developed and most of them can be thought as special scenarios, modifications, or extensions of the family-based association test framework [20]. In addition, bootstrap and multiple imputation have been considered [21,22]. More discussions regarding how to handle missing parental genotypes can be found in [23]. Missingness in both offspring and parental genotypes has also been considered [24,25].

When analyzing data with missing values, both "complete-case" analyses and "available-case" analyses often yield erroneous conclusions when missingness depends on the underlying true genotypes [26]. Such genotype-specific missingness affects family-based association tests more than case-control association tests [27,28]. However, identifying missing mechanism is usually

difficult, and often impossible. For data composed of unrelated subjects, incorporating genotype-specific missingness into statistical analysis at a single marker is not possible, as the genotype-specific call rate parameters are not identifiable [29]. When there are multiple markers that are close to each other, the linkage disequilibrium among them can be used to evaluate whether there is genotype-specific missingness [29-31]. For case-parents data, methods that incorporate informative missingness have been proposed [8,11]. These methods focus on the missingness resulted from the unavailability of parental DNA samples. Recently, Yu [24] proposed to incorporate genotype-specific call rates into a likelihood ratio test to reduce the bias caused by informative missingness in genotype calling rather than sampling.

Despite the previous efforts to incorporate informative missingness into association tests in case-parents studies, the degree of informative missingness has not been systematically investigated. Guo et al. [32] assessed the missing mechanism of parental genotypes by testing whether the conditional distributions of parental genotypes given offspring's are the same between case-parents trios and case-parent pairs. In their method, the missing mechanism of offspring's genotypes is not studied and offspring's missing status does not contribute information. For example, consider a trio with AA father, AA mother, and missing offspring genotype. This trio is excluded from their test. However, it is obvious that the offspring has the AA genotype. In family-based association studies where all the three members of each trio are successfully recruited, an important question is whether the three genotypes at a SNP have the same missing rate. Although Hao and Cawley [27] studied whether missingness relies on genotypes, they did not test informative missingness directly; instead, they tested Hardy-Weinberg equilibrium (HWE) to assess equal missingness. Using experimental methods, Fu et al. [33], recently reported informative missingness in genotyping. However, they only examined a few SNPs. Here we propose a likelihood ratio test that can be used to test whether there is genotype-specific missingness using offspring-parents data. In addition, we also propose a method to estimate call rates of the three genotypes at a SNP.

Methods

Consider a SNP with alleles A and B. Let $G_i=(G_{iF}, G_{iM}, G_{io})$ denote the vector of true genotypes for the ith trio in a data set of n offspring-parents trios, where the subscripts F, M, and O indicate father, mother, and offspring, respectively. Under the assumption of no Mendelian errors, there are fifteen possible trio types (genotype vectors), as shown in **Table 1** and we use $\theta = (\theta_1,...,\theta_{15})$ to denote the frequencies of the trio types in the sampled population. The observed genotypes for the ith trio is denoted by $g_i=(g_{iF}, g_{iM}, g_{io})$, which can contain up to three missing values. Let the vector of call rates be $c = (c_{AA}, c_{AB}, c_{BB})$.

Table 1: Trio types.

father genotype	mother genotype	offspring genotype	frequency parameter

AA	AA	AA	θ_1
AA	AB	AA	θ_2
AA	AB	AB	θ_3
AA	BB	AB	θ_4
AB	AA	AA	θ_5
AB	AA	AB	θ_6
AB	AB	AA	θ_7
AB	AB	AB	θ_8
AB	AB	BB	θ_9
AB	BB	AB	θ_{10}
AB	BB	BB	θ_{11}
BB	AA	AB	θ_{12}
BB	AB	AB	θ_{13}
BB	AB	BB	θ_{14}
BB	BB	BB	θ_{15}

Given the trio type frequencies θ, G_i follows a multinomial distribution, and given call rates, the event whether a genotype at a subject is observed or not follows a Bernoulli distribution. Thus, the complete- and observed-data likelihood functions contributed by the *i*th trio are

$$L_{i,\text{comp}}(c,\theta) = P(g_i, G_i \mid c,\theta) = P(G_i \mid \theta)P(g_i \mid G_i, c)$$
$$= \prod_{k=1}^{15}(\theta_k)^{I(G_i \text{ belongs to trio type } k)} \prod_{j\in\{F,M,O\}}(c_{G_{ij}})^{I(g_{ij} \text{ observed})}(1-c_{G_{ij}})^{I(g_{ij} \text{ unobserved})},$$
$$L_i(c,\theta) = P(g_i \mid c,\theta) = \sum_{G_i} L_{i,\text{comp}}(c,\theta),$$

respectively.

We derived the expectation-maximization (EM) steps according to the EM algorithm [34]:

E-step:

$$P(G_i \mid g_i, c^{\text{old}}, \theta^{\text{old}}) = \frac{P(G_i, g_i, c^{\text{old}}, \theta^{\text{old}})}{\sum_{G_i} P(G_i, g_i, c^{\text{old}}, \theta^{\text{old}})} = \frac{P(g_i \mid G_i, c^{\text{old}})P(G_i \mid \theta^{\text{old}})}{\sum_{G_i} P(g_i \mid G_i, c^{\text{old}})P(G_i \mid \theta^{\text{old}})}.$$

M-step:

$$c_{\text{geno}}^{\text{new}} = \frac{\sum_{\text{i,j}} I(g_{ij} \text{ observed})\text{P}(G_{ij} = \text{geno} \mid g_i, c^{\text{old}}, \theta^{\text{old}})}{\sum_{\text{i,j}} \text{P}(G_{ij} = \text{geno} \mid g_i, c^{\text{old}}, \theta^{\text{old}})},$$

$$\theta_k^{\text{new}} = \frac{\sum_i \Pr(G_i \text{ belongs to trio type } k \mid g_i, c^{\text{old}}, \theta^{\text{old}})}{n}.$$

where $i = 1,...,n$, $j \in \{AA, AB, BB\}$, geno $\in \{AA, AB, BB\}$, $k = 1,...,15$, and $I(\bullet)$ is the indicator function that is 1 when the condition in the parentheses is true and 0 otherwise. The maximum likelihood estimates of call rates are obtained once the algorithm converges.

To test whether the missingness is genotype-specific, we then conduct the EM algorithm under the null hypothesis of equal call rates, i.e., $c_{AA}=c_{AB}=c_{BB}$. The E and M steps are similar to those above except that the call rates are updated by the following formula:

$$c_{\mathrm{AA}}^{\mathrm{new}} = c_{\mathrm{AB}}^{\mathrm{new}} = c_{\mathrm{BB}}^{\mathrm{new}} = \frac{\sum_{\mathrm{i,j,geno}} I(g_{ij} \text{ observed}) \mathrm{P}(G_{ij} = \mathrm{geno} \mid g_i, c^{\mathrm{old}}, \theta^{\mathrm{old}})}{\sum_{\mathrm{i,j,geno}} \mathrm{P}(G_{ij} = \mathrm{geno} \mid g_i, c^{\mathrm{old}}, \theta^{\mathrm{old}})}.$$

Based on the above maximized likelihood, the likelihood ratio test with two degrees of freedom can be used to test the null hypothesis of equal call rates. Under the null hypothesis, it has an asymptotic chi-square distribution with two degrees of freedom. Note that we estimate the frequencies of all the 15 possible trio types (with the order of parental genotypes matters), rather than make assumptions on genetic models and random mating. Thus, our methods are not biased by model misspecifications or population stratifications.

Simulations and results

We first simulate offspring-parents trio data to examine the accuracy of the maximum likelihood estimation of call rates and the type I error rate of the likelihood ratio test. In each of 1000 simulations, 1000 trios from a random mating population are simulated under a combination of allele frequency and call rate parameters: the frequency of allele A varies from 0.1 to 0.5; the true call rates of the three genotypes are either all equal to 0.95 or all equal to 0.99. The calculation of type I error is based on the p-value cutoff of 0.05.

Figure 1 shows the difference between the estimated call rates and the true call rates using box plots. The estimation accuracy at the true call rate of 0.95 (upper panel of **Figure 1**) is lower than at 0.99 (lower panel of **Figure 1**), when the frequency of allele A is the same and the genotype group is the same. When the true call rate is the same, the estimation accuracy is low for rare genotypes and high for common genotypes. These results are expected, as it is more difficult to estimate parameters when there are more missing values or less data points.

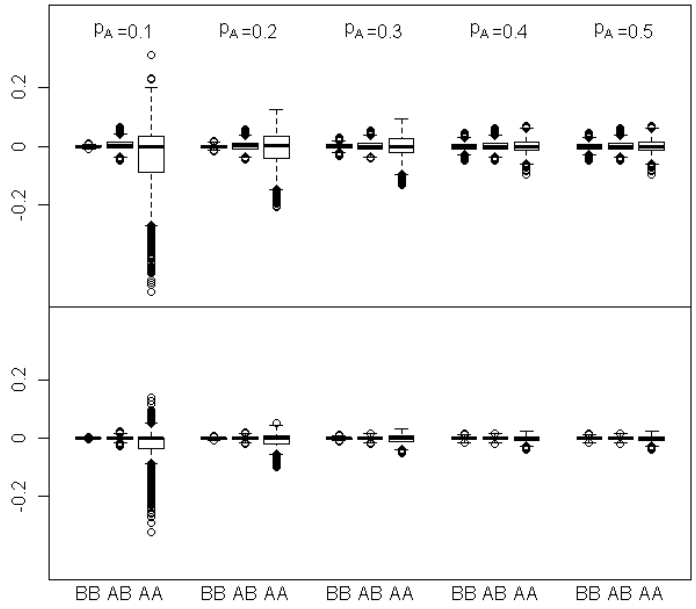


Figure 1: Box plots of the difference between the estimated call rates and the true call rates. Upper: when the true call rate is 0.95; lower: when the true call rate is 0.99. Here "AA" refers to the minor homozygous group, "AB" refers to the heterozygous group, and "BB" refers to the major homozygous group. p_A is the frequency of allele A.

The type I error rates of the likelihood ratio test are presented in **Table 2**. The values are generally not far away from 0.05 except that the test seems slightly conservative when call rates are 0.99 and liberal when call rates are 0.95.

Table 2: Type I error rates of the test for equal missingness.

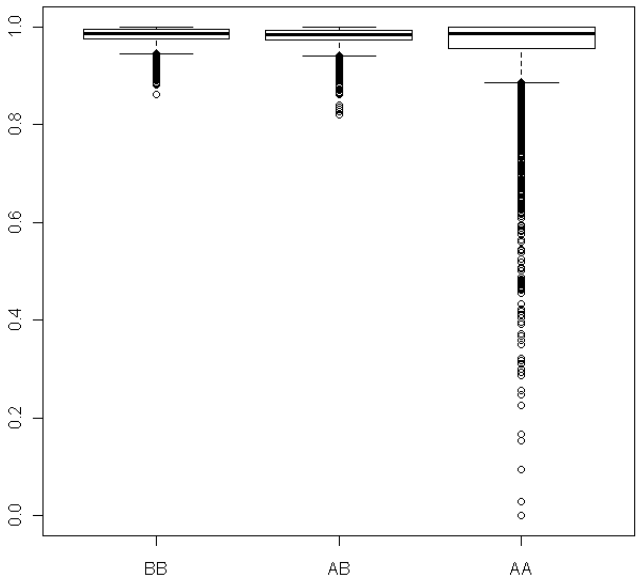
		p_A=0.1	p_A=0.2	p_A=0.3	p_A=0.4	p_A=0.5
call rate	0.95	0.049	0.049	0.053	0.062	0.066
	0.99	0.037	0.039	0.047	0.048	0.047

Application to the oral clefts study

We then apply our proposed methods to a subset of data collected by International Consortium to Identify Genes and Interactions Controlling Oral Clefts. The original study has been reported before [35] and here we focus on 889 case-parents trios of European descendents. Genotypes from 596,292 SNPs were measured using Illumina Human610. To remove low-quality SNPs, we only keep SNPs with minor allele frequency no less than 0.05, the number of Mendelian errors no more than 1, the overall call rate no less than 0.95, and the HWE chi-square no greater than 10. Note that these filters are calculated based on the observed genotypes. For SNPs with overall call rates very close to 1 we found

that all the estimated genotype-specific call rates are close to 1; thus, we further filter out SNPs with overall call rates greater than or equal to 0.99. With all these restrictions, 5,039 SNPs remain in our analysis. To examine the trend of call rates with genotype frequencies, we code the rare homozygous genotype to "AA", the common homozygous genotype to "BB", and the heterozygous genotype to "AB".

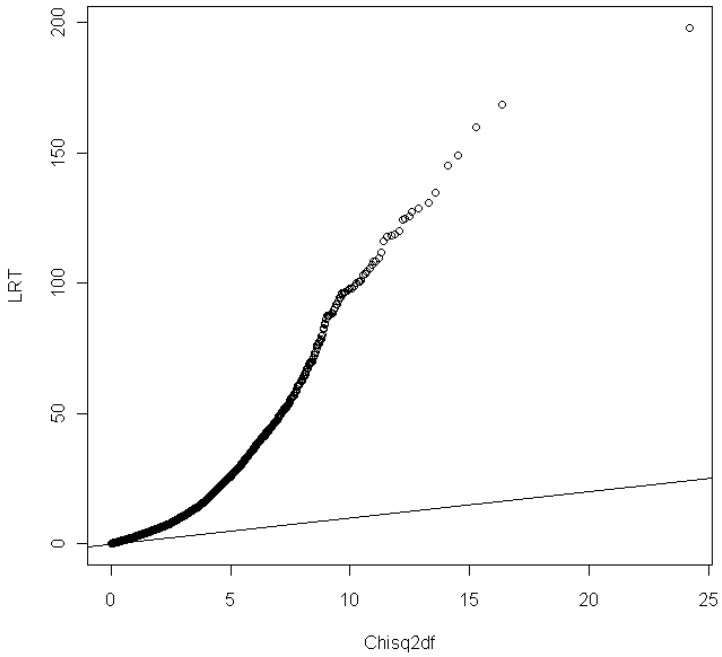
Figure 2: Box plots of the estimated call rates using SNPs from the oral clefts study. Here "AA" refers to the minor homozygous group, "AB" refers to the heterozygous group, and "BB" refers to the major homozygous group.



The box plots of estimated genotype-specific call rates are presented in **Figure 2**. Clearly the call rates decrease with the genotype frequencies. Although all the SNPs used in **Figure 2** have overall rates between 0.95 and 0.99, we see that the call rates for the rare genotype group can be as low as below 0.5. This result agrees with what have been seen from previous experimental studies [33].

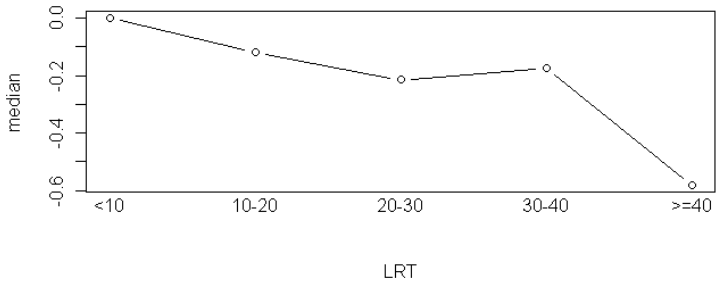
We next examine the null hypothesis of equal missingness across genotypes by plotting a quantitle-quantile plot of the likelihood ratio test statistics against the chi-square distribution with two degrees of freedom. As shown in **Figure 3**, the observed data are quite far away from the straight line with an intercept of 0 and a slope of 1, which indicates that there are an excessively large number of SNPs that show genotype-specific missingness.

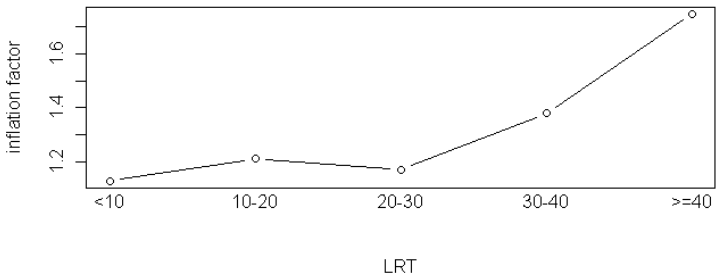
Figure 3: The quantile-quantile plot of the observed likelihood ratio test statistics for equal missingness against the chi-square distribution with two degrees of freedom. The intercept and the slope of the straight line are 0 and 1, respectively.



The median of the TDT statistic, calculated using complete trios, is -0.066. Note that in our calculation we use the common allele as the reference allele and the rare allele as the risk allele. Thus, this negative median implies that the bias due to informative missingness is toward predicting the common allele as the risk allele, as have been observed in [24]. To examine false positives, we calculate the inflation factor [36], which is defined as the ratio of the observed median of the squared TDT statistic to the median of the chi-square distribution with one degree of freedom. The inflation factor is calculated to be 1.16, which indicates inflated Type I errors. To examine the relationship between the magnitude of bias in association tests and the degree of informative missingness, we present the medians and inflation factors stratified by the values of the likelihood ratio test statistics: <10 (3853 SNPs), 10-20 (612 SNPs), 20-30 (234 SNPs), 30-40 (120 SNPs), and ≥40 (220 SNPs). The upper panel and the lower panel of **Figure 4** present the stratified medians and the stratified inflation factors, respectively. The overall trend is that both the bias and the inflation increase with the likelihood ratio test statistic, which shows that the identified informative missingness can have a negative impact on association tests for case-parents studies. Note that we have not seen a TDT statistic that reaches significance at the genome-wide level. Thus, informative missingness is likely to cause a small but systematic bias for genome-wide association studies using the case-parents design.

Figure 4: The medians and inflation factors of TDT with complete trios, stratified by the likelihood ratio test statistic for equal missingness.





Discussion

In this article we proposed methods to study the missing mechanism of genotypes that is caused by imperfect genotyping technologies. The family structure imposes restrictions on genotypes of the members of a family thus provides useful information to make statistical inference for missing mechanisms of genotypes. Based on this observation, we proposed a method to estimate genotype-specific call rates. We also proposed a likelihood ratio test to examine whether missingness during genotype calling is informative. We found that informative missingness during genotype calling is probably the norm rather than the exception.

In our methods we assume that the DNA samples of all subjects in a study were processed together. If this is not true, for example when parental genotypes were collected at a later time, pulling data from samples processed at different sites or different time points might introduce biases [37]. In this situation, we can use different call rates for samples in different batches. Multiple types of DNA specimens, such as blood or saliva, were often collected in the same study. If the difference in different types is a concern, we can also use different call rate parameters for different types of specimens.

We analyzed real genotype data measured using Illumina Human610. The methods we proposed can also be applied to data generated by other genotyping platforms, such as more recent Illumina genotyping platforms and genotyping platforms based on other technologies. The methods can also be used to evaluate the genotype missing mechanisms for data from next-generation sequencing data when both offspring and parents are sequenced. Our methods are developed for

offspring-parents data, and similar strategies can be used to develop methods for other study designs, although there are some challenges. For example, offspring-mother pairs can be used, although a large number of pairs might be needed to obtain reliable information, as the genotypic restriction in offspring-mother pairs is not as strong as that in offspring-parents trios. All family members in nuclear families can also be used, but some assumptions, such as genetic models and HWE, might be needed, which could lead to biased results when population stratification presents.

Acknowledgements

The author was partially supported by grant NIH/R01 HG004960.

We sincerely thank all of the patients and families who participated the study entitled “International Consortium to Identify Genes and Interactions Controlling Oral Clefts”.
Datasets used for the analyses described in this manuscript were obtained from dbGaP at
[http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000094.v1.p1] through dbGaP accession number phs000094v1. The genome-wide association study is part of the Gene Environment Association Studies (GENEVA) program of the trans-NIH Genes, Environment and Health Initiative [GEI] supported by U01-DE-018993. Genotyping services were provided by the Center for Inherited Disease Research (CIDR), fully funded by HHSN268200782096C. Assistance with genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01-HG-004446) and by the National Center for Biotechnology Information (NCBI).

References

1 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). Am J Hum Genet 1993;52:506-516.
2 Terwilliger JD, Ott J: A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 1992;42:337-346.
3 Schaid DJ, Sommer SS: Genotype relative risks: Methods for design and analysis of candidate-gene association studies. Am J Hum Genet 1993;53:1114-1126.
4 Weinberg CR, Wilcox AJ, Lie RT: A log-linear approach to case-parent-triad data: Assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 1998;62:969-978.
5 Self SG, Longton G, Kopecky KJ, Liang KY: On estimating hla/disease association with application to a study of aplastic anemia. Biometrics 1991;47:53-61.
6 Curtis D, Sham PC: A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 1995;56:811-812.
7 Weinberg CR: Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 1999;64:1186-1193.

8 Allen AS, Rathouz PJ, Satten GA: Informative missingness in genetic association studies: Case-parent designs. Am J Hum Genet 2003;72:671-680.
9 Clayton D: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 1999;65:1170-1177.
10 Cervino AC, Hill AV: Comparison of tests for association and linkage in incomplete families. Am J Hum Genet 2000;67:120-132.
11 Chen YH: New approach to association testing in case-parent designs under informative parental missingness. Genet Epidemiol 2004;27:131-140.
12 Chen JH, Cheng KF: A robust tdt-type association test under informative parental missingness. Stat Med 2011;30:291-297.
13 Guo CY, DeStefano AL, Lunetta KL, Dupuis J, Cupples LA: Expectation maximization algorithm based haplotype relative risk (em-hrr): Test of linkage disequilibrium using incomplete case-parents trios. Hum Hered 2005;59:125-135.
14 Schaid DJ, Li H: Genotype relative-risks and association tests for nuclear families with missing parental data. Genet Epidemiol 1997;14:1113-1118.
15 Bergemann TL, Huang Z: A new method to account for missing data in case-parent triad studies. Hum Hered 2009;68:268-277.
16 Zhou JY, Lin SL, Fung WK, Hu YQ: Detection of parent-of-origin effects in complete and incomplete nuclear families with multiple affected children using multiple tightly linked markers. Hum Hered 2009;67:116-127.
17 Hsieh HJ, Palmer CGS, Sinsheimer JS: Allowing for missing data at highly polymorphic genes when testing for maternal, offspring and maternal-fetal genotype incompatibility effects. Hum Hered 2006;62:165-174.
18 Dudbridge F: Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered 2008;66:87-98.
19 Yang Y, Wise CA, Gordon D, Finch SJ: A family-based likelihood ratio test for general pedigree structures that allows for genotyping error and missing data. Hum Hered 2008;66:99-110.
20 Rabinowitz D, Laird N: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 2000;50:211-223.
21 Croiseau P, Genin E, Cordell HJ: Dealing with missing data in family-based association studies: A multiple imputation approach. Hum Hered 2007;63:229-238.
22 Allen AS, Collins JS, Rathouz PJ, Selander CL, Satten GA: Bootstrap calibration of transmit for informative missingness of parental genotype data. Bmc Genet 2003;4 Suppl 1:S39.
23 Laird NM, Lange C: Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 2006;7:385-394.
24 Yu Z: Family-based association tests using genotype data with uncertainty. Biostatistics 2012;First published online: December 8, 2011
25 Sebastiani P, Abad MM, Alpargu G, Ramoni MF: Robust transmission/disequilibrium test for incomplete family genotypes. Genetics 2004;168:2329-2337.
26 Little RJA, Rubin DB: Statistical analysis with missing data, second edition. New York, Wiley, 2002.
27 Hao K, Cawley S: Differential dropout among snp genotypes and impacts on association tests. Hum Hered 2007;63:219-228.

28 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6:95-108.
29 Liu NJ, Bucala R, Zhao HY: Modeling informatively missing genotypes in haplotype analysis. Commun Stat-Theor M 2009;38:3445-3460.
30 Liu NJ, Beerman I, Lifton R, Zhao HY: Haplotype analysis in the presence of informatively missing genotype data. Genet Epidemiol 2006;30:290-300.
31 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: Plink: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-575.
32 Guo CY, Cupples LA, Yang Q: Testing informative missingness in genetic studies using case-parent triads. European Journal of Human Genetics 2008;16:992-1001.
33 Fu WQ, Wang Y, Wang Y, Li R, Lin R, Jin L: Missing call bias in high-throughput genotyping. Bmc Genomics 2009;10:106.
34 Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc B Met 1977;39:1-38.
35 Beaty TH, Murray JC, Marazita ML, Munger RG, Hetmanski IRJB, Liang KY, Wu T, Murray T, Fallin MD, Redett RA, Raymond G, Schwender H, Jin SC, Cooper ME, Dunnwald M, Mansilla MA, Leslie E, Bullard S, Lidral AC, Moreno LM, Menezes R, Vieira AR, Petrin A, Wilcox AJ, Lie RT, Jabs EW, Wu-Chou YH, Chen PK, Wang H, Ye XQ, Huang SZ, Yeow V, Chong SS, Jee SH, Shi B, Christensen K, Melbye M, Doheny KF, Pugh EW, Ling H, Castilla EE, Czeizel AE, Ma LA, Field LL, Brody L, Pangilinan F, Mills JL, Molloy AM, Kirke PN, Scott JM, Arcos-Burgos M, Scott AF: A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near mafb and abca4. Nat Genet 2010;42:727-727.
36 Devlin B, Roeder K: Genomic control for association studies. Biometrics 1999;55:997-1004.
37 Plagnol V, Cooper JD, Todd JA, Clayton DG: A method to address differential bias in genotyping in large-scale association studies. Plos Genet 2007;3:e74.