Biostatistics (2012), **13**, 2, *pp.* 228–240
doi:10.1093/biostatistics/kxr045
Advance Access publication on December 8, 2011

Family-based association tests using genotype data with uncertainty

ZHAOXIA YU*
Department of Statistics, University of California, Irvine, CA 92697, USA
yu.zhaoxia@uci.edu

SUMMARY

Family-based association studies have been widely used to identify association between diseases and genetic markers. It is known that genotyping uncertainty is inherent in both directly genotyped or sequenced DNA variations and imputed data in silico. The uncertainty can lead to genotyping errors and missingness and can negatively impact the power and Type I error rates of family-based association studies even if the uncertainty is independent of disease status. Compared with studies using unrelated subjects, there are very few methods that address the issue of genotyping uncertainty for family-based designs. The limited attempts have mostly been made to correct the bias caused by genotyping errors. Without properly addressing the issue, the conventional testing strategy, i.e. family-based association tests using called genotypes, can yield invalid statistical inferences. Here, we propose a new test to address the challenges in analyzing case-parents data by using calls with high accuracy and modeling genotype-specific call rates. Our simulations show that compared with the conventional strategy and an alternative test, our new test has an improved performance in the presence of substantial uncertainty and has a similar performance when the uncertainty level is low. We also demonstrate the advantages of our new method by applying it to imputed markers from a genome-wide case-parents association study.

Keywords: Case-parents design; Family-based association tests; Genotype-specific missingness; Genotyping uncertainty; Imputed genotypes.

1. INTRODUCTION

Genotyping uncertainty is inherent in both directly genotyped/sequenced DNA variations and imputed data *in silico*. In directly genotyped data, the fully automated clustering algorithms adopted by current high-throughput technologies are unavoidable to errors and missingness in assigning genotypes when the clouds of fluorescence signals are not perfectly separated or when a rare cluster has few data points. Analyzing genotype data with uncertainty is also increasingly encountered with the growing popularity of testing association using low-coverage sequencing data (Li *and others*, 2011) and imputed genotypes (Browning and Browning, 2009; Li *and others*, 2010; Lin *and others*, 2008; Marchini *and others*, 2007; Nicolae, 2006; Servin and Stephens, 2007; Zaitlen *and others*, 2007).

Genotyping uncertainty causes both genotyping errors and missingness. The consequence of genotyping errors on genetic association studies has been extensively studied (Pompanon *and others*, 2005).

*To whom correspondence should be addressed.

Random genotyping errors, i.e. errors not relying on disease status, may reduce the power (Gordon and Ott, 2001) but are unlikely to affect the false-positive rates of case–control studies. For family-based studies, however, genotyping errors not only affect the power but also inflate the false-positive rate (Gordon *and others*, 2001; Heath, 1998; Mitchell *and others*, 2003). As a result, different approaches have been proposed to correct the inflation due to genotyping errors for the case-parents design (Cheng and Chen, 2007; Gordon *and others*, 2001; Morris and Kaplan, 2004) and for general pedigrees (Gordon *and others*, 2004).

Genotyping errors can be prevented by using a stringent cutoff in genotype calling and it is believed that the impact of genotyping errors on different genetic studies is unlikely to be an important factor with the improvement of genotyping technologies (Laird and Lange, 2006). However, genotype-specific missingness cannot be eliminated by using a stringent cutoff in genotype calling. Despite the high genotyping accuracy and the overall high call rate of currently used genotyping and sequencing technologies, genotype-specific missingness still exists (Fu *and others*, 2009; Illumina, 2010). Compared with the impact of genotyping errors that of genotype-specific missingness is much less studied. For family-based association studies, there are only a few published articles related to the problem of genotype-specific missingness. (Hirschhorn and Daly, 2005) found that genotype-specific call rates can cause false positives in the case-parents design, and the impact is the largest when the minor allele frequency (MAF) is small enough to avoid Hardy–Weinberg disequilibrium (HWE) but large enough to cause false positives. Other relevant work includes methods that aim to reduce the bias caused by informative missingness of parental data (Allen *and others*, 2003; Chen, 2004). However, the missingness they consider is caused by the unavailability of parental DNA samples.

Because the bias caused by genotyping errors has been well studied, here we examine the bias caused by genotype-specific missingness in family-based association studies. Consider a single nucleotide polymorphism (SNP) with alleles A and B. For simplicity, we assume that case-parents trios are sampled from a random mating population. (Curtis and Sham, 1995) suggested that, to avoid bias, trios with one missing parent should be excluded from the calculation of the transmission disequilibrium test (TDT) (Spielman *and others*, 1993) statistic. Let b denote the number of heterozygous parents who transmitted allele A to his/her offspring; let c denote the number of heterozygous parents who transmitted allele B to his/her offspring. The TDT, which is a McNemar test, is defined as $(b-c)/\sqrt{b+c}$. Among the 3 genotypes of a SNP, the rare homozygous genotype is usually more difficult to be called because it has fewer data points, and the common homozygous genotype is usually easier to be called because it has more data points. Under these conditions, $(b-c)$ has a positive expectation when $p > 1/2$ and a negative expectation when $p < 1/2$, where p is the frequency of allele A. Full details of the direction of bias are in Section A of the supplementary material available at *Biostatistics* online. This implies that the unequal missingness leads to overtransmission of the common allele. (Mitchell *and others*, 2003) found that genotyping errors also lead to overtransmission of the common allele. This suggests that the bias reported by (Mitchell *and others*, 2003) is caused probably not only by genotyping errors but also by genotype-specific missingness.

The bias caused by genotype-specific missingness and genotyping errors demonstrates that it is of great importance to develop association testing strategies that can take into account uncertainty in genotyped, sequenced, or imputed genetic data for family-based association studies. Methods to incorporate genotyping uncertainty have been proposed and studied for samples with unrelated subjects (Allen *and others*, 2010; Guan and Stephens, 2008; Kutalik *and others*, 2011; Lin *and others*, 2008; Marchini *and others*, 2007). However, there are few family-based association methods that systematically and efficiently take uncertainty into consideration in imputed data. One exception is (Chen and Abecasis, 2007), which deals with the situation that only a subset of individuals in each pedigree was genotyped in a high resolution. As both genotyping error and genotype-specific missingness caused by genotyping uncertainty can inflate the false-positive rates of family-based association tests (FBATs), a strategy that can simultaneously address the 2 difficulties would be desirable. In this article, we discuss strategies to analyze uncertain genotype data for family-based studies and propose a new FBAT that incorporates genotyping uncertainty. Using simulated data, we demonstrate that our new method improves Type I error rates and power in the presence of substantial uncertainty. We also demonstrate the advantage of our new method by applying it to imputed markers from a genome-wide association study using case-parents trios. The software that implements our new method is available upon request.

2. Methods

In this section, we first briefly describe the original FBAT for genotype data without uncertainty. We will then provide testing strategies that can incorporate genotyping uncertainty.

2.1 *The original FBAT*

Suppose there are n_i assayed members in the ith family and we use $G_i = (G_{i1}, \ldots, G_{in_i})$ to denote the genotype vector of the assayed members, with $i = 1, \ldots, n$. Let T_{ij} be the trait value, and $X_{ij}(G_i)$ be the count of the risk alleles of the jth nonfounder in the ith family, with $j = 1, \ldots, n_i'$. Here, we assume that the nonfounders are a subset of the assayed members in a family. For example, for the ith case-parents trio, the number of assayed members n_i is 3 and the number of nonfounders n_i' is 1. Here, the genotype vector G_i is required to be free from Mendelian errors, and we call such genotype vectors as compatible genotypes. To conduct the original FBAT (Horvath *and others*, 2001; Rabinowitz and Laird, 2000), we first compute the score U and the variance V for each family, respectively:

$$U_{0,i}(G_i) = \sum_{j=1}^{n_i} (X_{ij}(G_i) - E_0(X_{ij}(G_i))),$$

$$V_{0,i}(G) = \mathrm{Var}_0\left(\sum_{j=1}^{n_i} T_{ij} X_{ij}(G_i)\right),$$

where $E_0(\cdot)$ and $\mathrm{Var}_0(\cdot)$ denote the expectation and variance, respectively, under the null hypothesis of no association between the testing marker and the trait. The mathematical formula of the null expectation of the score can be found in (Horvath *and others*, 2001). The formula of the null variance depends on whether association is tested in the absence or presence of linkage (Horvath *and others*, 2001; Lazzeroni and Lange, 1998; Martin *and others*, 2000; Rabinowitz and Laird, 2000). When testing association in the absence of linkage, its formula for general nuclear families was given by (Horvath *and others*, 2001). To test association in the presence of linkage, one can use the empirical variance (Lake *and others*, 2000), i.e. $V_{0,i}(G_i) = U_{0,i}^2(G_i)$. Once the scores and variances are computed, the test statistic can be computed as follows:

$$\mathrm{FBAT} = \frac{\sum_{i=1}^{n} U_{0,i}(G_i)}{\sqrt{\sum_{i=1}^{n} V_{0,i}(G_i)}}.$$

Under the null hypothesis of no genetic effects, the test statistic follows a standard normal distribution approximately. Because we define $X_{ij}(G_i)$ as the count of the risk alleles of the jth nonfounder in the ith family, our tests assume log-additive genetic effects. Other genetic effects can be tested by changing the definition of $X_{ij}(G_i)$.

2.2 *The FBAT that takes uncertainty into account by weighting (FBAT$_{weight}$)*

The first method we examine is to take uncertainty into account by weighting. The idea of incorporating weights into likelihood has been proposed for testing imputed markers using unrelated subjects (Guan and Stephens, 2008; Kutalik *and others*, 2011; Marchini and Howie, 2010). In these methods, quantifying uncertainty, such as computing genotype probabilities, is first conducted, then these probabilities are treated as known values in the association testing stage using weighted likelihood. Here, we adopt the idea of weighting to studies using families. In the presence of genotyping uncertainty, such as imputed genotypes, genotypes are usually not directly observed. Instead, it is the individual genotype probabilities that are available. Let $P(G_{ij}|S_{ij})$ be the genotype probability for the jth assayed member of the ith family, where S_{ij} denote the observed signal intensities in genotype data, read counts in sequence data, or the information that allows to impute an ungenotyped SNP such as genotyped SNPs and the linkage disequilibrium pattern of an independent panel of samples. These individual genotype probabilities are provided by many genotype calling packages (Bravo and Irizarry, 2010; Browning and Yu, 2009; Carvalho *and others*, 2010; Teo *and others*, 2007; WTCCC Consortium, 2007) or packages that impute ungenotyped SNPs (Browning and Browning, 2009; Li *and others*, 2010; Marchini *and others*, 2007). For a compatible genotype of the ith family $G_i = (G_{i1}, \ldots, G_{in_i})$ with the corresponding $S_i = (S_{i1}, \ldots, S_{in_i})$, the genotype probability $P(G_i|S_i)$ equals the product of individual genotype probabilities, subject to standardization by a constant:

$$P(G_i|S_i) = \prod_{j=1}^{n_i} P(G_{ij}|S_{ij}) \Bigg/ \sum_{\text{compatible } G_i^*} \prod_{j=1}^{n_i} P(G_{ij}^*|S_{ij}).$$

As derived in Section B of the supplementary material available at *Biostatistics* online, using weighted likelihood for the ith family leads to the following score and variance:

$$U_{0,i}(S_i) = \sum_{G_i} U_{0,i}(G_i)P(G_i|S_i),$$

$$V_{0,i}(S_i) = \sum_{G_i} V_{0,i}(G_i)P(G_i|S_i) - \sum_{G_i} U_{0,i}^2(G_i)P(G_i|S_i) + \left(\sum_{G_i} U_{0,i}(G_i)P(G_i|S_i)\right)^2. \quad (2.1)$$

We denote the FBAT test based on the above score and variance as FBAT$_{\text{weight}}$, which uses the test statistic FBAT$_{\text{weight}}$ $= \sum_{i=1}^{n} U_{0,i}(S_i) \Big/ \sqrt{\sum_{i=1}^{n} V_{0,i}(S_i)}$. When the null hypothesis is true, assuming that the weights are known and the uncertainty level is independent of genotypes, this test statistic is expected to follow the standard normal distribution asymptotically.

2.3 *A likelihood ratio test that assumes genotype-specific call rates (FBAT$_{LRT}$)*

Motivated by the fact that both genotype errors and genotype-specific call rates can cause bias, we propose a likelihood ratio test that uses genotype-specific call rates as nuisance parameters. At a single SNP, estimating genotype-specific call rates is an unidentifiable problem for unrelated subjects. Fortunately, with the familial structure and the resulting constraints on the genotypes of family data, genotype-specific call rates can be incorporated into likelihood function. Here, we present the method applicable to the case-parents design.

For data with uncertainty, we first use a cutoff value of 0.9 to make genotype calls. This step protects against large genotyping error rates of calls when data quality is not high, such as when imputed genotypes are used. Let g_i and G_i denote the observed and the underlying true genotype vectors for the ith trio, respectively. Note that g_i can contain up to 3 missing genotypes. To follow the FBAT framework, we assume the log-additive genetic effects. Let BB genotype be the baseline genotype, and β be the log of the relative risk (Schaid and Sommer, 1993) of the AB genotype. The log-additive assumption implies that

$$e^{\beta} = \frac{P(T_{ij} = 1|G_{ij} = \mathrm{AA})}{P(T_{ij} = 1|G_{ij} = \mathrm{AB})} = \frac{P(T_{ij} = 1|G_{ij} = \mathrm{AB})}{P(T_{ij} = 1|G_{ij} = \mathrm{BB})}.$$

Let $c = (c_{\mathrm{AA}}, c_{\mathrm{AB}}, c_{\mathrm{BB}})$ denote the vector of call rates, and $\mu = (\mu_1, \ldots, \mu_9)$ denote the 9 mating type probabilities. To make the presentation of the method easier, we let $\theta = (\theta_1, \ldots, \theta_{15})$ be the frequencies of the 15 possible trio types (see Table S2 of the supplementary material available at *Biostatistics* online). As shown in Table S2 of the supplementary material available at *Biostatistics* online, θ is not an additional vector of parameters, as it is completely determined by μ and β. Assuming no genotyping error, then the complete- and observed-data likelihood functions contributed by the ith trio are, respectively,

$$L_{i,\mathrm{comp}}(c, \mu, \beta) = P(g_i, G_i|c, \mu, \beta) = P(G_i|\mu, \beta)P(g_i|G_i, c),$$

$$L_i(c, \mu, \beta) = P(g_i|c, \mu, \beta) = \sum_{G_i} L_{i,\mathrm{comp}}(c, \mu, \beta),$$

where

$$P(G_i|\mu, \beta) = \prod_{k=1}^{15} (\theta_k)^{I(G_i \text{ belongs to trio type k})},$$

$$P(g_i|G_i, c) = \prod_{j=1}^{3} \left(c_{G_{ij}}\right)^{I(g_{ij} \text{ observed})} \left(1 - c_{G_{ij}}\right)^{I(g_{ij} \text{ unobserved})},$$

and $I(\cdot)$ is the indicator function. The EM algorithm (Dempster *and others*, 1977) is used to estimate parameters. We derive the expectation and maximization steps. The details of the steps can be found in Section C of the supplementary material available at *Biostatistics* online.

Let Λ denote the likelihood ratio statistic, defined as the ratio of the maximized likelihood under the null hypothesis to that obtained under the alternative. To examine the direction of bias, we use $\mathrm{FBAT}_{\mathrm{LRT}} = \mathrm{sign}(\hat{\beta})\sqrt{-2\log\Lambda}$, where $\hat{\beta}$ is the maximum likelihood estimate of β. Under the null hypothesis, the test statistic follows the standard normal distribution asymptotically, as -2logΛ follows the chi-square distribution with 1 degree of freedom asymptotically. The method does not make assumptions of genotype probabilities at population level, such as HWE. Thus, it is fully robust against population stratification.

3. Simulations and results

3.1 *Methods of simulations*

To compare the performance of different methods, we simulate case-parents genotype data under the null hypothesis H_0: $\beta = 0$ or under an alternative hypothesis H_1: $\beta = \ln(1.2)$. The genotypes of case-parents trios are simulated from a random mating population with the frequency of the allele A varying from 0.1 to 0.9. To mimic the situation of genotyping uncertainty, we consider normally distributed 1-dimension signal data with fixed cluster means: 0 for AA genotype, 1 for AB genotype, and 2 for BB genotype. We assume that the 3 genotype clusters have the same variance and use $\sigma^2 = 0.03$ and $\sigma^2 = 0.05$ to reflect different levels of genotyping uncertainty. Using 1-dimension simulation is justified by the fact that signal data from commonly used platforms can be transformed into 1-dimension data. The assumption of equal variance across genotypes may be violated in practice; however, when that happens, more genotyping errors and/or genotype-specific missingness are often expected. Once signal data are generated, we use the EM algorithm (Dempster *and others*, 1977) to fit a 3-component normal mixture model. After the EM algorithm converges, the posterior probabilities computed in the E-step are used as genotype data with uncertainty to examine the performance of the difference methods.

To evaluate the performance of $\text{FBAT}_{\text{weight}}$ and FBAT_{LRT}, we compare them with $\text{FBAT}_{\text{call}}$, which denotes the association test based upon called genotypes. In $\text{FBAT}_{\text{call}}$, each signal data point is either assigned to a genotype call or no-call, depending on whether the largest posterior probability is greater than a prespecified cutoff (here, we use 0.9). Then, families with Mendelian errors or at least one no-call are removed and the remaining “cleaned” data are used to perform FBATs. We also provide the “gold standard” by conducting association test using the true genotypes ($\text{FBAT}_{\text{true}}$). The computation of the Type I error rate or power is based on 1000 simulations, 1000 trios sampled from a random mating population, and the nominal p value cutoff of 0.05.

3.2 *Simulation results*

We first examine call rates and accuracy under the 2 different levels of genotyping uncertainty. The results for data simulated under the null hypothesis are summarized in Tables 1 and 2. With the high threshold (0.9) we use to make genotype calls, the genotyping accuracy is satisfying: >99.9% and 99.5% for $\sigma^2 = 0.03$ and $\sigma^2 = 0.05$, respectively, for all allele frequencies. The overall call rates are >99% and 94% for $\sigma^2 = 0.03$ and $\sigma^2 = 0.05$, respectively, for all allele frequencies. Despite the high call rates and genotyping accuracy, notable difference exists among genotype-specific call rates. For example, when allele frequency p is 0.1 and $\sigma^2 = 0.05$, the call rate of BB genotype is 98.23%, whereas that of AA genotype is only 89.25%. As we have shown in Section 1, this genotype-specific missingness can lead to bias in FBATs.

In the respect of Type I error rates, at a lower level of genotyping uncertainty, the Type I error rates of the 3 different tests are all close to those of the gold standard as illustrated in Figure 1. However, at a higher level of genotyping uncertainty, the Type I error rates of the 3 different tests are quite different: FBAT_{LRT} is close to the gold standard $\text{FBAT}_{\text{true}}$ at all allele frequencies with the maximum difference being 0.01.

Table 1. *Accuracy and call rates (%) for simulated data when* $\sigma^2 = 0.03$, *for different allele frequencies*

Allele frequency	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
Accuracy	99.95	99.94	99.93	99.92	99.92	99.92	99.93	99.94	99.95
Call rate	99.54	99.38	99.30	99.24	99.22	99.24	99.29	99.39	99.54
Call rate	(BB) 99.73	99.60	99.50	99.37	99.24	99.12	98.93	98.70	98.33
by genotype	(AB) 98.76	99.04	99.15	99.19	99.21	99.19	99.14	99.03	98.75
	(AA) 98.29	98.65	98.94	99.08	99.22	99.36	99.49	99.61	99.73

Table 2. *Accuracy and call rates (%) for simulated data when* $\sigma^2 = 0.05$, *for different allele frequencies*

Allele frequency	$p = 0.1$	$p = 0.2$	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
Accuracy	99.72	99.62	99.56	99.52	99.51	99.52	99.56	99.61	99.71
Call rate	96.99	95.87	95.25	94.91	94.78	94.91	95.24	95.89	97.00
Call rate	(BB) 98.23	97.33	96.53	95.82	94.93	94.13	92.94	91.41	89.25
by genotype	(AB) 91.87	93.50	94.24	94.52	94.61	94.49	94.22	93.55	91.86
	(AA) 89.25	91.37	93.00	94.03	94.98	95.81	96.54	97.34	98.24

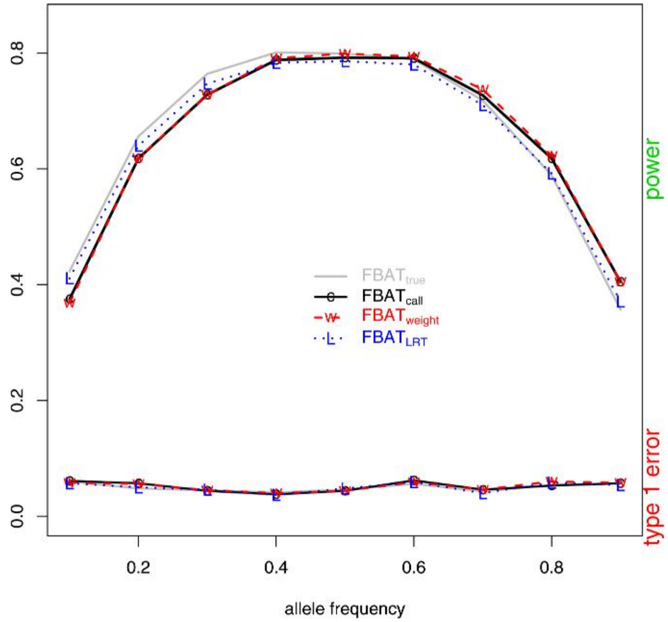

Fig. 1. The power (top) and Type I error rates (bottom) of the four tests when $\sigma^2 = 0.03$ for different frequencies of the risk allele. FBAT$_{\text{true}}$, TDT using true genotypes; FBAT$_{\text{call}}$, TDT using complete trios; FBAT$_{\text{weight}}$, TDT using the weighting strategy; and FBAT$_{\text{LRT}}$, the likelihood ratio test that assumes genotype-specific call rates.

In contrast, the Type I error rates of FBAT$_{\text{call}}$ and FBAT$_{\text{weight}}$ are >0.1 for small MAF as illustrated in Figure 2. FBAT$_{\text{call}}$ and FBAT$_{\text{weight}}$ tend to be negative when the frequency of the risk allele is <0.5 and positive when the frequency is >0.5. Box plots of the test statistics under the null hypothesis $\beta = 0$ (Figure S1 of the supplementary material available at *Biostatistics* online**)** show that both FBAT$_{\text{call}}$ and FBAT$_{\text{weight}}$ are biased toward overtransmission of the common allele. All these results indicate that the common allele appears to be overtransmitted using FBAT$_{\text{call}}$ or FBAT$_{\text{weight}}$. Mitchell et al. (Mitchell *and others*, 2003) reported that undetected genotyping errors also tend to predict the common allele to be overtransmitted. Thus, genotyping errors and genotype-specific missingness can not only cause transmission distortions but also pull transmission distortion toward the same direction.

In the respect of statistical power, FBAT$_{\text{LRT}}$ also has better performance than the other 2 tests. It has closer power to the gold standard FBAT$_{\text{true}}$ than both FBAT$_{\text{call}}$ and FBAT$_{\text{weight}}$ in most situations. The difference is especially clear when the frequency of the risk allele is small and the genotyping uncertainty is high. For example, when $p = 0.1$ and $\sigma^2 = 0.05$ (Figure 2), the power of FBAT$_{\text{true}}$, FBAT$_{\text{LRT}}$, FBAT$_{\text{weight}}$, and FBAT$_{\text{call}}$ is 0.42, 0.34, 0.13, and 0.15, respectively. With the decrease of genotyping uncertainty, the power of the 4 methods tends to be closer (Figure 1) as expected. However, the performance of FBAT$_{\text{LRT}}$ is still better than FBAT$_{\text{call}}$ and FBAT$_{\text{weight}}$ in most situations. Note that when the risk allele is the common allele, FBAT$_{\text{call}}$ and FBAT$_{\text{weight}}$ seem to have higher power than FBAT$_{\text{true}}$. However, the seemingly “higher” power in FBAT$_{\text{call}}$ and FBAT$_{\text{weight}}$ should not be interpreted as greater efficiency than the gold standard. Rather, the results once again confirm that the resulted bias is toward predicting the common allele to be overtransmitted.

One may argue that these problematic SNPs, i.e. SNPs with high genotyping uncertainty, can be filtered out by testing HWE on the called genotypes. However, we found that the inflation in false positives

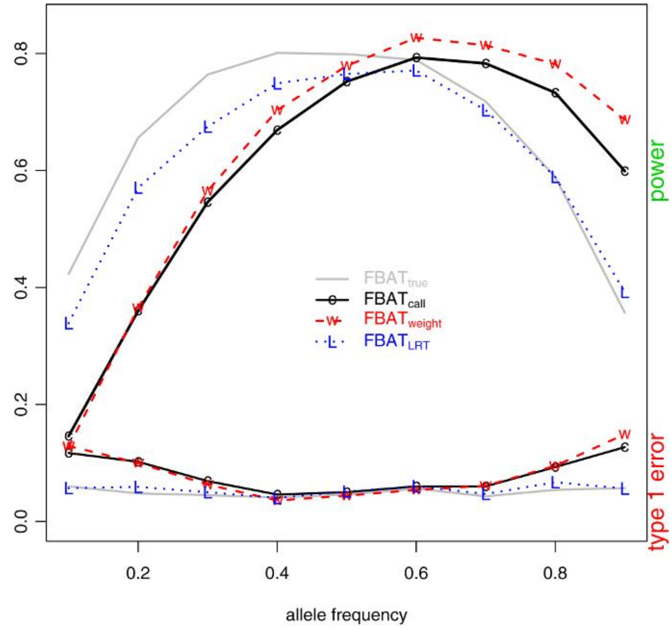

Fig. 2. The power (top) and Type I error rates (bottom) of the four tests when $\sigma^2 = 0.05$ for different frequencies of the risk allele. FBAT$_{\text{true}}$, TDT using true genotypes; FBAT$_{\text{call}}$, TDT using complete trios; FBAT$_{\text{weight}}$, TDT using the weighting strategy; and FBAT$_{\text{LRT}}$, the likelihood ratio test that assumes genotype-specific call rates.

and the incorrect power still persist even after filtering out those data sets that show deviation from HWE at level 0.05. In addition, filtering out SNPs can lead to loss of a substantial proportion of SNPs.

Since the methods we examine here do not make assumptions with regard to population genotype frequencies, such as HWE, they are all robust against population stratification. To confirm this, we simulated trios from 2 random mating subpopulations. The results (presented in Section D of the supplementary material available at *Biostatistics* online) demonstrate that they are indeed robust against population stratification.

4. Application to a real data set

We applied FBAT$_{\text{LRT}}$, FBAT$_{\text{weight}}$, and FBAT$_{\text{call}}$ to data collected by the International Consortium to Identify Genes and Interactions Controlling Oral Clefts. The genome-wide analysis of assayed SNPs was reported in (Beaty *and others*, 2010). Besides genotype data of assayed SNPs, imputed genotypes obtained using BEAGLE (Browning and Browning, 2009) are also provided by the database of Genotypes and Phenotypes (dbGaP). According to “OralClefts_imputation_report_final.pdf” downloaded from the dbGaP, phased HapMap genotype data from Phase III release 2, with base pair positions of SNPs based on NCBI build 36, were used for the imputation. The report also states that relatedness among subjects was ignored in the imputation. We used the 47 475 imputed markers on chromosome 8 from 856 case-parents trios of European descendants. Because the imputation was based on only 234 phased haplotypes in the HapMap Phase III, the imputation accuracy for SNPs with small MAF is likely to be low; thus, we only considered imputed SNPs with $\geqslant$5% MAF. We also exclude SNPs showing large deviation from HWE (chi-square statistic $>$10), with more than 1 Mendelian error or having $<$90% call rate. The remaining 23 524 SNPs are analyzed. Here, the MAF and other characteristics, such as HWE chi-square statistic, number of Mendelian errors, and call rate of the imputed SNPs are computed using called genotypes based on the cutoff 0.9.

To show the difference of the 3 tests in false positives, we present results from SNPs with call rates between 0.9 and 0.99 as SNPs with call rates below 0.9 are not reliable and SNPs with call rates above 0.99 do not show much uncertainty. One useful measure of false positives in analyzing real data is the inflation factor introduced by (Devlin and Roeder, 1999). However, inflation factors are based on squared statistics and they do not show the direction of bias. To investigate the direction of bias, we compute the FBAT test statistics by treating the minor allele as the risk allele and examine the medians of the 3 tests. In our way of computing test statistics, a negative value implies predicting the common allele to be the risk allele. The results of MAF-stratified medians are presented in Figure 3, where the median for a specific MAF category is the median of the test statistics of all SNPs that fall into this category. All the medians of $\text{FBAT}_{\text{weight}}$ are negative and the smaller the MAF, the larger the bias. Although $\text{FBAT}_{\text{call}}$ has less bias than $\text{FBAT}_{\text{weight}}$, most of its medians are still negative. In contrast, FBAT_{LRT} has both negative and positive medians and most of them are close to 0. These results not only agree with our theoretical prediction that genotyping uncertainty can lead to overtransmission of the common allele but also imply that our new approach FBAT_{LRT} can reduce bias caused by genotyping uncertainty.

Next, we examine the 3 most significant SNPs among the imputed SNPs. The results in Table 3 indicate that the 3 tests provide similar statistic values. The small difference is presumably due to low uncertainty level at the 3 SNPs as reflected by the large call rates. Nevertheless, compared with the 2 most significant SNPs that are directly assayed, i.e. rs1519847 and rs12542837 (with the same TDT statistic 5.83), the 3 SNPs in Table 3 have greater values. Note that the top 2 assayed and the top 3 imputed SNPs are within a 50 kb region at chromosome 8q24, indicating that they are probably in the same functional region. These results reveal the efficiency of imputing genetic markers and incorporating genotyping uncertainty into FBATs.

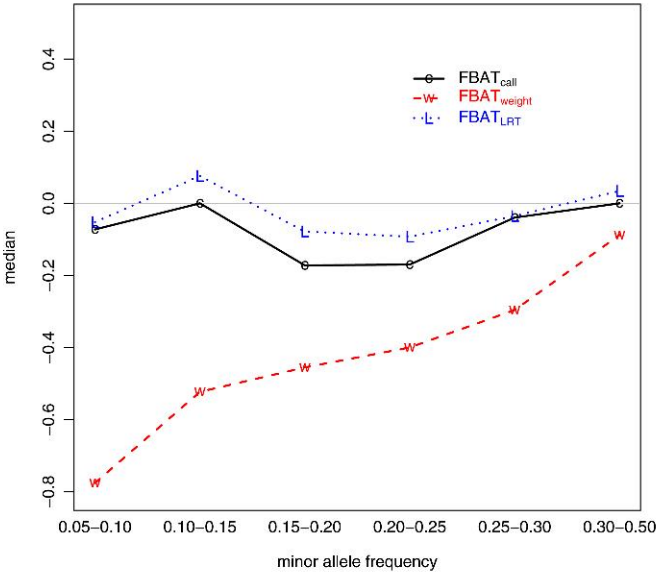

Fig. 3. MAF-stratified medians of the three tests on the imputed SNPs from the Oral Clefts project. $\text{FBAT}_{\text{call}}$, TDT using complete trios; $\text{FBAT}_{\text{wight}}$, TDT using the weighting strategy; and FBAT_{LRT}, the likelihood ratio test that assumes genotype-specific call rates.

Table 3. *The top three imputed SNPs*

Marker	MAF	Position (bp)	Call rate (%)	$\text{FBAT}_{\text{call}}$	$\text{FBAT}_{\text{weight}}$	FBAT_{LRT}
rs997310	0.31	129989462	99.96	6.59	6.59	6.69
rs17241908	0.24	130014058	99.84	6.78	6.80	6.78
rs17242358	0.24	130034055	98.60	6.77	6.86	6.85

5. Discussion

In this article, we investigated 3 FBATs for uncertain genotype data. Among them, $\text{FBAT}_{\text{call}}$ is the conventional method that uses called genotypes; $\text{FBAT}_{\text{weight}}$ is a weighted test with the idea of weighting borrowed from association tests for studies using unrelated subjects. It is surprising to find that these ideas, which have been widely used and shown to perform well in studies using unrelated subjects, have unsatisfactory performance for the case-parents design. $\text{FBAT}_{\text{call}}$ can be biased by genotype-specific call rates even when a stringent threshold is used to rule out most genotyping errors. $\text{FBAT}_{\text{weight}}$ can be biased by both genotype-specific uncertainty levels and genotyping errors that are Mendelian consistent and therefore cannot be eliminated by weighting. The novelty of our new test, FBAT_{LRT}, is that by using genotype calls with a high quality and incorporating genotype-specific call rates, we can simultaneously correct bias due to both genotyping errors and genotype-specific missingness. Simulation results show that our new method FBAT_{LRT} reduces the false-positive rates of the methods using called genotypes ($\text{FBAT}_{\text{call}}$) and the weighting approach ($\text{FBAT}_{\text{weight}}$).

The proposed method may be extended in several respects. First, because FBAT_{LRT} concentrates on the case-parents design, it would be helpful to develop methods that are applicable to families of arbitrary structures and other types of traits. One difficulty in extending the case-parents design to families with multiple children is that the number of family types, which are defined as combinations of genotypes and phenotypes, can be very large; as a result, assumptions such as HWE might need to be used (Dudbridge, 2008), which can lead to inflated false positives in the presence of population stratification. Second, FBAT_{LRT} can be extended to analyze other types of traits, such as quantitative traits. To do so, genotype risks need to be replaced by density functions for a quantitative trait. For families of arbitrary structures and general types of traits, we are currently developing methods that can efficiently correct the uncertainty-caused bias with minimum sacrifice of robustness.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank Dr Wesley Johnson for helpful discussion and Dr Brian Browning for the suggestion of using imputed SNPs of the Oral Clefts project. Datasets used for the analyses described in this article 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. *Conflict of Interest:* None declared.

Funding

National Institute of Health (grant R01-HG004960).

References

Allen, A. S., Rathouz, P. J. and Satten, G. A. (2003). Informative missingness in genetic association studies: case-parent designs. *American Journal of Human Genetics* **72**, 671–680.

Allen, A. S., Satten, G. A., Bray, S. L., Dudbridge, F. and Epstein, M. P. (2010). Fast and robust association tests for untyped SNPs in case-control studies. *Human Heredity* **70**, 167–176.

Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Hetmanski, I. R. J. B., Liang, K. Y., Wu, T., Murray, T., Fallin, M. D., Redett, R. A., *and others* (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature Genetics* **42**, 727–727.

Bravo, H. C. and Irizarry, R. A. (2010). Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics* **66**, 665–674.

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics* **84**, 210–223.

Browning, B. L. and Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *American Journal of Human Genetics* **85**, 847–861.

Carvalho, B. S., Louis, T. A. and Irizarry, R. A. (2010). Quantifying uncertainty in genotype calls. *Bioinformatics* **26**, 242–249.

Chen, W. M. and Abecasis, G. R. (2007). Family-based association tests for genomewide association scans. *American Journal of Human Genetics* **81**, 913–926.

Chen, Y. H. (2004). New approach to association testing in case-parent designs under informative parental missingness. *Genetic Epidemiology* **27**, 131–140.

Cheng, K. F. and Chen, J. H. (2007). A simple and robust TDT-type test against genotyping error with error rates varying across families. *Human Heredity* **64**, 114–122.

Curtis, D. and Sham, P. C. (1995). A note on the application of the transmission disequilibrium test when a parent is missing. *American Journal of Human Genetics* **56**, 811–812.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39**, 1–38.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.

Dudbridge, F. (2008). Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Human Heredity* **66**, 87–98.

Fu, W. Q., Wang, Y., Wang, Y., Li, R., Lin, R. and Jin, L. (2009). Missing call bias in high-throughput genotyping. *BMC Genomics* **10**, 106.

Gordon, D., Haynes, C., Johnnidis, C., Patel, S. B., Bowcock, A. M. and Ott, J. (2004). A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *European Journal of Human Genetics* **12**, 752–761.

Gordon, D., Heath, S. C., Liu, X. and Ott, J. (2001). A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *American Journal of Human Genetics* **69**, 371–380.

Gordon, D. and Ott, J. (2001). Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pacific Symposium on Biocomputing* **6**, 18–29.

Guan, Y. T. and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genetics* **4**, e1000279.

HEATH, S. (1998). A bias in TDT due to undetected genotyping errors. *American Journal of Human Genetics Suppl* **63**, A292.

HIRSCHHORN, J. N. AND DALY, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108.

HORVATH, S., XU, X. AND LAIRD, N. M. (2001). The family based association test method: strategies for studying general genotype-phenotype associations. *European Journal of Human Genetics* **9**, 301–306.

ILLUMINA. (2010). Genotyping rare variants. *Illumina Technical Note: Data Analysis*. San Diego, CA: Illumina.

KUTALIK, Z., JOHNSON, T., BOCHUD, M., MOOSER, V., VOLLENWEIDER, P., WAEBER, G., WATERWORTH, D., BECKMANN, J. S. AND BERGMANN, S. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics* **12**, 1–17.

LAIRD, N. M. AND LANGE, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* **7**, 385–394.

LAKE, S. L., BLACKER, D. AND LAIRD, N. M. (2000). Family-based tests of association in the presence of linkage. *American Journal of Human Genetics* **67**, 1515–1525.

LAZZERONI, L. C. AND LANGE, K. (1998). A conditional inference framework for extending the transmission/disequilibrium test. *Human Heredity* **48**, 67–81.

LI, Y., SIDORE, C., KANG, H. M., BOEHNKE, M. AND ABECASIS, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome Research* **21**, 940–951.

LI, Y., WILLER, C. J., DING, J., SCHEET, P. AND ABECASIS, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**, 816–834.

LIN, D. Y., HU, Y. AND HUANG, B. (2008). Simple and efficient analysis of disease association with missing genotype data. *American Journal of Human Genetics* **82**, 444–452.

MARCHINI, J. AND HOWIE, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**, 499–511.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. AND DONNELLY, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913.

MARTIN, E. R., GILBERT, J. R., LAI, E. H., RILEY, J., ROGALA, A. R., SLOTTERBECK, B. D., SIPE, C. A., GRUBBER, J. M., WARREN, L. L., CONNEALLY, P. M., *and others* (2000). Analysis of association at single nucleotide polymorphisms in the APOE region. *Genomics* **63**, 7–12.

MITCHELL, A. A., CUTLER, D. J. AND CHAKRAVARTI, A. (2003). Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *American Journal of Human Genetics* **72**, 598–610.

MORRIS, R. W. AND KAPLAN, N. L. (2004). Testing for association with a case-parents design in the presence of genotyping errors. *Genetic Epidemiology* **26**, 142–154.

NICOLAE, D. L. (2006). Testing untyped alleles (TUNA)—applications to genome-wide association studies. *Genetic Epidemiology* **30**, 718–727.

POMPANON, F., BONIN, A., BELLEMAIN, E. AND TABERLET, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* **6**, 847–859.

RABINOWITZ, D. AND LAIRD, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity* **50**, 211–223.

SCHAID, D. J. AND SOMMER, S. S. (1993). Genotype relative risks: methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics* **53**, 1114–1126.

SERVIN, B. AND STEPHENS, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3**, 1296–1308.

SPIELMAN, R. S., MCGINNIS, R. E. AND EWENS, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**, 506–516.

TEO, Y. Y., INOUYE, M., SMALL, K. S., GWILLIAM, R., DELOUKAS, P., KWIATKOWSKI, D. P. AND CLARK, T. G. (2007). A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741–2746.

WTCCC CONSORTIUM. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

ZAITLEN, N., KANG, H. M., ESKIN, E. AND HALPERIN, E. (2007). Leveraging the HapMap correlation structure in association studies. *American Journal of Human Genetics* **80**, 683–691.

[*Received June 10, 2011; revised November 7, 2011; accepted for publication November 8, 2011*]