International Journal of Biomedical Data Mining

Li et al., Biomedical Data Mining 2013, 3:1
http://dx.doi.org/10.4172/2090-4924.1000104

Research Article Open Access

Likelihood Ratio Test of Hardy-Weinberg Equilibrium Using Uncertain Genotypes for Sibship Data

Qiong Li, Hélène Massam and Xin Gao*

Department of Mathematics and Statistics, York University, Toronto, ON, M3J 1P3, Canada

Abstract

Testing for Hardy-Weinberg equilibrium of genotype frequencies is a crucial first step in the study of population genetics. In this paper, we develop an Expectation-Maximization algorithm to estimate the genotype frequencies for sibship data with genotype uncertainty. We also develop a likelihood ratio test of Hardy-Weinberg equilibrium for sibships with no parental genotypes available and with possible genotyping errors. Simulations show that our likelihood ratio test maintains valid control of the type I error rate and good statistical power. Finally, the likelihood ratio test is extended across strata when a sample is stratified by multiple ethnic populations with different genotype frequencies.

Keywords: Hardy-Weinberg Equilibrium; Genotyping errors; Dependent population

Introduction

The genetic composition of a population can be influenced by various factors such as migration, mutation, inbreeding, natural selection and gene flow. Investigations of these mechanisms through behavioral, ecological, breeding and other studies are usually preceded by testing of the given population for Hardy-Weinberg equilibrium (HWE) [1]. If the allele frequencies in a population with two alleles a_1 and a_2 at a locus are p_{a_1} and p_{a_2}, then testing HWE is to check whether the observed genotype frequency for genotype (a_1,a_2) equals $2p_{a_1}p_{a_2}$ and the observed genotype frequency for genotype (a_s,a_s), s=1,2, equals $p_{a_s}^2$. Departures from HWE at particular markers may indicate problems which include genotyping errors, sample mishandling, inconsistencies within family pedigrees and population structure [2]. Therefore, testing HWE of marker genotype frequencies has been used for screening genetic markers in association studies [3], and for fine-scale localization of a disease-susceptibility locus [4]. Some statistical tests to detect and measure deviation from HWE have been derived in the twentieth century (see for examples [5-7]). Wigginton et al. [8] provided a general algorithm to perform an exact test of HWE in large or small samples. The works mentioned above were done under the assumptions of independence of individuals and no genotyping errors in the sample.

Bourgain et al. [9] presented a new test for HWE suitable for samples containing dependent individuals, when the genealogy of the population is available. They used quasi-likelihood which allows them to work with large pedigrees. However, their method can't accommodate genotyping errors. Genotyping errors mean that there is a probability of observing a_s when $a_{s'}$ is the true allele. It is well known that genotyping errors can have significant effects on linkage analysis: see Gordon and Ott [10] for effects on estimated REcombination fraction, Gordon et al. [11], Leal [12] and Cox and Kraft [13] for effects on the analysis of single nucleotide polymorphisms (SNP).

A likelihood ratio test (LRT) for testing HWE has been developed which takes into account potential disease-genotype association[14]. Yu et al. [15] presented another LRT using both case and control samples. Shriner [16] first offered approximate χ^2 and exact tests of HWE using uncertain genotypes. However, his method can't handle dependent individuals. In this paper, we offer a LRT of HWE for dependent individuals with possibility of genotyping errors. Our method is designed for sibship data which is used widely in genetic association studies [17-20] and gene mapping of complex quantitative traits [21]. We apply our method effectively for two generations with an arbitrary number of siblings. In principle, our method can be extended to arbitrary pedigree structures. Increasing the number of generations, however, increases significantly the computational burden. Bourgain et al. [9] can handle large pedigrees but no genotyping errors. Shriner [16] can handle genotyping errors but no dependent individuals. We can handle genotyping errors within sibship data.

In fact, our paper goes somewhat further: we accommodate genotype data with dependent individuals, genotyping errors and across strata. Considering genotype data sampled from several populations allowing for different marker allele frequencies, Haldane [5] first developed a HWE test for stratified data. Schaid et al. [22] proposed an exact stratified test for diallelic markers for independent individuals. However, their methods work under the assumptions of independent individuals and no genotyping errors.

Through our likelihood ratio method, we give a test for stratified sibship data with possible genotyping errors. Through simulation studies, we evaluate the performance of the proposed tests. The results show that they have the correct type I error rate and strong power. An R function is developed and it can be obtained at http://xingao.info.yorku.ca/.

Methods

Suppose N sibships are genotyped at a multiallelic marker. We assume the genotypes of their 2*N* parents are not available. Let $g_i=(g_{i1}, g_{i2})$ and $G_i=(G_{i1},G_{i2})$ represent the genotypes of the sibships and their parents in the i-th family, respectively. Let a_s,s=1,…,k, be alleles and

***Corresponding author:** Xin Gao, Department of Mathematics and Statistics, York University, Toronto, ON, M3J 1P3, Canada, Tel: 416-736-2100, ext: 66097; E-mail: xingao@mathstat.yorku.ca

Received August 22, 2013; **Accepted** November 18, 2013; **Published** November 26, 2013

Citation: Li Q, Massam H, Gao X (2013) Likelihood Ratio Test of Hardy-Weinberg Equilibrium Using Uncertain Genotypes for Sibship Data. Biomedical Data Mining 3: 104. doi: 10.4172/2090-4924.1000104

$\mathcal{A}=\{a_s a_{s'}; s,s'=1,\cdots,k\}$ be the set of genotypes. Let $p_{a_s}, s=1,\cdots,k$, be allele frequencies and q_m, $m\in A$, be genotype frequencies. We denote their estimators by $\hat{p}_{a_s}$, $s=1,\cdots,k$ and $\hat{q}_m, m\in\mathcal{A}$. To test HWE, we want to test

$$H_0: \ q_{a_s a_{s'}} = \begin{cases} 2p_{a_s}p_{a_{s'}}, & a_s \neq a_{s'}; \\ p_{a_s}^2, & otherwise. \end{cases}$$

First, let us consider the problems for sibling dependence but no genotyping errors. When both g and G are known, the complete likelihood function is

$$L(q;G,g)=\prod_{i=1}^{N}\prod_{(m,\ m')\in\mathcal{A}^2}[q_m q_{m'}P(g_i\mid G_i=(m,m'))]^{I(G_i=(m,m'))}, \quad (1)$$

where $(m,m')\in\mathcal{A}^2$ is a pair of genotypes and $I(G_i=(m,m'))$ is an indicator function, which is defined by $I(G_i=(m,m'))=\begin{cases}1, & \text{if the genotype } G_i \text{ is}(m,m');\\ 0 & \text{otherwise.}\end{cases}$. The observed genotypes g of siblings in each family constitute the observed data, while the unknown genotypes G of parents constitute the missing data. The Expectation-Maximization (EM) algorithm [23] will be used to estimate $q=(q_m, m\in\mathcal{A})$ with missing data.

E-step: The E-step uses the expectation with respect to G of the complete log likelihood as follows

$$\begin{aligned} Q(q|g,q^{(t)}) &= E\left[\ln L(q;G,g)\mid q^{(t)},g\right] \\ &= \sum_{i=1}^{N}\sum_{(m,m')\in\mathcal{A}^2} E\left[I(G_i=(m,m'))|q^{(t)},g_i\right][\ln q_m+\ln q_{m'}+\ln P(g_i\mid G_i=(m,m'))] \\ &= \sum_{i=1}^{N}\sum_{(m,m')\in\mathcal{A}^2}\omega_{i(m,m')}^{(t)}[\ln q_m+\ln q_{m'}+\ln P(g_i\mid G_i=(m,m'))], \end{aligned}$$

with

$$\begin{aligned} \omega_{i(m,m')}^{(t)} &= E\left[I(G_i=(m,m'))|q^{(t)},g_i\right] \\ &= \frac{q_m^{(t)}q_{m'}^{(t)}P(g_i\mid G_i=(m,m'))}{\sum_{(m,m')\in\mathcal{A}^2}q_m^{(t)}q_{m'}^{(t)}P(g_i\mid G_i=(m,m'))}, \end{aligned} \quad (2)$$

where i=1,…,N, and $q^{(t)}$ is the estimated parameter obtained after t iterations.

M-step: Given the expected probability $\omega_{i(m,m')}^{(t)}=P(G_i=(m,m')\mid q^{(t)},g_i)$, we maximize $Q(q\mid g,q^{(t)})$ to obtain the (t+1)- updated value of q. The parameters $q^{(t+1)}$ are determined according to the estimate of the variables $\omega_{i(m,m')}^{(t)}, i=1,\cdots,N, m\in\mathcal{A}$. That yields $q_m^{(t+1)}=W_m^{(t)}/\sum_{m\in\mathcal{A}}W_m^{(t)}, m\in\mathcal{A}$, where $W_m^{(t)}=\sum_{i=1}^{N}\left(2\omega_{i(m,m)}^{(t)}+\sum_{(m,m')\in\mathcal{A}^2,m\neq m'}\omega_{i(m,m')}^{(t)}\right)$.

The estimates are updated until convergence.

Next, let us allow for random genotyping errors in g by which we mean any miscoding of a sibling's correct marker genotype [24]. This model is realistic for automated SNP calling, in which a machine makes the allele calls without human intervention [11]. We assume the SNP loci have two alleles, a_1 and a_2. Genotying errors occur when allele a_1 is incorrectly coded as a_2 and vice versa. Let ε_1 be the probability of incorrectly coding a_1 as a_2, and ε_2 be the probability of incorrectly coding a_2 as a_1. The probability P (observed genotype | true genotype) is referred to as the penetrance of the genotype [25]. Given error rates ε_1 and ε_2, Table 1 presents the matrix of penetrances (Table 1).

The genotying errors will affect $P(q\mid G_i=(m,m'))$ in formula (2). Let $g_i^*, i=1,...,N$, denote the true genotype of siblings in family i which could be different from the observed g_i. Then, for $g_i^*\in\mathcal{A}^2$, i=1,…,N,

$$\begin{aligned} P\left(g_i|G_i=(m,m')\right) &= \sum_{g_i^*\in\mathcal{A}^2}P\left(g_i,g_i^*|G_i=(m,m')\right) \\ &= \sum_{g_i^*\in\mathcal{A}^2}P\left(g_i|g_i^*,G_i=(m,m')\right)P\left(g_i^*|G_i=(m,m')\right) \\ &= \sum_{g_i^*\in\mathcal{A}^2}P\left(g_i|g_i^*\right)P\left(g_i^*|G_i=(m,m')\right), \end{aligned}$$

where the last equation is due to the fact that g is independent of G given g^*. The probability $P(g_i\mid g_i^*)$ can be computed from Table 1. With this modified probability for $P(g_i\mid G_i=(m,m'))$, we repeat the E-step and M-step as described before to estimate q. So far, we have estimated q under no assumption of HWE. We can proceed in exactly the same way to estimate p_{a_s} and $p_{a_{s'}}$ under H_0. We replace $q_{a_s a_{s'}}$ by $2p_{a_s}p_{a_{s'}}$ when $s\neq s'$ and replace $q_{a_s a_s}$ by $p_{a_s}^2$ in our likelihood equation (1) above.

The likelihood ratio test statistic is $-2\ln(L_0/L_1)=-2(\ln L_0-\ln L_1)$, where

$$\ln L_1=\sum_{i=1}^{N}\ln\sum_{(m,m')\in\mathcal{A}^2}\left[\hat{q}_m\hat{q}_{m'}P\left(g_i|G_i=(m,m')\right)\right],$$

and

$$\ln L_0=\sum_{i=1}^{N}\ln\sum_{m=(a_1,a_2),m'=(a_1',a_2')}[\hat{q}_{a_1a_2}\hat{q}_{a_1'a_2'}P\left(g_i|G_i=(m,m')\right)],$$

with

$$\hat{q}_{a_1a_2}=\begin{cases}2\hat{p}_{a_1}\hat{p}_{a_2}, a_1\neq a_2;\\ \hat{p}_{a_1}^2, otherwise.\end{cases}$$

Under HWE, the test $-2(\ln L_0-\ln L_1)$ has an approximate Chi-square distribution with $k(k+1)/2-k=k(k-1)/2$ degrees of freedom. Finally, we extend our testing procedure to accommodate a stratified population where each stratum has different genotype frequencies. In what follows, we will propose a likelihood ratio test for multiple populations with different genotype frequencies. Suppose the genotype data is sampled from L known populations. Let n_l be the number of families in the l-th, l=1,…,L, population. The likelihood ratio test for HWE is given by $-2\ln\left(\prod_{l=1}^{L}L_0^l/\prod_{l=1}^{L}L_1^l\right)$, where

$$\ln L_1^{\,l}=\sum_{i=1}^{n_l}\ln\sum_{(m,m')\in\mathcal{A}^2}\left[\hat{q}_m^l\hat{q}_{m'}^l P\left(g_i|G_i=(m,m')\right)\right],$$

and

$$\ln L_0^{\,l}=\sum_{i=1}^{n_l}\ln\sum_{m=(a_1,a_2),m=(a_1',a_2')}\left[\hat{q}_{a_1a_2}^l\hat{q}_{a_1'a_2'}^l P\left(g_i|G_i=(m,m')\right)\right].$$

Under HWE, the test has an approximate Chi-square distribution with $lk(k-1)/2$ degrees of freedom.

Simulations

In this section, we conduct simulation studies to assess the performance of the proposed method. We focus on showing that the likelihood ratio test has appropriate type I error rates and good power.

Observed Genotypes	True Genotypes		
	AA	Aa	aa
AA	$(1-\varepsilon_1)^2$	$\varepsilon_2(1-\varepsilon_1)$	ε_2^2
Aa	$2\varepsilon_1(1-\varepsilon_1)$	$\varepsilon_1\varepsilon_2+(1-\varepsilon_1)(1-\varepsilon_2)$	$2\varepsilon_2(1-\varepsilon_2)$
aa	ε_1^2	$\varepsilon_1(1-\varepsilon_2)$	$(1-\varepsilon_2)^2$

Table 1: Penetrances of genotypes with 2 alleles.

We will also compare our LRT with the approximate χ^2 and exact test proposed by Shriner [16].

To assess type I error rate, we simulate independent families under the null hypothesis of HWE with each family having 2 non-genotyped parents and 2 genotyped siblings. We perform 1000 simulation runs at a significance level α=0.05. Firstly, we investigate how type I error rate varies across different genotyping errors at a fixed genotype frequency of parents. We choose two specific genotype frequencies for parents: (0.25, 0.5, 0.25) and (0.04, 0.32, 0.64) with genotyping errors $\varepsilon_1 = \varepsilon_2 = 0.01$ or $\varepsilon_1 = \varepsilon_2 = 0.05$. Simulation results for three methods, LRT, approximate and exact are given in Table 2. Table 2 demonstrates LRT approximates α best, with increased accuracy as the sample size increases. Secondly, we investigate how type I error rate varies across different minor allele frequencies (MAF), family sizes N, and with or without genotyping errors. For simulations without genotyping errors $(\varepsilon_1 = \varepsilon_2 = 0)$, the MAF ranges from 0.01 to 0.5. For simulations with genotyping errors $(\varepsilon_1 = \varepsilon_2 = 0.01)$, the MAF ranges from 0.05 to 0.5. Figure 1 shows that without genotyping errors, LRT maintains valid control of type I error rate across different MAFs and sample sizes. As the family size and MAF increase, type I error rate of LRT is further improved. In the presence of genotyping error, LRT still controls type I error rate close to the nominal level except for extremely low MAFs. This means that the extremely low MAFs coupled with genotying errors lead to larger type I error inflation. For example, with MAF=0.01, genotying errors $\varepsilon_1 = \varepsilon_2 = 0.01$, and N=500, the type I error rate is 0.193.

We further investigate the impact of MAFs and inbreeding coefficient f on the power of the proposed LRT [8,26-28]. We simulate independent parents under the alternative hypothesis with $f \neq 0$. We perform 1000 simulation runs at a significance level α=0.05. Firstly, we examine the impact of genotying errors on the power. We simulate two scenarios of genotyping errors $\varepsilon_1 = \varepsilon_2 = 0.01$ or $\varepsilon_1 = \varepsilon_2 = 0.05$ with genotype frequencies set to be (0.1, 0.2, 0.7) or (0.04, 0.1, 0.86). Table 3 shows that LRT has a higher power compared to the approximate and exact tests, and the power increases as the sample size increases and genotyping errors decrease. Secondly, we investigate how the power varies across different MAFs, family sizes N, and inbreeding coefficient f. The results are summarized in Figures 2 and 3. The MAF ranges from 0.05 to 0.5 and genotyping errors are $\varepsilon_1 = \varepsilon_2 = 0.01$. Figures 2 and 3 demonstrate that the power of LRT increases with the sample size, MAF and the inbreeding coefficient f.

We further evaluate the performance of the proposed LRT across strata in terms of type I error rate and power. We run experiments with 1000 repetitions and L=2 known populations at a significance

				Type I error rate		
ε_1	ε_2	N	G	LRT	Approximate χ^2	Exact
0.01	0.01	100	(0.25, 0.5, 0.25)	0.057	0.078	0.084
		200		0.051	0.098	0.098
0.05	0.05	100		0.071	0.298	0.323
		200		0.044	0.592	0.626
0.01	0.01	100	(0.04, 0.32, 0.64)	0.025	0.080	0.075
		200		0.051	0.102	0.101
0.05	0.05	100		0.049	0.334	0.376
		200		0.047	0.675	0.691

Table 2: Type I error rate using uncertain genotypes for two alleles.

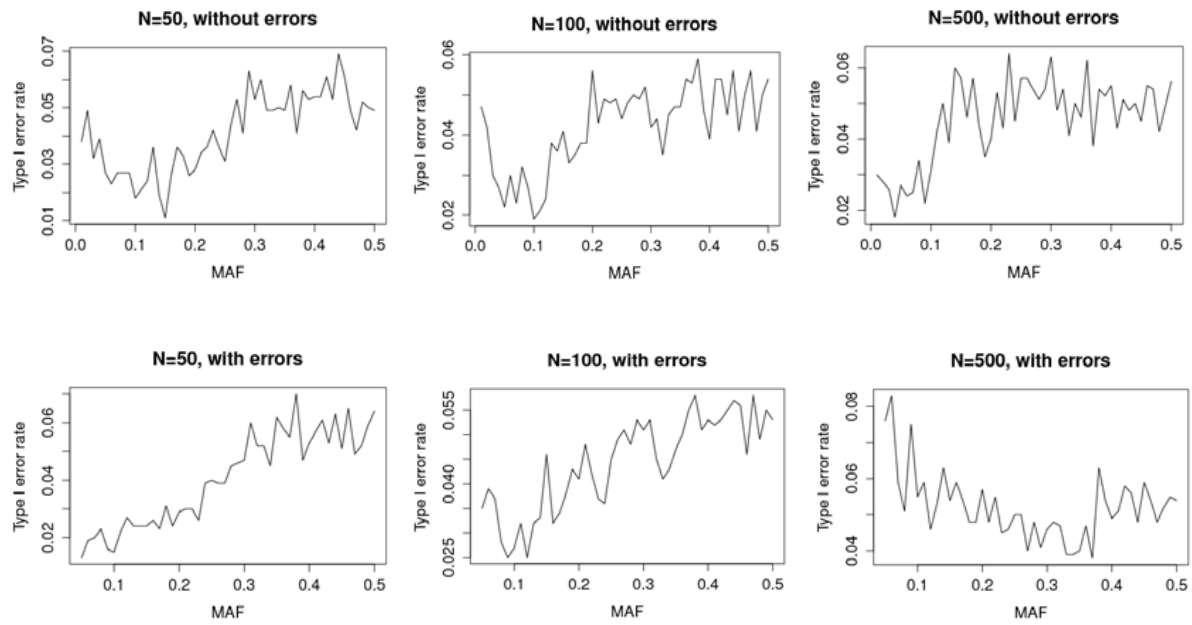

Figure 1: Type I error rate of LRT against MAF for different family size (50, 100 and 500). The labels "with errors" and "without errors" represent the simulations with genotyping errors and without genotyping errors, respectively.

				Power		
ε_1	ε_2	N	G	LRT	Approximate χ^2	Exact
0.01	0.01	100	(0.1, 0.2, 0.7)	0.598	0.097	0.087
		200		0.894	0.126	0.116
0.05	0.05	100		0.548	0.328	0.365
		200		0.836	0.666	0.677
0.01	0.01	100	(0.04, 0.1, 0.86)	0.384	0.088	0.109
		200		0.725	0.102	0.125
0.05	0.05	100		0.498	0.392	0.433
		200		0.847	0.758	0.766

Table 3: Power using uncertain genotypes for two alleles.

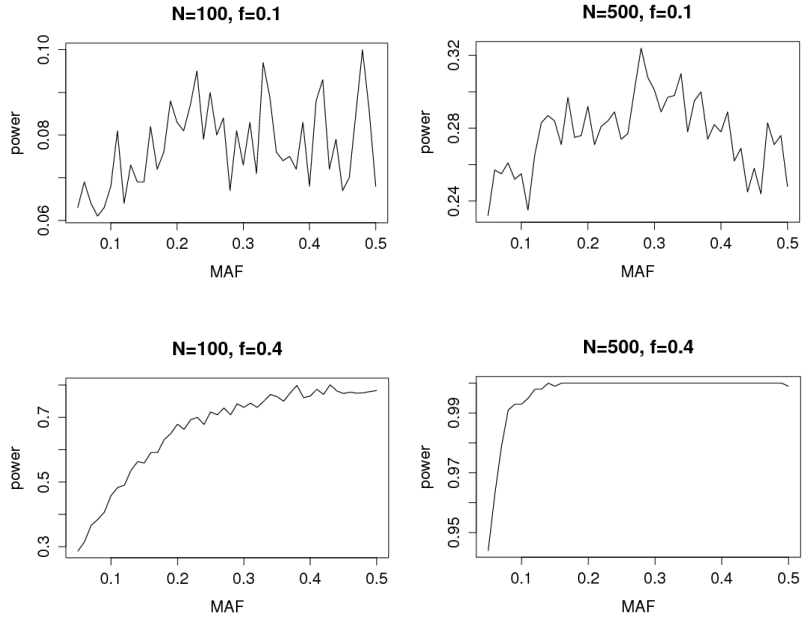

Figure 2: Power against MAF for different family size (100 and 500) with *f*=0.1 or *f*=0.4.

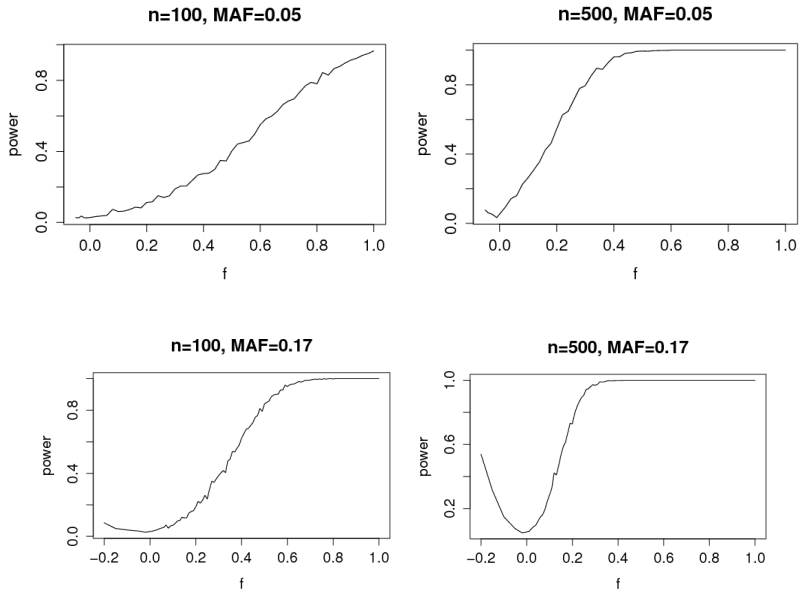

Figure 3: Power against f for different family size (100 and 500) with MAF=0.05 or MAF=0.17. For MAF=0.05, f ranges from -0.02 to 1; for MAF=0.17, f ranges from -0.2 to 1.

			Under the null hypothesis		Under the alternative hypothesis	
ε_1	ε_2	n_l	G	Type I error rate	G	Power
0.01	0.01	100	(0.25, 0.5, 0.25)	0.043	(0.04, 0.32, 0.64)	0.918
		100	(0.16, 0.48, 0.36)		(0.2, 0.2, 0.6)	
0.01	0.01	200	(0.25, 0.5, 0.25)	0.048	(0.04, 0.32, 0.64)	1.000
		200	(0.16, 0.48, 0.36)		(0.2, 0.2, 0.6)	
0.05	0.05	100	(0.25, 0.5, 0.25)	0.046	(0.04, 0.32, 0.64)	0.753
		100	(0.16, 0.48, 0.36)		(0.2, 0.2, 0.6)	
0.05	0.05	200	(0.25, 0.5, 0.25)	0.051	(0.04, 0.32, 0.64)	0.980
		200	(0.16, 0.48, 0.36)		(0.2, 0.2, 0.6)	

Table 4: Type I error rate and power across strata.

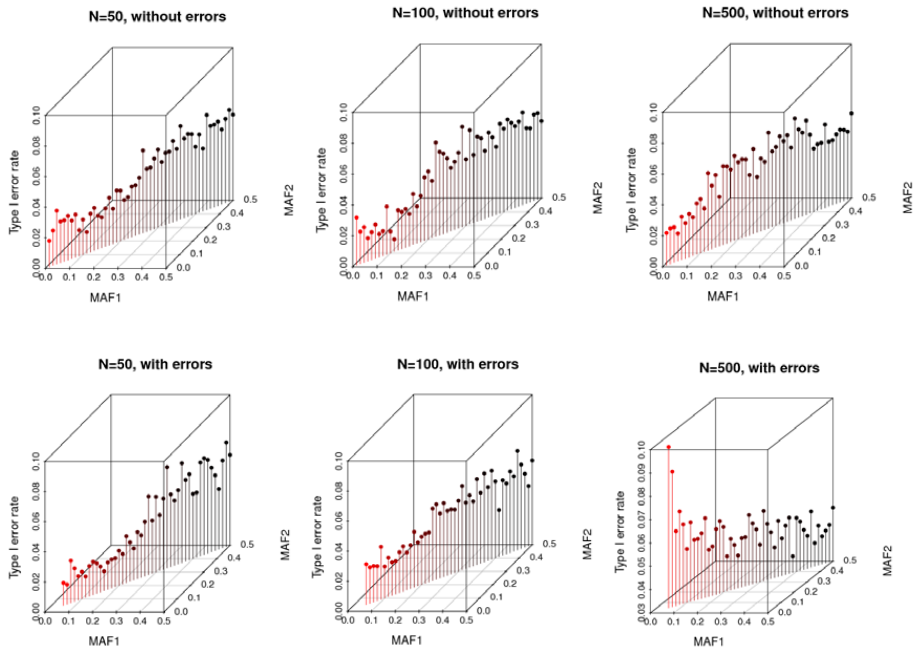

Figure 4: Type I error rate of LRT against MAFs for different family size (50, 100 and 500). The labels "with errors" and "without errors" represent the simulations with genotyping errors and without genotyping errors, respectively.

level α=0.05. Firstly, we assess type I error rate and power for different genotyping errors. To assess type I error rate, we set the genotype frequencies for parents to be (0.25, 0.5, 0.25) for population 1 and (0.16, 0.48, 0.36) for population 2. To assess the power of LRT, we set the genotype frequencies for parents to be (0.04, 0.32, 0.64) for population 1 and (0.2, 0.2, 0.6) for population 2. In each population, we choose the same error rates $\varepsilon_1 = \varepsilon_2 = 0.01$ or $\varepsilon_1 = \varepsilon_2 = 0.05$. In Table 4, the type I error rate and power are summarized. It can be seen that our LRT are valid, in the sense that they maintain appropriate control of the type I error rate. The power of the LRT increases as the sample size increases and genotyping errors decrease. Next, we investigate how type I error rate varies across MAF1 (MAF in the population 1) and MAF2 (MAF in the population 2), the family size, and with or without genotyping errors. For the simulation without genotyping errors $(\varepsilon_1 = \varepsilon_2 = 0)$, the range of MAF1=MAF2 is from 0.01 to 0.5. For the simulation with genotyping errors $(\varepsilon_1 = \varepsilon_2 = 0.01)$, the range of MAF1=MAF2 is from 0.05 to 0.5. Figure 4 shows that the LRT can approximate α better as the family size and MAFs increase. Genotyping errors lead to larger type I error inflation when MAFs are extremely low. Finally, we also investigate the effect of MAFs, inbreeding coefficient *f* and the sample size on power by Figures 5 and 6. The range of MAF1=MAF2 is from 0.05 to 0.5 and the genotyping errors are $\varepsilon_1 = \varepsilon_2 = 0.01$. It is observed that the power of LRT increases with the sample size, MAF and the inbreeding coefficient *f*.

Real Data Analysis

We analyze the human CEPH genotype dataset (V10) which is available online (www. cephb.fr/cephdb). The dataset contains DNA marker genotypes for 32,356 marker loci from 65 families.

Within each family, the second generation contains 1 to 15 genotyped siblings. We perform HWE test on the collection of all second generation siblings. We select a few markers D1S16, D1S20, D1S14 and D1S70 on the chromosome 1 with MAF ranges from 0.119 to 0.454. We obtain the p-values on these markers and compare them with the p-values from the approximate and exact tests. The results are summarized in Table 5. For the markers D1S16, D1S20 and D1S14, the decision of the LRT to accept or reject HWE is in agreement with the other two methods. However, for marker D1S70, the LRT accepts the HWE, while the approximate and exact tests both reject the HWE. The

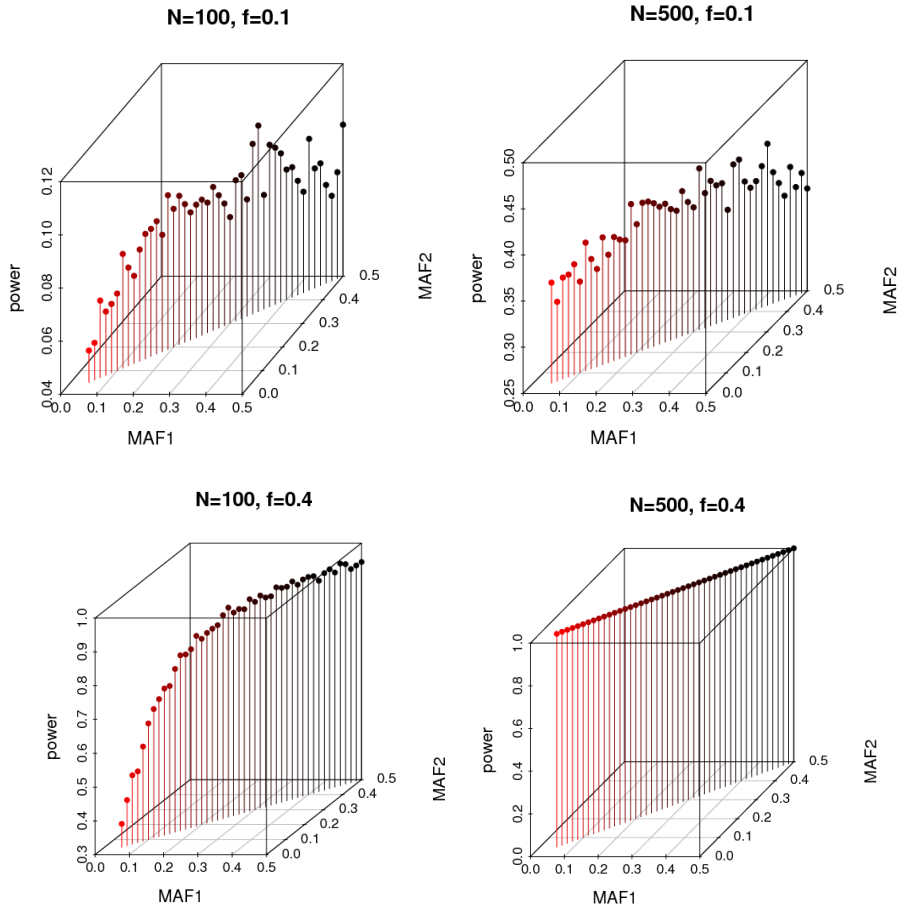

Figure 5: Power against f for different family size (100 and 500) with MAFs=0.05 or MAFs=0.17.

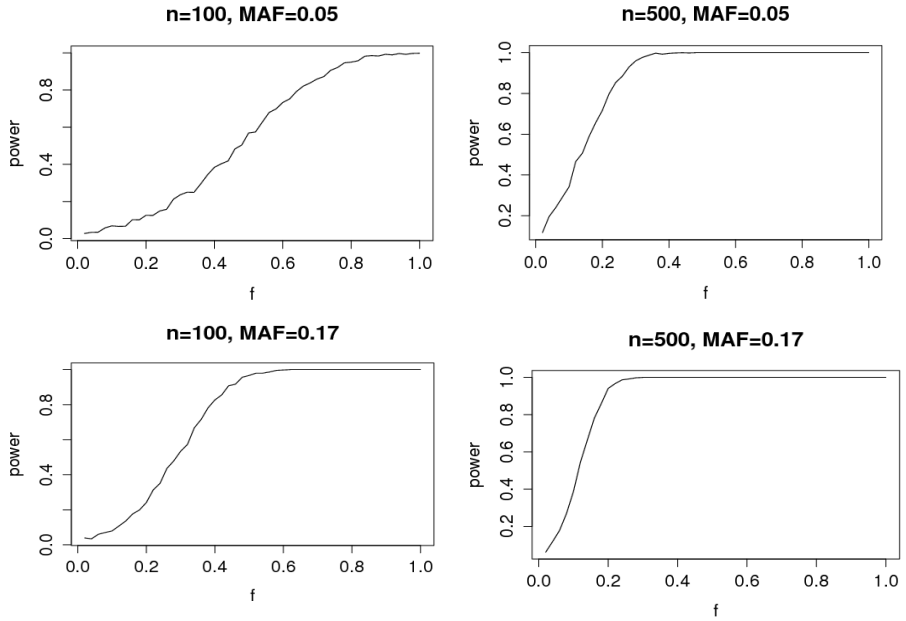

Figure 6: Power against MAFs for different family size (100 and 500) with *f*=0.1 or *f*=0.4.

	Minor allele	Family	Number of	*p*-values		
Markers	Frequencies	Size	Siblings	LRT	Approximate χ^2	Exact
D1S16	0.277	32	213	0.052	0.030	0.039
D1S20	0.119	24	139	0.326	0.112	0.217
D1S14	0.454	37	273	0.118	0.869	0.903
D1S70	0.323	39	296	0.694	0.0002	0.0003

Table 5: *p*-values for 4 markers on chromosome 1.

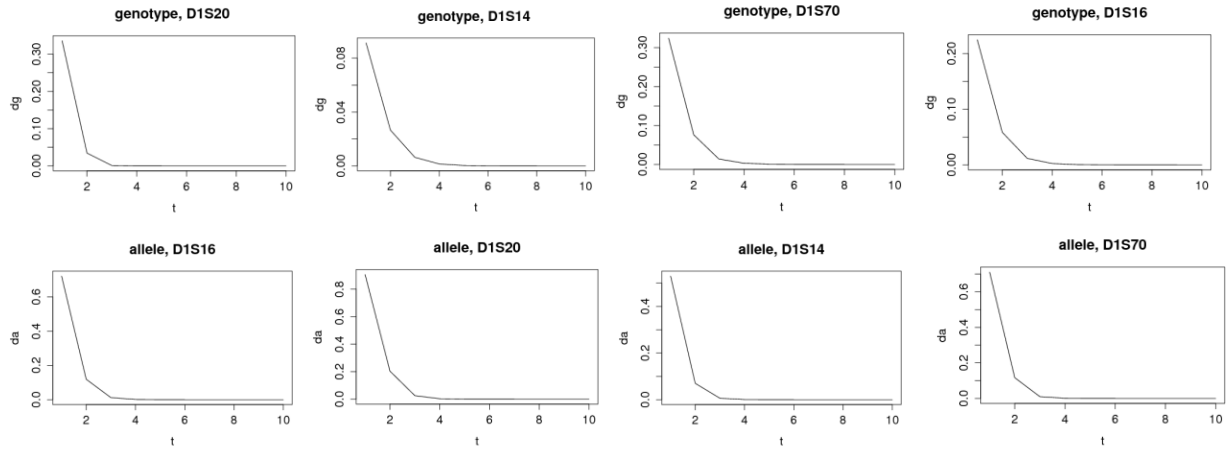

Figure 7: Difference in successive updates of parameters against the iteration number t for markers D1S16, D1S20, D1S14 and D1S70 on chromosome 1, with $d_a = \left(\sum_{s=1}^{k} \left(\hat{a}_s^{(t+1)} - \hat{a}_s^{(t)} \right)^2 \right)^{\frac{1}{2}}$ and $d_g = \left(\sum_{m \in A} \left(\hat{q}_m^{(t+1)} - \hat{q}_m^{(t)} \right)^2 \right)^{\frac{1}{2}}$.

estimates of genotype frequency and allele frequency based on our EM algorithm are (0.49, 0.4, 0.11) and (0.693, 0.307), respectively, which does not manifest a severe deviation from HWE. This discrepancy is due to the fact that our method takes into account the dependencies among the multiple siblings and estimates the founders' genotype frequencies through EM algorithm. In contrast, blindly applying the other two methods on dependent siblings leads to the use of the siblings' genotype frequencies as the founders' genotype frequencies. The convergence of EM algorithm is demonstrated through Figure 7, which plots the difference between successive updates of parameters against iteration number. It is observed that the EM algorithm converges quickly after a few iterations.

Conclusion

In this paper, we develop a likelihood ratio test for HWE for sibship data with random genotying errors. One constraint of our method is that we need to know the genotyping error rates ε_1 and ε_2 accurately. Our method is shown to maintain the type I error rate better and to be more powerful than previous tests that do not take into account dependence of siblings and genotyping errors.

Acknowledgements

The research was supported by Canadian NSERC Grant of Hélène Massam and Xin Gao. The authors would like to thank Dr. Daniel Shriner for providing R programs of the approximate χ^2 and exact tests, and thank Dr. Lei Sun for helpful discussions. In particular, the authors thank reviewers for their insightful comments and suggestions.

References

1. Weir WS (1996) Genetic data analysis II. MA: Sinauer Associates, Sunderland.
2. Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, et al. (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet 12: 395-399.
3. Lee WC (2003) Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. Am J Epidemiol 158: 397-400.
4. Nielsen DM, Ehm MG, Weir BS (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 63: 1531-1540.
5. Haldane JBS (1954) An exact test for randomness of mating. Journal of Genetics 52: 631-635.
6. Hernandez JL, Weir BS (1989) A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 45: 53-70.
7. Levene H (1949) On a matching problem arising in genetics. Ann Math Stat 20: 91-94.
8. Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76: 887-893.
9. Bourgain C, Abney M, Schneider D, Ober C, McPeek MS (2004) Testing for Hardy Weinberg Equilibrium in samples with related individuals. Genetics 168: 2349-2361.
10. Gordon D, Ott J (2001) Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput: 18-29.
11. Gordon D, Finch SJ, Nothnagel M, Ott J (2002) Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 54: 22-33.
12. Leal SM (2005) Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet Epidemiol 29: 204-214.
13. Cox DG, Kraft P (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. Hum Hered 61: 10-14.

14. Li M, Li C (2008) Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. Genet Epidemiol 32: 589-599.

15. Yu C, Zhang S, Zhou C, Sile S (2009) A likelihood ratio test of population Hardy-Weinberg equilibrium for case-control studies. Genet Epidemiol 33: 275-280.

16. Shriner D (2011) Approximate and exact tests of Hardy-Weinberg Equilibrium using uncertain genotypes. Genet Epidemiol 35: 632-637.

17. Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66: 279-292.

18. Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64: 259-267.

19. Minică CC, Dolan CV , Hottenga JJ, Willemsen G, Vink JM, et al. (2013) The use of imputed sibling genotypes in sibship-based association analysis: on modeling alternatives, power and model misspecification. Behav Genet 43: 254-266.

20. Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, et al. (2007) Genome partitioning of genetic variation for height from 11,214 sibling pairs. Am J Hum Genet 81: 1104-1110.

21. Alcaïs A, Abel L (2000) Linkage analysis of quantitative trait loci: sib pairs or sibships? Hum Hered 50: 251-256.

22. Schaid DJ, Batzler AJ, Jenkins GD, Hildebrandt MA (2006) Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. Am J Hum Genet 79: 1071-1080.

23. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39: 1-38.

24. Gordon D, Heath SC, Liu X, Ott J (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 69: 371-380.

25. Ott J (1999) Analysis of human genetic linkage. John Hopkins University Press, Baltimore.

26. Emigh TH (1980) A comparison of tests for Hardy-Weinberg equilibrium. Biometrics 36: 627-642.

27. Graffelman J, Moreno V (2013) The mid p-value in exact tests for Hardy-Weinberg equilibrium. Stat Appl Genet Mol Biol 12: 433-448.

28. R. V. Rohlfs, B. S. Weir (2008) Distributions of Hardy-Weinberg equilibrium test statistics. Genetics 180: 1609-1616.

Citation: Li Q, Massam H, Gao X (2013) Likelihood Ratio Test of Hardy-Weinberg Equilibrium Using Uncertain Genotypes for Sibship Data. Biomedical Data Mining 3: 104. doi: 10.4172/2090-4924.1000104