ASSESSMENT AND MANAGEMENT OF SINGLE NUCLEOTIDE POLYMORPHISM GENOTYPE ERRORS IN GENETIC ASSOCIATION ANALYSIS

DEREK GORDON AND JÜRG OTT

Laboratory of Statistical Genetics, Rockefeller University
1230 York Avenue, New York, NY 10021-6399

Single nucleotide polymorphisms (SNP) may be used in case-control designs to test for association between a marker (the SNP) and a disease. However, such designs usually assume that the genotype data are reported without error. We propose a method, the reduced penetrance model method (RPM) that allows for errors in a case-control design, as compared to the full penetrance model method (FPM), that assumes data are errorless. Pearson's χ^2 applied to a 2×2 contingency table is the test statistic considered. Additionally, we provide a likelihood method to estimate error rates using SNP genotype data in CEPH pedigrees. We test our method (RPM) against the standard method (FPM) using simulated data. All SNP loci are assumed to have two alleles, coded *1* and *2*.

We consider three pairs of error rates, two different sample sizes, and two sets of allele frequencies for the SNP locus. SNP genotype data in two populations are simulated under a null hypothesis (allele frequencies equal in both populations) and under an alternative hypothesis (allele frequencies differ between two populations). The total number of simulations is 24; 12 simulations under the null hypothesis, and 12 simulations under the alternative. The significance level threshold is 5%.

For the null case, 9/12 (75%) of the simulations show no increase in type I error under RPM, while 3/12 (25%) show a slight increase (rejecting the null for at most 7% of the replicates). There is no increase in the type I error rate for FPM method, which can also be shown analytically. For the alternative case (power), there is a consistent increase in power for the RPM method as compared to FPM method, and average increase of 0.02 for the simulations considered. When sample sizes are large there is virtually no difference in power between RPM and FPM methods. Also, the RPM method provides consistently more accurate allele frequency estimates for the various populations.

Our likelihood method to estimate error rates with CEPH pedigrees provides good estimates on average. The largest difference between a true error rate and our average estimated error rate is 0.006. However, there is a fair amount of variability in the estimates, suggesting the need for multiple experiments or larger numbers of CEPH pedigrees.

Researchers may use the methods presented in this paper to (1) estimate error rates for their automated genotyping process, and (2) allow for such errors in association analyses, thereby increasing power to detect differences between allele frequencies in case and control populations when errors are present.

1 Introduction

There is growing interest in the use of single nucleotide polymorphisms (SNP) for the genetic dissection of complex human diseases[1]. Some reasons are: 1) SNPs are significantly more abundant than microsatellite polymorphisms (about one SNP for every 500-1000 base pairs[2] and therefore are potentially more powerful in detecting linkage disequilibrium (LD) around disease loci; 2) high throughput genotyping of large numbers of SNP markers is possible with the use of microarray technologies;

3) some SNP mutations may be causative of disease phenotype; 4) the completion of the human genome reference sequence in the not-too-distant future should pave the way for discovery of many of the common polymorphisms should be possible[3].

To take advantage of the greater expected LD between SNP loci and disease loci, statistical methods being considered are population-based tests of association (case-control studies). Much work has been done to determine the statistical properties of such tests, including validity and power of such tests under different genetic models of disease[4]. However, it is almost always assumed in these analyses that the genetic data considered are without errors. By errors, we mean any miscoding of a person's correct marker genotype. Sources of error include non-paternity, sample-swaps in the lab, or genotyping errors. While there has been much written on methods to detect errors[5-12], there is only one very recent set of papers[13-16] that consider methodology allowing for errors in linkage and/or LD analysis, even though it is well known that errors in genetic data can have significant effects on linkage analyses. Such effects include increase in the estimated recombination fraction between markers or between marker and trait (more generally, an inflation of the map distance for multiple markers), an increase in type I error rate, and a decrease in power[17-19]. The purpose of this work is the assessment of errors on the validity (type I error rate) and power of specific population-based association tests, and the proposal of statistical methods that allow for errors in the analysis. In addition, with the use of currently available software, the effectiveness of these methods will be demonstrated empirically.

2 Materials and Methods

2.1 Error Model

For all our analyses, it will be assumed the SNP loci in question have two alleles, coded as *1* and *2*. Also, we assume that each *1* allele has a constant probability ε_1 of being incorrectly coded as a *2* allele, and likewise, each *2* allele has a constant probability ε_2 of being incorrectly coded as a *1* allele. Thus, the number of *observed 1* or *2* alleles (as opposed to the actual number of such alleles) in a sample is still a binomial variable, so that it is possible to compute means and variances for observed numbers, even when errors are present.

2.2 Data and Test Statistic

The data selected for use with population-based tests is a SNP locus that has two alleles in the population. These alleles are coded as *1* and *2*.

Table 1 - 2 $\times$ 2 contingency table

	1 Allele	2 Allele	Row Totals
Cases	N_{11}	N_{12}	N_{1*}
Controls	N_{21}	N_{22}	N_{2*}
Column Totals	N_{*1}	N_{*2}	N

N_{ij} = number of j alleles observed in the i population (i = 1 - cases; i = 2 - controls)
$N_{*j} = N_{1j} + N_{2j}$
$N_{i*} = N_{i1} + N_{i2}$
$N = N_{11} + N_{12} + N_{21} + N_{22}$

      The statistical test chosen is Pearson's χ^2 on 2 $\times$ 2 contingency tables (see Table 1 for an example). The rows of the table are cases (affected individuals) and controls (unaffected individuals) randomly sampled from a population. The columns of the contingency table are the counts of *1* alleles and *2* alleles in each population (cases and controls). Using the notation from Table 1, the Pearson Chi-square statistic for a sample is:

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(N_{ij} - M_{ij})^2}{M_{ij}} \ , \qquad (1)$$

where $M_{ij} = N_{i*} N_{*j} / N.$

2.3 Reduced Penetrance Model Method (RPM)

As a means of allowing for errors in population-based SNP data, we propose use of a reduced penetrance model (RPM) as implemented in the ILINK program from the FASTLINK suite of programs[20-23]. The ILINK program has the flexibility of allowing for "reduced penetrance" marker genotypes. This step is achieved by recoding the marker locus as an affection status locus. For an example, see reference no. 21, Section 10.2. For error rates ε_1 and ε_2, the matrix of penetrances used in the ILINK runs is presented in Table 2.

Table 2 - Penetrances for observed genotypes with 2-allele locus assuming errors

	True Genotype		
Observed Genotype	*1/1*	*1/2*	*2/2*
1/1	$(1 - \varepsilon_1)^2$	$\varepsilon_2 (1 - \varepsilon_1)$	ε_2^2
1/2	$2\varepsilon_1 (1 - \varepsilon_1)$	$\varepsilon_1 \varepsilon_2 + (1 - \varepsilon_1)(1 - \varepsilon_2)$	$2\varepsilon_2 (1 - \varepsilon_2)$
2/2	ε_1^2	$\varepsilon_1 (1 - \varepsilon_2)$	$(1 - \varepsilon_2)^2$

      One thousand replicates of genotype data for two populations (labeled in this study as cases and controls) are simulated using the SIMULATE program[24]. For

simplicity, simulations use either population of $N = 250$ or 500 individuals in each population. The pedigree structure used in the SIMULATE simped.dat file is a triad (father, mother, child) in which the parents are assumed to have unknown genotypes. Allele frequencies for the *1* allele in each population are entered in the SIMULATE simdata.dat file (see reference no. 21, section 28.5 for an example of the simped.dat and simdata.dat files). Because the SIMULATE program simulates null data, genotype data are created independent of disease status (i.e., recombination fraction between marker and disease locus is 0.5). For these simulations, we consider the both different allele *1* frequencies in the case and control populations (power) and also allele *1* frequencies that are the same in case and control populations (null or type I error). For power simulations, allele *1* frequencies simulated in (case, control) populations are (0.2, 0.3) and (0.4,0.5). For the null simulations, allele *1* frequencies simulated are (0.3, 0.3) and (0.5, 0.5). Errors for the *1* and *2* alleles are introduced using a C program according to error rates ε_1 and ε_2. It is assumed that errors are introduced randomly and independently into a sample of alleles, and that error introduction is independent of population, i.e., errors are introduced into the set of alleles for both populations according to the same error rates. For these simulations, we assume error rates of 0.01and 0.05 for each of the ε_i ($i = 1, 2$), for a total of 3 joint error rates (0.01, 0.05), (0.05, 0.01), and (0.05, 0.05).

The genotype data with errors introduced are recoded so that they are data for an affection status locus (a necessary step with the ILINK program). The ILINK program is run assuming full penetrance (i.e., assuming that $\varepsilon_1 = \varepsilon_2 = 0$). We shall call this analysis "the full penetrance model method" (FPM). Penetrances are then determined by substituting the correct error rates ε_1 and ε_2 into the matrix (Table 2). We call this analysis "the reduced penetrance model method". Output from one iteration of the ILINK program consists of frequency estimates for the *1* and *2* alleles. These estimates, for each population, are multiplied by the number of alleles in each population, and placed in the respective cells (Table 1). For example, if we label the first population as the "case" population, sample size N_1, and ILINK produces an estimate of p_1 for the *1* allele under full penetrance, then the value $2 \times N_1 \times p_1$ is entered in the upper-left cell of Table 1, and $2 \times N_1 \times (1- p_1)$ is entered in the upper-right cell. Likewise for the other two cells. These values are then placed into formula (1) and the chi-square statistic is computed. The proportions of chi-square statistics greater than 3.84 (corresponding to a *p*-value of 0.05) for the data analyzed under the FPM method, and under the RPM method, for all sets of allele frequencies, and for different sample sizes ($N = 250$ or 500 individuals in each population) are reported in Table 4.

2.4 Estimation of Error rates

One potential limitation of the RPM method presented above is the requirement that accurate error rates be specified in the analysis. To make accurate estimates of the

parameters ε_1 and ε_2 we propose the use of CEPH family genotypes[25]. It is possible to estimate error rates for an automated genotyping process by genotyping a certain number of CEPH pedigrees, and by computing the relative likelihoods of the genotypes for different values of ε_1 and ε_2 in the ILINK program, under the assumption that the recombination fraction between marker and "trait" is 0.5 (the unaffected affection status is provided for each CEPH individual). Evaluation of genotype likelihood for each setting of the pair $(\varepsilon_1, \varepsilon_2)$ is performed in the same manner as is described above (Section 2.3). The settings of ε_1 and ε_2 that provide the largest likelihood for the CEPH genotype data become the maximum likelihood estimates (MLEs) for ε_1 and ε_2, denoted by $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ respectively.

To determine the effectiveness of this method for estimating error rates, we performed several simulation studies. One set of 15 CEPH pedigrees, with a total of 212 individuals, was selected. A list of the pedigrees selected may be found in Table 3. Allele frequency pairs (p_1, p_2) at one locus chosen for the simulations were (0.3, 0.7) and (0.5, 0.5). Error rate pairs $(\varepsilon_1, \varepsilon_2)$ selected were as above, namely (0.01, 0.05), (0.05, 0.05), and (0.05, 0.01). Genotype data were simulated and errors were introduced as described above (Section 2.3). The grid of parameter values $(\varepsilon_1, \varepsilon_2)$ for which likelihoods were evaluated ranged from (0.0, 0.0) to (0.10, 0.10), in increments of 0.005 for each coordinate. A total of $21^2 = 461$ likelihoods were therefore computed in each simulation. The values of $(\varepsilon_1, \varepsilon_2)$ that provided the maximum likelihood for each simulation were recorded. For each pair of allele frequencies and error rates (one simulation), a total of 100 replicates were created and analyzed. The sample mean and standard deviation of each simulation are reported in Table 6.

Table 3 - List of CEPH pedigrees selected for simulation study on estimating error rates

1326	13291	1353
1327	13292	1354
1328	1347	1355
13281	1349	1356
1329	1350	1357

We choose CEPH pedigrees because error rates should be better estimated by using extended pedigrees instead of nuclear families. Ehm et al.[6] pointed out that errors are more precisely and easily detected in extended pedigrees. Also, Gordon et al.[7] showed analytically that error detection rates are, on average, at most 58% for nuclear families genotyped at a SNP locus, in which there are at most 3 children.

3 Results

3.1 Type I Error Rate and Power - FPM Method

Bross[26] proved that there is no increase in type I error rate for the statistic X^2 applied to 2×2 contingency tables, when the errors in diagnosis (mislabeling a case as a control and vice versa) occur independently of the particular population. For our purposes, assume that we have collected genotypes for a SNP locus on cases and controls, and further assume that errors occur in the genotype data according to the error model presented above. Further assume that the error model is valid *independent of disease status (case or control).* It follows immediately from Bross's proof[26] and from the fact that the Chi-square statistic X^2 is invariant under permutation of indices (i.e., rows and columns may be interchanged) that *there is no increase in type I error rate for the Chi-square statistic X^2 applied to the 2×2 contingency table (Table 1) using the FPM method, when errors are introduced according to our error model.* Our simulations (Table 4) also show this result.

Table 4 - Power for case-control samples for SNP genotype data with errors analyzed under (1) FPM (assuming no errors) and (2) RPM using correct values of error rates

Sample Size	*1* allele frequency - pop1	*1* allele frequency - pop2	Error Rate ε_1	Error Rate ε_2	Power-FPM	Power-RPM	True power w/o errors
POWER AT 5% SIGNIFICANCE LEVEL							
250	0.2	0.3	0.01	0.05	0.912	0.943	0.956
250	0.4	0.5	0.01	0.05	0.853	0.880	0.893
250	0.2	0.3	0.05	0.01	0.932	0.943	0.956
250	0.4	0.5	0.05	0.01	0.857	0.871	0.893
250	0.2	0.3	0.05	0.05	0.885	0.937	0.956
250	0.4	0.5	0.05	0.05	0.814	0.857	0.893
500	0.2	0.3	0.01	0.05	0.997	0.998	1.000
500	0.4	0.5	0.01	0.05	0.985	0.988	0.996
500	0.2	0.3	0.05	0.01	1.000	1.000	1.000
500	0.4	0.5	0.05	0.01	0.989	0.995	0.996
500	0.2	0.3	0.05	0.05	0.997	0.999	1.000
500	0.4	0.5	0.05	0.05	0.985	0.989	0.996
TYPE I ERROR FOR 5% SIGNIFICANCE LEVEL							
250	0.3	0.3	0.01	0.05	0.050	0.070	0.05
250	0.5	0.5	0.01	0.05	0.046	0.056	0.05
250	0.3	0.3	0.05	0.01	0.055	0.063	0.05
250	0.5	0.5	0.05	0.01	0.041	0.063	0.05
250	0.3	0.3	0.05	0.05	0.037	0.059	0.05
250	0.5	0.5	0.05	0.05	0.044	0.068	0.05
500	0.3	0.3	0.01	0.05	0.036	0.060	0.05
500	0.5	0.5	0.01	0.05	0.036	0.050	0.05
500	0.3	0.3	0.05	0.01	0.046	0.050	0.05
500	0.5	0.5	0.05	0.01	0.054	0.061	0.05
500	0.3	0.3	0.05	0.05	0.038	0.061	0.05
500	0.5	0.5	0.05	0.05	0.038	0.069	0.05

However, as the saying goes, there is no free lunch. The price one pays for errors in genotype data is a potential loss in power. The extent of loss (or gain) in power depends on the values ε_1 and ε_2. In what follows, we use the notation p_1 and p_2 to represent the 1 allele frequency in cases and controls, respectively. The null hypothesis is $H_0 : p_1 = p_2$, and the alternative hypothesis is $H_1 : p_1 \neq p_2$. Even in the case of no errors, power to reject a false null hypothesis depends on the sample size, and the magnitude of the difference[27] $\left| p_1 - p_2 \right|$. Errors affect the power of the test in that they cause the apparent magnitude of the difference to be altered, as well as affecting the variance of the magnitude.

Power assuming no errors can be computed using exact methods or approximations[27]. We compute an estimate of the power by means of simulations. In each simulation, we specify the sample size in cases and controls, the exact error rates ε_1 and ε_2, the allele frequencies p_1 and p_2, the number of simulations to be performed, and a random seed. Allele 1 and 2 counts are then determined by invoking a random number generator. We simulated 20,000 replicates to determine an estimate of the true power under no errors. Our simulation results agree with those of Patnaik[27] in the case of 18 and 12 individuals at the 2% level up to 2 digits (Patnaik only considered power at 2% and 10% significance levels).

In studying Table 4, we see that there is a definite and consistent loss of power for all sample sizes and all error rates when comparing data analyzed under FPM, as opposed to the true power with errorless data. If we consider the difference (true power without errors - power assuming FPM) then, the average differences are 0.05 and 0.01 for sample sizes of 250 and 500 respectively. The largest difference occurs for a sample of 250 individuals, error rates (0.05, 0.05), and for allele frequencies of 0.4 and 0.5, a difference of 0.079. The smallest difference occurs for a sample of 500 individuals, error rates (0.05, 0.01), and for allele frequencies of 0.2 and 0.3, a difference of 0.0. These results suggest that, with increasing sample size, the difference in power between the true power and the power assuming a FPM decreases.

3.2 Type I Error Rate and Power - RPM Method

Table 4 also contains the results from simulations in which SNP data were analyzed under the RPM method. Considering type I error rate, we note that for an observed p-value of 0.064 (= 64/1000) the lower limit of the 95% confidence interval is less than 5 percent[28]. Therefore, let us declare an inflation in type I error when the observed p-values are greater than 0.064. With this threshold, we note that 9/12 (75%) of the simulations showed no increase in type I error, while 3/12 simulations (25%) did show some increase. The most extreme increase occurs for a sample size of 250 individuals, allele frequencies of 0.3, and error rates (0.01, 0.05). The increase was 0.07 - 0.05 = 0.02. Note also that the type I error rate for the RPM

method is always greater than the FPM method. One reason for this increase is that the variance of the allele frequency estimates from the RPM method is always equal to or greater than the FPM method (Table 5).

Regarding power, we note that for each simulation, power under the RPM method is *always* greater than or equal to power under the FPM method (Table 4). If we consider the difference (power under RPM - power under FPM), then the average difference over sample sizes of 250 and 500 individuals are 0.03 and 0.003 respectively. The largest difference is 0.052, which occurs for a sample size of 250, allele frequencies (0.2, 0.3), and error rates (0.05, 0.05).

Additionally, we note that the RPM method provides more accurate estimates of the allele frequencies for each population. Bross[26] noted that when errors are present, precise estimates of population frequencies are problematic. As Table 5 indicates, our simulations show that, under the full penetrance model, the estimates of allele frequencies are biased toward 0.5 on average. That is, when errors are present, more extreme allele frequencies tend to be estimated as being closer to 0.5. This observation also explains why there is a loss in power when analyzing under the FPM method, namely, the difference is allele frequencies appears to be smaller. It should also be noted that the bias persists even for large samples (500 individuals). In constrast, under the RPM method, the average of the allele frequency estimates are much closer to the true values.

3.3 Estimation of Error Rates

Table 6 provides summary statistics for the estimation of error rates from 15 CEPH pedigrees (listed in Table 3). We note that the average of the estimates $\hat{\varepsilon}_i$ ($i = 1,2$) are reasonably accurate. The largest difference between an average estimated parameter and the true parameter is 0.006, occurring for the error rates (0.05, 0.01) and for the allele frequencies (0.3, 0.7). We also note that the standard deviations of the MLEs $\hat{\varepsilon}_i$ can be fairly large, ranging from 0.009 to 0.036. This result indicates the variability in the MLEs. One way to reduce this variability is to perform, say, N genotyping experiments, and take the average of the MLEs over the N experiments as the estimates of the error rates. In this way, the standard error becomes the measure of the variability. Since the standard error is given by $\sigma / \sqrt{N}$, $(\sigma = \text{standard deviation})$, it reduces the variability of the MLEs by a factor of $\sqrt{N}$. However, several genotyping experiments must then be performed, increasing the cost of the study. Another way to decrease variability is to increase the number of CEPH families genotyped initially, thereby increasing sample size.

Table 5 - Summary statistics - case-control samples for SNP genotype data with errors analyzed under (1) FPM (assuming no errors) and (2) RPM using correct values of error rates

Size	*1*allele - pop 1	*1*allele - pop 2	ε_1	ε_2	Population 1				Population 2			
					FPM		RPM		FPM		RPM	
					Avg. *1* allele	SD *1* allele	Avg. *1* allele	SD *1* allele	Avg. *1* allele	SD *1* allele	Avg. *1* allele	SD *1* allele
					POWER							
250	0.2	0.3	0.01	0.05	0.238	0.019	0.200	0.020	0.332	0.021	0.300	0.023
250	0.4	0.5	0.01	0.05	0.427	0.021	0.401	0.022	0.520	0.022	0.500	0.023
250	0.2	0.3	0.05	0.01	0.198	0.018	0.200	0.019	0.292	0.020	0.300	0.021
250	0.4	0.5	0.05	0.01	0.386	0.022	0.400	0.023	0.480	0.022	0.500	0.023
250	0.2	0.3	0.05	0.05	0.230	0.019	0.200	0.021	0.320	0.021	0.300	0.023
250	0.4	0.5	0.05	0.05	0.411	0.022	0.402	0.025	0.501	0.023	0.502	0.025
500	0.2	0.3	0.01	0.05	0.239	0.013	0.201	0.014	0.333	0.014	0.301	0.015
500	0.4	0.5	0.01	0.05	0.427	0.015	0.401	0.016	0.520	0.016	0.500	0.017
500	0.2	0.3	0.05	0.01	0.198	0.013	0.200	0.013	0.293	0.014	0.301	0.015
500	0.4	0.5	0.05	0.01	0.387	0.016	0.401	0.017	0.481	0.016	0.501	0.017
500	0.2	0.3	0.05	0.05	0.230	0.013	0.200	0.015	0.320	0.015	0.301	0.017
500	0.4	0.5	0.05	0.05	0.411	0.016	0.401	0.018	0.501	0.015	0.501	0.017
					NULL							
250	0.3	0.3	0.01	0.05	0.332	0.020	0.300	0.021	0.331	0.020	0.299	0.022
250	0.5	0.5	0.01	0.05	0.521	0.021	0.501	0.022	0.520	0.022	0.500	0.023
250	0.3	0.3	0.05	0.01	0.297	0.020	0.305	0.022	0.297	0.020	0.305	0.021
250	0.5	0.5	0.05	0.01	0.489	0.021	0.510	0.023	0.489	0.022	0.510	0.023
250	0.3	0.3	0.05	0.05	0.324	0.019	0.305	0.021	0.324	0.020	0.305	0.022
250	0.5	0.5	0.05	0.05	0.509	0.021	0.510	0.024	0.510	0.021	0.511	0.024
500	0.3	0.3	0.01	0.05	0.337	0.014	0.305	0.015	0.336	0.014	0.304	0.015
500	0.5	0.5	0.01	0.05	0.522	0.015	0.502	0.016	0.523	0.015	0.503	0.016
500	0.3	0.3	0.05	0.01	0.294	0.015	0.302	0.015	0.294	0.014	0.303	0.015
500	0.5	0.5	0.05	0.01	0.484	0.016	0.504	0.017	0.485	0.015	0.505	0.016
500	0.3	0.3	0.05	0.05	0.326	0.014	0.306	0.016	0.325	0.014	0.306	0.015
500	0.5	0.5	0.05	0.05	0.506	0.015	0.507	0.017	0.507	0.015	0.507	0.017

Table 6 - Sample mean and standard deviations for maximum likelihood estimates of error rates ε_1 and ε_2 using CEPH pedigree structures - 100 Replicates

Error Rate ε_1	Error Rate ε_2	*1* Allele Frequency	Average $\hat{\varepsilon}_1$	Std. Dev. $\hat{\varepsilon}_1$	Average $\hat{\varepsilon}_2$	Std. Dev. $\hat{\varepsilon}_2$
0.01	0.05	0.3	0.013	0.023	0.046	0.020
0.01	0.05	0.5	0.009	0.010	0.048	0.024
0.05	0.01	0.3	0.044	0.031	0.010	0.009
0.05	0.01	0.5	0.049	0.028	0.010	0.013
0.05	0.05	0.3	0.046	0.036	0.050	0.023
0.05	0.05	0.5	0.051	0.029	0.049	0.031

4   Summary and Discussion

In this work, we present a method (RPM) that allows for random errors in the SNP genotype data of cases and controls. The RPM method is more powerful at detecting allele frequency differences in different populations (e.g., cases and controls) than the method that assumes errorless genotype data (FPM), and provides more accurate estimates of the allele frequency parameters on average. The main requirement of the RPM method is that accurate estimates of error rates ε_i are needed.

We also provide a likelihood-based method for estimation of the error rates ε_i using CEPH pedigrees. The average of the estimates are good, but the variability from replicate to replicate suggests the need for multiple retests or larger initial sample sizes. At present, genotyping CEPH pedigrees for the purpose of error estimation would only be feasible (cost-wise) for one or a few SNP loci, and not for whole-genome scans.

While it is true that for some the simulations, there was an observed increase in the type I error rate for the RPM method, the increase was relatively small (at most 7% of the replicates rejected the true null hypothesis at the 5% significance level), and the increase only occurred for 25% of the simulations. One possible solution to the increase in type I error would be an increase in the threshold used in the RPM method (e.g., a cutoff of 4.0 rather than 3.84 to declare significance at the 5% level). This is work in progress.

Acknowledgments

Electronic Database Information

The freeware program LINKAGE is available via ftp from the URL ftp://linkage.rockefeller.edu/software/linkage/. The freeware program FASTLINK is available via ftp from the URL ftp://fastlink.nih.gov/pub/fastlink/. Pedigree structures for the CEPH pedigrees may be downloaded from the URL http://www.cephb.fr/cephdb/.

References

1. F.M. De La Vega and M. Kreitman in *Pacific Symposium on Biocomputing 2000*, "Human genome variation: analysis, management, and application of SNP data." Eds. RB Altman, AK Dunker, L Hunter, K Lauderdale, and TE Klein (World Scientific, Singapore, 2000).
2. A. Chakravarti, "Population genetics - making sense out of sequence" *Nat Genet Suppl* **21**, 56 (1999)

3. F.S. Collins, L.D. Brooks, and A. Chakravarti, "A DNA polymorphism discovery resource for research on human genetic variation" *Genome Res* **8**,1229 (1998)
4. P. Sham in *Statistics in Human Genetics*, "Association analysis using a case-control design." (J Wiley and Sons, New York, 1998).
5. L.M. Brzustowicz, C. Merette, X. Xie, L. Townsend, T.C. Gilliam, and J. Ott, "Molecular and statistical approaches to the detection and correction of errors in genotype databases" *Am J Hum Genet* **53**, 1137 (1993)
6. M.G. Ehm, M. Kimmel, and R.W. Cottingham Jr., "Error detection for pedigree data, using likelihood methods" *Am J Hum Genet* **58**, 225 (1996)
7. D. Gordon, S.M. Leal, S.C. Heath, and J. Ott in *Pacific Symposium on Biocomputing 2000*, "An analytic solution to Single Nucleotide Polymorphism error-detection rates in nuclear families: implications for study design." Eds. RB Altman, AK Dunker, L Hunter, K Lauderdale, and TE Klein (World Scientific, Singapore, 2000).
8. S.E. Lincoln and E.S. Lander, "Systematic detection of errors in genetic linkage data" *Genomics* **14**, 604 (1992)
9. K.L. Lunetta, M. Boehnke, K. Lange, and D.R. Cox, "Experimental design and error detection for polyploid radiation hybrid mapping" *Genome Res* **5**, 151 (1995)
10. J.R. O'Connell and D.E. Weeks, "PedCheck: a program for identification of genotyping incompatibilities in linkage analysis" *Am J Hum Genet* **63**, 259 (1998)
11. J. Ott, "Detecting marker inconsistencies in human gene mapping" *Hum Hered* **43**, 25 (1993)
12. H.M. Stringham and M. Boehnke, "Identifying marker typing incompatibilities in linkage analysis" *Am J Hum Genet* **59**, 946 (1996)
13. H.H.H. Göring and J. D. Terwilliger, "Linkage analysis in the presence of errors I: Complex-valued recombination fractions and complex phenotypes" *Am J Hum Genet* **66**, 1095 (2000)
14. H.H.H. Göring and J. D. Terwilliger, "Linkage analysis in the presence of errors II: Marker-locus genotyping errors modeled with hypercomplex recombination fractions" *Am J Hum Genet* **66**, 1107 (2000)
15. H.H.H. Göring and J. D. Terwilliger, "Linkage analysis in the presence of errors III: Marker loci and their map as nuisance parameters" *Am J Hum Genet* **66**, 1298 (2000)
16. H.H.H. Göring and J. D. Terwilliger, " Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* **66**, 1310 (2000)
17. J. Ott, "Linkage analysis with misclassification at one locus" *Clin Genet* **12**, 110 (1977)

18. S.C. Heath, "A bias in TDT due to undetected genotyping errors" *Am J Hum Genet* **63**, A292 (1998)
19. J.D. Terwilliger, D.E. Weeks, and J. Ott, "Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross overestimates of the recombination fraction" *Am J Hum Genet* **47**, A201 (1990)
20. G.M. Lathrop, J.M. Lalouel, C. Julier, and J. Ott, "Strategies for multilocus linkage analysis in humans" *Proc Natl Acad Sci USA* **81**, 3443 (1984)
21. J.D. Terwilliger and J. Ott in *Handbook of Human Genetic Linkage*, "Running the LINKAGE programs MLINK and ILINK." (Johns Hopkins University Press, Baltimore, 1994).
22. R.W. Cottingham Jr., R.M. Idury, and A.A. Schäffer, "Faster sequential genetic linkage computations" *Am J Hum Genet* **53**, 252 (1993)
23. A.A. Schäffer, S.K. Gupta, K. Shriram, and R.W. Cottingham Jr., "Avoiding recomputation in linkage analysis" *Hum Hered* **44**, 225 (1994)
24. J.D. Terwilliger and J. Ott in *Handbook of Human Genetic Linkage*, "Computer Simulation Methods." (Johns Hopkins University Press, Baltimore, 1994).
25. J. Dausset, H. Cann, D. Cohen, M. Lathrop, J.M. Lalouel, and R. White, "Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative genetic mapping of the human genome" *Genomics* **6**, 575 (1990)
26. I. Bross, "Misclassification in 2×2 tables" *Biometrics* **10**, 478 (1954)
27. P.B. Patnaik, "The power function of the test for the difference between two proportions in a 2×2 table" *Biometrika* **35**, 157 (1948)
28. J. Ott in *Analysis of Human Genetic Linkage*, *3rd Edition,* "Interval Estimation." (Johns Hopkins University Press, Baltimore, 1999)