

Identifying disease polymorphisms from case–control genetic association data

L. Park

Received: 5 March 2010 / Accepted: 27 September 2010 / Published online: 14 October 2010
© Springer Science+Business Media B.V. 2010

Abstract In case–control association studies, it is typical to observe several associated polymorphisms in a gene region. Often the most significantly associated polymorphism is considered to be the disease polymorphism; however, it is not clear whether it is the disease polymorphism or there is more than one disease polymorphism in the gene region. Currently, there is no method that can handle these problems based on the linkage disequilibrium (LD) relationship between polymorphisms. To distinguish real disease polymorphisms from markers in LD, a method that can detect disease polymorphisms in a gene region has been developed. Relying on the LD between polymorphisms in controls, the proposed method utilizes model-based likelihood ratio tests to find disease polymorphisms. This method shows reliable Type I and Type II error rates when sample sizes are large enough, and works better with re-sequenced data. Applying this method to fine mapping using re-sequencing or dense genotyping data would provide important information regarding the genetic architecture of complex traits.

Keywords Linkage disequilibrium · Case–control association · Common polymorphism

Introduction

After the successes of the HapMap project (Frazer et al. 2007; Altshuler et al. 2005), genome-wide association studies (GWAS) have enabled the identification of more

candidate genes related to diseases (Wellcome Trust Case Control Consortium 2007; Easton et al. 2007; Maris et al. 2008; Scott et al. 2009; Erdmann et al. 2009; Barrett et al. 2009; Cho et al. 2009). Reviews regarding GWAS presented current concerns and future research strategies, which emphasized the importance of obtaining common, truly functional polymorphisms influencing expression in a locus or protein structure (Altshuler et al. 2008; Hardy and Singleton 2009; Manolio et al. 2009). It is now of primary and imperative interest to find the functional polymorphisms associated with diseases through dense genotyping or re-sequencing of target regions. Typically, several polymorphisms simultaneously show associations in the target locus (Yasuda et al. 2008; Unoki et al. 2008; Wrensche et al. 2009; Shete et al. 2009; Song et al. 2009). The significant associations of many markers in target gene loci are primarily due to linkage disequilibrium (LD) between the markers and the real disease polymorphisms. It is important to use the LD information to separate true associations, but there is no actual method that primarily makes use of LD information to discriminate the associations of real disease polymorphisms from the associations of marker polymorphisms.

Early efforts to discriminate real disease polymorphisms from several associated marker polymorphisms in case–control association studies have made use of conventional statistical solutions to deal with confounders (Wrensche et al. 2009; Nicodemus et al. 2004), and the usual conclusion from current statistical methods is that the most significantly associated polymorphism is the disease polymorphism. Recent efforts include more advanced statistical approaches, such as various regression, ensemble, and network methods (Szymczak et al. 2009; Charoen et al. 2007). However, these recent studies in case–control associations are more focused on detecting main effects or gene–gene interaction in genome-wide association data

L. Park (✉)
Natural Science Research Institute, Yonsei University,
134 Shinchon-Dong, Seodaemun-Ku, Seoul 120-749, Korea
e-mail: lypark@yonsei.ac.kr

than distinguishing real disease polymorphisms in a genetic locus, and they have demonstrated that there are problems identifying disease polymorphisms using current methods. In family based association studies, there have been efforts for identifying polymorphisms that explain a linkage signal (Biernacka and Cordell 2009). Similar efforts were partially demonstrated in case–control association studies based on a step-wise regression, but there were difficulties in differentiating between potentially causal polymorphisms and other polymorphisms (Biernacka et al. 2007; Cordell and Clayton 2002).

The classical methods in statistics for dealing with confounding factors may work less efficiently in the discrimination of real disease polymorphisms from associated markers in LD compared to methods primarily based on the actual LD relationship. Based on the LD relationship, a study was previously tested whether the positive association of most polymorphisms in the *APOE* gene region with Alzheimer's disease comes from the single disease polymorphism, encoding ApoE ϵ 4 (Park 2007). This study found that these associations are difficult to explain with the single disease polymorphism. Expanding this previous effort, a method for identifying disease polymorphisms from the genotypes of cases and controls was developed in the current study. Since there is no information regarding which polymorphisms are the disease polymorphisms, the likelihood ratio tests for various models should be considered to find the most likely set of disease polymorphisms from the data. If a model is accepted, then it probably indicates the set of true disease polymorphisms and can tell us how the gene influences disease presentation.

Methods

Likelihood ratio tests based on models

In this study, all of the genotyped polymorphisms were considered potential disease polymorphisms. To find the most likely set of disease polymorphisms, all possible tests using the genotype data were conducted, ranging from the model with one disease polymorphism to the model with the true number of disease polymorphisms, one of which is usually accepted because it is the correct model. In actual case–control association studies, the tests starts from the model with one disease polymorphism and continued until accepting a model at a certain number of disease polymorphisms, which is considered as the true number of disease polymorphisms. For each model, the expected allele frequencies in cases were calculated based on the control allele frequency data, the case allele frequency data of the targeted disease polymorphism(s), and the LD

between the markers and the disease polymorphism in controls. Likelihood ratio tests were then conducted for the expected case allele frequencies using the case genotype data.

When testing the model with only one disease polymorphism, the calculation is simple. It involves adding the portion changed due to the disease polymorphism based on the LD relationship to the control marker allele frequency. When there is one real disease polymorphism, the odds ratio of the disease allele is directly estimated; the actual odds ratio of the disease allele is the same as the observed odds ratio. When the model contains two or more disease polymorphisms, the calculation of the expected allele frequencies of polymorphisms in cases is more complicated. In the cases, the observed allelic odds ratios of disease polymorphisms calculated from the genotype data are not the same as the real odds ratio, given that the frequency of a disease polymorphism is influenced by the frequencies of other disease polymorphism(s) due to LD between disease polymorphisms. Therefore, the independent odds ratios of disease polymorphisms should be estimated first. Based on the independent odds ratios, the expected frequencies of markers were estimated from the LD relationship with the disease polymorphisms. The general expression of marker allele frequencies in cases is expressed as shown in Eq. 1.

$$p_{M'_i} = p_{M_i} + \sum_j \Delta_j \frac{p_{M_i d_j} - p_{M_i} p_{d_j}}{p_{d_j}(1 - p_{d_j})} \quad (1)$$

In Eq. 1, $p(M'_i)$ indicates the marker allele frequency in cases; $p(M_i)$ indicates the marker allele frequency in controls; $p(M_i d_j)$ is the frequency of haplotype $M_i d_j$; $p(d_j)$ indicates the disease allele frequency in controls; delta indicates the genuine differences in disease allele frequency between controls and cases, in which the effects of all other disease polymorphisms in LD are excluded. Therefore, the disease allele frequency in cases can be used to derive the real odds ratio of the disease polymorphism from delta. By solving the following matrix as shown in Eq. 2, the delta values can be obtained.

$$\begin{pmatrix} p_{d'_1} - p_{d_1} \\ p_{d'_2} - p_{d_2} \\ \vdots \end{pmatrix} = \begin{pmatrix} 1 & \frac{p_{d_1 d_2} - p_{d_1} p_{d_2}}{p_{d_2}(1 - p_{d_2})} & \cdots \\ \frac{p_{d_1 d_2} - p_{d_1} p_{d_2}}{p_{d_1}(1 - p_{d_1})} & 1 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \end{pmatrix} \quad (2)$$

In some cases, there may be no solution for delta or the estimated allele frequencies and/or delta may be out of the appropriate range. These cases indicate that the underlying model with the targeted disease polymorphisms is not appropriate. It should therefore be excluded in the search for the correct model with the real disease polymorphisms.

The hypotheses for a set of selected potential disease polymorphisms can be tested using likelihood ratio tests over the case genetic data. For a given hypothesis, the frequency changes in other polymorphisms are due to frequency changes in the selected polymorphism(s) and the LD relationship with the selected polymorphism(s). Based on this hypothesis, the case allele frequencies are calculated as previously indicated in Eq. 1. It should be noted that the allele frequencies in controls are not real control population frequencies because the control allele frequencies come from a sampling of the target population. Therefore, there should be a correction term for the likelihood ratio tests.

As indicated, because parameters of the underlying hypothesis in likelihood ratio tests are variable due to sampling, correction for variance is necessary in the likelihood ratio tests. When the binomial distribution is approximated as a normal distribution, the following Eq. 3 is derived. In this equation, θ_0 is the parameter derived from the selected model indicated previously; $np(1-p)$ is the variance of the maximum likelihood estimate; σ^2 is the variance of θ_0 . The sum of two normal random variables, θ_0 and $\hat{\theta}$, is normally distributed, and the last term of the multiplication in the top equation in Eq. 3 converges to a chi-squared distribution with one degree of freedom. Therefore, the bottom equation has an approximately chi-squared distribution with one degree of freedom.

$$\begin{aligned}
 -2 \log \left(\frac{L(\theta_0)}{L(\hat{\theta})} \right) &= \frac{(\hat{\theta} - \theta_0)^2}{np(1-p)} = \frac{np(1-p) + \sigma^2}{np(1-p)} \\
 &\times \frac{(\hat{\theta} - \theta_0)^2}{np(1-p) + \sigma^2} \quad (3) \\
 -2 \log \left(\frac{L(\theta_0)}{L(\hat{\theta})} \right) &\times \frac{np(1-p)}{np(1-p) + \sigma^2} \sim \chi_1^2
 \end{aligned}$$

The variance of θ_0 , σ^2 , can be estimated approximately through simulations. For simulations, the observed parameters from the sample were used instead of the real parameters from the population. Simulations were repeated 1,000 times to estimate the approximate variances. This simulation was performed for each polymorphism in a model with selected disease polymorphisms. Therefore, there were n chi-squared distributions, with n being the number of polymorphisms excluding the disease polymorphism(s). The sum of the distributions is the chi-squared distribution with n degrees of freedom, and it was the actual test in this method. The software that implements the computation of likelihood ratio test statistics is available in the R statistical package under the package name “IFP” (Identifying Functional Polymorphisms).

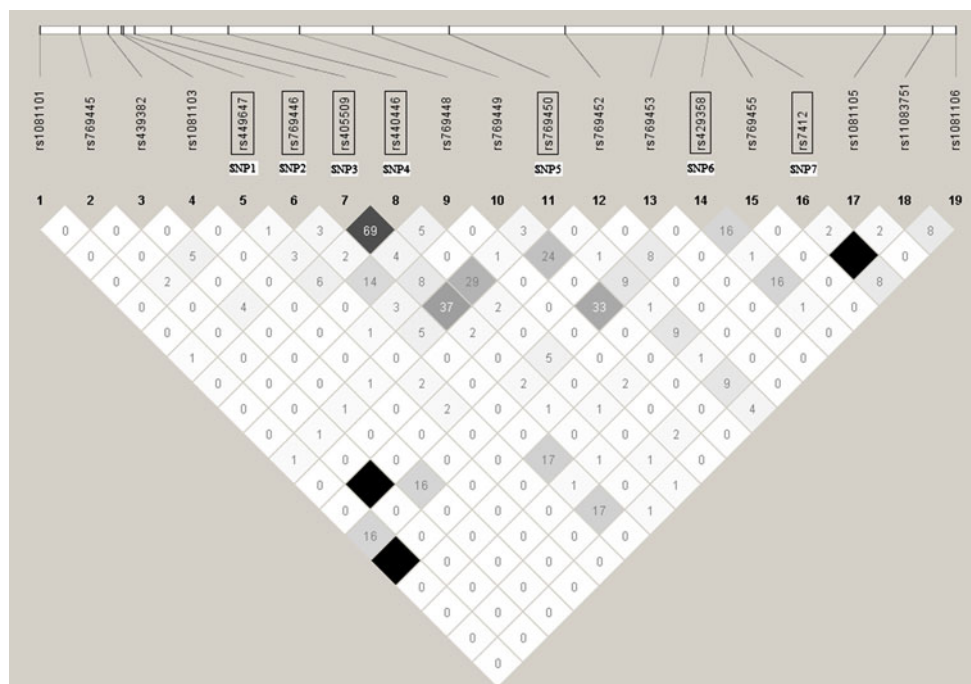
Simulation studies for estimating the type I and II error rates

To generate the virtual data set, the re-sequenced public data of the *APOE* region were used (<http://www.droog.gs.washington.edu/mdecode/data/apoe/>) (Nickerson et al. 2000). The data provided sequencing results for 48 individuals, and there were 19 polymorphic sites (Fig. 1; Qin et al. 2002; Barrett et al. 2005). Three pairs of these sites were in complete LD with one another, as indicated in black in Fig. 1. Each one of the pairs in complete LD was eliminated for this study because they produced a duplicate result for another SNP in complete LD. When the control sample size is large enough, the probability of observing polymorphisms in complete LD decreases without requiring the removal of portions of the data. Among the 16 polymorphic sites remaining after the elimination, only seven single nucleotide polymorphisms (SNPs) had a minor allele frequency higher than 0.1. These polymorphisms in the gene region showed both high and low LD between polymorphisms. Therefore, the effect of both high and low LD could be examined in this study. The seven polymorphisms with frequencies higher than 0.1 were used to generate primary data, and the 16 polymorphisms were used for examining the effect of using re-sequenced data.

For the simulation, control sample data were derived from binomial sampling using the frequencies of the *APOE* data, depending on the control sample sizes, to obtain the control allele frequencies and LD between the polymorphisms. The haplotype frequencies of the *APOE* data were estimated using haplo.stats in the R package (Yu and Schaid 2007; Schaid et al. 2002). To generate case data, various sets of disease polymorphisms were selected from the seven polymorphisms. Many recent results of GWAS discovered associated polymorphisms with odds ratios of ~ 1.5 (Manolio et al. 2009). Since it is expected that further fine mapping using dense genotyping or re-sequencing might reveal associations with odds ratio higher than 1.5, the independent odds ratios of each disease polymorphism were fixed at 2.0. Using Eq. 2, the observed differences in allele frequencies of disease polymorphisms between cases and controls were calculated based on the original differences in each polymorphism (independently of other disease polymorphisms). Additionally, using Eq. 1, the expected allele frequencies of other polymorphisms in cases were calculated. These frequencies and the case sample size were used to generate case data from binomial sampling.

Determination of the correct model is based on the results of tests with various models; therefore, both Type I and II error rates are important in this approach, concerning the intrinsic multiple testing in this method. To estimate the

Fig. 1 LD (r^2) between all polymorphic sites from re-sequenced *APOE* data (the seven SNPs with minor allele frequencies higher than 0.1 are indicated as *squares*)



Type I and Type II error rates, the generation of case and control data was repeated 1,000 times for each set of disease polymorphism(s) with a fixed odds ratio(s). The likelihood ratio tests were applied to each set of data for various models, assuming sets of the true disease polymorphism(s). For instance, assume that the true model is that SNP2 is the only disease polymorphism among seven SNPs with an allelic odds ratio of 2.0. Seven possible models can be tested for a single disease polymorphism. Among these seven models, the true model is the one in which the disease polymorphism is SNP2. The $-2 \log$ (likelihood ratio) with variance correction is distributed approximately as a chi-squared distribution (Eq. 3), from which the Type I error rate can be estimated. The other six models are incorrect models for the data; their likelihood ratio tests are rejected in most cases, providing reasonably low Type II error rates.

The model with real disease polymorphism(s) was tested to obtain the Type I error rate, and all other possible models were tested to obtain the Type II error rate. In these tests it is inevitable that the value of the $-2 \log$ (likelihood ratio) with variance correction becomes small enough to accept the model whenever the model involves all of the real disease polymorphisms. Some models with a greater number of disease polymorphisms than the real number include all of the real disease polymorphisms; the acceptance of these models results in Type II errors. Therefore, the tests for identifying the appropriate model should stop at the actual number of disease polymorphism(s), at which point the right model is usually accepted. In inferring the appropriateness of the accepted model, examination of all

of the likelihood tests could be helpful from one disease polymorphism to N disease polymorphisms, where N is the total number of polymorphic sites in the region.

The closest previous method for identifying real disease polymorphisms, the stepwise regression method (Biernacka et al. 2007; Cordell and Clayton 2002), was performed using the simulated data to compare with the current method. Simulated data were generated based on the same *APOE* data assuming causal polymorphisms with fixed odds ratios, similar to the generation of the virtual data for the current method. The original application of stepwise regression to case–control association study was based on the genotype effect rather than allelic effect. Since the alleles should be the variables in the regression instead of genotypes for the comparison with the current method, the relationship between alleles (haplotypes) should be known. Haplotypes based on the *APOE* data were obtained using the R statistical package, haplo.stats (Yu and Schaid 2007; Schaid et al. 2002), and used for generating the control and case data, in which the alleles were coded as -1 and 1 for the regression. Sample sizes were 500 cases and 500 controls (1,000 case haplotypes and 1,000 control haplotypes), and simulations were repeated 1,000 times.

The effects of the sample size and odds ratio were examined with regard to error rates. Currently, a large control sample size is common due to huge population genetic studies and large-scale case–control studies. A variable number of controls (from 500 to 10,000) and cases (from 500 to 5,000) were examined to determine the appropriate sample sizes for circumstances with high error rates. Considering recent GWAS results, variable odds

ratios from 1.2 to 3.0 were applied to examine the effect of odds ratio on error rates. Although computation time is longer, using fully re-sequenced data might reduce Type II error by providing more data to test the appropriateness of the model. Therefore, these types of data were also examined for Type I and Type II error rates. Finally, a study of the *APOE* association with Alzheimer's disease was examined using the current approach.

Results

Number of disease polymorphisms in a gene region

Various situations with a varying number of disease polymorphisms were examined using a sample size of 500 for both cases and controls. Each situation of all possible sets of disease polymorphism(s) was simulated in this study for one, two, and three disease polymorphism(s). The actual disease polymorphism(s) in each simulation is indicated in bold characters in Tables 1, 2, 3, and the independent allelic odds ratios were fixed as 2.0 for each polymorphism. When there was only one disease polymorphism, the observed odds ratio of the disease polymorphism was the same as the true odds ratio; in this case the most significantly associated polymorphism was the disease polymorphism (Table 1). When there was more than one disease polymorphism, however, the observed odds ratios differed from the actual odds ratios as calculated by Eq. 2 in the “Materials and methods” section (Tables 2 and 3). These differences result from the LD relationship between disease polymorphisms, and they are substantially reduced or increased in several cases. Moreover, polymorphisms not associated with diseases often show more significant observed associations than disease polymorphisms.

The estimated Type I error rates were reasonable overall, although their average was slightly higher than the nominal error rates of $\alpha = 0.05$ and $\alpha = 0.01$ (Tables 1, 2, 3). This might be because the variations in the model parameters were estimated by simulations based on observed data rather than true data. Therefore, the limited sample sizes likely produced less variability in simulations with extreme frequencies. The Type II error rates were low overall but varied depending on the set of true disease polymorphisms (Tables 1, 2, 3), indicating that the power to identify the set of true disease polymorphism(s) could be quite high for most genetic association data.

Table 1 shows that the Type II error rates were higher when SNP7 was the disease polymorphism than they were for the other SNPs. SNP1 is in moderate LD with SNP2 (r^2 : 0.178), and the model in which SNP2 is the disease polymorphism looks slightly similar to the model in which SNP7 is the disease polymorphism. However, the value of the $-2 \log$ (likelihood ratio) with variance correction of the true model is almost always lower than that of the false model (91.5% probability from simulations). Therefore, if those values are considered for identifying disease polymorphisms, the actual Type II error rates can be greatly reduced when SNP7 is the disease polymorphism and even for other true models. The Type II error rates when SNP2 is the disease polymorphism are not as high as the rates when SNP7 is the disease polymorphism because of the higher minor allele frequency and moderately high D' of SNP2 with other polymorphisms.

As shown for several cases in Tables 2 and 3, the Type II error rates become large when there is more than one real disease polymorphism. The representative case occurs when SNP3 and SNP4 are the disease polymorphisms. The r^2 between these two SNPs is high (0.693); therefore, the minor alleles of the two SNPs are highly associated with each other. An odds ratio of 2.0, however, was applied to

Table 1 Estimates (Based on 1,000 replicated samples) of the Type I and Type II error rates for nominal Type I Error α and Type II Error β when there is only one disease polymorphism with an odds ratio of 2.0 and sample size of 500 for both cases and controls

Index	Minor allele frequency of the disease SNP	Odds ratio of each SNP							Type I error		Type II error	
		SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	$\alpha = .05$	$\alpha = .01$	$\beta = .05$	$\beta = .01$
1	0.28 ^a	2.00	1.11	1.11	0.86	1.03	1.03	0.97	0.054	0.012	0.002	0.014
2	0.13	1.09	2.00	0.89	1.10	0.87	1.03	0.71	0.054	0.011	0.023	0.042
3	0.33 ^a	1.13	0.85	2.00	0.54	1.41	0.87	0.93	0.074	0.021	0.092	0.138
4	0.28	0.83	1.17	0.56	2.00	0.67	0.76	0.88	0.045	0.008	0.078	0.116
5	0.43	1.04	0.76	1.64	0.63	2.00	0.67	1.48	0.046	0.008	0.002	0.008
6	0.11	1.02	1.03	0.91	0.86	0.83	2.00	1.13	0.058	0.008	0.001	0.006
7	0.11 ^a	0.99	0.79	0.97	0.95	1.11	1.07	2.00	0.068	0.013	0.164	0.388

Bold Odds ratio of the real disease SNP

^a The odds ratio is applied to the major allele

Table 2 Estimates (Based on 1,000 replicated samples) of the Type I and Type II error rates for nominal Type I Error α and Type II Error β when there are two disease polymorphisms, each with an odds ratio of 2.0 and sample size of 500 for both cases and controls

Index	Set of disease polymorphisms	Odds ratio of each SNP							Type I error		Type II error	
		SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	$\alpha = .05$	$\alpha = .01$	$\beta = .05$	$\beta = .01$
1	1,2	2.29	2.13	0.99	0.96	0.89	1.06	0.70	0.044	0.006	0.011	0.029
2	1,3	2.41	0.95	2.33	0.44	1.45	0.89	0.91	0.052	0.014	0.045	0.103
3	1,4	1.54	1.28	0.62	1.78	0.69	0.78	0.86	0.061	0.016	0.046	0.109
4	1,5	2.13	0.86	1.88	0.52	2.06	0.70	1.43	0.047	0.013	0.009	0.027
5	1,6	2.06	1.14	1.01	0.73	0.86	2.03	1.10	0.036	0.005	0.009	0.025
6	1,7	1.97	0.89	1.08	0.82	1.14	1.10	1.91	0.061	0.015	0.080	0.150
7	2,3	1.24	1.81	1.72	0.62	1.23	0.90	0.67	0.065	0.017	0.056	0.090
8	2,4	0.90	2.21	0.51	2.17	0.57	0.79	0.64	0.056	0.014	0.056	0.088
9	2,5	1.14	1.69	1.43	0.71	1.73	0.70	0.95	0.048	0.005	0.003	0.011
10	2,6	1.12	2.04	0.82	0.96	0.72	2.04	0.78	0.052	0.010	0.004	0.011
11	2,7	1.08	1.74	0.87	1.06	0.96	1.10	1.16	0.048	0.010	0.162	0.278
12	3,4	0.93	1.01	0.98	1.28	0.97	0.63	0.82	0.056	0.015	0.407	0.626
13	3,5	1.18	0.61	4.49	0.26	2.90	0.55	1.34	0.079	0.021	0.061	0.104
14	3,6	1.15	0.87	1.76	0.44	1.18	1.83	1.04	0.072	0.019	0.123	0.171
15	3,7	1.11	0.64	1.92	0.51	1.57	0.93	1.77	0.076	0.019	0.240	0.362
16	4,5	0.86	0.92	0.85	1.41	1.37	0.45	1.24	0.052	0.008	0.055	0.119
17	4,6	0.85	1.20	0.52	1.78	0.55	1.70	0.98	0.069	0.018	0.170	0.253
18	4,7	0.82	0.95	0.55	1.92	0.75	0.82	1.60	0.085	0.023	0.264	0.373
19	5,6	1.06	0.79	1.47	0.52	1.67	1.60	1.76	0.066	0.015	0.067	0.129
20	5,7	1.03	0.56	1.58	0.59	2.23	0.74	4.76	0.062	0.013	0.094	0.184
21	6,7	1.01	0.82	0.89	0.82	0.93	2.09	2.51	0.048	0.004	0.062	0.134

Bold Observed odds ratio of the disease SNP

the major allele of SNP3 and the minor allele of SNP4. As a result, the observed odds ratios of SNP3 and SNP4 were diminished to 0.98 and 1.28, respectively. These values are lower than most other SNPs due to the counterbalancing effect of each of these SNPs on the other (Table 2).

With three real disease polymorphisms, a similar phenomenon can be seen when including both SNP3 and SNP4 as the disease polymorphisms (Table 3). Therefore, there are certain situations that make it difficult to detect associations and identify the disease polymorphisms. It should be noted that this situation would be the worst-case scenario, which might not be very common. In the opposite of this situation, the Type II error rates can be low. The high r^2 between SNP3 and SNP4 is also responsible for increasing the Type II error rates in several cases involving one of the SNPs as a disease polymorphism. However, this effect has a less significant impact on increasing the Type II error rates (Tables 2 and 3).

For the same situations in Table 1 and 2, the stepwise regressions were conducted for the comparison with the current method. As suggested (Cordell and Clayton 2002), the backward stepwise regression were performed, and the final model was examined compared to the true model. Frequencies of selecting the correct model as the final

model were overall very low ranging from 0 to 0.243 based on 1,000 simulations. When each regression result was examined one by one, the real disease polymorphism usually showed the most significant coefficient when there was one real disease polymorphism. For two disease polymorphisms, one or both of the disease polymorphisms were often eliminated in the final model. The result confirms that the current method is more efficient for identifying the real disease polymorphisms.

Effect of sample size

Increasing the sample size is very helpful to reduce Type II errors in this test. Since the overall Type II error rates were low, the situations with high Type II errors were of primary interest. The control sample sizes were increased from 500 to 10,000, and the case sample sizes were increased from 500 to 5,000. Three different combinations of disease SNPs were examined for their Type II error rates at various sample sizes. Two of these had the worst Type II error rates for two or three true disease polymorphisms; these were caused by the counterbalancing effects of SNP3 and SNP4. The other combination of disease SNPs studied for sample size effects had moderately high Type II error rates with

Table 3 Estimates (Based on 1,000 replicated samples) of the Type I and Type II error rates for nominal Type I Error α and Type II Error β when there are three disease polymorphisms, each with an odds ratio of 2.0 and sample size of 500 for both cases and controls

Index	Set of disease polymorphisms	Odds ratio of each SNP							Type I error		Type II error	
		SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	$\alpha = .05$	$\alpha = .01$	$\beta = .05$	$\beta = .01$
1	1,2,3	2.81	1.93	1.97	0.51	1.26	0.92	0.66	0.064	0.013	0.063	0.131
2	1,2,4	1.72	2.35	0.56	1.93	0.59	0.81	0.63	0.052	0.013	0.063	0.133
3	1,2,5	2.44	1.82	1.62	0.59	1.78	0.73	0.93	0.050	0.013	0.012	0.034
4	1,2,6	2.36	2.17	0.90	0.82	0.74	2.07	0.76	0.039	0.005	0.012	0.034
5	1,2,7	2.24	1.86	0.96	0.91	0.99	1.13	1.13	0.056	0.014	0.098	0.159
6	1,3,4	1.80	1.11	1.09	1.12	0.99	0.66	0.80	0.053	0.022	0.205	0.346
7	1,3,5	2.58	0.71	5.98	0.17	3.00	0.57	1.30	0.101	0.057	0.078	0.161
8	1,3,6	2.49	0.98	2.03	0.34	1.22	1.86	1.02	0.059	0.020	0.111	0.188
9	1,3,7	2.36	0.74	2.22	0.41	1.61	0.96	1.69	0.091	0.026	0.176	0.284
10	1,4,5	1.62	1.02	0.94	1.23	1.41	0.47	1.21	0.041	0.015	0.041	0.105
11	1,4,6	1.58	1.31	0.57	1.57	0.56	1.73	0.95	0.043	0.011	0.129	0.230
12	1,4,7	1.52	1.05	0.60	1.71	0.77	0.85	1.54	0.077	0.019	0.183	0.291
13	1,5,6	2.19	0.89	1.66	0.41	1.72	1.63	1.69	0.054	0.012	0.050	0.108
14	1,5,7	2.09	0.65	1.80	0.48	2.30	0.76	4.31	0.075	0.024	0.073	0.130
15	1,6,7	2.02	0.92	0.98	0.69	0.96	2.12	2.38	0.053	0.010	0.056	0.113
16	2,3,4	1.01	2.01	0.88	1.40	0.84	0.66	0.61	0.051	0.022	0.300	0.430
17	2,3,5	1.30	1.51	3.48	0.32	2.48	0.58	0.89	0.044	0.013	0.056	0.100
18	2,3,6	1.27	1.84	1.53	0.51	1.03	1.87	0.73	0.054	0.009	0.088	0.134
19	2,3,7	1.22	1.55	1.65	0.59	1.36	0.96	1.07	0.066	0.017	0.212	0.302
20	2,4,5	0.94	1.89	0.76	1.54	1.19	0.48	0.84	0.046	0.012	0.049	0.103
21	2,4,6	0.92	2.25	0.47	1.93	0.46	1.74	0.70	0.051	0.014	0.138	0.209
22	2,4,7	0.89	1.94	0.49	2.09	0.64	0.85	1.00	0.052	0.011	0.255	0.358
23	2,5,6	1.16	1.73	1.29	0.59	1.45	1.63	1.07	0.070	0.019	0.047	0.085
24	2,5,7	1.13	1.45	1.38	0.67	1.93	0.77	1.83	0.049	0.015	0.064	0.113
25	2,6,7	1.10	1.77	0.79	0.91	0.80	2.13	1.33	0.041	0.007	0.084	0.151
26	3,4,5	0.97	0.77	1.60	0.85	1.93	0.33	1.14	0.087	0.058	0.264	0.371
27	3,4,6	0.95	1.04	0.89	1.12	0.80	1.54	0.91	0.026	0.012	0.553	0.694
28	3,4,7	0.92	0.80	0.95	1.22	1.07	0.69	1.44	0.046	0.014	0.588	0.771
29	3,5,6	1.20	0.64	3.63	0.17	2.39	1.44	1.57	0.092	0.041	0.211	0.308
30	3,5,7	1.16	0.42	4.18	0.23	3.28	0.61	3.68	0.131	0.063	0.189	0.278
31	3,6,7	1.14	0.67	1.70	0.41	1.31	1.92	2.16	0.078	0.018	0.228	0.325
32	4,5,6	0.88	0.95	0.78	1.24	1.15	1.32	1.44	0.044	0.011	0.258	0.407
33	4,5,7	0.85	0.71	0.82	1.35	1.52	0.51	3.07	0.083	0.032	0.148	0.228
34	4,6,7	0.84	0.98	0.50	1.71	0.62	1.78	1.93	0.065	0.021	0.290	0.383
35	5,6,7	1.05	0.59	1.42	0.48	1.86	1.68	8.67	0.070	0.026	0.131	0.209

Bold Observed odds ratio of the disease SNP

three disease polymorphisms. This moderately high Type II error rate came from the high LD between SNP3 and SNP4. As shown in Table 4a, the Type II error rates in this case were reduced quickly as sample size increased. Increasing the control sample size was more effective for reducing the Type II error rates than increasing the case sample size. As the sample sizes increased, the two combinations of disease SNPs with the worst initial Type II

error rate for two or three true disease polymorphisms also showed fairly large improvements. In those cases, however, the Type II error rates were substantially reduced when both the case and control sample sizes were large enough. In summary, by increasing the sample sizes to a certain extent depending on situation, this method provided a reliable test identifying true disease polymorphisms, even for the worst-case situations.

Table 4 Type II error rates for various case and control sample sizes: (a) when there are three disease polymorphisms (SNP2, SNP4, and SNP7) with moderately high Type II error; (b) when there are two disease polymorphisms with the highest Type II error (SNP3 and SNP4); (c) when there are three disease polymorphisms with the highest Type II error (SNP3, SNP4, and SNP7)

Control sample size	500		1,000		2,000		5,000		10,000	
	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$
(a)										
500	0.26	0.36	0.16	0.25	0.11	0.18	0.08	0.14	0.07	0.12
1,000	0.22	0.31	0.11	0.18	0.06	0.11	0.03	0.06	0.02	0.04
2,000	0.19	0.27	0.09	0.15	0.04	0.07	0.01	0.03	0.01	0.02
5,000	0.18	0.25	0.08	0.13	0.03	0.05	0.01	0.01	0.00	0.00
(b)										
500	0.41	0.61	0.27	0.45	0.22	0.36	0.18	0.30	0.18	0.29
1,000	0.31	0.50	0.17	0.28	0.11	0.18	0.08	0.14	0.08	0.13
2,000	0.24	0.41	0.12	0.20	0.07	0.11	0.04	0.07	0.03	0.06
5,000	0.21	0.37	0.09	0.15	0.04	0.07	0.01	0.03	0.00	0.01
(c)										
500	0.58	0.77	0.44	0.64	0.32	0.51	0.27	0.43	0.24	0.40
1,000	0.47	0.67	0.27	0.44	0.17	0.30	0.12	0.20	0.10	0.17
2,000	0.41	0.61	0.20	0.33	0.10	0.18	0.05	0.09	0.04	0.07
5,000	0.35	0.54	0.14	0.24	0.05	0.10	0.01	0.03	0.00	0.01

Effect of the odds ratio

The test can be affected by the odds ratios of disease polymorphisms. As previous association studies resulted in associated SNPs with varying odds ratios, it is worthwhile to examine the effect of the odds ratios in the current

method. As shown in Table 5, several sets of disease polymorphisms with high Type II error rates were tested with changes in odds ratios ranging from 1.2 to 3.0. A set of disease SNPs with a high Type II error rate (SNP3 and SNP4) and a set with a moderately high Type II error rate (SNP2, SNP4, and SNP7) were examined for variable odds

Table 5 Type II error rates for various odds ratios when there are two disease polymorphisms (2,000 controls and 2,000 cases were used for simulation; OR: odds ratio): (a) the set of SNP3 and SNP4 as disease polymorphisms with the highest Type II error; (b) the set of SNP2, SNP4, and SNP7 as disease polymorphisms with moderately high Type II error (the odds ratio of SNP7 is fixed at 2)

OR of SNP4	1.2		1.5		2.0		2.5		3.0	
	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$
(a)										
OR of SNP3	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$
1.2	0.79	0.91	0.28	0.37	0.20	0.25	0.20	0.23	0.19	0.23
1.5	0.32	0.45	0.31	0.48	0.09	0.15	0.08	0.13	0.07	0.12
2.0	0.21	0.26	0.08	0.15	0.07	0.11	0.03	0.05	0.03	0.05
2.5	0.20	0.25	0.06	0.10	0.09	0.15	0.03	0.05	0.02	0.03
3.0	0.19	0.24	0.05	0.09	0.03	0.07	0.03	0.06	0.01	0.03
(b)										
OR of SNP4	1.2		1.5		2.0		2.5		3.0	
OR of SNP2	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$
1.2	0.23	0.32	0.23	0.33	0.23	0.32	0.24	0.33	0.26	0.35
1.5	0.09	0.15	0.10	0.18	0.11	0.18	0.12	0.19	0.13	0.22
2.0	0.05	0.08	0.04	0.08	0.04	0.07	0.04	0.07	0.05	0.08
2.5	0.06	0.09	0.05	0.08	0.04	0.07	0.04	0.07	0.05	0.08
3.0	0.07	0.11	0.06	0.10	0.04	0.08	0.05	0.08	0.05	0.08

ratios. For the set of three disease SNPs, the odds ratios of SNP2 and SNP4 varied; the odds ratio of SNP7 was fixed. As the examined models had high Type II error rates, this method is expected to reliably identify disease polymorphisms with various odds ratios higher than 1.5 (Table 5). Low odds ratios (e.g., 1.2) resulted in high Type II error rates. Increased odds ratios reduced the Type II error rates, but the decrements were not as consistent as those in response to changes in the sample size. Reductions of Type II error rates impeded at a certain point as the odds ratios were increased (data not shown). Overall, when the Type II error rate was low enough, increased odds ratios did not affect the Type II error rates very much. For sets of disease SNPs with high Type II error rates (Table 5), however, increased odds ratio decreased the Type II errors. Although the trend is not always true, it is notable that high observed odds ratios or large differences in allele frequencies between cases and controls may be more important for reducing the Type II error rates than high independent odds ratios for each polymorphism; this is consistent with the data shown in the previous section with variable sets of true disease polymorphisms (Tables 1, 2, 3). In comparison to sample size, therefore, changes in the odds ratios are not as crucially important for the accuracy of the tests.

Using re-sequenced data

High Type II error rates in several bad situations can be improved by obtaining more information from fully re-sequenced data. Re-sequenced data include rare polymorphic sites and common polymorphic sites. These rare polymorphisms are usually in complete D' and low r^2 with other polymorphisms, and this LD information can be very useful for these model-based likelihood ratio tests. Using all 16 SNPs from the re-sequenced data (Table 6), similar tests were conducted for various sets of disease polymorphisms.

First, to examine the reduction in Type II error rates resulting from the use of re-sequenced data, simulation tests were conducted using 500 cases and 2,000 controls. Increased numbers of controls were used to ensure stable results for normal approximation. As in previous tests, each allelic odds ratio was fixed at 2.0. Table 7a demonstrates that the use of 16 instead of seven SNPs resulted in a substantial reduction in Type II error rates, even though the reductions differed depending on the set of disease SNPs. As previously indicated, examination of the lowest value of the $-2 \log$ (likelihood ratio) with variance correction can also be helpful to obtain the correct model. The reduced Type II error rates result from the inclusion of more data in the tests. Expanding the re-sequenced region would produce better outcomes by providing more data with appropriate frequencies.

Table 6 Minor allele frequencies of the 16 SNPs tested for examining re-sequenced data and their frequencies in comparison to results from an *APOE* association study with Alzheimer's disease

Index	rs number	Minor allele frequency	Alleles ^b	Association	
				Controls	LOAD ^c
1	rs439382	0.021 ^a	A:G	0.004	0.005
2	rs1081103	0.010 ^a	C:T	0.000	0.000
3	rs449647	0.281 ^a	A:T	0.187	0.114
4	rs769446	0.125	T:C	0.090	0.091
5	rs405509	0.333 ^a	G:T	0.500	0.567
6	rs440446	0.281	G:C	0.379	0.250
7	rs769448	0.021 ^a	C:T	0.030	0.034
8	rs769449	0.042	G:A	0.121	0.319
9	rs769450	0.427	G:A	0.377	0.288
10	rs769452	0.010	T:C	0.004	0.003
11	rs769453	0.010 ^a	C:G	0.000	0.000
12	rs429358	0.115	T:C	0.155	0.427
13	rs769455	0.021 ^a	C:T	0.000	0.003
14	rs7412	0.115 ^a	C:T	0.084	0.028
15	rs1081105	0.021 ^a	A:C	0.030	0.088
16	rs1081106	0.031	T:C	0.084	0.073

Bold seven SNPs with a minor allele frequency higher than 0.1

^a An odds ratio of 2.0 is applied to the major allele

^b The second allele is the minor allele

^c Late-Onset Alzheimer's Disease

Caution regarding the control sample size should be applied when using re-sequenced data. Depending on the model, including very rare polymorphisms might decrease parameter variances estimated using simulations since there are situations that do not have many possible options for the simulated sets. In addition, the successful normal approximation of binomial variables depends on both frequency and sample size. Therefore, extremely low minor allele and haplotype frequencies may result in increased Type I error rates (Table 7b) if the sample sizes are not sufficiently large. Polymorphic sites with adequate allele and haplotype frequencies that can be handled properly using the given control sample size from re-sequenced data should be included in the tests as data. The inclusion of many polymorphisms increases the number of testing models. Even though the Type II error rates were reduced in re-sequenced data, the actual number of false positive results might be increased proportional to the increased number of testing models. Therefore, depending on sample sizes, extremely rare polymorphisms may be excluded in the data unless they were suspected as disease polymorphisms.

To examine the error rates when rare polymorphisms are the real disease polymorphisms, tests for sets of rare disease polymorphisms as well as rare and common disease

Table 7 Results using re-sequenced data (SNP numbers in the SNP set indicate the same SNP number as previously presented using the seven common polymorphisms, except in (c))

(a)				
Disease SNP set	7 SNPs		All SNPs	
	$\beta = .05$	$\beta = .01$	$\beta = .05$	$\beta = .01$
3,4	0.223	0.359	0.068	0.150
4,7	0.092	0.187	0.024	0.060
2,3,7	0.106	0.167	0.023	0.046
3,4,6	0.431	0.571	0.084	0.028
3,4,7	0.331	0.509	0.080	0.188
4,5,6	0.127	0.226	0.029	0.075

(b)								
Disease SNP set	1,000		2,000		5,000		10,000	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
2,7	0.055	0.015	0.049	0.019	0.047	0.009	0.044	0.007
3,4	0.072	0.039	0.069	0.035	0.078	0.036	0.084	0.045
3,6	0.185	0.077	0.128	0.048	0.060	0.017	0.059	0.019
3,7	0.435	0.274	0.315	0.172	0.174	0.069	0.086	0.032
4,6	0.069	0.021	0.062	0.015	0.057	0.024	0.050	0.017
4,7	0.089	0.029	0.067	0.018	0.039	0.006	0.041	0.009

(c)					
Type of Disease SNP	Disease SNP set	Type I error		Type II error	
		$\alpha = .05$	$\alpha = .01$	$\beta = .05$	$\beta = .01$
Rare only	15, 16	0.034	0.010	0.074	0.101
	8, 16	0.177	0.064	0.000	0.000
	2, 7	0.198	0.081	0.512	0.784
	1, 7, 10	0.111	0.036	0.067	0.145
	10, 13, 15	0.134	0.040	0.064	0.128
Rare & common	8, 14	0.060	0.016	0.001	0.002
	9, 10	0.171	0.060	0.004	0.008
	1, 8, 14	0.109	0.040	0.021	0.026
	5, 10, 13	0.219	0.086	0.025	0.037
	4, 7, 14	0.092	0.031	0.023	0.027
Common only	8, 9, 14	0.196	0.087	0.000	0.001
	5, 6	0.073	0.032	0.004	0.010
	4, 5, 6	0.090	0.023	0.001	0.002
	4, 5, 14	0.191	0.080	0.002	0.004

(a) Comparison of Type II error rates for the 7 common polymorphisms versus all 16 polymorphisms (500 cases and 2,000 controls were used for the simulation) for various sets with high Type II error. (b) Type I error rates using all 16 polymorphisms for various sets of seven common polymorphisms with high Type II errors (a case sample size of 500 was used for the simulation). (c) Type I and II error rates when rare polymorphisms serve as the disease polymorphisms (2,000 cases and 2,000 controls were used for the simulation; SNP numbers in the SNP set indicate the same SNP number as noted in Table 6)

polymorphisms were conducted using control and case sample sizes of 2,000. As shown in Table 7c, the patterns of error rates when one or more disease polymorphisms were rare were not different from the patterns observed when only common polymorphisms were the real disease

polymorphisms. Although not always the case, extremely rare polymorphisms serving as disease polymorphisms may result in high Type I and/or Type II error rates (Table 7c, the set of SNP 2 and 7). In these situations, the increment or decrement of the actual disease polymorphisms was not

obvious compared to the frequency changes of other common polymorphisms in LD with the disease polymorphisms. As previously indicated, several sets of disease polymorphisms showed increased Type I error rates, which could be decreased as the control sample size was increased. These results show that this method is applicable to re-sequenced data involving rare disease polymorphisms.

Application to the *APOE* association with Alzheimer's disease

The current method was applied to the *APOE* association with Alzheimer's disease. The results of many previous association studies were available, and the study examining the largest number of SNPs near the *APOE* region was selected (Yu et al. 2007). They typed 50 SNPs using 232 controls and 193 cases. As shown in Table 6, the frequencies of rare SNPs in the association study do not match well with the re-sequenced data used in this study. Therefore, the genotype data for the seven common SNPs primarily used in the current study were selected for the tests. Since their study did not provide specified linkage disequilibrium data, based on the same data previously used in the current study, 10,000 controls were generated for the test. Likelihood ratio tests were conducted on the models ranging from a single disease SNP to seven disease SNPs. When the models with five disease SNPs were tested, two models showed acceptance at the level of $\alpha = 0.01$. The set of SNP1, 2, 4, 5, and 6 resulted in a *P* value of 0.96, and the set of SNP1, 2, 4, 6, and 7 resulted in a *P* value of 0.98. One of the models with six disease SNPs was finally accepted with the level of $\alpha = 0.05$ (*P* value of 0.81). The set contained SNP1, 2, 3, 5, 6, and 7. The model with all seven SNPs as disease polymorphisms was rejected in this test, since the model could not provide an appropriate solution for the original differences in disease allele frequencies between cases and controls.

This result confirmed that a few common SNPs might not explain the strong and consistent association between *APOE* and Alzheimer's disease. It is possible that, as in this test, many common SNPs can be actual disease polymorphisms. However, this result merits some skepticism because (1) the controls in the test were not actual controls for late-onset Alzheimer's disease and (2) the actual sample sizes were not large enough for either cases or controls. Therefore, the true disease polymorphisms responsible for the *APOE* association with Alzheimer's disease may differ from this result. A recent study indicated that rare disease polymorphisms with very high odds ratios (allele frequency between 0.005 and 0.02; genotype relative risk = 4) can create synthetic associations of common polymorphisms (Dickson et al. 2010). This is highly probable if there are a

few rare polymorphisms with high odds ratios in the gene region. In the previous study of the *APOE* association with Alzheimer's disease (Yu et al. 2007), however, there was no such rare SNP with high odds ratios. Because the data in their association study was not re-sequenced data, it is still possible that rare SNPs with high odds ratios exist in the region. Further investigation of the region with increased sample sizes may help to identify real disease polymorphisms.

Discussion

The advantages and limitations of GWAS have been indicated recently, and identifying several causal polymorphisms in a gene region using fine mapping or re-sequencing has been emphasized (Altshuler et al. 2008; Hardy and Singleton 2009; Manolio et al. 2009). With the goal of obtaining true associations connecting gene association results with a disease presentation, the current study provides the first practical method for finding actual disease polymorphisms from case-control genotype data using LD information in a given gene region. In contrast to other conventional statistical approaches for dealing with confounders, this method uses the actual LD relationship between markers and disease polymorphisms instead of statistically treated relationships. The current method can provide more concrete conclusions from which to infer the real disease polymorphisms in a gene region.

Concerning the intrinsic multiple testing in the method, both Type I and Type II error rates were examined thoroughly. As shown in the Results section, the Type I error rates were consistently reasonable and the Type II error rates usually low for most tests. In situations of disease polymorphisms in very high LD through r^2 with other polymorphisms, the Type II error rates increased; however, increasing the control sample size remedied the error rates. For unusual situations in which the true disease polymorphisms were in very high LD through r^2 and the effects of their odds ratios were opposed, increasing both the control and case sample sizes reduced the Type II error rates to a great extent. This method is also applicable to multi-allelic polymorphisms, where it considers each allele as a separate SNP in complete linkage disequilibrium through D' . Therefore, the effect of each allele can be independently tested in this method. In summary, the current method is expected to provide valid results for most sets of disease polymorphisms with odds ratios higher than 1.5 and a sufficient sample size.

When there are polymorphisms in complete LD in both cases and controls, it is not possible to distinguish one from another. As indicated previously, increasing sample sizes could mitigate the complete LD, but high LD between

those polymorphisms could still hamper the identification of real disease polymorphisms if one of them is the disease polymorphism. In the case, comparisons with case–control association studies using different population could be helpful. It is well known that there are clear population differences in allele frequencies and linkage disequilibrium patterns (Frazer et al. 2007). Therefore, the polymorphisms in complete LD in a certain population might not be in strong LD in other populations, and it would be possible to distinguish the real disease polymorphism among the polymorphisms in complete LD in the previous population. If the polymorphisms are in complete LD only in either case or control samples, there is no need to exclude the polymorphisms for the analysis.

As shown in the Results section, the use of re-sequenced data substantially reduces Type II error rates even in bad situations. However, as indicated previously, the actual Type II errors might slightly be increased due to the increment of possible testing models using re-sequenced data. In addition, a large control sample size or polymorphisms with appropriate minor allele and haplotype frequencies should be used as data to obtain stable Type I error rates. Therefore, if the control sample sizes are large enough and/or sufficiently frequent polymorphisms are used as data, better outcomes are expected using re-sequenced data. Since these methods are also useful for polymorphisms in low LD, expanding the region to include more data would be helpful to obtain better results (even though the expanded regions are not related to disease presentation). Another advantage of using re-sequencing data is that a more accurate set of disease polymorphisms can be derived. Including all possible polymorphisms reduces the chance of missing the actual disease polymorphisms in the model set. As re-sequencing takes a greater role in the next generation of genetic methodologies, the method presented here is expected to be even more advantageous than existing methods.

This is the first report of a method to identify actual disease polymorphism(s) based primarily on the LD relationship between polymorphisms using case–control genotype data. Having information about truly associated disease polymorphisms permits better understanding of the role of associated genes in complex traits and better modeling of gene–gene interactions. The method would improve our knowledge of the genetic architecture of complex traits. From an evolutionary point of view, it is beneficial for two deleterious polymorphisms not to exist in a haplotype. Therefore, the worst-case scenario of SNP3 and SNP4 might be more common than expected. Applying the method developed in this study to dense association mapping or re-sequenced data would provide valid interpretations for the results of GWAS. This would improve our understanding of the true genetic effects underlying

disease presentation and could provide a better explanation for purifying selection pressures in human genome.

Acknowledgments This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-532-C00017) and by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2009-353-C00061). The key calculations were performed using the supercomputing resource at the Korea Institute of Science and Technology Information (KISTI), supported by grant No. KSC-2009-S01-0003 from KISTI.

References

- Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. *Nature* 437(7063):1299–1320
- Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322(5903):881–888. doi:[322/5903/881](https://doi.org/10.1126/science.1156409)[pii][10.1126/science.1156409](https://doi.org/10.1126/science.1156409)
- Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics* 21(2):263–265
- Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C et al (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet*. doi:[ng.381](https://doi.org/10.1038/ng.381) [pii][10.1038/ng.381](https://doi.org/10.1038/ng.381)
- Biernacka JM, Cordell HJ (2009) A composite-likelihood approach for identifying polymorphisms that are potentially directly associated with disease. *Eur J Hum Genet* 17(5):644–650. doi:[ejhg.2008.242](https://doi.org/10.1038/ejhg.2008.242)[pii][10.1038/ejhg.2008.242](https://doi.org/10.1038/ejhg.2008.242)
- Biernacka JM, Charoen P, Cordell HJ (2007) Joint linkage and association analysis for identification of potentially causal polymorphisms in gaw15 data. *BMC Proc* 1(Suppl 1):S36
- Charoen P, Biernacka JM, Cordell HJ (2007) Linkage and association analysis of gaw15 simulated data: fine-mapping of chromosome 6 region. *BMC Proc* 1(Suppl 1):S23
- Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, Yoon D, Lee MH, Kim DJ, Park M et al (2009) A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 41(5):527–534. doi:[ng.357](https://doi.org/10.1038/ng.357)[pii][10.1038/ng.357](https://doi.org/10.1038/ng.357)
- Cordell HJ, Clayton DG (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to hla in type 1 diabetes. *Am J Hum Genet* 70(1):124–141. doi:[S0002-9297\(07\)61288-9](https://doi.org/10.1086/338007)[pii][10.1086/338007](https://doi.org/10.1086/338007)
- Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol* 8(1):e1000294. doi:[10.1371/journal.pbio.1000294](https://doi.org/10.1371/journal.pbio.1000294)
- Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struwing JP, Morrison J, Field H, Luben R et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447(7148):1087–1093. doi:[nature05887](https://doi.org/10.1038/nature05887)[pii][10.1038/nature05887](https://doi.org/10.1038/nature05887)
- Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, Hall AS, Linsel-Nitschke P, Kathiresan S, Wright B, Tregouet DA et al (2009) New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet* 41(3):280–282. doi:[ng.307](https://doi.org/10.1038/ng.307)[pii][10.1038/ng.307](https://doi.org/10.1038/ng.307)
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM et al (2007) A

- second generation human haplotype map of over 3.1 million snps. *Nature* 449(7164):851–861. doi:[nature06258](https://doi.org/10.1038/nature06258)[pii]10.1038/nature06258
- Hardy J, Singleton A (2009) Genomewide association studies and human disease. *N Engl J Med* 360(17):1759–1768. doi:[NEJMra0808700](https://doi.org/10.1056/NEJMra0808700)[pii]10.1056/NEJMra0808700
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753. doi:[nature08494](https://doi.org/10.1038/nature08494)[pii]10.1038/nature08494
- Maris JM, Mosse YP, Bradfield JP, Hou C, Monni S, Scott RH, Asgharzadeh S, Attiye EF, Diskin SJ, Laudenslager M et al (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N Engl J Med* 358(24):2585–2593. doi:[NEJMoa0708698](https://doi.org/10.1056/NEJMoa0708698)[pii]10.1056/NEJMoa0708698
- Nickerson DA, Taylor SL, Fullerton SM, Weiss KM, Clark AG, Stengard JH, Salomaa V, Boerwinkle E, Sing CF (2000) Sequence diversity and large-scale typing of snps in the human apolipoprotein e gene. *Genome Res* 10(10):1532–1545
- Nicodemus KK, Stenger JE, Schmechel DE, Welsh-Bohmer KA, Saunders AM, Roses AD, Gilbert JR, Vance JM, Haines JL, Pericak-Vance MA et al (2004) Comprehensive association analysis of apoe regulatory region polymorphisms in alzheimer disease. *Neurogenetics* 5(4):201–208
- Park L (2007) Controlling linkage disequilibrium in association tests: revisiting apoe association in Alzheimer's disease. *Genomics Inform* 5(2):61–67
- Qin ZS, Niu T, Liu JS (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* 71(5):1242–1247
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70(2):425–434. doi:[S0002-9297\(07\)63956-1](https://doi.org/10.1086/338688)[pii]10.1086/338688
- Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, Upmanyu R, Tozzi F, Li JZ, Burmeister M, Absher D et al (2009) Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc Natl Acad Sci USA* 106(18):7501–7506. doi:[0813386106](https://doi.org/10.1073/pnas.0813386106)[pii]10.1073/pnas.0813386106
- Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, Delattre JY et al (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* 41(8):899–904. doi:[ng.407](https://doi.org/10.1038/ng.407)[pii]10.1038/ng.407
- Song H, Ramus SJ, Tyrer J, Bolton KL, Gentry-Maharaj A, Wozniak E, Anton-Culver H, Chang-Claude J, Cramer DW, DiCiccio R et al (2009) A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat Genet* 41(9):996–1000. doi:[ng.424](https://doi.org/10.1038/ng.424)[pii]10.1038/ng.424
- Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, Zhang H, Sun YV (2009) Machine learning in genome-wide association studies. *Genet Epidemiol* 33(Suppl 1):S51–S57. doi:[10.1002/gepi.20473](https://doi.org/10.1002/gepi.20473)
- Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jorgensen T et al (2008) Snps in *kcq1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* 40(9):1098–1102. doi:[ng.208](https://doi.org/10.1038/ng.208)[pii]10.1038/ng.208
- Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661–678. doi:[nature05911](https://doi.org/10.1038/nature05911)[pii]10.1038/nature05911
- Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S et al (2009) Variants in the *cdkn2b* and *rte11* regions are associated with high-grade glioma susceptibility. *Nat Genet* 41(8):905–908. doi:[ng.408](https://doi.org/10.1038/ng.408)[pii]10.1038/ng.408
- Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y et al (2008) Variants in *kcq1* are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* 40(9):1092–1097. doi:[ng.207](https://doi.org/10.1038/ng.207)[pii]10.1038/ng.207
- Yu Z, Schaid DJ (2007) Sequential haplotype scan methods for association analysis. *Genet Epidemiol* 31(6):553–564. doi:[10.1002/gepi.20228](https://doi.org/10.1002/gepi.20228)
- Yu CE, Seltman H, Peskind ER, Galloway N, Zhou PX, Rosenthal E, Wijsman EM, Tsuang DW, Devlin B, Schellenberg GD (2007) Comprehensive analysis of apoe and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. *Genomics* 89(6):655–665. doi:[S0888-7543\(07\)00050-X](https://doi.org/10.1016/j.ygeno.2007.02.002)[pii]10.1016/j.ygeno.2007.02.002