Methodological Issues

Human Heredity

Hum Hered 2000;50:133–139

Received: December 17, 1998
Revision received: February 15, 1999
Accepted: February 23, 1999

Model-Free Analysis and Permutation Tests for Allelic Associations

Jing Hua Zhao[a] David Curtis[b] Pak Chung Sham[a]

[a]Department of Psychological Medicine, Institute of Psychiatry, and [b]Joint Academic Department of Psychological Medicine, St. Bartholomew's and Royal London School of Medicine and Dentistry, London, UK

Key Words
Case-control design · Allelic association · Haplotype analysis · Permutation test · Model-free association analysis

Abstract
In this short report, we address some practical problems in performing likelihood-based allelic association analysis of case-control data. Model-free statistics are proposed and their properties assessed by simulation, and procedures based on permutation tests are described for marker-marker as well as marker-disease associations. A memory-efficient algorithm is developed which enables several highly polymorphic markers to be analysed.

Introduction

Allelic associations refer to both marker-marker and marker-disease association, for which likelihood methods based on haplotype frequences have been implemented in the EH program [1, 2]. For marker-disease association, EH requires a single-locus disease model, parametrized by allele frequencies and genotype-specific penetrances (f_0, f_1, f_2), to be specified.

We have encountered several problems while using EH in the analysis of real data. Firstly, the contingency table of genotypes may be very sparse for highly polymorphic marker loci (such as those in the HLA system) so that the usual asymptotic chi-squared approximation may become inaccurate. Secondly, the mode of inheritance is often uncertain for complex diseases, making it uncertain what parameter values should be specified. Thirdly, although EH can flexibly deal with an arbitrary number of loci each with an arbitrary number of alleles, in practice its application is limited by the large memory requirement when marker loci are highly polymorphic.

We have written a program to overcome these problems. The program implements model-free statistics similar to those in MFLINK [3], with permutation tests to obtain empirical p values. It also incorporates an algorithm that can be used to reduce the storage requirement from the counts of all possible combinations of genotypes to those actually present in the sample.

Test Statistics

For case-control data, EH outputs three log-likelihoods, $\ln L_0$, $\ln L_1$ and $\ln L_2$, which correspond to the hypotheses H_0 (no associations allowed), H_1 (marker-marker but not marker-disease associations allowed), and H_2 (marker-marker and marker-disease associations al-

KARGER
Fax + 41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2000 S. Karger AG, Basel
0001–5652/00/0502–0133$17.50/0

Accessible online at:
www.karger.com/journals/hhe

Jing Hua Zhao
Department of Psychological Medicine
Institute of Psychiatry, De Crespigny Park
London SE5 8AF (UK)
Tel. +44 171 919 25 24, Fax +44 171 701 90 44, E-Mail j.zhao@iop.kcl.ac.uk

lowed). Supposing there are n marker loci, where the ith locus has a_i alleles, the number of parameters for the three hypotheses are

$$N_0 = \sum_i (a_i - 1),\ N_1 = \prod_i a_i - 1 \quad \text{and } N_2 = 2N_1.$$

i = 1,2,...,n, respectively. A likelihood ratio test for the presence of marker-disease associations, under the user-specified genetic model, is given by $2(\ln L_2 - \ln L_1)$, which is asymptotically chi-squared with N_1 degrees of freedom. This statistics assumes the use-specified model, and is denoted as T1.

The new program computes four other tests. A statistic is obtained under a Mendelian recessive model ($f_0 = f_1 = 0$, $f_2 = 1$), which is defined as T_2. Similarly a statistic is obtained under a Mendelian dominant model ($f_0 = 0$, $f_1 = f_2 = 1$), which is defined as T3. In both cases, allele frequencies are calculated from the disease prevalence implied by the user-specified model. These two tests represent the two extremes of a single-locus Mendelian transmission model.

The T4 statistic is obtained by treating disease allele frequencies and penetrances as nuisance parameters. To calculate T4, the case-control option of the EH program is used as in the calculation of T1, T2, and T3. However, instead of being fixed at a single set of values, the disease model parameters are allowed to vary over a certain range. The maximum log-likelihood ratio statistic over this range of parameter values is defined as the T4 statistic. As in the 'model-free' method for linkage analysis [3], the disease model parameters are constrained to produce the population prevalence (K) implied by the user-defined model, and only certain fully dominant ($f_1 = f_2$) and recessive ($f_0 = f_1$) models are considered. If the parameter space is represented in three dimensions as the coordinates (f_0, f_1, f_2), then the likelihood is evaluated only for disease models represented by points on the straight lines joining (0, 0, 1) to (K, K, K) and joining (K, K, K) to = (0, 1, 1). The program varies f_1 from 0 to 1, and calculates both f_0 and f_2 in terms of f_1: if $f_1 < K$, then $f_0 = f_1$, $f_2 = f_1 (K - 1)/K + 1$, otherwise $f_2 = f_1$, $f_0 = (1 - f_1)K/(1 - K)$. The disease model is therefore characterized by a single nuisance parameter, namely f_1. The standard method for dealing with a nuisance parameter, namely to maximize the log-likelihood over it under both the null and the alternative hypotheses, is not applicable because the log-likelihood is independent of f_1 under the null hypothesis of linkage equilibrium. We show by simulation that T4 can be considered conservatively to have a chi-square with $N_1 + 1$ degrees of freedom. Instead, one can obtain an empirical significance level T4 by computer-intensive methods (see below).

Table 1. The five genetic models used to stimulate data

Model	f_0	f_1	f_2	p_2	K
Null (H0)	0.005	0.500	0.5	0.0050	0.005
Rare recessive (RR)	0	0	1	0.0316	0.001
Rare dominant (RD)	0	1	1	0.0005	0.001
Common recessive (CR)	0.005	0.005	0.5	0.1000	0.010
Common dominant (CD)	0.005	0.500	0.5	0.0050	0.010
Minor gene (MG)	0.050	0.200	0.8	0.1300	0.100

The T5 statistic provides a nonparametric test for homogeneity in allele frequencies between cases and controls. To calculate T5, the EH program is used three times, once on the cases alone, once on the controls alone, and once for the cases and controls pooled together. In each analysis, allele frequencies are estimated and the maximum log-likelihood calculated. Denoting these maximum log-likelihoods as $\ln L_{case}$, $\ln L_{control}$ and $\ln L_{combine}$, T5 is defined as $2(\ln L_{case} + \ln L_{control} - \ln L_{combine})$, which is asymptotically chi-squared with N_1 degrees of freedom.

Properties of the Test Statistics [4]

Disease Models

We assume a biallelic disease locus A with alleles A_1, A_2 occurring at frequencies p_1, p_2. The probability of the disease in an individual with i copies of the A_2 allele is denoted by the penetrance parameter f_i, for i = 0, 1, 2. The population risk of the disease if given by $K = f_0 p_1^2 + 2 f_1 p_1 p_2 + f_2 p_2^2$. We consider 6 different single gene disease models (table 1), in order to examine the properties of the tests under a range of conditions.

Probability Model of Marker Genotypes

We consider a marker locus, denoted here as B, with n alleles. Conditional on an individual's affection status, the probability distribution of the individuals' genotype at locus B is given by Sham [5]. The overall likelihood of a sample of cases and controls is the product of such probabilities, one for each observation of an affection status and the associated genotype at the marker locus.

Data Simulation

Haplotype frequencies were derived from Oudet et al. [6] on fragile X syndrome. The frequencies of the seven

Table 2. Mean and standard deviation of chi-squared statistics from 500 replicates of 500 cases and 500 controls

Model	T1	T2	T3	T4	T5
H0	5.4 (2.9)	6.1 (3.4)	6.0 (3.2)	6.4 (3.5)	6.1 (3.4)
RR	321.9 (33.8)	321.9 (33.8)	239.9 (23.8)	322.0 (33.8)	321.9 (33.8)
RD	108.6 (18.5)	95.4 (17.1)	108.6 (18.5)	108.7 (18.5)	95.4 (17.1)
CR	85.6 (18.7)	77.5 (17.8)	59.5 (13.2)	86.6 (18.8)	77.5 (17.8)
CD	30.5 (10.5)	30.4 (10.6)	31.3 (10.7)	32.1 (10.8)	30.4 (10.6)
MG	31.0 (10.4)	32.0 (10.7)	31.1 (10.3)	32.9 (10.8)	32.0 (10.8)

alleles at DXS548 were 0, 42, 32, 1, 1, 29, 1 on fragile X chromosome and 2, 117, 23, 1, 1, 15, 2 on normal chromosomes. These conditional probabilities were multiplied by the marginal probabilities of the disease alleles (which are different for different disease models) to give the joint haplotype frequencies of the marker and disease loci. For each model, 500 replicate samples of 1,000 subjects (500 cases and 500 controls) and 500 replicate samples of 10,000 subjects (5,000 cases and 5,000 controls, data not shown), were simulated in order to investigate the accuracy of the asymptotic chi-squared distribution as a function of sample size.

Table 3. Estimated sample sizes (cases and controls combined) required for 90% power at 0.05 significance level

Model	T1	T2	T3	T4	T5
RR	56	56	75	59	56
RD	170	195	170	180	195
CR	219	244	325	230	244
CD	712	713	688	729	713
MG	697	671	695	706	670

Comparison of Test Statistics

The properties of each statistic under each model were investigated using the simulated samples. The mean value of the test statistics over the replicates is an estimator of its theoretical expectation (which is the sum of the non-centrality parameter and degree of freedom). This allows the non-centrality parameter to be estimated. Under the null hypothesis (H_0), the non-centrality parameter should be 0. A value greater than 0 implies an increase in the false-positive rate, while a value less than 0 indicates that the test is conservative. Under an alternative hypothesis, the non-centrality parameter determines the power of the test. At 5% significance level, the values of non-centrality parameter required for 90% power are 17.4 and 18.3, for 6 and 7 degrees of freedom, respectively. Since the non-centrality parameter is proportional to sample size, the required sample size can be extrapolated from the non-centrality parameter estimates for the desired level or power.

The results are shown in tables 2 and 3. Since DXS548 shows seven alleles in the dataset, we have $N_1 = 6$ and $N_2 = 12$. Under H_0, the empirical means are close to their theoretical values, with the exception of T1 and T4, which have empirical means of 5.4 and 6.4 when the theoretical means are 6 and 7, respectively. T4 was thought to have an extra degree of freedom because a maximisation over f_1 was conducted in order to obtain the likelihoods assuming marker-disease association. The result indicates that the asymptotic distribution of T4 is closer to chi-squared with 6 degrees of freedom than chi-squared with 7 degrees of freedom. As expected, the empirical means under alternative hypotheses from simulated samples of size 10,000 (data not shown) are approximately 10 times the corresponding values from that of size 1,000.

No single test is uniformly most powerful among the 5 tests over all 5 models. T1, which is obtained by specifying the parameters used in the simulation, has the best performance overall. The powers of T2 and T5 appear to be equivalent. T4 is more powerful than T5 in some situations. For minor a gene model, T5 appears to be subsantially more powerful than T4. If T4 were considered to have a chi-squared distribution with 6 degrees of freedom, then T4 and T5 would be equally powerful even in this situation, but then T4 would be slightly anti-conservative.

The simulation showed that, when the disease model is unknown, the standard χ^2 test of homogeneity of allele frequencies (T5) provides nearly optimal power, especially for a minor-gene model. If the null distribution of the parametric 'model-free' test (T4) is accurately known, then this may be preferred to T5. Overall, our preferred test for routine use on case-control data is the standard χ^2

test of homogeneity of allele frequencies. If one were to extend this approach to data involving related individuals, however, it is likely that parametric 'model-free' methods will have a greater degree of superiority over non-parametric methods.

Permutation Tests

The reliance on asymptotic theory for obtaining significance levels is potentially problematic for two reasons. The first is that asymptotic theory may become inaccurate when the number of possible genotypic combinations is very large in relation to the sample size. The second is that the T4 statistic has a complicated distribution. Both these problems may be overcome by the use of permutation procedures to obtain empirical p values. An efficient algorithm due to Fisher and Yates, as described in Knuth [7] and Weir [8], can be adapted to permute case-control labels and blocks of markers. Here we propose such procedures for both marker-marker and marker-disease associations. We have written a C program to implement these procedures.

Marker-Marker Association

The proposed permutation procedure examines whether a block of markers is in linkage equilibrium with a second block of markers. A block may contain several markers, or just a single marker, as specified by the user. Denoting the log-likelihood with associations within each block but independence between blocks as $\ln L_1$, and the log-likelihood with associations between all markers as $\ln L_2$, then the statistic $2(\ln L_2 - \ln L_1)$ provides a test for linkage disequilibrium between the two blocks of markers, taking into account possible associations between markers within the same block. The degrees of freedom of this test is equal to $(h_1h_2 - 1) - (h_1 - 1) - (h_2 - 1)$, where h_1 and h_2 are the number of haplotypes in blocks 1 and 2, respectively. This number will be very large if one of both blocks contain several highly polymorphic markers, so that for realistic sample sizes the standard chi-squared approximation may be inaccurate. The proposed permutation procedure randomly reassigns all the genotypes of block 2 to the genotypes at block 1, preserving allelic associations for markers within the same block, but destroying allelic associations between markers in different blocks. For each permuted replicate, the statistic $2(\ln L_2 - \ln L_1)$ is calculated, so that after a large number of replicates an empirical sampling distribution of the statistic is obtained.

Marker-Disease Association

We simply permute the case-control labels and calculate each of the five test statistics. The test statistics from a large number of permuted replicate samples are then used to obtain empirical sampling distributions of the test statistics under the null hypothesis.

Modifying EH for Highly Polymorphic Loci

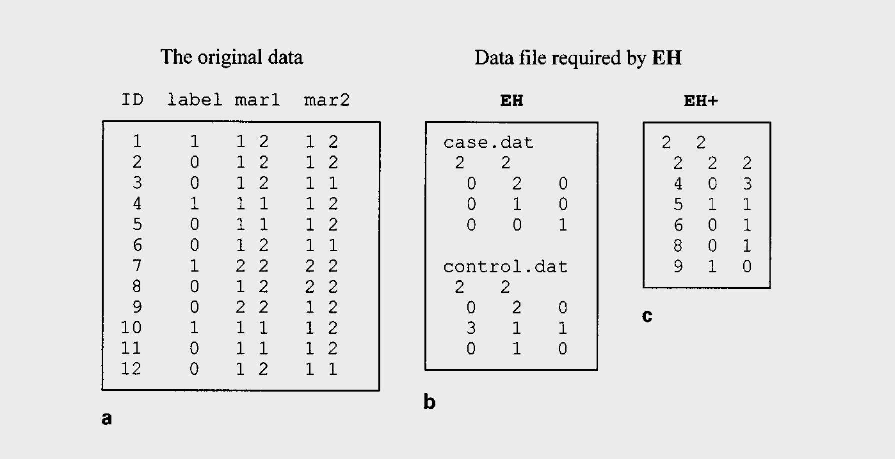
Genotype counts are presented to EH in the form of a 2-way contingency table, with each row representing a possible combination of genotypes at all loci except the last, and each column representing a genotype at the last locus. These counts are internally represented as a linear array, the size of which is

$$A = \prod_i \frac{a_i\,(a_i - 1)}{2},$$

where a_i, i = 1,...,n, is the number of alleles at locus i. In EH, each genotype combination is indexed by an identifier between 1 and A, and there is a one-to-one correspondence between the identifiers and the combinations of genotypes, as defined by the linenum function.

The input format and the data representation in EH have obvious disadvantages for highly polymorphic markers. An input file has to be prepared from the raw genotype data, perhaps by using statistical software such as SAS. These packages tend to omit any row or column containing only zero counts so that all these rows and columns must be added to the output file to make an input file for EH. This is easy for an analysis involving two or three biallelic loci, but extremely tedious for an analysis involving a greater number of markers or of highly polymorphic markers. For instance, with three markers having 25, 10, 15 alleles, respectively, the input file will have 17,875 rows and 120 columns, regardless of the size of the sample. Internally, the storage of these 214,500 counts, most of which will be 0 for realistic sample sizes, is very memory inefficient.

To overcome these problems, we first create a simplified input file for EH. The first line of this file gives the number of alleles for each locus. Each subsequent line contains an identifier (which is related to a particular combination of genotypes by the linenum function), and the observed counts for cases and controls (the count for cases being necessarily 0 for a study involving marker data only). The file can be created easily from a data file containing the case-control status and genotypes of each subject, using a 'search-and-insert' scheme. The marker



```
     The original data                Data file required by EH

 ID  label mar1  mar2                 EH                 EH+

 1    1    1 2   1 2          case.dat              2  2
 2    0    1 2   1 2           2   2                 2  2  2
 3    0    1 2   1 1            0   2   0            4  0  3
 4    1    1 1   1 2            0   1   0            5  1  1
 5    0    1 1   1 2            0   0   1            6  0  1
 6    0    1 2   1 1                                 8  0  1
 7    1    2 2   2 2          control.dat            9  1  0
 8    0    1 2   2 2           2   2
 9    0    2 2   1 2            0   2   0          c
 10   1    1 1   1 2            3   1   1
 11   0    1 1   1 2            0   1   0
 12   0    1 2   1 1
                             b
 a
```

Fig. 1. Data files required by EH before and after revision. **a** The raw data set contains subject id, case/control label (1 = case, 0 = control), two biallelic markers mar1 and mar2. **b** Data files for EH (case.dat and control.dat). The first row indicates number of alleles for markers mar1 and mar2, and the following rows contain the observed counts of genotype combinations. **c** The data file for the modified program, columns after the first row are as follows: 1 genotype identifier; 2 count of cases; 3 count of controls. Note lines with identifiers 1, 3, 7 are omitted.

genotypes of each individual are used to calculate the identifier; if this identifier is already present in the file, then the appropriate count (whether case or control) in that line is increased by 1, otherwise an extra line with that identifier and count one is inserted. When all observations have been read, the input file will contain one line for each combination of genotypes that is present in the sample. Even if no two individuals have the same combination of genotypes, the number of lines is still only equal to the sample size.

We have modified EH to accept this new format. This is achieved by altering the linenum function in EH to refer to the appropriate record in the simplified input file. Whenever EH requires the count of a combination of genotypes, it computes the associated identifier and a search is made to get the appropriate information. To speed up this search, we initially considered a hash table, indexed by modulo operation on the identifiers with respect to a moderately large (usually comparable to the sample size) prime number, a method originally due to Dumey [9].

However, we then realised we could achieve the same effect simply by using the array of the observed identifiers to index the array containing observed genotype counts. We have found this to be as efficient as the hash table method. Both options are available in the program.

A schematic representation of the various data files for an analysis of just 12 individuals is shown in figure 1.

As a by-product of this work, a C version of EH has been obtained by modifying the output from the p2c utility. We have also made EH print out the Freeman-Tukey [10] residuals, as a measure of the discrepancy between observed and expected counts.

The program implementing the model-free statistics and the modified EH program, called EHPLUS, are freely available from the first author (e-mail: j.zhao@iop.kcl.ac.uk).

Example: Association between Schizophrenia and HLA Markers

This is a case-control association study of schizophrenia (Dr. Padraig Wright, pers. commun.]. Three highly polymorphic markers DRB, DQA, and DQB in the HLA region of chromosome 6 were examined in 94 schizophrenic patients and 177 controls. The numbers of reasonably common alleles of these three markers are 25, 10, and 15, respectively. We were not able to use EH to get the appropriate log-likelihoods for marker-marker and disease-marker associations involving all three markers.

For marker-marker association analysis of all three markers in the overall sample (patients and controls) the modified EH program computed a log-likelihood of –3,180.84 under the null hypothesis that the markers are independent, and –1,594.75 allowing for the presence of marker-marker associations. The number of free parameters associated with these two hypotheses are 47 and 3,749, respectively. The likelihood ratio chi-squared statistic for marker-marker association is therefore 2 (3,180.84 – 1,594.75) = 3,172.18, with 3,702 degrees of freedom. This yields a p value that is close to 1. Yet the p value based on asymptotic theory for such a large number of degrees of freedom is quite unreliable.

The likelihood evaluation incorporating all three markers took several hours to do a single analysis under DEC Alpha, therefore it would be quite time-consuming to do permutation test involving all three markers. Here we only used DQA in the test and illustrated it with the case-control option. A recessive model with disease allele frequency 0.1, penetrances 0.005, 0.005, 0.5 was assumed [11]. The user-specified, recessive, dominant, model-free and heterogeneity chi-square statistics are 89.96, 92.90, 57.24, 97.78 and 101.90, respectively, which correspond to asymptotic p-values smaller than 0.000001 for chi-squared distribution with degrees of freedom 9 and 10. With the permutation option none of the 10,000 replicates produced any statistics which exceeded chi-squared values from the observed data. We thus conclude that there is association between DQA and schizophrenia based on the empirical evidence.

Discussion

We have implemented several modifications to EH which should make it an even more useful program for the analysis of marker-marker and marker-disease association data. The program may be particularly useful for single-nucleotide polymorphisms, where it is important to combine information from several tightly linked loci in order to increase the power to detect disease-associated haplotypes. The original EH program implements a maximum-likelihood approach to the estimation of haplotype frequencies and the testing of hypotheses concerning allelic associations. The present modifications have simplified the input data files, increased the number of loci and alleles that can be handled, provided model-free test statistics suitable for complex disorders, and given the option to obtain empirical p values by permutation tests.

The saving in memory requirement is achieved at the cost of some extra CPU time, and the permutation procedure remains slow for large problems. Further improvements of EH are therefore desirable, possibly in the sense that the EM algorithm processes data by person at each iteration, rather than by phenotype. The original genotypes could alternatively be used as multiple identifiers in the simplified input file so that output produced by statistical packages could be re-used. For example, the list format produced by SAS PROC FREQ might be appropriate.

We have implemented test statistics which can be used when the disease model is uncertain. Our results from simulation studies indicate that a non-parametric test of heterogeneity of haplotype frequencies (T5) is nearly optimal in most circumstances. Interestingly, this test is also almost equivalent to a test assuming Mendelian recessive inheritance (T2).

One issue not addressed by this report is whether it is desirable to group together certain alleles in order to reduce the size of the data set and the number of parameters, and possibly to increase power [12, 13]. Another issue is that with multiple loci, many individuals may have genotype data for some but not all loci. The present program discards observations with partially missing data, and it will be desirable to develop algorithms that will make full use of all available data.

A number of other programs are also available for allelic association analysis, including 3LOCUS [14], the VAX/VMS-based HAPLO [15], GDA [8, 16–18] and Arlequin [19]. We have not explored fully the features of these programs but none appear to be directly applicable to case-control data.

Acknowledgments

We thank authors of EH for their pioneering work, Prof. Jürg Ott for helpful comments, and Dr. Jeffrey Long for information concerning 3LOCUS. Thanks also to Drs. Tao Li and Maria Arranz, and other colleagues at the Institute of Psychiatry, for their continuous interest in case-control association analyses. J.H.Z. is supported by Wellcome Trust grant No. 043279.

References

1 Xie X, Ott J: Testing linkage disequilibrium between a disease gene and marker loci. Am J Hum Genet 1993;53:1107.

2 Terwilliger J, Ott J: Handbook of Human Genetic Linkage. Baltimore, The Johns Hopkins University Press, 1994.

3 Curtis D, Sham PC: Model-free linkage analysis using likelihoods. Am J Hum Genet 1995; 57:703–716.

4 Zhao JH, Sham PC: Model-free allelic association analysis of case-control studies. Am J Med Genet (Neuropsychiatric Genet) 1997;74:604.

5 Sham PC: Statistics in Human Genetics. Londo, Edward Arnold, 1998, p 159.

6 Oudet C, Mornet E, Serre JL, Thomas F, Lentes-Zengerling S, Kretz C, Deluchat C, Tejada I, Boue A, Mandel JL: Linkage disequilibrium between the fragile X mutation and two closely linked CA repeats suggests that fragile X chromosomes are derived from a small number of founder chromosomes. Am J Hum Genet 1993;52:297–304.

7 Knuth DE: The Art of Computer Programming. London, Addison-Wesley, 1998, vol 2: Seminumerical Algorithms, ed 3.

8 Weir BS: Genetic Data Analysis: Methods for Discrete Population Genetic Data. Sunderland, Sinauer, 1990.

9 Knuth DE: The Art of Computer Programming. London, Addison-Wesley, 1998, vol 3: Sorting and Searching, ed 2.

10 Freeman MF, Tukey JW: Transformations related to the angular and square root. Ann Math Stat 1950;21:607–611.

11 Murray R: Schizophrenia; in Murray R, Hill P, McGuffin P (eds): The Essentials of Postgraduate Psychiatry. London, Cambridge University Press, 1997, pp 281–309.

12 Sham PC, Curtis D: Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 1995;59: 97–105.

13 Cox A, Camp NJ, Nicklin MJH, di Giovine FS, Duff GW: An analysis of linkage disequilibrium in the interleukin-1 gene cluster, using a novel grouping method for multiallelic markers. Am J Hum Genet 1998;62:1180–1188.

14 Long JC, Williams RC, Unbanek M: An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 1995;56: 799–810.

15 Hawley ME, Kidd KK: HAPLO: A program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 1995;86:409–411.

16 Weir BS: Genetic Data Analysis II. Sunderland, Sinauer, 1996.

17 Zaykin D, Zhivotovsky L, Weir BS: Exact tests for association between alleles at arbitrary numbers of loci. Genetica 1995;96:169–178.

18 Lewis PO, Zaykin D: Genetic Data Analysis: Computer program for the analysis of allelic data. http://chee.unm.edu/gda/1997.

19 Schneider S, Kueffer JM, Roessli D, Excoffier L: Arlequin: A software for population genetic analysis. http://anthropologie.unige.ch/arlequin. 1998.