Original Paper

Human Heredity

Hum Hered 2002;53:36–41

Received: September 10, 2001
Revision received: December 6, 2001
Accepted: December 10, 2001

Faster Haplotype Frequency Estimation Using Unrelated Subjects

Jing Hua Zhao[a] Pak Chung Sham[a,b,c]

[a]Department of Psychological Medicine, [b]Department of Biostatistics and Computing, [c] Social, Genetic, Developmental Psychiatry Research Centre, Institute of Psychiatry, London, UK

Key Words
EM algorithm · Gene counting · Linkage disequilibrium · Case-control association analysis · Permutation tests · Contingency table analysis

Abstract
Linkage disequilibrium (LD) between tightly linked loci provides fine mapping information of disease-predisposing allelic variants. The most common method of LD analysis involves unrelated cases and controls. We have previously proposed model-free and permutation tests for diseases with unknown mode of inheritance that can be applied to several highly polymorphic loci. However, performing such analyses remained computer intensive. In this report we propose a speed-up of both the gene-counting procedure and the permutation procedure. We demonstrate the improved method with an analysis of schizophrenia and human leucocyte antigen markers, and an analysis of alcoholism and mitochondrial aldehyde dehydrogenase markers. Our implementation also allows the rapid calculation of permutation-based LD measures and related statistics.

Introduction

Haplotype analysis of unrelated individuals is widely used for examining linkage disequilibrium (LD) between a set of marker loci in one or more populations [1–5], for association studies between a proposed disease locus and markers [6, 7] and for providing information for family haplotype analysis [7–10]. Haplotype frequency estimation is commonly achieved by gene counting, which is a simple form of the EM algorithm [11–16]. Multiple samples from one or more populations can be subject to a heterogeneity analysis [17]. If samples of cases and controls are involved, then a putative disease locus can also be incorporated into the analysis, as has been implemented in the computer program EH (Estimating Haplotypes) [6].

Although the EM algorithm for haplotype frequency estimation is simple in principle, it must be implemented efficiently in order to deal with data that involve a large number of haplotypes and multilocus genotypes. A standard method of summarising categorical data is to tabulate the individuals into a multidimensional contingency table according to their multilocus genotypes. However, not only is this computationally inefficient for a large number of marker loci or any number of highly polymorphic loci, because the resulting contingency table will be very sparse, but the test statistics may not conform to standard asymptotic distributions.

In a previous report [18], we discussed model-free analysis and permutation tests for case-control data, and proposed a modification of the recursive implementation in EH to handle large problems. We implemented our method in a C program called EHPLUS. Despite these improvements, this program remained time consuming; a single case-control analysis can take minutes or hours. This restricts the use of permutation tests, which involve analysing a large number of replicate samples.

KARGER

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

0001–5652/02/0531–0036$18.50/0

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

Dr. Jing Hua Zhao
Dept. of Psychological Medicine, Developmental Psychiatry Research Center
Institute of Psychiatry, De Crespigny Park
Denmark Hill, London, SE5 8AF (UK)
E-Mail j.zhao@iop.kcl.ac.uk

Here we propose a further speed-up of both the gene-counting and the data preparation procedure. After describing the proposed speed-ups, we demonstrate the new method with two analyses: schizophrenia and human leucocyte antigen (HLA) data given in our previous report, and alcoholism and mitochondrial aldehyde dehydrogenase (ALDH2) as reported in Koch et al. [19]. Finally, we give some general results and a brief discussion.

Methods

Previous Modifications to EH

We use two biallelic markers to illustrate our previous modifications. Let the alleles of each biallelic marker be 1 and 2, and the genotypes be 1/1, 1/2, and 2/2. These genotypes are commonly designated as 1, 2 and 3. A sample of individuals genotyped at these two loci can then be tabulated into a 3 × 3 genotype table, with each cell in the table corresponding to a two-locus genotype. These two-locus genotypes can be collectively identified with numbers $(i-1)*3+j$, where i, j = 1, 2, 3, are the genotype identifiers at loci 1 and 2. To obtain the haplotype counts, EH goes through each cell of the 3 × 3 table. For large problems, we may expect some cells to have empty counts; such cells do not contribute to the haplotype counts or the likelihoods. Figure 1 of Zhao et al. [18] gave an example of 12 individuals genotyped at two biallelic loci, the 3 × 3 genotype table has zero counts for cells with two-locus genotype identifiers 1, 4, 7, corresponding to marker genotypes 1/1-1/1, 1/2-1/1, 2/2-1/1.

The method implemented in EHPLUS was to construct a sorted list for non-empty cells during data preparation, 2, 3, 5, 6, 8 and 9 in the example, with each item of the list containing a two-locus genotype identifier and number of cases and controls associated with it. This sorted list was achieved by a standard linked list, which is a dynamic data structure built at run-time, so that a pre-defined constant for list length is not needed. It differs from array representation in that each item in the list is created when the program is running and has a pointer to the next item. Starting from an empty list, an insert or delete operation is performed dynamically by allocating or de-allocating memory and tuning appropriate pointers. De-allocation is appropriate when replicate analyses are performed. When gene counting goes through identifiers 1–9, matches in the list are located via several search methods. The binary search [20] was first chosen. To search for a record in a sorted list of records, the middle record is picked up and its key compared to the key to be searched. If the middle key is smaller, then we expect our match to be greater than the middle key so we do another search in the second half; alternatively if the key of the middle record is bigger than our search key we continue the search in the first half. This comparison is repeated until either a match is found or no match is found (indicating an empty cell). As for the current example, EHPLUS used the two-locus genotype identifier as the key. The second search method was hashing [20]. It proceeded by setting up a 'home bucket' holding hash keys obtained from subjecting genotype identifiers to modulo operation to a prime number, comparable to the sample size and being the number of home buckets, plus other information such as the number of cases and controls. Matches were then searched with respect to hash keys. Any collisions (i.e., more than two records with the same hash key) were resolved by further binary search. Since the prime number was fairly large, these hash keys were almost evenly spaced in these buckets and we expect few collisions. The third method involved setting up a sentinel variable. Since the counting process proceeds through the genotypes in order, a counter was created to keep track of the items of the list already used. It either took the value of the counter (to index the item in the non-empty list to be used for haplotype count updating), or –1 (a signal to abort haplotype count updating). In the current example the counter value was initialised to 0. Since cell 1 had an empty count, the counter remained unchanged, indicating that no item in the sorted list was used; and the sentinel variable took value –1. Cell 2 was non-empty so the sentinel variable took value 0 and the first item of our list was used to update haplotype counts, and the counter was then increased by 1. The process was repeated until cell 9, when our counter became 5 and we had exhausted all the observed data.

Further Speed-Ups

A linked-list is easy to build but slow for insert operation, for the position of insertion is determined by a sequential search from the start of the list and iterating through its pointer to next term of the list. This is now replaced with a more efficient scheme using a tree structure [20]. A tree is a collection of nodes and edges connecting them, and beginning from any node in the tree and traversing along the nodes and edges to another node does not return to the starting node without using certain edge(s) twice. Drawn top-down, the top node of the tree is root, and nodes below it are called its children. Each node of the tree can also have its own children. We have chosen a binary search tree, with which each node has at most two children, for it is the simplest yet performs well for our purpose. As permutation tests involve generating large number of replicates, such an improvement will have a pronounced effect on computing time. Finally, permutation of a case-control sample can in principle be achieved by randomly assigning genotypes to be cases, with the rest being re-labelled as controls.

Seeing that both searching and sentinel variable method have to go through all genotypes, some of which are not used, we might be able to loop over the non-empty cells kept in our sorted list directly. However, this would not be possible without knowing which marker genotypes give rise to the current multilocus identifier. Recovering marker genotypes from the two-locus genotype identifier would be expensive, yet this suggests we need keep track of the original marker genotypes as part of the items in the list. Formally, these alleles serve as the cache for the counting procedure. Returning to our two-locus example, we now directly iterate over genotype identifiers 2, 3, 5, 6, 8 and 9, their genotypes being 1/1-1/2, 1/1-2/2, 1/2-1/2 and 1/2-2/1, 1/2-2/2, 2/2-1/2, 2/2-2/2 as from the sample. Now, in calculating the likelihood, empty cells can be discarded since they have no contributions to the likelihood. The genotypic probabilities associated with non-empty cells could be done as usual by considering all possible phases given current haplotype estimates.

A Summary of the New Procedure and Some Further Remarks

We can summarise the proposed changes in a three-step procedure.

Step 1. Build a list of observed multilocus genotypes, the number of their occurrences, and the genotypes of the individual markers.

Step 2. Loop through each item in the list, use cached alleles to update counts of haplotype frequencies.

Step 3. Obtain haplotype frequencies and log likelihood and repeat step 2 until there is no appreciable change in log likelihoods.

At step 1, individuals with the same multilocus genotype are collapsed together in the binary search tree. Each item in the tree contains the genotype identifier as key, the number of cases and controls, and the genotypes of the constituent markers, plus a disease phenotype if a putative disease locus is incorporated. The genotype identifier for a single locus can be calculated as l + u(u – 1)/2, l ≤ u. For example, an individual with marker genotype 1/11 corresponds to identifier 1 + 11(11 – 1)/2 = 56). The total number of genotype identifiers at one locus with a alleles is a(a + 1)/2 (this corrects an error in Zhao et al. [18]). A multilocus genotype identifier can be built in a similar manner. The size of the multilocus genotype table is the product of genotype identifiers at individual loci. The length of the list is at most the number of individuals in the sample.

Step 2 is the basic gene-counting procedure that is equivalent to the EM algorithm. The E-step obtains the probability for each phase given current haplotype frequency estimates, while the M-step updates the haplotype counts based on these probabilities. If the number of heterozygous loci in a multilocus genotype is h, then there are 2^{h-1} possible phases, referring to the 2^{h-1} possible pairs of haplotypes which could give rise to the observed genotype. Unlike resorting to a standard search problem by either binary search or hashing or sentinel variable method, under the caching scheme we directly iterate over the observed genotypes.

We have noticed that the convergence criteria used at step 3 are almost equivalent to setting appropriate criteria with respect to the estimated haplotype frequencies as used in EH.

However, for disease marker analysis involving a putative disease locus, we were not able to get the likelihoods in this manner, despite many of our examples showed the log likelihoods to be smaller by a constant quantity and the likelihood ratio statistics to be virtually the same (analytic details involving one or two loci are available upon request). While this is a good attempt to model the phenotype-genotype relationship, implementing the correct method would make the calculation slower and less attractive for the permutation procedure. We therefore keep the simple χ^2 test of heterogeneity for case-control analysis in the current implementation, for this has almost the same power as an analysis assuming the correct model for an explicit disease locus [18].

The speed-up allows for a larger number of replicates for the permutation procedure, which can be used to assess information of allelic association from multiple multiallelic markers. Asymptotically, the log likelihood ratio test statistic of allelic association will have non-central χ^2 distribution with a non-centrality parameter $N\xi$, where N is the number of subjects and ξ is the overall measure of deviation from random association. The log likelihood ratio test statistic from the observed data, here denoted as t, will have mean and $f + N\xi$ variance $2(f+2\,N\xi)$, with f being its degrees of freedom. Let the mean and variance of the likelihood ratio test statistic from its empirical distribution obtained by permutation be μ and σ^2, the LD measure as proposed in Zhao et al. [21] is estimated from replicate samples as $\hat{\xi} = \sqrt{2f}\,[(t - \mu)/\sigma]/N$. This measure uses the scaled standardised difference of observed log likelihood ratio test statistic and its empirical mean without heavy reliance on asymptotic result. The sample variance of $\hat{\xi}$, $2(f + 2N\hat{\xi})/N^2$, can be used to construct confidence intervals. For a set of markers from a chromosome segment, we can also use the measure to examine variations of LD on different marker subsets [22], as will be illustrated below.

Example and Application

We illustrate the new method with two analyses. The first is a re-analysis of the HLA data set in Zhao et al. [18] which demonstrates the relative performance of the different schemes of gene counting described above. The second analysis illustrates extraction of LD information from the likelihoods of permutation replicates. We used the same set-up as EH for both examples.

Example 1: Association between Schizophrenia and HLA Markers

HLA markers are located in a gene-rich region on chromosome 6. In Zhao et al. [18], markers DRB, DQA and DQB were used for allelic association in a study of 94 schizophrenic patients and 177 controls. Markers DRB, DQA and DQB were supposed to have 25, 10 and 15 alleles. One patient with incomplete genotypes was left out from the analysis. While potentially there are 2,145,000 possible three-locus genotypes, only 163 of them were actually observed in the sample.

Three hypotheses of allelic association were considered. H0: No association between three markers so that haplotype frequencies are simply the equilibrium frequencies, being the products of frequencies from constituent alleles. Each locus therefore contributes number of alleles-1 free parameter(s), and the total number of free parameters is equal to the sum of the contributions from all loci (including the disease locus). H1: Association between markers but not with disease locus. The number of free parameters is now the number of haplotypes-1. H2: Both markers and disease locus are associated. The number of free parameters is twice the number of haplotypes-2. In the sample, DRB and DQA had only 24 and 9 alleles, respectively, so that number of free parameters for the three hypotheses was adjusted accordingly. These are given as (24 – 1) + (9 – 1) + (10 – 1) = 40, (24)(9)(15) – 1 = 3,239, and (2)(24)(9)(15) – 2 = 6,478. We also obtained heterogeneity statistics between cases and controls as in Zhao et al. [18]: –2(log likelihood[cases + controls] – log likelihood[controls] – log-likelihood[cases]), which has (24)(9)(15) = 3,239 degrees of freedom.

The original analysis was conducted on DEC Alpha and Sun SPARC stations, which were comparable in speed with a Pentium 500 HMz PC with 256 MB memory running MicroSoft Windows 98. A comparison of different implementations is given in table 1. As our original interests were to conduct case-control analysis involving all three markers, this was extensively tested for different search methods. Assuming a disease allele frequency of

Table 1. A comparison of different implementations for HLA data

Type of analysis	Platform	Method	Time	
Case-control with an explicit disease model	Single analysis	DEC Alpha, Sun SPARC station, or Pentium 500MHz PC	binary search	3–4 days
		DEC Alpha	hashing	4 h
		Pentium 500MHz PC	sentinel variable	36 min
		Pentium 500MHz PC	caching	6 min
	10,000 permutations	Pentium 500MHz PC	caching	15 min
Marker-marker analysis	10,000 permutations	Pentium 500MHz PC	caching	23 min

0.1 and penetrances of 0.005, 0.005, 0.5 [23], the log likelihoods under H2 and H1 were −1518.48 and −1594.75, respectively, yielding a log likelihood ratio χ^2 statistic testing H2 against H1 of 152.54 (nominally on 6,478 – 3,239 degrees of freedom). As the log likelihoods thus obtained were model-dependent and asymptotic χ^2 approximation was probably unreliable, we used a heterogeneity statistic and our new implementation for 10,000 permutations. The heterogeneity statistic was 196.83 with 3,239 degrees of freedom, giving a p value of near 1.0 based on asymptotic theory, yet none of these replicates yielded a heterogeneity statistic as extreme as 196.83. The estimated empirical p value was therefore less than 0.00001 (i.e. 1/10,000). log likelihoods without correction under H2 and H1 were −1949.00 and −2,025.26, respectively leading to a likelihood ratio χ^2 statistic 152.52. With respect to H1 and H0 we obtained a log-likelihood ratio test statistic of 3172.18 from the combined data of cases and controls involving three markers, with 3,199 degrees of freedom. To obtain a nominal p value we invoked a permutation procedure with 10,000 replicates, and none of the permuted samples yielded any test statistic value as extreme as 3,172.18, suggesting that the result is at least significant at level 0.00001.

Example 2: Association between Alcoholism and ALDH2 Region

The ALDH2 locus is located on chromosome 12 and plays an important role in ethanol metabolism. In oriental populations, ALDH2 exists in two forms that differ in activity due to a G → A mutation in exon XII resulting in a lysine-for-glutamine substitution. Our study was conducted to examine association between alcoholism and alleles of several simple sequence repeat polymorphisms and a single nucleotide polymorphism in the ALDH2 region. The sample consisted of 130 alcoholics and 136 controls. Six microsatellite markers and two single nucleotide polymorphisms were genotyped in the ALDH2 region: D12S2070, D12S839, D12S821, D12S1344, EXON XII, EXON1, D12S2263, D12S1341; the number of alleles at these loci were 8, 8, 13, 14, 2, 2, 13, 10. The physical distances (in base pairs) of these polymorphisms, relative to EXON XII, are as follows.

>450,000, >450,000, ~400,000, 83,853, 0, 37,335, 38,927, >450,000

Both single marker and haplotype analyses revealed strong disequilibria in this region. Haplotype analysis gave stronger evidence than single marker analysis. Permutation tests usually provided smaller p values than asymptotic results. Here we repeated part of the analysis on two markers on either side of the functional locus EXON XII to obtain permutation-based LD measures, as our interests were in detection of a functional gene using LD on the premise that we do not a priori know the functional locus. They were conveniently numbered 1, 2, 3, 4 and all subsets they formed are listed as follows.

I	12	D12S821-D12S1344
II	13	D12S821-EXON1
III	14	D12S821-D12S2263
IV	23	D12S1344-EXON1
V	24	D12S1344-D12S2263
VI	34	EXON1-D12S2263
VII	123	D12S821-D12S1344-EXON1
VIII	124	D12S821-D12S1344-D12S2263
IX	134	D12S821-EXON1-D12S2263
X	234	D12S1344-EXON1-D12S2263
XI	1234	D12S821-D12S1344-EXON1-D12S2263

The results are shown in table 2. It is interesting that for two-locus sets, LD estimates are usually between 0 and

Table 2. All-subset LD measures for D12S821-D12S1344-EXON1-D12S2263

Subset	Permutation-based LD measures							
	case only		controls only		cases + controls		heterogeneity	
	$\hat{\xi}$	SE	$\hat{\xi}$	SE	$\hat{\xi}$	SE	$\hat{\xi}$	SE
Two-locus set								
I	1.55	0.27	0.93	0.22	1.33	0.17	0.20	0.11
II	0.53	0.15	0.54	0.16	0.58	0.11	0.06	0.05
III	0.35	0.16	0.32	0.18	0.36	0.11	0.39	0.12
IV	0.99	0.20	0.77	0.18	0.92	0.13	0.05	0.04
V	0.73	0.20	0.24	0.16	0.51	0.12	0.33	0.11
VI	0.86	0.18	0.57	0.16	0.77	0.12	0.21	0.07
Three-locus set								
VII	3.47	0.43	2.66	0.40	3.26	0.29	0.19	0.15
VIII	3.92	0.63	2.80	0.61	3.77	0.42	0.79	0.35
IX	1.74	0.33	1.28	0.35	1.82	0.24	0.40	0.16
X	2.85	0.39	2.03	0.36	2.53	0.25	0.55	0.16
Four-locus set								
XI	7.92	0.94	6.72	0.99	8.38	0.66	1.10	0.52

$\hat{\xi}$ = Permutation-based global LD measure; SE = standard error of $\hat{\xi}$.

1, being slightly larger in cases than in controls. For three-locus and four-locus sets LD tends to be stronger in cases than in controls. The heterogeneity LD measure and its sample variance were also calculated from $\hat{\xi} = \sqrt{2f}[(t - \mu)/\sigma]/N$ and $(2f + 2N\hat{\xi})/N^2$, where t is now the sample heterogeneity statistic similar to example 1 and f its degrees of freedom, while μ and σ^2 are the mean and variance of the test statistic from replicate samples. The heterogeneity LD measure might be conceived as a measure of effective size for discrepancy of case-control haplotype frequencies, given that association is detected in both cases and controls.

Discussion

We describe the effect of using a more efficient algorithm, based on genotype caching, in haplotype analysis of unrelated individuals. We also present a faster implementation for providing data to the algorithm. Both improvements make the permutation-based methods more feasible. We also implemented the permutation-based disequilibrium measure [21].

Binary search trees have been described in Zaykin et al. [24] and the caching method in linkage analysis has been described in Cottingham et al. [25]. Binary search normally has complexity of order O(log(N)) (N being the number of items), but is very slow for the whole analysis. Hashing has an order of O(1), but uses more computer memory. Linked list is easy to build but slower than search trees, a factor becoming important when the sample size is large or replicate analyses are necessary. Binary search trees roughly also have order of O(log(N)) for N individuals with complete genotypic information. As the pattern of multilocus genotypes for individuals is usually irregular, if not totally random, a binary search tree would suffice. This avoids the need to consider more sophisticated data structures such as skip lists [26] and balanced trees [20]. Since our proposal in EHPLUS was to use permutation tests and each permutation generated a new set of individuals, the improvement would have significant effect when the number of replicates is large. The speed-up via a caching scheme was achieved by significant reduction of the number of loops, as in the first example application the potential number of loops changed from 2,145,000 to 163.

The improvements described here would increase the size of problems that can be analysed for case-control association using permutation methods. Nevertheless, the computational demands for very large problems may remain daunting. For example, the alcoholism and ALDH2 data involve 8 × 8 × 13 × 13 × 14 × 2 × 2 × 10 = 6,056,960 possible haplotypes and $8(8 + 1)\ 8(8 + 1)\ 13(13 + 1)\ 13(13 + 1)\ 14(14 + 1)\ 2(2 + 1)\ 2(2 + 1)\ 10(10 + 1)/2^8$ possible genotypes, a method for efficiently dealing with very rare haplotypes is then desirable. Although we have chosen to compare our implementations using practical examples, the differences are actually clear in terms of their complexities. Let M and N be numbers of possible and observed genotypes, h be the maximum number of heterozygotes for these genotypes. For the previous procedure, the complexity was dominated by both M and O($N2^h$) while the new procedure is O($N2^h$) only. Not only M is large relative to N, but also previous methods involve search through such a large array or use of a sentinel to keep track of the genotypes. In general, the number of multiply heterozygous markers would depend on their heterozygosities and the underlying haplotype patterns. We expect our implementation will be appropriate for problems of moderate size, and the Markov chain Monte Carlo approach is an important alternative for very large problems [27–29].

A program that incorporates the current implementation is freely available from the first author (E-Mail: j.zhao@iop.kcl.ac.uk) or our website (http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.stm).

Acknowledgements

We are grateful to Prof. Jürg Ott and Dr. David Curtis for critical comments, Dr. Padraig Wright for providing the schizophrenia and HLA data, Prof. Ian Craig for the alcoholism and ALDH2 data, Dr. Stefan Jenisch for pointing out similar work of his team on HLA, Dr. Frank Dudbridge for discussing an issue of impossible haplotypes in case-control data with an illuminating example, Dr. T. Li, Dr. J. Tang, Dr. Kay Chapman for real-data analysis, and Dr. Dimitri Zaykin for suggestions on improvements of the program. Finally, we thank two anonymous reviewers for their helpful comments. This work was supported by the Wellcome Trust project grant 055379, NIH grant EY-12562 and MRC grant G9901258.

References

1 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.
2 Schneider S, Kueffer JM, Roessli D, Excoffier L: Arlequin: A software for population genetic analysis. http://anthopologie.unige.chh/arlequin/. 2000.
3 Lewis PO, Zaykin D: Genetic data analysis: Computer program for the analysis of allelic data. http://lewis.eeb.uconn.edu/lewishome/software.html. 2001.
4 Jenisch S, Westphal E, Nair RP, Stuart P, Voorhees JJ, Christophers E, Kronke M, Elder, JT, Tilo, H: Linkage disequilibrium analysis of familial psoriasis: Identification of multiple disease-associated MHC haplotypes. Tissue Antigens 1999;53:135.
5 Kidd JR, Pakstis AJ, Zhao H, Lu R-B, Okonofua FE, Odunsi A, Grigorenko E, Bonne-Tamir B, Friedlaender J, Schulz, LO, Parnas J, Kidd, KK: Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of population. Am J Hum Genet 2000;66:1882–1899.
6 Xie X, Ott J: Testing linkage disequilibrium between a disease gene and marker loci. Am J Hum Genet 1993;53:1107.
7 Terwilliger JD, Ott J: Handbook of Human Genetic Linkage. Baltimore, The Johns Hopkins University Press, 1994.
8 Thompson EA, Deeb S, Walker D, Motulsky AG: The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. Am J Hum Genet 1988;42:113–124.
9 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.
10 Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK: Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 2000;67: 936–936.
11 Ceppellini R, Siniscalco M, Smith CAB: The estimation of gene frequencies in a random mating population. Ann Hum Genet 1955;20: 97–115.
12 Smith CAB: Counting methods in genetical statistics. Ann Hum Genet 1957;21:254–276.
13 Hill WG: Tests for association of gene frequencies at several loci in random mating diploid populations. Biometrics 1975;31:881–888.
14 Dempster AP, Laird N, Rubin DB: Maximum likelihood from incomplete data via the MX algorithm. J R Stat Soc B 1977;39:1–38.
15 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12: 921–927.
16 Long JC, Williams RC, Urbanek M: An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 1995;56: 799–810.
17 Workman PL, Niswander JD: Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Am J Hum Genet 1970;22:24–39.
18 Zhao JH, Curtis D, Sham PC: Model-free analysis and permutation tests for allelic associations. Hum Hered 2000;50:133–139.
19 Koch HG, McClay J, Loh E-W, Higuchi S, Zhao J-H, Sham P, Ball D, Craig IW: Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. Hum Mol Genet 2000; 9:2993–2999.
20 Knuth DE: The Art of Computer Programming. Vol. 3 Sorting and Search. Reading, Addison-Wesley, 1998, pp 409–423, 426–454, 513–549.
21 Zhao H, Pakstis AJ, Kidd JR, Kidd KK: Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. Ann Hum Genet 1999;63: 167–179.
22 Furnival GM, Wilson RW Jr: Regression by leaps and bounds. Technometrics 1974;16: 499–511.
23 Murray R: Schizophrenia; in Murray R, Hill P, McGuffin P (eds): The Essentials of Postgraduate Psychiatry, ed 3. Cambridge University Press, 1997, pp 281–309.
24 Zaykin D, Zhivotovsky L, Weir BS: Exact tests for association between alleles at arbitrary numbers of loci. Genetica 1995;96:169–178.
25 Cottingham RWJ, Idury RM, Schaffer AA: Faster sequential genetic linkage computations. Am J Hum Genet 1993;53:252–263.
26 Pugh W: Skip lists: A probabilistic alternative to balanced trees. Comm Assoc Comput Mach 1990;33:668–676.
27 Lazzeroni LC, Lange K: Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. Ann Statist 1997;25:138–168.
28 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68:978-989.
29 Niu T, Qin ZS, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 2002;70:157–169.