BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 8 2004, pages 1325–1326
DOI: 10.1093/bioinformatics/bth071

2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis

Jing Hua Zhao

Department of Epidemiology and Public Health, University College London, 1-19 Torrington Place, London WC1E 6BT, UK

Received on August 31, 2003; revised on November 7, 2003; accepted on November 17, 2003
Advance Access publication February 10, 2004

ABSTRACT

Summary: Computer programs are introduced which calculate pair-wise linkage disequilibrium statistics and conduct haplotype frequency estimation, including X chromosome data, and using a heuristic algorithm to handle multiple genetic markers and missing data.

Availability: Programs 2LD, GENECOUNTING and HAP are available on Internet from http://www.hgmp.mrc.ac.uk/~jzhao and http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/software.shtml

Contact: jzhao@hgmp.mrc.ac.uk

Linkage disequilibrium (LD) refers to the dependence of alleles from neighbouring loci and can provide information on population histories and disease mapping. A widely used statistic measuring pairwise LD between single nucleotide polymorphisms (SNPs) and/or multiallelic markers is Hedrick's D', which is based on two-locus haplotype frequencies. Moreover, haplotype frequency estimation involving multiple loci and unrelated individuals is often necessary, e.g. for examining haplotype effects in a case–control study. We encountered several difficulties in our haplotype analysis of unrelated individuals. First, the sampling variance of D' was not easy to obtain, especially when multiallelic markers were involved. Second, we were not able to find a computer program to conduct analysis of X chromosome data containing both males and females. Third, the gene-counting method was quite limited by the number of markers due to the large amount of computing time required.

I have developed computer programs to address these limitations: the sampling variance of D' as derived by Zapata *et al*. (2001)—which has complicated form—has been implemented in 2LD; an algorithm and facility for a general gene counting algorithm including X chromosome data have been included in GENECOUNTING (Zhao *et al*., 2002); an algorithm that handles both SNPs and multiallelic markers are implemented in HAP, based on SNPHAP (http://www-gene.cimr.cam.ac.uk/clayton/software). I now give a brief description of these programs.

2LD is a two-locus LD calculator for two multiallelic markers, including SNPs. It gives the D' estimate and standard error as well as a number of association tests. It was designed to avoid the need for users to specify a parameter file and automatically recognizes three types of inputs: phased haplotype frequencies, two-locus genotype table and raw genotype data. This allows for the direct use of haplotype frequencies from other packages, as well as estimation from raw data by the program. For raw genotype data there are several tests of association available: a test of independence between two markers based on the observed genotype table without assuming Hardy–Weinberg equilibrium, a log-likelihood ratio χ^2 statistic assuming allelic association and Hardy–Weinberg equilibrium, the residual χ^2 statistic between the two and χ^2 statistic based on estimated haplotype counts, which is also used to obtain Cramer's V statistic.

GENECOUNTING employs the standard EM algorithm for haplotype frequency estimation and now extends its algorithm to handle data on X chromosome, so that for such data instead of using only males to avoid phase ambiguity, data from both sexes can be used in a single framework. In the new algorithm, haplotypes of females are counted as before (Zhao *et al*., 2002), while those from males are added directly to the total haplotype counts; the haplotype frequency estimates for that particular iteration of the EM algorithm are obtained by dividing the total haplotype counts by the total number of chromosomes. By assuming missing at random, missing genotype data in both males and females can readily be utilized by listing all haplotypes compatible with the observed with contribution to the total counts weighted by the genotype probabilities.

HAP employs the EM algorithm and multiple imputations for haplotype estimation involving multiallelic markers. The algorithm adds one locus at a time according to a given order and drops some haplotypes with frequencies below certain trimming threshold. The counting of haplotypes is via sorting and collecting unphased genotypes enumerated and stored in computer memory from the last iteration. For heterozygous loci, the algorithm lists all possible haplotypes and the sorting has time complexity of $O(N \ln N)$ with N being number of all listed haplotypes at that iteration. This avoids indexing individual haplotypes using Horner's algorithm (Zhao and Sham, 2003), so the computing time is considerably reduced for a large problem. For instance, in a PC with 256 MB RAM, it can handle well over 100 SNPs quickly. In case of no trimming, it yields results comparable to that obtained using GENECOUNTING. This is a utility program that collects output from multiple imputations to SAS data step programs.

Next, I present some results of applying these programs to experimental and real data.

For SNPs, 2LD yields identical results obtained from expressions specific for SNPs (Zapata *et al*., 1997) and for multiallelic markers. Zapata *et al*. (2001) reported that for a small number of alleles the analytical expressions were undesirable according to Monte Carlo experiments. Data in some of their experiments had been obtained via a QBASIC program. For instance, the first four-allele example in their Table 1 had asymptotic variance 0.0017 and Monte Carlo 0.0014, but 2LD yielded 0.001422, much closer to the Monte Carlo result.

In a study of Parkinson's disease, a total of 183 patients and 157 controls (150 males, 190 females) were available, together with five markers in monoamine oxidase A (MAOA) region with alleles 12, 9, 6, 5, 3, and the first three markers were genotyped in all individuals while the fourth and fifth were genotyped for 294 and 304 individuals. Thirty-one iterations using GENECOUNTING are completed in 1 min on a HGMP (http://www.hgmp.mrc.ac.uk) Unix workstation *tin* yielding log-likelihoods of -2554.88 and -2099.27 with and without assuming linkage equilibrium, respectively.

Data on alcoholism (130 alcoholics and 133 controls) reported elsewhere were used to compare GENECOUNTING and HAP. Eight markers, D12S2070, D12S839, D12S821, D12S1344, EXON XII, EXON1, D12S2263 and D12S1341, were geno-typed in the ALDH2 region; the number of alleles at these markers were 8, 8, 13, 14, 2, 2, 13, 10, respectively, leading to $8 \times 8 \times 13 \times 14 \times 2 \times 2 \times 13 \times 10 = 6\,056\,960$ possible haplotypes. Of the 263 individuals, 93 had incomplete information. It was time-consuming to use missing genotype data. GENECOUNTING took roughly a month on the HGMP Unix workstation iron to analyse all eight markers, compared to only several minutes on the same machine with the threshold for posterior haplotype trimming being 10^{-3}.

Note that the EM algorithm is for unrelated individuals and differs from other programs that use family data. It also assumes Hardy–Weinberg equilibrium. Some further remarks are worthwhile. First, although limited to two loci, 2LD is very easy to use and offers more statistics than most of the other programs available; it has been used in junction with other haplotype analysis program Hyun *et al*. (2003). Second, GENECOUNTING is appropriate for analysis of sets of markers via slide-windows (Zaykin *et al*., 2002). A shell program to GENECOUNTING has been written to obtain permutation and haplotype specific tests. Third, as HAP uses a heuristic algorithm, the log-likelihood under haplotype trimming can be smaller than that from a brute-force algorithm in GENECOUNTING but more appropriate for including many SNPs in one analysis. The algorithm is more time-consuming for a large sample size if not collapsing for individuals with similar multilocus genotypes. Finally, although they have been incorporated in an R package in conjunction with analyses described by Schaid *et al*. (2002) and Zaykin *et al*. (2002), as individual programs they are appealing to many in particular applications.

ACKNOWLEDGEMENTS

I wish to thank Dr Carlos Zapata for communications on 2LD, Professor Pak Sham and other colleagues for collaborative work on earlier version of GENECOUNTING, Dr Helen Latsoudis, Professors David Collier and Ian Craig for Parkinson's and alcoholism data. The work is partly supported by NIA grant (AG13196) to the Whitehall II study.

REFERENCES

Hyun,C., Filippich,L.J., Lea,R.A., Shepherd,G., Hughes,I.P. and Griffiths,L.R. (2003). Prospectus for whole genome linkage disequilibrium mapping in domestic dog breeds. *Mamm. Genome*, **14**, 640–649.

Schaid,D.J., Rowland,C.M., Tines,D.E., Jacobson,R.M. and Poland,G.A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **70**, 425–434.

Zapata,C., Alvarez,G. and Carollo,C. (1997). Approximate variance of the standardized measure of gametic disequilibrium D'. *Am. J. Hum. Genet.*, **61**, 771–774.

Zapata,C., Carollo,C. and Rodriguez,S. (2001) Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. *Ann. Hum. Genet.*, **65**, 395–406.

Zaykin,D.V., Westfall,P.H., Young,S.S, Karnoub,K.M., Wagner,M.J. and Ehm,M.G. (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum. Hered.*, **53**, 79–91.

Zhao,J.H. Lissarrague,S., Essioux,L. and Sham,P.C. (2002) GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics*, **18**, 1694–1695.

Zhao,J.H. and Sham,P.C. (2003). Generic number system and haplotype analysis. *Comput. Meth. Prog. Biomed.*, **70**, 1–9.