BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 14 2005, pages 3168–3170
doi:10.1093/bioinformatics/bti453

Sequence analysis

Association mapping and fine mapping with TreeLD

Sebastian Zöllner*, Xiaoquan Wen and Jonathan K. Pritchard

Department of Human Genetics, The University of Chicago, 920 East 58th Street–CLSC 507, Chicago, IL 60637, USA

Received on January 31, 2005; revised on April 12, 2005; accepted on April 13, 2005
Advance Access publication April 26, 2005

ABSTRACT

Summary: The program package TreeLD implements a unified approach to association mapping and fine mapping of complex trait loci and a novel approach to visualizing association data, based on an inferred ancestry of the sample. Fundamentally, the TreeLD approach is based on the idea that the evidence for association at a particular position is contained in the ancestral tree relating the sampled chromosomes at that position. TreeLD provides an easy-to-use interface and can be applied to case–control, TDT trio and quantitative trait data.

Availability: The program TreeLD is available on the Internet from http://pritch.bsd.uchicago.edu/software.html

Contact: szoellne@genetics.bsd.uchicago.edu

INTRODUCTION

Recent advances in genotyping technology make it feasible to generate dense single nucleotide polymorphism (SNP) data for use in linkage disequilibrium (LD) mapping of complex trait loci. However, there is currently a lack of effective computational tools for analyzing such data. Thus, we announce the release of the program package TreeLD 1.0, a tool for multipoint LD-analysis that applies a unified approach to test for the presence of a disease locus and to fine-map its location. Furthermore, the package provides a novel way of visualizing the association signal in a sample. TreeLD can be applied to case–control data and TDT trio data and it is also the first program of its kind that can map quantitative trait loci (QTLs). The package has a user-friendly point-and-click interface, allowing an in-depth exploration of the data. Executables for any of Windows 2000/XP, Mac OS X, or Linux and Solaris operating systems and extensive documentation for TreeLD can be downloaded from the website given in the summary. The source code is available from the authors upon request. In the following, we describe the conceptual basis for the algorithm, give an overview of the program and its features and describe its application to a set of quantitative trait data.

OVERVIEW OF THE METHOD

The algorithm employed in TreeLD makes use of the observation that association between a phenotype and a marker is the result of shared ancestry among cases in the vicinity of that marker. Indeed under certain conditions, the ancestry of the disease locus incorporates the entire evidence about the location of the disease mutation that is contained in the marker data. Thus, it is natural to estimate the ancestry of a region-of-interest based on the marker information and to explore this ancestry for non-random distribution of phenotypes among its terminal nodes, indicating the presence of a disease mutation (Fig. 1). For an in-depth description of these ideas and of the mathematical algorithm, please refer to Zöllner and Pritchard (2005).

TreeLD is designed for analyzing a sample of unrelated individuals that have either been classified as cases and controls or measured for a quantitative trait of interest. TDT data can also be analyzed by considering the non-transmitted chromosomes as sampled from unrelated control individuals. The input file for TreeLD consists of one text file containing the marker locations, phenotype information and marker haplotypes for all individuals in the sample. The program in its present form is designed for bi-allelic markers. Furthermore, it can only be used on phased data; therefore, we suggest generating haplotypes using an estimation method such as PHASE 2.1 (Stephens *et al*., 2001; Stephens and Donnelly, 2003). A Perl script that parses the output of PHASE 2.1 into the input format for TreeLD is included in our program package.

After the input data are read in, TreeLD estimates the coalescent ancestry of the sampled chromosomes at each of a series of positions ('focal points') across the region-of-interest. The number and spacing of focal points determines the resolution of the final analysis and is selected by the user. Although the algorithm estimates the ancestry of the chromosome at distinct focal points, the analysis makes full multipoint use of the SNP data, since SNPs at varying distances contain information about the ancestry at different levels in the tree. Our model does not assume that the data can be described by a series of distinct haplotype blocks, as in some other methods.

At each focal point, the evidence for association is assessed using a sample of trees that are consistent with the entire marker data, and using a Markov chain Monte Carlo (MCMC) algorithm (Gilks *et al*., 1996) to sample ancestries from the distribution Pr(Tree|Marker data). To evaluate convergence of the MCMC, the program plots the probability of the generated trees conditional on the marker data. After a user-assigned burn-in period, a number of trees are sampled with a fixed thinning interval. This set of likely ancestries is displayed for the user to examine, providing a novel way to visualize the information in the data. By identifying which clades carry an excess of case chromosomes, the user can determine which affected individuals are most likely to carry mutations at this locus and which are more likely to be phenocopies.

For a formal analysis, the likelihood of the phenotype data, given the set of trees sampled at each focal point, can be computed using a peeling algorithm (Felsenstein, 1981). For this purpose, the user can either enter the penetrances of the susceptibility and the wild-type alleles, or choose to integrate over a set of penetrance parameters. For each tree sampled from the MCMC, the program then calculates the probability of the observed phenotypes, assuming that this tree represents the ancestry of a disease locus. By averaging over all trees at one focal point, a term that is proportional to the posterior probability of that focal point being the locus of the disease mutation is calculated. The results can be transformed into a Bayesian posterior distribution for the location of the disease mutation(s) by normalizing over all focal points. Based on this distribution, credible regions for the disease locus can be generated and the maximum of the distribution can be chosen as a point estimator (Fig. 2).

*To whom correspondence should be addressed.

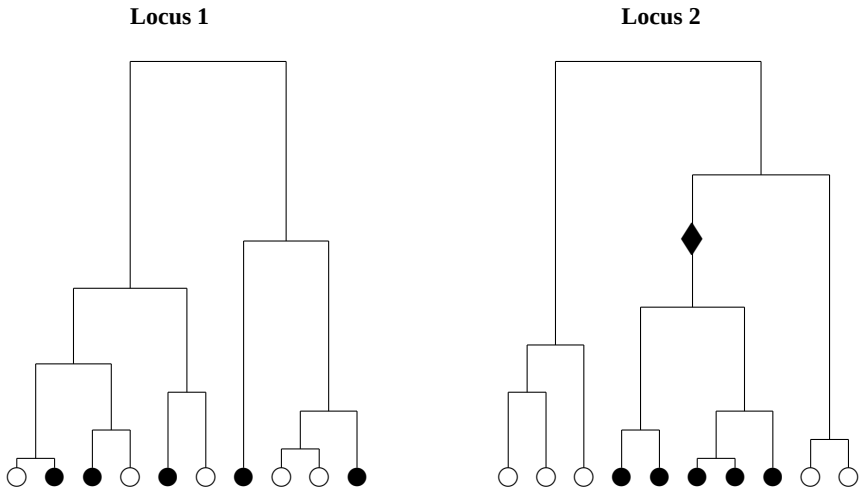

Fig. 1. Hypothetical example of coalescent genealogies at two different loci for a case–control sample of 10 chromosomes. Each circle at the bottom of the tree represents a sampled chromosome; closed circles representing chromosomes from cases and open circles representing chromosomes from controls. The lines indicate the ancestral relationships among the chromosomes. On the tree of Locus 1, the case and control chromosomes appear to be randomly distributed among the tips of the tree, while on the tree at Locus 2 all case chromosomes are descendants of one branch (marked by the diamond). This indicates that Locus 2 is more likely to be close to the disease locus as the pattern of cases can be explained by a single disease mutation event. In contrast, the tree at Locus 1 shows less indication that Locus 1 is near a disease locus, as it would require multiple mutations and/or low phenotype–genotype correlation to explain the observed distribution of case chromosomes. In practice, for a complex trait, we expect the degree of clustering to be less pronounced than illustrated here.

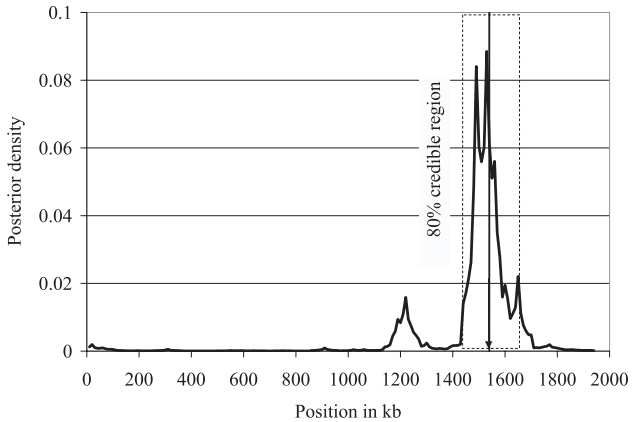

Fig. 2. Posterior distribution for the location of a QTL. The graph displays the posterior distribution, the vertical arrow indicates the true location of the QTL, the dashed box contains the 80% credible region reaching from 1485 to 1655 kb. The 95% credible region extends from 1185 to 1655 kb.

To perform a significance test for association based on the generated likelihoods, two options are available in TreeLD: a likelihood-ratio test based on a χ^2-approximation which is fast but slightly conservative and a more powerful permutation test that automatically corrects for multiple tests. Applied to a set of 50 datasets, this method has shown 25% higher power than a single marker test using Pearson's χ^2-test for independence (Zöllner and Pritchard, 2005).

Given that some steps in the analysis are computationally intensive, it is desirable to run TreeLD in parallel on multiple machines. To this end, the front-end generates scripts that can be run independently on separate processors. The largest dataset we have analyzed with TreeLD consisted of 838 chromosomes typed at 36 markers. On a small cluster, the program can be used to analyze datasets of up to 500 individuals and up to 200 markers in a few days. Information about expected runtimes can be found in the documentation file.

RESULTS FOR A QUANTITATIVE TRAIT

We applied TreeLD to a quantitative trait dataset, generated by simulating a coalescent ancestry for a sample of 100 diploid individuals and 100 bi-allelic markers distributed over 2 Mb. At a randomly selected QTL-locus within those 2 Mb, we generated mutations in the ancestry with a scaled mutation rate of $\nu = 0.4$. Depending on the resulting genotype, each individual in the sample was randomly assigned a phenotype. These phenotypes were generated from a normal distribution with a standard deviation of 2.0 and a mean of 0.0 for a homozygote wild-type, 2.0 for a heterozygote and 8.0 for a homozygote mutant.

The generated dataset was analyzed with TreeLD, and after a burn-in of 3×10^6 updates per focal point, we sampled 100 trees at each of 200 focal points and generated the posterior distribution, based on those ancestries (Fig. 2). The peak of the posterior distribution is within 10 kb of the real functional variant, indicating that TreeLD is able to precisely estimate the location of this QTL. Furthermore, most of the 2 Mb region has essentially zero posterior probability of carrying the disease mutation. Thus, the resulting credible region is a relatively narrow window of 170 kb. This is consistent with earlier observations that TreeLD generates precise estimates and narrow credible regions compared with other methods of fine-mapping (Zöllner and Pritchard, 2005).

In summary, the package TreeLD provides a set of powerful tools for the analysis of association data with an easy to use interface. The structure of the algorithm facilitates extensions, and we plan to release additional features for TreeLD to provide further options as free tools to the research community.

ACKNOWLEDGEMENT

This work was supported by grant HG 2772 to J.K.P. from the National Institutes of Health.

REFERENCES

Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Gilks,W.R., Richardson,S. and Spiegelhalter,D.J. (1996) Introducing Markov chain Monte Carlo. In Gilks,W.R., Richardson,S. and Spiegelhalter,D.J. (eds), *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, pp. 1–19.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Stephens,M. *et al*. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.

Zöllner,S. and Pritchard,J.K. (2005) Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*, **169**, 1071–1092.