Statistical software for gene mapping by admixture linkage disequilibrium

Giovanni Montana and Clive Hoggart

Received (in revised form): 7th June 2007

Abstract

Admixture mapping is a statistical methodology that detects genetic variants in recently admixed populations that are responsible for ethnic differences in disease risk. Three software packages are now available for admixture mapping and we provide a brief overview of the statistical methods and other principal features they implement.

Keywords: gene mapping; admixture linkage disequilibrium; hidden Markov models

INTRODUCTION

Empirical evidence shows that, in human populations formed by relatively recent mixing of distinct ancestral groups, linkage disequilibrium (LD) is observed over greater distances than in other, less heterogeneous populations. For example, in African-Americans, weak LD may persist over distances as large as 20 cM [1]. For diseases which vary in prevalence between two or more ancestral populations, the long range LD observed in admixed populations can be exploited to search for genetic variants responsible for the ethnic difference in disease risk. The resulting methodology is referred to as admixture mapping or mapping by admixture linkage disequilibrium (MALD).

The principle underlying a MALD study is simple; loci in LD with a locus responsible for an ethnic difference in disease risk will have a greater than expected proportion of markers with ancestry from the high risk population. In these studies, a sample of diseased individuals from the admixed population under study is collected and the probability of ancestry of each marker to each ancestral population and the proportion of ancestry of each individual to the ancestral populations is statistically determined. An unexpectedly large bump of ancestry in a localized genomic region then suggests that the region may harbour a disease-bearing gene.

When compared to more traditional association studies, MALD requires far fewer markers with the extra advantage of protecting against allelic heterogeneity. Extensive reviews illustrating advantages and disadvantages of this promising gene mapping strategy have been written [2–5]. Further guidelines are also available [6–8]. In this review we describe the three main publicly available software packages specifically designed for admixture mapping: the suite of programs STRUCTURE2/MALDsoft [8, 9], ADMIXMAP [6] and ANCESTRYMAP [7].

DOWNLOAD AND DATA INPUT

STRUCTURE2 and MALDsoft are executable programs available for Windows, Linux and Solaris and can be downloaded with documentation at http://pritch.bsd.uchicago.edu/software.html. They are meant to be used jointly, and they both share the same data format. Careful error checking is implemented in these programs to make sure that the data set is in the correct format.


Corresponding author. Giovanni Montana, Imperial College London, Department of Mathematics, South Kensington Campus, 180 Queen's Gate, London SW7 2AZ. Tel: +44 (0) 207-594-8577; E-mail: g.montana@imperial.ac.uk

Giovanni Montana is a Lecturer of Statistics in the Mathematics Department, Imperial College London, UK. He is interested in data mining and statistical pattern recognition.

Clive Hoggart is a Research Fellow in Statistical Genetics in the Department of Epidemiology and Public Health, Imperial College London, UK. He is interested in statistical genetics and Bayesian methodology.

© The Author 2007. Published by Oxford University Press. For Permissions, please email: journals.permissions@oxfordjournals.org
When importing the data, ANCESTRYMAP checks for duplicate individuals and misspecification of marker position. Both ANCESTRYMAP and ADMIXMAP can test for Hardy–Weinberg equilibrium at each marker and discrepancy in the ancestry specific allele frequencies in the admixed and unadmixed populations.

All programs have varying input formats. However, code to translate between ADMIXMAP and ANCESTRYMAP format and from STRUCTURE2 format to ADMIXMAP is available from http://www.ucd.ie/genepi/admixmap/tools.html. All programs are executed from the command line, but a graphical front-end is available for STRUCTURE2.

STATISTICAL MODELS
The three programs implement similar statistical methodologies. In particular, the same hierarchical model for population admixture, individual admixture and locus ancestry is implemented. They also share the same probability model for the stochastic variation of ancestry along the chromosomes; it assumes that chromosomes can be broken into blocks of common ancestry, breakpoints between adjacent blocks occur as a Poisson process and transitions between adjacent ancestral blocks are governed by a Markov model.

Using MCMC simulations, the three programs can produce posterior summaries of allele frequencies, individual ancestry proportions, and a measure of the time since admixture. ADMIXMAP can model any number of ancestral populations, whereas MALDsoft and ANCESTRYMAP are limited to modelling admixture between two ancestral populations only. ANCESTRYMAP allows only for biallelic markers, whereas the other programs can handle markers with any number of alleles. ADMIXMAP allows for allelic association between markers by modelling the unobserved haplotypes and thus tightly linked loci can be included in the analysis.

All programs allow prior distributions for allele frequencies to be specified. In ANCESTRYMAP and ADMIXMAP these priors are specified in a separate file containing allele counts from unadmixed individuals, whereas in STRUCTURE/MALDsoft admixed and unadmixed individuals are included in the same file and their status is specified by an indicator variable. The default setting of STRUCTURE/MALDsoft assumes that markers have not been selected to be informative for ancestry and thus a prior that imposes correlation between the allele frequencies in each subpopulation at each marker is specified. ADMIXMAP and ANCESTRYMAP do not in general assume any correlation between the allele frequencies in each subpopulation. ANCESTRYMAP can model X chromosome marker data from males and females, whereas the other programs are restricted to autosomal data.

TESTING FOR LINKAGE
All three programs can test for linkage in an affected-only and case-control study design. The latter is generally less powerful than an affected-only study [2, 6, 8], but the former can be sensitive to misspecified allele frequencies. ADMIXMAP and ANCESTRYMAP implement a one stage test for linkage in which the test is calculated within the MCMC sampler for admixture conditional on the current realized values of the model parameters, whereas MALDsoft calculates tests for linkage using the final output of STRUCTURE2; the process of importing STRUCTURE2 output into MALDsoft is automated. All three programs produce an output file containing a test statistic for each locus.

The affected-only test statistic used in MALDsoft is defined as the observed minus expected locus ancestry summed across all individuals where the expected locus ancestry is given by the estimated individual admixture. If controls are available, the user has the option of running a case-control test in which it is tested whether the observed minus expected locus ancestry differs between cases and controls at each locus. The standard deviation required for the computation of the test statistics is estimated via parametric bootstrap, based on a modified Baum–Welch algorithm. Under the null hypothesis of no association, each test has a normal distribution. In order to control the experiment-wise significance level, MALDsoft produces the empirical distribution of the most extreme tests.

ADMIXMAP tests for linkage using score tests. The affected-only test is similar to that used in MALDsoft; the test statistic is the same, however, applying a score test, which assumes normality, the variance of the test statistic is minus the second derivative of the log-likelihood. The test for linkage in a case-control study design is implemented by testing for no affect of locus ancestry in a logistic regression model conditional on individual admixture. This setup allows the test to be conditioned on other covariates. A test for linkage to a quantitative trait can be calculated as a score test in a linear
regression model. ADMIXMAP can also test for residual LD not accounted for by the admixture model. No correction for multiple testing is provided.

In an affected-only analysis ANCESTRYMAP tests for linkage at each locus using a likelihood ratio test. The likelihood ratio test is calculated for the null hypothesis of no marker ancestry effect versus alternative hypotheses specified by the user. The hypotheses are specified in terms of the risk ratio associated with the number of alleles from the high risk population. The user can specify one or more alternative hypotheses. The case-control test is fundamentally the same as that used in MALDsoft, but a $t$-statistic is calculated for a difference in means between cases and controls. A Bayesian genome-wide test for significance is provided.

**Key Points**
- This article reviews the main features of all existing software packages for gene mapping by admixture linkage disequilibrium.
- It provides an overview of the underlying statistical methods and testing procedures implemented by each package.
- It provides a concise introduction to the principle of admixture mapping.

**References**