ADMIXMAP - a program to model admixture using marker genotype data

USER MANUAL

Introduction

ADMIXMAP is a general-purpose program for modelling admixture, using marker genotypes and trait data on a sample of individuals from an admixed population (such as African-Americans), where the markers have been chosen to have extreme differentials in allele frequencies between two or more of the ancestral populations between which admixture has occurred. The main difference between ADMIXMAP and classical programs for estimation of admixture such as ADMIX is that ADMIXMAP is based on a multilevel model for the distribution of individual admixture in the population and the stochastic variation of ancestry on hybrid chromosomes. This makes it possible to model the associations of ancestry between linked marker loci, and the association of a trait with individual admixture or with ancestry at a linked marker locus.

Possible uses of the ADMIXMAP program

1. Modelling the distribution of individual admixture values and the history of admixture (inferred by modelling the stochastic variation of ancestry along chromosomes).
2. Case-control, cross-sectional or cohort studies that test for a relationship between disease risk and individual admixture
3. Localizing genes underlying ethnic differences in disease risk by admixture mapping
4. Controlling for population structure (variation in individual admixture) in genetic association studies so as to eliminate associations with unlinked genes
5. Reconstructing the genetic structure of an ancestral population where unadmixed modern descendants are not available for study

ADMIXMAP can model admixture between more than two populations, and can use data from multi-allelic or biallelic marker polymorphisms. The program has been developed for application to admixed human populations, but can also be used to model admixture in livestock or for fine mapping of quantitative trait loci in outbred stocks of mice.

A manual for the program is available which describes the statistical model in more detail. Downloads of the program compiled for various platforms are also available. We recommend that before trying to run the program, you consult us first about your requirements.

Description of program:

ADMIXMAP is designed to analyse datasets that consist of trait measurements and genotype data on a sample of individuals from an admixed or stratified population. Although the name of the program reflects its origins as a program designed for admixture mapping, it has wider uses, especially in genetic association studies. The study design can be a cross-sectional survey of a quantitative trait or binary outcome, a case-control study or a cohort study. For admixture to be modelled efficiently, at least some of the loci typed should be "ancestry-informative markers": markers chosen to have large allele frequency differentials between the ancestral subpopulations that underwent admixture. The program can deal with any number of ancestral subpopulations and any number of linked marker loci. In its present version, the program handles only data from samples of unrelated individuals.

The program is written in C++, and is freely available with source code under a GPL. Offers to help with development of the program are welcome. The current version runs only on a single processor, and computation time is a serious limitation on large datasets. We are developing a parallel version that will be able to run on a computing cluster, using the MPICH implementation of the MPI message-passing standard.

The program is based on a hybrid of Bayesian and classical approaches. A Bayesian full probability model is specified, assigning vague prior distributions to parameters for the distribution of admixture in the population and the stochastic variation of ancestry along hybrid chromosomes. The posterior distribution of all unobserved variables given the observed genotype and trait data, is generated by Markov chain Monte Carlo simulation. These unobserved variables include the ancestry at each locus and the ancestry-specific allele frequencies at each locus. For a description of the theory underlying this approach, see the following papers:-

McKeigue, P.M., Carpenter, J., Parra, E.J., Shriver, M.D.. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annals of Human Genetics* 2000;**64**: 171-86.

Hoggart, C.J., Parra, E.J., Shriver, M.D., Bonilla, C., Kittles, R.A., Clayton, D.G. and McKeigue, P.M. Control of confounding of genetic associations in stratified populations. *Am J Hum Genet*. 2003; **72**:1492-1504.

Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G. and McKeigue, P.M. Design and analysis of admixture mapping studies.*Am J Hum Genet*. 2004; **74**:965-978.

McKeigue, PM; Prospects for admixture mapping of complex traits*. Am J Hum Genet* 2005; in press

Back to top

Applications of the program:

1. Modelling the dependence of a disease or quantitative trait upon individual admixture
For a binary trait, such as presence of disease, the program fits a logistic regression model of the trait upon individual admixture, mean of admixture proportions of both parents. For a continuous trait, such as skin pigmentation, the program fits a linear regression model of the trait value on individual admixture. Covariates such as age, sex and socioeconomic status can be included in the regression model. The program output includes posterior means and 95% credible intervals for the regression coefficients. Alternatively, the program can be used to test a null hypothesis of no association of disease risk or trait level effect with individual admixture as described below.

2. Controlling for confounding of genetic associations in stratified populations
For more details of this application, see Hoggart et al (2003). The program calculates a score test for association of the disease or trait with alleles or haplotypes at each locus, adjusting for individual admixture and other covariates in a regression model. Where there is evidence for association of a trait with individual admixture, the posterior distribution of the regression coefficient can be estimated in a further analysis. For this application, the dataset should include at least 30 markers informative for ancestry.

3. Admixture mapping: localizing genes that underlie ethnic differences in disease risk.
For more details of this application, see Hoggart et al (2004). Where differences in disease risk have a genetic basis, testing for association of the disease with locus ancestry by conditioning on parental admixture can localize genes underlying these differences. This approach is an extension of the principles underlying linkage analysis of an experimental cross. To exploit the full power of admixture mapping, 1000 or more markers informative for ancestry across the genome are required.

4. Detecting population stratification and identifying admixed individuals
Where no information about the demographic background of the population under study is available, ADMIXMAP can be used to test for population stratification, to determine how many subpopulations are required to model this stratification, and to identify admixed individuals. This is useful when assembling panels of unadmixed individuals to be used for estimating allele frequencies. We emphasize that when the program is run without supplying prior information about allele frequencies in each subpopulation, the subpopulations are not identifiable in the model. Thus inference should be based only on the posterior distribution of variables that are unaffected by permuting the labels of the subpopulations.

5. Testing for associations of a trait with haplotypes and estimating haplotype frequencies from a sample of unrelated individuals
Where two or more loci in the same gene have been typed, ADMIXMAP will model the unobserved haplotypes, conditional on the observed unordered genotypes. Score tests for association of haplotypes with the trait can be obtained, and samples from the posterior distribution of haplotype frequencies can be obtained. This application of the program is not limited to admixed or stratified populations: for a population that is not stratified, the user can simply specify the option *populations*=1

For each of these applications, score tests of the appropriate null hypotheses are built into the program.

Modelling admixture and trait values:

Any run of two or more loci for which the distance between loci is specified as zero is modelled as a single "compound locus". Thus if L "simple loci" (SNPs, insertion/deletion polymorphisms or microsatellites) have been typed, and three of these simple loci are in the same gene, the model will have L - 2 compound loci. The program assumes that on any gamete, the ancestry state is the same at all loci within a compound locus. The program allows for allelic association within any compound locus that contains two or more simple loci, and models the unobserved haplotypes at this compound locus.

For each parent of each individual, admixture proportions are defined by a vector with *K* co-ordinates, where *K* is the number of ancestral subpopulations that contributed to the admixed population under study. For instance, in a Caribbean population it may be possible to model the gene pool of the admixed population as a mixture of three subpopulations: African, European and Native American. The model of admixture is described by the following hierarchy:-

1. The population distribution from which the parental admixture proportions are drawn is modelled by a Dirichlet distribution with parameter vector of length *K*.

2. The allele or haplotype frequencies at each compound locus are modelled by a Dirichlet distribution, with prior parameters specified by the user.

3. Locus ancestry is modelled by a multinomial distribution, with cell probabilities specified by the admixture of both parents.

4. The probabilities of observing each allele or haplotype at each locus on each gamete, given the ancestry of the gamete at that locus, are modelled by a multinomial distribution, with parameters given by the ancestry-specific allele (or haplotype) frequencies.

5. The stochastic variation of ancestral states along the chromosomes transmitted from each parent is modelled as a mixture of independent Poisson

arrival processes with intensities α, β, γ per Morgan (for three-way admixture). For given values of parental admixture, it is only necessary to specify a single parameter ρ for the sum of intensities: $\rho = \alpha + \beta + \gamma$.

If an outcome variable is supplied, ADMIXMAP fits a regression model (logistic regression for a binary trait, linear regression for a quantitative trait) with individual admixture proportions and any covariates supplied by the user as explanatory variables.

Modelling allele/haplotype frequencies:

The program can be run with ancestry-specific allele / haplotype frequencies specified either as fixed or as random variables. If one of the two options *allelefreqfile* or *fixedallelefreqs* is specified, the allele frequencies are specified as fixed at the values supplied. If option *populations* is specified, the allele frequencies are specified as random variables with reference (uninformative) prior distributions. If option *priorallelefreqsfile* is specified, the allele frequencies are specified as random variables with a prior distribution given by the values in this file. This option is used where allele frequencies have been estimated from samples of unadmixed modern descendants of the ancestral subpopulations that contributed to the admixed population under study. For instance, in a study of a population of mixed European and west African ancestry, allele frequencies at some or all of the loci typed may have been estimated in samples from modern unadmixed west African and European populations. The program will use this information to estimate the ancestry-specific allele frequencies from the unadmixed and admixed population samples simultaneously, allowing for sampling error.

If no information about allele frequencies in the ancestral subpopulations is provided, the ancestry-specific allele frequencies are estimated only from the admixed population under study. If no information about allele frequencies is provided at any locus, the subpopulations are not identifiable in the model. This does not matter when the program is used only to control for confounding by hidden population stratification, as described in Hoggart et al. (2003).

The file *priorallelefreqfile* specifies the parameters of a Dirichlet prior distribution for the allele frequencies at each locus in each subpopulation. Where the alleles or haplotypes have been counted directly in samples from unadmixed modern descendants, these parameter values should be specified by adding 0.5 to the observed counts of each allele or haplotype in each subpopulation. These parameter values specify the Dirichlet posterior distribution that we would obtain by combining a reference prior with the observed counts. Using this as a prior distribution when analysing data from the admixed population is equivalent to estimating the allele frequencies simultaneously from the admixed and unadmixed population samples, with a reference prior.

For compound loci, where haplotype frequencies have been estimated from unordered genotypes rather than by counting phase-known gametes, the user cannot specify the prior distribution simply by adding 0.5 to the observed counts of each haplotype in a sample of unadmixed modern descendants, because the counts have not been observed. Instead, we can compute the posterior distribution of haplotype frequencies in the unadmixed population, given a reference prior and the observed unordered genotype data. In accordance with the principles of Bayesian inference, we can then use this posterior distribution to specify the prior distribution when modelling data from the admixed population. To simplify the computation, we generate a large sample from the posterior distribution of haplotype frequencies in the unadmixed population and calculate the parameters of the Dirichlet distribution that most closely approximates this posterior distribution. The parameters of this Dirichlet distribution are then entered in file *priorallelefreqfile*.

This can be implemented by running ADMIXMAP with genotype data from each unadmixed population sample, specifying options *populations*=1 and *allelefreqoutputfile* to sample the posterior distribution of the haplotype frequencies. The parameters of the Dirichlet distribution that most closely approximates this posterior distribution can then be computed, and substituted into the file *priorallelefreq* as input to the program when modelling data from the admixed population. This is straightforward: for each locus and each subpopulation, the posterior expectations and the posterior covariance matrix of the allele frequencies are evaluated using the samples from the posterior distribution in *allelefreqoutputfile*. The Dirichlet parameters α_i that approximate the posterior distribution are then computed by equating the posterior expectations of the allele frequencies to the ratios $\alpha_i / \Sigma\alpha_i$, and equating the determinant of the posterior covariance matrix to the determinant of the covariance matrix of the Dirichlet distribution. If ADMIXMAP is run with the option *allelefreqoutputfile* and the R script *AdmixmapOutput.R* is run to process the output files, the Dirichlet parameters will be computed and written to a file in the correct format for use in subsequent analyses as *priorallelefreqfile*.

With options *populations*, *allelefreqfile* or *priorallelefreqfile*, the program fits a model in which the allele frequencies in modern unadmixed descendants of the ancestral subpopulations are identical to the corresponding ancestry-specific allele frequencies in the admixed population under study. The option *dispersiontestfile* will generate a diagnostic test of this assumption.

With option *historicallelefreqfile*, the program fits a more general model in which there is dispersion of allele frequencies between the unadmixed and admixed populations.

With option *correlatedallelefreqs* a correlated allele frequency model is fitted, in which the allele frequency prior parameters are the same across subpopulations and specified as vectors of proportions and a sum, common to all loci.

Inference:

There are various approaches to statistical inference and hypothesis testing using the ADMIXMAP program:-

(1) A model based on the null hypothesis can be fitted, and this null hypothesis can be tested against alternatives with a score test computed by averaging over the posterior distribution of the missing data. For a description of the theory underlying this approach, see Hoggart et al. (2003). Several score tests are built into the program, and are described below. Additional score tests can be constructed by the user.

(2) The effect under study can be included in the regression model, so that the posterior distribution of the effect parameter is estimated. In large samples, the posterior mean and 95% credible interval for the effect parameter are asymptotically equivalent to the maximum likelihood estimate and 95% confidence interval that would be obtained by classical methods.

(3) The log-likelihood function for the effect of a parameter can be computed: not yet implemented

(4) The marginal likelihood of the model can be evaluated using Chib's algorithm or thermodynamic integration. Chib's algoritm is implemented only for a single individual (the first listed in the genotypes file) and for a model with no outcome variable but thermodynamic integration is implemented for all models.

In addition to these methods for formal inference, model diagnostics, based on the posterior predictive check probability, are provided to detect population stratification not accounted for in the model, lack of fit of the allele frequencies to the specified model and for departure from Hardy-Weinberg equilibrium.

Back to top

Comparison with other programs for modelling admixture:

The program STRUCTURE (available from http://pritch.bsd.uchicago.edu/) fits a similar hierarchical model for population admixture, given genotype data on admixed and unadmixed individuals, if you specify the "popalphas" option (see documentation for this program at http://pritch.bsd.uchicago.edu/software/readme_structure2.pdf).

The main differences between ADMIXMAP and STRUCTURE are:-

1. STRUCTURE does not model the dependence of the outcome variable on individual admixture and thus cannot adjust for the effect of individual admixture on the outcome variable.
2. STRUCTURE does not allow the user to supply prior distributions for the allele frequencies. To use allele frequency data from unadmixed individuals in STRUCTURE, individuals sampled from unadmixed and admixed populations have to be included in the same model. This is not recommended, either with STRUCTURE or ADMIXMAP, because the model assumes a unimodal distribution of individual admixture values in the population. If samples from both unadmixed and admixed populations are included in the same model, the distribution of individual admixture values will generally not be unimodal, and the fit of the model will be poor.
3. STRUCTURE does not allow for allelic association (other than that generated by admixture), and is therefore unsuitable for analysis of datasets in which two or more tightly-linkedloci (for instance SNPs in the same gene) have been typed. ADMIXMAP allows for allelic association (if the distance between loci is coded as zero) and models the unobserved haplotypes.
4. STRUCTURE assumes that the markers have not been selected to be highly informative for ancestry, and the authors recommend that a model which allows for correlation between the allele frequencies in each subpopulation is specified. ADMIXMAP does not in general assume any correlation between the allele frequencies in each subpopulation.

The program ANCESTRYMAP (available from http://genepath.med.harvard.edu/~reich/) is similar to ADMIXMAP but is restricted to diallelic simple loci and two populations.

Tools for converting data from ANCESTRYMAP format to ADMIXMAP format and vice-versa as well as from STRUCTURE format to ADMIXMAP format are available here (http://www.ucd.ie/genepi/admixmap/tools.html).

Installing and running the program

ADMIXMAP runs as a console application. We are no longer supporting the Windows graphical front-end. Binaries (compiled executable files) are provided for Linux and Windows via SourceForge.

The source code is available as a tarball from the same page. Compilation instructions are provided. The source code is also available from our Subversion repository. Instructions for accessing the repository are here.

If you have downloaded a binary, or compiled the program yourself, and placed the ADMIXMAP executable in a directory that lies on your search path, the program can be invoked directly by typing

```
admixmap <optionsfile>
```

where *optionsfile* contains statements of the form *optionname=value*, one line per optionname. A sample optionsfile, *testArguments.txt* is provided with the test files.

To generate summary statistics, tables and graphics from the output files generated by ADMIXMAP you will need to run the script *AdmixmapOutput.R*. This requires the R statistical package to be installed.

If your output files are in a directory other than 'results', you will need to supply R with the directory name. If running R interactively (Rterm, Rgui), use the command

`Sys.putenv("RESULTSDIR" = <resultsdirname>)`

If you are running the script from the command line, first set the RESULTSDIR environment
variable.
To do this in Linux use: "`export RESULTSDIR=resultsdirname`"
and in Windows use: "`set RESULTSDIR=resultsdirname`"

Then run the R script using:
`R --quiet --no-save --no-restore <AdmixmapOutput.R >resultsdirname/Rlog.txt.`
The most convenient way to run the entire analysis is to use a Perl script. If you are working in Windows, you should install ActivePerl if not already installed. The Perl script specifies the user options, writes the to file, runs ADMIXMAP, and invokes R in batch mode to run the script *AdmixmapOutput.R* to process the output files. See, for example, the Perl script *admixmap.pl* provided with the test files. User options can be changed by editing this file. This requires only minimal knowledge of Perl. The Perl script is especially convenient when running a batch of different analyses, as each run may require only a few options to be changed.

The Perl script admixmap.pl should be invoked from a directory that has a sub-directory named "data" containing the data files. Results will be output to the directory specified in the *resultsdir* option or to one named "results" if none is specified.

User options

The program requires a list of options to be specified by the user either as command-line arguments, or in a text file the name of which is given as a single argument to the program . As explained above, the most convenient way to specify these arguments is to use a Perl script (see "admixmap.pl"). A list of these options is given in the following table. Required arguments are in **bold**.

General Options

samples	Integer specifying total number of iterations of the Markov chain, including burn-in. With strong priors and informative markers, a run of about 500 should suffice for inference. Otherwise, a run of at least 20 000 iterations may be necessary. See here for how to determine if the run has been long enough.
burnin	Integer specifying number of iterations for burn-in of the Markov chain, before posterior samples are output. A burn-in of at least 50 iterations is recommended for inference. For analyses requiring long runs, a burn-in of up to 500 may be required.
every	Integer specifying the "thinning" of samples from the posterior distribution that are written to the output files, after the burn-in period. For example, if *every*=10, sampled values are written to the output files every 10 iterations. We recommend using a value of 5 to keep down the size of the output files. Sampling more frequently than this does not much improve the precision of results, because successive draws are not independent. Thinning the output samples does not affect the calculation of ergodic averages or test statistics, which are based on all sampled values. Note that *every* must be no greater than (samples - burnin) / 10 or some output files may be empy.
numannealedruns	If *thermo*=0, this specifies the number of "annealing" runs during burnin. This usually improves mixing. If *thermo*=1, this specifies the number of "temperatures" at which to run in order to estimate the marginal likelihood by thermodynamic integration. Default is 20.

displaylevel	0 - silent mode; Only start and finish times output to screen. 1 - quiet mode; Model specification, priors, test results and diagnostics written to screen. 2 - normal mode; more verbose information and an iteration counter output to screen. >2 - monitor mode; population-level parameters also written to screen with frequency specified by *every.*
resultsdir	Path of directory for output files. Default is 'results'.
logfile	Name of log file written by the program. Default is 'logfile.txt',
seed	can be used to specify a seed for the random number generator.

Allele / Haplotype Frequency Model

The program requires **one** of the following four options, any one of which specifies the number of subpopulations in the model: populations, allelefreqfile, priorallelefreqfile, or historicallelefreqfile. These options are mutually exclusive.

populations	Integer specifying number of subpopulations that have contributed to the admixed population under study. If specified as 1, the program fits a model based on a single homogeneous population. This option is not required (and is ignored) if information about allele frequencies is supplied in *allelefreqfile*, *priorallelefreqfile*, or *historicallelefreqfile*, as the number of columns in any of these files defines the number of subpopulations in the model. If none of these files are specified, the parameters of the Dirichlet priors for allele or haplotype frequencies default to 1/n, where n is the number of alleles or haplotypes at each compound locus.

allelefreqfile	Pathname of file containing the allele frequencies of the genotyped loci for each subpopulation. When this option is specified, the model treats the allele frequencies as fixed constants. This option is obsolete, and retained only for backward compatibility. Instead, use option priorallelefreqfile to specify the allele frequencies, and specify option fixedallelefreqs=1. This allows you to use the same format for the allele frequency file, whether the allele frequencies are fixed, have a prior distribution with no dispersion, or are specified with a dispersion model.
priorallelefreqfile	Pathname of file containing parameters of the Dirichlet prior distributions for allele frequencies (or haplotype frequencies) at each compound locus in each subpopulation. Where allele frequencies have been estimated from a sample of unadmixed individuals, the prior distribution parameters for the corresponding subpopulation should be specified as the observed allele counts plus 0.5. Where no allele frequency data are available, specify the prior parameters as 0.5 for each allele ("reference" prior). When this option is specified, the program fits a model in which the allele frequencies in each subpopulation are estimated simultaneously from the unadmixed samples and the admixed sample under study
historicallelefreqfile	Pathname of file containing observed allele counts at the genotyped loci from samples of unadmixed individuals in each subpopulation. When this option is specified, the program fits a model that allows the "historic" allele frequencies in the unadmixed population to vary from the corresponding ancestry-specific allele frequencies in the admixed population under study

Data Files

Details of file formats are under Input files

locusfile	path to file containing information about each locus typed
genotypesfile	path to file containing genotypes for each individual typed
outcomevarfile	path to file containing values of outcome variables
coxoutcomevarfile	path to file containing data for a Cox regression
covariatesfile	path to file containing covariates for a regression model
targetindicator	Integer specifying column in *outcomevarfile* that contains the first outcome variable to be modelled. This column number should be specified as an offset from column 1: thus to select the variable in column 1, specify *targetindicator*=0. The default is 0.
outcomes	valid only with *outcomevarfile.* Integer specifying the number of columns of the outcomevarfile to use, starting with *targetindicator*.
reportedancestry	not fully tested or documented: allows prior information about each individual's ancestry to be specified in the model
testgenotypesfile	specifies genotypes for offline score tests at loci that have not been included in the model.

Model Specification

indadmixhiermodel	0 - Model for a collection of individuals in which the admixture proportions of each individual's parents, and the sum of intensities on each parental gamete, are statistically independent given the priors on these parameters. This option is useful in two situations: (1) when you already have strong prior information about the distribution of admixture in the population from which the individuals have been sampled, and want to specify a Dirichlet prior for each individual's parental admixture proportions using the option initalpha0; or (2) when you want to calculate the marginal likelihood of the model given the genotype data on each individual. 1- Hierarchical model on individual admixture The default is 1.
randommatingmodel	0 - assortative mating model (admixture proportions the same in both parents) 1 - random mating model The default is 0.
globalrho	0 - the sum of intensities parameter ρ is allowed to vary between individuals, or between gametes if a random mating model is specified). This specifies a hierarchical model, with a gamma distribution for the variation of ρ between individuals specified as below. 1 - the sum of intensities ρ is modelled as a global parameter, set to be the same on all parental gametes The default is 1

fixedallelefreqs	1 specifies that *priorallelefreqfile* contains fixed allele frequencies 0 otherwise default is 0
correlatedallelefreqs	valid only with '*populations'* or '*priorallelefreqfile'* options 1 specifies a correlated allele frequency model 0 otherwise default is 0

Prior Specification

sumintensitiesprior globalsumintensitiesprior	In a model with global sumintensities or without a hierarchical model of individual admixture, the sum of intensities parameter has a Gamma(a, b) prior specified as "*globalsumintensitiesprior="a,b"* ". Default values for a and b are 3 and 0.5, giving a prior mean of 6 and prior variance of 12. Otherwise (indadmixhiermodel=1 and globalrho=0), the sum of intensities parameter ρ has a Gamma(a,b) prior distribution and the scale parameter b has a beta hyperprior with parameters β_0 and β_1. This specifies a "GammaGamma" prior, which has mean E()ρ = βα$_1$ β(/$_0$)1 – and variance E(()ρE()1+)ρ (/β$_0$.)2– The three parameters of this prior are specified with sumintensitiesprior. The three values must be enclosed by quotes and separated by commas e.g "*sumintensitiesprior="*2,3,4" ". Thus, for instance, to model an African-American population, for which we have prior

	parameter is about 6 per morgan, we could specify sumintensitiesprior = "6,40,39" This specifies the prior for the sum of intensities parameter ρ as Gamma(6, 1) which has mean 6 and variance 1. "0,1,0" specifies a flat prior on log ρ "1,1,0" specifies a flat prior on ρ The default, if this option is not specified, is "4,3,3" Where there is not enough data for reliable inference of the sum of intensities parameter, it is often useful to specify that the prior distribution should be truncated at some upper limit of plausible values, using the option *truncationpoint*.
etapriormean, etapriorvar	Specify the prior mean and variance of the dispersion parameter(s), ,η in a dispersion or correlated allele frequency model.
etapriorfile	File containing parameters of the gamma prior distribution specified for the allele frequency dispersion parameter η in each subpopulation. This option can be used only when a dispersion model has been specified with the option *historicallelefreqfile*. This is useful when there are not enough data for the dispersion parameter to be inferred from the data, and we want to use prior information from population genetics. This file has one row for each subpopulation (in the same order as the order of subpopulations by columns in historicallelefreqfile, and two columns specifying the shape and location parameters of the gamma distribution. Thus, for a sample from an African-American population, in which *historicallelefreqfile* contains counts of alleles in samples of modern west Africans (in the first column) and Europeans (in the second

	column), we might specify an etaprior file containing these two lines:- 50 1 500 1 This specifies a prior with mean 50 for the parameter for dispersion of allele frequencies between modern unadmixed west Africans and the African gene pool in African-Americans, and a prior with mean 500 and variance 500 for the parameter for dispersion of allele frequencies between modern unadmixed Europeans and the European gene pool in African-Americans. The dispersion parameter is related to the fixation index F_{ST} by $\xi = (1 + F_{ST}) / F_{ST}$, so values of 50 and 500 for ξ correspond roughly to values of 0.02 and 0.002 for F_{ST}.
admixtureprior, admixtureprior1	When *indadmixhiermodel* = 0, each of these two options can be used to specify a Dirichlet parameter vector for parental admixture proportions. The parameter vector is specified as a string of numbers separated by commas. For instance, with a model based on 3 subpopulations:- admixtureprior = “2, 8, 3.5” would specify the prior for parental admixture proportions (or the maternal gamete if option randommatingmodel=1 has been specified) with parameter vector c(2, 8, 3.5). admixtureprior1 can be used similarly to specify the prior for paternal admixture proportions if optionrandommatingmodel=1 has been specified. For example, "*admixtureprior = 1,1,0*" and "*admixtureprior1 = 1,1,1*" would specify that one parent has 2-way admixture (between subpopulations 1 and 2) and the other has 3

	way admixture between subpopulations . If *indadmixhiermodel* =1, admixtureprior can be used to specify initial values for the population admixture Dirichlet parameters.
regressionpriorprecision	Prior precision (1 / variance) of regression parameters
popadmixproportionsequal	Specifies that the population-level admixture proportions are to be kept equal

Output Files

Pathnames of output files, details of file formats in Output files.

paramfile	Population-level admixture and sum-of-intensities
regparamfile	Regression parameters
dispparamfile	Allele/haplotype frequency dispersion in historicallelefreqs model
indadmixturefile	Individual-level admixture proportions and sum-of-intensities
allelefreqoutputfile	Name of output file containing samples from the posterior distribution of ancestry-specific allele frequencies. Valid only when the allele frequencies are specified as random variables, i.e. when one of the two options priorallelefreqfile or historicallelefreqfile is specified and fixedallelefreqs is 0.
ergodicaveragefile	Ergodic averages of population-level parameters and of the mean and variance of the deviance.

Tests and Diagnostics

The options below specify additional tests or output,but do not change the model itself

chib	1 - Calculate marginal likelihood for the first individual using Chib algorithm. 0 - default
thermo	1 - Use thermodynamic integration to compute marginal likelihood. 0 - default
testoneindiv	1 - compute marginal likelihood for the first individual listed in the genotypes file. This individual will not be included as part of the sample and should not be included in an outcomevarfile or covariatesfile. 0 - default
indadmixmodefile	Name of output file containing posterior estimates of the modes of individual admixture proportions and individual-level sumintensities (if globalrho=0).

admixturescorefile	Pathname of file to which results of a score test for the association of the trait with individual admixture will be written. This option is valid only if an outcome variable has been specified. This option is used only to obtain a formal test of the null hypothesis of no association between the trait and individual admixture. If *admixturescorefile* is specified, the regression model will not include individual admixture proportions as explanatory variables, and tests for allelic association or linkage will not be adjusted for the effect of individual admixture. Provided an *outcomevarfile* is specified and unless option *admixturescorefile* is specified the program will fit a regression model with the outcome variable as dependent variable and individual admixture proportions (plus any covariates specified in inputfile) as explanatory variables.
allelicassociationscorefile	Name of output file containing score tests for association of the outcome variable with alleles at each simple locus, adjusting for individual admixture.
residualellelicassocscorefile	Name of output file containing score tests for residual allelic association between pairs of unlinked loci.
haplotypeassociationscorefile	Name of output file containing score tests for association of the outcome variable with haplotypes for all compound loci containing haplotypes, adjusting for individual admixture.
ancestryassociationscorefile	Name of output file containing score tests at each compound locus for linkage with genes underlying ethnic variation in the trait. This is a test for association of the trait with locus ancestry, adjusting for individual admixture and covariates. This test should be used in a cross-sectional or cohort study design. For a case-control study of a rare disease, the

	affected-only test below has greater statistical power.
affectedonlyscorefile	Name of output file containing score tests at each compound locus for linkage with ancestry, based on comparing the observed and expected proportions of gene copies at this locus that have ancestry from each subpopulation. This test is calculated from affected individuals only: individuals are their own controls. Even when the sample includes both cases and controls, this test is more powerful than the regression model score test in *ancestryassociationscorefile* if the disease is rare.
likratiofile	Name of output file containing likelihood ratios for the affecteds-only score test at values of 0.5 and 2 for the ancestry risk ratio.
allelefreqscorefile	Name of output file containing score tests of mis-specified ancestry specific allele frequencies. This option is valid only when the allele frequencies are fixed, i.e. when option *allelefreqfile* is specified or *fixedallelefreqs* is 1. There is a test for each population at each locus as well as a summary chi-squared test across populations.
hwscoretestfile	Name of outputfile containing score tests for heterozygosity across loci, as a test for departure from Hardy-Weinberg equilibrium. These can be used to detect genotyping errors.

stratificationtestfile	Name of output file containing test for residual population stratification (stratification not accounted for by the fitted model).
dispersiontestfile	Name of output file containing test for dispersion of allele frequencies between the unadmixed populations sampled and the corresponding ancestry-specific allele frequencies in the admixed population under study. This is evaluated for each subpopulation at each locus, and as a global test over all loci. This option is valid only if option *priorallelefreqfile* is specified. The results are "Bayesian p-values", as above.
fstoutputfile	This option is used only with option historicallelefreqfile (which specifies a dispersion model for allele frequencies). Under a dispersion model, the allele frequencies in unadmixed modern descendants are allowed to vary from the corresponding ancestry-specific allele frequencies in the admixed population. The variance of allele frequencies at a locus can be measured by Wright's "fixation index subpopulation-total" (F_{st}). In Wright's terminology, the unadmixed modern descendants and the pool of genes of corresponding ancestry in the admixed population are "subpopulations", and the "historic" population from which both these gene pools are derived is the "total" population. This differs from the terminology used in this manual, in which K "subpopulations" are specified in the model as ancestors of the admixed population. For each locus, and each subpopulation, specifying the option fstoutputfile will make the program output the ergodic average of the F_{st} value. These values can be examined as a diagnostic: a locus with an unusually large F_{st} value may indicate errors in coding, errors in typing, or possibly that allele frequencies in unadmixed modern descendants have diverged from the corresponding allele frequencies in the admixed population as a result of recent selection pressure.

Input File Formats

genotypesfile

The first row of the file is a header row listing locus names, enclosed in quotes and separated by spaces. Locus names should be exactly the same as in the file locusfile. Loci must be ordered by their map positions on the genome. Each subsequent row contains genotype data for a single individual. Each line contains the individual ID, the individual's sex, coded as 1 for male, 2 for female, 0 for missing, followed by observed genotypes at each locus, optionally enclosed in quotes. The sex column may be omitted if none of the loci are on the X chromosome. Haploid genotypes (including X chromosome genotypes for males) are coded as single integers. Diploid genotypes are coded as pairs of integers separated by a comma. Where there are *a* alleles at a locus, the alleles should be coded as numbers from 1 to *a*. Missing genotypes are coded as "0,0" (or "0" for haploid genotypes).

For compatibility with existing datasets, we plan to change this file format to one more similar to the PEDFILE format used with LINKAGE.

testgenotypesfile

This file contains genotypes for each individual in the genotypesfile at diallelic loci not included in the model due to large haplotypes not being modelled. The format is the as for the genotypesfile above except that genotypes should be coded as 0 for "1,1", 1 for "1,2" and 2 for "2,2". Missing genotypes should be coded as NA. The file is not used by the program itself but will indicate that, provided there is a regression model, "offline" score tests are to be carried out in the R script.

locusfile

File contains information about each simple locus: that is, each locus that is typed. The first row of the file is ignored by the program, and can be used as a header. Each subsequent row contains values of four variables: locus name; number of alleles at this locus; genetic map distance in Morgans, centimorgans or megabases between this locus; and the previous locus and the name of the chromosome where the locus is located. The last column is optional if none of the loci lie on the X chromosome. If distances are supplied in centimorgans, the header of the distance column must

contain "cm" or cM". If the distances are supplied in megabases, the header should contain "mb" or "Mb". Loci must be ordered by their map positions on the genome. Locus names should contain only alphanumeric characters (no spaces, dots or hyphens). If the previous locus is unlinked, the genetic map distance should be coded as "NA", "#" or "." . Loci considered too far apart to be linked may also be treated as unlinked. For two or more loci that are so close together that they should be analysed as a single compound locus (as with DRD2Bcl and DRD2Taqd in the tutorial), map distance should be coded as 0.

The website http://actin.ucd.ie/cgi-bin/rs2cm.cgi can be used to obtain the genetic map positions (in cM)
of a list of SNPs, which, once converted to distances, may be specified in the locusfile.

allelefreqfile

This file contains the ancestry-specific allele frequencies at each compound locus in each ancestral subpopulation. The first row contains headers in quotes, separated by spaces. The first string in this row is ignored. Subsequent strings in the first row specify the names of the ancestral subpopulations contributing to the admixed population under study. Subsequent rows specify the ancestry-specific allele frequencies (usually estimated by from sampling modern descendants of the subpopulations that underwent admixture. The first column in each row gives the name of the compound locus, in quotes.

For diallelic loci, only the frequency of allele 1 in each population is specified. For each locus with a alleles, there are (a - 1) rows specifying frequencies of alleles 1 to (a - 1).

Where two or more loci are to be analysed as a single haplotype, the ancestry-specific frequency of each haplotype must be specified. Thus in the example files below, there are two SNPs in the DRD2 gene, giving four possible haplotypes) and four lines specifying the ancestry-specific frequencies of haplotypes 11, 12, 21, 22. The loci in the haplotype are ordered by their map position on the genome, and the haplotypes are ordered by incrementing a counter from right to left. For instance if there were three loci A, B, C, with 4, 2 and 3 alleles respectively, the haplotypes would be listed in the following order: 111, 112, 113, 121, 122, 123, 211, ..., 422, 423.

Note: Use of this file is not recommended and is supported only for backward

compatibility with previous versions. Instead you can specify the allele frequencies in priorallelefreqfile or historicallelefreqfile as fixed, with option fixedallelefreqs = 1

priorallelefreqfile

This file contains parameter values for the Dirichlet prior distribution of the allele or haplotype frequencies at each compound locus in each subpopulation. At each compound locus with k alleles or k possible haplotypes, a Dirichlet prior distribution is specified by a vector of k positive numbers. Where these alleles or haplotypes have been counted directly in samples from an unadmixed subpopulation, the parameter values should be specified as 0.5 plus the observed counts of each allele. Where no information is available about allele or haplotype frequencies at a compound locus in a subpopulation, or no copies of the allele have been observed in the sample from that subpopulation, specify 0.5 in the corresponding cells. Specifying 0.5 in all cells, with columns for *b* subpopulations, is equivalent to specifying the option *populations = b* .

Where haplotype frequencies at a compound locus have been estimated from unordered genotypes, the user should supply the parameters of the Dirichlet distribution that most closely approximates the posterior distribution of haplotype frequencies given the observed genotypes and a reference prior, as described above.

The first row is a header row, consisting of strings in quotes, separated by spaces. The first string in this row is ignored, and the subsequent strings specify the names of the ancestral subpopulations contributing to the admixed population).

After the header row, there is one row for each allele (or haplotype) at each compound locus. The first column in each row gives the name of the compound locus in quotes. Subsequent columns give the prior parameters for the frequency of the allele (or haplotype) in each subpopulation, separated by a single space.

If the compound locus consists of two or more simple loci, (see notes above), the rows list prior parameters for the haplotypes in the order defined by incrementing a counter from right to left. For instance if there were three loci A, B, C, with 4, 2 and 3 alleles respectively, the haplotypes would be listed in the following order: 1-1-1, 1-1-2, 1-1-3, 1-2-1, 1-2-2, 1-2-3, 2-1-1, ..., 4-2-2, 4-2-3. Estimated counts should be given for all possible haplotypes, however rare. The program will include all possible haplotypes in the model, but will omit rare haplotypes when constructing test statistics.

historicallelefreqfile

This file contains observed counts of alleles or haplotypes at each compound locus in samples from unadmixed subpopulations. The format of this file is exactly the same as the format of priorallelefreqfile described above. The only difference between the two files is that in historicallelefreqfile, 0.5 is not added to the observed counts.

outcomevarfile

This file contains values of one or more outcome variables. After the header row, the file has one row per individual. Binary variables should be coded as 1 = affected, 0 = unaffected. Missing values are coded as #. The header row contains the variable labels in quotes separated by spaces. If the file contains more than one outcome variable, the column containing the first variable of interest should be specified by the command-line option targetindicator. The number of columns to be used (1 or 2) can be specified with the *outcomes* option.

covariatesfile

This file contains values of covariates to be included in the regression model. It is used only if an *outcomevarfile* has been specified, and is optional even then. The header row contains covariate names in quotes, separated by spaces. Subsequent rows contain the observed values of these variables. For computational reasons, the values of the covariates should be centred about their sample means. Missing values are coded as #.

coxoutcomevarfile

This file contains surivival data for a Cox regression model. After the header row, the file has one row per individual and there are three columns. The first contains the times when each individual began to be observed; the second contains the times the individuals ceased being observed and the last column contains the number of events that occurred during the obseerved period (usually 0 or 1). The start and finish times must be numeric and relative to the same point in time, (usually the first start time).

Output files:

Output files are formatted as either tab-delimited tables with a header line (for 2-way arrays) or as R objects (for 3- or 4-way arrays). Output files are written to the directory specified by *resultsdir.*

paramfile - Posterior draws of the following at intervals determined by option every: -

1. Parameters of the Dirichlet distribution for parental admixture: one for each subpopulation
2. Sum of intensities for the stochastic process of transitions of ancestry on hybrid chromosomes

regparamfile - Posterior draws of intercept, slope and precision (the inverse of the residual variance) parameters in the regression model, at intervals determined by option every.

dispparamfile - Posterior draws of allele frequency dispersion parameters, one for each subpopulation, at intervals determined by option every. These are written only if option *historicallelefreqfile* has been specified or *correlatedallelefreqs* = 1.

Median and 95% credible intervals for these parameters are written to the file PosteriorQuantiles.txt.

indadmixturefile - Posterior draws of individual/gamete level variables, at intervals determined by option every written as an R object. The outputs to this file are, in the following order;

1. gamete admixture proportions, ordered by subpopulations and then by gamete if a random mating model is specified. If an assortative mating model is specified only individual admixture proportions will be output.
2. gamete/individual sum-of-intensities if globalrhoindicator is false
3. predicted value of the outcome variable in the regression model
4. paternal and maternal haplotypes at this locus.

These values are written out for every individual at every iteration This file is formatted to be read into R as a three-way array (indexed by variables, individuals, draws).

allelefreqoutputfile - Posterior draws of the ancestry-specific allele or haplotype frequencies for each state of ancestry at each compound locus, at intervals determined by option every. These results can be used to compute new parameters for the prior distributions specified in priorallelefreqfile which can be used in subsequent studies with independent samples

ergodicaveragefile - Cumulative posterior means over all iterations ("ergodic averages") for the variables in paramfile, regparamfile and dispparamfile, output at intervals of 10 ×every iterations. Monitoring these ergodic averages allows the user to determine whether the sampler has been run long enough for the posterior means to have been estimated accurately.

The output files *admixturescorefile, allelicassociationscorefile, ancestryassociationscorefile, affectedsonlyscorefile* contain results of score tests obtained by averaging over the posterior distribution. Each table of score test results, based on cumulative averages for the score and information over all posterior samples obtained after the burn-in period, is

output at intervals of 10 × every. Monitoring these repeated updates allows the user to determine when the sampler has been run long enough for the test results to be computed accurately. For inference, only the last table, which is output separately and which is based on the entire posterior sample, is used. All these files are formatted to be read into R as a three-way array (indexed by loci, test statistics, output number).

For univariate null hypotheses (testing the effect of one allele, one haplotype, or one subpopulation against all others) the test statistic is the score divided by the square root of the observed information, which has a standard normal distribution under the null hypothesis. The percent of information extracted (the ratio of observed information to complete information) measures the information obtained about the parameter under test, in comparison the information that would be obtained if individual admixture, haplotypes at each locus, and gamete ancestry at each locus were measured without error.

For the affected-only and ancestry association score tests, the missing information can be partitioned into two components: missing information about locus ancestry, and missing information about model parameters(parental admixture) . These components are tabulated separately.

For composite null hypotheses, the score **U** is a vector, the observed information **V** is a matrix, and the test statistic ($\mathbf{UV}^{-1}\mathbf{U}^{\prime}$) has a chi-squared distribution under the null hypothesis.

admixturescorefile - test for association of trait with individual admixture. The null hypothesis is no effect of individual admixture in a regression model, with covariates as explanatory variables if specified. The test statistic is computed for the effect of each subpopulation separately, with a summary chi-square test over all subpopulations if there are more than two subpopulations.

allelicassociationscorefile - tests for allelic association at each locus. The null hypothesis is no effect of the alleles or haplotypes in a regression analysis with individual admixture (and covariates if specified) as explanatory variables. The test statistic is computed for each allele or haplotype separately, with a summary chi-square statistic over all alleles or haplotypes at each locus if there are more than two alleles or haplotypes. Rare alleles or haplotypes are grouped together.
This test is appropriate when testing for association of the trait with alleles or haplotypes in a candidate gene.

ancestryassociationscorefile - tests for linkage of each locus with genes underlying ethnic variation in disease risk or trait values. This is a test for association of the trait with ancestry at each compound locus, conditional on parental admixture. The null hypothesis is no effect of locus ancestry in a regression analysis with individual admixture (and covariates if specified) as explanatory variables. The test statistic is computed for the effect of each subpopulation separately, with a summary chi-square statistic over all subpopulations at each locus if there are more than two subpopulations. The proportion of information extracted depends upon the information content for

ancestry of the marker locus and other nearby loci. This test is appropriate when the objective of the study is to exploit admixture to localize genes underlying ethnic variation in the trait value, using ancestry-informative markers rather than candidate gene polymorphisms.

affectedsonlyscorefile - tests for linkage of each locus with genes underlying the ethnic difference in disease risk, using only the affected individuals. The null hypothesis is that the risk ratio between populations that the locus accounts for is 1. This test statistic is computed for the effect of each subpopulation at each locus. The test compares at each locus the observed and expected proportion of gene copies that have ancestry from the high-risk subpopulation. This is the only test that can be used if the sample consists only of affected individuals. Even if a control group has been typed, for a rare disease the affected-only test is more efficient than the test given in ancestryassociationscorefile based on a regression model. This is because for a rare disease, the observed and expected proportion of gene copies that have ancestry from the high-risk subpopulation will not differ by very much in unaffected individuals.

allelefreqscorefile - tests for mis-specification of ancestry-specific allele frequencies.
This test is computed only if allele frequencies have been specified as fixed with option allelefreqfile. For each compound locus and each subpopulation, a score test is computed for the null hypothesis that the frequencies of all alleles have been specified correctly. A summary test over all k subpopulations is also computed at each locus.

args.txt - a list of the options used by the program. This is used by the R script to identify output files and other information. This is writen to *resultsdir*.

An R script (AdmixmapOutput.R) is supplied that processes these output files to produce tables of posterior quantiles, frequency plots of the posterior distribution, and plots of the cumulative posterior means for the variables that are output to paramfile. The R script also calculates a summary slope parameter for the effect of admixture from each subpopulation, versus the others. This R script is run automatically from the Perl script (admixmap.pl) that is supplied as a wrapper for the program.

Back to top

Interpretation of output from the program

These notes are based on the output produced by using the Perl script admixmap.pl to run the main program. Output files produced by the main are processed by the R script AdmixmapOutput.R. This produces several text files, and a file plots.ps containing graphs in postscript format

Evaluating the sampler

The adequacy of the burn-in period can be evaluated by the Geweke diagnostics in the R output. If the burn-in period is adequate, the numbers in this table should have approximately a standard normal distribution.

The mixing of the MCMC sampler can be evaluated by examining the autocorrelation plots. Autocorrelation extending beyond 20 iterations (2 thinned draws if every = 10) indicates slow mixing.

Acceptance rates for the Metropolis-Hastings samplers used by the program are printed to screen and logfile.

The adequacy of the total number of iterations can be evaluated by examining a plot of the statistic of interest calculated from all iterations since the end of the burn-in period, against the iteration number. Where inference is based on the mean of a parameter, this statistic is an ergodic (cumulative) average over all iterations to that point. Plots of ergodic averages of the population-level parameters are given in file ErgodicAveragePlots.ps.

Evaluating the fit of the model

The file *stratificationtestfile* contains results of a diagnostic test for residual population stratification that is not explained by the fitted model. For details of how this test is calculated, and a discussion of how to interpret it, see Hoggart (2003). The test is based on testing for allelic association between unlinked loci that is not explained by the model. The results is a "Bayesian p-value": $p < 0.5$ indicates lack of fit. The "Bayesian p-value" calculated by this test is more conservative than a classical p-value. Our experience has been that a test p-value of 0.3 or less is fairly strong evidence for residual stratification. Where this statistic yields evidence of lack of fit, the model should be specified with more subpopulations, unless there is some other reason for lack of fit such as mis-specified allele frequencies.

The file *dispersiontestfile* contains results of a diagnostic test for variation between the allele frequencies in the unadmixed populations that have been sampled to calculate the prior parameter values in priorallelefreqfile and the corresponding ancestry-specific allele frequencies in the admixed population under study. Again the results are "Bayesian p-values", for which the deviation of the test p-value from its expected value of 0.5 does not provide an absolute measure of the strength of evidence for lack of fit. For each subpopulation, the test statistic is calculated as a summary test over all loci and for each locus separately. Examination of the test statistic for each locus may reveal errors in coding, or errors in specifying the prior allele frequencies.

The option *dispersiontestfile* is valid only where option priorallelefreqfile has been specified. Where allele frequencies have been specified as fixed, option *allelefreqscorefile* should be specified and the output file should be examined.

No diagnostic test for lack of fit of the distribution of individual admixture proportions to the model is yet implemented. However the plots in file Plots.ps can be examined to compare the estimated distribution of individual admixture proportions (based on the the posterior means for individual admixture) with an estimate for the distribution of individual admixture values in the population (based on the posterior means for the Dirichlet parameters of this distribution).

The deviance and Deviance Information Criterion (DIC) are computed each time.

For an analysis of a single individual, with option *chib*, the log marginal likelihood, also known as the log evidence, is computed.

With option *thermo=1,* the marginal likelihood is approximated for any model. The greater the value of *numannealedruns*, the more accurate will be the approximation, but the longer the program will take to run.