Table of Contents

Table of Contents

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.

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 license. Offers to help with development of the program are welcome. The current version runs only on a single processor, and computation time can be a serious limitation on large datasets.

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, P.M. Prospects for admixture mapping of complex traits. *Am J Hum Genet* 2005; **76** (1):1-7.

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 a, b, g per Morgan (for three-way admixture). For given values of parental admixture, it is only necessary to specify a single parameter r for the sum of intensities: r = a + b + g.

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 $\mathbf{a}_i$ that approximate the posterior distribution are then computed by equating the posterior expectations of the allele frequencies to the ratios $\mathbf{a}_i / a\mathbf{S}_i$, 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 algorithm 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.

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-linked loci (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.

ADMIXMAP installation instructions

These instructions supersede any older version in the README file included with the program.

Obtaining and installing the binary package

The binary packages and the source code can be downloaded from https://pm2.phs.ed.ac.uk/downloads/.

Windows

Open the file admixmap-X.XX.XXXX.zip and unzip the contents into a suitable directory.

Admixmap may have problems if the filenames or directory paths contain spaces. For best results, extract the contents of the zip file to a directory for which the path does not contain any spaces.

Linux

Unpack the file admixmap-X.XX.XXXX-<arch>.tar.gz, where <arch> is the architecture of your machine (one of *i686*, *x86_64* or *x86_64-openmp*), in a suitable directory (tar -xzf admixmap*.tar.gz). You may need to change the permissions in order to run the program. Use the command *chmod a=x admixmap* to do this.

You can move the executable admixmap.exe to any directory that is in your PATH. The R script (AdmixmapOutput.R) should not be moved, unless you edit the perl scripts to specify its new location.

Other platforms

The source code is available as a tarball as admixmap-X.XX.XXX-src.tar.gz. Instructions for compiling it are provided. The source code is also available from our Subversion repository: instructions for accessing the repository are here.

Requirements for running ADMIXMAP

For the program to be usable, R and Perl should be installed.

R is a free statistical analysis program that can be downloaded from http://www.r-project.org. If you are using Windows, you will need to add the path for R to your PATH. For Windows XP, go to Start-Control Panel-System. On the Advanced tab, click on Environment Variables. Select the PATH variable and click on Edit. Add the path of the R binary directory to the end of the path variable. This is usually something like ";C:\Program Files\R\bin".

Perl is included in Linux distributions. Windows users can download it from http://www.activestate.com/Products/activeperl.

Contents of the installed directory

Distributed with the admixmap executable are the following:

- data: a folder with test data and input files. This test dataset is based on a cross-sectional sample of Hispanic Americans in San Luis Valley, California, typed at 21 marker loci, with measurements of a binary variable (diabetes) and a continuous variable (skin pigmentation, measured as skin reflectance).
- inddata: a folder with data on a single individual. This dataset is based on a single (unknown) individual with no phenotype data.
- testArguments.txt: a text file containing a list of options to test the programme using the test data
- admixmap.pl: a sample perl script to use with the test data
- tutorial.pl: a script to go with the tutorial
- AdmixOutput.R: an R script to process the output of the admixmap executable
- README.txt: this file
- COPYING: a copy of the GNU Public Licence (GPL)

Running the program

From the command line, specifying an options file

This is the simplest (but not the most user-friendly) way to run the program.

If you have 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 option name. The complete list of options supported can be obtained by typing

```
admixmap --options
```

A sample optionsfile, *testArguments.txt* is provided with the test files. To try this, using the test dataset and the options file testArguments.txt (supplied as an example), open a console window (Start -> Run ->cmd in Windows), navigate to the admixmap directory, and type

```
admixmap testArguments.txt
```

Note: Do not use the name 'args.txt' for your options file as this may be overwritten by the program. If the admixmap executable is in your current directory and you are running Linux, you may have to prefix the word admixmap with "./" (the current directory is not in your search path by default).

The output will be written to a directory named results.

To generate summary statistics, tables and graphics from the output files generated by ADMIXMAP you will need to run the script *AdmixmapOutput.R*. The easiest way to do this is to run the Perl script as described below. If for some reason you prefer to run the R script from the command line or from an R session, you can do this as follows:

If your output files are in a directory other than 'results', you will need to supply R with the directory name.You can also optionally specify which format you want the graphical output. Available formats are PostScript(default), pdf, jpeg and png. Suppose your results are in a subdirectory of the working directory called 'myresults' and you want pdf graphics.

If you are running R from an interactive session (Rterm, Rgui)

Set the parameters as environment variables by using the R commands:

```
Sys.setenv("RESULTSDIR" = "myresults")
Sys.setenv("RPLOTS" = "pdf")
```

If you are running the R script in batch mode

There are two ways to specify the parameters:

1. Using environment variables

First set the RESULTSDIR environment variable.

To do this in Linux use: `export RESULTSDIR=myresults`

and in Windows use: `set RESULTSDIR=myresults`

Then run the R script using either:

```
R --vanilla <AdmixmapOutput.R >myresults/Rlog.txt
```

or

```
R CMD BATCH --vanilla AdmixmapOutput.R myresults/Rlog.txt
```

2. Using command-line options

```
R --vanilla --args myresults pdf <AdmixmapOutput.R >myresults/Rlog.txt
```

Alternatively, if you have version 2.5 or later of R, you can use the 'Rscript' frontend:

```
Rscript AdmixmapOutput.R myresults pdf >resultsdir/Rlog.txt
```

Invoking the program from a Perl script

If you are running several analyses with different options, it is easiest to use a Perl script, based on the example scripts supplied, to automate the above steps. For each analysis, the Perl script

1. specifies the program options (making it easy to see which options are changed in successive analyses)
2. writes these analyses to a file
3. starts admixmap with the options file .
4. sets the environment variable RESULTSDIR
5. runs the R script Consult the documentation for details of the options.

Viewing and editing Perl scripts is easier if you use a text editor that provides syntax highlighting, such as emacs (all platforms) or Crimson Editor (windows only).

To run the Perl script admixmap.pl, open a console window and type

```
perl admixmap.pl
```

You can now view the contents of the results directory. The file Rlog.txt contains a log file written by R.

If you want to do more processing of the program output, you can set the RESULTSDIR environment variable as described above, start an R session and run the R script by typing

```
source("AdmixmapOutput.R")
```

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").

If the options are specified in a text file, they must be indicated as `optionname=optionvalue`.

1. 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 empty.

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

Sets a seed for the random number generator.

2. Options for allele frequency Model

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

The *allelefreqfile* option is no longer supported: instead, specify the options *priorallelefreqfile* and *fixedallelefreqs=1*, and supply the allele frequencies in priorallelefreqfile format (without adding 0.5 to each cell).

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.

priorallelefreqfile

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.

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 *S* alleles or *S* possible haplotypes, a Dirichlet prior distribution is specified by a vector of *S* 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 (this is equivalent to combining the counts with a reference prior). 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 *K* subpopulations, is equivalent to specifying the option *populations = K*.
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. 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.

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.

3. Options that specify data files

locusfile

This 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 webpage 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.

genotypesfile

Specifies the path to a file containing genotypes for each individual typed. 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.

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(s) containing the variable(s) of interest should be specified by the *outcomevarcols* option, otherwise all columns are used. For example, 'outcomevarcols=1,3' uses the first and third columns.

coxoutcomevarfile

This file contains survival 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 observed 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).

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 #. If the file contains more than one covariate, the column(s) containing the covariate(s) of interest should be specified by the *covariatecols* option, otherwise all columns are used.

outcomevarcols

Valid only with outcomevarfile.

Vector of integer specifying the columns of the outcomevarfile to use. For example, "outcomevarcols = 3, 1" specifies to regress on the third and first variables, in that order. If not specified, all columns are used.

covariatecols

Valid only with both outcomevarfile and covariatesfile.

Vector if integers specifying the columns of the covariatesfile to use. If not specified, all columns are used.

reportedancestry

Not fully tested or documented: allows prior information about each individualâ s ancestry to be specified in the model.

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.

4. 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 individualâ 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 (default) - Hierarchical model on individual admixture

randommatingmodel

0 (default) - assortative mating model (admixture proportions the same in both parents)

1 - random mating model

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 (default) - the sum of intensities ρ is modelled as a global parameter, set to be the same on all parental gametes

globalpsi

0 - individual-level odds ratio female/male between ancestral populations (psi)

1 (default) - population level odds ratio psi

The prior parameters on log psi are set through the 'oddsratiosprior' option.

fixedallelefreqs

1 specifies that priorallelefreqfile contains fixed allele frequencies

0 (default) otherwise

correlatedallelefreqs

valid only with 'populations' or 'priorallelefreqfile' options

1 specifies a correlated allele frequency model

0 (default) otherwise

poplabels

A list of strings specifying the labels for the subpopulations in the model. For example, 'poplabels = Afr, Eur'. It is ignored unless the *populations* options is used and must have length equal to the value of *populations*.

5. 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 $\beta 0$ and $\beta 1$. This specifies a "GammaGamma" prior, which has mean

$E(r) = \alpha\beta_1 / (\beta_0 - 1)$ and variance $E(r)(E(r)+1) / (\beta_0-2)$.

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 information that the sum of intensities 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"

oddsratioprior

In a model with X chromosome, this option allows to set the prior on the log of the odds ratio parameter psi. If we are modelling a population-level odds ratio female/male founders by setting the 'globalpsi=1' option, we can specify

oddsratiosprior = a, b

where a and b are the mean and precision (inverse variance) of a Gaussian prior on the natural log of the odds ratio. So, setting a=2 and b=100 will specify that log psi is expected to be close to 2, which corresponds to an odds ratio of e^2. The value for the precision must be at least 0.1.

For a model with individual-level odds ratios (where 'globalpsi=0'), the option requires four arguments

oddsratiosprior = a, b, c, d

where a and b have the same meaning as above, while c and d are the shape and rate parameters of a Gamma prior on the psi precision (c and d must be positive).

etapriormean, etapriorvar

Specify the prior mean and variance of the dispersion parameter(s) 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 FST by

$\xi = (1 + FST) / FST$, so values of 50 and 500 for ξ correspond roughly to values of 0.02 and 0.002 for FST.

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 the 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 the option 'randommatingmodel=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

Specifies the prior precision (1 / variance) of the regression parameters.

popadmixproportionsequal

Specifies that the population-level admixture proportions are to be kept equal.

6. Output Files

Output files are written to the directory specified by *resultsdir*.

An R script (AdmixmapOutput.R) is supplied that processes these output files to produce tables of posterior quantiles, frequency plots of the posterior distributions, convergence diagnostics and plots of the cumulative posterior means. 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

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

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).

indadmixmodefile

Name of output file containing posterior estimates of the modes of individual admixture proportions and individual-level sumintensities (if globalrho=0).

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. 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. 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 as well as the mean and variance of the deviance, 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.

locusancestryprobs

Posterior marginal probabilities of each ancestry state (0, 1, or 2 copies from *k*-th population) at each locus in each individual. The output file name is *LocusAncestryPosteriorProbs.txt*. The file is formatted as an R object. To read it into R, use the command

```
dget(file="LocusAncestryPosteriorProbs.txt")
```

This will create a 4-way array indexed by individuals, loci, populations, ancestry states. You can use this to construct tests for allelic association stratified by locus ancestry, or to test for allelic association conditional on locus ancestry.

7. Tests and Diagnostics

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

Model diagnostics

chib

Set to1 to calculate the marginal likelihood for the first individual using the Chib algorithm.

thermo

Set to 1 to use thermodynamic integration to compute marginal likelihood.

testoneindiv

Set to 1 to compute the 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.

stratificationtest

Set to 1 to perform a test for residual population stratification (stratification not accounted for by the fitted model).

dispersiontest

Set to 1 to perform tests 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.

fstoutput

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 fstoutput=1 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.

Score tests

The *allelicassociationtest*, *haplotypeassociationtest*, *ancestryassociationtest*, *affectedsonlytest* and *residualldtest* each produce two files containing results of score tests obtained by averaging over the posterior distribution: a p-value file and a final table. The *admixtureassoctest*, *allelefreqtest* and *hwtest* only produce final tables.

The p-values, based on cumulative averages for the score and information over all posterior samples obtained after the burn-in period, are output at intervals of 10 Ã every. Monitoring these repeated allows the user to determine when the sampler has been run long enough for the test results to be computed accurately. These files are formatted to be read into R as arrays.

The final tables, which are based on the entire posterior sample, are used for inference.

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 asociation 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 ($\mathbf{UV^{-1}U'}$) has a chi-squared distribution under the null hypothesis.

admixtureassoctest

Set to 1 to perform a score test for the association of the trait with individual admixture. This option is valid only if an outcome variable has been specified. 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.

If *admixtureassoctest* 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.

allelicassociationtest

Set to 1 to perform score tests for association of the outcome variable with alleles at each simple locus, adjusting for individual admixture. 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.

haplotypeassociationtest

Set to 1 to perform score tests for association of the outcome variable with haplotypes for all compound loci containing haplotypes, adjusting for individual admixture. Valid only with *allelicassociationtest*.

residualldtest

Set to 1 to perform score tests for residual allelic association between pairs of unlinked loci.

ancestryassociationtest

Set to 1 to perform score tests at each compound locus for linkage with genes underlying ethnic variation in disease risk or trait values. This is a test for association of the trait with locus ancestry, adjusting for individual admixture and covariates. 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. 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.

affectedsonlytest

Set to 1 to perform score tests for linkage with ancestry at each compound locus, based on comparing the observed and expected proportions of gene copies at each locus that have ancestry from each subpopulation. This test is calculated from affected individuals only: individuals are their own controls. This is the only test that can be used if the sample consists only of affected individuals. Even when the sample includes both cases and controls, this test is more powerful than the regression model score test in *ancestryassociationtest* if the disease is rare. 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.

In addition to the p-values and final table, this test produces a file called "AffectedsOnlyLikRatios.txt", containing likelihood ratios for the affecteds-only test at values of 0.5 and 2 for the ancestry risk ratio.

allelefreqtest

Set to 1 to perform 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. 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 ch-squared test over all subpopulations is also computed at each locus.

hwtest

Set to 1 to perform score tests for heterozygosity across loci, as a test for departure from Hardy-Weinberg equilibrium. These can be used to detect genotyping errors.

Interpreting output from ADMIXMAP

The output files produced by ADMIXMAP should be processed by running the R script AdmixmapOutput.R. This produces several text files and graphs (by default in postscript format). The R script is run automatically if you use the Perl script to invoke ADMIXMAP.

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 *stratificationtest* outputs 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 result 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 *dispersiontest* outputs 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 *dispersiontest* is valid only where option priorallelefreqfile has been specified. Where allele frequencies have been specified as fixed, option *allelefreqtest* 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 plot DistributionIndividualAdmixture can be examined to compare the estimated distribution of individual admixture proportions (based on 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.

Tutorial on using ADMIXMAP to model genotype and phenotype data from an admixed or stratified population

If you have problems getting the program to run, email david.odonnell (please supply any error messages and logfiles if available)

To feed back comments on this tutorial, email paul.mckeigue

Append @ed.ac.uk to the email addresses given above.

Getting started

This tutorial is based on data from a sample of 446 Hispanic-Americans resident in Colorado, typed at 32 loci. ADMIXMAP is designed to analyse larger datasets with more markers, but the analysis of this small dataset illustrates all the methods. The following data files have been prepared for input to ADMIXMAP. Before starting, open these files to view them. They are most easily viewed with a program such as Excel that will interpret tabs as column separators.

Filename	**Contents**
outcomevars.txt	column 1: diabetes, coded as 0=unaffected, 1=affected. column 2: skin reflectance, scored as a quantitative trait
covariates2std.txt	age and sex, standardized about their sample means.
covariates3std.txt	age, sex and income group, standardized about their sample means.
genotypes.txt	genotypes at 32 SNP loci. These include 2 -4 SNPs in each of three candidate genes for diabetes (*CAPN10*, *PPARG*, and *SUR1*), and one SNPs that is in a candidate gene for skin pigmentation (*TYR*). The first column is the ID number. The names of the other columns (given in the header row) must match the first column of the locusfile (*loci.txt*). For each genotype, the two alleles are separated by a comma. The alleles must be numbered 1, 2, ... N, and this numbering must correspond to the sequence of rows in *priorallelefreqfile.*
loci.txt	locus description file, with locus name, number of alleles, and map distance from last locus. This file was generated from the file *LociChr.txt*, which gives the chromosome number and estimated genetic map position (in cM) for each locus. Map distances are given in morgans. If the locus is not linked to the last locus, the distance from last locus is coded as 100. If the locus is very close to the last locus (< 100 kb), the distance from last locus is coded as 0. Thus, for instance, the four SNPs (**simple loci**) in the CAPN10 gene are modelled as a single **compound locus**, with 16 possible haplotypes. Thus the 32 SNPs ("simple loci") will be grouped into 24 compound loci.
priorallelefreqs.txt	parameters for Dirichlet prior distributions of ancestry-specific allele frequencies (European, Native American, west African) at each compound locus. This file has one row for each possible haplotype at each compound locus. If the compound locus contains only one SNP, the number of possible haplotypes is of course just 2. Haplotypes are ordered by incrementing a counter from the right: for instance the 16 possible haplotypes at a compound locus consisting of four SNPs are ordered 1-1-1-1, 1-1-1-2, 1-1-2-1, ..., 2-2-2-2. Where the compound locus contains only one simple locus, the prior parameters are

	calculated simply by adding 0.5 to the observed allele counts in samples of unadmixed individuals. Where no data from unadmixed individuals are available, the prior parameters are specified simply as 0.5 (a "reference" prior). Where data from unadmixed individuals are available, the compound locus contains two or more simple loci, the parameters for the prior on haplotype frequencies are obtained by using ADMIXMAP with a single population model to generate the posterior distribution of haplotype frequencies from data on unadmixed individuals. For an example of how to do this, run the script *HapFreqs.pl* in the folder HapFreqs. This will generate the tables of parameter values that were used to specify the prior parameters for the compound loci DRD2, CAPN10, PPARG, and SUR1. These runs will take only a few seconds.
etapriors.txt	Parameters for gamma prior distribution on allele frequency dispersion parameters. See below for explanation

These notes assume that you have installed Perl (ActivePerl is the windows version, the R statistical package, and a viewer for postscript files (such as Ghostview).

First, download and install ADMIXMAP. The Perl script *tutorial.pl* has been provided for this tutorial. Edit the Perl script where indicated to specify the location of the ADMIXMAP executable, the R executable (*Rcmd.exe* on a Windows platform) and the R script (AdmixmapOutput.R) that processes the output from the ADMIXMAP executable.

You are now ready to start the Perl script from your working directory. To do this, open the console shortcut and type "*perl tutorial.pl*". This script will run the program six times with different models, calling the main ADMIXMAP program each time with appropriate command-line options. The command-line options are stored by Perl in an array or "hash". The Perl script provides a convenient means of running several analyses with different options in batch mode. After each run of the ADMIXMAP program, the Perl script will run the R script *AdmixmapOutput.R* to analyse the output files, and move all output files to a folder (subdirectory) named for the type of analysis that has been run. Five new subdirectories will be created, each containing files output by ADMIXMAP and the R script. On an ordinary PC, these analyses will take about 20 minutes. You can inspect the results of each analysis, as described below, to determine if a longer run (with more samples from the posterior distribution) is needed. The results quoted in the tutorial below are from a long run, so may not correspond exactly to those obtained with a shorter run.

We now examine the results of each ADMIXMAP analysis.

SinglePopResults

This folder contains results of analysis with command-line option *populations=1*. This analysis does not exploit ADMIXMAP's ability to model population structure, but does exploit its ability to fit regression models and to model association with haplotypes given unphased genotype data. For this purpose, ADMIXMAP is an alternative to HAPLOSCORE, a program which uses a similar score test for association with haplotypes in a regression model. Other uses of the single population model are to test for population stratification, and to estimate haplotype frequencies in samples from unadmixed populations. These haplotype frequency estimates can be used to specify priors for the analysis of data from admixed populations.

To specify that the second column of *outcomevarfile* contains the outcome variable of interest (skin reflectance), we specify the option *targetindicator=1* (i.e. offset by 1 from column 0). The program automatically determines that this is a continuous variable. No information about allele frequencies or haplotype frequencies is supplied, so the program generates the posterior distribution of haplotype frequencies from the data, given a reference prior. The program fits a linear regression model with skin reflectance as

outcome variable, and age, sex and body mass index as explanatory variables. As the program does not have to model population structure, it requires only a short run: a total of 1100 iterations, including a burn-in of 100, is ample for all test statistics to be computed.

The file *RegressionParamConvergenceDiagnostics.txt* contains results of a simple test for the adequacy of the burn-in period, attributed to Geweke (1992). If burn-in is adequate and the sampling run after burn-in is long enough, these test statistics should have a standard normal distribution. Extreme values of the test statistics (indicated by small p-values) imply that a longer burn-in, and a longer sampling run after the burn-in, should be used. For this simple model, there is no evidence of lack of convergence.

Now examine the log file. This contains the result of a test for residual population stratification, based on testing for allelic association between unlinked loci. See the main program documentation for more details of how this test statistic is calculated. Only 14 loci are used in this calculation, as only unlinked loci can be included. The result is reported as a **posterior predictive check probability** or "Bayesian p-value". This is the frequency with which, over the posterior distribution of model parameters, a simulated dataset gives results more extreme than the observed dataset. We can interpret the "p-value" of 0.04 as evidence for stratification: there are more associations between unlinked loci than expected by chance. In general, posterior predictive check probabilities are more conservative than classical p-values: a p-value of 0.05 is fairly strong evidence of lack of fit.

The file *args.txt* contains one line for each of the options specified on the command-line. This includes default values for some of the command-line options that were not specified in the Perl script. An alternative way to invoke the program with the options specified in this file is simply to type

```
[path to admixmap executable/]admixmap [path to args.txt/]args.txt
```

The file *HardyWeinbergTest.txt* contains results of tests for Hardy-Weinberg equilibrium at each of the 32 simple loci. This test is based on averaging over the posterior distribution of allele frequencies, whereas the classic test for Hardy-Weinberg equilibrium conditions on the observed counts of alleles. For a single population model, we expect the score test to give very similar results to the classical test. Positive scores indicate that the proportion of homozygotes is higher than expected given the allele frequencies. Two of the loci - FY and MID52 - show evidence of departure from Hardy-Weinberg equilibrium. For locus FY there are too few copies of allele 2 for the test to be valid. Possible explanations for deviation from Hardy-Weinberg equilibrium are genotyping error, a higher proportion of missing genotypes (failure to call the genotype) in heterozygotes, or population stratification accompanied by non-random mating. If population stratification accounts for the departure from Hardy-Weinberg equilibrium, we expect that there will be no evidence of departure from Hardy-Weinberg equilbrium when the test is repeated with a model of admixture between two or more subpopulations.

The file *PosteriorQuantiles.txt* contains summary statistics for the posterior distribution of the parameters of the regression model. Ignore the rows for Dirichlet parameter and sumIntensities, which are irrelevant to a single population model. The posterior means and 95% credible intervals for the regression coefficients for age and sex are given. The "precision" parameter is the inverse of the residual standard deviation. All these estimates will be very similar to those that would be obtained with a standard regression program: classical 95% confidence intervals are equivalent to Bayesian 95% credible intervals where (as in this application) the sample size is large and the priors on the regression coefficients are non-informative.

Each simple locus is tested for allelic association with the outcome variable (skin reflectance), and the haplotypes at each compound locus are tested for allelic association with hypertension. These score tests test the null hypothesis =0, where is the regression coefficient for the effect of number of copies of the allele (or haplotype) on the outcome variable, in a model with the covariates (age, sex and income in this case). The

test statistic is calculated by dividing the score (gradient of the log-likelihood at =0) by the square root of the observed information (curvature of the log-likelihood at =0). This statistic has a standard normal distribution under the null hypothesis. Positive score values indicate that the most likely value of is greater than zero. The proportion of information extracted is a measure of how much information about we have, in comparison with a dataset in which all variables were observed directly.

First examine the file *TestsAllelicAssociationFinal.txt,* which contains the results of score tests for association of the outcome variable with allele 1 at each simple locus. For ease of viewing, open this file with a program such as Excel that will interpret the tabs as column separators. The p-values in this table will be practically identical to those that would be obtained by testing for association with number of copies of each allele in a classical logistic regression analysis, adjusting for age, sex and income. Note that there are three loci for which the p-values are less than 0.01. Of these three, TYR192 is in a candidate gene for skin pigmentation, and CYP19e2 is closely linked to a gene (SLC24a5) that has recently been shown to account for some of the ethnic variation in skin pigmentation.

However we expect these test results to be confounded by population stratification, as this has not been accounted for in the statistical model.

The file *TestsHaplotypeAssociationFinal.txt* contains tests for association with haplotypes at each compound locus. All haplotypes with frequency < 1% are grouped together as "others". The other haplotypes are tested one at a time for association, and also with a summary chi-square test of the null hypothesis that all haplotype effects are equal. A more appropriate chi-square test would exclude the "others" category: this will be fixed in a later release.

Note that the proportion of information extracted is typically at least 70% for each haplotype - this is what we would expect when inferring haplotypes from unphased genotype data in the presence of strong allelic association between the simple loci within each compound locus.

TwoPopsResults

This folder contains results of analysis with command-line option *populations=2*. This specifies a model with admixture between two subpopulations. For this analysis we specify the allele frequencies in these two subpopulation as unknown, with (uninformative) reference priors. The program fits a linear regression model with individual admixture proportion, together with age, sex, and income category as explanatory variables. The program attempts to infer the structure of the population from the allelic associations between markers and from the association of marker alleles with the outcome variable. Even if your objective is only to study population structure, including an outcome variable (such as skin reflectance) that is strongly related to individual admixture proportions helps the program to learn about individual admixture proportions and population stratification.

With no information about allele frequencies, and only 21 ancestry-informative marker loci in this dataset, the program requires very long runsto explore the posterior distribution of the population admixture parameters. Examine the tests for convergence in the file *PopAdmixParamConvergenceDiags.txt*: unless you have specified a very long run (*samples = 10000,* or more), the p-values will indicate that the sampler has not run long enough. To see why long runs are required, view the postscript file *PopAdmixAutocorrelations.ps*. This shows that the sampling of the Dirichlet parameters for the distribution of admixture proportions in the population mixes slowly, with autocorrelation of 0.5 or more up to a lag of 50 iterations. Mixing is much faster when prior information about allele frequencies is supplied, and when larger numbers of ancestry-informative marker loci have been typed. Larger datasets with hundreds of ancestry-informative markers usually require only a few hundred iterations for reliable inference.

The log file shows that the posterior predictive check probability in the test for residual stratification is 0.51 - there is no evidence of residual stratification not accounted for by a model with two subpopulations.

The model now contains four extra parameters: two parameters for the (Dirichlet) distribution of admixture proportions in the population, a "sum-intensities" parameter that is equivalent to the effective number of generations since admixture, and a coefficient for the effect of admixture on the outcome variable in the regression model. The file *regparams.txt* contains samples from the posterior distribution of the regression coefficients. The file *PosteriorQuantiles.txt* contains summary statistics for the posterior distribution of the model parameters. The program infers the population admixture proportions as about 2 to 1, and that skin reflectance is inversely related to proportionate admixture from the subpopulation that contributes less to the ancestry of the admixed population. As the next analysis shows, these results from a model with no information about allele frequencies are close to those obtained when we supply information about allele frequencies in Europeans, Native Americans and west Africans. As the two subpopulations are not identifiable in the model, their labelling as 1 and 2 is arbitrary, and the labels of the subpopulations that are associated with higher and lower skin reflectance might be reversed when the program is run with a different seed on the random number generator. In principle, the labelling could even reverse during a single run of the sampler. However the inverse association of skin reflectance with proportionate admixture from the subpopulation that makes the smaller contribution to the admixed population should be consistent.

To determine whether the sampler has been run long enough to estimate the p-value accurately for the tests that we are interested in, we can examine the postscript files *TestsAllelicAssociation.ps* and *TestsHaplotypeAssociation.ps*. These show the results of successive evaluations of the score test, evaluated over all posterior samples obtained since the end of the burn-in period and written to file every 50 iterations. After a few hundred iterations, most of the p-values are stable.

As the regression model includes individual admixture proportion as a predictor variable, tests of allelic association will be adjusted for whatever population stratification is inferred by the model. The p-values for association of skin reflectance with tyr192 and cyp19e2 are still statistically significant (at $p = 0.007$ and $p = 0.00005$ respectively), from which we can infer that the associations seen in the single-population model are unlikely to be accounted for by confounding effects of population stratification. Note that the proportion of information extracted is now only about 80% at most loci, compared with nearly 100% in the single population model. This is because there are only 21 ancestry-informative markers in the study, and the score statistic (which is based on adjusting for individual admixture in the regression model) consequently varies over the posterior distribution. The proportion of information extracted can be interpreted as a measure of the efficiency of the test. With more markers, the proportion of information extracted in the score test would be higher.

Estimates of individual admixture proportions obtained in this analysis are not meaningful because unless at least some information about allele frequencies or individual ancestry is supplied, the subpopulations are not identifiable in the model. This does not matter if we simply want to adjust for population stratification, because the score tests for allelic association are valid even if the labelling of the subpopulations is permuted (thus reversing the sign of the corresponding regression coefficient.). To rank individuals by their "degree of admixture", we can examine the posterior mean of the "ancestry diversity" for each individual, in the file *IndAdmixPosteriorMeans.txt*. This is calculated as the probability that locus ancestry differs between two gene copies drawn at random from unlinked loci in the individual under study. It does not depend upon the labelling of the subpopulations, and so can be computed even when the subpopulations are not identifiable. If you are using the program to identify admixed individuals, without using prior information about allele frequencies, we recommend that you use this statistic.

PriorFreqResultsSkin

This folder contains result of analysis with command-line option *priorallelefreqfile=priorallelelefreqs.txt*. With this option, the program fits a model in which allele frequencies in the unadmixed populations sampled are assumed to be identical to the corresponding ancestry-specific allele frequencies in the admixed population. The file priorallelefreqs.txt has three columns specifying priors on the allele frequencies in Europeans, Native Americans and west Africans. Specifying a prior distribution for the allele frequencies, rather than specifying them as fixed constants, allows for the uncertainty in estimates of allele frequencies that are based on samples of finite size. At loci where no prior information about allele frequencies is available for a given subpopulation, a reference prior (all parameters equal to 0.5) is specified. Where haplotype frequencies have been estimated from phase-unknown genotypes, the uncertainty in these estimates can also be accounted for in the parameters of the prior distribution.

The file *PosteriorQuantiles.txt* contains the posterior mean, posterior median and central 95% credible interval for population-level variables: Dirichlet parameters for the distribution of admixture, sum-intensities, regression coefficients, and the population admixture proportions. If the study sample size is large, the posterior mode and 95% credible interval are asymptotically equivalent to the maximum likelihood estimate and 95% confidence interval. If the sample size is large, the posterior distribution will be approximately normal and thus the posterior mode will be approximately equal to the mean (or median). This approximation is closest if the variable has been transformed to lie on the real line (between minus infinity and plus infinity). Plots of the posterior densities of all population-level parameters are given in the postscript file *PosteriorDensities.ps*. The table below briefly explains what the variables tabulated in this file mean.

Variable	**Explanation**
Dirichlet parameters for distribution of admixture (Eur, NAm, Afr) in the population	The ratios between the Dirichlet parameters determine the average admixture proportions in the population, and the sum of the Dirichlet parameters determines the variance of admixture proportions between individuals. A large value of the sum of Dirichlet parameters implies that the variance of admixture proportions between individuals is small.
Sum of intensities	This parameter measures the frequency with which transitions between states of European, African and Native American ancestry occur along the chromosomes in this population. This parameter is assumed to be the same in all individuals unless the option globalrho=0 is specified. The parameter can be interpreted as the average number of generation back to unadmixed ancestors. Note that in this dataset, with only a few linked markers, we don't have much information from which to infer the sum of intensities parameter: the 95% credible interval is from about 5 to about 19.
Regression coefficients for intercept, covariates, and individual admixture	These are linear regression coefficients. For a model with *K* subpopulations, *K - 1* regression coefficients for individual admixture proportions are displayed (the population given in the first column of the allele frequency file is taken as the baseline category).
Population admixture proportions	Population admixture proportions are calculated by dividing the Dirichlet parameters by their sum.
Precision	Inverse of residual variance in regression model.

Posterior means of individual admixture are in the file *IndAdmixPosteriorMeans.txt.* These values can be plugged into other types of genetic analysis, but should not be used to test for a relationship with skin pigmentation because the association between individual admixture and the skin pigmentation has already been used by the program to learn about individual admixture. If you want estimates of individual admixture that you can plug into a regression model to test for association with the outcome variable, run the program

without an outcomevarfile.

The file *DistributionIndividualAdmixture.ps* contains histograms of the distribution of posterior means of individual admixture. For comparison, the distribution of individual admixture specified by the posterior means of the Dirichlet parameters is shown as a curve on the same plot. This file can be examined to test if there is any obvious lack of fit of the distribution of individual admixture proportions to the model: for instance a bimodal distribution of admixture proportions, which is not compatible with a Dirichlet distribution.

The file *TestsForDispersion.txt* contains cumulative results of a test for variation of allele frequencies between the unadmixed populations (in this example European, Native American and west African) that were sampled to obtain prior parameters in *priorallelefreqsfile*, and the corresponding ancestry-specific allele frequencies in the admixed population (in this case Hispanic-Americans in Colorado). The numbers given in the file are "posterior predictive check probabilities" or "Bayesian p-values". Small posterior predictive check probabilities indicate lack of fit. There are no loci at which the p-values are small, indicating good fit of the data to the allele frequencies given in priorallelefreqfile. In this small dataset, we do not have enough information to detect small departures from the "no dispersion" model. However the unadmixed Native American populations that were sampled (from north and south America) are unlikely to be exactly representative of the ancestral Native American subpopulation that contributed genes to this admixed population in Colorado. If there is evidence of dispersion of allele frequencies, we can deal with this by fitting a dispersion model as described in the next section.

The results of tests of association of skin reflectance with alleles at each locus are similar to those obtained in the TwoPopsResults analysis, in which no prior information about allele frequencies was supplied.

In this analysis, where the subpopulations are identifiable, we can test also for association of the outcome variable with ancestry at each locus. This is the basis of **admixture mapping**, an approach that exploits admixture to localize genes underlying ethnic variation in disease risk. The results of these tests are in the file *TestsAncestryAssociationFinal.txt*, with plots of successive evaluations in the postscript file *TestsAncestryAssociation.ps*. The program runs more slowly when these tests are specified. We can ignore the tests for association with African ancestry, as the proportion of African admixture in this population is too low for such tests to be meaningful. We can also ignore any tests for which the observed information is very small, as where there is not enough information in the data, the asymptotic properties of the score test in large samples (approximation of the log-likelihood to a quadratic function) will not hold. At some loci, such as GNB3, the observed information from the test for association with Native American ancestry is evaluated as negative - probably because the true value is close to zero and the sampler has not been run long enough to estimate it accurately, although it is formally possible for the information to be negative (log-likelihood not concave downwards) in small samples.

Note that the efficiency of this test (proportion of information extracted) is much lower than the efficiency of the test for allelic association. For association with Native American ancestry, the highest proportion of information extracted is at marker locus mid52. This marker locus is highly informative for Native American ancestry. There is evidence of linkage to genes underlying the ethnic differences in disease risk at TYR192 and CYP19e2. The scores for association with Native American ancestry are negative, implying that average skin reflectance is inversely related to the proportion of gene copies that are of Native American ancestry at this locus, as we would expect if the trait locus accounts for some of the ethnic difference in skin reflectance. The tests for linkage with ancestry, unlike the test for allelic association, use the genotype data from all loci on each chromosome to extract information about ancestry at each locus. With these tests, we expect any evidence of linkage to be detectable over a broad region: 20 cM or more. To establish whether linkage with ancestry at TYR192 and CYP19e2 can be confirmed, the next step would be to type more markers informative for Native American versus European ancestry in these regions.

The file *TestsAncestryAssociation.ps* also contains a plot of the proportion of information extracted at each locus, across all loci. With larger marker sets, this plot can be used to evaluate the coverage of each chromosome by the marker set.

PriorFreqResultsDiabetes

This folder contains results of an analysis with diabetes as outcome variable. As diabetes is a binary variable, the program fits a logistic regression model. The logistic regression coefficients are log odds ratios, and can be transformed to odds ratios by taking exponents. With age and sex as the only other covariates in the regression model, the posterior mean for the log odds ratio for the effect of unit change in Native American admixture is 3.0, with a 95% credible interval from 0.15 to 7.0. As this interval does not overlap 0, we can interpret this as a statistically significant ($p<0.05$) association of diabetes with Native American admixture, adjusted for age and sex. We cannot, however, exclude the possibility that the association is accounted for by an unmeasured confounder that is associated with the proportion of Native American admixture and is independently associated with diabetes. If we run the analysis again with age, sex and income in the model as covariates (edit the Perl script *tutorial.pl* or the file *args.txt* to specify option *covariatesfile='covariates3std.txt'*), the posterior mean for the adjusted log odds ratio associated with Native American admixture falls to 2.0, with a 95% credible interval that overlaps 0. This is because income is associated both with Native American admixture and diabetes. The results are thus compatible with an environmental explanation for the association of diabetes with Native American admixture in this population. A larger sample size, more ancestry-informative markers, and more extensive measurements of environmental covariates would be required to investigate this.

These concerns about confounding by environmental factors do not apply to the tests for allelic association or to the tests for linkage with locus ancestry. With these tests, it is sufficient to adjust for individual admixture proportions to guarantee that confounding by environmental factors will be eliminated. There is weak evidence of association with two candidate genes: PPARG (one of two SNPs) and SUR1 (one of the sixteen possible haplotypes). These results are consistent with associations described in other studies. As none of the markers in candidate genes for diabetes are informative for ancestry, and no other nearby ancestry-informative markers have been typed, we cannot assess the possible contribution of these candidate genes to ethnic variation in diabetes risk.

As diabetes is a binary trait, we can evaluate both affected-only and case-control tests for linkage with locus ancestry. The fit of these test results to the distribution expected under the null is given by the QQ plots. These show that the affected-only and case-control test statistics are a good fit to a standard normal distribution. The affected-only test is more powerful than the case-control test, but assumes that the ancestry state frequencies do not vary systematically across the genome within the admixed population. The case-control test is robust to violation of this assumption.

HistoricAlleleFreqResults

This folder contains results of analysis with command-line option *historicallelefreqfile=priorallelefreqs.txt*. With this option, the program fits a "dispersion" model for the allele frequencies, which allows the allele frequencies in the unadmixed populations to vary from the corresponding ancestry-specific allele frequencies in the admixed population under study. A single allele frequency dispersion parameter η (eta) is estimated for each subpopulation. The option *outcomes=2* specifies that regression models should be fitted for both outcome variables simultaneously. Thus if we are interested in evaluating associations with diabetes, we can still use the skin reflectance values to provide additional information about individual admixture.

The dispersion parameter can be interpreted as follows. Imagine that the variation between allele frequencies in the modern unadmixed west African populations sampled and the allele frequencies in the pool of genes of

African ancestry in the African-American population of Philadelphia had been generated by drawing two independent equal-sized samples from an ancestral total population. The allele frequencies in the two samples would differ as a results of sampling variation, and the variance of the sample frequencies between these two samples would depend upon the size of the sample that was drawn. The parameter η is equivalent to this sample size. Small values of η (<100) indicate dispersion of allele frequencies. η is related to Wright's F_{ST} (fixation index subpopulation-total) by the relation $F_{ST}=1/(1+\eta)$.

As the program does not have much information from which to estimate these parameters, we have to specify strong priors. These are specified in the file *data/etapriors.txt*. There is one row for each subpopulation, with rows ordered as in the columns of *historicallelefreqfile*. Each row specifies a shape parameter and a rate parameter for a gamma prior distribution. The prior mean is shape/rate, and the prior variance is shape/rate^2. We have specified the prior means on the dispersion parameters as 500, 50 and 50 for European, Native American and African allele frequencies respectively. based on estimates of F_{ST} between subpopulations within these continental groups. The prior variances are specified as 5, 0.5, and 0.5 respectively. These very small prior variances effectively constrain the dispersion parameters to be close to their prior means: there is not enough information in the dataset for us to be able to estimate the dispersion parameters from the data. With this dataset, the point of running a dispersion model is simply to examine whether relaxing the assumption of no dispersion of allele frequencies alters the results.

The posterior distributions of the allele frequency dispersion parameters (labelled as eta.Afr and eta.Eur for African and European subpopulations respectively) are plotted in the file *ParameterPosteriorDensities.ps*, and the means and medians are given in the file *PosteriorQuantiles.txt.*Posterior means of the standardized variance of allele frequencies (expressed as F_{ST}) between unadmixed and admixed subpopulations within each continental group are in the file *lociFst.txt*.

Estimates of the dispersion parameter have implications for the sample size required when estimating allele frequencies in unadmixed populations in order to model admixture. In general, there is no point in using a sample size that is an order of magnitude larger than the value of eta for a given subpopulation, because the allele frequencies in the unadmixed populations will not accurately predict the corresponding ancestry-specific allele frequencies in the admixed population.

The results of tests for allelic association and linkage with ancestry are similar to those obtained with the previous analysis using a model with no dispersion, suggesting that these analyses are fairly robust to assumptions about allele frequencies.

Other possible models

You can edit the Perl script or the *args.txt* file to specify other options.

randommatingmodel=1 would specify that the admixture proportions on the two parental gametes are independent draws from the Dirichlet distribution in the population. In general this option is only useful if you have at least 100 ancestry-informative markers, allowing the program to infer the admixture proportions of each parental gamete separately. The default is *randommatingmodel=0*

globalrho=0 would specify a hierarchical model for the sum-intensities parameter. Again this option is only useful if you have at least 100 ancestry-informative markers, allowing the program to infer the sum-intensities parameter for each gamete. The default is *globalrho=1*.

indadmixhiermodel=0 would eliminate the hierarchical model for individual admixture, allowing you to specify a Dirichlet prior on individual or gamete admixture directly with *initalpha0* and *initalpha1*. This

option is occasionally useful when you do not want the "shrinkage" effect of a hierarchical model, in which outlying observations are pulled towards the population mean.

Frequently Asked Questions

ADMIXMAP executable

Are there any differences between the Windows and Linux versions?

No. They are identical, just compiled for different platforms.

Is there a Windows interface / GUI?

No. An early version had a Windows frontend but we are no longer using this. The console version runs fine under Windows.

Is there a version for UNIX?

There is no precompiled version for UNIX but you can easily compile one yourself. Instructions on how to obtain and compile the source code are here.

Will it work with a different processor?

Probably. If not, you can compile a version optimised for your platform. See instructions here.

The program won't run!

Check you have execute permission for the executable (ls -l admixmap). If not, use the command chmod +ux.

The program has crashed!

Examine any error messages and ensure the options are valid. If it still won't run, send us the data and the exact options you are using and include any error messages.

What is best way to put in the genotypes from the gene we want to test for association with is. Is it okay to put them in the genotypes file? And how should it be declared in the locus file?

ADMIXMAP makes no distinction between markers and candidate loci. All loci are tested for association.

Should we specify ancestral frequencies where they are available or should we not specify and let the program do its analysis anway?

In the prior allele frequency file, enter 0.5 (reference prior) for cells where you have no allele frequency data from the parentals.

Can I use microsatellite data?

Yes. Code the different alleles as integers from 1 to the number of alleles.

When should I use a correlated allele frequency model?

The correlatedallelefreqs option should be used only if you are unable to supply priors on the ancestry-specific allele freqs. Specifying this option introduces another layer into the model - instead of a prior on the allele

frequencies, the allele frequencies are modelled as drawn from a Dirichlet distribution parameterized as (p, eta) where p is a proportion vector (varies across loci) and eta is the dispersion parameter (same for all loci). This helps the program to infer subtle stratification/admixture, but should be used only as a last resort when prior information on the allele freqs is not available.

How do I specify a prior for the dispersion parameter in a dispersion ('historicallelefreqs') model?

The mean and variance you specify for eta should reflect your subjective evaluation of the range of values that are plausible. If you are unsure, you can use the default.

How do I run an admixture mapping case-only analysis?

Just run the program without supplying an outcome file, and specify the 'affectedsonlytest' option: it will assume that all individuals are cases.

R script (AdmixmapOutput.R)

Error message #1:

```
'R' is not recognized as an internal or external command,
operable program or batch file.
```

You need to put the path to R in your PATH environment variable. Instructions on how to do this are in the manual.

Error message #2:

```
unable to open input file
```

R cannot find the script. Check the path to the script (relative to the working directory) is correct. The simplest arrangement is to have the perl script, R script and executable all in the same place, one level above a subdirectory for results.

Error message #3:

```
Error in xy.coords(x, y, xlabel, ylabel, log) :
        (subscript) logical subscript too long
Execution halted
```

Probably a bug. Send us your output and we'll try to fix it.

Error message #4:

```
Error in FUN(newX[, i], ...) : 'x' is empty
```

R is trying to read an empty file. You are probably using an old version of the R script. Get the latest version and if the problem persists, let us know.

Error message #5:

```
Error in file(file, "r") : unable to open connection
```

R cannot find your results. Check your paths are correct.

Error message #6:

```
unable to open output file
```

The resultsdir you specified is invalid. Check the path is correct.

The R script crashes when reading large output files. Are they too big?

Try increasing R's memory limit using the function memory.limit or the command-line argument max.-memory-size. If this doesn't work try thinning the output by specifying a larger value of *every*.

How do I view the graphical output?

You need a postscript viewer e.g. Ghostview.

How do I get pdf format graphics?

Set the environment variable 'RPLOTS=pdf' or run the script with 'Rscript' (R version 2.5 or later):

```
Rscript AdmixmapOutput.R <resultsdir> pdf
```

Tools for converting between data formats

Note: Windows may block your browser from downloading executable files, including the Perl scripts linked from this page. If so, you can download the zipped versions from here.

Convert data from ANCESTRYMAP format to ADMIXMAP format

C++ program

anc2adm.zip: Windows (45kb), Linux (30kb) or source (14kb)

Instructions are built in.

PERL Script (right-click on link and choose "Save Link As")

ANCtoADM.pl (Thanks to Arti Tandon)

Convert data from ADMIXMAP format to ANCESTRYMAP format

C++ program

adm2anc.zip: Windows (45kb), Linux (29kb) or source (14kb)

Instructions are built in. Header fields should be delimited by double quotes. Converting the locus file and allele freqs file may require a separate run of the program (skipping conversion of genotypes file).

Convert data from STRUCTURE format to ADMIXMAP format

PERL script (right-click on link and choose "Save Link As") to create an ADMIXMAP genotypes file. (Thanks to Indrani Halder)

STRtoADM.pl sample inputfile

Run as : "perl STRtoADM.pl <inputfile> [<genotypesfile>]". The genotypesfile name is optional and defaults to "genotypes.txt".

Note that you will also need to create a locusfile in order to use ADMIXMAP.

Simulating data from an admixed population in ADMIXMAP format

R script

SimAdmixture.R

Contact Information

Paul McKeigue: paul DOT mckeigue AT ed DOT ac DOT uk

If you have problems running the program, please supply as much information as possible, including any on-screen error messages, the logfile and Rlog if available.