NAME

poolmapB − Statistical analysis for gene mapping using data from DNA pools

SYNOPSIS

TO COMPUTE THE NONPARAMETRIC LIKELIHOOD RATIO TEST STATISTIC

poolmapB −P −S −C [**−d***datafile*] [**−e***errormodelfile*]

TO COMPUTE POSTERIOR USING QUADRATURE

poolmapB −P −S −F [**−d***datafile*] [**−e***errormodelfile*] [**−c***outputfileroot*]

TO COMPUTE POSTERIOR USING MCMC

[**GSL_RAN_SEED=***seedvalue*] **poolmapB −C −F** [**−d***datafile*] [**−e***errormodelfile*] [**−n***samples*] [**−s***outputfile*] [*options*]

TO COMPUTE PROFILE LIKELIHOOD

poolmapB −S −C −F [**−d***datafile*] [**−p***outputfile*]

DESCRIPTION

poolmapB analyses genetic data to estimate the position, age and prevalence of a disease risk factor.

The data are allele frequency estimates at a number of marker SNPs (single nucleotide polymorphisms), for a group of cases (with the disease) and a group of controls (without the disease).

The error model specifies how accurate the allele frequency estimates are.

At present three statistical methods are implemented, and the (slightly odd) default behaviour of the program is to apply all three methods. A fully Bayesian analysis can be performed using Cartesian product quadrature and/or using MCMC (Markov chain Monte Carlo). A more approximate but faster analysis can be performed by calculating a profile likelihood. This document only explains how to perform these analyses using the poolmapB program; for an explanation of the statistical methodology consult the references below.

INPUT

All methods require the same input data file. By default this is called *poolmap_data*, and is changed using the **−d** option.

The fully Bayesian methods (quadrature and MCMC) use a file describing an error model for the allele frequency estimates, and/or a file of precomputed emmission probabilities. By default these are called *poolmap_error* and *poolmap_emitprob*, and are changed using the **−E**, **−e**, **−Q** and **−q** options.

The formats for these files are described below.

OUTPUT

As the program runs, progress is reported to standard error. Tests for the presence of a QTL are also reported to standard error.

The profile likelihood is written to a file *poolmap_proflik*, unless changed using the **−p** option. The fast version of the profile likelihood is written to a file *poolmap_viterbi*, unless changed using the **−f** option. The samples generated by the MCMC are written to a file *poolmap_samps*, unless changed using the **−s** options. The posterior computed using Cartesian product quadrature is written to a file *poolmap_quad_all*, and the marginal posteriors are written to files *poolmap_quad_mu*, *poolmap_quad_mu_flat*, *poolmap_quad_tau* and *poolmap_quad_rho*. The root of these filenames can be changed using the **−c** option.

The emmission probabilities computed for the Bayesian (MCMC and Cartesian product quadrature) methods are written to a file *poolmap_emitprob*, unless changed using the **−q** option. These may be read in to speed up subsequent runs of the program on the same data.

MCMC TUNING

The **−n** option determines the number of samples that will be generated.

It would be very unwise to assume that the MCMC analysis will work using the default initial point and proposal distribution. The performance may be radically improved by adjustment of the proposal distribution.

The initial point is set using the **−a** or **−m**, **−r** and **−t** options. The proposal distribution is set using the **−A** or **−M**, **−R** and **−T** options. For details see below.

OPTIONS

The available options are as follows.

−h Display this documentation, then exit.

−v Display version information for the program, then exit.

−l Display the GNU General Public Licence, then exit.

−A[*value*]
Automatically set the standard deviation parameter for the position proposal distribution for the MCMC analysis, by using *value* (or 1 if omitted) times the standard deviation of the normalised profile likelihood. This overrides any setting made using the **−M** option.

−a Automatically set the initial position for the MCMC analysis, by using the expectation of normalised profile likelihood. This overrides any setting made using the **−m** option.

−C Skip the Bayesian analysis using Cartesian product quadrature. The default is not to skip it.

−c*filename*
Set the root part of the names of the files to write output from the samples from the quadrature analysis. The suffixes *_all*, *_mu*, *_tau* and *_rho* will be appended. The default is *poolmap_quad*.

−d*filename*
Set the name of the data file, which contains the data to be analysed. The default is *poolmap_data*.

−E Assume the data are exact counts with no error. The values in the data file are rounded to the nearest integer if necessary. The default behaviour is to assume the error model described by the error file. Using this option is equivalent to providing a *poolmap_error* file with a single **e** on every line.

−e*filename*
Set the name of the error file, which describes the error model for the allele frequency estimates. This is used to compute the emmission probabilities for the Bayesian (quadrature and MCMC) analyses. The default is *poolmap_error*.

−F Skip computing the fast (Viterbi−nonparametric) approximation to the Bayes factor. The default is not to skip it.

−f*filename*
Set the name of the file to write the fast (Viterbi−nonparametric) approximation to the posterior for position of the disease risk factor. This is identical to the profile likelihood but uses the emmission probabilities calculated by taking into account errors in allele frequency estimation. The default is *poolmap_viterbi*.

−G Read the design points and prior for quadrature analysis from files instead of using the defaults (see below). The file names are set using the **−g** option.

−g Set the root part of the names of the files containing design points for quadrature analysis. The suffixes *_mu*, *_tau* and *_rho* will be appended. The default is *poolmap_design*.

−I*value*
Set the maximum of the prior for the position parameter to value. A uniform prior between the minimum and maximum is assumed. Defaults to the position of the last marker in the data file.

−i*value*
Set the minimum of the prior for the position parameter to value. A uniform prior between the minimum and maximum is assumed. Defaults to the position of the first marker in the data file.

−M*value*
Set the standard deviation parameter for the position proposal distribution for the MCMC analysis to *value*. The default is 0.1 times the marker span.

−m*value*
Set the initial position for the MCMC analysis to *value*. The default is the mid-point of the marker span.

−n*value*
Generate *value* samples in the MCMC analysis. This defaults to 10,000, which may take a long time for large data sets on a slow computer.

−P Skip the profile likelihood analysis. The default is not to skip it.

−p*filename*
Set the name of the file to write the profile likelihood to. The default is *poolmap_proflik*.

−Q Read the emission probabilities from a file instead of computing them. The file defaults to *poolmap_emitprob* unless changed using the **−q** option.

−q*filename*
Set the name of the file to write the emission probabilities to. The default is *poolmap_emitprob*

−R*value*
Set the parameter for the rate parameter proposal distribution for the MCMC analysis. At each update the proposed rate is beta distributed with parameters (*value* times rate) and (*value* times (1−rate)). The default is 5.

−r*value*
Set the initial value of the rate parameter in the MCMC analysis to *value*. The default is 0.5.

−S Skip the MCMC analysis. The default is not to skip it.

−s*filename*
Set the name of the file to write samples from the MCMC analysis to. The default is *poolmap_samples*

−T*value*
Set the shape parameter for the age parameter proposal distribution for the MCMC analysis. At each update the proposed age is gamma distributed with shape parameter *value* and mean equal to the current value of the age parameter. The default is 10.

−t*value*
Set the initial value of the age parameter in the MCMC analysis to *value*. The default is 100.

−x*value*
Set the largest value of x in any analysis to be *value*. This may speed up the analysis, especially the propagation algorithm that is run for every sample in the MCMC analysis, which may take O(x^2) operations. Artificial lowering of the maximum for x may be a good approximation when the disease variant is at low frequency within the group of cases. If not set, it defaults to the largest value consistent with the data, i.e. the calculation is done exactly.

RETURN VALUE

The program returns 0 upon successful completion. Other return values indicate a problem with the command line arguments or input files, and are accompanied by a diagnostic message.

DIAGNOSTICS

Email me if you do not understand the messages output by the program. During calculation of emmission probabilities or sampling, progress can be checked by examining the file *poolmap_emitprob* or *poolmap_samples*.

EXAMPLES

Compute just the profile likelihood for the data in file hosking_27SNP

poolmapB −S −C −dhosking_27SNP −phosking_proflik

Generate the posterior like the one shown in the poster

poolmapB −C −dhosking_27SNP −shosking_postprob −E −n50000

ENVIRONMENT

GSL_RNG_TYPE

The type of random number generator used in the MCMC analysis can be set at run time to any value supported by your libgsl. The default is mt19937. Other (faster) simulation quality generators are taus and gfsr4.

GSL_RNG_SEED

The seed for the random number generator used in the MCMC analysis is set using this environment variable. If it is not set, the program will generate identical output on every run. To obtain independent sets of samples you must set this to a different value for each run.

REFERENCES

Johnson, T. (2005) *Multipoint linkage disequilibrium mapping using multilocus allele frequency data.* The Annals of Human Genetics **69**:474−498. http://dx.doi.org/10.1046/j.1529−8817.2005.00178.x

Johnson, T. (2005) *Bayesian method for disease QTL detection and mapping, using a case and control design and DNA pooling.* (in prep)

Liu, J. S., C. Sabatti, J. Teng, B. J. Keats and N. Risch (2001) *Bayesian analysis of haplotypes for linkage disequilibrium mapping.* Genome Research **11**:1716−1724. http://dx.doi.org/10.1101/gr.194801

McPeek, M. S. and A. Strahs (1999) *Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping.* Am. J. Hum. Genet. **65**:858−875.

Morris, A. P., J. C. Whittaker and D. J. Balding (2000) *Bayesian fine-scale mapping of disease loci, by hidden Markov models.* Am. J. Hum. Genet. **67**:155−169.

GNU Scientific Library http://www.gnu.org/software/gsl

Blitz++ class library http://www.oonumerics.org/blitz

BUGS

Options that take an optional argument (only **−A**) require that argument to be given immediately after the option, without any intervening space. Thus **−A 1** is wrong but **−A1** is right. This is a feature of using GNU getopt.

A more sensible default behaviour would be to just do the quadrature analysis since it is best, rather than all four.

Map positions are (mostly) output to 8 decimal places, which corresponds to 1 bp if one cM per Mb is assumed. There should be better control of this. Perhaps the whole program should deal with physical positions?

The input files contain only data and no headers.

An option to use an initial point sampled from the prior would be a useful option for running multiple chains and using Gelman's convergence diagnostic.

A quiet mode and a summary statistic mode would be useful.

The Viterbi-like DP algorithm that is better PAVA should be used to completely replace the old profile likelihood method.

The propagation algorithm used in the Cartesian product quadrature has had improvements in speed and numerical stability that have not been incorporated into the corresponding parts of the MCMC analysis.

The default prior for the age integrates to slightly less than one using the default design points for quadrature (see below).

AUTHOR

Toby Johnson toby.johnson@ed.ac.uk

FILE FORMATS

The input files read by poolmapB only data and no headers. This is a poor feature of the program.

THE DATA FILE

The data file contains one row per SNP. Each row contains five (floating point) numeric fields. In order, they are (1) the map position of the SNP in (assumed) Morgans, (2) the estimated count of the 0 allele in cases, (3) the estimated count of the 1 allele in cases, (4) the estimated count of the 0 allele in controls, and (5) the estimated count of the 1 allele in controls. Fields 2 and 3 should add up to the same total for every line (SNP). Each row of the file could be the data for a chi-square test for association using a 2 by 2 table.

If you have real counts and missing data, you must estimate the total counts and then specify that you did this in the error model file.

THE ERROR MODEL FILE

The error model file contains one row per SNP, describing the error model specific to that SNP. On each row the first character of the first field determines the type of error model. Subsequent fields on that row contain parameters for the model. The available error models are

e Assume that the allele counts are exactly correct when rounded to the nearest integer. No parameters are required for this model.

c Assume that the allele frequencies derive from counts with missing-at-random data. The two parameters are the number of observations used in the cases and in the controls respectively. These numbers must not be greater than the total sizes of the respective pools.

g Assume that the allele frequencies are derived from kinetic PCR experiments where the error in lag estimation is Gaussian with constant and known standard deviation. The three parameters are this standard deviation and the number of replicate experiments over which the lag was averaged for the cases and in the controls respectively.

When an error function with large variance is used, for large numbers of cases and/or large numbers of SNPs, computing the emmission probabilities may be time consuming. It is possible to save these emmission probabilities so that multiple samplers can be run without duplicating this calculation. These are stored in a file with three fields per line. These are (1) the SNP number, (2) the value of x, and (3) the (natural) log of the emmission probability. If emmission probabilities are read from a file, any error file is ignored. The count data in the datafile are also ignored, but the map positions are still read from this file.

THE DESIGN FILES

The three files (by default called *poolmap_design_mu*, *poolmap_design_tau* and *poolmap_design_rho*) can specify design points and the prior for quadrature analysis. The settings in this file override any other specification of the prior (e.g. using the **–i** and **–I** options). Each file contains one line per design point. Each line contains three fields, which are (1) the location (value of mu, tau or rho) of the design point, (2) the weight for the design point, and (3) the density of the prior at the design point.

The product of (2) and (3), summed over all design points, must equal one and the program will abort with an error if it does not. That is, the prior must integrate to one when the design is used. This may be an undesirable feature and an option should be included to turn off this check.

If no design is specified, a default design is used.

DEFAULT DESIGN FOR QUADRATURE

The default design for mu is 100 points at positions 0.005, 0.015, 0.025, ..., 0.995 on a scale where 0 and 1 are the minimum and maximum of the prior for mu. Each point has a weight of 0.01 times the width of the prior.

The default design for tau is 100 points. The i–th point is at exp(i/11) and has weight (1/11)exp(i/11), where i = 0, 1, 2, ... 99. Thus the points span an interval from 1 to 8103 generations. Note that the default

prior (exponential with mean 1000) integrates to only 0.9988 using this design. This is a bug.

The default design for rho is 100 points at positions 0.005, 0.015, 0.025, ..., 0.995. Each point has a weight of 0.01.

COPYRIGHT

LICENCE AND WARRANTY