BIMBAM user manual

Yongtao Guan and Matthew Stephens
Baylor College of Medicine and University of Chicago

Version 1.0
Revised on 25 June 2015

Contents

1 Copyright

What BIMBAM can do

- IMPUTATION: to fill in missing genotypes or untyped genotypes. The output can be genotype distribution (-wgd), mean genotype (-wmg), or best guess genotypes (-wbg). However, mean genotypes are recommended for computing Bayes factors.
  - Take the panel (e.g. HapMap) and cohort (study sample genotypes) as input. This mainly uses LD information in the dense genotyped panel.
  - Take the cohort alone as input. This only uses LD information in the cohort genotypes.
- SINGLE-SNP SSSOCIATION: to compute single SNP Bayes factors. The input files can be either raw genotype (with/without missing values), mean genotype, or genotype distributions.
  - For quantitative phenotypes, compute single SNP Bayes factors for each SNP. The priors of additive and dominant effect sizes can be specified by users (-A, -D, -df).
  - For binary phenotypes, compute single SNP Bayes factors for each SNP via Laplace approximation for logistic regression (-cc).
  - Compute p-values for each SNP via permutation of phenotypes (-pval).
  - Compute single-SNP Bayes factors using importance sampling for imputed genotypes (-i). Warning: computation intensive.
- MULTI-SNP SSSOCIATION: to compute multi-SNP Bayes factors. The input can be either mean genotypes or genotype distributions.
  - This shall be used only for candidate gene studies, because it exhaustively goes through all possible combinations of SNPs with specified maximum number of SNPs (-l).
- COMBINE STUDIES USING SNP SUMMARY DATA
  - Produce SNP summary data (-psd).

- Use SNP summary data to compute single SNP Bayes factors (-ssd).

- GENOTYPE MANIPULATION
  - Exclude SNPs based on missing proption, or minor allele frequencies, or if SNPs have an entry in position files (-exclude-miss, -exclude-maf, -exclude-nopos).
  - Specify a genomic region and only keep SNPs in the region (-gene, -gf).
  - Write exact numerically coded genotypes without imputation with missing genotypes denoted by NA (-weg).
  - Support non-genotype file, for example, microarray intensity data (-notsnp).

2 Introduction

BIMBAM implements methods for "Bayesian IMputation-Based Association Mapping". It is suitable for single-SNP analyses of large studies (e.g. genome scans) and multi-SNP analyses of smaller studies (candidate regions or genes). The software is written by Yongtao Guan, based on work from Scheet and Stephens (2006) and Servin and Stephens (2007) and Guan and Stephens (2008). Users of the software are assumed to be somewhat familiar with the papers of Servin and Stephens (2007) and Guan and Stephens (2008). Examples of applying BIMBAM to data analysis can be found in Reiner et al. (2008) and Barber et al. (2010).

Please send bug reports and requests for help to `bimbam_help@googlegroups.com`. Before sending a request check out `http://groups.google.com/group/bimbam_help` to see if someone else has already asked the same question. If you use BIMBAM for imputation, please cite Scheet and Stephens (2006). If you use BIMBAM for association mapping, please cite Servin and Stephens (2007). For practical issues, please cite Guan and Stephens (2008).

2.1 The model

Bayes Factors are computed under linear or logistic regression of phenotypes on genotypes. Specifically, for quantitative phenotypes the BFs are computed under the model

$$Y_i = \mu + aX_i + dI(X_i = 1) + \epsilon_i \tag{1}$$

where Y_i denotes the phenotype for individual i, X_i denotes the genotype for individual i (coded as 0, 1 or 2), a denotes the additive effect, d denotes the dominance effect, and ϵ_i denotes an error term (assumed to be iid normal). The BFs are computed using the prior D2 from Servin and Stephens (2007), averaging over $\sigma_a = 0.05, 0.1, 0.2, 0.4$ and $\sigma_d = \sigma_a/4$.

Similarly, for binary (0/1) phentypes the BFs are computed under a logistic regression model,

$$\log(\Pr(Y_i = 1)/\Pr(Y_i = 0)) = \mu + aX_i + dI(X_i = 1). \tag{2}$$

The BFs are computed under the same priors for μ, a and d as in prior D2 from Servin and Stephens (2007), using a Laplace approximation to perform the necessary integration.

Note that the above models are both "prospective", and so BFs computed from these models are appropriate for prospective studies, but not strictly appropriate for retrospective designs (e.g. case-control designs, or where genotype data are collected on individuals whose quantitative phenotypes lie in the tails of the population distribution). Most published analyses of case-control designs use

prospective models, and is known that, asymptotically, maximum likelihood parameter estimates based on these models converge to the correct values. For typed SNPs, results from Seaman and Richardson (2004) provide conditions for the equivalence of prospective and retrospective Bayesian analysis. Although these results do not apply directly to imputed SNPs, we anticipate that even for these SNPs, using BFs from prospective models to analyse case-control data will not be grossly misleading.

2.2 Transforming quantitative traits

For quantitative traits an important assumption underlying the methods implemented in BIMBAM is that the phenotype has a normal distribution within each genotype class. Based on unpublished data (M Barber and M Stephens) we suggest using a normal quantile transformation to transform the phenotype to be normal before running BIMBAM . For example, in `R`, this can be accomplished using `xtransformed = qqnorm(x,plot.it =F)$x`.

This quantile transformation does not fully solve the problem (it ensures that the phenotype is normal overall, but not necessarily normal within each genotype class). However, with the small effect sizes typical in genetic assocation studies it appears to be a simple sensible way to guard against strong departures from modelling assumptions. If you have other covariates that may be important predictors of phenotype (e.g. Age, Sex) we suggest first regressing the phenotype on these covariates using standard multiple linear regression software, and then running BIMBAM on the residuals from this regression (after applying a normal quantile transformation to these residuals).

3 Input file formats

In most cases, the users must supply two input files: a genotype file and a phenotype file. Optionally, a SNP location file can also be specified (if this is missing then the physical locations of the SNPs will be assumed to be in the same order as they occur in the Genotype file). If data are available on multiple chromosomes, we suggest analyzing each chromosome separately.

Notes on input file conventions:

1. Input files should be saved as plain text files.
2. The input files can be comma-delimited, space-delimited, tab-delimited, and semi-colon delimited, or mixed use of those. (i.e. entries can be separated by commas, spaces, semi-colons or tabs).
3. All input files can contain empty lines, and comment lines: lines starting with # are ignored by BIMBAM .

The following sections describe the format of each input file in more detail. The software distribution also includes example files (`test.geno.txt, test.pheno.txt` etc.) in the `input` subdirectory.

3.1 Basic genotype file format

Here is an example of basic genotype file, with 5 individuals and 4 SNPs:

```
5
4
IND, id1, id2, id3, id4, id5
rs1, AT, TT, ??, AT, AA
rs2, GG, CC, GG, CC, CG
rs3, CC, ??, ??, CG, GG
rs4, AC, CC, AA, AC, AA
```

Genotypes should be for bi-allelic SNPs, all on the same chromosome. The number on the first line indicates the number of individuals; the number in the second line indicates the number of SNPs. Optionally, the third row can contain individual ID: this line should begin with the string IND, with subsequent strings indicating the identifier for each individual in turn. Subsequent rows contain the genotype data for each SNP, with one row per SNP. In each row the first column gives the SNP ID (which can be any string, but might typically be an rs number), and subsequent columns give the genotypes for each individual in turn. Genotypes must be coded in ACGT while missing genotypes can be indicated by NN or ??.

Note that plink can convert genotype files from plink format to bimbam format. The option is `--recode-bimbam`.

3.2 Phased genotype file format

By default BIMBAM assumes that the genotypes in the basic genotype file are *unphased*. If one has data where the phase information is known, or can be accurately estimated (e.g. from trio data, as in the HapMap data), then this can be specified by putting an "=" sign at the end of the first line, after the number of individuals. In this case, the order of the two alleles in each genotype becomes significant: the first allele of each genotype should correspond to the alleles along one haplotype, and the second allele of each genotype should correspond to the alleles along the other haplotype. For example, in the following input file, the haplotypes of the first individual are AGCA and TCCC:

```
5 =
4
IND, id1, id2, id3, id4, id5
rs1, AT, TT, ??, AT, AA
rs2, GC, CC, GG, CC, CG
rs3, CC, ??, ??, CG, GG
rs4, AC, CC, AA, AC, AA
```

Note: accidentally treating phased data as unphased is less harmful than accidentally treating unphased panel as phased, so make sure it is phased genotype before you put "=" sign!

3.3 Mean genotype file format

This file has a *different* form from the genotype input file. There are no number of individual line or number of SNPs line. The first column of the mean genotype files is the SNP ID, the second and third columns are allele types with minor allele first. The rest columns are the mean genotypes of different individuals – numbers between 0 and 2 that represents the (posterior) mean genotype, or

dosage of the minor allele. An example of mean genotypes file of two SNPs and three individuals follows.

```
rs1, A, T, 0.02, 0.80, 1.50
rs2, G, C, 0.98, 0.04, 1.00
```

To feed mean genotype file to BIMBAM , one need to use `-gmode 1` in addition to `-g`.

3.4 Genotype distribution file format

This file is similar to the mean genotype file. The first three columns are identical to those of the mean genotype file. The only difference is that each SNP represented by two adjacent columns instead of one. The first of the two columns denotes the posterior probability of SNP being 0 and the second column for probability being 1. An example of genotype distribution file of two SNPs and three individuals follows.

```
rs1, A, T, 0.98, 0.01, 0.60, 0.38, 0.90, 0.06
rs2, G, C, 0.80, 0.14, 1.00, 0.00, 0.55, 0.20
```

To feed genotype distribution file to BIMBAM , one need to use `-gmode 2` in addition to `-g`.

3.5 Non-genotype input file format

BIMBAM can take non-genotype input files, for example, microarray intensity data. The input files should be prepared in the same format as mean genotype files described above. Users should put the first three column in with arbitrary SNP ID and allele codings using ACGT.

In addition to `-gmode 1` and `-g`, one need to use `--notsnp`, otherwise, BIMBAM will exclude report covariates that has out of range allele frequency.

3.6 Phenotype file format

In the phenotype input file, each line is a number indicating the phenotype value for each individual in turn, in the same order as in the Genotype file. Missing phenotypes should be denoted as NA. The number of lines should be equal to the number of individuals in genotype file (N), otherwise the program will either throw away the values after N or append "NA" at the end to observe N values. In either case, a warning will be printed.

Example Phenotype file with 5 individuals:

```
1.2
NA
2.7
-0.2
3.3
```

If the phenotypes are binary (e.g. in a case-control study) then the format is the same, but each entry should be 0, 1 or NA. It does not matter which group is denoted 0 and which denoted 1.

3.7 Multiple phenotype file format

One can include multiple phenotype in a phenotype file, with each column corresponds to one phenotype, and each row corresponds to an individual. This feature comes handy for microarray expression data when there are many phenotypes. This multiple phenotype feature is only acceptable for single SNP Bayes factor calculations. For this to work, users use `-f` option to specify the number of phenotypes. Example Phenotype file with 5 individuals each with 3 phenotypes:

```
1.2   -0.3   -1.5
NA     1.5    0.3
2.7    1.1    NA
-0.2  -0.7    0.8
3.3    2.4    2.1
```

Note, however, one can not mix the quantitative phenotypes with binary phenotypes in a single file unless one wants to treat binary phenotype as quantitative.

3.8 SNP location file format

The file contains two, or three, columns, with the first column being the SNP name, and the second column being its physical location. The optional but highly recommended third column should contain chromosome number of the SNPs. Note, it is OK if the rows are not ordered according to position, but the file must contain all the SNPs in the genotype files. If the genotype files contain SNPs across different chromosome, BIMBAM will sort SNPs based on its chromosome and position.

Example file:

```
rs1, 1200,  1
rs2, 1000,  1
rs3, 3320,  1
rs4, 5430,  1
```

Note: This file is strictly needed only if the order of the SNPs in the genotype file is not the same as the order of their physical locations along the chromosome, or if multiple genotype and phenotype files are used (see below).

3.9 Use of multiple genotype and phenotype files

In some cases it may be convenient to provide genotypes (and corresponding phenotypes) in multiple files. For example, in a genome-wide study, it may be helpful to have one genotype file containing the HapMap data, and a second genotype file containing the study data. Or, in a candidate gene study where resequencing data are available for a panel of individuals as well as tag-SNP data are available for a study sample, it may be convenient to provide one genotype file for the panel and a second for the tag-SNP data. BIMBAM allows for this use of multiple input files. When using multiple genotype files BIMBAM does not require that the same SNPs be present in both files (although if the same SNP is present in both files then the SNP identifier should be the same in both files, to convey this information). However, to allow for this flexibility, when using multiple genotype files a SNP location file *must* be provided to specify the locations of the SNPs.

When using multiple genotype files, the user must also provide multiple phenotype files, with each phenotype file corresponding to the individuals in a genotype file. The exception to this is that, if all the individuals in a genotype file have no phenotype data available (as might be the case if the genotypes are from the HapMap individuals for example) then this can be specified using `-p 0`. The phenotype files must be specified in the same order as the genotype files to which they correspond.

3.9.1 The strand issue

When merging genotypes from different studies, there arises the issue of whether or not the genotypes for a SNP were obtained on the same strand. In some cases this can be checked easily: for example, if a SNP in one study is A/G, and in the other is T/C, we infer that the two studies used different strands, and we can flip one of the SNPs to correct this. BIMBAM performs these kinds of flip automatically. However, if a SNP is A/T, or C/G, one cannot tell whether the strandedness is the same or different across studies without external information. Currently BIMBAM assumes that genotypes for a single SNP in multiple input files refer to the same strand.

Note: if genotypes at a SNP are not compatible with the SNP being bi-allelic, even after strand flips, then the SNP is considered to be "bad" and BIMBAM will make all the genotypes of that SNP missing.

4 Running BIMBAM : imputation and EM

Some general comments:

1. BIMBAM is a command line based program. The command should be typed in a terminal window, in the directory in which bimbam executable exists.

2. The command line should be all on one line: the line-break in the example is only because the line is too large to fit on one page.

3. Unless otherwise stated, the "options" (`-g -p -pos -o`, etc.) are all case-sensitive.

Imputation usually involves two genotype input files: panel and cohort. Here panel refer to the densely genotyped individuals, e.g. HapMap, 1000 Genomes. A link to files containing panel data, in BIMBAM format, will be updated on the BIMBAM website.

```
./bimbam -g input/panel.txt -p 0 -g input/cohort.txt -p input/pheno.txt -pos
   input/pos.txt  -e 10 -w 20 -s 1 -c 15 --nobf  -o pref -wgd -wmg
```

Numeric 0 after -p denotes the matching genotype file is panel. This command line asks BIMBAM to run EM 10 times, each EM runs 20 steps on panel data alone, and additional 1 step on cohort data. The number of cluster is 15, at the end of the run do not compute Bayes factors. After imputation, output both genotype distribution and mean genotype files, the name of output files start with pref.

One can also perform imputation without panel, when only the LD of the cohort is used to infer the missing genotypes. Note this is not recommended.

```
./bimbam -g input/cohort.txt -p input/pheno.txt   -e 10 -s 20 -c 15
          --nobf  -o pref  -wmg
```

This command line asks BIMBAM to run EM 10 times, each EM run 20 steps. After imputation, output mean genotypes.

4.1 Saving results from EM runs

One can save EM results (`-sem`) and reuse it later with `-rem`, followed by the the name of the file used to store the EM results. For example:

```
./bimbam -g input/cohort.txt -p input/pheno.txt -g input/panel.txt -p 0
    -pos input/pos.txt -o pref1 -sem -i 1

./bimbam -g input/cohort.txt -p input/pheno.txt -g input/panel.txt -p 0
    -pos input/pos.txt -o pref2  -rem output/pref1.em -i 1000
```

The first command line ask bimbam to save EM results, which produce a `pref1.em` file in the `output` directory. The second command line use this EM results to perform important sampling.

Notes:

- When restoring the results of an EM run, the genotype file used must be the same as that used when the results were saved.
- When using `-rem` the user can request further EM iterations to be performed, starting from the saved parameter values, by using the `-s` option. (E.g. `-s 5` would perform an additional 5 iterations for each EM run).
- We do not recommend to use `-sem -rem` anymore, one should save the mean genotype and/or genotype distribution files.

5 Running BIMBAM : computing Bayes factors and p-values

In the software package there are examples files in the `input` directory.

5.1 Calculation of Single-SNP BFs

Here is an example to compute single-SNP Bayes factor:

```
./bimbam -g input/cohort.txt -p input/pheno.txt -pos input/pos.txt -o pref3
```

This command line ask BIMBAM to compute single-SNP Bayes factors using exact genotypes, ignoring the individuals with missing genotypes or phenotypes.

5.2 Calculation of multi-SNP BFs

The `-l` option can be used to instruct BIMBAM to compute multi-SNP BFs for all subsets of up to L SNPs, where L is user-defined. For example:

```
./bimbam -g input/cohort.txt -p input/pheno.txt -pos input/pos.txt -o pref4 -l 3
```

This command line ask BIMBAM to compute multi-SNP BFs for all subsets of size 1, 2 and 3 SNPs (i.e. $L = 3$). Since BIMBAM looks at *all* subsets of size up to L in the multi-SNP BF calculation, this option can be computationally very intensive. We suggest initially using $L = 2$, and, if the results seem interesting, increasing L to 3 or 4.

5.3 Calculation of imputation-based BFs

A natural way to compute imputation based BFs is to perform imputation first, then feed either imputed mean genotype file or imputed genotype distribution file to BIMBAM to compute Bayes factors. BIMBAM provides an integrated approach to perform imputation and computing Bayes factors, either using mean genotype, or sampling genotypes based on genotype distributions. Both of which are invoked using the `-i` option.

The recommended approach, which is invoked by `-i 1`, involves estimating the genotype of each individual by the posterior mean, and then computing a BF for each SNP as if this single estimate were in fact the observed genotype. This approach ignores the uncertainty in the estimated genotype, but it is fast, and in simulation experiments provides results very similar to the conventional approach of averaging over multiple imputations Guan and Stephens (2008). For example,

```
./bimbam -g input/cohort.txt -p input/pheno.txt  -g input/panel.txt -p 0
           -pos input/pos.txt  -o pref5 -i 1 -wmg
```

Note the `-p 0` option to include a "panel" of individuals for whom no phenotype data are available. The above command line also save the imputed mean genotypes, which is highly recommended.

If the user prefers to compute BFs by averaging over multiple imputations, this can be achieved by specifying the number of imputations after the `-i`. However, although this was default behavior in an early release of this software, we no longer recommend this as it is not only very time consuming but, unless the number of imputations is very large, there is a risk that the results may actually be worse than using `-i 1`.

5.4 P-value calculation: `-pval` option

BIMBAM can compute p values assessing the "significance" of observed BFs (see Servin and Stephens, 2007). To invoke this feature, use the `-pval` option, followed by the number of permutations to use. For example,

```
./bimbam -g input/cohort.txt -p input/pheno.txt -o pref6 -pval 10000
```

This command line will compute p-values for each SNP (being the proportion of permutations whose single-SNP BFs for that SNP exceeds that of the observed data) using 10000 random permutations of phenotype. BIMBAM will compute a p-value for the region (being the proportion of permutations whose sum of BF exceeds that of the observed data) as well.

Note: p-value calculations can be very slow, since it multiplies BF calculation times by the number of permutations used (partly because we have not yet taken the smart approach of limiting the number of permutations used for non-significant p values). To speed calculation of p values, *bimbam* computes BFs using a single prior pair $\sigma_a = 0.2, \sigma_d = 0.05$, and expected genotypes, as in the `-i 1` option described above.

5.5 Specify priors on genetic effects: the `-A -D` options

BIMBAM allows user to specify priors for additive effects and dominant effects, or more specifically, to specify values for σ_a and σ_d (see Servin and Stephens). The `-A` and `-D` must be used in pair, and BIMBAM allow multiple usage of `-A -D`, in which case, reported BFs are averages of all prior pairs. For example, to compute BFs by averaging over $(\sigma_a, \sigma_d) = (0.2, 0.1)$ and $(0.1, 0.05)$, one would use

```
./bimbam -g input/cohort.txt -p input/pheno.txt -o pref7 -A 0.2 -D 0.05 -A 0.4 -D 0.1
```

Users are invited to investigate how different σ_a and σ_d affect the Bayes factors. If user chooses not to use `-A -D` options, default values for σ_a, σ_d will be used in BF calculations.

5.6 Combining studies: the `-ssd -psd` options

In some settings, it may be desirable to combine results for multiple studies without sharing individual level genotype and phenotype data. BIMBAM facilitate this by inputting and outputting summary level data that can be shared among investigators and used to perform combined analyses.

To accomplish this, each investigator should first run BIMBAM on their own data using `-psd` option to produce a summary data file. Note if -psd option follow by a string, then the generated SNP summary file will have the string as its name, otherwise, a default name `prefix.ssd` will be used. For example, to produce a summary data file with the name "test.ssd.txt" use:

```
./bimbam -g input/cohort.txt -p input/pheno.txt -o pref8 -psd test.ssd.txt
```

Results from multiple studies can then be combined by running BIMBAM on summary data files. For example, to combine analysis of two studies whose summary data files are test.ssd.1.txt and test.ssd.2.txt, use:

```
./bimbam -ssd input/test1.ssd -ssd input/test2.ssd -o test
```

The file format for the summary data file output by `-psd` and input by `-ssd` is as follows:

```
SNP A1 A2 STRAND ni sg sg2 sgd sd sy sy2 syg syd
rs1162 A G NA 661 550.00 790.00 310.00 310.00 331.00 331.00 319.00 165.00
rs3764 A G NA 662 432.00 566.00 298.00 298.00 331.00 331.00 253.00 161.00
rs1750 C T NA 557 235.00 323.00 147.00 147.00 287.00 287.00 150.00 92.00
rs2215 G A NA 661 276.00 326.00 226.00 226.00 331.00 331.00 117.00 99.00
rs4690 A G NA 662 308.00 384.00 232.00 232.00 331.00 331.00 184.00 136.00
rs1447 C G NA 655 619.00 925.00 313.00 313.00 329.00 329.00 338.00 166.00
```

Each SNP is summarized in a row. The first four columns are SNP id, minor and major allele, and strand information (not in use for the moment). Suppose g_i, y_i are genotype and phenotype of individual i respectively, let $d_i = Pr(g_i = 1)$. From the fifth column on, ni = number of

individuals, $sg = \sum g_i$, $sg2 = \sum g_i^2$, $sgd = \sum g_i d_i$, $sd = \sum d_i$, $sy = \sum y_i$, $sy2 = \sum y_i^2$, $syg = \sum y_i g_i$, $syd = \sum y_i d_i$.

Notes:

1. There are many things to worry about when combining data across studies. e.g., differential recruitment criteria, or systematic DNA genotyping biases. BIMBAM simply analyses all the data as if it came from a single study, so care is required when preparing input files (e.g. phenotype definition) and interpreting results.

2. The information in the SNP summary data file is essentially equivalent to the within genotype class counts, phenotype means and variances (see Guan and Stephens,2008).

5.7 Binary (0/1) hhenotype: the `-cc` option

For binary (case-control) phenotypes, BFs can be calculated with the `-cc` option. For example,

```
./bimbam -g input/cohort.txt -p input/pheno.cc -o pref  -cc

./bimbam -gmode 1 -g case_mgt.txt -p 1 -g ctrl_mgt.txt -p z -pos pos.txt
                  -o pref  -A 0.2 -D 0.05 -cc
```

The first example, pheno.cc contain binary phenotype. In the second example, the `-p 1` assign all individuals in the matching genotype (case_mgt.txt in the example) as 1, and `-p z` assign all individuals in the matching genotype (ctrl_mgt.txt in the example) as 0. Recall `-p 0` denotes the matching genotypes are panel.

Note with `-cc` option BFs are calculated under a logistic regression model, using a Laplace approximation to perform the necessary integration. This is slower than the analytic calculations that can be performed for quantitative phenotypes. In preliminary investigations we have found that treating binary phenotype as quantitative phenotype gives similar results (i.e., with a binary phenotype, the BFs obtained with `-cc` option are similar to without `-cc`). Since the calculations are faster for quantitative phenotypes, a sensible strategy may be to initially perform analyses treating the 0/1 phenotypes as quantitative, and then to follow up interesting regions using the `-cc` option.

6 Output files

BIMBAM will create output files in a directory names `output/`. (If this directory does not exist then it will be created.) Output files will be produced, each with a name beginning with "prefix" that was specified by the `-o` option. We now describe the contents of these output files.

6.1 Log file: `prefix.log`

A log file includes details of the run parameters used and any warnings generated. When sending in a bug report, it is important to include the log file as an attachment.

6.2 Single-SNP Bayes factors: `prefix.single.txt`

This output file contain 10 columns. The first column contains the SNP identifier. The second column contains the physical location of the SNP (or the physical order along the chromosome, if no SNP location file is specified). The third column contains which chromosome the SNP is in. The fourth column is $\log_{10}$ of the single-SNP Bayes factors (averaged over imputations, where these are performed). The fifth column contains the $\log_{10}$ of the standard error of these BFs across the imputations (unless multiple imputation is used, this column is set to `NA`). The sixth column contains the rank of the SNP among all single SNP BFs, if `-sort` is used, otherwise, this column is the physical order along the chromome. The seventh column is p-value for each SNP obtained from the permutation test. (If the `-pval` option is used, otherwise this column becomes `NA`.) The last three columns contain posterior mean of coefficients in Bayesian regression. By default, the rows of these file are sorted according to SNP physical location. To sort by the single-SNP BF values (i.e. highest BF first), use `-sort` when running BIMBAM .

If importance sampling are performed, it is important to check that the standard error of the BFs is small enough that the estimated BFs are reliable. If a SNP has a high BF in the second column, but also a high standard error in the third column, then the high BF may be due to inadequate iterations in the imputation step, and the program should be rerun with more imputations. As a rough guide, we suggest performing more imputations if the $\log_{10}$ standard error (fifth column) is larger than (fourth column-1).

6.3 Single-SNP Bayes factors for binary phenotype

When `-cc` option is used to compute single SNP BFs for binary phenotype, the `prefix.single.txt` changes slightly in that it no longer contains the parameter estimates for μ, a, d.

6.4 Single-SNP Bayes factors for multiple phenotypes

When `-f` option is used to compute single SNP BFs for many phenotypes, the `prefix.single.txt` changes. Suppose there are 3 phenotypes, the output file will contain $5(3+2)$ columns with each row is a SNP. The first column is the SNP ID and the second columns is the SNP location, the rest of the columns are $\log_{10}$BFs of the single SNP BFs for each phenotype.

6.5 Multi-SNP Bayes factors: `prefix.multi.txt`

This file is produced only if the user asks for multi-SNP BFs to be computed (see the `-l` option above). In this file, each SNP is identified by its rank in the single-SNP BF calculations (the 6th column in the single-SNP output file) when `-sort` were used, by default this column is the order of SNP physical location. To make description easier, we use an example output file obtained with the `-l 4` option, which means we calculate up to 4 SNPs combinations.

```
bf       se        snp1    snp2    snp3    snp4
+6.214   +5.207       1      NA      NA      NA
+7.842   +5.734       1      2       NA      NA
......
+0.031   -2.802       16     18      19      20
```

In each row, the first column gives a $\log_{10}$ multi-SNP BF, the second column gives a $\log_{10}$ standard error (NA if not available), and remaining columns identify the combination of SNPs that give rise to that BF. For example,

```
+7.842   +5.734       1       2       NA      NA
```

means that the model with SNPs 1 and 2 having non-zero effect on phenotype has a BF of $10^{7.842}$ compared with the null model of no SNPs having an effect.

Interpreting the results of this file will typically require post-processing (e.g. in R). Some helpful R functions for visualising the results of this file will be made available from the BIMBAM resources site, accessible from `http://stephenslab.uchicago.edu/software.html`.

6.6 Summary of results: `prefix.summary.txt`

This file starts by giving the ($\log_{10}$ of the) overall BF for association between genetic variants in the region and the phenotype, and, if requested, a corresponding permutation-based p value. These should be considered as measures of the evidence against the "global" null hypothesis that there is no association between genetic variation in the region and phenotype; as such they probably only really make sense in a candidate gene study where this might be considered a sensible null.

Note: The overall BF is computed assuming that, under the alternative hypothesis, the prior on the number of SNPs $p(l) \propto 0.5^l$ for $l = 1, \ldots, L$. If $L = 1$ then this is the overall BF computed in the power studies from Servin and Stephens (2007), which should be consulted for more details.

The remainder of the file concentrates on summarising the evidence for *which* variants in the region are associate with phenotype, assuming that the global null is false. So the remainder of the file is generally of interest only if the evidence against the global null is non-negligible.

- The $\log_{10}$(BF) values for l-SNP models, and the posterior probabilities of l under the prior specified above (conditional on $l > 0$). These should be viewed as helping to indicate whether there is evidence for multiple SNPs affecting phenotype in the region.
- A matrix containing 1-SNP and 2-SNP $\log_{10}$(BF) values for the top M SNPs, in order of their physical location. (So the i, jth entry gives the $\log_{10}(BF)$ for the pair of SNPs labelled i and j in the `multi` file; the diagonal entries give the single-SNP BFs).
- The corresonding matrix of posterior probabilities on 1-SNP and 2-SNP models, using the prior $p(l)$ above, conditional on $l \in \{1, 2\}$.

The wordy lines start with ## to ease reading in the statistical package R.

6.7 Output for combined studies `prefix.ssd-bf.txt`

The file contains two columns, the first column is the SNP ID and the second columns is the $\log_{10}(BF)$ of the combined study.

7 Other options

7.1 Restricting the multi-SNP calculations: the -m option

To restrict multi-SNP calculations to only the M SNPs with the largest single-SNP BFs, use the `-m` option.

Example: to compute BFs for all subsets of up to $L = 5$ SNPs, among the $m = 15$ SNPs with the highest single-SNP BFs,

```
./bimbam -g input/cohort.txt -p input/pheno.txt
    -pos input/pos.txt -o test -l 5 -m 15
```

7.2 Restricting analyses to subsets of the data: the -gene and -GF option

In a large study (e.g. a whole-genome scan) one may be iterested in analyzing some subsets of the data (e.g. genes or candidate regions) in detail. BIMBAM allows the user to specify a number of regions for analysis by providing a "gene file". Each line of this file specifies a region to be analyzed, with the first column giving a name for the region, and subsequent columns giving the chromosome number, and the start and end positions:

```
genename1  chr_num1  start_pos1  end_pos1
genename2  chr_num2  start_pos2  end_pos2
...
```

To use this option the user must supply a location file specifying a position for each SNP in the study. Currently the chromosome number is ignored, the regions in a gene file should all be on the same chromosome, and the user must ensure that the genotype data provided are on the same chromosome as the regions specified.

This option is helpful for performing multi-SNP analyses, with or without imputation, of candidate genes (say) in a genome-wide study, without having to develop a separate input file for each candidate gene. When performing such analyses, it may be desirable to include all SNPs within some distance of the gene, rather than only in the gene itself. To do this, the `-GF` option can be used to specify a length of flanking region to include (symmetric, upstream and downstream). This length is subtracted from the start position and added to the end position specified in the gene file.

For example,

```
./bimbam -g input/cohort.txt -p input/pheno.txt -g input/panel.txt -p 0 -pos
   input/test.pos.txt -gene input/genefile.txt -GF 20000 -o test2.out -l 2 -i 1000
```

would perform imputation-based multi-SNP (2-SNP) analysis of each gene in `genefile.txt`, including 20kb upstream and downstream of each gene.

7.3 Genotype data screening

Often it is desirable to exclude SNPs of small minor allele frequency and/or large missingness. BIMBAM provides options `-exclude-maf` and `-exclude-miss` to accommodate such requirements. For example, if one wants to exclude SNPs whose MAF < 0.01 and missing proportion > 0.10 one may use

```
./bimbam -g geno.txt -p pheno.txt -exclude-maf 0.01 -exclude-miss 0.10 -o test
```

One may also choose to exclude certain SNPs by using option `-exclude-nopos`. One can comment out certain SNP positions (by putting # at the beginning of the corresponding lines in the position file), and those SNPs that has no position information will be excluded in the analysis if `-exclude-nopos` is used.

References

Barber, M., L. Mangravite, C. Hyde, D. Chasman, J. Smith, C. McCarty, X. Li, R. Wilke, M. Rieder, P. Williams, P. Ridker, A. Chatterjee, J. Rotter, D. Nickerson, M. Stephens, and R. Krauss (2010). Genome-wide association of lipid-lowering response to statins in combined study populations. *PloS one 5*(3).

Guan, Y. and M. Stephens (2008, 12). Practical issues in imputation-based association mapping. *PLoS Genet 4*(12), e1000279.

Reiner, A. P., M. J. Barber, Y. Guan, P. M. Ridker, L. A. Lange, D. I. Chasman, J. D. Walston, G. M. Cooper, N. S. Jenny, M. J. Rieder, J. P. Durda, J. D. Smith, J. Novembre, R. P. Tracy, J. I. Rotter, M. Stephens, D. A. Nickerson, and R. M. Krauss (2008, May). Polymorphisms of the hnf1a gene encoding hepatocyte nuclear factor-1 alpha are associated with c-reactive protein. *Am J Hum Genet 82*(5), 1193–1201.

Scheet, P. and M. Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet 78*, 629–644.

Seaman, S. R. and S. Richardson (2004). Equivalence of prospective and retrospective models in the bayesian analysis of case-control studies. *Biometrika 91*, 15–25.

Servin, B. and M. Stephens (2007). Efficient multipoint analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics 3*.

Appendix A: Command line examples

- Imputation
  - Imputation with panel. Here panel refer to the densely genotyped individuals, e.g. HapMap, 1000 Genomes.

    ```
    ./bimbam -g input/panel.txt -p 0 -g input/cohort.txt -p input/pheno.txt
      -pos input/pos.txt -e 10 -w 20 -s 1 -c 15 --nobf  -o pref -wgd -wmg
    ```

    Numeric 0 after -p denotes the matching genotype file is panel. This command line asks BIMBAM to run EM 10 times, each EM runs 20 steps on panel data alone, and additional 1 step on cohort data. The number of cluster is 15, at the end of the run do not compute Bayes factors. After imputation, output both genotype distribution and mean genotype files, the name of output files start with pref.
  - Imputation without panel.

    ```
    ./bimbam -g input/cohort.txt -p input/pheno.txt -e 10 -s 20 -c 15 -o pref
           -wmg
    ```

    This command line asks BIMBAM to run EM 10 times, each EM run 20 steps. After imputation, output mean genotypes.
- Single-SNP association
  - Compute single SNP Bayes factor.

    ```
    ./bimbam -g input/cohort.txt -p input/pheno.txt  -o pref -A 0.2 -D 0.05
            -A 0.4 -D 0.1
    ```

    This command line asks BIMBAM to compute Bayes factors with two sets of priors and the output is the average of the Bayes factors obtained with two sets of priors.

    ```
    ./bimbam -gmode 1 -g input/mgt.txt -p input/pheno.txt  -o pref
    ```

    This command line asks BIMBAM to compute Bayes factors with default priors. The input is the mean genotype.
  - Binary phenotypes

    ```
    ./bimbam -gmode 1 -g case_mgt.txt -p 1 -g ctrl_mgt.txt -p z -pos pos.txt\
                    -o pref  -A 0.2 -D 0.05 -cc
    ```

    This command line asks BIMBAM to compute Bayes factors with Laplace approximation. The '-p 1' assign all individuals in the matching genotype (case_mgt.txt in the example) as 1, and '-p z' assign all individuals in the matching genotype (ctrl_mgt.txt in the example) as 0. Recall '-p 0' denotes the matching genotypes are panel.
  - Compute p-values.

    ```
    ./bimbam -gmode 1 -g input/mgt.txt -p input/pheno.txt -o pref -pval 100000
    ```

    This command line asks BIMBAM to compute single SNP Bayes factors, and compute p-values for each SNP via permuation phenotypes 100000 times.

– Importance sampling

```
./bimbam -g panel.txt -p 0 -g cohort.txt -p pheno.txt  -pos pos.txt
          -e 10 -w 20 -s 1 -c 15  -o pref -wgd -i 10000
./bimbam -gmode 2 -g input/gdens.txt -p input/pheno.txt -i 10000 -o pref
```

This first command line asks BIMBAM impute, output genotype distribution, and compute single-SNP BFs by sampling genotypes 10000 times. Note if use '-i 1' then BIMBAM use mean genotypes to compute Bayes factor without sampling. The second command line read in a genotype distribution file and do importance sampling.

- MULTI-SNP SSSOCIATION

```
./bimbam  -g input/cohort.txt -p input/pheno.txt  -o pref  -l 3
```

This command line asks BIMBAM to compute all combinations of SNPs up to and include 3 SNPs, using default priors.

```
./bimbam -gmode 1 -g chr16.txt -p pheno.txt -pos chr16.pos -o pref  -l 3  \
            -gene gene_file.txt -gf 50000
```

This command line asks BIMBAM to compute all combinations of SNPs up to and include 3 SNPs, but only using SNPs that in the region (specified in gene_file.txt) ±50kb. The gene_file.txt may contain multiple entries and BIMBAM will compute those regions separately.

- COMBINE STUDIES USING SNP SUMMARY DATA

– Produce SNP summary data.

```
./bimbam -g genotype.txt -p pheno.txt  -o pref1  -psd
./bimbam -gmode 1 -g mgt.txt -p ph.txt  -o pref2  -psd
```

Both command lines generate SNP summary files, pref1.psd and pref2.psd.

– Use SNP summary data to compute single SNP Bayes factors (-ssd).

```
./bimbam -ssd pref1.psd -ssd pref2.psd -A 0.2 -D 0.05 -A 0.4 -D 0.1 \
            -o combined
```

This command line takes two SNP summary data and compute single SNP Bayes factors using two sets of priors.

Appendix B: BIMBAM Options

Unless otherwise stated, *arg*implies the argument is a string, *num*implies the argument is a number.
FILE I/O RELATED OPTIONS:

- -g *arg* can use multiple times, must pair with -p.
- -p *arg* can use multiple times, must pair with -g. *arg* can be a file name; 0, which indicates the pairing genotypes are panel; z or 1, which indicates the pairing genotype individuals have phenotype 0 or 1.
- -pos *arg* can use multiple times. *arg* is a file name.
- -f *num* specify number of phenotypes (columns in the phenotype files).
- -o *arg* *arg* will be the prefix of all output files, the random seed will be used by default.
- -weg *num* write exact genotype, missing denote by NA. 0 (default value), when BIMBAM write cohort genotype in numerical format; 1, when BIMBAM write cohort genotype in bimbam format.
- -wmg *num* write mean genotype. 0 (default value), write cohort only; 1, write both panel and cohort.
- -wbg *num* write best guess genotype in ACGT+- format. 0 (default value), write cohort only; 1, write both panel and cohort.
- -wgd *num* write genotype distribution, pr(0), pr(1) for each genotype. 0 (default value), write cohort only; 1, write both panel and cohort.

BAYES FACTOR RELATED OPTIONS:

- -a(A) *arg* repeatable, specify priors for additive effect, must pair with -d.
- -d(D) *num* repeatable, specify priors for dominant effect, must pair with -a.
- -df *num* 1, additive effect model; 2 (default), additive and dominance effect model.
- -pval *num* calculate p-values via permuations.
- -sort sort single SNP Bayes factors.
- -cc calc bf of logit regression on binary phenotype.

MULTI-SNP RELATED OPTIONS:

- -i *num* specify number of samplings to compute BF via importance sampling. 0, no imputation; 1, use mean genotype; > 100, importance sampling.
- -m *num* specify number of SNPs for multiple SNP study. The default value is total number (n) of SNPs. If *num*is smaller than n, then SNPs with high single SNP BF will be used.
- -l *num* specify maximum number of SNPs in all combinations. It's l as in lambda.

- -gene *arg* to read gene file that specify regions of interests.
- -gf(GF) *num* pair with -gene to specify gene flanking region in kb.

Combine studies:

- -psd *arg* convert genotype and pheotype to summary statitics and save to a file.
- -ssd *arg* take (multiple) summary data and calculate BF after combining them.

EM related options:

- -e(em) *num* specify number of EM runs, default 10.
- -w(warm) *num* specify steps of warm up EM run, default 10.
- -s(step) *num* specify steps of each EM run, default 1.
- -c *num* specify number of clusters in EM algorithm, default 20.
- -r *num* specify random seed, system time by default.
- -sem *arg* to save EM results, if *arg*is missing prefix.em will be used.
- -rem *arg* to read EM results.

Other options:

- -v(ver) print version and citation
- -h(help) print this help
- -exclude-maf *num* exclude SNPs whose maf ¡ *num*, default 0.01.
- -exclude-miss *num* exclude SNPs whose missing rate ¿ *num*, default 1.
- -exclude-nopos *num* exclude SNPs that has no position information, 1 = yes (default), 0 = no
- –notsnp tell bimbam to allow any numerical values as covariates.
- –nobf tell bimbam not to compute Bayes factors.
- –silence no terminal output.