

# BIMBAM User Manual

Yongtao Guan and Matthew Stephens

February 19, 2009

## Contents

<b>1</b>	<b>What's new in this release version 0.99</b>	<b>3</b>
1.1	New features . . . . .	3
<b>2</b>	<b>Introduction</b>	<b>3</b>
<b>3</b>	<b>Preliminaries: Transforming Quantitative Traits</b>	<b>4</b>
<b>4</b>	<b>Input file format</b>	<b>5</b>
4.1	Genotype file format . . . . .	5
4.1.1	Phased genotype file format . . . . .	6
4.2	<a href="#">Mean genotype and genotype distribution file as input</a> . . . . .	6
4.2.1	<a href="#">Non-genotype input files</a> . . . . .	7
4.3	Phenotype file format . . . . .	7
4.4	SNP location file format . . . . .	7
4.5	Use of multiple genotype and phenotype files . . . . .	8
4.5.1	The strand issue . . . . .	8
<b>5</b>	<b>Running BIMBAM</b>	<b>9</b>
5.1	Calculation of multi-SNP BFs . . . . .	9
5.2	Calculation of imputation-based BFs . . . . .	10
5.3	P-value calculation: <code>-pval</code> option . . . . .	10
5.4	Specify priors: <code>-A -D</code> option . . . . .	11
5.5	Combining studies: <code>-ssd -psd</code> option . . . . .	11
<b>6</b>	<b>Output Files</b>	<b>12</b>
6.1	Single-SNP Bayes Factors: <code>prefix.single.txt</code> . . . . .	12
6.2	Multi-SNP Bayes Factors: <code>prefix.multi.txt</code> . . . . .	13
6.3	Summary of results: <code>prefix.summary.txt</code> . . . . .	13

6.4	Log file: <code>prefix.log</code> . . . . .	14
<b>7</b>	<b>Other Options</b>	<b>14</b>
7.1	Binary (0/1) phenotype: the <code>-cc</code> option . . . . .	14
7.2	Restricting the multi-SNP calculations: the <code>-m</code> option . . . . .	15
7.3	Restricting analyses to subsets of the data: the <code>-gene</code> and <code>-GF</code> option . . . . .	15
7.4	Saving results from EM runs . . . . .	16
7.5	<a href="#">Outputting imputed genotypes</a> . . . . .	16
7.6	<a href="#">Genotype data screening</a> . . . . .	17
7.7	Miscellaneous other options . . . . .	17
<b>8</b>	<b>Parallel computing version (no longer supported in version 0.99)</b>	<b>18</b>

# 1 What's new in this release version 0.99

## 1.1 New features

- Support new input file formats, including inputting posterior genotype distributions and mean genotypes as input genotype files via `gmode` option.
- Support input files that are quantitative variables rather than genotypes, also via `gmode` option. This allows one use BIMBAM to do Bayesian variable selection using more general covariates (eg using gene expression values to predict phenotype).
- New data management functions, excluding SNPs according to missing proportion, minor allele frequency, whether SNPs have entries in position file.
- New options for writing mean genotypes and genotype distributions, so that one can choose to write cohort SNPs only, or write both cohort and panel SNPs.

## 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). New findings regarding practical issues of imputation based association mappings are reported in Guan and Stephens (2008). Please send bug reports and requests for help to [bimbam\\_help@googlegroups.com](mailto:bimbam_help@googlegroups.com). Before sending a request check out [http://groups.google.com/group/bimbam\\_help](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 Guan and Stephens (2008) and Scheet and Stephens (2006). If you use BIMBAM for association mapping, please cite Servin and Stephens (2007). If you use it for both imputation and association mapping, please cite all three papers.

To briefly explain the rationale for imputation-based methods, consider the “tag-SNP” design for association studies, where SNPs are first identified (e.g. by resequencing) in a panel of individuals, and then a subset of these SNPs (“tags”) are typed in the study sample. The imputation-based approach exploits the fact that tag SNPs are often good predictors for the other (non-tag) SNPs, to first “impute” the genotypes of all individuals at all non-tag SNPs, and then assesses the strength of the association between the imputed genotypes and the phenotype. The idea is that this both improves power to detect associations, and interpretability of results (by assessing which SNPs, both tag and non-tag, are the best candidates for causally affecting the phenotype).

Imputation-based methods are also extremely helpful in combining data from multiple studies that have typed different SNPs in the same region (e.g. genome-wide scans using different genotyping platforms). Here, the idea is to use known patterns of correlation among the two sets of markers (e.g. from the HapMap data) to impute genotypes at all markers in all individuals, allowing the data from both studies to be used when assessing correlation between phenotype and each marker.

BIMBAM computes both single-SNP Bayes Factors (BFs) for each SNP, and, optionally, multi-SNP BFs for combinations of SNPs. The latter allows one to assess the potential that multiple SNPs in a data sets are combining to influence phenotype, and is intended for use in small genomic regions (e.g. candidate genes). The imputation is performed using the algorithm used in fastPHASE. Bayes Factors are computed under linear or logistic regression of phenotypes on genotypes. Specifically, for quantitative phenotypes the BFs are computed under the model

$$Y_i = \mu + aX_i + dI(X_i = 1) + \epsilon_i \quad (1)$$

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  $\epsilon_i$  denotes an error term (assumed to be iid normal). The BFs are computed using the prior D2 from [?], averaging over  $\sigma_a = 0.05, 0.1, 0.2, 0.4$  and  $\sigma_d = \sigma_a/4$ .

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

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

The BFs are computed under the same priors for  $\mu$ ,  $a$  and  $d$  as in prior D2 from [?], 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.

### 3 Preliminaries: 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 association 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).

## 4 Input file format

The user 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 analysing each chromosome separately.

Notes on input file conventions:

1. Input files should be saved as plain text (`.txt`) files.
2. Our example files here are all comma-delimited, but space-delimited, tab-delimited, and semi-colon delimited are also fine. (i.e. columns 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.

### 4.1 Genotype file format

Genotypes should be for bi-allelic SNPs, all on the same chromosome. The first two lines should each contain a single number. 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 identifiers for each individual whose genotypes are included: 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 SNPs “name” (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 `??`. Example 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

```

#### 4.1.1 Phased genotype file format

By default BIMBAM assumes that the genotypes in the 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!

## 4.2 Mean genotype and genotype distribution file as input

The `-gmode` option allows BIMBAM can read in mean genotype or genotype distributions files produced by the `-wmg` or `-wgd` options (see *Outputting imputed genotypes* section below for details on these options, and the file formats). Use `-gmode 1` to input mean genotype files and `-gmode 2` for genotype distribution files. Multiple input files are acceptable and will be merged based on the “SNP ID”. However, multiple files should be the same type, either all mean genotypes or all genotype distributions. Here is an example to input feed two mean genotype files into BIMBAM .

```

./bimbam -gmode 1 -g case_genotype.txt -p case_pheno.txt
          -g ctrl_genotype.txt -p ctrl_pheno.txt -o test

```

This option can be helpful if you want to separate out the imputation process from the mapping analysis. It can also be used to apply BIMBAM to perform multi-SNP mapping analyses using genotype imputations produced by other imputation software.

#### 4.2.1 Non-genotype input files

BIMBAM can also be used to perform Bayesian variable selection regression with covariates that are not genotypes (e.g. microarray intensities). The input files should be prepared in the same format as mean genotype files described in the *Outputting imputed genotypes* section.

For non-genotype variables, use both the `-gmode 1` option, and an additional option `-notsnp`. (Without this second option BIMBAM will report an error for an invalid SNP value). Here is an example.

```
./bimbam -gmode 1 -g microarray.dat -p pheno.txt -o test -notsnp
```

### 4.3 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.

### 4.4 SNP location file format

The file contains two columns, with the first column being the SNP name, and the second column being its physical location. 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.

Example file:

```
rs1, 1200  
rs2, 1000
```

rs3, 3320  
rs4, 5430

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).

## 4.5 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.

### 4.5.1 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.

## 5 Running BIMBAM

The simplest way to run BIMBAM is to supply a single genotype file , a phenotype file, and optionally a position file. Examples of these (`test.geno.txt`, `test.pheno.txt`, and `test.pos.txt`) are included in the distribution in the `input` subdirectory. To run BIMBAM on these files, use

```
./bimbam -g input/test.geno.txt -p input/test.pheno.txt  
-pos input/test.pos.txt -o test1.out
```

Notes:

1. The above command line should be typed in a terminal window, in the directory in which `bimbam` executable exists.
2. The user may find the supplied html file, `comgen.html`, helpful to generate command lines. Simply load this file into your browser (e.g. by double-clicking it in most systems), fill in the appropriate boxes, and click "Generate". This will generate a command line which can be cut and paste into a terminal window.
3. The command line should be all on one line: the line-break above is only because the line is too large to fit on one page.
4. The "options" (`-g -p -pos` and `-o`) are all case-sensitive, so you must use `-g` and not `-G`.

BIMBAM will create output files in a directory names `output/`. (If this directory does not exist then it will be created.) Four output files will produced, each with a name beginning with "test1.out", or whatever prefix was specified by the `-o` option. The contents of these output files is described in section 6 below.

There are three main options that the user may wish to add to the above command line. One is to perform multi-SNP analyses allowing for multiple causal variants; another is to use imputation to perform BF computations (e.g. for SNPs that were not typed in the study sample); the final one is to use the BFs to compute  $p$  values (by permutation). These are not performed by default: the user must specifically ask for them to be performed as described below.

Our suggestion is that for genome-wide association studies you might begin by running BIMBAM , either with or without imputation, but using only single-SNP analyses, and then follow-up interesting regions and genes (of size of the order of 100kb) with multi-SNP analyses.

### 5.1 Calculation of multi-SNP BFs

By default BIMBAM will compute only the single-SNP BFs. The `-1` 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. These multi-SNP BFs allow one to assess the evidence for multiple SNPs affecting phenotype (currently assuming effects combine additively across SNPs, with no interactions).

Example: to compute multi-SNP BFs for all subsets of size 1, 2 and 3 SNPs (i.e.  $L = 3$ ) use:

```
./bimbam -g input/test.geno.txt -p input/test.pheno.txt
        -pos input/test.pos.txt -o test2.out -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. (See also the `-m` option below to restrict the exhaustive search to only those SNPs with a high single-SNP BF.)

## 5.2 Calculation of imputation-based BFs

By default BIMBAM does not perform imputation: it computes the single-SNP and multi-SNP BFs using only those individuals with phenotype data and complete genotype data for the relevant SNPs. To perform imputation, the user must use the `-i` option, as we now describe.

BIMBAM has two ways for performing imputation, 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, submitted).

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`.

The following example illustrates how to compute BFs using the recommended approach to imputation, and also the use of multiple input files, and the `-p 0` option to include a “panel” of individuals for whom no phenotype data are available.

```
./bimbam -g input/test.geno.txt -p input/test.pheno.txt
        -g input/test.panel.txt -p 0 -pos input/test.pos.txt
        -o test3.out -i 1
```

One potential set of “panel” data that we anticipate many users will want to use are the data from the International HapMap project. A link to files containing these data (phased, for unrelated individuals), in BIMBAM format, will be made available from the BIMBAM resources website, accessible from <http://stephenslab.uchicago.edu/software.html>.

## 5.3 P-value calculation: `-pval` option

BIMBAM can compute  $p$  values assessing the “significance” of observed BFs (see [?]). To invoke this feature, use the `-pval` option, followed by the number of permutations to use. For example, using `-pval 1000` will compute  $p$ -values using 1000 random permutations.

BIMBAM will compute a  $p$ -value for the region (being the proportion of permutations whose overall BF exceeds that of the observed data) and also for each SNP (being the proportion of permutations whose single-SNP BFs for that SNP exceeds that of the observed data).

**Note:**  $p$ -value calculations can be very slow if  $M$  and/or  $L$  are large, 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.

#### 5.4 Specify priors: `-A -D` option

BIMBAM allows user to specify priors for additive effects and dominant effects, or more specifically, to specify values for  $\sigma_a$  and  $\sigma_d$  (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/test.geno.txt -p input/test.pheno.txt -A 0.2 -D 0.1 -A 0.1 -D 0.05
```

If user chooses not to use `-A -D` options, default values for  $\sigma_a, \sigma_d$  will be used in BF calculations.

#### 5.5 Combining studies: `-ssd -psd` option

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.snp.summary.data.ssd` will be used. For example, to produce a summary data file with the name “test.ssd.txt” use:

```
./bimbam -g input/test.geno.txt -p input/test.pheno.txt -o test4.out -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/test.ssd.1.txt -ssd input/test.ssd.2.txt -o test4.out
```

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

The information in this file is essentially equivalent to the within genotype class counts, phenotype means and variances.

## 6 Output Files

The program generates four output files as follows:

### 6.1 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 imputations 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).

## 6.2 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 `-1` 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 `-1 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
+4.149	+3.811	2	NA	NA	NA
.....					
+7.842	+5.734	1	2	NA	NA
+5.729	+4.205	1	3	NA	NA
.....					
+0.031	-2.802	16	18	19	20
+0.025	-1.327	17	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>.

## 6.3 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  $l = 1, \dots, L$ . If  $L = 1$  then this is the overall BF computed in the power studies from [?], which should be consulted for more details.

The remainder of the file concentrates on summarising the evidence for *which* variants in the region are associated 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.

It contains the following information, which essentially summarises the results in the `prefix.multi.txt` file described above.

- The  $\log_{10}(\text{BF})$  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  $\log_{10}(\text{BF})$  values for the top  $M$  SNPs, in order of their physical location. (So the  $i, j$ th entry gives the  $\log_{10}(\text{BF})$  for the pair of SNPs labelled  $i$  and  $j$  in the `multi` file; the diagonal entries give the single-SNP BFs).
- The corresponding 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.

## 6.4 Log file: `prefix.log`

The last file is a log file, and includes details of the run parameters used and any warnings generated.

# 7 Other Options

## 7.1 Binary (0/1) phenotype: the `-cc` option

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

## 7.2 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/test.geno.txt -p input/test.pheno.txt
        -pos input/test.pos.txt -o test2.out -l 5 -m 15
```

## 7.3 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 interested in analysing 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 analysed, 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/test.geno.txt -p input/test.pheno.txt -g input/test.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.

## 7.4 Saving results from EM runs

One of the more time-consuming aspects of running BIMBAM with imputation is the initial fitting of the model used to perform imputation. This fitting is performed using the EM algorithm. To save time in future runs the results of this fitting can be stored in a file, using the `-sem` option. The file used to store the results will be called `outputprefix.em`.

To restore the results of a previous EM run, use the `-rem` option, followed by the the name of the file used to store the EM results.

Example:

```
./bimbam -g input/test.geno.txt -p input/test.pheno.txt -g input/test.panel.txt -p 0  
-pos input/test.pos.txt -o test.i1 -sem -i 1
```

```
./bimbam -g input/test.geno.txt -p input/test.pheno.txt -g input/test.panel.txt -p 0  
-pos input/test.pos.txt -o test.i1000 -rem output/test.i1.em -i 1000
```

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.
- The use of `-sem` with the `-gene` option is not recommended.
- 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).

## 7.5 Outputting imputed genotypes

If you want BIMBAM to output a summary of the imputed genotypes, use either `-wmg` or `-wgd`.

The option `-wmg` will cause BIMBAM to output the posterior mean for each genotype in an additional output file `prefix.mean.genotype.txt`. This file has a *different* form from the genotype input file. There is 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 remaining columns are the mean genotypes of different individuals – numbers between 0 and 2 that represents the (posterior) mean genotype, where genotypes 0,1 and 2 denote the number of copies of the minor allele (where the minor allele at each SNP is determined from those individuals for whom genotype data are available). An example of mean genotypes file with two SNPs and three individuals follows.

```
rs1, A, T, 0.02, 1.80, 1.50  
rs2, G, C, 1.98, 0.04, 1.00
```

The `-wgd` option will cause BIMBAM to output the posterior probabilities for each genotype in an additional output file `prefix.genotype.distribution.txt`. The format of this file is similar to the mean genotype file. The only difference is that each SNP is represented by two adjacent columns instead of one. The first of the two columns denotes the posterior probability of the genotype being 0 and the second column denotes the probability of being 1. (Of course the probability of the genotype being 2 is obtained by 1 minus the sum of these two numbers.) 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
```

For both `-wmg` and `-wgd` options, one may choose to append a number chosen from  $\{0, 1\}$ . If 0 is appended after the options (e.g. `-wmg 0`), BIMBAM will output SNPs that are only in cohort, while if 1 is appended, BIMBAM will output SNPs that are in either cohort or panel.

BIMBAM provides an option to output “best guess” genotypes with `-wbg` option. This will produce an additional output file `prefix.best.guess.genotype.txt`. The best guess genotype is the genotype that has maximum posterior probability. However, the use of best guess genotypes for association studies is not recommended. Both mean genotype and best guess genotype are obtained from genotype distributions that are averaged over multiple runs of the EM algorithm. The best guess genotypes are coded in 0, 1, 2 and corresponding allele coding (A,C,G,T,+,-) can be found in `prefix.log.txt`.

## 7.6 Genotype data screening

In some cases it may be desired 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.

## 7.7 Miscellaneous other options

A summary of other available options (e.g. controlling number of iterations in the EM algorithm; setting the seed for the random number generator) can be obtained by using the `-h` or `-help` option:

```
./bimbam -help
```

## 8 Parallel computing version (no longer supported in version 0.99)

BIMBAM can take advantage of parallel computing, specifically the MPI package, to substantially reduce run-times. If one have cluster access, the MPI BIMBAM can speed up calculations dramatically, especially for large scale genome wide studies. Even for a single computer, the MPI version of BIMBAM can take advantage of multiple-core and multiple CPUs to speed up computations.

To run the MPI version of BIMBAM, one needs first to install two freely-available packages: openMPI and the GSL (gnu scientific library). If you are running BIMBAM on a cluster, then contact the systems administrator to help with this. If you are running BIMBAM on a multi-core or multi-processor desktop, then you may be able to manage this yourself: we give brief instructions, and encourage you to contact a local expert if you need further help.

To install MPI:

- Download openMPI from <http://www.open-mpi.org/software/ompi/v1.2/>.
- Build and install MPI libraries.

```
shell$ gunzip -c openmpi-1.2.3.tar.gz | tar xf -
shell$ cd openmpi-1.2.3
shell$ ./configure --enable-static --prefix=/path/to/mpi/
<...lots of output...>
shell$ make all install
```

Note if one doesn't supply `--prefix` option in the `./configure` line then MPI will be installed in `/usr/local/` directory.

To install GSL:

1. Download the GSL from <ftp://ftp.gnu.org/gnu/gsl/>. Scroll down to the bottom of the page to find the most recent version (at time of writing, this is v1.9).
2. Unzip and untar it as above. Then install iusing

```
./configure
make
sudo make install
```

By default, the library will be installed in `/usr/local/lib`.

After installation of MPI and GSL, download BIMBAM source code and unzip untar it. Then one needs to make minor changes in the Makefile in the `/src` directory (not the simple one in mother directory of `/src`). Here is what top part of Makefile look like.

```

*****
#       user configuration
# set mpi = yes for the parallel version
MPI ?= yes
MPICC = /path/to/mpi/mpic++
#
# set debug = yes to debug
DEBUG ?= no
#
# set readline = yes for additional interactive feature
# note: you must set MPI ?= no to use interactive mode;
READLINE ?= no
#
# set impute = yes for mask/imputation feature
IMPUTE ?= no
#
# end user configuration;
*****

```

One needs to make sure the lines `MPI ?= yes`, and `MPICC = /path/to/mpi/mpic++` are correct. If not make corresponding changes. To compile use

```

shell$ make clean
shell$ make

```

If you are running BIMBAM on a single multi-core or multi-processor desktop, then it is easy to run MPI executable: use

```

mpirun -np 8 ./bimbam ...

```

where here `-np` stands for number of processes, and we recommend to be set to the number of cores you wish BIMBAM to use, although over-subscribe is possible.

If you are running BIMBAM on a cluster, then you will need to create a machine file to tell BIMBAM which machines to use. This specification is system-dependant, depending on use of `qsub` and other issues: contact your systems administrator for help.