BFMAP by Jicai Jiang

Bayesian Fine-Mapping and Association for Population and Pedigree Data

BFMAP is a software tool for genomic analysis of quantitative traits, with a focus on fine-mapping, SNP-set association, and functional enrichment. It uses phenotypes and genotypes, and works for samples with population structure and/or relatedness. BFMAP currently supports the following analyses.

- Estimating SNP-heritability;
- Genome-wide single-marker/SNP-set association analysis;
- Fine-mapping by forward selection or shotgun stochastic search;
- Estimating causal-variant enrichment of a functional annotation (i.e., functional enrichment analysis);
- Incorporating functional enrichment into fine-mapping.

Installation

BFMAP is statically compiled with the Intel Math Kernel library for the Unix/Linux environment. After download the static executable, be sure to make the file executable.

Options on the command line can be recognized so long as they provide a **prefix** of the option name that matches exactly one of the accepted options. For example, the option `--trait` will match as `--trait_name`, but `--covariate` (either `--covariate_file` or `--covariate_names`) will not match uniquely, so an error will be raised.

Phenotype File

```
--phenotype_file <CSV file> --trait_name <column header>
```

```
--phenotype_file <CSV file> --trait_name <column header> --error_weight_name <column
header>
```

The phenotype file is a comma-delimited values (CSV) file with the first row being header. The first column must be individual ID, and the following columns are phenotypes (one trait per column) or individual-specific weights for error variance (one trait per column). One can use column header to specify trait and error weight.

Leave missing values empty in the phenotype file. Individuals with missing phenotype or missing error weight (if specified) will not be used in analysis.

When the error weight option is not specified, the weights will be all set to 1.

Genotype Files

PLINK Binary File

BFMAP uses PLINK binary files (***bed***/***bim***/***fam***) for genotypes. Refer to the PLINK website for description of the file formats. Note that BFMAP only uses within-family ID in ***fam*** file.

BFMAP Binary File

BFMAP also uses a simple binary file (with extension ***bin***) to store individual-major genotype data, in which each genotype is stored in a byte. The ***bin*** file is accompanied by ***indi*** and ***mrk*** text files. The ***indi*** file lists all the individuals in the binary file in the same order. The ***mrk*** file lists all the SNP markers in the binary file in the same order, of which the five columns are SNP ID, chromosome and physical position, allele 1, and allele 2, respectively. Neither of the files has header.

BFMAP has an option to convert a CSV genotype file to ***bin***, ***indi*** and ***mrk*** files, as shown below. Three files, foo.bin, foo.indi, and foo.mrk, would be generated.

```
./bfmap --csv_genotype_file foo.csv --binary_genotype_file foo
```

In the CSV genotype file, the first five rows list SNP IDs, SNP chromosomes, physical positions, allele 1, and allele 2, respectively, and the first column lists individual IDs. When the CSV genotype file option is used, BFMAP will do file conversion only. To do other analyses, the binary genotype file option must be given.

Option to Read Genotype Files

```
--binary_genotype_file <filename prefix>
```

With this option, BFMAP will first search for BFMAP binary files (bin, indi, and mrk). If they are not found, it will further search for PLINK binary files (bed, bim, and fam).

BFMAP assumes that there is no missing genotype and that SNPs have been sorted by physical positions.

Covariate File

```
--covariate_file <CSV file>
```

```
--covariate_file <CSV file> --covariate_names all
```

```
--covariate_file <CSV file> --covariate_names <covar1>,<covar2>,<covar3>
```

The covariate file is a CSV file with the first row being header. The first column must be individual ID, and the following columns are covariates (one per column). When one wants to use all the covariates in the covariate file, typing ***all*** is sufficient.

If the covariate file option is not specified, BFMAP will search for covariates in the phenotype file. In such case, do not use the key word ***all***. If the covariate names option is not specified, BFMAP will automatically use only intercept. However, if the covariate names option is specified, BFMAP will not intentionally add intercept into covariates, so one has to put a column of **1**'s for intercept in the covariate file.

Leave missing values empty in the covariate file. Individuals with any missing value for the specified covariates will not be used in analysis.

SNP Info File

```
--snp_info_file <CSV file>
```

```
--snp_info_file <CSV file> --snp_set_name <column header>
```

```
--snp_info_file <CSV file> --snp_weight_name <column header>
```

```
--snp_info_file <CSV file> --snp_set_name <column header> --snp_weight_name <column
header>
```

The SNP info file is a CSV file with the first row being header. The first column must be SNP ID. There may be additional columns specifying SNP sets or SNP weights for variance of effects.

When the SNP set option (**--snp_set_name**) is not specified, all SNPs will be considered in one set called ***NULL***. When the SNP weight option (**--snp_weight_name**) is not given, all SNPs will be given a weight of **1**. Leave missing values empty in the SNP info file. Individuals with missing SNP set or missing SNP weight (if specified) will not be used in analysis.

BFMAP only uses the SNPs listed in the SNP info file, so that one can easily specify which SNPs to be used in fine-mapping or association tests without changing big genotype file.

MAF-based SNP weights

```
--beta_weight <a,b>
```

In this option, *a* and *b* are two parameters of the beta distribution. The weights are computed exactly using the probability density function of the beta distribution. For example, one may use **--beta_weight 1,10** to assign more weight to rare variants. The default is 1,1 (i.e., $a = b = 1$).

BFMAP automatically replaces all SNP weights specified by **--snp_weight_name** with the beta weights.

Note that the weights are assigned to unscaled genotypes (i.e., 0, 1, and 2).

Analysis Set

To generate the set of subjects used for analysis, BFMAP takes only the individuals whose phenotype, genotypes, and covariates (if specified) are all present. One can edit the CSV phenotype file to easily control the analysis set.

Filtering Options

BFMAP can filter SNPs by minor allele frequency (MAF) or Hardy-Weinberg equilibrium (HWE) exact test.

```
--min_maf <maf> --min_hwe_pval <p-value> --midp
```

--min_maf filters out all variants that have MAF smaller than or equal to the provided threshold. The default value is 0.

--min_hwe_pval filters out all variants which have HWE exact test *p*-value below the provided threshold. The default value is 0. **--midp** is optional and enables a mid-p adjustment.

These filtering options work whenever a binary genotype file is read.

Multithreading

```
--num_threads <num>
```

BFMAP can use multiple threads to speed up for computing GRM, estimating heritability, fine-mapping, and association tests.

Genomic Relationship Matrix (GRM)

Computing GRMs

```
./bfmap --compute_grm <1|2> --binary_genotype_file <filename prefix> --snp_info_file
<CSV file> --output_file <GRM filename prefix> --subject_set <text file>
```

--compute_grm computes a GRM given genotypes. One needs to choose between two GRM forms: **1** and **2** refer to equations 1 and 2, respectively. In the equations, *Z* represents centered (but not scaled) genotypes, and *p* and *q* represent the frequencies of two alleles at a biallelic locus.

$$G = \frac{ZZ'}{\sum_{i=1}^{m} 2p_i q_i} \tag{1}$$

$$G = Z \begin{pmatrix} 2p_1 q_1 & 0 & \dots & 0 \\ 0 & 2p_2 q_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 2p_m q_m \end{pmatrix} Z' \tag{2}$$

--snp_info_file is required and specifies which SNPs to be used to compute GRM. Currently, SNP weighting by **--snp_weight_name** is not supported for computing GRM.

BFMAP generates a binary file (**.grm.bin**) and a single-column text file (**.grm.indi**) to save a GRM. The binary file stores GRM elements with double-precision, floating-point numbers. The text file lists all the subjects in the GRM. **--output_file** specifies the filename prefix of the two files.

--subject_set <text file> is optional and specifies a single-column text file with no header that will control the individuals included in the analysis. If this option is not present, all individuals in the genotype file will be included in computation.

Combining GRMs

```
./bfmap --combine_grms <text file> --output_file <GRM filename prefix> --subject_set
<text file>
```

This command combines GRMs, which is useful for leave-one-chromosome-out (LOCO) association tests.

--combine_grms <text file> specifies a single-column text file with no header which lists the filename prefixes of GRMs to be combined. One GRM per line.

--output_file <GRM filename prefix> sets the filename of the combined GRM.

--subject_set <text file> is the same as described above.

Note that the GRMs to be combined should be based on the same equation (either 1 or 2).

--subject_set <text file> is available only for computing a GRM and combining GRMs.

Estimating SNP-Heritability

BFMAP uses an eigendecomposition approach to estimate SNP-heritability, like EMMA. However, σ_e^2 (the error variance) in the likelihood function is integrated out in BFMAP rather than treated as a parameter of interest in EMMA. As a result, EMMA reports the estimates of σ_e^2 and σ_g^2 (the variance explained by SNPs), while BFMAP only reports the estimate of their ratio (σ_g^2/σ_e^2).

In addition, BFMAP can take individual-specific weights for error variance by **--error_weight_name**. This feature is useful in animal genetics studies where phenotypes are often breeding values and of varied reliability. In such a scenario, what BFMAP gets is not really SNP-heritability.

```
./bfmap --varcomp --phenotype <CSV file> --trait <column header> --binary_grm_file <GRM
filename prefix> --output <filename>
```

```
./bfmap --varcomp --phenotype <CSV file> --trait <column header> --error_weight_name
<column header> --binary_grm_file <GRM filename prefix> --output <filename>
```

--varcomp initializes the estimation.

--binary_grm_file <GRM filename prefix> sets a GRM file.

--output <filename> writes output into a CSV file. Below is an example. The first line shows σ_g^2/σ_e^2, and the second line shows $\sigma_g^2/(\sigma_e^2+\sigma_g^2)$ (SNP-heritability or the proportion of variance explained by SNPs). The next two lines show the likelihood-ratio test for $H_0: \sigma_g^2/\sigma_e^2 = 0$.

```
variance ratio,3.67805
proportion,0.786236
LLR test statistic,31537.2
LLR p-value,0
```

The SNP-heritability estimate may be used for the **--heritability** option in fine-mapping.

Forward Selection for Fine-Mapping

In this approach, BFMAP first identifies independent association signals within a genomic region by forward selection. For each independent signal, BFMAP determines an inclusive list of variants (which are supposed to include the causal one), assigns a posterior probability of causality (PPC) to each variant, and then sorts the variants in a descending order of PPC. BFMAP subsequently selects top *n* variants so that the sum of their PPC exceeds a given threshold (e.g., 95%). Accordingly, a credible set is formed by the selected variants for each independent association signal. If the threshold is set to 95%, the resulting credible set is >95% likely to contain the causal variant.

Running Forward Selection

The following command works for unrelated samples.

```
./bfmap --phenotype_file <CSV file> --trait_name <column header> --snp_info_file <CSV
file> --binary_genotype_file <filename prefix> --output <filename>
```

One may include a GRM in BFMAP by the following options to account for population structure and relatedness in fine-mapping.

```
--binary_grm_file <GRM filename prefix> --heritability <estimate>
```

The SNP-heritability set by **--heritability** must correspond with the GRM set by **--binary_grm_file**. One should first estimate the heritability (by **--varcomp**) and then set it as the argument of **--heritability**. Note that **--binary_genotype_file** is present in fine-mapping, while it is not in SNP-heritability estimation. As in SNP-heritability estimation, BFMAP can take individual-specific weights for error variance by **--error_weight_name** in fine-mapping.

```
--error_weight_name <column header>
```

If the option is present in fine-mapping, the **--heritability** argument must be from the heritability estimation in which **--error_weight_name** is also present.

Options Specific to Forward Selection

Shown below are additional options that are specific to the forward-selection fine-mapping. To better understand what these options are meant for, refer to the beginning of this section.

```
--meff <num> --prob_threshold <probability> --prob_min_ld_r2 <LD r-squared>
```

--meff <num> sets the effective number of SNPs. The number is used to determine the stopping criterion of forward selection. If the option is not present, BFMAP will estimate the number using the method by Li & Ji (2005). To make the stopping criterion more stringent, one may set the argument to the actual number of SNPs or larger ones.

--prob_threshold <probability> sets the threshold of posterior probability of causality for a credible set. For example, an argument of 0.95 results in a credible set of variants which are >95% likely to contain the causal one. The default value is 0.95.

--prob_min_ld_r2 <LD r-squared> sets a threshold of LD r^2 (the square of the correlation coefficient) for determining an inclusive list of variants for each independent association signal. A variant is included in the list, only if the r^2 between the variant and the lead variant exceeds the threshold. The default value is 0.3. A smaller value (like 0.25 or 0.2) may also work well.

Output File

BFMAP generates a CSV output file for the forward-selection fine-mapping. There are 20 columns in the CSV output file, shown in the table below.

Column header	Description
signal	An integer starting from 0 for specifying an association signal
SNPindex	An integer starting from 0 for specifying a variant in an association signal
SNPname	Variant ID
Chr	Chromosome
Pos	Physical position

Column header	Description
Allele1	Allele 1
Allele2	Allele 2
MAF	Minor allele frequency
HWE_Pval	*P*-value for HWE test
sample_size	Sample size
effect	Effect size estimate of a variant conditional on other association signals
log_sBF	Natural log of scaled Bayes factor (H_0: The tested variant has no effect.)
lambda	A coefficient determining the null distribution of scaled Bayes factor (=1)
Pval	*P*-value corresponding to H_0
logProb	Natural log of *P*(Data|Model)
penalty	Natural log of *P*(Model), equal to 0 for equal prior of models
penalized_logProb	Natural log of posterior probability
rel_logProb	Penalized_logProb of a model relative to the variants-excluded model
normedProb	Posterior probability of causality between 0 and 1
R	Genotype correlation between a variant and the lead variant

---

Shotgun Stochastic Search (SSS) for Fine-Mapping

A small number of variants within a region can form a large number of models, which often results in infeasible computation. Only a limited number of high-probability models, however, matter for computation of variant inclusion probabilities. Shotgun stochastic search works well for fine-mapping, because it can efficiently identify (relatively-)high-probability models in high-dimensional model space.

Running SSS

--sss enables shotgun stochastic search in BFMAP. **--sss** works for unrelated samples as well as samples with population relatedness, with the same usage as forward selection. For example, the following command works for samples with population relatedness.

```
./bfmap --sss --phenotype_file <CSV file> --trait_name <column header> --snp_info_file
<CSV file> --binary_genotype_file <filename prefix> --binary_grm <GRM filename prefix> -
-heritability <estimate> --output <filename prefix>
```

Shown below are additional options for **--sss**.

```
--variance_ratio <number> --beta_weight <a,b>
```

--variance_ratio <number> sets a common prior of effect size for all variants. It is the ratio of variance of a variant to the error variance (σ_e^2). The default value is 0.01.

--beta_weight sets MAF-based SNP weights. Note that weights can also be specified by SNP Info File.

Let γ and w denote the variance ratio and weight for a variant, respectively. $\gamma w \sigma_e^2$ is thus the actual variance assigned to a variant. Since $\sigma_p^2 = \sigma_g^2 + \sigma_e^2 = \sigma_e^2/(1-h^2)$, $\gamma w(1-h^2)$ is the proportion of phenotypic variance explained by a variant. When $w = 1$, the default of **--variance_ratio** (i.e., $\gamma = 0.01$) actually assigns a moderate prior effect size to variants for a moderate-heritability trait.

Options Specific to SSS

```
--sa_initial <num> --per_temp_iterations <num> --num_iterations <num>
```

BFMAP uses simulated annealing with a linear cooling scheme to improve shotgun stochastic search. In the linear cooling scheme, $T_{k+1} = T_k - \Delta T$, ΔT is fixed to 1, and the final temperature (T_f) is fixed to 1.

--sa_initial <num> sets the initial temperature (T_0). The default is 100.

--per_temp_iterations <num> sets the number of SSS iterations at each temperature larger than T_f. The default is 10.

--num_iterations <num> sets the number of SSS iterations at T_f. The default is 100.

With the default setting, BFMAP will run a total of 1090 SSS iterations. One can set **--sa_initial 1** not to use annealing.

```
--max_num_causals <num> --prior_num_causals <num>
```

--max_num_causals <num> sets the maximum number of causal variants which limits model size. The default is 5.

--prior_num_causals <num> sets a prior of number of causal variants. The default is 1.

```
--max_logbf_difference <num> --min_model_prob_output <num>
```

--max_logbf_difference <num> sets a threshold of $log(BF)$ relative to the largest $log(BF)$. If the difference of $log(BF)$ between the best model and a model exceeds the threshold, the model will be disregarded when computing variant inclusion probability. The default is 30.

--min_model_prob_output <num> sets a threshold of posterior model probability. Models that have a posterior model probability smaller than the threshold are not written to output file. The default is 1E-8.

Output Files

BFMAP generates two CSV output files for the SSS fine-mapping. One has a filename suffix of **model.csv**, and the other has a filename suffix of **pip.csv**.

The **model.csv** file lists models and has $k+2$ columns, where k is the number of putative causal variants in the largest model. In each row, the first k columns show what variants a model consists of, and the penultimate column and the last column show $log(BF)$ and posterior model probability of the model, respectively. Many models have less than k variants, for which **NA** is used to fill up.

The **pip.csv** file lists variants and shows their posterior inclusion probability (PIP). Column headers have the same description as in the forward selection output.

SNP-Set Association

--assoc enables single-marker/SNP-set association tests in BFMAP. **--assoc** works for unrelated samples as well as samples with population relatedness, with the same usage as forward selection.

It is critical to use SNP Info File and **--snp_set_name** to specify SNP-sets of interest in association tests, while **--snp_set_name** is not applicable to fine-mapping. When **--snp_set_name** is not present, all SNPs will be considered in one set called ***NULL***. In such a case, BFMAP will do an association test for ***NULL***. If this set is very large, an memory error may occur.

Single-marker association is a special case of SNP-set association. Note that the first column is variant ID in SNP Info File. So one can specify **--snp_set_name <the header of the first column>** to enable single-marker association tests.

Additionally, it is critical to use SNP-specific weights in SNP-set association. One can specify MAF-based SNP weights by **--beta_weight** , or use previously-obtained weights in SNP Info File by **--snp_weight_name**. Use of SNP weights should be coupled with **--variance_ratio <number>**, as described in Running SSS.

The following command works for samples with population relatedness.

```
./bfmap --assoc --phenotype_file <CSV file> --trait_name <column header> --snp_info_file
<CSV file> --snp_set_name <column header> --beta_weight <a,b> --variance_ratio <number>
--binary_genotype_file <filename prefix> --binary_grm <GRM filename prefix> --
heritability <estimate> --output <filename>
```

BFMAP generates a CSV output file for SNP-set association. There are seven columns in the CSV output file, shown in the table below.

Column header	Description
SNPset	SNP-set ID as in the SNP info file
size	SNP-set size, i.e., number of variants
logBF	Natural log of Bayes factor, $log(BF)$ (H_0: The tested SNP-set has no effect.)
logBFMean	Mean of the null distribution of $log(BF)$

Column header	Description
logsBF	Natural log of scaled Bayes factor
Pval	*P*-value corresponding to H_0
lambda	Coefficients determining the null distribution of $log(BF)$

Functional Enrichment

Functional enrichment is currently implemented in two R scripts. **sss_annot.R** estimates causal-variant enrichment of a functional annotation. **update_sss.R** incorporates functional enrichment into fine-mapping.

Estimating Functional Enrichment

```
Rscript sss_annot.R <model-file> <variant-annotation-file> <annotation-name> <category-prob-file> <cumulative-model-prob-cutoff> <output-filename>
```

<model-file> is a tab-delimited text file created from fine-mapping output files. This model file should contain multiple loci. It can be generated by merging multiple SSS model files. Its first column lists locus ID to indicate what the model in a line comes from, and the rest columns are the same as in an SSS model file.

<variant-annotation-file> is a tab-delimited text file containing functional annotations of variants. The first row lists column headers. The first column must be variant ID, and the following columns show functional annotations.

<annotation-name> specifies the functional annotation to be analyzed. It must be a column header in **<variant-annotation-file>**.

<category-prob-file> is a tab-delimited text file containing three columns. The first row is header. The first column lists categories of the functional annotation in the analysis. The second column lists starting value for the probability of a causal variant being in each category. The last column lists the probability of a non-causal variant being in each category, which can be easily estimated by the whole-genome frequency. Basically, this analysis aims to estimate the second column.

<cumulative-model-prob-cutoff> sets a threshold of cumulative model probability. Low-probability models are disregarded in the analysis. Generally, 0.9 works well.

<output-filename> specifies a output filename. The output file is the same as **<category-prob-file>** except that its second column has been updated.

Incorporating Functional Enrichment Into Fine-Mapping

```
Rscript update_sss.R <model-file> <variant-annotation-file> <annotation-name> <category-prob-file> <cumulative-model-prob-cutoff> <output-filename>
```

All the arguments in the command above are the same as those for **sss_annot.R**. Note, however, that **<category-prob-file>** for **update_sss.R** is actually the output file of **sss_annot.R**.

In addition, **<cumulative-model-prob-cutoff>** had better be 1.

Two output files are generated. One is updated model file, and the other is updated PIP file. The new PIP file lists posterior inclusion probabilities of variants with incorporation of a functional annotation.

References

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, *95*(3), 221.