BFMAP by Jicai Jiang

Bayesian Fine-Mapping and Association for Population and Pedigree Data

Installation

BFMAP is statically compiled with the Intel Math Kernel library for the Unix/Linux environment.

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 effect size.

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.

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)

```
./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.

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

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

--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 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.

```
./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 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 sets the filename of the combined GRM.

--subject_set is the same as described above.

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

--subject_set 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 <GRM
filename prefix> --output <filename>
```

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

--varcomp initializes the estimation.

--output 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

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

The command above works for unrelated samples. 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 <number>
```

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.

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

As in SNP-heritability estimation, BFMAP can take individual-specific weights for error variance by **--error_weight_name** in fine-mapping. If the option is used in fine-mapping, the **--heritability** argument must be from the heritability estimation in which **--error_weight_name** is also present.

```
--prob_threshold 0.95 --prob_min_ld_r2 0.3 --repos_min_ld_r2 0.1
```

Shotgun Stochastic Search for Fine-Mapping

SNP-Set Association

Functional Enrichment

Example Commands

- Converting a CSV Genotype File to Binary
- GRM Calculation

```
./bfmap --compute_grm --binary_genotype_file geno --snp_info_file chr1_snps.csv --
output chr1
[--subject_set subject_set.csv]
[--min_maf 0.01 --min_hwe_pval 1e-6 --midp]
```

- Combining GRMs

```
./bfmap --combine_grms grm_list.txt --output autosome
[--subject_set subject_set.csv]
```

- Estimating Heritability

```
./bfmap --varcomp
```

- Fine-Mapping (Forward Selection)

```
./bfmap --phenotype_file phen.csv --trait_name trait1 --snp_info_file
snp_info.region1.csv --binary_genotype_file geno --output out.region1
[--binary_grm_file autosome --heritability 0.3 --error_weight_name pta_r2]
[--covariate_file covar.csv --covariate_names sex,age]
[--min_maf 0.01 --min_hwe_pval 1e-6 --exact_hwe_test]
[--prob_threshold 0.95 --prob_min_ld_r2 0.3 --repos_min_ld_r2 0.0625 --min_log_sbf
5 --num_threads 2]
[--num_permut 1000]
```

- Fine-Mapping (Shotgun Stochastic Search)

```
./bfmap --sss
```

- Single-Marker GWAS

  ```
  ./bfmap --scan
  ```

- Functional Enrichment

  ```
  Rscript
  ```

TODO