# Documentation for Software Package FM-QTL

Yeji Lee, Xiaoquan Wen

May 7, 2015

# Contents

# 1 Introduction

FM-QTL is a software package to fine mapping potentially multiple independent QTL signals across many heterogeneous subgroups. FM-QTL has been applied in applications of fine mapping expression quantitative trait loci (eQTLs) across multiple tissues (**?**) and multiple population groups (**?**). The key statistical methodology implemented in FM-QTL is detailed in **?**. Here we introduce some important concepts/terminologies, and provide a general, non-mathematical overview of the methods.

## 1.1 Subgroup Structure

FM-QTL deals with genetic association data that naturally form subgroups. Depending on the experiment design, there are two types of subgroup structures that can be described by two different types of linear model systems:

1. Multivariate linear regression (MVLR) Model. This model is typically applied when multiple phenotypes (or a single phenotype in different environments) from a common set of unrelated individuals are collected for a joint analysis. One example of this sort is the multiple-tissue eQTL analysis, where for each target gene, the expression levels of the same set of samples are measured in different tissues.

2. System of simultaneous linear regression (SSLR) Model. This model is commonly used to describe the setting where phenotype-genotype data are collected from multiple cohorts. A meta-analysis of genetic association study is a typical example.

In both cases, the "subgroup" are naturally defined by the designs, where the key difference is whether the different subgroups share the same genotype data.

## 1.2 Multi-SNP Association Mapping

FM-QTL implements a Markov Chain Monte Carlo (MCMC) based Bayesian variable selection algorithm that simultaneously identifies potentially multiple association signals across all subgroups. One of the key features is that the linkage disequilibrium (LD) pattern between genetic markers are explicitly considered in each subgroup, and there is no assumptions requiring consistent or inconsistent LD patterns across subgroups.

For each candidate genetic variant, FM-QTL computes a posterior probability of association, known as posterior inclusion probability or PIP (i.e., the posterior probability that the target SNP is included into the joint association model) in each subgroup. Larger value of PIP indicates higher plausibility

of association. Nevertheless, given observed genotype-phenotype data, we may not be able to fully resolve LD and uniquely determine the causal association variant. Consider 4 SNPs are in complete LD (i.e., $r^2 = 1$), even we are certain one of the 4 SNPs drives the causal association, without additional information, the 4 SNPs are not identifiable. As a consequence, each of the 4 SNPs should be assigned PIP = 0.25. In the end, for each subgroup, FM-QTL reports independent association signal "clusters", where each cluster consists of correlated genetic variants. The PIPs are provided at both cluster and variant levels.

### 1.3 Configuration vs. Meta-analysis Models

There are typically two competing scenarios when considering the association patthern of a causal QTL across multiple subgroups. In the first case, the genuine association signal is expected to be consistent across subgroups, although the actual effect sizes may display some level of heterogeneity in different subgroups. An example of this sort is the signals identified in meta-analysis of genetic associations. In the second case, the causal QTL association may display a subgroup specific pattern, e.g., the causal variant is associated with the trait only in *some* subgroups. Such case is typically resulted from some strong interactions between the causal genetic variant and subgroups. A typical example is tissue specific eQTLs (**?**). Consider $s$ subgroups, each possible association pattern is called a *configuration* and there are $2^s - 1$ possible configurations for a causal QTL (i.e., showing association in at least one subgroup). From this point on, we call the former and latter cases meta-analysis and configuration models, respectively. In general, the meta-analysis model should be viewed as a (much) constrained configuration model.

FM-QTL accommodates both models with the general configuration model as the default. Under the configuration model, FM-QTL also infers the configuration for each causal QTL. Due to the computational cost, we do not recommend to run configuration model with more than 7 subgroups.

### 1.4 Special Cases and Related Software

FM-QTL is designed for multiple QTL mapping in multiple subgroups. When applied with a single subgroup, the behavior of FM-QTL is similar to standard Bayesian variable selection algorithm implemented in BVSR (**?**). When used in single variant analysis, FM-QTL yields identical output as software packages MeSH (**?**) and eQTLBMA (**?**).

### 1.5 Cite FM-QTL

## 2 Installation

We provide pre-compiled binary executables for some UNIX based operating systems. Alternatively, one can directly download the source code and compile the desired binary executable. To do so, a GNU C++ compiler,e.g. GNU g++, and GNU Scientific Library (GSL) are required. To download and install GSL on your system, please refer to http://www.gnu.org/gsl/ for details.

# 3 Input Files

FM-eQTL processes each gene separately. For each gene, the required input files are

1. a grid file describing prior magnitude and heterogeneity of genetic effects across subgroups
2. a data file containing expression levels and genotypes of *cis*-SNPs

The grid file is typically shared across all genes, and the data file needs to be prepared for each gene. The detailed format descriptions for these two files are given in the following subsections.

**We also provide software utility that automatically converts multiple gene data files from MatrixEQTL input format, the details are explained in section ??.**

## 3.1 Grid File (Required)

The grid file contains prior specifications for various prior models. In all cases, the grid file always contains a two-column data matrix: the first column always represents the heterogeneity parameter ($\phi$) and the second column is used to specify the average effect size parameter ($\omega$). Each row of the grid data matrix provides a unique prior model and different rows can be used to describe different prior heterogeneity levels and/or prior average effect sizes.

The following sample of the grid file is used by **?** to perform tissue-specific eQTL analysis. The grid assumes five levels of overall prior effects ($\sqrt{\omega^2 + \phi^2} = 0.1, 0.2, 0.4, 0.8, 1.6$ values and two degrees of heterogeneity ($\phi^2/\omega^2 = 0, 0.1$), which provides a comprehensive coverage of many possible scenarios. The order of the grid can be arbitrary.

```
0.0000  0.1000
0.0000  0.2000
0.0000  0.4000
0.0000  0.8000
0.0000  1.6000
```

## 3.2 Data File (Required)

The data file contains expression levels and genotype information required by FM-eQTL analysis. The formats are slightly different for MVLR and SSLR because of their difference in experiment design. In the former case, the genotypes are shared across all subgroups, whereas in the latter case, the genotypes across subgroups are completely non-overlapping. In both cases, the data files are in text format and

contains two consecutive sections for expression levels (phenotype) and genotype, respectively. Missing values in phenotype and genotype data are allowed, and they should be represented by "NA".

### 3.2.1 Input for MVLR Model

For MVLR model in the phenotype section, each line contains a list of expression levels of all individuals in a subgroup, i.e.,

```
pheno group_id exp_ind_1 exp_ind_2 ... exp_ind_n
```

The leading "pheno" is a keyword that indicates the line encodes expression levels. The group_id field is a character string that uniquely labels a specific group. Note that the group IDs should be consistently used in the data file. The following entries are numerical values of expression levels of individual 1 to $n$ for the target gene in the indicated subgroup.

The genotype section directly follows the phenotype section. Each line contains the genotypes of a SNP, i.e.,

```
geno snp_id  geno_ind_1 geno_ind_2 ... geno_ind_n
```

The leading "geno" is a keyword that indicates the line encodes genotypes. The snp_id field contains a character string that denotes the ID of a SNP. The remaining entries are genotypes of the target SNP for individual 1 to $n$ coded in dosage format (i.e, 0,1 or 2, or any fractional numbers between $[0, 2]$ if the genotype is imputed).

For $s$ subgroups, $p$ SNPs and $n$ individual samples, the data file for MVLR model should contain $(s+p)$ lines each with $(n+2)$ entries.

### 3.2.2 Input for SSLR Model

For SSLR model in the phenotype section, each line contains a list of expression levels of all individuals in a subgroup, i.e.,

```
pheno  pheno_id group_id exp_ind_1 exp_ind_2 ... exp_ind_n
```

The leading "pheno" is a keyword that indicates the line encodes expression levels. The pheno_id field contains a character string that denotes the name of the phenotype. (Note this entry is NOT in MVLR model input data) The group_id field is a character string that uniquely labels a specific group. Note

that the group IDs should be consistently used in the data file. The following entries are numerical values of expression levels of individual 1 to $n$ for the target gene in the indicated subgroup. Note, because each subgroup can have different number of individuals, the length of each line in this section generally differs.

The genotype section directly follows the phenotype section. Each line contains the genotype information of a SNP for samples in a subgroup, i.e.,

```
geno snp_id  group_id geno_ind_1 geno_ind_2 ... geno_ind_n
```

The leading "geno" is a keyword that indicates the line encodes genotypes. The snp_id field contains a character string that denotes the ID of a SNP. The additional group_id field indicates the particular subgroup in which genotypes are measured. The remaining entries are genotypes of the target SNP for individual 1 to $n$ in the subgroup coded in dosage format (i.e, 0,1 or 2, or any fractional numbers between $[0, 2]$ if the genotype is imputed).

For $s$ subgroups, $p$ SNPs and $n$ individual samples, the data file for MVLR model should contain $s \times (p + 1)$ lines. The lines are not in equal length: for a line describing a genotype/phenotype for a subgroup with $n_s$ samples, the length should be $(n_s + 3)$.

### 3.3 Prior Specification File (Required)

## 4 eQTL Mapping with FM-eQTL

### 4.1 Gene-level Testing

The commands to compute the gene-level Bayes factor for a specified gene are as follows: for SSLR

```
sbams_sslr -d data_file -g grid_file -ga -n gene_name [ -meta | -p prior_file ] -o output
```

for MVLR, the command is almost identical, except that the name of the executable "sbams_sslr" needs to be replaced by "sbams_mvlr". The command line options are

- *-ga*: required for gene-level analysis
- *-d data_file*: required, specification of the data file
- *-g grid_file*: required, specification of the grid file
- *-meta* | *-p prior_file* : required, specification of priors

- *-o output_file*: optional, specification of the output file. By default, the program output goes to screen.

To generate the commands for all interrogated genes, users can initiate the command from `fm_eqtl` as follows

```
fm_eqtl -ga -p params.ga
```

where parameter file "params.ga" specifies the necessary input options. An example of the parameter file for gene-level analysis is given below:

```
<Contents of params.ga>
FMEQTL_BIN /usr/bin/fmeqtl/bin/
CLUSTER_MODE batch
MODEL SSLR

EQTL_DATA /net/home/eqtl_data/*.dat
GRID_FILE /net/home/grid_file
USE_META  1
```

In this example, the required data files of all the genes are put into directory `/net/home/eqtl_data/`. For each gene, the gene name is coded in the data file name and is represented by the wild card in `EQTL_DATA` option. The output from `fm_eqtl` contains the gene-level analysis commands for all the genes specified, and ready to be executed either in a cluster or a single workstation.

### 4.2 Multiple SNP Fine Mapping Analysis

The command to run the MCMC algorithm for fine mapping *cis*-eQTLs for a given gene using SSLR model is as follows

```
sbams_sslr -d data_file -g grid_file -fm [-meta | -p prior_file]  -b burnin -r repeats -o outpu
```

For MVLR, replace `sbams_sslr` with `sbams_mvlr`.

### 4.3 Program Parameters

The program parameters listed in this section should be provided in the command line.

#### 4.3.1 Input/Output Options

#### 4.3.2 Model Options

- *-es* | *-ee* | *-cef_es* | *-cef_ee* : optional, specification of the prior model. By default, ES model is assumed. Note, when *-min_info* option is set, i.e., the data file is in minimum information format, only the EE (*-ee*) and the CEF_EE (*-cef_ee*) models are eligible.

- *-abf*: optional, specification of the algorithm to compute Bayes factors. By default, this flag is NOT set, and the numerical optimization based Laplace method is used. If the flag is set, the Bayes factors are computed using analytic approximations (also by Laplace method). When *-min_info* option is set, this option is automatically set.

- *-use_config*: optional. If this flag is set, the genetic association in each subgroup is modeled as either "active" ($\beta_s \neq 0$) or "silent" ($\beta_s = 0$), an scenario quite useful in some cases of gene-environment interactions. For $s$ given subgroups, the program computes Bayes factors for all $2^s - 1$ possible configurations. Some useful applications of this model can be found in **??**.

#### 4.3.3 Miscellaneous Options

- *-no_adjust*: optional. This option is used with *-abf* option, by default, the analytic approximation of Bayes factors adjusts for small sample sizes (recommended) and makes the results more accurate. Setting this flag avoids the adjustment, this flag is automatically set when "*min_info*" flag is automatically set. (in such case, there is no enough information to perform the adjustment).

- it -print_subgrp: optional. When this flag is set, the program outputs summary statistics for each individual subgroups, namely $\hat{\beta}$ and se($\hat{\beta}$).

- *-prep_hm*: optional. When this option is specified, the program prepares input Bayes factors suitable for the Hierarchical models implemented in software packages eqtlbma and bridge.

### 4.4 Output from Program

In the simplest case, the program outputs the Bayes factor for each SNP included in the data file in each line. If "*-print_subgrp*" option is specified, following the Bayes factor result, $\hat{\beta}$ and se($\hat{\beta}$) for each subgroup are also displayed.

By default, the results are displayed on screen (stdout). If "-output_file" is specified, all the results are recorded in the specified output file.

# 5 Utilities

### 5.0.1 Genotype Information File

The file having genotype information should be separated by white-space or tabs. The first row is assumed to have sample IDs, which are corresponding to IDs in a phenotype information file and a subgroup file. After the first line, each row should contain genotype dosages for each SNP. The first entry of a row denotes SNP's name, which is same with one in maps file. The following entries can take values in numerical or decimal form, while $NA$ indicates a missing value. Missing values will be imputed by the mean of each SNP, but more sophisticated imputation beforehand is strongly recommended. The following is the example of genotype information file :

```
ID id001 id002 id003 id004 id005 ···
snp01 0 0 2 1 NA  ···
snp02 NA 0.5 1.5 2.0 1.0  ···
···
```

FM-eQTL can take either uncompressed or compressed file with gzip (filename extension is ".gz") as input. Also, separated genotype files for each chromosome are recommended.

### 5.0.2 Phenotype Information File

The phenotype information file is also assumed to be space or tab-delimited, and the sample IDs' are located at the first row. The followings contains expression levels of one gene per each row and the first column consists of genes' name. Expression levels are expected to be adjusted by possible confounders and appropriately transformed to return valid results. Also, files can be separated by each subgroup. Same with genotype information file, missing values can be denoted by $NA$ and the mean imputation will be performed. However, it is still recommended to handle missing data before using FM-eQTL. The example of phenotype information file is as follows :

```
ID id001 id002 id003 id004 id005 ···
gene01 -0.8414491 -1.4945668  0.1981837  NA -1.8755480  ···
gene02 0.9864335 -0.2313034 NA -0.3701193 -2.1197308  ···
···
```

### 5.0.3 Map File

Map files for genes and SNPs without a header are required, and they should be separated by chromosome. Both map files for genes and for SNPs are required 3 columns : gene name (SNP name), chronosome and basepair position. For each gene(SNP), only one position must be assigned. For position of genes, using the transcription start sites of genes are recommended.

```
<Example of a map file for genes>
gene01 chr1 100000
gene02 chr1 100050
gene03 chr1 100070
...
<Example of a map file for SNPs>
snp01 chr1 100003
snp02 chr1 100057
snp03 chr1 100081
...
```

### 5.0.4 Subgroup File

Subgroup file should have 2 columns and no header. The first column must have sample IDs in genotype information file and phenotype information file, but the order of ID does not need to be the same with them. The second one has subgroup information.

```
id001 group01
id002 group02
id003 group03
...
```

# 6 Data Preparation

Once all files in "Input file" section has been prepared, FM-eQTL provides way to assemble data for each gene, which consists SNPs in *cis*-region of the gene. These can be done with the following commands in Unix or Linux shell such as bash or tch.

### 6.1 Pick *cis*-SNPs for each gene

The following line makes a command file called "batch_cis.cmd" in the current directory.

```
run_sbams -p myparameters01.txt -cis_def
 in "bash" this is fine, but need to run "./run_sbams" in "tcsh"
```

myparameters01.txt takes the following lines as parameters. One line should only contain only one parameter.

```
<Contents of myparameters01.txt>
SBAMS_BIN /net/home/FM-eQTL
#CLUSTER_MODE batch
CIS_RD 200000
GENE_MAP /net/home/gene_map/chr*.gene.map
SNP_MAP /net/home/gene_map/chr*.snp.map
CIS_DEF_DIR /net/home/cis_map/
```

- SBAMS_BIN : specification of the directory that FM-eQTL has been installed. not required at this stage?
- CLUSTER_MODE : add later
- CIS_RD : definition of *cis*-region in the unit of basepair(bp). Any SNP located within ±value in this parameter will be picked and saved as *cis*-SNP.
- GENE_MAP : specification of map file for genes. Wildcard character is allowed when multiple chromosomes are handled simultaneously.
- SNP_MAP : specification of map file for SNPs. Same as GENE_MAP parameter, wildcard character is allowed.
- CIS_DEF_DIR : specification of the directory that saves the output files, which saves *cis*-pairs of SNP and gene. If the directory does not exist, then the program will automatically make one.

"batch_cis.cmd" has command lines that can run jobs in parallel. Each command reads position information from map files, and saves the list of genes and corresponding *cis*-SNPs in new files separated by each chromosome. The names of output files are in form of "chr*.cis.snp". add example - but it should be quite different depends on system....?

### 6.2 Assemble File For Analysis

After picking *cis*-SNPs, genotype and phenotype information need to be merged and reassembled, so each file only contains phenotype information of one gene and genotype information of the corresponding *cis*-SNPs. The following command produces a file called "`batch_assemble.cmd`".

```
run_sbams -p myparameters02.txt -assemble
 in "bash" this is fine, but need to run "./run_sbams" in "tcsh"
```

`myparameters02.txt` takes the following lines as parameters. One line should only contain only one parameter, same as `myparameters01.txt`

```
<Contents of myparameters02.txt>
SBAMS_BIN /net/home/FM-eQTL  no need?
#CLUSTER_MODE batch
CIS_DEF_MAP /net/home/cis_map/chr*.cis.map
EXPR_DATA /net/home/gene_map/exp.*.dat
GENO_DATA /net/home/gene_map/geno.chr*.gz
SUBGRP_DEF /net/home/cis_map/all.id
ASSEMBLE_DIR /net/home/files_for_analysis
```

- `CIS_DEF_MAP` : specification of files that defines *cis*-regions. This file should be generated from the previous step. Wildcard character is allowed to denote multiple chromosomes.
- `EXPR_DATA` : specification of phenotype information file. Wildcard character can be used to indicate several subgroups. Variables in wildcard character should be the same with subgroup names in `SUBGRP_DEF`.
- `GENO_DATA` : specification of genotype information file. Wildcard character can denote multiple chromosomes.
- `SUBGRP_DEF` : specification of subgroup file.
- `ASSEMBLE_DIR` : specification of the directory that saves the output files. If the directory does not exist, then the program will automatically make one.

"`batch_assemble.cmd`" has command lines that can assemble files that are ready to be used in FM-eQTL. The names of output files are in form of "*genename*.dat", and internal format looks like this :

```
<Contents of gene01.dat>
pheno gene01 group01 -0.8414491 -1.4945668  0.1981837 ···
pheno gene01 group01 0.2074784 -1.8755480  0.9864335 ···
pheno gene01 group01 -0.2313034 -0.9277150 -0.3701193 ···
geno snp01 group01 0 1 2 ···
geno snp01 group02 1 2 0 ···
geno snp01 group03 1 1 0 ···
geno snp02 group01 2 0 1 ···
geno snp02 group02 0 0 1 ···
geno snp02 group03 0 2 0 ···
...
```

?possible to run both steps in one command file?

# 7 Running Program

## 7.1 Gene-level Analysis

FM-eQTL can perform gene-level analysis, which also requires for fine-mapping analysis. Results of gene-level analysis will provides the most significant eQTL and the corresponding Bayes Factor, as well as the posterior probability and Bayes factor of the gene. The command line below will generates a command file called "`batch_ga.cmd`" in the current directory.

```
run_sbams -p myparameters03.txt -ga
 in "bash" this is fine, but need to run "./run_sbams" in "tcsh"
 commands in batch_ga.cmd also has same issue - maybe need to generate "./sbams_sslr" when shell
```

The example of `myparameters03.txt` are as follows :

```
<Contents of  myparameters03.txt>
SBAMS_BIN /net/home/FM-eQTL  changing this part does not change batch_ga.cmd
#CLUSTER_MODE batch
EQTL_DATA /net/home/files_for_analysis/*.dat
GRID_FILE /net/home/grid   any default grid file?
GA_DIR    /net/home/ga_results
```

- `EQTL_DATA` : specification of assembled files, generated by running "`batch_assemble.cmd`" in the previous step. Wildcard denotes names of gene.
- `GRID_FILE` : specification of grid file having prior values of subgroup heterogeneity and average effect size.
- `GA_DIR` : specification of directory for output files. If the directory does not exist, then the program will automatically make one.

Running "`batch_ga.cmd`" performs association analyses in gene level and generates an output file for each gene called "*genename*.ga.rst" in the output directory. The output file has the following format.

```
<Contents of  gene01.ga.rst>
 change header later / second column means...?
gene ? bf_gene pos_prob most_sig_snp bf_snp
```

```
gene01 1065 13.420 0.8 snp02 15.211
```