eQtlBma

for version 1.3.1a, 15 April 2015

T. Flutre and X. Wen (eqtlbma-users@googlegroups.com)

This manual is for eQtlBma (version 1.3.1a, 15 April 2015), which implements Bayesian methods for eQTL detection.

Table of Contents

1 Overview

In genetics, “QTL” stands for quantitative trait locus. It corresponds to a genotype-phenotype relationship for which a proportion of the variation in phenotype can be ascribed to deviation of the genotype from the mean genetic value (Lewontin, 2006). For the moment, this package focuses on the case where genotypes come from single nucleotide polymorphisms (SNPs) and phenotypes are gene expression levels, thus explaining the “e” in “eQTL” (but other phenotypes can be handled as well).

This package provides implementations of Bayesian methods with two goals in mind:

- to detect eQTLs;
- to interpret them.

The implemented methods allow to jointly analyze data sets from multiple subgroups. Here *subgroups* can be different tissues, populations, platforms, treatments, etc. Currently three main tools are available:

- `eqtlbma_bf` can compute summary statistics in each subgroup, Bayes factors for the joint analysis using default hyperparameters, as well as perform permutations at the gene-level;
- `eqtlbma_hm` can fit the hierarchical model via an EM algorithm (maximum likelihood), and thus provide “empirical Bayes” estimates of hyperparameters;
- `eqtlbma_avg_bfs` can perform Bayesian model averaging using the raw Bayes factors from `eqtlbma_bf` weighted by the estimates from `eqtlbma_hm`, and then compute various posterior probabilities of interest.

The details of the model are freely available online in the article by Flutre *et al* (PLoS Genetics, 2013). See also Wen’s PhD thesis (2011), the article by Wen & Stephens (Annals of Applied Statistics, 2014) and the article by Wen (Biometrics, 2014).

To install the package, it is recommended to read this web page. In case of trouble during installation or usage of the package, questions can be posted to this mailing list.

2 Tutorial

The `eqtlbma` package implements a hierarchical model based on multivariate linear regressions in a Bayesian framework fitted via an EM algorithm ("empirical Bayes"). The package can be used to answer numerous questions (see articles referenced in the "Overview" section) and has a fairly large set of options (see next sections). As such, it is quite general and thus powerful, yet can be overwhelming at first for users, even if most options are set by default. As concrete example, this tutorial hence describes a whole analysis aiming at finding eQTLs by jointly analyzing multiple tissues. We hope that it provides a clear-enough case that any user can then adapt to its own need.

As you can read this manual, we assume that the `eqtlbma` package was successfully installed on your machine. You should thus have several new programs in your PATH: `eqtlbma_bf`, `eqtlbma_hm` and `eqtlbma_avg_bfs`, as well as `tutorial_eqtlbma.R`. Following the tradition, all programs in the `eqtlbma` package show a help message with options `-h` and `--help`.

If you are not accustomed to:

- running command lines in a terminal (that is, without graphical interface), you can learn more here;
- text manipulation on UNIX-based systems, you can learn more here;
- bash scripting, you can learn more there;

but you should be able to follow the tutorial without extensive knowledge of all this.

Let's start by creating a directory for this tutorial:

```
cd $HOME
mkdir tutorial_eqtlbma
cd tutorial_eqtlbma
```

Instead of using "real" data, we will simulate "realistic" data using the widely-used free software R (version >= 3.0). Let's imagine that N=200 individuals are sampled. For each, mRNA levels are measured for G=1000 genes in S=3 tissues. Moreover, each gene has on average 50 SNPs in *cis*, for which we know the genotypes.

```
tutorial_eqtlbma.R \
--pkg ~/src/eqtlbma \
>& stdout_tutorial_eqtlbma.txt &
```

where "~/src/eqtlbma" is supposed to be the path corresponding to where the package is present (you may have to adapt it to your particular situation).

You can track the progress of the program by looking into the file "stdout_tutorial_eqtlbma.txt". Simulating data with R can be slow, but you can use option `--cores` to speed this up (adapt to the number of cores available on your machine). Note also that the script "tutorial_eqtlbma.R" contains many options one can play with to make the data more realistic (e.g. having SNPs with low MAF via option `--rare`).

As you can now see in the directory, several files were generated in such formats that it should be easy for you to prepare your own data similarly.

We can now launch the first program, `eqtlbma_bf`, to compute the Bayes factors assessing the support in the data for each gene-SNP pair being an eQTL:

```
eqtlbma_bf \
--geno list_genotypes.txt \
--scoord snp_coords.bed.gz \
--exp list_explevels.txt \
--gcoord gene_coords.bed.gz \
--anchor TSS \
--cis 1000 \
--out out_eqtlbma \
--analys join \
--covar list_covariates.txt \
--gridL grid_phi2_oma2_general.txt \
--gridS grid_phi2_oma2_with-configs.txt \
--bfs all \
--error mvlr \
>& stdout_eqtlbma_bf.txt &
```

You can track the progress of the program by looking into the file “stdout_eqtlbma_bf.txt”. Upon completion, the output file “out_eqtlbma_l10abfs_raw.txt.gz” contains the Bayes factor for each configuration of each gene-SNP pair (in rows) and each grid point (in columns):

```
zcat out_eqtlbma_l10abfs_raw.txt.gz | head
```

We can now feed this file to the second program, `eqtlbma_hm`, to fit the hierarchical model with an EM algorithm and get maximum-likelihood estimates of hyper-parameters, most importantly the configuration probabilities:

```
eqtlbma_hm \
--data "out_eqtlbma_l10abfs_raw.txt.gz" \
--nsubgrp 3 \
--dim 7 \
--ngrid 10 \
--out out_eqtlbma_hm.txt.gz \
>& stdout_eqtlbma_hm.txt &
```

You can track the progress of the program by looking into the file “stdout_eqtlbma_bf.txt”. Upon completion, the output file “out_eqtlbma_hm.txt.gz” contains the estimates as meta-data (commented lines starting with a hashtag “#”).

We now need to extract these estimates before calculating the posterior probabilities of interest:

```
zcat out_eqtlbma_hm.txt.gz | grep "#grid" | cut -f2 > grid_weights.txt
zcat out_eqtlbma_hm.txt.gz | grep "#config" \
| awk '{split($1,a,"."); print a[2]"\t"$2}' > config_weights.txt
```

Finally we can launch the third program, `eqtlbma_avg_bfs`. To obtain the posterior probabilities, we need an estimate of the probability for a gene to have no eQTL in any tissue, π_0. As this quantity is hard to estimate accurately with the EM algorithm, we usually perform permutations using `eqtlbma_bf` (see below). But for now, as this can take some time, we will use its true value, as indicated in the output file “stdout_tutorial_eqtlbma.txt”.

```
eqtlbma_avg_bfs \
--in "out_eqtlbma_l10abfs_raw.txt.gz" \
--gwts grid_weights.txt \
--nsubgrp 3 \
--dim 7 \
--cwts config_weights.txt \
--save post \
--pi0 0.3 \
--post a+b+c+d \
--bestdim \
--alldim \
--out out_eqtlbma_avg_bfs.txt.gz \
>& stdout_eqtlbma_avg_bfs.txt &
```

Upon completion, the output file “out_eqtlbma_avg_bfs.txt.gz” contains the posterior probability for the gene to have an eQTL in at least one tissue (column 3), the posterior for a SNP to be “the” eQTL (column 4), the posterior for the eQTL to be active in a given tissue (columns 6-8) and the posterior for the eQTL to be active in a given configuration (columns 9-15).

You should now be able to perform a similar analysis with your own data. Of course, you will surely need more details. See the next sections about input and output formats, program options, parallelization, etc.

The remaining of this section briefly describes how to estimate π_0, via gene-level p-values obtained by permutations followed by the qvalue package, or via the EBF/QBF procedures proposed in this preprint by Wen (2013).

As permutations take time, we will run them in parallel, which hence requires to prepare the data in batches (see Chapter 7 [Frequently asked questions], page 20 for more explanations).

First we keep only the transcription start site (TSS) of each gene:

```
zcat gene_coords.bed.gz \
| awk 'BEGIN{OFS="\t"} \
{printf "%s\t%s\t%s", $1,$2,$2+1; \
for(i=4;i<=NF;++i)printf "\t%s", $i; printf "\n"}' \
| gzip > gene_coords_TSS.bed.gz
```

Then we split the genes in batches:

```
nbBatches="100"; rm -rf lists_genes; mkdir lists_genes; cd lists_genes; \
zcat ../gene_coords_TSS.bed.gz | split \
-l $(echo "scale=0; $(zcat ../gene_coords_TSS.bed.gz | wc -l)/${nbBatches}" | bc -l) \
--suffix-length=3 --numeric-suffixes=1 --additional-suffix=.bed \
--filter='gzip > $FILE.gz' - list_genes_; cd ..
```

Then we make a file containing all *cis* SNPs per batch of genes:

```
rm -rf lists_snps; mkdir lists_snps; \
seq -w 1 $(ls lists_genes/* | wc -l) | \
parallel 'i={}; bedtools window -w 1000 \
-a lists_genes/list_genes_${i}.bed.gz \
-b snp_coords.bed.gz | \
cut -f10 | sort -V | uniq | gzip > lists_snps/list_snps_${i}.txt.gz'
```

Finally we draw one seed per batch (to make permutations reproducible):

```
nbSeeds=$(ls lists_genes/* | wc -l); \
echo "set.seed(1859); x <- sample.int(n=1000000, size=${nbSeeds}); \
write(x, gzfile(\"list_seeds.txt.gz\"), 1)" \
| R --vanilla --quiet
```

For the gene-level p-values, we can now perform permutations in parallel (note also in the command-line below that each batch is run on 4 threads):

```
echo "eqtlbma_bf_parallel.bash --p2b ~/bin/eqtlbma_bf --geneD lists_genes \
--snpD lists_snps --seedF list_seeds.txt.gz --geno list_genotypes.txt \
--scoord snp_coords.bed.gz --exp list_explevels.txt --anchor TSS --cis 1000 \
--out out_eqtlbma --analys join --covar list_covariates.txt \
--gridL grid_phi2_oma2_general.txt --gridS grid_phi2_oma2_with-configs.txt \
--bfs all --error mvlr --nperm 10000 --trick 2 --tricut 10 --pbf all --thread 4" \
| qsub -cwd -j y -V -l h_vmem=8g -N stdout_eqtlbma_bf-perm -t 1-100 -pe simple_pe 4
```

Once all jobs are finished, let us check that they all ran successfully:

```
ls stdout_eqtlbma_bf-perm.o*.* | while read f; do \
if [ $(grep -c "END" $f) != 2 ]; then echo $f; fi; done
```

We can then open R and estimate π_0 with the `qvalue` package:

```
setwd("~/tutorial_eqtlbma")
f <- Sys.glob("out_eqtlbma_[0-9][0-9][0-9]_joinPermPvals.txt.gz")
d <- do.call(rbind, lapply(f, function(fi){
  read.table(fi, header=TRUE, stringsAsFactor=FALSE)
}))
hist(d$join.perm.pval, xlim=c(0,1), breaks=100)
library(qvalue)
qobj <- qvalue(p=d$join.perm.pval, fdr.level=0.05, robust=TRUE, pi0.method="smoother")
qobj$pi0
called.nulls.pval <- ! qobj$significant
```

We can now use this estimate of π_0 to compute the various posterior probabilities of interest with `eqtlbma_avg_bfs` as shown above.

For the EBF procedure, we only need the gene-level Bayes factors averaged over the grid and configuration weights (estimated via the EM algorithm):

```
eqtlbma_avg_bfs \
--in "out_eqtlbma_l10abfs_raw.txt.gz" \
--gwts grid_weights.txt \
--nsubgrp 3 --dim 7 \
--cwts config_weights.txt \
--save bf \
--out out_eqtlbma_avg_bfs_genes.txt.gz \
>& stdout_eqtlbma_avg_bfs_genes.txt &
```

Once we have these Bayes factors, we can estimate π_0 with the EBF procedure implemented in an R file in the `eqtlbma` package:

```
gene.bfs <- read.table("out_eqtlbma_avg_bfs_genes.txt.gz", header=TRUE)
source("~/src/eqtlbma/scripts/utils_eqtlbma.R")
pi0.ebf <- estimatePi0WithEbf(log10.bfs=gene.bfs$gene.log10.bf[!duplicated(gene.bfs$gene)],
                              verbose=1)
called.nulls.ebf <- ! controlBayesFdr(log10.bfs=gene.bfs$gene.log10.bf[!duplicated(gene.bfs$gene)],
                                      pi0=pi0.ebf, fdr.level=0.05, verbose=1)
```

We can now use this estimate of π_0 to compute the various posterior probabilities of interest with `eqtlbma_avg_bfs` as shown above.

For the QBF procedure, the `eqtlbma` package performs permutations to get the median Bayes factors under the null, but much less than to get p-values:

```
echo "eqtlbma_bf_parallel.bash --p2b ~/bin/eqtlbma_bf --geneD lists_genes \
--snpD lists_snps --seedF list_seeds.txt.gz --geno list_genotypes.txt \
--scoord snp_coords.bed.gz --exp list_explevels.txt --anchor TSS --cis 1000 \
--out out_eqtlbma_qbf --analys join --covar list_covariates.txt \
--gridL grid_phi2_oma2_general.txt --gridS grid_phi2_oma2_with-configs.txt \
--bfs all --error mvlr --nperm 250 --pbf all --thread 4" \
| qsub -cwd -j y -V -l h_vmem=8g -N stdout_eqtlbma_bf-permed -t 1-100 -pe simple_pe 4
```

We check that all jobs ran successfully:

```
ls stdout_eqtlbma_bf-permed.o*.* | while read f; do \
if [ $(grep -c "END" $f) != 2 ]; then echo $f; fi; done
```

Once we have these median Bayes factors, we can estimate π_0 with the QBF procedure implemented in an R file in the `eqtlbma` package:

```
f <- Sys.glob("out_eqtlbma_qbf_[0-9][0-9][0-9]_joinPermPvals.txt.gz")
perms.qbf <- do.call(rbind, lapply(f, function(fi){
  read.table(fi, header=TRUE, stringsAsFactor=FALSE)
}))
source("~/src/eqtlbma/scripts/utils_eqtlbma.R")
pi0.qbf <- estimatePi0WithQbf(log10.bfs=perms.qbf$med.perm.log10.bf[!duplicated(perms.qbf$gene)],
                              gamma=0.5, verbose=1)
called.nulls.qbf <- ! controlBayesFdr(log10.bfs=gene.bfs$gene.log10.bf[!duplicated(gene.bfs$gene)],
                                      pi0=pi0.qbf, fdr.level=0.05, verbose=1)
```

We can now use this estimate of π_0 to compute the various posterior probabilities of interest with `eqtlbma_avg_bfs` as shown above.

As shown in the preprint of Wen (2013), the EBF/QBF procedures are much less computationally intensive than obtaining p-values by permutations, hence being tractable even for large datasets, while still conservatively controlling a given FDR level.

As a quick example, the table below summarizes results for the tutorial data obtained with all commands listed above, for which the true π_0 is 0.287, the true number of eGenes is 713 and the FDR is controlled at 0.05:

Method	**π_0 (estim.)**	**eGenes (#)**	**FDP**	**FNP**
p-values (with trick)	0.439	602	0.0266	0.3191
p-values (without trick)	0.394	609	0.0328	0.3171
EBF	0.476	617	0.0324	0.3029
QBF	0.400	627	0.0447	0.3056

3 Computing Bayes factors

Typing `eqtlbma_bf --help` or `eqtlbma_bf -h` gives the list of options. As the help message is long, we may prefer to type `eqtlbma_bf -h | less` instead.

Most importantly, for each gene-SNP pair, the `eqtlbma_bf` program can compute the Bayes factors (for each configuration and each grid point). Such Bayes factors can be computed from “raw” data or from “summary statistics” per subgroup. The `eqtlbma_bf` program can also perform permutations at the gene-level.

3.1 Inputs and options

3.1.1 Genotypes

The option `--geno` requires a file as argument. This file has two columns separated by a space or a tabulation, and one line per subgroup. The first column is the identifier of the subgroup. The second column is the path to the file containing the genotypes for this subgroup. Here is an example:

```
Fibroblasts /data/genotypes.vcf.gz
LCLs /data/genotypes.vcf.gz
T-cells /data/genotypes.vcf.gz
```

As you can see, the genotypes can all be in the same file, useful for instance if subgroups share some or all individuals. But of course it is also possible to have one file per subgroup.

If you want to skip one subgroup, simply add a hashtag at the begining of the line, like this:

```
#Fibroblasts /data/genotypes.vcf.gz
```

The files containing the genotypes can be in three possible formats. Even though these formats can handle genetic variants other than SNPs, we focus here on SNPs. Moreover, for each format, the names of the individuals have to be indicated and they need to be the same as in the files containing the gene expression levels (see next section).

SNPs with missing genotypes are skipped with a warning. It is therefore advised to impute them first with packages such as IMPUTE2 or BLIMP.

The program `eqtlbma_bf` recognizes the original VCF format (it only requires the “GT” keyword in the FORMAT column). See the specificiations on the website of the 1000 Genomes project here.

The program `eqtlbma_bf` can also handle a format very similar to the genotype format used by the IMPUTE program. The exact specification of this format is described here. The only difference is that a header line is required. Here is an example:

```
chr name coord a1 a2 <ind1>_a1a1 <ind1>_a1a2 <ind1>_a2a2 <ind2>_a1a1 ...
```

where the “<ind1>”’s have to be replaced by the name of the individuals in the given data set.

Finally, the program `eqtlbma_bf` also reads genotypes as allele dose, that is 0, 1, 2 or NA. This format is also read by the R package MatrixEQTL. Here is an example:

```
id ind1 ind2 ind3 ...
snp1 0 2 1 ...
```

```
snp2 0 1 0 ...
...
```

The VCF and IMPUTE-like formats contain information about SNP coordinates, thus they should not be used with `--scoord`. However, the allele-dose format do need the option `--scoord`, followed by a file containing the SNP coordinate in the BED format. This means that the start coordinate is 0-based, there is no header line and the column separator is a tabulation.

When parallelizing an analyzis over genes, we may want to only load the SNPs in *cis* of the genes in the given job. In order to speed-up this, the eQtlBma package uses some code from TABIX described in this paper (Li, 2011). More specifically, if a bgzip-compressed BED file named "snp_coords.bed.gz" is given to the option `--scoord`, the code will look for a tabix-indexed file named "snp_coords.bed.gz.tbi" and use it to only load the SNPs in *cis* of the genes specified by the option `--gcoord`. If the index file is not present, all SNPs will be loaded which will be slower and use more memory. See Chapter 7 [Frequently asked questions], page 20 to know how to build the index for the BED file.

The option `--covar` requires a file as argument. This file has two columns separated by a space or a tabulation, and one line per subgroup. The first column is the identifier of the subgroup. The second column is the path to the file containing the covariates for this subgroup.

Each covariate file has to be in the following format:

```
id ind1 ind2 ...
covar1 0.32 0.11 ...
covar2 -1.0 0.8 ...
...
```

Note that this format is also read by the R package MatrixEQTL. Here also, no missing covariate is allowed. However, in eQtlBma, the covariates are assumed to be additive. This is fine for continuous covariates (say, principal components to account for some population structure) or binary (say, gender). However, if the covariates are categorical, we recommend to regress them out beforehand, for instance by using R where you can encode them as factors.

3.1.2 Expression levels

As for the option `--geno`, the option `--exp` requires a file as argument. This file has two columns separated by a space or a tabulation, and one line per subgroup. The first column is the identifier of the subgroup. The second column is the path to the file containing the genotypes for this subgroup. Here is an example:

```
Fibroblasts /data/phenotypes_Fibroblasts.txt.gz
LCLs /data/phenotypes_LCLs.txt.gz
T-cells /data/phenotypes_T-cells.txt.gz
```

The program `eqtlbma_bf` uses the term "gene" as the generic term for the entities for which we have measurements. Besides genes, they could be exons, transcripts, proteins, metabolites, etc, but we stick to genes in this manual. (Note also that the program implements a model with a specific prior meaningful for genes but which may not be appropriate for some other entities.)

The actual files containing the expression levels have the following format:

```
ind1 ind2 ind3 ...
gene1 2.0495 1.0947 1.9924 ...
gene2 0.1928 -0.873 0.5284 ...
...
```

Here again, this format is read by the R package MatrixEQTL.

More importantly, the sample identifiers should be the same between genotype and expression files. The order of the columns is not important, but the fact that the identifiers should be the same between files is an effort to avoid forgetting which column correspond to which sample, as can easy happen when data sets are shared between collaborators.

Note that, starting with version 1.3, missing expression levels are allowed. They should be encoded with "NA", "na", "NaN" or "nan". Missing data can arise from various reasons. The consequence can be that given individuals are completely missing from some subgroups, or only some genes from given individuals are missing in some subgroups. If the individuals are different in each subgroup, then one doesn't need to allow the errors in the multivariate regression to be correlated, and there is no problem. However, if the individuals are partially overlapping between subgroups, we need to explicitly handle the missing data. In such a case, we assume that the data are "missing at random", that is, the fact that a gene expression level is missing is a priori not associated with the genotypes at any SNP. Then, the only issue is about estimating the covariance between the errors in each pair of subgroups (off-diagonal elements of the covariance matrix). Our solution is straightforward: for a given pair of subgroup, we simply use the individuals present in these two subgroups.

As the files with phenotypes don't contain the gene coordinates, we also need to use the option `--gcoord` to specify gene coordinates in the BED format. Genes with no coordinates will be skipped (useful when launching the program in parallel, see below).

The option `--qnorm` allows the program to transform the expression level of each gene into the quantiles of a standard Normal distribution. This is done just before performing the linear regressions. Otherwise, an FAQ entry at the end of this document indicates how to do this beforehand in R (better because ties are randomly broken in the R code).

3.1.3 *Cis* region

The `eqtlbma_bf` program focuses on detecting associations between SNPs and genes, restricting itself to SNPs in a *cis* candidate region of each gene. The option `--cis` precises the length of half of the *cis* region (i.e. the radius), in base pairs.

Following the convention in BEDTools, the definition of the *cis* region uses `<=` and `>=` instead of `<` and `>`.

Note that, for the moment, only the first four columns of the BED file are used, assuming that the start and end coordinates correspond to the TSS. At some point we will have to improve the code to also handle the strand, if specified.

3.1.4 Possible analyzes

The option `--out` requires a character string which will be used as a prefix to name the output files. Moreover, all output files are directly written in a compressed mode using zlib. That is, all output files are readable by `gzip` and `zcat`.

The program `eqtlbma_bf` can perform several analyzes. The option `--analys sep` means that the gene-SNP pairs will be tested for association using the subgroup-by-subgroup analysis ("separate" analysis). The option `--analys join` means that the gene-SNP pairs will be tested for association using all subgroups jointly, which is more powerful in the context of eQTL studies, as shown in Flutre *et al.*

At the beginning of each step, summary statistics are computed in each subgroup (estimates of effect sizes, standard errors, p-values, etc). If the option `--outss` is not specified, the summary statistics won't be saved. This can be useful in some cases: for instance, when we want to run a set of jobs with `--permsep 1` and another set of jobs with `--permsep 2`, both sets of jobs in the same directory. We would typically use option `--outss` with the first set of jobs but not with the second, otherwise both sets of jobs may overwrite each other's files. However, note that we need the summary stats per subgroup if we want to later make meta-analysis-like forest plots.

If the option `--outw` is not specified, only the raw Bayes factors will be saved, as they are needed to fit the hierarchical model with the `eqtlbma_hm` program. If it's specificed, the program will also save the Bayes factors per configuration, averaged over the grid using uniformly equal weights.

When using option `--analys join`, we need to specify the options `--gridL` and `--gridS`, along with two files containing the grids over which the Bayes factors are averaged. A grid has two columns, the first contains values of ϕ^2 (prior variance of the standardized effects b_s's in each subgroup) and the second values of ω^2 (prior variance of the average standardized effect $\bar{b}$).

The option `--gridL` specifies a "large" grid. It is typically used for meta-analysis (see Wen & Stephens, AoAS 2014), or for the BMAlite analysis (see Flutre *et al*, 2013). The option `--gridS` specifies a "small" grid used with configurations (see Flutre *et al*). An FAQ entry below shows how to produce such files in R.

The option `--bfs` allows to specify which Bayes factors we want to compute. The acronym "abf" is used to mean "approximated Bayes factor" because the Bayes factor can't be calculated analytically and therefore has to be approximated. The article from Wen & Stephens in AoAS 2014 detailed different ways to do that. However, `--bfs` only implements the ABF with the ES model using Laplace's method. Moreover, the small sample size correction is also implemented (Wen & Stephens, AoAs 2014, appendix C).

Specifying `--bfs gen` computes the "general" BF corresponding to the consistent configuration using the large grid. This "general" BF is useful in a meta-analysis setting, and fixed-effect and maximum-heterogeneity BFs are also calculated (see Wen & Stephens, AoAS 2014).

Specifying `--bfs sin` also computes the BF for each singleton, that is for configurations (100), (010) and (001) if there are 3 subgroups. Also, the average of the "general" BF and each "singleton" BF is reported under the name "abf.gen.sin". It corresponds to "BF_BMAlite" in Flutre *et al.*

Finally, specifying `--bfs all` computes the BF for each configuration. The weighted average of all these BFs is also reported under the name "abf.all". It corresponds to "BF_BMA" in Flutre *et al.* Using `--bfs all` can be too costly when the number of subgroups exceeds 15 or 20.

The model implemented in this package is based on a linear regression, and there are several ways of specifying the variance-covariance matrix of the errors. If the subgroups contain different individuals, we can choose `--error uvlr`, for univariate linear regression, as in Wen & Stephens, AoAS 2014. If the subgroups contain exactly the same individuals, we can choose `--error mvlr`, for multivariate linear regression, as in Wen, Biometrics 2014. If the subgroups contain some individuals in common and some not, we can choose `--error hybrid`. For the latter, the effect sizes and their variance are estimated using all individuals in each subgroup, whereas their covariance are estimated using each pair of subgroups with only individuals in common.

When using `--error mvlr` or `--error hybrid`, the option `--fiterr` is set by default at 0.5. See Wen, Biometrics 2014, for the rationale. Also, with `--error mvlr`, the summary statistics per subgroup are not exported (in theory it's possible but the current code doesn't allow it easily). So in this case, if we want to make forest plots, we will have to also launch `eqtlbma_bf` with option `--error uvlr`.

3.1.5 Summary statistics

Sometimes it is not possible to access the "raw" data, as in human genetics where genotypes often are confidential. In such a case, `eqtlbma_bf` can still computes the Bayes factors using summary statistics. Note that, for the moment, it only works with `--error uvlr`.

The option `--inss` requires a file as argument. This file has two columns separated by a space or a tabulation, and one line per subgroup. The first column is the identifier of the subgroup. The second column is the path to the file containing the summary statistics for this subgroup. Here is an example:

```
Fibroblasts /results/sstats_Fibroblasts.txt.gz
LCLs /results/sstats_LCLs.txt.gz
T-cells /results/sstats_T-cells.txt.gz
```

The actual files containing the summary statistics need a header line containing the following words: gene, snp, n, sigmahat, betahat.geno and sebetahat.geno (in any order). Let's consider the following linear regression of mRNA levels at gene g in subgroup s on the genotypes at SNP p: $\forall i \in \{1, \dots, n\},\ y_{gsi} = \mu_{gs} + \beta_{gps} x_{psi} + \epsilon_{gpsi}$ with $\epsilon_{gpsi} \sim N(0, \sigma^2_{gps})$. As a result, the n column should contain the number of samples in the linear regression; the *sigmahat* column should contain the estimate of the standard deviation of the errors, σ_{gps}; the *betahat.geno* column should contain the estimate of the effect size of the genotype, β_{gps}; and the *sebetahat.geno* column should contain the standard error of this estimate. In the end, the file should have the following format:

```
gene    snp     n       sigmahat        betahat.geno    sebetahat.geno
gene1   snp26   200     7.843116e-01    8.091162e-02    8.258911e-02
...
```

3.1.6 Permutations

Genes having different numbers of SNPs in *cis*, with different patterns of linkage disequilibrium, we implemented a permutation procedure *at the gene level* (see Flutre *et al*). Such a procedure provides a p-value for each gene, required to control the FDR at the gene level, hence allowing statements such as "there are X genes having at least one eQTL at an FDR of x%".

The option `--nperm` allows to specify how many permutations will be performed. We recommend 10,000. In practice, we permute the individual labels (not the sample labels). As individuals can be present in several subgroups, we recommend to use `--permsep 1` to preserve such correlation structure when doing a subgroup-by-subgroup analysis. We can also specify the initialization of the random number generator with the option `--seed` in order to be able to replicate the results exactly.

To speed-up the permutations, we also recommend to use the option `--trick 1`. Indeed, when it is clear that there is no association between the given gene-SNP pair, it is not necessary to perform 10,000 permutations, a much smaller number is enough, and this option implements this adaptively for each gene-SNP pair. It requires another random number generator, which also uses `--seed`. The output file will contain the total number of permutations performed. The option `--tricut` allows to tune the speed gain of the trick: the smaller the faster (i.e. less permutations are performed when there is no association). In our experience, using `--tricut 10` gives good results.

If we want to compare the two approaches ("separate" versus "joint" analysis), we may want to use the exact same permutations for both. Yet we may also want to use the "trick". Specifying `--trick 2` allows to do just that and is therefore recommended in this setting.

Finally, the option `--pbf` specifies which BF is used as a test statistic when `--analys join`. The BF called "BMA" in Flutre *et al* corresponds to `--pbf all`, and the BF called "BMAlite" corresponds to `--pbf gen-sin`.

However, permutations to obtain gene-level p-values are computationally intensive, perhaps prohibitively so for large datasets. See Chapter 2 [Tutorial], page 2, in the end of the chapter, for efficient, yet powerful alternatives, namely the EBF and QBF procedures proposed in this preprint by Wen (2013).

3.2 Computing in parallel

For a small analysis, the command-line for `eqtlbma_bf` given in the tutorial is enough. However, when dealing with many genes (20,000) and SNPs (5 millions), we recommend to split the analysis in batches and launch them in parallel. To simplify this and avoid the burden of creating new input files with genotypes and expression levels, we can simply have several BED files with different subsets of genes (one per batch).

If we want 100 batches, we only need to split all the gene coordinates into 100 lists. An FAQ entry below indicates how to do this easily.

Then, we can use the script `eqtlbma_bf_parallel.bash`. After installation of the package, it should be in your PATH. Otherwise it is in the directory `scripts/` of the package.

A typical command-line looks like this (works with Sun Grid Engine):

```
qsub -cwd -j y -V -l h_vmem=2g -N job_eqtlbma -t 1-100 \
eqtlbma_bf_parallel.bash \
--p2b ~/bin/eqtlbma_bf \
--geneD lists_genes \
--snpD lists_snps \
--seedF list_seeds.txt.gz \
--geno list_genotypes.txt \
--scoord snp_coords.bed.gz \
--exp list_expressions.txt \
--out out_eqtlbma \
--analys join \
--covar list_covariates.txt \
--gridL grid_phi2_oma2_general.txt.gz \
--gridS grid_phi2_oma2_with-configs.txt.gz \
--bfs all
--error mvlr \
--nperm 10000 \
--trick 2 \
--pbf all
```

Note that you can also use the option `--snp` if you want to analyse only a subset of all SNPs per batch, e.g. only those in *cis* of the genes in the corresponding batch. An FAQ entry shows how to find SNPs in *cis* for each gene.

Another FAQ entry shows how to generate a file of seeds, to make each batch reproducible when doing permutations.

Once all jobs are finished, see Chapter 7 [Frequently asked questions], page 20 for details on how to concatenate all output files of a given kind, for instance to have all “_sumstats_<subgroup>.txt.gz” batch files into a single file.

If the cluster doesn't work with SGE but with SLURM, here is the equivalent command-line:

```
for i in $(seq -w 1 100); do echo $i; \
echo -e '#!/usr/bin/env bash\n'''eqtlbma_bf_parallel.bash \
--p2b ~/bin/eqtlbma_bf \
--geneD lists_genes \
--snpD lists_snps \
--seedF list_seeds.txt.gz \
--task ${i} \
--geno list_genotypes.txt \
--scoord snp_coords.bed.gz \
--exp list_expressions.txt \
--out out_eqtlbma \
--analys join \
--covar list_covariates.txt \
--gridL grid_phi2_oma2_general.txt \
--gridS grid_phi2_oma2_with-configs.txt \
--bfs all \
--error mvlr \
--nperm 10000 \
--trick 2 \
--pbf all'' \
| sbatch -J job_eqtlbma_bf -o stdout_eqtlbma_bf-perm-${i}.o%j \
--mem-per-cpu=10000; sleep 1; done
```

3.3 Reading the outputs

The program `eqtlbma_bf` creates several output files, all starting with the character string given to option `--out`, e.g. “out_eqtlbma” (remember to include the batch number when you parallelize yourselves, otherwise `eqtlbma_bf_parallel.bash` does it automatically).

All output files contain a header line, which should make it easy to understand what each file contains, as well as load each file into R.

If `--outss` is set and `--error mvlr` is not, one file is created per subgroup with some summary statistics. They have the suffix "_sumstats_<subgroup>.txt.gz". These summary statistics can be used to draw forest plots. For steps 2 and 5, there will also be file(s) with the results of the permutations. If `--permsep 1` was given, there will be one such file, with suffix "_sepPermPvals.txt.gz". If `--permsep 2` was given, there will be one file per subgroup, with suffix "_sepPermPvals_<subgroup>.txt.gz".

One file will contain all the "raw" BFs, i.e. one per config per grid point, with suffix "_l10abfs_raw.txt.gz". Such files are necessary to run the hierarchical model with `eqtlbma_hm` (see below).

If option `--outw` was given, there will also be one file containing all the BFs averaged over the grid, with suffix "_l10abfs_avg-grids.txt.gz". Also, for steps 4 and 5, there will also be a file with the results of the permutations, with suffix "_jointPermPvals.txt.gz".

4 Fitting the hierarchical model

The `eqtlbma_hm` program can take several options, available in the command line via `eqtlbma_hm -h`.

The option `--data` requires the input file with the Bayes factors, typically the output file from `eqtlbma_bf` with suffix "_l10abfs_raw.txt.gz". We can also give a file pattern (a glob), such as `--data "out_eqtlbma_[0-9][0-9][0-9]_l10abfs_raw.txt.gz"`, where "`[0-9][0-9][0-9]`" corresponds to the batch numbers (e.g. 001, 002, ..., 100).

The option `--nsubgrp` requires the number of subgroups, e.g. 3.

The option `--dim` requires the number of configurations to considered (i.e. the dimension of the latent space). More specifically, it corresponds to the number of active configurations, e.g. 7 if there are 3 subgroups.

The option `--ngrid` requires the number of grid points to consider. For instance, if we launched `eqtlbma_bf` with a "small" grid of 10 points (ϕ_l^2, ω_l^2), we need to specify `--ngrid 10`.

The option `--out` requires the name of the output file, which will be gzipped. The first lines start with a hashtag and correspond to the estimates of the hyperparameters, along with their confidence intervals (if option `--getci` was given). By default, these lines will be considered as comments by R and won't be loaded. Then, if the option `--getbf` was given, the rest of the file contains averaged Bayes factors for each gene and gene-SNP pair.

The option `--init` can take an initialization file. It should have 3 columns separated by a tabulation and one line per parameter. The first column should contain the name of the parameter, such as "config.1-2-3" or "grid.1". The second column should contain the value of the parameter. The third column should contain a boolean, encoded as TRUE or FALSE, indicating if the parameter should be kept fixed or not. Note that all parameters should be present in the file.

The option `--rand` can be used to randomly initialize the parameters at the beginning of the EM. To make inference replicable, we can use the option `seed`.

The option `--tresh` can be given the threshold to terminate the EM algorithm. That is, if the log-likelihood increases less than this threshold, the iterations stop. The default value is set at 0.05.

The option `--maxit` can be used to fix the maximum number of iterations to be performed before stopping the EM algorithm. It can be useful if `eqtlbma_hm` is executed on a computer cluster with a wall-time limit shorter than the running time of the EM algorithm. For instance, if `eqtlbma_hm` is killed after 27 iterations, we can launch it a first time with `--maxit 25`; then extract the current estimates of the parameters and format them appropriately into an initialization file (see Chapter 7 [Frequently asked questions], page 20); finally launch `eqtlbma_hm` a second time with `--init`.

The option `--sq` makes use of the SQUAREM procedure to speed-up the EM algorithm (Varadhan & Roland, 2008). It was implemented by Chaoxing Dai.

In order to speed-up the computations greatly, the option `--thread` can be given a number of threads (the code uses OpenMP).

If we want to fit the hierarchical model using only a pair of subgroups, we can use the option `--configs`. For instance, among 3 subgroups, to only load the Bayes factors corresponding to subgroups 1 and 3, we would do `--configs "1|3|1-3"`.

If we want to fit the hierarchical model according to the BMAlite approach, we can use the option--keepgen. With --keepgen, the raw ABFs with "gen" in the "config" column of the output file from eqtlbma_bf will be kept, and those with gen-fix/gen-maxh will be ignored. On the contrary, without --keepgen (i.e. the default), all raw ABFs with gen/gen-fix/gen-maxh in the "config" column will be ignored.

The option --getci can be set in order to compute and return 95% confidence intervals using the profile likelihood. However, this is not multi-threaded and can therefore be quite long. Otherwise, only the maximum-likelihood estimates of the hyperparameters are returned.

By default, only the estimates of the hyperparameters (gene-level π_0, grid and configuration weights) are returned in the output file. We can use option --getbf in order to also get the Bayes factor for each gene and each gene-SNP pair, as well as the BF for each configuration, which can take a lot of time to compute and result in a big file. We would hence surely prefer to use the eqtlbma_avg_bfs program (see below) which offers more flexibility about which quantities to compute (averaged Bayes factors, various posteriors).

The gene-level π_0 (the probability for a gene to have no eQTL) is hard to estimate accurately with the EM algorithm, therefore it can be useful to estimate it by another method (e.g. permutations, EBF/QBF procedure) and set it manually to see how it impacts the estimates of the other hyperparameters. We can easily do it using option --pi0, meaning that pi0 won't be updated by the EM algorithm. Another way is to use a file with --init, but in that case all other parameters should also be present in the file.

Finally, the command-line will typically look like this:

```
eqtlbma_hm \
--data "out_eqtlbma_*_l10abfs_raw.txt.gz" \
--nsubgrp 3 \
--dim 7 \
--ngrid 10 \
--out out_eqtlbma_hm.txt.gz \
--thread 4
```

After launching the eqtlbma_hm program, we can follow the EM iterating on stdout.

Then, if option --getbf was not given, we can use the eqtlbma_avg_bfs program to compute the final quantities of interest, e.g. posteriors.

5 Computing the posteriors

The `eqtlbma_avg_bfs` program can take several options, available in the command line via `eqtlbma_avg_bfs -h`.

The option `--in` requires the input file with the Bayes factors, typically the output file from `eqtlbma_bf` with suffix "_l10abfs_raw.txt.gz". We can also give a file pattern (a glob), such as `--data "out_eqtlbma_[0-9][0-9][0-9]_l10abfs_raw.txt.gz"`, where "`[0-9][0-9][0-9]`" corresponds to the batch numbers (e.g. 001, 002, ..., 100).

The option `--gwts` requires the path to a file containing the grid weights. There should be one value per line. For instance, with the default grid in file "grid_phi2_oma2_with-configs.txt.gz" as generated by the R code in the FAQ, there are 10 lines. See Chapter 7 [Frequently asked questions], page 20 to know how to extract the grid weights from the output of `eqtlbma_hm`.

If you want to only keep a subset of the Bayes factors, for instance only those corresponding to lines 1, 3 and 5 of the grid, you can use the option `--gtk`, such as `--gtk 1+3+5`.

The option `--nsubgrp` requires the number of subgroups, and the option `--dim` requires the dimension of the model, that is the number of active configurations (7 if there are 3 subgroups).

The option `--cwts` requires the path to a file containing the configuration weights. This file should have two columns, the identifier of the configuration and its probability. There should then be one configuration per line. In the end, the file should look like this:

```
1         0.13
2         0.19
1-2       0.68
```

See Chapter 7 [Frequently asked questions], page 20 to know how to extract the configuration weights from the output of `eqtlbma_hm`.

The option `--save` is used to indicate which quantity(ies) should be saved in the output file. Using `--save bf` means that only Bayes factors will be saved, `--save post` means that only posteriors will be saved, and `--save bf+post` means both. Note that saving the posteriors also requires specifying the options `--pi0` and `--post` (see below).

The option `--pi0` requires the value of the probability for a gene to have no eQTL in any subgroup. If not provided, Bayes factors will be saved instead of posteriors. As π_0 is hard to estimate accurately with the EM algorithm, it can be useful to estimate it by another method (e.g. permutations, EBF/QBF procedure).

The option `--post` requires the kind(s) of posteriors to save. Using `--post a` corresponds to the posterior for a gene to have at least one eQTL in at least one subgroup, `--post b` corresponds to the posterior that the SNP is "the" eQTL for the gene (i.e. "eQTN"), in at least one subgroup, given that the gene has exactly one eQTL and assuming all cis SNPs are equally likely; `--post c` corresponds to the posterior that the SNP is 'an' eQTL for the gene, in at least one subgroup, given that the gene contains at least one eQTL and that the SNPs are independent; and `--post d` corresponds to the posterior that the SNP is an eQTL in subgroup s, given that it is "the" eQTL for the gene (i.e. the configurations are marginalized).

The option `--gene` requires the path to a file with a subset of gene(s) to keep. There should be one gene per line.

The option `--snp` requires the path to a file with a subset of SNP(s) to keep. There should be one SNP per line. Caution about this option because, as not all cis SNPs are kept, this will change gene-level Bayes factors and posteriors.

The option `--gene-snp` requires the path to a file with a subset of gene-SNP pair(s) to keep. There should be two columns, the first for the gene and the second for the SNP. As for `--snp`, as not all cis SNPs are kept, this will change gene-level Bayes factors and posteriors.

The option `--bestsnp` requires the kind of best SNP(s) to save. The default, `--bestsnp 0` means that all cis SNPs will be saved. Using `--bestsnp 1` means only the best SNP is saved (pick one if tie), based on the proba for a SNP to be "the" eQTL. Using option `--bestsnp 2` means that, possibly several, best SNPs are reported so that the sum of their proba to be "the" eQTL just exceeds 0.95.

The option `--bestdim` is used to report the best configuration per SNP, as well as its Bayes factor and/or posterior, whereas the option `--alldim` reports the Bayes factors and/or posteriors of all configurations (caution, this can be a lot).

The option `--thread` requires the number of threads to use (default is 1).

Finally, the command-line will typically look like this:

```
eqtlbma_avg_bfs \
--in "out_eqtlbma_*_l10abfs_raw.txt.gz" \
--gwts grid_weights.txt \
--nsubgrp 3 \
--dim 7 \
--cwts config_weights.txt \
--save post \
--pi0 0.783629 \
--post a+b+c+d \
--bestdim \
--alldim \
--out out_eqtlbma_avg_bfs.txt.gz \
--thread 4
```

6 For developers

The eQtlBma package is freely available under the GNU General Public License version 3 or later (GPL-3+). The code is versioned with git and the official repository is available on Github, so you can fork it and let us know of any pull request. The core of the package is written in C++, but parts of it are also in R and bash. In all languages, we are using the camel case notation with an uppercase first letter for classes and a lowercase first letter for methods, functions and variables. Functional tests are implemented, even though, of course, unitary tests would be preferable (see the TODO file). Importantly, variable and function name are chosen so as to be as explicit as possible! For those using Emacs, here is the configuration:

```
(setq-default indent-tabs-mode nil)
(setq-default tab-width 2)
(setq c-default-style "bsd"
      c-basic-offset 2
      tab-width 2
      indent-tabs-mode t)
(setq sh-basic-offset 2
      sh-indentation 2)
```

In terms of versioning, the eQtlBma package follows the Semantic Versioning guidelines. For each release, a git tag is created, and then pushed to GitHub.

7 Frequently asked questions

- **How do I cite this package?**

  Flutre T, Wen X, Pritchard J, Stephens M (2013) A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLoS Genet 9(5): e1003486. doi:10.1371/journal.pgen.1003486

  This article is freely available online.

- **Who funded this work?**

  As described in Flutre et al., Jonathan Pritchard, Matthew Stehens and Xiaoquan Wen were supported by NIH grant MH090951. Timothée Flutre was also supported by the Institut National de la Recherche Agronomique (INRA) as ASC.

- **How do I make the file(s) for the grid(s)?**

  See the function `makeGrid` in the file `scripts/utils_eqtlbma.R`. A typical code would look like this:

```
gridL <- makeGrid("general")
write.table(x=gridL, file=gzfile("grid_phi2_oma2_general.txt.gz"),
            quote=FALSE, row.names=FALSE, col.names=FALSE)
gridS <- makeGrid("configs")
write.table(x=gridS, file=gzfile("grid_phi2_oma2_with-configs.txt.gz"),
            quote=FALSE, row.names=FALSE, col.names=FALSE)
```

- **How do I transform my phenotypes beforehand into the quantiles of a standard Normal?**

  See the function `transformGeneExpInStdNormal` in the file `scripts/utils_eqtlbma.R`. Ties can be broken randomly (particularly useful with RNA-seq).

- **How do I make the tabix index for the BED file with SNP coordinates?**

  Start by installing the TABIX package. Then sort the BED file and compress it with the `bgzip` program (part of the TABIX package). Finally, make the index with the `tabix` program. All this can be done with the following commands:

```
cat snp_coords.bed | sort -k1,1V -k2,2g | bgzip > snp_coords.bed.gz
tabix -p bed snp_coords.bed.gz
```

  The option `-V,--version-sort` of GNU sort allows to sort chromosome names in alpha-numeric order, i.e. “chr10” after “chr2”. It is available at least in version 8.17 of GNU coreutils or later.

- **How do I split the BED file of gene coordinates in 100 batches?**

  Using GNU tools and assuming the coordinates are in a file named “gene_coords.bed.gz”:

```
nbBatches="100"; rm -rf lists_genes; mkdir lists_genes; cd lists_genes; \
zcat ../gene_coords.bed.gz | split \
-l $(echo "scale=0; $(zcat ../gene_coords.bed.gz | wc -l)/${nbBatches}" | bc -l) \
--suffix-length=3 --numeric-suffixes=1 --additional-suffix=.bed \
--filter='gzip > $FILE.gz' - list_genes_; cd ..
```

This will create approximately 100 files in a directory, such as “lists_genes/list_genes_001.bed.gz”, “lists_genes/list_genes_002.bed.gz”, etc.

- **How do I get a file with SNPs in *cis* for each batch of genes?**

  Using BEDTools, it’s quite easy. As it can take some time, we can use GNU parallel to speed this up:

  ```
  rm -rf lists_snps; mkdir lists_snps; \
  seq -w 1 $(ls lists_genes/* | wc -l) | \
  parallel 'i={}; bedtools window -w 100000 \
  -a lists_genes/list_genes_${i}.bed.gz \
  -b snp_coords.bed.gz | \
  cut -f10 | sort -V | uniq | gzip > lists_snps/list_snps_${i}.txt.gz'
  ```

  Each SNP file contains the identifiers (e.g. rs number) of the SNPs in cis of the genes in the corresponding batch. Change the "100000" into "1000000" if you want a 1Mb radius instead of a 100Kb radius for the *cis* window. Note also that, to have the *cis* region centered on the TSS only (i.e. neglecting the TES), you will first have to modify the file "gene_coords.bed.gz":

  ```
  zcat gene_coords.bed.gz \
  | awk 'BEGIN{OFS="\t"} \
  {printf "%s\t%s\t%s", $1,$2,$2+1; \
  for(i=4;i<=NF;++i)printf "\t%s", $i; printf "\n"}' \
  | gzip > gene_coords_TSS.bed.gz
  ```

- **How do I make the file of seeds when using eqtlbma_bf_parallel.bash?**

  Before launching `eqtlbma_bf_parallel.bash` to do permutations, use the following command-line (requires R):

  ```
  nbSeeds=$(ls lists_genes/* | wc -l); \
  echo "set.seed(1859); x <- sample.int(n=1000000, size=${nbSeeds}); \
  write(x, gzfile(\"list_seeds.txt.gz\"), 1)" \
  | R --vanilla --quiet
  ```

- **How do I easily concatenate the output files from all batches?**

  When launching `eqtlbma_bf` in parallel, you will get several output files for each batch. For a given kind of output files, for instance the summary statistics of a given subgroup, it may be easier to deal with a single file. Below are simple bash commands to concatenate all batch files of a same kind into a single file and compress it:

  ```
  sbgrp="Tissue3"; i=0; \
  ls out_eqtlbma_[0-9][0-9][0-9]_sumstats_${sbgrp}.txt.gz | while read f; do \
  let i=i+1; echo $i; \
  if [ $i -eq "1" ]; then zcat $f > out_eqtlbma_sumstats_${sbgrp}.txt; \
  else zcat $f | sed 1d >> out_eqtlbma_sumstats_${sbgrp}.txt; fi; done
  gzip out_eqtlbma_sumstats_${sbgrp}.txt
  ```

  You will have to adapt this command for the other kinds of output files.

- **How do I extract the grid weights from eqtlbma_hm's output?**

  Before using `eqtlbma_avg_bfs`, use the following command-line:

  ```
  zcat out_eqtlbma_hm.txt.gz | grep "#grid" | cut -f2 > grid_weights.txt
  ```

- **How do I extract the configuration weights from eqtlbma_hm's output?**

  Before using `eqtlbma_avg_bfs`, use the following command-line:

  ```
  zcat out_eqtlbma_hm.txt.gz | grep "#config" \
  | awk '{split($1,a,"."); print a[2]"\t"$2}' > config_weights.txt
  ```

- **How do I format the parameter estimates from eqtlbma_hm's output to feed –init?**

  Use the following command-line:

  ```
  zcat out_eqtlbma_hm.txt.gz | grep "#" | sed 's/#//g' | sed 1d \
  | cut -f1,2,5 > init_hm.txt
  ```

- **How do I extract ABFs for a subset of genes after eqtlbma_bf?**

  First, you have to write in a file the list of genes you are interested in. This file, below assumed to be called "list_genes_tokeep.txt", should have a single column and a single gene identifier per line.

  Then, assuming that you ran eqtlbma_bf in parallel, use the following command-line to (1) list all ABF files, and then, for each of them, (2) retrieve the file prefix, (3) extract the genes of interest into another file, and (4) compress it:

  ```
  ls *_l10abfs_raw.txt.gz | while read f; do \
    echo $f; f2="${f%.txt.*}"; echo $f2; \
    rm -f ${f2}_filter.txt; \
    zcat $f | head -1 > ${f2}_filter.txt; \
    zcat $f | sed 1d | grep -f list_genes_tokeep.txt >> ${f2}_filter.txt;
    gzip ${f2}_filter.txt
  done
  ```

  As it can take some time, we can use GNU parallel to speed this up:

  ```
  ls *_l10abfs_raw.txt.gz | \
  parallel 'f={}; f2="${f%.txt.*}"; rm -f ${f2}_filter2.txt; \
  zcat $f | head -1 > ${f2}_filter2.txt; \
  zcat $f | sed 1d | grep -f list_genes_tokeep.txt >> ${f2}_filter2.txt;
  gzip ${f2}_filter2.txt'
  ```

- **Is this packaged tested?**

  We implemented some R code in order to perform functional tests on `eqtlbma_bf` and `eqtlbma_hm`. Launching them is automatized via `make check` (requires R >= 2.15).

  You can also find in the `src/` directory the code used to simulate data as in Flutre *et al* (2013). To compile it, enter into the `src/` directory and run `grep "g++" simul_flutre_et_al.cpp` to see how to do it. As usual, a help message is available with the option `-h`.

  If you find a bug, please don't hesitate to contact us, thanks in advance!

Appendix A GNU Free Documentation License

Version 1.3, 3 November 2008

0. PREAMBLE

   The purpose of this License is to make a manual, textbook, or other functional and useful document *free* in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

   This License is a kind of “copyleft”, which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

   We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

   This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The “Document”, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as “you”. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

   A “Modified Version” of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

   A “Secondary Section” is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document’s overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

   The “Invariant Sections” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “Cover Texts” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “Transparent” copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “Opaque”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “Title Page” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

The “publisher” means any person or entity that distributes copies of the Document to the public.

A section “Entitled XYZ” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “Acknowledgements”, “Dedications”, “Endorsements”, or “History”.) To “Preserve the Title” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

   If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

   If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

   If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

   It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

   You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

   A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

   You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

   The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

   In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. COLLECTIONS OF DOCUMENTS

   You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

   You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

   A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

   If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

   Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

   If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

   You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

   However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

   Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

   Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

    The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

    Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

    "Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

    "CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

    "Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

    An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

    The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

```
Copyright (C)  year  your name.
Permission is granted to copy, distribute and/or modify this document
under the terms of the GNU Free Documentation License, Version 1.3
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover
Texts.  A copy of the license is included in the section entitled ``GNU
Free Documentation License''.
```

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with. . . Texts." line with this:

```
with the Invariant Sections being list their titles, with
the Front-Cover Texts being list, and with the Back-Cover Texts
being list.
```

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Index