Documentation of eQTLHap (V0.1)

Ziad Al Bkhetan

Email: Ziad.albkhetan@gmail.com

Last updated: 10 July 2020

Contents

1 Introduction

eQTLHap is a software to conduct a comprehensive Expression Quantitative Trait Loci (eQTL). eQTLHap investigate genomics variations considering three different representations:

1. Single SNP: Similar to standard eQTL analysis, the association between gene expression and SNPs (individually) is assessed statistically using linear regression.
2. Haplotype blocks: Association between gene expression and haplotype blocks is assessed statistically using multiple linear regression.
3. Genotype blocks: Association between gene expression and genotype blocks is assessed statistically using ANOVA regression.

eQTLHap is implemented in R and it depends on matrices operations to calculate correlation coefficients similar to the ultra-fast Matrix eQTL (`http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/`) to achieve high speed.

2 License

3 Running eQTLHap with default parameters

```
$ Rscript eQTLHap.R
            -f path_to_phased_haplotypes_in_shapeit_format
            -g path_to_gene_expression_file_bed_format
            -b path_to_haplotype_blocks_plink_det_format
            -o path_to_output_files
```

By default, eQTLHap does the following:

1. eQTLHap decides on the processed chromosome as the first chromosome in the haplotype/genotype file. However, this chromosome can be changed using the option **--chrm**.

2. SNPs and gene expression are filtered to keep data related to the chromosome chosen in (1).

3. eQTLHap filters individual keeping the common ones between haplotype, gene expression, and covariate (if provided) files. individual IDs are obtained as follow:

   - Haplotype/genotype file: Individuals' IDs in the header line of VCF file that starts with #CHROM. Individuals' IDs in the second column of .sample file for SHAPEIT format input file.
   - Gene expression file: Individuals' IDs in the header line (the first line).
   - Covariate file: Individuals' IDs in the header line (the first line).

4. Blocks are determined for each gene when they are located within this range [gene_start - scanning_window, gene_start + scanning_window]. Where gene_start is the column provided in gene expression file and scanning window is provided through the option **-w, --window** (default is 1000,000).

5. SNPs with MAF < 0.01 are ignored. This can be customised through the option **--smf**.

6. The statistical assessment is conducted for each block (all SNPs between the first and last SNPs of each block). The assessment considers the following: The block represented through its haplotypes, The block represented through its genotypes, Block's SNPs individually. For block assessment all haplotypes/genotypes with frequency less than 0.02 are considered the same. This will reduce the unique haplotype/genotype within each block. The threshold can be tuned using the options: **--hmf** and **--gmf**.

7. For each gene, p-values of blocks and SNPs are corrected using Benjamini-Hochberg (BH) method and associations with pvalue < 0.05 are recorded for the output files. correction method can be changed using the option **--mtc**.

4 Required parameters

Option	Type	Details
-f or --haps	String	Phased haplotypes in SHAPEIT format (.haps/.sample). The complete path of the .haps file should be provided, however, .sample file should be in the same location and with the same name as .haps file. Users can also provide a VCF file but it requires the flag --vcf to be enabled. Files can be gzipped (.gz).
-g or --genes	String	The path for gene expression file in bed format.
-b or --blocks	String	The path for haplotype blocks file. This parameter is mandatory when block assessment is required.
-o or --out	String	The path for output RData files.

5 Optional parameters

Option	Type	Default	Details
-c , --cov	String	""	The path for covariates file.
--chrm	Number	-1	The path for covariates file.
--mtc	String	Benjamini-Hochberg (BH)	Multiple test correction method. It accepts any value from p.adjust.methods.
-p, --permutation	Number	1,000	The number of permutations for permutation-based multiple test correction.
-w, --window	Number	1000,000	scanning window up/down transcription start site (TSS).
-a, --assessment	String	HSG	Assessment type, eQTLHap takes any combination of the letters S, G and H. where S: single SNP assessment. G: block's genotype. H: block's haplotype. The main purpose of eQTLHap is to conduct block assessment. If you are interested in single SNP eQTL, eQTL Matrix (`http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/`) can be better solution as it is faster.
--vcf	Boolean	FALSE	Flag to process VCF file for phased haplotype file provided by -f.
--smf	Number	0.01	SNP minimum frequency to be included in the analysis.
--hmf	Number	0.02	Haplotype minimum frequency to be included in the analysis.
--gmf	Number	0.02	Genotype minimum frequency to be included in the analysis.
--maxPval4Perm	Number	0	Maximum p-value for an association to be passed to permutation-based multiple test correction. t the threshold 0 means no permutation test will be applied.
--rmvIndividuals	Boolean	FALSE	When block assessment is applied and this option is enabled, individuals with rare haplotypes (freq < --hmf), rare genotypes (freq < --gmf) will be eliminated from the assessment.
--minIndividuals	Number	50	Minumum individuals count to perform a statistical assessment.
--outSignifcancePval	Number	0.05	Maximum p-value for the association to be reported in the output files.
--outSignifcanceQval	Number	1	Maximum corrected p-value for the association to be reported in the output files.
--outSignifcancePerm	Number	1	Maximum permutation p-value for the association to be reported in the output files.
--customBlocks	Boolean	FALSE	A flag to provide a custom block (a subset of the SNPs within the block) instead of considering the complete block (all SNPs).
--unphased	Boolean	FALSE	It is needed when there is no haplotype-based eQTL analysis and input VCF file is unphased.

6 Input file format

6.1 Haplotype/Genotype input file

Phased haplotype can be provided to eQTLHap in VCF and SHAPEIT format. Files can be compressed (.gz).
Details about SHAPEIT format are available at: `https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#hapsample`.
Details about VCF format are available at: `https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/`.

By default, eQTLHap assumes that alleles in the input file are phased and it is important to separate VCF file alleles using "|" separator. However, VCF file can be provided unphased ("/" separator) to conduct eQTL analysis based on the genotypes of SNPs and blocks as explained in eQTLHap paper. In this case, it is important to enabling the flag **--unphased**.

When dealing with SHAPEIT format (.haps and .sample files), eQTLHap assumes that both files are in the same location with the prefix (prefix.haps and prefix.sample).

6.2 Gene expression input file

Gene expression file is assumed to have bed format, where lines represent genes and columns represent individuals. There must be four columns for gene information at the beginning of each line. For eQTLHap, the following columns are mandatory: chromosome, gene name, and start position. eQTLHap assumes the following order for these columns: Chromosome, gene name, start, ignored-column. In case, the provided gene expression file has these columns in a different order, the user can adjust the order list (the variable gene_expression_info_columns) in the eQTLHap script. For example, if the provided file contains these columns in this order: (start, fourth_columns, chromosome, gene name), the user should change the order list to the following:
$gene\_expression\_info\_columns = c(3, 4, 1, 2)$.
By doing this, eQTLHap reorders these columns before starting the analysis. See the gene expression file in the provided example and replicate the same format if needed.

6.3 Block input file

By default, eQTLHAP accepts block file (plink.blocks.det) obtained from PLINK software: `https://www.cog-genomics.org/plink/1.9/ld#blocks`. eQTLHap defines the block including all SNPs between the first and last SNPs in the column "SNPS". Blocks can be customised by enabling the flag **--customBlocks**. This flag makes eQTLHap includes only the SNPs provided in the column "SNPS". The SNPs in the column "SNPS" must be separated by "|".

It is also possible to provide another block file as long as it has the following columns: CHR, SNPS. eQTLHAP ignores other columns. "SNPS" columns should contain the block's SNPs separated by "|". SNPs names must match the names written in the haplotype/genotype file.

See the examples of different block files "./comparison/example/blks.det" and "./comparison/example/blks_customised.det"

6.4 Covariates input file

This file is optional. When provided, it is assumed to have the covariates in each line and the individuals in the columns. The first column is the covariate name.
See the example of covariate file at "./comparison/example/cov_test".

7 Output file format

By default, the output is three RData files: *-block-haplotype.RData, *-block-genotype.RData, and *-snp.RData. Each file contains the significant associations for each assessment (block's haplotype, block's genotype and SNPs individually).

7.1 Block-based results

The data frame of block results contain the following information:

1. **block_id**: The of the block as a number representing the order of the block in the block file provided as an input. The block_id = 1 is the first block in the file and so on.

2. **predictors**: the number of unique haplotype/genotypes considered in the regression model.
3. **individuals**: The number of individuals included in the association assessment. By default, it is equivalent to the individual count included in the data, however, it can be less than the total individual number if the flag **--rmvIndividuals** is enabled.
4. **MR2**: Multiple R squared calculated for the association.
5. **pval**: The significance p-value of the association.
6. **qval**: The corrected p-value using the correction method provided as an option. By default, it is based on the Benjamini-Hochberg (BH) method.
7. **gene**: the gene name or id provided in the gene expression file.
8. **perm_pval**: Permutation based p-value if permutation test is enabled.

7.2 SNP-based results

The data frame of SNP-based results contains the following information:

1. **snp**: The SNP ID as written in the haplotype/genotype file.
2. **R2**: R squared calculated for the association.
3. **ttest**: t test calculated for the association.
4. **pval**: The significance p-value of the association.
5. **qval**: The corrected p-value using the correction method provided as an option. By default, it is based on Benjamini-Hochberg (BH) method.
6. **gene**: the gene name or id provided in the gene expression file.
7. **perm_pval**: Permutation based p-value if permutation test is enabled.

8 eQTLHap customisation

8.1 Include covariates in the analysis

The covariates file should contain the covariates in rows and the individuals in columns. The first column is the covariate ID. the first row is the header that contains the same individuals from the haplotype/gene expression files. to conduct comprehensive eQTL analysis including the covariates, run the following command:

```
$ Rscript eQTLHap.R
            -f path_to_phased_haplotypes_in_shapeit_format
            -g path_to_gene_expression_file_bed_format
            -b path_to_haplotype_blocks_plink_det_format
            -c path_to_covariates
            -o path_to_output_files
```

Considering the provided example, run the following:

```
$ Rscript eQTLHap.R
            --vcf
            -f ./comparison/example/snp_test.vcf
            -g ./comparison/example/gene_test.bed
            -b ./comparison/example/blks.det
            -c ./comparison/example/cov_test
            -o ./comparison/example/test
```

8.2 Multiple test correction through permutation test

Permutation pvalue can be calculated through eQTLHap by adjusting two options:

- **-p, --permutation**: The number of permutations t be applied, bu default it is 1,000.
- **--maxPval4Perm**: Associations with p-value greater than this threshold will not be considered for permutation test. Just to reduce computational time. We recommend 0.05 for this threshold.

When permutation test is enabled, a new column "perm_pval" will be added to the final output datasets. Perm_pval is calculated as following:

1. calculate p-value of the associations based on the correlation as descried in the manuscript (call it $pval_0$).
2. shuffle the gene expression data for (-p or –permutation) times. In each iteration, calculate the p-value of the association based on the shuffled data (call it $pval_i$).
3. calculated permutation based pvalue as the number of $pval_i < pval_0$ divided by the iteration number.

In general, permutation test is time consuming. It can be applied as follows:

```
$ Rscript eQTLHap.R
            -f path_to_phased_haplotypes_in_shapeit_format
            -g path_to_gene_expression_file_bed_format
            -b path_to_haplotype_blocks_plink_det_format
            -c path_to_covariates
            -o path_to_output_files
            --maxPval4Perm threshold
            -p permutation_number
```

Considering the provided example, run the following:

```
$ Rscript eQTLHap.R
            --vcf
            -f ./comparison/example/snp_test.vcf
            -g ./comparison/example/gene_test.bed
            -b ./comparison/example/blks.det
            -c ./comparison/example/cov_test
            -o ./comparison/example/test
            --maxPval4Perm 0.05
            -p 1000
```

8.3 Customised blocks

By default, eQTLHap determine the block based on the first and last SNPs in the column SNPS within the provided block file. For example, if the value of SNPS column is $a \mid b \mid e \mid f \mid g \mid z$, eQTLHAP will consder all SNPs between a and z within the haplotype/genotype file even if they include other SNPs that are not mentioned in SNPs column. If a user wants to limit the block to the specific SNPs mentioned in SNPS column, the flag **--customBlocks** should be enabled

```
$ Rscript eQTLHap.R
            -f path_to_phased_haplotypes_in_shapeit_format
            -g path_to_gene_expression_file_bed_format
            -b path_to_haplotype_blocks_plink_det_format
            -c path_to_covariates
            -o path_to_output_files
            --customBlocks
```

Considering the provided example, run the following:

```
$ Rscript eQTLHap.R
            --vcf
            -f ./comparison/example/snp_test.vcf
            -g ./comparison/example/gene_test.bed
            -b ./comparison/example/blks.det
            -c ./comparison/example/cov_test
            -o ./comparison/example/test
            --customBlocks
```

8.4 Eliminating rare haplotypes/genotypes

When conducting block-based eQTL assessment, eQTLHap considers all haplotypes/genotypes with frequency less than 0.02 (within the same block) the same. However, there is an option to eliminate such rare haplotypes and genotypes from the assessment by enabling the flag **--rmvIndividuals**. When this flag is enabled, eQTLHap detrmine the rare haplotypes/genotypes, then remove in individual who has one of these rare haplotypes/genotypes. If the remaining individuals are less than a threshold determined based on the option **--minIndividuals** (default is 50), eQTLHap skip this block from the assessment. Here is an example of how to use these options.

```
$ Rscript eQTLHap.R
            -f path_to_phased_haplotypes_in_shapeit_format
            -g path_to_gene_expression_file_bed_format
            -b path_to_haplotype_blocks_plink_det_for
            mat
            -c path_to_covariates
            -o path_to_output_files
            --rmvIndividuals
            --minIndividuals threshold
```

Considering the provided example, run the following:

```
$ Rscript eQTLHap.R
            --vcf
            -f ./comparison/example/snp_test.vcf
            -g ./comparison/example/gene_test.bed
            -b ./comparison/example/blks.det
            -c ./comparison/example/cov_test
            -o ./comparison/example/test
            --rmvIndividuals
            --minIndividuals 100
```

8.5 Output filtration

eQTLHap allows to filter the output results based on pvalue (default threshold is 0.05), qvalue (default threshold is 1) and permutation-based pvalue (default threshold is 1). This filtration can be tuned using the command below:

```
$ Rscript eQTLHap.R
            -f path_to_phased_haplotypes_in_shapeit_format
            -g path_to_gene_expression_file_bed_format
            -b path_to_haplotype_blocks_plink_det_format
            -c path_to_covariates
            -o path_to_output_files
            --outSignifcancePerm threshold1
            --outSignifcancePval threshold2
            --outSignifcanceQval threshold3
```

Considering the provided example, run the following:

```
$ Rscript eQTLHap.R
            --vcf
            -f ./comparison/example/snp_test.vcf
            -g ./comparison/example/gene_test.bed
            -b ./comparison/example/blks.det
            -c ./comparison/example/cov_test
            -o ./comparison/example/test
            --outSignifcancePerm 0.015
            --outSignifcancePval 0.05
            --outSignifcanceQval 0.05
```