eQTLA

(Expression Quantitative Trait Loci Analysis)

Version: **1.1**

Date: **17 June, 2013**

Author: **Boyko Kabakchiev**

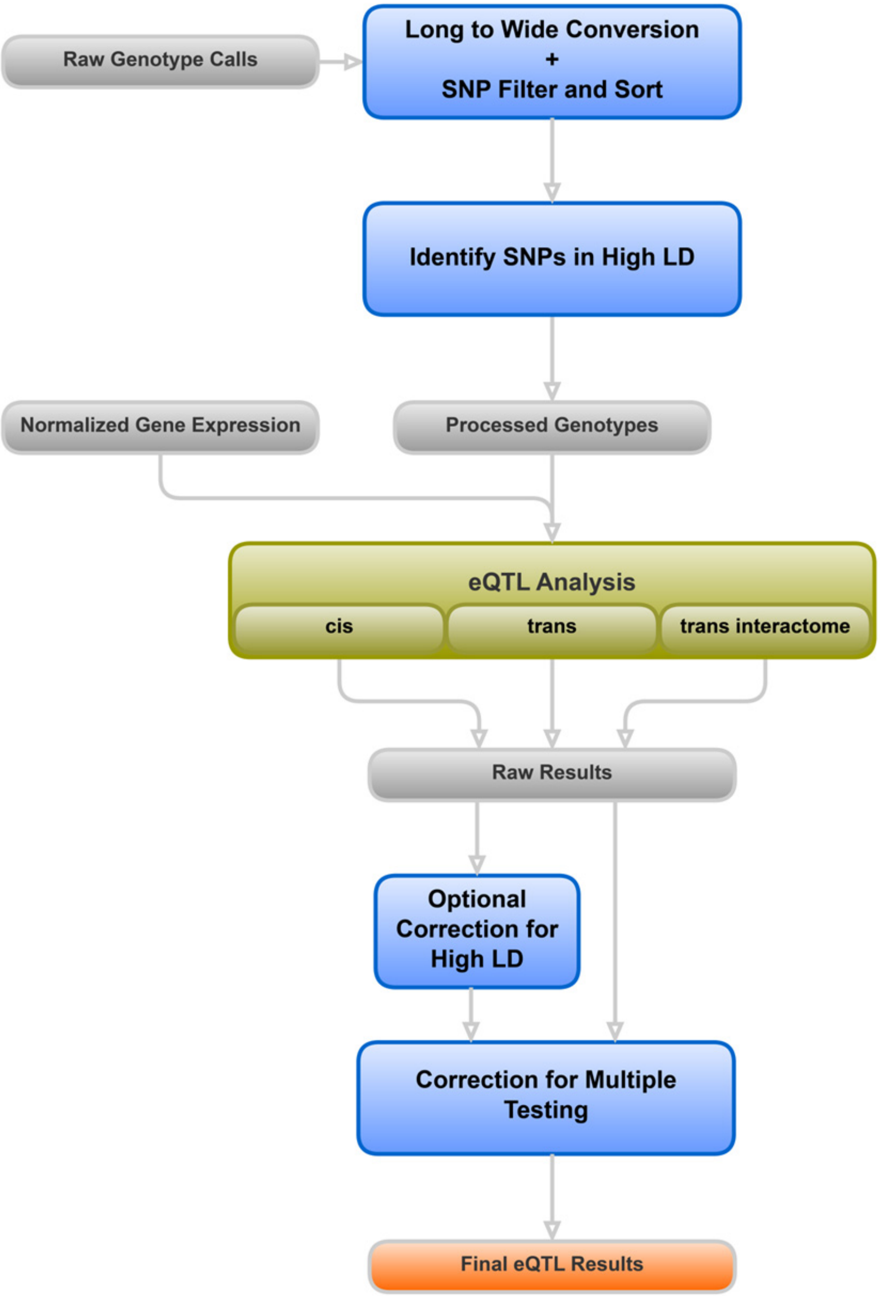
Introduction

The core motivation behind the design of eQTLA was to create a tool which can efficiently perform eQTL analysis on a single workstation with limited hardware resources.

eQTLA is capable of performing both cis- and trans-eQTL analysis given a set of raw genetic marker data and expression values. Currently a C++ implementation of the non-parametric Kruskal–Wallis one-way analysis of variance is the test of significance applied to genotype comparison, forgoing the need for normally distributed data. Nevertheless, the addition of other statistical tests to the framework of eQTLA is straightforward due to its object-oriented nature. Raw p-values are adjusted for multiple testing by either Bonferroni or Benjamini and Hochberg False Discovery Rate (FDR) correction.

An optional feature in eQTLA can identify markers in close proximity which are highly correlated due to apparent linkage disequilibrium (LD). This is accomplished by pairwise calculation of either D` or r^2 values between markers within a given genomic interval. A C++ implementation of Gregory Warnes's *LD* function from the R package *genetics* is used for this operation. Each group of markers in high LD is assigned a unique locus id. Post-hoc adjustment can be applied to eQTL data by amalgamating markers from the same locus and assigning the median $-\log_{10}$ converted p-value to the cluster.

Typical dataflow in eQTLA.



Contents

File Types

MAR

Files ending with **.mar** are one of the basic input containers for eQTLA and the product of running the *longtowide* command. These data structures contain all of the necessary genotype information used in later steps in a tabulation-separated format. The first row is a mandatory header, the first 3 columns of which must be *MarkerID*, *MarkerChr* and *MarkerPos*, followed by any number of sample IDs. Every row following the header corresponds to a single genetic marker. The rows must be sorted by chromosome and position of the markers in an ascending order. Listed genotypes can conform to either the ATCG notation (e.g. AA, TG, CT, --, -A, etc.) or AB notation (e.g. AA, AB, BB, --), but only one style can be used per file. An example is given below:

MarkerID	MarkerChr	MarkerPos	Sample1	Sample2	Sample3	Sample4	…
rs6426833	1	20171860	AA	AG	GG	AA	…
rs7702331	5	72551134	GG	AG	AA	AG	…
rs4836519	5	130017287	CC	TT	TC	TT	…
rs2413583	22	39659773	TC	TC	CC	TC	…
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	

MarkerID Marker IDs, such as rs numbers, should be listed in this column. These values must be unique

MarkerChr The corresponding chromosome for each genetic marker is listed in this column. Allowed values include 1, 2, 3, 4 ... , X, Y, M. If unknown, 0 can be used instead

MarkerPos The corresponding chromosomal position for each genetic marker is listed in this column. Allowed values include all positive integers. If unknown, 0 can be used instead

LMAR

Files ending with **.lmar** contain semi-processed genotype data by eQTLA. These data structures are the result of successfully running the *ldloci* command. The format of LMAR files is similar to that of MAR files with a couple of notable exceptions: the 4th column is reserved for a header called *Locus* and the genotypes are converted to digits. An example is given below:

MarkerID	MarkerChr	MarkerPos	Locus	Sample1	Sample2	Sample3	Sample4	…
rs6426833	1	20171860	1	2	1	5	2	…
rs7702331	5	72551134	25	5	1	2	1	…
rs4836519	5	130017287	25	4	3	1	3	…
rs2413583	22	39659773	80	1	1	4	1	…
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	

Locus A unique identifier for each set of markers belonging to the same locus based on a given LD cut-off and a base pair limit (see *ldloci* for more information)

RES

Files ending with **.res** contain the raw, tabulation-separated output from *ciseqtl*, *transeqtl* or *htranseqtl*. The format is the same for files generated by *ciseqtl* and *transeqtl*, but differs for files generated by *htranseqtl*. The first row is a mandatory, unmodifiable header. Examples of both formats are given below.

Generated by *ciseqtl* and *transeqtl*:

Gene	GChr	Marker	MChr	Position	Locus	Pval
2056	7	rs1734907	7	100315517	38	0.797273
3107	6	rs9264942	6	31274380	31	0.176425
55652	12	rs11168249	12	48208368	55	0.503319
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Gene — Gene IDs taken from the expression file are listed in this column

GChr — The corresponding chromosome for each gene is listed in this column

Marker — Marker IDs taken from the **.lmar** file are listed in this column

MChr — The corresponding chromosome for each genetic marker is listed in this column

Position — The corresponding chromosomal position for each genetic marker is listed in this column

Locus — Locus IDs carried over from the **.lmar** file are listed in this column

Pval — Raw p-values from the associative test between genotype and expression

Generated by *htranseqtl*:

IGene	IGChr	OGene	OGChr	Marker	MChr	Position	Locus	OPval	IPval
2770	7	10636	5	rs12654812	5	176794191	29	0.00037315	0.250698
2771	3	10636	5	rs12654812	5	176794191	29	0.00037315	0.737876
2773	1	10636	5	rs12654812	5	176794191	29	0.00037315	0.599989
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.

IGene Gene IDs of the interacting genes, taken from the expression file are listed in this column (see *htranseqtl* for more information)

IGChr The corresponding chromosome for each interacting gene is listed in this column

OGene Gene IDs of the original genes, taken from the expression file are listed in this column (see *htranseqtl* for more information)

OGChr The corresponding chromosome for each original gene is listed in this column

Marker Marker IDs taken from the **.lmar** file are listed in this column

MChr The corresponding chromosome for each genetic marker is listed in this column

Position The corresponding chromosomal position for each genetic marker is listed in this column

Locus Locus IDs carried over from the **.lmar** file are listed in this column

OPval Raw p-values from the associative test between genotype and the expression of the original gene

IPval Raw p-values from the associative test between genotype and the expression of the interacting gene

LRES

Files ending with **.lres** contain the LD-corrected, tabulation-separated output from *ldcorr*. The format of LRES files is similar to that of RES files. An example is given below:

Gene	GChr	Markers	LChr	Positions	Size	Locus	Pval
1	19	rs7257719	19	58819113	1	352451	0.734089
1	19	rs7359922/rs16988665	19	58830261/58840905	2	352452	0.846256
10	8	rs10103029/rs10088333/rs7816847	8	18267338/18267878/18275189	3	178687	0.100682
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Gene
Gene IDs taken from the **.res** file are listed in this column

GChr
The corresponding chromosome for each gene is listed in this column

Markers
Marker IDs taken from the **.res** file are listed in this column. The referenced markers belong to the same locus and are separated by a forward slash

LChr
The corresponding chromosome for each locus is listed in this column

Positions
The respective chromosomal positions for all genetic markers within a locus are listed in this column and separated by a forward slash

Locus
Locus IDs carried over from the **.res** file are listed in this column

Pval
LD-adjusted p-values are listed in this column. (see *ldcorr* for more information)

CRES

Files ending with **.cres** contain the multiple comparisons corrected, tabulation-separated output from *mtcorr* or *amtcorr*. The format of CRES files is identical to that of RES files, except that the rows are sorted in ascending order by raw *Pval* and an extra column is added to accommodate the corrected p-values. An example is given below:

Gene	GChr	Marker	MChr	Position	Locus	Pval	FDR_Pval
1521	11	rs2231884	11	65656564	52	4.11E-14	6.33E-12
254122	11	rs2231884	11	65656564	52	5.66E-13	4.36E-11
2524	19	rs516246	19	49206172	72	6.32E-13	3.25E-11
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Bonferroni_Pval / FDR_Pval
P-values adjusted for multiple comparisons by either the Bonferroni or FDR method are listed in this column (see *mtcorr* or *amtcorr* for more information)

EXP

Files ending with **.exp** are one of the basic input containers for eQTLA. These data structures contain all of the necessary gene expression information used in later steps in a tabulation-separated format. The first row is a mandatory header, the first column of which must be *Genes*, followed by the same number of sample IDs and in the exact same order as listed in the respective **.mar** file. Every row following the header corresponds to a gene or a transcript. An example is given below:

Gene	Sample1	Sample2	Sample3	Sample4	…
143384	9.726943	9.392158	9.672458	9.377743	…
7515	8.650414	8.829617	8.545529	8.743195	…
9525	9.025473	8.826889	9.234246	8.782236	…
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	

Gene

Gene or transcript IDs are listed in this column. These values are always treated as strings of characters. While it is not required that the listed IDs be unique, it is recommended in order to avoid bias later in the pipeline

GEN

Files ending with **.gen** are one of the basic input containers for eQTLA. These data structures contain all of the necessary gene coordinate information used in later steps in a tabulation-separated format. The first row is a mandatory header. Every row following the header corresponds to a gene or a transcript and its respective coordinates. An example is given below:

GeneID	Chr	ChrStart	ChrStop
100287934	1	721320	722513
310	10	75135189	75173841
6747	3	156257929	156272973
.	.	.	.
.	.	.	.
.	.	.	.

GeneID — Gene or transcript IDs are listed in this column. These values are always treated as strings of characters and should be unique

Chr — The corresponding chromosome for each gene is listed in this column.

ChrStart — The corresponding chromosomal position where each gene begins is listed in this column

ChrStop — The corresponding chromosomal position where each gene ends is listed in this column

GRA

Files ending with **.gra** are one of the basic input containers for eQTLA. These data structures contain all of the necessary gene interaction information used in later steps in a tabulation-separated format. The first row is a mandatory header. Every row following the header corresponds to a pair of genes defining an interaction. Each interaction is considered only once and duplicates are discarded. In addition, directionality is not taken into account (e.g GeneA-GeneB is equivalent to GeneB-GeneA). An example is given below:

GeneA	GeneB
1	10321
246565	32268
37433	43229

GeneA Gene or transcript IDs are listed in this column. These values are always treated as strings of characters

GeneB Gene or transcript IDs are listed in this column. These values are always treated as strings of characters

Commands

All commands in eQTLA can be executed from a compatible command-line interpreter, such as Command Prompt in Windows or a UNIX shell. This standard syntax can be evoked in the directory where eQTLA is located:

```
eqtla command –paramter1 “value1” –parameter2 “value2” …
```

Possible commands include: *longtowide*, *ldloci*, *ciseqtl*, *htranseqtl*, *transeqtl*, *ldcorr*, *mtcorr* and *amtcorr*. The following pages detail information on their function and use. Note that it is not necessary to enclose values in quotation marks unless punctuated by white spaces or tabulations and it is not necessary to supply parameters in a specific order.

longtowide

This command is not required in the eQTLA pipeline, but has been added to assist with conversions of long datasets into wide format. It is common for software capable of processing raw genotype data from Illumina, Affymetrix and other platforms to export results into fairly large files. These files are typically repetitions of each genotype once per every sample in the project, possibly with additional quality control information, spanning many lines of text. If the data is already sorted by sample ID and genetic marker ID, *longtowide* and its parameters can be used to convert such a file into an eQTLA compatible **.mar** file. (see File Types for more information)

Parameters:

-in Name of file containing long data

-out Name of output **.mar** file to which wide data will be written. Default value if omitted: “default.mar”

-header Position of header row. Value must be >0. Default value if omitted: “1”

-a1 Column of first allele. Value must be >0. Default value if omitted: “5”

-a2 Column of second allele. Value must be >0. Default value if omitted: “6”

-samples Name of file containing sample IDs. If provided, only listed sample IDs are included, otherwise all IDs are included. This file should contain one sample ID per line and a mandatory header at row one

-markers Name of file containing genetic marker IDs. If provided, only listed marker IDs are included, otherwise all IDs are included. This file should contain one marker ID per line and a mandatory header at row one

-sid Sample ID column. Value must be >0. Default value if omitted: “4”

-mid Genetic marker ID column. Value must be >0. Default value if omitted: “1”

-mchr Genetic marker chromosome column. Value must be >0. Default value if omitted: “2”

-mpos Genetic marker chromosomal position column. Value must be >0. Default value if omitted: “4”

Example of a long dataset:

```
[Header]
GSGT Version      1.8.4
Processing Date   12/29/2011 2:08 PM
Content           HumanOmniExpress-12v1_H.bpm
Num SNPs          730525
Total SNPs        730525
Num Samples       1852
Total Samples     1872
[Data]
      SNP Name   Chr    Position   Sample ID   Allele1 - Forward   Allele2 - Forward   GC Score
     rs1000000    12   126890980        4258                   T                   C     0.8228
     rs1000002     3   183635768        4258                   -                   -          0
    rs10000023     4    95733906        4258                   T                   T      0.836
```

Example of *longtowide* use:

```
eqtla longtowide –in “mylongdata.txt” –out “mywidedata.mar” –header 15 –a1 5 –a2 6 –samples
“use_these_samples.txt” –markers “use_these_markers.txt” –sid 4 -mid 1 –mchr 2 –mpos 3
```

ldloci

Commands performing associative analysis in later steps of the eQTLA workflow require that genotyping data be inputted through **.lmar** files. Therefore, *ldloci* should be run for each **.mar** file. (see File Types for more information) In essence, *ldloci* computes and adds locus IDs to each genetic marker. The markers which are given the same locus ID are determined by the apparent linkage disequilibrium and distance between them.

Parameters:

-in Name of **.mar** file

-out Name of output **.lmar** file to which data will be written. Default value if omitted: “default.lmar”

-cutoff Cut off for parameter determined by *–measure*. Values greater than or equal to this number will allocate two genetic markers on the same locus unless the distance between them is greater than *–bplim*. Value must be ≥0.0 and ≤1.0. Default value if omitted: “0.0”

-type Designates data type in **.mar** file. Allowed values are “G” for ATCG notation and “A” for the AB notation. Default value if omitted: “G”

-measure Determines type of LD measure to be used. Allowed values are “R” for r^2 and “D” for D’. Default value if omitted: “R”

-bplim Maximum size of locus in base pairs. Genetic markers farther away from each other than this distance will not be included on the same locus regardless of LD. Value must be greater than “0”. It is recommended to supply values equal to or greater than the average gene length added to twice the window size used in later eQTL analyses. Default value if omitted: “1000000000”

Example of *ldloci* use:

```
eqtla ldloci –in “mywidedata.mar” –out “mywidedata.lmar” –cutoff 0.5 –type “G” –measure “R”
–bplim 200000
```

ciseqtl

In order to perform a cis-eQTL analysis given a set genetic markers and gene expression values, the *ciseqtl* command should be used. For this type of analysis, only genetic markers positioned within the chromosomal coordinates of a gene or within extensions upstream and downstream of these coordinates as defined by the –*window* parameter are considered. Additional quality control measures can be set through built-in parameters. Currently, Kruskal–Wallis one-way analysis of variance is used to test for association.

Parameters:

-min Name of **.lmar** file

-ein Name of **.exp** file (see File Types for more information)

-gin Name of **.gen** file (see File Types for more information)

-out Name of output **.res** file to which data will be written. Default value if omitted: “default.res”

-window Determines maximum distance upstream and downstream from gene coordinates to include genetic markers. Value must be ≥0. Default value if omitted: “50000”

-crmin Sets minimum call rate for included genetic markers. Markers with values below this limit are skipped. Value must be ≥0.0 and ≤1.0. Default value if omitted: “0.95”

-mafmin Sets minimum minor allele frequency for included genetic markers. Markers with values below this limit are skipped. Value must be ≥0.0 and ≤1.0. Default value if omitted: “0.05”

-hwchi2max Sets maximum allowed Hardy-Weinberg Equilibrium χ^2 value. Markers with values greater than this limit are skipped. Value must be ≥0.0. Default value if omitted: “6.635”

Example of *ldloci* use:

```
eqtla ciseqtl –min “mywidedata.lmar” –ein “myexpression.exp” –gin “human.gen” –out
“myciseqtlresult.res” –window 50000 –crmin 0.95 –hwchi2max “19.511”
```

htranseqtl

This form of trans-eQTL analysis builds upon the cis-eQTL principle. Only genetic markers positioned within the chromosomal coordinates of a given gene or within extensions upstream and downstream of these coordinates as defined by the *–window* parameter are considered. However, associative analysis using Kruskal–Wallis one-way analysis of variance is performed between these markers and the expression of all genes with which the original one interacts. These interactions must be supplied through the *–graph* parameter. Additional quality control measures can be set through built-in parameters.

Parameters:

-min Name of **.lmar** file

-ein Name of **.exp** file (see File Types for more information)

-gin Name of **.gen** file (see File Types for more information)

-out Name of output **.res** file to which data will be written. Default value if omitted: “default.res”

-window Determines maximum distance upstream and downstream from gene coordinates to include genetic markers. Value must be ≥0. Default value if omitted: “50000”

-crmin Sets minimum call rate for included genetic markers. Markers with values below this limit are skipped. Value must be ≥0.0 and ≤1.0. Default value if omitted: “0.95”

-mafmin Sets minimum minor allele frequency for included genetic markers. Markers with values below this limit are skipped. Value must be ≥0.0 and ≤1.0. Default value if omitted: “0.05”

-hwchi2max Sets maximum allowed Hardy-Weinberg Equilibrium χ^2 value. Markers with values greater than this limit are skipped. Value must be ≥0.0. Default value if omitted: “6.635”

-graph Name of **.gra** file (see File Types for more information)

Example of *htranseqtl* use:

```
eqtla htranseqtl –min “mywidedata.lmar” –ein “myexpression.exp” –gin “human.gen” –out
“myhtranseqtlresult.res” –window 50000 –crmin 0.95 –hwchi2max “6.635” –graph “human.gra”
```

transeqtl

This command performs conventional trans-eQTL analysis. Only genetic markers positioned outside of the chromosomal coordinates of a given gene and outside of extensions upstream and downstream of these coordinates as defined by the *–window* parameter are considered. Additional quality control measures can be set through built-in parameters. Currently, Kruskal–Wallis one-way analysis of variance is used to test for association.

Parameters:

-min Name of **.lmar** file

-ein Name of **.exp** file (see File Types for more information)

-gin Name of **.gen** file (see File Types for more information)

-out Name of output **.res** file to which data will be written. Default value if omitted: “default.res”

-window Determines minimum distance upstream and downstream from gene coordinates to include genetic markers. Value must be ≥0. Default value if omitted: “50000”

-crmin Sets minimum call rate for included genetic markers. Markers with values below this limit are skipped. Value must be ≥0.0 and ≤1.0. Default value if omitted: “0.95”

-mafmin Sets minimum minor allele frequency for included genetic markers. Markers with values below this limit are skipped. Value must be ≥0.0 and ≤1.0. Default value if omitted: “0.05”

-hwchi2max Sets maximum allowed Hardy-Weinberg Equilibrium χ^2 value. Markers with values greater than this limit are skipped. Value must be ≥0.0. Default value if omitted: “6.635”

Example of *transeqtl* use:

```
eqtla transeqtl –min “mywidedata.lmar” –ein “myexpression.exp” –gin “human.gen” –out
“mytranseqtlresult.res” –window 50000 –crmin 0.95 –hwchi2max “19.511”
```

ldcorr

Optional correction for high LD can be applied to **.res** files generated by *ciseqtl* and *transeqtl* with the *ldcorr* command. Genetic markers on the same locus as determined by *ldloci* are amalgamated and the median p-value is assigned to the cluster. For even-sized clusters, the median is computed by the following formula: $\tilde{p} = \left(\frac{\log_{10}(p_i)+\log_{10}(p_{i+1})}{2}\right)^{10}$, where p_i and p_{i+1} are the two middle ranking p-values.

Parameters:

-in Name of **.res** file

-out Name of output **.lres** file to which data will be written. Default value if omitted: “default.lres”

Example of *ldcorr* use:

```
eqtla transeqtl –in “myciseqtlresult.res” –out “my_ldcorr_ciseqtlresult.lres”
```

mtcorr

Optional correction for multiple testing can be applied to **.res** and **.lres** files generated by *ciseqtl*, *htranseqtl*, *transeqtl*, and *ldcorr* with the *mtcorr* command. An extra column containing corrected p-values is added to the input files.

Parameters:

-in Name of **.res** or **.lres** file

-out Name of output **.cres** file to which data will be written. Default value if omitted: “default.cres”

-method Determines correction method to be used. Allowed values are “B” for Bonferroni and “F” for FDR. Default value if omitted: “B”

Example of *mtcorr* use:

```
eqtla mtcorr –in “mytranseqtlresult.res” –out “my_FDRcorr_transeqtlresult.cres” –method “F”
```

amtcorr

This command performs identically to *mtcorr*. The only difference is that *amtcorr* is capable of correcting a fraction of the total number of p-values. In certain situations, it may not be possible to read all raw p-values into the computer's memory. By supplying a p-value cut-off, *amtcorr* corrects only raw p-values below this limit, but takes into account the total number of p-values during the correction process.

Parameters:

-in Name of **.res** or **.lres** file

-out Name of output **.cres** file to which data will be written. Default value if omitted: "default.lres"

-method Determines correction method to be used. Allowed values are "B" for Bonferroni and "F" for FDR. Default value if omitted: "B"

-pcutoff Sets raw p-value cut-off. Value must be ≥0.0 and ≤1.0. Default value if omitted: "0.05"

Example of a*mtcorr* use:

```
eqtla amtcorr –in "mytranseqtlresult.res" –out "my_FDRcorr_transeqtlresult.cres" –method "F"
–pcutoff 0.05
```

Troubleshooting

In order to guarantee maximum compatibility, is it highly recommended that users of eQTLA recompile the application on each new system. However, if the provided precompiled, 32bit executable for Windows-based systems is used, it is possible that error messages pertaining to missing .dll files will be generated by the operating system. These issues can be resolved by installing **MinGW** from http://www.mingw.org/ .

For questions and suggestions regarding eQTLA, please contact Boyko Kabakchiev at kabakchiev (at) lunenfeld (dot) ca .