

eQTL Analysis by Linear Model

Wei Sun

November 19, 2009

1 Overview

```
> library(eMap)
```

This vignette describes eQTL mapping, visualization and other analysis tools in **R/eMap** package. To improve efficiency, we implemented the computational intensive parts of our algorithms by C using GSL library (<http://www.gnu.org/software/gsl/>). Therefore in order to install this R package, GSL library need to be installed. Users may need to change the Makevars file in src folder to provide the correct location for header file and library file of GSL.

2 eQTL mapping

2.1 Linear regression

To maintain a reasonable power given limited sample size and multiple testing correction in eQTL studies, the smallest model with only additive genetic effect is often used to map eQTL [Stranger et al., 2007]:

$$y = a + bx + \epsilon,$$

where y indicates a gene expression trait and x indicates the additive genetic effect, which can be coded by the number of minor alleles, and ϵ is the residual error. Sometime two genetic effects need to be modeled simultaneously. For example, if we want to model both additive or dominant effects, or total copy number and allelic copy number effects:

$$y = a + b(nA + nB) + c(nA - nB) + \epsilon,$$

where nA and nB indicates the number of A and B alleles of a SNP.

This linear regression approach with up to tow genetic effects is implemented in function **eMap1**:

```
eMap1(me, mm1, mm2=NULL, output.tag, p.cut, cis.only=FALSE, cis.distance=1e6,  
      eChr=NULL, ePos=NULL, mChr=NULL, mPos=NULL, tol=1e-7,  
      nna.percent=0.75)
```

Here **me**, **mm1**, and **mm2** are matrices of gene expression data and two types of genotype data respectively, with each row corresponds to one gene/marker and each column corresponds to one individual. The result of this function are written into two files, **tag_eqtl.txt** and **tag_freq.txt**, where tag is specified by **output.tag**. Only the results with p-values smaller than **p.cut** are kept. The following is an example of **tag_eqtl.txt**:

geneID	markerID	a	a_p	b	b_p	N
1	4516	-0.374	7.21e-03	0.661	7.74e-05	59
1	18632	-0.420	3.88e-03	0.687	8.83e-05	63
...						

where `geneID` and `markerID` are the row IDs of genes and markers in matrices `me` and `mm` respectively, `a_b` and `b_p` are p-values for coefficient a and b respectively, and N is the sample size after excluding missing values. The file `tag_freq.txt` includes the frequencies of the p-values so that although only those p-values smaller than `p.cut` are recorded, we still have the distribution information for all the p-values.

If `cis.only=TRUE`, only cis-eQTL computation is carried out, and an eQTL is defined as cis-eQTL if the distance between the gene and the marker is smaller than `cis.distance`. In order to identify cis-eQTL, we need the chromosome locations of all the genes and markers, which are specified by `eChr`, `ePos`, `mChr`, and `mPos`, respectively. Notice the expression/genotype data are matched to their location information based on row ID. Therefore, `eChr` should be a vector of which the length equals to the number of rows of `me`. Similar length restrictions apply to `ePos`, `mChr`, and `mPos`. It is assumed that `mm1` and `mm2` have the same number of rows and the same row of `mm1` and `mm2` corresponds to the same genetic marker. It is also assumed that both `eChr` and `mChr` are integer vectors, so chromosome X, Y should be coded as numbers.

2.2 Linear regression with permutation

Multiple testing correction is implemented in two levels. First, given a gene, multiple tests across markers are corrected by evaluating permutation p-value. Secondly, multiple tests across genes are corrected by choosing a permutation p-value cutoff based on False Discovery Rate (FDR) estimates. In this R package, we focus on the calculation of permutation p-value. Once we obtain the permutation p-values, FDR can be estimated by existing method such as `R/qvalue`.

The function that carry out permutations is defined as follows:

```
eMap1P(me, mm1, mm2=NULL, ptype, output.tag, n.permute=10,
p.cut=1e-4, cis.only=FALSE, cis.distance=1e6, eChr=0, ePos=0,
mChr=0, mPos=0, trace.it=TRUE, tol=1e-7, nna.percent=0.75,
np.max=100000, np=c(100, 1000, 5000, 10000, 50000),
aim.p=c(0.1, 0.05, 0.02, 0.01, 0.002), confidence.p=0.0001,
permute.grp=NULL)
```

In addition to the parameters in function `eMap1`, this function has a few additional parameters for permutations. If `ptype=1`, it simply runs the function `eMap1` K times using permuted gene expression data, where $K = \text{n.permute}$. In each permutation, both `tag_eqtl.txt` and `tag_freq.txt` are written out. If `ptype=2`, we carry out adaptive permutations with the rules specified by parameters `np.max`, `np`, and `aim.p`. Specifically, we stop the permutation if there is strong evidence that the permutation p-value is bigger than `aim.p[i]` after `np[i]` permutations, otherwise permute at most `np.max` times. The strong evidence means the binomial p-value of the test

$$H_0 : \text{permutation p-value} = \text{aim.p}[i] \text{ vs. } H_1 : \text{permutation p-value} > \text{aim.p}[i]$$

is smaller than `confidence.p`. When `ptype=2`, the output is also written into a file. Here is an example of the output:

geneID	markerID	pval	permuteP	npermute
1	42729	1.4e-05	0.75	20
2	63293	8.6e-07	0.114	1000
...				

where `geneID` and `markerID` are the row IDs of genes and markers in matrices `me` and `mm` respectively, `pval` and `permuteP` are p-values and permutation p-values respectively, and `npermute` is the number of permutations used.

User can also specify a categorical parameter `permute.grp` such that all the permutations are carried out with each category of this parameter. This conditional permutation is useful to control some confounding effect such as race or sex effects.

2.3 Multiple linear regression

For eQTL studies with enough samples, both additive and dominant genetic effects can be considered. Specifically, we refer to the coding “(AA, AB, BB) $\rightarrow$ (-1, 0, 1)” as additive effect, and the coding “(AA, AB, BB) $\rightarrow$ (-1, 1, -1)” as dominant effect. The additive and dominant effects are complementary to each other. The additive effect is actually a linear effect proportional to the number of either A or B alleles, and the dominant effect captures the deviation from the linear effect. Note the dominant effect does not correspond to the dominant inheritance, for example, dominant inheritance of allele B corresponds to the coding of “(AA, AB, BB) $\rightarrow$ (a, b, b)”, where $a \neq b$. A dominant inheritance can be represented by a combination of additive and dominant effects.

As a confounding variable, sex effect often need to be taken into account in eQTL studies. Besides the binary variable sex, which is present in every model to control the sex main effect, there are four other possible predictors, two genetic main effects, additive effect (add) and dominant effect (dom), and two interaction effects, add:sex and dom:sex. Here we use X:Y to indicate the interaction between X and Y. Thus there are altogether 15 models with at least one genetic effect. The significance of each model can be assessed by comparing it with the baseline model $y = \mu + \beta(\text{sex}) + \epsilon$ by likelihood ratio test.

We reduce the 15 models to 5 models (equation (1)-(5)) based on the following three principles: (1) If an interaction term is included in the model, the corresponding main effects should also be included. (2) If the dominant effect is included in the model, the additive effect should also be included. (3) If the interaction term sex:dom is included in the model, the sex:add effect should also be included. The first principle is a commonly used criteria for model selection. Principle (2) and (3) are biologically meaningful because the dominant effect by itself means the two homozygous genotypes have the same effect which is different from the effect of heterozygous genotype. This is rarely the case, and any deviation from this situation requires the presence of additive effect. The commonly encountered dominant, recessive, or co-dominant inheritance patterns all require both the additive and dominant effects.

$$\text{model 1: } y = \mu_1 + \beta_{11}(\text{sex}) + \beta_{12}(\text{add}) + \epsilon_1 \quad (1)$$

$$\text{model 2: } y = \mu_2 + \beta_{21}(\text{sex}) + \beta_{22}(\text{add}) + \beta_{24}(\text{sex:add}) + \epsilon_2 \quad (2)$$

$$\text{model 3: } y = \mu_3 + \beta_{31}(\text{sex}) + \beta_{32}(\text{add}) + \beta_{33}(\text{dom}) + \epsilon_3 \quad (3)$$

$$\begin{aligned} \text{model 4: } y = \mu_4 + \beta_{41}(\text{sex}) + \beta_{42}(\text{add}) + \beta_{43}(\text{dom}) \\ + \beta_{44}(\text{sex:add}) + \epsilon_4 \end{aligned} \quad (4)$$

$$\begin{aligned} \text{model 5: } y = \mu_5 + \beta_{51}(\text{sex}) + \beta_{52}(\text{add}) + \beta_{53}(\text{dom}) \\ + \beta_{54}(\text{sex:add}) + \beta_{55}(\text{sex:dom}) + \epsilon_5 \end{aligned} \quad (5)$$

In our implementation (function `lmEQTL`), we either output the results of all the five models or select model by a backward procedure (Figure 1). Motivated by the observation that in many cases, the additive and/or sex:add effects are significant, while the dominant and sex:dom effects are not [Sun, 2007], we also allow the option of only considering additive and sex:add effects, i.e., model (1) and (2).

Currently adaptive permutation has not been implemented for this linear regressions with sex effects. A fixed number of permutations can be carried out by functions `eMap2P` and `lmEQTL.byChr.permute`. Merging permutations across genes saves a significantly amount of computation time, and can provide an initial estimate of the overall significance. However, as suggested by empirical studies, a large number of gene-specific permutations is more appropriate [Carlborg et al., 2005].

3 eQTL visualization

The function, `ePlot` which visualize the eQTL results, is defined as follows

```
ePlot(geneID, markerID, scores, scuts, cols, eChr, ePos, mChr,
```

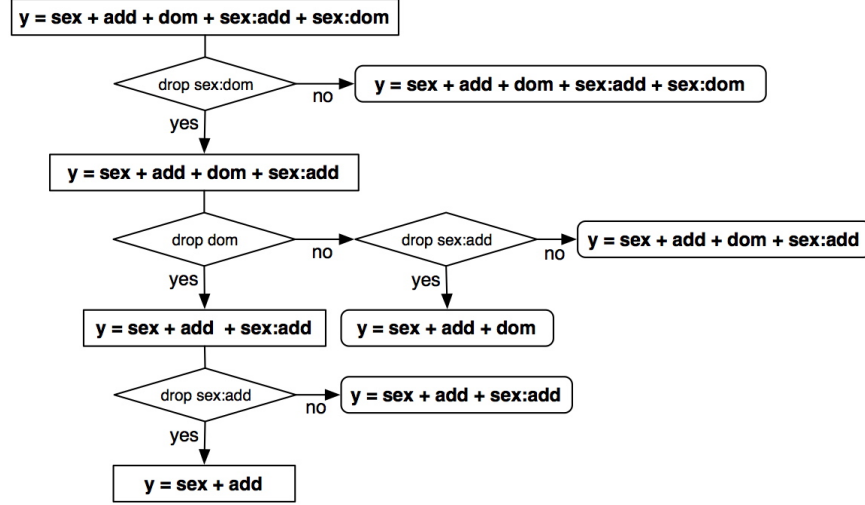


Figure 1: Backward linkage model selection. The default model is model 1: $y = \mu_1 + \beta_{11}(\text{sex}) + \beta_{12}(\text{add}) + \epsilon_1$. Comparing with the forward model selection procedure in a previous publication [Wang et al., 2006], our backward model selection procedure aims to find smaller models, e.g., model 1 instead of model 3, model 2 instead of model 4. The decision of dropping a term is made based on regular likelihood ratio test for comparing nested models.

```

mPos, chroms, xlab="QTL Location", ylab="Transcript Location",
plot.hotspots=TRUE, hotspots.cut=10, score.type="p-value"),

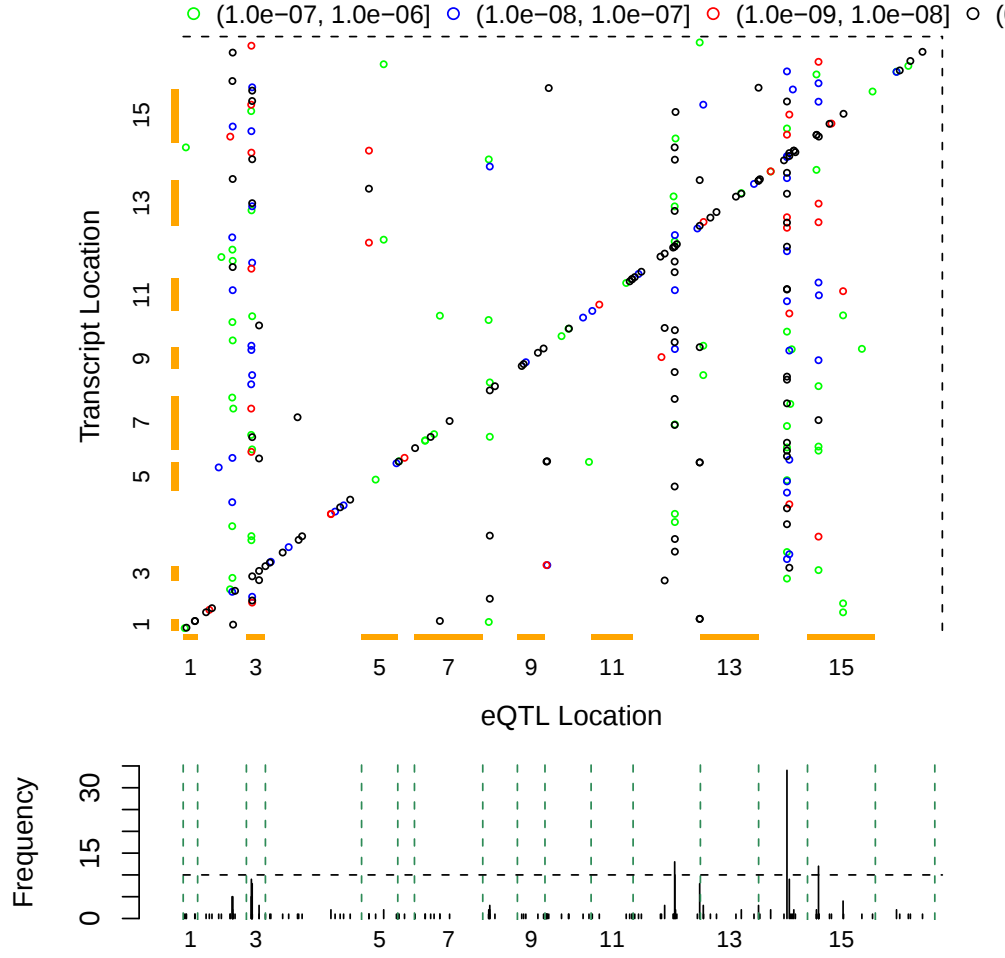
```

where `geneID`, `markerID` and `scores` are three vectors of same length. One triplet of `geneID[i]`, `markerID[i]`, and `scores[i]` indicates an association between a gene ($\text{ID}=\text{geneID}[i]$) and a marker ($\text{ID}=\text{markerID}[i]$), with p-value equals to `scores[i]`. Note here ID is just row ID of gene/marker in the gene expression/marker genotype data matrix, and `eChr`, `ePos`, `mChr`, and `mPos` are vectors including location information for all the genes/markers, with the same order as the original data matrices. For example, the information of gene with $\text{ID}=\text{geneID}[i]$ is stored at `eChr[geneID[i]]`, and `ePos[geneID[i]]`. The scores can be either p-value or regression coefficients, which is specified by parameter `score.type`. Parameter `scuts` and `cols` specify the colors for eQTL results of different significance levels. For example, if `score.type = "p.value"`, `scuts = c(1e-4, 1e-5)`, and `cols = c("green", "blue")`, then all the associations with p-values within $(1e-5, 1e-4]$ are plotted as green points and all the associations with p-values smaller or equal to $1e-5$ are plotted as blue points. An example is listed below using the eQTL data of Brem and Kruglyak (2005).

```

> data(eqt1.y112)
> data(eInfo.y112)
> data(mInfo.y112)
> eq = eqt1.y112
> mI = mInfo.y112
> eI = eInfo.y112
> ePlot(eq$geneID, eq$markerID, eq$pValue, scuts = c(1e-06, 1e-07,
+ 1e-08, 1e-09), cols = c("green", "blue", "red", "black"),
+ eChr = eI$chr, ePos = 0.5 * (eI$start + eI$end), mChr = mI$chr,
+ mPos = mI$start, chroms = 1:16, xlab = "eQTL Location", ylab = "Transcript Location",
+ plot.hotspots = TRUE, hotspots.cut = 10)

```



As shown in Figure 2, the eQTL results are visualized a scatter plot with marker location on the X-axis and gene location on the Y-axis (function `eplot`). Different levels of significance are plotted using different colors. At the bottom of the scatter plot, we also plot the number of genes linked to each marker.

4 eQTL module

An eQTL hotspot is a small segment of DNA sequence that harbors the eQTL of multiple genes, and we define an eQTL module as an eQTL hotspot together with the associated genes. eQTL modules can be detected by moving window of constant size or by a hypothesis testing method which allows variable hotspot sizes [Sun et al., 2007] (function `eModule`). Another question is how to quantify the significance of an eQTL module. If all the genes in an eQTL module are independent, the joint significance can be measured by the product of individual p-values. However, the independence assumption is obviously not correct and it exaggerates the significance level. Suppose there are T genes in an eQTL module, of which the expressions are highly correlated regardless of the DNA variation. Then as long as one of them is associated with a marker by chance, the other genes are probably associated with the marker as well. Based on this concern, we propose to evaluate the significance of an eQTL module as follows. First, identify a representative genotype profile in the eQTL hotspot; record it as m_c . Then find the gene that is most significantly associated with m_c ; record it as $e_{(1)}$ and the corresponding p-value as p_1 . Next, from all the remaining genes, find the gene with the most significant relation with m_c , given $e_{(1)}$ (by linear model $y = \beta_0 + \beta_1 e_{(1)} + \gamma m_c + \epsilon$); record it as $e_{(2)}$ and the corresponding p-value as p_2 . Then include both $e_{(1)}$ and $e_{(2)}$ into the model to identify the third gene. Repeat this procedure until a p-value cutoff q (e.g., $q=0.01$) is met, then we can calculate a statistics for the eQTL module as $\sum_{p_i \leq q} \log(p_i)$ (function `moduleScore`). An example is shown below for

an eQTL module on chromosome 15 including 122 genes.

```
> data(eData.y112)
> data(mData.y112)
> data(mInfo.y112)
> data(eqt1.y112)
> eqt1 = eqt1.y112
> eqt1 = eqt1[eqt1$pValue <= 1e-06, ]
> module = eModule(eData.y112, mData.y112, mInfo.y112, eqt1.y112,
+   p.binom = 0.05, plrt.cut = 0.1, prop.cut = 0.2, haploid = TRUE)
> mIDs = module$chr15$mod1$mID
> mc = mData.y112[mIDs[1], ]
> eIDs = module$chr15$mod1$eID
> eD = eData.y112[eIDs, ]
> moduleScore(eD, mc)

$geneID
[1] 10 17 39 16 37

$pval
[1] 2.645702e-14 1.434475e-03 2.524796e-03 6.901860e-03 1.798608e-02

$score
[1] -52.78593
```

5 Integrated study of eQTL and other variables

Integrated studies of eQTL and other variables is of great interest because the ultimate goal of an eQTL study is often beyond the relation between gene expression and genetic markers. Interesting questions include how does a genetic variation affect gene expression [Sun et al., 2007], what is the relation between an eQTL and a clinical outcome [Schadt et al., 2005]? To answer these questions, we often need to dissect the relation between DNA, gene expression, and another quantitative variable, e.g., identifying the following causal relations: $\text{DNA} \rightarrow \text{Transcription factor} \rightarrow \text{gene expression}$, or $\text{DNA} \rightarrow \text{gene expression} \rightarrow \text{clinical outcome}$. In this software, we implemented a likelihood testing method (function `cauLR`) to dissect the possible relations between three variables: $X \rightarrow Y \rightarrow Z$, $X \rightarrow Z \rightarrow Y$, and $Z \leftarrow X \rightarrow Y$. As demonstrated by [Sun et al., 2007], this likelihood testing method is effective to identify the transcription factors that mediate the eQTL modules. These basic network structures including only three elements can also be the basis to build larger networks. An example using simulated data is shown below. The underlying true model is $y_0 \rightarrow y_1 \rightarrow y_2$, i.e., a causal model. The likelihood ratio test does correctly identify this relation.

```
> y0 = rnorm(100)
> y1 = y0 + rnorm(100, 0, 1)
> y2 = y1 + rnorm(100, 0, 1)
> cauLR(y0, y1, y2)

$lrr
[1] -9.867328e+01 -1.181957e+02 -1.355016e+02 3.827654e+00 6.468527e-05

$label
[1] "causal"
```

It is worth to mention that to compare the three models, because they are not nested to each other, we employ likelihood ratio tests for non-nested models [Vuong, 1989]. Sometime it is also desired to compare one of the three models to a complete model where any two of the variables are related. This is can be done by commonly used likelihood ratio test of nested models.

References

- [Brem and Kruglyak, 2005] Brem, R. B. and Kruglyak, L., 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, **102**(5):1572–1577.
- [Carlborg et al., 2005] Carlborg, O., De Koning, D. J., Manly, K. F., Chesler, E., Williams, R. W., and Haley, C. S., 2005. Methodological aspects of the genetic dissection of gene expression. *Bioinformatics*, **21**(10):2383–2393. Evaluation Studies.
- [Schadt et al., 2005] Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., *et al.*, 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, **37**(7):710–717.
- [Stranger et al., 2007] Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., *et al.*, 2007. Population genomics of human gene expression. *Nat Genet*, **39**(10):1217–1224.
- [Sun, 2007] Sun, W., 2007. Statistical strategies in eQTL studies. *UCLA, Department of Statistics*, .
- [Sun et al., 2007] Sun, W., Yu, T., and Li, K.-C., 2007. Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics*, **23**(17):2290–2297.
- [Vuong, 1989] Vuong, Q. H., 1989. Likelihood ratio test for model selection and non-nested hypotheses. *Econometrica*, **57**(2):307–333.
- [Wang et al., 2006] Wang, S., Yehya, N., Schadt, E. E., Wang, H., Drake, T. A., and Lusis, A. J., 2006. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet*, **2**(2):e15.