iBMQ: An Integrated Hierarchical Bayesian Model for Multivariate eQTL Mapping

Greg Imholte, MariePier Scott-Boyer, Aurélie Labbe,
Christian F. Deschepper and Raphael Gottardo

August 9, 2012

1 Download

`iBMQ` is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation. For citation purposes, please refer to the first reference in the bibliography.
In order to properly compile iBMQ, the GNU Scientific Library (GSL) is required. GSL is free and can be downloaded at http://www.gnu.org/software/gsl/ . A tutorial on how to configure and use GSL and R can be found at http://wiki.rglab.org/. For calculations with multiple processors (Mac or Linux only), OpenMP needs also to be installed (http://openmp.org/wp/).

2 Introduction

Recently, mapping studies of expression quantitative loci (eQTL), where expression levels of thousands of genes are viewed as quantitative traits, have been used to provide greater insight into the biology of gene regulation. Current data analysis and interpretation of eQTL studies involve the use of multiple methods and applications, the output of which is often fragmented. We present an integrated hierarchical Bayesian model that jointly models all genes and SNPs to detect eQTLs. We propose a model (named iBMQ) that is specifically designed to handle a large number G of gene expressions, a large number S of regressors (genetic markers) and a small number n of individuals in what we call a large G, large S, small n paradigm [1]. This method incorporates genotypic and gene expression data into a single model while 1) specifically coping with the high dimensionality of eQTL data (large number of genes), 2) borrowing strength from all gene expression data for the mapping procedures, and 3) controlling the number of false positives to a desirable level. One distinct feature of our model is that we calculate one weight parameter for each Gene g and each SNP j.

3 Example dataset: data from mouse Recombinant Inbred Strains (RIS)

This example uses data generated by Williams and Lu, and available from the Gene Network website (genenetwork.com). This dataset consists of the mRNA profiles of whole eye tissue from n = 68 BXD recombinant inbred mouse strains, as measured using Affymetrix M430 2.0

microarrays [2]. To ease calculation and facilitate comparisons, we will use a set of G = 1000 probes and 1700 single nucleotide polymorphic markers (SNPs).

4 Formatting the data

The next subsection describes the format of the data. To run the eQTL model we need two type of data: a genotype data frame and an expression data frame. To visualize the data, we also need a data frame with the genomic positions of each SNP and of each gene/probe.

4.1 The genotype data frame

The genotype matrix consists of genotype data for each member of the population. Each row corresponds to an individual and each column represents a SNP. In the current implementation, the method has been designed for data from RIS, with genotype being homozygotes for each SNP. The genotype must be coded with 0 and 1 for RIS data. Otherwise, for heterozygous sites, the genotype most be coded 0,1,2 and the parameter RIS of the "eqtl.mcmc" most be equal to FALSE. The column names must be the SNP names.

Example:

	SNP1	SNP2	SNP3	SNP4
Ind1	0	0	1	1
Ind2	0	1	0	1
Ind3	0	1	1	0

4.2 The gene expression data frame

The expression data frame consists of expression data for each member of the population. Each row and column corresponds to an individual and a gene, respectively. We have assumed that the gene expression have been appropriately normalized and analyzed using specialized statistical methods.

	Gene1	Gene2	Gene3
Ind1	10.2	11.4	12.2
Ind2	6.7	5.6	7.7
Ind3	13.1	14.5	12.3

4.3 The SNP position data frame

For the genome-wide eQTL mapping plot we need a data frame specifying the genomic locations of all SNPs with following columns: SNP name, chromosome number, the SNP location (in base pair).

SNP	Chr	pos
Snp1	1	17098
Snp2	1	1029012

4.4 The gene position data frame

A data frame specifying the genomic locations of each gene/probe with the following columns: gene name, chromosome number, the start location (in base pairs) and the stop location (in base pairs).

Gene	Chr	start	end
Gene1	1	10290	10460
Gene2	1	18989	19069

5 Preparing the workspace

We will first load the library iBMQ, the SNP data and gene expression data.

To load the `iBMQ` package:

> library(iBMQ)

To load the SNP data:

> data(snp)

To load gene expression data:

> data(gene)

6 Running the eQTL model:

This function computes the MCMC algorithm to produce Posterior Probability of Association for eQTL mapping. This function takes time: please be patient! On a Mac 2* 3.2 Ghz Quad-Core Intel Xeon computer using 6 core it takes 2.44 minutes.

Do not run if you have a slow computer. We will load a pre-computed PPA table later on.

> PPA= eqtl.mcmc(snp, gene, n.iter=100,burn.in=100,n.sweep=20,nproc=6, RIS=TRUE)

In this example, we will perform only 100 iterations in the interest of time. Please note that an optimal analysis requires a greater number of iterations (preferably in the order of 100,000, with a burn-in of 50,000). The result is a matrix with Posterior Probabilities of Association for each gene (row), and each SNP (column). You can load the PPA matrix previously calculated with 100,000 iterations with the following command line:

> data(PPA)

7 Computing the FDR Threshold for eQTL identification:

Our ultimate goal is to identify gene/SNP associations, which can be done using parameter estimates from our model. An eQTL for gene g at SNP j is declared significant if it's corresponding marginal posterior probability of association (PPA) is greater than a given threshold. In the context of multiple testing and discoveries, a popular approach is to use a common threshold leading to a desired false discovery rate (FDR)[3].

To calculate the threshold:

> cutoff=calculateThreshold(PPA, 0.1)

In this example the PPA optimal cutoff is 0.740. We can calculate how many eQTLs have PPA above the cutoff:

> length(which(PPA> cutoff))

A total of 759 eQTLs are identified.

8 Visualizing the result:

The plot.eqtl function allows us to visualize the genome-wide eQTL mapping plot. For the plotting function, we need a data frame specifying the genomic locations of the SNPs with columns and a data frame specifying the genomic locations of the gene/probe (see section on data format).

Load the data containing the position for each SNP:

> data(snppos)

Load the data containing the position for each gene/probe:

> data(genepos)

To visualize each chromosome along the whole mouse genome axis, we need to define the length of each mouse chromosome in base pairs.

> mouseChr=c(1.97e+08, 1.82e+08, 1.60e+08, 1.56e+08, 1.53e+08, 1.50e+08, + 1.53e+08, 1.32e+08,1.24e+08, 1.30e+08,1.22e+08,1.21e+08,1.20e+08, 1.25e+08,1.03e+08, 9.80e+07, 9.50e+07, 9.10e+07, 6.10e+07)

If you do not have the chromosome length for our organism, the option chr=FALSE of the function will assume that all chromosomes have the same length in the plot.
Plot the significant eQTLs with the respect of the gene/snp position:

> plot.eqtl(PPA,cutoff,snppos, genepos,mouseChr)

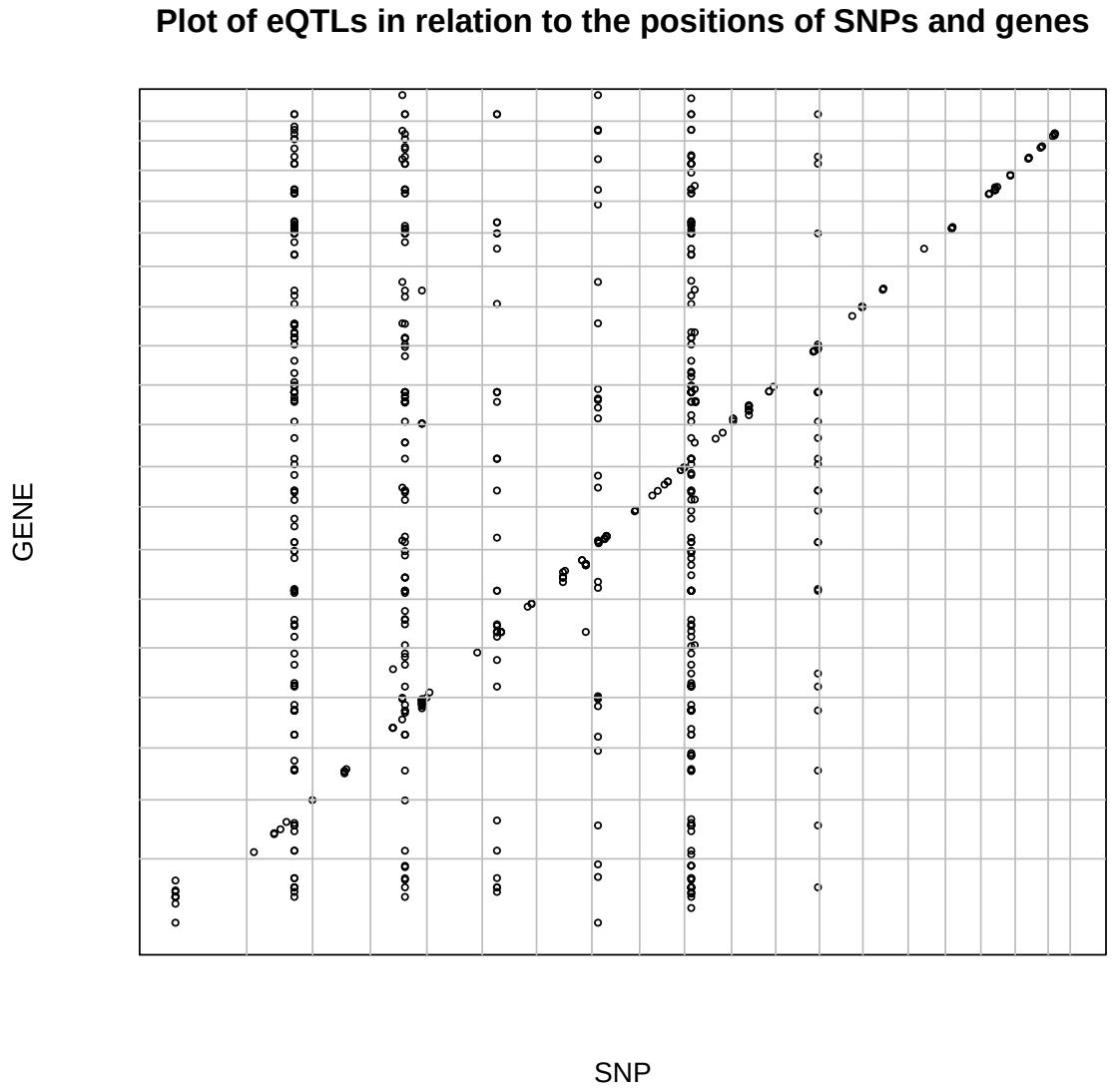

Figure 1: Genome-wide distribution of eQTLs found by iBMQ for the test using 1000 probes for whole eye tissue from 68 BXD mouse recombinant inbred strains. The x-axis gives the position of each eQTL along the genome; the y-axis gives the position of the probe set target itself. The grey lines mark chromosome boundaries. cis-QTLs form a diagonal line. Vertical bands represent groups of transcripts linked to same trans-eQTL.

References

[1] Scott-Boyer, MP., Tayeb, G., Imholte, Labbe, A., Deschepper C., and Gottardo R. An integrated Bayesian hierarchical model for multivariate eQTL mapping (iBMQ). Statistical Applications in Genetics and Molecular Biology Vol. 11, 2012.

[2] Geisert, EE., Lu, L., Freeman-Anderson, NE., Templeton, JP., Nassr, M., Wang, X.,

Gu, W., Jiao, Y., Williams, RW. (2009): "Gene expression in the mouse eye: an online resource for genetics using 103 strains of mice." *Mol Vis.*,15, 1730-63

[3] Newton, MA., Noueiry, A., Sarkar, D. and Ahlquist, P. (2004): "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biometrics*, 5(2), 155-176