Statistical eQTL analysis of five genes in lymphoblastoid cells lines from 344 individuals

Christopher Jenness

May 7, 2016

Abstract

As part of the 1000 Genome Project, single nucleotide polymorphisms (SNPs) have been identified throughout the human genome. Here, we look at 50,000 SNPs in 344 individuals to identify genetic loci that affect gene expression of 5 different genes: MARCH7, FAHD1, PEX6, ERAP2, and GFM1. Using a genetic linear regression model incorporating population structure as a covariate, we identify single hits for ERAP2, PEX6, and FAHD1 in our expression quantitative trait locus (eQTL) analysis. We note that the hits contain the gene-of-origin, indicating that the causal loci likely fall within promoters or other regulatory elements and act in cis.

1 The Data

Five genes (MARCH7, FAHD1, PEX6, ERAP2, and GFM1) were chosen for eQTL analysis. To identify loci in the genome that affect each gene, mRNA levels were measured in 344 individuals from 5 populations (Table 1). The raw mRNA levels were quantile normalized, giving a normal distribution of all five genes (Figure 1). For each individual 50,000 SNPs were measured across chromosomes 1-22.

Table 1: Populations analyzed in this study

Population	Abbreviation	Individuals Measured
Utah	CEU	78
Finland	FIN	89
Great Britain	GBR	85
Tuscany	TSI	92

Additional filtering was performed on the data. Individuals with missing data were omitted, and SNPs with minor allele frequencies < 0.05 were removed to reduce the incidence of false positives.

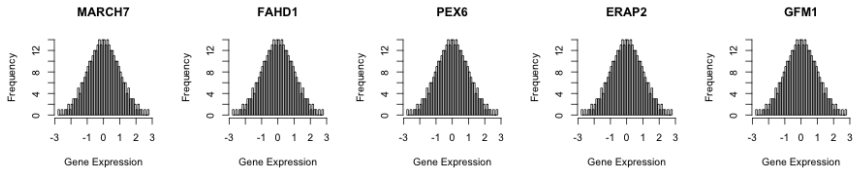

Figure 1: Distribution of mRNA expression levels of 5 genes in 344 individuals.

2 The Model

The goal of eQTL analysis is to find loci in the genome where $Cov(X, Y) \neq 0$, where Y is the mRNA level for the gene of interest and X is a marker in the genome. We model the genetic system using multiple regression (equation 1).

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \tag{1}$$

where X_a and X_d encode genotypes as per equations 2 and 3.

$$\begin{aligned} X_a(A_1A_1) &= -1 \\ X_a(A_1A_2) &= 0 \\ X_a(A_2A_2) &= 1 \end{aligned} \tag{2}$$

$$\begin{aligned} X_d(A_1A_1) &= -1 \\ X_d(A_1A_2) &= 1 \\ X_d(A_2A_2) &= -1 \end{aligned} \tag{3}$$

Maximum likelihood estimators of parameters $\beta_\mu, \beta_a, \beta_d$ from equation 1 are obtained using the closed form solution in equation 4.

$$MLE(\hat{\boldsymbol{\beta}}) = (\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y} \tag{4}$$

To determine causal genetic loci, hypothesis testing was performed on each marker by calculating an F statistic using the following null and alternative hypotheses:

$$\begin{aligned} H_0 &: \beta_a = 0 \cap \beta_d = 0 \\ H_A &: \beta_a \neq 0 \cup \beta_d \neq 0 \end{aligned} \tag{5}$$

Finally, since we conducting multiple hypothesis tests (50,000. One for each genetic marker), we adjust our type one error using a Bonferroni correction.

3 Results

Conducting hypothesis tests for each genetic marker as described above, we plotted p values as Manhattan plots for each of the five genes analyzed. For three genes (MARCH7, FAHD1, and PEX6), we identified single peaks of p values < 0.000001, indicating a chromosomal region in linkage disequilibrium, likely containing a causal locus (Figure 2). For two genes (ERAP2, GFM1), we could not identify any causal loci (Figure 3). The chromosomal positions containing p values < 0.000001 are listed in Table 2.

Table 2: Chromosomal Markers with $p < 0.000001$

Gene	Chromosome	Start position	End position
ERAP2	5	96774230	97110808
PEX6	6	42873885	43108015
FAHD1	16	1524250	1929366

Under the null hypothesis, P values should be uniformly distributed from 0 to 1. To asses our model and to identify potential covariates, we constructed QQ plots, plotting the expected P values against our observed P values for each gene of interest (Figure 4). Our non-significant p values were fairly linear, indicating good model selection.

4 Including Population Structure as a Covariate

Since population structure is the most common unaccounted for covariates in GWAS analysis, we incorporated our population structure into our model using two approaches. First, we used given population structure (GCEU, FIN, BGR, and TSI) encoded as dummy variables in our linear regression model (Figure 5, top half). Alternatively, we learned the population structure from the genotype data using PCA and incorporated the first principal component into our linear regression model (Figure 5, bottom half). Importantly, we still identified the same hits as before after accounting for population structure. However, the non-significant p values were more uniformly distributed as seen in the QQ plots of GFM1 and MARCH7.

Finally, after including gender as a covariate, no improvement in QQ plots or differences in hits were obvserved (data not shown).

5 Biological Significance

Using the USCSC genome browser, we looked at genes that were contained within the causal loci identified for ERAP2, PEX6, and FAHD1 (Table 3). Interesting, for each gene of interest, the same gene was contained within the causal locus. This indicates that there may be SNPs neighboring the gene that

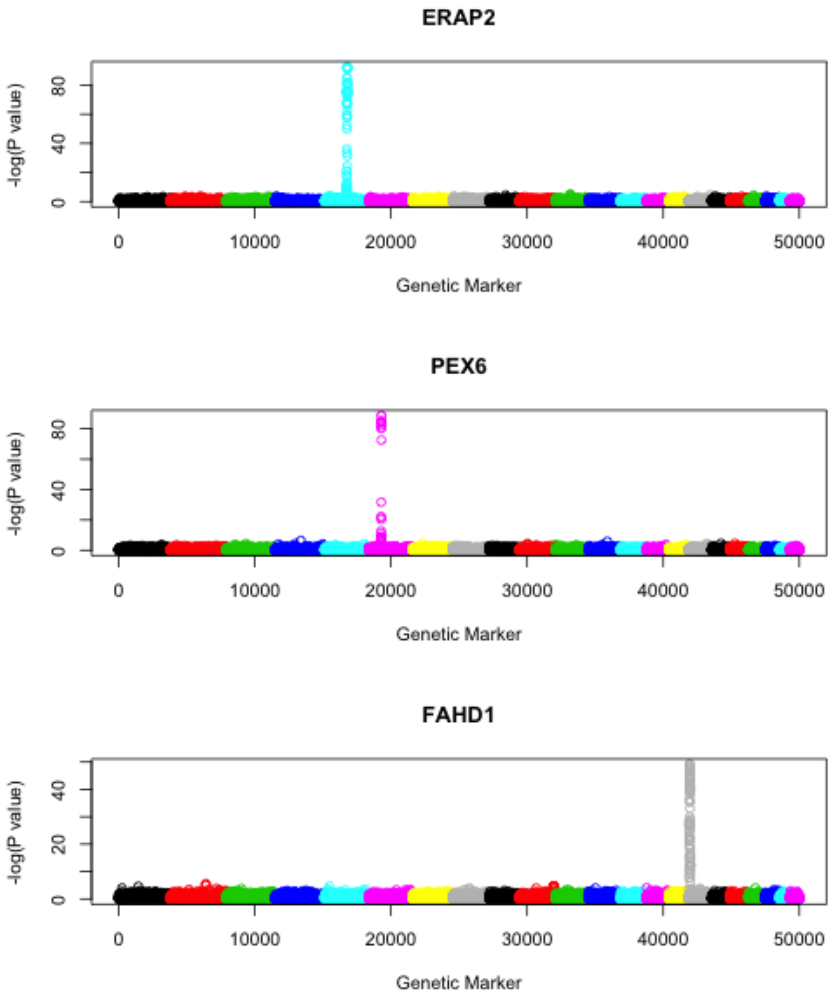

Figure 2: Manhattan plot of 50,000 genetic markers for three genes (MARCH7, FAHD1, PEX6) with identified causal loci. Colors indicate different chromosomes (1-22)

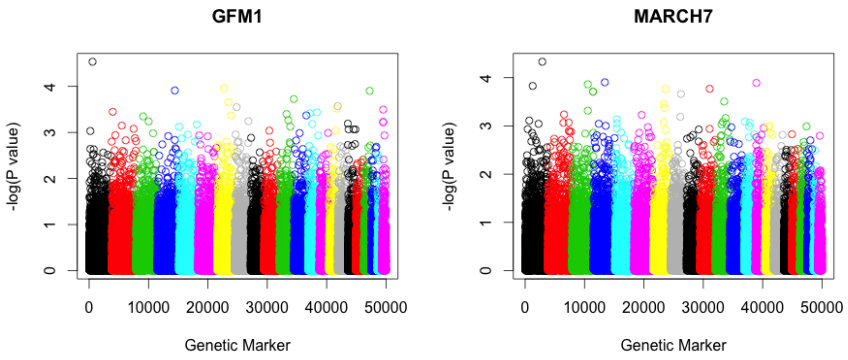

Figure 3: Manhattan plot of 50,000 genetic markers for two genes (ERAP2, GFM1) with no identified causal loci. Colors indicate different chromosomes (1-22)

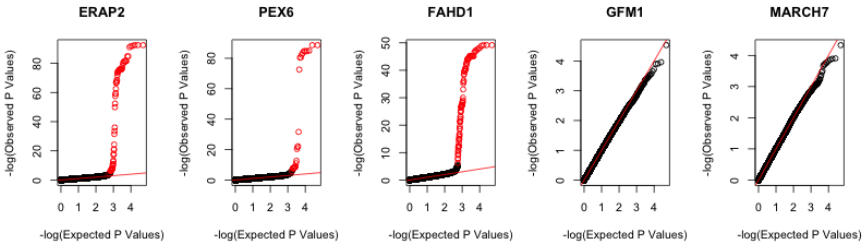

Figure 4: Quantile-Quantile plots for 5 genes using the genetic linear regression model (equation 1). Positive hits from manhattan plots in figure 2 are colored red. The red line indicates the expected p values under the null hypothesis.

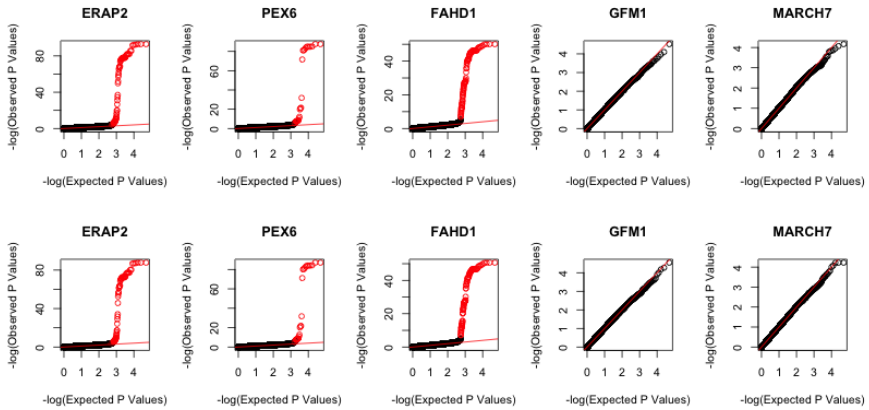

Figure 5: Quantile-Quantile plots for 5 genes using the genetic linear regression model (equation 1) and incorporating the given population structure (top half) or the learned population structure from PCA (bottom half). Positive hits from manhattan plots in figure 2 are colored red. The red line indicates the expected p values under the null hypothesis.

affect its transcription in cis. These causal SNPs may be in promoters or other regulatory elements that surround the gene.

Table 3: Genes contained within significant loci

Quantative Trait Gene	Genes contained within causal loci
ERAP2	ERAP1, ERAP2, LNPEP, LIX1
PEX6	GNMT1, PEX6, PPPD2R53 (among others)
FAHD1	HN1L, MRPS34, FAHD1 (among others)

6 Conclusion

Here we identified causal loci affecting the expression of three genes (ERAP2, PEX6, and FADH1). Importantly, acceptable QQ-plots could be generated after incorporating population structure as a covariate. We note that the causal loci are likely acting in cis to affect gene expression since they are found near the gene of interest.

Finally, we were unable to detect causal loci affecting the expression of GFM1 and MARCH7. This indicates that there may not be SNPs present in the popuoations we examined that affect the expression of these genes.