Project report

Yuanyuan Wu

Highlights:

After the analysis of expression quantitative trait loci(eQTLs) of 5 genes for 344 samples of 4 different populations, 3 regions of the genome were found to be possibly causal for the expression of 3 genes respectively. The functionality of the uncovered regions ware partially double checked by the linkage disequilibiurm(LD).

Methods and results:

1. Quality control of the data and stratification:
   Expression of five genes were tested and the information of the genes are as follow:

PROBE	CHROMOSOME	START	END	SYMBOL
ENSG00000136536.9	2	159712456	159768582	MARCH7
ENSG00000180185.7	16	1827223	1840206	FAHD1
ENSG00000124587.9	6	42963872	42979242	PEX6
ENSG00000164308.12	5	96875939	96919702	ERAP2
ENSG00000168827.9	3	158644496	158692571	GFM1

Take the expression levels of the 5 genes as phenotype, the distribution of the phenotypes were analyzed to see if it is normally distributed as required by GWAS. For all 5 genes, it was perfectly normally distributed with similar mean and sd.

A. *B.*

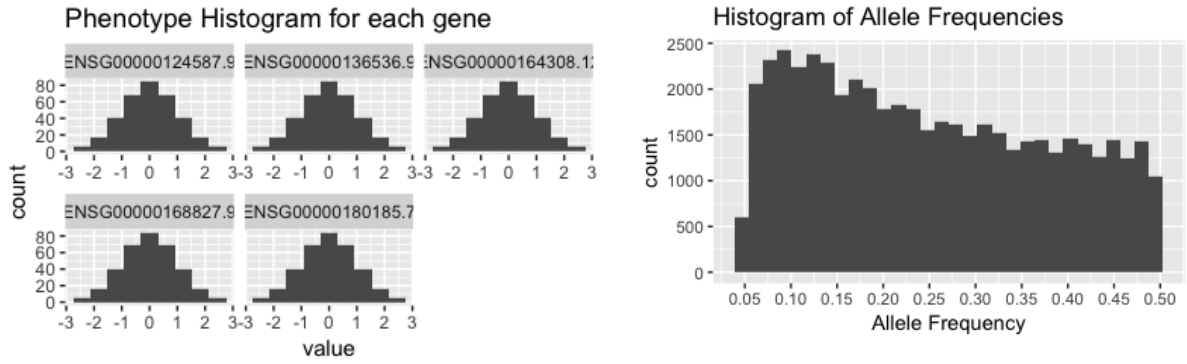


Figure 1 histogram of the 5 genes(1A) and MAF of the SNPs(1B)

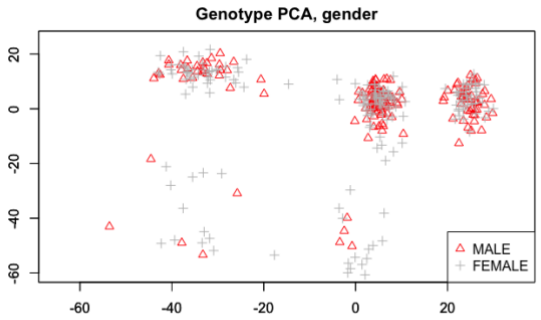
Next, the quality of the genotype data was also tested. The criteria for the genotypes were as following:

a) Individuals with missing genotypes greater than 10% were filtered out. No missing data were found in the given dataset.
b) SNPs with <0.05 MAF will be filtered out. In this study, no MAF<0.05 were found.

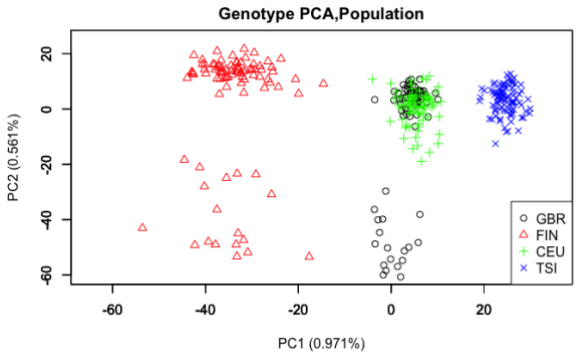
After careful filtering, no snps were filtered.

Given the substructure of the population, stratification was done to uncover the covariates and clustering of the population. Principle component analysis was done based on the genotypes and grouping of the data based on gender or population were shown as below. There is no grouping pattern for the MALE/FEMALE (fig 2B)while clustering of the populations were obvious (fig 2B). The 'FIN' and 'GBR' highly overlapped with each other which may result from the possible common ancestry for both populations. The variance explained by the first 10 PCs were demonstrated by figure 2C.

A.



B.



C.

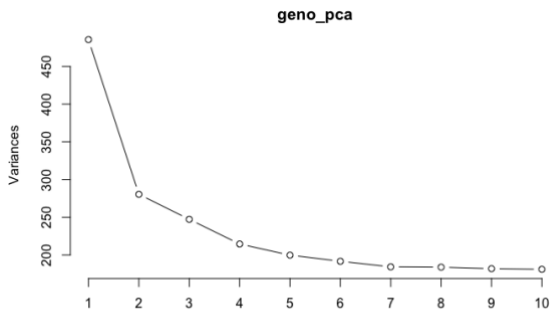
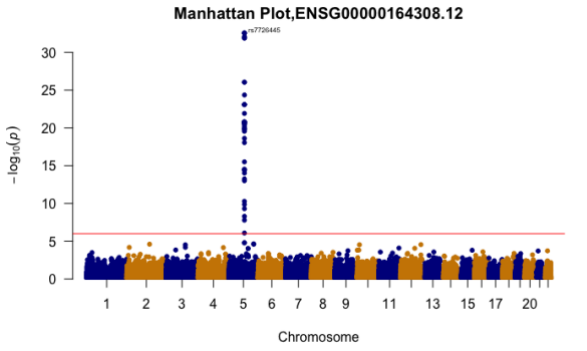
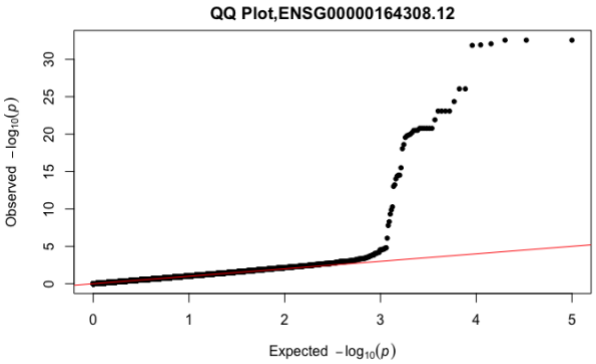


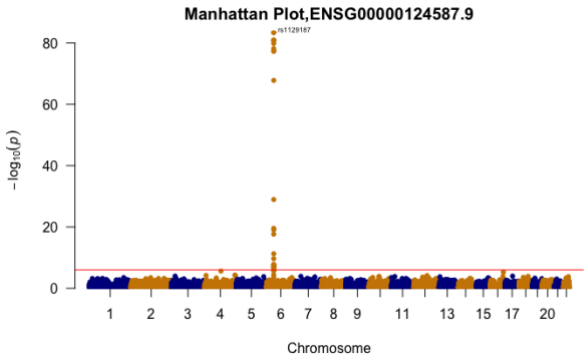
Figure 2. PCA for the genotypes and variance explained by first 10 PCs

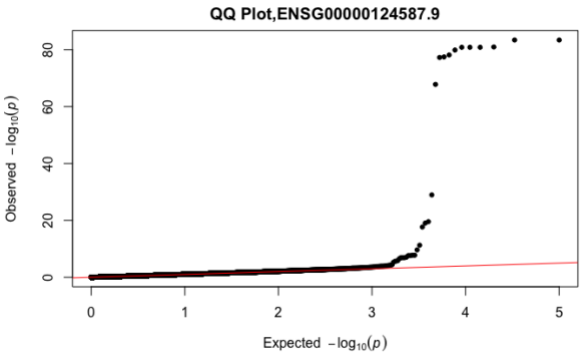
2. GWAS

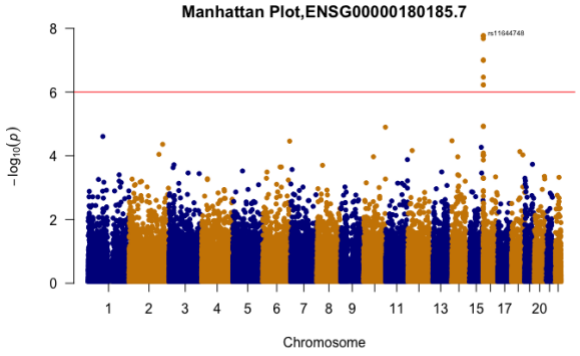
Linear regression with gender and population, as well as first 10 PCs as covariates were done for each of the gene. Given the genotype information, qqman package was used to plot the Manhattan plot across the chromosome. Bernoulli corrected p-val was used to avoid false positive detections. QQ plot was listed beside the Manhattan plots of the 5 genes respectively.

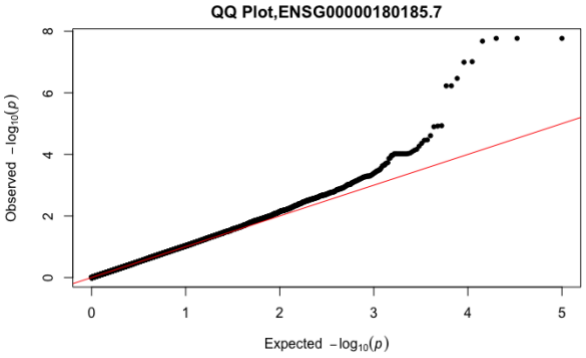

Manhattan Plot,ENSG00000164308.12
rs7726445
−log10(p)
Chromosome

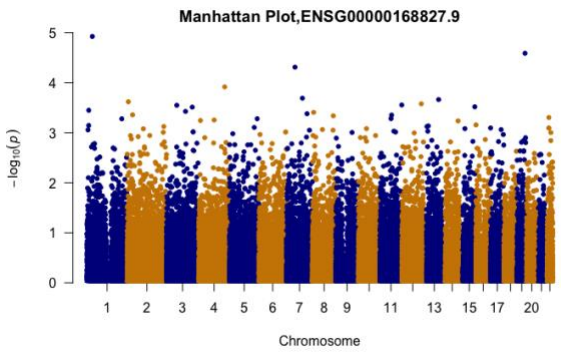

QQ Plot,ENSG00000164308.12
Observed −log10(p)
Expected −log10(p)

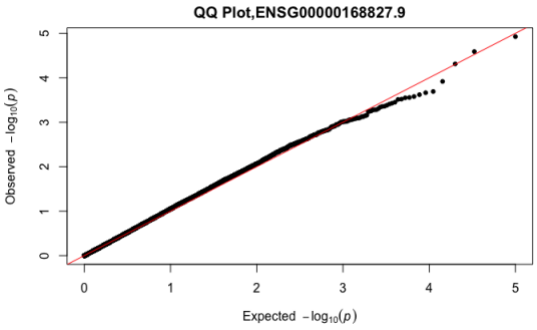

Manhattan Plot,ENSG00000124587.9
rs1129187
−log10(p)
Chromosome


QQ Plot,ENSG00000124587.9
Observed −log10(p)
Expected −log10(p)


Manhattan Plot,ENSG00000180185.7
rs11644748
−log10(p)
Chromosome


QQ Plot,ENSG00000180185.7
Observed −log10(p)
Expected −log10(p)


Manhattan Plot,ENSG00000168827.9
−log10(p)
Chromosome


QQ Plot,ENSG00000168827.9
Observed −log10(p)
Expected −log10(p)

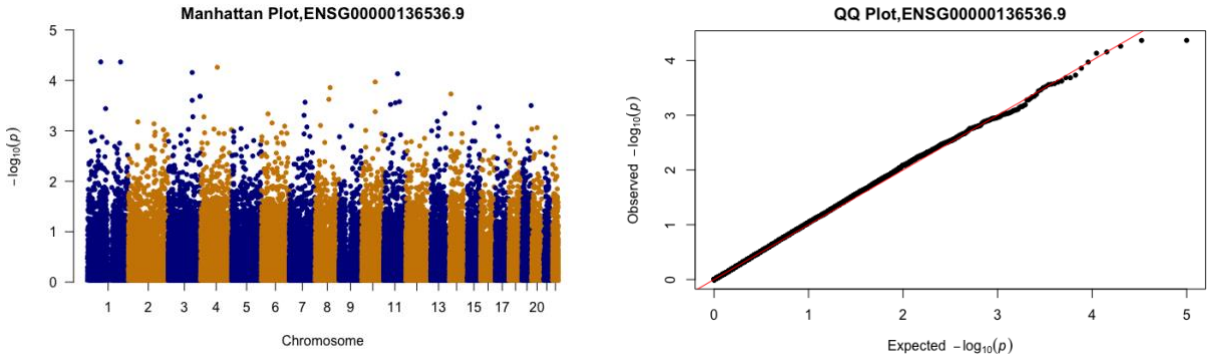


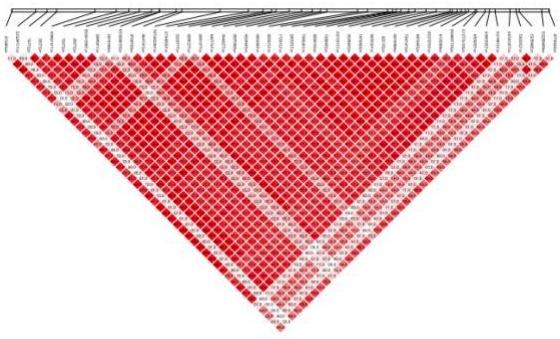
Figure 3. Manhattan plots and QQ plots for each gene

It is confident that there are associated SNPs for the expression level of the gene ENSG00000164308.12 (ERAP2) in chromosome 5 and the gene ENSG00000124587.9(PEX6). A marginal confident hot site was detected for the gene ENSG00000168827.9(GFM1). No positive results were found for the rest two genes.

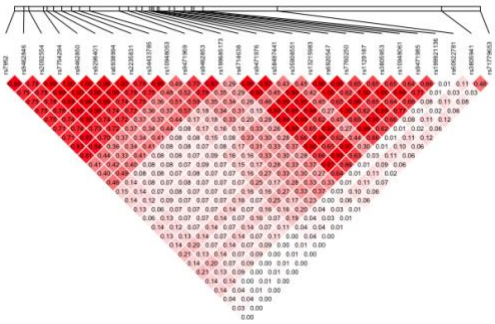
3. LD for functionality

The positive SNPs for gene ERAP2(3A) range from 96.87MB to 97.04MB in the chromosome 5. The gene itself sits in approximate range, which indicates it is likely a cis-eQTL. A LD map was done for the region of the positive snp region. The highly linked pattern indicated by the map proved the functionality of the region. LD plots for the positive region of the gene PEX6(3B) and the marginal significant region GFM1(3C) are also listed below.

A.



B.



C.

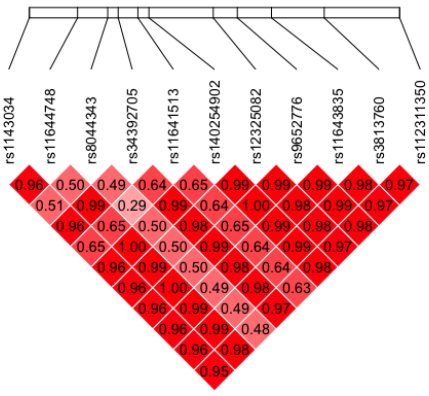


Figure 3. LD plot for the reported positive regions

Confirmed by the strong LD in the regions reported positive snps, it is intriguing to say that are the discovered regions are functional and possible candidates to be causal to the change in the gene expression of ERAP2, PEX6 and GEM1 genes.