BTRY 6830 Project

Weilin Xu

DATASETS & QUALITY CONTROL

The datasets used in this study were from part of the Genetic European ariation in Health and Disease (gEUVADIS), which contained samples of 344 individuals. Each of these individuals were measured on 50000 single nucleotide polymorphisms (SNPs), and the mRNA levels for 5 genes were quantified using RNA-Seq:

Probe	Chromosome	Start Position	End Position	Symbol
ENSG00000136536.9	2	159712456	159768582	7-Mar
ENSG00000180185.7	16	1827223	1840206	FAHD1
ENSG00000124587.9	6	42963872	42979242	PEX6
ENSG00000164308.12	5	96875939	96919702	ERAP2
ENSG00000168827.9	3	158644496	158692571	GFM1

Phenotype Data

The phenotype dataset contained mRNA expression levels of 5 different genes. In order to conduct a genome-wide association study (GWAS), the phenotypes of expression levels should be normally distributed. In order to check for normality of the expression levels, histograms for expression levels of each gene, along with the p values of the Shapiro-Wilk tests (for tests of standard normal distribution), were used. It can be shown that the expression level of each gene was normally distributed with mean 0 and variance 1. To reduce the influence of outliers on regression, for each gene all individuals with expression levels greater than 3 or less than -3 were removed from further analysis. After all these steps, 344 individuals remained for each gene. [Figures and p values not shown due to restrictions on page counts.]

Genotype Data & Population Stratification

The following steps were adapted from Reed et al., 2015 [DOI: 10.1002/sim.6605] and Jason's Lecture Slides [Lecture 20: Minimal GWAS Analysis Steps]

- Filtering SNPs based on missing rate. We should remove SNPs with missing rates greater than 0.05 across all individuals. In this study, the genotype file was complete and no SNP was removed.
- Filtering SNPs based on Minor Allele Freqeuncy (MAFs). To ensure the power of association tests, all SNPs with MAF < 0.05 were removed; in this project, none was removed and all SNPs had MAFs in the range of 0.05 to 0.50.
- Filtering samples based on missing rate. We should remove individuals with missing SNPs greater than 0.10 across the genomes. In this study, the genotype file was complete and no SNP was removed.
- Filtering samples based on heterozygosity. Low heterozygosity within an individual in the study might indicate inbreeding, while high heterozygosity might suggest some genotyping error. To filter samples based on heterozygosity, we have that Inbreeding Coefficient $F = 1 - \frac{O}{E}$, where O and E represent observed and expected counts of heterozygous loci for a given individual. Then individuals with $|F| > 0.10$ were removed. In this study, none of the individuals had a absolute inbreed value greater than 0.10.
- Filtering samples based on relatedness. One assumption in population-based GWAS is that the individuals in the sample should be unrelated. In general, the identify by descent (IBD) kinship coefficient will be used to quantify the relatedness. However, in this study no information on kinship was provided, so we could not filter samples with high relatedness.
- Discovering population stratification and filtering samples based on ancestry. It was reported that the sampled individuals were from four populations: CEU (Utah residents with European ancestry), FIN (Finns), GBR (British) and TSI (Toscani). However, self-reported ethnicity and ancestry might not be compatible with the genetic background. From the first two components of the Principal Component Analysis (PCA) results, the FIN and the GBR populations were respectively divided into two sub-populations. Surprisingly, the CEU population had

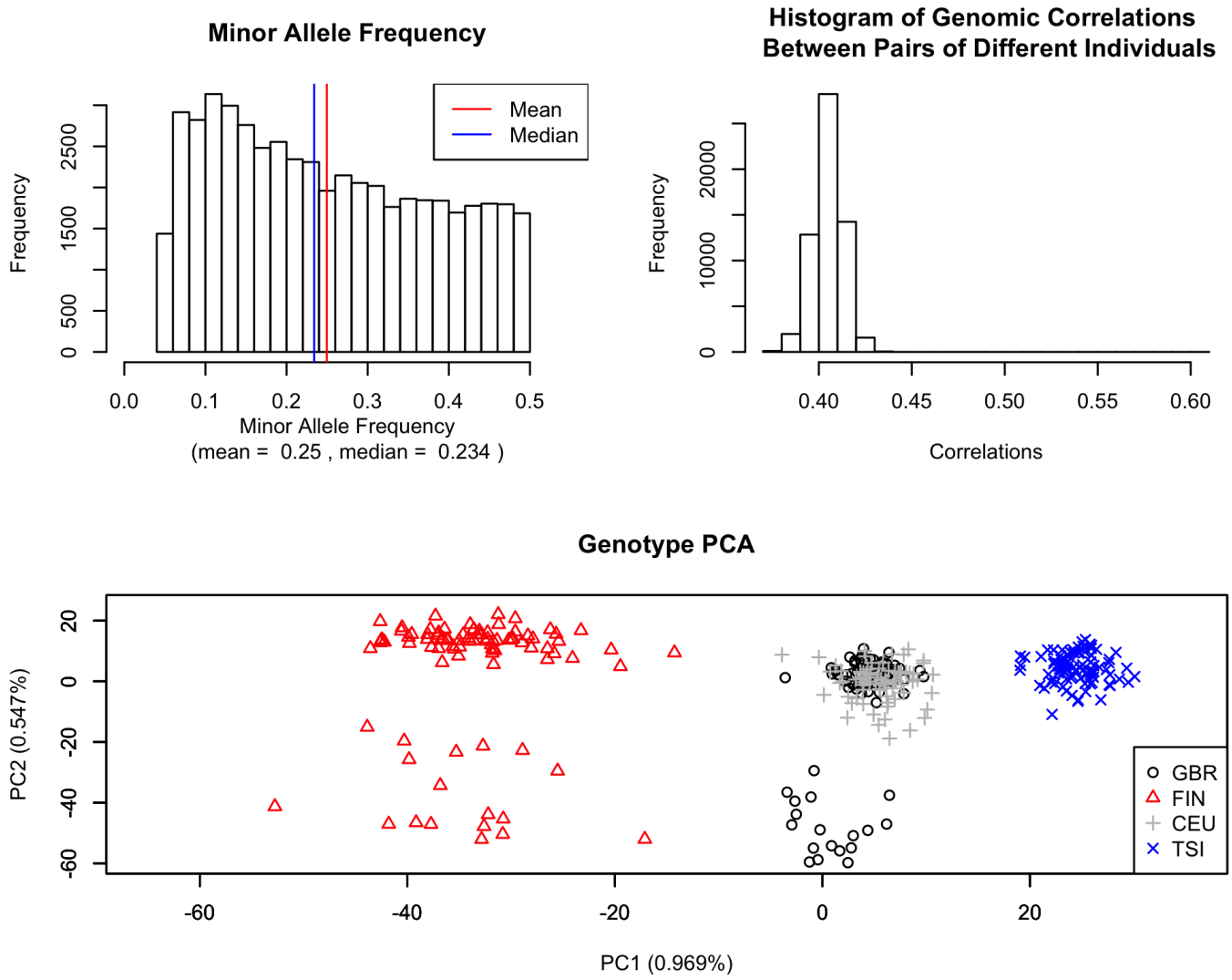

Figure 1: Top Left: Distribution of minor allele frequency. Top Right: Histogram of genomic correlations between pairs of different individuals. Bottom: Principal Component Analysis of Genotypes

substantial overlapping with one of the GBR sub-populations. Despite the geographical distance between the two populations, the PCA results indicated that there might be population admixture between them, which could be supported by the fact that the CEU population was of European ancestry. In this step, we could not remove any individual because there was no individual with incompatible ancestry.

- Filtering out SNPs based on Hardy-Weinberg Equilibrium. To further support the existence of stratification, we should assess whether there was deviation from Hardy-Weinberg Equilibrium. From the Manhattan Plot of HWE p values, we can see that there were genomic positions where substantial deviation from HWE might exist. Three major factors might lead to the deviation from HWE: population stratification, population admixture and genotyping errors. In this step, we remove all SNPs with a corresponding p value less than 10^{-7}, to avoid the influence of genotyping error on the association tests.

After all these steps, 49753 SNPs remained for downstream analysis.

GENOME-WIDE ASSOCIATION ANALYSIS

A single-trait GWA study was conducted for expression levels of ENSG00000164308.12 while controlling for genders and populations. Meanwhile, the first 10 principal components were also used as covariates. There was a significant association

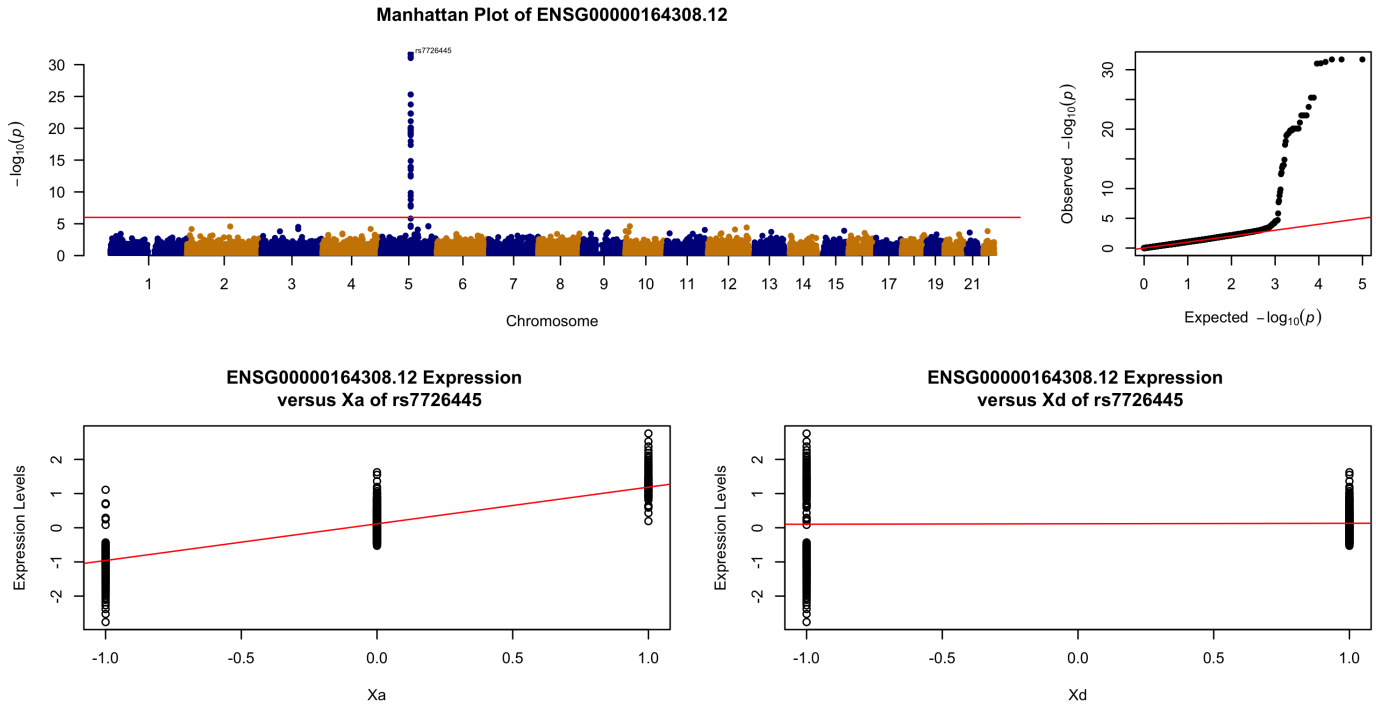

Figure 2: GWAS result of ENSG00000164308.12. Top: Manhattan Plot and QQ Plot. Bottom: Expression levels versus X_a and X_d for top SNP rs7726445.

peak on Chromosome 5. The associated peak was located at about 96.86 Mb to 97.04 Mb, with the top SNP being rs7726445 at 96945338 BP ($p < 1.87 \times 10^{-32}$). The gene ENSG00000164308.12 is located on Chromosome 5, from 96875939 to 96919702 BP, which substantially overlaps with the association peaks, indicating the existence of *cis*-eQTLs. However, the top SNP was not inside the gene.

Due to restrictions on page counts, here let us focus on the association peak on Chromosome 5 (corresponding to ENSG00000164308.12), especially the top SNP.

Linkage Disequilibrium

LocusZoom is an online tool for analysis of GWAS result (http://locuszoom.org/). Using LocusZoom, we can "zoom in" the region upstream and downstream of the top SNP for 400 Kb respectively (Figure 3, Left Panel). LocusZoom also provides statistics on linkage disequilibrium (LD) in the form of r^2 as well as recombination rates based on data from 1000 Genomes (version hg19), although we could obtain LD plots of the surrounding region more directly using genotype data (Figure 3, Right Panel).

From the regional association plot, the association peak was in a LD block of size ~0.2 Mb, where the recombination rates were relatively low. Besides, the SNPs in the association peak were correlated with the top SNP; the correlations between other SNPs and the top SNPs increase as their physical distances in the genome decrease. All these provided evidence for the functionality of this eQTL. In addition, there were light strips within the LD plot of significant SNPs, suggesting the existence of genotypeing errors in the association peaks.

Functioal Studies

In addition to the NCBI dbSNP database, the Online Mendelian Inheritance in Man (OMIM, https://www.omim.org/), and the Gene Ontology Consortium (GO, http://www.geneontology.org/), provide easy access to in-depth information on the genes and SNPs. From Figure 3, the association peak straddled two genes: *ERAP2* and *LNPEP*. *ERAP2* stands for "endoplasmic reticulum aminopeptidase 2", which codes for a aminopeptidase that uses its N-terminus for hydrolysis of protein or peptide substrates (adapted from Online Mendelian Inheritance in Man, OMIM). *LNPEP* is another type of aminopeptidase, Leucyl and cystinyl aminopeptidase, of placenta. The top SNP rs7726445 is in the intron of *LNPEP*,

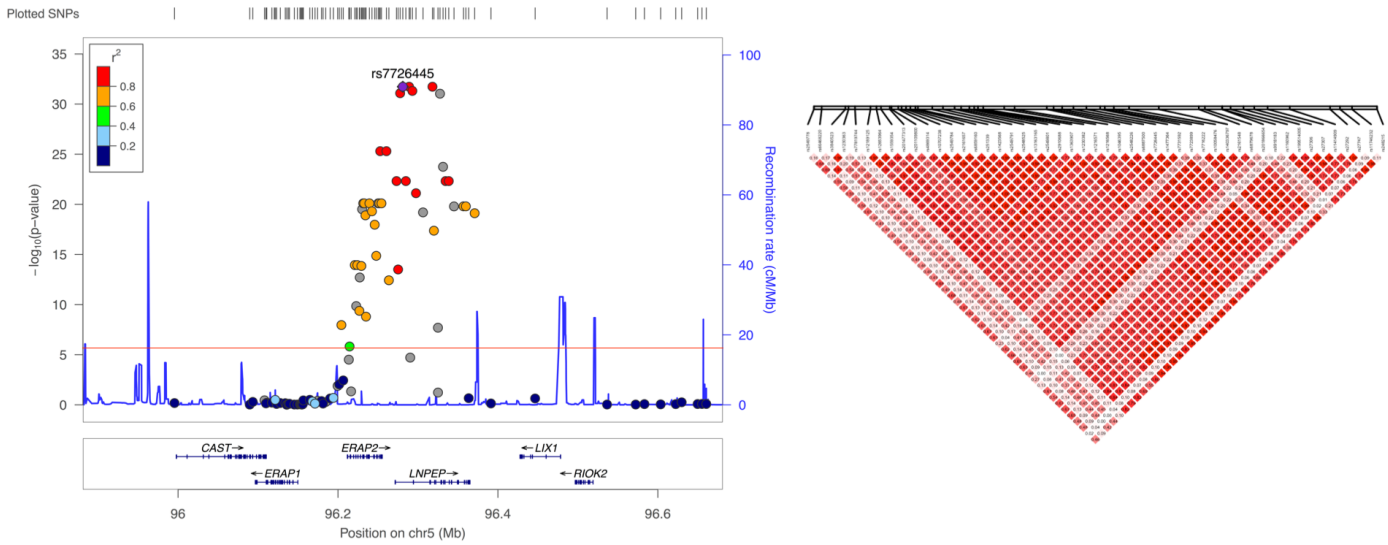

Figure 3: Left: LocusZoom view of the top SNP and the region ± 400 Kb. Right: LD plot of significant SNPs

which is downstream of *ERAP2*. If the resolution was limited to all Bonferroni-corrected significant SNPs, most of the 42 significant SNPs are located in the introns of *ERAP2* and *LNPEP* (based on information from dbSNP), which did not by themselves indicate any functional role.

Heritability

Narrow-sense heritability is defined as the fraction of additive genetic variance and phenotypic variance. With narrow-sense heritability, we can not only predict the chance of successful GWAS, but also look for genomic sites in the genome that might have genetic effects on the phenotype. For ENSG00000164308.12, we can see from the Manhattan plot of narrow-sense heritability that SNPs in the association peak still had very high values of h^2 (Figure 4). In addition, on Chromosome 11 there was also a peak of narrow-sense heritability. This indicated that even though in single-trait GWAS the SNPs on Chromosome 11 did not stand out, they might have some additive genetic effects on the expression of the gene ENSG00000164308.12. We can also conduct heritability analysis on the other genes to predict whether GWAS would be successful on these genes, the results of which were omitted here.

MULTIVARIATE GWAS

The correlation matrix between each gene pair showed relatively low correlations of expression levels, so we could conduct multivariate GWAS on the five genes by conducting single-trailt GWAS on each of the gene respectively. Among the five genes provided, only three of them were found associated with some peaks of significant SNPs.

- ENSG00000124587.9: There was a significant association peak on Chromosome 6. The associated peak was located at about 42.87 Mb to 43.11 Mb, with the top SNP being rs1129187 at 42964461 BP ($p = 2.36 \times 10^{-87}$). ENSG00000124587.9 is located from 42963872 to 42979242; it straddles the association peak, which suggested the existence of *cis*-eQTLs. In this case, the top SNPs was in the gene. (Figure 5)
- ENSG00000180185.7: There was a significant association peak on Chromosome 16. The associated peak was located at about 1.82 Mb to 1.84 Mb, with the top SNP being rs11644748 at 1829958 BP ($p = 3.29 \times 10^{-8}$).

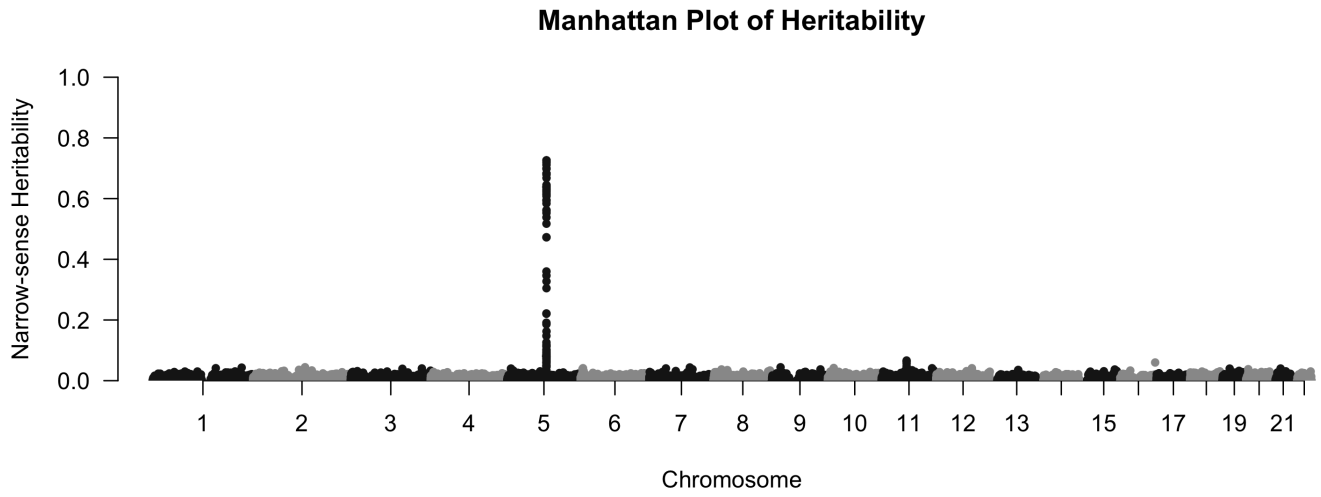

Figure 4: Narrow-sense heritability along the autosomal chromosomes for expression levels of ENSG00000164308.12.

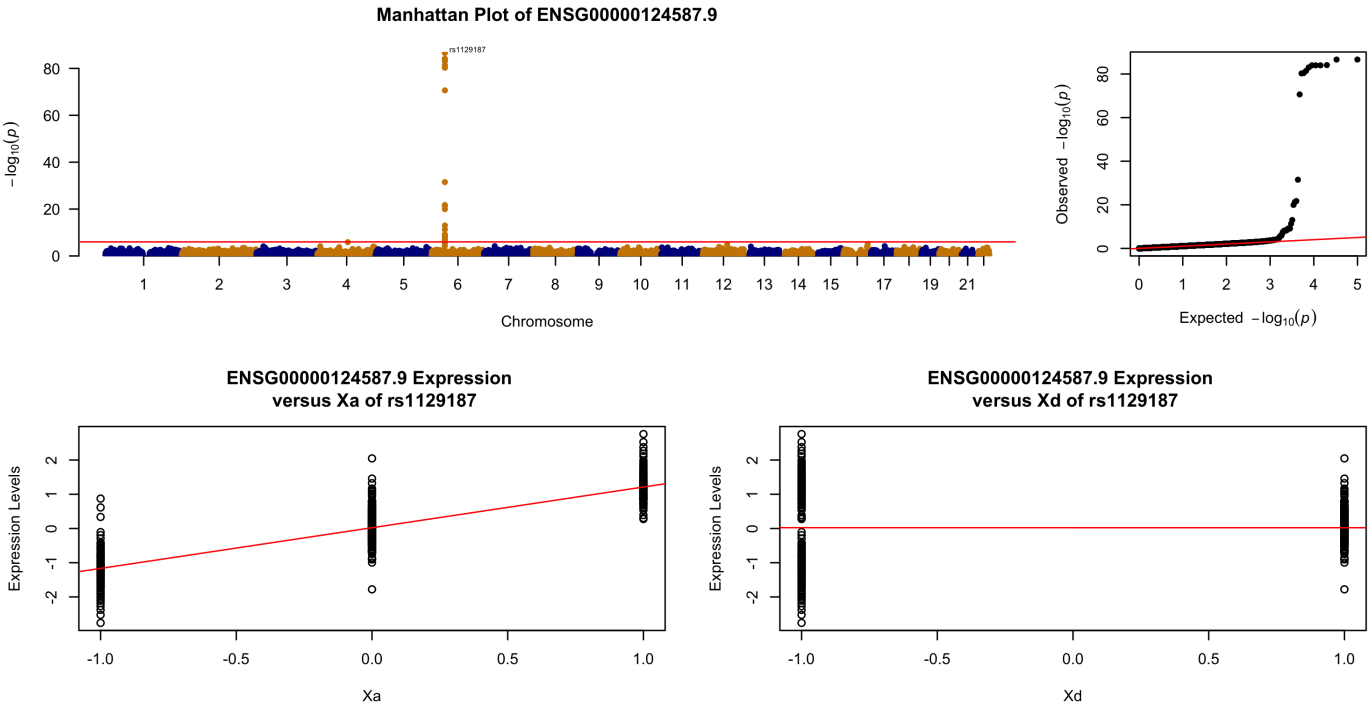

Figure 5: GWAS result of ENSG00000124587.9. Top: Manhattan Plot and QQ Plot. Bottom: Expression levels versus X_a and X_d for top SNP rs1129187.

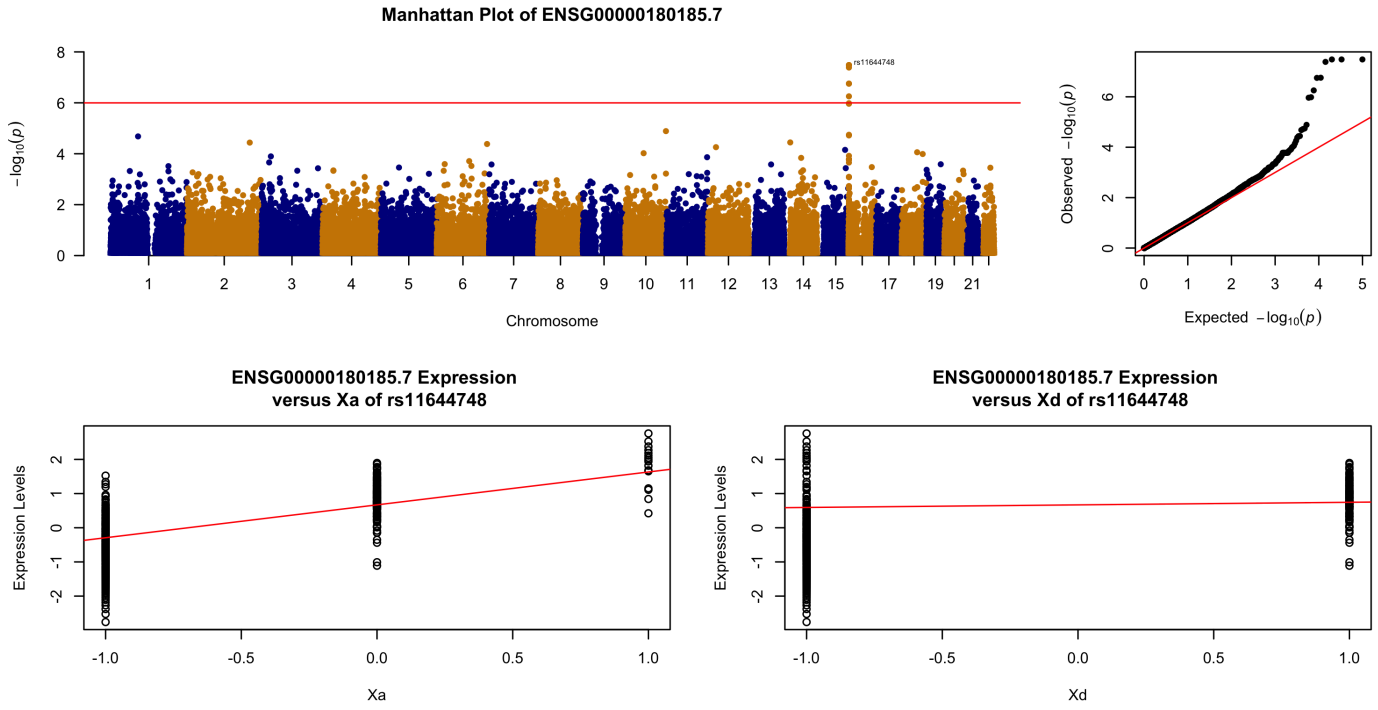

Figure 6: GWAS result of ENSG00000180185.7. Top: Manhattan Plot and QQ Plot. Bottom: Expression levels versus X_a and X_d for top SNP rs11644748.

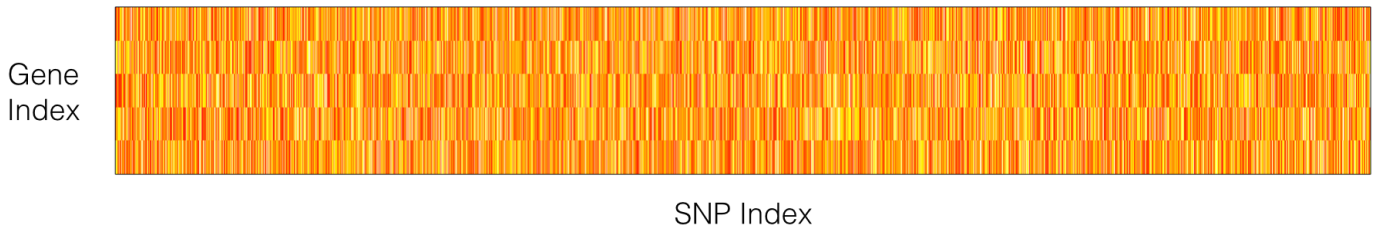

Figure 7: Heatmap of GWAS p values

ENSG00000180185.7. spans from 1827223 to 1840206 BP on Chromosome 16 overlapping with the association peak, which indicated that there were *cis*-eQTLs. The top SNP was also inside the gene. (Figure 6)

All three top SNPs showed significant additive effects on the expression levels, which made biological sense: if one allele of the SNPs would lead to increase/decrease in expression of the gene, then the existence of two copies of the allele would double the increase/decrease in expression relative to having only one copy of the allele. In addition, the heatmap of p values from the association studies of five genes showed no horizontal or vertical "strips", which suggested that the covariates had been controlled. (Figure 7)

REFERENCES

- Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M.P. and Foulkes, A.S., 2015. A guide to genome-wide association analysis and post-analyticnterrogation. Statistics in medicine, 34(28), pp.3769-3792.
- Mezey, J., 2018. Lecture Slides of Quantitative Genetics and Genomics.
- Foulkes, A.S., 2009. Applied statistical genetics with R: for population-based association studies. Springer Science & Business Media.