QG20 Final Project

Genome-Wide Association Analysis of Subset of Data from
Genetic European Variation in Health and Disease (gEUVADIS) Consortium

Jess White

5/12/2020

Contents

Introduction

The Genetic European Variation in Disease (gEUVADIS) consortium produced genome and RNA sequencing data from a subset of samples from the 1000 Genomes project.[1] This included data from 5 populations surveyed, including the CEPH (CEU) or Utah residents with Northern and Western European ancestry, Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI).[2]

Here, a subset of the gEUVADIS data containing 50,000 SNPs and 344 individuals was analyzed. An expression quantitative trait loci (eQTL) analysis was performed using the expression levels of 5 transcripts, MARCHF7, FAHD1, PEX6, ERAP2, GFM1.

Data pre-processing

The initial genotype matrix contained the following coding: 0 (homozygote 1; A_1A_1), 1 (heterozygote; A_1A_2 or A_2A_1), and 2 (homozygote 2; A_2A_2). This required conversion to a X_a (-1: minor allele homozygote, 0: heterozygote, 1: major allele homozygote) and X_d (-1: homozygotes, 1: heterozygotes) matrices.

Exploratory data analysis

Missing data

To ensure that the eQTL analysis is adequately powered, missing data must be accounted for. Normally, individuals with >10% missing data across all genotypes and genotypes with >5% missing data across the entire population would be removed from the analysis. There were no missing data in this data set.

Distribution of phenotypes

The phenotypes or the expression levels of the selected five genes appears to have already been normalized, as their summary statistics are identical, as shown below. Therefore, no pre-processing was necessary.

	Min	Quartile_1	Median	Mean	Quartile_3	Max
ERAP2	-2.759042	-0.6699413	0	0	0.6699413	2.759042
FAHD1	-2.759042	-0.6699413	0	0	0.6699413	2.759042
GFM1	-2.759042	-0.6699413	0	0	0.6699413	2.759042
MARCHF7	-2.759042	-0.6699413	0	0	0.6699413	2.759042
PEX6	-2.759042	-0.6699413	0	0	0.6699413	2.759042

Below is a table summarizing the locations and the functions of the genes of interest.[3]

Symbol	Description	Chr	Start	End
ERAP2	Encodes a zinc metalloaminopeptidase in the endoplasmic reticulum	5	96,875,939	96,919,716
FAHD1	Encodes oxaloacetate decarboxylase	16	1,827,224	1,840,207
GFM1	Encodes one of the mitochondrial translation elongation factors	3	158,644,527	158,695,581
MARCHF7	Encodes membrane-bound E3 ubiquitin ligase	2	159,712,494	159,771,027
PEX6	Encodes a member of the ATPases associated with diverse cellular activities family of ATPases	6	42,963,865	42,980,224

Demographic and sex covariates

We can see that this subset includes 4 of the 5 populations contained in the Geuvadis data set, which are summarized below.

[1] Lappalainen, T., et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511 (2013).

[2] Resource, T.I.G.S. Geuvadis. (2020).

[3] Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 May 12]. Available from: https://www.ncbi.nlm.nih.gov/gene/

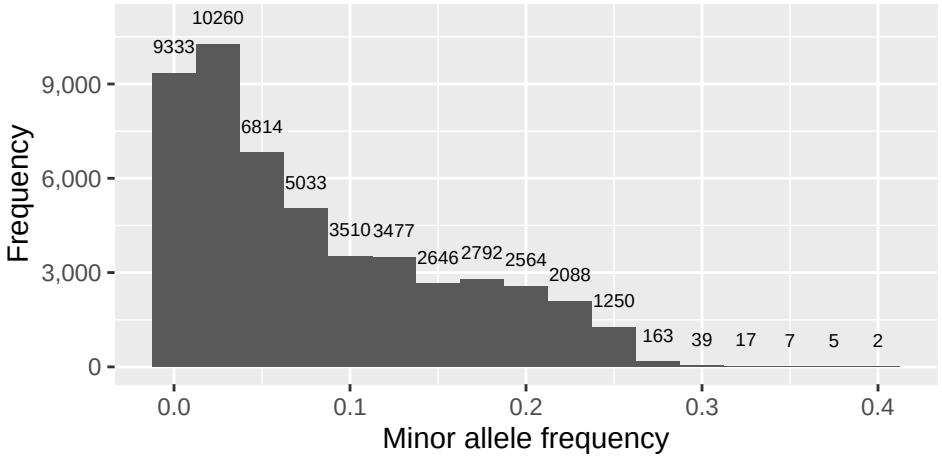
Code	Population	Description
GBR	British	British in England and Scotland
CEU	CEPH	Utah residents (CEPH) with Northern and Western European ancestry
FIN	Finnish	Finnish in Finland
TSI	Toscani	Toscani in Italy

Additionally, we can see that these populations are represented in the data set in roughly similar proportions. Finally, while the sex cohorts are similarly balanced overall, in the Finish cohort, there are over 50% more females subjects than male subjects.

Additionally, I ensured the sex and population data was formatted as factors, enabling their inclusion in the downstream analyses without additional coding.

	FEMALE	MALE	Total	% Total
CEU	37	41	78	22.67%
FIN	55	34	89	25.87%
GBR	46	39	85	24.71%
TSI	43	49	92	26.74%
Total	181	163	344	
% Total	52.6%	47.4%		

Minor allele frequency (MAF)



In an eQTL analysis, MAFs that are too low result in a reduction in the statistical power of the test. As such, they should be removed prior to the analysis.

Given that 23,685 individual SNPs fall below the threshold of 5% given by Dr. Mezey, I removed these SNPs for the purposes of the downstream analysis.

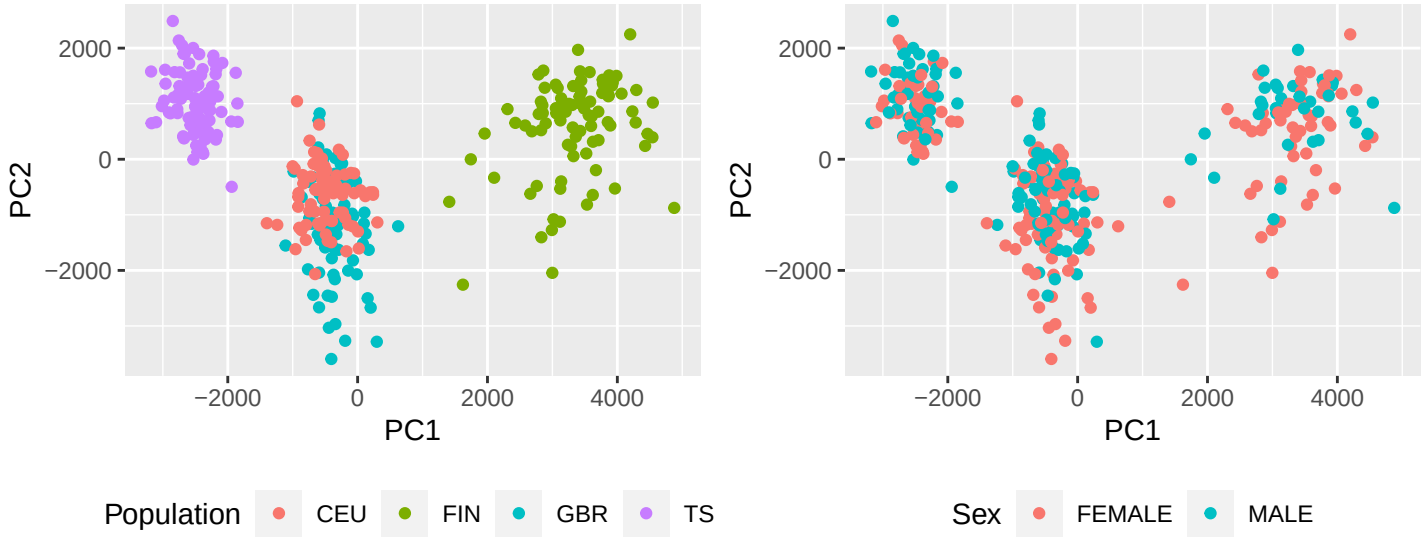
Principal components analysis (PCA)

Initially, I performed a PCA without removing SNPs whose MAF fell below the 5% threshold (*not shown*). In this analysis, the clustering was more spread out, particularly for the PC2 of the GBR and FIN populations, which appeared to show a potential sex effect with female subjects extending far in the negative direction of PC2.

After removing SNPs with low MAFs, the clusters are much more confined (*shown below*). According to the 2000 census, Utah had the highest percentage of English ancestry of any state at 29%.[4] Given the preponderance of British ancestry among this cohort, the degree of overlap between the CEU and the GBR clusters is not suprising. The TSI cluster seems to be the most self-contained, with the Finish cohort being the most spread out.

This observation illustrates the importance of pre-processing the data to remove SNPs with low MAFs.

[4]Brittingham, A. & de la Cruz, G.P. Ancestry: 2000. (2004).



eQTL Analysis

For the various eQTL analyses below, I followed the same procedures. I used the `lm()` function to perform a multiple linear regression of the level of each transcripts' expression for the five given genes at each polymorphic site on the X_a matrix and the X_d matrix. For the analyses noted below, I also included additional covariates, including sex and/or population. Then I extracted the F-statistic and calculated an associated p-value. Finally, I performed multiple hypothesis test correction using the more stringent Bonferroni method.

For each combination of covariates tested, Manhattan plots were generated to visualize the sites of significant SNPs. Additionally, QQ-plots were generated to ensure that the significant SNPs can be interpreted as potential causal polymorphisms. For reasons discussed in the **Summary of eQTL analysis with covariates**, these plots are only shown for the model using sex and population as covariates.

No Covariates

Initially, an eQTL analysis was performed on the five transcript expression levels at each polymorphic site without the inclusion of any additional covariates. The result was the identification of potential sites of causal polymorphisms for ERAP2, FAHD1, and PEX6, but not for the GFM1 and MARCHF7 genes. This was a consistent observation across all models.

Following the multiple linear regression without covariates and with multiple hypothesis correction via the Bonferroni method, there are 64 signficant polymorphisms for ERAP2, 45 signficant polymorphisms for FAHD1, 29 signficant polymorphisms for PEX6. There were no identified statistically significant polymorphisms for GFM1 or MARCHF7.

Sex as a covariate

Following the multiple linear regression without sex as a covariate and with multiple hypothesis correction via the Bonferroni method, there are 64 signficant polymorphisms for ERAP2, 44 signficant polymorphisms for FAHD1, 27 signficant polymorphisms for PEX6. There were no identified statistically significant polymorphisms for GFM1 or MARCHF7.

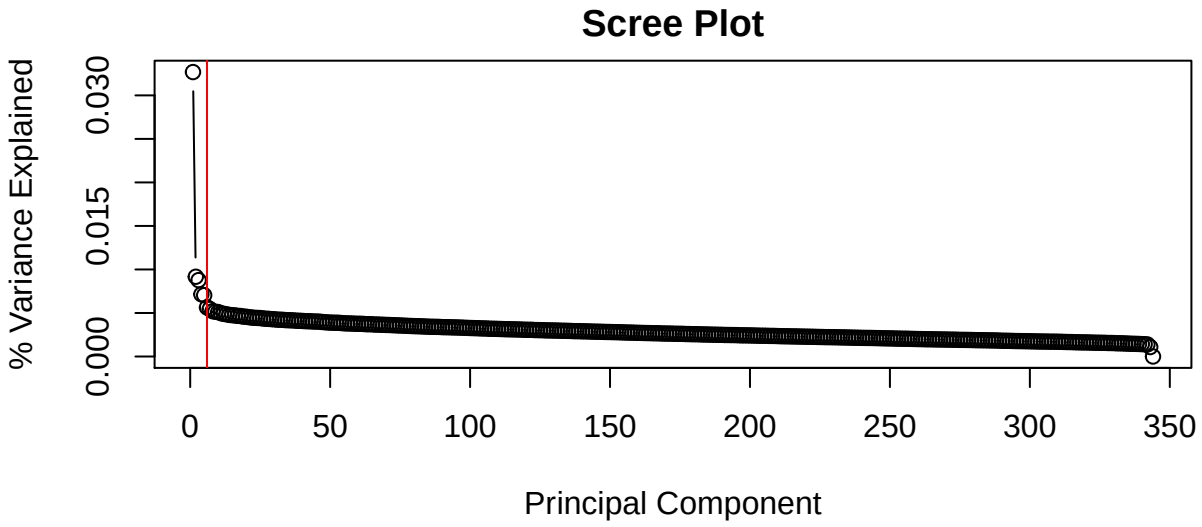
Population as a covariate

Following the multiple linear regression with population as a covariate and with multiple hypothesis correction via the Bonferroni method, there are 66 signficant polymorphisms for ERAP2, 45 signficant polymorphisms for FAHD1, 26 signficant polymorphisms for PEX6. There were no identified statistically significant polymorphisms for GFM1 or MARCHF7.

PC as covariate

To assess the number of PCs to include as covariates, I used two common methods: assessing how many PCs were needed to explain 80% of the variance and the "elbow method." To visualize the relative contribution of each principal component to the percent variance explained, I constructed a scree plot, shown below.

Given how little of the variance is explained unless many principal components are included, I opted not include the PCs as a covariate in the model.

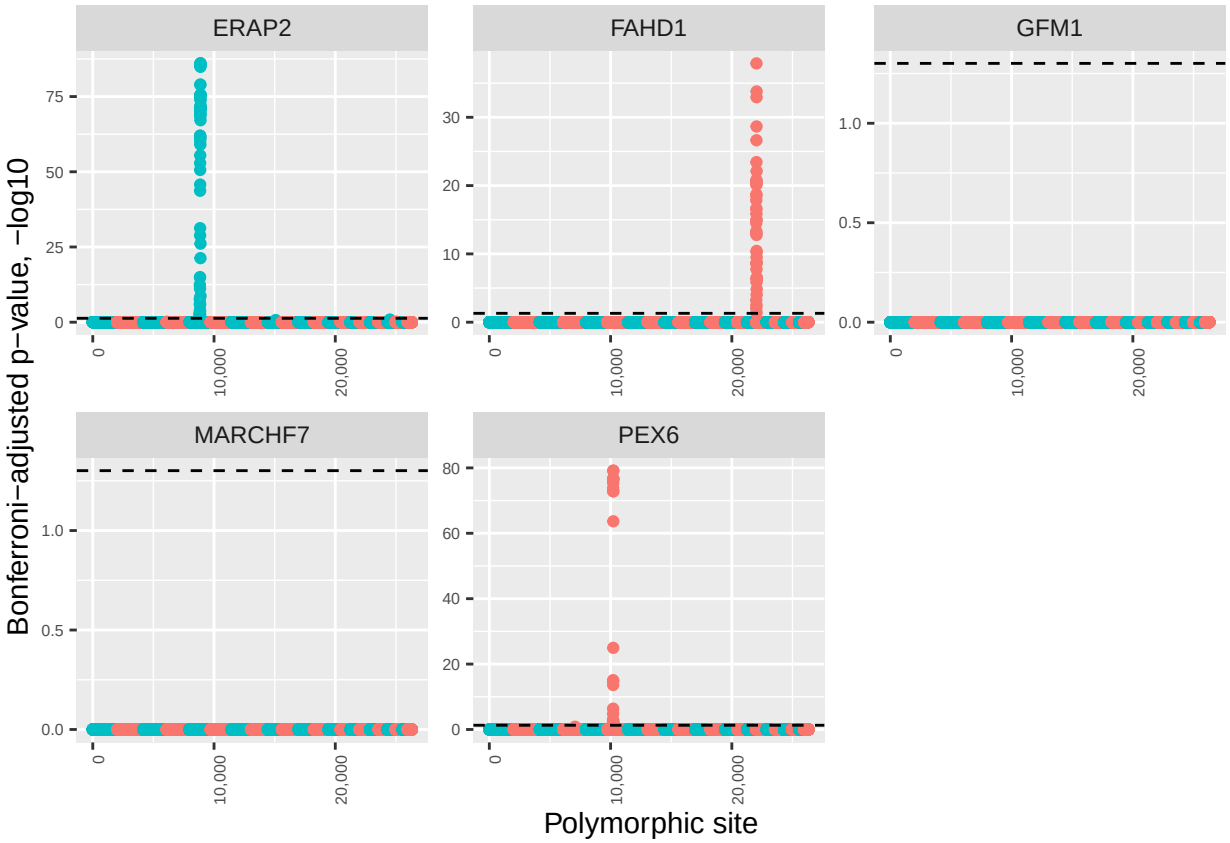


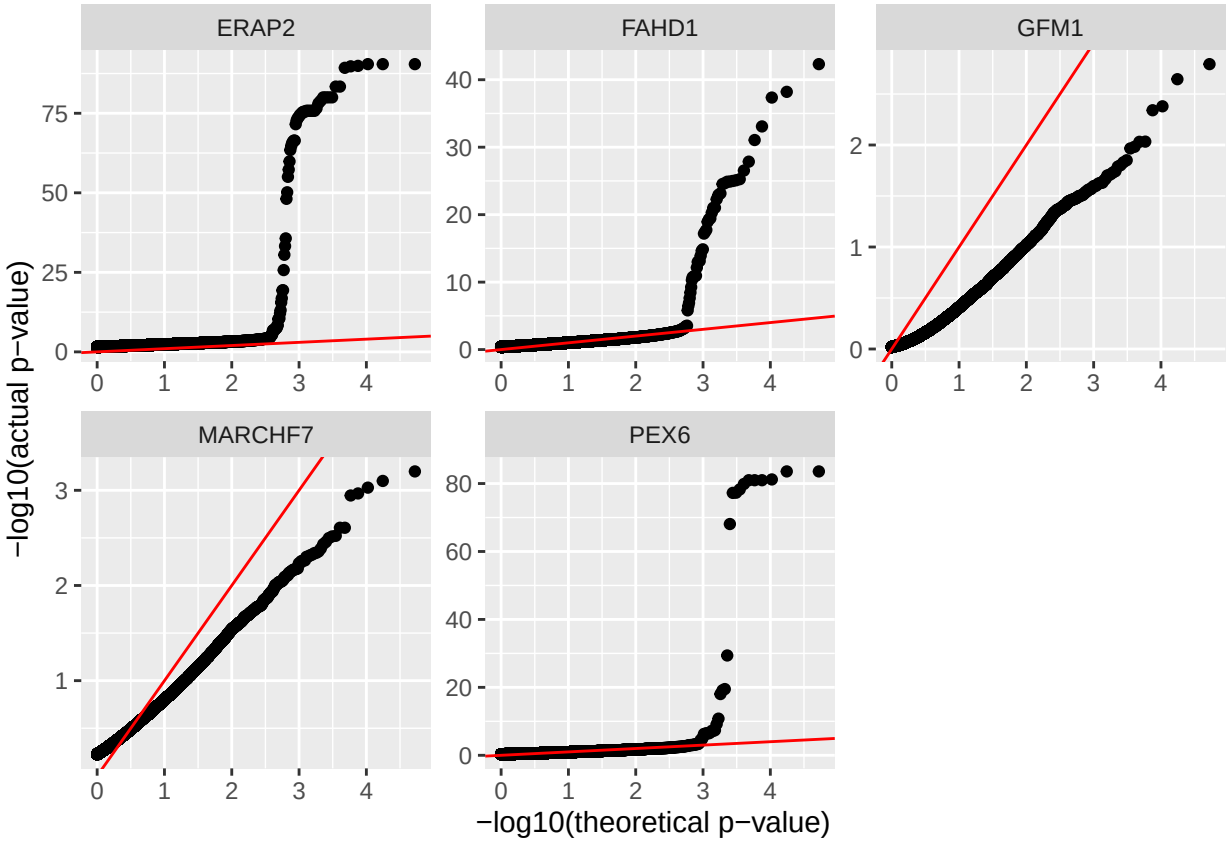

```
## 80% of the variance is explained by the first 234 principal components.
```

```
## 7% of the variance is explained by the first 6 principal components.
```

Sex and population as covariates

Ultimately, I chose to to include both sex and population as covariates, a decision discussed further in the **Summary of eQTL analysis with covariates** section. Below is the corresponding Manhattan plot for each gene of interest. This indicates no potential causal polymorphisms were detected for the GFM1 and MARCHF7 genes. For ERAP2, FAHD1, and PEX6, potentially causal polymorphisms were detected in single, constrained regions of the genome.






```
## There are 66 significant Bonferroni-corrected SNPs associated with the level of ERAP2 expression
## There are 26 significant Bonferroni-corrected SNPs associated with the level of PEX6 expression
## There are 45 significant Bonferroni-corrected SNPs associated with the level of FAHD1 expression
## There are 0 significant Bonferroni-corrected SNPs associated with the level of GFM1 expression
## There are 0 significant Bonferroni-corrected SNPs associated with the level of MARCHF7 expression
```

The QQ-plots above confirm that the significant loci detected for ERAP2, FAHD1, and PEX6 can be interpreted as potentially causal polymorphisms. This is because most of the p-values observed follow a uniform distribution and do not deviate from this line until the significant "tail." In contrast, the QQ-plots for GFM1 and MARCHF7 indicate fewer significant p-values than would normally be expected. Since there were no statistically significant causal polymorphisms found for either GFM1 or MARCHF7, this does not hinder interpretation.

Notably, these observations were true for all combinations of covariates included in the analysis. With one exception, discussed further in **Summary of eQTL analysis with covariates**, all of the statistically significant polymorphisms identified for each gene were confined to a constrained area on a single chromosome, and this region was identical for all analyses, as illustrated with successive Manhattan plots. Statistically significant polymorphisms were never observed for GFM1 or MARCHF7, and the resulting QQ-plots revealed fewer statistically signficant findings than would be expected given a uniform distribution of p-values.

Summary of eQTL analysis with covariates

To the right are heatmaps illustrating how the inclusion of various covariates alters the number of significant SNPs observed. Yellow indicates polymorphic sites that were found to be statistically significant given the procedure described above, and red indicates analyses in which the sites were not found to be statistically significant. The rows indicate which covariates, if any, were included in the multiple regression analysis. The columns indicate the chromosome and the position at which these polymorphisms were found.

In general, the inclusion of covariates is associated with a reduction in the observation of false positive causal polymorphisms. However, the inclusion of the covariates of either sex alone or sex and population actually led to the inclusion of additional significant SNPs both within and at the end of the region already revealed to contain potential causal polymorphisms.

For FAHD1, the inclusion of sex as a covariate eliminated a single statistically significant causal polymorphism from within the identified region. This did not apply when both sex and population were used as covariates.

Finally, and most importantly, for PEX6, when either no covariate or sex as a covariate were included, significant polymorphisms were found on both chromosome 4 and chromosome 6. With the addition of either population or sex and population as covariates, the likely false positive polymorphism on chromosome 4 was no longer statistically signficant, leaving all other potentially causal polymorphisms on chromosome 6, which will be discussed further in **Analysis of potential location of causal polymorphisms**.

For these reasons, for the purposes of additional downstream analyses, I used the model containing sex and population as covariates as the base model.

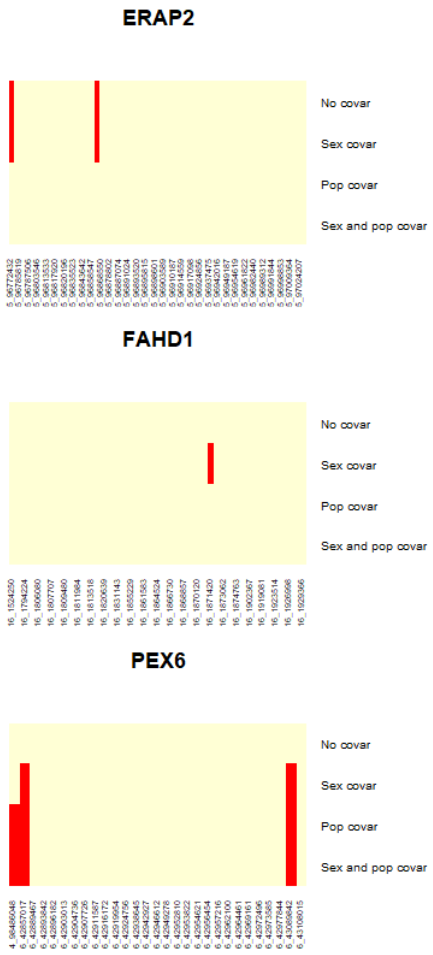


Figure 1: Heatmap of statistically significant polymorphic sites for various covariates

Analysis of potential location of causal polymorphisms

Following the multiple linear regression using sex and population as covariates and with multiple hypothesis correction via the Bonferroni method, there are 66 signficant polymorphisms for ERAP2, 45 signficant polymorphisms for FAHD1, 26 signficant polymorphisms for PEX6.

The figure below illustrates the empirically identified polymorphisms from the eQTL analysis relative to the known position of the genes of interest per the GRCh38 reference genome.[5] All statistically significant polymorphisms occur on the same chromosome as the gene of interest. this also provides us with context that the additional polymorphisms associated with ERAP compared to the other two genes with potentially causal polymorphisms identified can be attributed to the relative size of the gene (43.777) and the fact that numerous SNPs included in the analysis lie directly in the gene body.

The two red markers denote the start and stop site of the known gene locations. Note that in the below the -log10(p-value) values for the actual locations is shown for illustrative purposes only.

To disaggregate the effects of linkage disequilibrium as compared to functional relationships between the polymorphisms, I attempted to characterize the relationship between the significant polymorphisms. To do this, I determined the Pearson correlation of the X_a. I took the square of this value since negative correlations still imply a meaningful relationship in the context of the X_a matrix (e.g., a minor allele of one polymorphism could be associated with a major allele of another polymorphism, implying a direct relationship, but the value would be negative).

[5]Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 May 12]. Available from: https://www.ncbi.nlm.nih.gov/gene/

The values of the resulting coefficients of determination for the various significant polymorphisms are illustrated in the form of an upper triangular heatmap, shown below. Yellow denotes relationships characterized by linkage disequilibrium (i.e., highly correlated), and red denotes relationships that are relatively uncorrelated and suggest the alleles are not in linkage disequilibrium. This suggests that the resulting relationships may in fact be functional in nature, and that the statistically significant relationship between gene expression and polymorphisms at these sites is attributable to something other than linkage disequilbrium.

Indeed, for ERAP2, there is a known downstream gene, LNPEP, which starts at position 96,934,859 and continues through position 97,037,513 that is known to be in linkage disequilibrium with the ERAP2 gene, particularly in the CEU European descendants population. Similarly, there is an upstream regulatory region that is shared by ERAP2 and ERAP1, which is further upstream, but is not in linkage disequlibrium with the ERAP2 gene. This regulatory relationship could explain the finding that they are statistically significantly related without seeming to be in linkage disequilibrium.[6]

Conceivably, similar relatioships could exist for the other genes and the statistically significant identified polymorphisms. The distal, upstream elements are in linkage disequilbrium with their nearest neighbors, but this doesn't seem to extend to the body of the gene of interest, which could indicate a regulatory relationship. For PEX6, neither the upstream nor downstream distal elements seem to be in linkage disequlibrium with SNPs in the gene body, which could similarly indicate a true functional relationship, regulatory or otherwise.

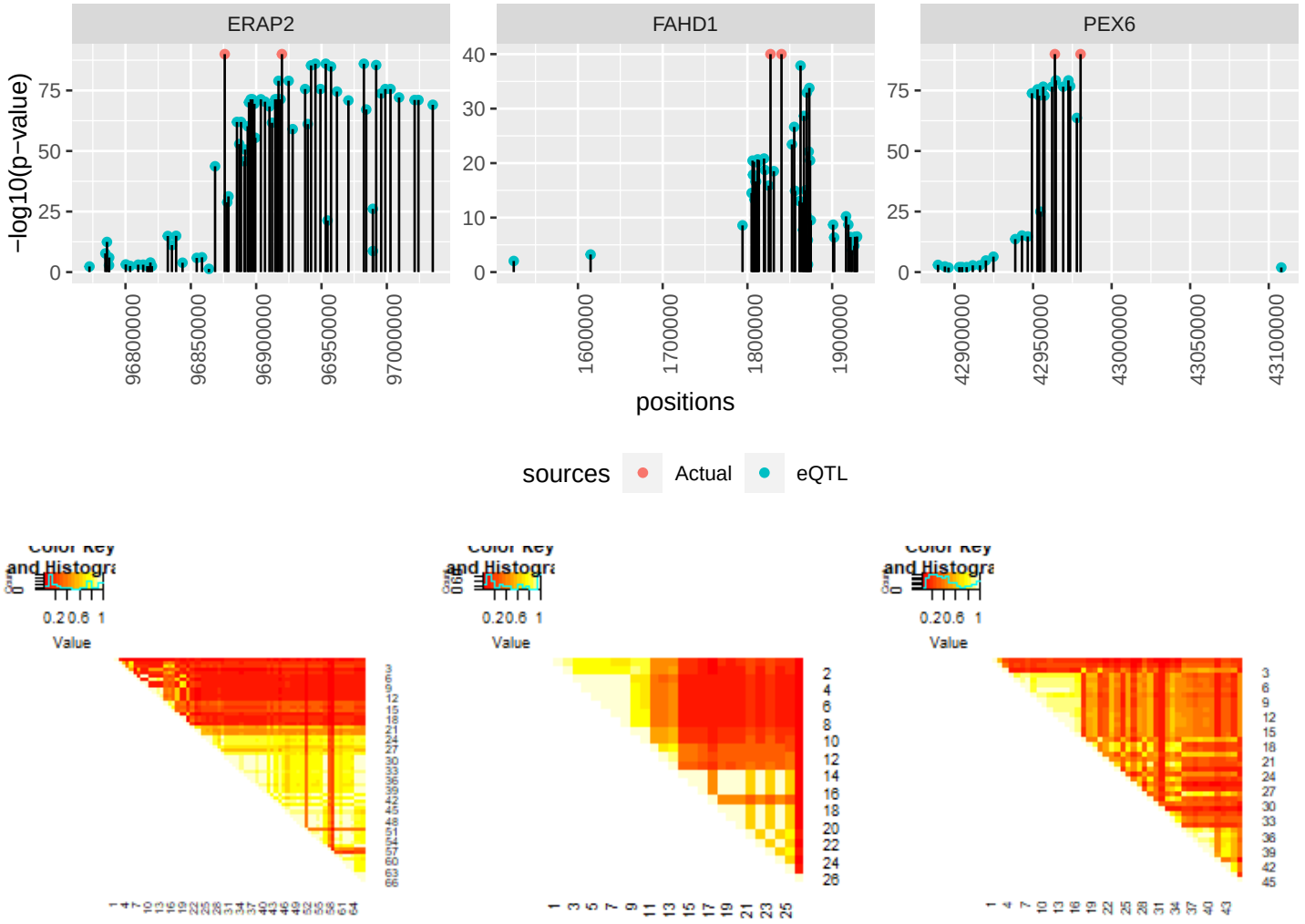


Figure 2: Location of statistically significant causal polymorphisms relative to known gene locations and heatmap of the coefficient of determination for Xa matrix for potential causal polymorphic sites

[6]Andrés, A.M., et al. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genet 6, e1001157 (2010).

References

1. Andrés, A.M., et al. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genet 6, e1001157 (2010).
2. Brittingham, A. & de la Cruz, G.P. Ancestry: 2000. (2004).
3. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 May 12]. Available from: https://www.ncbi.nlm.nih.gov/gene/
4. Lappalainen, T., et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511 (2013).
5. Resource, T.I.G.S. Geuvadis. (2020).