Quantitative Genomics and Genetics 2018 Project

Darya Akimova

May 8, 2018

All of the provided data files were imported successfully and the quality of the data was accessed as follows to ensure that the data is in the expected format.

```
# Are there any missing entries in the data?
anyNA(list(geno, pheno, covars, snp_info, gene_info))
```

```
[1] FALSE
```

```
# Are there approximately equal numbers of people in each covariate group? Is the design balanced?
table(covars$Population)
```

```
CEU FIN GBR TSI
 78  89  85  92
```

```
table(covars$Sex)
```

```
FEMALE   MALE
   181    163
```

```
# Is the coding of the genotypes as expected across all of the data? Any unusual values?
table(as.matrix(geno))
```

```
      0       1       2
8181444 5811217 3207339
```

```
# Are there any genotypes with a minor allele frequency below 5% that need to be removed?
geno_sums <- map_dbl(geno, function(x) sum(x) / (nrow(geno) * 2))
# Do any genotypes have a MAF < 5%
any(geno_sums < 0.05 | geno_sums > 0.95)
```

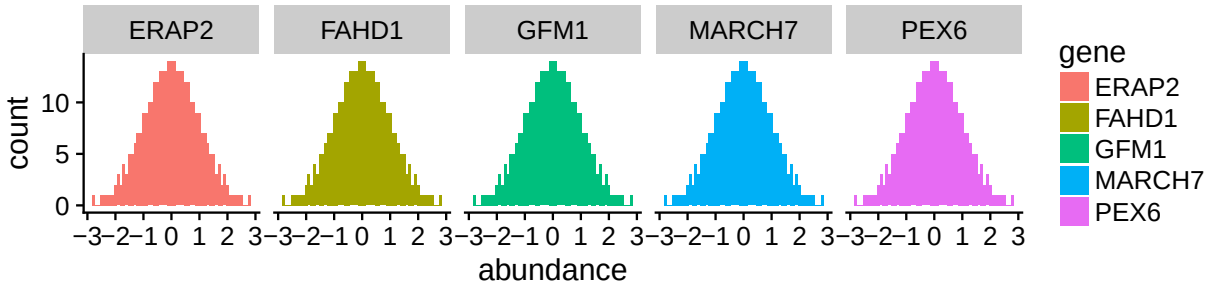
```
[1] FALSE
```

```
# Are there any genotypes without associated phenotypes, or vice versa?
all.equal(rownames(geno), rownames(pheno))
```

```
[1] TRUE
```

In summary, there were no missing values across all of the data and information files provided. There are approximately equal numbers of males and females in the study, and approximately equal numbers of individuals in each of the population groups. Most importantly, none of the covariate groups had a small n and the representation across groups was balanced. The genotypes were coded as expected, with no unusual values. All genotypes were found to have a minor allele frequency of greater than 5%, which indicated that no alleles at any genotype position were too rare for analysis. Lastly, all of the individuals in the genotype dataset were found in the phenotype dataset, and no samples needed to be removed.

Each of the gene expression level distributions were visualized below with a number of methods. In general, the expression level of each gene followed a normal distribution and no extreme outliers were found that needed to be excluded from analysis.



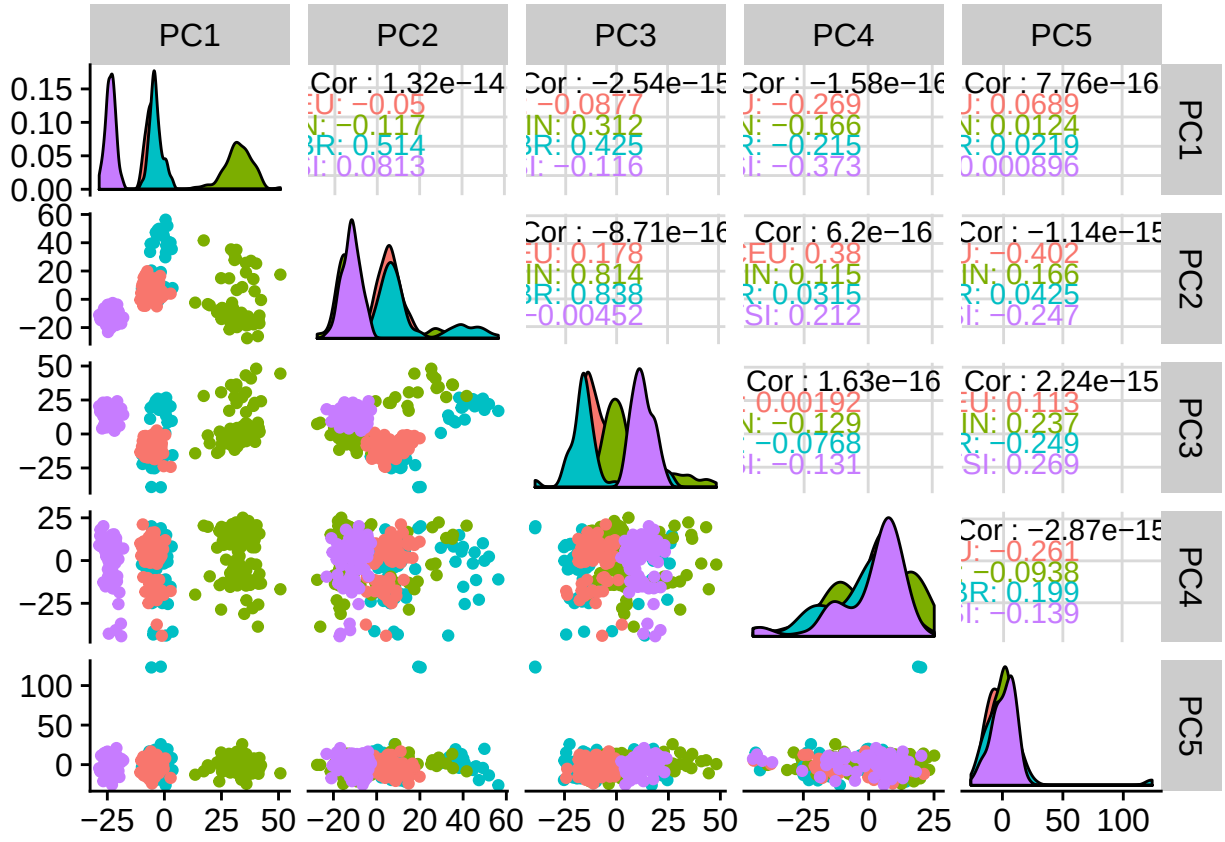
A principal component analysis (PCA) was performed on the phenotype data as an alternative means to determine if there were any outliers or an unusual structure in the data. PCA is a useful method for multi-dimensional data, such as the case here with multiple phenotypes or genotypes per sample. Each principal component represents the maximum explained variance from the original observations, starting with maximum variance explained by the first component and decreasing with subsequent components. Principal components can be plotted to reveal underlying structure in the data and also used for covariate modeling in later analysis.

In the case of the phenotype data, the PCA analysis did not reveal any unusual patterns. Furthermore, none of the phenotype genes were found to be correlated with each other and, therefore, each one can be analyzed in a one by one pairing with each SNP. To visually investigate if the known population and sex covariates could have a potential relationship with the phenotypes, principal components from the phenotype PCA were plotted by a number of methods and the points were colored based on the provided population and sex covariates. These plots were not included in the final report, but the contents of code block 6 can be run in order to access the results. The plots did not suggest a relationship between either covariate.

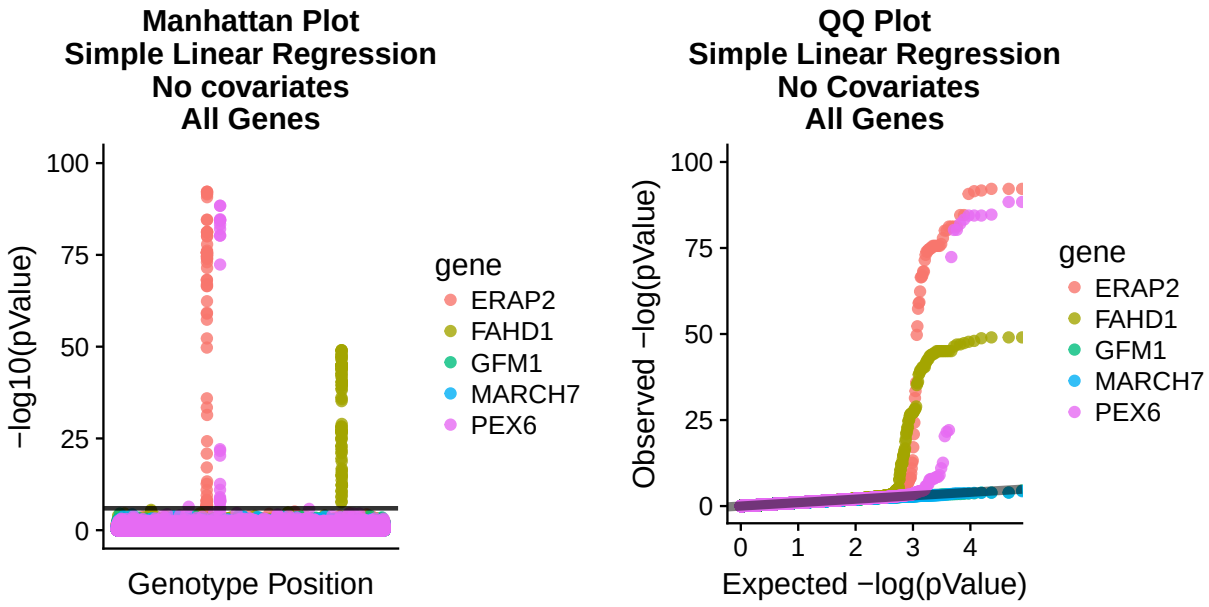
Each SNP position was tested for Hardy-Weinberg Equilibrium using the exact test described in Wigginton *et al* (2005). In this analysis, the test was used as implemented by the HardyWeinberg R package: HardyWeinberg Package on CRAN. The result determined if the proportion of the alleles at a particular genotype position were as expected. A statistically significant result could indicate a problem with the sequencing process or a population structure, both of which can lead to misleading GWAS results.

It was found that 3,436 SNPs, out of the original 50,000, failed the Hardy-Weinberg Equilibrium test with a p-value < 0.05 cut-off. Code block 7 contains the entirety of the Hardy-Weinberg testing process, along with a histogram of the combined distribution of p-values at each SNP position. A Manhattan plot of the p-values, which indicated that the SNPs that reached stasitical significance were scattered throughout the genome. All of these genotypes were removed from further analysis.

PCA analysis was performed on the genotype data, after filtering for SNPs that had failed the Hardy-Weinberg test, in order to determine if there was unusual structure in the genotype data that would need to be accounted for. PCA analysis on the genotype data was conducted in a number of ways: by analyzing the full genotype data, by analyzing a dataset where only every 10th genotype was included, and by an alternative method that was less computationally intensive. Scatterplots were created in code block 8 in order to better understand the results, but only a matrix of plots will be included here. Both the scatterplots below the diagonal, where the first 5 principal components were plotted against each other and colored by the population group, and the histograms of the distribution of each principal component grouped by population, indicate that there is significant structure in the data related to ancestry. The numbers above the diagonal indicate the strength of the relationship between each principal component, but in this case these numbers are not useful and can be ignored. Interestingly, plots of the PCAs in code block 8 suggest that the sex covariate did not seem to play a significant role in the genotype structure.



Although the PCA analysis showed that there was structure in the genotype data, the first analysis that was conducted did not consider any covariates, either provided or derived from the PCA analysis. Briefly, the genotype data were coded into the X_a and X_d matrices, as standard for a GWAS analysis in order to model the additive and dominance genetic effects respectively. The statistical significance of the relationship between each genotype and each phenotype was tested using a simple linear regression. The results for each gene were used to produce a Manhattan Plot, a plot of p-values against each SNP, and a QQ Plot to assess the quality of the model.



Out of all SNPs tested, 66 were found to have a statistically significant association with ERAP2, 82 with FAHD1, and 27 with PEX6, after Bonferroni correction for multiple hypothesis to reduce the number of false positives. No SNPs were found to be associated with the expression of the GFM1 and MARCH7 genes. The R code in block 9 can be run fully to explore all of the results, including on what chromosome the hits were found and what proportion of the hits were inside the genes themselves. However, the odd structure of the QQ Plot, particularly the leveling off of the lowest p-values at the end, suggested that the model should be modified to include at least one covariate.

In order to determine if the provided population, sex, or both covariates should be accounted for, each covariate was tested for a relationship with each phenotype alone, not including the genotype data. The tests varied in how the covariate population is coded, because there were 4 groups and there were a number of ways that this information can be converted into a numerical matrix that could be incorporated into the model.

In summary, the sex covariate, alone or in combination with population, did not have a statistically significant relationship with any of the phenotypes with a p-value cut-off of 0.05. This was not a surprising result, after the PCA analysis of the phenotype in code block 6. The only phenotype that had a statistically significant relationship was the gene ERAP2 with the population covariate. Therefore, the next modeling approach incorporated the population covariate in the test, which was applied to all of the phenotype genes for consistency. A modified version of the test used in the previous modeling attempt was used.

Although population was a promising candidate as a covariate, the results of including it in the analysis carried out in block 11 were identical to the results of the analysis that incorporate covariate at all, with the exception of one fewer hit for PEX6. The resulting QQ plots for ERAP2, FAHD1, and PEX6 still suggested that there was unaccounted for structure in the data. It was possible that the issue was with the coding of the Population covariate, or perhaps a continuous method for including for population information would remove some of the structure in the data. Therefore, the first principal component from the genotype PCA analysis was included as a covariate in the next method because it led to the most defined separation of the population groups, yet the similar CEU and GBR populations overlapped.

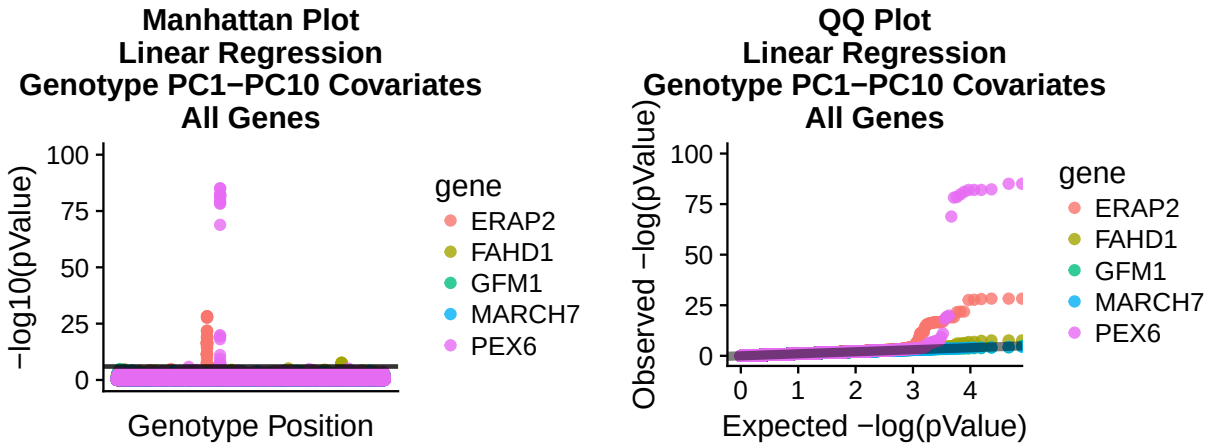
The result of incorporating the first genotype principal component was identical to the previous two approaches, including the number of hits and the structure in QQ Plot. So far, for each analysis, each gene was treated the same, but it may have been more appriate to modify the covariate matrix for each individual phenotype based on their relationship with the covariate. In code block 13, a linear regression was performed on each phenotype as a function of the first four principal components.

The principal components for which there was a statistically signficiant relationship between one of the genes

were included as covariates for that gene in the analysis carried out in code block 14. However, the results were identical to the previous modeling attempts. In an effort to determine if the including more principal components in general would improve the quality of the results, principal components 1 to 10 were included as covariates in the last linear regression modeling attempt.

Including more principal components and covariates did not completely resolve the issues with the behavior of the p-values, as demonstrated by the QQ Plot. A linear mixed model, an alternative to the linear regression which can incorporate a relationship matrix between individuals, was applied in block 16 in order to determine if the population structure suggested by the previous QQ Plots could be modeled out. The first three principal components from the genotype PCA analysis were included in the model to account for the ancestry of the individuals in the study.

The statistically significant SNP positions and the QQ Plot were, again, largely identical to the previous models and this attempt did not succeed in improving the quality of the results. The QQ Plot produced by the analysis that included PC1-PC10 as covariates appeared to be the most normal, and those results will be interpreted here.



In conclusion, 41 SNPs were found to be related to ERAP2 expression, 44% of which were in the coding region of the gene, and all were on the same chromosome in nearby regions. All of the 6 FAHD1 hits were within the gene coding region. Lasly, 15 SNPs were found to have a statistically significant effect on PEX6 expression. The last section of this code can be uncommented to export the results to a csv for further analysis.