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)
```

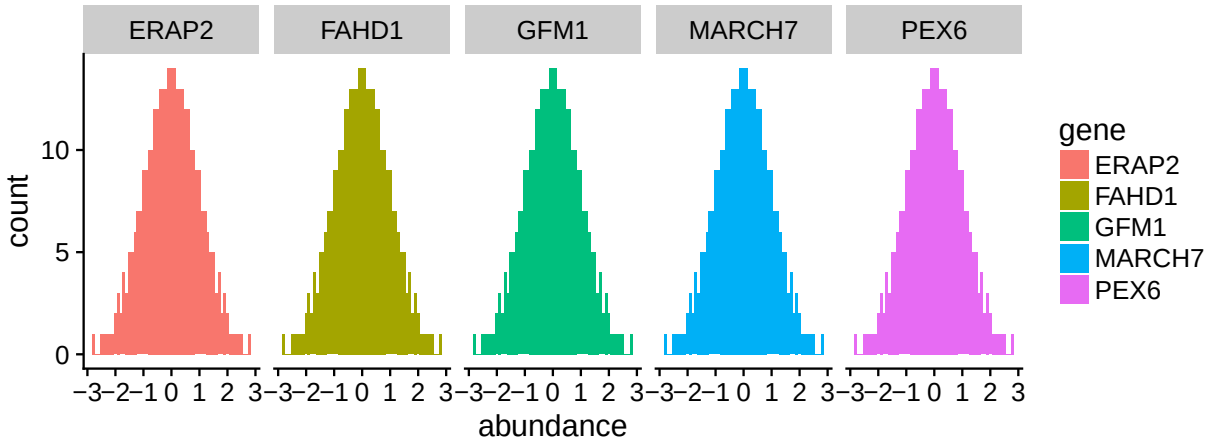
```
[1] FALSE
```

```
# 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 of groups is balanced. The genotype data were coded as expected, with no unusual values. All genotypes were found to have a minor allele frequency of greater than 5%, which indicates that no alleles at any genotype position were too rare for analysis. Lastly, all of the individuals in the genotype dataset are found in the phenotype dataset and no samples need to be filtered out on this criteria.

Each of the gene expression levels included in the phenotype data were visualized below. 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.

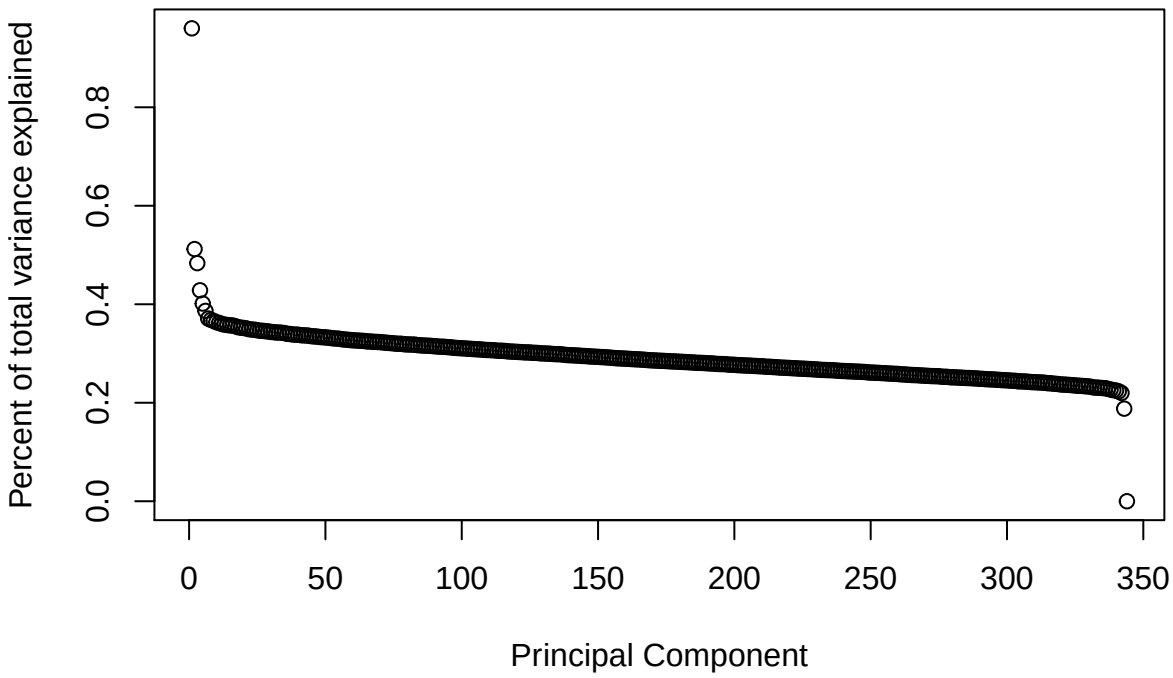


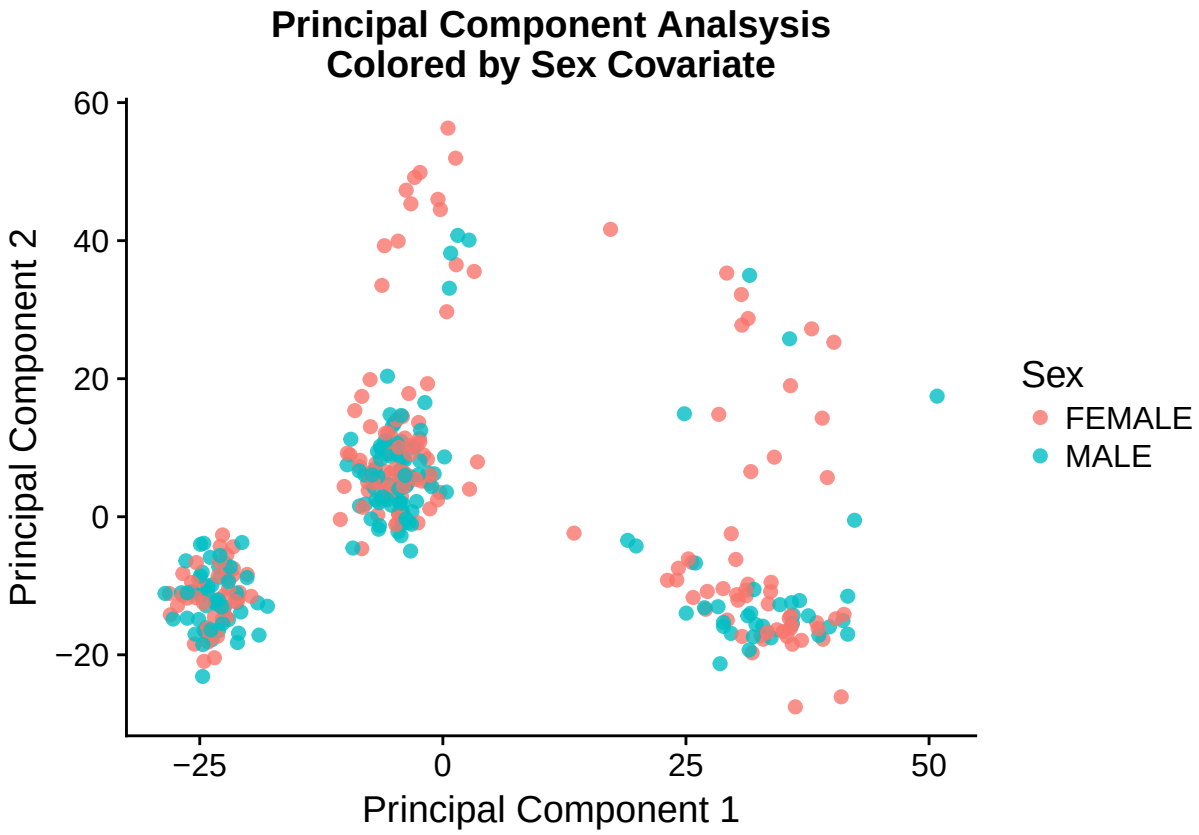
A principal component analysis (PCA) was performed on the phenotype data as an alternative means to determine if there are any outliers or an unusual structure. In brief, PCA is a useful method for multi-dimensional data, such as the case here with multiple phenotypes or genotypes per sample. Each principal component contains maximum shared information from the original observations in a step-wise fashion from each component to the next. Principal components can be plotted to reveal underlying structure in the data and also used for covariate modeling in later analysis.

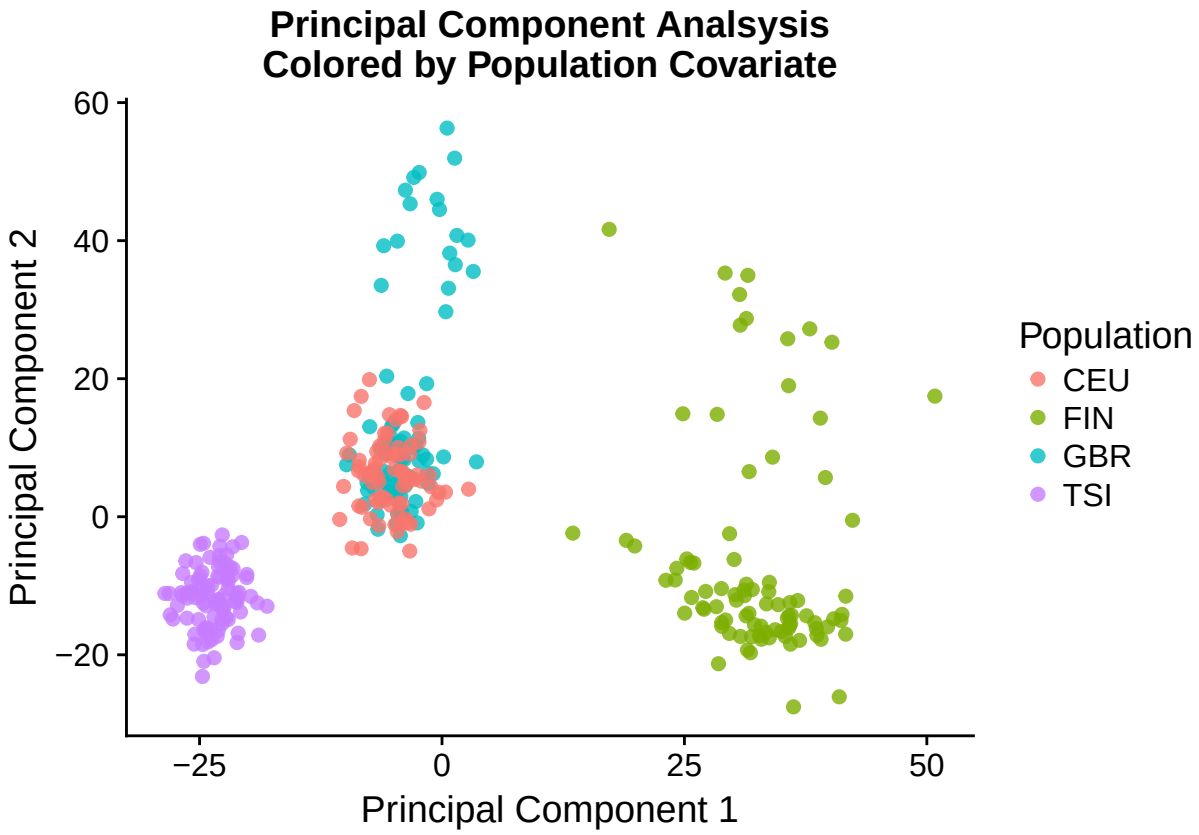
In the case of the phenotype data, the PCA analysis did not reveal any unusual patterns. Furthermore, none of the phenotype genes were found to be correlated with each other to a meaningful degree and, therefore, each one can be analyzed in a one by one pairing with each genotype. To visually investigate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

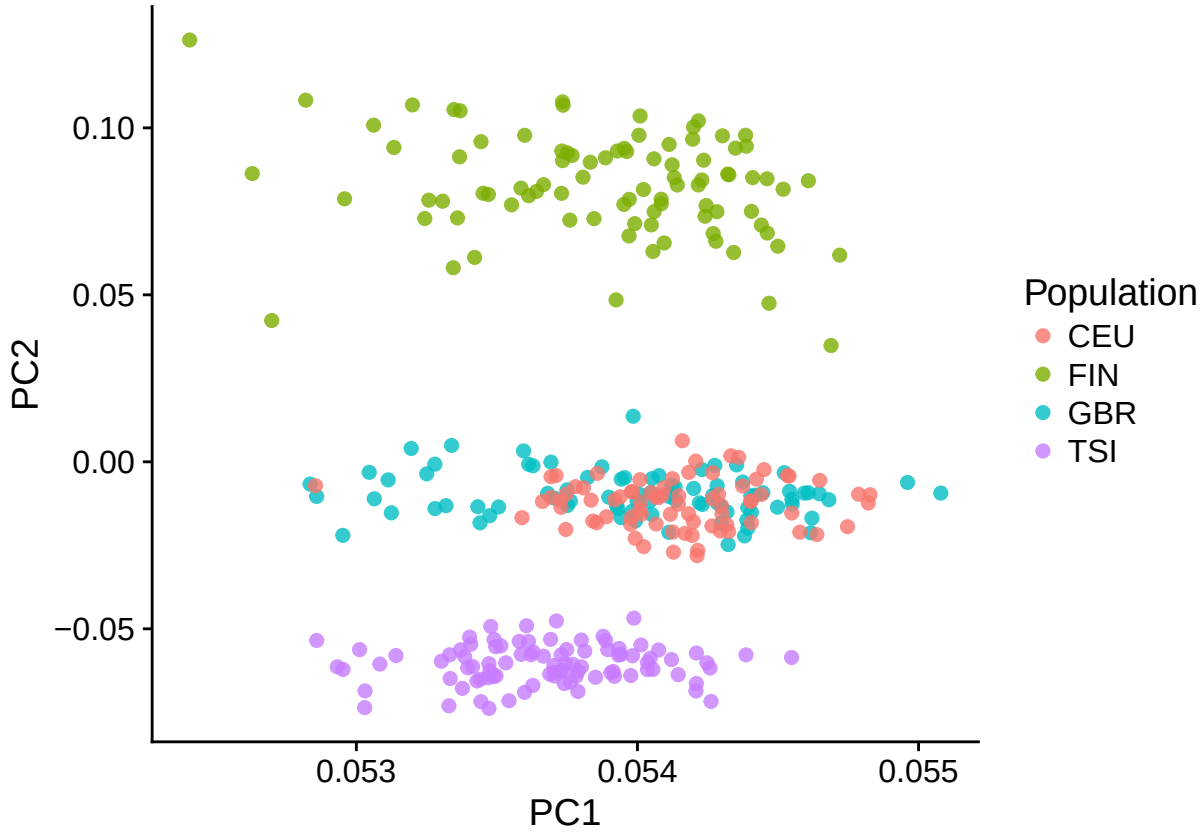
Each SNP position was tested for Hardy-Weinberg Equilibrium using the exact test published by Wigginton *et al* (2005). This is the same test that is performed by default by the Genome Association Analysis software PLINK. In this analysis, the test was used as implemented by the HardyWeinberg R package: HardyWeinberg Package on CRAN. This test determined if the proportion of alleles at a particular genotype position were as expected. A statistically significant result can indicate a problem with the sequencing process or a population structure, both of which can result in misleading GWAS results.

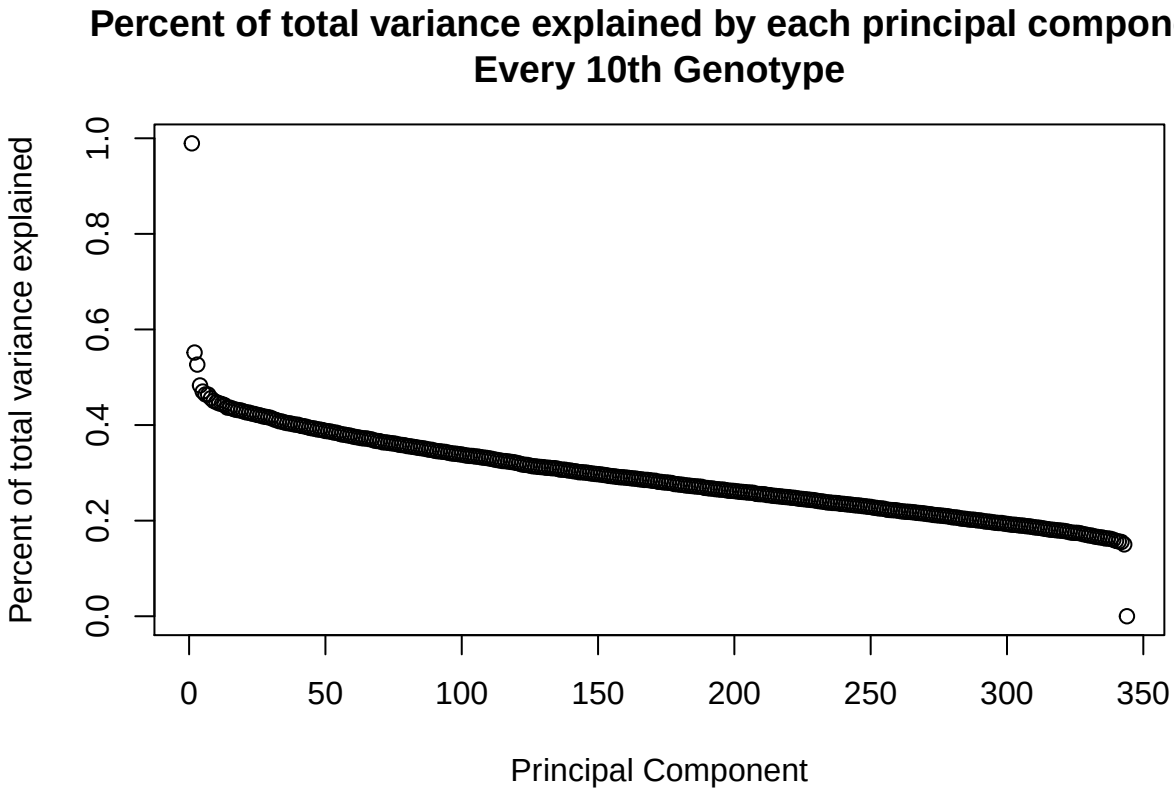
It was found that 3,436 SNPs, out of the original 50,000, failed the Hardy-Weinberg Equilibrium test with a p-value < 0.05 cut-off. Code block 7 contains the entirety of the Hardy-Weinberg testing process, along with a histogram of the p-values for each SNP position and a Manhattan plot of the p-values,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

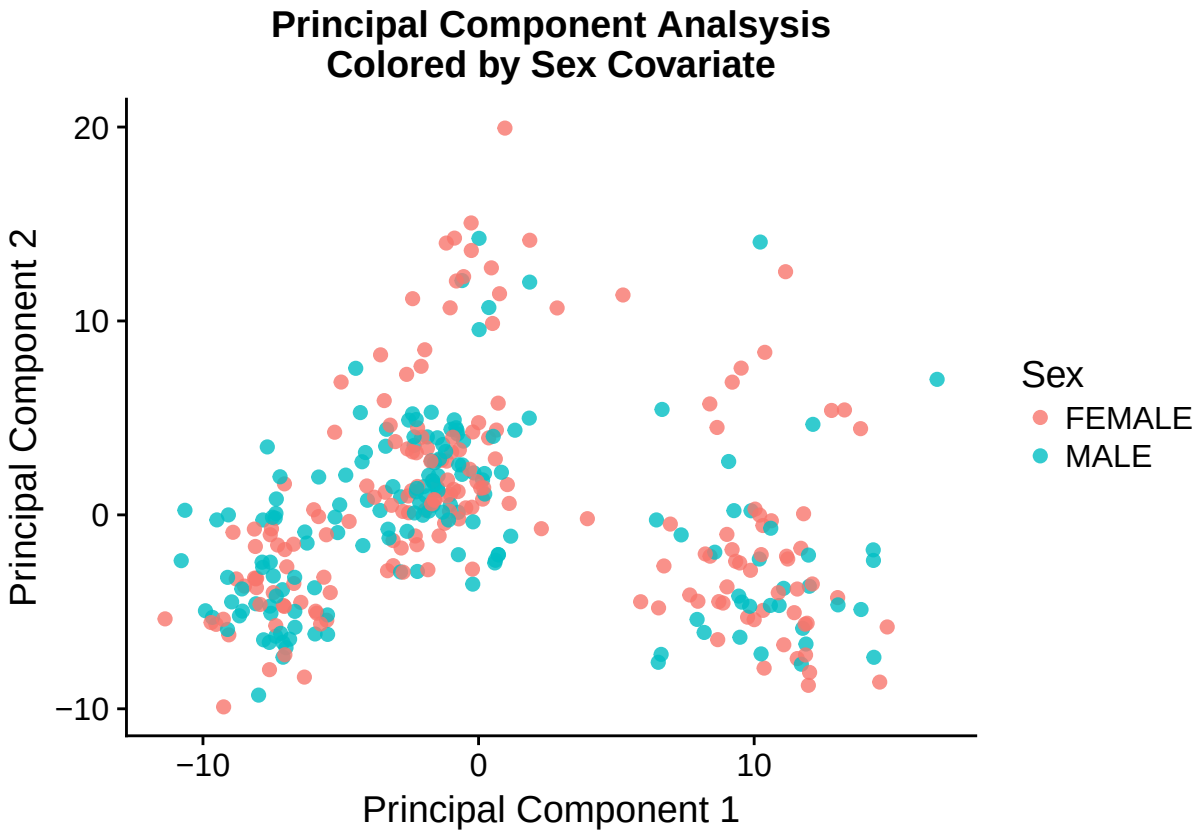

Percent of total variance explained by each principal component
Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
0
50
100
150
200
250
300
350
Principal Component

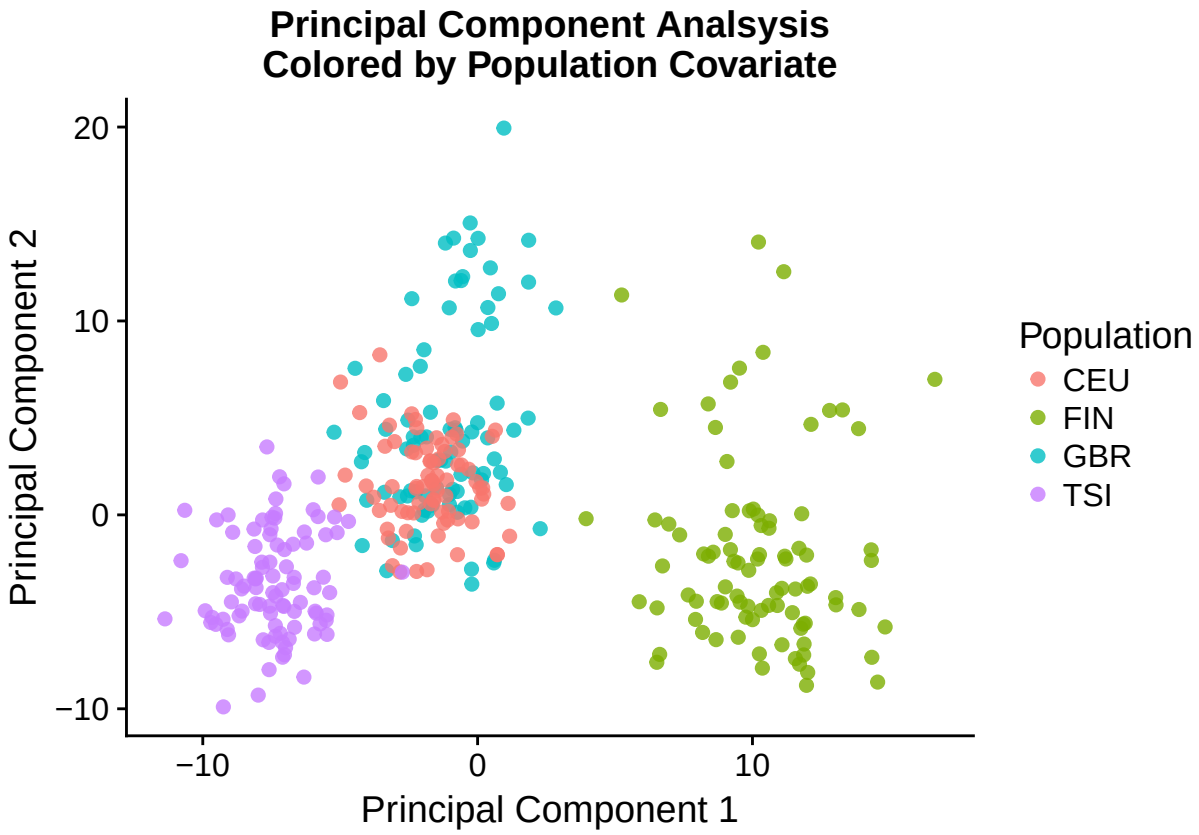

Principal Component Analsysis
Colored by Sex Covariate
Principal Component 2
60
40
20
0
−20
−25
0
25
50
Principal Component 1
Sex
FEMALE
MALE

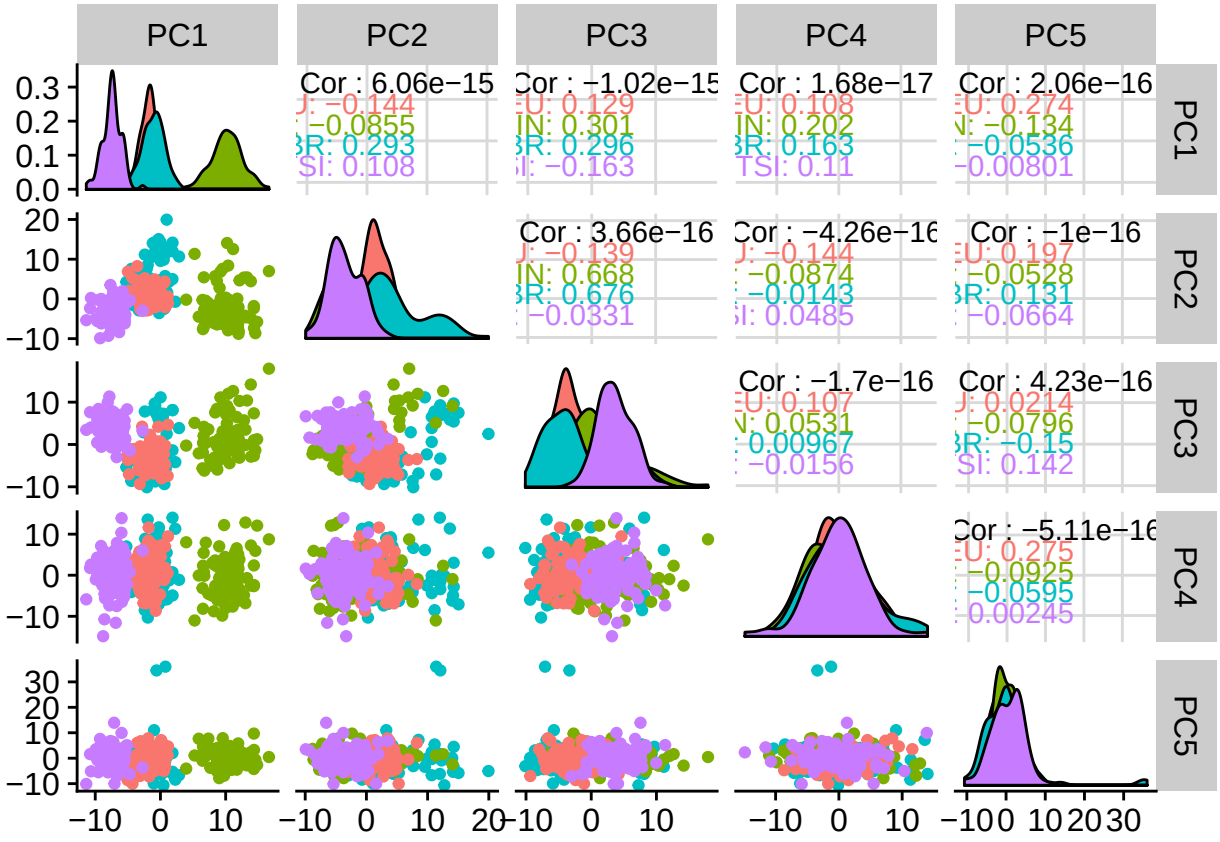

Principal Component Analsysis
Colored by Population Covariate
Principal Component 2
60
40
20
0
−20
−25
0
25
50
Principal Component 1
Population
CEU
FIN
GBR
TSI

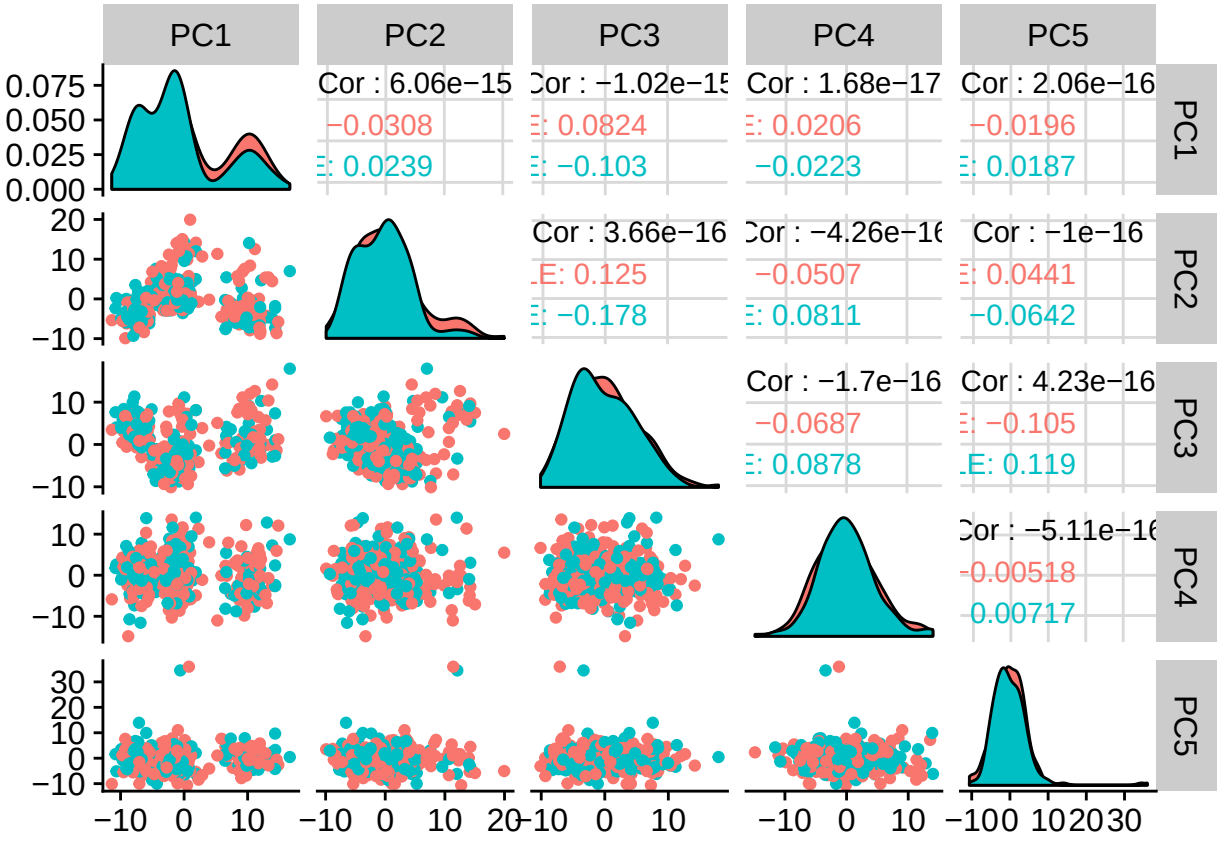

0.10
0.05
0.00
−0.05
PC2
0.053
0.054
0.055
PC1
Population
CEU
FIN
GBR
TSI

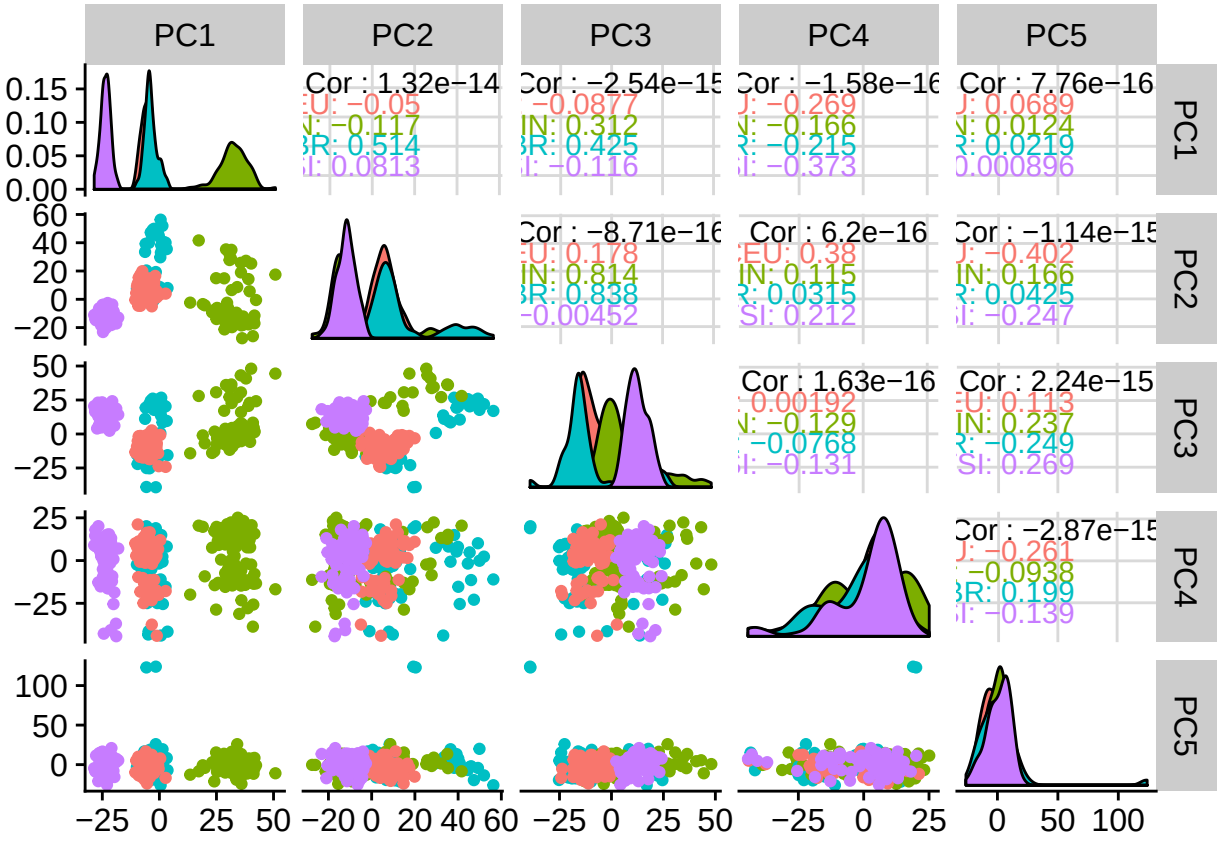

Percent of total variance explained by each principal component
Every 10th Genotype
Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Principal Component

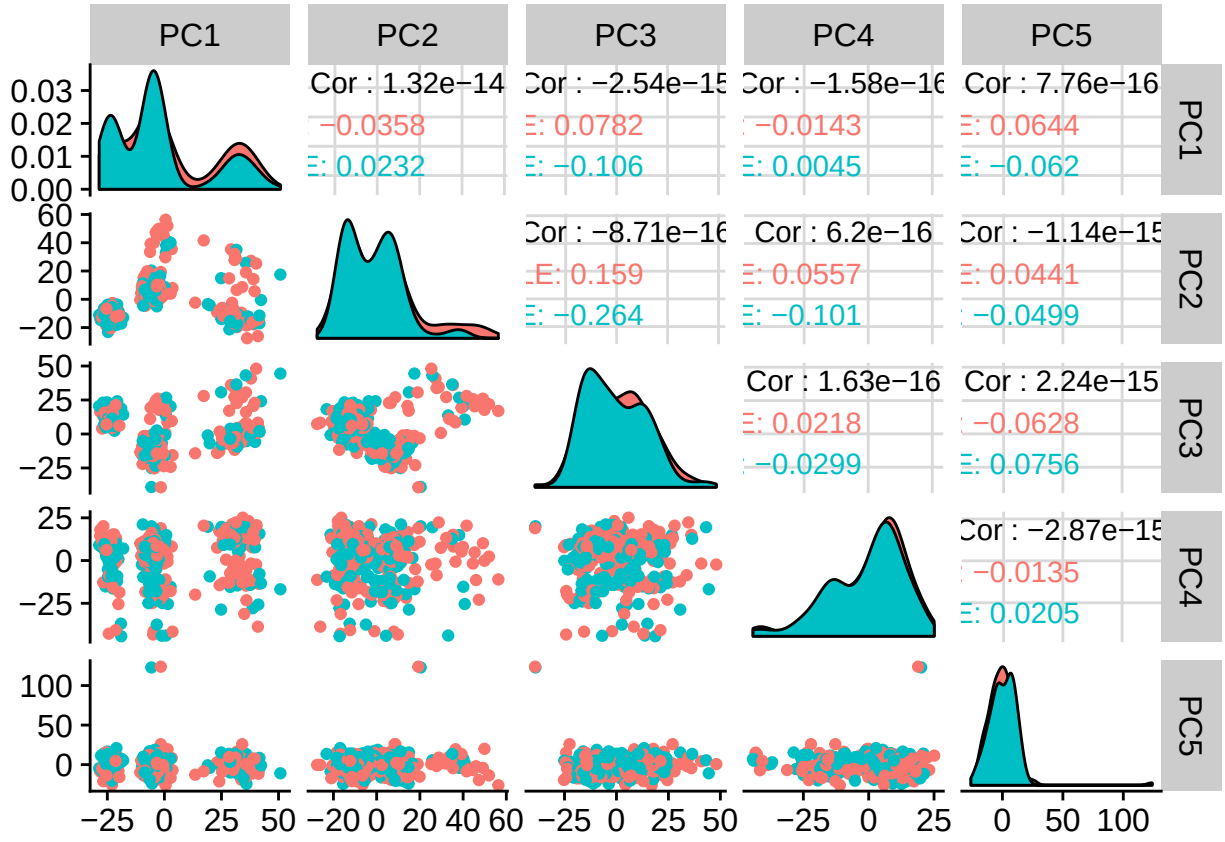

Principal Component Analsysis
Colored by Sex Covariate
Principal Component 2
20
10
0
−10
−10
0
10
Principal Component 1
Sex
FEMALE
MALE


Principal Component Analsysis
Colored by Population Covariate
Principal Component 2
20
10
0
−10
−10
0
10
Principal Component 1
Population
CEU
FIN
GBR
TSI


PC1
PC2
PC3
PC4
PC5
Cor : 6.06e−15
Cor : −1.02e−15
Cor : 1.68e−17
Cor : 2.06e−16
Cor : 3.66e−16
Cor : −4.26e−16
Cor : −1e−16
Cor : −1.7e−16
Cor : 4.23e−16
Cor : −5.11e−16

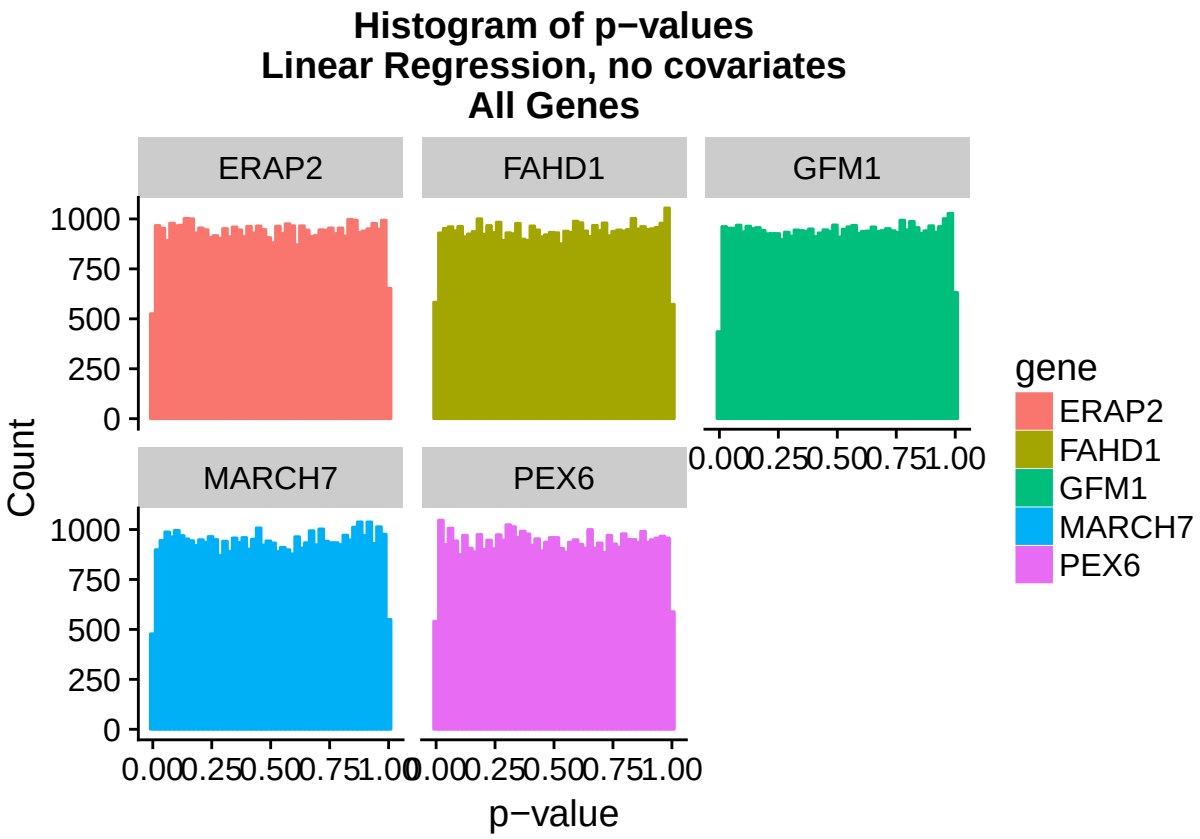

PC1
PC2
PC3
PC4
PC5
Cor : 6.06e−15
−0.0308
: 0.0239
Cor : −1.02e−15
: 0.0824
: −0.103
Cor : 1.68e−17
: 0.0206
−0.0223
Cor : 2.06e−16
−0.0196
: 0.0187
Cor : 3.66e−16
E: 0.125
: −0.178
Cor : −4.26e−16
−0.0507
: 0.0811
Cor : −1e−16
: 0.0441
−0.0642
Cor : −1.7e−16
−0.0687
: 0.0878
Cor : 4.23e−16
: −0.105
E: 0.119
Cor : −5.11e−16
−0.00518
0.00717

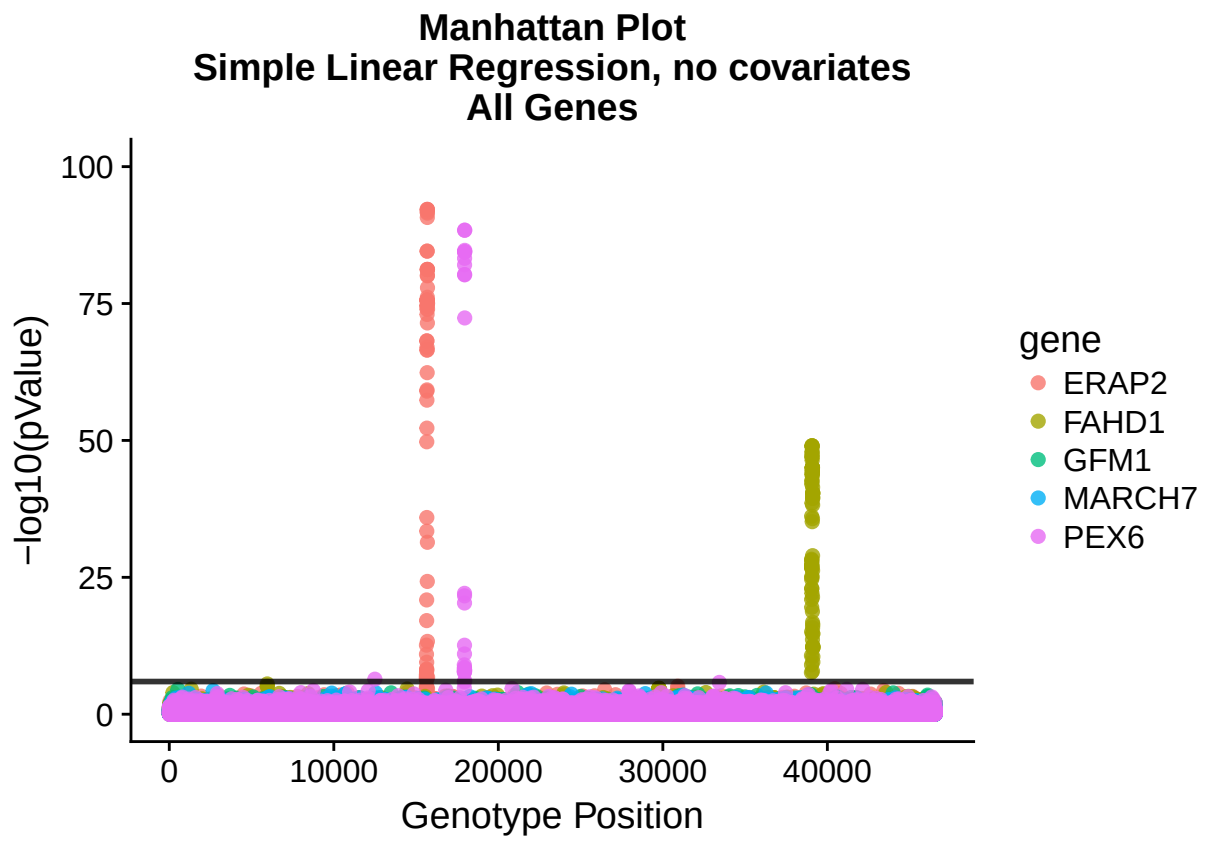

PC1
PC2
PC3
PC4
PC5
Cor : 1.32e−14
Cor : −2.54e−15
Cor : −1.58e−16
Cor : 7.76e−16
Cor : −8.71e−16
Cor : 6.2e−16
Cor : −1.14e−15
Cor : 1.63e−16
Cor : 2.24e−15
Cor : −2.87e−15



Based on on the principal component analysis of the genomes, colored by the population of origin, it is possible to see that there is clearly population structure. Interestingly, the principal component analysis suggests that the Sex covariate does not seem to play a significant role in the genotype structure. This makes sense since the genotype data does not include the sex chromosomes. The first principal component seems to carry most of the population-related variance, regardless of how the principal components are calculated.

The first type of analysis that was conducted did not consider any covariates, either provided or derived from the PCA analysis. The genotype data that was used, was the set filtered on the Hardy-Weinberg Equilibrium test significant genes.


Histogram of p−values
Linear Regression, no covariates
All Genes
ERAP2
FAHD1
GFM1
MARCH7
PEX6
Count
1000
750
500
250
0
0.00
0.25
0.50
0.75
1.00
p−value
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6




```
# A tibble: 3 x 2
# Groups:   gene [3]
  gene      n
  <chr> <int>
1 ERAP2    66
2 FAHD1    82
3 PEX6     27
```

```
[1] 31.81818
```

```
[1] 5
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
96774230 96844130 96901461 96903313 96954278 97110808
```

```
[1] 12.19512
```

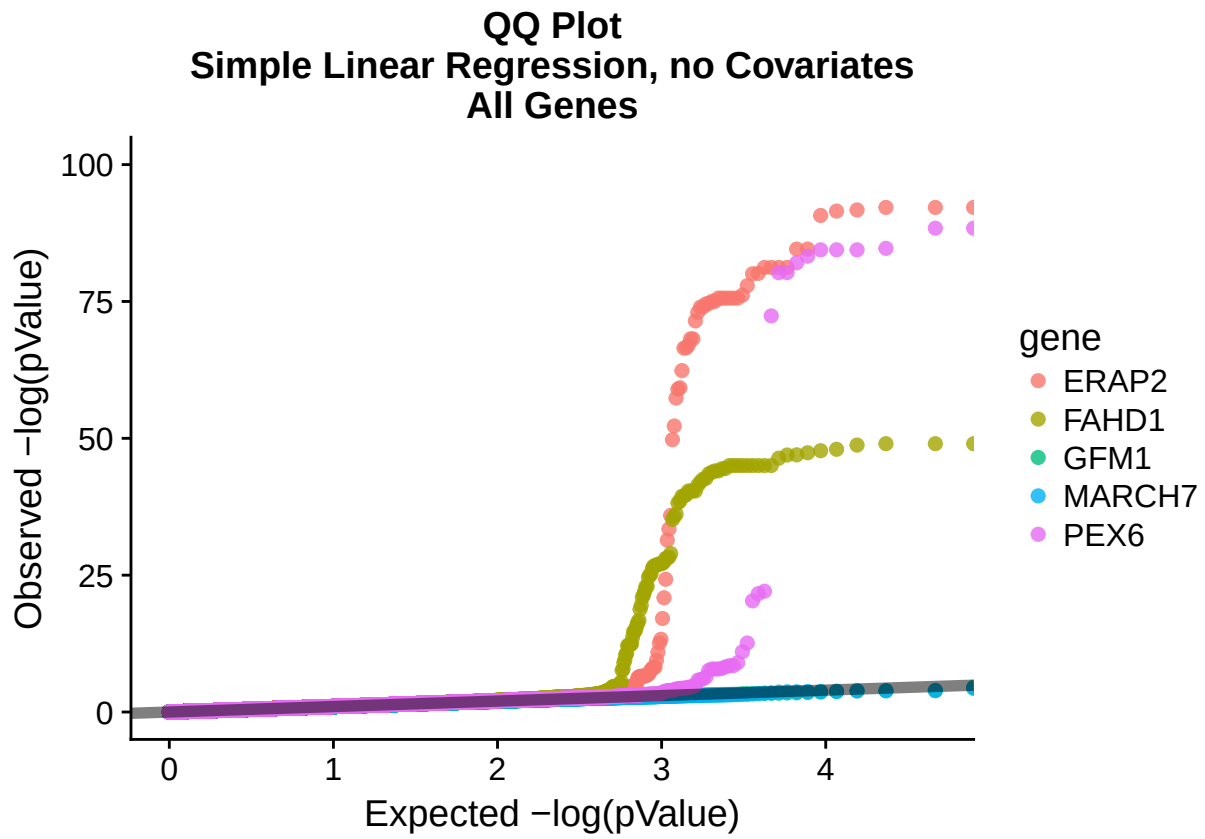
```
[1] 16
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1524250 1816976 1847491 1842589 1871763 1929366
```

```
[1] 18.51852
```

```
[1] 4 6
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
42889467 42912733 42947945 42949335 42963871 43108015
```



A number of SNPs were found to be associated with the expression levels of ERAP2, FAHD1, and PEX6. No SNPs were found to be associated with the expression of the GFM1 and MARCH 7 genes. The R code in Block 8 can be run in the notebook to explore all of the results. Based of the PCA analysis, it appears that the Population may be an important covariate to include in the models.

But first, let's test the relationship of each phenotype with each covariate:

```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2 0.00799
```

```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2 0.00914
```

```
[1] FALSE
```

```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2 0.00787
```

```
# A tibble: 3 x 6
  gene  term                           estimate std.error statistic p.value
  <chr> <chr>                             <dbl>     <dbl>     <dbl>   <dbl>
1 ERAP2 (Intercept)                       0.399    0.140       2.86 0.00454
2 ERAP2 as.numeric(factor(Population~    -0.124    0.0474     -2.63 0.00901
```

```
3 ERAP2 factor(Sex)MALE                -0.172    0.105      -1.64 0.103
```

```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2 0.00678
```

```
# A tibble: 5 x 6
  gene  term                  estimate std.error statistic p.value
  <chr> <chr>                    <dbl>     <dbl>     <dbl>   <dbl>
1 ERAP2 (Intercept)              0.273     0.123     2.21  0.0278
2 ERAP2 factor(Population)FIN   -0.207     0.152    -1.37  0.172
3 ERAP2 factor(Population)GBR   -0.0606    0.153    -0.397 0.692
4 ERAP2 factor(Population)TSI   -0.458     0.150    -3.06  0.00241
5 ERAP2 factor(Sex)MALE         -0.172     0.106    -1.63  0.105
```

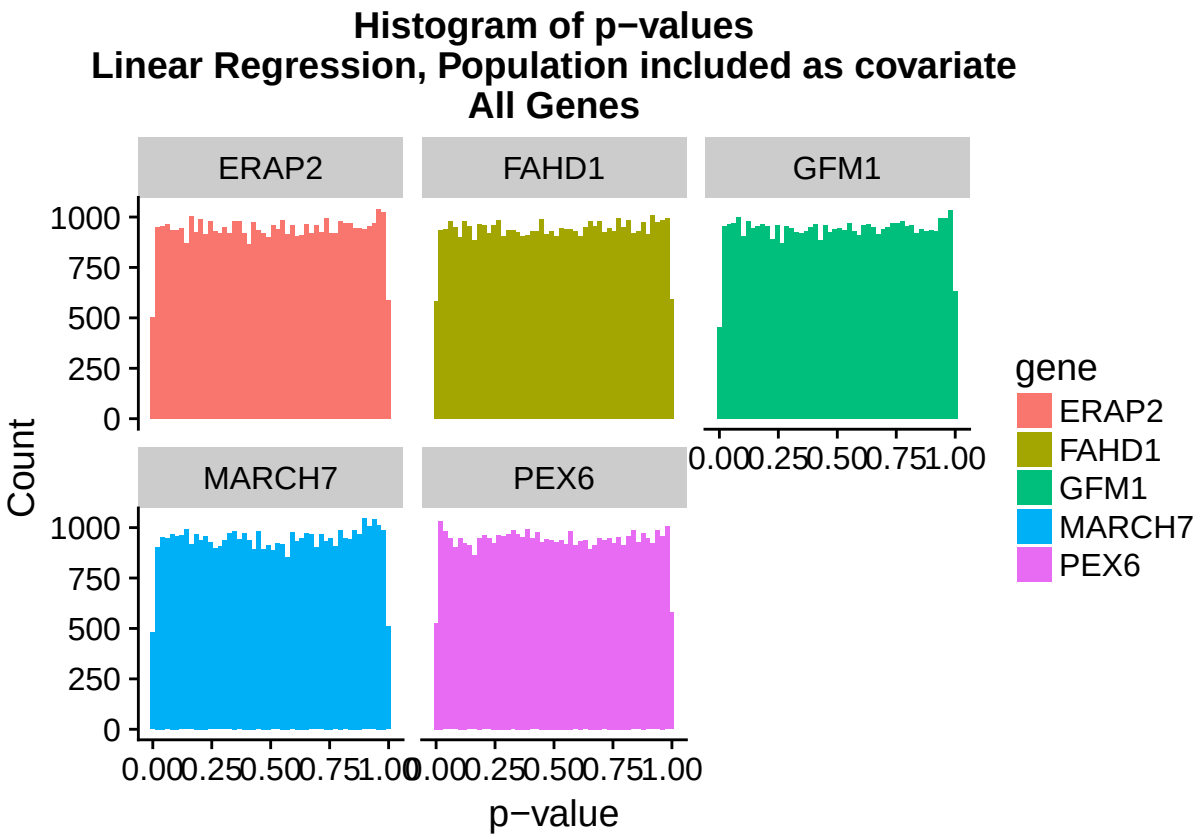
```
[1] FALSE
```

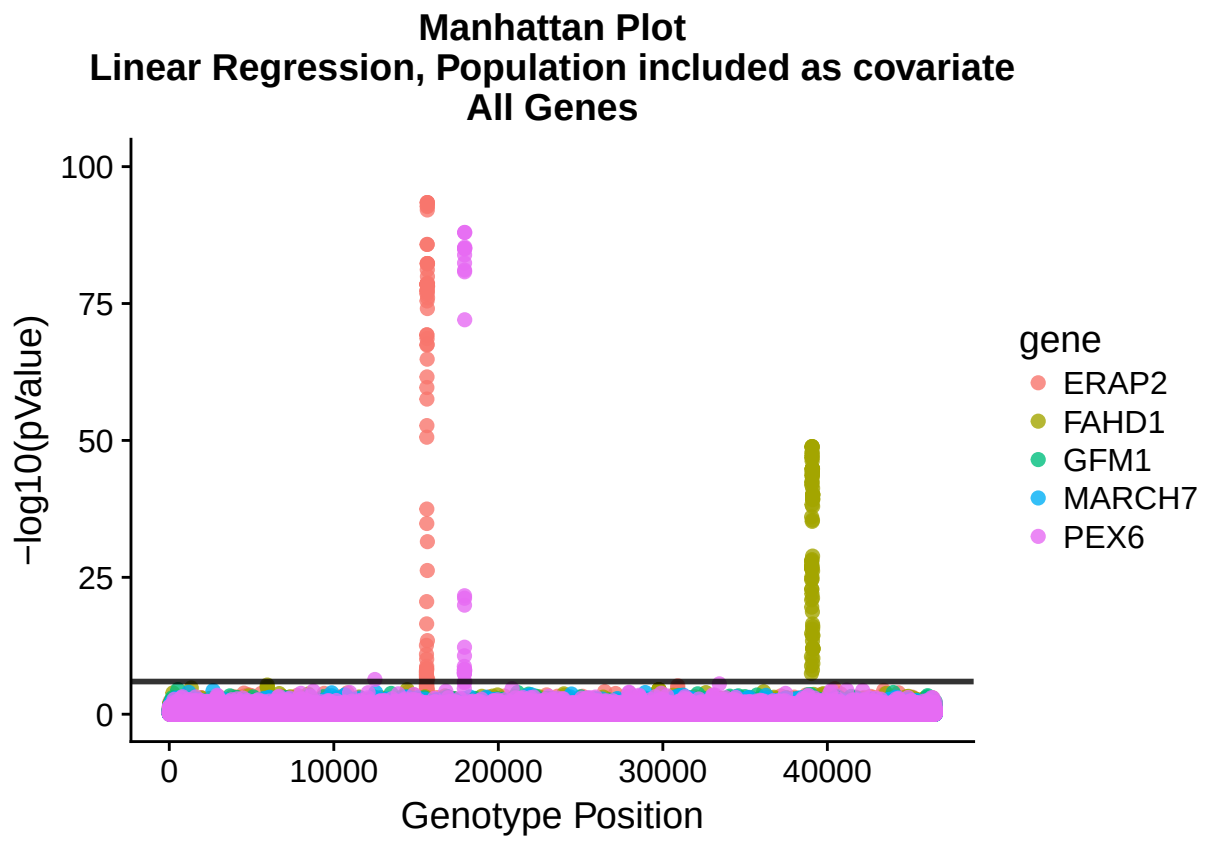
```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2 0.00326
```

```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2 0.00286
```

```
# A tibble: 4 x 6
  gene  term                  estimate std.error statistic p.value
  <chr> <chr>                    <dbl>     <dbl>     <dbl>   <dbl>
1 ERAP2 (Intercept)              0.0644    0.111     0.583  0.561
2 ERAP2 factor(Population)TSI   -0.250     0.145    -1.72   0.0856
3 ERAP2 factor(Population)WEU    0.176     0.128     1.37   0.173
4 ERAP2 factor(Sex)MALE         -0.170     0.106    -1.61   0.108
```

Including Population as given as a covariate:


Histogram of p−values
Linear Regression, Population included as covariate
All Genes
ERAP2
FAHD1
GFM1
MARCH7
PEX6
Count
1000
750
500
250
0
0.00
0.25
0.50
0.75
1.00
p−value
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6




```
# A tibble: 3 x 2
# Groups:   gene [3]
  gene      n
  <chr> <int>
1 ERAP2    66
2 FAHD1    82
3 PEX6     26
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15628   15646   15666   15665   15683   15701
```

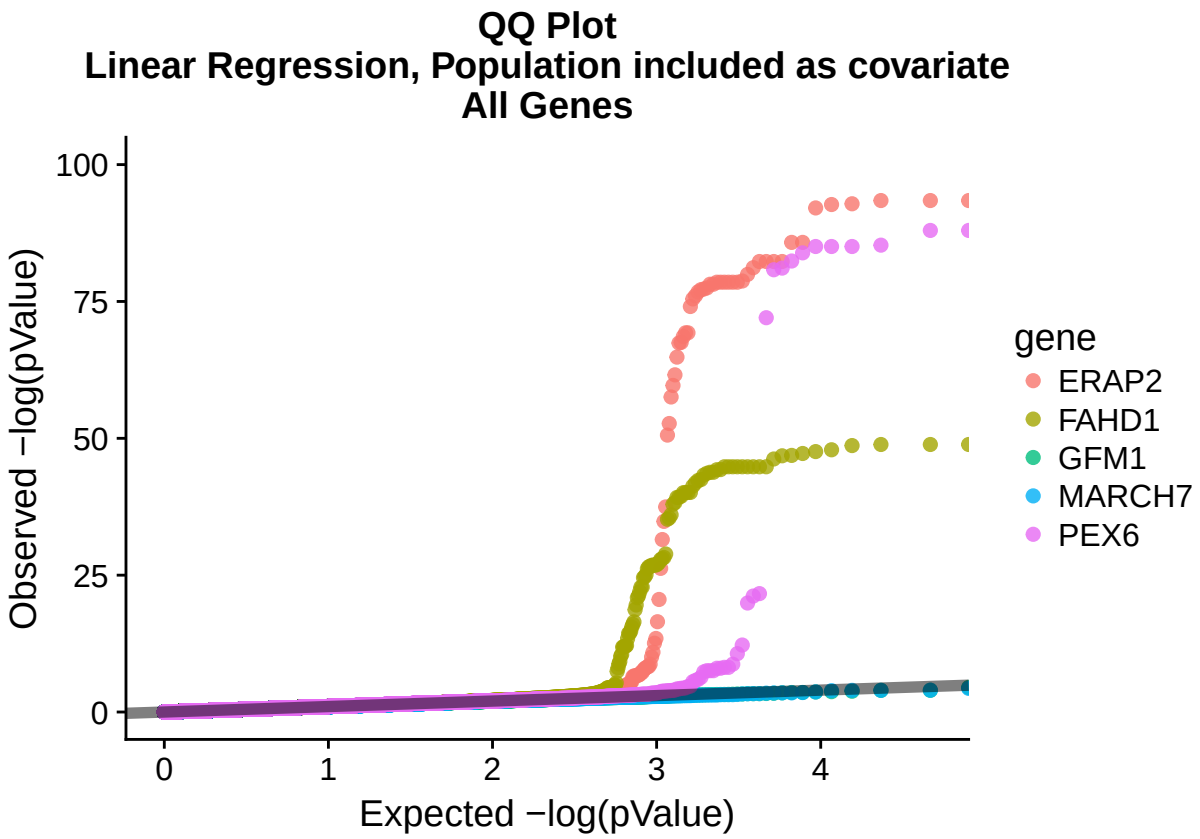
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15628   15646   15666   15665   15683   15701
```

```
[1] TRUE
```

```
[1] TRUE
```

```
[1] 4 6
```

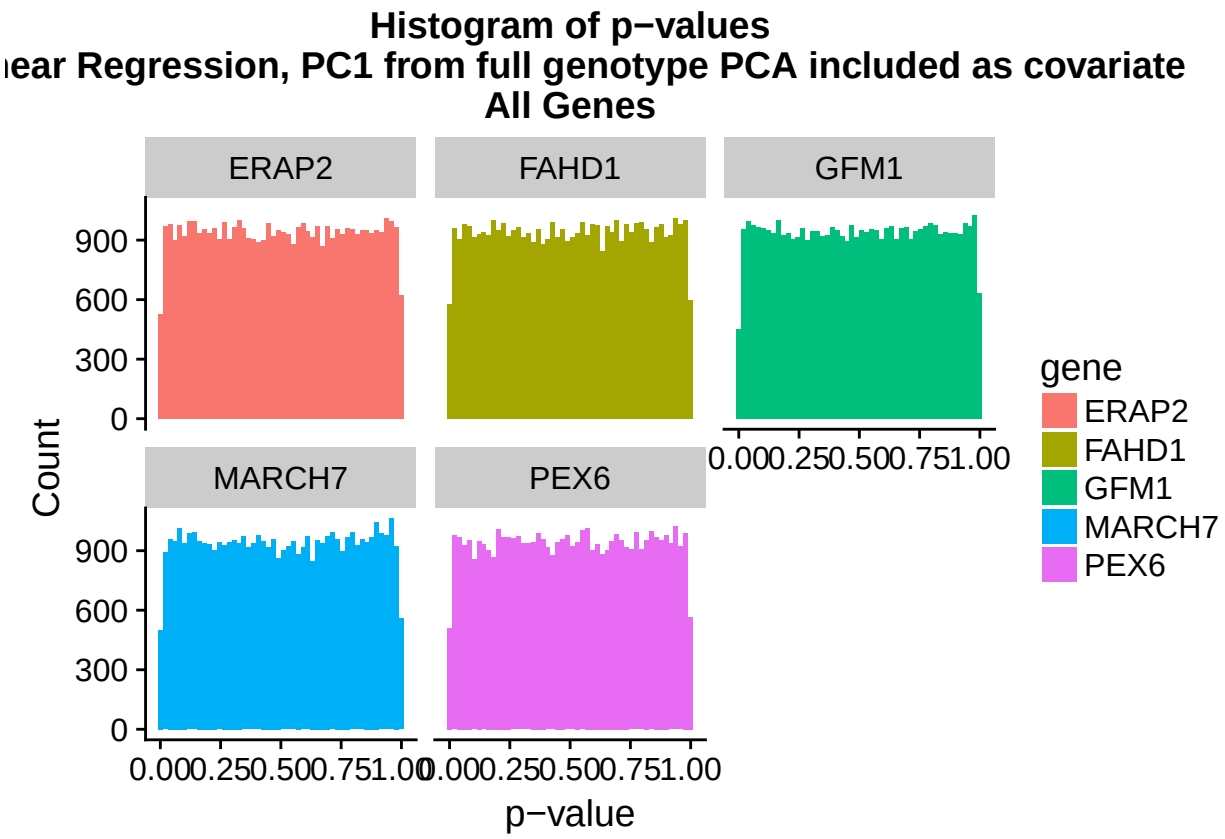
```
          id chromosome position in_gene
26 rs3805941          6 43089842   FALSE
```

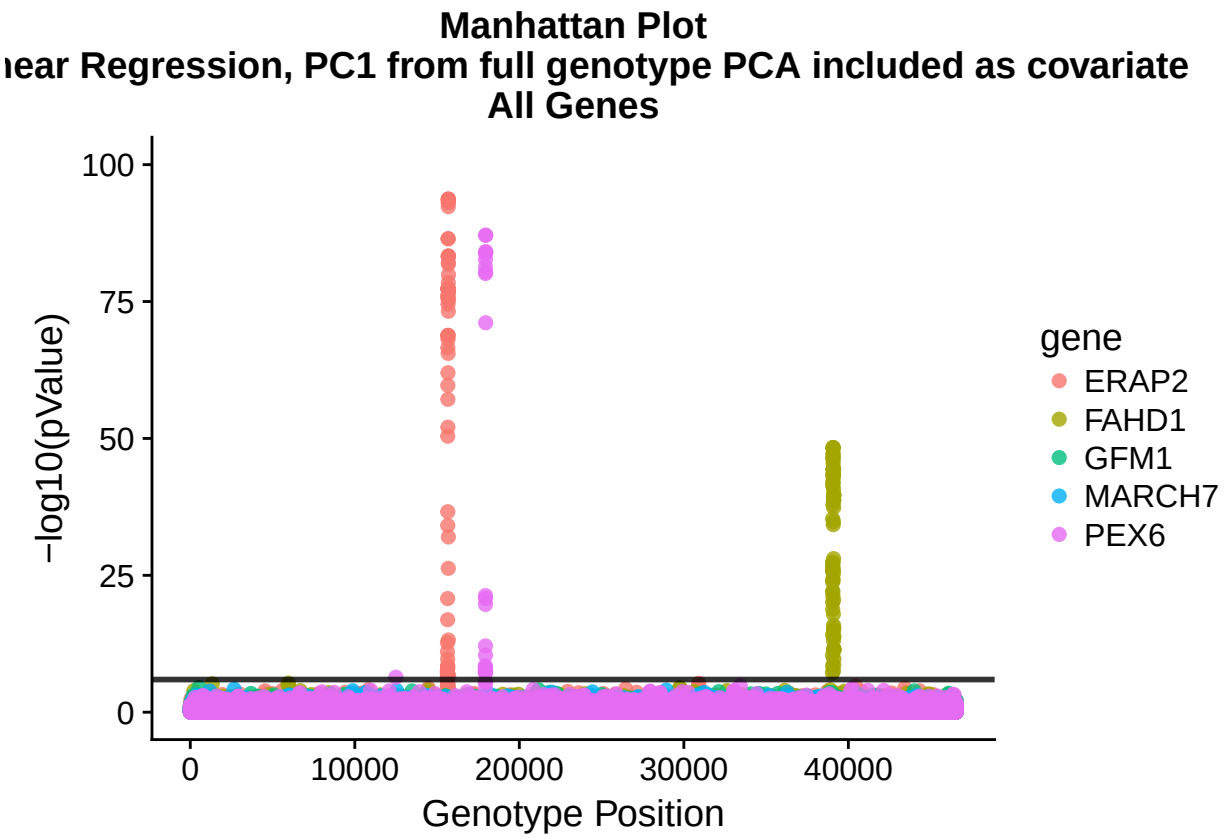


Try PC1 as a covariate:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15627   15646   15666   15664   15682   15701

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15628   15646   15666   15665   15683   15701
```

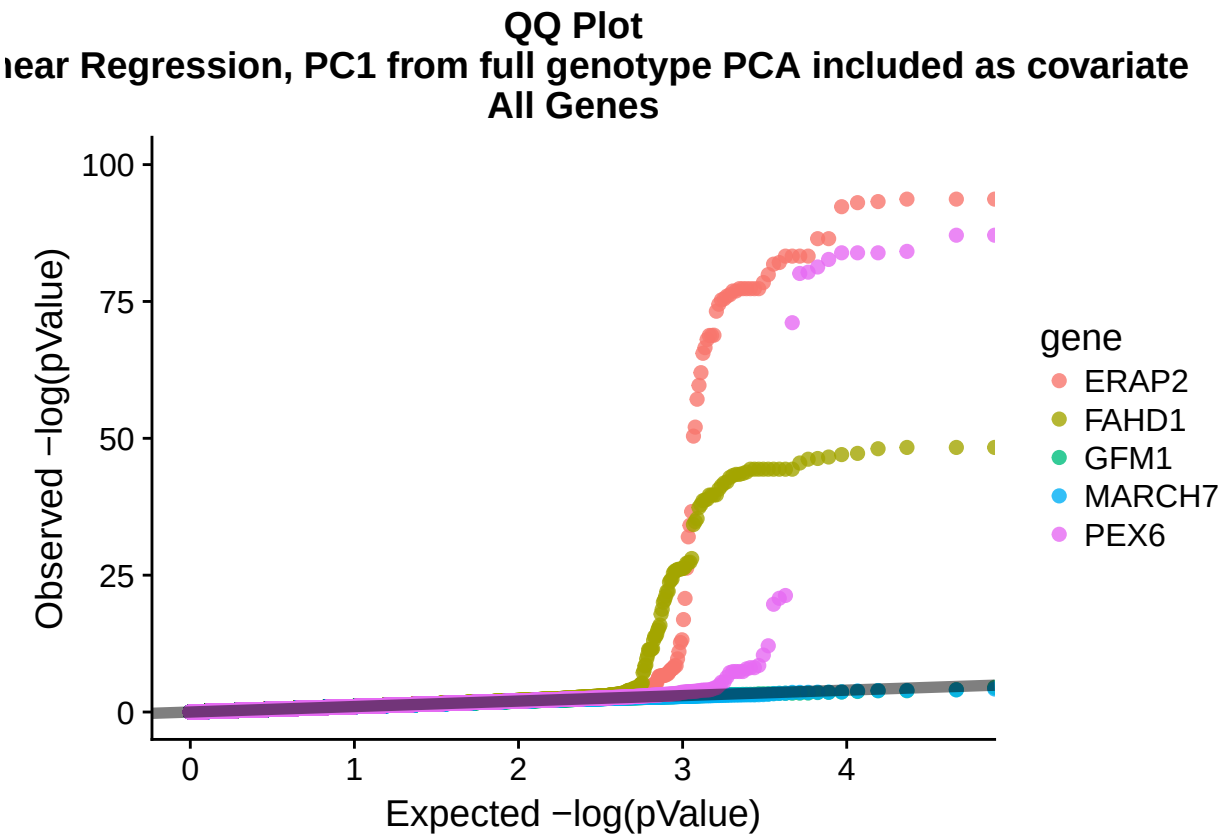

Histogram of p−values
ıear Regression, PC1 from full genotype PCA included as covariate
All Genes
ERAP2
FAHD1
GFM1
MARCH7
PEX6
Count
900
600
300
0
0.00 0.25 0.50 0.75 1.00
p−value
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6




```
# A tibble: 3 x 2
# Groups:   gene [3]
  gene      n
  <chr> <int>
1 ERAP2    67
2 FAHD1    82
3 PEX6     26
```

```
[1] TRUE
```

```
[1] TRUE
```



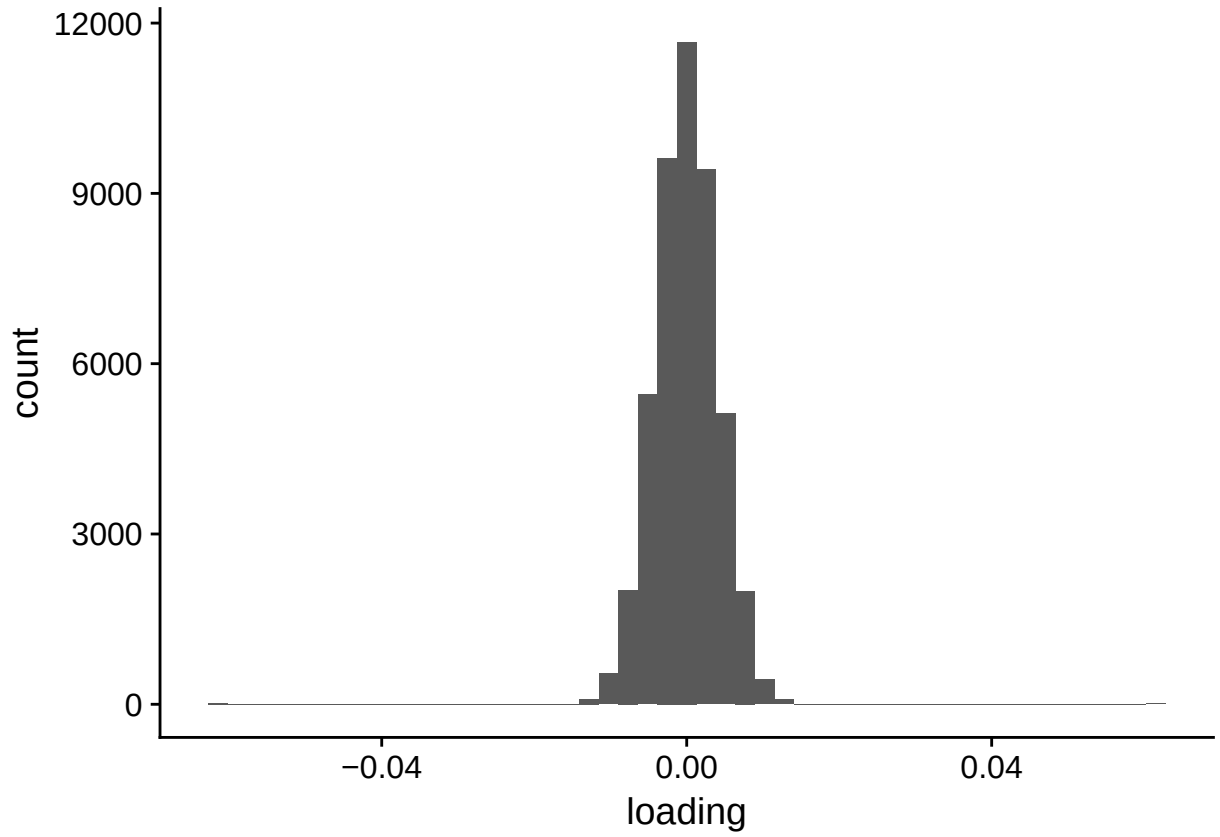
So far, for each analysis, each gene was treated the same. What is the relationship of the expression level of each gene with the first few principal components?

```
# A tibble: 2 x 2
  gene  p.value
  <chr>   <dbl>
1 FAHD1  0.0393
2 PEX6   0.0145
```

```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2  0.0117
```

```
# A tibble: 2 x 2
  gene    p.value
  <chr>     <dbl>
1 ERAP2  0.000147
2 MARCH7 0.0216
```

```
# A tibble: 1 x 2
  gene   p.value
  <chr>    <dbl>
1 FAHD1 1.94e-44
```

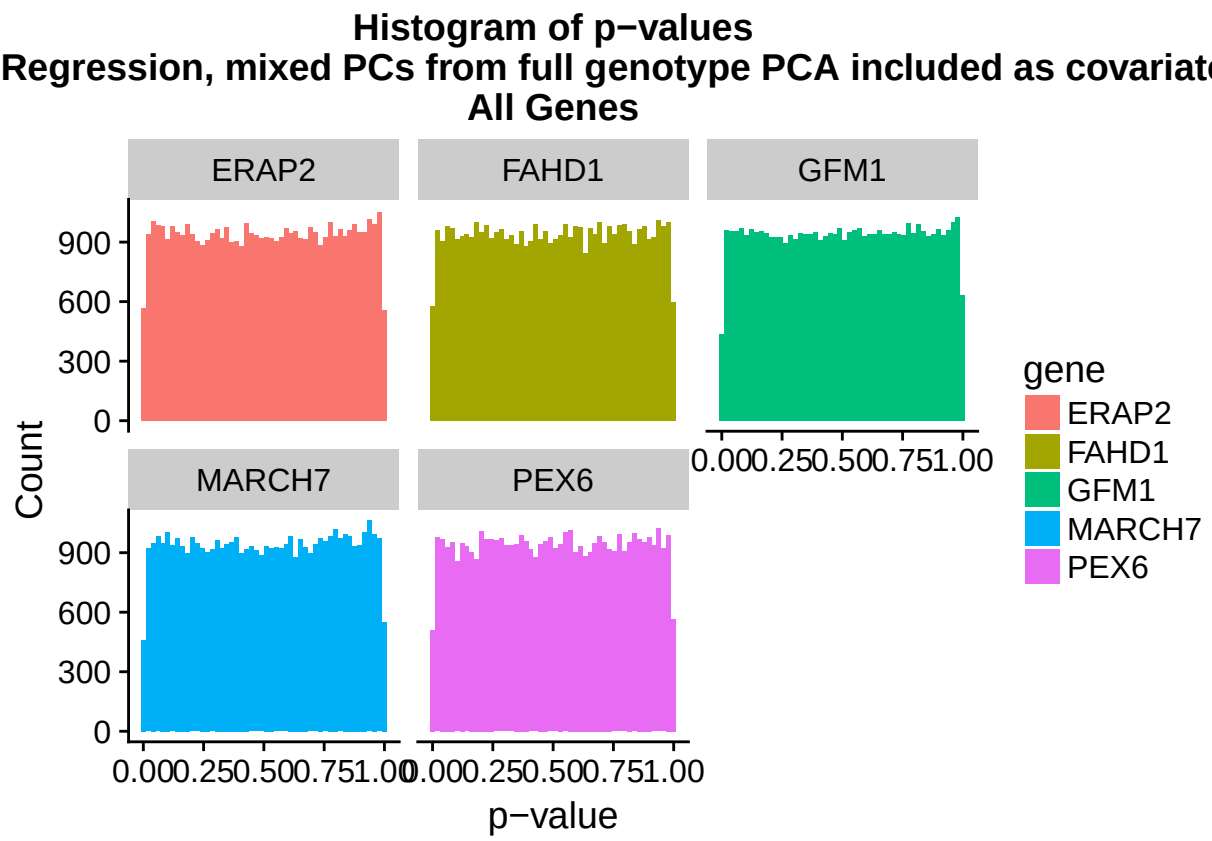


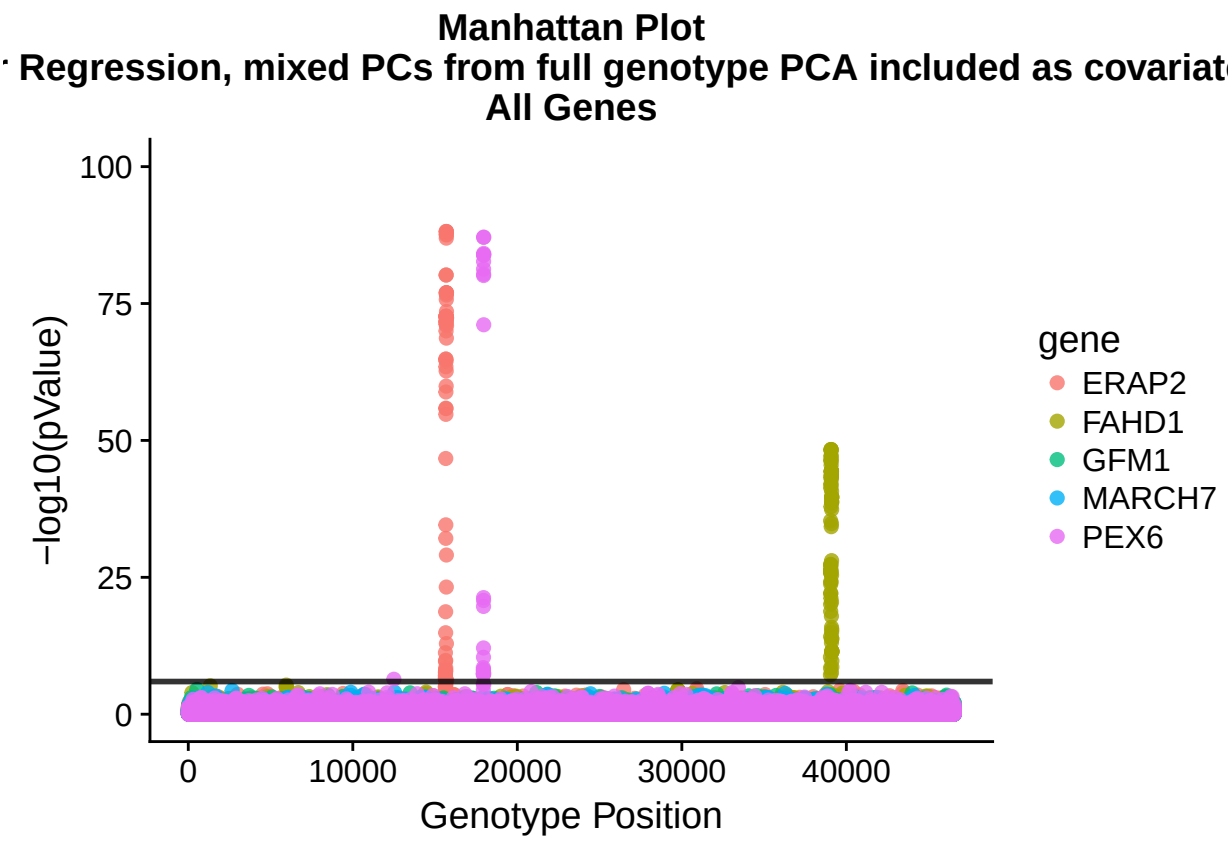

```
[1] 16
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1780619 1818728 1845504 1846450 1868307 1921599
```

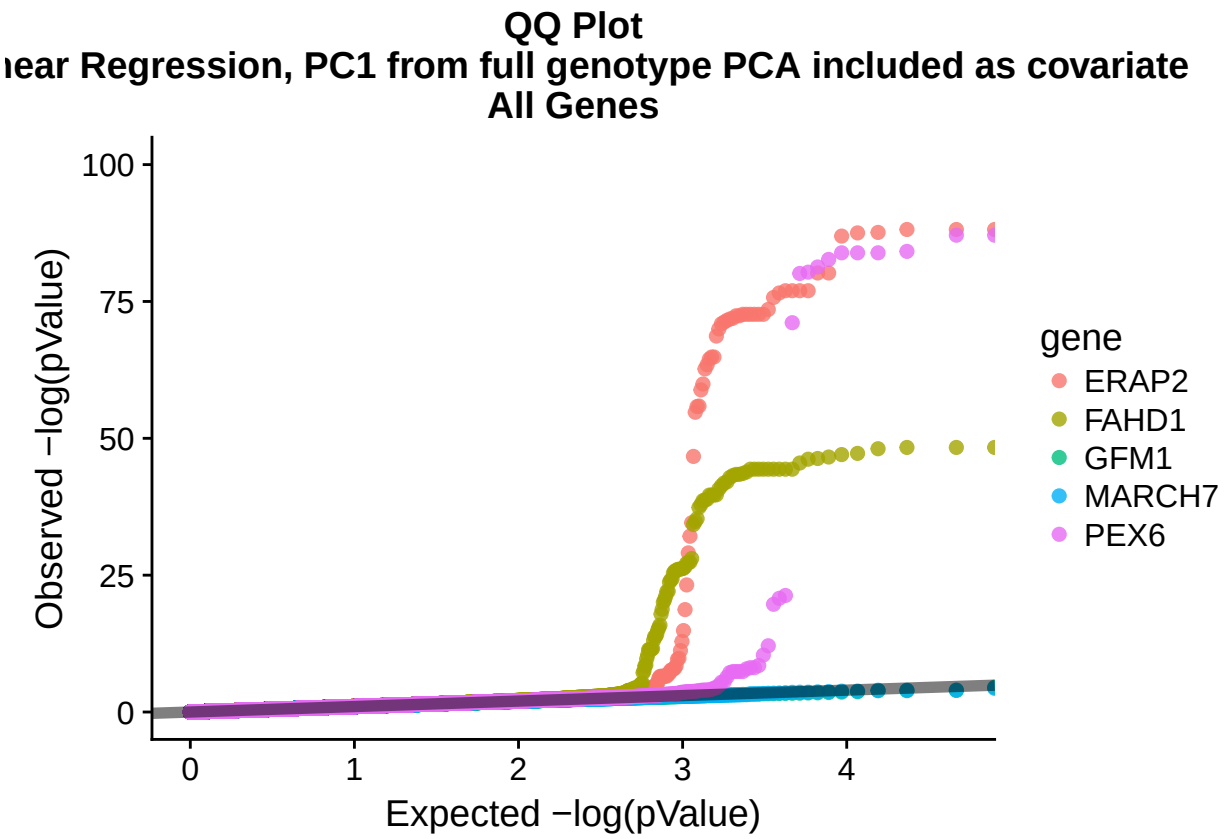
```
[1] 13.88889
```

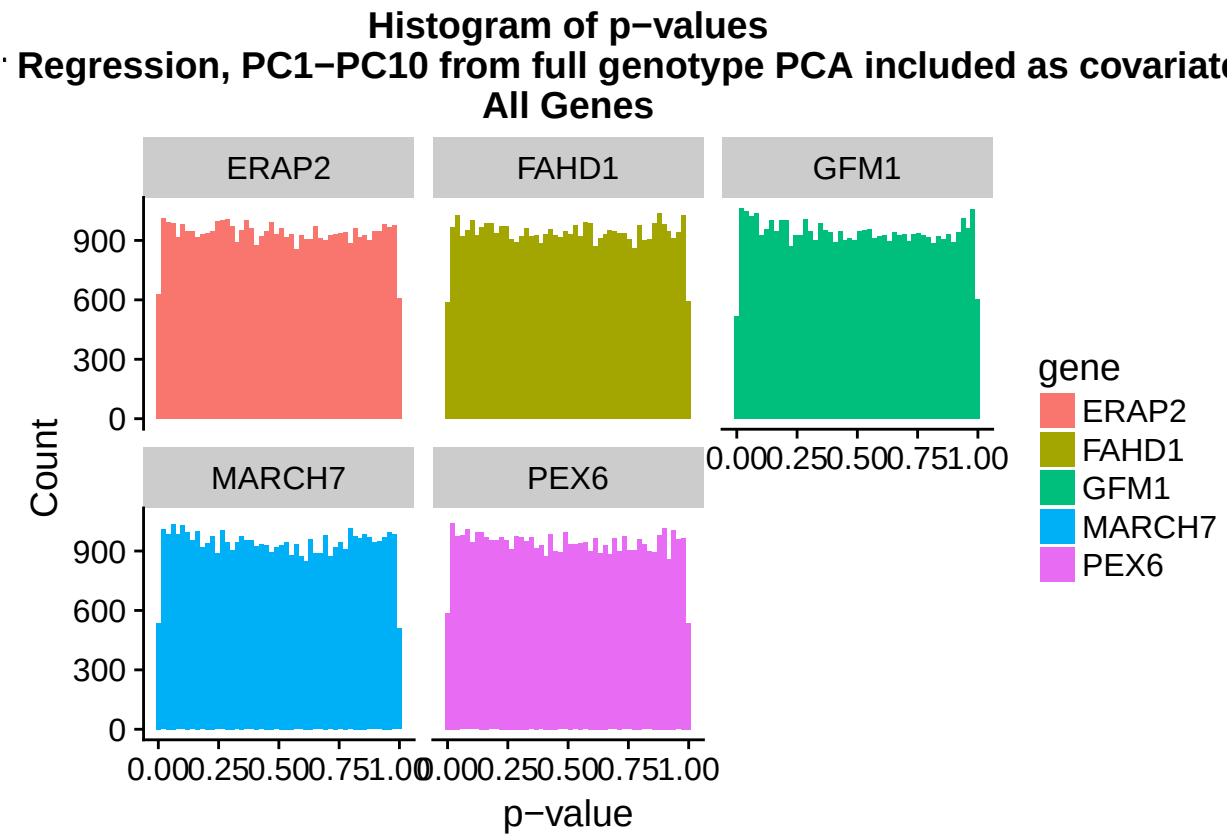
```
[1] 87.80488
```

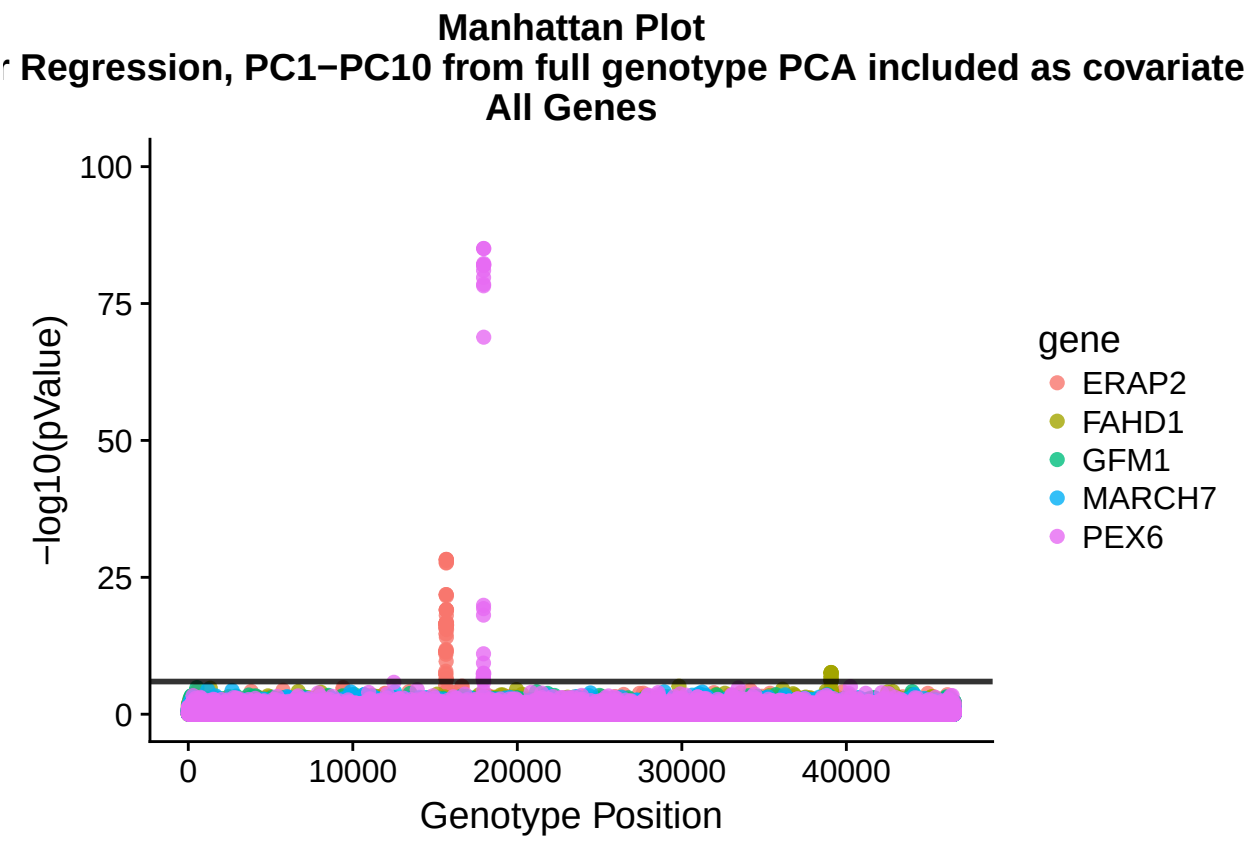

Histogram of p−values
Regression, mixed PCs from full genotype PCA included as covariate
All Genes
ERAP2
FAHD1
GFM1
MARCH7
PEX6
Count
900
600
300
0
0.00 0.25 0.50 0.75 1.00
p−value
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6



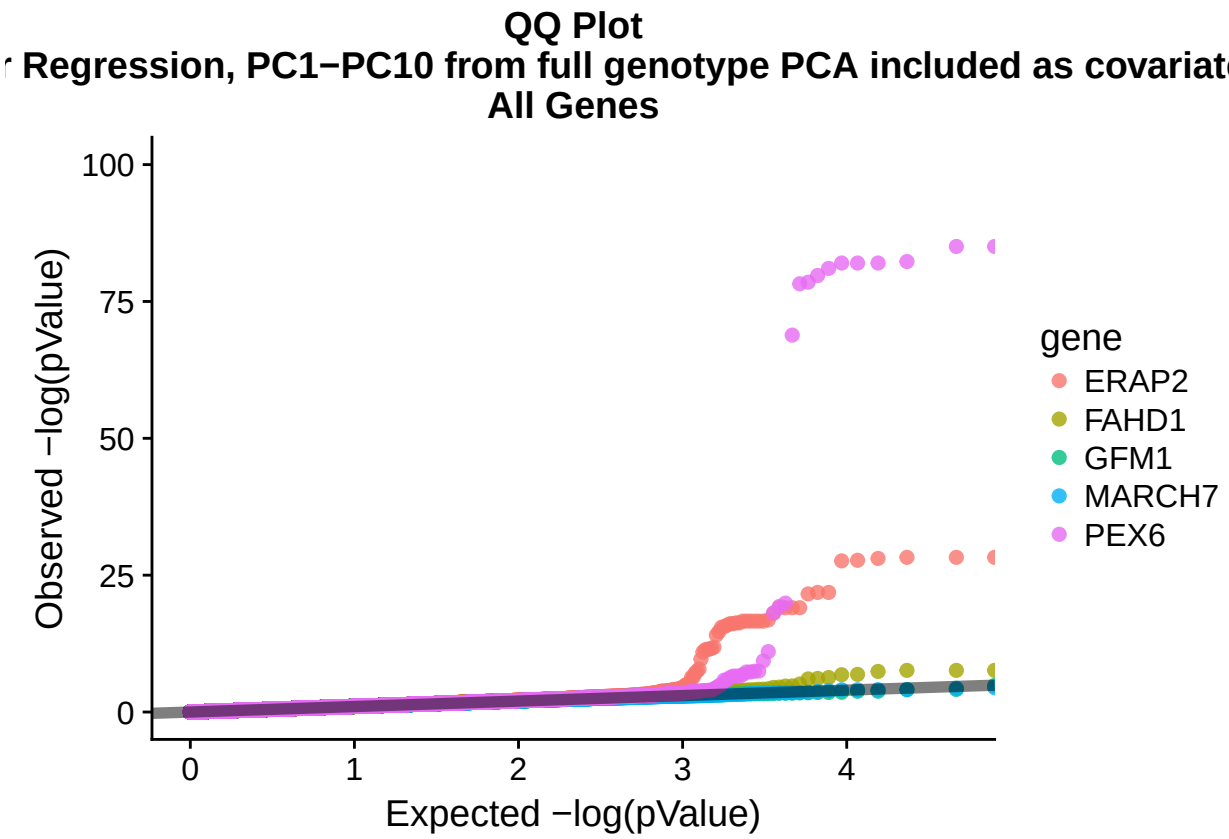

```
# A tibble: 3 x 2
# Groups:   gene [3]
  gene      n
  <chr> <int>
1 ERAP2    66
2 FAHD1    82
3 PEX6     26
```

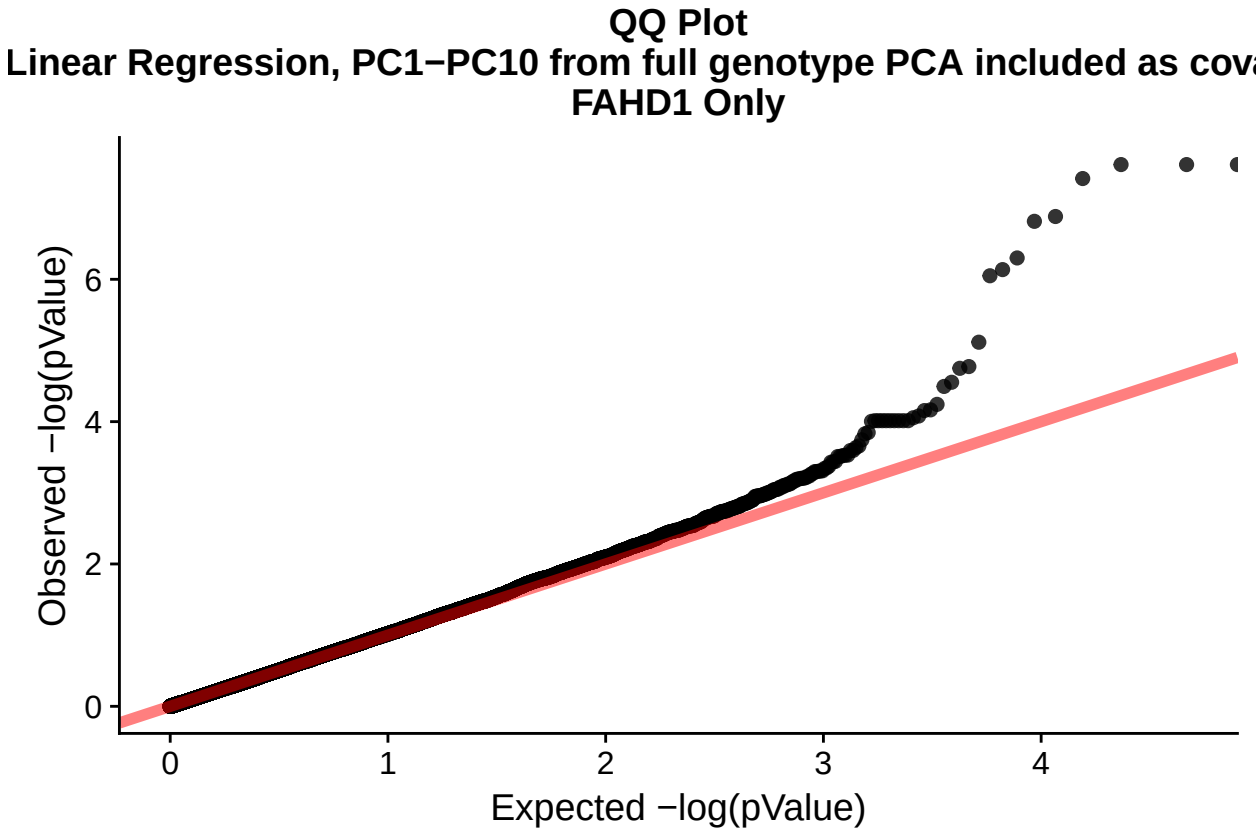

QQ Plot
ıear Regression, PC1 from full genotype PCA included as covariate
All Genes
Observed −log(pValue)
Expected −log(pValue)
100
75
50
25
0
0
1
2
3
4
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

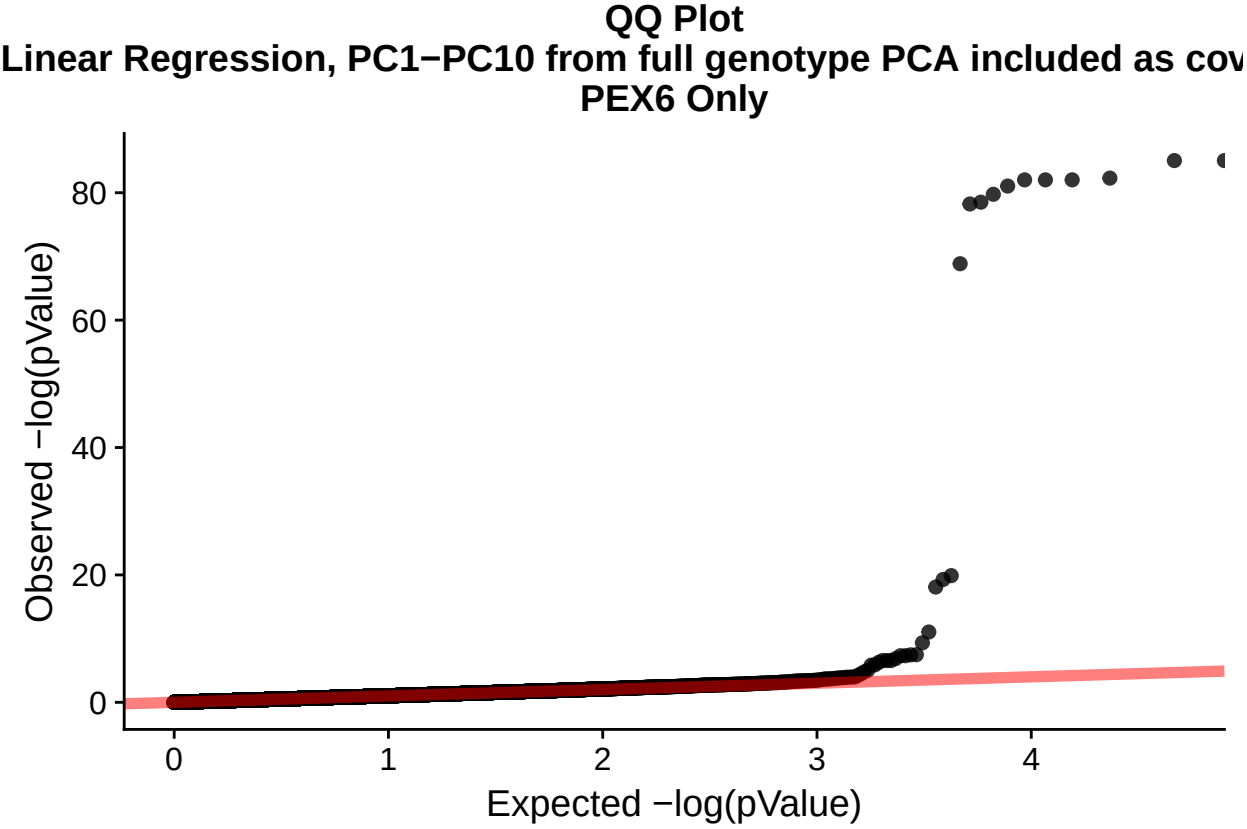

Histogram of p−values
Regression, PC1−PC10 from full genotype PCA included as covariate
All Genes
ERAP2
FAHD1
GFM1
MARCH7
PEX6
Count
900
600
300
0
0.00
0.25
0.50
0.75
1.00
p−value
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

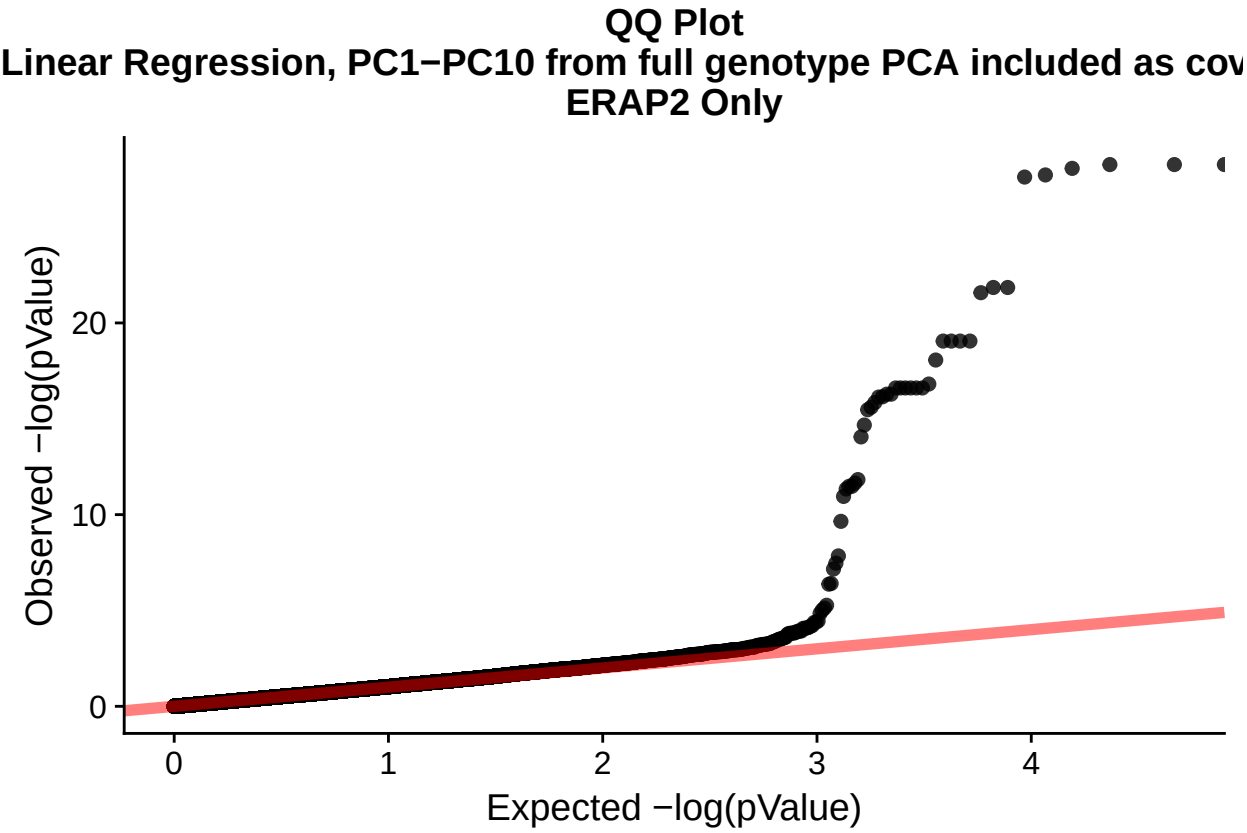



```
# A tibble: 3 x 2
# Groups:   gene [3]
  gene      n
  <chr> <int>
1 ERAP2    42
2 FAHD1     9
3 PEX6     25
```


QQ Plot
r Regression, PC1−PC10 from full genotype PCA included as covariate
All Genes
Observed −log(pValue)
Expected −log(pValue)
100
75
50
25
0
0
1
2
3
4
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6


QQ Plot
Linear Regression, PC1−PC10 from full genotype PCA included as cov
FAHD1 Only
Observed −log(pValue)
Expected −log(pValue)
0
2
4
6
0
1
2
3
4


QQ Plot
Linear Regression, PC1−PC10 from full genotype PCA included as cov
PEX6 Only
Observed −log(pValue)
Expected −log(pValue)
0
20
40
60
80
0
1
2
3
4



Including more principal components and covariates does not seem to resolve the issues with the QQ Plot. Perhaps it is best to try another modeling approach.

```
library(gaston)
geno_bed_matrix <- as(as.matrix(geno_HW_filt), "bed.matrix")
geno_A <- cov(t(geno_HW_filt))
pheno_test <- pheno_names$ERAP2
covariate_lmm <- as.matrix(base::cbind(rep(1, nrow(pheno)), pheno_with_geno_pca_x[, 7:9]))
### ERAP2 ###
lmm_ERAP2 <- association.test(
  x = geno_bed_matrix, Y = pheno_names$ERAP2, X = covariate_lmm,
  method = "lmm", eigenK = eigen(geno_A), test = "lrt", response = "quantitative"
  )
### FAHD1 ###
lmm_FAHD1 <- association.test(
  x = geno_bed_matrix, Y = pheno_names$FAHD1, X = covariate_lmm,
  method = "lmm", eigenK = eigen(geno_A), test = "lrt", response = "quantitative"
  )
### GFM1 ###
lmm_GFM1 <- association.test(
  x = geno_bed_matrix, Y = pheno_names$GFM1, X = covariate_lmm,
  method = "lmm", eigenK = eigen(geno_A), test = "lrt", response = "quantitative"
  )
### MARCH7 ###
lmm_MARCH7 <- association.test(
  x = geno_bed_matrix, Y = pheno_names$MARCH7, X = covariate_lmm,
  method = "lmm", eigenK = eigen(geno_A), test = "lrt", response = "quantitative"
```

```
  )
### PEX6 ###
lmm_PEX6 <- association.test(
  x = geno_bed_matrix, Y = pheno_names$PEX6, X = covariate_lmm,
  method = "lmm", eigenK = eigen(geno_A), test = "lrt", response = "quantitative"
  )
### Combined data frame
pval_lmm_PC1_to_PC3_covar <- as.data.frame(
  cbind(
    x = 1:ncol(geno_HW_filt),
    ERAP2 = lmm_ERAP2$p,
    FAHD1 = lmm_FAHD1$p,
    GFM1 = lmm_GFM1$p,
    MARCH7 = lmm_MARCH7$p,
    PEX6 = lmm_PEX6$p
    )
  )
row.names(pval_lmm_PC1_to_PC3_covar) <- colnames(geno_HW_filt)
### p-value distribution ###
pval_lmm_PC1_to_PC3_covar %>%
  gather("gene", "pval", 2:6) %>%
  ggplot(aes(x = pval, color = gene, fill = gene)) +
  geom_histogram(position = "identity", bins = 50) +
  facet_wrap(~ gene) +
  xlab("p-value") +
  ylab("Count") +
  ggtitle("Histogram of p-values\nLinear Mixed Model, PC1-PC3 as Covariates\nAll Genes")
### Manhattan plot ###
pval_lmm_PC1_to_PC3_covar %>%
  gather("gene", "pval", 2:6) %>%
  ggplot(aes(x = x, y = -log(pval, base = 10), color = gene)) +
  geom_point(alpha = 0.8, size = 2) +
  ggtitle("Manhattan Plot\nLinear Mixed Model, PC1-PC3 as Covariates\nAll Genes") +
  xlab("Genotype Position") +
  ylab("-log10(pValue)") +
  geom_hline(yintercept = -log(0.05 / ncol(geno_HW_filt), base = 10), size = 1, alpha = 0.8) +
  ylim(0, 100)

### number of statistically significant positions after Bonferroni correction ###
pval_lmm_PC1_to_PC3_covar %>%
  rownames_to_column("snp") %>%
  dplyr::select(-x) %>%
  gather("gene", "pval", 2:6) %>%
  filter(pval < (0.05 / ncol(geno_HW_filt))) %>%
  group_by(gene) %>%
  count()

### QQ Plot to assess data quality ###
pval_lmm_PC1_to_PC3_covar_qqplot <- data.frame(
  cbind(
    expected = sort(-log10(seq(from = 0,to = 1, length.out = length(pval_lmm_PC1_to_PC3_covar$x)))),
    ERAP2 = sort(-log10(pval_lmm_PC1_to_PC3_covar$ERAP2)),
    FAHD1 = sort(-log10(pval_lmm_PC1_to_PC3_covar$FAHD1)),
```

```
    GFM1 = sort(-log10(pval_lmm_PC1_to_PC3_covar$GFM1)),
    MARCH7 = sort(-log10(pval_lmm_PC1_to_PC3_covar$MARCH7)),
    PEX6 = sort(-log10(pval_lmm_PC1_to_PC3_covar$PEX6))
    )
  )
pval_lmm_PC1_to_PC3_covar_qqplot %>%
  gather("gene", "neglog10_pval", 2:6) %>%
  ggplot(aes(x = expected, y = neglog10_pval, color = gene)) +
  geom_point(alpha = 0.8, size = 2) +
  geom_abline(intercept = 0, slope = 1, col = "black", alpha = 0.5, size = 2) +
  xlab("Expected -log(pValue)") +
  ylab("Observed -log(pValue)") +
  ggtitle("QQ Plot\nLinear Mixed Model, PC1-PC3 as Covariates\nAll Genes") +
  ylim(0, 100)
pval_lmm_PC1_to_PC3_covar_qqplot %>%
  ggplot(aes(x = expected, y = FAHD1)) +
  geom_point(alpha = 0.8, size = 2) +
  geom_abline(intercept = 0, slope = 1, col = "red", alpha = 0.5, size = 2) +
  xlab("Expected -log(pValue)") +
  ylab("Observed -log(pValue)") +
  ggtitle("QQ Plot\nLinear Mixed Model, PC1-PC3 as Covariates\nFAHD1 Only")
pval_lmm_PC1_to_PC3_covar_qqplot %>%
  ggplot(aes(x = expected, y = PEX6)) +
  geom_point(alpha = 0.8, size = 2) +
  geom_abline(intercept = 0, slope = 1, col = "red", alpha = 0.5, size = 2) +
  xlab("Expected -log(pValue)") +
  ylab("Observed -log(pValue)") +
  ggtitle("QQ Plot\nLinear Mixed Model, PC1-PC3 as Covariates\nPEX6 Only")
pval_lmm_PC1_to_PC3_covar_qqplot %>%
  ggplot(aes(x = expected, y = ERAP2)) +
  geom_point(alpha = 0.8, size = 2) +
  geom_abline(intercept = 0, slope = 1, col = "red", alpha = 0.5, size = 2) +
  xlab("Expected -log(pValue)") +
  ylab("Observed -log(pValue)") +
  ggtitle("QQ Plot\nLinear Mixed Model, PC1-PC3 as Covariates\nERAP2 Only")
```