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)
```

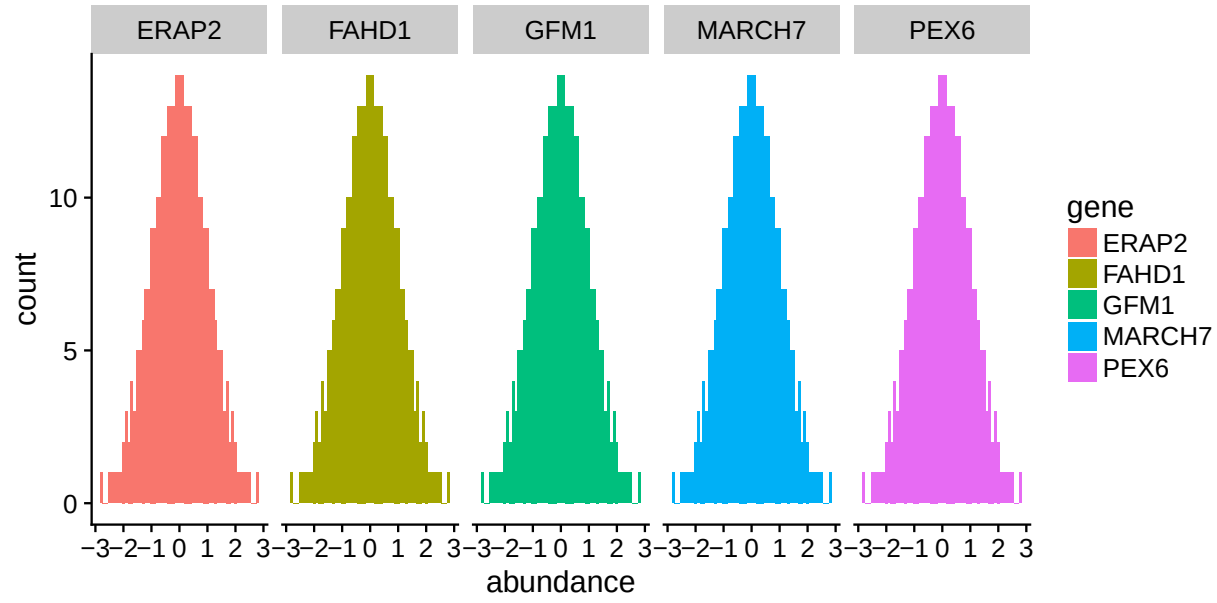
```
[1] FALSE
```

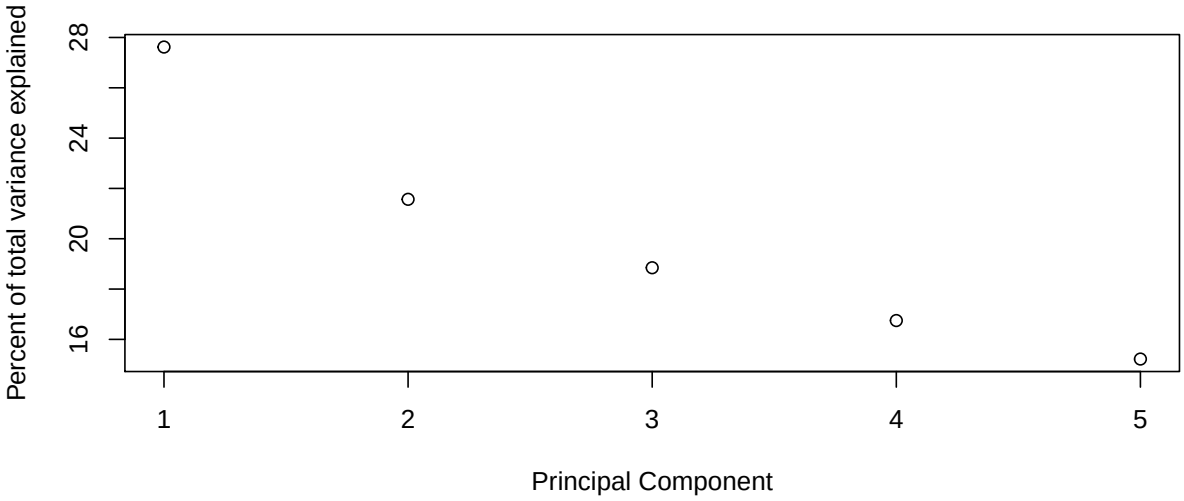
```
# Are there any genotypes without associated phenotypes, or vice versa?
all.equal(rownames(geno), rownames(pheno))
```

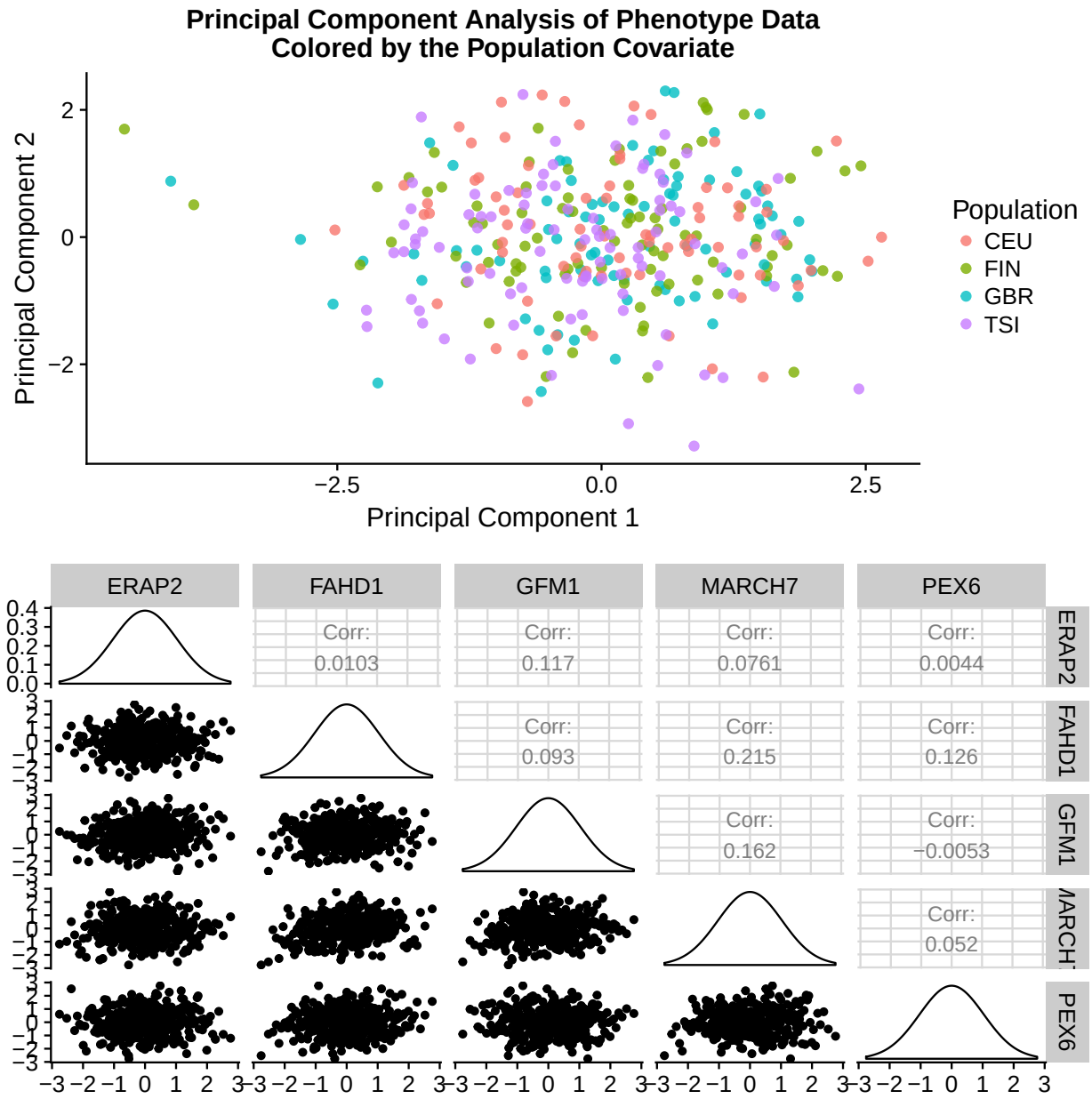
```
[1] TRUE
```

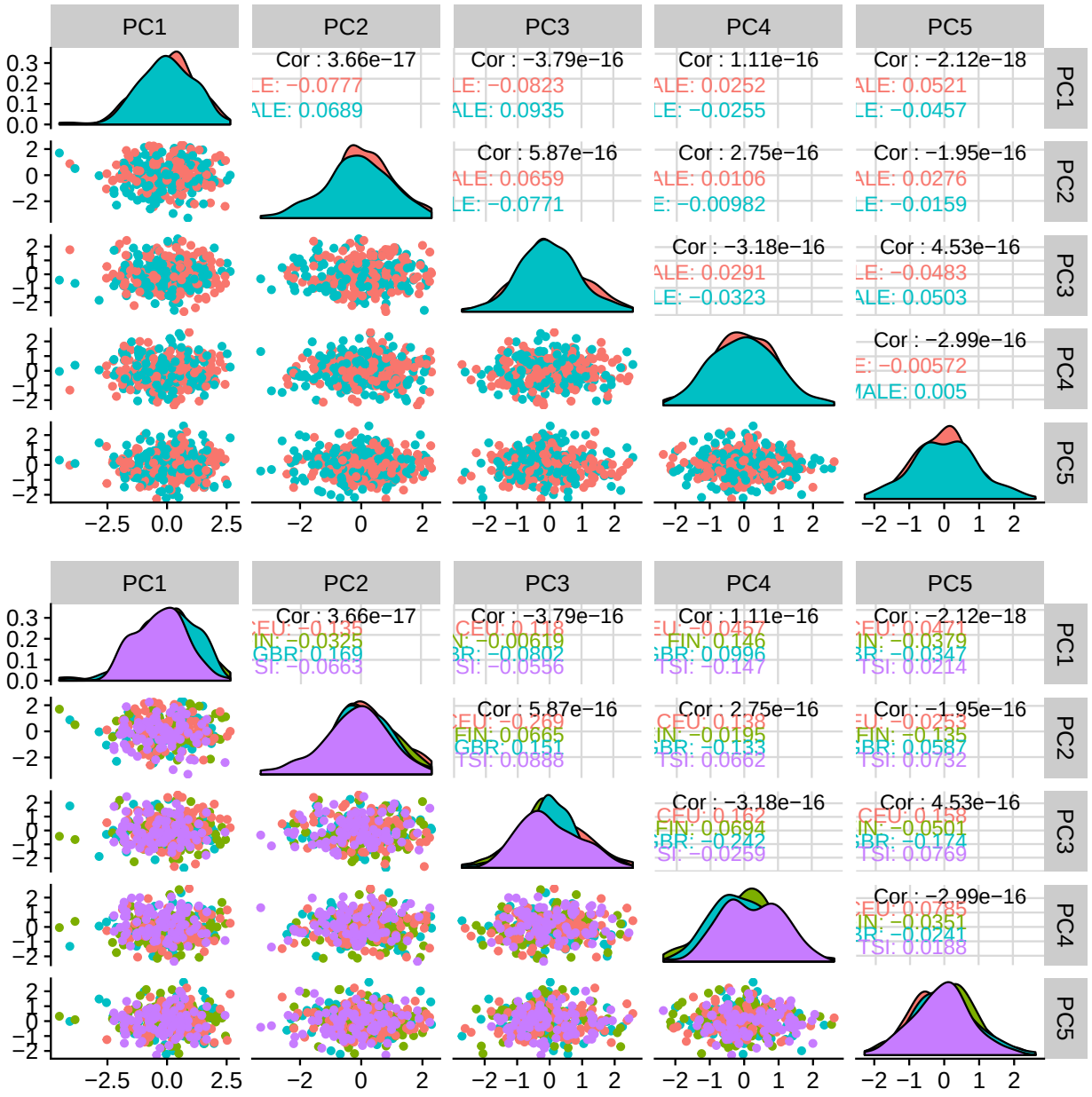
Overall, there were no missing values across all of the data and information provided. There are approximately equal numbers of males and females in the study, and approximately equal numbers of individuals in each of the population groups. Most importantly, none of the covariate groups had a small n. The genotype data were coded as expected, with no unusual values. No genotypes had a minor allele frequency below 5%, which indicates that no SNPs need to be removed from analysis based on that criteria. Lastly, all of the individuals in the genotype dataset are found in the phenotype dataset and no samples need to be filtered out on that criteria.

The phenotype data was visualized and a PCA analysis was performed on the phenotype data in order to determine if there are any outliers in the phenotype data or unusual structure.


ERAP2
FAHD1
GFM1
MARCH7
PEX6
count
10
5
0
−3 −2 −1 0 1 2 3
abundance
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

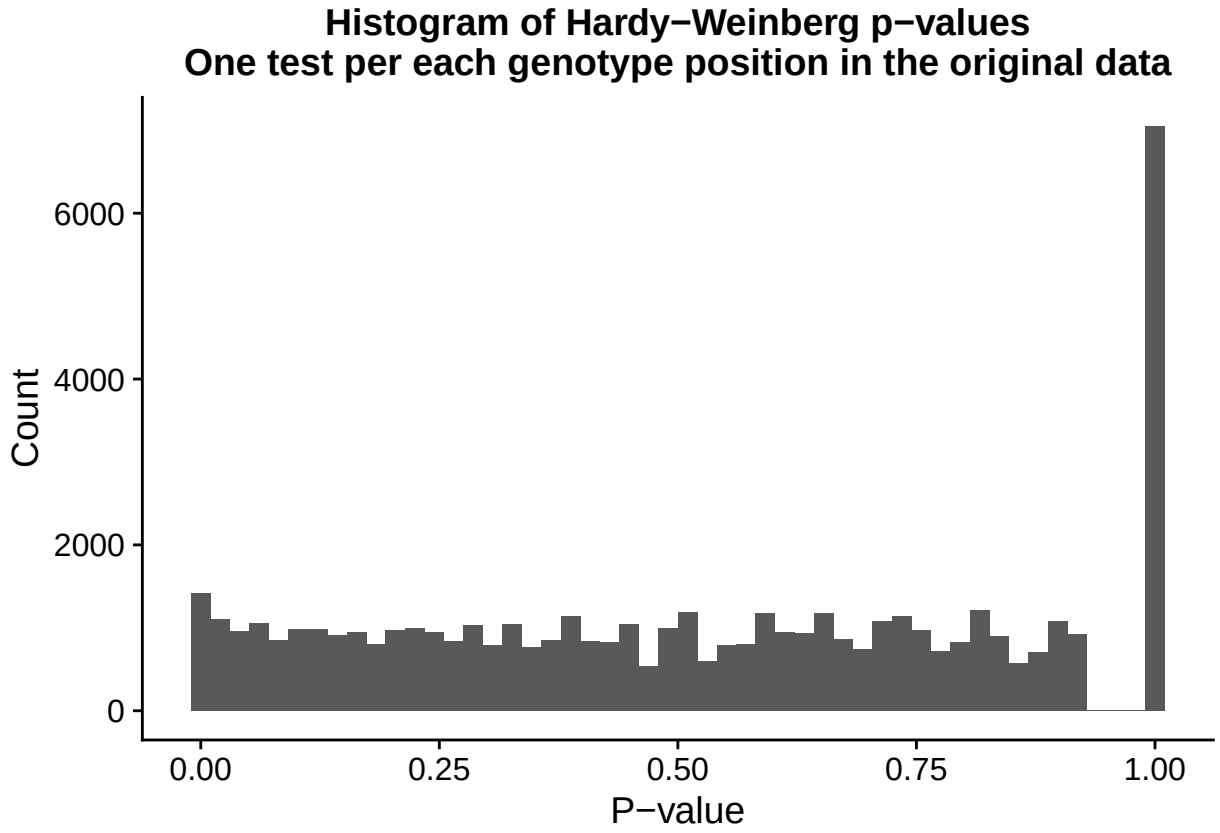

Percent of total variance explained
28
24
20
16
1
2
3
4
5
Principal Component

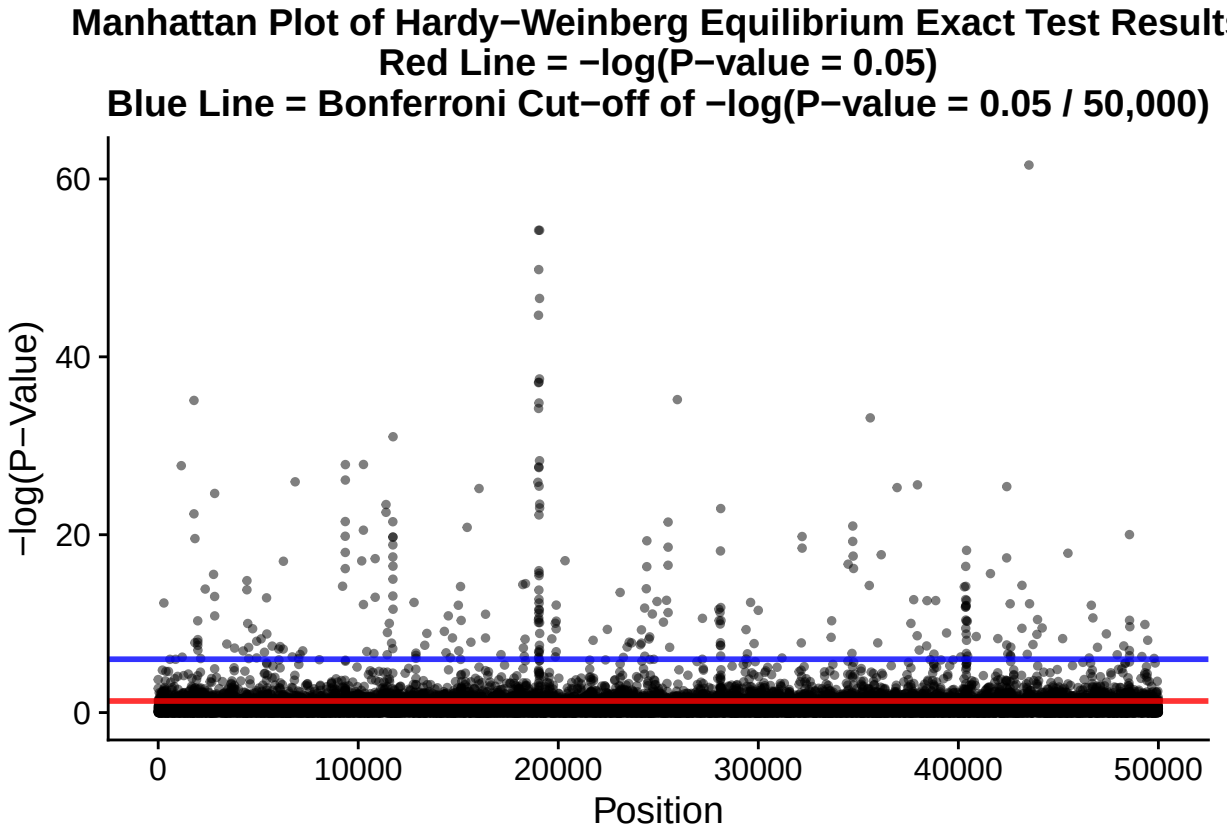

Principal Component Analysis of Phenotype Data
Colored by the Population Covariate
Principal Component 2
Principal Component 1
Population
CEU
FIN
GBR
TSI
ERAP2
FAHD1
GFM1
MARCH7
PEX6
Corr:
0.0103
Corr:
0.117
Corr:
0.0761
Corr:
0.0044
Corr:
0.093
Corr:
0.215
Corr:
0.126
Corr:
0.162
Corr:
−0.0053
Corr:
0.052



The abundances of each gene were found to be normally distributed and the PCA analysis did not reveal any unusual patterns. Further more, 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.

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, it was used as implemented in the HardyWeinberg R package: HardyWeinberg Package on CRAN. This test determines if the proportion of alleles at a particular genotype position is 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.

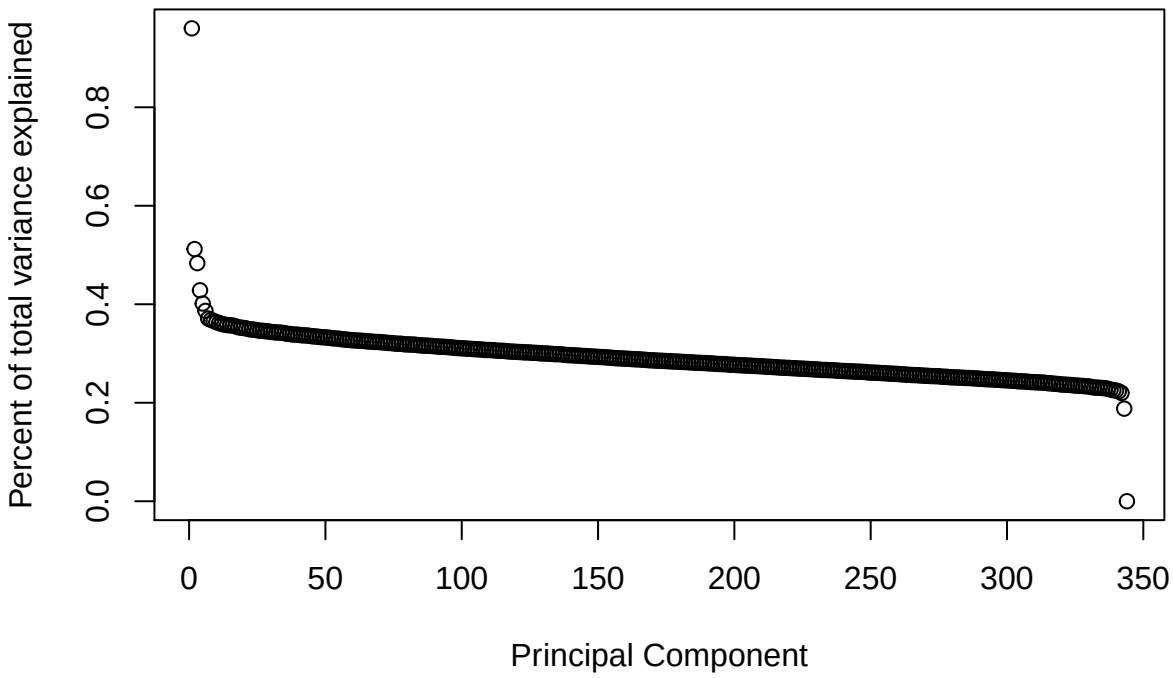

Histogram of Hardy−Weinberg p−values
One test per each genotype position in the original data
Count
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
P−value

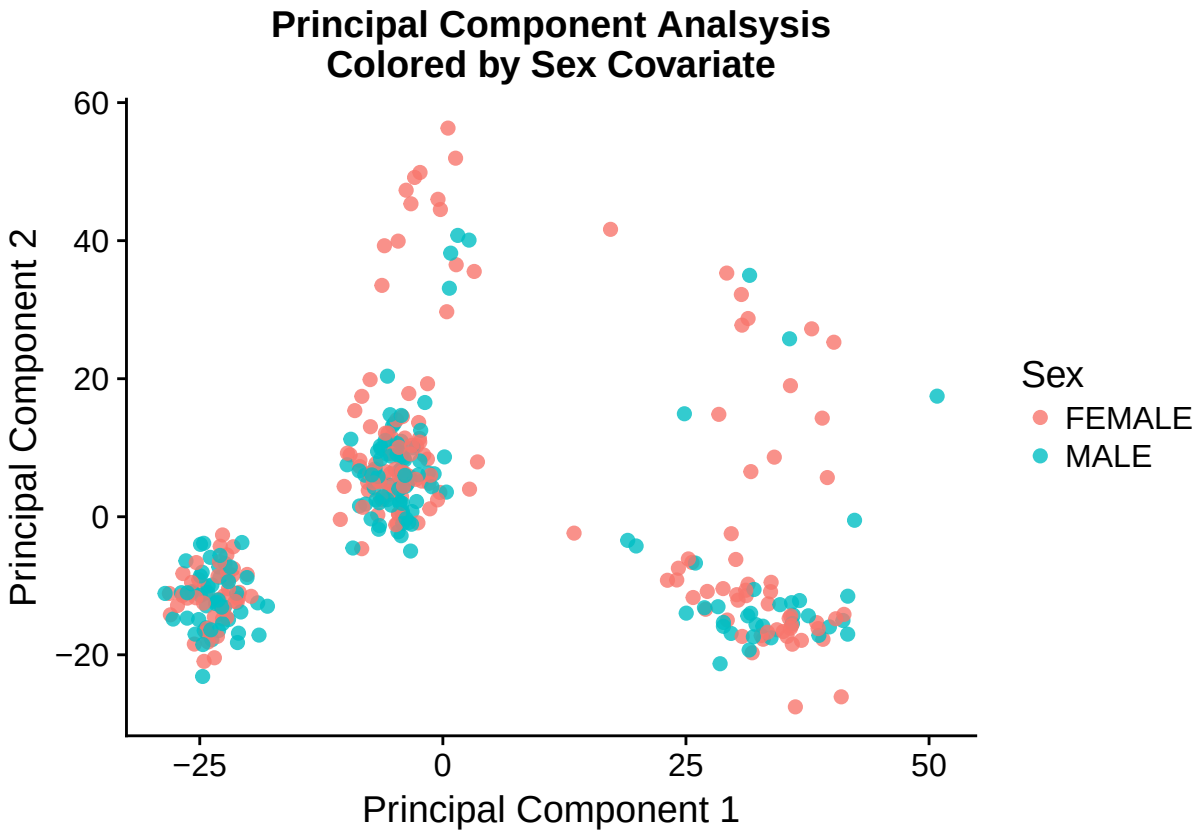


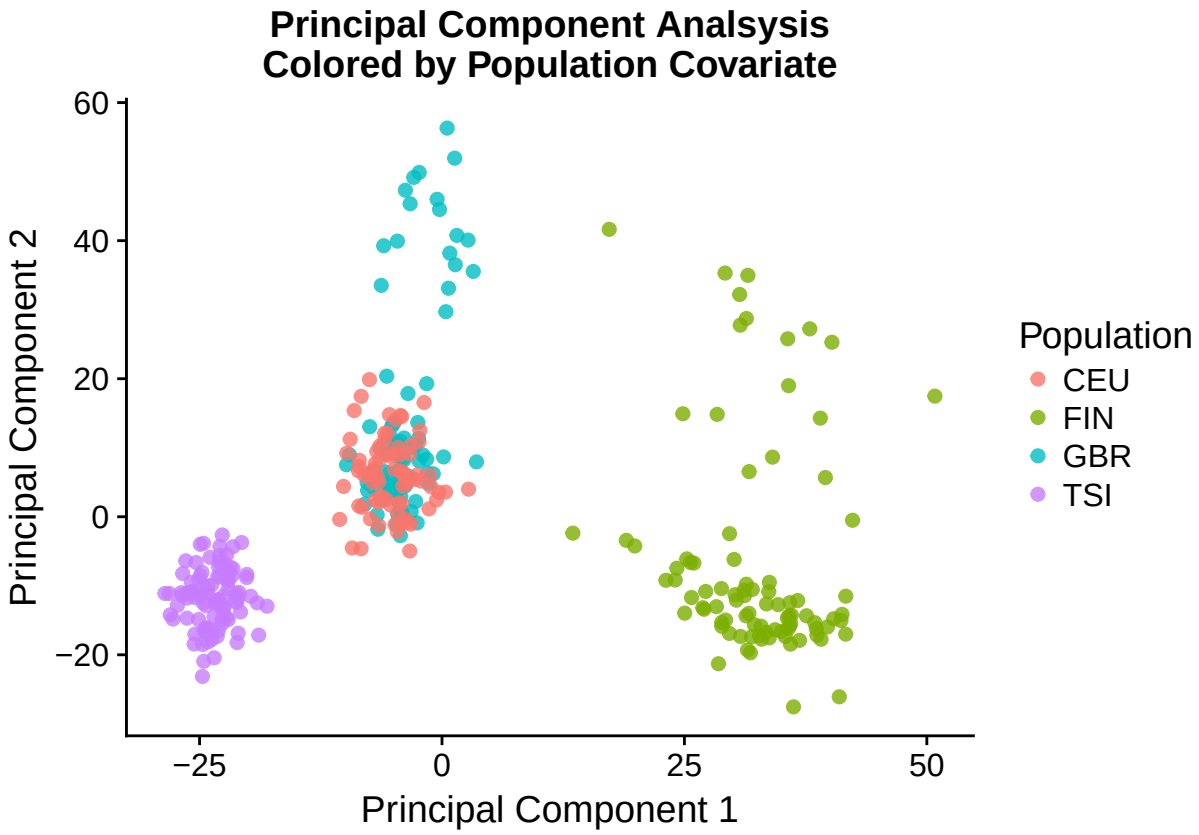

```
[1] 277
```

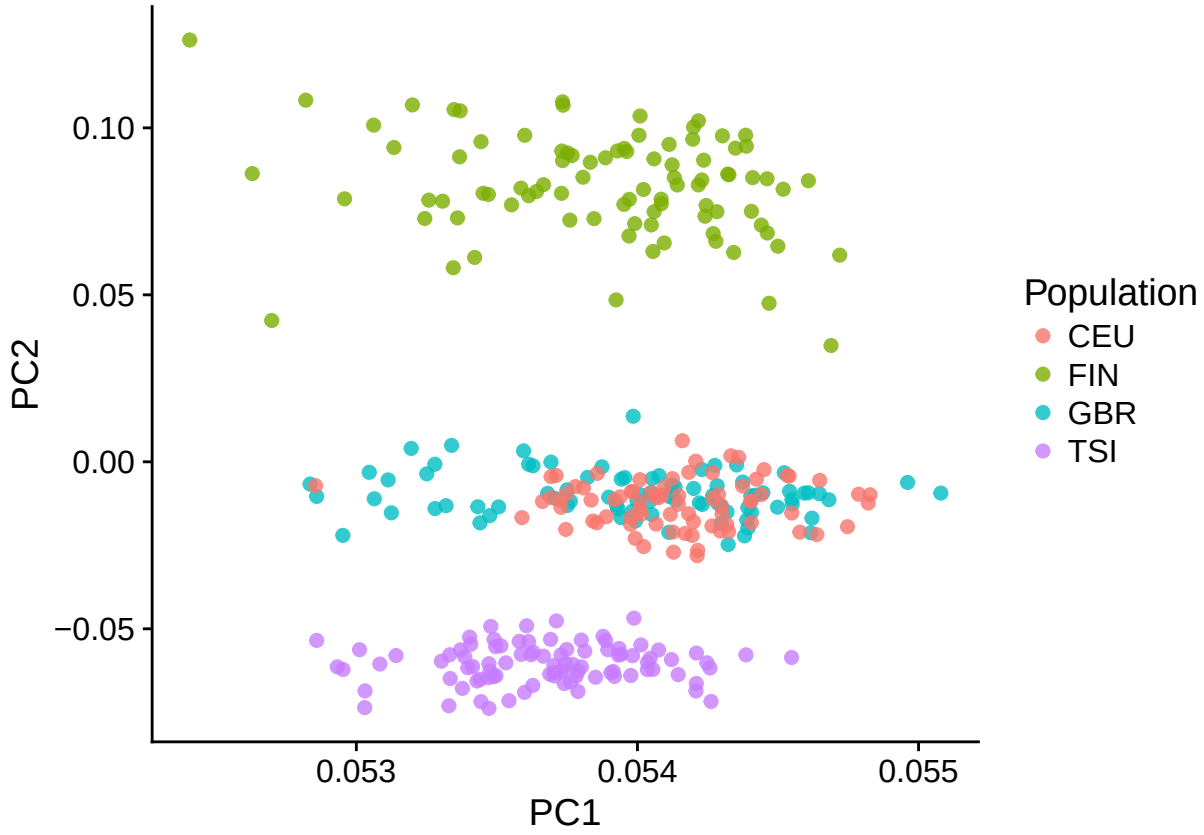
```
[1] 3436
```

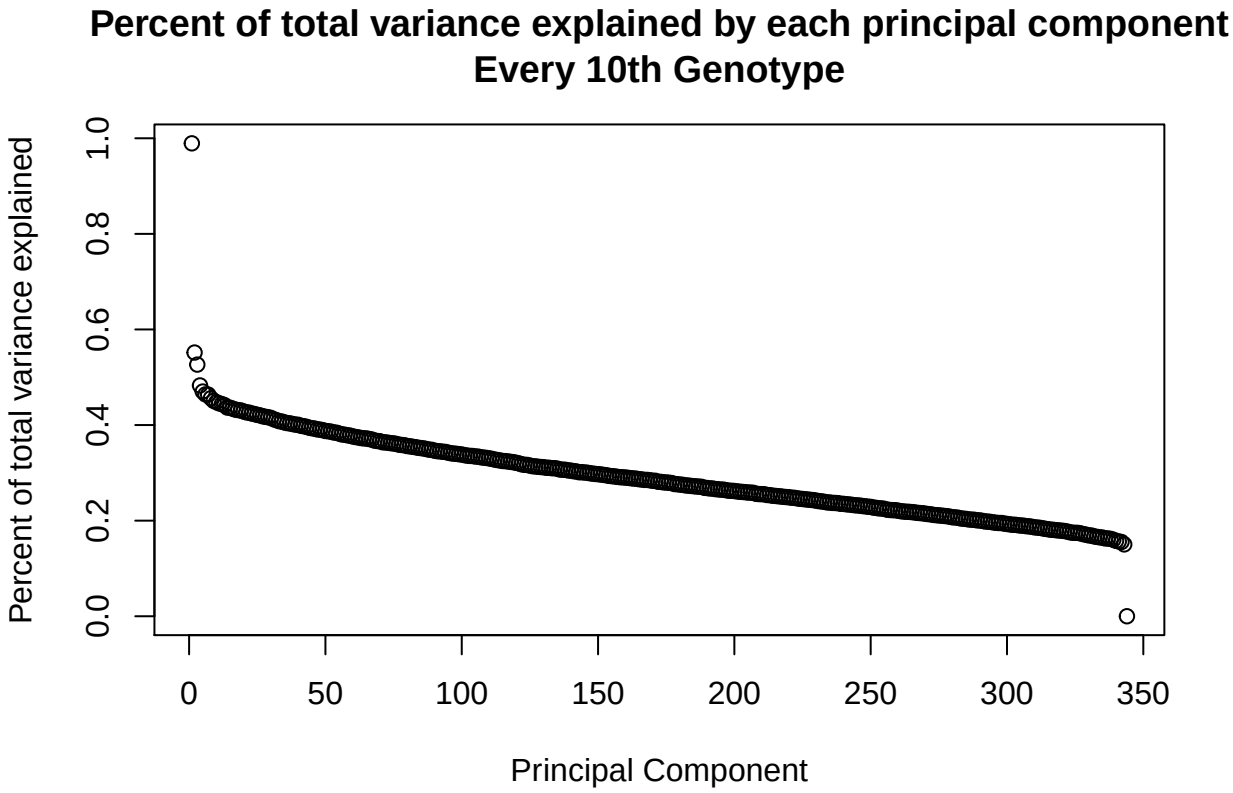
To be conservative, all genotypes that were found to fail the Hardy-Equilibrium exact test with a p-value $<$ 0.05 were removed from consideration.

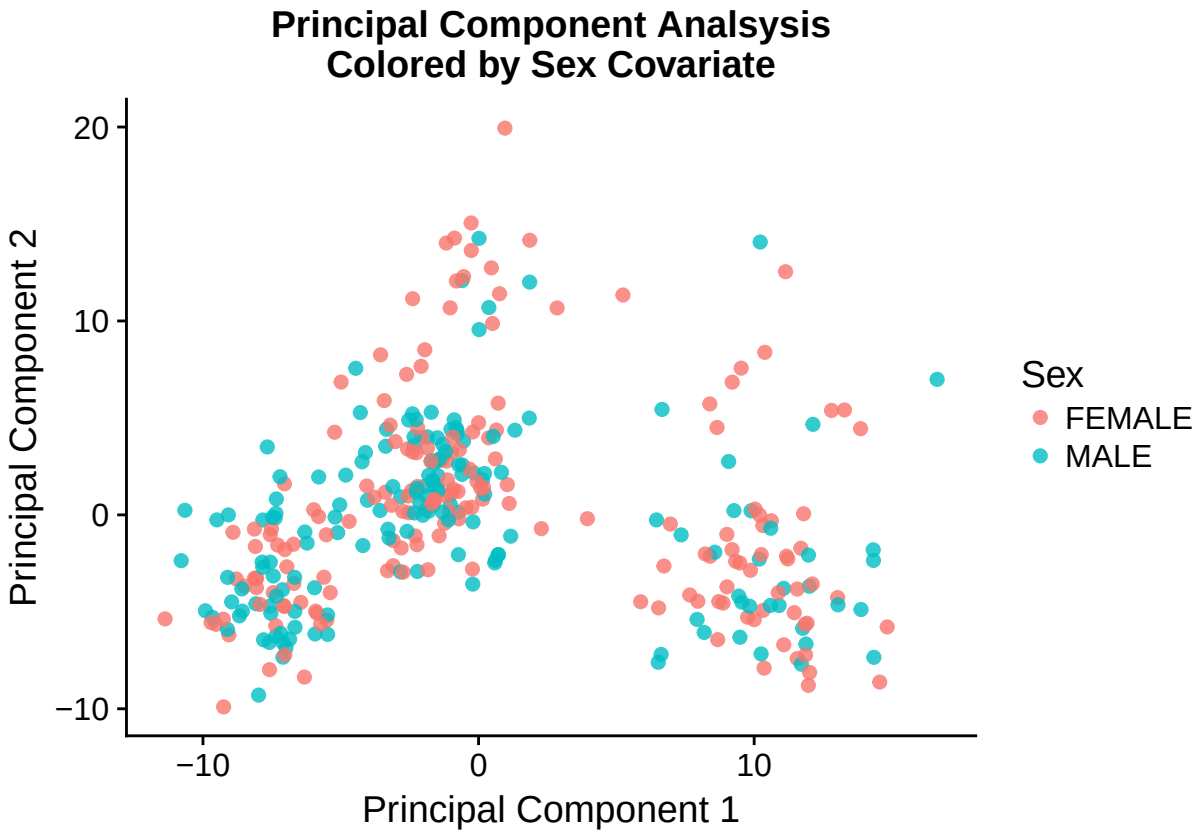

Percent of total variance explained by each principal component
Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
0
50
100
150
200
250
300
350
Principal Component

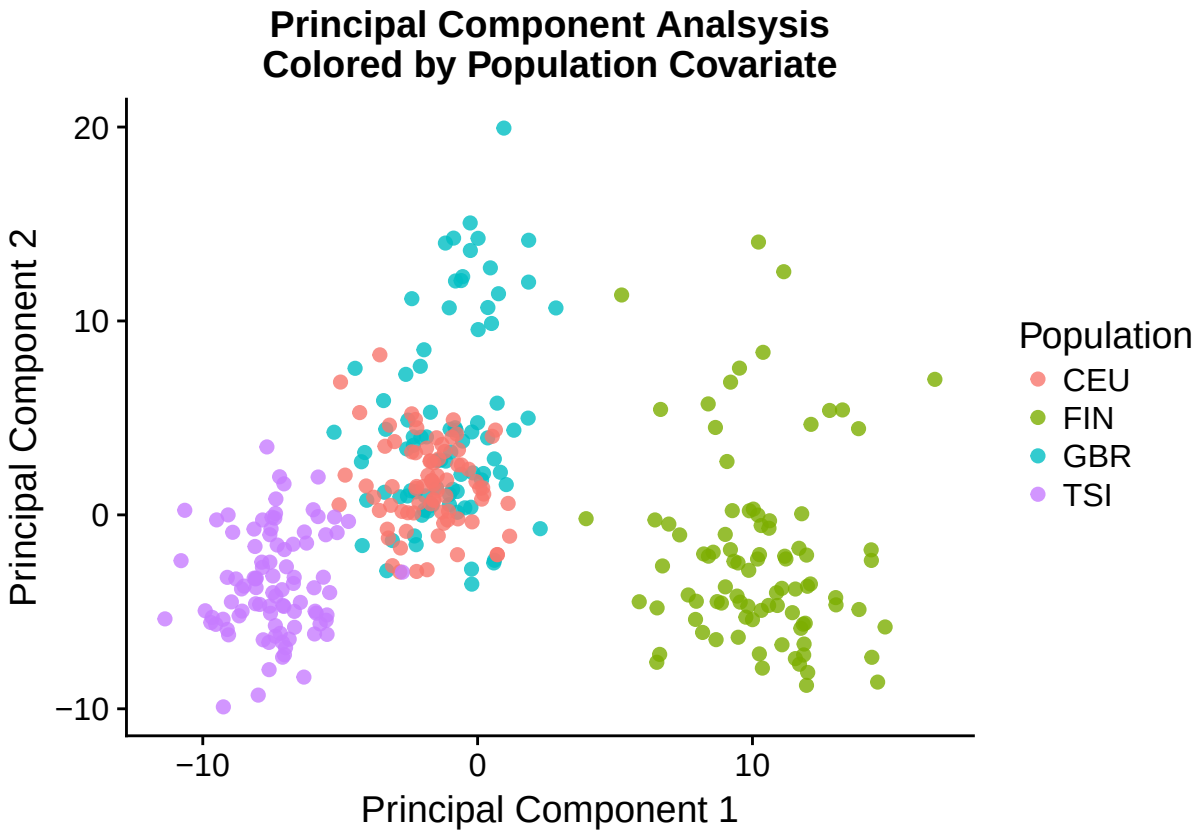

Principal Component Analsysis
Colored by Sex Covariate
Principal Component 2
60
40
20
0
−20
−25
0
25
50
Principal Component 1
Sex
FEMALE
MALE

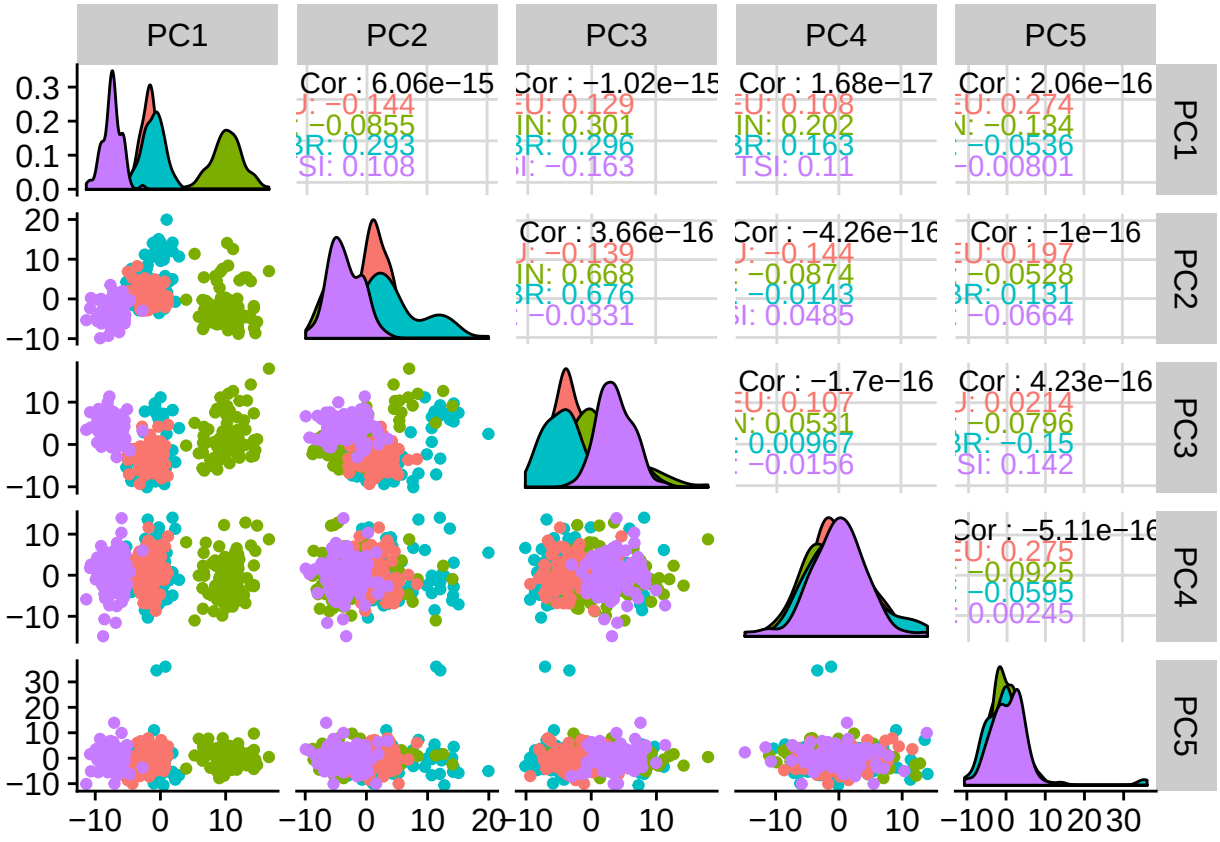

Principal Component Analsysis
Colored by Population Covariate
Principal Component 2
60
40
20
0
−20
−25
0
25
50
Principal Component 1
Population
CEU
FIN
GBR
TSI

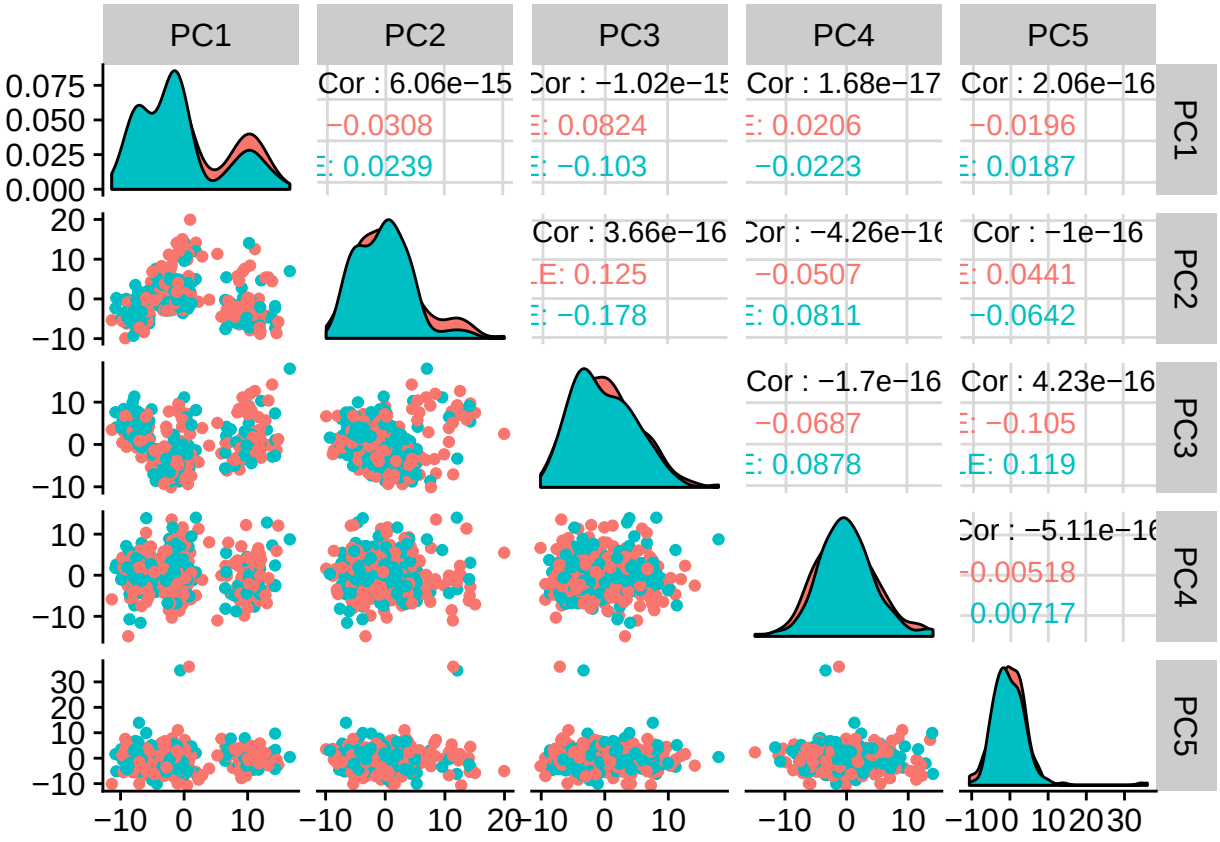

0.10
0.05
0.00
−0.05
PC2
0.053
0.054
0.055
PC1
Population
CEU
FIN
GBR
TSI

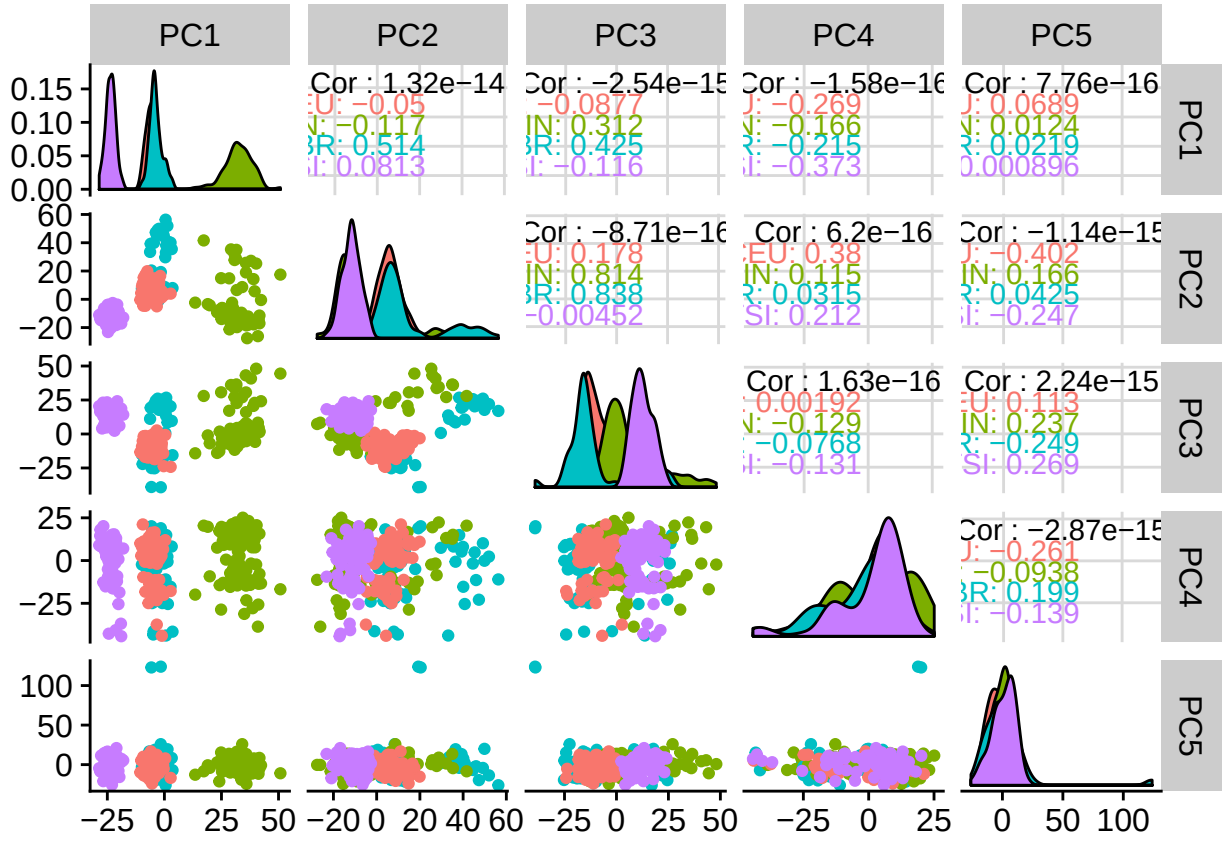

Percent of total variance explained by each principal component
Every 10th Genotype
Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Principal Component

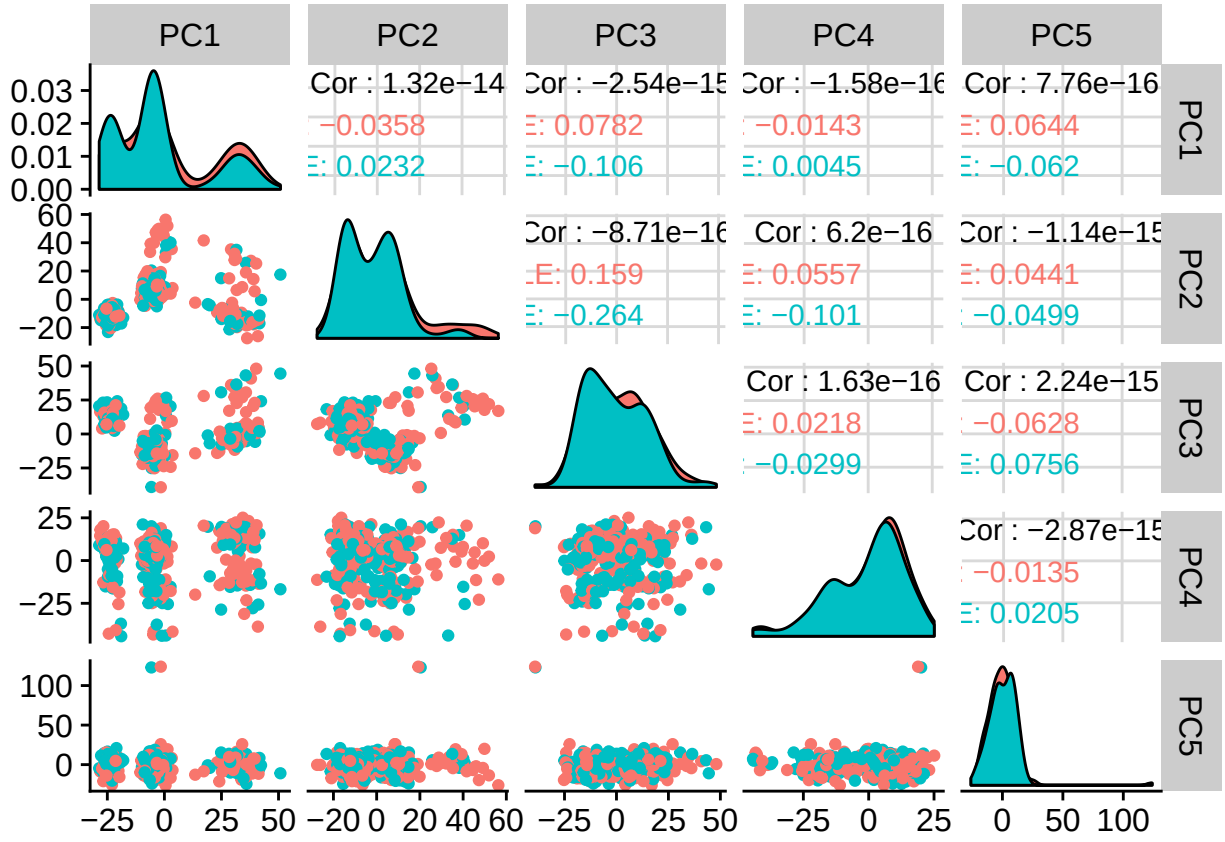

Principal Component Analsysis
Colored by Sex Covariate
Principal Component 2
20
10
0
−10
−10
0
10
Principal Component 1
Sex
FEMALE
MALE


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

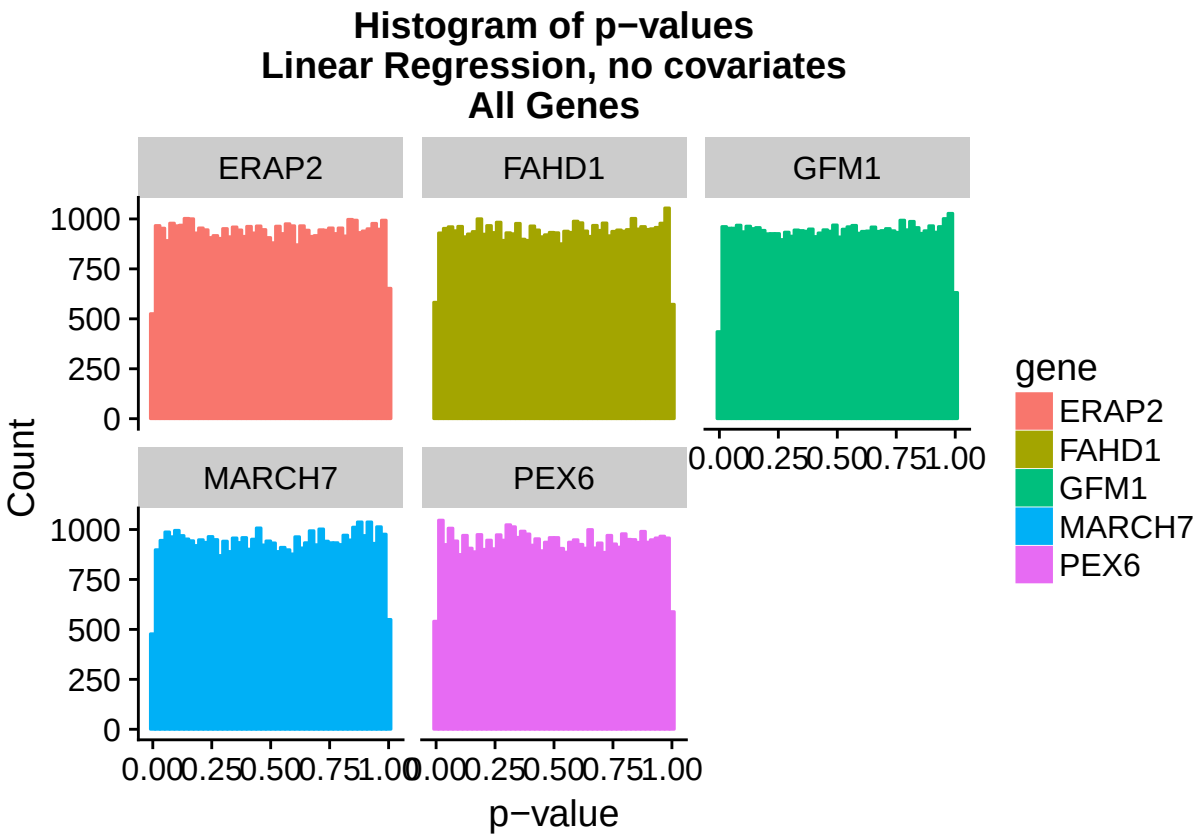

PC1
PC2
PC3
PC4
PC5
Cor : 6.06e−15
−0.0308
: 0.0239
Cor : −1.02e−15
: 0.0824
: −0.103
Cor : 1.68e−17
: 0.0206
−0.0223
Cor : 2.06e−16
−0.0196
: 0.0187
Cor : 3.66e−16
E: 0.125
: −0.178
Cor : −4.26e−16
−0.0507
: 0.0811
Cor : −1e−16
: 0.0441
−0.0642
Cor : −1.7e−16
−0.0687
: 0.0878
Cor : 4.23e−16
: −0.105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

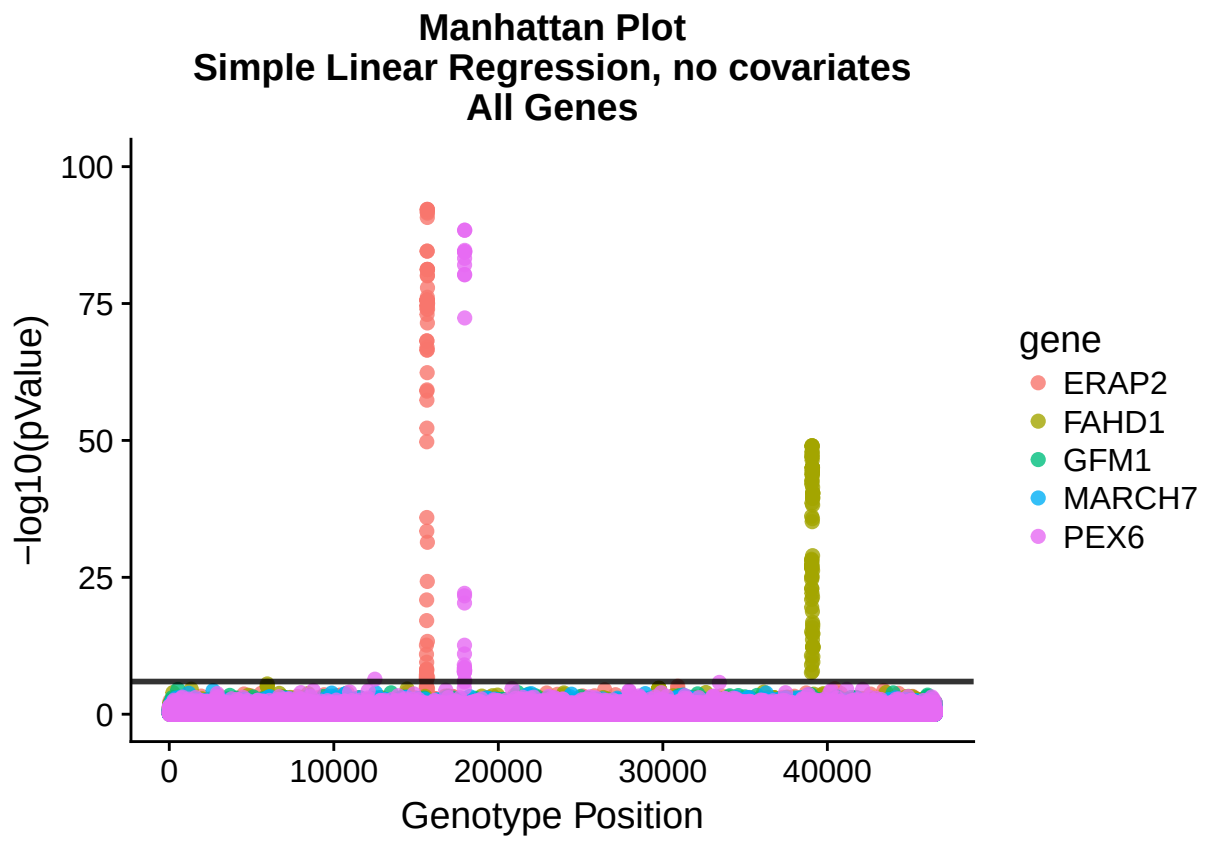

PC1
PC2
PC3
PC4
PC5
Cor : 1.32e−14
Cor : −2.54e−15
Cor : −1.58e−16
Cor : 7.76e−16
Cor : −8.71e−16
Cor : 6.2e−16
Cor : −1.14e−15
Cor : 1.63e−16
Cor : 2.24e−15
Cor : −2.87e−15
0.15
0.10
0.05
0.00
60
40
20
0
−20
50
25
0
−25
25
0
−25
100
50
0
−25 0 25 50
−20 0 20 40 60
−25 0 25 50
−25 0 25
0 50 100



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
```

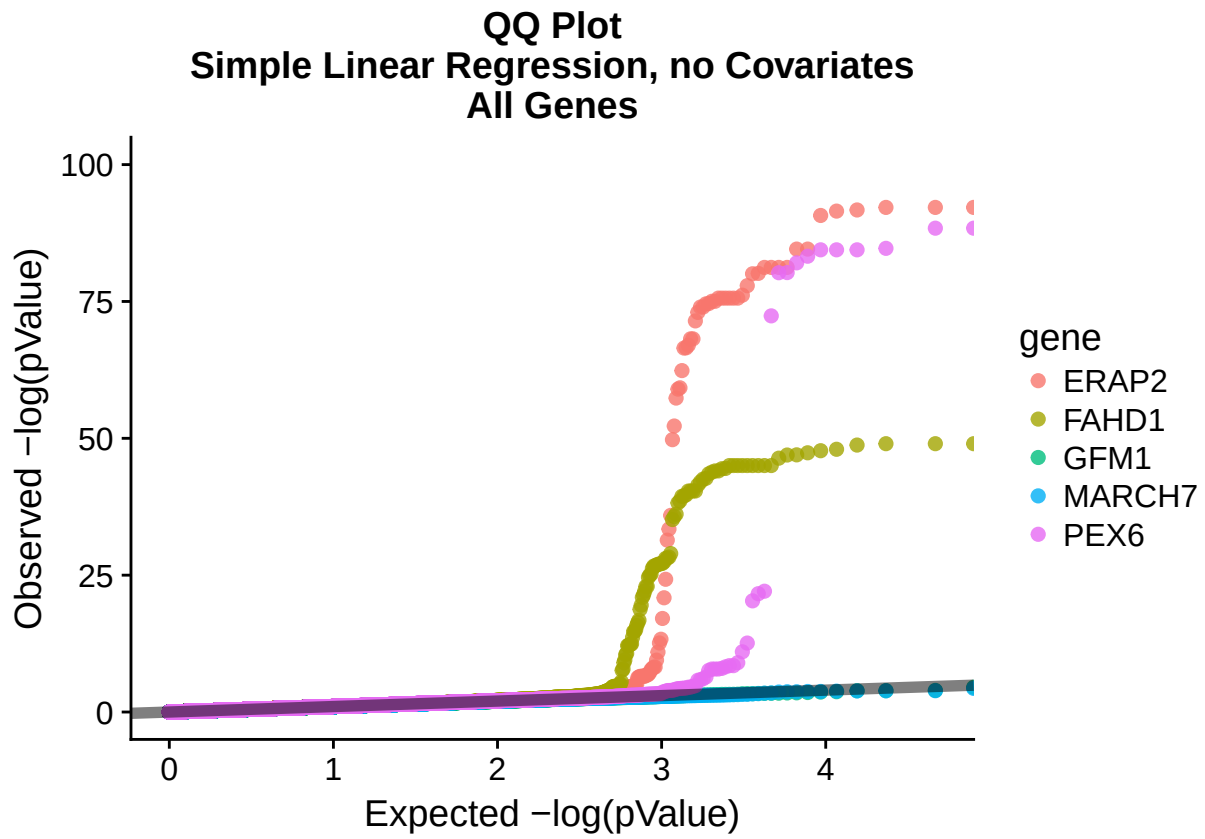
```
[1] 16
```

```
   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. 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
```

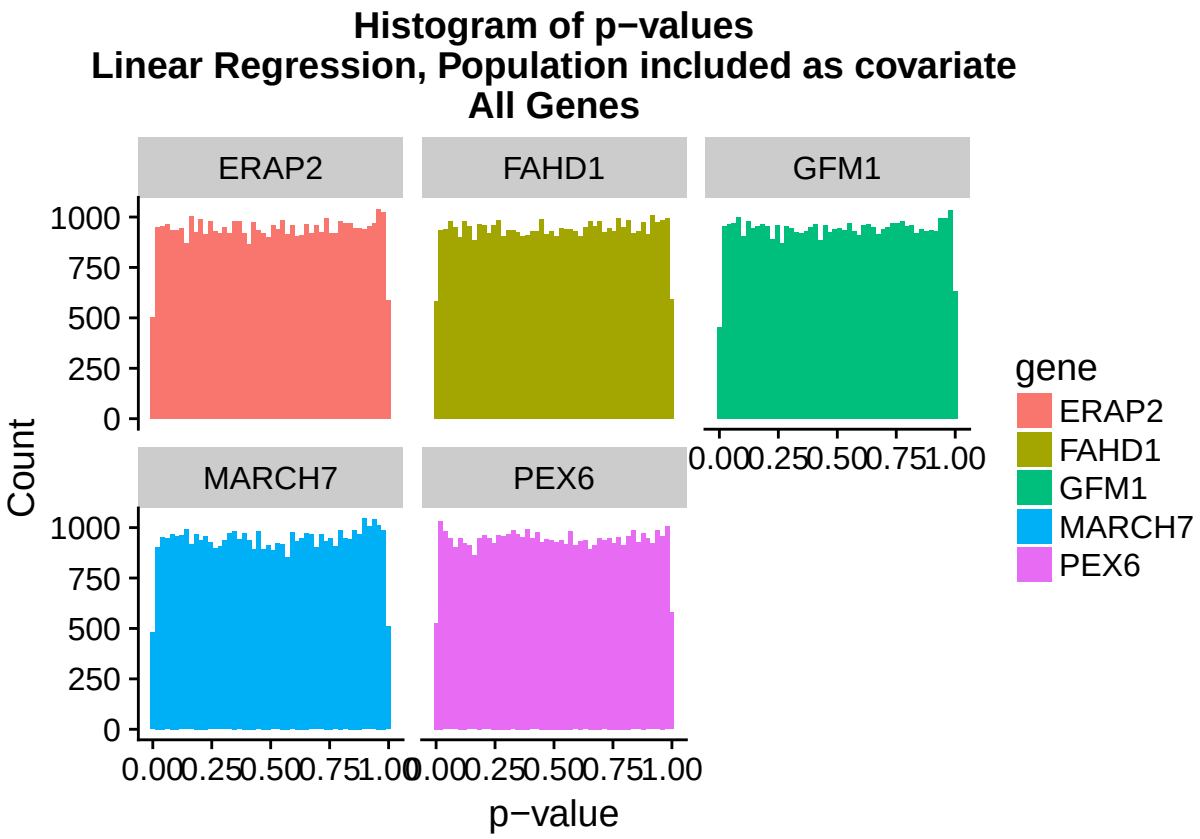
```
[1] FALSE
```

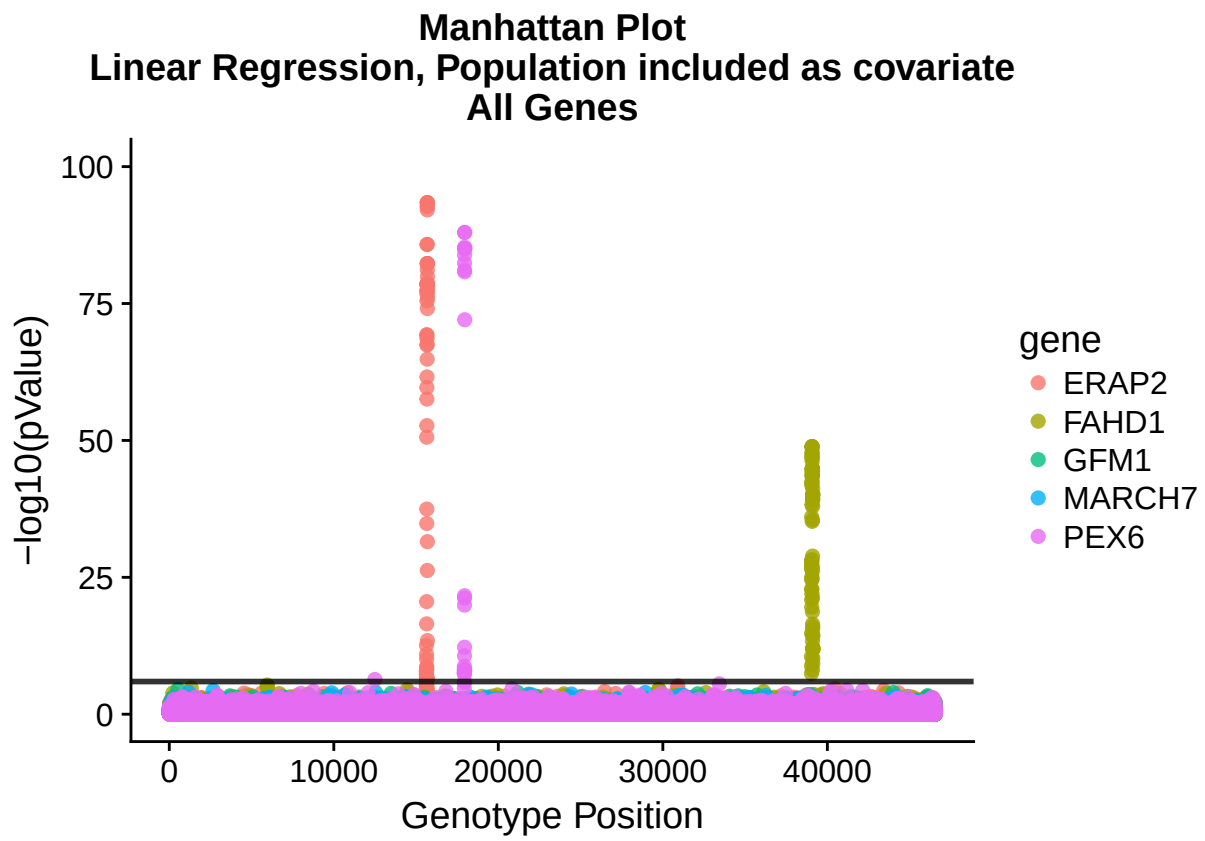
```
# A tibble: 1 x 2
  gene  p.value
  <chr>   <dbl>
1 ERAP2 0.00326
```

```
# 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
```

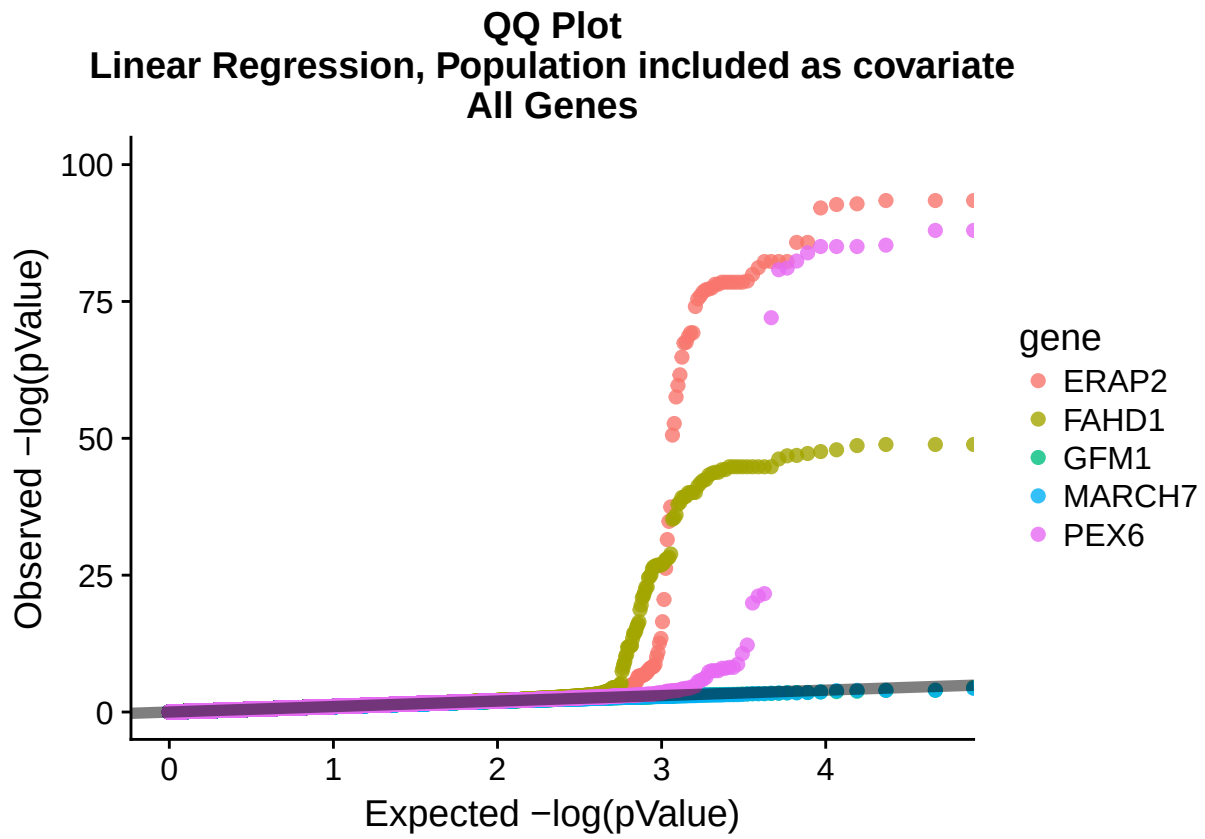
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15628   15646   15666   15665   15683   15701
```

```
[1] TRUE
```

```
[1] TRUE
```

```
[1] 4 6
```

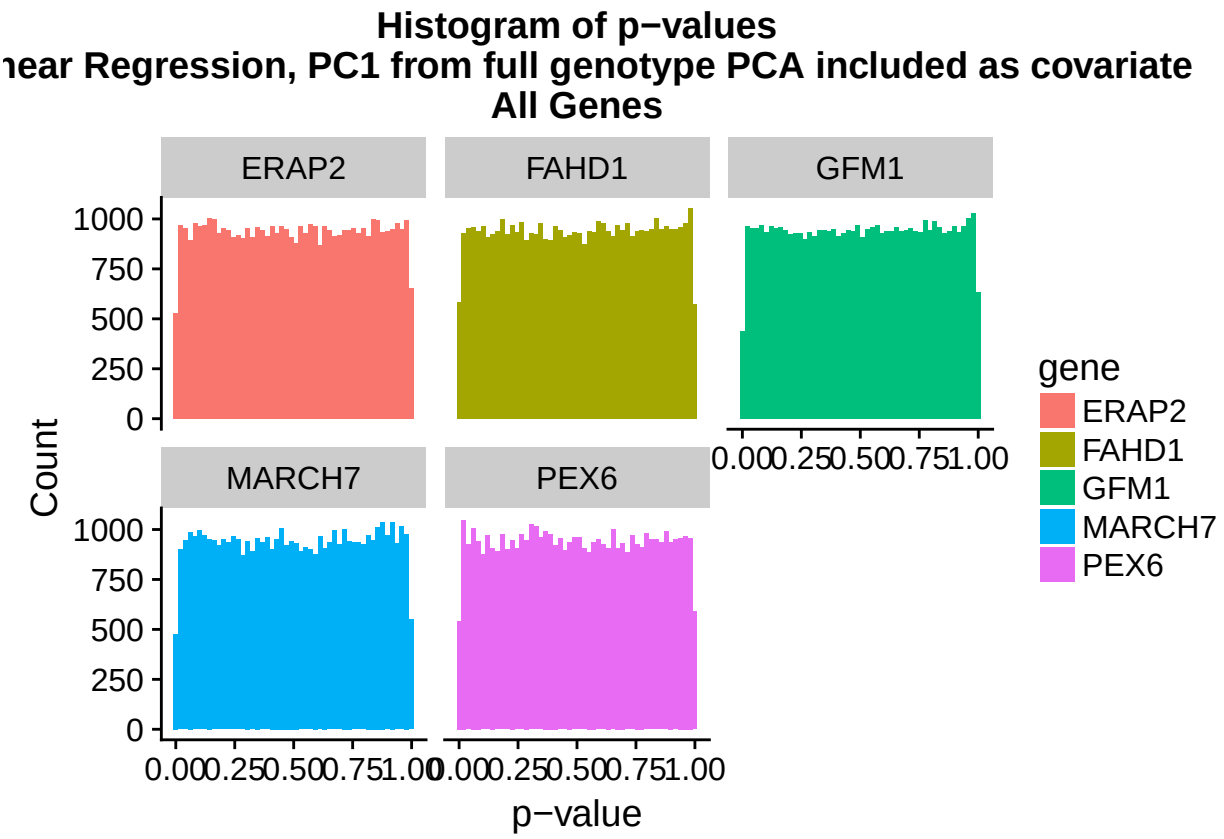
```
          id chromosome position in_gene
26 rs3805941          6 43089842   FALSE
```

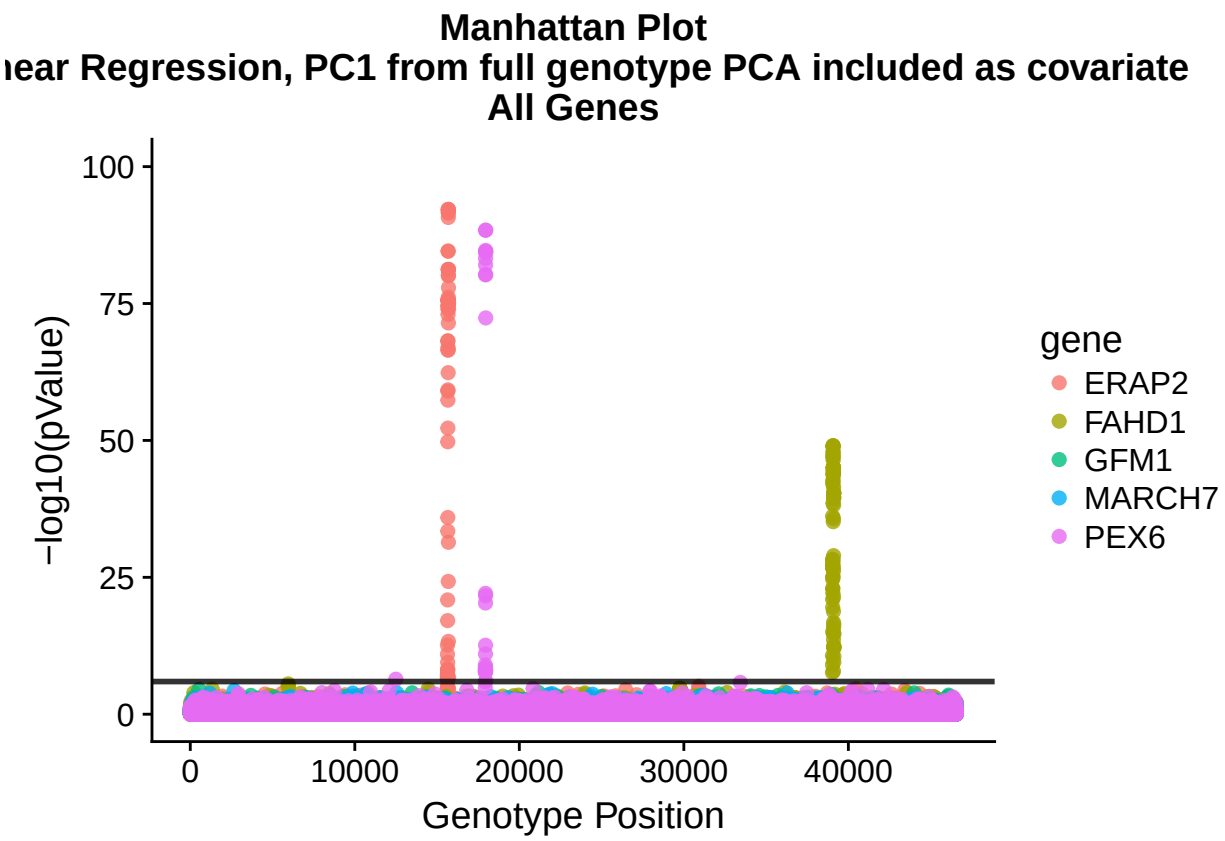


Try PC1 as a covariate:

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

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

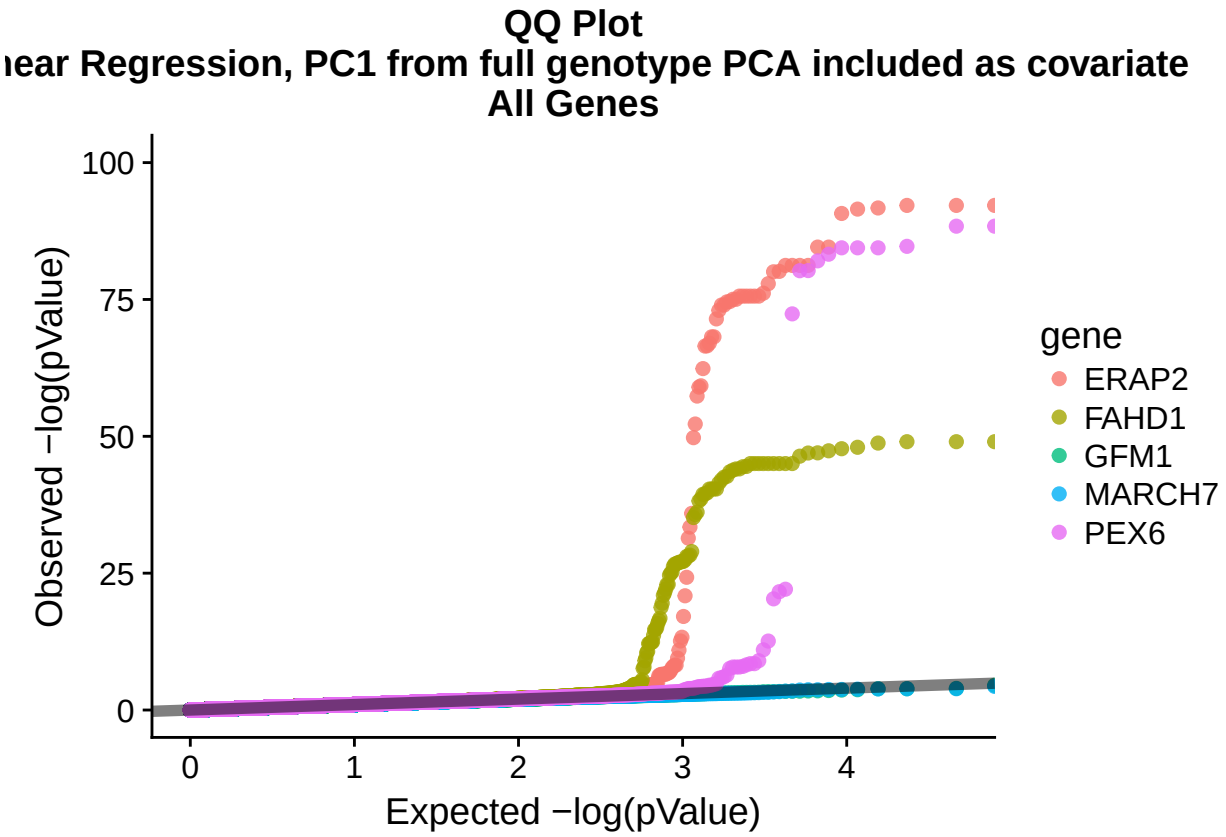

Histogram of p−values
ıear Regression, PC1 from full genotype PCA included as covariate
All Genes
ERAP2
FAHD1
GFM1
MARCH7
PEX6
Count
1000
750
500
250
0
0.00
0.25
0.50
0.75
1.00
p−value
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6




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

```
[1] TRUE
```

```
[1] "Numeric: lengths (27, 26) differ"
```



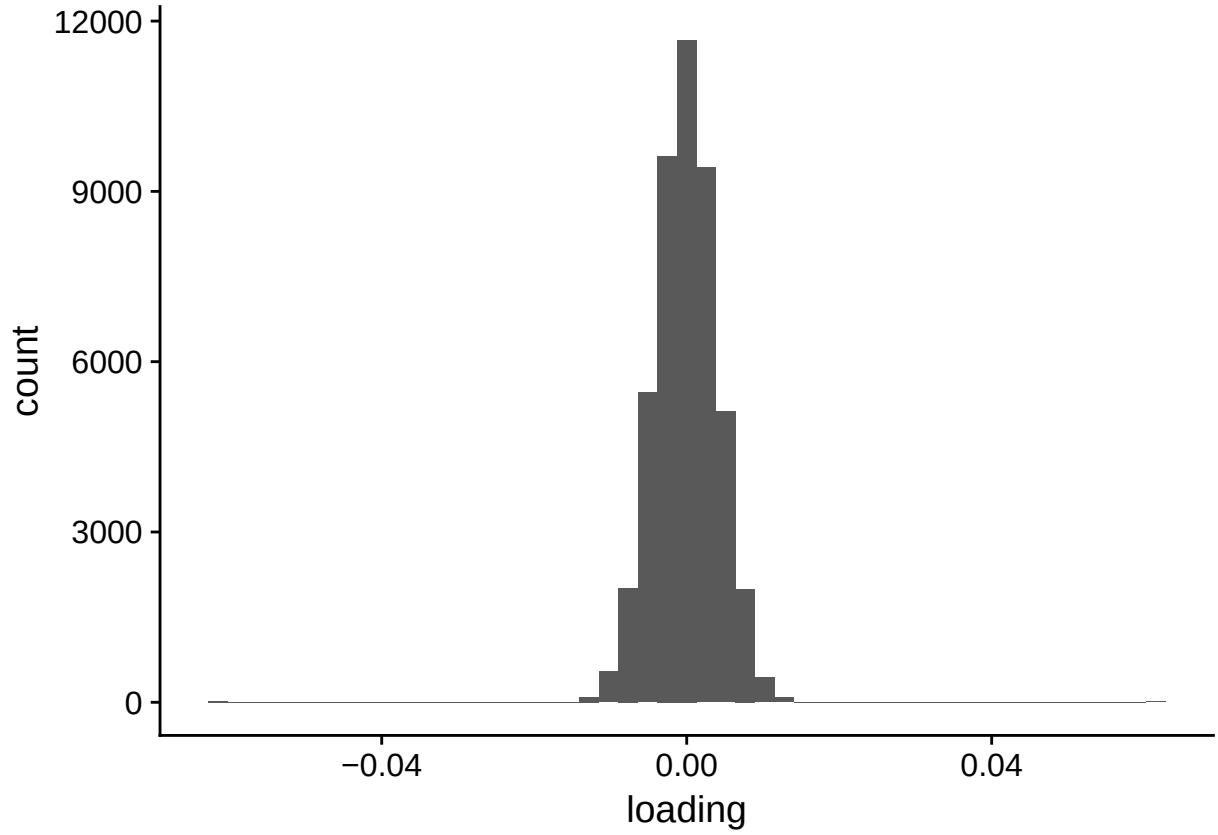
So far, for each analysis, each gene was treated the same. What is the relationship of the expression level of each gene with the first few principal components?

```
# 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
```

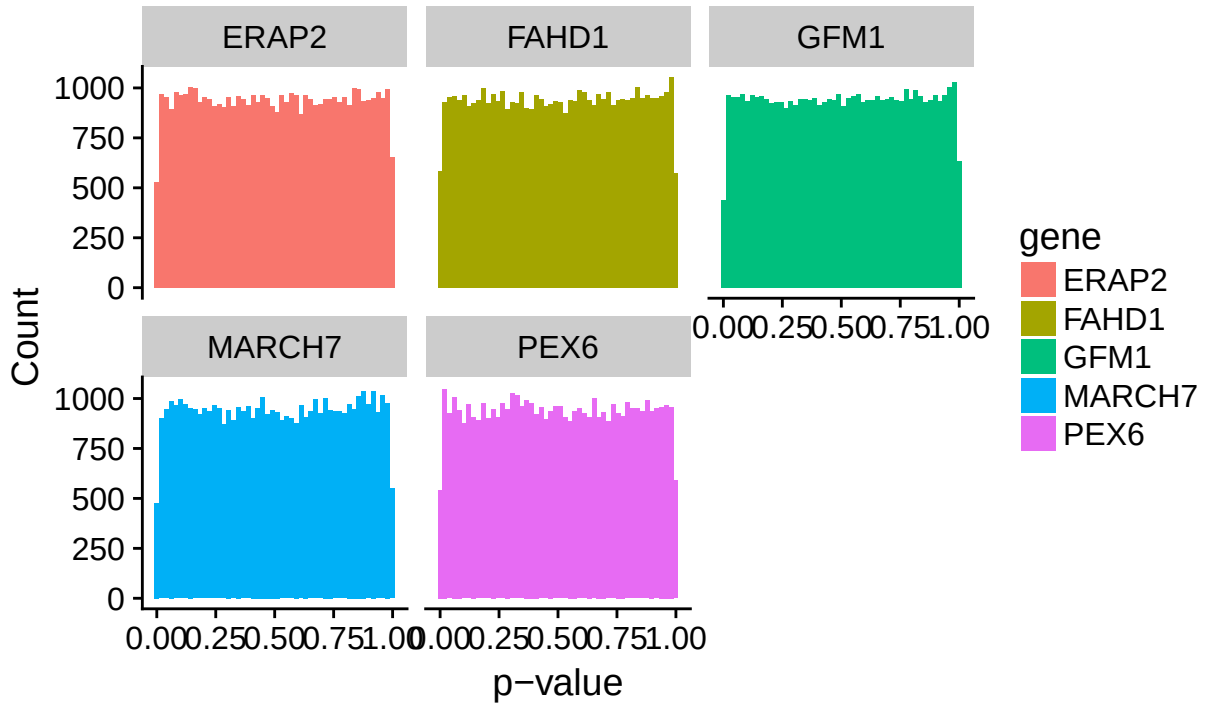


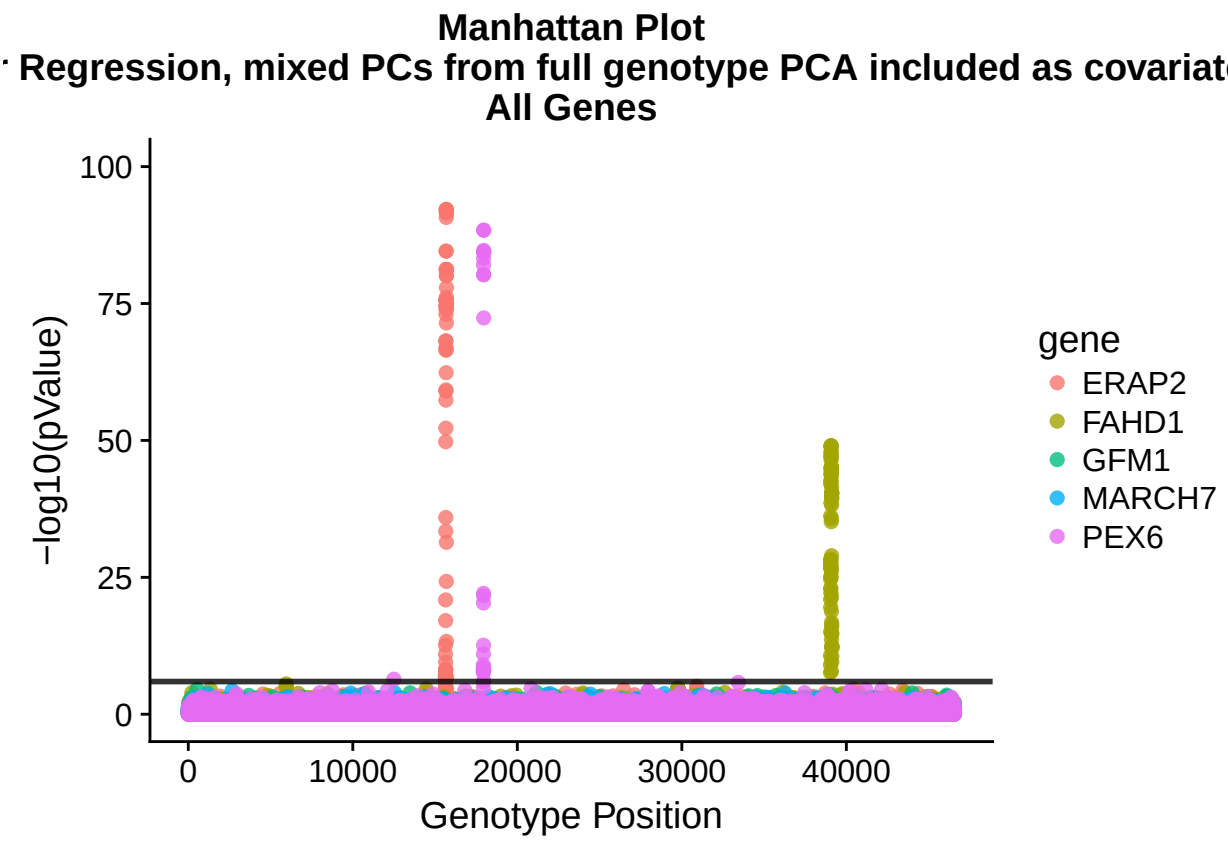

```
[1] 16
```

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

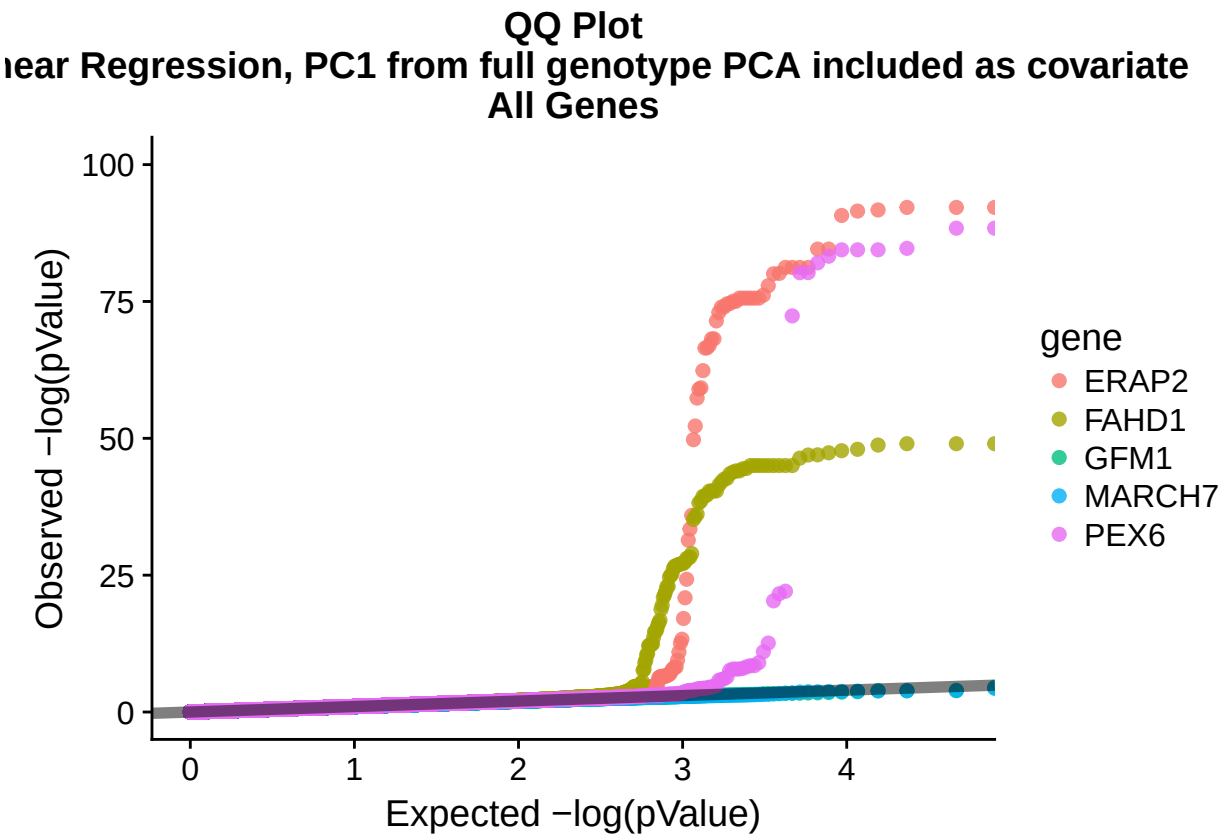
```
[1] 13.88889
```

```
[1] 87.80488
```


Histogram of p−values
r Regression, mixed PCs from full genotype PCA 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     27
```



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.

```
lm_test <- lm(pheno$ENSG00000164308.12 ~ x_a[, 1000] + x_d[, 1000] + factor(covars$Population))
lm_tidy <- tidy(lm_test)
fstat <- summary(lm_test)$fstatistic
fstat_2 <- glance(lm_test)$statistic
pf(fstat[1], fstat[2], fstat[3], lower.tail = FALSE)

#Including Population and Sex as Covariates:
x_c_pop_and_sex <- cbind(
  as.numeric(factor(pheno_with_covars$Population)),
  as.numeric(factor(pheno_with_covars$Sex))
  )
pval_ERAP2_w_pop_and_sex_covar <- map2(
  data.frame(x_a), data.frame(x_d),
  lin_reg_fstat_w_covar,
  pheno_names$ERAP2, x_c_pop_and_sex
  ) %>%
  flatten_dbl()
summary(which(pval_ERAP2_w_pop_and_sex_covar < 0.05 / ncol(geno)))
summary(which(pval_ERAP2 < 0.05 / ncol(geno)))

pval_FAHD1_w_pop_and_sex_covar <- map2(
  data.frame(x_a), data.frame(x_d),
  lin_reg_fstat_w_covar,
  pheno_names$FAHD1, x_c_pop_and_sex
```

```
  ) %>%
  flatten_dbl()
pval_GFM1_w_pop_and_sex_covar <- map2(
  data.frame(x_a), data.frame(x_d),
  lin_reg_fstat_w_covar,
  pheno_names$GFM1, x_c_pop_and_sex
  ) %>%
  flatten_dbl()
pval_MARCH7_w_pop_and_sex_covar <- map2(
  data.frame(x_a), data.frame(x_d),
  lin_reg_fstat_w_covar,
  pheno_names$MARCH7, x_c_pop_and_sex
  ) %>%
  flatten_dbl()
pval_PEX6_w_pop_and_sex_covar <- map2(
  data.frame(x_a), data.frame(x_d),
  lin_reg_fstat_w_covar,
  pheno_names$PEX6, x_c_pop_and_sex
  ) %>%
  flatten_dbl()
pval_pop_and_sex_covar <- as.data.frame(
  cbind(
    x = 1:ncol(geno),
    ERAP2 = pval_ERAP2_w_pop_and_sex_covar,
    FAHD1 = pval_FAHD1_w_pop_and_sex_covar,
    GFM1 = pval_GFM1_w_pop_and_sex_covar,
    MARCH7 = pval_MARCH7_w_pop_and_sex_covar,
    PEX6 = pval_PEX6_w_pop_and_sex_covar
    )
  )
row.names(pval_pop_and_sex_covar) <- colnames(geno)
pval_pop_and_sex_covar %>%
  gather("gene", "pval", 2:6) %>%
  ggplot(aes(x = pval, fill = gene)) +
  geom_histogram(position = "identity", alpha = 0.5, bins = 20)
pval_pop_and_sex_covar %>%
  gather("gene", "pval", 2:6) %>%
  ggplot(aes(x = x, y = -log(pval, base = 10), color = gene)) +
  geom_point(alpha = 0.8, size = 2) +
  ggtitle("Manhattan Plot\nAll genes after Including Population and Sex as Covars") +
  xlab("Genotype Position") +
  ylab("-log10(pVal)") +
  ylim(0, 100)
pval_pop_and_sex_covar_qqplot <- data.frame(
  cbind(
    expected = sort(-log10(seq(from = 0,to = 1, length.out = length(pval_pop_and_sex_covar$x)))),
    ERAP2 = sort(-log10(pval_pop_and_sex_covar$ERAP2)),
    FAHD1 = sort(-log10(pval_pop_and_sex_covar$FAHD1)),
    GFM1 = sort(-log10(pval_pop_and_sex_covar$GFM1)),
    MARCH7 = sort(-log10(pval_pop_and_sex_covar$MARCH7)),
    PEX6 = sort(-log10(pval_pop_and_sex_covar$PEX6))
    )
  )
```

```r
pval_pop_and_sex_covar_qqplot %>%
  gather("gene", "neglog10_pval", 2:6) %>%
  ggplot(aes(x = expected, y = neglog10_pval, color = gene)) +
  geom_point(alpha = 0.5) +
  geom_abline(intercept = 0, slope = 1, col = "black", alpha = 0.5, size = 2) +
  xlab("Expected -log(pValue)") +
  ylab("Observed -log(pValue)") +
  ggtitle("QQ Plot\nAll Genes") +
  ylim(0, 100)


xa_pca <- prcomp(x_a, center = TRUE, scale. = TRUE)
plot((xa_pca$sdev^2/ sum(xa_pca$sdev^2)) * 100, xlab = "Principal Component", ylab = "Percent of total v
xa_pca_x <- data.frame(xa_pca$x)
xa_pca_x %>%
  rownames_to_column("sample") %>%
  left_join(covars %>% rownames_to_column("sample"), by = "sample") %>%
  ggplot(aes(x = PC1, y = PC2, color = Population)) +
    geom_point(size = 2, alpha = 0.8) +
    ggtitle("First two principal components of X_a matrix, colored by Population")
plot(geno_HW_filt_pca_x$PC1, xa_pca_x$PC1)
abline(a = 0, b = 1)
plot(geno_HW_filt_pca_x$PC2, xa_pca_x$PC2)
abline(a = 0, b = 1)

xd_pca <- prcomp(x_d, center = TRUE, scale. = TRUE)
plot((xd_pca$sdev^2/ sum(xd_pca$sdev^2)) * 100, xlab = "Principal Component", ylab = "Percent of total v
xd_pca_x <- data.frame(xa_pca$x)
xd_pca_x %>%
  rownames_to_column("sample") %>%
  left_join(covars %>% rownames_to_column("sample"), by = "sample") %>%
  ggplot(aes(x = PC1, y = PC2, color = Population)) +
    geom_point(size = 2, alpha = 0.8) +
    ggtitle("First two principal components of X_d matrix, colored by Population")
plot(geno_HW_filt_pca_x$PC1, xd_pca_x$PC1)
abline(a = 0, b = 1)
plot(geno_HW_filt_pca_x$PC2, xd_pca_x$PC2)
abline(a = 0, b = 1)
```