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

```
# is there any missing data?
anyNA(list(geno, pheno, covars, snp_info, gene_info))
```

```
[1] FALSE
```

```
# are there approximately equal numbers of people in each covariate group?
table(covars$Population)
```

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

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

```
FEMALE   MALE
   181    163
```

```
# does the coding of the genotypes makes sense across all the data?
table(as.matrix(geno))
```

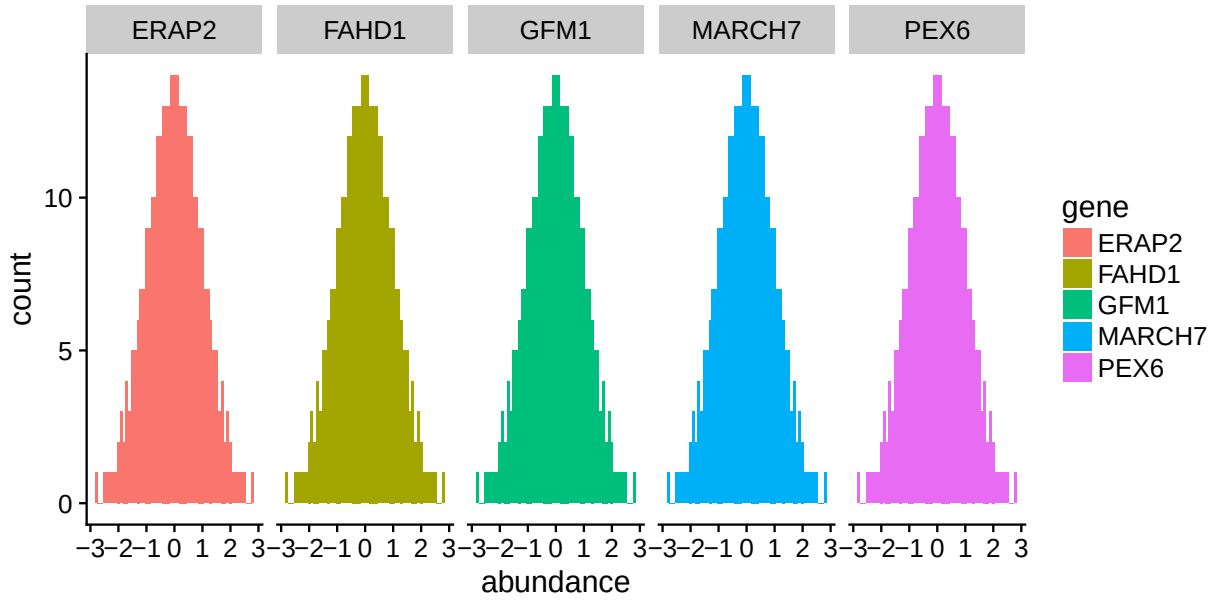
```
      0       1       2
8181444 5811217 3207339
```

```
# calculate allele frequency
geno_sums <- map_dbl(geno, function(x) sum(x) / (nrow(geno) * 2))
# any genotypes need to be removed because of MAF < 5%?
any(geno_sums < 0.05 | geno_sums > 0.95)
```

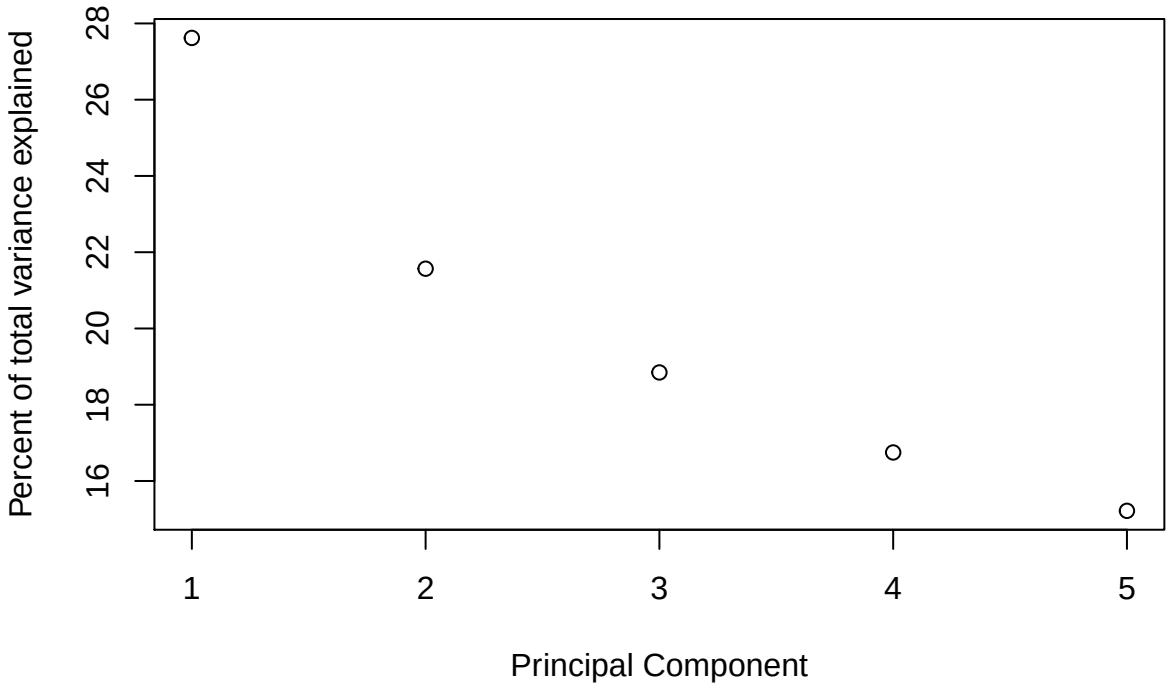
```
[1] FALSE
```

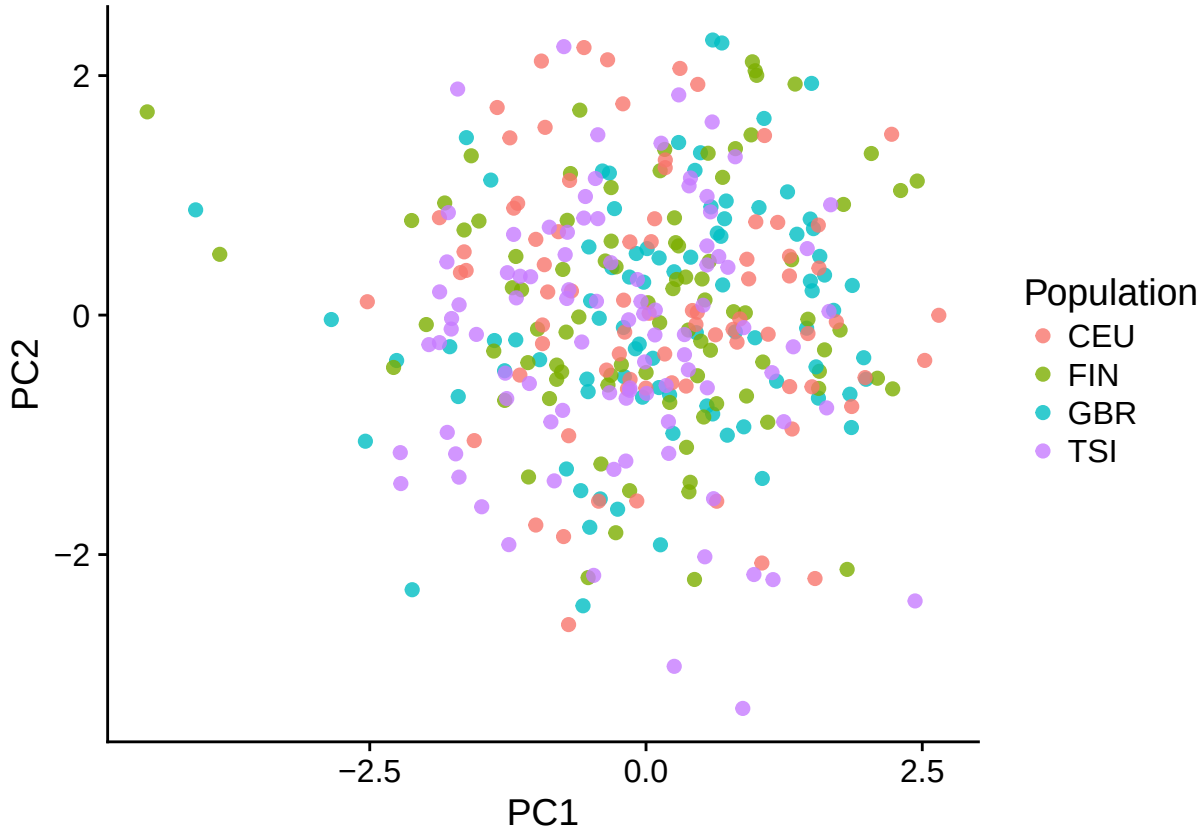
```
# no genotypes need to be removed because of MAF < 5%
# are there any genotypes without associated phenotypes, or vice versa?
all.equal(rownames(geno), rownames(pheno))
```

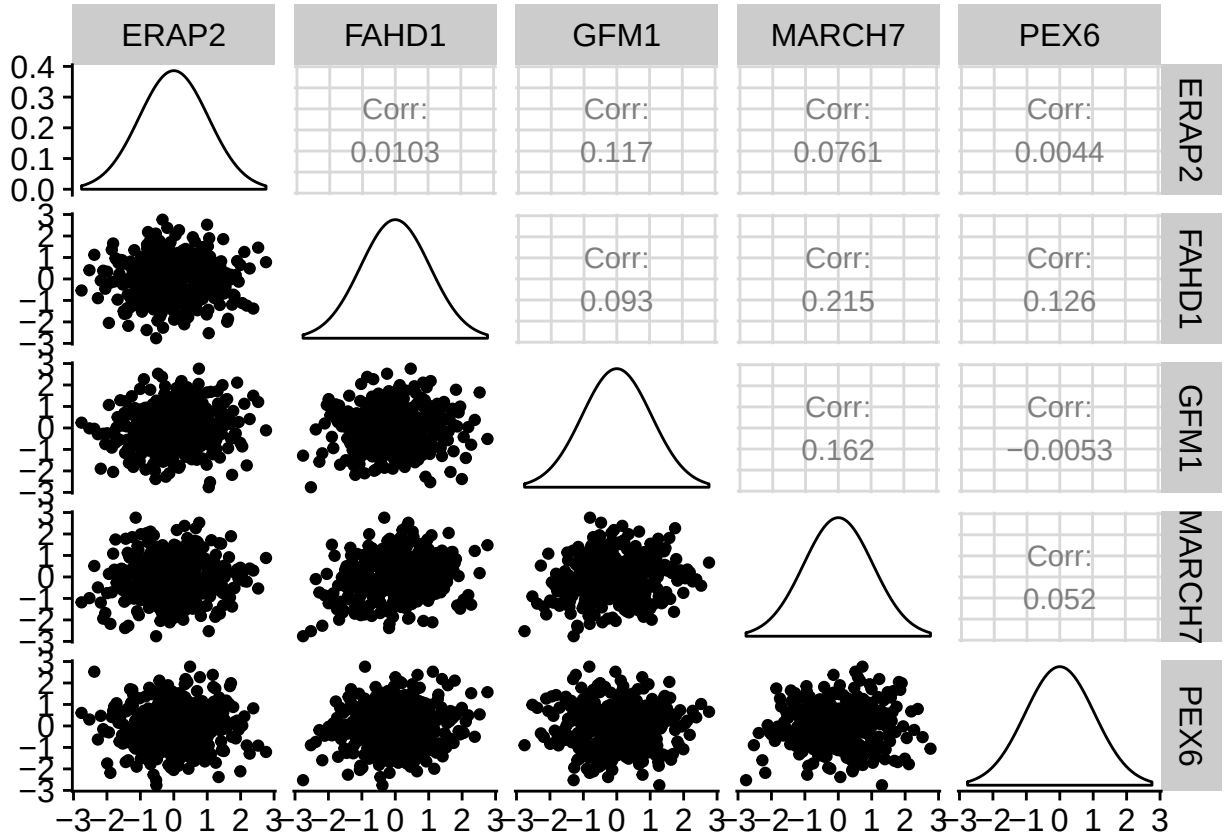
```
[1] TRUE
```

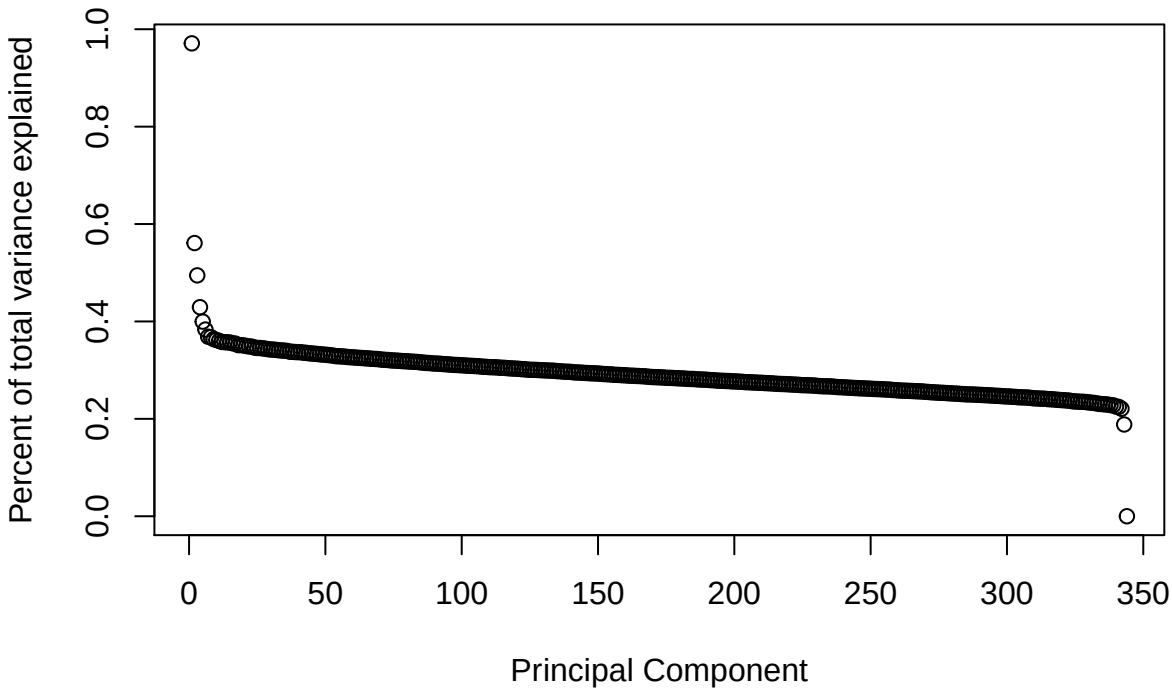


A PCA analysis was performed on the phenotype data

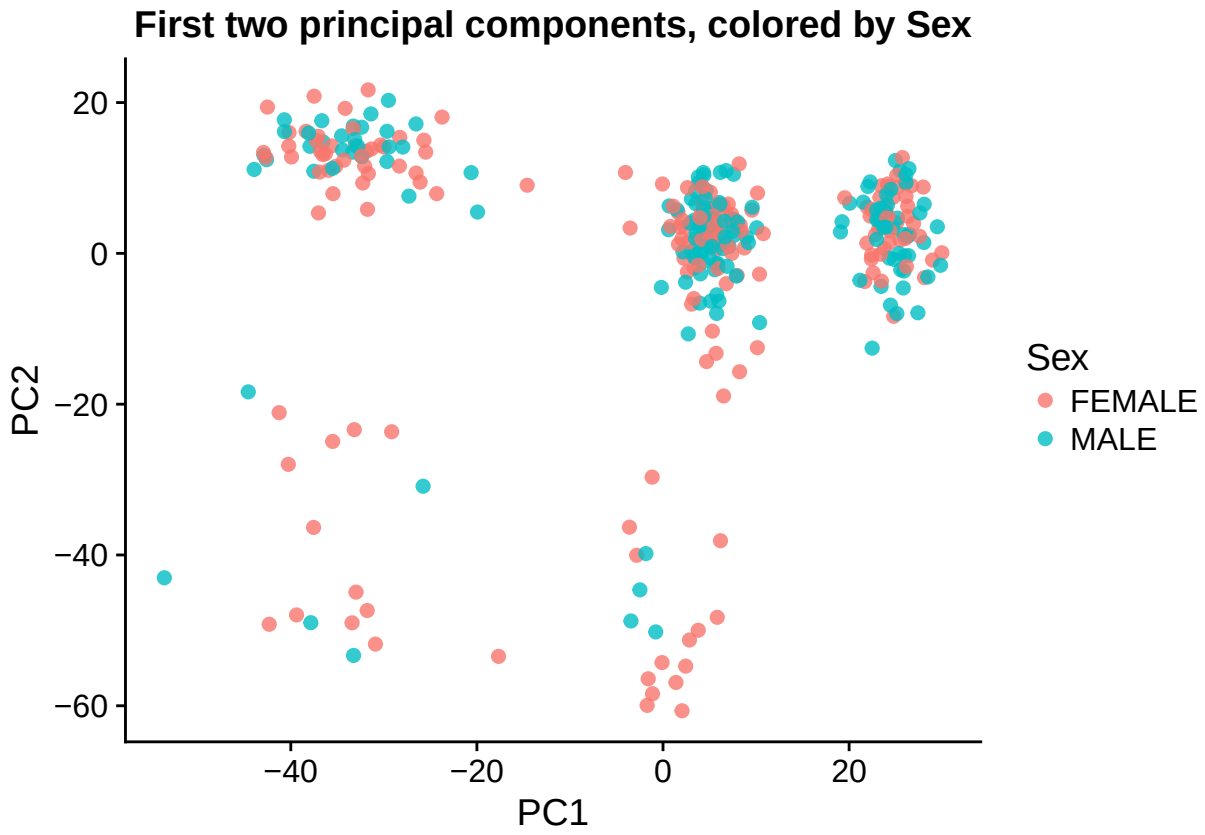


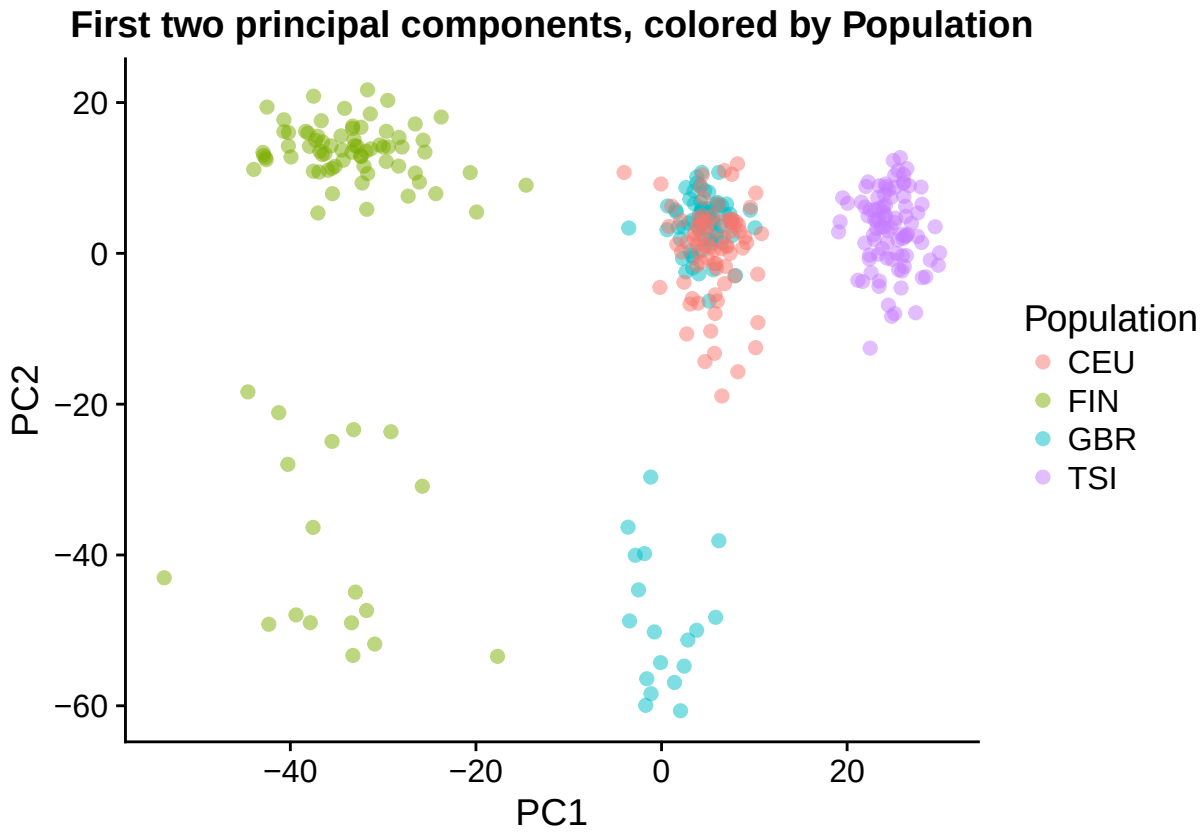

2
0
−2
PC2
−2.5
0.0
2.5
PC1
Population
CEU
FIN
GBR
TSI

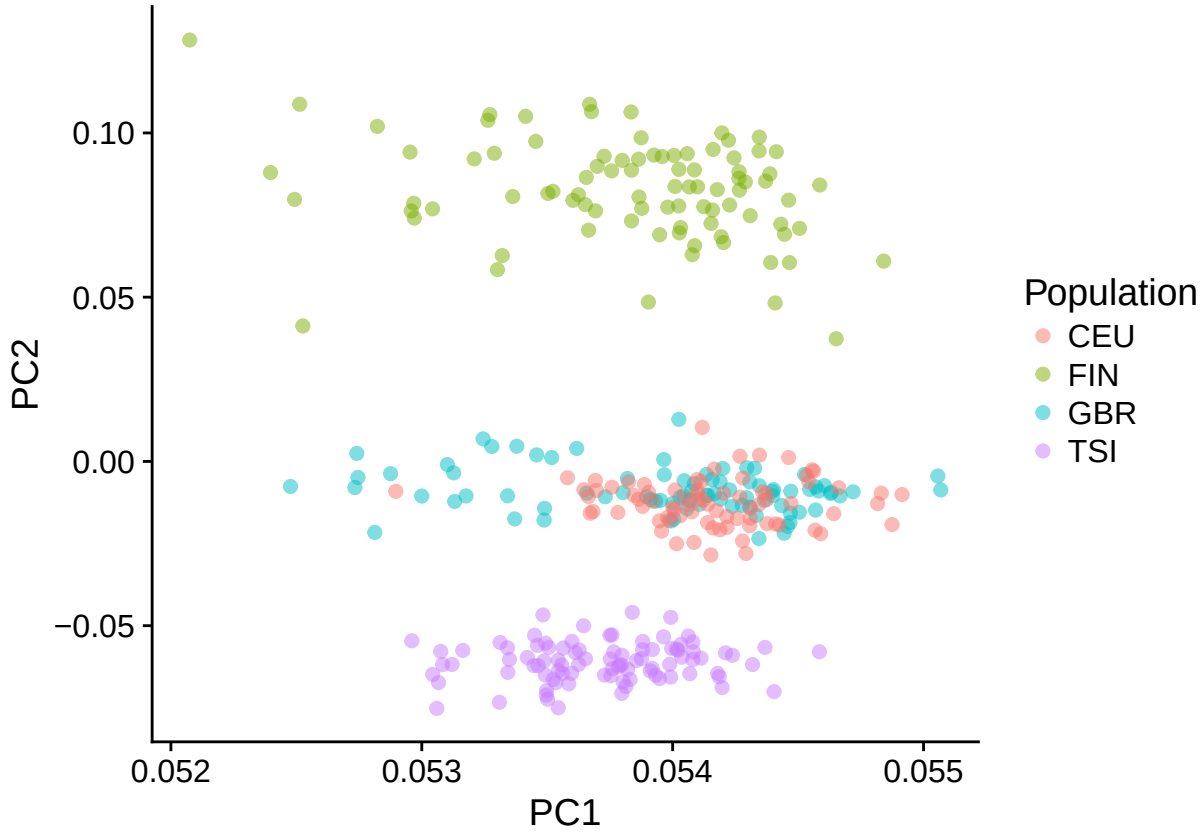

ERAP2
FAHD1
GFM1
MARCH7
PEX6
Corr:
0.0103
Corr:
0.117
Corr:
0.0761
Corr:
0.0044
Corr:
0.093
Corr:
0.215
Corr:
0.126
Corr:
0.162
Corr:
−0.0053
Corr:
0.052

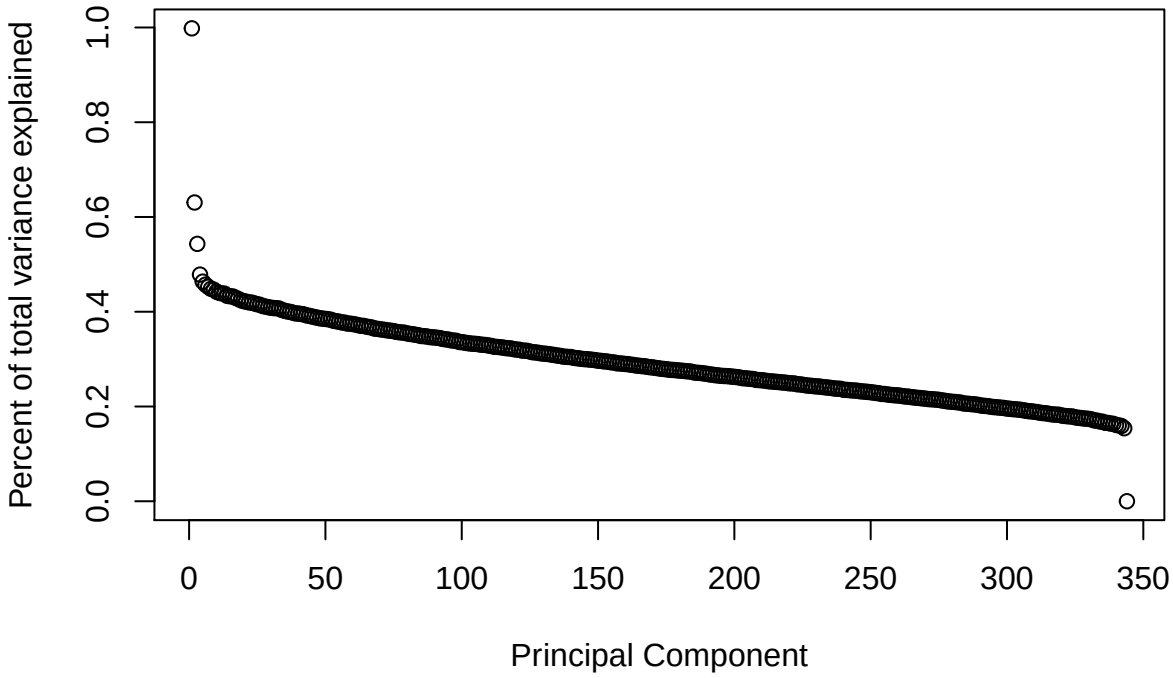


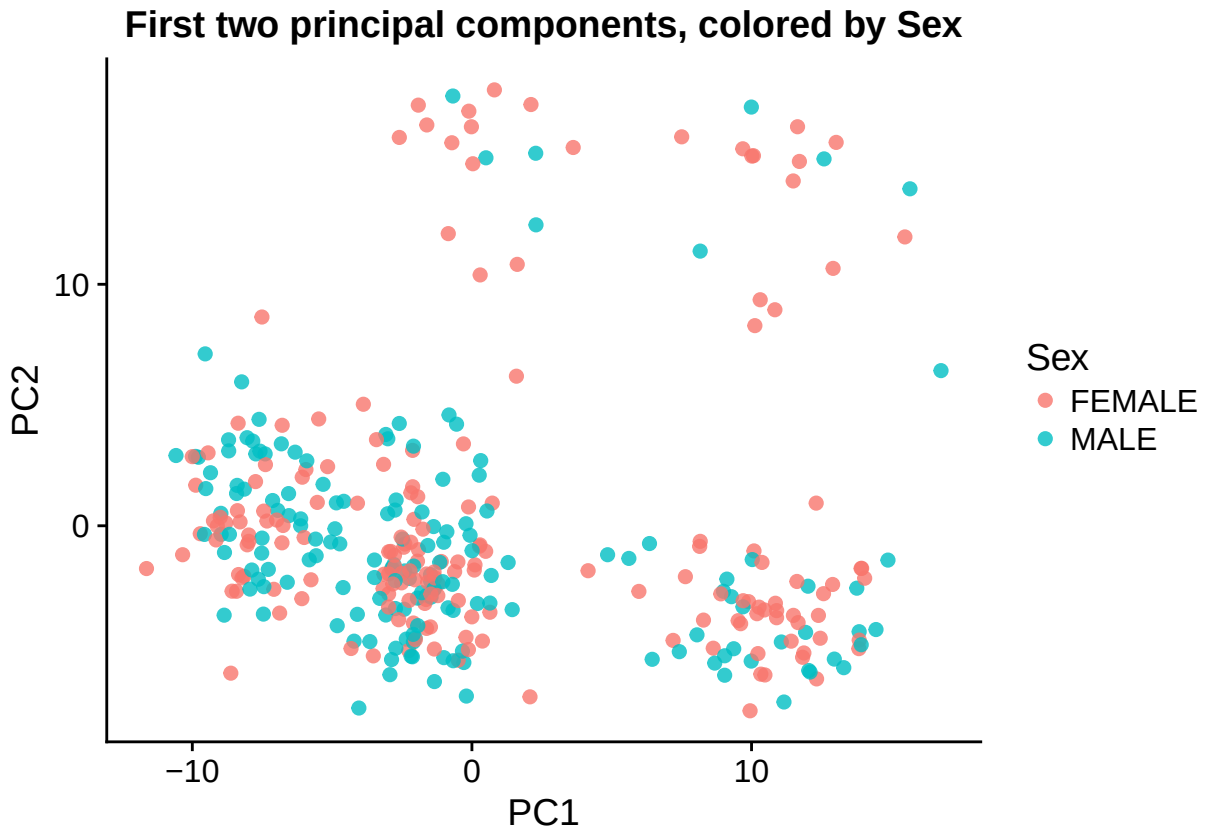
Genotype plots:

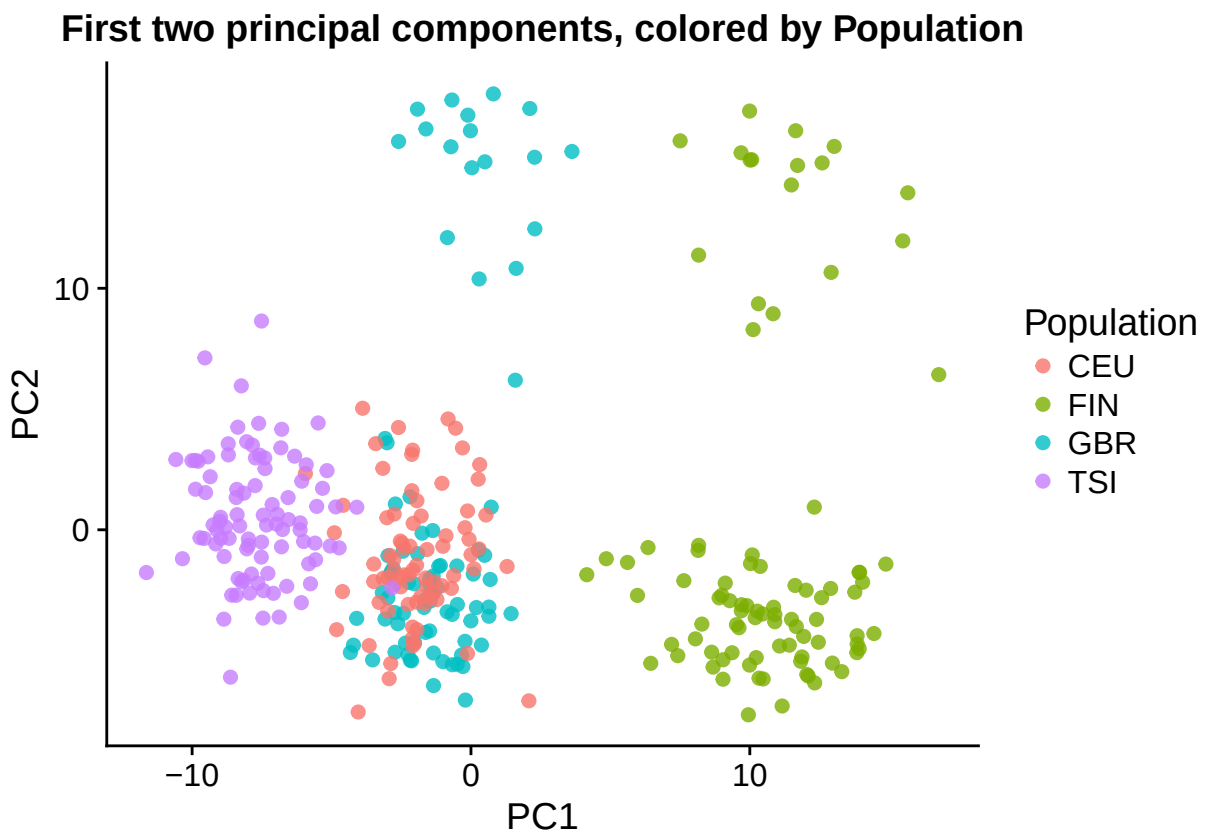

First two principal components, colored by Sex
20
0
−20
−40
−60
PC2
−40
−20
0
20
PC1
Sex
FEMALE
MALE


First two principal components, colored by Population
PC2
PC1
20
0
−20
−40
−60
−40
−20
0
20
Population
CEU
FIN
GBR
TSI


0.10
0.05
0.00
−0.05
PC2
0.052
0.053
0.054
0.055
PC1
Population
CEU
FIN
GBR
TSI

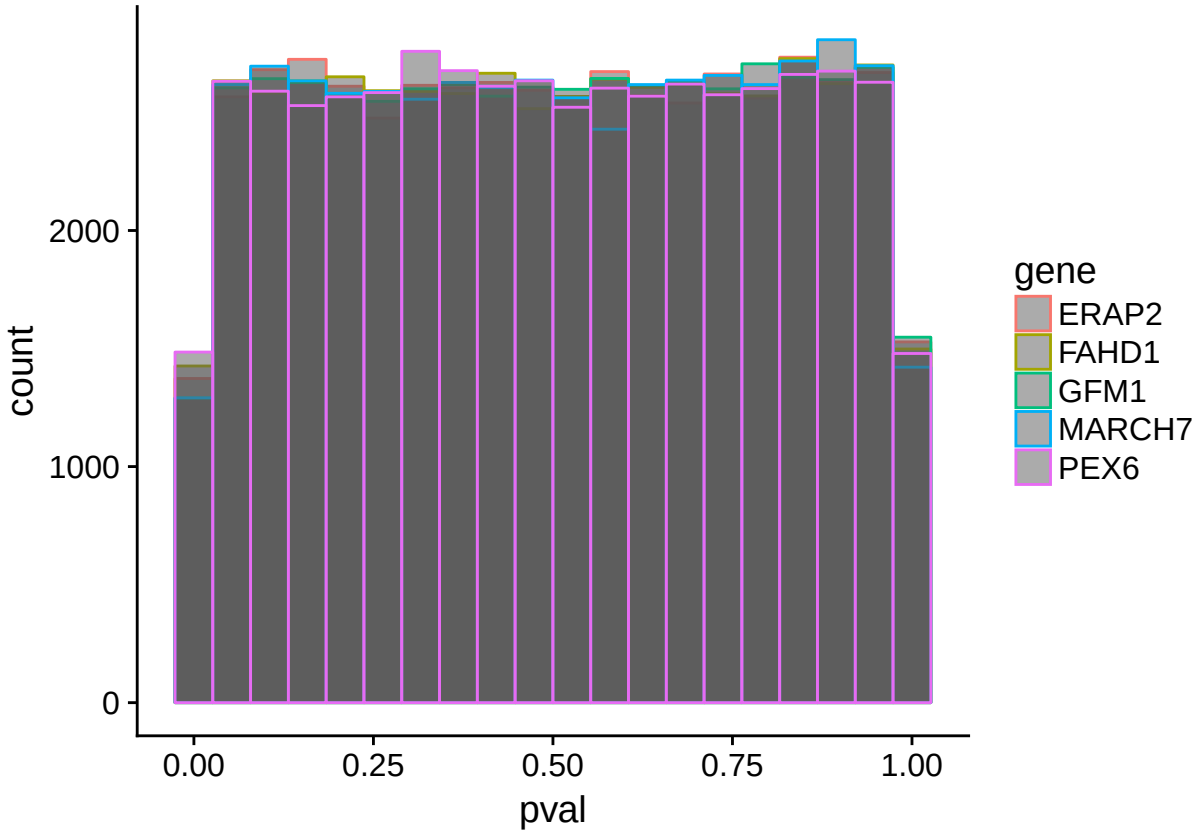

Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Principal Component

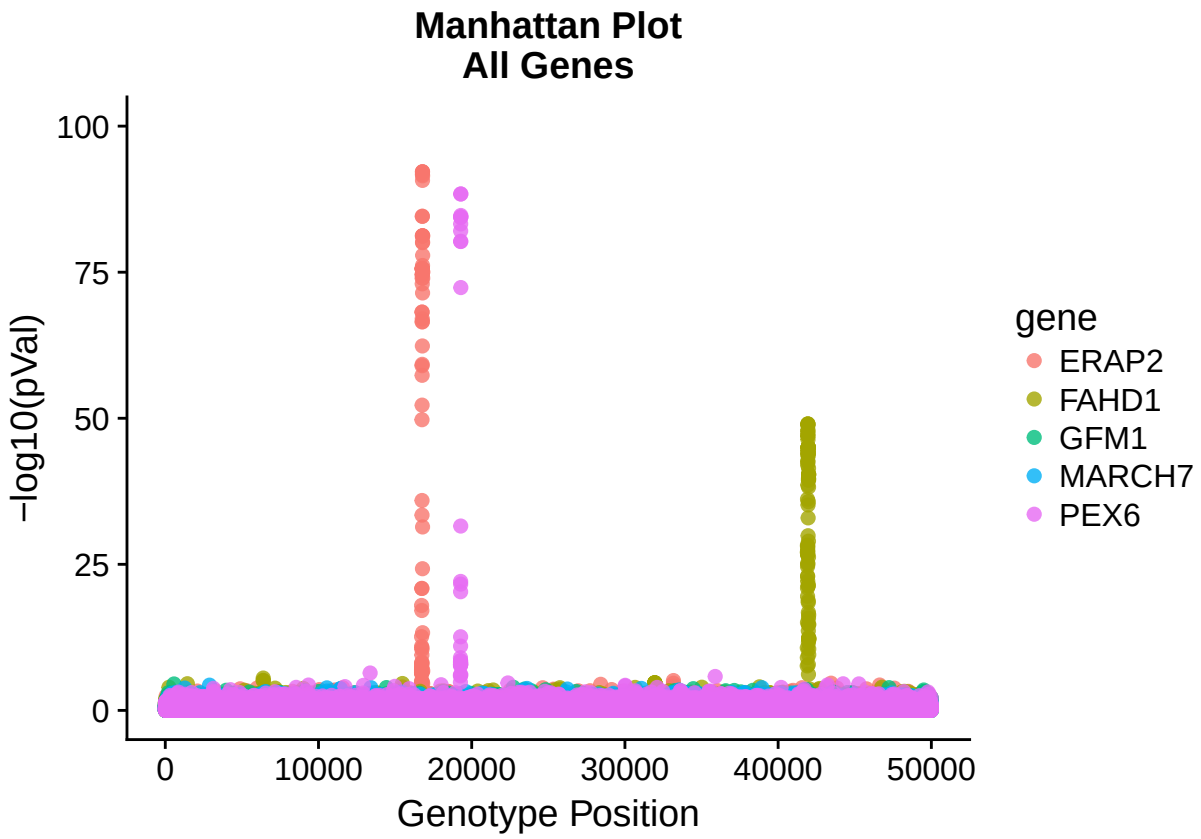

First two principal components, colored by Sex
PC2
10
0
−10
0
10
PC1
Sex
FEMALE
MALE



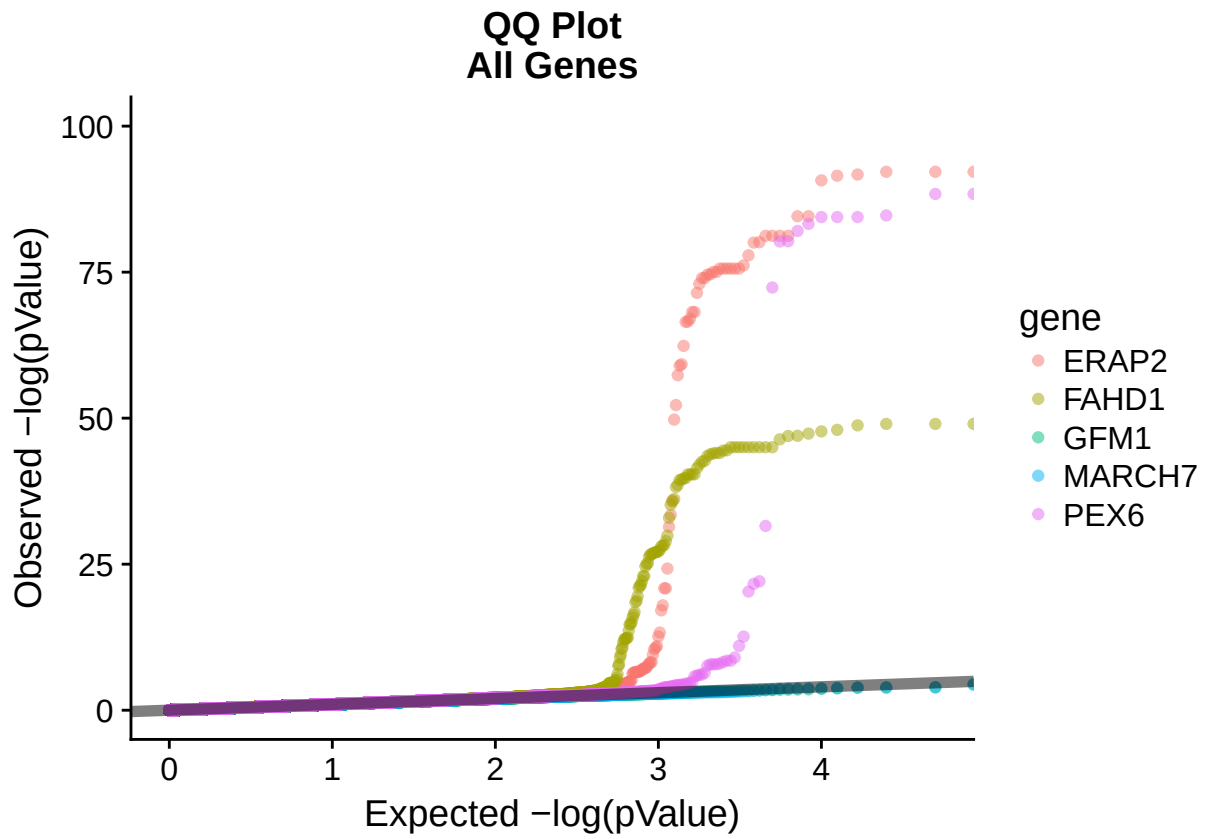
Based on on the principal component analysis of the genomes, colored by the population of origin, it is possible to see that there is clearly population structure.

Test each covariate individually.


count
2000
1000
0
0.00
0.25
0.50
0.75
1.00
pval
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6




```
# A tibble: 3 x 2
# Groups:   gene [3]
  gene      n
  <chr> <int>
1 ERAP2    75
2 FAHD1    93
3 PEX6     31
```



Seems like there is a problem with just the base model. 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
```

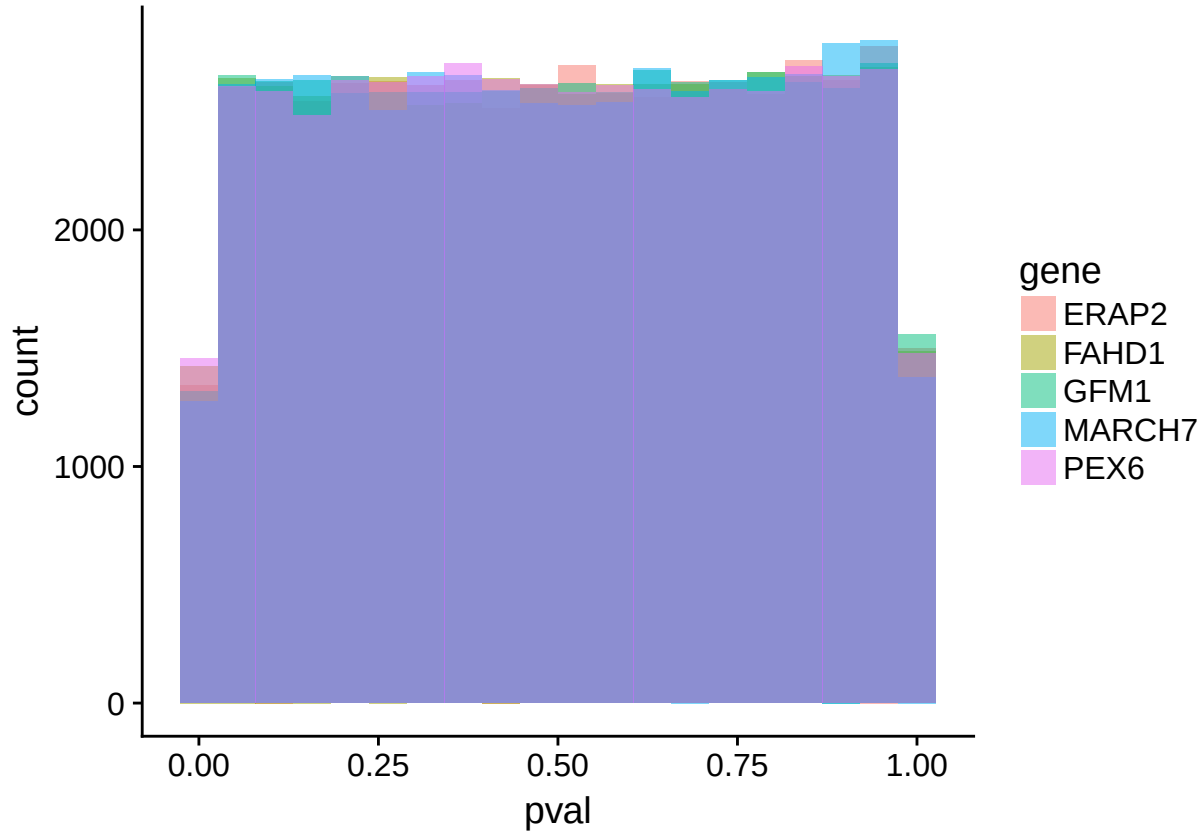
```
## [1] FALSE
```

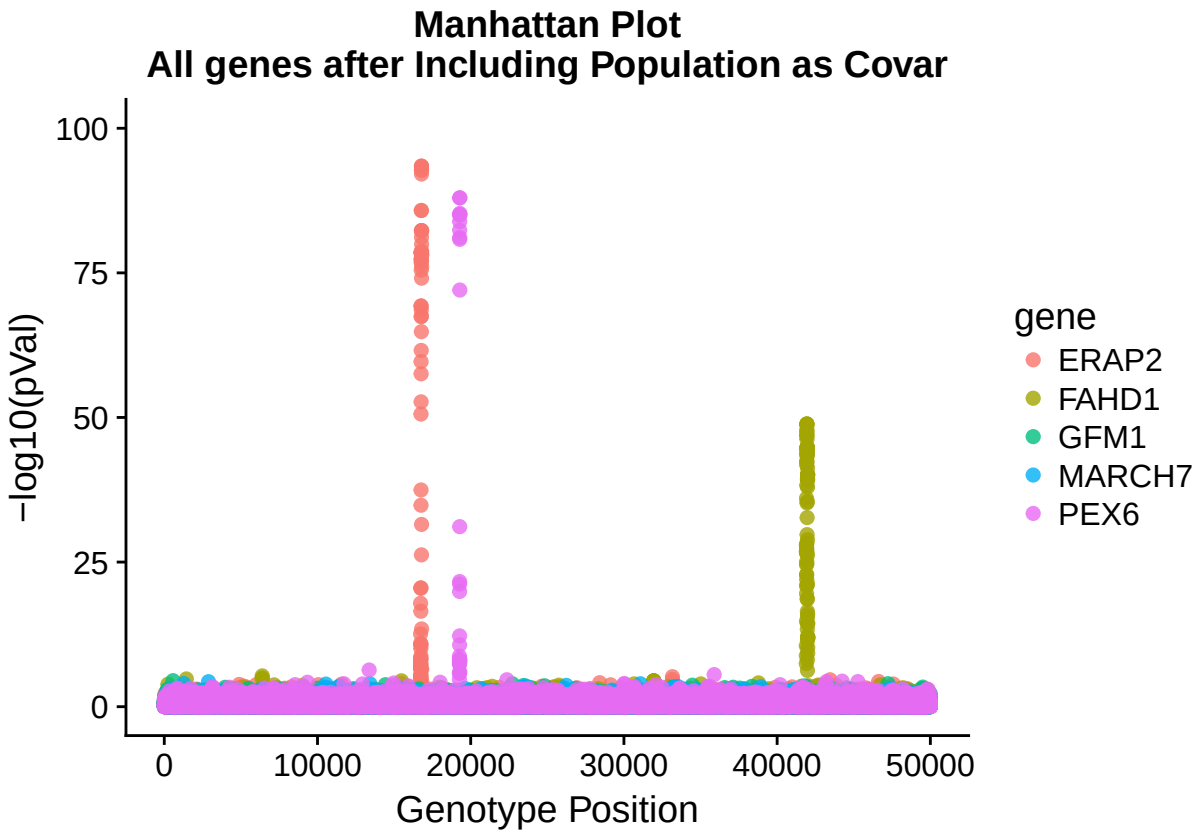
```
## # A tibble: 1 x 2
##   gene  p.value
##   <chr>   <dbl>
## 1 ERAP2 0.00326
```

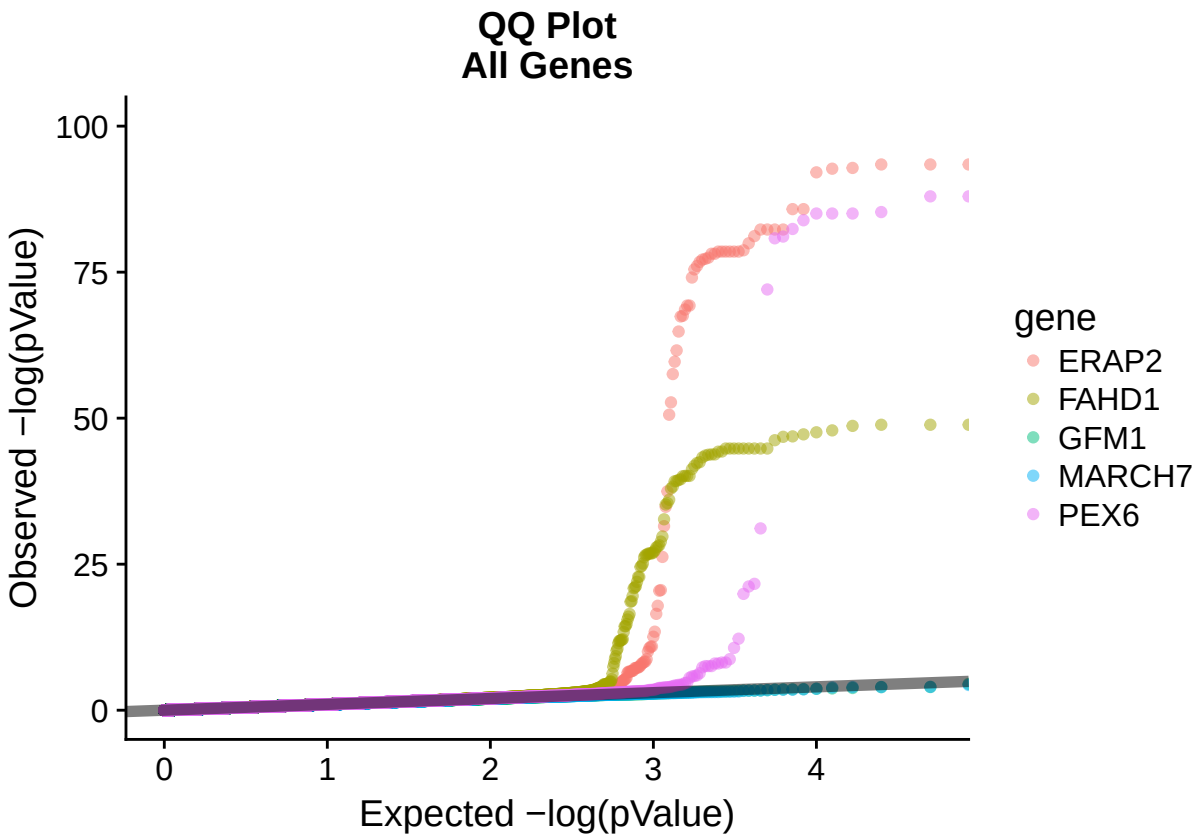
Including Population as covariate:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805
```



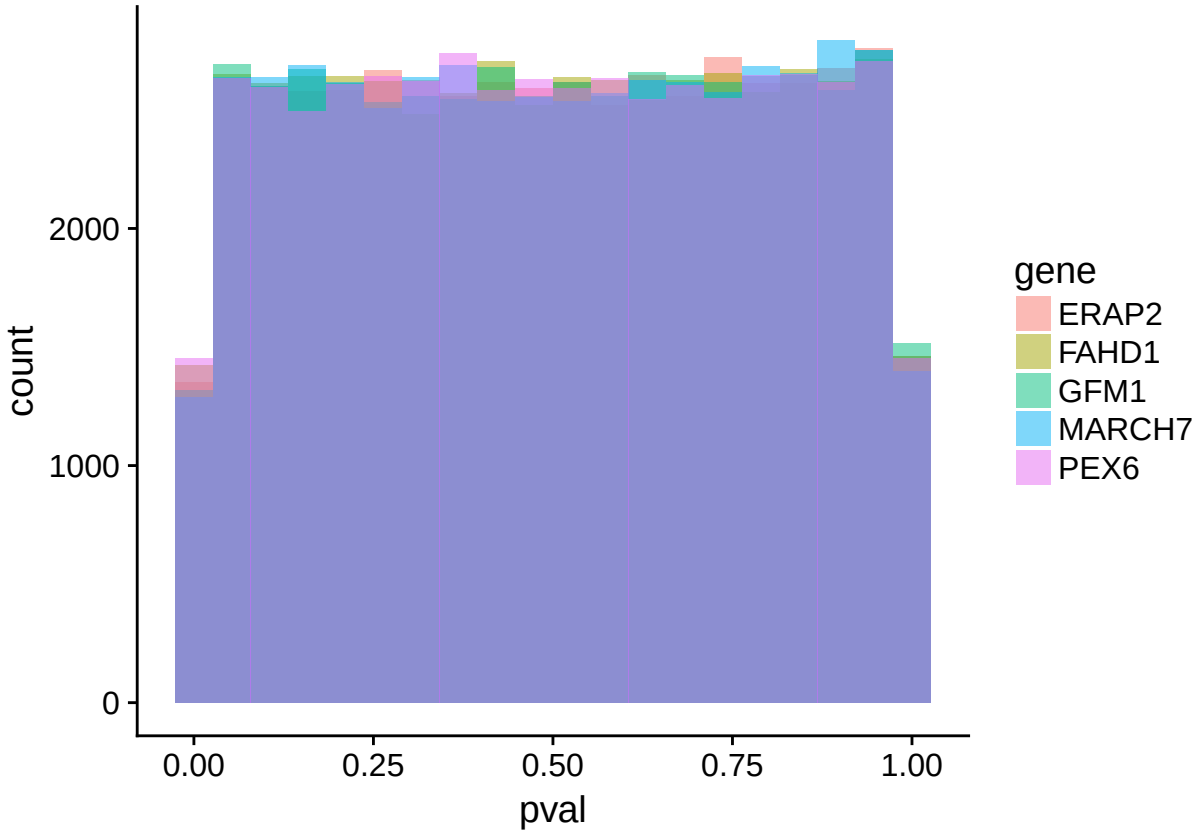

Manhattan Plot
All genes after Including Population as Covar
−log10(pVal)
100
75
50
25
0
0
10000
20000
30000
40000
50000
Genotype Position
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

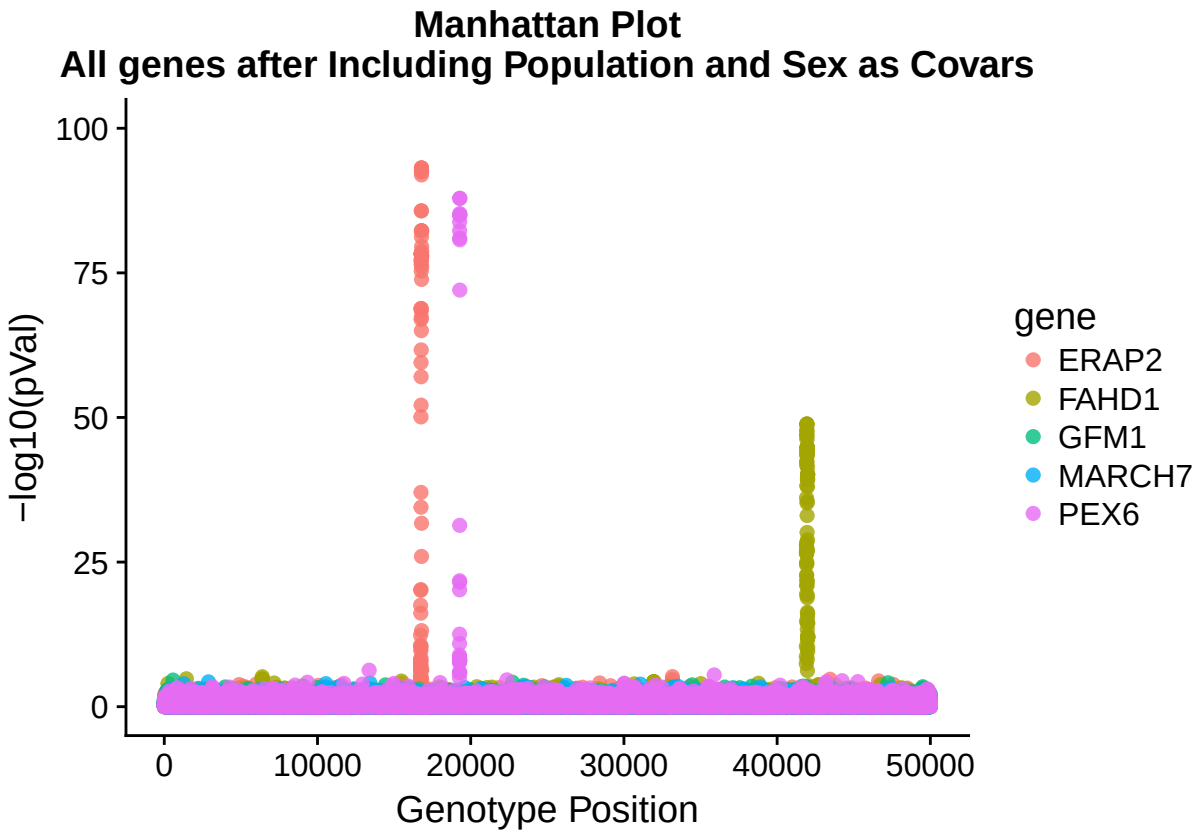


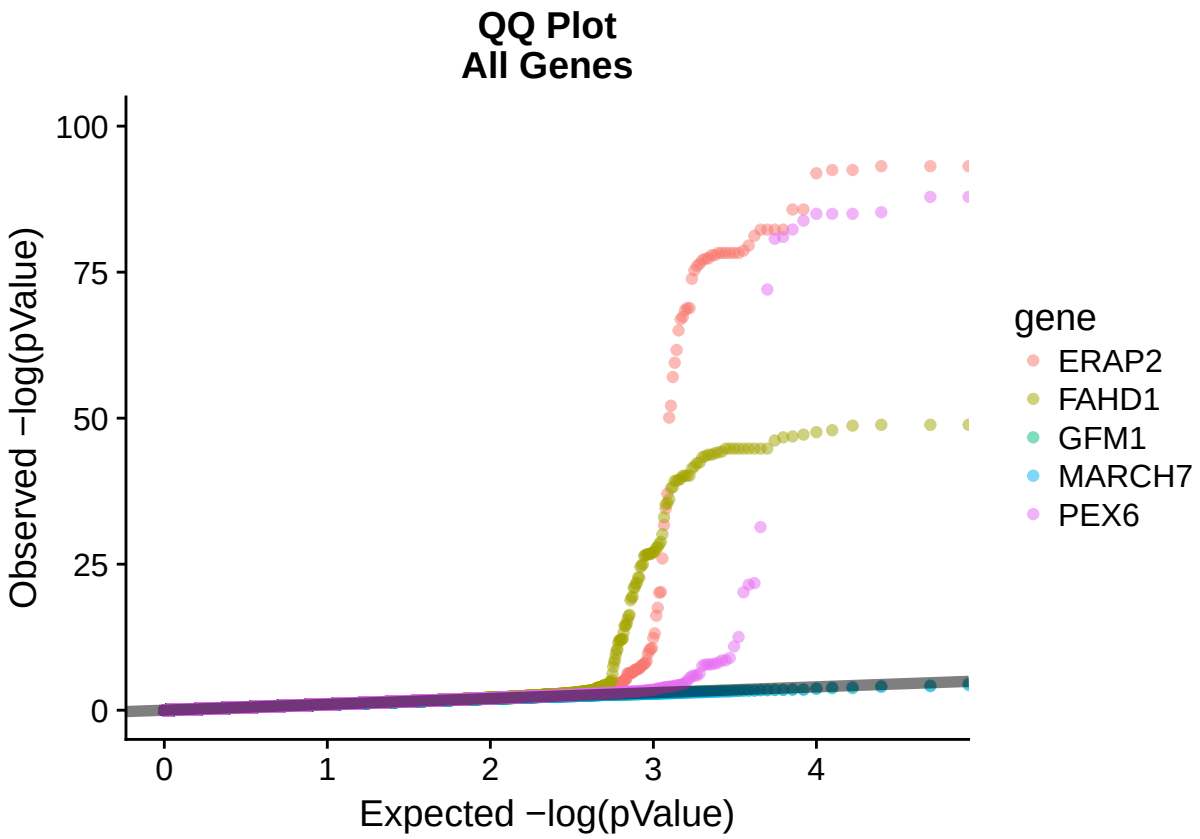
Including Population and Sex as Covariates:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805
```


count
2000
1000
0
0.00
0.25
0.50
0.75
1.00
pval
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

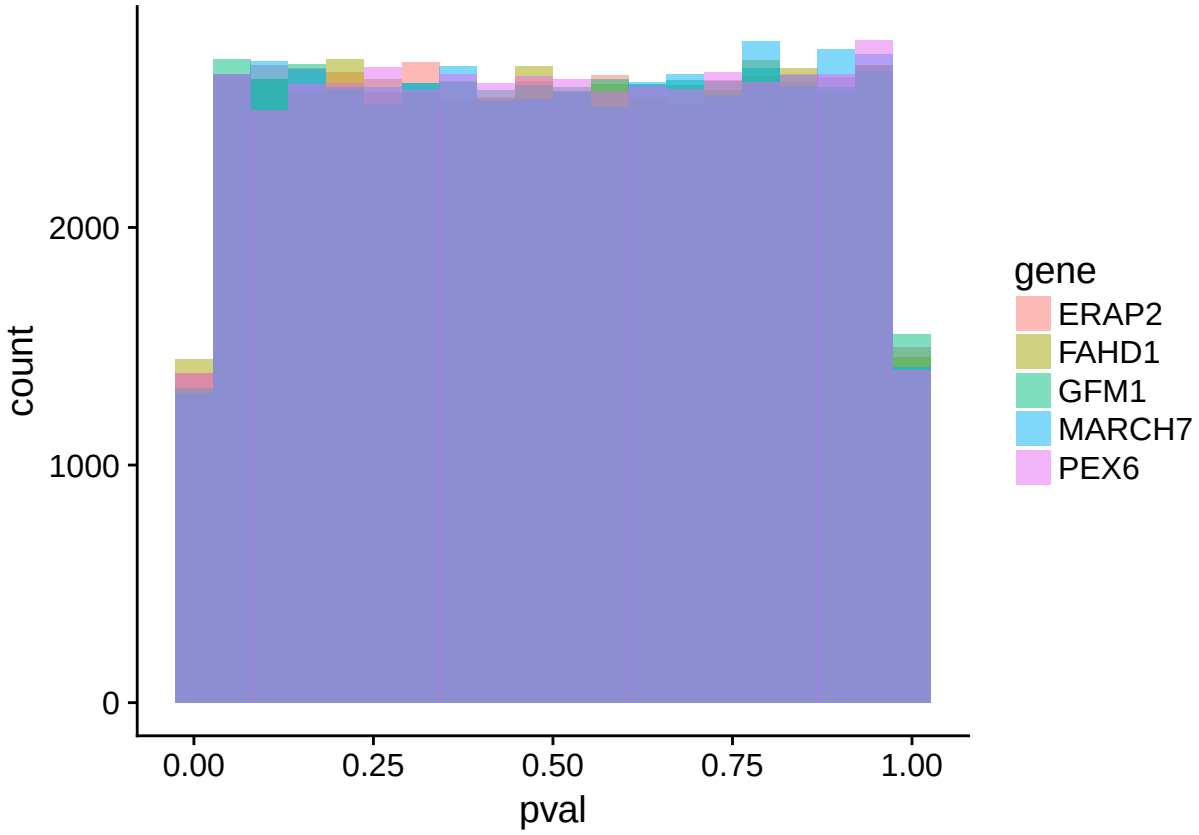

Manhattan Plot
All genes after Including Population and Sex as Covars
100
75
50
25
0
−log10(pVal)
0
10000
20000
30000
40000
50000
Genotype Position
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

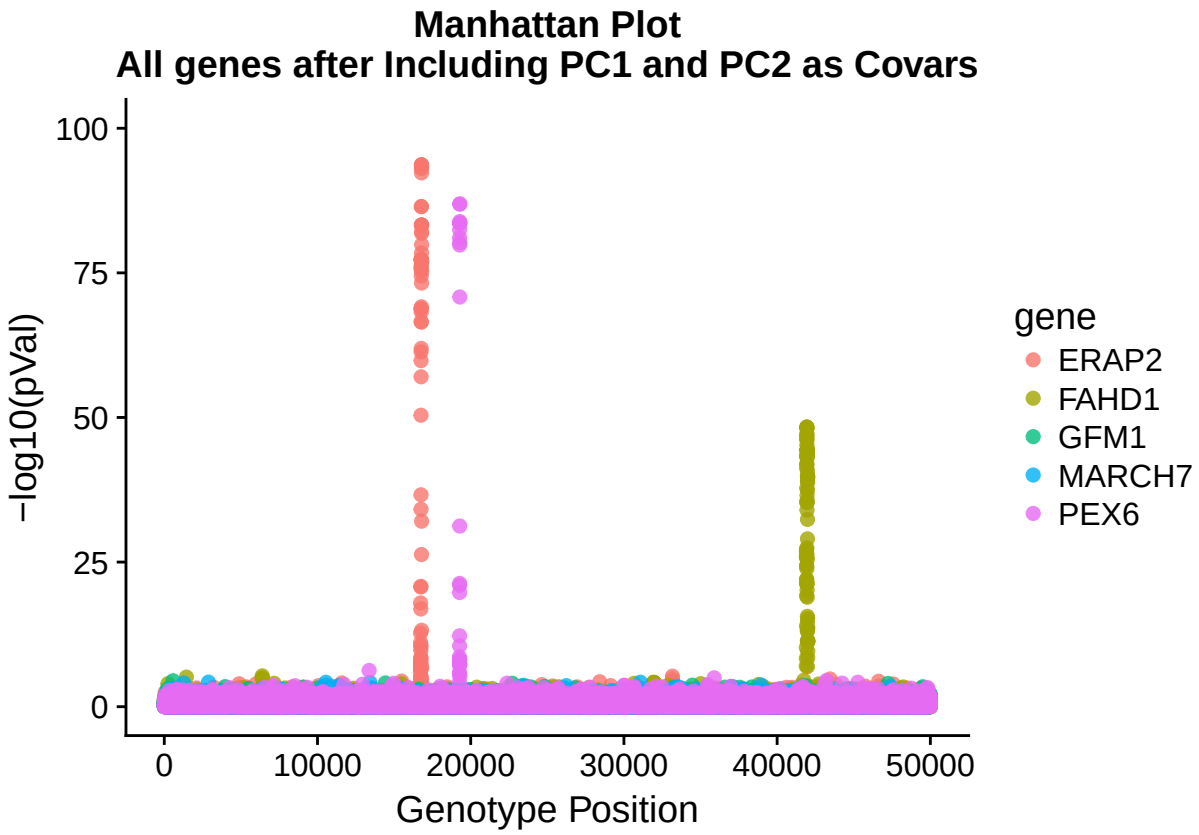


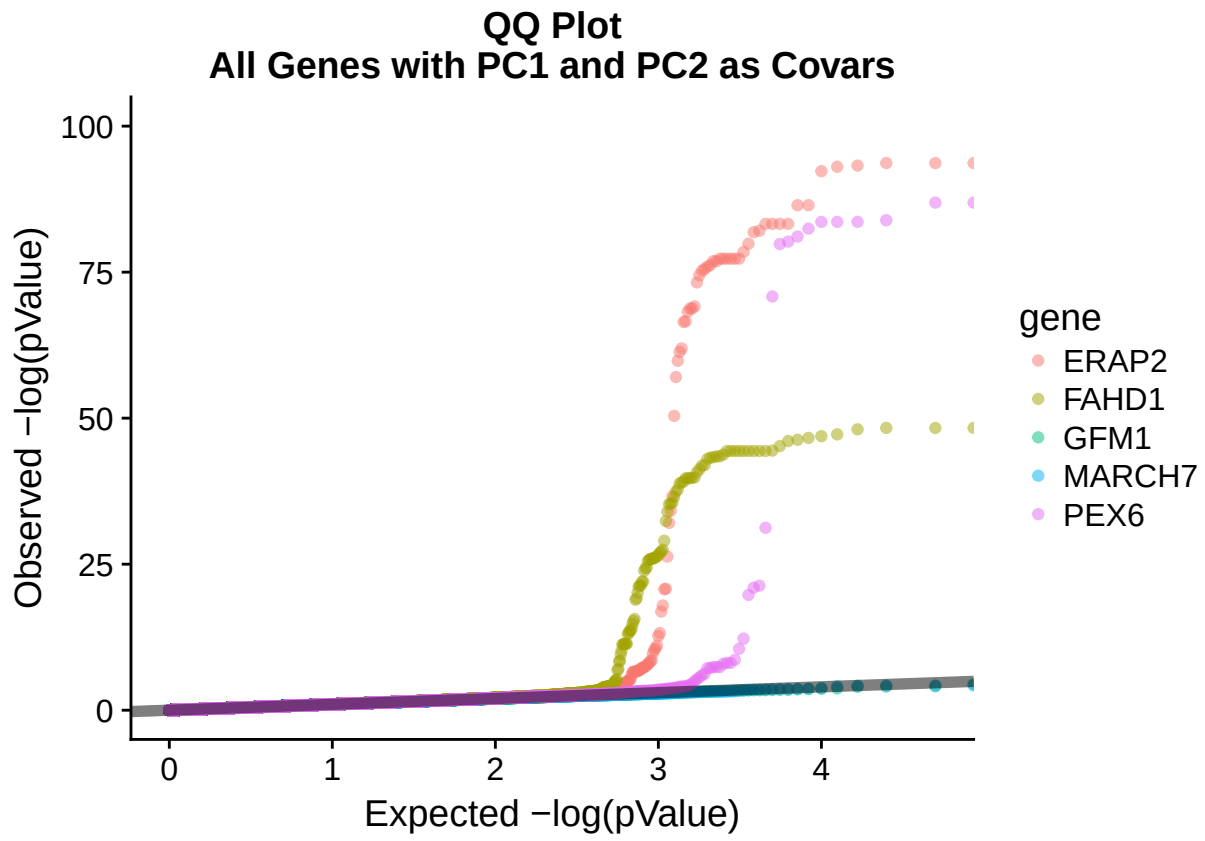
Try PC1 and PC2 as covars:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805
```


count
2000
1000
0
0.00
0.25
0.50
0.75
1.00
pval
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

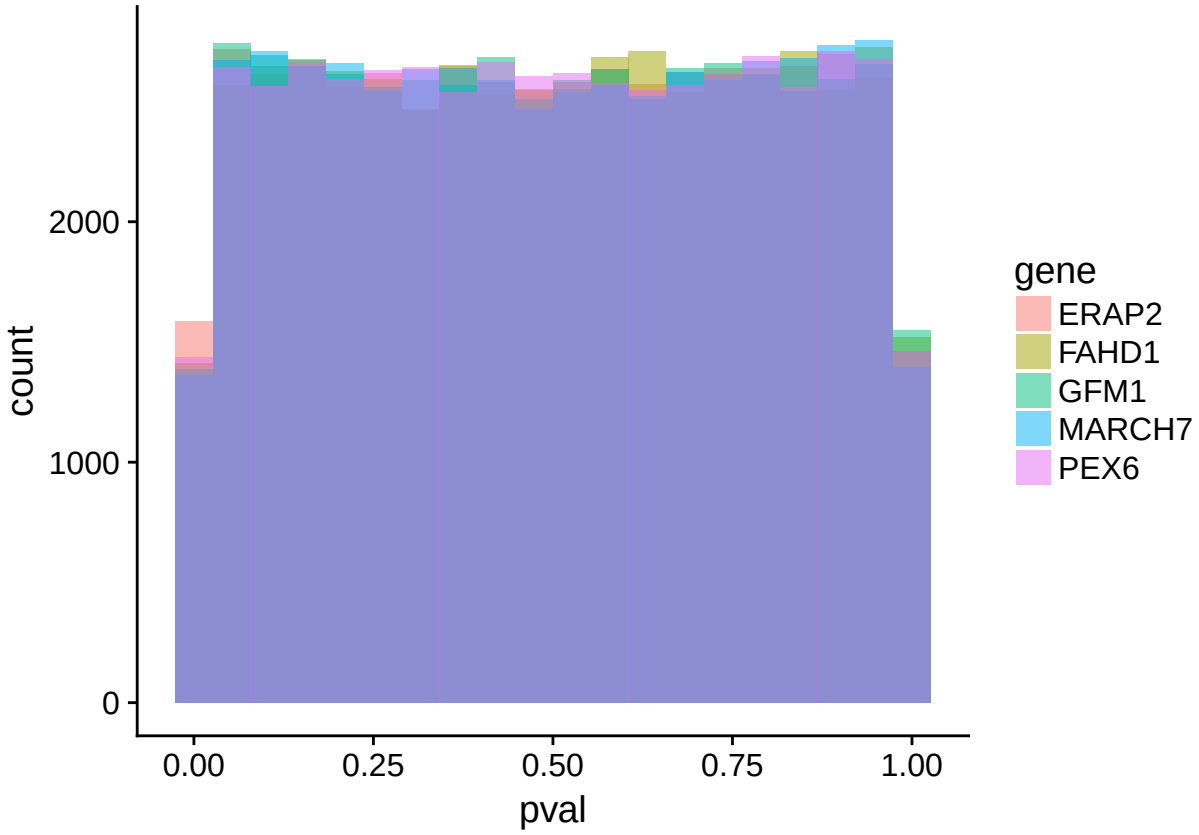

Manhattan Plot
All genes after Including PC1 and PC2 as Covars
−log10(pVal)
100
75
50
25
0
0
10000
20000
30000
40000
50000
Genotype Position
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

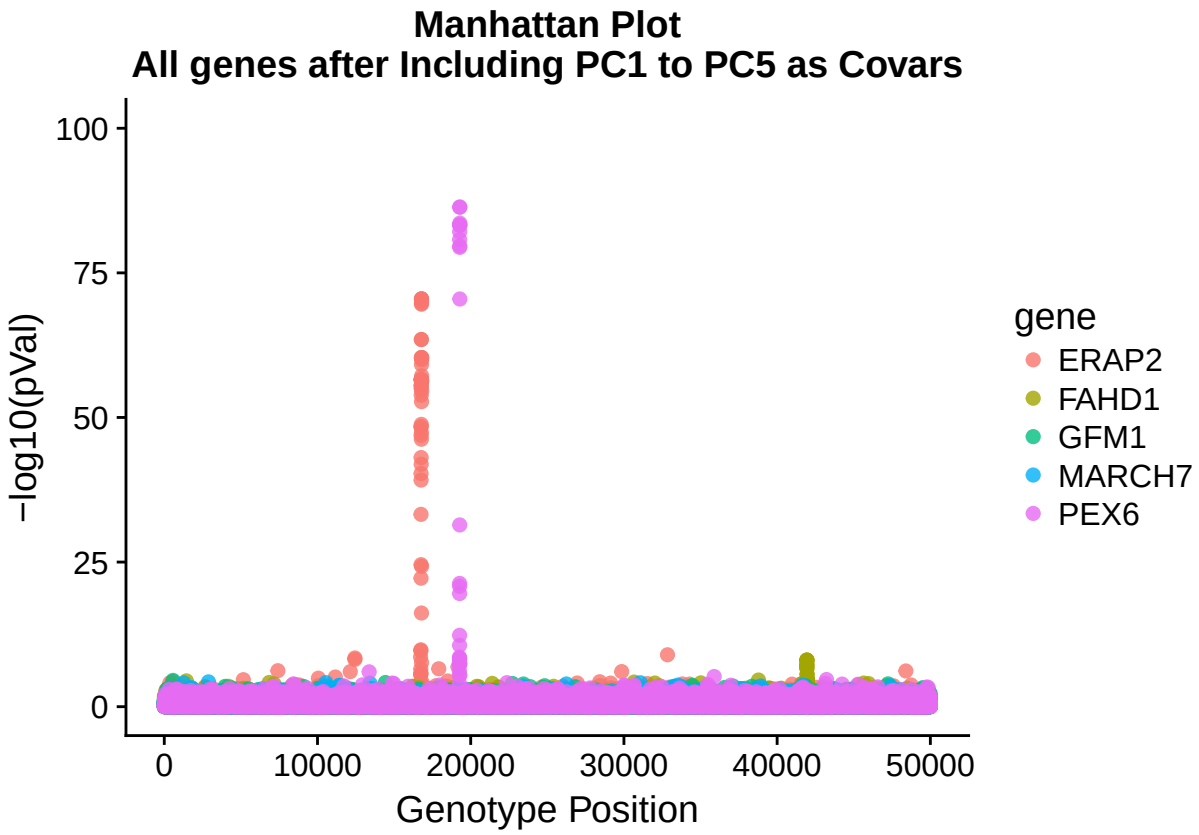


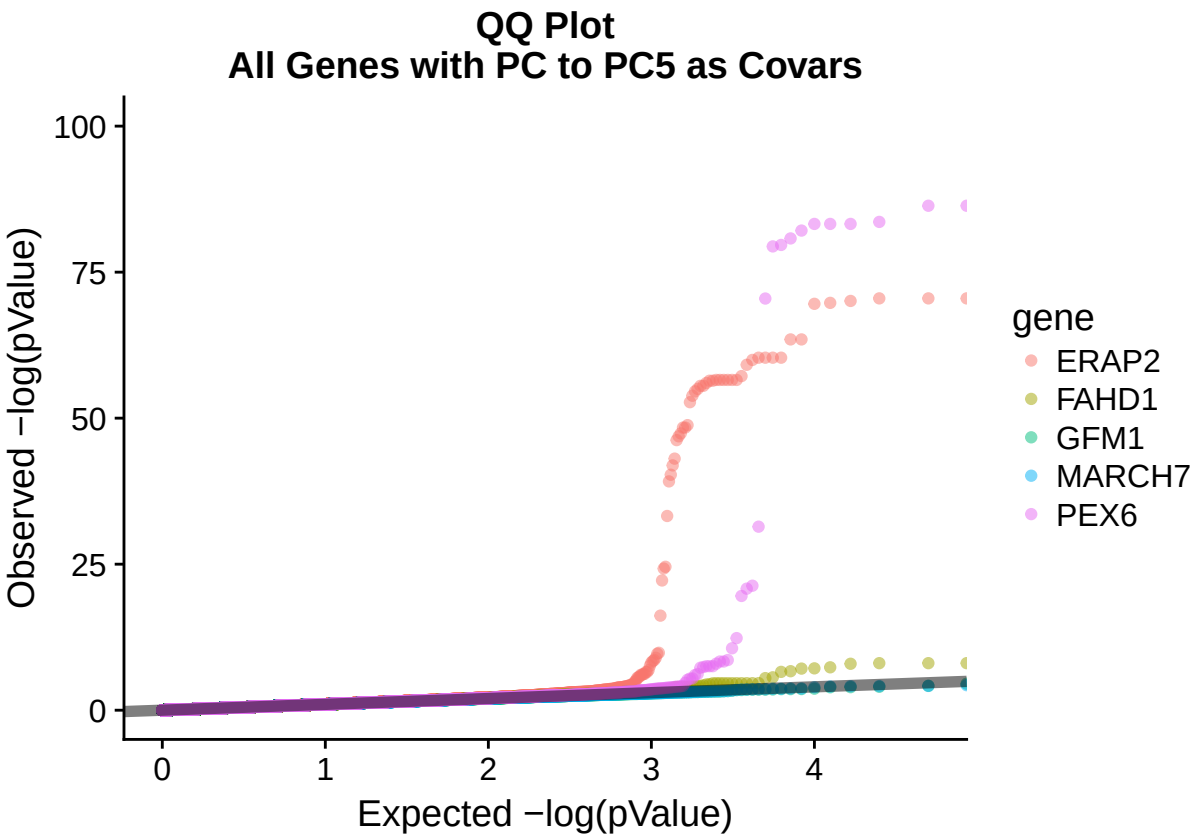
Even more PCs

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16725   16745   16767   16765   16785   16805
```

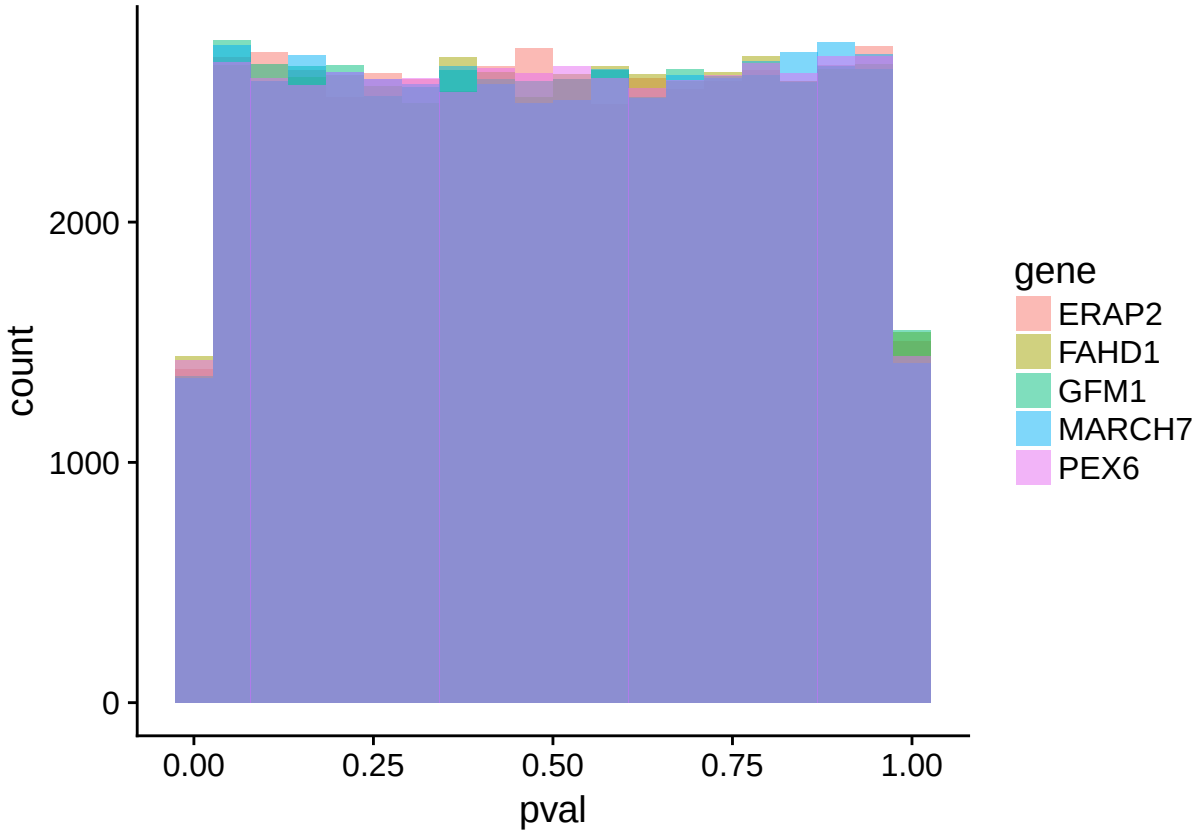

count
2000
1000
0
0.00
0.25
0.50
0.75
1.00
pval
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

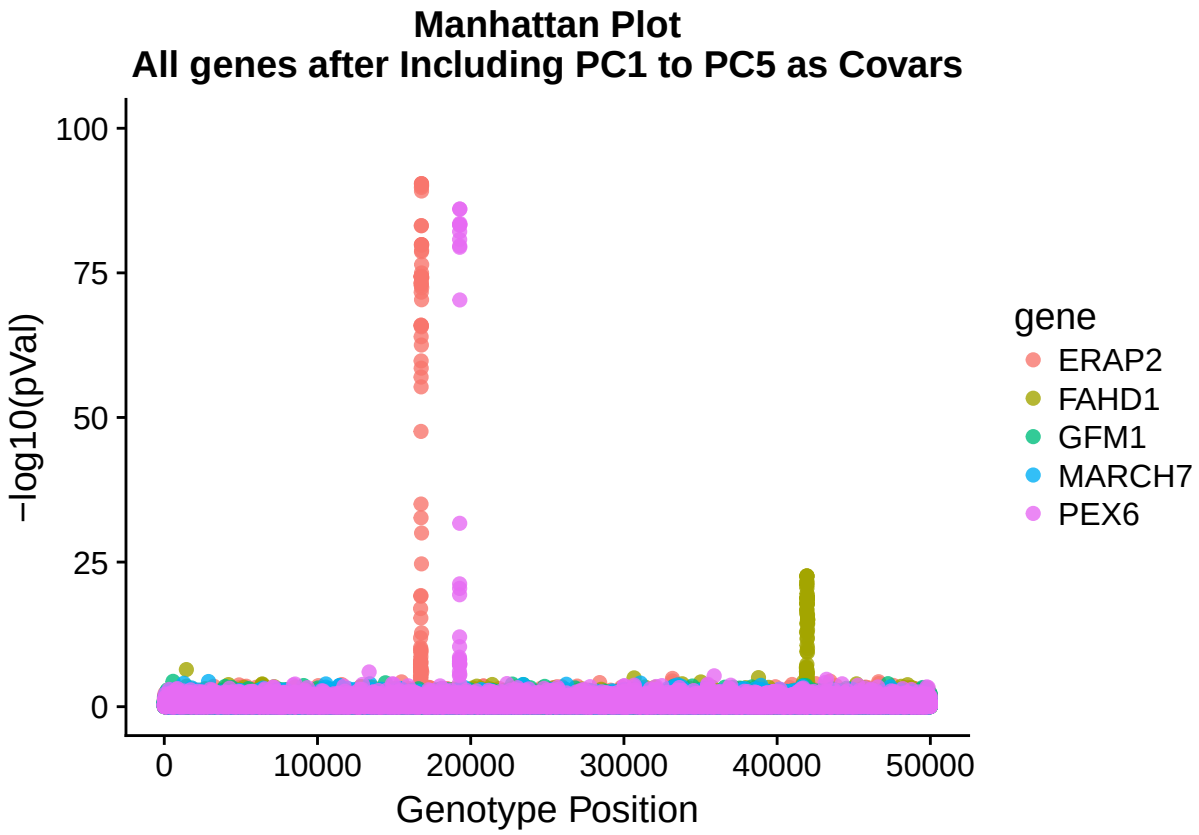

Manhattan Plot
All genes after Including PC1 to PC5 as Covars
−log10(pVal)
100
75
50
25
0
0
10000
20000
30000
40000
50000
Genotype Position
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

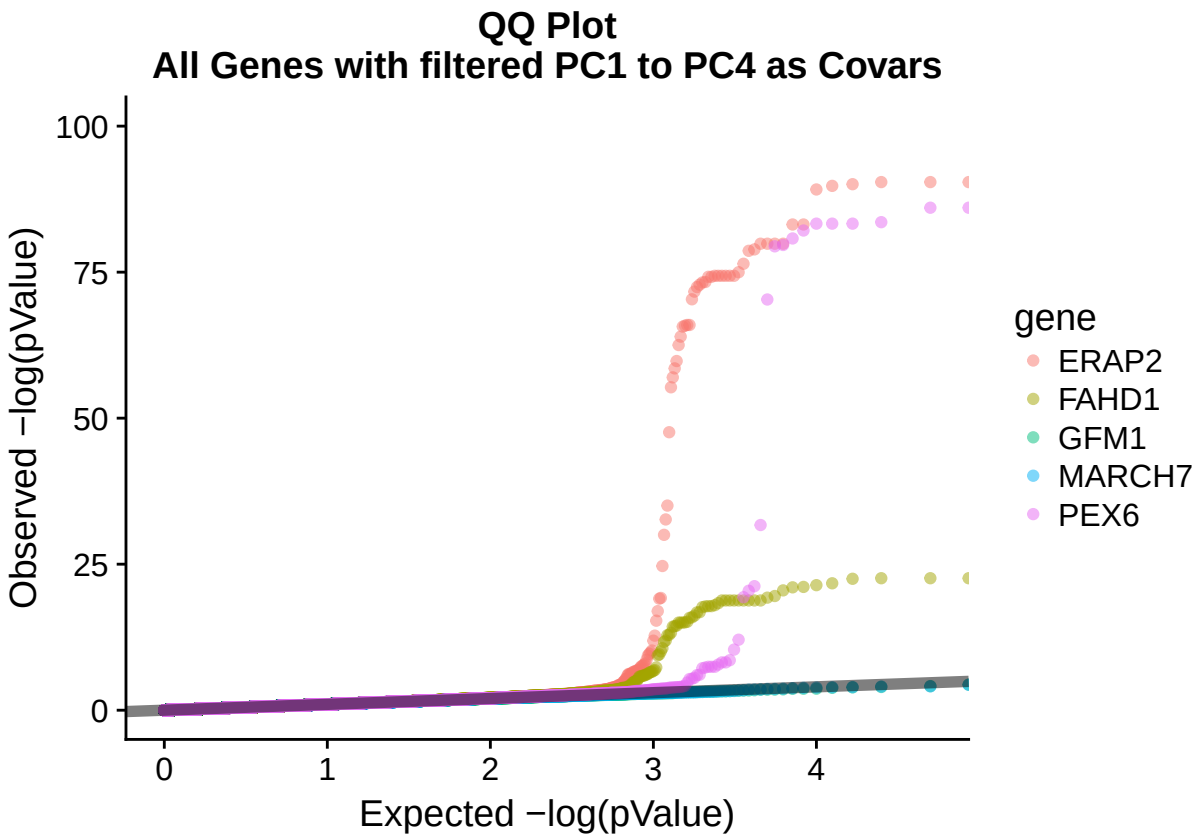


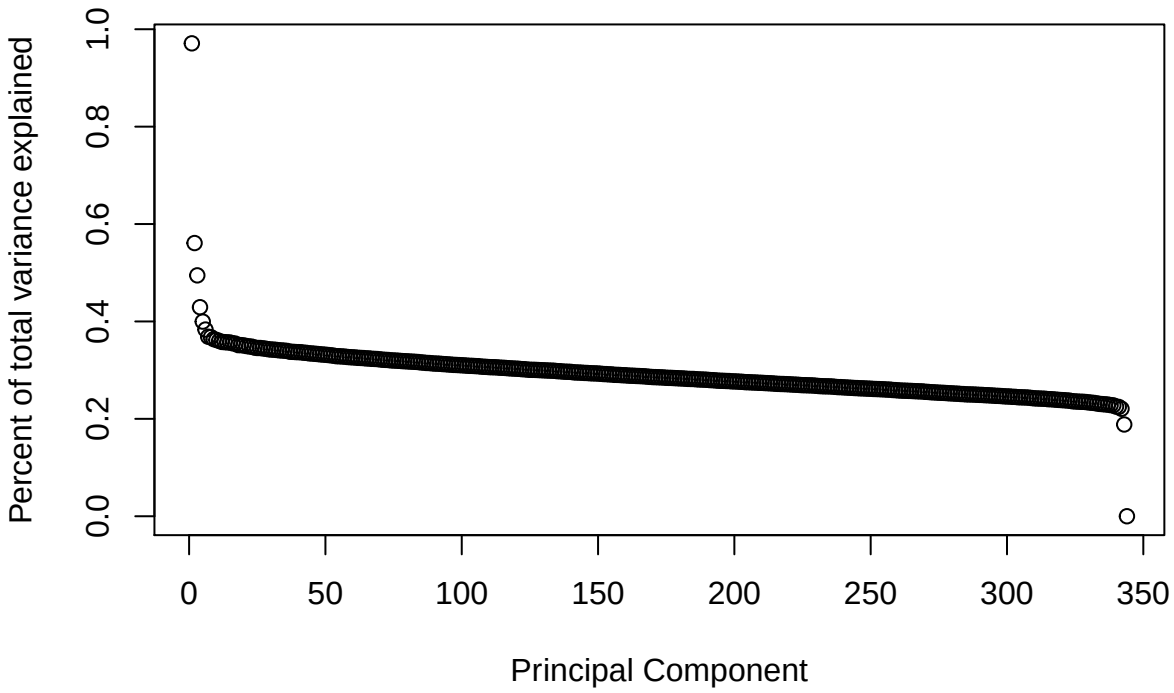

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16725   16744   16767   16765   16784   16803
```

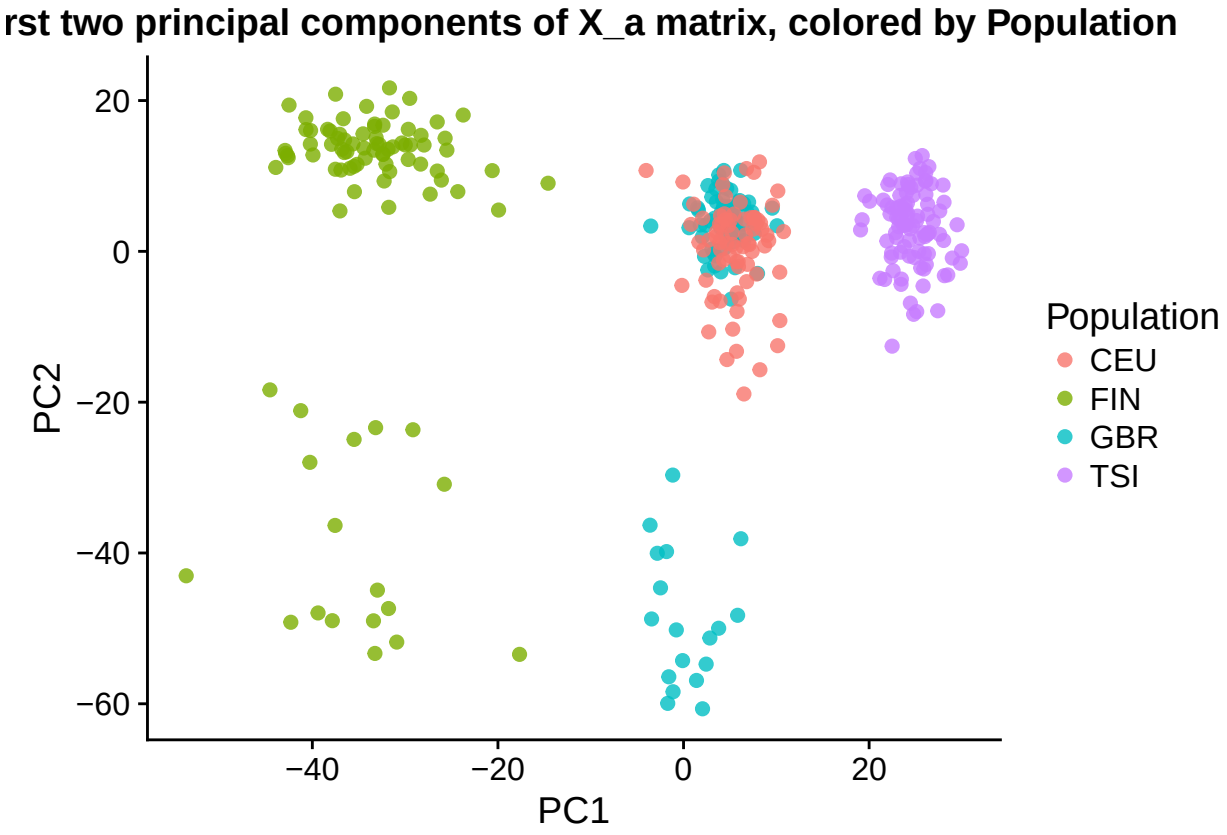
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16725   16745   16767   16765   16785   16805
```

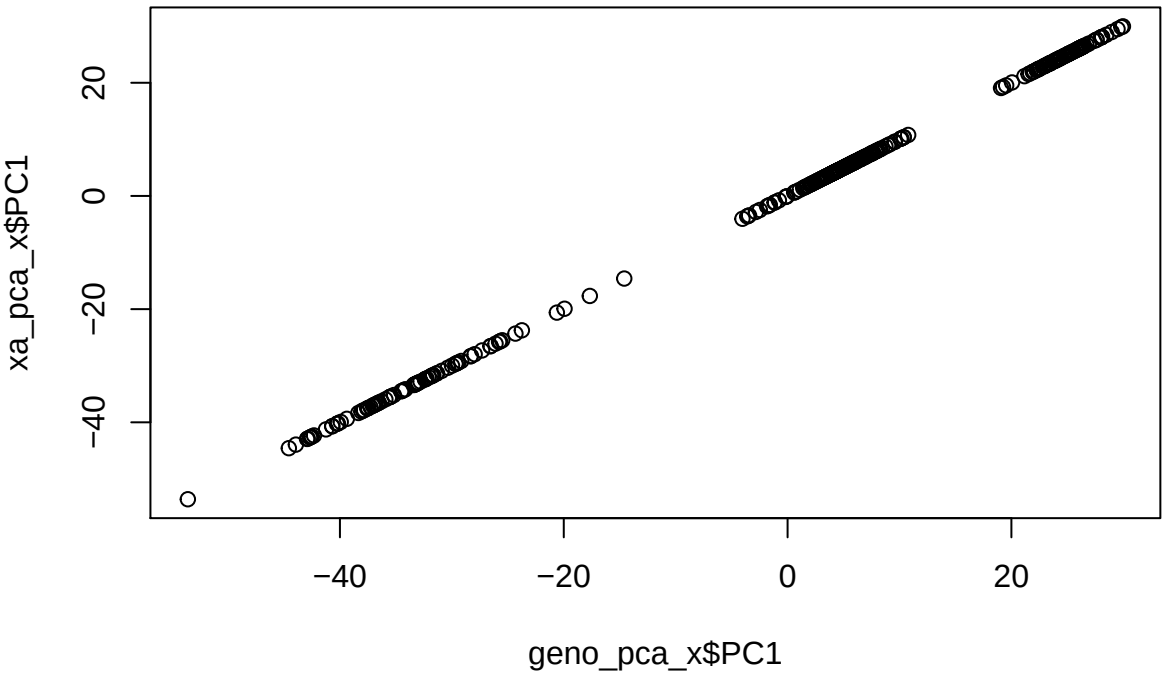

count
2000
1000
0
0.00
0.25
0.50
0.75
1.00
pval
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

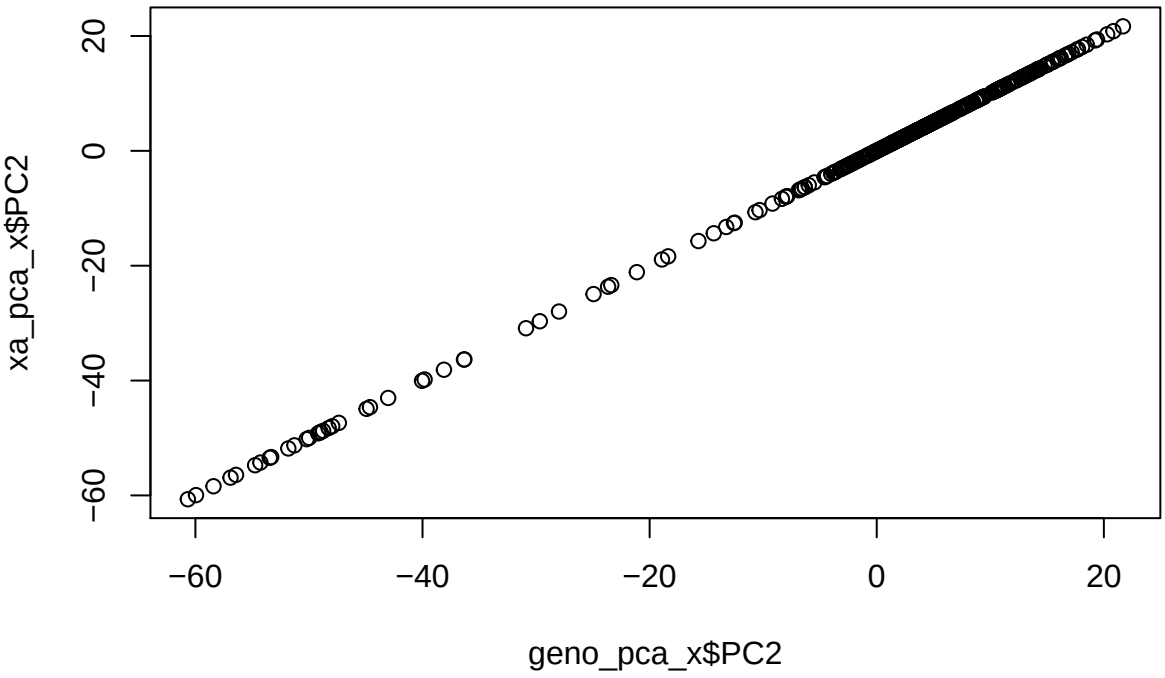

Manhattan Plot
All genes after Including PC1 to PC5 as Covars
−log10(pVal)
100
75
50
25
0
0
10000
20000
30000
40000
50000
Genotype Position
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

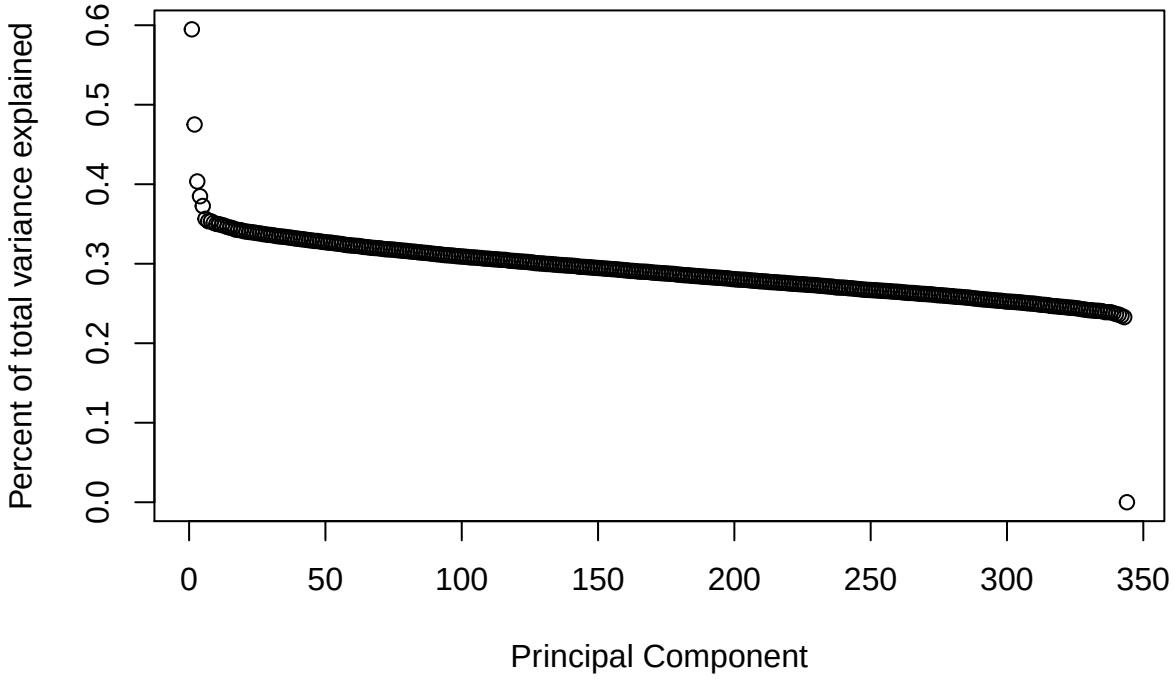

QQ Plot
All Genes with filtered PC1 to PC4 as Covars
Observed −log(pValue)
Expected −log(pValue)
100
75
50
25
0
0
1
2
3
4
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

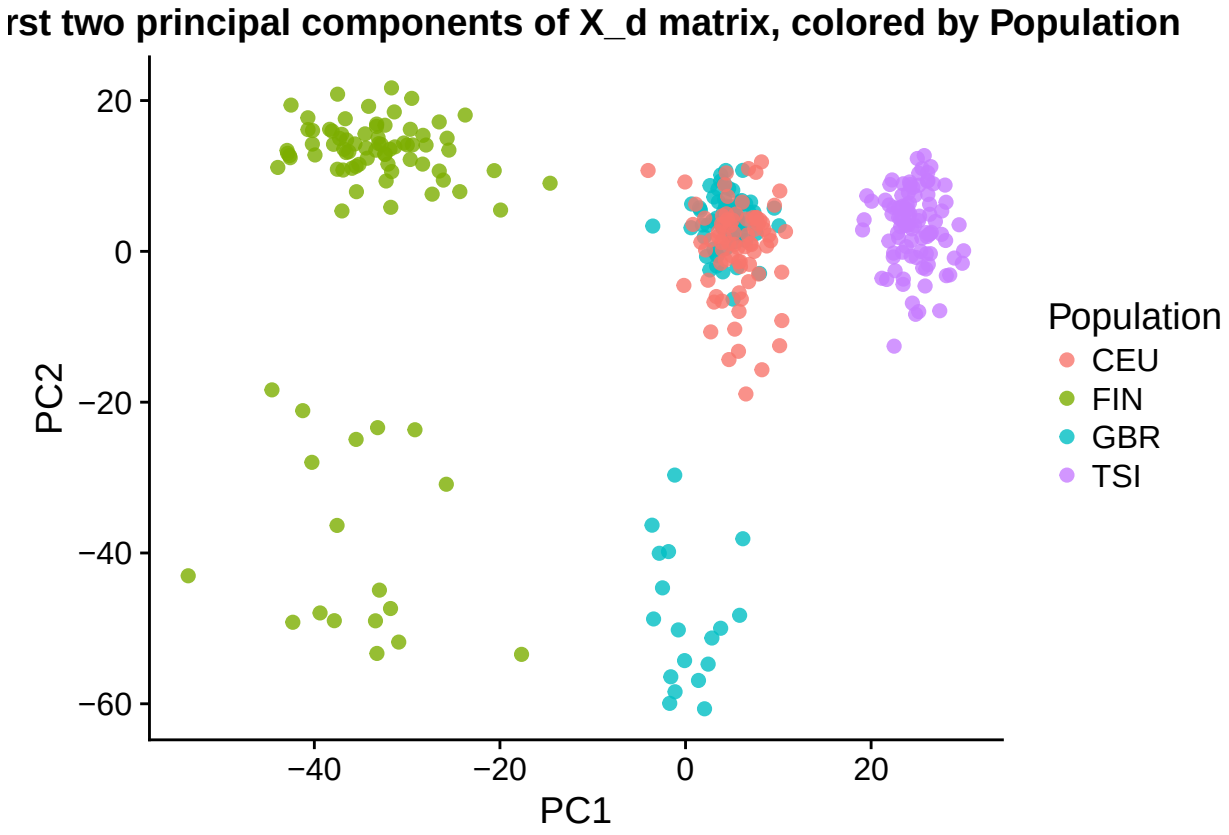

Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Principal Component

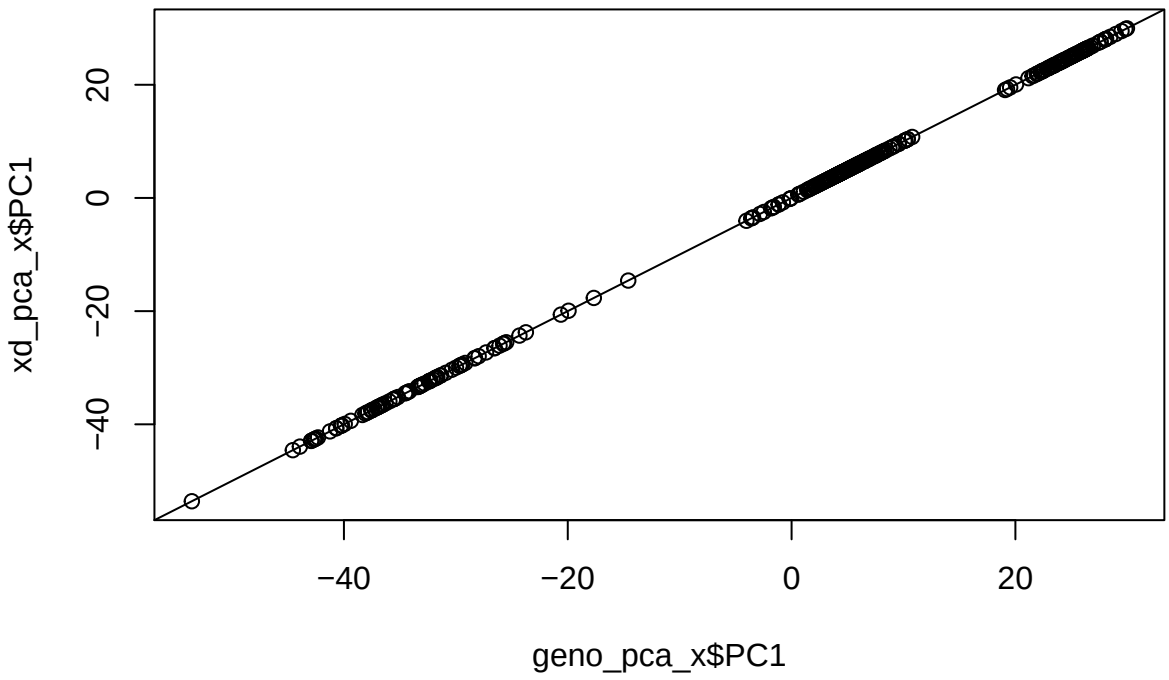

rst two principal components of X_a matrix, colored by Population
20
0
−20
−40
−60
PC2
−40
−20
0
20
PC1
Population
CEU
FIN
GBR
TSI


xa_pca_x$PC1
20
0
−20
−40
−40
−20
0
20
geno_pca_x$PC1

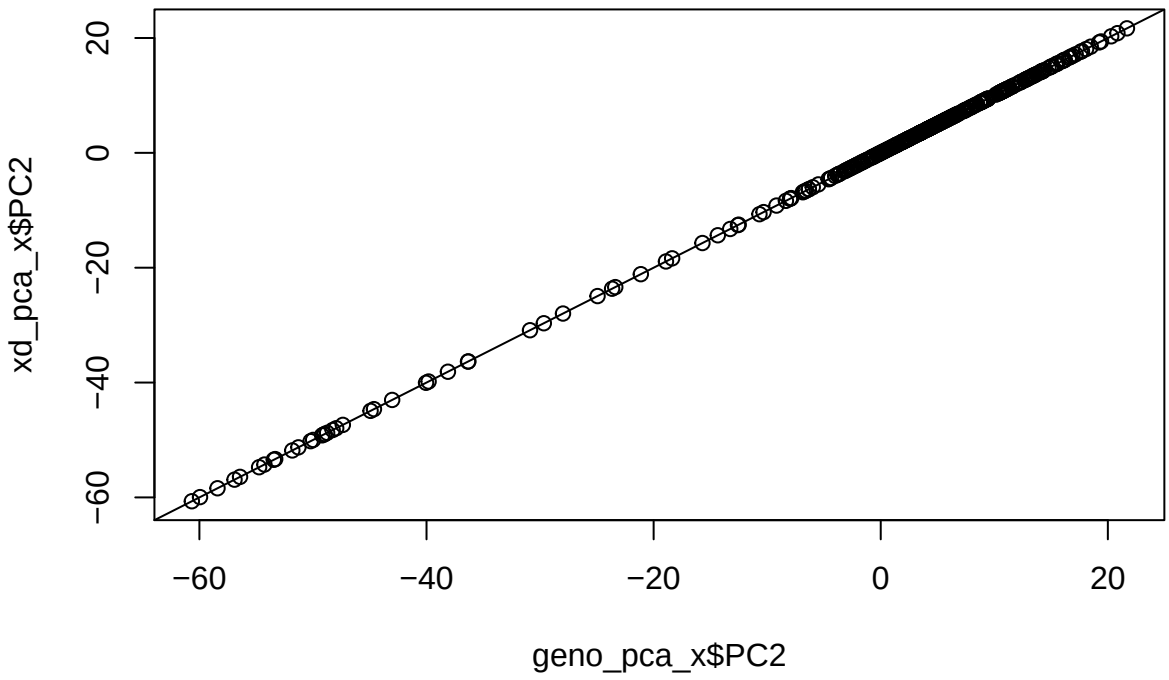

xa_pca_x$PC2
20
0
−20
−40
−60
−60
−40
−20
0
20
geno_pca_x$PC2


Percent of total variance explained
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
350
Principal Component


rst two principal components of X_d matrix, colored by Population
PC2
PC1
20
0
−20
−40
−60
−40
−20
0
20
Population
CEU
FIN
GBR
TSI




```
integer(0)
```



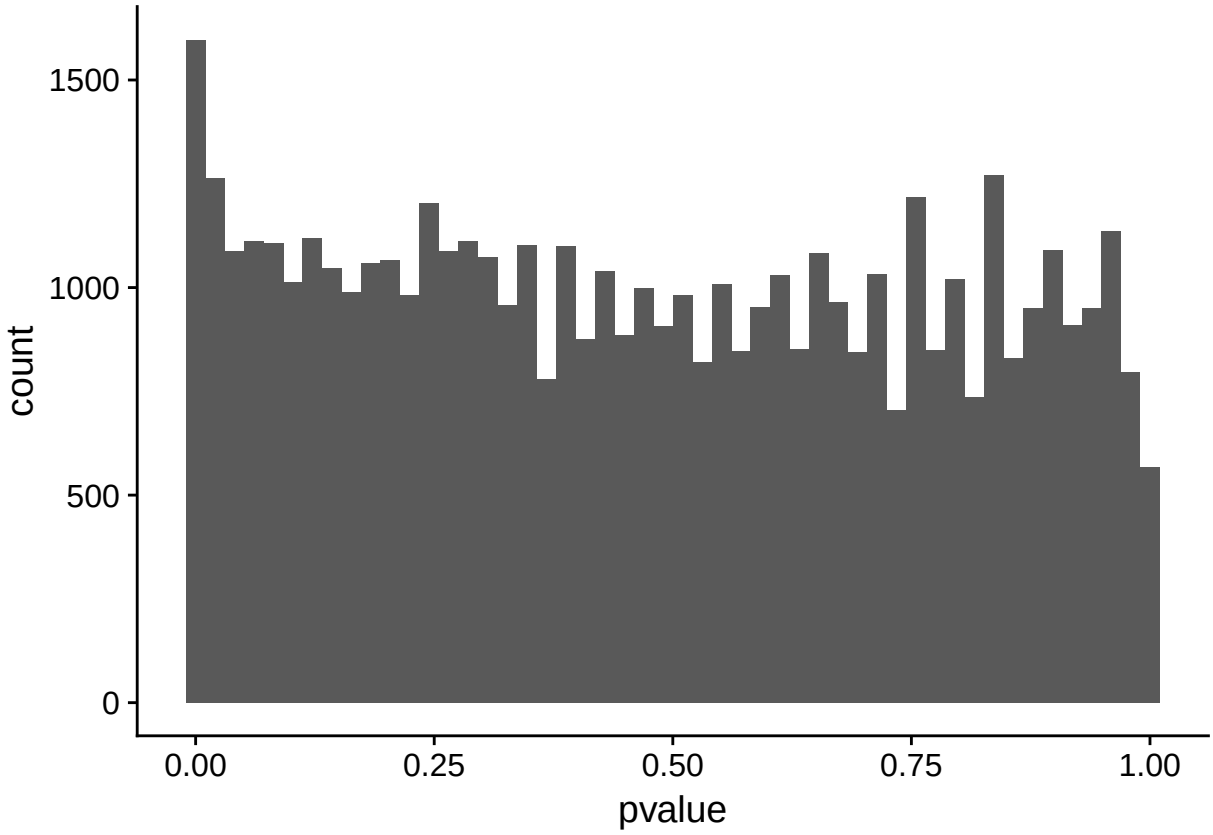

```
integer(0)
```

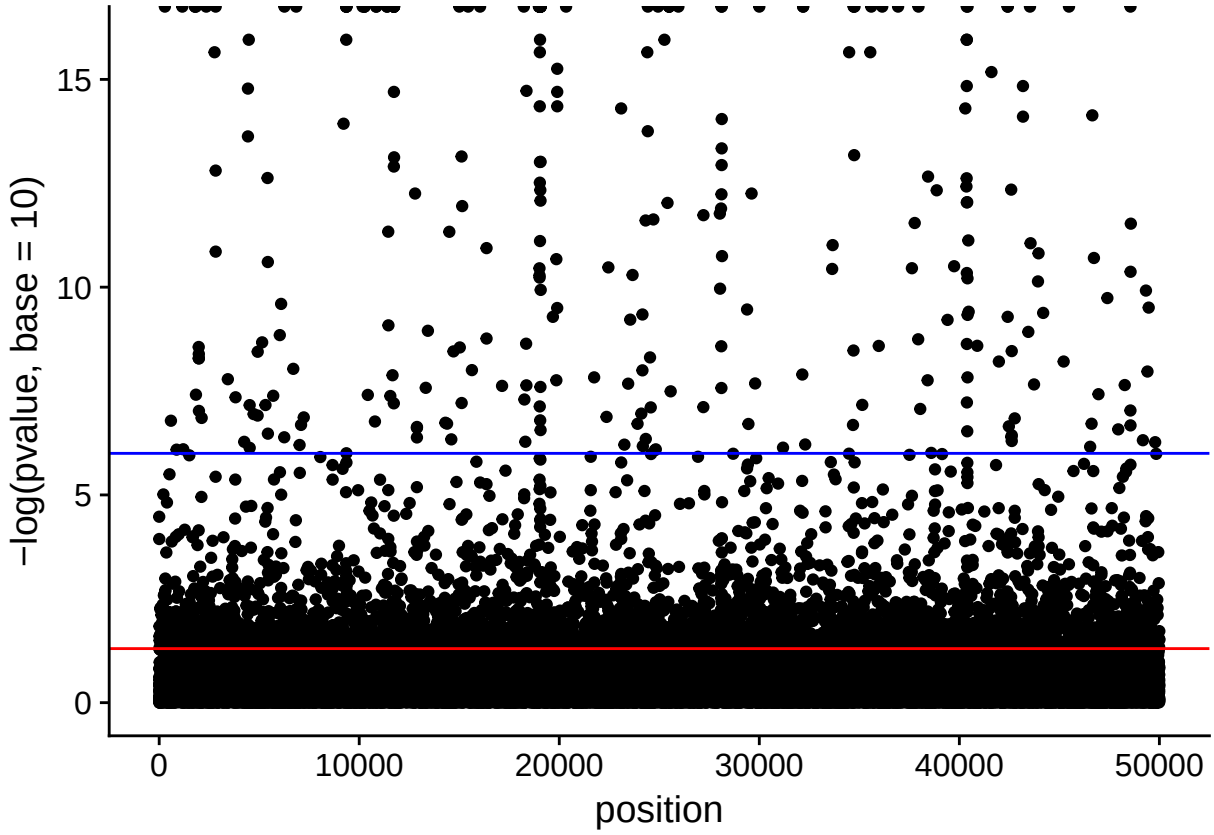
```
Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
Chi2 =  15.45815 DF =  1 p-value =  8.435235e-05 D =  -8.499273 f =  0.2236694
```

```
[1] 1.381394
```

```
[1] 0.2398638
```

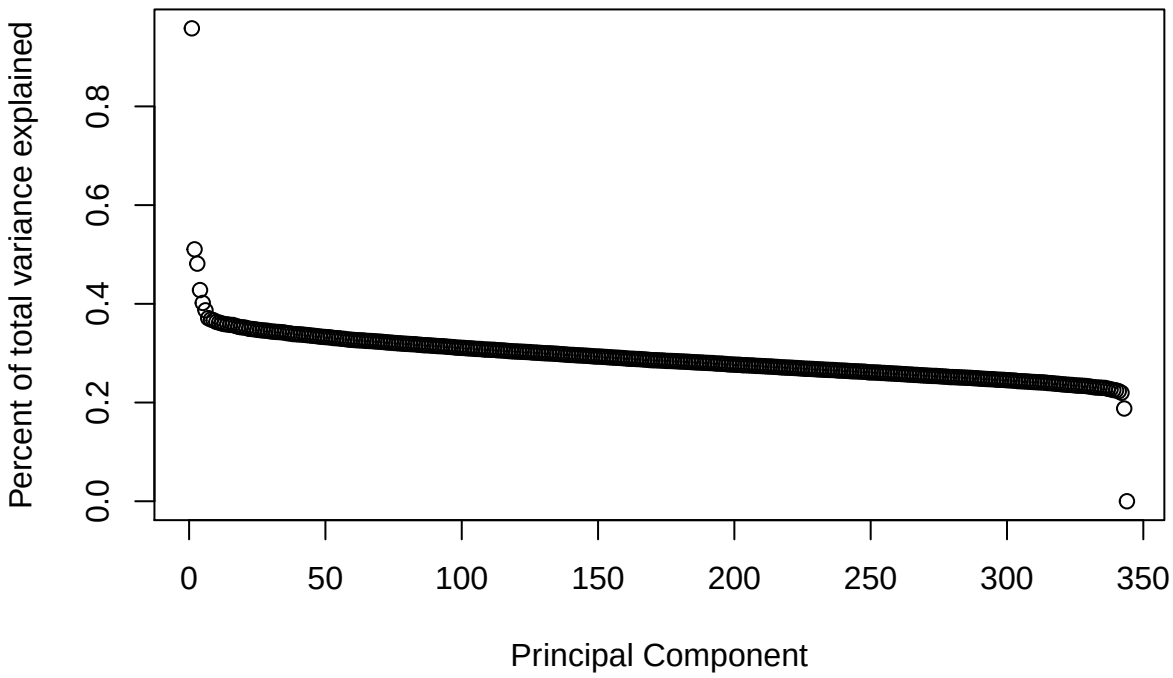
```
Chi-square test for Hardy-Weinberg equilibrium (autosomal)
Chi2 =  1.381394 DF =  1 p-value =  0.2398638 D =  1.221657 f =  -0.0633694
```

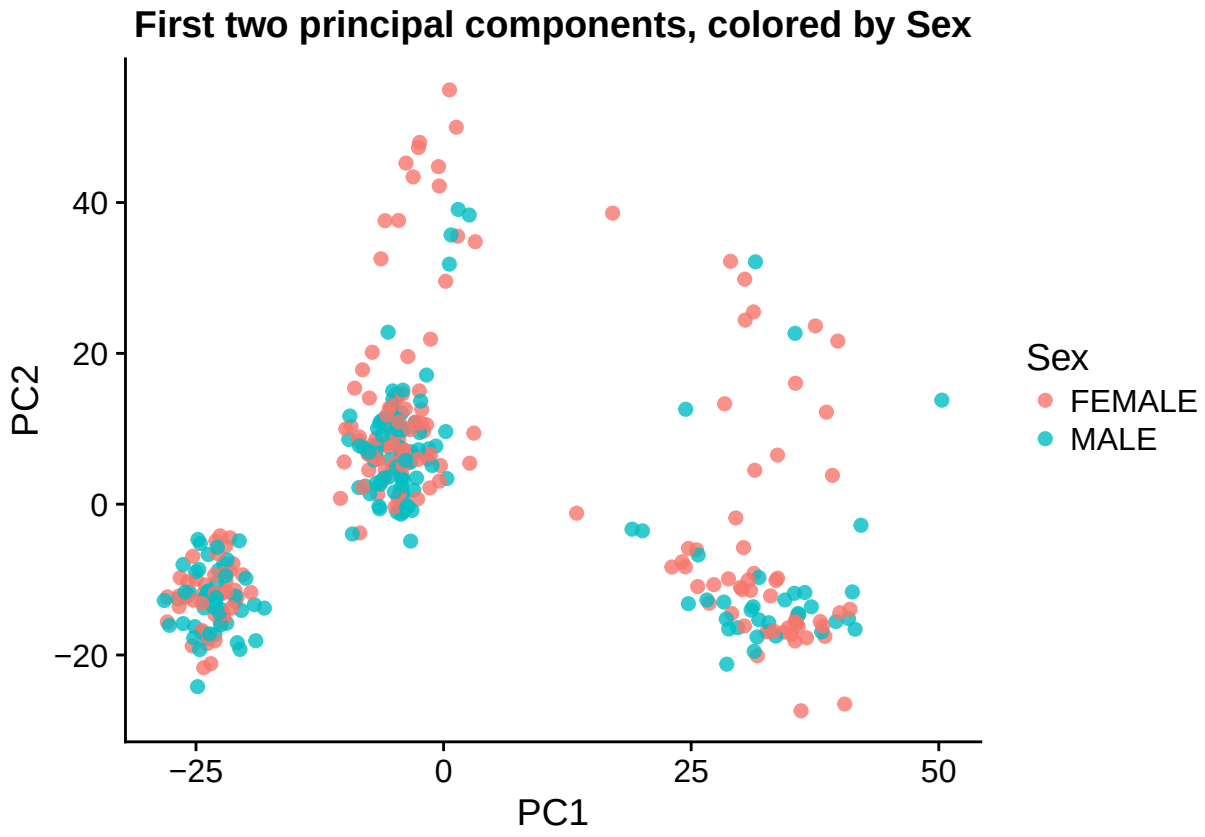

1500
1000
500
0
count
0.00
0.25
0.50
0.75
1.00
pvalue

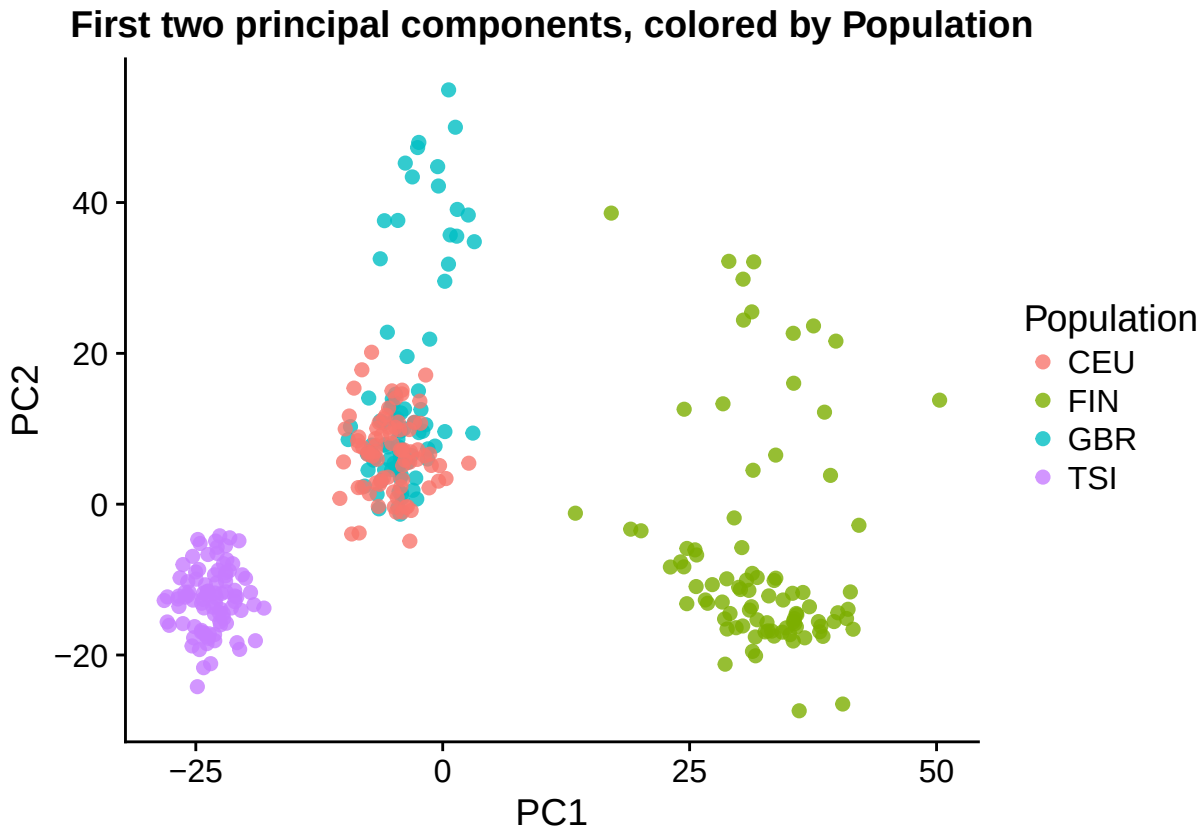

15
10
5
0
−log(pvalue, base = 10)
0
10000
20000
30000
40000
50000
position

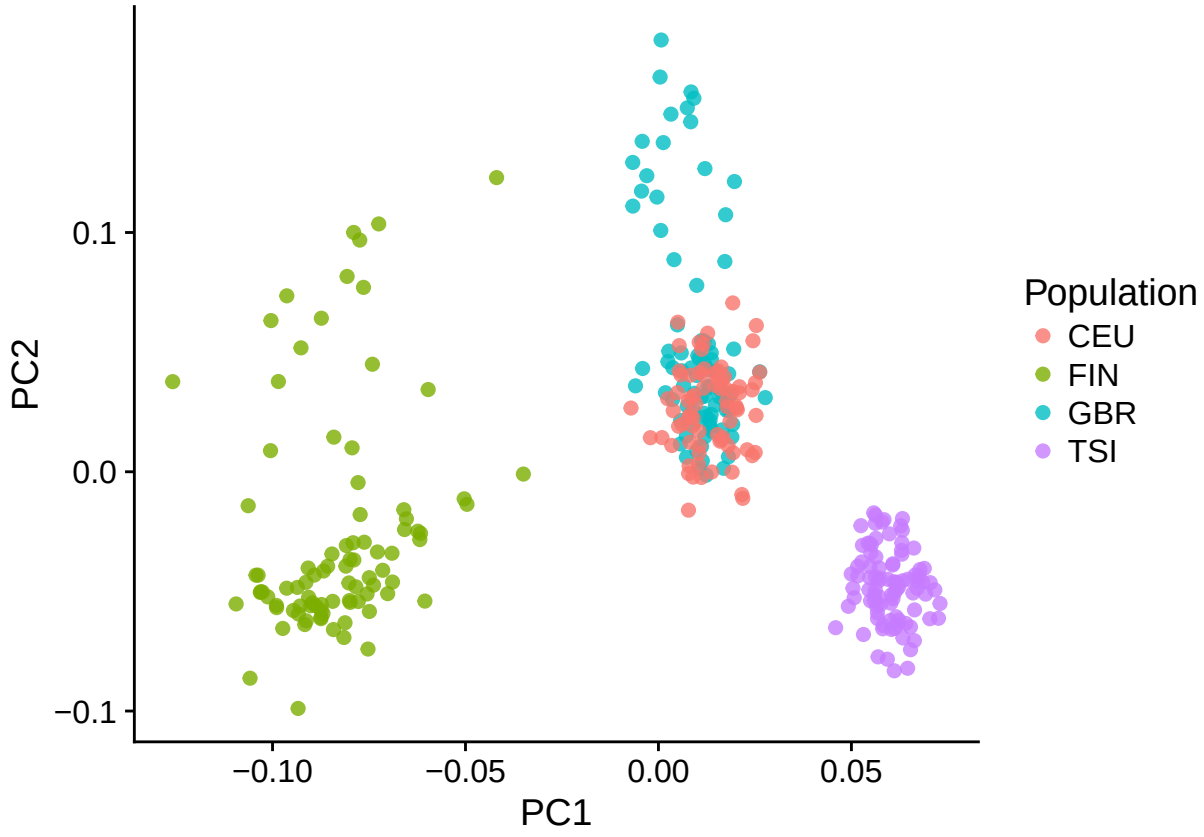
```
[1] 296
```

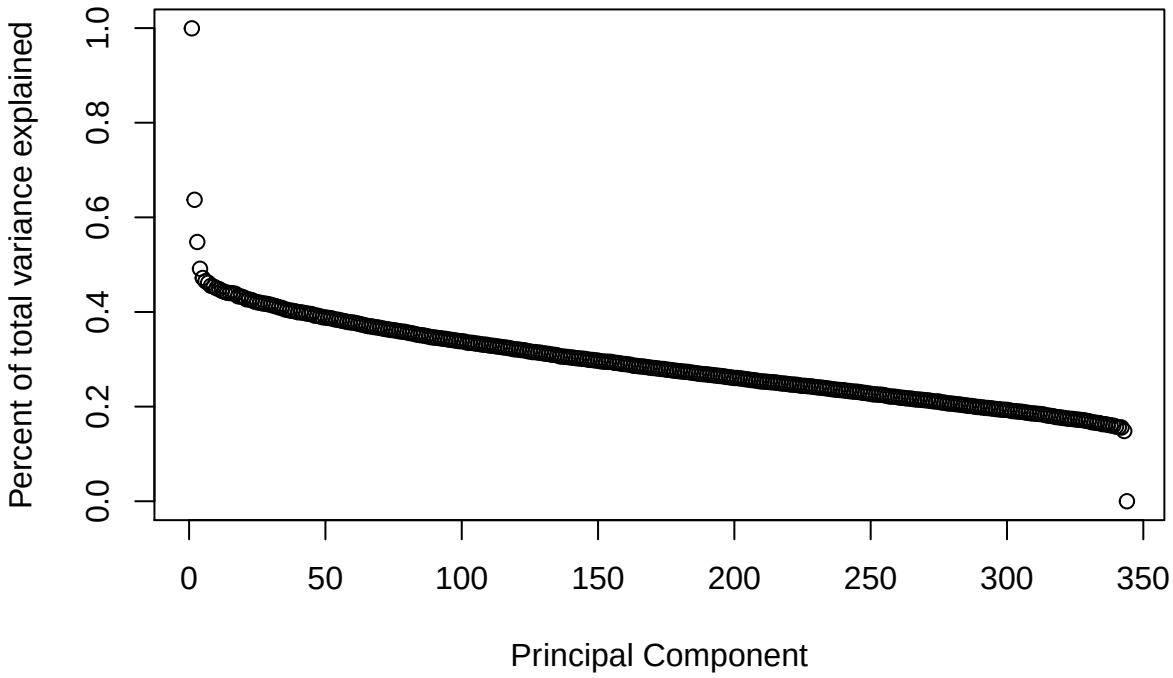
```
[1] 3908
```

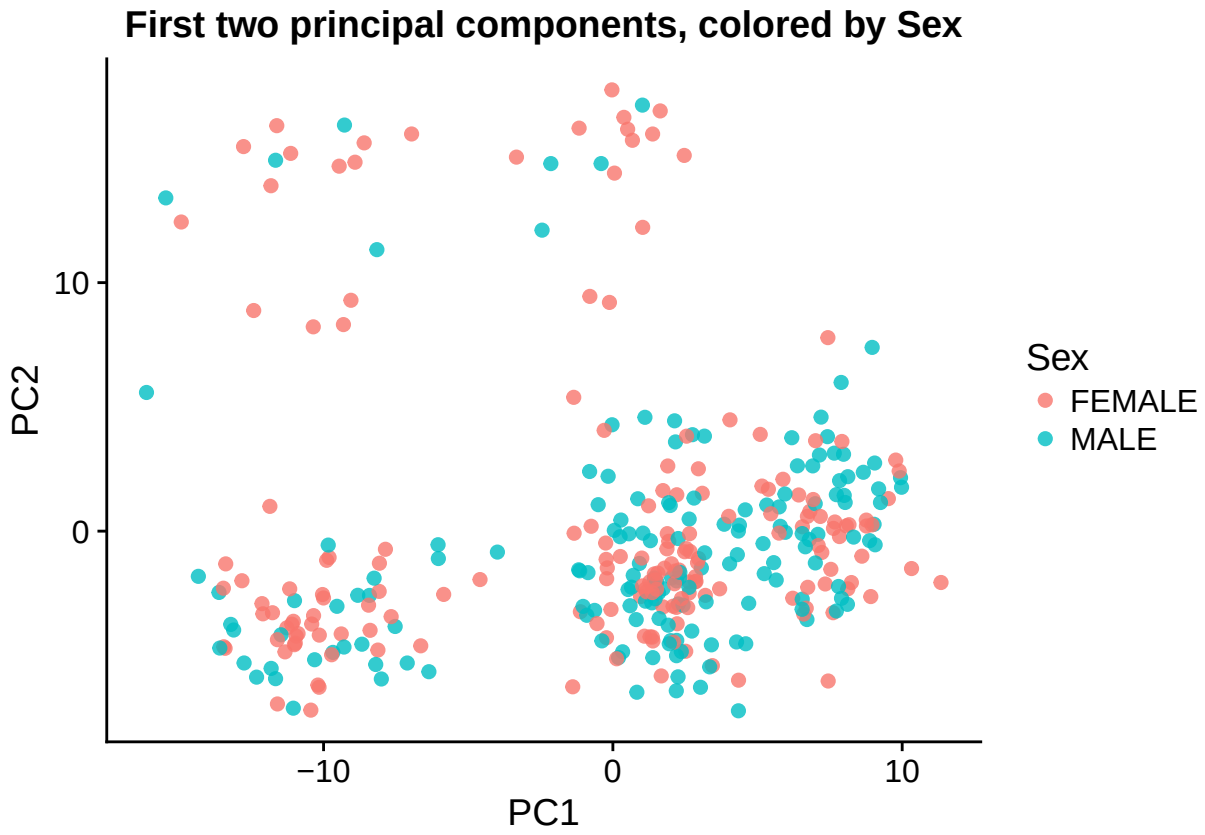

Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
0
50
100
150
200
250
300
350
Principal Component

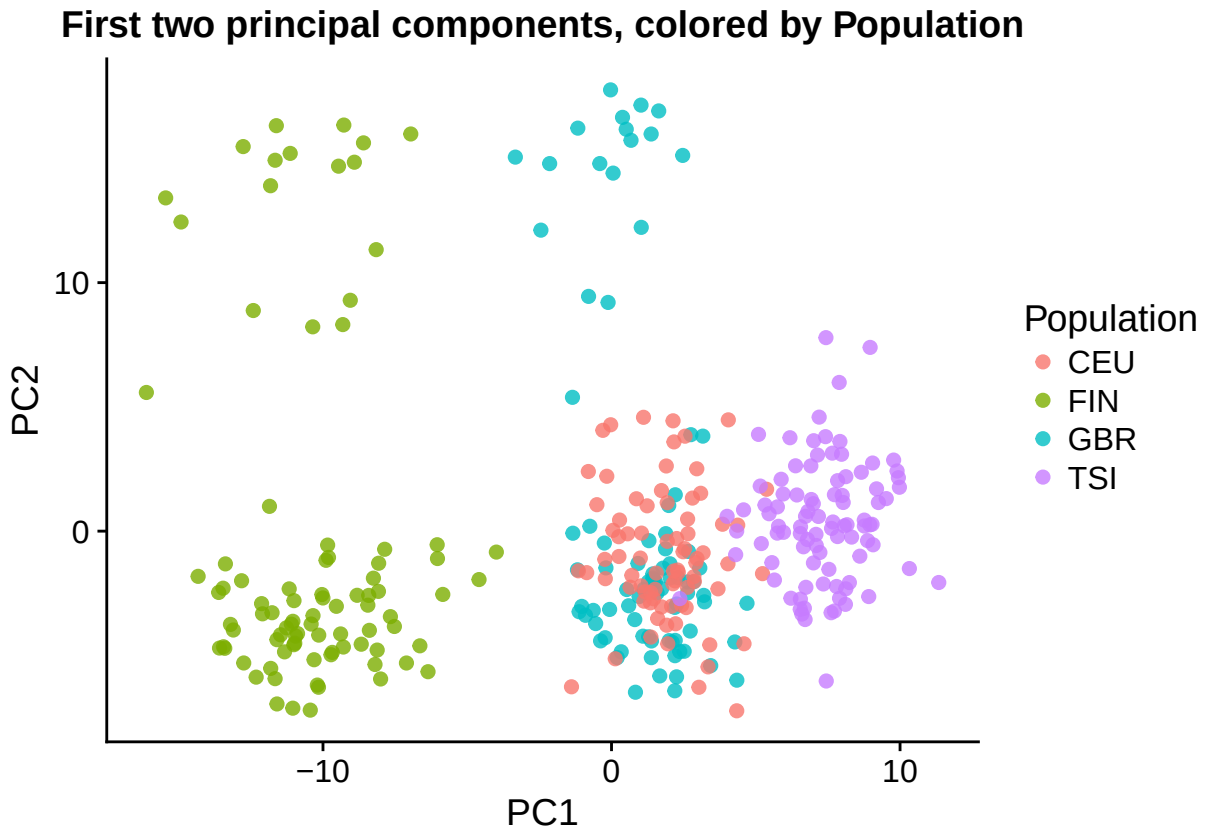

First two principal components, colored by Sex
PC2
40
20
0
−20
−25
0
25
50
PC1
Sex
FEMALE
MALE


First two principal components, colored by Population
PC2
40
20
0
−20
−25
0
25
50
PC1
Population
CEU
FIN
GBR
TSI

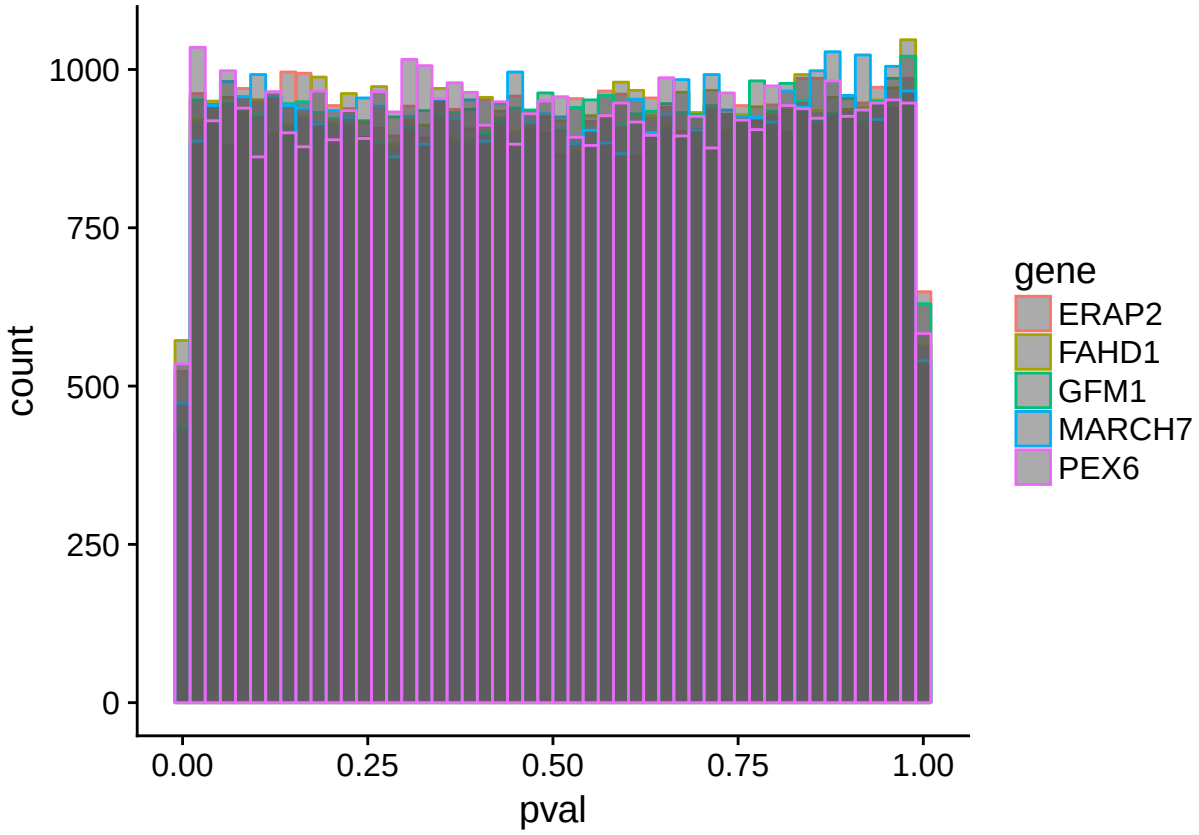

PC2
0.1
0.0
−0.1
−0.10
−0.05
0.00
0.05
PC1
Population
CEU
FIN
GBR
TSI

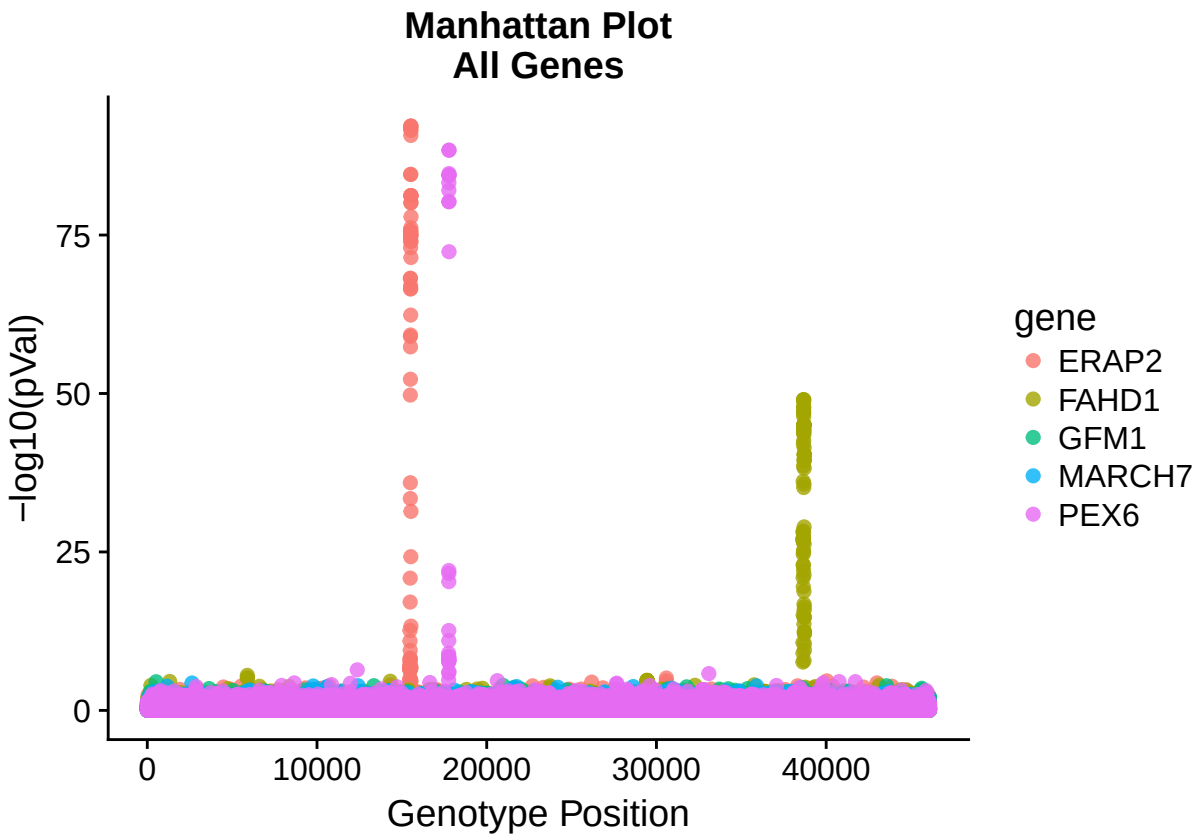

Percent of total variance explained
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
Principal Component

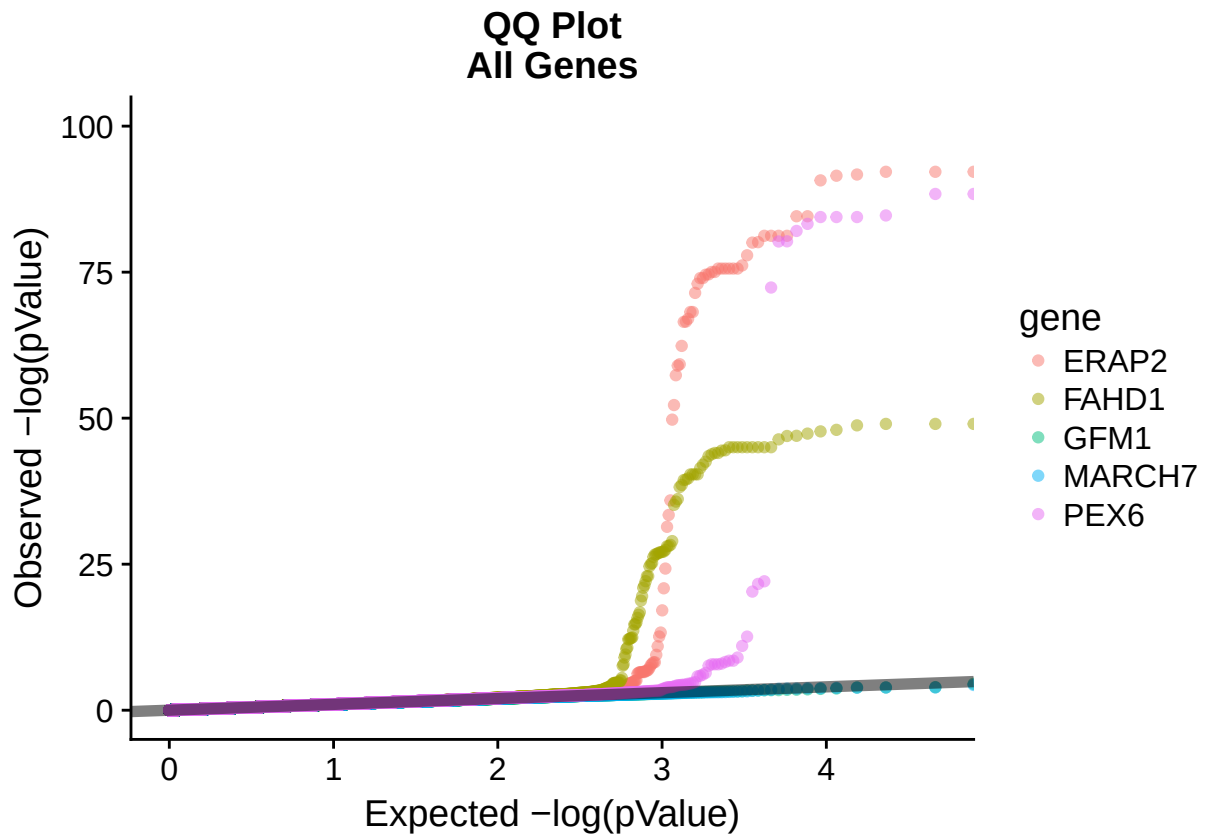

First two principal components, colored by Sex
PC2
10
0
−10
0
10
PC1
Sex
FEMALE
MALE



Try just a lin reg and F-test with no covars after filtering.


1000
750
500
250
0
count
0.00
0.25
0.50
0.75
1.00
pval
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6

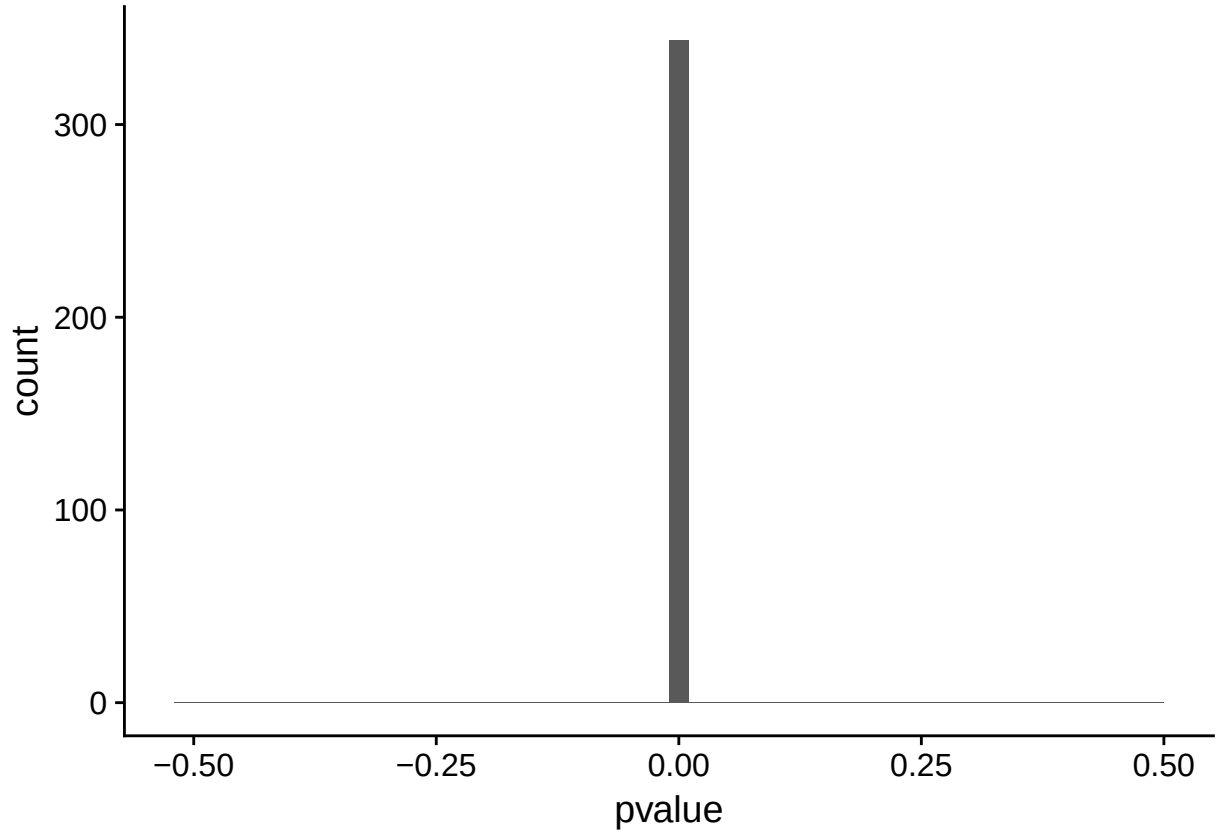

Manhattan Plot
All Genes
−log10(pVal)
75
50
25
0
0
10000
20000
30000
40000
Genotype Position
gene
ERAP2
FAHD1
GFM1
MARCH7
PEX6



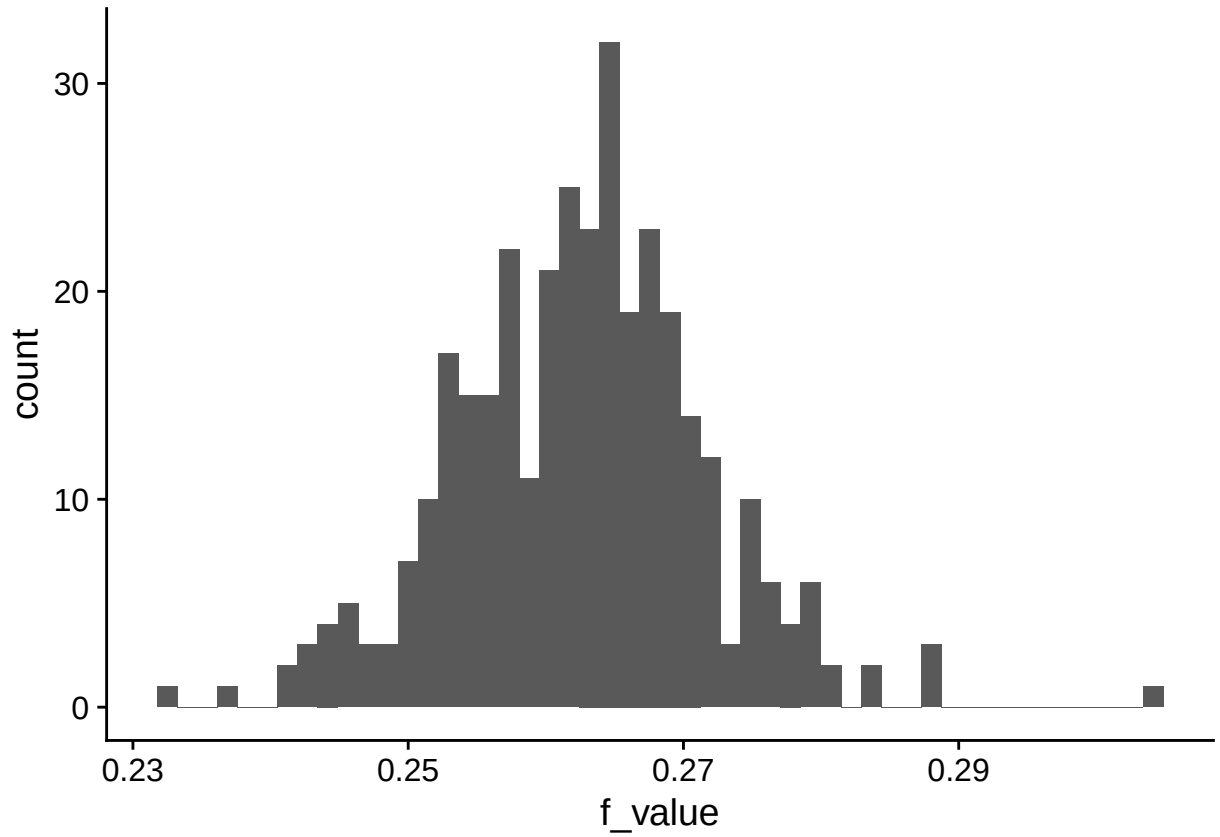

```
indv_HW <- map(data.frame(t(geno)), function(x) HWChisq(table(factor(x, levels = c(0, 1, 2))), cc = 0, v
HWChisq(table(factor(t(geno[ 1,]), levels = c(0, 1, 2))), cc = 0.5)
```

```
## Chi-square test with continuity correction for Hardy-Weinberg equilibrium (autosomal)
## Chi2 =  3361.967 DF =  1 p-value =  0 D =  -2980.787 f =  0.259339
```

```
indv_HW2 <- purrr::transpose(indv_HW )
indv_HW_simple <- data.frame(unlist(indv_HW2$pval))
indv_HW_f <- data.frame(unlist(indv_HW2$f))
colnames(indv_HW_simple) <- "pvalue"
colnames(indv_HW_f) <- "f_value"
indv_HW_simple$position <- 1:nrow(indv_HW_simple)
indv_HW_f$position <- 1:nrow(indv_HW_f)
indv_HW_simple %>%
  ggplot(aes(x = pvalue)) +
  geom_histogram(bins = 50)
```




```
indv_HW_f %>%
  ggplot(aes(x = f_value)) +
  geom_histogram(bins = 50)
```




```
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)
```