quant_gen_project

Darya Akimova

April 20, 2018

```
# 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?
knitr::kable(table(covars$Population))
```

Var1	Freq
CEU	78
FIN	89
GBR	85
TSI	92

```
knitr::kable(table(covars$Sex))
```

Var1	Freq
FEMALE	181
MALE	163

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

```
[1] 2
```

```
min(as.matrix(geno))
```

```
[1] 0
```

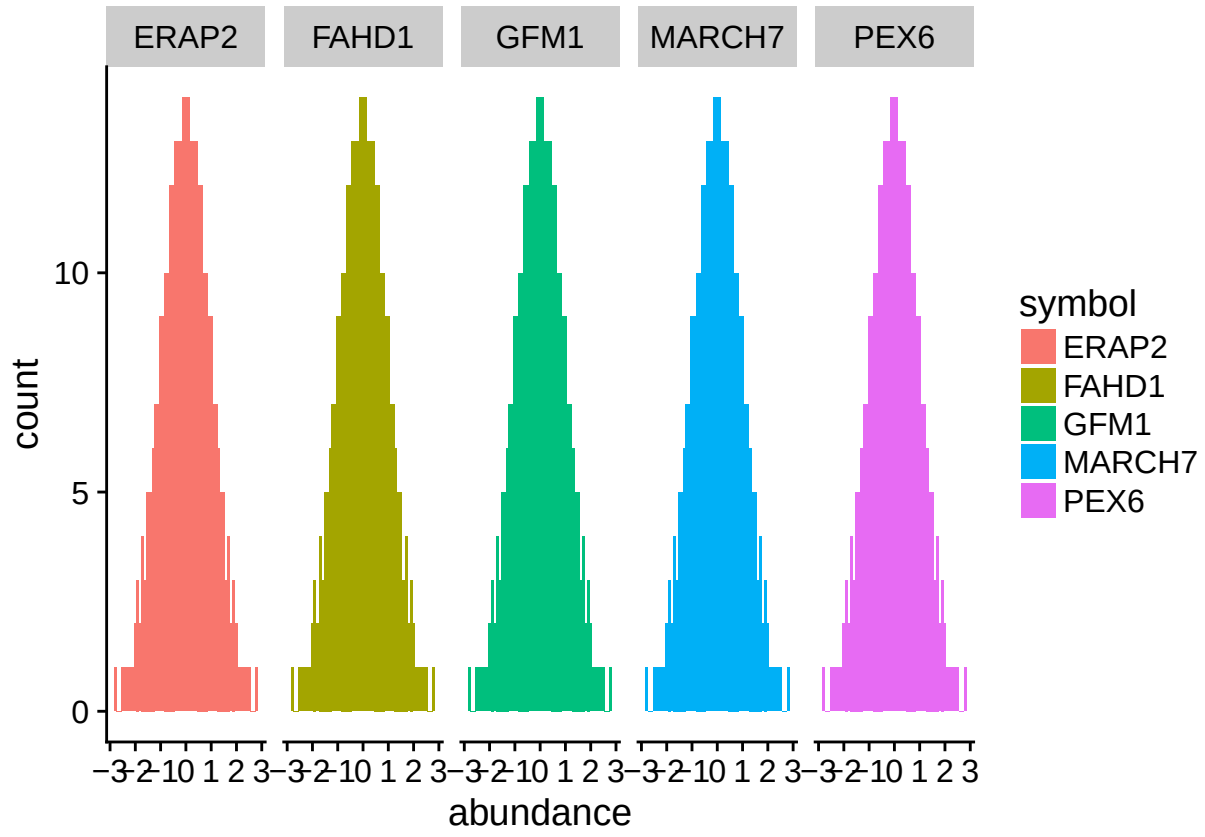
```
# calculate allele frequency
geno_sums <- map_dbl(geno, function(x) sum(x) / (nrow(geno) * 2))
# any genotypes need to be removed because of MAF < 5%?
any(geno_sums < 0.05 | geno_sums > 0.95)
```

```
[1] FALSE
```

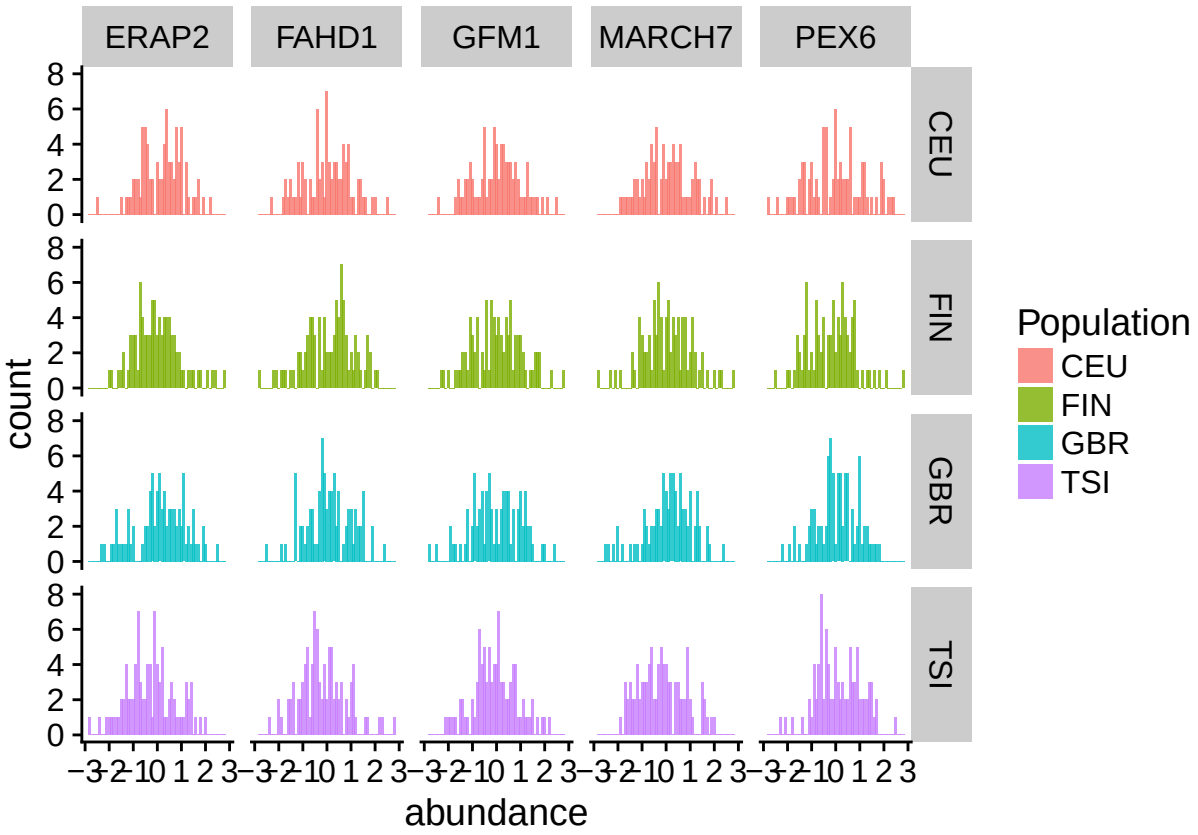
```
# no genotypes need to be removed because of MAF < 5%
```

Phenotype plots:

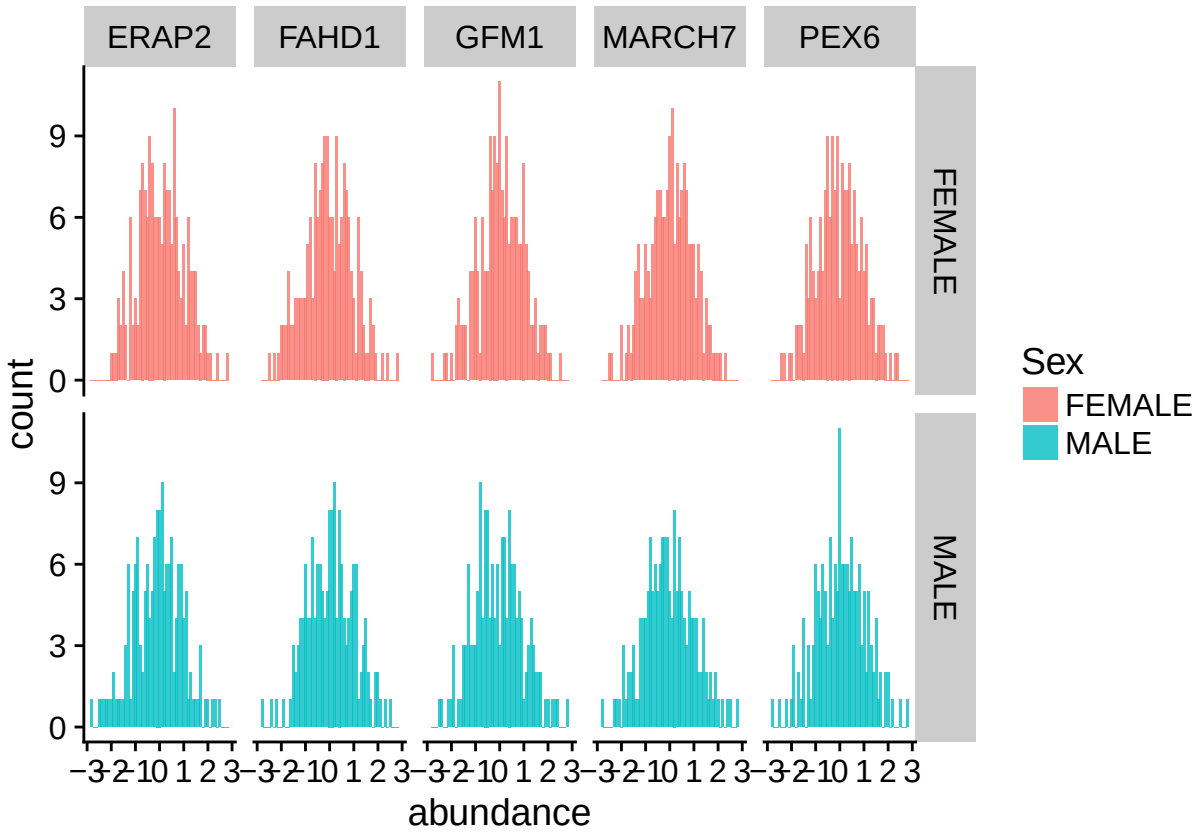
```
pheno %>%
  rownames_to_column("sample") %>%
  gather("probe", "abundance", 2:6) %>%
  left_join(gene_info, by = "probe") %>%
  ggplot(aes(x = abundance, fill = symbol)) +
  geom_histogram(binwidth = 0.1) +
  facet_wrap(~ symbol, nrow = 1)
```



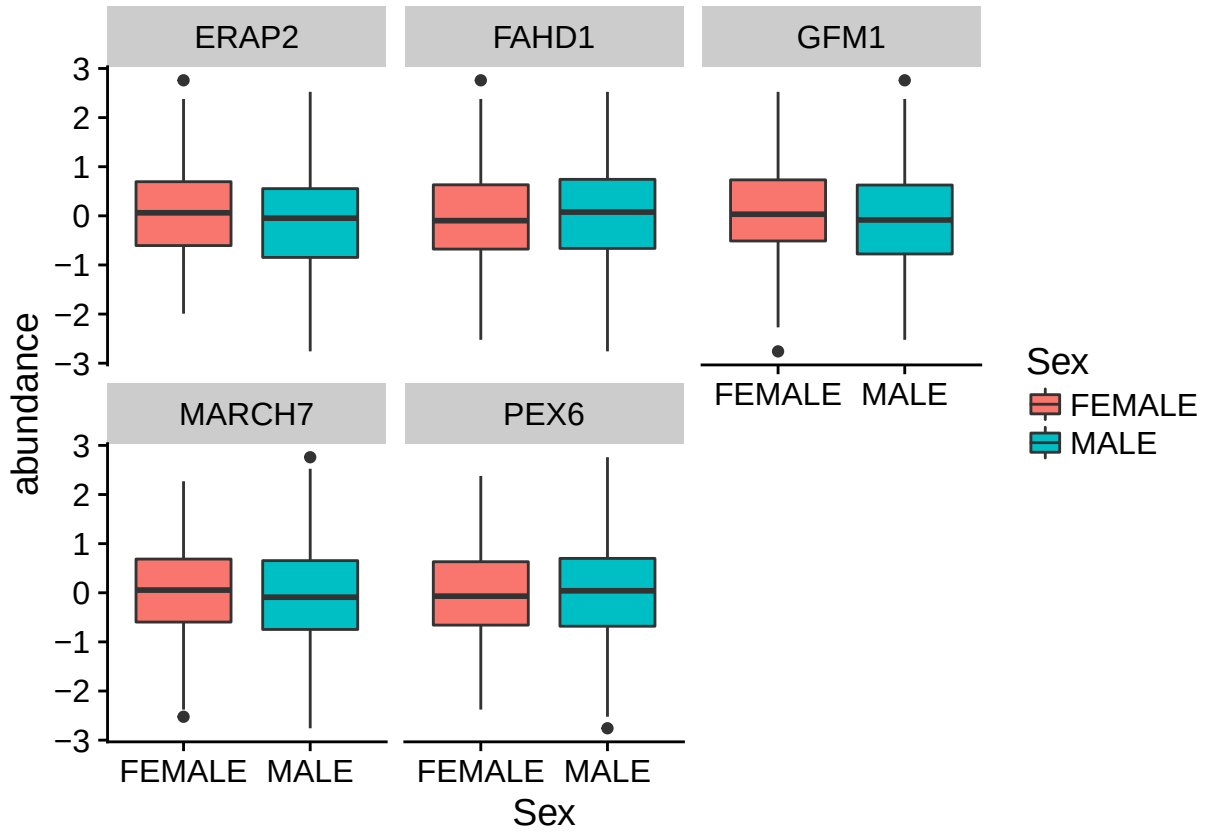

```
pheno %>%
  rownames_to_column("sample") %>%
  gather("probe", "abundance", 2:6) %>%
  left_join(gene_info, by = "probe") %>%
  left_join(
    covars %>% rownames_to_column("sample"),
    by = "sample") %>%
  ggplot(aes(x = abundance, fill = Population)) +
  geom_histogram(binwidth = 0.1, alpha = 0.8, position = "identity") +
  facet_grid(Population ~ symbol)
```



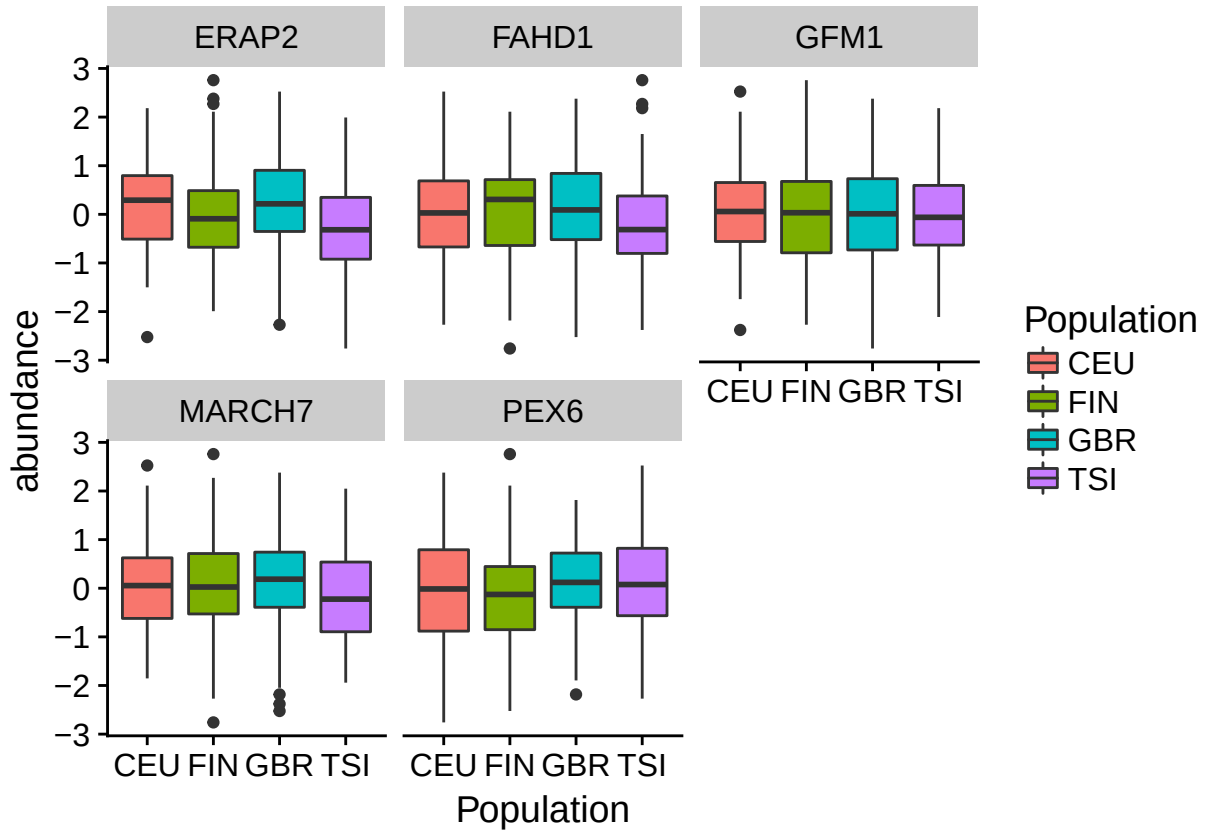

```
pheno %>%
  rownames_to_column("sample") %>%
  gather("probe", "abundance", 2:6) %>%
  left_join(gene_info, by = "probe") %>%
  left_join(
    covars %>% rownames_to_column("sample"),
    by = "sample") %>%
  ggplot(aes(x = abundance, fill = Sex)) +
  geom_histogram(binwidth = 0.1, alpha = 0.8, position = "identity") +
  facet_grid(Sex ~ symbol)
```




```
pheno %>%
  rownames_to_column("sample") %>%
  gather("probe", "abundance", 2:6) %>%
  left_join(gene_info, by = "probe") %>%
  left_join(
    covars %>% rownames_to_column("sample"),
    by = "sample") %>%
  ggplot(aes(x = Sex, y = abundance, fill = Sex)) +
  geom_boxplot() +
  facet_wrap(~ symbol)
```



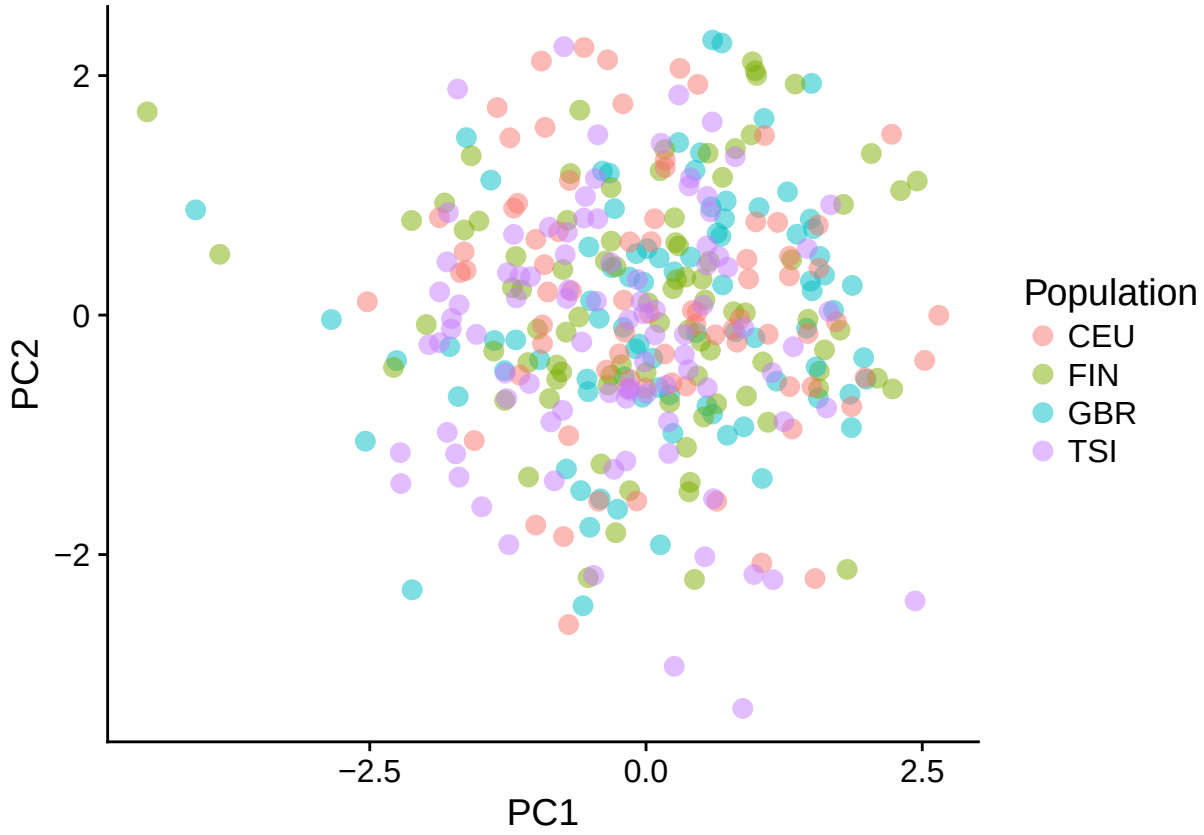

```
pheno %>%
  rownames_to_column("sample") %>%
  gather("probe", "abundance", 2:6) %>%
  left_join(gene_info, by = "probe") %>%
  left_join(
    covars %>% rownames_to_column("sample"),
    by = "sample") %>%
  ggplot(aes(x = Population, y = abundance, fill = Population)) +
  geom_boxplot() +
  facet_wrap(~ symbol)
```



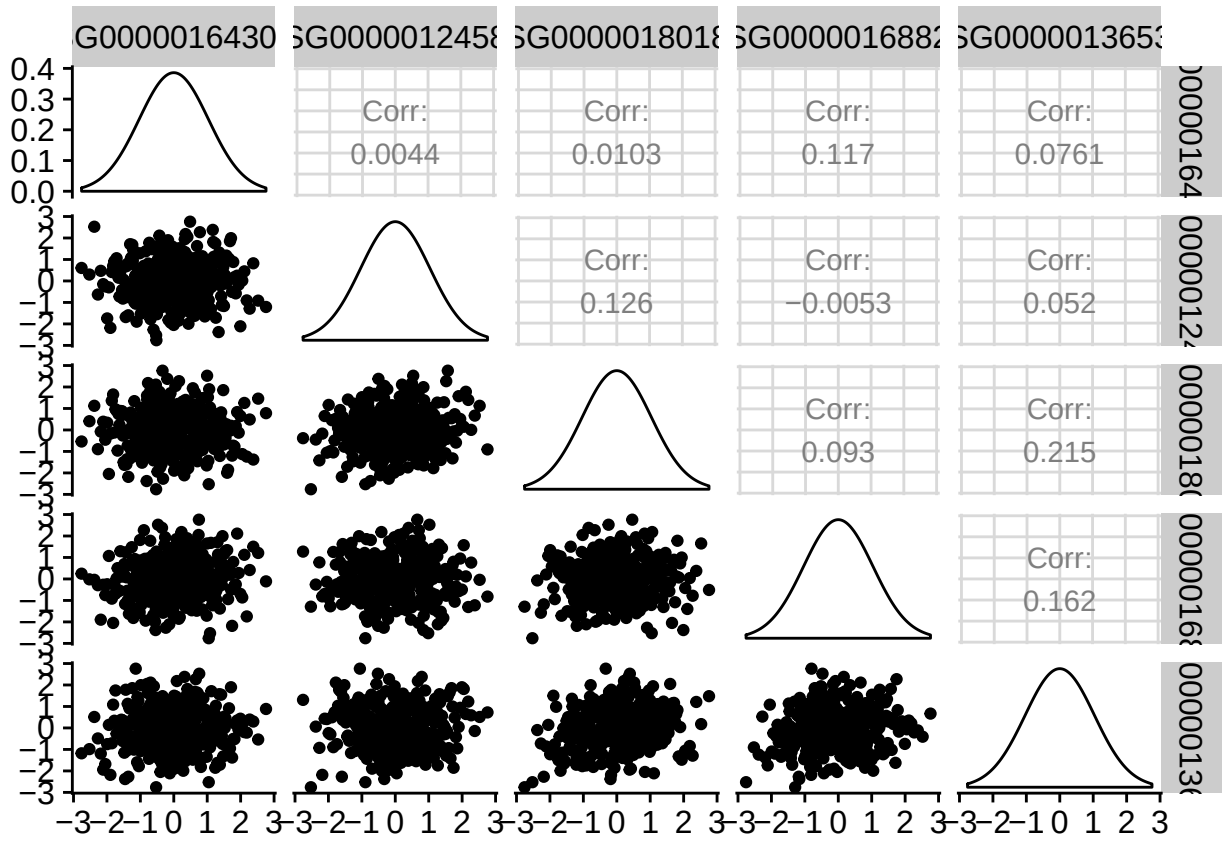

```
# are there any genotypes without associated phenotypes, or vice versa?
all.equal(rownames(geno), rownames(pheno))
```

```
[1] TRUE
```

```
pheno_pca <- prcomp(pheno)
pheno_pca_x <- as.data.frame(pheno_pca$x)
pheno_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 = 3, alpha = 0.5)
```

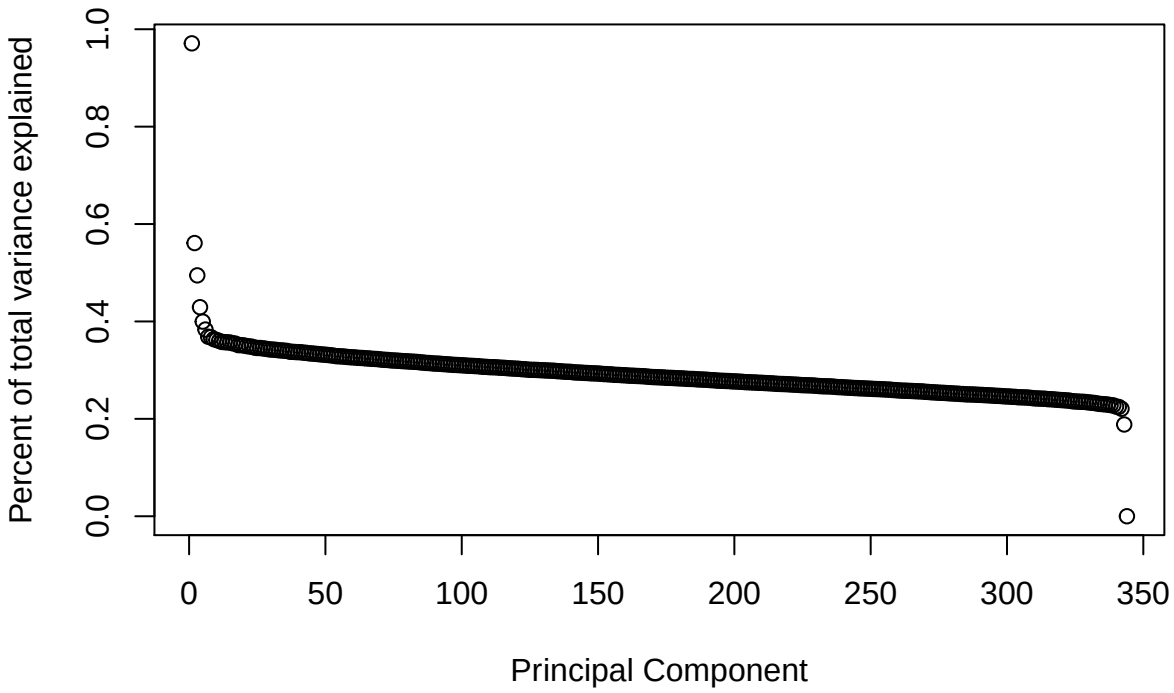



```
ggpairs(pheno)
```

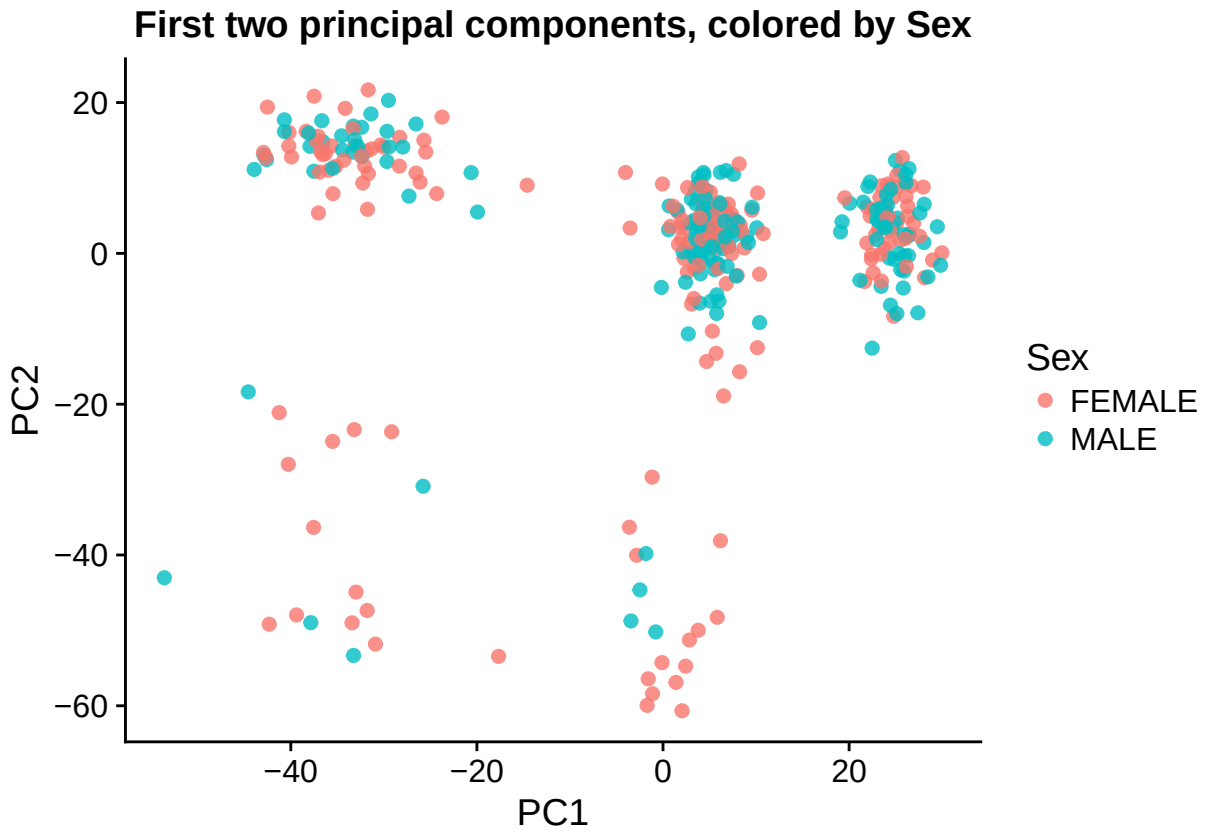


Genotype plots:

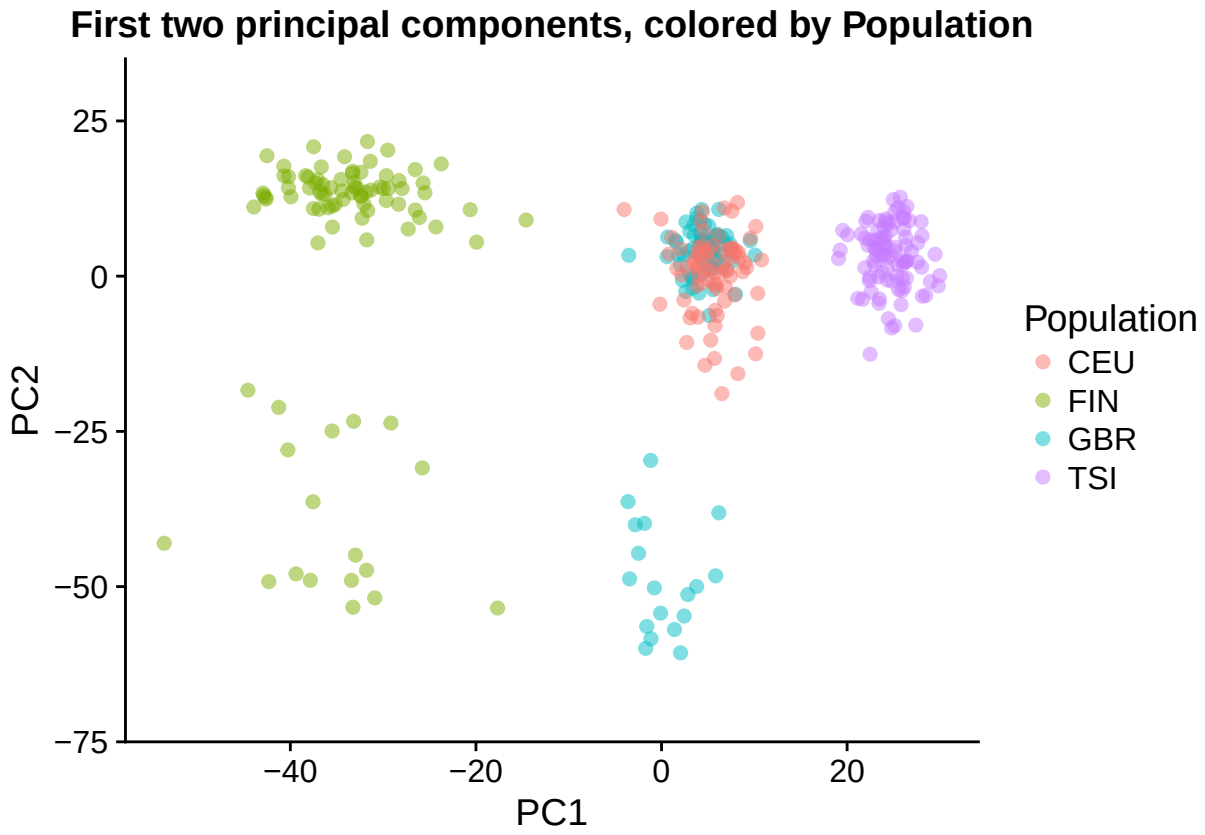
```
# PCA analysis of genetic data
geno_pca <- prcomp(geno, center = TRUE, scale. = TRUE)
plot((geno_pca$sdev^2/ sum(geno_pca$sdev^2)) * 100, xlab = "Principal Component", ylab = "Percent of to
```




```
geno_pca_x <- data.frame(geno_pca$x)
geno_pca_x %>%
  rownames_to_column("sample") %>%
  left_join(covars %>% rownames_to_column("sample"), by = "sample") %>%
  ggplot(aes(x = PC1, y = PC2, color = Sex)) +
    geom_point(size = 2, alpha = 0.8) +
    ggtitle("First two principal components, colored by Sex")
```




```
geno_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 = 0.5) +
    ggtitle("First two principal components, colored by Population") +
  ylim(-70, 30)
```




```
#indv_pca <- prcomp(t(geno), center = TRUE, scale. = TRUE)
#indv_pca_x <- indv_pca$x
```

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.

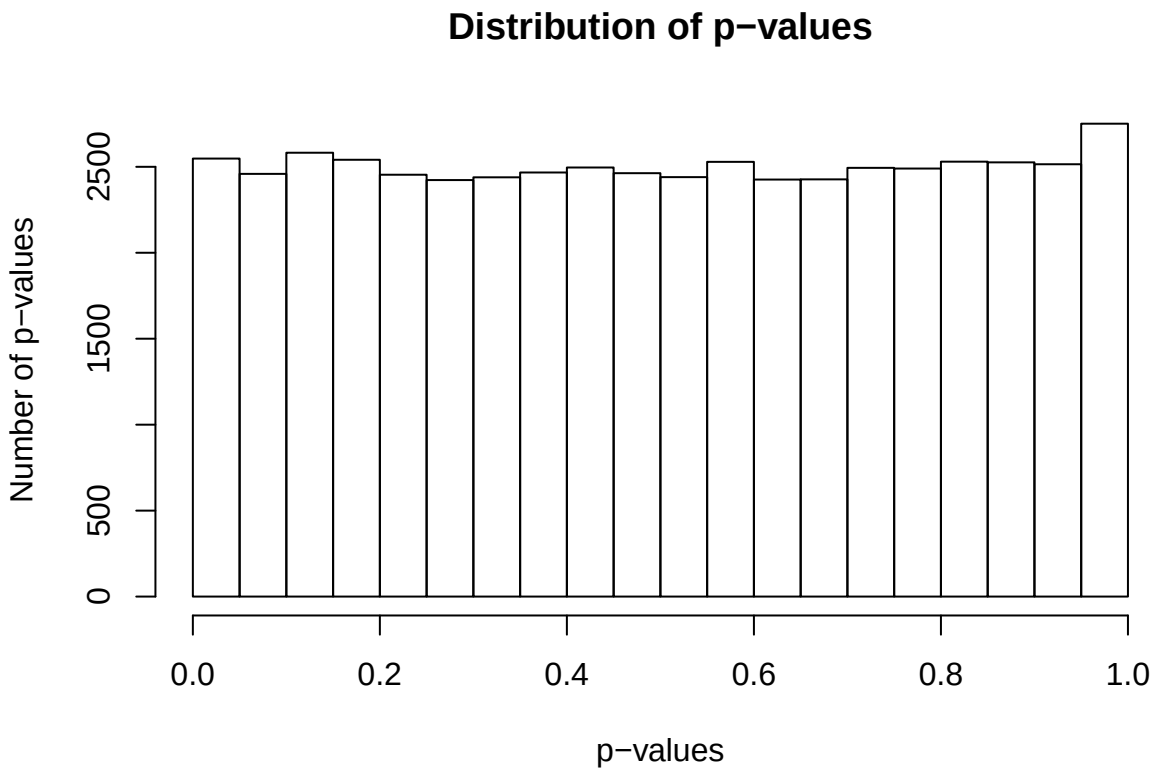
```
x_a <- as.matrix(geno - 1)
x_d <- replace(as.matrix(geno), which(as.matrix(geno) == 2 | as.matrix(geno) == 0), -1)
MLE <- function(xa, xd, y_mat) {
  X_mat <- cbind(1, xa, xd)
  beta_hat <- ginv(t(X_mat) %*% X_mat) %*% t(X_mat) %*% y_mat
}
fstat_calc <- function(xa, xd, MLE, y_mat) {
  X_mat <- cbind(1, xa, xd)
  y_hat <- X_mat %*% MLE
  SSM <- sum((y_hat - mean(y_mat)) ^ 2)
  SSE <- sum((y_mat - y_hat) ^ 2)
  df_M <- 2
  df_E <- length(xa) - 3
  Fstat <- (SSM / df_M) / (SSE / df_E)
  return(Fstat)
}

# gene 1
mle_gene1 <- map2(data.frame(x_a), data.frame(x_d), MLE, as.numeric(as.matrix(pheno[, 1])))
```

```
pval_gene1 <- pmap(
  list(
    xa = data.frame(x_a),
    xd = data.frame(x_d),
    MLE = mle_gene1
    ),
  fstat_calc, pheno[, 1]
  ) %>%
  flatten_dbl() %>%
  pf(2, 344 - 3, lower.tail = FALSE)
# gene 2
mle_gene2 <- map2(data.frame(x_a), data.frame(x_d), MLE, as.numeric(as.matrix(pheno[, 2])))
pval_gene2 <- pmap(
  list(
    xa = data.frame(x_a),
    xd = data.frame(x_d),
    MLE = mle_gene2
    ),
  fstat_calc, pheno[, 2]
  ) %>%
  flatten_dbl() %>%
  pf(2, 344 - 3, lower.tail = FALSE)
# gene 3
mle_gene3 <- map2(data.frame(x_a), data.frame(x_d), MLE, as.numeric(as.matrix(pheno[, 3])))
pval_gene3 <- pmap(
  list(
    xa = data.frame(x_a),
    xd = data.frame(x_d),
    MLE = mle_gene3
    ),
  fstat_calc, pheno[, 3]
  ) %>%
  flatten_dbl() %>%
  pf(2, 344 - 3, lower.tail = FALSE)
# gene 4
mle_gene4 <- map2(data.frame(x_a), data.frame(x_d), MLE, as.numeric(as.matrix(pheno[, 4])))
pval_gene4 <- pmap(
  list(
    xa = data.frame(x_a),
    xd = data.frame(x_d),
    MLE = mle_gene4
    ),
  fstat_calc, pheno[, 4]
  ) %>%
  flatten_dbl() %>%
  pf(2, 344 - 3, lower.tail = FALSE)
# gene 5
mle_gene5 <- map2(data.frame(x_a), data.frame(x_d), MLE, as.numeric(as.matrix(pheno[, 5])))
pval_gene5 <- pmap(
  list(
    xa = data.frame(x_a),
    xd = data.frame(x_d),
    MLE = mle_gene5
```

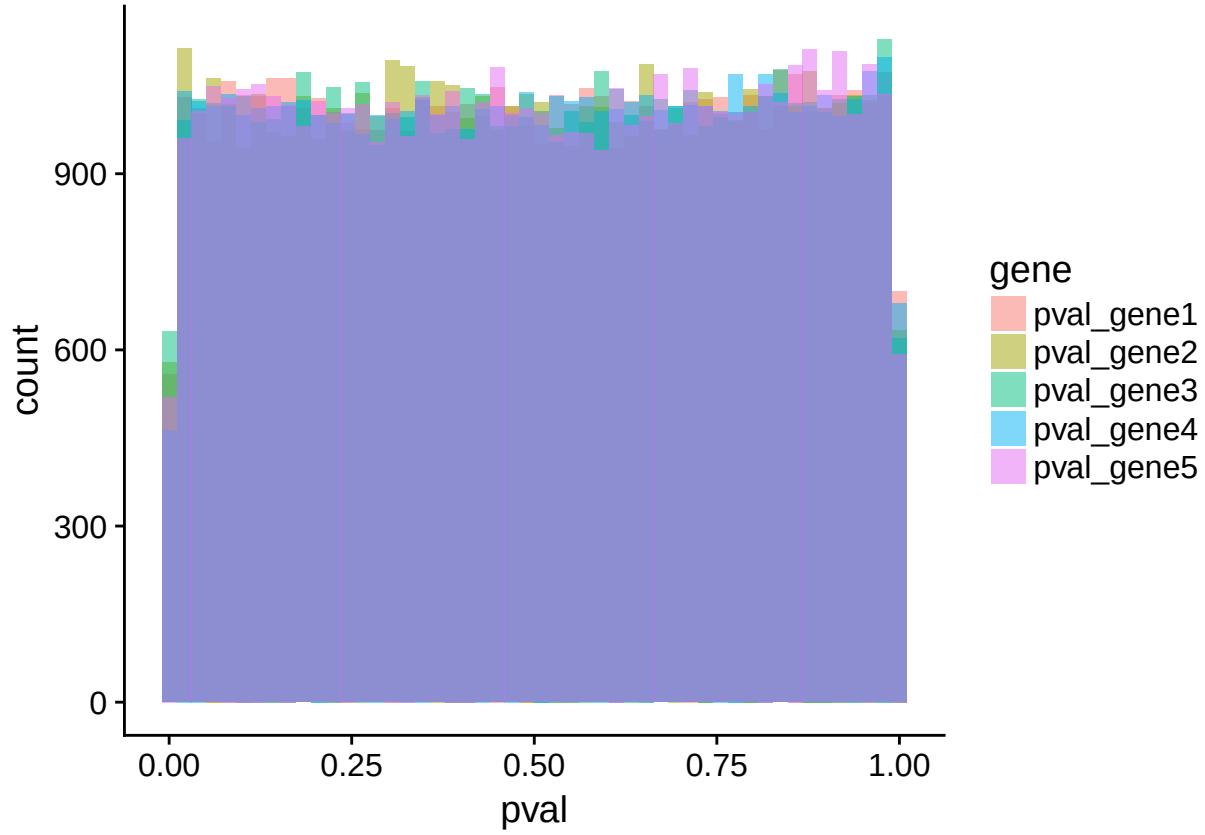
```
    ),
  fstat_calc, pheno[, 5]
  ) %>%
  flatten_dbl() %>%
  pf(2, 344 - 3, lower.tail = FALSE)


hist(pval_gene1, breaks = 20,
     xlab = "p-values",
     ylab = "Number of p-values",
     main = "Distribution of p-values")
```

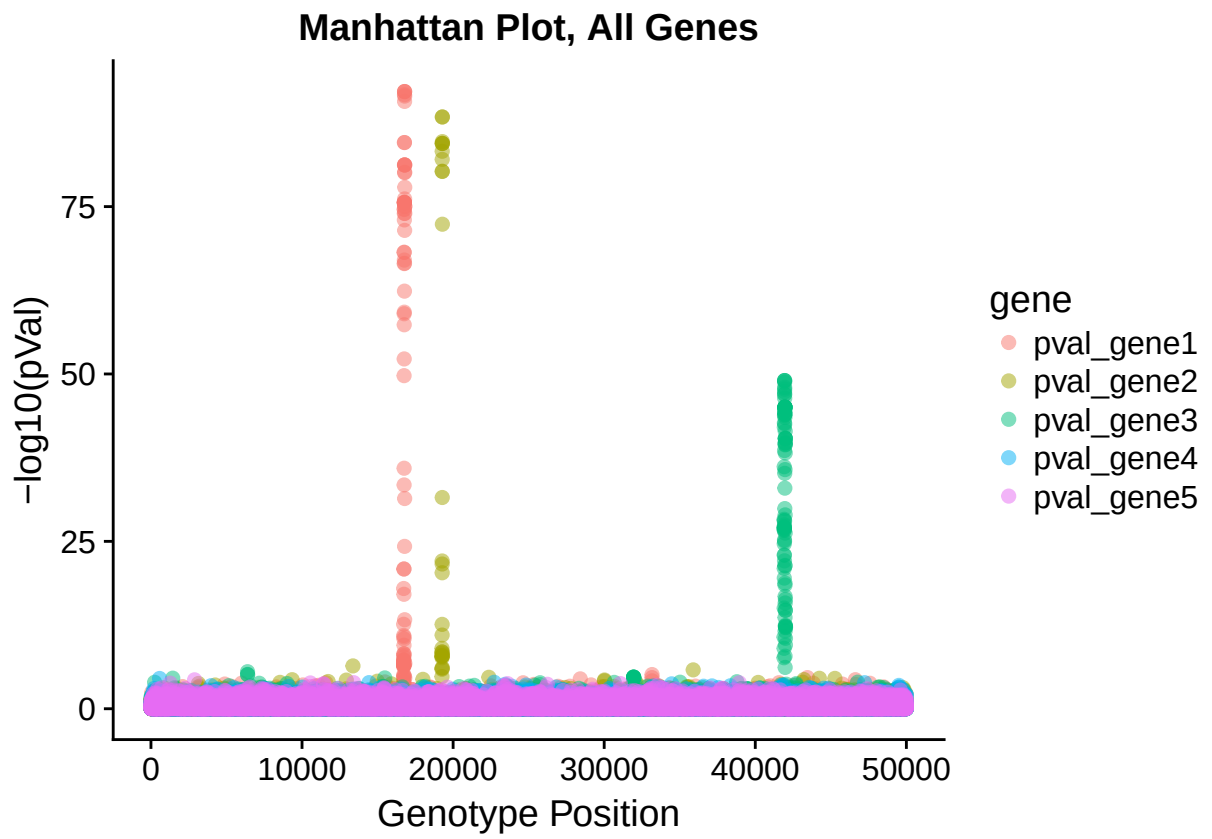



```
pval_full_data <- as.data.frame(
  cbind(
    x = 1:ncol(geno),
    pval_gene1,
    pval_gene2,
    pval_gene3,
    pval_gene4,
    pval_gene5
    )
  )
row.names(pval_full_data) <- colnames(geno)
pval_full_data %>%
  gather("gene", "pval", 2:6) %>%
  ggplot(aes(x = pval, fill = gene)) +
```

```
  geom_histogram(position = "identity", alpha = 0.5, bins = 50)
```




```
pval_full_data %>%
  gather("gene", "pval", 2:6) %>%
  ggplot(aes(x = x, y = -log(pval, base = 10), color = gene)) +
  geom_point(alpha = 0.5, size = 2) +
  ggtitle("Manhattan Plot, All Genes") +
  xlab("Genotype Position") +
  ylab("-log10(pVal)")
```




```
pval_full_data %>%
  rownames_to_column("snp") %>%
  dplyr::select(-x) %>%
  gather("gene", "pval", 2:6) %>%
  filter(pval < (0.5 / ncol(geno)))
```

```
           snp       gene         pval
1   rs200641494 pval_gene1 9.148167e-06
2       rs27433 pval_gene1 1.269621e-07
3   rs146925065 pval_gene1 3.066873e-07
4       rs27043 pval_gene1 5.827862e-08
5   rs200528525 pval_gene1 2.423748e-13
6      rs469783 pval_gene1 1.098376e-18
7      rs469367 pval_gene1 3.872843e-07
8      rs246455 pval_gene1 1.124749e-11
9       rs27640 pval_gene1 1.850944e-07
10      rs26491 pval_gene1 5.340844e-07
11      rs28048 pval_gene1 4.855907e-08
12     rs149189 pval_gene1 1.923361e-08
13      rs26653 pval_gene1 3.058964e-07
14      rs28050 pval_gene1 7.235482e-08
15      rs27778 pval_gene1 7.235482e-08
16      rs34756 pval_gene1 2.928521e-07
17      rs34761 pval_gene1 2.928521e-07
18      rs40090 pval_gene1 6.325407e-09
19      rs40604 pval_gene1 2.928521e-07
```

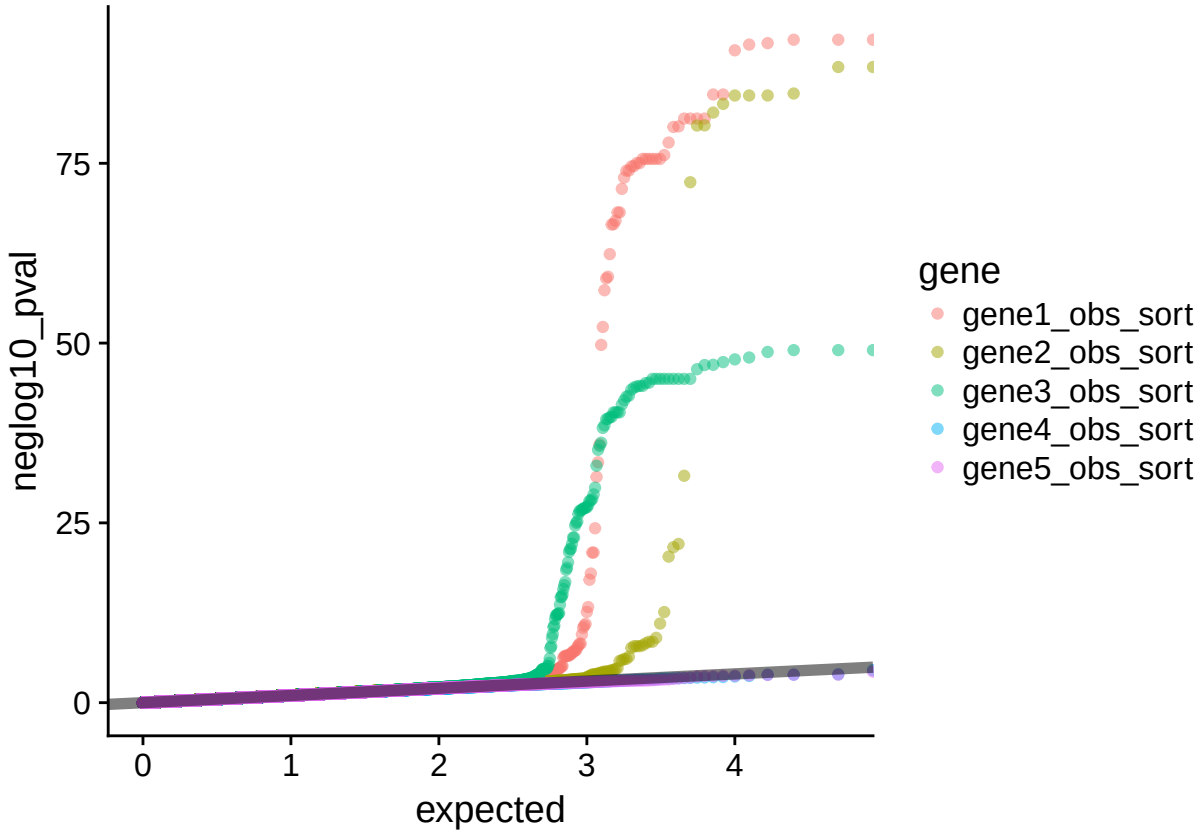
```
20      rs34737 pval_gene1 1.336235e-21
21    rs2549802 pval_gene1 8.056264e-18
22     rs709669 pval_gene1 1.350231e-21
23    rs2911138 pval_gene1 1.065107e-08
24    rs2927620 pval_gene1 3.293978e-10
25   rs70981851 pval_gene1 6.136687e-09
26    rs2059247 pval_gene1 3.841600e-11
27    rs6868625 pval_gene1 1.859734e-11
28    rs2549778 pval_gene1 1.777997e-50
29    rs3840523 pval_gene1 3.741563e-34
30    rs1230363 pval_gene1 1.194293e-36
31   rs12189125 pval_gene1 6.716777e-69
32   rs12653964 pval_gene1 1.011421e-59
33    rs1559354 pval_gene1 6.716777e-69
34  rs201477313 pval_gene1 5.657318e-53
35  rs201109800 pval_gene1 4.464857e-58
36    rs4869314 pval_gene1 3.277003e-67
37   rs10707238 pval_gene1 2.369039e-75
38    rs2549784 pval_gene1 2.452171e-76
39    rs2161657 pval_gene1 2.452171e-76
40    rs6859160 pval_gene1 1.033638e-74
41     rs251339 pval_gene1 5.889106e-60
42    rs1423568 pval_gene1 2.452171e-76
43    rs2549791 pval_gene1 2.732271e-75
44    rs2548525 pval_gene1 9.781632e-74
45   rs13163165 pval_gene1 9.708860e-68
46    rs2549801 pval_gene1 2.452171e-76
47    rs2910688 pval_gene1 2.452171e-76
48    rs1363907 pval_gene1 2.704396e-85
49    rs1230382 pval_gene1 2.452171e-76
50    rs1216571 pval_gene1 2.704396e-85
51    rs1216568 pval_gene1 4.246055e-63
52    rs1046395 pval_gene1 5.988364e-82
53    rs2548226 pval_gene1 2.927206e-67
54    rs6887500 pval_gene1 1.935424e-92
55    rs7726445 pval_gene1 6.561995e-93
56    rs1477364 pval_gene1 5.988364e-82
57    rs7731592 pval_gene1 6.561995e-93
58    rs7723899 pval_gene1 5.636683e-25
59    rs7716222 pval_gene1 1.937345e-91
60   rs10058476 pval_gene1 8.744376e-81
61  rs140336797 pval_gene1 7.136110e-77
62    rs2161548 pval_gene1 6.561995e-93
63    rs6879678 pval_gene1 3.565862e-72
64  rs201866654 pval_gene1 4.011525e-32
65    rs9918183 pval_gene1 5.067325e-14
66    rs1160962 pval_gene1 3.101499e-92
67  rs199514005 pval_gene1 7.406371e-81
68      rs27306 pval_gene1 5.988364e-82
69      rs27307 pval_gene1 5.988364e-82
70   rs11414909 pval_gene1 1.305277e-78
71      rs27292 pval_gene1 9.325835e-76
72      rs27747 pval_gene1 9.325835e-76
73     rs248215 pval_gene1 1.050026e-74
```

```
74   rs72777610 pval_gene1 1.745556e-07
75     rs731842 pval_gene1 7.866467e-06
76    rs6854915 pval_gene2 4.021951e-07
77    rs9462840 pval_gene2 1.268501e-06
78  rs200242944 pval_gene2 8.354577e-07
79       rs7952 pval_gene2 9.637984e-10
80    rs9462846 pval_gene2 5.566909e-09
81    rs2092554 pval_gene2 2.404863e-08
82    rs7754294 pval_gene2 1.389429e-08
83    rs9462850 pval_gene2 1.389429e-08
84    rs9296401 pval_gene2 1.389429e-08
85    rs6938994 pval_gene2 3.359650e-09
86    rs2235831 pval_gene2 3.359650e-09
87   rs34433785 pval_gene2 9.928791e-12
88   rs10948053 pval_gene2 2.505439e-13
89    rs9471969 pval_gene2 4.905110e-21
90    rs9462853 pval_gene2 8.497103e-23
91  rs199685173 pval_gene2 2.414732e-22
92    rs4714638 pval_gene2 8.830077e-83
93    rs9471976 pval_gene2 5.428222e-84
94   rs58497441 pval_gene2 5.297122e-81
95   rs35908551 pval_gene2 2.814398e-32
96   rs13215983 pval_gene2 3.683014e-85
97    rs6920547 pval_gene2 5.599943e-81
98    rs7760250 pval_gene2 3.683014e-85
99    rs1129187 pval_gene2 4.098120e-89
100   rs3805953 pval_gene2 3.683014e-85
101  rs10948061 pval_gene2 4.098120e-89
102   rs9471985 pval_gene2 1.967219e-85
103 rs199921136 pval_gene2 4.285489e-73
104   rs3805941 pval_gene2 8.065379e-07
105  rs71779653 pval_gene2 1.065528e-08
106  rs61927910 pval_gene2 1.567648e-06
107  rs75747449 pval_gene3 2.946579e-06
108  rs62172878 pval_gene3 7.532754e-06
109 rs200834180 pval_gene3 8.341361e-06
110   rs2235639 pval_gene3 2.421310e-08
111  rs75276217 pval_gene3 1.962786e-11
112   rs7201813 pval_gene3 9.466001e-10
113   rs1178435 pval_gene3 7.774525e-28
114   rs2575352 pval_gene3 9.295452e-28
115   rs1065663 pval_gene3 5.010139e-29
116   rs3751893 pval_gene3 7.759308e-29
117   rs2745205 pval_gene3 9.206052e-16
118   rs2575345 pval_gene3 7.035690e-37
119   rs2745195 pval_gene3 1.099358e-21
120 rs142797272 pval_gene3 1.424277e-27
121 rs145338194 pval_gene3 2.347974e-25
122   rs2473469 pval_gene3 1.984595e-27
123   rs1657107 pval_gene3 3.046090e-20
124   rs1629534 pval_gene3 1.108691e-23
125   rs7186249 pval_gene3 1.108691e-23
126   rs8046750 pval_gene3 9.067516e-43
127   rs9935687 pval_gene3 1.030610e-27
```

128	rs9928566	pval_gene3	2.230755e-27
129	rs2575357	pval_gene3	6.821591e-28
130	rs62038378	pval_gene3	2.065160e-43
131	rs2076455	pval_gene3	3.624326e-28
132	rs2268674	pval_gene3	6.308651e-26
133	rs3760040	pval_gene3	1.509559e-44
134	rs1657152	pval_gene3	8.054322e-23
135	rs28364708	pval_gene3	2.729566e-39
136	rs1143034	pval_gene3	1.043991e-47
137	rs11644748	pval_gene3	9.564535e-50
138	rs8044343	pval_gene3	9.595524e-26
139	rs34392705	pval_gene3	4.417002e-48
140	rs11641513	pval_gene3	8.619659e-29
141	rs140254902	pval_gene3	9.564535e-50
142	rs12325082	pval_gene3	1.003780e-48
143	rs9652776	pval_gene3	9.564535e-50
144	rs11643835	pval_gene3	1.703923e-49
145	rs3813760	pval_gene3	1.833812e-48
146	rs112311350	pval_gene3	1.125363e-47
147	rs141551352	pval_gene3	4.257928e-47
148	rs4598914	pval_gene3	3.584927e-45
149	rs62038431	pval_gene3	2.815762e-43
150	rs62038440	pval_gene3	6.902053e-36
151	rs57530073	pval_gene3	3.186165e-45
152	rs138033584	pval_gene3	9.428227e-46
153	rs2575359	pval_gene3	9.428227e-46
154	rs2492884	pval_gene3	1.270439e-30
155	rs1742423	pval_gene3	9.428227e-46
156	rs1657118	pval_gene3	9.428227e-46
157	rs5006374	pval_gene3	1.132210e-33
158	rs9746488	pval_gene3	4.756848e-22
159	rs1742432	pval_gene3	9.428227e-46
160	rs410465	pval_gene3	2.989535e-44
161	rs449530	pval_gene3	8.986569e-45
162	rs388627	pval_gene3	8.986569e-45
163	rs433268	pval_gene3	9.428227e-46
164	rs366223	pval_gene3	1.717161e-19
165	rs1657127	pval_gene3	1.074919e-44
166	rs2492879	pval_gene3	9.428227e-46
167	rs2475043	pval_gene3	3.414030e-19
168	rs2492881	pval_gene3	9.428227e-46
169	rs391543	pval_gene3	2.351764e-14
170	rs150843657	pval_gene3	1.913703e-36
171	rs368188	pval_gene3	1.413490e-08
172	rs1742446	pval_gene3	9.428227e-46
173	rs4451947	pval_gene3	3.873116e-22
174	rs2437747	pval_gene3	1.676823e-17
175	rs2575335	pval_gene3	3.859583e-40
176	rs12926045	pval_gene3	2.477514e-12
177	rs12924741	pval_gene3	6.767286e-07
178	rs2369275	pval_gene3	1.113960e-29
179	rs6600179	pval_gene3	1.822134e-40
180	rs2815297	pval_gene3	5.692892e-27
181	rs6600180	pval_gene3	1.582825e-16

```
182   rs3952970 pval_gene3 3.540373e-42
183   rs1657100 pval_gene3 6.249310e-39
184   rs2974857 pval_gene3 4.185499e-41
185   rs2982447 pval_gene3 2.096186e-40
186   rs2974860 pval_gene3 4.185499e-41
187   rs2917517 pval_gene3 4.185499e-41
188   rs2906903 pval_gene3 1.910529e-15
189   rs1632124 pval_gene3 3.479223e-13
190   rs2982433 pval_gene3 3.340945e-40
191   rs1657139 pval_gene3 4.321624e-41
192   rs8059871 pval_gene3 4.592033e-17
193    rs740466 pval_gene3 2.065603e-15
194   rs1742401 pval_gene3 4.731150e-13
195    rs344364 pval_gene3 2.971985e-10
196    rs344366 pval_gene3 3.075476e-11
197    rs344369 pval_gene3 4.559263e-13
198   rs2575358 pval_gene3 7.109386e-13
199  rs58530366 pval_gene3 7.394539e-13
```

```r
data_for_qqplot <- data.frame(
  cbind(
    expected = sort(-log10(seq(from = 0,to = 1, length.out = length(pval_full_data$pval_gene1)))),
    gene1_obs_sort = sort(-log10(pval_full_data$pval_gene1)),
    gene2_obs_sort = sort(-log10(pval_full_data$pval_gene2)),
    gene3_obs_sort = sort(-log10(pval_full_data$pval_gene3)),
    gene4_obs_sort = sort(-log10(pval_full_data$pval_gene4)),
    gene5_obs_sort = sort(-log10(pval_full_data$pval_gene5))
    )
  )
data_for_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)
```




```
#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)
```

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:

```
gene1_pop_lm <- lm(pheno[, 1] ~ as.numeric(factor(covars$Population)))  # significant
gene2_pop_lm <- lm(pheno[, 2] ~ as.numeric(factor(covars$Population)))  # ns
gene3_pop_lm <- lm(pheno[, 3] ~ as.numeric(factor(covars$Population)))  # ns
gene4_pop_lm <- lm(pheno[, 4] ~ as.numeric(factor(covars$Population)))  # ns
gene5_pop_lm <- lm(pheno[, 5] ~ as.numeric(factor(covars$Population)))  # ns


nested_pheno <- pheno %>%
  rownames_to_column("samples") %>%
  left_join(
    covars %>%
      rownames_to_column("samples"),
    by = "samples"
  ) %>%
  gather("gene", "abundance", 2:6) %>%
  nest(-gene)
```

```
num_factor_pop_lm <- nested_pheno %>%
  mutate(model = map(data, ~lm(abundance ~ as.numeric(factor(Population)), data = .))) %>%
  mutate(glance_model = map(model, glance)) %>%
  unnest(glance_model) %>%
  dplyr::select(gene, p.value)
num_factor_pop_lm[which(num_factor_pop_lm$p.value < 0.05), ]
```

```
## # A tibble: 1 x 2
##   gene               p.value
##   <chr>                <dbl>
## 1 ENSG00000164308.12 0.00799
```

```
# what if Population was treated as a factor and R was allowed to code it
cat_factor_pop_lm <- nested_pheno %>%
  mutate(model = map(data, ~lm(abundance ~ factor(Population), data = .))) %>%
  mutate(glance_model = map(model, glance)) %>%
  unnest(glance_model) %>%
  dplyr::select(gene, p.value)
cat_factor_pop_lm[which(cat_factor_pop_lm$p.value < 0.05), ]
```

```
## # A tibble: 1 x 2
##   gene               p.value
##   <chr>                <dbl>
## 1 ENSG00000164308.12 0.00914
```

```
# what about the Sex covariable?
covar_sex_lm <- nested_pheno %>%
  mutate(model = map(data, ~lm(abundance ~ factor(Sex), data = .))) %>%
  mutate(glance_model = map(model, glance)) %>%
  unnest(glance_model) %>%
  dplyr::select(gene, p.value)
any(covar_sex_lm < 0.05)
```

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

```
# none are significant on the sex covariable
# what about a combination?
covars_sex_num_pop_lm <- nested_pheno %>%
  mutate(model = map(data, ~lm(abundance ~ as.numeric(factor(Population)) + factor(Sex), data = .))) %>%
  mutate(glance_model = map(model, glance)) %>%
  unnest(glance_model) %>%
  dplyr::select(gene, p.value)
covars_sex_num_pop_lm[which(covars_sex_num_pop_lm$p.value < 0.05), ]
```

```
## # A tibble: 1 x 2
##   gene               p.value
##   <chr>                <dbl>
## 1 ENSG00000164308.12 0.00787
```

```
nested_pheno %>%
  mutate(model = map(data, ~lm(abundance ~ as.numeric(factor(Population)) + factor(Sex), data = .))) %>%
  mutate(tidy_model = map(model, tidy)) %>%
  unnest(tidy_model) %>%
  filter(gene %in% covars_sex_num_pop_lm[which(covars_sex_num_pop_lm$p.value < 0.05), 1])
```

```
## # A tibble: 3 x 6
```

```
##   gene                term              estimate std.error statistic p.value
##   <chr>               <chr>                <dbl>     <dbl>     <dbl>   <dbl>
## 1 ENSG00000164308.12 (Intercept)          0.399    0.140       2.86 0.00454
## 2 ENSG00000164308.12 as.numeric(fact~    -0.124    0.0474     -2.63 0.00901
## 3 ENSG00000164308.12 factor(Sex)MALE     -0.172    0.105      -1.64 0.103
```

```
# indicates that the relationship between population only is significant
covars_sex_cat_pop_lm <- nested_pheno %>%
  mutate(model = map(data, ~lm(abundance ~ factor(Population) + factor(Sex), data = .))) %>%
  mutate(glance_model = map(model, glance)) %>%
  unnest(glance_model) %>%
  dplyr::select(gene, p.value)
covars_sex_cat_pop_lm[which(covars_sex_cat_pop_lm$p.value < 0.05), ]
```

```
## # A tibble: 1 x 2
##   gene                p.value
##   <chr>                 <dbl>
## 1 ENSG00000164308.12 0.00678
```

Including covariates:

```
library(lmtest)

lr_likelihood <- function(y, x_input = NULL){
    n_samples <- length(y)
    X_mx <- cbind(matrix(1, nrow = n_samples, ncol = 1), x_input)
    MLE_beta <- ginv(t(X_mx) %*% X_mx) %*% t(X_mx) %*% y
    y_hat <- X_mx %*% MLE_beta
    var_hat <- sum((y - (y_hat))^2) / (n_samples - 1)
    log_likelihood <- -((n_samples / 2) * log(2 * pi * var_hat) ) - ((1/ (2*var_hat)) * sum((y - (y_hat
    return(log_likelihood)
}

LRT_test <- function(logl_H0, logl_HA, df_test){

    LRT<-2*logl_HA-2*logl_H0 #likelihood ratio test statistic
    #likelihood ratio test statistic for every genotype
    pval <- pchisq(LRT, df_test, lower.tail = F)
    return(pval)
}

set.seed(2018)
x = sample(c(-1,0,1), 100, replace = TRUE)
y = 0.9 * x + rnorm(100)
h0_nocovar <- lr_likelihood(y)
h1_nocovar <- lr_likelihood(y, x)
LRT_test(h0_nocovar, h1_nocovar, df_test = 1)
```

```
## [1] 1.319385e-11
```

```
x_c = sample(c(0,1), 100, replace = TRUE)
y2 = y + 0.8 * x_c

h0_withcovar <- lr_likelihood(y2)
h1_withcovar <- lr_likelihood(y2, x)
LRT_test(h0_withcovar, h1_withcovar, df_test = 1)
```

```
## [1] 2.03641e-11
```

```
h0_includecovar <- lr_likelihood(y2, x_c)
ha_includecovar <- lr_likelihood(y2, cbind(x,x_c))
LRT_test(h0_includecovar, ha_includecovar, df_test = 1)
```

```
## [1] 1.592204e-11
```

```
# y2 = y, x_c = covar, x = x

X_mat_null <- cbind(1, x_c)
beta_hat_null <- ginv(t(X_mat_null) %*% X_mat_null) %*% t(X_mat_null) %*% y2
X_mat_alt <- cbind(1, x, x_c)
beta_hat_alt <- ginv(t(X_mat_alt) %*% X_mat_alt) %*% t(X_mat_alt) %*% y2
y_hat_null <- X_mat_null %*% beta_hat_null
y_hat_alt <- X_mat_alt %*% beta_hat_alt
SSE_null <- sum((y2 - y_hat_null) ^ 2)
SSE_alt <- sum((y2 - y_hat_alt) ^ 2)
fstat <- ((SSE_null - SSE_alt) / 2) / (SSE_alt / (length(x) - 3))

pf(fstat, 2, length(x) - 3, lower.tail = FALSE)
```

```
## [1] 2.714078e-10
```

```
# testing with real data

X_mat_null <- cbind(matrix(1, nrow = nrow(x_a), ncol = 1), NULL)
beta_hat_null <- ginv(t(X_mat_null) %*% X_mat_null) %*% t(X_mat_null) %*% pheno[, 2]
X_mat_alt <- cbind(1, x_a[, 2], x_d[, 2])
beta_hat_alt <- ginv(t(X_mat_alt) %*% X_mat_alt) %*% t(X_mat_alt) %*% pheno[, 2]
y_hat_null <- X_mat_null %*% beta_hat_null
y_hat_alt <- X_mat_alt %*% beta_hat_alt
SSE_null <- sum((pheno[, 2] - y_hat_null) ^ 2)
SSE_alt <- sum((pheno[, 2] - y_hat_alt) ^ 2)
fstat <- ((SSE_null - SSE_alt) / 2) / (SSE_alt / (nrow(geno) - 3))
pf(fstat, 2, nrow(geno) - 3, lower.tail = FALSE)
```

```
## [1] 0.8845043
```