Analysis of Lymphoblastoid Cell Lines (LCL) mRNA Levels from Four European Populations Based on GWAS Models

BTRY 4830 Final Project Report

Zhongyi (James) Guo

5/10/2022

Because of length limit, most code will be hidden from the report. Please refer to the attached RMD file for source code or result replication.

Introduction

In this project, I performed GWAS analysis on the lymphoblastoid cell lines (LCL) mRNA levels quantified through RNA sequencing from 4 different European populations. I also analyzed genotype and phenotype data by employing two different strategies: excluding covariates or including covariates with 50,000 of the SNP genotypes for 344 samples from the `CEU` (Utah residents with European ancestry), `FIN` (Finns), `GBR` (British) and, `TSI` (Toscani) population.

I am interested in studying if population and gender could cause the expression of lymphoblastoid cell lines (LCL) mRNA levels to be different. Hereby, I want to raise my research question: - Would population and gender as covariates influence the GWAS analysis result?

Data Cleaning and Exploratory Data Analysis (EDA)

Here is the link to the data source: Genetic European Variation in Health and Disease (gEUVADIS) (http://www.internationalgenome.org/data-portal/data-collection/geuvadis/)

The first step is to import all the datasets we need and rename the row names of each dataset.

I noticed all data in `covars` are of type string, which can cause trouble for downstream analysis. I will convert them to integer categorical data.

Here is a documentation of `covars` data: - `Population`: `GBR`: 3; `FIN`: 2; `CEU`: 1; `TSI`: 4 - `Sex`: `MALE`: 1; `FEMALE`: 0

Then, I will check if any NA values or infinite values are present in each dataset.

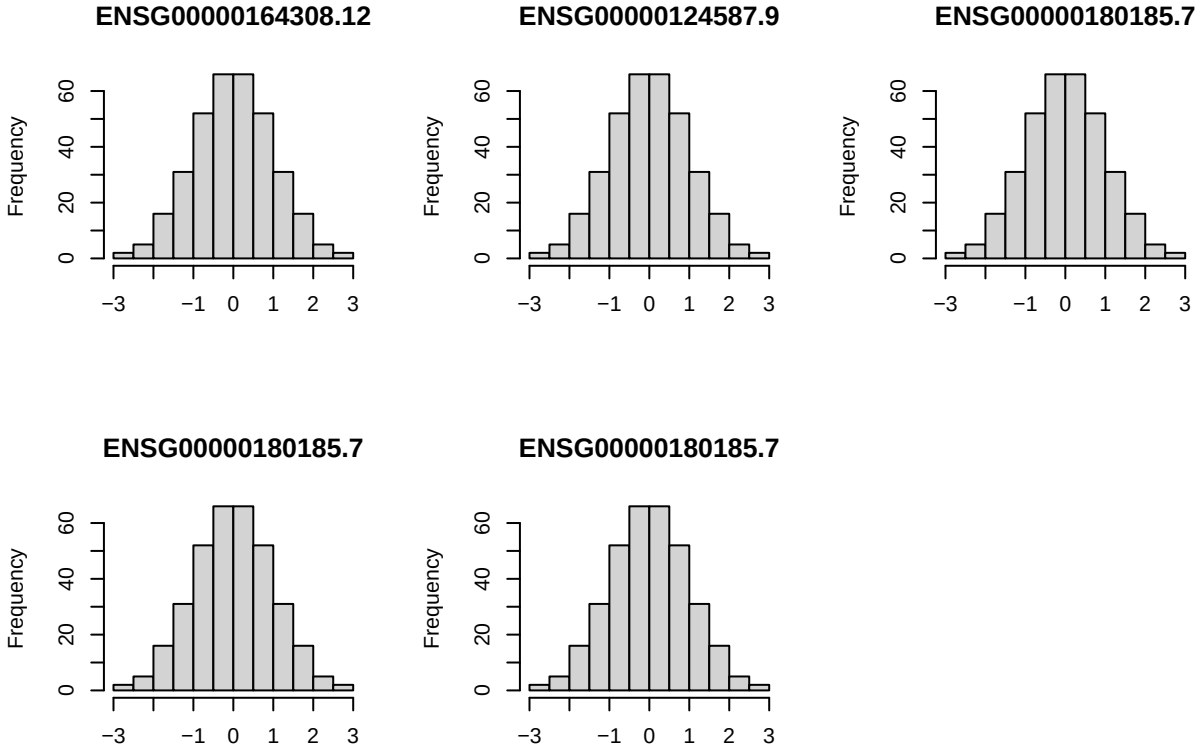
```
## [1] "Number of NA values or infinite values in genotype: 0"
```

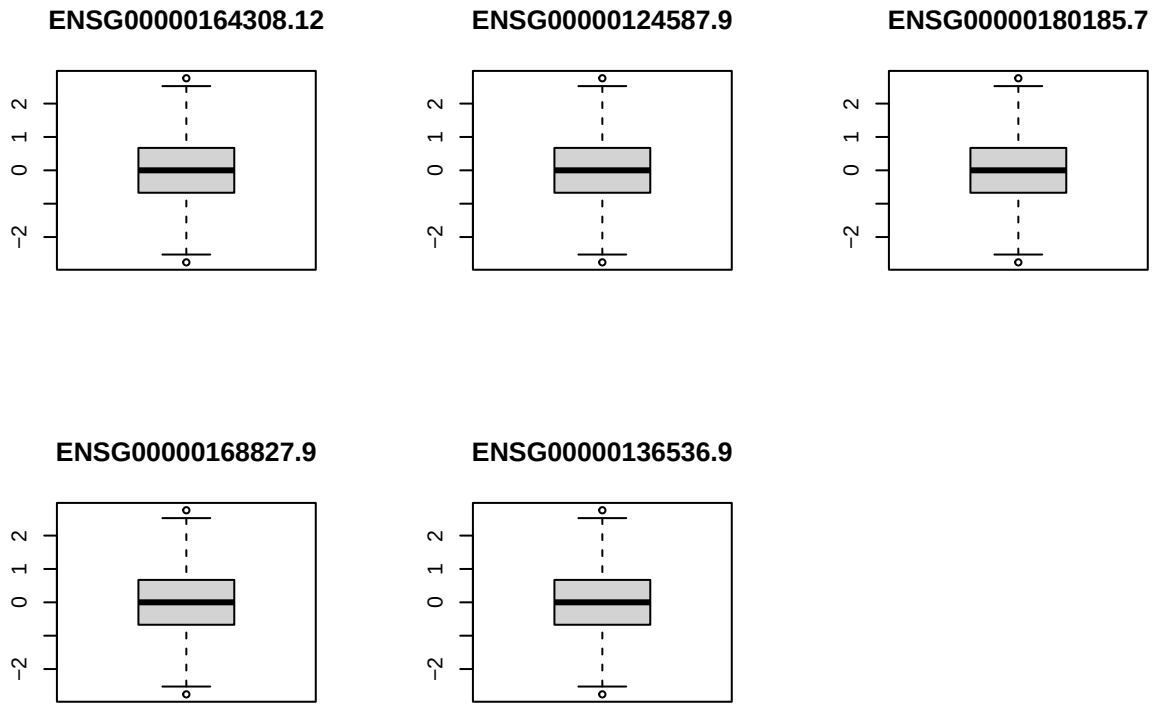
```
## [1] "Number of NA values or infinite values in covariates: 0"
```

```
## [1] "Number of NA values or infinite values in phenotype: 0"
```

There are no NA values or infinite values in `geno`, `covars`, and `pheno`.

Next, I will check if phenotype data are normally distributed using histogram and see if there are odd phenotypes or outliers, which need to be removed, using boxplot.





All phenotypes are (approximately) normally distributed, and no odd phenotypes or outliers were detected. I think the data are ready for downstream analysis.

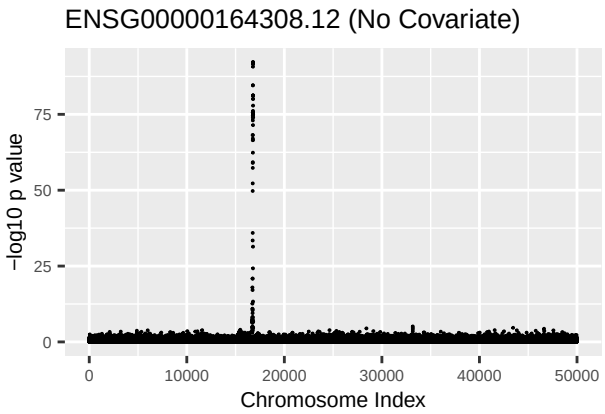
```
##  tidyverse data.table       MASS   ggthemes
##       TRUE       TRUE       TRUE       TRUE
```

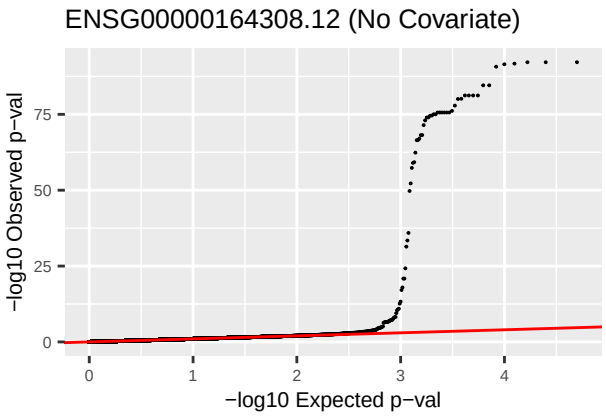
Model Building, Data Visualization, and Step-by-Step Analysis

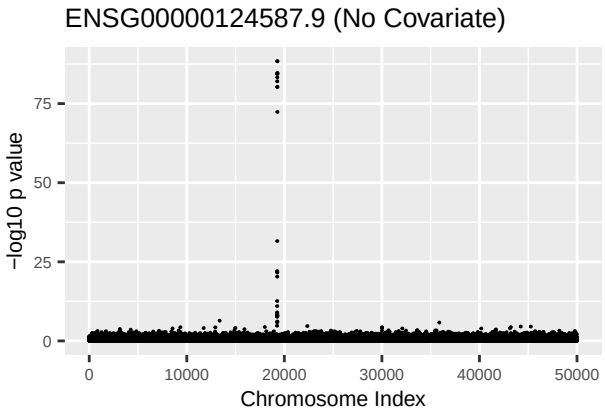
I will calculate `xa_matrix` and `xd_matrix` respectively from `geno` data.

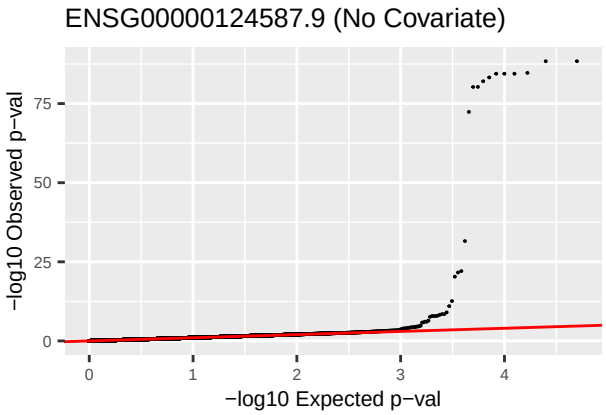
Linear Model excluding covariates

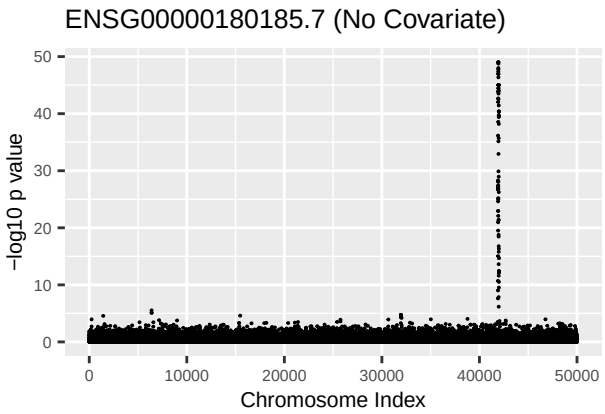
I will continue to calculate the results of each phenotypes with no covariates, and then create Manhattan plots and QQ plots for each of the phenotype.

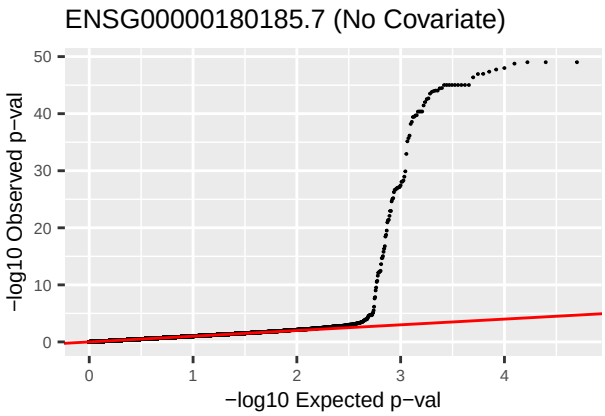

ENSG00000164308.12 (No Covariate)
−log10 p value
0
25
50
75
0
10000
20000
30000
40000
50000
Chromosome Index

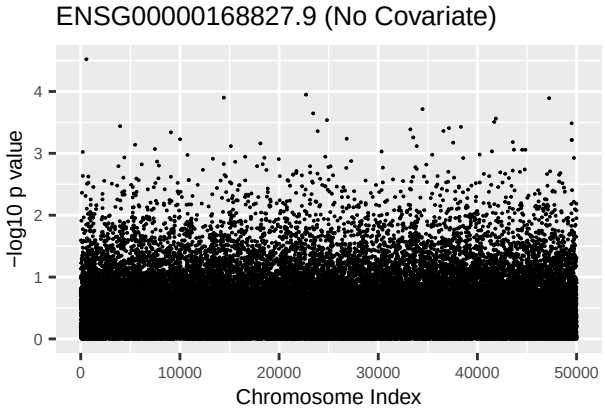

ENSG00000164308.12 (No Covariate)
−log10 Observed p−val
0
25
50
75
0
1
2
3
4
−log10 Expected p−val

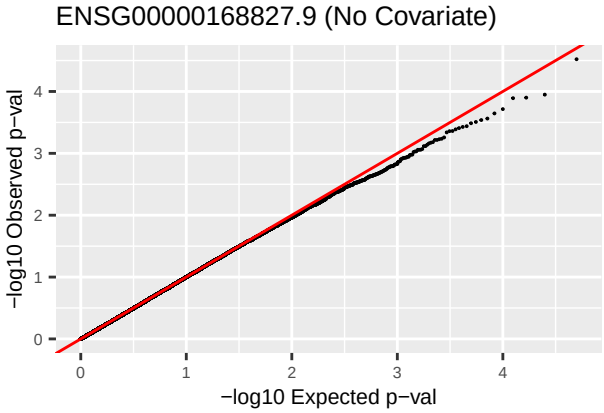

ENSG00000124587.9 (No Covariate)
−log10 p value
0
25
50
75
0
10000
20000
30000
40000
50000
Chromosome Index

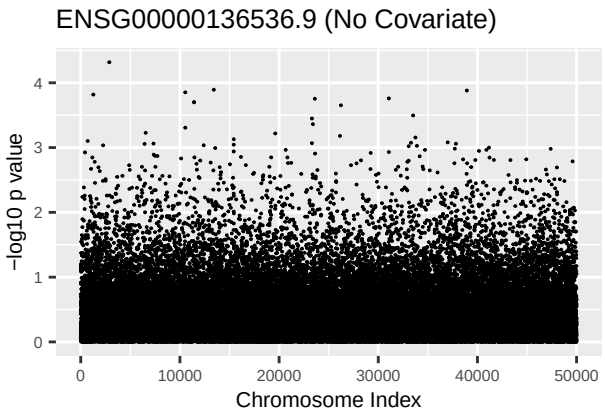

ENSG00000124587.9 (No Covariate)
−log10 Observed p−val
0
25
50
75
0
1
2
3
4
−log10 Expected p−val

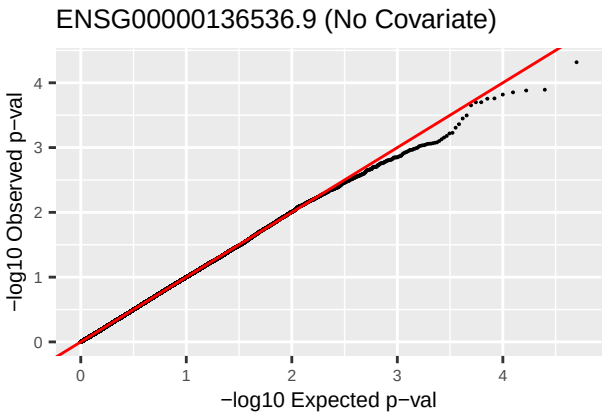

ENSG00000180185.7 (No Covariate)
−log10 p value
Chromosome Index
0
10000
20000
30000
40000
50000
0
10
20
30
40
50


ENSG00000180185.7 (No Covariate)
−log10 Observed p−val
−log10 Expected p−val
0
1
2
3
4
0
10
20
30
40
50


ENSG00000168827.9 (No Covariate)
−log10 p value
Chromosome Index
0
10000
20000
30000
40000
50000
0
1
2
3
4


ENSG00000168827.9 (No Covariate)
−log10 Observed p−val
−log10 Expected p−val
0
1
2
3
4





For phenotype named `ENSG00000164308.12`, `ENSG00000124587.9`, and `ENSG00000180185.7`, the QQ plots have huge and uplifting tails, which indicate causal polymorphisms. I will locate the chromosome index where the causal mutations that possibly occur.

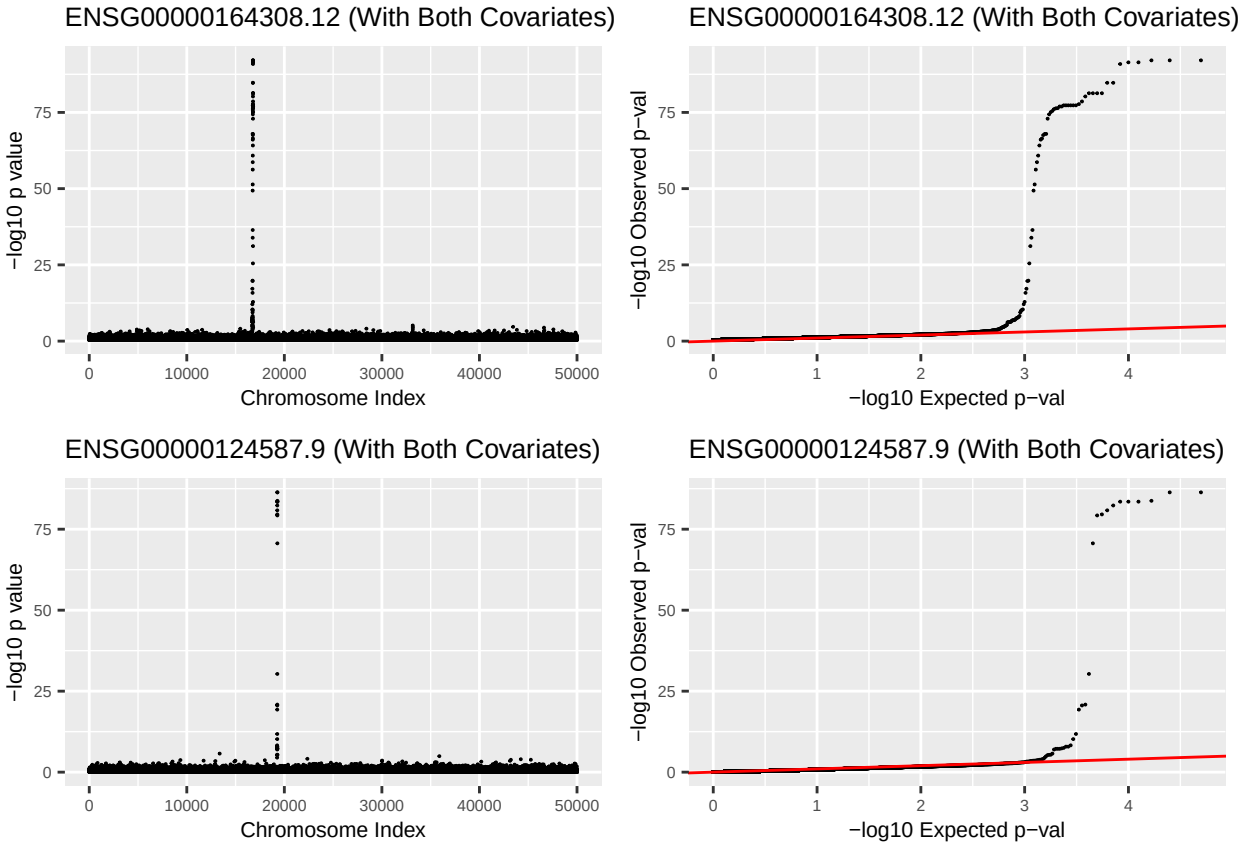
```
## [1] "ENSG00000164308.12: 54 possible causal polymorphism sites"

##  [1] 16728 16729 16731 16745 16746 16747 16751 16752 16755 16757 16758 16760
## [13] 16761 16762 16763 16764 16765 16766 16767 16768 16769 16770 16771 16772
## [25] 16773 16774 16775 16776 16777 16778 16779 16780 16781 16782 16783 16784
## [37] 16785 16786 16787 16788 16789 16790 16791 16792 16793 16794 16795 16796
## [49] 16797 16798 16799 16800 16801 16803

## [1] "ENSG00000124587.9: 17 possible causal polymorphism sites"

##  [1] 19274 19275 19276 19277 19278 19279 19280 19281 19282 19283 19284 19285
## [13] 19286 19287 19288 19289 19290

## [1] "ENSG00000180185.7: 85 possible causal polymorphism sites"

##  [1] 41904 41911 41912 41913 41914 41915 41916 41917 41918 41919 41920 41921
## [13] 41922 41923 41924 41925 41926 41927 41928 41931 41932 41934 41935 41936
## [25] 41937 41938 41939 41940 41941 41942 41943 41944 41945 41946 41947 41948
## [37] 41949 41950 41951 41952 41953 41954 41955 41956 41957 41958 41959 41960
## [49] 41961 41962 41963 41964 41965 41966 41967 41968 41969 41970 41971 41973
## [61] 41974 41975 41976 41977 41979 41980 41981 41982 41983 41984 41985 41986
## [73] 41987 41988 41989 41990 41991 41992 41994 41995 41996 41998 41999 42000
## [85] 42001
```

For phenotype of `ENSG00000168827.9` and `ENSG00000136536.9`, the QQ plots suggest that I should not interpret any of significant p-values as indicating locations of causal polymorphisms. This phenomenon can also possibly occur because we excluded covariates. So the next model we will build includes covariates.

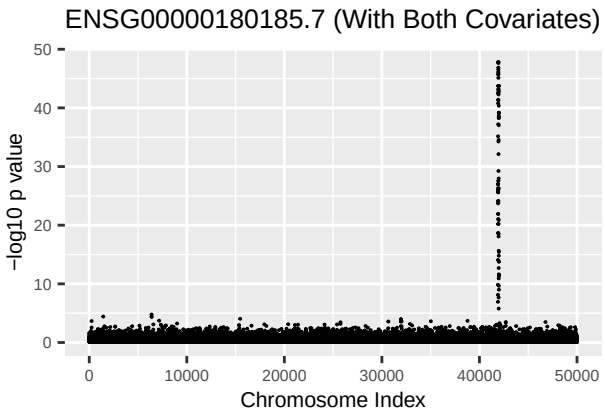
Linear Model including covariates `Population` and `Sex`

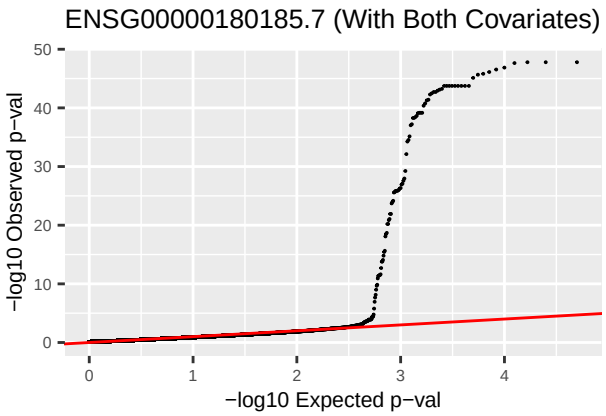
I originally planned to do `Population` or `Sex` separately, however that version of report exceeded the maximum length expected for this project, and also, they do not reveal very significant discovery. So I decided to stop doing `Population` or `Sex` separately but instead, include them both.

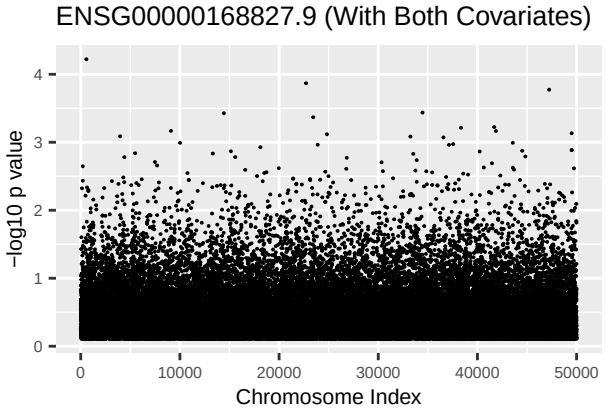
```
## $`1`
```

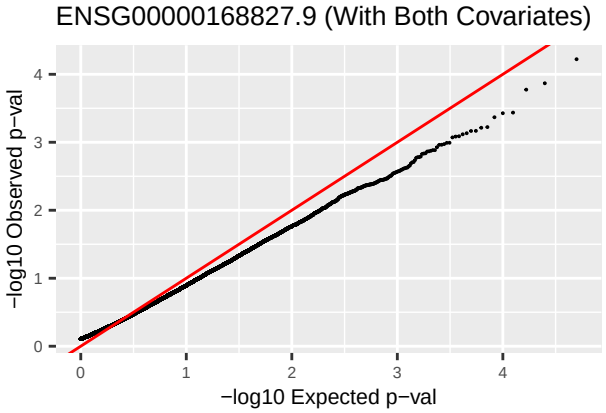



```
##
## $`2`
```

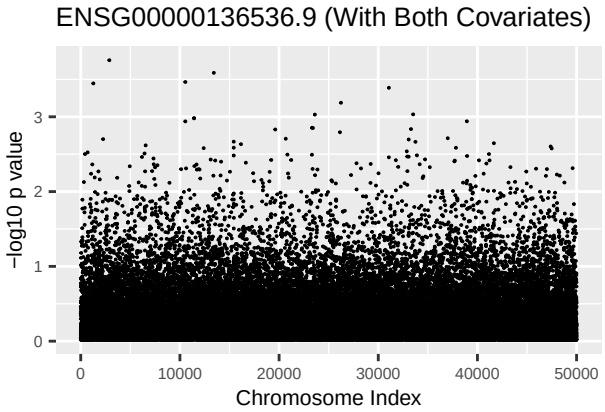

ENSG00000180185.7 (With Both Covariates)
−log10 p value
Chromosome Index
0
10000
20000
30000
40000
50000
0
10
20
30
40
50

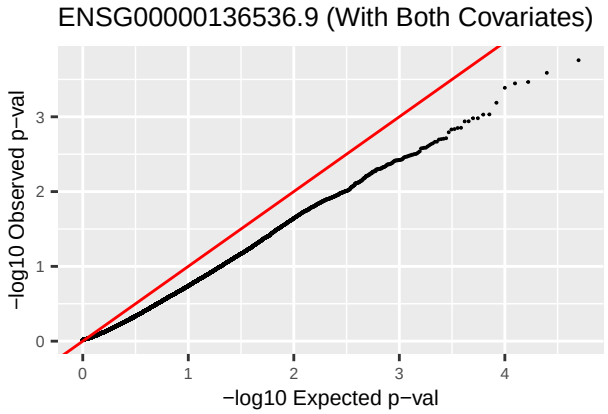

ENSG00000180185.7 (With Both Covariates)
−log10 Observed p−val
−log10 Expected p−val
0
1
2
3
4
0
10
20
30
40
50


ENSG00000168827.9 (With Both Covariates)
−log10 p value
Chromosome Index
0
10000
20000
30000
40000
50000
0
1
2
3
4


ENSG00000168827.9 (With Both Covariates)
−log10 Observed p−val
−log10 Expected p−val
0
1
2
3
4
0
1
2
3
4

```
##
## $`3`
```






```
##
## attr(,"class")
## [1] "list"      "ggarrange"
```

Similarly to the model with no covariates, I observed huge uplifting tails in the QQ plots of phenotypes `ENSG00000164308.12`, `ENSG00000124587.9`, and `ENSG00000180185.7`. That suggests covariates (`Population` and `Sex` do not contribute much to the mRNA levels of lymphoblastoid cell lines (LCL)) among 4 populations. I will print out index and count of possible causal polymorphism sites.

```
## [1] "ENSG00000164308.12: 54 possible causal polymorphism sites"
```

```
##  [1] 16728 16729 16731 16745 16746 16747 16751 16752 16755 16757 16758 16760
## [13] 16761 16762 16763 16764 16765 16766 16767 16768 16769 16770 16771 16772
## [25] 16773 16774 16775 16776 16777 16778 16779 16780 16781 16782 16783 16784
## [37] 16785 16786 16787 16788 16789 16790 16791 16792 16793 16794 16795 16796
## [49] 16797 16798 16799 16800 16801 16803
```

```
## [1] "ENSG00000124587.9: 17 possible causal polymorphism sites"
```

```
##  [1] 19274 19275 19276 19277 19278 19279 19280 19281 19282 19283 19284 19285
## [13] 19286 19287 19288 19289 19290
```

```
## [1] "ENSG00000180185.7: 83 possible causal polymorphism sites"
```

```
##  [1] 41911 41912 41913 41914 41915 41916 41917 41918 41919 41920 41921 41922
## [13] 41923 41924 41925 41926 41927 41928 41931 41932 41934 41935 41936 41937
## [25] 41938 41939 41940 41941 41942 41943 41944 41945 41946 41947 41948 41949
## [37] 41950 41951 41952 41953 41954 41955 41956 41957 41958 41959 41960 41961
## [49] 41962 41963 41964 41965 41966 41967 41968 41969 41970 41971 41973 41974
## [61] 41975 41976 41977 41979 41980 41981 41982 41983 41984 41985 41986 41987
## [73] 41988 41989 41990 41991 41992 41994 41995 41996 41999 42000 42001
```

For `ENSG00000168827.9` and `ENSG00000136536.9`, compared to the QQ plots with no covariates, I saw tails occured, which suggests covariates improve the performance of my GWAS model.

Discussion

There are some possible further research that we can carry out in the future. For example, we can refer to UCSC genome browser for nucleotide/codon letters and then hypothesize the mRNA levels. This will require a lot of biochemistry backgrounds.

Conclusion

I found no significant change before or after including covariates. In my data analysis, `Population` and `Sex` as covariates do not seem to impact the GWAS analysis result.