Project

Yuqi Cheng

2021/5/13

Contents

Background **1**

Loading data & preprocessing **2**

Exploratory analysis **2**

Missing value test 2

Descriptive statistics 2

Minor allele frequency 3

Principal components analysis 4

Genome-wide association studies with covariates **4**

GWAS with no covariate 5

GWAS with Sex as a covariate 5

GWAS with Population as a covariate 6

GWAS with Sex and Population as covariates 7

Summary **8**

Background

Recently, a large-scale human genomics resource *Genetic European Variation in Health and Disease* (gEUVADIS) is published.[1] This research includes samples from 4 different European populations (5 populations total). Each of these individuals was part of the 1000 Genomes project and their genomes were sequenced and analyzed to identify SNP genotypes.

Here, we are provided parts of the gEUVADIS resources, which is 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. Along with the detailed gene and SNP annotation information, we perform a thorough Genome-Wide Association Analysis (GWAS) to these data.

[1] Lappalainen, T., et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506-511 (2013).

Loading data & preprocessing

We first load the five datasets and check the format, then convert genotype matrix to X_a matrix and X_d matrix.

The genotype dataset provided contains 0, 1 and 2, represent A_1A_1, A_1A_2 and A_2A_2 respectively. Our mission is to convert this dataset to X_a (-1: minor allele homo, 0: hetero, 1: major allele homo) and X_d (-1: homo, 1, hetero) matrix.

Exploratory analysis

Before starting the downstream analysis, we first do exploratory analysis on our dataset.

Missing value test

The missing value is very common among various datasets and may lead to a wrong result if it is not treated reasonably. So we first examine if there were some NA values.

```
## The number of the missing value in phenotype dataset is: 0 in genotype is: 0 in covars is: 0
```

No missing value is detected.

Descriptive statistics

It is very important to know the basic statistics and distribution of our data, thus we try to do some descriptive statistics on the 5 genes we used (ERAP2, FAHD1, GFM1, MARCHF7, PEX6).

probe	chromosome	start	end	symbol
ENSG00000136536.9	2	159712456	159768582	MARCH7
ENSG00000180185.7	16	1827223	1840206	FAHD1
ENSG00000124587.9	6	42963872	42979242	PEX6
ENSG00000164308.12	5	96875939	96919702	ERAP2
ENSG00000168827.9	3	158644496	158692571	GFM1

ERAP2	PEX6	FAHD1	GFM1	MARCH7
Min. :-2.7590	Min. :-2.7590	Min. :-2.7590	Min. :-2.7590	Min. :-2.7590
1st Qu.:-0.6699	1st Qu.:-0.6699	1st Qu.:-0.6699	1st Qu.:-0.6699	1st Qu.:-0.6699
Median : 0.0000	Median : 0.0000	Median : 0.0000	Median : 0.0000	Median : 0.0000
Mean : 0.0000	Mean : 0.0000	Mean : 0.0000	Mean : 0.0000	Mean : 0.0000
3rd Qu.: 0.6699	3rd Qu.: 0.6699	3rd Qu.: 0.6699	3rd Qu.: 0.6699	3rd Qu.: 0.6699
Max. : 2.7590	Max. : 2.7590	Max. : 2.7590	Max. : 2.7590	Max. : 2.7590

Then we try to get the detailed information of these 5 genes from NCBI.[2]

- **ERAP2**

[2] Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2020 May 12]. Available from: https://www.ncbi.nlm.nih.gov/gene/

  - *Function*: This gene encodes a zinc metalloaminopeptidase of the M1 protease family that resides in the endoplasmic reticulum and functions in N-terminal trimming antigenic epitopes for presentation by major histocompatibility complex (MHC) class I molecules.
  - *Location*: Chromosome 5

- **PEX6**
  - *Function*: This gene encodes a member of the AAA (ATPases associated with diverse cellular activities) family of ATPases.
  - *Location*: Chromosome 6

- **FAHD1**
  - *Function*: This gene encodes fumarylacetoacetate hydrolase domain containing 1.
  - *Location*: Chromosome 16

- **GFM1**
  - *Function*: This gene encodes one of the mitochondrial translation elongation factors.
  - *Location*: Chromosome 3

- **MARCH7**
  - *Function*: This gene encodes membrane associated ring-CH-type finger 7.
  - *Location*: Chromosome 2

We also look at the format of the `covars` dataset which shows a distribution of the sample population and gender.

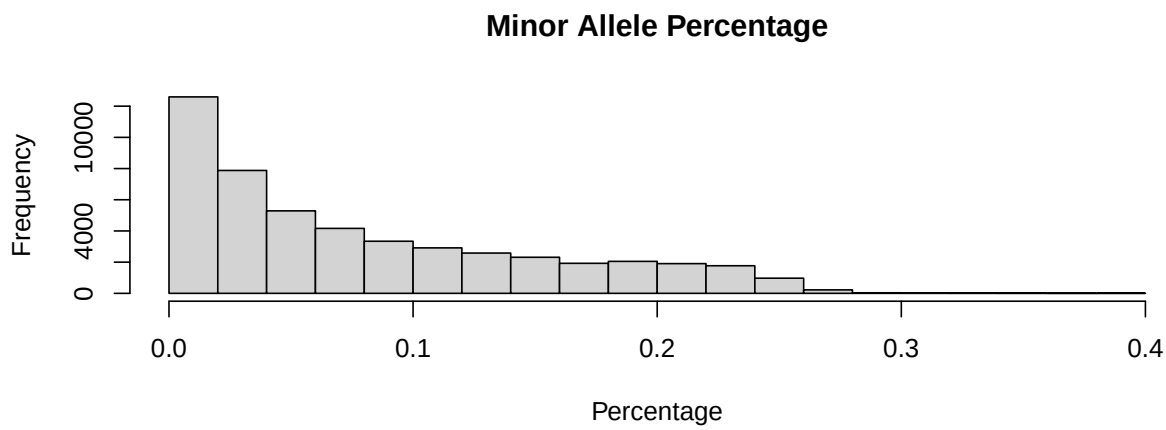
```
## GBR has male sample: 39 and female sample: 46 . Total: 85
## FIN has male sample: 34 and female sample: 55 . Total: 89
## CEU has male sample: 41 and female sample: 37 . Total: 78
## TSI has male sample: 49 and female sample: 43 . Total: 92
```

To make it clear, we show the popluation information here,

Abbr.	**Population**	**Description**
GBR	British	British
CEU	CEPH	Utah residents with European ancestry
FIN	Finnish	Finnish
TSI	Toscani	Toscani

Minor allele frequency

We then calculate the distribution of the minor allele frequency which is critical for the eQTL analysis. The histogram below shows the distribution of the minor allele frequency. We set the threshold at 0.05 and remove the SNP which maf<0.05.

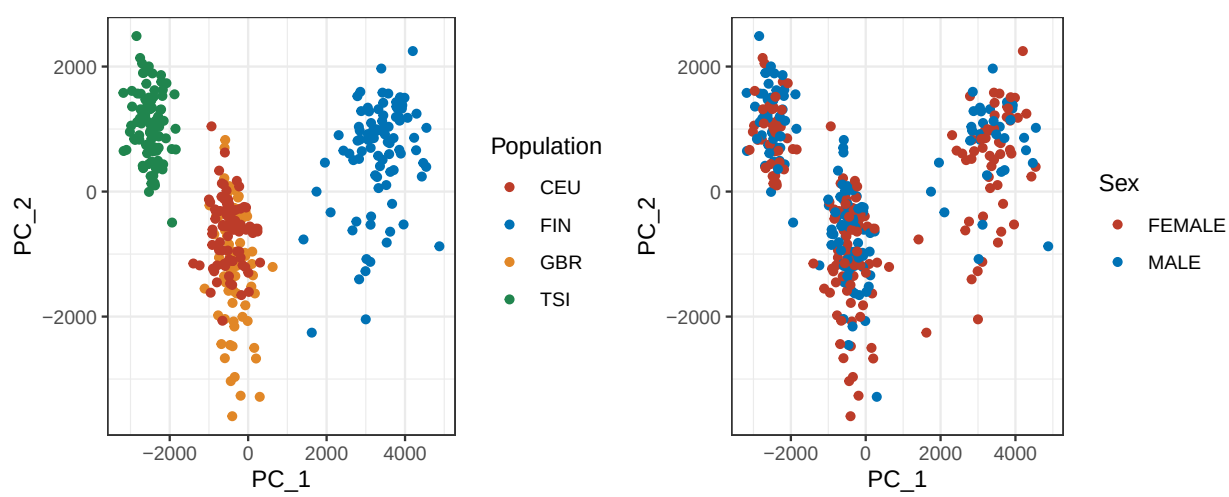



```
## Dimension of the matrix (trimmed) is: 344 26315
```

Principal components analysis

In order to know more about our data, We use trimmed X_a matrix to run PCA and obtain the figure below. From the PCA plot, we can find that CEU and GBR have a near distance, which may indicate that Utah residents with European ancestry population and British poplutaion have a more similar genetic pattern (or we can say a kind of relationship). However, with the result shown in the right panel, we can say that gender is not a factor to distinguish the human genetics resource.

Meanwhile, we also test the effect of our preprocessing. We firstly perform a PCA without removing SNPs whose MAF fell below the 0.05 threshold, and the clustering result is much spread out. Then we run PCA with the trimmed data, after removing SNPs with low MAFs, the clusters are much more confined (See the figure below). The result indicates the importance of preprocessing the data to remove SNPs with low MAFs.



Genome-wide association studies with covariates

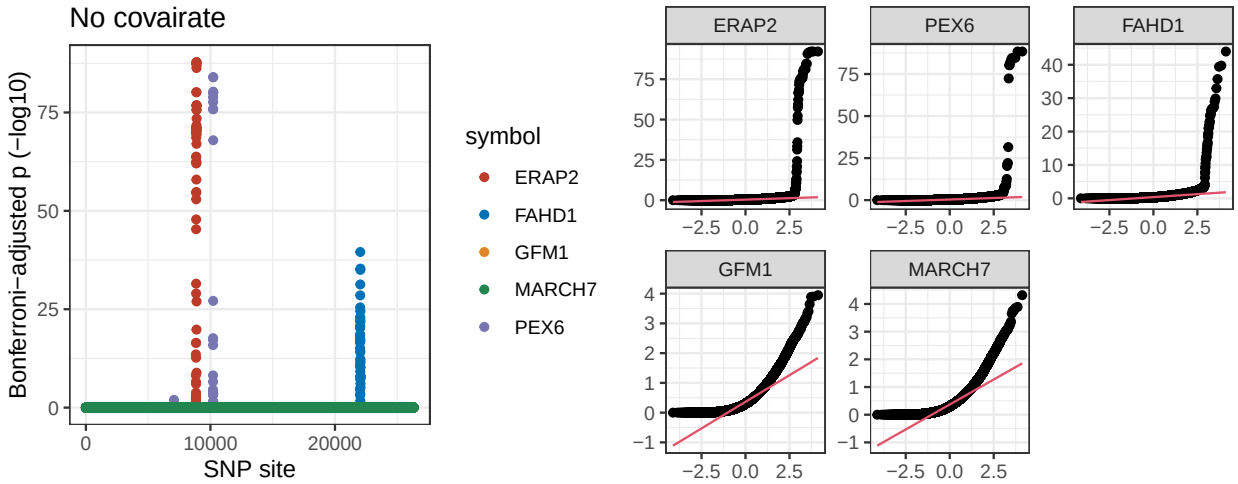
After completing the descriptive analysis, we have obtained the general insight of our data. Here, we start using our trimmed data to do GWAS with covariates. Different from what we used in the lab, we apply the

R base function `lm()` to calculate the p-value. To make our analysis more comprehensive, we try GWAS with different covariates, including no covariate, single covariate, and combined covariate.

For each of our tests, we generate a Manhattan plot and QQ-plot, the Manhattan plot is to visualize the significant SNP site and the QQ plot is to make sure that the significant SNP can be considered as potential causal polymorphisms.

GWAS with no covariate

Firstly, we try GWAS without any covariates, we only use X_a matrix and X_d matrix as the varibles to do linear regression. We use `lm()` to run the regression and `pf()` to do F test.

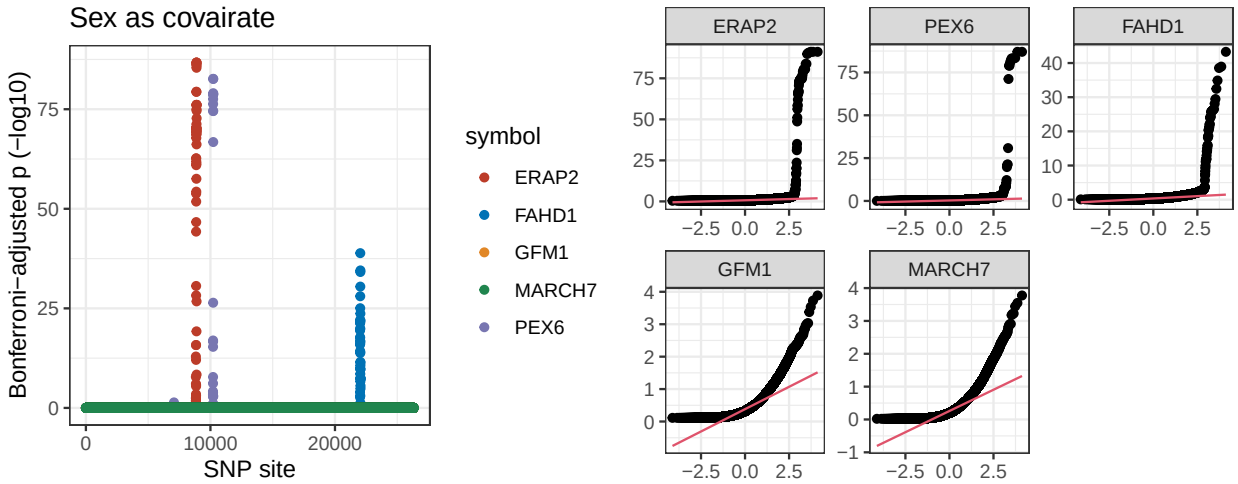


From the Manhattan plot, we can find that ERAP2, FAHD1, and PEX6 genes have potential sites of causal polymorphisms, however, GFM1 and MARCHF7 genes do not show significant potential sites. Based on the Bonferroni method, the number of potential causal polymorphisms of each gene is:

```
## ERAP2 has 64 potential signficant polymorphisms.
## PEX6 has 29 potential signficant polymorphisms.
## FAHD1 has 45 potential signficant polymorphisms.
## GFM1 has 0 potential signficant polymorphisms.
## MARCH7 has 0 potential signficant polymorphisms.
```

GWAS with Sex as a covariate

Then we add Sex as a covariate and follow the same pipeline with no covariate GWAS.

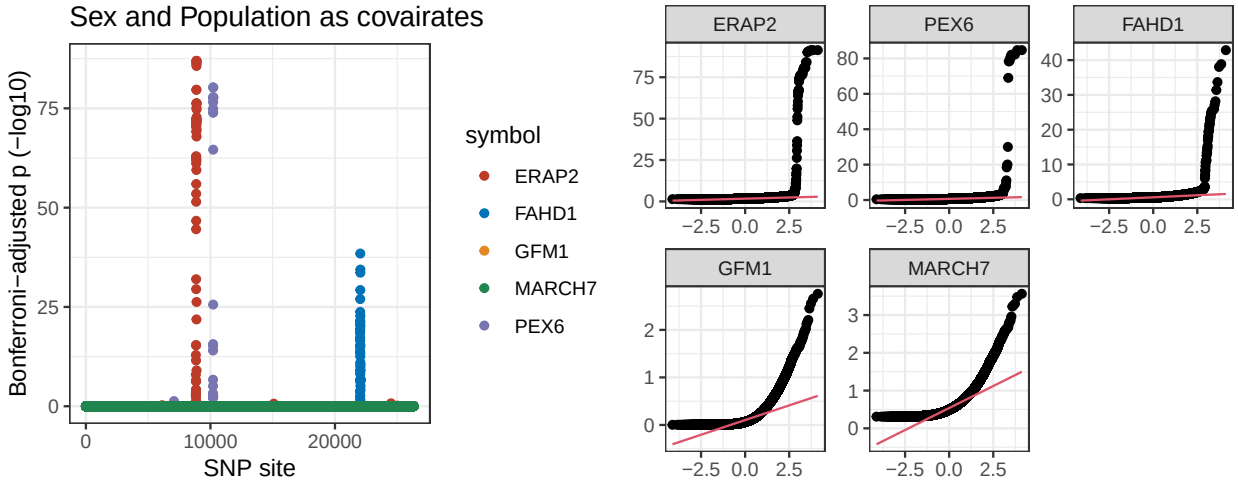


Same with what we get in section 4.1, we can find that ERAP2, FAHD1, and PEX6 genes have potential sites of causal polymorphisms, however, GFM1 and MARCHF7 genes do not show significant potential sites. Based on the Bonferroni method, the number of potential causal polymorphisms of each gene is:

```
## ERAP2 has 64 potential signficant polymorphisms.
## PEX6 has 27 potential signficant polymorphisms.
## FAHD1 has 44 potential signficant polymorphisms.
## GFM1 has 0 potential signficant polymorphisms.
## MARCH7 has 0 potential signficant polymorphisms.
```

GWAS with Population as a covariate

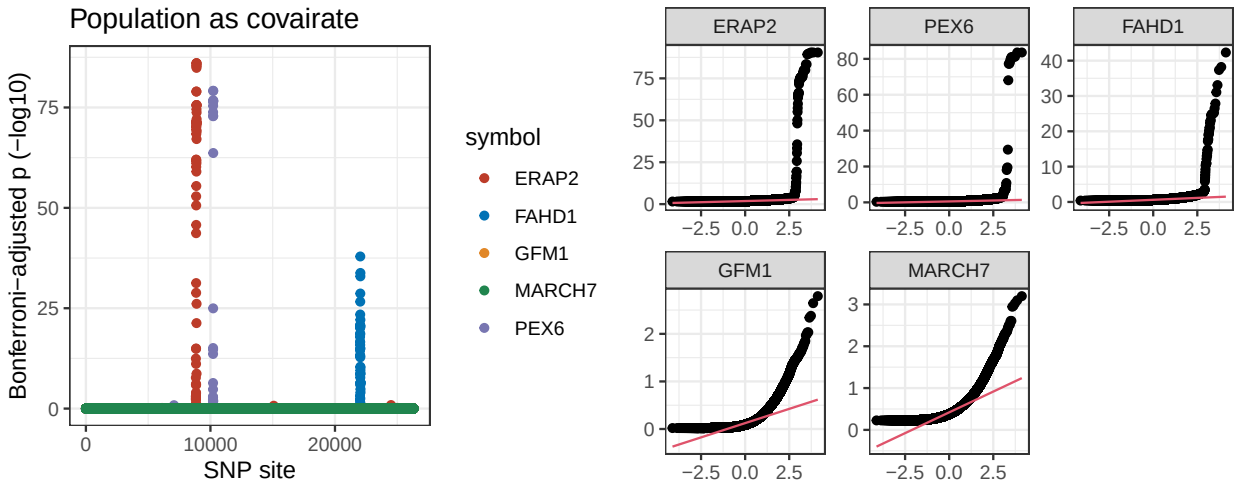
Following the multiple linear regression with population as a covariate and with multiple hypothesis correction via the Bonferroni method, we can find that ERAP2, FAHD1, and PEX6 genes have potential sites of causal polymorphisms, however, GFM1 and MARCHF7 genes do not show significant potential sites.




```
## ERAP2 has 66 potential signficant polymorphisms.
## PEX6 has 26 potential signficant polymorphisms.
## FAHD1 has 45 potential signficant polymorphisms.
## GFM1 has 0 potential signficant polymorphisms.
## MARCH7 has 0 potential signficant polymorphisms.
```

GWAS with Sex and Population as covariates

After testing the 1 covariate situation, we choose Sex and Population togethor as covariates to do the multiple linear regression and multiple hypothesis correction (Bonferroni method).

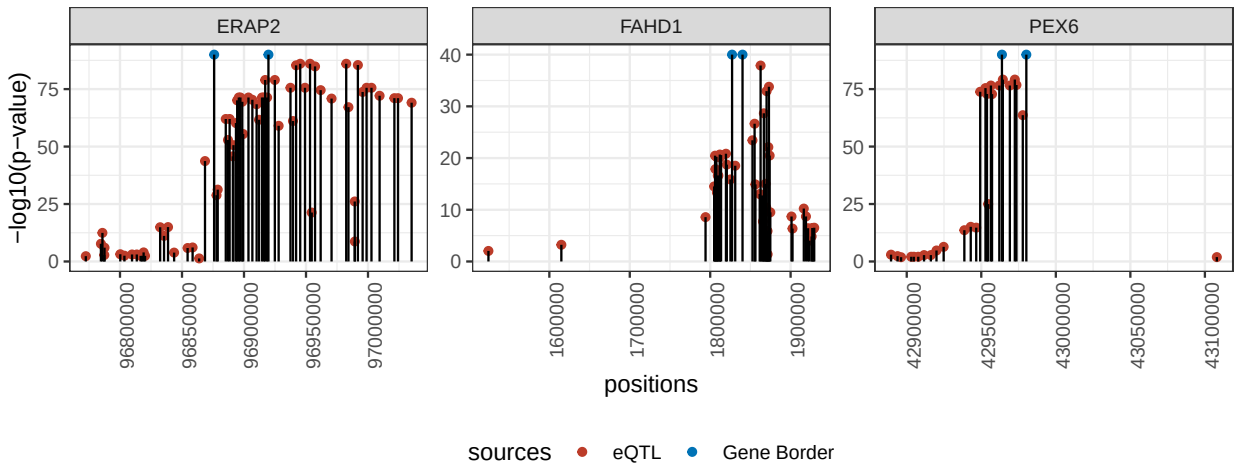



```
## ERAP2 has 66 potential signficant polymorphisms.
## PEX6 has 26 potential signficant polymorphisms.
## FAHD1 has 45 potential signficant polymorphisms.
## GFM1 has 0 potential signficant polymorphisms.
## MARCH7 has 0 potential signficant polymorphisms.
```

The QQ plots above confirm that the significant SNP sites detected for ERAP2, FAHD1, and PEX6 can be interpreted as potentially causal polymorphisms. We can see from the QQ plots of ERAP2, FAHD1 and PEX6, most of the p-values follow a uniform distribution (on the red line) and do not deviate from this line until the end. In contrast,however, the QQ-plots for GFM1 and MARCHF7 show fewer significant p-values than would normally be expected, which is corresponding with the result no significant causal polymorphisms found for either GFM1 or MARCHF7.

To further understand the exact location of the potential causal polymorphisms, we draw the distribution of the causal polymorphisms based on their location on the chromosome. We use the information from `SNP_info.csv` to locate the SNPs and also use the information from `gene_info.csv` to show the gene border.

Not all potential causal polymorphisms locate in the target gene regions, which may indicate that there has a linkage disequilibrium effect. The causal polymorphisms affect the target gene by affecting its linked gene directly. For example, for ERAP2, there might be a linked gene (nearby or regulatory-related) that is in linkage disequilibrium with the ERAP2 gene. The same situation also happens with FAHD1 and PEX6, we should pay more attention to their linked gene when we try to do analysis on these genes, especially their nearby genes and regulatory network.



Summary

In this project, we practice a completed GWAS with a covariates analysis pipeline. We first convert the raw data (genotypes) to X_a and X_d matrix before starting the analysis. In order to know more about our data, we then conduct the exploratory analysis (missing value test, descriptive statistics, MAF filter, and PCA). To limit the influence of the NA value, the dataset should be tested if there are many missing values, so we use `is.na` function to do this. Then we run the descriptive analysis on our datasets, we mainly use `summary` and `unique` functions to do the analysis. From the descriptive statistics, we obtain a basic insight into what our data can give us and how do we use our data. To prepare for the downstream analysis, we do MAF filtering to remove the SNP which maf<0.05 because MAFs that are too low result in a reduction in the statistical power of the test. Finally, PCA is applied to illustrates the importance of pre-processing the data to remove SNPs with low MAFs and show the potential effect of the covariates factors.

Then we conducted GWAS with covariates. We divide our analysis into three parts: No covariates, with 1 covariates and with 2 covariates. We have seperately shown the figures in our report. Our analysis shows that ERAP2, FAHD1, and PEX6 genes have potential sites of causal polymorphisms, however, GFM1 and MARCHF7 genes do not show significant potential sites.

In general, the inclusion of covariates is associated with a reduction in the observation of false positive causal polymorphisms. In our analysis, however, we find more potential causal polymorphisms are added into the ERAP2 gene when we apply GWAS with 2 covariates, which may indicates that we need to pay more attention on the false positive causal SNPs when we do research on ERAP2.

For FAHD1, the inclusion of sex as a covariate eliminate a single significant causal polymorphism (from 45 to 44). This did not apply when both sex and population are used as covariates. This result may shows that the causal polymorphisms do not have much false positive.

For PEX1, it shows a normal reaction to the inclusion of more covariates. The number of the causal polymorphisms decrease from 29 to 26.

In the end, we try to draw the distribution of the causal polymorphisms based on their location on the chromosome. We find that not all potential causal polymorphisms locate in the target gene regions, which may indicate that there has a linkage disequilibrium effect. The causal polymorphisms affect the target gene by affecting its linked gene directly.