PHARMACOGENOMICS OF HUMAN LEUKOCYTES ANTIGEN (HLA) VARIANTS IN 4 ASIAN GROUPS

Introduction

The HLA system, also known as the human version of the major histocompatibility complex (MHC) that is found in many animals, is a gene complex located on the short arm of Chromosome 6. These polymorphic genes code for HLA molecules which are primarily responsible for presenting processed peptide antigens.

HLA's have multiple other responsibilities within the human body. HLA Class I group, present peptides from inside the cell. HLA Class II presents antigens from outside of the cell to T-lymphocytes, whilst HLA corresponding to the MHC Class III encodes components of the complement system.

Clinically, the HLA system is important in hematopoietic stem cell transplantation and is also associated with certain diseases such as cancer, type 1 diabetes, systemic lupus erythematosus. In cancer, although the HLA often plays a protective role, it has been seen to exhibit both pro and anti cancer properties.

Relevance

The HLA loci are some of the most genetically variable loci in mammals. Hence, this project aims to compare the HLA variants in 4 different Asian population groups - Dai (CDX), Han (CHB), and Southern Han Chinese (CHS) and Vietnamese (KHV). The results from this analysis infers possible biological implications associated with the identified Asian HLA variants particularly in drug response.

Contributors

This project was executed by a team of 6, namely Ismat, Damilola, Omotoyosi, Oreoluwa, Tolani and Mayowa

Methodology

Data Collection

A tab delimited file containing the ID of each sample and the population code was downloaded directly from the complete 1000 genomes database as well as binary plink files asia.bim, asia.bed.gz

& asia.fam. The dataset was downloaded directly from this github repository using the '**wget**' command on the linux terminal and the compressed dataset "asia.bed.gz" was unzipped using the '**gunzip**' command.

The complete 1000 genome sample dataset is a large database of different human genetic variation obtained from 26 populations representing Europe, East & South Asia, West Africa, and America e.t.c.

Principal Component Analysis (PCA)

This is done to decompose the structure of the data and identify the different populations in the data. PCA was used to visualize the data into readable and pictorial 2D plots to identify the different populations and view clearly to what extent the 4 Asian populations within our genome dataset vary or intercept. The first step was to generate eigenvalues by running the plink command below . Eigenvalue shows the importance of the direction of spread within the data.

```
plink --bfile asia --pca
```

During the analysis the chr-set and no-xy parameters were not used as our samples are human chromosomes, which plink is preset on.

To create a PCA plot, the eigenvalues were downloaded into a PC then imported to RStudio. After specifying the directory containing the datasets, we set eigenvec to pca1. Since eigenvec is separated into multiple columns and does not have a header, this command was used:

```
pca1 <- read.table("plink.eigenvec",sep=" ",header=F).
```

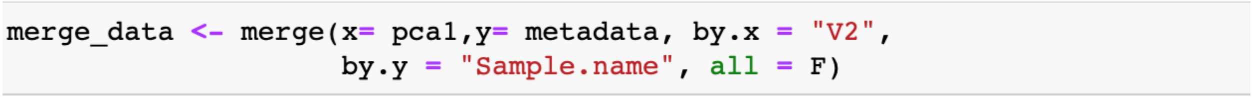
Using library("ggplot2") to load ggplot, we created a preliminary plot with pca1 using the default parameters.

```
ggplot(data=pca1, aes(V3,V4)) + geom_point()
```

To explain the properties of the 1000 genomes list, a metadata table was created using this command

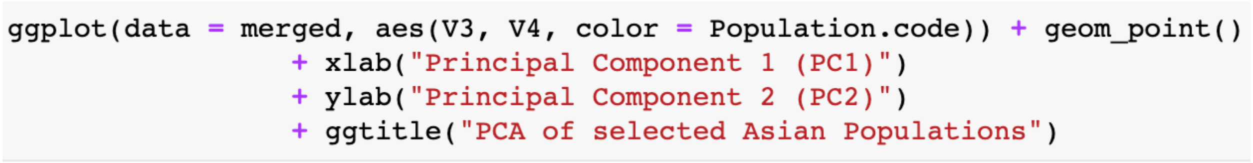
```
metadata <- read.table("complete_1000_genomes_sample_list_.tsv",
                       sep="\t",header=TRUE)
```

The next step was to merge pca1 and metadata using a common column in both dataset.



```
merge_data <- merge(x= pca1,y= metadata, by.x = "V2",
                    by.y = "Sample.name", all = F)
```

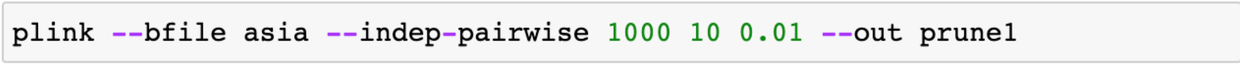
This was done to highlight the Asian populations in the complete 1000 genome list. To generate a final PCA plot and color by population, we ran the command below:



```
ggplot(data = merged, aes(V3, V4, color = Population.code)) + geom_point()
                + xlab("Principal Component 1 (PC1)")
                + ylab("Principal Component 2 (PC2)")
                + ggtitle("PCA of selected Asian Populations")
```

Multidimensional Scaling (MDS) Analysis

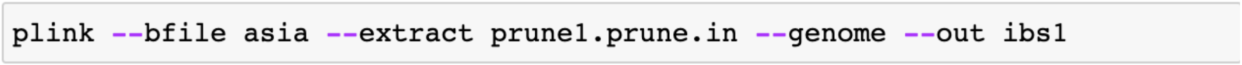
We performed this analysis in a linux terminal using plink.



```
plink --bfile asia --indep-pairwise 1000 10 0.01 --out prune1
```

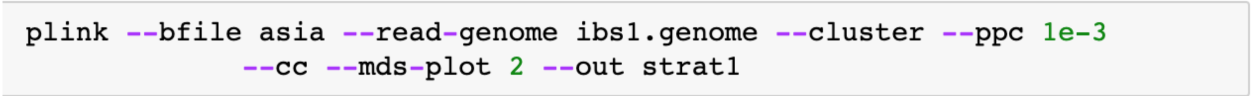
We created a pruned set of markers that are not highly correlated using whole genome SNP binary fileset (asia.bed, asia.bim, asia.fam) as the input . The set filtering values removes any SNP that has r-squared > 0.01 with another SNP within a 1000-SNP window; this window is shifted across the chromosome 10 SNPs at a time.

We then calculated genome-wide identity by descent score (allelic similarity) on the pruned marker list using:



```
plink --bfile asia --extract prune1.prune.in --genome --out ibs1
```

Finally, using the previous .ibs result, we performed population stratification by clustering individuals into homogeneous groups and performing multidimensional scaling analysis. To place constraints on the clusters, we used Pairwise Population Concordance (PPC) test in the command



```
plink --bfile asia --read-genome ibs1.genome --cluster --ppc 1e-3
             --cc --mds-plot 2 --out strat1
```

To visualize the MDS analysis, MDS component 1 (C1) was plotted against MDS Component 2 (C2) from the strat1.mds file using ggplot in RStudio. After setting the right working directory and launching ggplot2, we set strat1.mds as mdsdata using

```r
mdsdata <- read.table("strat1.mds", header = TRUE)
```

Next, we created a metadata table as earlier stated and merged mdsdata with metadata using

```r
merged_mds <- merge(x = mdsdata, y = metadata, by.x = "FID",
                    by.y = "Sample.name", all = F)
```

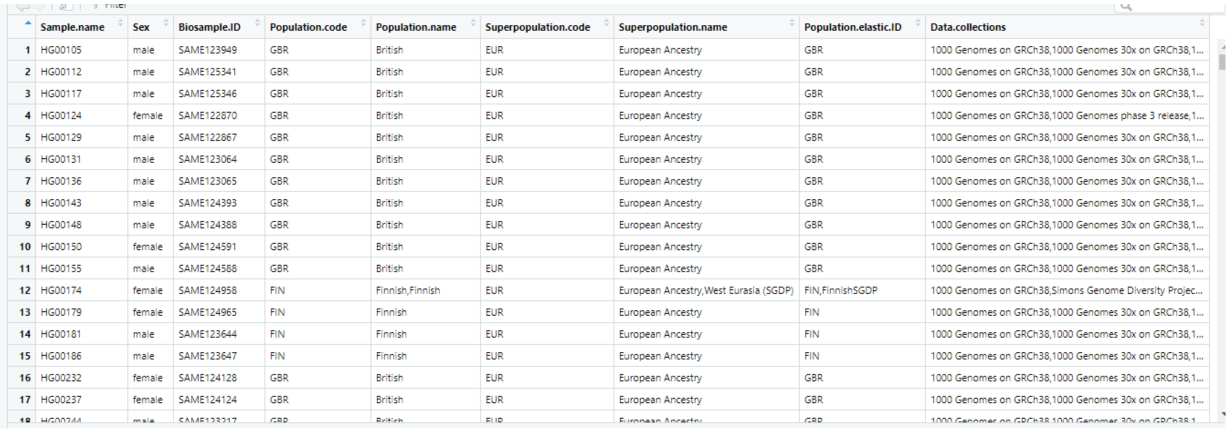
Finally, we created a scatterplot color-coded by population codes using the command

```r
ggplot(data=merged_mds, aes(C1,C2, color = Population.code))
       + geom_point() + ggtitle("Multidimensional Scaling Analysis")
```

Results and Discussion

Metadata

Metadata provides simplified details on the structure, nature and context of a dataset. Here, the metadata table was gotten from the complete 1000 genomes list and clearly shows attributes of the data sorted into ; sample name, sex, biosample ID, population code, population name, superpopulation name, superpopulation code, population elastic ID and data collection.
This table was instrumental in principal component analysis and selecting columns during color plotting by population.



	Sample.name	Sex	Biosample.ID	Population.code	Population.name	Superpopulation.code	Superpopulation.name	Population.elastic.ID	Data.collections
1	HG00105	male	SAME123949	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
2	HG00112	male	SAME125341	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
3	HG00117	male	SAME125346	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
4	HG00124	female	SAME122870	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes phase 3 release,1...
5	HG00129	male	SAME122867	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
6	HG00131	male	SAME123064	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
7	HG00136	male	SAME123065	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
8	HG00143	male	SAME124393	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
9	HG00148	male	SAME124388	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
10	HG00150	female	SAME124591	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
11	HG00155	male	SAME124588	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
12	HG00174	female	SAME124958	FIN	Finnish,Finnish	EUR	European Ancestry,West Eurasia (SGDP)	FIN,FinnishSGDP	1000 Genomes on GRCh38,Simons Genome Diversity Projec...
13	HG00179	female	SAME124965	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
14	HG00181	male	SAME123644	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
15	HG00186	male	SAME123647	FIN	Finnish	EUR	European Ancestry	FIN	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
16	HG00232	female	SAME124128	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
17	HG00237	female	SAME124124	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1...
18	HG00244	male	SAME123217	GBR	British	EUR	European Ancestry	GBR	1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1

Table 1: Metadata

Principal Component Analysis

Principal component analysis (PCA) is one of the most useful tools for population stratification. In this project, we carried out PCA on data from the 4 Asian populations using Plink and RStudio.

The plink analyses yielded the eigenvalues and eigenvectors of 20 principal components. In this analysis, all eigenvalues were greater than 1 and thus they all fulfilled the Kaiser Criterion. The eigenvectors with 2 highest eigenvalues (V3 and V4) were used to make a PCA plot (Figure 1) of the different populations with both accounting for approximately 17.9% of the total variation within the populations.

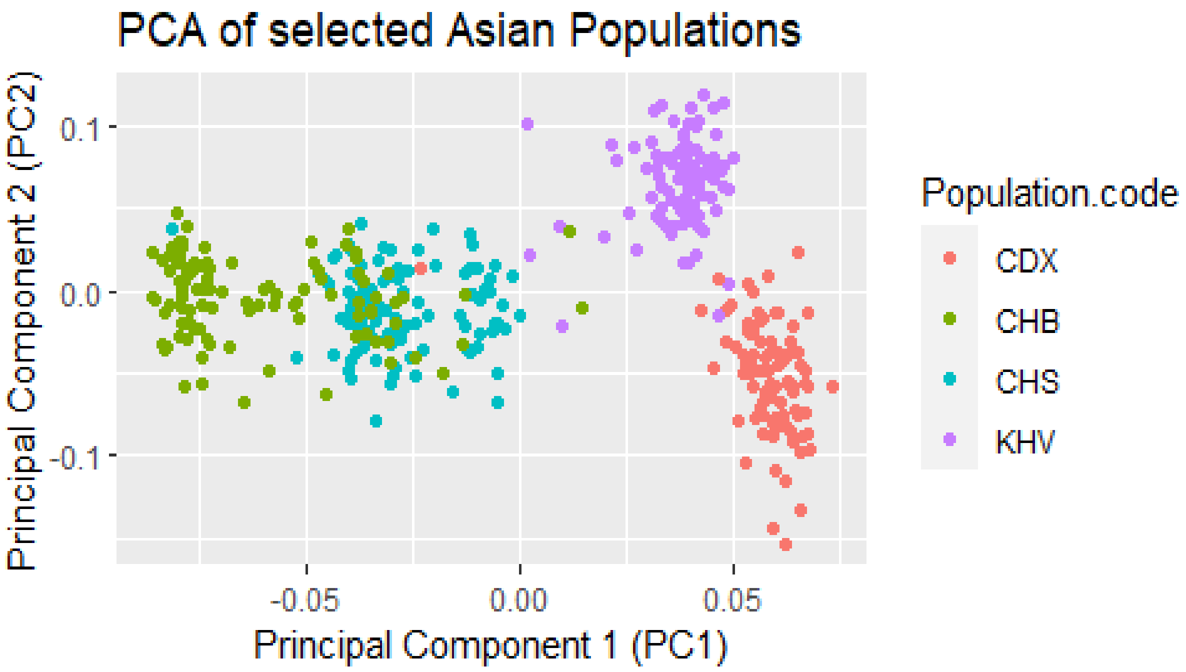


Figure 1: PCA plot

From the PCA plot, we can see 4 different clusters. There is an overlap between the CHB (Han Chinese) and CHS (Southern Han Chinese) population. By observing the distance between the clusters on the (PC1) axis, the CDX (Dai Chinese) population is more varied from the CHS and CHB population than from the KHV (Vietnamese) population. The Han Chinese (CHB & CHS) are separated from the southern population (CDX and KHV) by PC1. On PC2, CHB and CHS do not vary.

Multidimensional Scaling (MDS) Analysis

Multidimensional scaling is used to graphically depict the relationships between samples in a multidimensional space. It shows the degree of similarity or differences between the samples based on their proximity and gives no information about variables.

For our analysis, we first created a set of pruned markers (approx. 8700 SNPs) that were not highly correlated. Next, the identity by descent (IBD) scores were calculated for all pairs of individuals to determine the degree of similarity.. The IBD scores were then used to cluster individuals into homogeneous groups and also generate the first 2 MDS components for each individual (C1 and C2). These MDS components represent the position of each individual in first and second dimensions.

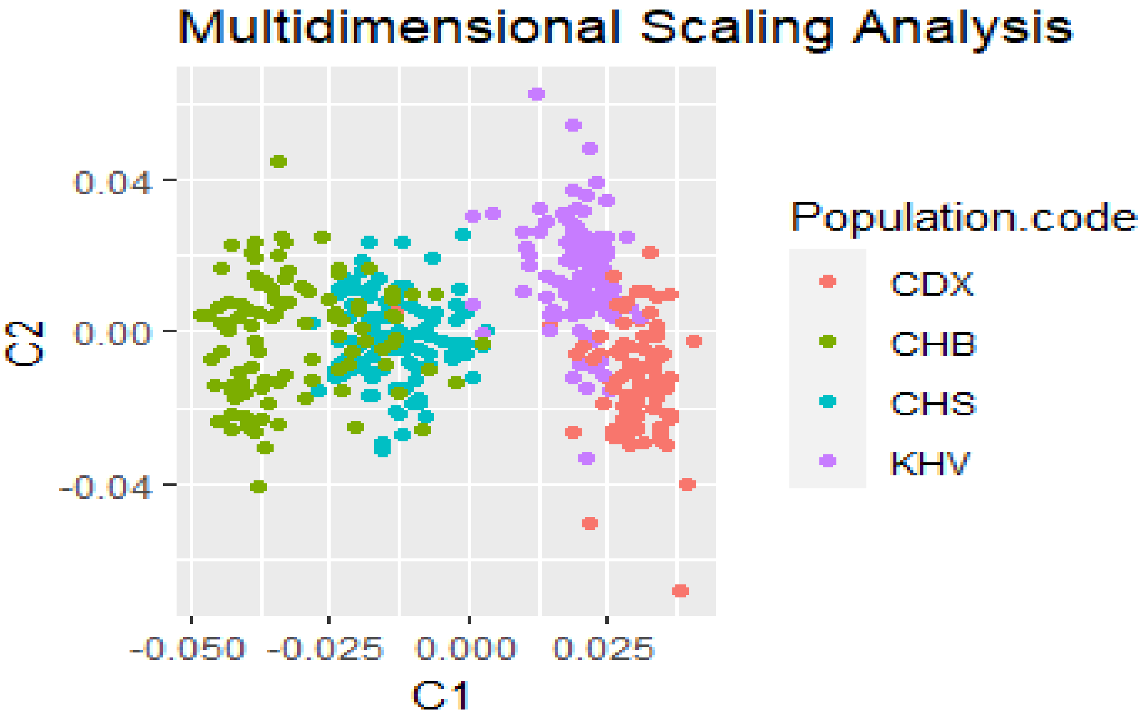


Figure 2: MDS plot

Plotting C1 against C2 and color-coding by population produced the scatter plot shown in Figure 2. There are 4 clusters in which individuals are closely packed together representing each Asian Population. The CHB and CHS also overlap significantly in both dimensions which suggests that both populations are more similar to each other than to CDX or KHV. Likewise, CDX and KHV appear more closely related based on the degree of overlap of both clusters.

Conclusion

Collectively , the results from PCA and MDS suggest that the CHS and CHB populations will show similar physiological responses to HLA associated drugs as both populations appear to be closely related while the CDX and KHV populations will have a distinct response to drugs as they slightly overlap on the MDS plot.

Reference Tutorials

- https://www-users.york.ac.uk/~dj757/popgenomics/workshop6.html
- http://hpc.ilri.cgiar.org/beca/training/data_mgt_2017/BackgroundMaterial/PlinkTutorial.pdf