Principal component analysis (PCA) on SNPs datasets with different single nucleotide polymorphism (SNPs) assignments.

Wayne Chi Wai Ng
Student ID: 12232148
Division of Life Science
The Hong Kong University of Science and Technology
cwngam@connect.ust.hk
Date: April 7, 2021

Abstract

The most common genetic variation in human genome is single nucleotide polymorphism (SNP). SNPs are large datasets and can be difficult to analysis. Principal component analysis (PCA) can reduce dimensionality of SNPs datasets. PCA is a tool that assumes no data distribution assumption, but SNPs datasets may be too skewed for PCA analysis in some cases.

1 Introduction

The simplest genetic variation in human genome is the single nucleotide polymorphism (SNP). SNP accounts for 90% of all human genetic variation. SNP occurs every 100 to 300 bases along the 3-billion-base human genome. SNP is the genomic variation between individuals at a particular nucleotide base position in a particular chromosome region. SNPs are large datasets that require special computer programs to analysis. To interpret SNPs, principal component analysis (PCA) is a technique to reduce dimensionality of SNPs datasets. PCA creates uncorrelated variables that successively maximize variance. In general, PCA is a tool that needs no data distributional assumptions (Jolliffe 2016). In this mini-project, PCA is performed using R. The assumptions of no data distribution on PCA is tested with SNPs datasets. SNPs are assigned as 0(AA), 1(AC), 2(CC) for genotyped data. As a result, SNPs are highly skewed datasets. PCA requires no Gaussian distribution of the datasets, but SNPs datasets may be too skewed for PCA analysis in some cases.

2 SNPs data explorations and processing before PCA

The SNPs dataset for this paper contains a data matrix of columns of SNPs (Single Nucleid Polymorphisms) and rows of peoples around the world. Each element is of three choices, 0 (for 'AA'), 1 (for 'AC'), 2 (for 'CC'). Genotyped data of the 1043 (n) subjects. 0(AA), 1(AC), 2(CC). Missing values are removed. The SNPs dataset are shared in google drive:

- Google drive link: https://drive.google.com/file/d/1a9I8_akfCMHBRrPMdnWkjyL9fKcQbJJq/view?usp=sharing
- Sample Information of 1043 subjects. Google drive link: https://drive.google.com/file/d/11Q-8B57WDQnrIV92b-h_WLqDGviiYsm2/view?usp=sharing

Here are top few lines of ceph_hgdp data. It contains snps number shown as rs, chromos number and position. SNPs are shown as 0,1,2. The SNPs matrix has dimension of 488919 x 1043. There are 1043 individuals and 488918 SNPs per individual.

> ceph_hgdp[1:5,1:5]

snp	chr	pos	HGDP00448	HGDP00479
rs10000929	4	131516474	1	0
rs10002472	4	159087423	2	1
rs10005550	4	128697858	2	2
rs10007576	4	59063992	2	0
rs10007998	4	35988597	0	0

> dim(ceph_hgdp)

[1] 488919 1043

The SNPs have values 0,1,2. A sample of distribution is shown below. Most samples have a mean of less than 1. They are highly skewed data.

> summary(ceph_hgdp_t_update$rs10000929)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.0000 0.0000 0.6117 1.0000 2.0000

> boxplot(ceph_hgdp_t_update$rs10000929)

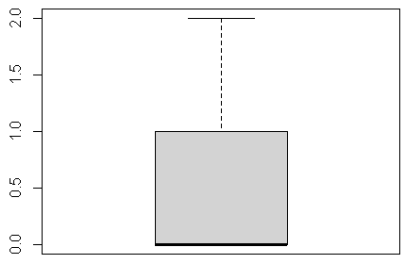

Figure 2.1 Boxplot of rs10000929 values

Before processing PCA, it is found that some columns are constant. The following R codes are used to remove constant/zero column.

```
# checking for constant/zero column

which(apply(ceph_hgdp_t, 2, var)==0)

# remove constant/zero column

ceph_hgdp_t_update<-ceph_hgdp_t[,which(apply(ceph_hgdp_t, 2, var)!=0)]
```

3 Basic PCA analysis

Here is result of PCA analysis. The Y-axis is percentage of variance. The X-axis is principle component. The first principle component is 6.48% of total variance.

```
fviz_eig(ceph_hgdp_t.pc)
```

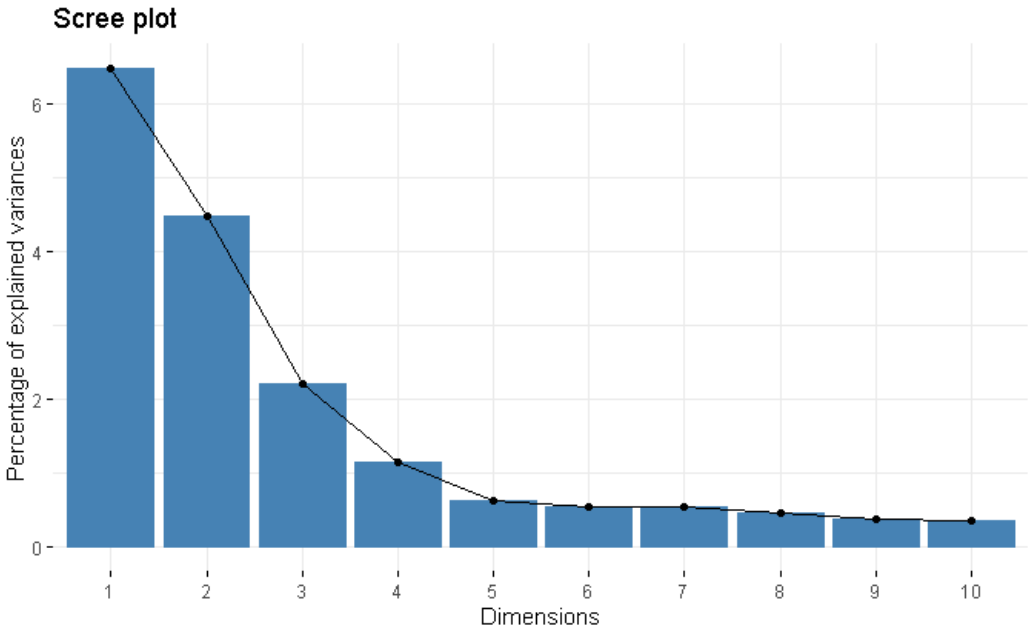

Figure 3.1 The percentage of variances explained by each principal component. The first principle component represents 6.4% variances. The first 10 principal components represent 17.3% variances.

```
> # first PC
> sum((ceph_hgdp_t.pc$sdev[1])^2)/sum(ceph_hgdp_t.pc$sdev^2)
[1] 0.06479549
> # first 3 PCs
> sum((ceph_hgdp_t.pc$sdev[1:3])^2)/sum(ceph_hgdp_t.pc$sdev^2)
[1] 0.1316289
> # first 10 PCs
> sum((ceph_hgdp_t.pc$sdev[1:10])^2)/sum(ceph_hgdp_t.pc$sdev^2)
[1] 0.1726876
```

The following diagram shows PCA plot of first two principle components. The PCA plot shows a clear 7 regions of population groupings.

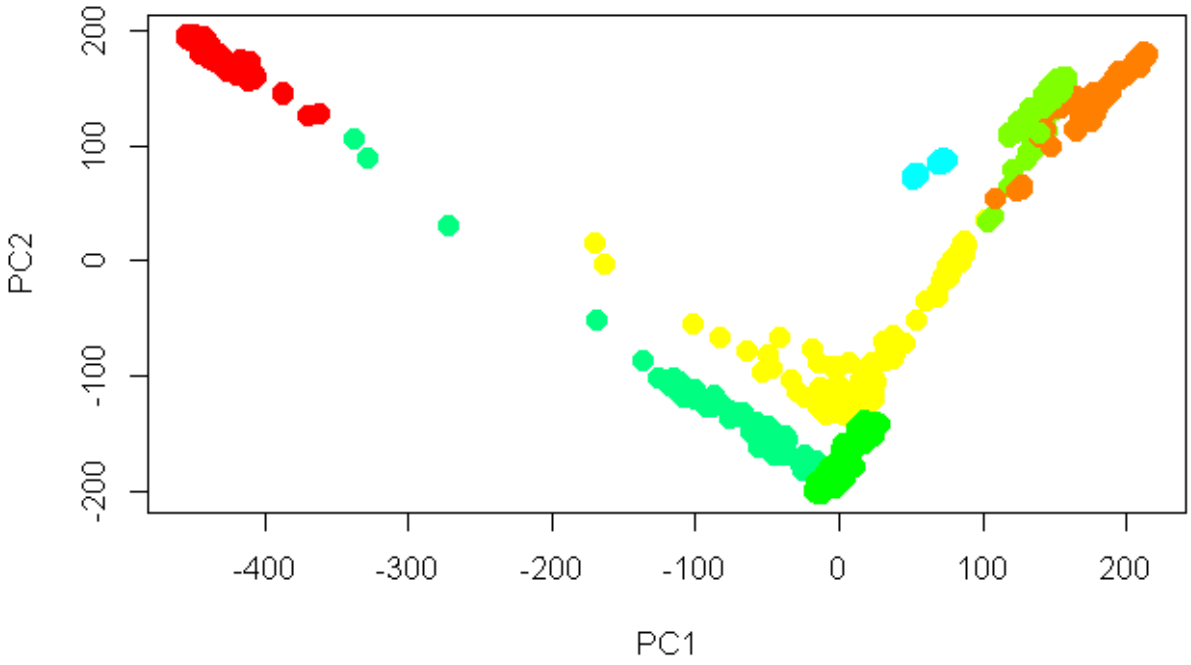

Figure 3.2 PCA plot of first two principle components. The PCA plot shows a clear 7 regions of population groupings.

4 Further PCA analysis

To test whether PCA can handle different skewness of datasets. The SNPs are reassigned as the following. Original SNP values of 2 are assigned to 3. Original SNP values of 0 are assigned to 2. As a result, the SNPs are changed to 1,2,3 with most common values are 2. The mean of SNPs is around 1.7.

```
# orignial snps values are 0,1,2. change snps 2->3, 0->2, at the end, we have 1,2,3, most common is 2

> use data.table

> require(data.table)

> dt.raw <- as.data.table(ceph_hgdp_t_update_snps, keep.rownames=T)

> setindex(dt.raw, rn)

> dt.raw <- replace(dt.raw, dt.raw==2, 3)

> dt.raw <- replace(dt.raw, dt.raw==0, 2)
```

> summary(ceph_hgdp_t_update_snps$rs10000929)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 1.000 2.000 1.725 2.000 3.000

> boxplot(ceph_hgdp_t_update_snps$rs10000929)

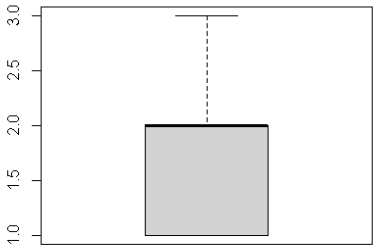

Figure 4.1 Boxplot of rs10000929 values

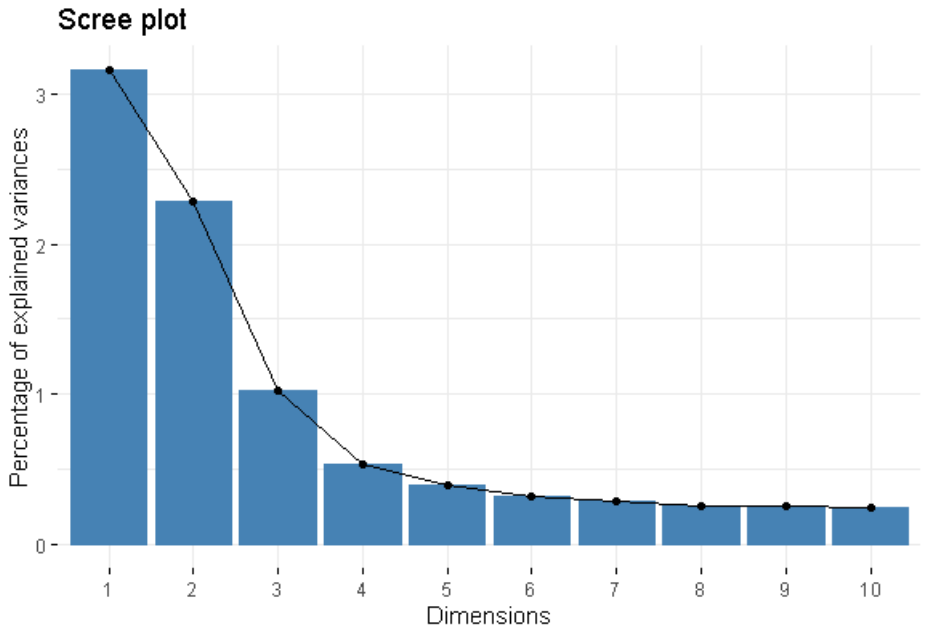

Figure 4.2 The percentage of variances explained by each principal component.

first PC

> sum((ceph_hgdp_t_update_snps.pc$sdev[1])^2)/sum(ceph_hgdp_t_update_snps.pc$sdev^2)

[1] 0.03158999

> # first 3 PCs

> sum((ceph_hgdp_t_update_snps.pc$sdev[1:3])^2)/sum(ceph_hgdp_t_update_snps.pc$sdev^2)

[1] 0.0647145

> # first 10 PCs

> sum((ceph_hgdp_t_update_snps.pc$sdev[1:10])^2)/sum(ceph_hgdp_t_update_snps.pc$sdev^2)

[1] 0.08765841

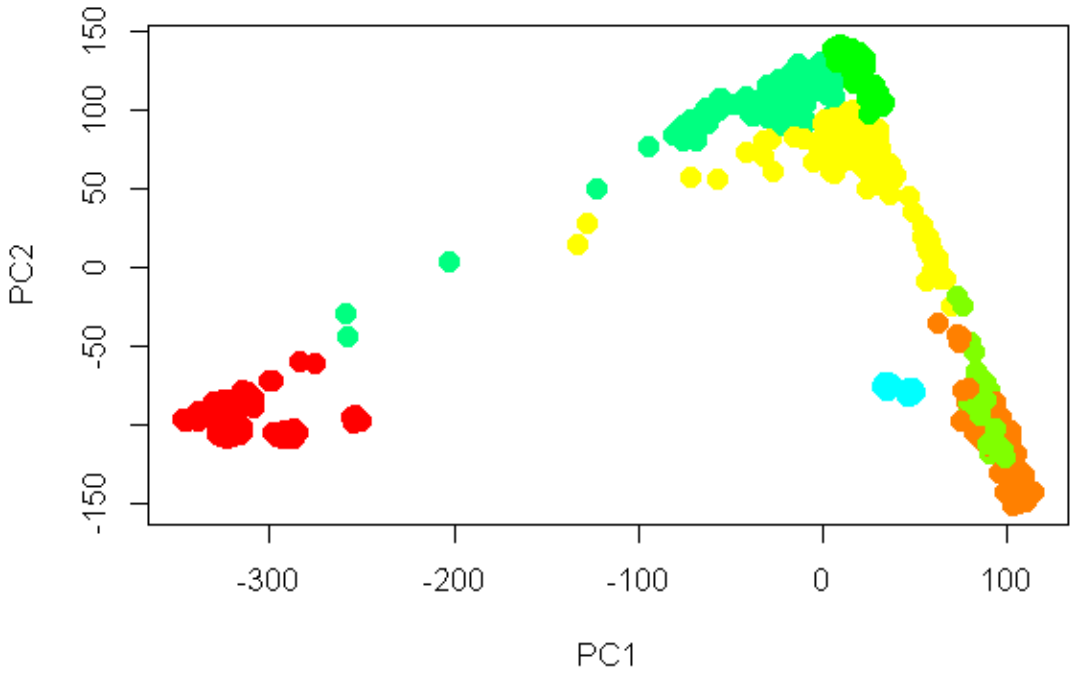

Figure 4.3 PCA plot of first two principle components. The PCA plot shows a clear 7 regions of population groupings.

From Figure 4.1 to 4.3, with reassignment of SNPs of 0,1,2 to 1,2,3 with most common of being 2. We can see that the first principle component variance is reduced from 6.4% to 3.2%. Although, figure 4.3 still shows 7 regions of population groupings, overall variances of PCs are reduced.

To further test whether PCA can handle different skewness of datasets. The SNPs are reassigned as the following. Original SNP values of 0 are assigned to 3. As a result, the SNPs are changed to 1,2,3 with most common values are 3. The mean of SNPs is around 2.1.

```
# orignial snps values are 0,1,2. change snps 0->3, at the end, we have 1,2,3, most common is 3

# use data.table

>require(data.table)

>dt.raw_3 <- as.data.table(ceph_hgdp_t_update_snps_3, keep.rownames=T)

>setindex(dt.raw_3, rn)

# dt.raw_3 <- replace(dt.raw_3, dt.raw_3==0, 3)


>summary(ceph_hgdp_t_update_snps_3$rs10000929)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  1.000   1.000   3.000   2.113   3.000   3.000
```

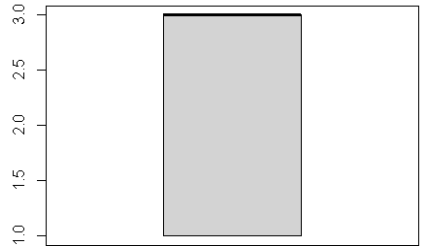

Figure 4.4 Boxplot of rs10000929 values

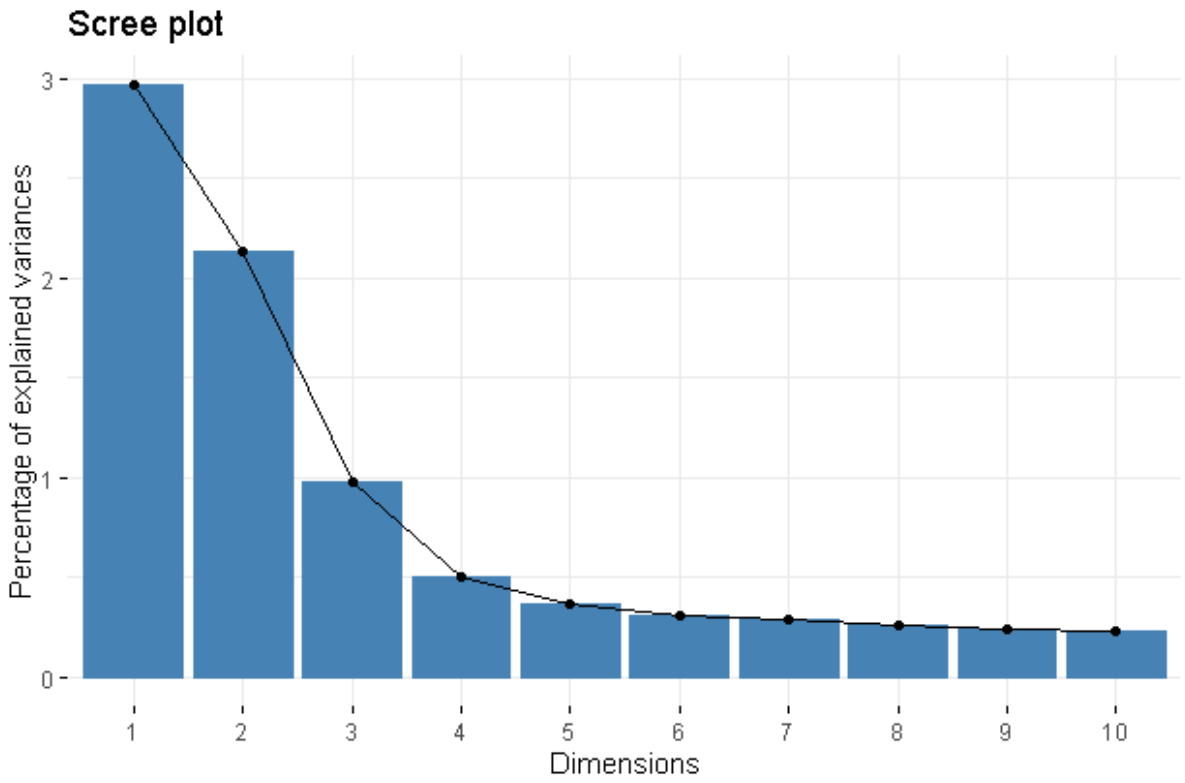

Figure 4.5 The percentage of variances explained by each principal component.

first PC

> sum((ceph_hgdp_t_update_snps_3.pc$sdev[1])^2)/sum(ceph_hgdp_t_update_snps_3.pc$sdev^2)

[1] 0.02967288

> # first 3 PCs

> sum((ceph_hgdp_t_update_snps_3.pc$sdev[1:3])^2)/sum(ceph_hgdp_t_update_snps_3.pc$sdev^2)

[1] 0.06079456

> # first 10 PCs

> sum((ceph_hgdp_t_update_snps_3.pc$sdev[1:10])^2)/sum(ceph_hgdp_t_update_snps_3.pc$sdev^2)

[1] 0.0828932

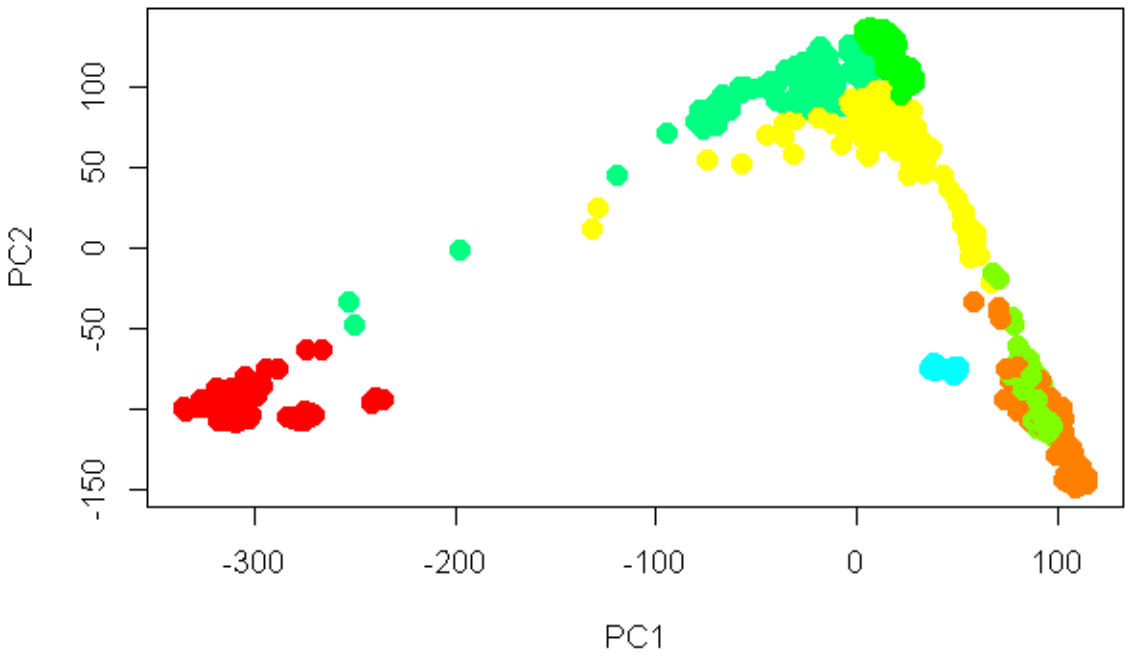

Figure 4.6 PCA plot of first two principle components. The PCA plot shows a clear 7 regions of population groupings.

From Figure 4.4 to 4.6, with reassignment of SNPs of 0,1,2 to 1,2,3 with most common of being 3. Again, we can see that the first principle component variance is reduced from 6.4% to 3.0%. Although, figure 4.6 still shows 7 regions of population groupings, overall variances of PCs are reduced.

5 Conclusions and discussions

Although PCA is a tool that assumes no Gaussian distribution of the datasets, a highly skewed datasets such as SNPs can obtain different PCA results. In this mini-project, PCA is performed using R. The first principle component of PCA is 6.48% of total variance. By changing SNPs of 0,1,2 to 1,2,3 with most common of being 2. We can see that the first principle component variance is reduced from 6.4% to 3.2%. Although, figure 4.3 still shows 7 regions of population groupings, overall variances of PCs are reduced. Another test using SNPs of 0,1,2 to 1,2,3 with most common of being 3. Again, we can see that the first principle component variance is reduced from 6.4% to 3.0%. In conclusion, PCA does not require Gaussian distribution of the datasets, but highly skewed datasets such as SNPs can reduce the performance of PCA. Further studies can be performed to improve PCA technique on handling highly skewed datasets.

Reference

1. Norrgard, K. (2008). *Genetic Variation and Disease: GWAS*. Www.Nature.Com. https://www.nature.com/scitable/topicpage/genetic-variation-and-disease-gwas-682/
2. http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/
3. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. doi:10.1098/rsta.2015.0202