PCA Functions User Manual

Chen Chun-Yu

February 7, 2017

Contents

1 Introduction

This manual shows you how to use the PCA function to perform the analysis and gives an example for demonstration.

2 Prerequest

To make the functions run successfully, please ensure you have installed PLINK in previous.
After you installed PLINK, please copy it from defult folder and paste it to /usr/local/bin.

```
$ cp ~/Bin/plink /usr/local/bin
```

PCA function will detect whether you have installed three required packages: "gdsfmt", "SNPRelate", and "randomcoloR", if you short of any of these it will install for you automatically.

3 PCA Functions

- my_pca(plink.file, data.clust, pop.list, output.name)
  - plink.file
    Please provide the name of your PLINK binary files(BED/BIM/FAM).
  - data.clust
    Please provide a list of indviduals with their population names which correspond to your FAM file. An example file format as below:

Family ID	Individual ID	Populations
1	TDC13	Paiwan
2	TDC117	Amis
3	TDC18	Bunun
4	TDC129	Amis
5	TDC49	Amis
6	TDC497	Puyuma

  - pop.list
    Please provide a list of populations that you are going to perform PCA. An example file format as below:

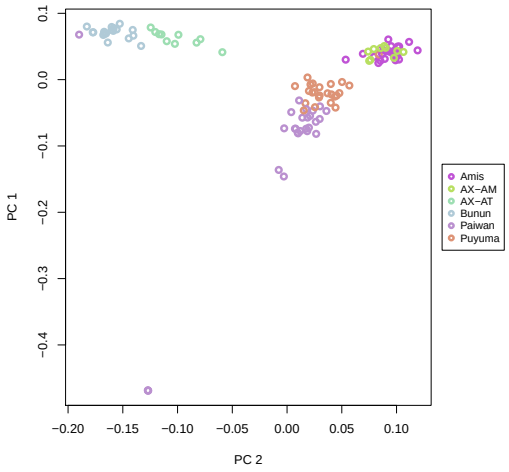
Populations
Paiwan
Amis
Bunun
Puyuma

  - output.name
    You need to name your output files. After the analysis, you will get three plots: PCA outcome("output.name_pca.pdf"), pair plot for the first four PCs("output.name_pairs.pdf"), and line plot that shows different variations capture by different components("output.name_var.pdf"). Besides, you will also get a table which shows the detail information of PCA, including values of PC1 and PC2.("output.name_pca.tab")

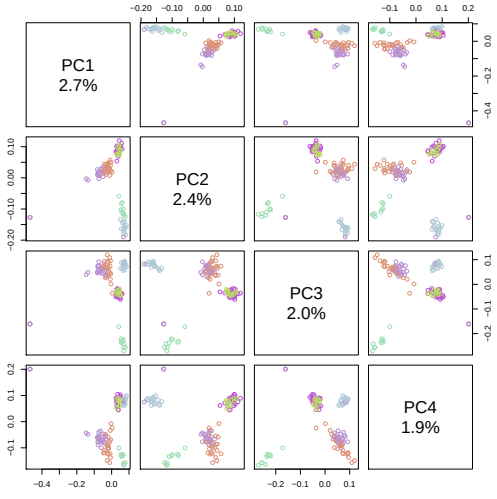
4 Examples

Here we give you some examples of applying PCA functions we have illustrated. We used three different sets of populations for analysis: Taiwanese aborigines, Han populations, and Philipine populations.

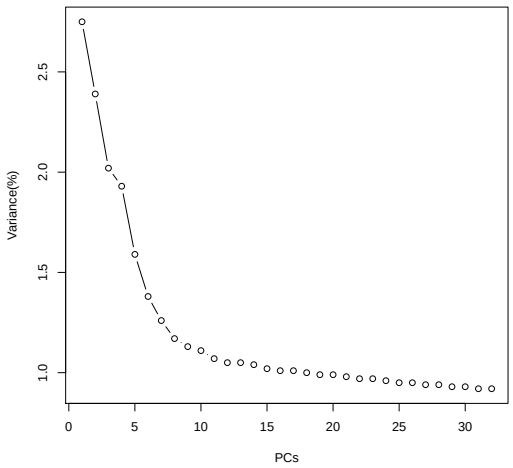
- Taiwanese aborigines



(a) PCA of Taiwanese aborigines



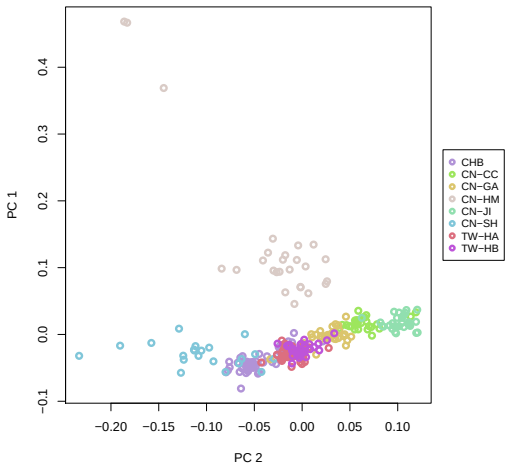
(b) Pair plot for the first four PCs.



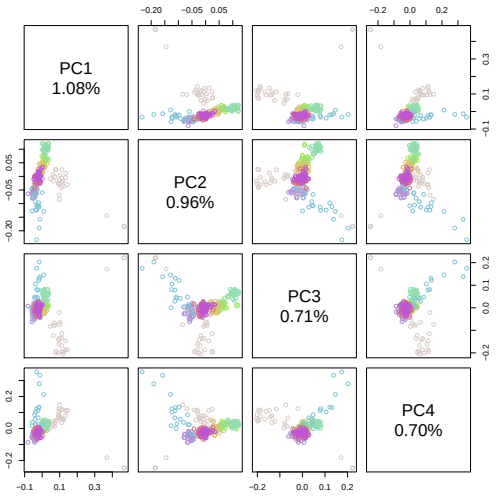
(c) Variation captured by different PCs.

Figure 1: Example 1 - Three kinds of output plots.

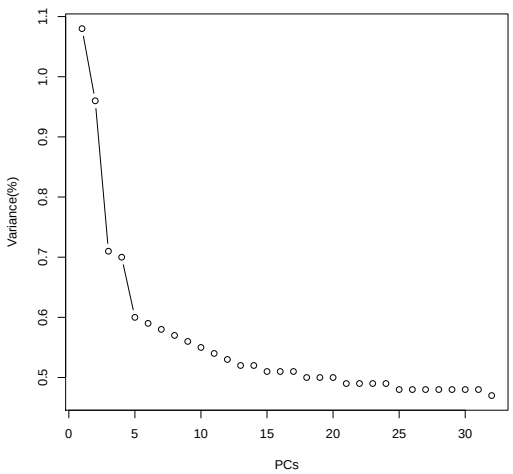
- Han populations



(a) PCA of Han populations



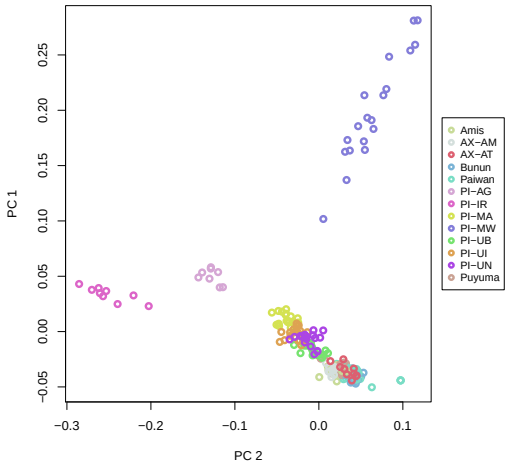
(b) Pair plot for the first four PCs.



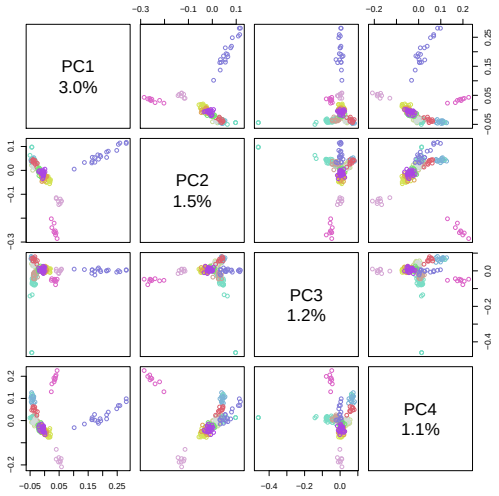
(c) Variation captured by different PCs.

Figure 2: Example 2 - Three kinds of output plots.

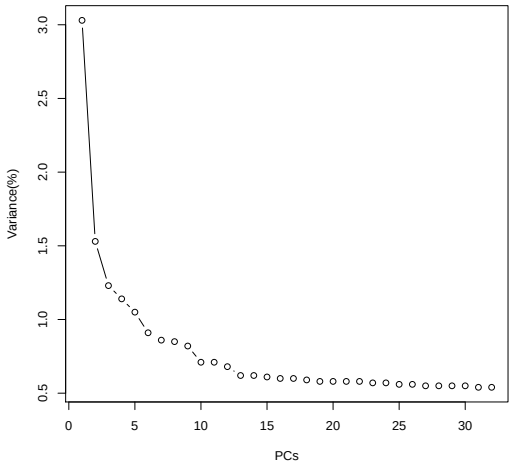
- Philipines populations



(a) PCA of Philipines populations



(b) Pair plot for the first four PCs.



(c) Variation captured by different PCs.

Figure 3: Example 3 - Three kinds of output plots.

References

[1] Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics.* 2007;81(3):559-575.

[2] Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics.* 2012;28(24):3326-3328. doi:10.1093/bioinformatics/bts606.