CSIC5011 Final Project: Exploring Genetic Variations Among Samples by Dimensional Reduction

Kewei Xiong[1], kxiongac@connect.ust.hk
[1]: Department of Life Science, School of Science, HKUST

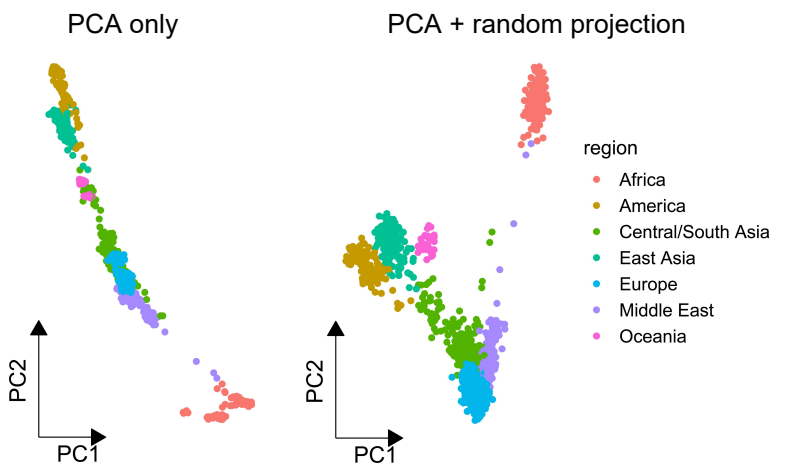
1. Introduction

Breast cancer risk is influenced by multiple variants in susceptibility genes, which are population-specific. To identify heterogeneous hierarchical genetic substructure, I performed random projections, principal component analysis, clustering, statistical analysis and co-localization. I found incorporating random projection into PCA outperformed PCA alone. Additionally, I found the most significantly differnetial SNP rs2250072 was located in SLC24A5 that is a susceptility gene in breast cancer.

2. SNP Dataset

This dataset contains 1034 samples and 488919 single-nucleotide polymorphism (SNPs) with ternary values (0: wild-type; 1: heterozygous mutation; 2: homogeneous mutation). I downloaded the processed data, where missing values have been removed and only autosomal SNPs were retained.

PCA vs. Random projections+PCA



3. Overall Analysis

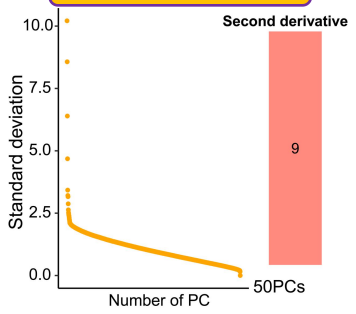
Methodology

- Sparse random projections
- Principal component analysis (PCA)
- Consensus K-means clustering
- Fisher's exact test

I performed PCA only and combined the results of random projections with PCA respectively to estimate whether random projection could improve the PCA outomes. Subsequently, I conducted the consensus K-means clustering to classify genetic clusters and identified differential SNPs across the clusters.

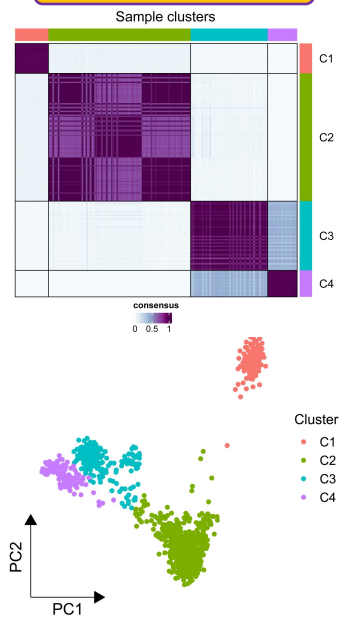
4. Clustering analysis

PCA elbowplot

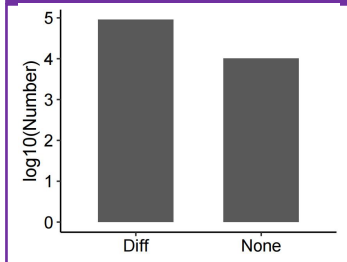


Firstly, I used the 'second derivative' strategy to determine the optimal number of principal components. Based on the PCA embedding, I then applied consensus K-means clustering by subsampling 80% samples per iteraction and with a total of 100 iterations.

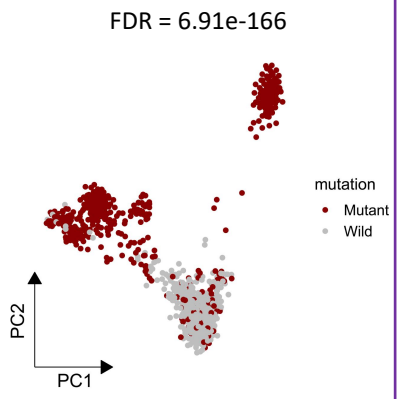
Consensus clustering



5. Differential SNPs and colocalization



Using Fisher's exact test to identify SNPs with significantly different mutation frequency across genetic clusters. P-values were corrected using false discovery rate.



6. Conclusion

By comparing PCA alone and incorporating random projections, I found that in this SNP dataset, the combining strategy outperformed PCA alone, which clearly demonstrated geographic variations.

Consensus clustering analysis accurately identified ethnic substructure based on random projection-derived PCA embedding, and Oceanian and East Asian share similar genetic profile.

rs2250072 is the most differential SNP across the genetic clusters located in cancer subseptbility gene SLC24A5 of breast cancer.

7. References

Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*. 2008.

Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*. 2017.