CSIC 5011 Mini-Project 1: Tracking Human Migration History via PCA on SNPs data

Zhihan Zhu[1], Dong Song[2], Jiabao Li[2] and Zongchao Mo[2] {zzhuay, dsongad, jligm, zmoad}@ust.hk
[1]: Department of Chemical and Biological Engineering, HKUST [2]: Department of Life Science, HKUST

Introduction

Single-nucleotide polymorphisms (SNPs) is a kind of variant that occurs at single nucleotide resolution in the genome when comparing with the reference genome. SNPs data is usually highly dimensional, especially when more and more human genomes were sequenced[1,2]. Therefore it's challenging to have insight into using the whole SNPs landscape directly. Principal component analysis (PCA) can reduce the dimension and keep most important information.

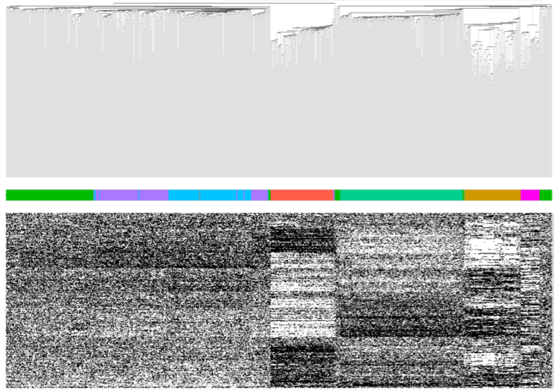
In this project, all of the SNP data from 1043 individuals were subjected to different dimension reduction methods to compare their performance to remove redundant and keep necessary information for different purposes of demonstration. The result showed that hierarchical clustering using raw SNPs data is computationally inefficient while classical PCA can well separate 7 regions and help to maintain the relative genome difference for evolution interpretation. Meanwhile, other methods could help to cluster population but lack the information for evolution analysis.

SNPs Dataset

The cleaned SNPs and clinical dataset we used is from Quanhua MU and Yoonhee Nam. There are 1043 samples from 7 regions, which has 488,919 SNPs marked by 0(AA), 1(AC) and 2(CC).

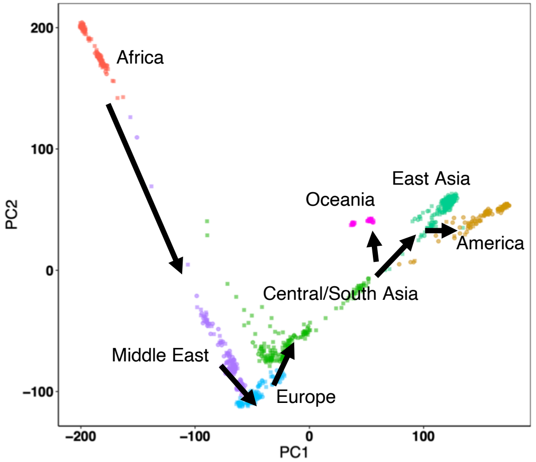
Results

1. Hierarchical clustering of SNPs dataset without dimensional reduction is sufficient to distinguish individuals that from different regions.



Hierarchical clustering was performed using 5000 SNPs with largest variations, using the average method and Euclidean distance. Via the hierarchical clustering, individuals that from similar geographic region were clustered together. However, we also found many cases were assigned to wrong groups, which indicated data noises from the high dimensional dataset affected the precision of clustering.

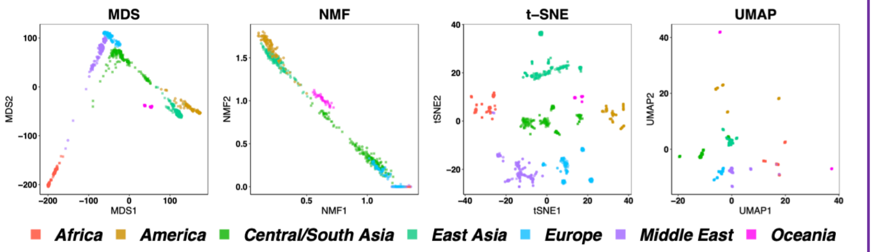
2. PCA of human genetic variance reproduced the migration trace during evolution.



To reduce the data noise and improve the precision of grouping, we used PCA to reduce the data dimension from 488919 to 2.

From the PCA plot, the people from 7 different regions are labelled by different color. Basically, we observed 7 dominant groups on the PCA plot that in accord with the geographic label. More importantly, according to the relative positions of each dots, we also found the continuous change on the PCA plots is highly consistent with the geographic structure of our planet. Assuming human initially originated from Africa, we can clearly find the human migration trace from Africa to Middle East and then other continents on the PCA plot.

3. Comparison among different dimensional reduction approaches.



4 other popular dimensional reduction methods MDS, NMF, t-SNE, and UMAP were also applied on this SNPs dataset. And we found only MDS remained linear relationship of variance among different individuals.

Discussion

In this study, we performed hierarchical clustering and various data dimension reduction methods in information mining of human SNPs dataset. Both hierarchical clustering and data dimension reduction methods can distinguish the different regions well. However, PCA and MDS showed more precise relationship among all individuals than other methods, and provided a reasonable inference of human migration trace during evolution.

References

1. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).
2. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020).

Contribution

Zhihan Zhu: Hierarchical Clustering, NMF, Poster
Dong Song: t-SNE, UMAP, visualization, Poster
Jiabao Li: PCA, Poster
Zongchao Mo: MDS, Poster