MATH5473 Mini-Project: Explore Geographical Variation of Human Gene with SNPs Data

Zhanmiao Huang, Wencan Xia, Yuanhui Luo {zhuangdj, wxiaab, yluocl}@connect.ust.hk

Department of Mathematics, HKUST

Introduction

Single-nucleotide polymorphisms (SNPs) is a substitution of a single nucleotide at a specific position in the genome, which is the most common type of human genetic variation and provides powerful tools for genetic studies.[4] However, SNPs data is usually high-dimensional, which brings great challenges to analysis and calls for efficient dimension reduction methods.

In this project, we first apply different dimension reduction methods on SNPs data to explore the relationship between the genetic variation and geographical variation. Then we investigate the SNPs with top importance and evaluate the number of important SNPs needed for good prediction on regions with statistical learning methods. In case study, we further focus on the genetic variability among populations in China and its neighboring areas and the conclusions are consistent with their geographical relationships.

Dataset

In this project, we use the cleaned dataset from Quanhua MU and Yoonhee Nam. The SNPs dataset contains 650,000 SNPs from 1064 people around the world with three types marked by 0(AA), 1 (AC) and 2 (CC), along with the geographical information of each person.

Methodology

To deal with the high dimensionality and extract crucial information from SNPs data, we consider following dimension reduction methods:

- Principal Component Analysis (PCA) is a technique that increases interpretability and preserves important information as possible by creating uncorrelated principal components that successively maximize variance[3].
- Multidimensional Scaling (MDS) is a method that provides a representation of similarity among objects and condenses large-scale data into a low-dimensional space while keeping the distance constant[2].
- Random Forest (RF) and Extra Trees (ET) are ensemble learning methods that construct multiple decision trees in two different ways and can identify key features for classification[1].

In addition, since the number of features is very large, random projection is adopted to reduce the dimension for comparison.

Result Analysis

1. PCA/MDS

We explore the genetic variation with geographic variations by PCA/MDS. As shown in Fig. 1-2, the high-dimensional SNPs data is reduced into two and three principal components (PC) respectively. We can see that both methods effectively distinguish the people from different regions and PCA gives a more intuitive segmentation than MDS. Due to the clear visualization, we use two principal components in later analysis without loss of generality.

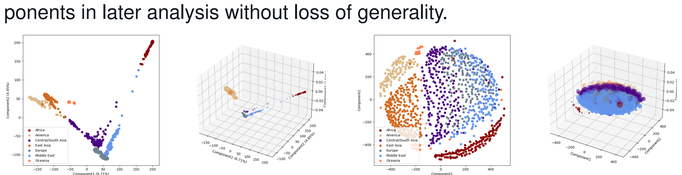
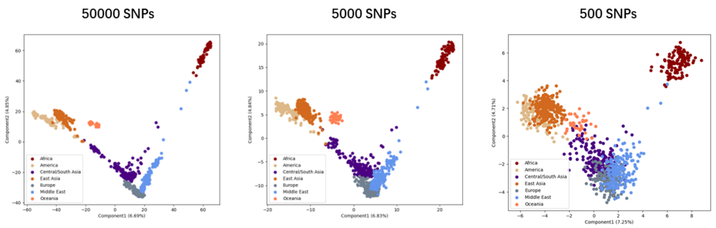


Fig. 1: PCA

Fig. 2: MDS

2. Random Projections

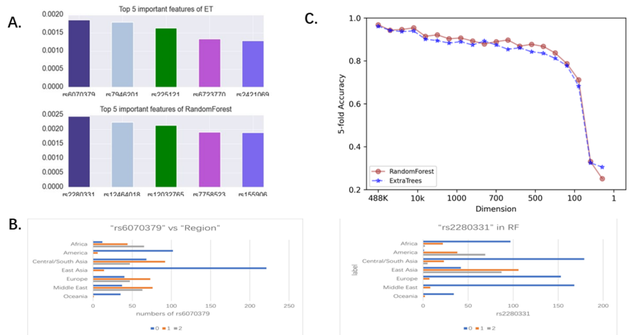
We randomly select 50000, 5000 and 500 SNPs as random projections and the PCA results are shown in the following figure. With a relatively large number of projections, PCA performs similarly to that with full dataset (Fig.1). As the number of projections decreases to 500, the performance of PCA degenerates significantly.



3. Statistical Learning Methods

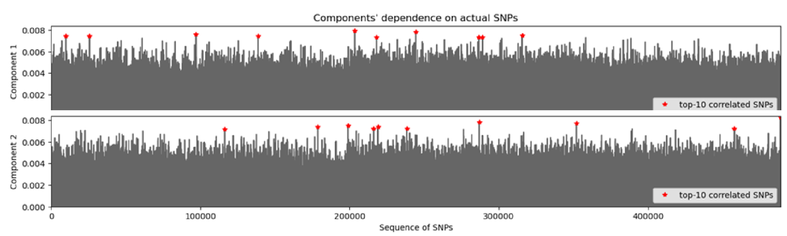
We use random forests and extra trees to find the 5-top important SNPs. As shown in fig. A, "rs2280331" is the most important SNP result from the random forest. However, according to fig. B, no strong signal is shown to tell the differences between genetic information from different regions. Meanwhile, extra trees give five totally different top SNPs, and fig. A shows that the most important one is "rs6970379", which tells the difference between genetic information from East Asia and other regions.

We also predict the region based on different numbers of SNPs features to investigate the minimal number that can ensure enough accuracy, and compare the efficiency of random forest and extra trees. As shown in fig. C, at least 500 important features are needed to ensure relatively good accuracy, extra trees predict regions better among the two tree methods. In sum, extra trees are more recommended for prediction.



Case Study

Here we focus on the genetic variability among the populations in China (Uygur, Lahu, Dai etc.), and compare with the neighboring East Asia areas (Japan, Russia, Cambodia).

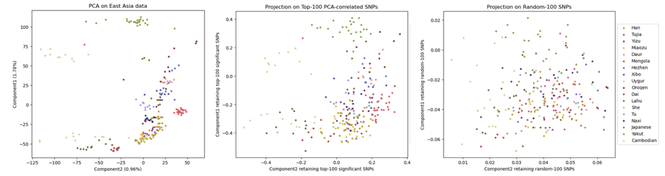


1. PCA of China and other East Asia regions suggest that the geographically close regions may share relevant genetic variations or similarities.

It is interesting to note that the point cloud data mirrors the actual locations of the regions on the map. For instance, Cambodia is genetically close to Lahu and Dai (southwest China). Japan shows more similarities with Han than the minority groups far away. Meanwhile, Uygur and Yakut (Russia) are isolated in East Asia PCA, indicating their probable origins from Central Asia and Europe, respectively.

2. Interpret PCs from mathematical abstractions to be a subset of PCA-correlated actual SNPs.

We selected the top-100 SNP components accounting for the largest proportion in PC1 and PC2, respectively (shown as bottom fig. in the middle column). The 2D projection of East Asia information onto the new PCs merely retaining top-100 SNPs basically keep the population structure exhibited above, while projection on randomly selected 100 SNPs does not.



Conclusion

In this project, we conduct PCA, MDS and random projections on SNPs dataset to explore the genetic variation with geographic variations. The result indicates that both PCA and MDS can separate people from different regions based on essential principal components of SNPs, which is more efficient than the random selection of SNPs. With random forest and extra trees, we find that an adequate number of SNPs with top importance can tell the difference between genetic information from different regions and predict the region where people come from. Further, we study the populations of China and its neighboring areas and reveal that the similarity of SNPs principal components can reflect the relationship between their geographical locations. In summary, there is a close relationship between human genes and geographic variation, which can be effectively detected by dimensionality reduction methods with SNPs data.

References

[1] Anne-Laure Boulesteix et al. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.6 (2012), pp. 493–507.

[2] Natalia Jaworska and Angelina Chupetlovska-Anastasova. "A review of multidimensional scaling (MDS) and its utility in various psychological domains". In: *Tutorials in quantitative methods for psychology* 5.1 (2009), pp. 1–10.

[3] Ian T Jolliffe and Jorge Cadima. "Principal component analysis: a review and recent developments". In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202.

[4] David G Wang et al. "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome". In: *Science* 280.5366 (1998), pp. 1077–1082.

Contribution

- PCA/MDS, Random Projections: Yuanhui Luo
- Statistical Learning Methods: Wencan Xia
- Case Study: Zhanmiao Huang
- Poster: All members