README: SNP-Based Population Classification Using Dimensionality Reduction

Project Overview

This project investigates the potential to classify individuals into populations based on their genetic information, specifically Single Nucleotide Polymorphisms (SNPs). The analysis involved using dimensionality reduction techniques such as PCA (Principal Component Analysis) and SVD (Singular Value Decomposition) to visualize and quantify genetic differentiation across French, Spanish, and Turkish populations.

Data Description

- geno: A genotype matrix (individuals x SNPs), with values 0, 1, or 2 representing the number of reference alleles.
- ind: Population labels for each individual (French, Spanish, Turkish, or Others).
- snp: Names of the SNPs used as features.

Methodology

1. The data was filtered to include only French, Spanish, and Turkish individuals.
2. The genotype matrix was normalized using z-score standardization.
3. Truncated SVD was applied to reduce the data to two dimensions.
4. The resulting projection was visualized to observe population structure.
5. PCA was used to analyze the variance structure and to determine how many components are needed to retain 50% of the variance.
6. Data reconstruction and Frobenius norm were used to assess information loss.

Key Findings

- The first two principal components explained only 1.40% of the variance but still revealed clear population clusters.
- A total of 301 components were required to explain 50% of the variance, reflecting the high dimensionality and dispersed nature of genetic data.
- The reconstruction error after using 301 components was approximately 2650, showing partial information loss but preservation of major patterns.

Conclusion

The study shows that despite the low variance captured by the first two principal components, they are effective in visualizing population-level genetic differences. Dimensionality reduction techniques such as PCA and SVD are powerful tools for uncovering structure in high-dimensional SNP data. This project supports the hypothesis that genetic variation, as measured through SNPs, can be used to classify individuals by population, and it highlights the importance of careful preprocessing and interpretation when working with genomic datasets.