Ancestry with MDS vs PCA

Deepika Dokuru

2024-06-03

Ancestry with MDS vs PCA

Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) are both dimensionality reduction techniques commonly used in population genetics to visualize genetic variation among individuals. PCA transforms the genetic data into a new coordinate system where the greatest variance by any projection of the data comes to lie on the first coordinate (principal component), the second greatest variance on the second coordinate, and so on. This method is particularly effective at capturing the major axes of genetic variation, making it useful for identifying population structure and detecting patterns such as population stratification or admixture.

On the other hand, MDS is a non-linear technique that aims to preserve the pairwise distances between samples as much as possible in the reduced dimensional space. It is based on a distance matrix, often derived from genetic distances, and is more flexible than PCA because it does not assume a linear relationship in the data. MDS is particularly useful when the genetic distances between individuals do not align well with the assumptions of PCA, such as in cases of complex population structures or non-linear genetic relationships. While both methods can reveal population structure, PCA tends to be more interpretable due to its linear nature, whereas MDS can better represent non-linear patterns but may be harder to interpret directly.
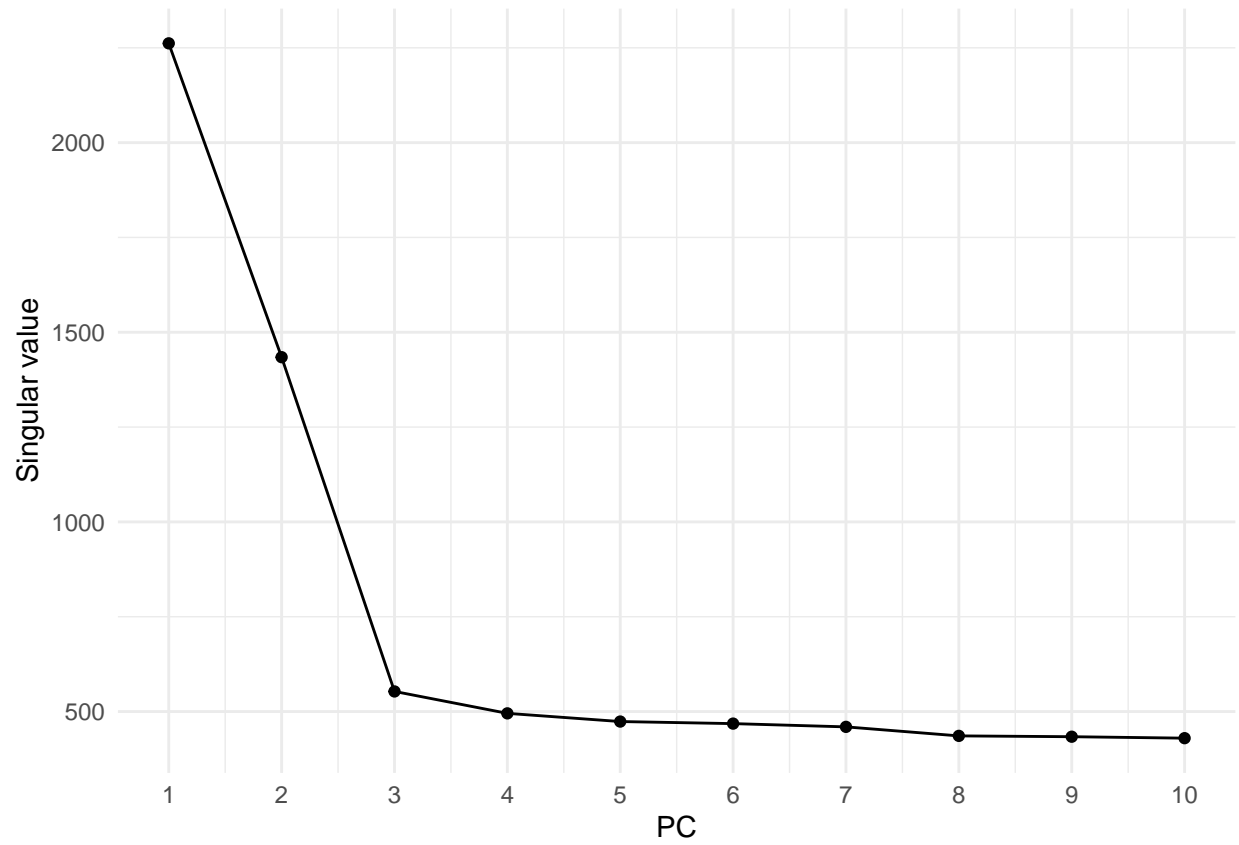
Load Necessary Files Load in the fam and population info files for a subset of 1000 genomes.

PC Analysis

Run PCA

Plot PCA Plotting PC 1 vs PC 2 and a Scree plot for the 10 PCs

PC1 vs PC2

Principal Component 1

Principal Component 2

Population

- ASW
- CEU
- MXL
- NAM
- YRI
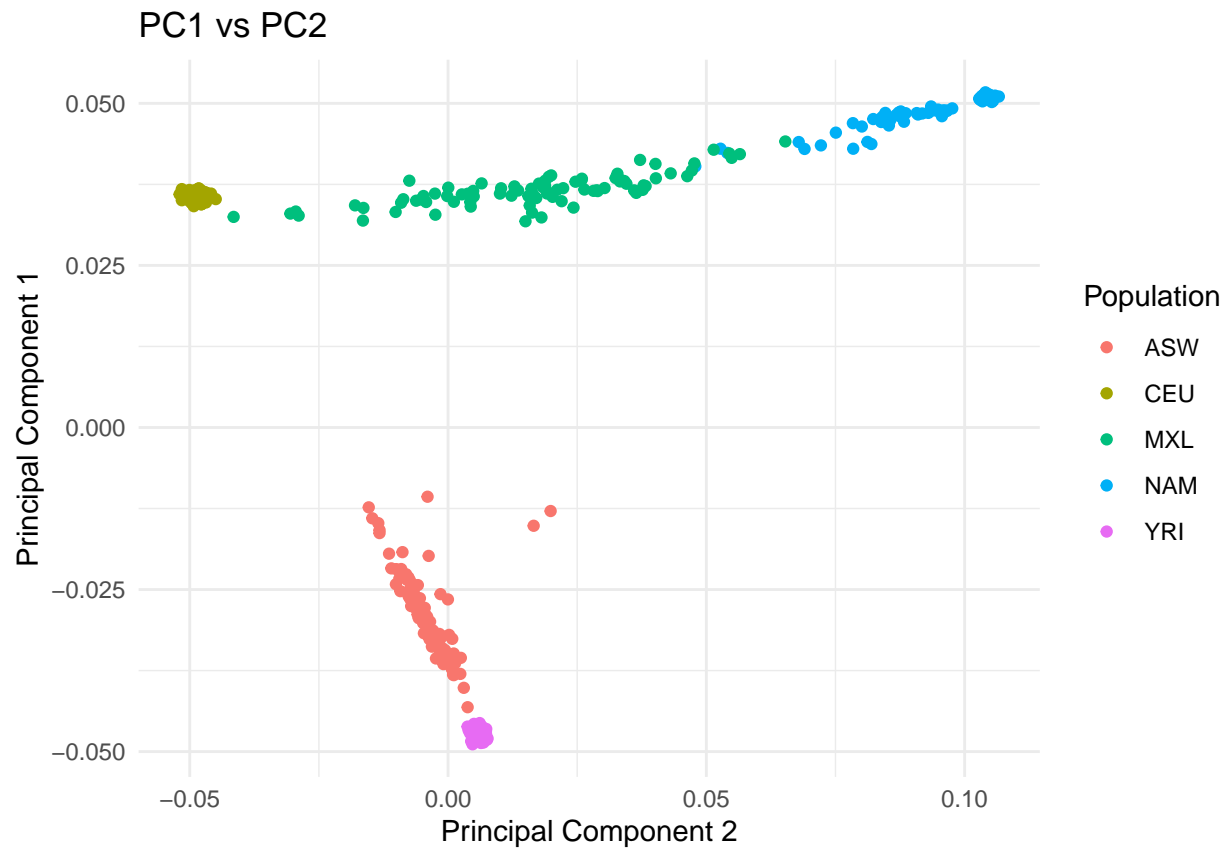
MDS Analysis

Run MDS

Plot MDS   Plotting MDS Dimension 1 vs MDS Dimension 2

MDS1 vs MDS2