Human Migration History: A Single Nucleotide Polymorphisms Perspective

Lue Shen
Department of Mathematics
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
lucas.shen@connect.ust.hk

Abstract

Human migration is a fascinating and complex topic that has been extensively studied in various fields, such as archaeology, genetics, and linguistics. One of the dominant models of the geographic origin and early migration of anatomically modern humans is the African origin hypothesis. In this paper, we aim to investigate the human migration history from a statistical perspective using genetic data. We employ Principle Component Analysis (PCA) to analyze the high-dimensional single nucleotide polymorphisms (SNPs) data and derive the importance of each SNP from the PCA results. We also explore the optimal model settings and also apply Multidimensional Scaling (MDS) and t-Distributed Stochastic Neighbor Embedding (t-SNE) as comparisons to the PCA method. Our results show that all the clustering results coincide with the African origin hypothesis. The study provides valuable insights into human migration patterns and contributes to the ongoing discussion on the origin and evolution of anatomically modern humans.

1 Introduction

The history of human migration has been a topic of interest for many years, and has been studied from various fields, including archaeology [1], genetics [2], and linguistics [3]. One of the most widely accepted models for the origin and early migration of anatomically modern humans is the African origin hypothesis [4]. According to this model, modern humans evolved in Africa and then migrated to other parts of the world, replacing other hominin species that existed at the time. This theory has been supported by various lines of evidence, including genetic data.

In past 40 years, advances in technology and analytical methods have enabled the analysis of large-scale genetic data [5], allowing researchers to investigate the genetic patterns of human migration in more detail than ever before. One of the most powerful approaches in this area is the analysis of single nucleotide polymorphisms (SNPs), which are variations in a single nucleotide that occur at a particular position in the genome [6]. The analysis of SNP data can be facilitated by statistical methods such as Principle Component Analysis (PCA), Multidimensional Scaling (MDS), and t-Distributed Stochastic Neighbor Embedding (t-SNE). These methods allow for the exploration of high-dimensional datasets and the identification of patterns that might not be immediately apparent to human eyes.

In this paper, we aim to study the human migration history from a statistical perspective on genetic data. Specifically, we have implemented PCA to analyze high-dimensional SNP data. By deriving the importance of each SNP from the PCA results and exploring optimal model settings, we hope to gain insights into the genetic patterns of human migration. Furthermore, we will compare the results obtained from PCA with those from MDS and t-SNE. All clustering results will be evaluated to determine if they coincide with the African origin hypothesis. Overall, this paper presents a

Preprint. Under review.

comprehensive approach to studying human migration history and sheds light on the origins and early movements of our species.

2 Theories

2.1 Principle Component Analysis

Principle Component Analysis (PCA) is a statistical method used to reduce data dimensions, especially suitable for high-dimensional data like SNPs. It will transform the original data into a new set of variables, say principle components, which aggregate feature information and capture the most significant patterns of variation in the data. These principle components are orthogonal to each other, because they are uncorrelated essentially. To compute the principle components, we first calculate the sample covariance matrix of the original data, given by:

$$\hat{\mathbf{C}} = \frac{1}{n-1}(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}}) \tag{1}$$

where n is the number of observations, $\mathbf{X}$ is the data matrix, and $\bar{\mathbf{X}}$ is the mean vector of the data.

Then we can calculate eigenvectors and eigenvalues of this covariance matrix by the following:

$$\hat{\mathbf{C}}\mathbf{v} = \lambda\mathbf{v} \tag{2}$$

where $\mathbf{v}$ is the eigenvector, representing the directions in which the data varies the most, and λ is the corresponding eigenvalue, measuring the amount of variance explained by each eigenvector.

The principle components are then computed by projecting the data onto the eigenvectors by the descending order of eigenvalues.

$$PC_i = \mathbf{v}_i^T\mathbf{x} \tag{3}$$

where $\mathbf{v}_i$ is the eigenvector corresponding to the i-th largest eigenvalue, and $\mathbf{x}$ is an observation in the data matrix. This produces a set of new variables that are uncorrelated and capture the most important patterns of variation in the data.

With the principle components, we can project the sample data onto a new coordinate system by the following:

$$\mathbf{y} = \mathbf{X}\mathbf{V}_k \tag{4}$$

where $\mathbf{X}$ is the original data matrix, and $\mathbf{V}_k$ is a matrix of the top k eigenvectors, say the number of principle components selected for dimensionality reduction. In this paper, k is set to 2 for simplicity. Matrix $\mathbf{y}$ contains the projected data points in the new coordinate system, say a 2-dimensional plane.

In our case, the final results can be visualized in a scatter plot, where each data point represents an observation in the original high-dimensional space. By plotting the final results, we can gain insights of the population structure and identify genetic patterns or clusters.

2.2 Multidimentional Scaling

Multidimensional Scaling (MDS) is used to demonstrate the similarity between objects based on a distance or dissimilarity matrix. It seeks to represent the objects in a lower-dimensional space while preserving the pairwise distances between them. MDS begins with a distance or dissimilarity matrix $\mathbf{D}$, where d_{ij} is the distance between objects i and j. The goal is to find a set of k points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ in a lower-dimensional space, such that the pairwise Euclidean distances between the points approximate the original distances in $\mathbf{D}$. Mathematically, we want to minimize the following stress function:

$$S(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k) = \sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij} - |\mathbf{x}_i - \mathbf{x}_j|)^2 \tag{5}$$

where n is the number of objects and $|\cdot|$ denotes the Euclidean norm.

MDS can be performed using either a classical or non-classical approach. In classical MDS, the k-dimensional points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are found by performing an eigendecomposition of a centered double-centered matrix $\mathbf{B}$, which is computed from the original distance matrix $\mathbf{D}$:

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}\mathbf{H}^T \tag{6}$$

where $\mathbf{H}$ is the centering matrix $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T = \mathbf{H}^T$.

The k-dimensional points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ are then given by the top k eigenvectors of $\mathbf{B}$, multiplied by the square root of their corresponding eigenvalues.

Non-classical MDS, on the other hand, does not require the computation of the centered double-centered matrix $\mathbf{B}$. Instead, it directly computes the k-dimensional points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ using an iterative algorithm, such as gradient descent.

2.3 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE) can be applied to visualize high-dimensional data in a low-dimensional space while preserving the pairwise similarities between the data points. Unlike PCA and MDS, t-SNE is a non-linear dimensionality reduction technique that can capture complex relationships in the data. It starts with a high-dimensional similarity matrix $\mathbf{P}$, where p_{ij} represents the similarity between data points i and j in the high-dimensional space. t-SNE then constructs a low-dimensional similarity matrix $\mathbf{Q}$, where q_{ij} represents the similarity between data points i and j in the low-dimensional space.

The low-dimensional similarity matrix $\mathbf{Q}$ is constructed by first computing the conditional probability $p_{j|i}$, which measures the similarity between data points i and j in the high-dimensional space relative to other points around i:

$$p_{j|i} = \frac{\exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|\mathbf{x}_i - \mathbf{x}_k\right\|^2 / 2\sigma_i^2\right)} \tag{7}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the high-dimensional representations of data points i and j, respectively, and σ_i is the variance of the Gaussian kernel used to compute $p_{j|i}$.

Next, t-SNE constructs a similar conditional probability $q_{j|i}$ in the low-dimensional space, using a t-distribution instead of a Gaussian distribution:

$$q_{j|i} = \frac{\left(1 + \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2\right)^{-1}}{\sum_{k \neq i} \left(1 + \left\|\mathbf{y}_i - \mathbf{y}_k\right\|^2\right)^{-1}} \tag{8}$$

where $\mathbf{y}_i$ and $\mathbf{y}_j$ are the low-dimensional representations of data points i and j, respectively.

The low-dimensional representations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are then optimized to minimize the divergence between $\mathbf{P}$ and $\mathbf{Q}$ using gradient descent. The optimization objective is given by the Kullback-Leibler divergence:

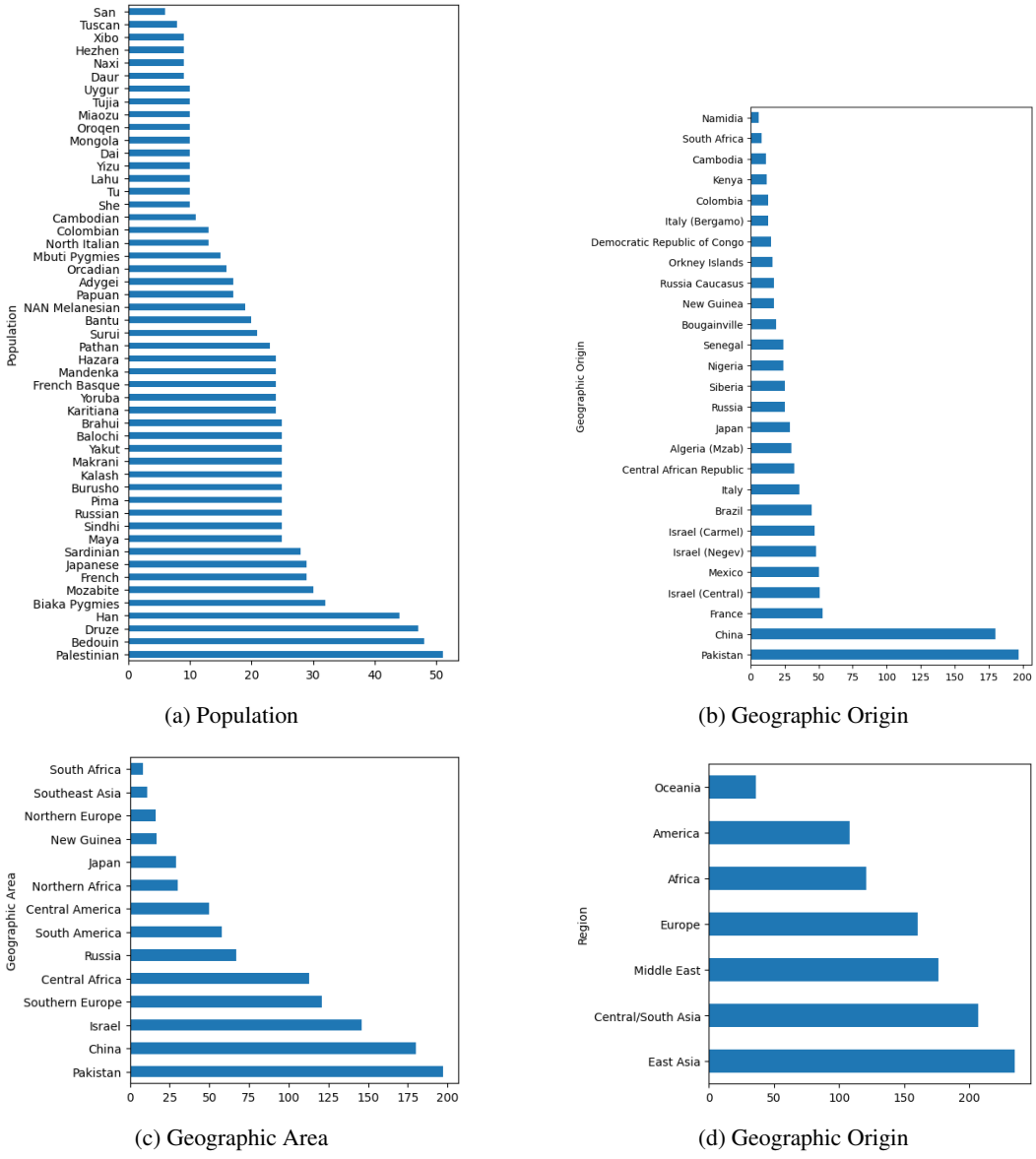
$$\mathrm{KL}(\mathbf{P} \| \mathbf{Q}) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{9}$$

3 Data Description

The dataset used in this study is based on the single nucleotide polymorphisms (SNPs) data from [5]. The dataset consists of a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, where $n = 1,064$ denotes the number of individuals from 51 populations worldwide, and $p = 650,000$ represents the number of SNPs. The values of the SNPs are represented as 0 for 'AA', 1 for 'AC', 2 for 'CC', and 9 for missing values. After removing all the missing values, the dataset was reduced to $n = 1,043$ and $p = 488,919$.

In addition to the raw genetic data, we also extracted population information for each individual in the dataset. The data was organized into four different levels of granularity: region, geographic area, geographic origin, and population labels. These levels allow for exploration of human migrations at various scales, from continent-level to country-level and even within individual populations. Figure 1 illustrates the distribution of the dataset at different granularities.

However, it is worth noting that the dataset is unevenly distributed among populations and countries, with China and Pakistan being the most represented. This may potentially introduce biases in the analysis and should be taken into consideration in our studies.



(a) Population

(b) Geographic Origin

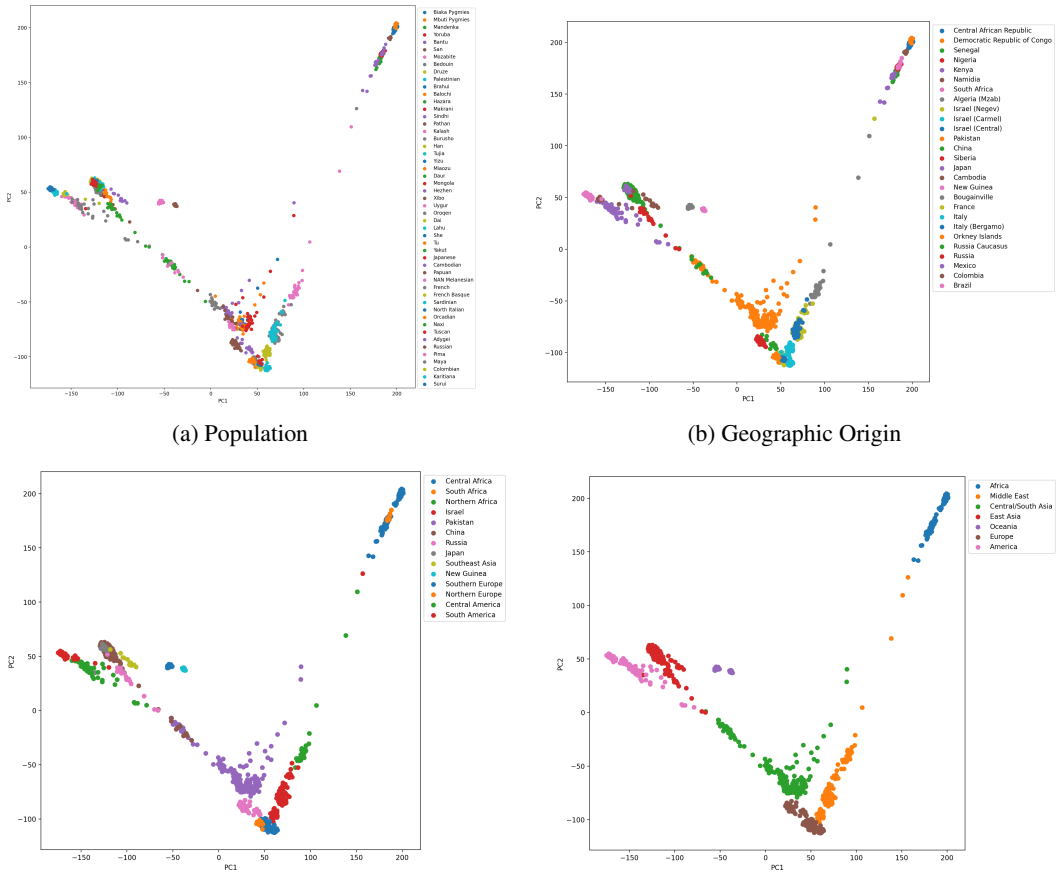
(c) Geographic Area

(d) Geographic Origin

Figure 1: Sample Data Distributions

4 Process and Results

To begin with, we performed data preprocessing on the SNPs dataset and extracted the sample distribution information through histogram plots. Subsequently, we employed PCA to reduce the data dimension to two, and plotted all data in the new coordinate system, as depicted in Figure 2. The four figures illustrate clear separations among different clusters. Figure 2d shows distinct separations between people from different continents, demonstrating that African people are genetically closest to Middle Eastern people. Moreover, Middle Eastern people are genetically closer to both European people and Central/South Asian people, which is consistent with the African origin hypothesis. The chart on the left of the figure depicts Central/South Asian people, who are connected to East Asian people, American people, and Oceanian people in terms of the SNPs data. However, it is difficult to determine which one of these three groups Central/South Asia is genetically closer to. Furthermore, Figure 2b and Figure 2c can also display discernible separations between clusters while Figure 2a cannot.



(a) Population

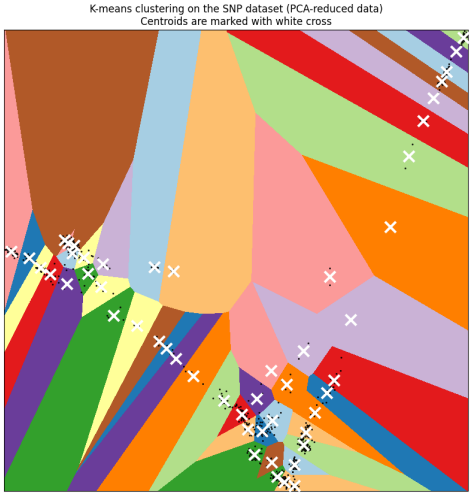
(b) Geographic Origin

(c) Geographic Area

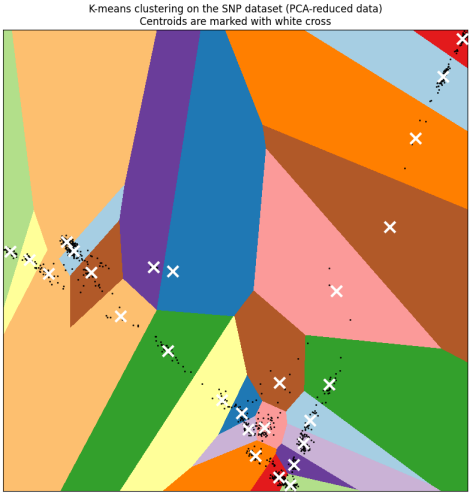
(d) Region

Figure 2: PCA Results

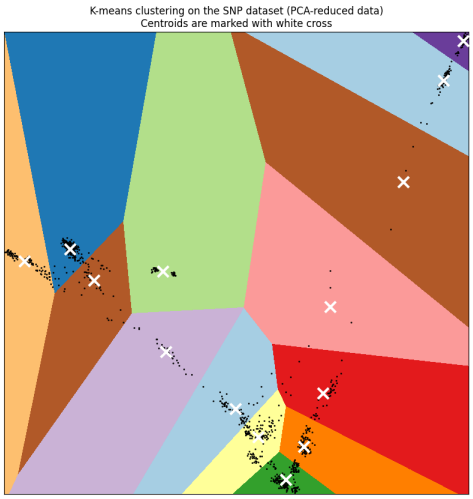
Next, we applied KMeans to find the centroid of each cluster and assigned different colors to visualize the data more effectively, as shown in Figure 3. Figure 3d indicates a potential migration path if we connect the central point of the African cluster to its closest point, so on and so forth. However, the accuracy of dividing the plane in this way is questionable as the Oceania cluster is located between the East Asia and Central/South Asia clusters. This discrepancy could be attributed to the limited size of the Oceania sample data, which only comprises 36 instances. To address this issue, we could consider more detailed granularity levels, such as country or population levels, as demonstrated in Figures 3c, 3b, and 3a. These figures reveal that the path from the Central/South Asia cluster to the East Asia cluster can be further broken down into more detailed subclusters, providing a more accurate representation of the human migration path.



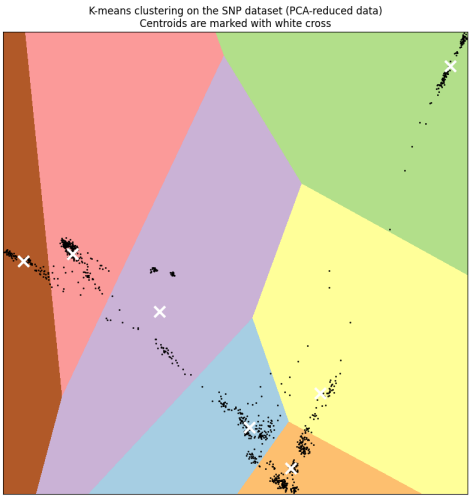
(a) Population



(b) Geographic Origin



(c) Geographic Area



(d) Region

Figure 3: KMeans Results

Thereafter, we conducted an importance analysis on the feature data and identified the top 10 SNPs with the highest importance for Principle Component 1 (PC1) and Principle Component 2 (PC2), as presented in Table 1. Interestingly, the two sets of top 10 SNPs have no overlap. Moreover, no individual SNP appears to dominate the corresponding principle component. Figure 4 provides a visual representation of the distribution of explanatory power for all SNPs on PC1 and PC2, revealing that over half of the SNPs have limited significance on the principle components, while only a few have relatively strong importance. These observations suggest that a single SNP may have little impact on an individual's population characteristics, but the combined differences in a significant number of SNPs may lead to the formation of a distinct population. This finding also highlights the stability and diversity of human genetics.

Lastly, we also explored other dimension reduction methods, such as MDS and t-SNE, and arrived at similar conclusions but from different angles. Figure 5 illustrates the MDS results, while Figure 6 shows the t-SNE results. In Figure 5d, we observe migration from the bottom right to the top left, i.e., from the Africa cluster to the American cluster. We also note some overlap between adjacent clusters, such as the Middle East and Europe or the Middle East and Central/South Asia, suggesting that people from Europe and Central/South Asia may have both migrated from the Middle East. Another interesting finding is that the Oceania cluster is split into two parts, one on the left and one on the right.

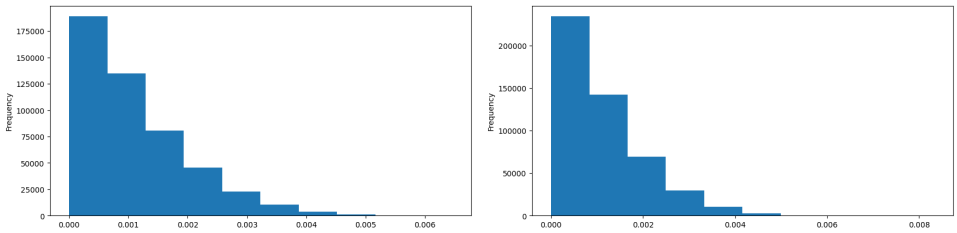
SNP	Importance
rs4918664	0.006456
rs10882168	0.006414
rs7091054	0.006382
rs11187300	0.006330
rs7556886	0.006310
rs12220128	0.006276
rs6583859	0.006270
rs4918924	0.006171
rs1834619	0.006151
rs4578856	0.006119

(a) PC1

SNP	Importance
rs1834640	0.008321
rs2250072	0.007733
rs260714	0.006920
rs10760260	0.006734
rs7531501	0.006731
rs2416899	0.006722
rs11637235	0.006701
rs618746	0.006655
rs595961	0.006609
rs10886189	0.006580

(b) PC2

Table 1: Top 10 SNPs



(a) PC1 (b) PC2

Figure 4: Feature Importance Distribution

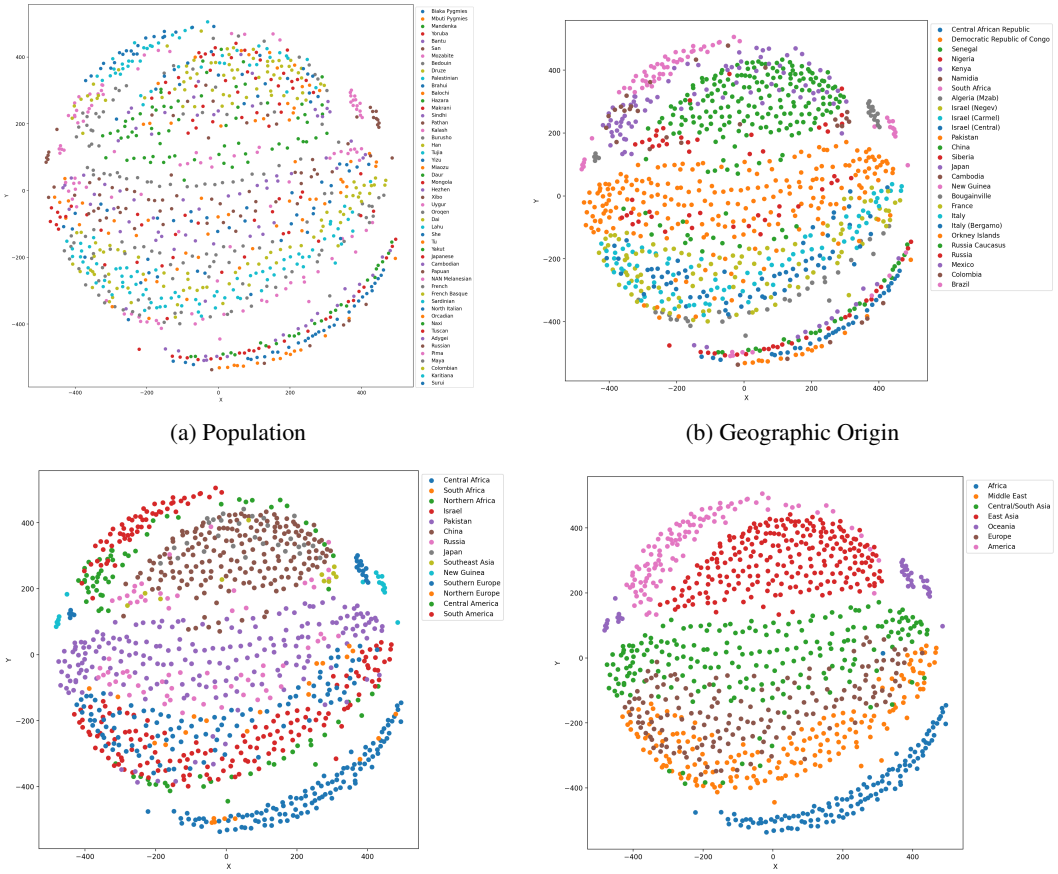
The left one is closer to the American cluster, while the right one seems closer to the East Asia cluster than the American cluster. By cross-checking the results from different data granularities, as shown in Figures 5b and 5a, we noticed that they still exhibit the same pattern. Hence, it is inappropriate to assume that some Oceania people migrated from East Asia while others migrated from America. However, the MDS clustering results may not be as good as the PCA method in addressing this issue. The root cause of this could be the insufficient amount of Oceania sample data, from which the KMeans-PCA method also suffers.

The t-SNE results presented in Figure 6 offer a more distinct clustering outcome. Although we cannot infer migration patterns from these results, we can establish certain relationships between various clusters. For instance, some individuals from the Central/South Asia cluster are closer to the East Asia cluster than their designated cluster. Upon further investigation, we found that these individuals are from certain populations in Pakistan, which is a neighboring country to China, an East Asian nation. Additionally, Figure 6c displays the Russian cluster as being distinctly separated, with some individuals situated in East Asia and others in Europe. This is consistent with the reality that Siberian people in Russia are essentially East Asian.

5 Discussions

In this study, we have explored the human migration history using SNP data and dimension reduction methods. Our results have shown the potential of using PCA, MDS, and t-SNE to reveal insights into the migration patterns of different human populations.

As shown in Figure 7, one interesting finding is that the first two principal components only explain a small portion of the variances in the data, which suggests that human genetics are highly diverse and complex. This diversity is also reflected in the clustering results, which show significant overlaps and separations between different populations. Another noteworthy result is the migration patterns of different populations. For example, the MDS results reveal human migration from Africa all the way to America, with some overlaps between adjacent clusters such as Europe and the Middle East. In contrast, t-SNE shows more separate clustering patterns, with some notable exceptions such as the proximity of some Central/South Asia individuals to the East Asia cluster.



(a) Population

(b) Geographic Origin

(c) Geographic Area

(d) Region

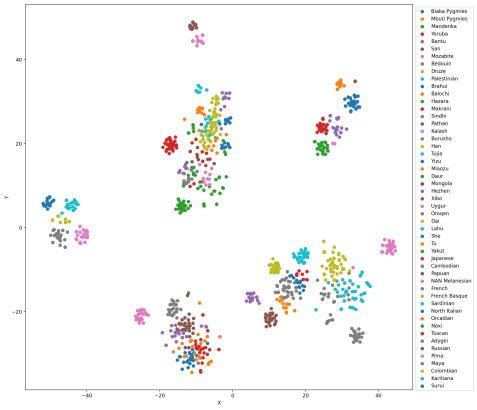
Figure 5: MDS Results

Our study also has some limitations. First, the sample size is relatively small compared to the dimensions of the features, which makes it challenging to directly extract information from the feature data. This limitation may affect the accuracy and generalizability of the results. Additionally, the study only focuses on SNPs data, while other genetic markers such as insertion-deletions (indels) and copy number variations (CNVs) may also play an essential role in human migration history.

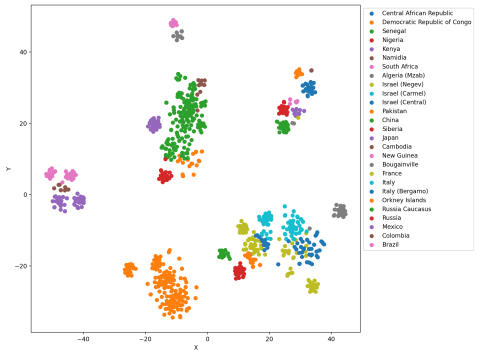
In short, our study sheds light on the complexity and diversity of human genetics and highlights the potential of using dimension reduction methods to study human migration history. Future studies can build upon our work and incorporate more comprehensive genetic data to obtain a more accurate and complete picture of human migration history.

6 Conclusions

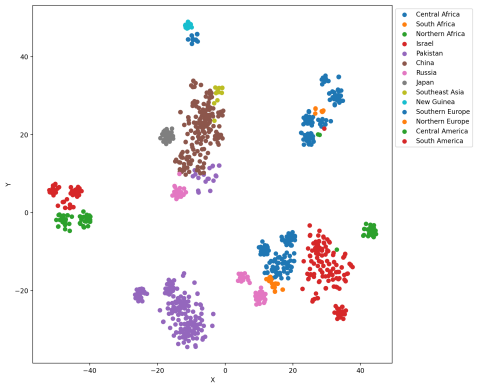
In conclusion, our study on human migration history using SNPs data and dimension reduction methods has yielded interesting results. We have identified several migration patterns, including the migration from Africa to America, and the possible migration of Europe and Central/South Asia populations from the Middle East. PCA has proven to be a suitable method for dimension reduction of SNPs data, as it can effectively capture the variation in the data and provide useful information for clustering and analysis. MDS and t-SNE also provide valuable insights into the relationships between different populations, although the clustering results may not be as clear as those obtained from PCA. Despite the limitations of our study, such as the small sample size and the lack of data from certain regions, we believe that our findings can contribute to the understanding of human migration history and the genetic diversity of human populations. Future studies could expand the sample size and incorporate additional genetic markers to further explore the complex relationships between



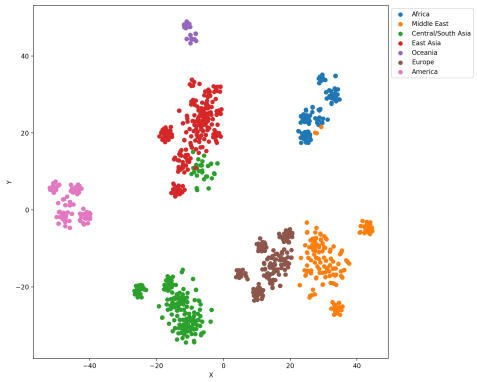
(a) Population



(b) Geographic Origin



(c) Geographic Area



(d) Region

Figure 6: t-SNE Results

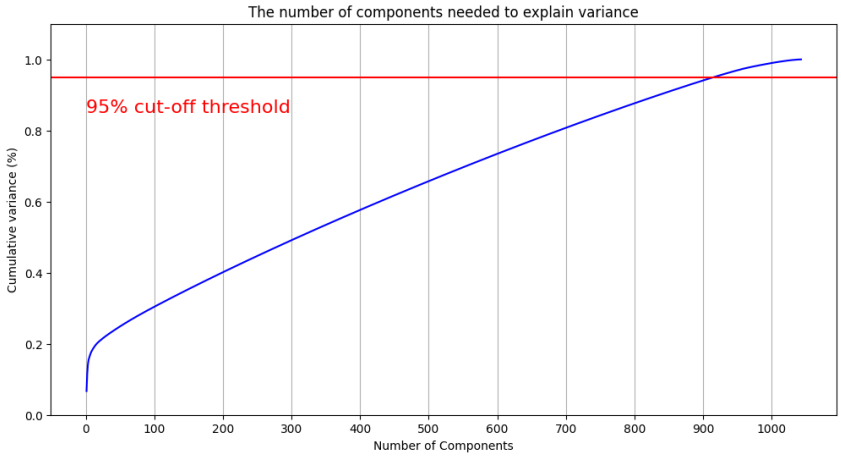


Figure 7: Number of Components vs Variance Explained

different populations. Overall, our study highlights the importance of using advanced statistical and computational methods to analyze complex genetic data and uncover new insights into the history of human migration.

References

[1] T Douglas Price, Corina Knipper, Gisela Grupe, and Václav Smrcka. Strontium isotopes and prehistoric human migration: the bell beaker period in central europe. *European journal of archaeology*, 7(1):9–40, 2004.

[2] Ananyo Choudhury, Shaun Aron, Laura R Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik Bensellak, Gordon Wells, Judit Kumuthini, Daniel Shriner, Yasmina J Fakim, et al. High-depth african genomes inform human migration and health. *Nature*, 586(7831):741–748, 2020.

[3] Paul Kerswill. Migration and language. *Sociolinguistics/Soziolinguistik. An international handbook of the science of language and society*, 3:1–27, 2006.

[4] Hua Liu, Franck Prugnolle, Andrea Manica, and François Balloux. A geographically explicit genetic model of worldwide human-settlement history. *The American journal of human genetics*, 79(2):230–237, 2006.

[5] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104, 2008.

[6] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.