Population Genetics

GPSmaf: Tracing Ancient Populations with Deep Learning

Arash Darzian[1,*]

[1*]Department of Biology, Lund University, 22362 Lund, Sweden

To whom correspondence should be addressed.

Abstract

Motivation: We present GPSmaf, a novel methodology for inferring the geographic population structure from ancient DNA data using deep learning techniques. Our approach can complements existing methods and aims to improve resolution of population inference. By utilizing data preprocessing, allele frequency calculation, training of autoencoders, and UMAP clustering analysis, we demonstrate the effectiveness of GPSmaf in capturing population structure patterns regardless of abundant missingness in ancient DNAs.

Results: The results reveal distinct clusters of countries based on geographic proximity, indicating the impact of continental relationships on genetic similarities. Additionally, within specific time periods, our analysis uncovers potential population subdivisions and genetic differentiation within countries.

Availability: To facilitate accessibility, we developed an interactive web application (https://gpsmaf.streamlit.app/) for visualizing and exploring the population structure findings. Also, Data and code presented in github repository (https://github.com/arash-darzian/Geographic_Population_Structure_GPS).

Contact: ar7363da-s@student.lu.se

1 Introduction

The study of human population history and migration patterns has been greatly enriched by the analysis of ancient DNA. By extracting genetic material from ancient remains, researchers have gained insights into the ancestral relationships, demographic changes, and geographic distributions of past populations. Inferring the geographic population structure from ancient DNA data is a fundamental aspect of understanding human evolution and dispersal (Nielsen et al., 2017).

Significant progress has been made in deciphering ancient population structure using various genetic markers, including single nucleotide polymorphisms (SNPs) (Elhaik et al., 2014). These genetic markers provide valuable information about the genetic variation within and between populations. By analyzing the minor allele frequencies (MAF), researchers have been able to discern population clusters and trace migration routes.

Previous studies have employed admixture analysis, a commonly used approach for inferring population structure. Admixture analysis estimates the proportions of ancestry derived from different ancestral populations. While this method has been successful in revealing broad-scale population patterns, it has limitations in capturing fine-grained population structure and subtle genetic variations.

While numerous studies have contributed to our understanding of ancient population structure using admixture analysis, several knowledge gaps remain. Firstly, the accuracy and resolution of population inference methods based on SNP data can be limited by the availability of comprehensive and representative ancient DNA datasets and missingness. Although substantial efforts have been made to increase the number of available ancient DNA samples, the field would greatly benefit from larger and more diverse datasets encompassing different geographic regions and time periods. Though, working with these data and reducing of the effect of missingness that is present in them needs to address and handle in a more comprehensive mannar.

Secondly, the application of deep learning techniques to ancient DNA analysis represents a promising avenue for enhancing population structure inference. Integrating deep learning approaches with ancient DNA analysis could potentially improve the efficiency of inferring geographic population structure by reducing the missingness effect.

In this study, our aim is to address these challenge by developing a methodology for inferring the geographic population structure from

allele frequency of SNPs in an ancient DNA datasets with huge number of missing data points. We harness the power of deep learning techniques, employing an ensemble of autoencoder and decoder to extract crucial features from the SNP data. By training these models on the comprehensive ancient DNA dataset, we aim to uncover hidden patterns and capture population structure information that may have been previously undetected.

based on the "date" column, retaining samples predating 15,000 years ago to ensure sufficient sample size.

2.2 Grouping of Samples

Samples were grouped based on current political entities (country) that they have found and corresponding BP date, labeled as "country_date." Groups with less than two samples were excluded to ensure robust population representation.

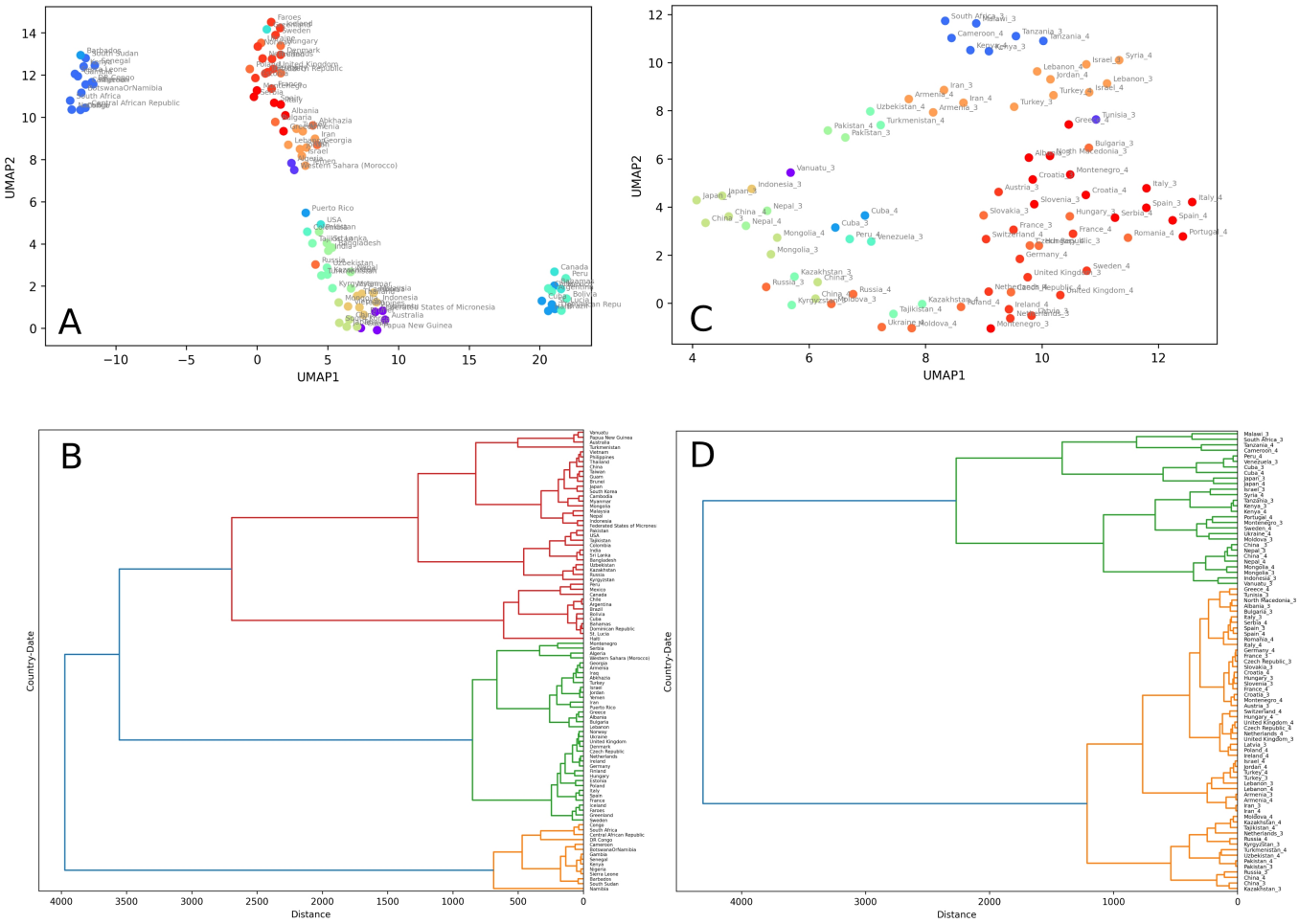

Figure 1: Visualization and Clustering Results for Different Time Periods **A)** UMAP projection of country-date groups available for the time period of 0-1000 years before present (bp). Each color represents a distinct region. **B)** Dendrogram of hierarchical clustering analysis performed on all country-date groups available for the time period of 0-1000 years bp. **C)** UMAP projection of country-date groups available for the time period of 2001-4000 years bp. Each color represents a distinct region. The labels follow the format: <CountryName>_<int(bp date / 1000)>. **D)** Dendrogram of hierarchical clustering analysis performed on all country-date groups available for the time period of 2001-4000 years bp. The labels follow the format: <CountryName>_<int(bp date / 1000)>.

2 Methods

2.1 Data Collection and Preprocessing

The dataset comprised PLINK files and annotation data obtained from the BINP29 course instructor, consisting of 16,465 samples and 1,233,013 SNPs. Missing data was filtered, removing samples with more than 70% missing data and SNPs with more than 20% missing data. Further filtering excluded samples with more than 10% missing data and SNPs with more than 6% missing data. The dataset was then filtered

2.3 Calculation of Allele Frequencies

Missing values in data were filled with the mode of the respective group's allele frequencies, or assigned a value of 0 if no mode was present. Allele frequencies for each country_date group were computed.

2.4 Feature Extraction and Clustring

An ensemble of autoencoders was trained to extract meaningful features from the SNP data. Autoencoders consisted of encoder and decoder layers, with progressively decreasing and increasing dimensions, respectively. Training utilized the Adam optimizer and MSE loss function. Early stopping prevented overfitting, retaining the best weights based on validation loss. Encoded data was extracted from each autoencoder, resulting in a concatenated numpy array. This lower-

dimensional representation captured distinct encodings from each autoencoder.
The encoded data array was subjected to the UMAP algorithm for dimensionality reduction and clustering analysis. A distance matrix was computed based on UMAP results to construct a dendrogram, visualizing hierarchical relationships between population groups.

2.5 Development of Interactive Web Application

An interactive web application was developed using the Streamlit framework to present and visualize clustering results. Users could explore population structure, navigate dendrograms, and examine genetic composition and relationships within specific country_date groups.

2.6 Performance Evaluation:

To evaluate the performance of our population structure inference, we assessed the agreement between the inferred genetic clusters and the geographical proximity of the country_date groups.

3 Results & Discussions

The UMAP plot revealed distinct clustering patterns based on continent, indicating a successful grouping of countries. Figure 1 provides a visual representation of the clustering results, highlighting the relationships between different regions.

In Figure 1A and 1B (the time period of bp = 0-1000 years), it can be observed that countries in North Africa are clustered closely with South Africa, indicating a genetic affinity between these regions. Additionally, the plot shows the proximity of different parts of Asia. Similarly, West Asia clusters closely with East Europe, implying shared genetic characteristics and potential historical interactions between these regions. South Asia is observed to cluster with East Asia and Oceania, suggesting genetic affinities among populations in these areas. Furthermore, the clustering analysis reveals distinct genetic clusters for USA in North America and South America, Central America and the Caribbean, aligning with populating of USA.

Figure 1C and 1D shows that within the time range of 2001-4000 BP (bp 3 and bp 4), some country_date groups exhibited distinct patterns. In particular, there were instances where a single country was represented by multiple groups, denoted as countryName_3 and countryName_4. The UMAP plot demonstrated variations in distance and clustering between these groups, suggesting potential population subdivisions or genetic differentiation.

4 Conclusion

Overall, this study provides evidence that the clustering analysis based on deep learning approach and encoded data successfully captures the population structure and relationships among different regions and can handle the missingness which is pretty abundant in ancient DNA. These findings highlight the genetic affinities and historical connections between populations across continents and provide valuable insights into ancient population dynamics.
It is important to note that the results presented in the results section represent a snapshot of the population structure at specific time periods and should be interpreted in the context of the study's limitations and the complex nature of human migration and genetic admixture. Further analysis and validation using additional genetic markers and comprehensive datasets would contribute to a more comprehensive understanding of ancient population structure.

References

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017 Jan 19;541(7637):302-10.

Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calò C, De Montis A, Atzori M, Marini M, Tofanelli S, Francalacci P, Pagani L. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nature communications. 2014 Apr 29;5(1):3513.