Mathematical Methods in Data Science and Signal Processing

Final Project

Nadav Marciano 305165698

The study ***"Genes Mirror Geography within Europe"***, led by John Novembre and colleagues, explores the connection between genetic variation and geography in European populations using computational methods for genetic data analysis. The main objective was to determine whether genetic data can accurately predict an individual's geographic origin and to assess the impact of population structure on genome-wide association studies (GWAS). A key question was whether European populations form distinct genetic clusters or if genetic variation follows a continuous geographic pattern. To investigate this, the researchers analyzed a dataset of 3,192 individuals from different European countries, collected through the POPRES project. Each individual was genotyped at 500,568 SNP loci using the Affymetrix 500K SNP chip. The geographic origin of participants was determined based on their grandparents' country of birth, or, if unavailable, their own reported country of birth. To ensure consistency, individuals of non-European origin were excluded, and the dataset was adjusted to balance sample sizes and remove extreme genetic outliers. PCA (Principal Component Analysis) was applied to identify genetic variation patterns in two dimensions, revealing whether genetic differences align with geographic locations. The PC1 and PC2 axes were rotated to optimize correlation with latitude and longitude using the following function:

$$f(\theta) = Cor(g(\theta, v_1, v_2), Long) + Cor(h(\theta, v_1, v_2), Lat)$$

where $g(\theta, v_1, v_2)$ and $h(\theta, v_1, v_2)$, return the rotated PC1 and PC2 coordinates, and Cor represents the correlation function. The optimal rotation angle found was 216°. After identifying genetic variation patterns through PCA, multiple linear regression was used to estimate latitude and longitude based on the rotated PC1 and PC2 values. This step aimed to evaluate how well genetic data could predict an individual's approximate location. Additionally, a discrete assignment approach was performed, where individuals were assigned to the country whose geographic center was closest to their estimated coordinates. Both methods produced similar results, reinforcing the idea that genetic data can effectively infer geographic origin The study demonstrated a strong correlation between genetic variation and geographic distances. The PCA analysis revealed a structure closely resembling a map of Europe, with individuals clustering according to their geographic proximity. In practice, 50% of individuals were assigned within 310 km of their reported location, and 90% within 700 km, indicating a high degree of accuracy in predicting geographic origin from genetic data. These findings emphasize the need to account for population structure in GWAS, as failing to do so may lead to false associations. The results support the idea that DNA data provides reliable information about geographic ancestry, with important applications in genetic ancestry testing, forensic science, and anthropology.

The objective of this project was to apply the algorithm described in the study *"Genes Mirror Geography within Europe"* to a different dataset and assess whether similar patterns of genetic variation and geographic distribution could be observed. For this purpose, the 1000 Genomes Project dataset was selected, which contains genetic information from 2,504 individuals from diverse populations. Given the study's focus on European populations, chromosome 22 was chosen as the initial dataset for analysis. Chromosome 22 is one of the smallest autosomal chromosomes and contains significant genetic markers used in population genetics studies. The genetic information in this dataset is encoded as SNP (Single Nucleotide Polymorphism) variations, which serve as key indicators of genetic diversity across populations. As a preprocessing step, the individuals were first sorted by country of origin. Only individuals from European populations were retained, specifically those belonging to the following groups: GBR (British), FIN (Finnish), IBS (Iberian Population in Spain), CEU (Utah residents of European ancestry), and TSI (Tuscans from Italy). After filtering, 503 individuals remained for further analysis. To mitigate the effect of linkage disequilibrium (LD), SNP pruning was performed using the *PLINK* tool. The filtering process removed 69,945 SNPs, leaving 26,739 variants from chromosome 22 for analysis. Following the preprocessing steps, Principal Component Analysis (PCA) was performed to identify the primary axes of genetic variation. PCA is a widely used technique in genetic studies that reduces dimensionality while preserving key variance structures in the data. To align the principal components with geographic coordinates, the axis rotation method described in the original study was applied. The optimal rotation angle was found to be 153.03°, which was determined using the optimization function provided in the study. Upon visualization of the results, it became evident that the findings did not align with those reported in the original paper (*See Figures 1 and 2*).

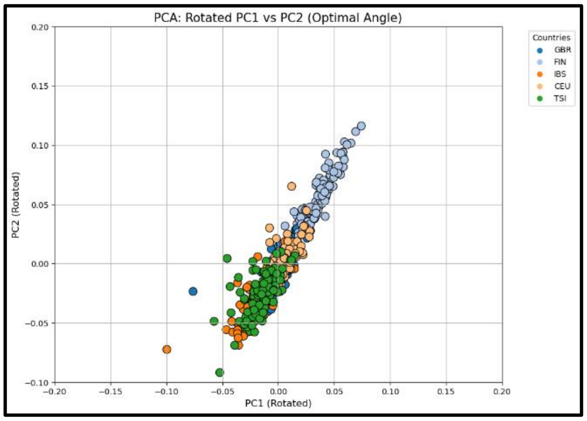

Figure 1 - PCA of Chromosome 22

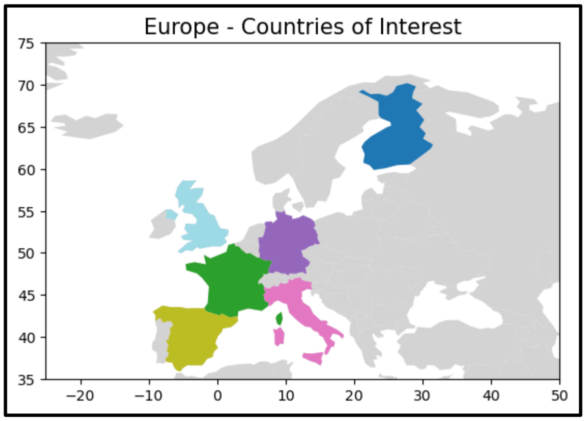

Figure 2 – Europe: Countries of Interest

The expected clustering of populations according to geographic proximity was not clearly observed. However, Finnish individuals appeared more distinct compared to the other European populations. This discrepancy could be attributed to several factors. One key issue is that the 1000 Genomes dataset does not provide information on the ancestral origins of the participants' grandparents, making it difficult to account for recent migration or population movements. Additionally, Finland is geographically and genetically more isolated compared to other European populations, which may explain why it showed a more distinct genetic pattern. proximity was not clearly observed. To further investigate whether the results were influenced by chromosome selection, the same analysis was repeated using chromosome 1, which is significantly larger and contains a greater number of genetic markers. The same preprocessing steps, including SNP pruning and PCA with axis rotation, were applied. The results from chromosome 1 were consistent with those from chromosome 22, showing that Finnish individuals remained the most distinct group, while other populations did not exhibit clear clustering patterns (*See Figure 3*). Several improvements could be made to refine the analysis. A larger and more diverse dataset with reliable information about individuals' ancestral origins, including their grandparents' birthplaces, would provide better control over population structure. Additionally, performing an initial PCA outlier filtering step using *SmartPCA* could help remove individuals who do not fit within the main population clusters. An attempt was made to use *SmartPCA* for this purpose, but compatibility issues arose due to discrepancies in the formatting of individual IDs. A broader dataset with better-documented ancestry information would allow for more robust comparisons and a clearer interpretation of genetic-geographic correlations. By refining the dataset and applying more sophisticated filtering techniques, it may be possible to improve the accuracy of population assignments and achieve results that more closely mirror those observed in the original study.

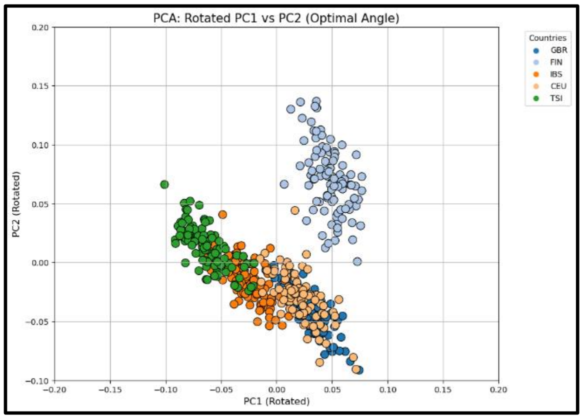

Figure 3 - PCA of Chromosome 1

Several improvements can be made to enhance the accuracy of genetic-geographic inference:

- **Expand and refine the dataset**: Use a larger dataset with detailed ancestral origins to reduce bias from recent migrations.
- **Improve PCA methodology**: Use alternative dimensionality reduction techniques (e.g., t-SNE, UMAP) and optimize axis rotation through direct geographic regression.
- **Integrate clustering algorithms**: Utilize ADMIXTURE or Gaussian Mixture Models (GMM) to better capture population structure beyond PCA.

These enhancements can improve the accuracy of genetic inference and provide more reliable insights into population structure.

Future research should focus on improving genetic-geographic inference by incorporating advanced machine learning techniques alongside traditional PCA-based methods. While PCA is widely used for dimensionality reduction, it assumes a linear structure in genetic variation, which may not fully capture the complexity of population genetics. To address this, alternative approaches such as t-SNE and UMAP should be explored, as they can detect non-linear population structures that PCA might overlook. Additionally, deep learning models could analyze large-scale genomic datasets to uncover hidden genetic patterns and refine ancestry estimation. The first step in such research is to compile a high-quality dataset with detailed genetic information from diverse populations, ensuring well-documented ancestral origins to minimize bias from recent migrations. Once established, a direct comparison between PCA and machine learning approaches (t-SNE, UMAP, and deep learning models) should be conducted to determine which method best captures genetic variation. Metrics such as clustering accuracy, explained variance, and silhouette scores should be used to assess effectiveness. Another approach involves using unsupervised clustering algorithms such as ADMIXTURE and Gaussian Mixture Models (GMM), which provide probabilistic assignments of individuals to populations rather than rigid classifications. Comparing these techniques with PCA-based classification can help determine whether machine learning enhances population structure analysis. Additionally, hybrid models that integrate PCA with machine learning clustering could further improve resolution. Since deep learning and clustering models require significant computational resources, future studies should focus on optimizing efficiency through GPU acceleration or cloud computing. Evaluating these models across different datasets and chromosomes would provide deeper insights into their robustness and generalizability. By integrating high-resolution datasets, non-linear dimensionality reduction techniques, and advanced clustering models, future research can significantly enhance the accuracy of genetic-geographic inference.

In conclusion, the findings suggest that the genetic-geographic inference algorithm does not perform consistently across all datasets and chromosomes. The accuracy of the results depends on both the genetic markers analyzed and the quality of population ancestry data. Ensuring a more comprehensive dataset, including ancestral origins of participants' grandparents, can significantly improve classification reliability.

This study highlights the strong correlation between genetic variation and geographic distribution, showing how historical migration and population structure shape genetic patterns across Europe. As Charles Darwin aptly stated:
"The inhabitants of each country become modified and adapted to their varied surroundings."
(*On the Origin of Species*, 1859).