Atlas Journal of Biology 2025, pp. 839–845
https://doi.org/10.5147/ajb.260

Published in the United States of America
https://www.atlas-publishing.org

Dissecting Continental and Intra-European Genetic Structure Using Chromosome 22 SNPs from The 1000 Genomes Project

My Abdelmajid Kassem

Plant Genomics and Bioinformatics Lab, Department of Biological and Forensic Sciences, Fayetteville State University, Fayetteville, NC 28301, USA.

Received: May 12, 2025 / Accepted: June 21, 2025

Abstract

Understanding the population structure of global human groups remains fundamental to population genetics and medical genomics. In this study, I analyzed genomic variation from Chromosome 22 using publicly available data from the 1000 Genomes Project, focusing on four populations: Yoruba (YRI), Iberian (IBS), Tuscan (TSI), and Utah residents of Northern/Western European ancestry (CEU). Using ~50,000 high-quality biallelic SNPs, I applied principal component analysis (PCA), ADMIXTURE, pairwise FST, multidimensional scaling (MDS), and phylogenetic clustering to characterize inter-population relationships. PCA revealed a strong continental split between African and European individuals, with minimal separation among the European subgroups. ADMIXTURE analysis (K = 4) confirmed this pattern, showing consistent European ancestry clusters and distinct divergence from African ancestry. Pairwise FST values highlighted low differentiation among the European groups (FST ≈ 0.0016–0.0030) and a much higher divergence from the YRI population (FST > 0.137). The site frequency spectrum was dominated by rare variants, in line with recent population expansions. This study demonstrates that even a single chromosome's SNP subset can robustly capture major axes of genetic structure, offering a scalable model for population genomics education and exploratory research.

Keywords: Human population structure, Chromosome 22, 1000 Genomes Project, Principal Component Analysis (PCA), ADMIXTURE, Genetic differentiation (FST), Population genomics, Ancestry inference.

* Corresponding author: mkassem@uncfsu.edu

1. Introduction

Understanding the genetic structure of human populations is central to population genomics, medical genetics, and anthropology. The spatial distribution of genetic variation among human groups reflects complex historical processes, including ancient migrations, geographic isolation, genetic drift, and admixture. Continental-scale divergence – such as that between African and non-African populations – is a well-documented outcome of the "Out of Africa" expansion of anatomically modern humans (Tishkoff et al., 2009; Pagani et al., 2016). At a finer scale, intra-continental variation – particularly in Europe – has revealed subtle but consistent population structure that often mirrors geography (Novembre et al., 2008; Lao et al., 2008).

The availability of large-scale, publicly accessible genomic datasets such as the 1000 Genomes Project (1KGP) has significantly enhanced the ability to explore human genetic diversity (The 1000 Genomes Project Consortium, 2015). These data sets provide dense, high-quality genotype information across hundreds of individuals from diverse populations, enabling robust analyses of genetic structure and ancestry patterns. Such analyses are essential not only for understanding evolutionary and demographic history but also for improving the accuracy of disease association studies and personalized medicine in diverse populations (Martin et al., 2017).

In this study, I analyze data from four representative populations in the 1000 Genomes Project: the Yoruba population from Nigeria (YRI), Iberians from Spain (IBS), Tuscans from Italy (TSI), and Utah residents with Northern and Western European ancestry (CEU). This population set captures both broad continental divergence (Africa vs. Europe) and fine-scale European variation. Previous research has consistently found a strong separation between African and non-African populations, attributed to the bottleneck effect and genetic drift following the migration out of Africa ~60,000 years ago (Reich et al., 2009; Rosenberg et al., 2002). European populations, while more genetically homogeneous compared to Africans, still exhibit discernible north-south and east-west clines in genetic variation (Novembre et al., 2008; Lao et al., 2008; Capocasa et al., 2014).

Although most population genomic studies analyze the entire genome, chromosome-specific analyses provide a practical alternative for focused studies. Chromosome 22 is one of the smallest human autosomes but is gene-rich and has been previously utilized to examine genetic diversity, recombination hotspots, and selective sweeps (McVean et al., 2009; Novembre and Ramachandran, 2011). Analyzing a representative autosome such as chromosome 22 allows researchers to reduce computational complexity while still capturing informative patterns of genetic structure, especially when using dense single nucleotide polymorphism (SNP) data.

In this work, I uzed a subset of ~50,000 biallelic SNPs randomly selected from chromosome 22 to assess population differentiation and structure among YRI, IBS, TSI, and CEU samples. I applied a suite of complementary analytical methods including Principal Component Analysis (PCA), ADMIXTURE clustering (Alexander et al., 2009), pairwise F_{ST} statistics (Weir and Cockerham, 1984), Multidimensional Scaling (MDS), and phylogenetic clustering. These tools are widely used in human genomics to visualize genetic relationships, infer ancestry proportions, and quantify genetic differentiation (Patterson et al., 2006; Rosenberg et al., 2002).

PCA and MDS are powerful techniques for reducing high-dimensional genotype data into interpretable axes of variation, often revealing broad population clusters that correlate with geography (Novembre et al., 2008). ADMIXTURE, a maximum-likelihood model-based method, is used to infer the proportion of an individual's genome originating from different ancestral populations. F_{ST} metrics, meanwhile, provide a quantitative measure of genetic differentiation between populations and can help interpret clustering and ancestry patterns in evolutionary terms.

I also explored the site frequency spectrum (SFS), which summarizes the distribution of allele frequencies across populations. The SFS is shaped by demographic history and natural selection and often exhibits a strong skew toward rare variants in expanding populations (Keinan and Clark, 2012). Additionally, I used hierarchical clustering to construct a simple phylogenetic representation of the genetic distances among individuals, providing an intuitive visualization of population relationships.

This chromosome-focused study has both scientific and practical motivations. Scientifically, it contributes to my understanding of continental and sub-continental genetic structure by confirming patterns observed in full-genome studies using a smaller genomic subset. Practically, it demonstrates how publicly available data, and lightweight computational pipelines can be used to teach core concepts in population genetics and to explore hypothesis-driven questions in resource-constrained settings.

By combining multiple lines of evidence, I aimed to (1) demonstrate clear population structure between African and European populations, (2) identify subtle genetic differentiation within Europe, and (3) show how a single-chromosome approach can replicate major population genomic patterns. This work thus serves as a scalable and reproducible framework for both educational and exploratory research in human population genetics.

2. Methods

2.1. Data Source and Sample Selection

I used variant call format (VCF) files from Chromosome 22 of the 1000 Genomes Project Phase 3 release (Danecek et al., 2011; The 1000 Genomes Project Consortium, 2015), specifically the high-coverage phased genotypes:

[ALL.chr22.phase3_shapeit2_mvncall_integrated_v5b.20130502.genotypes.vcf.gz]

Sample metadata were obtained from the corresponding 1000 Genomes annotation file 20130606_sample_info.txt. From the full dataset, four populations were selected based on geographic and ancestral relevance: CEU (Utah Residents with Northern and Western European Ancestry), TSI (Toscani in Italia), IBS (Iberian Population in Spain), YRI (Yoruba in Ibadan, Nigeria). A total of 421 individuals were retained after filtering for completeness and ancestry. Chromosome 22 was selected due to its relatively small size, high gene density, and frequent use in population genetic studies. Its manageable size allows for efficient computational analysis while still capturing relevant evolutionary signals, making it ideal for both educational use and exploratory genomic work (McVean et al., 2009; Novembre and Ramachandran, 2011).

2.2. VCF Processing and Genotype Extraction

I used bcftools v1.21 to subset the sample population and index the VCF file. The following command was used:

[bcftools view -S sample_ids.txt -Oz -o chr22_subset.vcf.gz ALL.chr22*.vcf.gz]

This step extracted the four target populations from the full chromosome 22 dataset. The resulting VCF file was then parsed in Python using the scikit-allel package (Miles et al., 2019). Only biallelic SNPs with no missing genotype calls were retained. To optimize computational performance, a filtered set of 50,050 SNPs was randomly sampled for downstream analyses including PCA and ADMIXTURE. No LD pruning or Hardy-Weinberg equilibrium filtering was applied in this analysis to preserve maximal variance for PCA and ADMIXTURE inference. All SNPs used were biallelic and had complete genotype calls.

2.3. Principal Component Analysis (PCA)

Genotype data were converted to alternate allele count matrices, transposed to a samples × variants format, and passed to scikit-learn's PCA module (Pedregosa et al., 2011). The top two principal components were visualized using matplotlib and seaborn. Additional PCA was performed stratified by gender to explore sex-specific variation.

2.4. Site Frequency Spectrum

I computed the site frequency spectrum (SFS) by summing alternate allele counts across all individuals and plotting the distribution using 100 histogram bins. The SFS was used to infer overall patterns of allele frequency, highlighting common vs. rare variant contributions.

2.5. ADMIXTURE Analysis

The [chr22_subset.vcf.gz] file was converted to PLINK format using PLINK --vcf (Purcell et al., 2007), followed by ADMIXTURE analysis with K=4 using a Dockerized ADMIXTURE v1.3.0 container (Alexander et al., 2009). Cross-validation error was recorded across K=2 to K=6 to assess optimal clustering resolution, with the minimum CV error observed at K = 4, supporting the choice to emphasize this value in the analysis.

Ancestry proportions (Q matrix) were merged with population and gender metadata and visualized using stacked bar plots. Gender-specific ancestry profiles were explored using grouped boxplots and Mann–Whitney U tests for significance.

2.6. F_{ST} Calculation

Pairwise Weir and Cockerham's F_{ST} statistics (Weir and Cockerham, 1984) were computed using scikit-allel. The weir_cockerham_fst function was applied across the filtered Genotype Array for each pair of populations, and results were aggregated to compute average divergence across chromosome 22. The resulting F_{ST} matrix was visualized as a heatmap using Seaborn.

2.7. Multidimensional Scaling (MDS)

Pairwise genetic distances between individuals were computed from the alternate allele count matrix using Euclidean distance via scipy.spatial.distance.pdist. A 2D MDS embedding was computed using scikit-learn's MDS implementation.

2.8. Phylogenetic Tree Construction

A hierarchical clustering dendrogram was generated using the same genetic distance matrix. Ward's linkage was used via scipy.cluster.hierarchy.linkage. The tree was visualized with individual sample IDs.

Figure 1. PCA of Selected 1000 Genomes Populations (50K SNPs).

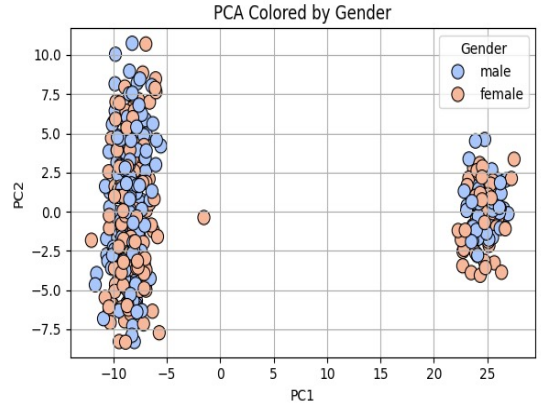

Figure 2. PCA Colored by Gender.

2.9. Software and Environment

All analyses were conducted in Python 3.11 (https://www.python.org/), using Anaconda on macOS with M1 Pro chip. Major packages include: Scikit-learn (Pedregosa et al., 2011), Scikit-allel (Miles et al., 2019), Numpy (Harris et al., 2020), Pandas (McKinney, 2010), Matplotlib (Hunter, 2007), Seaborn (Waskom, 2021), bcftools (Danecek et al., 2011, 2021), PLINK (Purcell et al., 2007), ADMIXTURE (Alexander et al., 2009), and Docker for containerized execution (Merkel, 2014).

All analyses were performed on a standard personal computer (MacBook Pro) running Mac OS Sonoma 14.4.1. All code and Jupyter Notebooks (Kluyver et al., 2016) are available upon request.

3. Results

3.1. Principal Component Analysis (PCA)

Figure 1 displays a principal component analysis of 421 individuals from four populations (IBS, CEU, TSI, YRI), based on 50,050 randomly sampled SNPs from chromosome 22. The first two principal components explain 11.88% and 0.78% of the total variance, respectively. PC1 clearly separates the African (YRI) population from the European clusters, while the three European groups—IBS, TSI, and CEU—exhibit substantial overlap, suggesting shared ancestry and limited substructure (Patterson et al., 2006). A few outlier individuals may reflect cryptic population substructure or recent admixture.

Figure 2 overlays gender on the same PCA coordinates. No visual evidence of sex-specific clustering was observed, indicating the absence of significant sex-biased allele frequency differences across these autosomal markers.

3.2. Site Frequency Spectrum (SFS)

Figure 3 presents the site frequency spectrum, which shows a pronounced skew toward low-frequency variants—i.e., SNPs with rare alternate alleles. This pattern is consistent with expectations from recent population growth and purifying selection acting on deleterious alleles (Keinan and Clark, 2012). The steep decline from rare to common variants suggests that the sampled SNPs effectively capture demographic signatures across the studied populations.

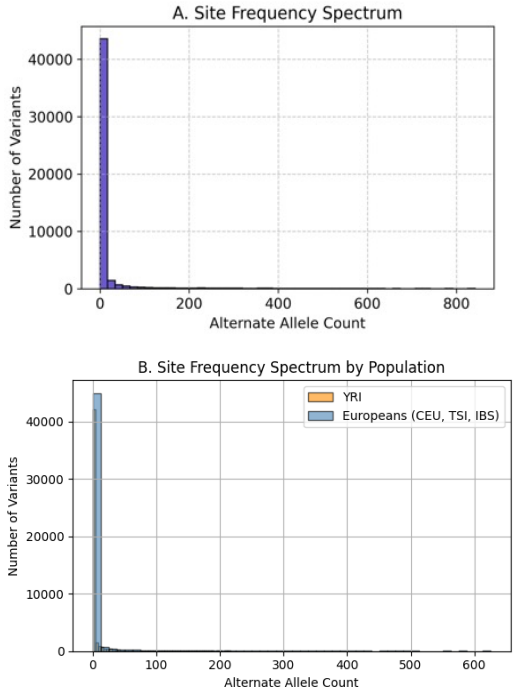

Figure 3. A. Site Frequency Spectrum. B. Site Frequency Spectrum by Population.

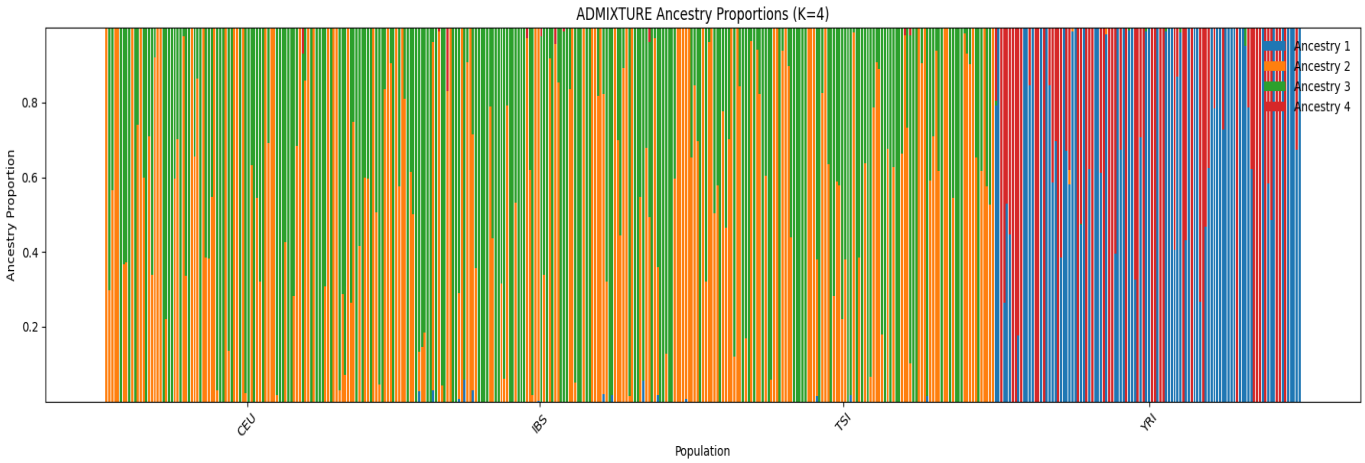

Figure 4. ADMIXTURE Ancestry Proportions (K=4).

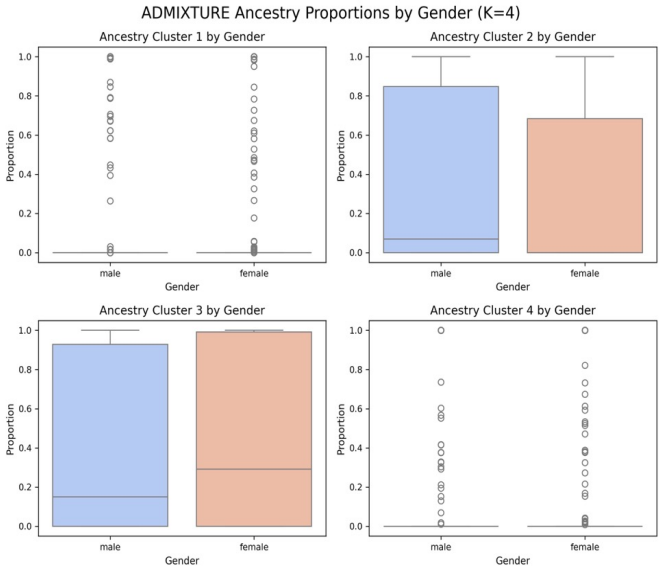

Figure 5. ADMIXTURE Ancestry Proportions by Gender (K=4).

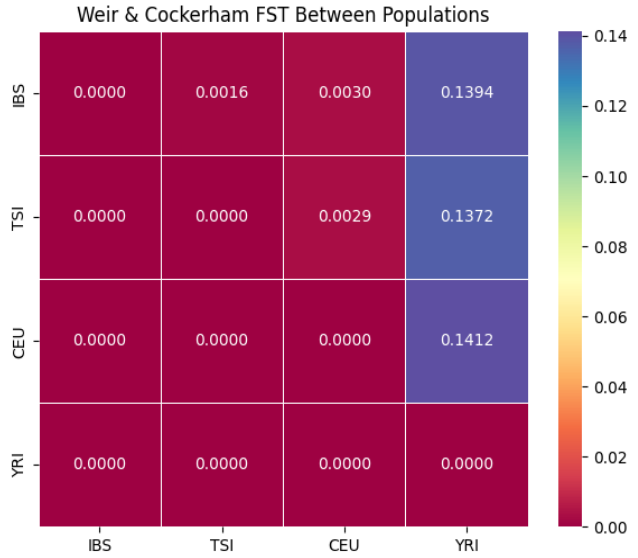

Figure 6. Weir & Cockerham FST Between Populations.

3.3. Population Structure via ADMIXTURE

Figure 4 shows ancestry estimates inferred using ADMIXTURE with K = 4. Individuals from the YRI population display nearly uniform membership in two distinct ancestry components (Clusters 1 and 4), whereas individuals from CEU, IBS, and TSI are primarily composed of Clusters 2 and 3. Among Europeans, TSI and IBS appear more similar to each other than to CEU, reflecting subtle intra-European differentiation. This pattern aligns with known geographic and historical relationships among these groups.

Figure 5 disaggregates ancestry proportions by gender for each cluster. Minor differences between males and females are visible but fall within expected stochastic variation. No consistent or statistically significant sex-specific ancestry patterns were detected, as expected for autosomal loci.

3.4. Pairwise Genetic Differentiation (F_{ST} Matrix)

Figure 6 summarize pairwise F_{ST} estimates computed using the Weir and Cockerham method (Weir and Cockerham, 1984). Genetic differentiation among the three European populations (IBS, TSI, CEU) is minimal, with F_{ST} values ranging from 0.0016 to 0.0030, indicating high genetic similarity and ex-

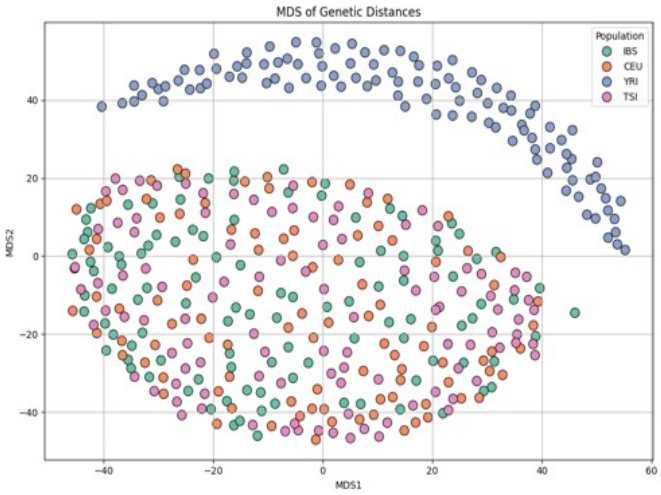

Figure 7. MDS of Genetic Distances.

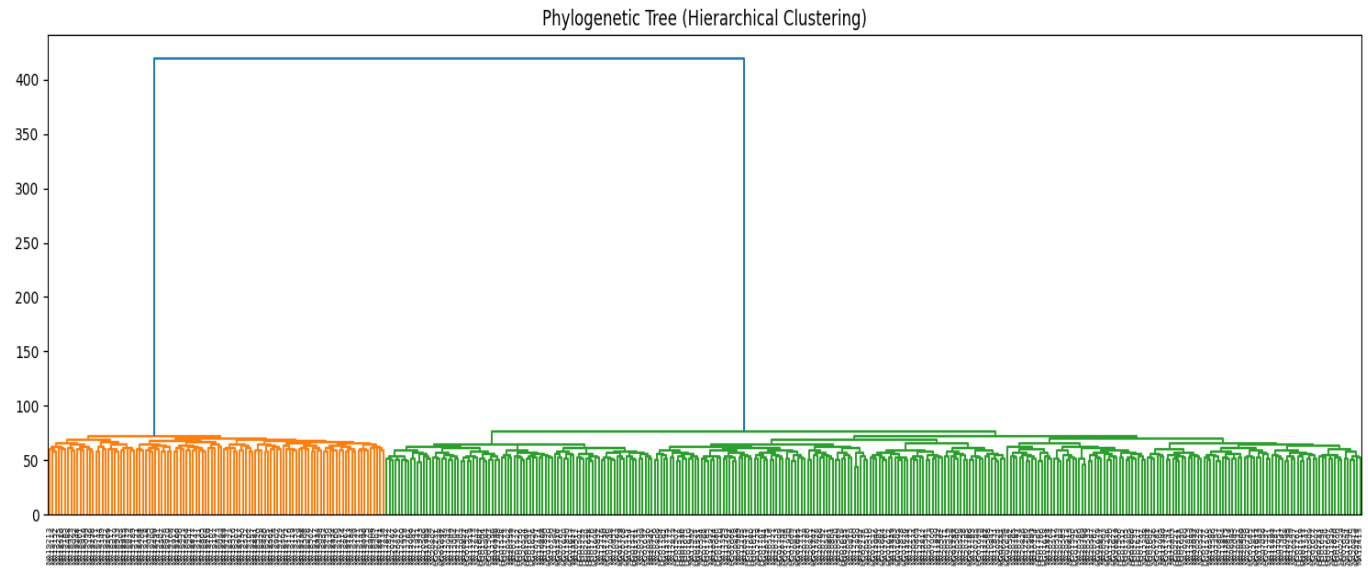

Figure 8. Phylogenetic Tree (Hierarchical Clustering).

tensive gene flow. In contrast, comparisons involving the YRI population yield substantially higher F_{ST} values (0.137–0.141), reflecting long-term geographic and evolutionary separation from the European groups (The 1000 Genomes Project Consortium, 2015).

These results corroborate PCA and ADMIXTURE findings, reinforcing the strong continental divergence between African and European populations and the fine-scale genetic continuity within Europe.

3.5. Multidimensional Scaling (MDS)

Figure 7 displays a two-dimensional embedding of pairwise genetic distances computed via multidimensional scaling. The MDS results mirror the PCA structure, with YRI individuals forming a distinct cluster separated from the European samples. Within Europe, CEU appears slightly shifted relative to TSI and IBS, which again cluster more closely together. This convergence of results across dimensionality reduction methods enhances confidence in the robustness of the inferred structure.

3.6. Phylogenetic Tree

Figure 8 presents a hierarchical clustering dendrogram constructed from pairwise genetic distances. Two major branches emerge: one composed entirely of YRI individuals and another containing the three European groups. Within the European clade, IBS and TSI cluster more closely, while CEU forms a distinct sub-branch. Although this method is not strictly phylogenetic in a population-genetic sense, it provides a visual summary of genetic distances shaped by historical demographic processes and geographic separation (Reich et al., 2009).

4. Discussion

This study leveraged publicly available whole-genome data from the 1000 Genomes Project to examine genetic structure and ancestry across four representative human populations—YRI, IBS, TSI, and CEU—using a subset of approximately 50,000 SNPs from chromosome 22. Despite being limited to a single chromosome, this targeted approach provided sufficient resolution to detect both broad continental divergence and subtle intra-European substructure, while also reducing computational demands.

4.1. Continental and Regional Structure

Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) consistently identified a dominant axis of genetic variation separating African and European populations, consistent with known human migration out of Africa (Tishkoff et al., 2009; Pagani et al., 2016). Within Europe, populations clustered tightly with slight differentiation, reflecting historically documented gene flow across the continent.

F_{ST} metrics and hierarchical clustering reinforced these patterns. Genetic distances between European populations were minimal ($F_{ST} > \approx 0.0016$–0.0030), whereas F_{ST} values between YRI and any European group were substantially higher (≈0.14), indicating deep evolutionary divergence (Rosenberg et al., 2002; Reich et al., 2009). Notably, TSI and IBS showed the closest affinity ($F_{ST} = 0.0016$), likely reflecting their shared Mediterranean ancestry and geographic proximity (Capocasa et al., 2014).

4.2. Site Frequency Spectrum and Rare Variants

The site frequency spectrum (SFS) exhibited a characteristic skew toward low-frequency alleles, a hallmark of recent population expansion and purifying selection (Keinan and Clark, 2012; Tennessen et al., 2012). This enrichment of rare variants supports the demographic history inferred from PCA and F_{ST} metrics and illustrates the utility of random SNP sampling in capturing global allele frequency distributions.

4.3. Ancestry Proportions and Admixture

ADMIXTURE analysis at K = 4 revealed clear differentiation between African and European ancestries, with YRI individuals displaying almost entirely distinct ancestry components from European populations. These clusters likely reflect deeper genetic structure within West Africa (Tishkoff and Williams, 2002; Gurdasani et al., 2015). European samples, particularly TSI and IBS, shared more mixed ancestry profiles, consistent with previously reported subcontinental structure (Nelson et al., 2008).

The similarity between TSI and IBS observed in both ADMIXTURE and PCA further supports historical gene flow across Southern Europe and the Mediterranean basin (Fiorito et al., 2016). The distinctiveness of CEU reflects its Northern/Western European background, in contrast to the more Mediterranean-affiliated Southern groups.

4.4. Sex-Specific Analyses

Analysis of ancestry proportions by gender revealed no significant differences across male and female individuals, confirming the expected uniformity of autosomal inheritance (Mathieson et al., 2015). While sex-biased admixture has been detected in studies using uniparentally inherited markers such as mtDNA and Y-chromosomes (Goldberg et al., 2017), such effects were not evident in

this autosomal analysis, highlighting the importance of genomic context when interpreting sex-specific genetic patterns.

4.5. Implications and Limitations

This study demonstrates the effectiveness of chromosome-specific SNP analysis in identifying both global and regional population structure. The reduced computational burden of using a single autosome, coupled with robust analytical methods, makes this approach suitable for educational purposes, preliminary research, and scalable genomic studies.

Nonetheless, several limitations merit discussion. First, results derived from chromosome 22 may not fully represent genome-wide patterns, especially those influenced by linkage disequilibrium or selection on other chromosomes. Second, while the sample sizes (~100 individuals per population) were sufficient to detect broad trends, more subtle signals (e.g., recent admixture or adaptive introgression) would require larger cohorts or whole-genome data. Finally, the study's focus on only four populations – though geographically informative – limits its applicability to broader global patterns of human genetic diversity.

Future work should consider expanding the population set, applying genome-wide SNP data, and integrating functional annotations to investigate population-specific variants with biomedical relevance.

5. Conclusion

This study presents a focused analysis of population genetic structure and ancestry using a subset of chromosome 22 SNPs from the 1000 Genomes Project. By integrating principal component analysis (PCA), ADMIXTURE clustering, F_{ST} statistics, multidimensional scaling (MDS), and hierarchical clustering, I uncover both broad continental divergence and fine-scale substructure among European populations. My key findings include:

- A clear genetic separation between African (YRI) and European (IBS, TSI, CEU) populations.
- Subtle yet consistent differences between Southern (IBS, TSI) and Northern/Western European (CEU) groups.
- A site frequency spectrum dominated by rare variants, consistent with recent population expansions.
- No evidence of significant sex-specific differences in ancestry distribution at the autosomal level.

These results demonstrate that even a single chromosome can yield robust insights into human population structure when paired with appropriate computational tools and population-genetic models. The analytical pipeline developed here is lightweight, reproducible, and adaptable, making it a valuable resource for exploratory research and population genomics education.

Future research may build upon this framework by incorporating genome-wide data, expanding the number of studied populations, and exploring the functional and biomedical relevance of population-specific genetic variation. Such efforts will be essential for improving our understanding of human evolutionary history and for ensuring that genomic research benefits diverse global populations.

Acknowledgments

I gratefully acknowledge the developers and maintainers of the open-source tools and packages used in this study, including bcftools, PLINK, ADMIXTURE, Docker, and the Python libraries scikit-allel, scikit-learn, numpy, pandas, matplotlib, and seaborn. Their contributions to the scientific community have made this research possible.

I also thank the 1000 Genomes Project Consortium for generating and publicly releasing the high-quality genomic data used in this analysis. Their work continues to be an invaluable resource for population genetics research and education.

References

Alexander, D.H.; Novembre, J.; Lange, K. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. Genome Research 2009, 19, 1655–1664. https://doi.org/10.1101/gr.094052.109.

Arauna, L.R.; Mendoza-Revilla, J.; Mas-Sandoval, A.; Izaabel, H.; Bekada, A.; Benhamamouch, S.; Fadhlaoui-Zid, K.; Zalloua, P.; Hellenthal, G.; Comas, D. Recent Historical Migrations Have Shaped the Gene Pool of Arabs and Berbers in North Africa, Molecular Biology and Evolution, 2017, 34(2), 318–329. https://doi.org/10.1093/molbev/msw218.

Arredi, B.; Poloni, E.S.; Paracchini, S.; Zerjal, T.; Fathallah, D.M.; Makrelouf, M.; Pascali, V.L.; Novelletto, A.; Tyler-Smith, C. A Predominantly Neolithic Origin for Y-Chromosomal DNA Variation in North Africa. Am. J. Hum. Genet. 2004, 75, 338–345. https://doi.org/10.1086/423147.

Bekada, A.; Arauna, L.R.; Deba, T.; Calafell, F.; Benhamamouch, S.; Comas, D. Genetic Heterogeneity in Algerian Human Populations. PLoS ONE, 2015, 10(9): e0138453. https://doi.org/10.1371/journal.pone.0138453.

Bekada, A.; Fregel, R.; Cabrera, V.M.; Larruga, J.M.; Pestano, J.; Benhamamouch, S.; Gonzalez, A.M. Introducing the Algerian Mitochondrial DNA and Y-Chromosome Profiles into the North African Landscape. PLoS ONE, 2013, 8(2): e56775. https://doi.org/10.1371/journal.pone.0056775.

Botigué, L.R.; Henn, B.M.; Gravel, S.; Maples, B.K.; Gignoux, C.R.; Corona, E.; Atzmon, G.; Burns, E.; Ostrer, H.; Flores, C.; Bertranpetit, J.; Comas D.; Bustamante, C.D. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. PNAS, 2013, 110(29), 11791-11796. https://doi.org/10.1073/pnas.1306223110.

Brisighelli, F.; Blanco-Verea, A.; Boschi, I.; Garagnani, P.; Pascali, V.L.; Carracedo, A.; Capelli, C.; Salas, A. Patterns of Y-STR variation in Italy. Forensic Science International: Genetics, 2012, 6, 6, 834-839. https://doi.org/10.1016/j.fsigen.2012.03.003.

Capocasa, M.; Anagnostou, P.; Bachis, V.; Battaggia, C.; et al. Linguistic, geographic and genetic isolation: a collaborative study of Italian populations. Journal of Anthropological Sciences, 2014, 92, 201-231. https://doi.org/10.4436/JASS.92001.

Cruciani, F.; La Fratta, R.; Santolamazza, P.; Sellitto, D.; Pascone, R.; Moral, P.; Watson, E.; Guida, V.; Colomb, E.B.; Zaharova, B.; Lavinha, J.; Vona, G.; Aman, R.; Calì, F.; Akar, N.; Richards, M.; Torroni, A.; Novelletto, A.; Scozzari, R. Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa. The American Journal of Human Genetics, 2004, 74(5), 1014–1022. https://doi.org/10.1086/386294.

Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; McVean, G.; Durbin, R. 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics, 2011, 27, 15, 2156–2158, https://doi.org/10.1093/bioinformatics/btr330.

Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; Li, H. Twelve years of SAMtools and BCFtools, GigaScience, 2021, 10, 2, giab008. https://doi.org/10.1093/gigascience/giab008.

Fadhlaoui-Zid, K., Martínez-Cruz, B., Khodjet-el-khil, H., Mendizabal, I.; Benammar-Elgaaied, A.; Comas, D. Genetic structure of Tunisian ethnic groups revealed by paternal lineages. American Journal of Physical Anthropology, 2011, 146(2), 271–280. https://doi.org/10.1002/ajpa.21581.

Fadhlaoui-Zid, K.; Haber, M.; Martínez-Cruz, B.; Zalloua, P.; Benammar-Elgaaied, A.; Comas, D. (2013) Genome-Wide and Paternal Diversity Reveal a Recent Origin of Human Populations in North Africa. PLoS ONE, 2013, 8(11): e80293. https://doi.org/10.1371/journal.pone.0080293.

Fiorito, G.; Di Gaetano, C., Guarrera, S.; Guarrera, S.; Rosa, F.; Feldman, M.W.; Piazza, A.; Matullo, G. The Italian genome reflects the history of Europe and the Mediterranean basin. European Journal of Human Genetics 2016, 24, 1056–1062. https://doi.org/10.1038/ejhg.2015.233.

Goldberg, A.; Gunter, T.; Rosenberg, N.A; Jacobson, M. Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. PNAS, 2017, 114(10), 2657–2662. https://doi.org/10.1073/pnas.1616392114.

Gurdasani, D.; Carstensen, T.; Tekola-Ayele, F.; Pagani, L.; Tachmazidou, I.; et al. The African Genome Variation Project shapes medical genetics in Africa. Nature 2015, 517, 327–332. https://doi.org/10.1038/nature13997.

Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585(7825):357-362. https://doi.org/10.1038/s41586-020-2649-2.

Henn, B.M.; Botigué, L.R.; Gravel, S.; Wang, W.; Brisbin, A.; Byrnes, J.K.; Fadhlaoui-Zid, K.; Zalloua, P.A.; Moreno-Estrada, A.; Bertranpetit, J., Bustamante, C.D., Comas, D. Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. PLoS Genetics, 2012, 8(1): e1002397. https://doi.org/10.1371/journal.pgen.1002397.

Hunter, J.D. Matplotlib: A 2D graphics environment. Computing in Science and Engineering 2007;9(3):90-95. https://doi.org/10.1109/MCSE.2007.55.

Keinan, A.; Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science, 2012, 336, 6082, 740-743. https://doi.org/10.1126/science.1217283.

Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonier, M.; Frederic, J.; Kelly, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.; Willing, C.. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, 2016:87-90. https://doi.org/10.3233/978-1-61499-649-1-87.

Lao, O.; Lu, T.T.; Nothnagel, M.; Junge, O.; Freitag-Wolf, S.; Caliebe, A.; Balascakova, M.; Bertranpetit, J.; Bindoff, L.A.; Comas, D.; Holmlund, G.; Kouvatsi, K.; Macek, M.; Mollet, I.; Parson, W.; et al. Correlation between genetic and geographic structure in Europe. Current Biology, 2008, 18(16), 1241–1248. https://doi.org/10.1016/j.cub.2008.07.049.

Maddy-Weitzman, B. "Notes". The Berber Identity Movement and the Challenge to North African States, New York, USA: University of Texas Press, 2011, pp. 211-254. https://doi.org/10.7560/725874-012.

Martin, A.R.; Gignoux, C.R.; Walters, R.K.; Wojcik, G.L.; Neale, B.M.; Gravel, S.; Daly, M.J., Bustamante, C.D.; Kenny, E.E. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. The American Journal of Human Genetics, 2017, 100, 4, 635-649. https://doi.org/10.1016/j.ajhg.2017.03.004.

Mathieson, I.; Lazaridis, I.; Rohland, N.; Mallick, S.; Patterson, N.; Alpaslan, S. Genome-wide patterns of selection in 230 ancient Eurasians. Nature 2015, 528, 499–503. https://doi.org/10.1038/nature16152.

McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. Austin, TX; 2010:51-56. https://doi.org/10.25080/Majora-92bf1922-00a.

McVean, G. A genealogical interpretation of principal components analysis. PLoS Genetics, 2009, 5(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686.

Merkel, D. Docker: lightweight linux containers for consistent development and deployment. Linux Journal, 2014, 239, 2.

Miles, A., Ralph, P., Rae, S., Pisupati, R. cggh/scikit-allel: v1.2.1. Zenodo, 2019. https://zenodo.org/record/3238280.

Montinaro, F; Busby, G.B.J.; Pascali, V.L.; Myers, S.; Hellenthal, G.; Capelli, C. Unravelling the hidden ancestry of American admixed populations. Nature Communications 2015, 6, 6596. https://doi.org/10.1038/ncomms7596.

Nelson, M. R.; Bryce, K.; King, K.S.; Indian, A.; Boyko, A.R.; et al. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. American Journal of Human Genetics, 2008, 83(3), 347–358. https://doi.org/10.1016/j.ajhg.2008.08.005.

Novembre, J.; Johnson, T.; Bryc, K.; Kutalik, Z.; Boyko, A.R.; Auton, A.; Indap, A.; King, K.S.; Bergmann, S.; Nelson, M.R.; Stephens, M.; Bustamante, C.D. Genes mirror geography within Europe. Nature 2008, 456, 98–101. https://doi.org/10.1038/nature07331.

Novembre, J.; Ramachandran, S. Perspectives on Human Population Structure at the Cusp of the Sequencing Era. Annual Reviews of Genomics and Human Genetics, 2011, 12, 245-274. https://doi.org/10.1146/annurev-genom-090810-183123.

Pagani, L.; Lawson, D.; Jagoda, E.; Morseburg, A.; Ericsson, A.; et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature 2016, 538, 238–242. https://doi.org/10.1038/nature19792.

Patterson, N.; Price, A.L.; Reich, D. Population Structure and Eigenanalysis. PLoS Genetics, 2006, 2, e190. https://doi.org/10.1371/journal.pgen.0020190.

Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.

Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; Sham, P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet., 2007, 81(3), 559–575. https://doi.org/10.1086/519795.

Reich, D.; Thangaraj, K., Patterson, N.; Price, A.L.; Singh L. Reconstructing Indian population history. Nature 2009, 461, 489–494. https://doi.org/10.1038/nature08365.

Rosenberg, N. A.; Pritchard, J.K.; Webber, J.L.; Cann, H.M.; Kidd, K.K.; Zhivotovsky, L.A.; Feldman, M.W. Genetic structure of human populations. Science, 2002, 298(5602), 2381–2385. https://doi.org/10.1126/science.1078311.

Rosenberg, N.A.; Mahajan, S.; Ramachandran, S.; Zhao, C.; Pritchard, J.K.; Feldman, M.W. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. PLoS Genetics, 2005, 1, e70. https://doi.org/10.1371/journal.pgen.0010070.

Semino, O.; Magri, C.; Benuzzi, G.; Lin, A.A.; Al-Zahery, N.; Battaglia, V.; Maccioni, L.; Triantaphyllidis, C.; Shen, P.; Oefner, P.J.; Zhivotovsky, L.A.; King, R.; Torroni, A.; Cavalli-Sforza, L.L.; Underhill, P.A.; Santachiara-Benerecetti, A.S. Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area. Am. J. Hum. Genet. 2004, 74, 1023–1034. https://doi.org/10.1086/386295.

Tennessen, J.A.; Bigham, A.W.; O'Connor, T.D.; Fu, W.; Kenny, E.E.; Gravel, S.; et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science, 2012, 337(6090), 64–69. https://doi.org/10.1126/science.1219240.

The 1000 Genomes Project Consortium. A Global Reference for Human Genetic Variation. Nature 2015, 526, 68–74. https://doi.org/10.1038/nature15393.

Tishkoff, S.A.; Reed, F.A.; Friedlaender, F.R.; Ehret, C.; Ranciaro, A.; Froment, A.; Hirbo, J.B.; Awomoyi, A.A.; Bodo, J.M.; Doumbo, O.; Ibrahim, M.; Juma, A.T.; Kotze, M.J.; Lema, G.; Moore, J.H.; Mortensen, H.; Nyambo, T.B.; Omar, S.A.; Powell, K.; Pretorius, G.S.; Smith, M.W.; Thera, M.A.; Wambebe, C.; Weber, J.L.; Williams, S.M. The genetic structure and history of Africans and African Americans. Science, 2009, 324, 5930, 1035–1044. https://doi.org/10.1126/science.1172257.

Tishkoff, S.A.; Williams, S.M. Genetic analysis of African populations: human evolution and complex disease. Nature Reviews Genetics, 2002, 3(8), 611–621. https://doi.org/10.1038/nrg865.

Waskom ML. Seaborn: statistical data visualization. Journal of Open Source SoftwareDa 2021;6(60):3021. https://doi.org/10.21105/joss.03021.

Weir, B.S.; Cockerham, C.C. Estimating F-Statistics for the Analysis of Population Structure. Evolution 1984, 38, 6, 1358–1370. https://doi.org/10.2307/2408641.