The effect of admixture on minor allele frequency distribution and ancestry consistency

Johnson Huang, Yujung Lin

Abstract

Human populations exhibit genetic diversity shaped by historical migrations, isolation, selection, and genetic drift. Admixed populations, formed through interpopulation interactions, retain signatures of ancestral groups within their genomes. This study investigates the effects of admixture on population genetic structures, focusing on linkage disequilibrium (LD), minor allele frequency (MAF) distributions, and ancestry composition. Using HapMap3 data from 11 populations, including five admixed and six non-admixed groups, we analyzed chromosome 1 data through PCA, local and global LD, MAF, ADMIXTURE, and LD decay evaluations. Results reveal that admixed populations exhibit faster LD decay due to increased recombination, intermediate MAF patterns, and a mix of ancestral components. Simulations highlight the influence of simplified models on LD measures, underscoring the importance of realistic assumptions. Comparisons with non-admixed populations, particularly African populations like YRI, demonstrate reduced LD levels and higher genetic variability due to diverse recombination histories. These findings emphasize the distinct genetic architectures of admixed and non-admixed populations, offering valuable insights into genetic diversity, admixture dynamics, and their implications for evolutionary, medical, and forensic research.

Introduction

Problem Formulation

Human populations vary genetically due to historical migrations, isolation, selection, and genetic drift. Admixed populations, formed when previously isolated populations come into contact, carry signatures of their ancestral groups in their genomes. Understanding how admixture affects genetic diversity, particularly the distribution of minor allele frequencies (MAFs), as well as ancestry consistency, is crucial for interpreting genetic data in evolutionary, medical, and forensic contexts.

Motivation and Related Works

Previous studies have shown that admixture can alter patterns of genetic variation and LD, influencing the detection of disease-associated variants and the inference of population structure [1–3]. For instance, Patterson et al. [4] demonstrated how principal components can reveal population substructure, while Alexander et al. [5] developed the ADMIXTURE software to estimate ancestry proportions, which was later employed by Reich et al. [6] to investigate the genetic history of admixed populations in combination with D-statistics to study gene flow patterns. Studies on the HapMap and 1000 Genomes projects have documented how population history shapes allele frequency distributions [1,7]. Our work extends these insights by comparing LD, MAF distributions, and variability metrics between admixed and non-admixed populations, and further evaluates the effects of simulated admixture events.

Overview of Analysis and Data

We utilize HapMap3 data from 11 populations (CEU, CHD, CHB, YRI, JPT, TSI, ASW, MKK, LWK, GIH, MXL), focusing on chromosome 1. Five of these populations are recognized as admixed (ASW, MKK, LWK, GIH, MXL), and six are considered non-admixed (CEU, CHD, CHB, YRI, JPT, TSI). After appropriate file format conversions for PLINK compatibility using Convertf [4], we perform PCA to visualize population structure, examine linkage disequilibrium (LD) patterns both locally and globally, and assess MAF distributions. We also simulate admixed populations by combining genetic data blocks from non-admixed populations to study the patterns of MAF and LD under controlled admixture scenarios. MAF distribution is further analyzed through variability measures (coefficient of variation) and the proportion of rare SNPs.

We also use ADMIXTURE to study the population composition of admixed populations. Here we focus on ASW, CEU, and YRI chromosome 1. We examine the composition at two different levels: Individual-level ancestry proportions and SNP-level ancestry estimates. Lastly, we evaluate LD decay patterns across three populations to quantify how admixture impacts the rate at which LD diminishes with increasing physical distance between SNPs.

Results

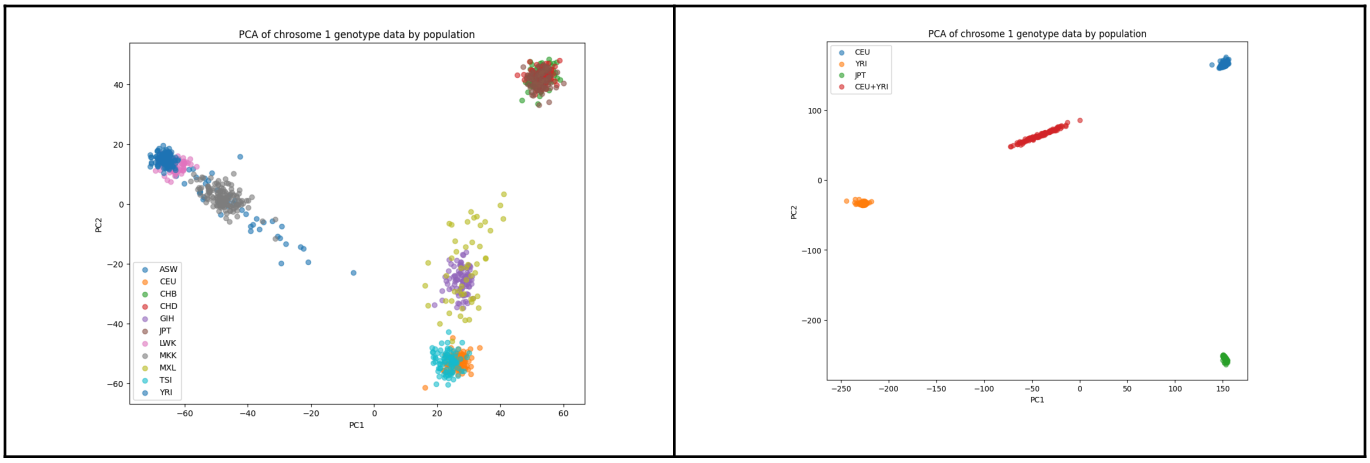


Figure 1. PCA plot of the 11 HapMap3 populations (left) and simulated admixed populations (right)

Our PCA analysis demonstrates clear genetic differentiation among populations. Non-admixed populations form distinct clusters, with East Asian populations (CHB, CHD, JPT) clustering tightly on the right side, reflecting more homogeneous genetic backgrounds. Admixed populations (ASW, MXL, LWK, MKK, GIH) occupy intermediate genetic positions, indicative of their mixed ancestry. The simulated admixed populations show a spread that loosely mimics the intermediate positions of the real admixed groups, though their distribution is more diffuse, likely due to simplifying assumptions in our simulation method.

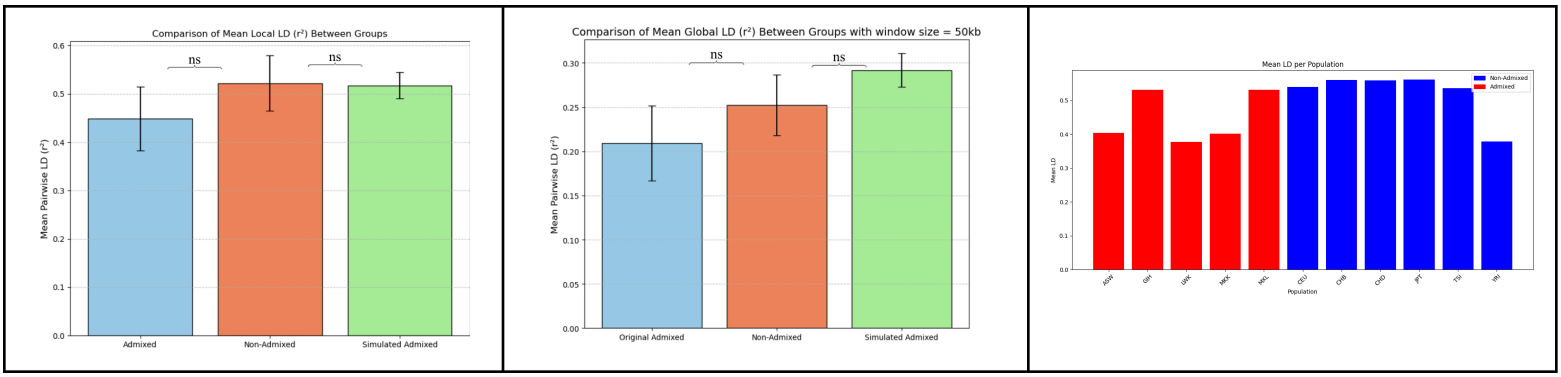


Figure 2. The local LD score (left) and global LD score (middle) of original and simulated populations, and the local LD score of each original population in the HapMap3 project (right)

Figure 2 presents LD analyses across different populations. Both local and global LD scores tend to be slightly higher in non-admixed populations compared to admixed populations, probably because admixture disrupted both short-range and long-range LD and did not create globally consistent haplotype background, although statistical tests (t-tests) do not show significant differences. YRI, a non-admixed African population, is an outlier, showing relatively low local LD, possibly due to extensive historical recombination events and higher genetic diversity. On the other hand, GIH and MXL, classified as admixed, display relatively higher local LD than other admixed populations—this may be due to limited intermarriage over a short timescale, resulting in less breakdown of LD. Simulated admixed populations exhibit inflated LD scores (both local and global), likely because the simulation model does not incorporate the complexity of real-world demographic events and recombination patterns.

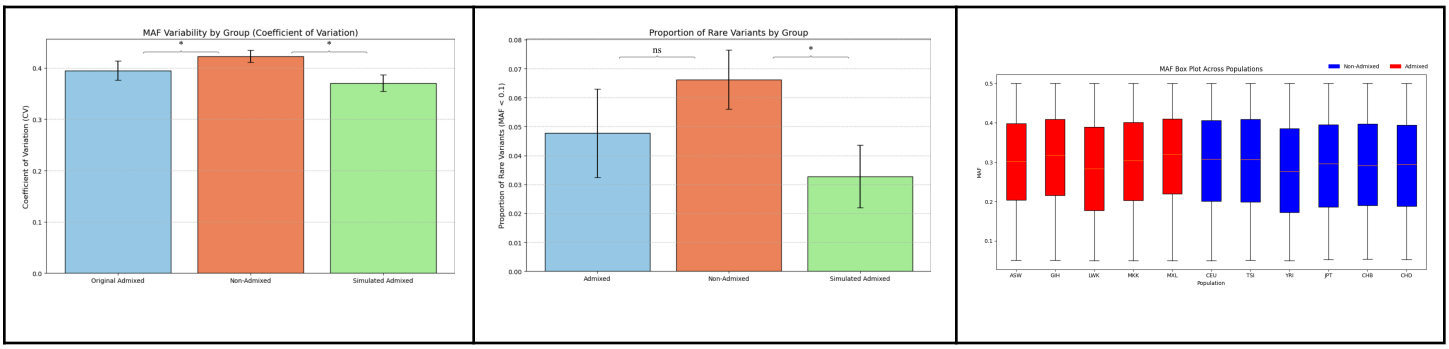


Figure 3. The MAF distribution of each original population in the HapMap3 project (right), and the variability (left), percentage of rare SNPs (middle) of original and simulated populations

In Figure 3, MAF distributions reveal that African populations (such as YRI and LWK) tend to have a higher proportion of low-frequency variants, reflecting their older population history and deep genetic diversity. Non-admixed populations exhibit broader MAF distributions, while some admixed populations show intermediate patterns. The bottom panels quantify variability using the coefficient of variation and assess the proportion of rare SNPs. Overall, non-admixed populations display greater variability and a higher proportion of rare SNPs compared to both

original and simulated admixed populations. This trend supports the notion that complex demographic histories and prolonged isolation have allowed non-admixed groups to accumulate unique low-frequency variants. The lower variability and fewer rare SNPs observed in admixed and simulated admixed populations may stem from the homogenizing effect of mixing and the limited time scale since admixture events.

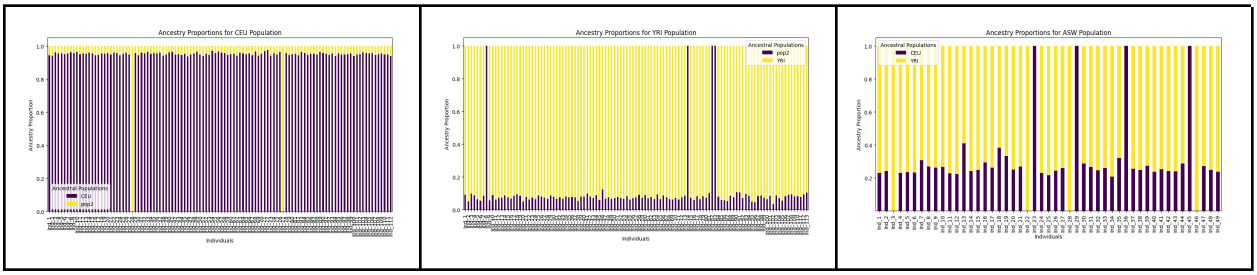
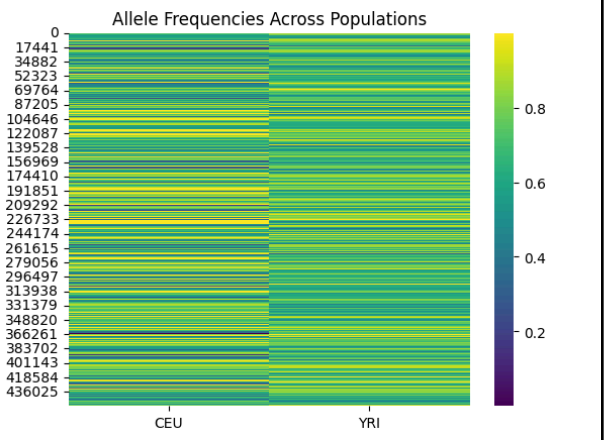


Figure 4. Ancestral population composition at the individual level of CEU (left), YRI (middle), and ASW (right) populations

In Figure 4, the population compositions at individual levels of ASW, CEU, and YRI reveal the clear distinctions between admixed and homogenous populations. In ASW, most individuals exhibit genomes composed of contributions from two ancestral populations, with proportions ranging from 0.15 to 0.85 for each population. Conversely, homogeneous populations, such as CEU and CHB, display ancestry composition dominated by a single population, with minor contributions from other populations likely due to sampling bias.



	CEU	YRI
mean	0.69	0.70
min	1e-5	0.25
max	0.99	0.99
std	0.21	0.14

Figure 5. Ancestral population composition at the SNP level of ASW (left) and the statistic of the composition (right)

In Figure 5, the SNP-level composition of ASW reveals the allele frequencies contributions from CEU and YRI. The summary statistic shows that CEU and YRI contribute similar proportions for most common SNPs, as reflected by their comparable mean allele frequencies (0.69 for CEU and 0.70 for YRI.) However, the range of contributions differs significantly between the two populations. CEU exhibits a broader range of allele frequencies, as seen in its higher standard deviation (0.21 compared to 0.14 for YRI), and a wider range. The heatmap further highlights

this variability, with more pronounced yellow and blue strips in CEU, suggesting greater fluctuation in CEU's contributions across different SNPs.

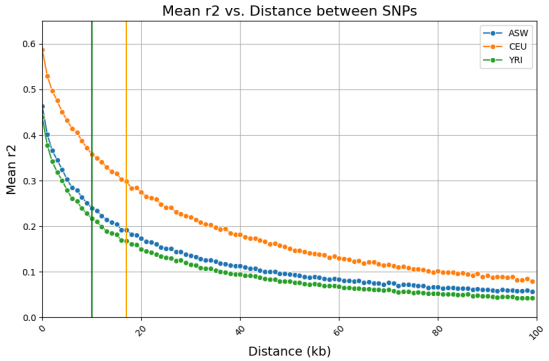


Figure 6. LD decay of ASW, YRI, and CEU populations

In Figure 6, the LD decay plot shows that ASW and YRI have a faster decay rate than CEU. The mean r^2 for ASW and YRi halves at around 10kn, while CEU's half-decay distance is closer to 20kb. This indicates that admixed populations, such as ASW, experience more rapid LD decay due to increased recombination events. Similarly, YRI, as an African population, exhibits fast LD decay attributed to its high genetic diversity and frequent recombination events, resulting in a decay pattern comparable to admixed populations.

Discussions

Our findings emphasize that admixture can markedly influence population genetic structures, MAF distributions, and patterns of LD. Non-admixed populations, shaped by longer and more isolated evolutionary histories, typically exhibit higher variability and a greater proportion of rare SNPs. African populations, exemplified by YRI, have reduced LD levels consistent with a history of numerous recombination events and genetic diversity. Admixed populations, both real and simulated, show intermediate patterns reflecting their mixed ancestry. However, our simulations highlight that simplified admixture models inflate LD measures, underscoring the need for more realistic simulations that account for recombination, migration rates, and selective pressures. These results can inform future studies aiming to understand the genetic basis of complex traits in admixed populations and improve imputation and association mapping methods in diverse genomic datasets.

These findings underscore the distinct genetic architecture of admixed and homogeneous populations. Admixed populations exhibit multiple ancestral components within individual genomes, reflecting historical gene flow and diverse admixture events. In contrast, homogeneous populations predominantly consist of a single ancestry, with minor contributions from others likely due to migration or sampling bias. At the SNP level, CEU shows greater variability in allele frequency contributions compared to YRI, as indicated by a higher standard deviation and broader range. This suggests that SNPs with large allele frequency differences can serve as population markers, aiding in the identification of selection events or gene flow. Together, these results highlight ADMIXTURE's utility in characterizing population structures and genetic diversity shaped by evolutionary processes.

The results also emphasize the faster LD decay in admixed populations due to increased recombination events. The similarity between ASW and YRI reflects the impact of African populations' genetic diversity and recombination history on LD patterns, highlighting the need to account for population-specific LD dynamics in genetic studies.

Limitations and Future Work
Key limitations include using only chromosome 1 and a subset of HapMap3 populations. Future studies should incorporate whole-genome data, additional ancestral populations, and more sophisticated admixture models. Incorporating better mutation and recombination rate parameters, realistic demographic histories, and selective pressure scenarios in simulations could provide more accurate insights. Methods that refine LD and ancestry inference (e.g., haplotype-based models) and integrating advanced machine learning techniques may further improve the understanding of population structure and complex admixture events.

Methods

Data
We utilize HapMap3 data from 11 populations (CEU, CHD, CHB, YRI, JPT, TSI, ASW, MKK, LWK, GIH, MXL), focusing on chromosome 1. Five of these populations are recognized as admixed (ASW, MKK, LWK, GIH, MXL), and six are considered non-admixed (CEU, CHD, CHB, YRI, JPT, TSI).

Simulations
Simulated admixed populations were generated using block-based ancestry segments. We selected genomic blocks of fixed size (e.g., 1,000,000 base pairs) and randomly sampled individuals from two non-admixed parental populations to create new admixed genotypes. While this approach provides controlled admixture scenarios, it does not capture the full complexity of real-world demographic history.

Analysis Methods

PCA: Performed to visualize and quantify population substructure using standard approaches [4, 8].

LD Analysis: We assessed both local (short-range) and global (long-range) LD using custom Python scripts and masked arrays to handle missing data. LD pruning and decay analyses were performed using PLINK and in-house code.

MAF Distribution Analysis: We computed the minor allele frequency for each SNP, comparing distributions across populations. Variability was assessed using the coefficient of variation (standard deviation / mean), and we quantified the proportion of rare SNPs to gauge how population history affects genetic diversity.

ADMIXTURE: The ADMIXTURE software [4, 5] was used to estimate ancestry components at individual and SNP levels.

Statistical Evaluation: Simple t-tests were performed to compare mean LD values between populations. Due to the complexity of data, p-values were used cautiously, acknowledging potential multiple-testing concerns.

References

1. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851–861.
2. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.
3. Reich D, et al. Reconstructing Native American population history. Nature. 2012;488:370–374.
4. Patterson N, et al. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.
5. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–1664.
6. Reich, David, et al. Dissecting the Pre-Columbian Genomic Ancestry of Native Americans along the Andes–Amazonia Divide. Molecular Biology and Evolution, vol. 36, no. 6, 2019, pp. 1254–1269. Oxford Academic, https://doi.org/10.1093/molbev/msz066.Reich DE, et al. Reduced mtDNA diversity in the Hadza of Tanzania. Science. 2003;300(5620):189–193.
7. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–909.
8. Excoffier L, et al. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013;9(10):e1003905.
9. Nielsen R, et al. Tracing the peopling of the world through genomics. Nature. 2017;541:302–310.
10. Rosenberg NA, et al. Genetic structure of human populations. Science. 2002;298(5602):2381–2385.
11. Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655–1664. doi:10.1101/gr.094052.109