A SIMULATION-BASED APPROACH TO UNDERSTANDING GENOMIC VARIANTS IN ADMIXED POPULATIONS USING GRAPH-BASED REFERENCE GENOMES

Anonymous authors
Paper under double-blind review

1 INTRODUCTION

Admixed populations represent a crucial aspect of global genetic diversity, yet their unique genomic architectures remain underrepresented in genomic studies Stankiewicz & Lupski (2010). Traditional genome-wide association studies (GWAS) and variant detection methods often rely on linear reference genomes, which do not fully capture the complexity of genomic variation in admixed individuals Gopalan et al. (2022). This study employs a simulation-based approach to generate synthetic genomic data from two ancestral populations, European (EUR) and South American (AMR), and assess the impact of structural variants (SVs) on disease risk in an admixed population. Furthermore, the study evaluates the effectiveness of pangenome Eizenga et al. (2020) graph-based variant detection compared to traditional linear reference models.

2 METHODS

The study is structured into three primary stages: (1) Human Admixture Simulation, (2) Pangenome Generation, and (3) Read Mapping and Variant Detection.

2.1 HUMAN ADMIXTURE SIMULATION

The study simulates genetic data for 1,000 EUR and 1,000 AMR samples using the chromosomal region chr20:30,000,000-32,000,000. The simulation proceeds over five generations (150 years), after which the minor allele frequency (MAF) of single nucleotide variants (SNVs) is computed. A subset of SNVs unique to AMR (i.e., absent in EUR) is identified, and 20 such SNVs are modified into structural variants:

- 10 SNVs are transformed by adding 1,000 bases of random sequence.
- 10 SNVs are altered by duplicating the 1,000 bases following the SNV site.

2.2 PANGENOME GENERATION

A representative subset of 20 admixed samples is selected to generate a high-quality reference using T2T (telomere-to-telomere) assemblies, following methodologies employed by the Human Pangenome Project. The GRCh38 and T2T-CHM13 references are incorporated to construct a pangenome graph of the chr20 region. This approach enables the evaluation of graph-based reference genomes in variant detection.

2.3 READ MAPPING AND VARIANT DETECTION

Short-read sequencing data (100bp paired-end reads) is simulated for 100 AMR samples and 100 admixed samples to assess variant detection performance. Reads are mapped to the chr20 region using BWA-MEM with GRCh38 as the reference. The following analyses are conducted:

- The percentage of high-quality mapped reads per sample is computed.
- The percentage of correctly identified variants is determined.

- The detection rates of AMR-specific SVs are assessed by comparing the two types of structural variants (random sequence vs. duplicated sequence).

The experiment is repeated using the pangenome reference to compare:

- Read mapping quality.
- Correctly identified variants.
- Detection rates of AMR-specific SVs in both AMR and admixed samples.
- Differences in detection accuracy between the two structural variant types.

2.4 Repeated Analyses

To account for the stochastic nature of graph construction, the pangenome is reconstructed multiple times using smoothing, varying the sample composition:

1. Using a different subset of 20 AMR samples.
2. Different sample sizes (n=10, n=50) were used to assess the impact of dataset size on pangenome quality.
3. Constructing a pangenome with a mix of admixed (n=10) and AMR (n=10) samples to determine whether including ancestral genomes improves variant detection and read mapping.

3 Results and Expected Impact

This study is expected to provide insights into the following:

1. The extent to which traditional linear reference genomes fail to detect population-specific structural variations in admixed individuals.
2. The effectiveness of graph-based reference genomes in improving variant detection, particularly for AMR-specific structural variants.
3. Pangenome construction parameters (sample composition and size) influence mapping accuracy and variant calling performance.
4. By integrating graph-based reference models, potential strategies for improving disease risk variant identification in admixed populations.

References

Jordan M Eizenga, Adam M Novak, Jonas A Sibbesen, Simon Heumos, Ali Ghaffaari, Glenn Hickey, Xian Chang, Josiah D Seaman, Robin Rounthwaite, Jana Ebler, Mikko Rautiainen, Shilpa Garg, Benedict Paten, Tobias Marschall, Jouni Sirén, and Erik Garrison. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.*, 21(1):139–162, 2020.

Shyamalika Gopalan, Samuel Pattillo Smith, Katharine Korunes, Iman Hamid, Sohini Ramachandran, and Amy Goldberg. Human genetic admixture through the lens of population genomics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 377(1852), 2022.

Paweł Stankiewicz and James R Lupski. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.*, 61(1):437–455, 2010.