Project 4- Rachana Pandey

Project Title- Polygenic Risk Score Analysis and Fine-Mapping of Height-associated Variants Across European and African Populations

Introduction

Height is a highly heritable polygenic trait influenced by thousands of genetic variants. The GIANT consortium has conducted large-scale GWASs that seek to identify genetic loci that modulates anthropometric trait including human height across diverse populations. There exist substantial differences between ancestries in terms of genetic architecture, allele frequencies, and linkage disequilibrium patterns. Understanding these differences between the populations remain critical for understanding genetic architecture and transferability of PRS models. This study aims to analyze publicly available GIANT consortium GWAS summary statistics for height to calculate PRS scores, compare them across European (EUR) and African (AFR) ancestry individuals in 1000 genome population, fine-map top loci, and visualize GWAS signals. I will use GWAS summary statistics for standing height from the GIANT consortium (Yengo 2022) and genotypes from the 1000 Genomes Project.

Objectives

General Objective

To investigate the genetic architecture of height differences between European and African ancestry groups

Specific Objectives:

1. To compare and visualize height-related genetic differences between European and African ancestry groups.
2. To generate Manhattan and Quantile-Quantile (QQ) plots to visualize genome-wide association signals and assess potential inflation.
3. To construct the polygenic risk scores (PRS) and compare PRS for height between European and African population.
4. To perform fine-mapping of top loci to identify credible causal variants underlying height variation in European ancestry population.

Hypothesis

Polygenic risk scores for height differ significantly between individuals of European and African ancestry due to underlying differences in allele frequencies, effect sizes, and linkage disequilibrium patterns.

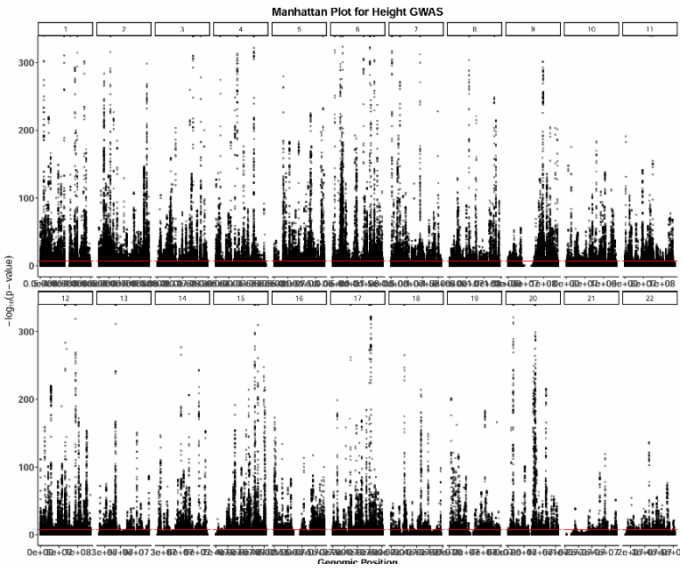
Methods and Analysis

Publicly available GWAS summary statistics for height from the Yengo 2022 (GIANT consortium) and Genotype data from 1000 Genomes Project Phase 3 were used for this study. For the GWAS SNPs visualization, Manhattan plots and QQ plots were generated from the summary statistics to visualize genome-wide association signals and assess p-value inflation from GIANT consortium. Effect size differences for height between European and African population from GIANT consortium was visualized using histogram. Then for the PRS calculation using the European summary statistics from Giant, independent SNPs were selected via clumping using plink2.1, with a p-value threshold of 5×10^{-8} and an LD threshold of $r^2<0.1$. 1000 Genomes Project Phase 3 genotypes were used for PRS computation and LD estimation, focusing on EUR and AFR populations. SNP effect sizes from the GIANT summary statistics were used to calculate PRS for each individual population. The PRS distribution differences were visualized using density plot. Similarly, for fine mapping a top SNP (rs11645785) with largest effect size difference between European and African population was selected, and a 500kb window was fine mapped using the susieR package to compute credible sets of likely causal variants.

Results and Discussion

1. **Genome-wide Visualization of Height-Associated SNPs in GIANT consortium (Yengo 2022)**

To capture the landscape of genetic associations with height, I generated a Manhattan plot using the GIANT consortium summary statistics (Yengo 2022), which includes individuals of diverse ancestry population. The plot reveals multiple genome-wide significant loci ($p < 5e{-}8$). This shows the polygenic nature of height (Figure 1A). Complementary to this plot, the QQ plot shows a strong deviation from the null hypothesis, indicating that a substantial number of SNPs are associated with height (Figure 1B). The overall curvature or the massive inflation at the top right tail suggest that there are not just few SNPs but large number of genome wide SNPs that have extremely small p values.



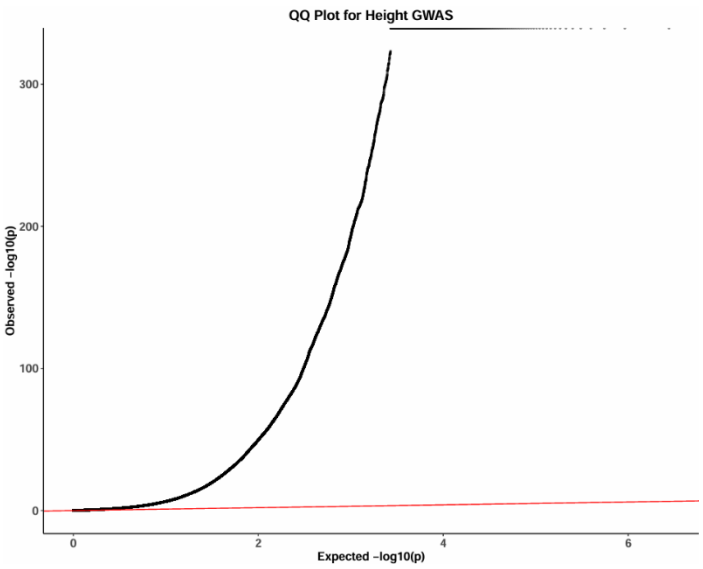


Figure 1- Manhattan plot (A) and quantile–quantile plot (B) to visualize GWAS results across all populations from Giant Consortium (Yengo 2022)

2. Height Effect Size Differences Between European and African Ancestry population from GIANT consortium

To quantify ancestry-specific genetic differences, I calculated the difference in SNP effect sizes between European and African populations. The distribution plot shows that majority of SNPs have small differences centered around zero (Figure 2A). To narrow it down, I ranked SNPs by the absolute difference in effect sizes between EUR and AFR individuals and visualized the top 10 SNPs with largest difference in effect size between European and African population (Figure 2B). Positive values indicate that the effect size for the SNP is greater in Europeans than in Africans. Negative values indicate that the effect size is greater in Africans than in Europeans.

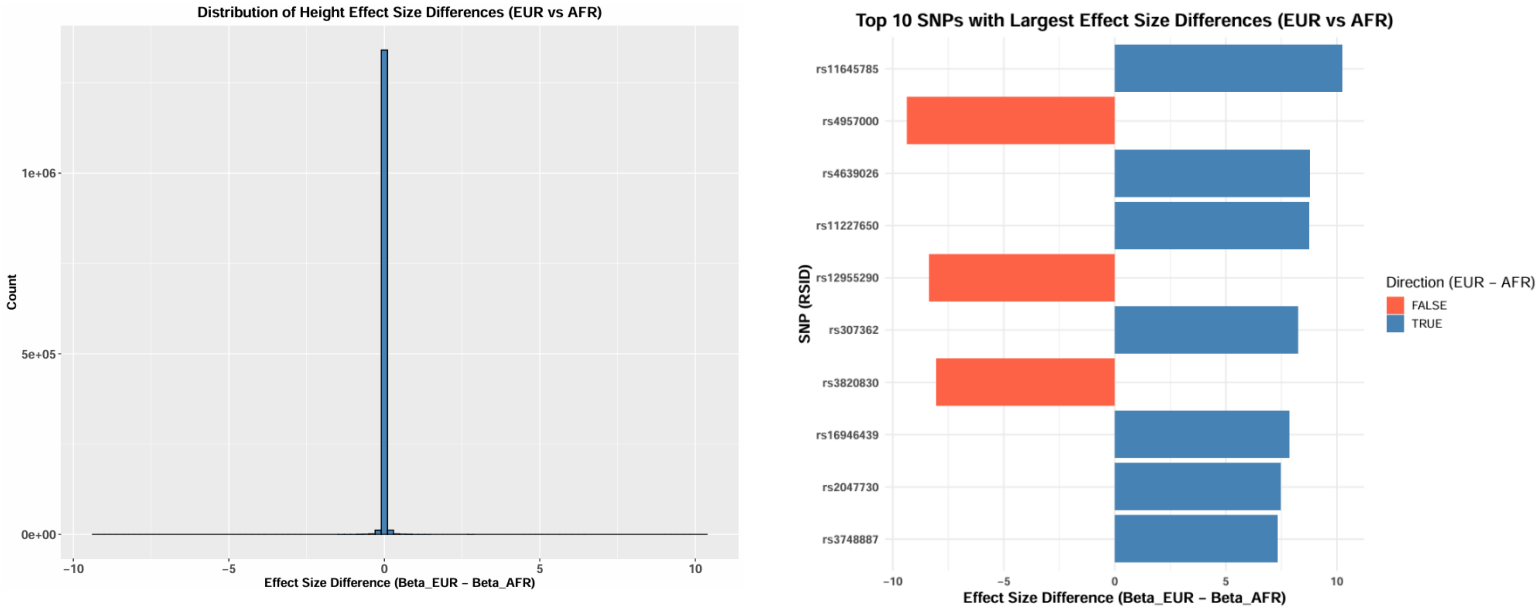


Figure 2- A- Distribution of height effect size difference between European and African population from Yengo 2022. B- Visualization of top 10 SNPs with the largest effect size difference between European and African population.

3. Polygenic Risk Score Construction and Transferability

Next, I built a PRS model based on European effect sizes from Yengo et al. (2022) and applied it to individuals from the 1000 Genomes Project (Figure 3). The key observation I noticed is that the **EUR population's PRS distribution is centered to the right** compared to the AFR

population. PRS scores for height in Europeans were significantly higher than for Africans. This high PRS score in Europeans were expected since the PRS model was trained on European GWAS summary statistics. However, the low PRS scores in African population does not mean that the African population has low genetic association with the height trait. Instead, it reflects the reduced transferability of PRS across different populations due to the difference in linkage disequilibrium, allele frequency, and the effect size across these populations.

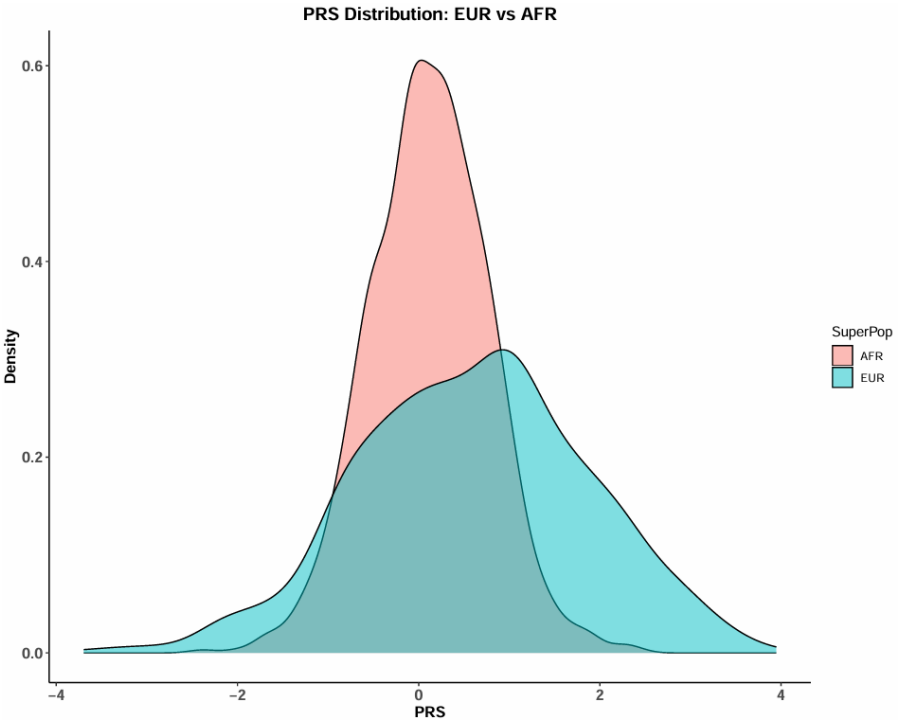


Figure 3- Polygenic risk score distribution between European and African population from 1000 genome populations.

4. Fine-Mapping of SNP with largest effect size difference between European and African population(rs11645785)

Next, I performed regional fine mapping for the SNP which showed largest difference in the effect size between European and African population from the GIANT consortium (Yengo 2022). I retrieved its genomic location from GTEx portal. The SNP rs11645785 had the largest effect size difference of 10.8. To identify the potential causal variants, I fine mapped a ±500 kb window around rs11645785 using susieR. The figure identified a broad credible set of casual variants associated with this SNP (Figure 4). The reason for this large number of casual variants could be because the European population has the larger linkage disequilibrium block. Because the variants within the LD block are highly correlated, the larger LD block makes it difficult to pinpoint the exact casual variant, providing us with a set of potential causal variants at a specific locus.

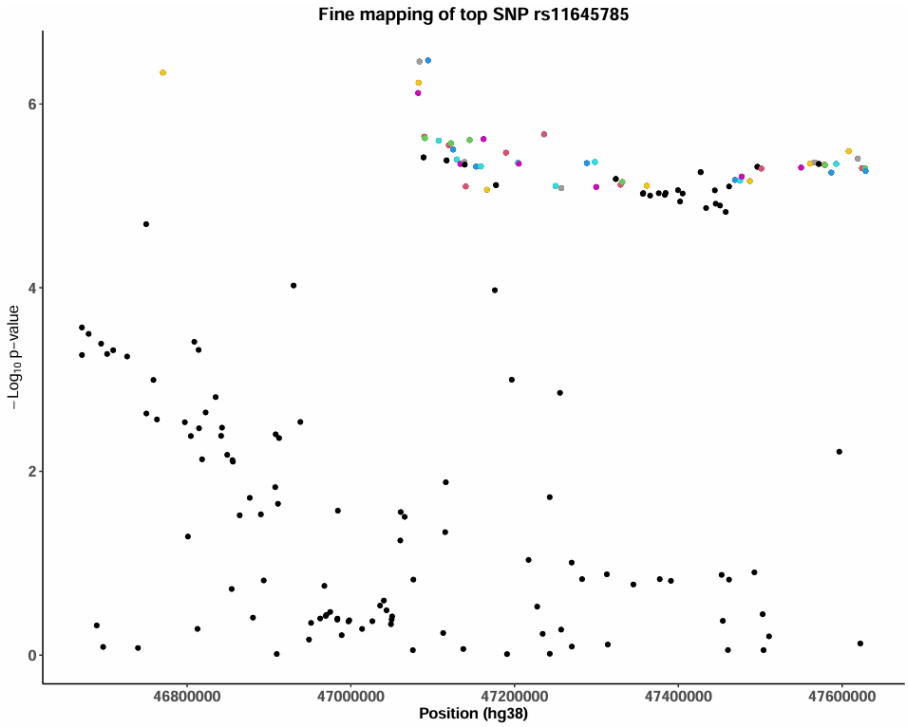


Figure 4- Statistical Fine mapping of rs11645785 Reveals Multiple Candidate Variants in ±500 kb Region in European population form Yengo 2022.

Conclusion

In this study I examined the ancestry-specific architecture of height by integrating GWAS summary statistics and 1000 Genomes genotypes. This study illustrates significant differences in height-associated genetic architectures between European and African ancestry groups. Our initial GWAS visualization confirmed robust genome wide association signals of height in Yengo 2022. Effect size comparisons revealed several SNPs with large ancestry-specific effects. PRS distributions differed substantially between European and African individuals, demonstrating reduced PRS transferability across populations. Similarly, Fine mapping of a top SNP (rs11645785) narrowed the association signal to a credible set of candidate causal variants offering insight on potential causal loci.

Limitations and Future Directions

In this report, we only investigated the genetic architecture for height for the two-ancestry population: European and African. However, this study can be extended to incorporate the other ancestries populations like south Asians and east Asians. Second is we can functionally annotate the fine mapped SNPs and leverage functional genomics datasets to validate identified causal variants. We can also develop ancestry-specific PRS models to improve predictive performance.