Population-Level Associations Between Polygenic Scores and Complex Traits Across Global Populations

1. Background

Polygenic scores (PGS), as quantitative measures of individual genetic risk, have been widely applied in studies of complex diseases at the individual level. Numerous studies have demonstrated that PGS can effectively predict an individual's susceptibility to various complex diseases, offering new perspectives for early disease warning and personalized prevention, and thereby advancing the development of precision medicine.

However, most existing research has focused on disease risk prediction at the individual level, with relatively limited attention to the performance of PGS at the population level. It remains unclear whether differences in average PGS among populations can accurately reflect corresponding differences in observed phenotypes or disease prevalence. Investigating the interpretability of PGS at the population level is crucial for understanding how genetic factors contribute to disease etiology across populations, as well as for assessing the generalizability of current PGS models across diverse ancestry groups.

This study aims to systematically examine the relationship between population-level average PGS and observed phenotypes or disease prevalence, and to evaluate the consistency of such associations across different sexes and traits. The results will provide empirical evidence for the applicability and interpretation of PGS in cross-population contexts.

2. Methods & Material

2.1 Base Data: Within-Family GWAS Summary Statistics

This study used the within-family genome-wide association study (GWAS) summary statistics published by Howe et al. (2022) as the base data for polygenic score (PGS) construction. The dataset includes 24 complex traits and provides effect size estimates (β) that are adjusted for family structure, thereby minimizing potential confounding from shared environmental factors and population stratification.

To ensure data consistency and reliability, a series of quality control (QC) procedures on the original summary statistics were performed. The analysis was restricted to autosomal

chromosomes (1–22), and only traits with SNP-based heritability (h^2_{snp}) greater than 0.05 were retained, resulting in 15 traits in total. Duplicate SNP records were identified and removed based on rsID or genomic position to avoid double-counting of variants during PGS calculation. Additionally, ambiguous SNPs with A/T or C/G alleles were excluded because their strand orientation cannot be reliably determined and may introduce allele-flip errors during base–target alignment.

Finally, cleaned GWAS summary statistics for 15 traits were obtained, each containing high-quality SNPs, allele annotations, and corresponding effect size estimates (β). These cleaned datasets served as the base input for subsequent polygenic score calculations.

2.2 Target Data: 1000 Genomes Project (2013 Release)

This study used whole-genome sequencing data from the 1000 Genomes Project Phase 3 (20130502 release) as the target data. The dataset comprises 2,504 individual samples from 26 populations, which can be grouped into five super-populations: African (AFR), European (EUR), East Asian (EAS), South Asian (SAS), and Admixed American (AMR) (Table 1).

To ensure consistency with the base data, a one-to-one mapping between the base and target datasets was established using genomic coordinates and allele information (CHR:BP:A1:A2). This procedure was used to supplement missing rsIDs and harmonize variant identifiers across datasets. In addition, sample sex information was incorporated based on the official panel file (sex annotation), enabling sex-stratified PGS calculation and subsequent comparisons between male and female groups.

Table 1: Mapping of 1000 Genomes Project Population Codes to Super-populations and Countries/Regions.

Population Code	Super-Population	Country/Region
ACB	AFR	Barbados
ASW	AFR	United States of America
BEB	SAS	Bangladesh
CDX	EAS	China
CEU	EUR	CEPH
CHB	EAS	China
CHS	EAS	China

CLM	AMR	Colombia
ESN	AFR	Nigeria
FIN	EUR	Finland
GBR	EUR	United Kingdom
GIH	SAS	India
GWD	AFR	Gambia
IBS	EUR	Spain
ITU	SAS	India
JPT	EAS	Japan
KHV	EAS	Viet Nam
LWK	AFR	Kenya
MSL	AFR	Sierra Leone
MXL	AMR	Mexico
PEL	AMR	Peru
PJL	SAS	Pakistan
PUR	AMR	Puerto Rico
STU	SAS	Sri Lanka
TSI	EUR	Italy

2.3 Polygenic Score Calculation

PLINK v1.9 was used to compute individual-level PGS. For each individual i, the PGS was defined as the weighted sum of effect sizes across all included variants:

$$PGS_i = \sum_{j=1}^{M} \beta_j \times \mathrm{G_{ij}}$$

Where M is the total number of SNPs used in the calculation, β_j represents the estimated effect size of the j-th SNP from the base GWAS data, and G_{ij} denotes the genotype dosage of the effect allele for individual i.

PGS was calculated separately for each of the five super-populations, and individual-level PGS values were output for each group. Subsequently, the mean and distribution of PGS

within each population were summarized to enable downstream comparisons and correlation analyses with corresponding phenotypic data.

2.4 Phenotype Data Collection

Five quantitative traits were selected for analysis in this study: height, body mass index (BMI), systolic blood pressure (SBP), low-density lipoprotein cholesterol (LDL), and high-density lipoprotein cholesterol (HDL). Phenotypic data were obtained from the NCD-RisC Global Database, using country-level estimates from 2013 to maintain temporal consistency with the 1000 Genomes Project (2013 release).

The data were extracted separately for males and females to enable sex-stratified analyses. For height, the mean values of individuals aged 18–19 years were used to minimize environmental variation associated with growth or aging. For the other traits (BMI, SBP, LDL, HDL), age-standardized national means provided by NCD-RisC were used.

Then country-level phenotypic means were calculated and matched with the corresponding population-level mean PGS to assess the relationship between genetic predictions and observed phenotypic variation across populations.

2.5 Allele Frequency Distribution Analysis

To assess genetic differences among populations and explore potential sources of variation in population-level PGS, we analyzed the allele frequency distributions of the five super-populations.

Allele frequencies were calculated for each population using PLINK v1.9, generating effect allele frequencies (EAF) for all SNPs. The EAF distributions across super-populations were then visualized and compared to evaluate differences in allele frequency spectra among groups.

2.6 Statistical Modeling: Association between PGS and Continuous Traits

To evaluate the population-level association between PGS and quantitative traits, we performed linear regression analyses using the following model:

$$Y = \beta_0 + \beta_1 \times PGS_z$$

where Y the standardized quantitative phenotype, and PGSz denotes the standardized (z-scored) mean polygenic score for each population. The regression coefficient β_1 reflects the direction and strength of the linear association between PGS and the trait of interest.

Analyses were conducted separately for five quantitative traits. To account for potential sex differences, sex-stratified models were fitted for males and females, respectively.

Modeling was performed at two hierarchical levels:

(1) **Super-population level analysis:**

All populations were combined to assess the overall relationship between PGS and phenotype across global populations. In this analysis, PGS values were standardized across all populations (global z-score).

(2) **Within super-population analysis:**

Independent regressions were performed within each super-population to examine the consistency and stability of PGS–trait associations across different ancestral groups. For these analyses, PGS values were standardized within each super-population (within-group z-score) to remove differences in scale.

Results were visualized using scatter plots with fitted regression lines, and the correlation coefficient (r), regression coefficient (β_1), and significance level (p-value) were reported to quantify the extent to which PGS explains population-level variation in the phenotypes.

3. Result

3.1 Allele Frequency Distribution Analysis **(Note: the SNPs used need to align to the base GWAS data, not the whole SNPs used in this figure)**

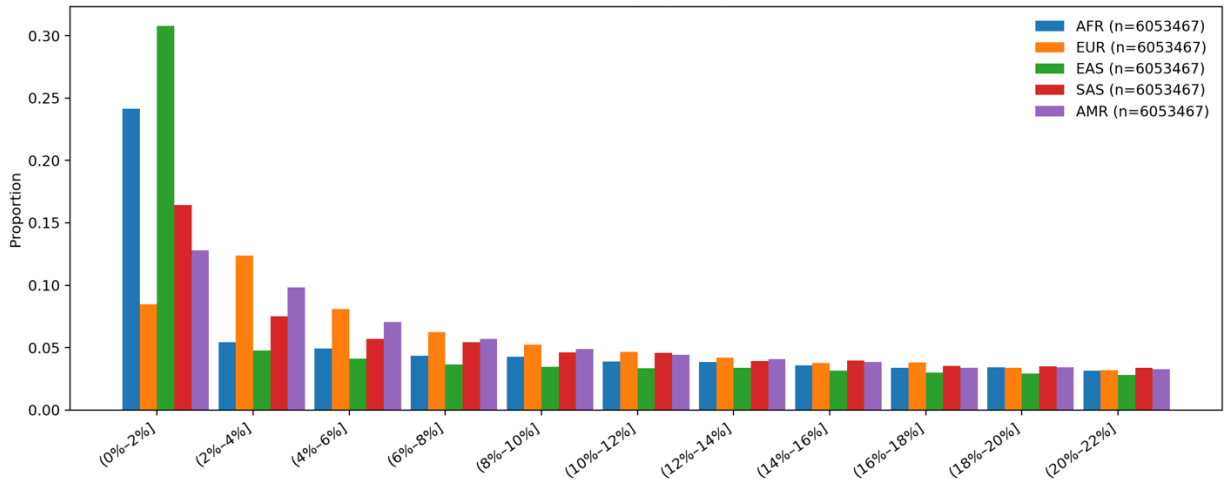
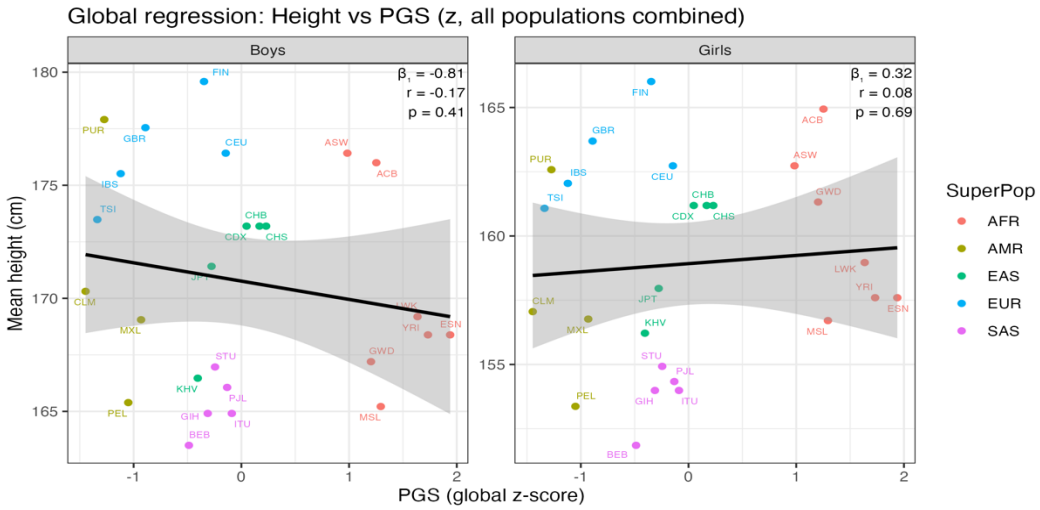
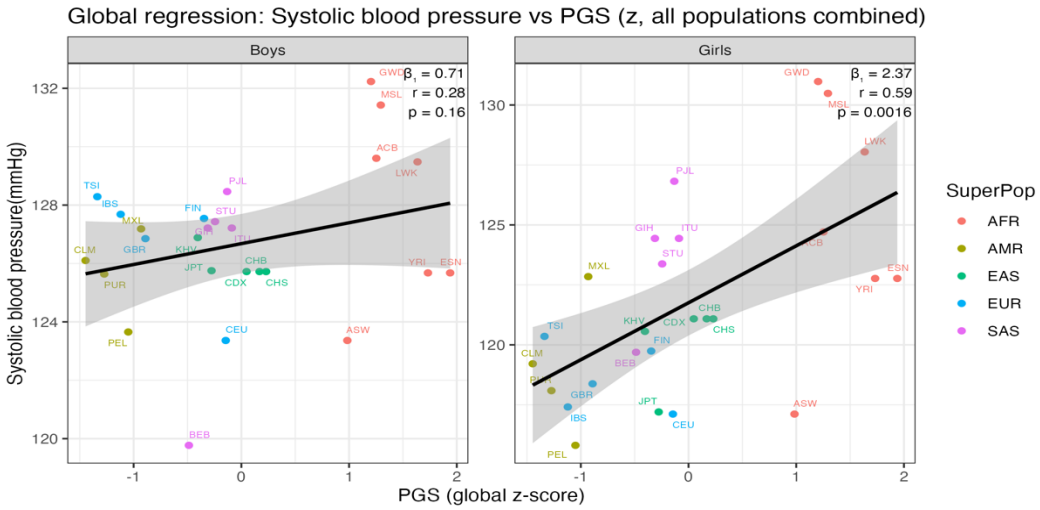
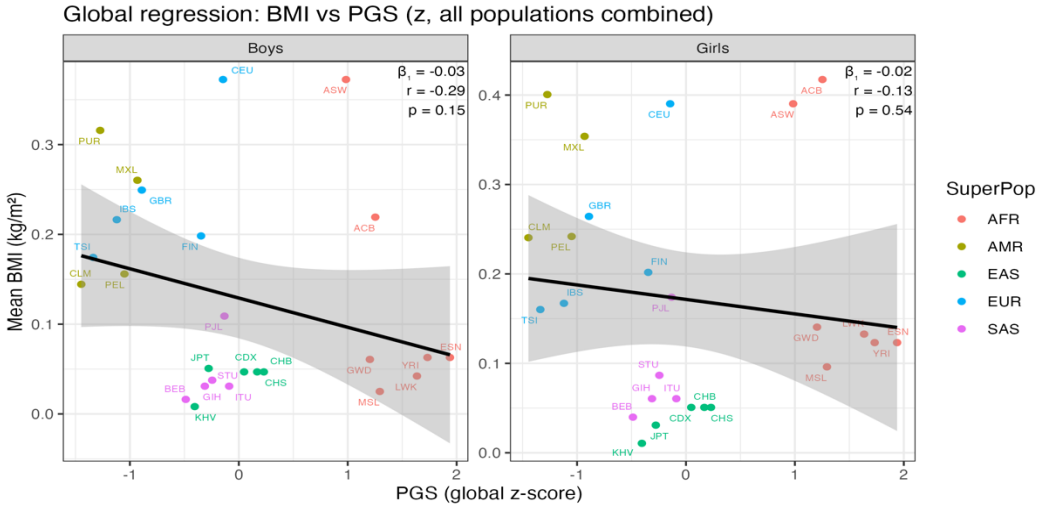


Figure 1. Distribution of Effect Allele Frequencies across Super-Populations for Height

3.2 Association between Polygenic Scores and Phenotypes across Super-Populations







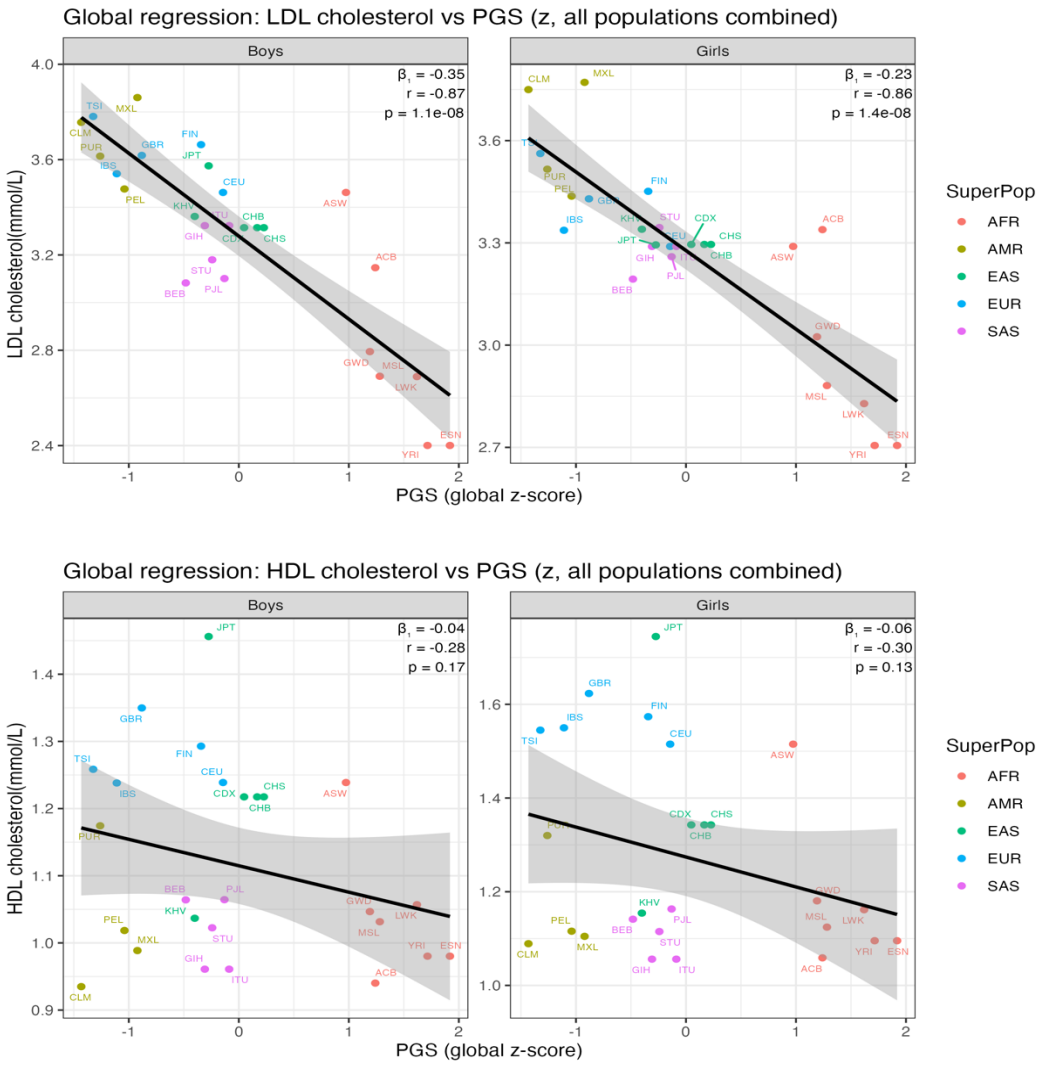


Figure2 Association between Polygenic Scores and Phenotypes across Super-Populations

From the figure, a significant positive correlation was observed between PGS and systolic blood pressure (SBP) among females ($r = 0.59$, $p = 0.0016$).

The relationship between LDL cholesterol with PGS was significantly negative in both sexes (males: $r = -0.87$, $p = 1.1 \times 10^{-8}$; females: $r = -0.86$, $p = 1.4 \times 10^{-8}$).

No notable associations were observed for the other traits.

3.3 Association between Polygenic Scores and Phenotypes across Super-Populations

The relationships between population-level polygenic scores (PGS) and quantitative phenotypes across five super-populations (AFR, AMR, EAS, EUR, and SAS) are shown in Figure 3 (see Attachment 1: Association between Polygenic Scores and Phenotypes across Super-Populations).

Across super-populations, several significant associations were identified. Among East Asian (EAS) populations, height and BMI were both negatively correlated with mean polygenic scores in females (height: $r = -0.97$, $p = 0.007$; BMI: $r = -0.94$, $p = 0.0016$).

For LDL cholesterol, PGS was negatively associated with measured LDL levels in both sexes (males: $r = -0.87$, $p = 0.011$; females: $r = -0.85$, $p = 0.016$).

No significant correlations were detected for the other traits across super-populations.

4. Discussion

Across all analyzed traits, we found that only a few traits exhibited significant correlations with PGS, and in some cases, the correlations were even negative.
At the same time, a consistent population-level pattern was observed: populations within each super-population (e.g., YRI, LWK, and GWD within AFR) clustered tightly together, whereas distinct super-populations were clearly separated. Specifically, AFR populations were consistently positioned toward the right end of the PGS axis, EUR and AMR populations toward the left, and EAS and SAS populations fell in between.

This stratified pattern can likely be explained by differences in EAF distributions across super-populations. As shown in Figure 1, the five super-populations exhibit systematic variation in their allele frequency structures:

- AFR populations show a higher proportion of alleles in the intermediate-to-high frequency range (2–10%), reflecting greater genetic diversity;
- EAS populations exhibit the highest proportion of low-frequency variants (0–2%), indicating that many effect alleles are rare in East Asian populations;
- EUR and AMR populations display the most similar frequency spectra, consistent with the substantial European ancestry component within AMR populations.

Because PGS are fundamentally the weighted sum of effect allele dosages, cross-population differences in allele frequency inevitably lead to systematic shifts in the overall PGS distributions. Even though the effect estimates (β) are derived from within-family GWAS—which theoretically eliminate confounding from population structure—the underlying allele frequency differences still cause mean PGS values to shift across populations.This population-level “PGS displacement” does not reflect true differences in genetic risk, but rather arises as a natural consequence of variation in frequency spectra among populations. Consequently, at the

global level, correlations between PGS and phenotypic means are often attenuated, as the true biological signal is masked by such systematic population-level shifts.

Although the PGS used in this study are derived from within-family GWAS β, which substantially mitigate confounding factors inherent to traditional GWAS—such as population stratification and shared family environment—the genetic component of the PGS can be considered relatively "clean."

However, the phenotypic data used for validation are based on country-level averages rather than individual-level measurements. Such aggregate data reflect not only genetic differences but also the influence of a wide range of environmental and sociocultural factors, including diet, nutrition, socioeconomic development, healthcare access, and lifestyle.

Consequently, even if a population exhibits a genetically higher predisposition for a given trait (e.g., a higher BMI tendency, indicated by higher PGS values), its observed population-level mean phenotype (e.g., mean BMI) may still be lower due to environmental counteracting effects—such as undernutrition or differing lifestyle patterns. In other words, the observed phenomenon of "high PGS but low phenotype" does not imply that the genetic effects act in the opposite direction; rather, it likely reflects the opposing modulation of environmental influences on the population mean phenotype.

What we were talking about last meeting:

The population-level ordering we observe is likely driven primarily by shared allele frequency structure rather than reflecting true differences in genetic predisposition to the trait. A reasonable hypothesis is that, if we extended this analysis to more traits, we would likely find that positive and negative correlations occur at roughly similar frequencies — suggesting that the direction of the correlation mainly reflects systematic patterns of allele frequency structure, rather than genuine polygenic signals for the trait.

5. Future Directions

This study has identified two major sources of influence in cross-population analyses of PGS:(1) systematic PGS displacement arising from differences in allele frequency structures among populations, and (2) substantial environmental and socioeconomic confounding in country-level phenotypic data.

Future work will focus on refining and expanding the analysis along these two dimensions to improve the accuracy and interpretability of cross-population comparisons.

5.1 Correction and Quantification of Population Frequency Differences

The observed variation in PGS distributions across super-populations largely reflects systematic differences in EAF structures rather than true differences in genetic risk. To address this issue, the next steps will include:

- Quantifying the contribution of allele frequency structure to PGS variation, for instance through simulation studies or variance decomposition analyses, to estimate how much of the between-population variance in PGS is attributable to frequency differences;
- Exploring and comparing different correction strategies, such as population-specific scaling, principal component (PCA) adjustment, or covariance-matrix–based normalization, to mitigate systematic displacement caused by frequency structure;
- Evaluating the performance of these correction methods, determining which approach best preserves the underlying genetic signal while minimizing frequency-driven bias.

5.2 Incorporating Environmental Covariates into Multivariate Models

Because the phenotypic data in this study are derived from country-level averages—strongly influenced by diet, nutrition, socioeconomic development, healthcare access, and lifestyle—it is essential to account for these environmental and social factors in future analyses. Specifically, I plan to:

- Construct multivariate regression models that include key environmental covariates to re-evaluate the relationship between PGS and phenotypic means under environmental control;
- Compare univariate and multivariate models to assess whether environmental adjustment enhances the robustness and explanatory power of genetic associations;

5.3 Extension to Disease Traits and Construction of Within-Family GWAS

While the present analysis focuses on quantitative traits such as height, BMI, SBP, LDL, and HDL, the ultimate goal is to understand how polygenic risk for complex diseases varies across populations. To move toward this goal, I plan to:

- Apply for or obtain existing within-family GWAS summary statistics for disease traits, to establish suitable base data for disease-focused analyses;

- Leverage the UK Biobank by utilizing family-relatedness information and disease phenotypes to construct within-family GWAS for selected diseases;
- Integrate disease-specific PGS with global disease prevalence datasets to systematically evaluate the concordance and divergence between genetic risk predictions and observed disease burdens, thereby gaining insight into how genetic and environmental factors jointly shape population-level disease risk.

6. Rotation Experience and Reflection

During this rotation, I gained an initial understanding of the fundamental concepts and theoretical framework of genetics. I also had hands-on experience working with GWAS data and the process of calculating and analyzing PGS. At the same time, I realized that I still tend to think like a "student" — waiting for the professor to make decisions instead of taking initiative to propose and evaluate my own ideas.

I also noticed major gaps in my knowledge. Without a solid background in genetics coursework, I sometimes struggle to identify key concepts or build a clear logical framework. While I can carry out the statistical and coding steps, my understanding of the underlying principles is still shallow. To address this, Instead of the biostatistics class, I use my free time to learn Linux-based data processing and programming, hoping to strengthen my skills in statistical modeling and genetic analysis in the future.

One of the most important lessons I learned is that knowing why you're doing something matters far more than simply doing something. At first, I was eager to get quick results, but I soon realized that just following tutorials doesn't really answer my own research questions—it only leads to detours and wasted time. Now I remind myself constantly to think about the purpose behind each step: What question am I trying to answer? What does each result mean? The process of thinking and understanding is much more valuable than the final outcome.

I'm very grateful to Professor Harpak for his thoughtful mentorship. Every one-on-one meeting was meaningful—he always pinpointed the core of the problem and guided me to think about the "why," instead of just giving me the "how." I'm also thankful to all the lab members for being so welcoming and supportive. The Monday ice-cream socials, the lively lunch conversations, and the fun discussions on Slack and in meetings made me feel that this is not only an intellectually stimulating lab, but also a warm, creative, and truly wonderful community.