

Supplementary Methods

S1 Reproducible workflow for SNP heritability estimation

To ensure reproducibility and cross-study comparability of SNP-based heritability estimation, we established a standardized analytical workflow comprising five key components: data quality control, genomic relationship matrix construction, heritability estimation, model diagnostics, and statistical interpretation. This workflow is applicable to both individual-level and summary-statistics-based analyses and explicitly accounts for differences in statistical estimands across methods.

S1.1 Quality control

In this study, to minimize systematic bias as much as possible and enhance the robustness of genetic parameter estimation, comprehensive quality control procedures were first applied to the raw genotype and phenotype data. For genotype data, filtering was conducted primarily from two aspects: the reliability and representativeness of variant sites. On the one hand, a minimum allele frequency (minor allele frequency, MAF) threshold ($MAF > 0.01$) was applied to remove variants with extremely low frequency in the population, thereby avoiding unstable estimates introduced by rare alleles. On the other hand, SNP missingness was controlled (typically limited to within 5%) to reduce the impact of missing data on analytical results. In addition, Hardy-Weinberg equilibrium tests were performed in unrelated individuals to identify potential genotyping errors or sequencing biases from the perspective of population genetic structure, further improving data accuracy and consistency. Through these multiple filtering steps, the interference of low-quality markers in subsequent analyses can be effectively eliminated.

At the individual level, quality control mainly focused on sample completeness and consistency. Specifically, individuals with significantly high missingness rates were excluded to prevent systematic distortion of the overall data structure. Meanwhile, by evaluating the distribution of individual heterozygosity, samples that deviated markedly from the population mean were identified and removed, as such outliers often indicate potential sequencing errors or contamination risks. In addition, consistency between genetically inferred sex and recorded sex was verified, and samples with clear mismatches were excluded. Where necessary, individuals with high levels of relatedness were further identified and removed to ensure that samples satisfy the basic assumption of independence required in statistical analyses, thereby improving the validity of model estimation.

Considering the potential impact of population structure on genetic effect estimation, principal component analysis (principal component analysis, PCA) was further introduced to identify and correct for population stratification. By applying dimensionality reduction to the genotype matrix, principal components reflecting genetic variation within the population were extracted, and the top 10 to 20 principal components were included as covariates in subsequent statistical models. This approach allows explicit control of underlying population structure during analysis, effectively reducing confounding effects caused by stratification and preventing systematic bias in heritability estimation and association results. Overall, these quality control and structural correction procedures provide a reliable data foundation for subsequent genetic analyses.

S1.2 Genomic relationship matrix construction

After completing rigorous data quality control, a genomic relationship matrix (GRM) was constructed based on the filtered high-quality SNP set to characterize the genetic similarity structure among individuals. Specifically, the standard GRM is obtained by centering and standardizing the genotype at each locus, and then calculating the weighted average of the genome-wide genetic covariance between individual i and individual j , yielding the following form of estimation:

$$G_{ij} = \frac{1}{M} \sum_{k=1}^M \frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k(1-p_k)}$$

Here, x_{ik} denotes the genotype coding of individual i at locus k , p_k represents the allele frequency at that locus, and M is the total number of SNPs included in the analysis. By standardizing genotypes with respect to allele