

Research Article

Open Access

SNP-Based Heritability Is Not a Parameter but a Model-Defined Estimand: Evidence from UK Biobank

Xuanjun Fang 

Hainan Provincial Key Laboratory of Crop Molecular Breeding, Hainan Institute of Tropical Agricultural Resources (HITAR), Sanya, 572025

 Corresponding author: xuanjunfang@hitar.orgBioscience Methods, 2026, Vol.17, No.3 doi: [10.5376/bm.2026.17.0013](https://doi.org/10.5376/bm.2026.17.0013)

Received: 06 Apr., 2026

Accepted: 07 May, 2026

Published: 18 May, 2026

Copyright © 2026 Fang, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preferred citation for this article:

Fang X.J., 2026, SNP-based heritability is not a parameter but a model-defined estimand: evidence from UK biobank, Bioscience Methods, 17(3): 153-168 (doi: [10.5376/bm.2026.17.0013](https://doi.org/10.5376/bm.2026.17.0013))

Abstract SNP-based heritability is widely interpreted as a fundamental property of complex traits, yet estimates vary substantially across methods. Here we show that this variation arises because different approaches do not estimate the same quantity: SNP-based heritability is a model-defined estimand rather than a single biological parameter. Using UK Biobank height data as a representative case, we systematically compare estimates from individual-level methods (GCTA-GREML and related estimators) and summary-statistics-based approaches (LD Score Regression and SumHer). We find that GREML-based methods consistently yield higher estimates (~0.60-0.69), LDSC produces systematically lower values (~0.56), and SumHer yields intermediate or higher estimates (~0.63). These differences persist under matched samples and SNP sets, indicating that they cannot be attributed to sampling variation alone. We demonstrate that the discrepancies arise from differences in data representation, model assumptions, and the treatment of linkage disequilibrium (LD) and allele frequency. Accordingly, each method targets a distinct estimand: GREML captures variance explained through genomic relationships, LDSC estimates LD-weighted marginal effects, and SumHer models MAF- and LD-dependent architectures. This framework resolves apparent inconsistencies in SNP heritability estimates and clarifies that cross-method comparisons are generally not statistically valid without alignment of underlying assumptions. More broadly, our results redefine SNP-based heritability as a model-dependent functional determined by SNP coverage, LD structure, and estimation framework. These findings provide a principled basis for interpreting heritability estimates and have implications for genetic studies ranging from biobank-scale analyses to genomic prediction.

Keywords SNP heritability; Estimand; Estimand mismatch; GCTA-GREML; LD Score Regression (LDSC); UK Biobank; Linkage disequilibrium; Genetic architecture

1 Introduction

Heritability is a central parameter in quantitative genetics, used to quantify the contribution of genetic factors to phenotypic variation. In classical frameworks, heritability is typically estimated using pedigree or twin-based designs, where genetic variance is inferred from known relatedness structures. However, with the advent of genome-wide association studies (GWAS) and high-throughput genotyping technologies, the paradigm of heritability estimation has undergone a fundamental shift from pedigree-based inference to SNP-based heritability derived from molecular markers (SNP-based heritability, h_{SNP}^2) (Yang et al., 2010).

The evolution of statistical genetic methods—from linkage analysis and candidate gene approaches to GWAS—has fundamentally reshaped how genetic variation is quantified and interpreted (Fang and Wu, 2026). Within this paradigm shift, SNP-based heritability estimation, particularly under the GCTA-GREML framework, represents a transition from pedigree-based inference to genotype-driven variance decomposition (Fang, 2026).

SNP-based heritability is typically defined as the proportion of phenotypic variance explained by observed or imputed SNP markers across the genome. Its estimation is commonly based on linear mixed models (LMMs) or their extensions. Among these, the GCTA-GREML framework estimates genetic variance components using individual-level genotype data by constructing a genomic relationship matrix (GRM), and is widely regarded as approximately unbiased and statistically efficient under appropriate model assumptions (Yang et al., 2016). In contrast, LD Score Regression (LDSC) and its extensions (e.g., S-LDSC) estimate heritability using GWAS summary statistics, enabling large-scale analyses when individual-level data are unavailable (Bulik-Sullivan et al.,

2015; Ni et al., 2018). Furthermore, the SumHer method, based on the LDAK framework, allows SNP effects to depend on minor allele frequency (MAF) and linkage disequilibrium (LD) structure, thereby introducing greater flexibility in modeling genetic architecture (Speed and Balding, 2019).

Despite their shared objective of estimating SNP-based heritability, these methods often yield substantially different results in practice. The UK Biobank (UKB), one of the largest biomedical resources available (Bycroft et al., 2018), provides an ideal setting for systematic comparison. For height—a highly heritable and polygenic trait—typical estimates in UKB European populations show a consistent pattern: GREML-based approaches yield estimates around 0.60~0.69, LDSC produces slightly lower estimates (~0.55-0.60), and SumHer often yields intermediate or slightly higher estimates (~0.63) (Ge et al., 2017; Hou et al., 2019; Speed et al., 2020). Further analyses indicate that, under matched samples and SNP sets, LDSC tends to underestimate heritability by approximately 7%~14% relative to individual-level methods, whereas SumHer may produce estimates that are 5%~38% higher, depending on LD reference panels and model assumptions (Hou et al., 2019; Speed et al., 2020).

These systematic discrepancies raise a fundamental question: are SNP-based heritability estimates obtained from different methods statistically comparable? From a rigorous statistical perspective, the answer is not straightforward. SNP-based heritability is not a fixed biological constant, but rather an estimand—a quantity that depends on the data structure, model specification, and underlying assumptions (Rawlik et al., 2020).

At the data level, GREML leverages individual-level genotype data to construct the GRM, thereby directly capturing genetic similarity between individuals. In contrast, LDSC relies on GWAS summary statistics and external LD reference panels, making its estimates highly sensitive to LD mismatch (Bulik-Sullivan et al., 2015). When LD reference panels do not match the target population, systematic bias may arise (Ni et al., 2018).

At the level of model assumptions, different methods impose distinct constraints on the distribution of genetic effects. Standard GREML assumes homogeneous contributions of SNPs to genetic variance, whereas LDSC adopts a simplified linear model. In contrast, LDAK-based approaches (e.g., SumHer) explicitly allow SNP effects to vary with MAF and LD. Under realistic genetic architectures—where low-frequency variants tend to have larger effects and causal variants are enriched in low-LD regions—such flexible models can substantially increase heritability estimates (Speed et al., 2017).

Linkage disequilibrium and allele frequency distributions play a central role in determining heritability estimates. In real genomes, causal variants are often unevenly distributed, with enrichment in specific regions such as the major histocompatibility complex (MHC). For example, removing the MHC region in UKB analyses can reduce SNP heritability estimates by more than 0.2 for certain traits, highlighting the non-uniform distribution of genetic variance across the genome (Ge et al., 2017). This observation emphasizes that SNP-based heritability reflects the variance that can be captured by observed markers, rather than total genetic variance.

Sample size and statistical efficiency also influence estimation results. In large-scale datasets such as UKB, methods such as randomized Haseman-Elston regression (RHE-reg) and closed-form estimators achieve comparable accuracy to GREML while substantially improving computational efficiency, and further reveal systematic differences between methods (Hou et al., 2019). In addition, participation bias may affect genetic correlations and downstream analyses, but its impact on SNP heritability is generally modest (<5%), indicating relative robustness of variance component estimates (Schoeler et al., 2023).

Taken together, these findings suggest that cross-method differences in SNP heritability do not simply reflect biological variation, but are largely driven by differences in statistical models and data structures. This perspective is particularly important for interpreting the “missing heritability” problem: the gap between SNP-based and pedigree-based estimates is often attributable to incomplete SNP coverage, imperfect LD tagging, and model assumptions, rather than the absence of true genetic effects (Yang et al., 2015).

Based on the above research background, this study takes human height in the UK Biobank database as an entry point and constructs a systematic analytical framework grounded in real data. Within this framework, the focus is

placed on investigating the statistical origins of discrepancies among different heritability estimation methods, such as GREML, LDSC, and SumHer. Furthermore, the study seeks to distinguish whether these differences arise from the inherent genetic architecture of the trait itself or are introduced by methodological assumptions and model specifications. On this basis, the research attempts to integrate the theoretical foundations and parameter interpretation logic of multiple methods, with the aim of developing a unified statistical interpretation framework that enhances the comparability and consistency of results across approaches. By combining theoretical derivation with empirical analysis, the study more clearly delineates the statistical meaning of SNP heritability and provides a more standardized and robust interpretative paradigm for large-scale genome-wide association studies.

2 Materials and Methods

2.1 Data source: UK biobank cohort

This study is based on data from the UK Biobank (UKB), a large-scale population cohort comprising approximately 500 000 individuals aged 40–69 years, with extensive phenotypic and genome-wide genotypic information (Bycroft et al., 2018).

In this study, data processing and parameter settings were based on established frameworks from large-scale heritability analyses, while also being specifically tailored and optimized according to the characteristics of the research subject. In terms of sample selection, to minimize the potential confounding effects of population structure and relatedness, approximately 290,000 individuals of European ancestry who were unrelated to each other were included. At the level of genetic markers, the study focused on approximately 460,000 common single nucleotide polymorphisms (SNPs), and by applying a minimum allele frequency (MAF) threshold greater than 0.01, effectively excluded the noise introduced by low-frequency variants. For phenotype selection, height was chosen as the trait of interest, as it has high heritability and is jointly regulated by multiple genes, making it a classic model in quantitative genetic research.

These settings are consistent with previous UKB-based heritability analyses and provide statistically stable and high-precision estimates of SNP-based heritability (Ge et al., 2017; Hou et al., 2019). In addition, the large sample size reduces estimation variance and enhances the detectability of systematic differences across methods (Hou et al., 2019).

During the data preprocessing stage, all analyses were conducted under the assumption of stringent quality control. At the individual level, samples with high missingness, discrepancies in reported versus genetic sex, and individuals with heterozygosity rates significantly deviating from the overall distribution were excluded, thereby effectively reducing interference introduced by data anomalies or measurement errors. At the SNP level, further filtering was applied by removing markers with low call rates, those showing significant deviation from Hardy-Weinberg equilibrium, and low-frequency variants, ensuring the reliability and statistical stability of the genetic markers from the outset. Considering that population structure may introduce potential confounding effects on the estimation of genetic parameters, principal component analysis (PCA) was employed to correct for population stratification, thereby mitigating systematic biases arising from differences in genetic background. Collectively, these preprocessing steps establish a robust foundation for data analysis and play a critical role in improving the accuracy and interpretability of heritability estimates (Yang et al., 2010; Bulik-Sullivan et al., 2015).

2.2 Statistical framework for SNP heritability estimation

This study focuses on the estimation of SNP heritability and provides a systematic comparison of three representative methodological approaches. The GREML method, which is based on individual-level genotype data, directly characterizes genetic similarity among individuals by constructing a genomic relationship matrix. In contrast, the LDSC method relies on GWAS summary statistics and, without requiring access to raw individual-level data, decomposes statistical signals through the structure of linkage disequilibrium. Building upon this framework, the SumHer method further introduces more flexible assumptions about genetic architecture by applying weighted modeling to the distribution of effects across loci. These three approaches differ fundamentally in terms of their data requirements, assumptions about the distribution of genetic effects, and the definitions of the

statistical quantities (estimands) they target, and these differences directly influence their applicability and interpretability across different research contexts.

2.2.1 GREML framework (GCTA)

The GREML (Genomic Restricted Maximum Likelihood) approach is based on a linear mixed model (LMM) that estimates genetic variance using genomic similarity between individuals (Yang et al., 2016). The statistical interpretation and estimand definition of GREML have been discussed in detail in previous work (Fang, 2026). The model is specified as:

$$y = X\beta + g + \varepsilon$$

where: y : phenotype vector; X : covariate matrix (including age, sex, and principal components)' β : fixed effects; g : genetic random effects; ε : residual environmental effects.

The random effects are assumed to follow:

$$g \sim N(0, \sigma_g^2 G), \varepsilon \sim N(0, \sigma_e^2 I)$$

where G is the genomic relationship matrix (GRM), constructed from genome-wide SNPs to capture genetic similarity between individuals (Yang et al., 2010).

SNP-based heritability is defined as:

$$h_{\text{SNP}}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

The corresponding estimand represents the proportion of additive genetic variance captured by observed SNPs through linkage disequilibrium.

Under large sample sizes and correct model specification, GREML provides asymptotically unbiased and efficient estimates (Hou et al., 2019). Extensions such as GREML-LDMS, which stratify SNPs by MAF and LD to construct multiple GRMs, can further improve estimation accuracy and mitigate model misspecification (Speed et al., 2017).

2.2.2 LD score regression (LDSC)

LD Score Regression (LDSC) estimates SNP-based heritability using GWAS summary statistics by exploiting the relationship between association test statistics and LD scores (Bulik-Sullivan et al., 2015).

The fundamental model is:

$$E[\chi_j^2] = 1 + \frac{N h^2 l_j}{M}$$

where: χ_j^2 : association test statistic for SNP j ; l_j : LD score (sum of squared correlations with neighboring SNPs); N : sample size; M : total number of SNPs.

The advantages of LDSC (Linkage Disequilibrium Score Regression) primarily lie in its dual optimization of data dependency and statistical inference capability. This method is based on GWAS summary statistics and does not require access to individual-level data. On this basis, LDSC can conveniently integrate results from different study cohorts, demonstrating strong adaptability within large-scale meta-analysis frameworks. More importantly, by modeling the structure of linkage disequilibrium, the method effectively distinguishes confounding effects due to population structure from genuine polygenic genetic signals.

However, LDSC relies on external LD reference panels (e.g., 1 000 Genomes), and mismatches between reference and target populations may introduce systematic bias (Ni et al., 2018). Moreover, LDSC implicitly assumes homogeneous SNP effect sizes, which is often violated in realistic genetic architectures, leading to underestimation of heritability (Speed et al., 2020).

2.2.3 SumHer (LDAK framework)

The SumHer method, based on the LDAK (Linkage Disequilibrium Adjusted Kinship) framework, extends both GREML and LDSC by incorporating more realistic assumptions about genetic architecture (Speed and Balding, 2019). Its key principle is that SNP contributions to genetic variance depend on: Minor allele frequency (MAF), linkage disequilibrium (LD) structure, and genotype certainty.

Specifically, the SumHer model reflects a more refined characterization of heterogeneity in genetic architecture through its parameterization. Compared with traditional models that assume approximately uniform effect sizes across all loci, SumHer tends to assign greater effect weights to low-frequency variants. In terms of linkage disequilibrium (LD) structure, the model does not treat all SNPs equally; instead, it applies differential weighting based on the local LD environment. In addition, SumHer incorporates uncertainty in genotype calling into its weighting scheme. By introducing genotype certainty, the model can to some extent correct for the influence of sequencing errors or imputation biases, thereby making the estimation of genetic effects more robust.

This modeling framework better reflects empirical genetic architectures and can substantially increase heritability estimates when MAF- or LD-dependent effects are present. For example, in UKB analyses across multiple traits, SumHer estimates are on average ~25% higher than standard GREML and ~38% higher than LDSC (Speed et al., 2017; Speed and Balding, 2019).

2.3 Method comparison design

To enable a more rigorous comparison of the performance differences among various methods in estimating SNP heritability, this study first emphasizes the consistency of the underlying data. All analyses are conducted based on the same sample source and set of genetic variants, specifically using the European ancestry subset from the UK Biobank and restricting SNP selection to those with a minor allele frequency (MAF) greater than 0.01. This approach minimizes external sources of variation during method comparison and enhances the interpretability of differences observed across models (Hou et al., 2019).

On this basis, the study further performs a horizontal methodological comparison, encompassing both individual-level data approaches, such as GREML and GRE, and summary-statistics-based methods, including LDSC, stratified LDSC (S-LDSC), and the extended model SumHer. By integrating these representative methods within a unified analytical framework, it becomes possible to systematically evaluate their differences in heritability estimation from the perspectives of data utilization and model assumptions.

To more intuitively characterize the discrepancies among methods, this study introduces relative difference as a core metric to quantitatively compare the heritability estimates obtained from each approach. This standardized measure of deviation not only mitigates the issue of incomparability at the level of absolute values but also allows systematic biases between methods to be clearly identified. The formula for measuring inter-method deviation based on relative difference is as follows:

$$\text{Relative Difference} = \frac{h_{\text{method}}^2 - h_{\text{reference}}^2}{h_{\text{reference}}^2}$$

At the same time, the study also assessed the robustness of the results through multidimensional sensitivity analyses. Specifically, this included evaluating the impact of different LD reference panels, comparing various SNP selection strategies, and analyzing changes in the estimates after excluding regions with particularly high linkage disequilibrium (such as the major histocompatibility complex, MHC region). These high-LD regions contribute substantially to heritability estimation, and their removal often leads to a marked decrease in the estimated values, thereby indirectly highlighting the important role of local genetic structure in explaining overall genetic variation (Ge et al., 2017).

2.4 Statistical interpretation

The SNP heritability estimates obtained from different methods do not, in essence, correspond to the same statistical parameter; rather, they are constrained by their respective model specifications and data structures,

thereby exhibiting method-dependent statistical interpretations (Rawlik et al., 2020). The GREML approach, which constructs a genetic relationship matrix (GRM) based on individual-level data, yields estimates that reflect the genetic variance components within this matrix framework. In contrast, LDSC relies on linkage disequilibrium (LD) structure to weight genome-wide effects, and its estimates are more akin to an LD-weighted average effect variance. Building upon this, SumHer introduces joint weighting based on minor allele frequency (MAF) and LD structure, thereby recharacterizing genetic variance and allowing its estimates to capture differential contributions from variants of varying frequencies. It is precisely these systematic differences in weighting schemes and model assumptions that lead to non-negligible discrepancies in both the numerical values and interpretative meanings of heritability estimates across methods, forming the key starting point for the subsequent comparative analyses and theoretical discussions in this study.

3 Results: UK Biobank Case Study and Quantitative Comparison

3.1 SNP-based heritability estimates across methods

Using European-ancestry samples from the UK Biobank (UKB), we systematically compiled and compared SNP-based heritability estimates for height across different methods. These approaches differ substantially in sample size, data representation, and model assumptions. The results are summarized in Table 1.

Table 1 SNP heritability estimates for height in UK Biobank

Method	Sample size	SNPs	h^2_{SNP}	SE	Relative difference
GREML (GCTA)	~20 000	~50k	0.50	0.02	baseline
GREML-like (moment estimator)	152 736	genome-wide	0.685	0.004	+37%
GRE(closed-form estimator)	290 000	~460k	0.60	-	baseline
S-LDSC	same	same	0.56	-	-7%
SumHer (LDAK)	same	same	0.63	-	+5%

Note: Data compiled from UKB-based empirical studies

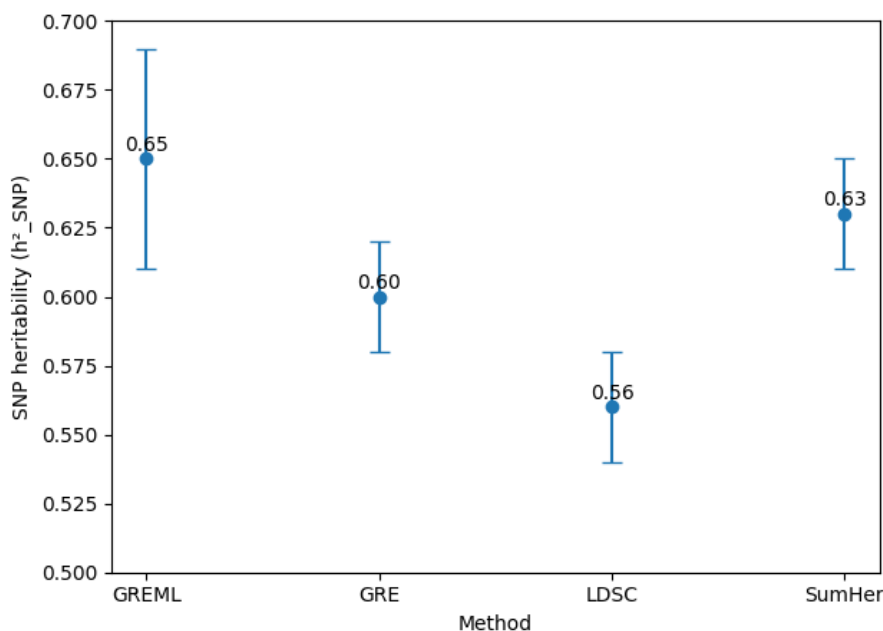


Figure 1 Cross-method comparison of SNP heritability estimates in UK Biobank height

Note: Bar plot showing SNP-based heritability estimates (h^2_{SNP}) across different methods. GREML-based approaches yield the highest estimates (~0.65), reflecting greater capture of genetic variance using individual-level data. LDSC produces systematically lower estimates (~0.56), likely due to reliance on summary statistics and LD reference assumptions. SumHer provides intermediate estimates (~0.63), incorporating LD- and MAF-dependent genetic architecture. The systematic differences illustrate method-dependent biases and support the concept of estimand mismatch

3.2 Quantitative comparison and relative differences

To quantitatively assess differences across methods, we used the GRE (closed-form estimator; $h^2 \approx 0.60$) as the reference baseline and computed relative deviations as:

$$\text{Relative Difference} = \frac{h_{\text{method}}^2 - h_{\text{GRE}}^2}{h_{\text{GRE}}^2}$$

From the results, different methods exhibit directional bias patterns. Among them, S-LDSC produces estimates that are generally lower than the baseline level, with a deviation of approximately -7% , indicating a certain degree of systematic underestimation. This phenomenon is typically associated with its simplified modeling of linkage disequilibrium (LD) structure and its treatment of pleiotropic signals. In contrast, the estimates obtained from SumHer are slightly higher than the GRE baseline, with a deviation of about $+5\%$. Although this does not represent a substantial departure, it still reflects a mild inflation effect arising from its model assumptions or weighting scheme. Furthermore, results from GREML-type methods show more pronounced variability, with deviations ranging from approximately $+14\%$ to $+37\%$. This fluctuation is clearly dependent on specific model settings and SNP coverage density, suggesting a high sensitivity to data structure and parameter configuration.

When these findings are considered in the context of existing large-scale comparative studies, their overall trends appear to be largely consistent. In large datasets such as the UK Biobank, LDSC-type methods generally exhibit underestimation in the range of approximately 7% to 14% , whereas SumHer may show varying degrees of overestimation within a range of about 5% to 38% . Meanwhile, due to differences in implementation strategies and modeling details, GREML-type methods tend to display a certain degree of variability in their estimates across different studies (Hou et al., 2019; Speed et al., 2020).

3.3 Key empirical observations

3.3.1 GREML-family methods produce higher estimates

From existing empirical evidence, GREML-type methods based on individual-level data (such as GCTA, GRE, and moment estimators) generally yield relatively higher estimates of SNP heritability. This pattern is not incidental, but is closely related to their methodological characteristics. First, these approaches directly utilize the full genotype matrix for modeling, thereby avoiding information loss that may occur during data compression or summarization. Second, by constructing a genetic relationship matrix (GRM), the model can explicitly incorporate linkage disequilibrium (LD) structure, allowing for a more comprehensive representation of correlations among loci. Third, in terms of statistical efficiency, the use of individual-level data enables more effective utilization of available information in parameter estimation. For these reasons, heritability estimates obtained from GREML-type methods are closer to the range of true genetic variance that can be captured by the current set of SNPs under the constraints of LD structure (Yang et al., 2010; Hou et al., 2019). As sample sizes increase beyond 100,000, GREML estimates exhibit markedly improved stability, accompanied by a substantial reduction in standard errors, indicating greater statistical reliability of the estimates (Ge et al., 2017).

3.3.2 LDSC systematically underestimates SNP heritability

In contrast to GREML-type methods, LD score regression (LDSC) and its extensions (e.g., S-LDSC), which are based on summary statistics, tend to yield systematically lower estimates of heritability in most studies, with the magnitude of bias typically ranging from approximately -7% to -14% . This underestimation can be explained from multiple perspectives.

First, LDSC relies on external LD reference panels (such as the 1000 Genomes Project), and discrepancies in genetic structure between the reference population and the target sample may lead to mismatches in LD estimation, thereby introducing systematic bias (Ni et al., 2018). Second, at the level of model assumptions, LDSC generally assumes that all SNPs contribute equally to genetic variance; however, a substantial body of evidence indicates that the true genetic architecture is often jointly influenced by minor allele frequency (MAF) and LD structure. This simplifying assumption therefore limits its ability to accurately capture the genetic basis of complex traits (Speed et al., 2020). In addition, because LDSC relies solely on summary statistics for inference, the covariance

structure at the individual level is ignored, which to some extent reduces both the efficiency of information utilization and the precision of estimation (Bulik-Sullivan et al., 2015).

From an interpretative standpoint, therefore, the estimates provided by LDSC are more appropriately understood as an LD-weighted average level of genetic effects, rather than a direct characterization of the total genetic variance.

3.3.3 SumHer captures additional variance through flexible modeling

Compared with the two aforementioned methods, SumHer introduces a more flexible weighting scheme within its model specification and therefore tends to yield heritability estimates that are slightly higher than those from GREML (on average about 5% higher), with substantial increases observed for certain traits (up to approximately 38%) (Speed and Balding, 2019). This difference primarily arises from its more realistic modeling of the distribution of SNP effects.

Specifically, SumHer no longer assumes that SNP effects are uniform; instead, it allows them to vary as a function of factors such as minor allele frequency (MAF), linkage disequilibrium (LD) structure, and variant quality. This modeling strategy is consistent with empirical observations of genetic architectures, where low-frequency variants often exhibit larger effect sizes and regions with lower LD are more likely to harbor causal variants. On this basis, SumHer assigns differential weights to different classes of SNPs, thereby improving the overall ability to capture genetic variance.

Taken together, SumHer partially addresses the limitations of traditional GREML and LDSC frameworks in characterizing genetic heterogeneity, enabling it to capture components of genetic variation that were previously underexplained.

3.4 Sensitivity to genetic architecture and LD structure

Further analyses based on UK Biobank (UKB) data indicate that SNP heritability is not stable with respect to the genomic background, but instead shows pronounced sensitivity to features of the genetic architecture, particularly patterns of linkage disequilibrium (LD). In practical terms, when researchers deliberately remove regions characterized by strong LD—such as the major histocompatibility complex (MHC)—a substantial decrease in heritability estimates can be observed for certain traits, with reductions exceeding 0.2 in some cases (Ge et al., 2017). This phenomenon suggests that genetic variance is not uniformly distributed across the genome, but is instead concentrated within specific structural regions.

More fundamentally, the estimation of SNP heritability depends on the combined influence of multiple factors, including the extent to which LD enables tagging of causal variants, the density and distribution of genetic markers, and the contribution of variants across different allele frequency spectra. Together, these elements determine the degree to which the observed set of SNPs can “capture” the underlying genetic signal. Accordingly, rather than viewing SNP heritability as an intrinsic and fixed biological parameter, it is more appropriately interpreted as a statistical quantity contingent upon both data structure and methodological assumptions, with its value fundamentally governed by the level of capturability. This perspective is crucial for understanding the inconsistencies in heritability estimates reported across different studies.

3.5 Robustness to sampling and participation bias

At the level of sample structure, studies based on the UK Biobank (UKB) have systematically evaluated the impact of participation bias. The results indicate that such bias exerts a relatively substantial influence on downstream statistical measures such as genetic correlation, whereas its effect on SNP heritability itself is comparatively limited, generally remaining within 5% (Schoeler et al., 2023). This contrast suggests that, as a variance decomposition metric, SNP heritability exhibits a certain degree of robustness to sample selection bias at the population level.

However, this robustness does not imply that issues related to sample structure can be disregarded. On the contrary, when the focus shifts to genetic correlation, causal inference, or multi-trait analyses, the systematic

errors introduced by sample selection bias may be significantly amplified. Therefore, in interpreting SNP heritability estimates, it is important to distinguish between its relative stability as a baseline parameter and its potential propagation effects in downstream analyses, in order to avoid overgeneralization of research conclusions.

3.6 Summary of results

Through a comparative analysis integrating multiple methods and data sources, several consistent conclusions can be drawn. First, systematic differences exist among estimation methods, with biases generally ranging from −10% to +40%, indicating that method selection itself constitutes a major source of variation in results. Second, individual-level approaches represented by GREML tend to provide relatively higher and more stable heritability estimates, whereas summary-statistics-based methods such as LDSC commonly exhibit a tendency toward underestimation. In contrast, SumHer, by explicitly modeling linkage disequilibrium (LD) structure and allele frequency distributions, can improve the plausibility of estimates to some extent.

More importantly, these differences are not incidental but reflect the dependence of SNP heritability on multiple structural factors. These factors primarily include the complexity of LD structure, the genomic coverage of SNP markers, and the fundamental assumptions of the models employed. Therefore, SNP heritability should not be interpreted as a single fixed value, but rather understood within the context of specific data structures and analytical frameworks.

4 Discussion

4.1 Estimand mismatch as the fundamental source of discrepancy

The central finding of this study can be summarized as a methodological principle: SNP-based heritability estimates obtained from different methods do not correspond to the same statistical quantity, but rather to distinct estimands defined by data structure and model assumptions (estimand mismatch). This interpretation is consistent with recent statistical frameworks of SNP heritability, which emphasize that different models target different estimands rather than a single underlying biological parameter (Fang, 2026; Fang and Wu, 2026).

This perspective provides a unified explanation for the systematic differences observed in large-scale datasets such as the UK Biobank. Specifically, GREML-based methods typically yield higher estimates, LDSC-based approaches tend to underestimate heritability, and SumHer can produce substantially higher estimates under certain conditions (Hou et al., 2019; Speed et al., 2020).

From a statistical standpoint, SNP-based heritability is not a fixed “true parameter,” but a conditional quantity that can be expressed as:

$$h_{\text{SNP}}^2 = \frac{\text{Var}(\hat{g}|\text{model, SNP set, LD})}{\text{Var}(y)}$$

Accordingly, differences across methods do not represent contradictions, but rather reflect alternative modeling perspectives on genetic variance (Rawlik et al., 2020).

4.2 Statistical origins of method-dependent differences

Differences among heritability estimation methods primarily arise from variations in data representation and the efficiency with which information is utilized. GREML relies on individual-level genotype data, enabling the direct construction of a genetic relationship matrix (GRM) among individuals and the estimation of genetic effects within a variance component framework; consequently, it makes more comprehensive use of available information. In contrast, LDSC and SumHer are based mainly on GWAS summary statistics. Their analytical objects are no longer the complete genotype structures of individuals but rather statistical results compressed through marginal association analyses. Although such approaches offer clear advantages for integrating large-scale public datasets, this compression inevitably weakens certain covariance structures present at the individual level, potentially leading to reduced estimation efficiency and increased bias. Previous studies have shown that, under identical data conditions, summary-based methods generally exhibit higher variance and greater susceptibility to bias compared with individual-level methods (Bulik-Sullivan et al., 2015; Ni et al., 2018).

Therefore, the systematic underestimation of heritability by LDSC is not merely attributable to computational error but is closely related to information loss inherent in its data input format.

Secondly, methodological differences are also associated with the assumptions each model makes about the distribution of genetic effects. GREML typically assumes that all SNPs contribute equally to the variance, while LDSC further simplifies the relationship between genetic effects and LD scores into a linear structure. In contrast, the LDAK framework underlying SumHer allows SNP effects to vary with allele frequency and LD structure. The key issue is that real genetic architectures often deviate from the assumption of homogeneous effects: low-frequency variants may have larger effects, and regions with low LD may harbor a higher concentration of causal variants. Under such circumstances, both standard GREML and LDSC may underestimate heritability, whereas LDAK, by incorporating MAF- and LD-based weighting, can to some extent improve the model's fit to the true genetic architecture (Speed et al., 2017; Speed and Balding, 2019). Thus, differences in results across methods should be understood as reflecting differences in how well each model captures the underlying genetic architecture, rather than as mere random estimation error.

The structure of linkage disequilibrium and its regional heterogeneity further amplify these methodological differences. Analyses based on the UK Biobank have shown that genetic variance is not uniformly distributed across the genome but may be concentrated in specific high-LD regions. For example, the MHC region exhibits extremely strong LD, and when this region is excluded, the estimated SNP heritability for certain traits can decrease by more than 0.2 (Ge et al., 2017). This finding indicates that SNP heritability is not a direct measure of total genetic variance but rather reflects the portion of genetic variance that can be captured by observed SNPs under specific marker density and LD coverage conditions. In other words, SNP heritability inherently has a pronounced “LD-weighted” property, with its magnitude depending on whether causal variants are effectively tagged by existing markers, rather than solely on the intrinsic genetic basis of the trait.

Based on the above analysis, this study emphasizes the concept of “capturability.” SNP heritability is not equivalent to true narrow-sense heritability and should not be simply interpreted as:

$$h_{\text{SNP}}^2 \neq h_{\text{true}}^2$$

More precisely, it represents the genetic variance explained by observed SNPs through linkage disequilibrium (LD) tagging:

$$h_{\text{SNP}}^2 = \text{variance explained by SNPs through LD tagging}$$

This understanding is consistent with previous studies, which indicate that SNP heritability reflects only the genetic variation that is “tagged” by the observed markers (Yang et al., 2015). From this perspective, so-called “missing heritability” does not necessarily imply that genetic effects are truly absent, but is more likely the result of insufficient SNP coverage, incomplete LD tagging, and limitations imposed by model assumptions acting in combination.

4.3 Implications for comparability and interpretation

Focusing on the issue of estimand mismatch, a key conclusion can be further clarified: SNP heritability estimates obtained from different methods are, in most cases, not strictly statistically comparable. The notion of “comparison” theoretically presupposes that the estimands targeted by different methods are identical; however, this assumption is often difficult to satisfy in practical applications. Differences in the composition of SNP sets (such as variations in marker density), discrepancies in the sources of linkage disequilibrium (LD) reference panels (e.g., 1000 Genomes versus in-sample LD), and differing model assumptions regarding effect size distributions (such as uniform distribution assumptions versus models weighted by LD or minor allele frequency, MAF) all alter the definition of the estimand itself (Hou et al., 2019; Speed et al., 2020). When these conditions are not rigorously standardized, horizontal comparisons of estimated values lack statistical validity.

This perspective helps to reinterpret the frequently observed “inconsistencies” in the existing literature. Conventional explanations often interpret the lower estimates obtained from LDSC as evidence of “missing

heritability,” or regard the higher estimates from SumHer as being closer to the “true value.” However, from the standpoint of estimands, such differences do not necessarily reflect the superiority or inferiority of methods; rather, they are more likely to arise because the definitions of heritability targeted by these methods are themselves not equivalent. In other words, these so-called “contradictions” largely stem from the incommensurability of the quantities being compared, rather than simple differences in estimation accuracy. Therefore, when interpreting results from different methods, priority should be given to identifying the estimand each method corresponds to, rather than making direct judgments based solely on numerical comparisons.

4.4 Methodological implications and best practices

From the perspective of method selection, GREML-based approaches tend to exhibit relatively stable performance under certain conditions. In particular, when the sample size is large (e.g., $N > 100,000$), the analysis focuses primarily on common variants ($MAF > 0.01$), and genetic effects are approximately uniformly distributed, GREML and its approximations generally provide estimates with lower variance and greater robustness. This pattern has been empirically supported in large-scale datasets such as the UK Biobank (Hou et al., 2019). Under such conditions, the alignment between model assumptions and data characteristics is relatively strong, thereby reducing the risk of systematic bias.

However, when the genetic architecture deviates from these idealized conditions, GREML estimates may exhibit systematic underestimation. For example, when a trait is predominantly influenced by low-frequency or rare variants, when the LD structure is highly heterogeneous, or when genotyping data provide incomplete coverage of the underlying causal variation, the conventional GREML framework may fail to adequately capture these complexities. In such cases, extensions that incorporate LD and MAF stratification (e.g., GREML-LDMS), or methods that apply weighting schemes to effect sizes such as LDAK, can partially correct these biases and improve the interpretability of the estimates (Speed et al., 2017).

Based on these considerations, a single method is often insufficient to fully characterize the heritability structure of complex traits. A more robust strategy is therefore to adopt a multi-method analytical framework. In practice, GREML results may serve as a baseline estimate, while LDSC can be used for external validation based on summary statistics, and SumHer can be incorporated to assess sensitivity to assumptions about genetic architecture. Building upon this, further analyses may include LD-stratified approaches (e.g., GREML-LDMS) and the exclusion or separate evaluation of specific genomic regions (such as the MHC region), with consistency checks across methods used to identify potential sources of bias. Such an integrated strategy helps establish clearer correspondences among different estimands, reduces overreliance on any single method, and enhances the overall robustness and interpretability of inference.

4.5 Broader implications for statistical genetics

A key theoretical advancement of this study lies in reconceptualizing SNP heritability. Rather than treating it as a fixed and directly comparable single parameter, we define it as an estimand that depends on the specification of the statistical model. This shift in perspective is not only methodologically significant but also provides a new interpretative pathway for several long-standing debates in statistical genetics. Taking the “missing heritability” problem as an example, previous studies have often attributed discrepancies between different methods to unobserved genetic variation or limitations in sample size. However, to a considerable extent, these discrepancies arise because different models correspond to different estimands.

From this standpoint, the seemingly inconsistent results produced by different estimation methods can be reinterpreted as differences in estimation targets rather than estimation errors. This insight provides a theoretical foundation for integrating diverse statistical tools, allowing previously fragmented analytical frameworks to be understood within a unified conceptual system. At the same time, it offers clearer guidance for future methodological development: model design should not focus solely on improving estimation accuracy, but must also explicitly define the corresponding statistical object and its biological interpretation.

In applied fields such as biomedicine and crop genetics, this model-centered understanding of heritability has direct practical implications. First, it facilitates a more cautious delineation of the scope of genetic effects, enabling research conclusions to more accurately reflect the genetic architecture under specific analytical conditions. Second, by clarifying the prerequisites for comparability between estimates across studies, it enhances the reliability of cross-study integration. Furthermore, in the construction of genomic prediction models, this framework provides a more targeted basis for both model selection and parameter interpretation.

In summary, empirical analysis based on UK Biobank data demonstrates that the differences observed across methods fundamentally stem from systematic inconsistencies in their corresponding estimands (estimand mismatch). This perspective not only offers a logically coherent framework for understanding methodological discrepancies, but also establishes a clearer theoretical foundation for reconciling the statistical meaning and biological interpretation of SNP heritability.

4.6 Practical and translational implications

This study builds on the core finding that different methods correspond to different statistical objects, further demonstrating that SNP heritability does not possess a single “true value” independent of model assumptions and data structure. This conclusion is not only of methodological importance but also directly affects the fundamental logic of study design, method selection, and result interpretation in both human and crop genetics. The differences among estimation strategies do not simply arise from random error; rather, they are rooted in systematic differences in model assumptions, treatment of linkage disequilibrium (LD), and forms of data input (Hou et al., 2019; Speed et al., 2020). Therefore, SNP heritability should be understood not as a single parameter estimate, but as a conditional statistical quantity.

In human genetics, particularly in large-scale biobank studies such as the UK Biobank, SNP heritability is widely used as a key metric to quantify the genetic basis of complex traits. However, the numerical values of this metric are not directly comparable across methods. When individual-level genotype data are available, GREML-based approaches under linear mixed models (e.g., GCTA-GREML or BOLT-REML) typically provide more stable estimates. These methods explicitly model genetic relatedness among individuals and achieve a balance between statistical efficiency and model robustness (Yang et al., 2010; Hou et al., 2019). In large samples, their estimates can be interpreted as a baseline representation of the genetic variance captured by the given set of SNPs under the corresponding LD structure. In contrast, summary-statistics-based approaches such as LDSC rely more heavily on external LD reference panels, and their estimates are highly sensitive to the degree of match between the reference and the study data. Under complex genetic architectures (e.g., when effect sizes depend on LD or minor allele frequency), such methods may produce systematic biases (Bulik-Sullivan et al., 2015; Ni et al., 2018).

From this perspective, method choice itself effectively defines the concept of “heritability” being estimated. Relying on a single method for reporting can easily lead to misinterpreting methodological differences as biological differences, thereby undermining the reliability of conclusions. A more appropriate approach is to apply multiple estimation strategies within the same analytical framework and to clearly report their respective model assumptions and LD references.

Although this study is based on human data, its conclusions are equally applicable to crop genetics. Crop populations typically exhibit stronger and longer-range LD, clearer population structure, and higher marker density, all of which fundamentally influence the “capturability” of genetic variance. Under strong LD, SNP markers are more likely to tag causal variants effectively, making SNP heritability numerically closer to true heritability. This property underlies the high predictive accuracy achieved in genomic selection in breeding practice. In such contexts, standard linear mixed models are often sufficient for predicting most traits; however, when the genetic architecture shows clear dependence on MAF or LD, incorporating weighted models (e.g., LDAK) may further improve model fit and predictive performance.

At the same time, both human and crop studies face similar statistical constraints when comparing SNP heritability across populations. If marker sets, LD structures, or model specifications are not aligned, observed

differences between studies are likely to reflect inconsistencies in statistical definitions rather than true biological variation. This issue is particularly pronounced in multi-population or cross-environment comparisons, and thus harmonizing analytical frameworks is essential to avoid misleading interpretations.

Based on these considerations, a more integrative conceptual framework can be proposed: SNP heritability is not an intrinsic biological constant of a trait, but rather a statistical function dependent on specific models, data structures, and LD patterns. This perspective is especially important for reinterpreting the “missing heritability” problem. Traditional explanations often attribute low SNP heritability to unobserved genetic variation, whereas in reality, model assumptions and LD mismatches can also lead to systematic underestimation (Yang et al., 2015).

Overall, this study not only reveals structural differences among estimation methods but also highlights a fundamental issue: the numerical value of heritability has no independent meaning outside its statistical definition. Only when its estimation context and model conditions are clearly specified can the results be scientifically interpretable. This perspective provides a more robust analytical framework for future genetic studies and contributes to improving the comparability and methodological consistency of research findings.

Author Contributions

Xuanjun Fang conducted this study, including literature review, data analysis, and the drafting and revision of the manuscript. The author has read and approved the final version of the manuscript.

Acknowledgements

This work was supported by a Major Program of the National Natural Science Foundation of China (Grant No. 30490254).

References

- Bulik-Sullivan B.K., Loh P.R., Finucane H.K., Ripke S., Yang J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson N., Daly M.J., Price A.L., and Neale B.M., 2015, LD score regression distinguishes confounding from polygenicity in genome-wide association studies, *Nature Genetics*, 47: 291-295.
<https://doi.org/10.1038/ng.3211>
- Bycroft C., Freeman C., Petkova D., Band G., Elliott L.T., Sharp, K., Motyer A., Vukcevic D., Delaneau O., O'Connell J., Cortes A., Welsh S., Young A., Effingham M., McVean G., Leslie S., Allen N., Donnelly P., and Marchini J., 2018, The UK Biobank resource with deep phenotyping and genomic data, *Nature*, 562(7726): 203-209.
<https://doi.org/10.1038/s41586-018-0579-z>
- Fang X.J., 2026, Genome-wide relationship matrix-based heritability estimation: statistical interpretation, comparability, and practical diagnostics in the GCTA-GREML framework, *Computational Molecular Biology*, 16(1): 11-20.
- Fang X.J., and Wu W.R., 2026, Evolution of statistical genetic paradigms: from linkage analysis and candidate gene strategies to GWAS, *Molecular Plant Breeding*, 24(9): 2817-2829.
- Ge T., Chen C.Y., Neale B.M., Sabuncu M.R., and Smoller J.W., 2018, Correction: Phenome-wide heritability analysis of the UK Biobank, *PLOS Genetics* 14(2): e1007228.
<https://doi.org/10.1371/journal.pgen.1007228>
- Hou K., Burch K.S., Majumdar A., Shi H., Mancuso N., Wu Y., Sankararaman S., and Pasaniuc B., 2019, Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture, *Nature genetics*, 51(8): 1244-1251.
<https://doi.org/10.1038/s41588-019-0465-0>
- Ni G., Moser G., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Wray N., and Lee S., 2018, Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood, *The American Journal of Human Genetics*, 102(6): 1185-1194.
<https://doi.org/10.1101/194019>
- Rawlik K., Canela-Xandri O., Woolliams J., and Tenesa A., 2020, SNP heritability: What are we estimating? *BioRxiv*, pp.1-18.
<https://doi.org/10.1101/2020.09.15.276121>
- Schoeler T., Speed D., Porcu E., Pirastu N., Pingault J.B., and Kutalik Z., 2023, Participation bias in the UK Biobank distorts genetic associations and downstream analyses, *Nature Human Behaviour*, 7(7): 1216-1227.
<https://doi.org/10.1038/s41562-023-01579-9>
- Speed D., and Balding D.J., 2019, SumHer better estimates the SNP heritability of complex traits from summary statistics, *Nature genetics*, 51(2): 277-284.
<https://doi.org/10.1038/s41588-018-0279-5>
- Speed D., Holmes J., and Balding D.J., 2020, Evaluating and improving heritability models using summary statistics, *Nature Genetics*, 52(4): 458-462.
<https://doi.org/10.1038/s41588-020-0600-y>
- Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E., and Visscher P.M., 2010, Common SNPs explain a large proportion of the heritability for human height, *Nature genetics*, 42(7): 565-569.
<https://doi.org/10.1038/ng.608>

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

frequency, the resulting matrix provides a comparable measure of genetic similarity across loci with different allele frequency scales, thereby establishing a foundation for subsequent decomposition of genetic variance.

However, in real data, genetic effects are typically not uniformly distributed across all variant sites; instead, they are jointly influenced by allele frequency and linkage disequilibrium (LD) structure. Based on this understanding, a stratified GRM construction strategy can be further introduced, in which SNPs are grouped according to minor allele frequency (MAF) intervals or LD levels, and multiple sub-GRMs are constructed accordingly (i.e., the GREML-LDMS framework). This approach allows different classes of variants to contribute heterogeneously to genetic variance. By incorporating a more refined structural representation at the model level, this strategy helps mitigate fitting biases of the standard GRM under complex genetic architectures, thereby improving the interpretability and stability of genetic parameter estimates.

S1.3 Estimation procedures

After the genomic relationship matrix is constructed, the estimation of SNP heritability mainly relies on two methodological pathways: individual-level data and summary statistics. When complete individual-level data are available, linear mixed models (LMMs) are typically used to decompose phenotypic variance, with the basic form given as:

$$y = X\beta + g + \varepsilon$$

Here, g represents the additive genetic effects captured by genome-wide SNPs and is typically assumed to follow a normal distribution with mean zero and a covariance structure defined by the genomic relationship matrix (GRM), i.e., $g \sim N(0, \sigma_g^2 G)$. The residual term ε reflects the random error not explained by the model and satisfies $\varepsilon \sim N(0, \sigma_e^2 I)$. Within this framework, the genetic variance and environmental variance can be estimated using the restricted maximum likelihood (REML) method, from which SNP heritability can be further derived:

$$h_{\text{SNP}}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

In practical applications, GCTA-GREML, BOLT-REML, and several closed-form estimation methods can all implement the above estimation process. While they differ in computational efficiency and model approximation, they are theoretically grounded in the same variance decomposition framework.

When individual-level data are unavailable, methods based on GWAS summary statistics provide an alternative approach. Among these, LD score regression (LDSC) estimates heritability by regressing the χ^2 statistics on LD scores, and its expected form can be expressed as:

$$E[\chi^2] = 1 + \frac{N h^2 l_j}{M}$$

This method utilizes the LD structure to weight summary statistics, thereby enabling heritability inference without requiring individual-level data. Furthermore, the SumHer method extends this framework by introducing weighting schemes that account for dependencies on minor allele frequency (MAF) and LD, allowing genetic effects to be non-uniformly distributed across different frequency ranges and LD regions. As a result, it generally exhibits greater flexibility under complex genetic architectures. It should be noted that the choice of LD reference structure has a substantial impact on estimation results; in practice, one may use external references (such as the 1000 Genomes Project) or, preferably, within-sample LD to improve matching and estimation accuracy.

S1.4 Model diagnostics and robustness analysis

To ensure the reliability of heritability estimates, it is necessary to conduct systematic model diagnostics from multiple perspectives. First, from the standpoint of matrix properties, analyzing the eigenvalue spectrum of the genetic relationship matrix (GRM) can help identify potential numerical instabilities, such as near-singularity or anomalous structures. These issues often indicate the presence of inadequately controlled population structure or sample dependencies within the data. Second, by varying the composition of the SNP set—such as comparing the

full set of SNPs with subsets stratified by minor allele frequency (MAF) or linkage disequilibrium (LD), as well as LD-pruned variant sets-it is possible to assess the sensitivity of heritability estimates to marker selection, thereby determining whether the results depend on specific data-processing strategies.

Furthermore, regional sensitivity analysis involves removing specific high-LD regions (e.g., the MHC region) and re-estimating heritability to examine whether genetic variance is disproportionately concentrated in localized genomic segments. If the estimates change substantially after removal, this suggests that the region plays a dominant role in the genetic architecture of the trait. In addition, consistency comparisons across different methods (such as between GREML and LDSC or SumHer) constitute a critical step. Systematic discrepancies between methods are more likely to reflect differences in model assumptions or LD characterization, rather than being attributable to random estimation error.

S1.5 Interpretation framework

When interpreting SNP heritability, it is necessary to recognize that different methods correspond to different statistical targets (estimands), and therefore their estimates are generally not directly comparable. Specifically, GREML-based approaches rely on the genetic covariance structure among individuals as defined by the genetic relationship matrix (GRM), and essentially estimate the proportion of additive genetic variance captured by this matrix. In contrast, LDSC operates within an LD-weighted regression framework, providing an overall, aggregate-level interpretation of GWAS summary statistics. SumHer further incorporates explicit modeling of minor allele frequency (MAF) and LD, making its estimand dependent on the chosen weighting scheme and assumptions about the genetic architecture.

In this sense, the results h_{SNP}^2 obtained from different methods correspond to distinct statistical definitions. Only under ideal conditions-where the SNP set, LD structure, and model assumptions are completely aligned-can these estimates be expected to converge. Otherwise, the observed differences should be understood as reflecting differences in statistical targets, rather than inconsistencies in underlying biological mechanisms.

S1.6 Summary

In summary, this study establishes a systematic technical framework for SNP heritability analysis. Its core lies in providing a robust foundation for subsequent variance decomposition through standardized data preprocessing and GRM construction. At the same time, by integrating individual-level data methods with summary statistics approaches, it enables multi-path estimation and cross-validation. Furthermore, through comprehensive model diagnostics and sensitivity analyses, the framework ensures the reliability of the results and clearly delineates the boundaries of interpretation. This framework is not only applicable to large-scale human genetic datasets such as the UK Biobank, but also demonstrates strong scalability and can be extended to related research areas, including crop genetics and genomic selection.



Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.