Molecular Plant Breeding 2026, Vol.17, No.1, 32-49
http://genbreedpublisher.com/index.php/mpb

Invited Article Open Access

Evolution of Statistical Genetic Paradigms: From Linkage Analysis and Candidate Gene Strategies to GWAS

Xuanjun Fang [1], Weiren Wu [2]

1 Hainan Provincial Key Laboratory of Crop Molecular Breeding, Hainan Institute of Tropical Agricultural Resources (HITAR), Sanya, 572025, Hainan, China
2 College of Agriculture, Fujian Agriculture and Forestry University, Fuzhou, 350002, Fujian, China

Co-corresponding authors: xuanjunfang@hitar.org; wuwr@fafu.edu.cn

Molecular Plant Breeding, 2026, Vol.17, No.1 doi: 10.5376/mpb.2026.17.0003
Received: 15 Apr., 2026
Accepted: 25 Apr., 2026
Published: 30 Apr., 2026

Preferred citation for this article:
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, 17(1): 32-49 (doi: 10.5376/mpb.2026.17.0003)

Abstract The genetic dissection of complex quantitative traits has long been constrained by polygenic architecture, small effect sizes, gene–environment interactions, and limited statistical power. In response to these challenges, statistical genetics has undergone a paradigm evolution from linkage analysis and candidate gene strategies to genome-wide association studies (GWAS). From a statistical genetic perspective, this study systematically reviews this methodological transition, with a focus on the changes in underlying assumptions, data structures, and statistical models, as well as their internal logical connections. Starting from classical assumptions such as Mendelian segregation and relatively stable recombination rates, we examine how factors including segregation distortion, genetic background heterogeneity, and genotyping errors may affect recombination rate estimation and likelihood distributions, thereby leading to systematic biases in linkage statistics and LOD curves. We further compare likelihood-based LOD statistics in linkage analysis with *p*-value–based significance measures derived from test statistics in GWAS, highlighting their differences in statistical foundations and significance assessment. It is emphasized that genome-wide significance thresholds generally require empirical calibration, such as permutation testing. With the development of high-density SNP data, large population samples, and mixed linear models, GWAS enables higher-resolution mapping by weakening locus-specific prior assumptions and explicitly modeling population structure and relatedness. Through a systematic comparison of different approaches and their respective limitations, this study argues that GWAS does not simply replace traditional methods, but represents a paradigm shift and extension in terms of statistical assumptions, model structures, and analytical scales. In the context of plant molecular breeding, the integration of GWAS with approaches such as eQTL analysis and genomic selection is expected to enhance the robustness of genetic inference and provide stronger statistical support for breeding decisions.

Keywords Statistical genetics; Paradigm evolution; Linkage analysis; Candidate gene strategies; Genome-wide association studies (GWAS); Mixed linear models

The genetic dissection of complex traits represents a central challenge shared by human health research and crop improvement, with profound implications for medicine, agriculture, and evolutionary biology. Complex traits are typically governed by polygenic effects, with genetic contributions arising not only from the cumulative action of numerous loci with small additive effects, but also from non-additive components (e.g., interactions among non-allelic loci, i.e., epistasis), as well as environmental influences and gene-environment interactions (G×E). In addition, phenotypic measurement errors, population evolutionary and breeding histories (such as ancestry structure and selection trajectories), and heterogeneity in linkage disequilibrium (LD) patterns across populations and genomic regions may further obscure true causal genetic signals (Watanabe et al., 2019). In both human and crop populations, this “polygenic-small-effect-environmental interaction” paradigm exhibits highly consistent structural characteristics at the level of statistical genetics. Therefore, establishing a robust statistical framework to uncover the genetic basis of complex phenotypes is essential for ensuring the reliability of disease prediction and breeding decisions.

Against this background, it is necessary to clarify the concept of statistical genetics. Statistical genetics generally refers to an interdisciplinary field that applies probabilistic models and statistical inference frameworks to analyze the relationships between genetic variation and phenotypes using population-level data. Its core objective is to

estimate genetic effects, characterize the genetic architecture of traits, and quantify uncertainty and predictive performance through explicit modeling. From a historical perspective, statistical genetics did not emerge independently from quantitative genetics, but rather represents a methodological extension built upon its theoretical foundations. Classical quantitative genetics focuses on variance decomposition and selection theory, using phenotypic correlations among relatives to describe the inheritance of polygenic traits. In contrast, statistical genetics, empowered by molecular markers and genomic data, further parameterizes and models these relationships, extending genetic analysis from "population-level variance description" to "locus-level and genome-wide statistical inference". Thus, the two fields are not substitutes, but reflect a paradigm evolution under a shared research objective: quantitative genetics provides the theoretical basis (e.g., heritability and polygenic models), while statistical genetics develops practical analytical frameworks—such as linkage analysis, variance component models, and genome-wide association studies (GWAS)—that enable a transition from locally hypothesis-driven approaches to genome-wide systematic modeling. This evolutionary trajectory constitutes the central theme of the present study.

From a methodological perspective, statistical genetics research in the 20th century gradually developed along two major analytical pathways. One is linkage analysis based on pedigrees and designed populations, which estimates recombination rates (θ) and LOD scores to localize quantitative trait loci (QTL) onto genetic linkage maps at the centimorgan (cM) scale (including populations such as RILs, BC, and DH), and has achieved success particularly for Mendelian traits or loci with large effects. The other pathway is based on Haseman–Elston (HE) regression and variance component methods, which utilize genetic relatedness or identity-by-descent (IBD) sharing for both heritability estimation and locus detection. The candidate gene approach, relying on prior biological knowledge, provides targeted insights but has been increasingly questioned due to hypothesis bias and limited reproducibility (Sebastiani et al., 2009; Zhang et al., 2019). These traditional approaches are constrained by limited marker density, small sample sizes, and insufficient statistical correction, making them inadequate for dissecting highly polygenic traits shaped by numerous small-effect loci and complex population structures, thereby contributing to the long-standing problem of "missing heritability" (Watanabe et al., 2019).

With the advent of high-throughput genotyping technologies and large-scale population resources, the research paradigm has shifted toward genome-wide, hypothesis-light scanning frameworks. Genome-wide association studies (GWAS) systematically evaluate hundreds of thousands to millions of variants across the genome, integrating mixed linear models (MLM) with genomic relationship matrices (GRM) to account for relatedness, while using principal component analysis (PCA) to correct for population structure. In addition, multiple testing corrections such as Bonferroni adjustment or false discovery rate (FDR) control are applied to address the issue of multiple comparisons (Pasaniuc and Price, 2016; Tibbs Cortes et al., 2021; Uffelmann et al., 2021). GWAS has identified thousands of genomic regions associated with traits in humans, crops, and model organisms, providing new insights into the polygenic and pleiotropic nature of complex traits.

However, GWAS also faces several challenges, including the interpretation of signals in non-coding regions, residual confounding from population stratification, and the gap between statistical associations and functional causality (Zhang et al., 2024). In crop research, multi-parent population designs such as nested association mapping (NAM) and multi-parent advanced generation intercross (MAGIC) populations serve as bridges between linkage and association mapping, enhancing the detection of small-effect loci and enabling cross-population validation. These developments have further promoted the integration of genomic selection (GS) and precision breeding strategies (Tibbs Cortes et al., 2021).

From the perspective of statistical genetics, this study systematically reviews the methodological evolution from linkage analysis and candidate gene strategies to genome-wide association studies (GWAS), with a particular focus on the transitions in assumptions, data structures, and statistical models. By doing so, it aims to elucidate the paradigm shift and complementary relationships among methods in the genetic dissection of complex traits.

1 Methodological Foundations of Linkage Analysis

1.1 Statistical framework and basic principles

Linkage analysis is a classical statistical approach that detects the co-segregation between genomic regions and traits based on genetic recombination information. Within the framework of quantitative genetics, its core principle is to utilize recombination events occurring during meiosis to convert the genetic distance between markers and putative causal loci into testable statistical quantities, thereby enabling the localization of trait-associated genomic regions across the genome (Fang et al., 2001; Xu et al., 2017). For quantitative traits, linkage analysis is typically implemented in the form of quantitative trait locus (QTL) mapping, where the positions and effects of putative QTL are inferred statistically by integrating individual-level genotype and phenotype data within a genetic map coordinate system (Meng et al., 2015; Zhang et al., 2015).

Within this framework, the recombination rate (θ) is the key parameter characterizing genetic relationships among markers. Using mapping functions (e.g., Haldane or Kosambi), θ can be transformed into genetic distance (in centimorgans, cM), thereby establishing a chromosomal coordinate system (Fang et al., 2001; Xu et al., 2017):

$$\text{Haldane}: d=-\frac{1}{2}\ln(1-2\theta)$$

$$\text{Kosambi}: d=\frac{1}{4}\ln(\frac{1+2\theta}{1-2\theta})$$

Where, d denotes genetic distance (cM) and θ represents the recombination rate. It should be emphasized that these mapping functions establish a functional relationship between recombination rate and genetic distance, providing a coordinate framework for downstream analyses, but do not themselves constitute the statistical inference procedure of linkage analysis.

1.2 Data basis and analytical implementation

Linkage analysis typically relies on specifically designed populations, such as biparental crosses (F_2, backcross populations) and derived populations including recombinant inbred lines (RILs) and doubled haploids (DH), which are generated through repeated selfing or haploid doubling. These populations provide observable segregation and recombination of alleles and thus form the basis for linkage analysis (Meng et al., 2015; Xu et al., 2017). In model species and major crops, the development of standardized population resources and high-density genotyping platforms has further enabled the use of multi-parent populations, such as multi-parent advanced generation intercross (MAGIC) populations and nested association mapping (NAM) populations, which increase recombination density and mapping resolution (Zheng et al., 2019; Qu et al., 2020).

Based on molecular marker genotyping data from these populations, genetic linkage maps can be constructed to describe the relative positions of markers and their recombination relationships. It is important to note that linkage map construction is a data preparation step that provides a spatial coordinate framework for subsequent statistical inference, rather than constituting linkage analysis itself (Fang et al., 2001; Meng et al., 2015). Building upon this framework, linkage analysis integrates marker genotypes and phenotypes of individuals to detect linkage between genomic regions and QTL. Therefore, QTL mapping can be regarded as the concrete implementation of linkage analysis in quantitative trait studies (Zhang et al., 2015; Xu et al., 2017).

From a statistical modeling perspective, linkage analysis methods can be broadly categorized into two classes. One is based on likelihood ratio testing, using the LOD (logarithm of odds) score as the core statistic, which evaluates evidence for linkage by comparing likelihoods under the “linkage” and “no linkage” hypotheses. The other is based on regression or correlation frameworks, such as Haseman–Elston (HE) regression, which detects linkage signals by modeling the relationship between identity-by-descent (IBD) sharing and phenotypic similarity (Sham and Purcell, 2001; Feingold, 2002; Chen, 2014). Despite their different forms, both approaches fundamentally rely on the statistical association between genetic sharing and phenotypic resemblance.

The most commonly used statistic in linkage analysis is the LOD score (log-likelihood ratio), defined as:

$$\mathrm{LOD}=\log_{10}\{\frac{L(\theta)}{L(0.5)}\}$$

Where, $L(\theta)$ is the likelihood under a given recombination rate θ, and $L(0.5)$ corresponds to the null hypothesis of no linkage. Traditionally, LOD ≥ 3 has been considered an empirical threshold for significant linkage, corresponding to approximately a 1 000:1 odds ratio (Fang et al., 2001; Zhang et al., 2015). However, in the context of high-density markers and genome-wide scans, fixed thresholds may not adequately account for marker correlation and population structure. Therefore, a more robust approach is to derive empirical thresholds via permutation testing to control the family-wise error rate (FWER) at the genome-wide level (Meng et al., 2015; Wang et al., 2024). It is also important to avoid directly equating LOD scores with $-\log_{10}(p)$, as they are only comparable under specific approximations; although LOD curves and Manhattan plots in GWAS appear similar in form, their statistical foundations and interpretations differ (Figure 1).

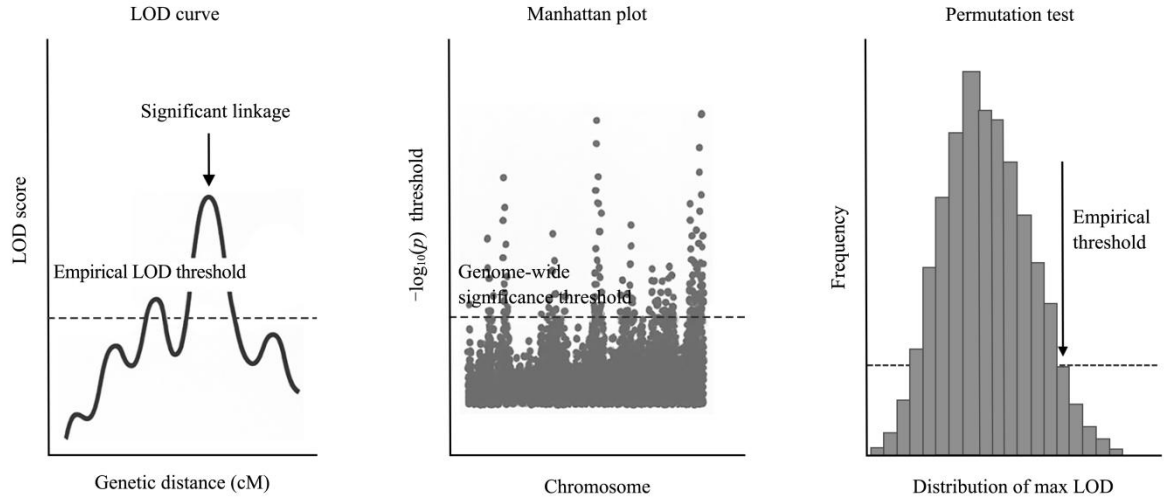

Figure 1 Comparison of statistical representations and significance thresholds in linkage analysis and association studies

Note: Figure 1 shows a schematic diagram of the genome-wide scan statistics and the method for determining the empirical significance threshold; Linkage analysis and genome-wide association studies (GWAS) rely on different test statistics when scanning the genome; In linkage analysis, evidence for linkage is summarized by LOD curves derived from likelihood ratios along genetic distance; In GWAS, association signals are typically visualized using Manhattan plots based on $-\log_{10}(p$ values) from single-marker tests; Although these statistics are not directly comparable, genome-wide significance thresholds for both frameworks are commonly determined empirically using permutation tests; Permutation procedures generate the null distribution of the maximum test statistic across the genome, from which appropriate thresholds are derived to control the family-wise error rate; Significant linkage or association signals are then identified by comparing observed scan statistics to these empirically determined thresholds。It should be noted that p-values are not test statistics themselves but are derived from the null distribution of test statistics; therefore, $-\log_{10}(p)$ values in Manhattan plots are not directly equivalent to LOD scores

1.3 Statistical assumptions, error propagation, and major limitations

Linkage analysis relies on several key statistical assumptions, including Mendelian segregation of alleles, relatively stable recombination rates across meioses, and the absence of strong structural effects that distort the null distribution of test statistics (Meng et al., 2015; Zhang et al., 2015). When these assumptions are violated, biases may propagate through different pathways and ultimately affect mapping results.

For example, segregation distortion (also referred to as transmission distortion) may arise from selection at the gametic or zygotic stage, thereby altering allele frequencies and affecting the estimation of θ. Genotyping errors may introduce spurious recombination events, leading to inflation or compression of genetic maps. Population structure effects may directly distort the likelihood function or the null distribution, resulting in inflated false

positives or underestimated effects (Taniguti et al., 2022; Wang et al., 2024). These sources of bias ultimately converge at the level of LOD curves, manifesting as peak shifts, inflation or deflation of significance, and increased false-positive or false-negative signals (Figure 2).

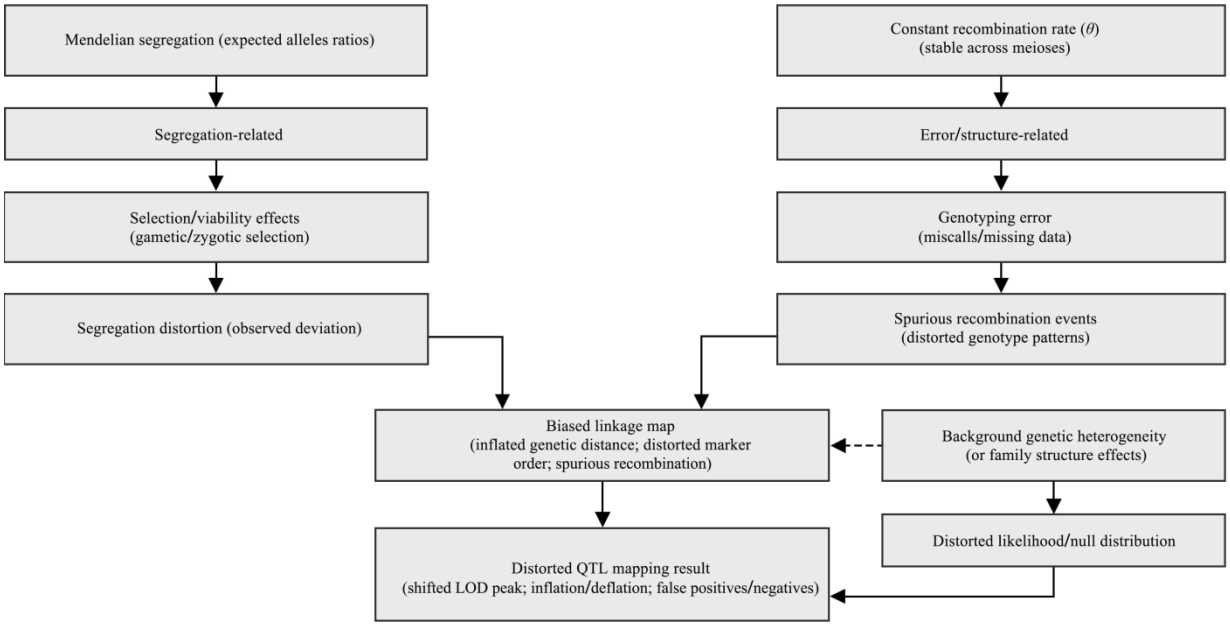

Figure 2 Logical pathways by which linkage-map construction biases propagate into QTL linkage statistics

Note: Figure 2 illustrates how major sources of bias arising during linkage-map construction may propagate into downstream QTL linkage statistics. At the upstream stage, deviations from Mendelian segregation, errors in recombination-rate estimation, and genotyping errors may distort linkage map construction by inflating genetic distances, introducing spurious recombination events, or affecting marker order. These distortions may subsequently propagate into QTL mapping, where they are manifested as peak shifts, inflation or deflation of LOD scores, and increased false-positive or false-negative signals. The figure highlights that some factors act indirectly through linkage-map construction, whereas others may also influence downstream statistical inference more directly

At the methodological level, approaches such as interval mapping (IM), composite interval mapping (CIM), and improved composite interval mapping (ICIM) enhance the power and stability of traditional linkage analysis by jointly estimating QTL positions and effects within the genetic map framework (Meng et al., 2015; Zheng et al., 2019). Nevertheless, the limitations of linkage analysis remain evident. First, mapping resolution is constrained by the number of recombination events within the population, typically remaining at the centimorgan (cM) scale. Second, statistical power depends heavily on population size, phenotypic replication, and population structure. Third, linkage results are population-specific, reflecting only alleles segregating between parental lines, and thus are difficult to generalize to broader genetic backgrounds (Xu et al., 2017; Qu et al., 2020). These limitations are particularly pronounced for complex traits governed by many small-effect loci and gene-environment interactions (Li et al., 2015; Wang et al., 2024).

It is precisely due to these limitations—restricted resolution, strong population dependence, and limited capacity to resolve complex genetic architectures—that research paradigms have gradually shifted toward genome-wide association studies (GWAS) based on natural populations and historical recombination, with the aim of dissecting complex traits at higher resolution and across broader genetic backgrounds.

2 Haseman–Elston Regression and Variance Component Methods

This section discusses Haseman–Elston (HE) regression and variance component methods, not as mainstream approaches for QTL mapping in crops, but from the perspective of statistical genetics development, to illustrate how regression frameworks based on genetic relatedness gradually evolved into mixed linear models and genome-wide analytical approaches.

2.1 Derivation of the Haseman–Elston test

The Haseman–Elston (HE) test, originally proposed by Haseman and Elston in 1972, is a pioneering method for detecting linkage between quantitative traits and genetic markers (Sham and Purcell, 2001; Feingold, 2002). Its fundamental idea is that if an additive QTL is located near a given marker, sibling pairs sharing a higher proportion of identity-by-descent (IBD) alleles at that marker are expected to exhibit more similar phenotypes. Therefore, linkage can be tested by regressing the squared phenotypic difference of sibling (or quasi-sibling) pairs on their IBD sharing proportion at a given marker. The original HE regression model can be expressed as:

$$S=(y_1-y_2)^2=\alpha-\beta\pi+\varepsilon$$

Where, y_1 and y_2 denote the trait values of a sibling pair, and π represents the proportion of IBD sharing at the locus. In the presence of a QTL, the expected slope satisfies $\beta>0$, indicating that higher genetic sharing corresponds to smaller phenotypic differences (Feingold, 2002; Chen, 2014).

A key advantage of the HE method is that it does not rely on the normality of the trait distribution, but only on the monotonic relationship that phenotypic differences decrease with increasing IBD sharing. This property made it an early low-cost and computationally simple genome-wide scanning method (Sham and Purcell, 2001). Subsequent extensions introduced alternative response variables, such as the cross-product or the sib-pair phenotype sum, which can reduce variance and improve statistical power (Wang and Elston, 2005).

The HE framework can also be extended to more complex pedigree structures, including half-sibs, cousins, and general pedigrees. By estimating genome-wide IBD or identity-by-state (IBS) sharing from pedigree or molecular marker data, sliding-window regression can be applied to generate linkage evidence curves along chromosomes (Chen, 2014; Sofer, 2017). This approach has been widely used in early mapping of human genetic diseases and low-cost preliminary screening of quantitative traits in crops. In modern genomics, it has been unified with variance component methods, leading to a family of extensions based on mixed linear models and genomic relationship matrices (GRM) (Liu and Chen, 2022).

2.2 Variance component models and heritability estimation

Variance component models represent an important extension of HE regression, providing a more general and powerful framework for dissecting the genetic basis of quantitative traits. These models are typically formulated within a mixed linear model (MLM) framework:

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

Where, y is the phenotype vector, $X\beta$ represents fixed effects (e.g., environmental covariates), $u\sim N(0,\sigma_g^2\mathrm{K})$ denotes the random genetic effect, with K being the kinship matrix or genomic relationship matrix (GRM), and $\varepsilon\sim N(0,\sigma_e^2\mathrm{I})$ is the residual term. Using restricted maximum likelihood (REML), one can estimate the genetic variance σ_g^2 and residual variance σ_e^2, and subsequently compute heritability:

$$h^2=\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$$

Compared with HE regression, variance component models have the advantage of directly modeling the covariance structure among individuals, thereby linking phenotypic similarity with genetic relatedness (Sofer, 2017; Xu et al., 2021). In fact, previous studies have shown that HE regression can be regarded mathematically as a simplified form of variance component models, indicating an inherent theoretical equivalence between the two approaches (Chen, 2014).

This unified perspective enables variance component models to be applied not only in traditional pedigree-based analyses but also as the theoretical foundation for mixed linear models in large-scale genome-wide association studies (GWAS) (Zhou, 2017; Jeong et al., 2024).

In crop studies, these models are primarily used for heritability estimation and background variance modeling. Variance component approaches have been widely applied to evaluate the heritability of quantitative traits such as yield, plant height, and disease resistance, and can be combined with high-density molecular markers and structured populations to accurately estimate additive genetic variance (Xu et al., 2021; Liu and Chen, 2022). Compared with pedigree-based kinship, GRMs constructed from molecular markers better capture realized genetic relationships and selection history, providing a parametric basis for genomic prediction and genomic selection (GS) (Chen, 2016; Liu and Chen, 2022).

Variance component models are also highly extensible. In multi-environment trials (METs), environmental effects and genotype-by-environment (G×E) interactions can be modeled as random effects, enabling the estimation of across-environment heritability and improving phenotype adjustment and QTL stability (Xu et al., 2021).

2.3 Limitations and methodological transition

Although HE regression and variance component methods play important roles in quantitative genetics, they have several limitations in practice. Their statistical performance strongly depends on sample size, phenotypic replication, and the accurate estimation of IBD sharing (π). When marker density is low, errors in estimating π increase, leading to reduced detection power, broader confidence intervals, and potential bias (Mei and Wang, 2016). It should also be noted that HE regression was originally designed for sibling pairs or simple pedigree structures, and its applicability to complex pedigrees or unrelated populations is limited (Wang and Elston, 2005).

For highly polygenic traits, the variance contribution of any single locus is typically small, making the slope in HE regression or local variance signals difficult to distinguish from background noise (Chen, 2016). The presence of epistasis, non-additive effects, or heteroscedastic phenotypes further violates model assumptions, potentially leading to systematic underestimation of heritability and exacerbating the “missing heritability” problem (Mei and Wang, 2016). These challenges highlight the inherent limitations of HE and variance component methods in modeling complex genetic architectures.

With the expansion of genomic data and the development of high-density marker platforms, research has gradually shifted toward a framework characterized by “genome-wide coverage + fixed-effect testing + random-effect modeling of background variation”. This framework is exemplified by mixed linear models in GWAS (GWAS-MLM), where candidate SNPs are treated as fixed effects, while the genomic relationship matrix (GRM) captures polygenic background effects and cryptic relatedness (i.e., unobserved but real genetic relationships among individuals), thereby correcting for population structure and relatedness (Zhou, 2017; Xu et al., 2021).

New statistical approaches, such as LD score regression and variance component estimation based on summary statistics, have further enabled efficient implementation in large-scale populations and across studies (Jeong et al., 2024). These developments mark a natural transition from variance decomposition frameworks centered on HE and variance components to GWAS-based prioritization of causal genomic regions, representing a paradigm shift in methodology.

This theoretical unification has laid the foundation for GWAS frameworks based on mixed linear models. The next section will further discuss how genome-wide association analysis has developed into a dominant approach within this statistical framework.

3 The Rise and Decline of Candidate Gene Approaches

3.1 Rationale and advantages of candidate gene selection

The candidate gene strategy emerged as a natural extension of advances in molecular biology and quantitative genetics. Its core logic is to start from explicit biological priors and prioritize genes and functional loci that are likely to influence target traits, based on pathway knowledge, mutant phenotypes, cross-species homology, differential expression, or metabolic evidence (Zhu and Zhao, 2007). This hypothesis-driven approach allows researchers to focus on genes that are theoretically linked to the phenotype of interest, such as key enzymes in

metabolic pathways, transcription factors, or structural proteins. In plant genetics, candidate genes have often been prioritized due to their known roles in critical agronomic processes such as starch biosynthesis, disease resistance, and stress responses (Stanton-Geddes et al., 2013; Raj and Nadarajah, 2022).

The typical workflow of this approach follows a sequence: "mechanistic hypothesis → candidate gene and marker selection → association testing in target populations → functional interpretation". Compared with genome-wide, hypothesis-free scans, the candidate gene strategy reduces the dimensionality of the search space, thereby lowering the requirements for sample size and genotyping costs. This makes it particularly suitable for resource-limited settings or for rapid validation of specific biological mechanisms (Zhu and Zhao, 2007). Before the widespread adoption of high-throughput genotyping technologies, this approach played a crucial role and was a practical choice in genetic studies of crops and model organisms.

A major advantage of the candidate gene approach lies in its specificity and interpretability. When a candidate variant is consistent with biochemical function, cellular pathways, or mutant phenotypes, statistical association can be directly linked to causal inference, forming a coherent evidence chain (Zhu and Zhao, 2007; Raj and Nadarajah, 2022). The technical requirements are relatively modest: validation can be achieved using low-throughput genotyping platforms such as KASP or TaqMan, combined with conventional field phenotyping. This facilitates replication across multiple environments or generations and enables rapid integration into breeding programs, particularly for marker-assisted selection (MAS) (Ibrahim et al., 2020; Kushanov et al., 2021).

3.2 Methodological biases and the reproducibility crisis

Although the candidate gene approach provided an early low-cost, hypothesis-driven framework for genetic analysis, its methodological characteristics also introduced risks of systematic bias and reproducibility challenges. While reliance on prior knowledge reduces the search space, it also leads to repeated testing of a limited set of genes, resulting in the overrepresentation of a few "hotspot genes" and neglect of broader genomic variation (Zhu and Zhao, 2007; Baxter, 2020). This focus increases detection efficiency but also amplifies publication bias and the "winner's curse", where significant results are more likely to be reported, while negative findings are underrepresented, leading to the accumulation of false or inflated associations in the literature (Baxter, 2020).

Candidate gene studies also suffer from limited statistical power. Due to small sample sizes and the lack of modeling for correlations among loci, significance levels often fail to reflect genome-wide error rates (Zhu and Zhao, 2007; Stanton-Geddes et al., 2013). When multiple loci or traits are tested simultaneously without rigorous correction for multiple comparisons, false-positive rates are systematically underestimated, further compromising the reliability of conclusions.

Population structure and cryptic relatedness are major sources of spurious associations. In crop studies, candidate genes are often tested in structured germplasm panels; without proper correction using principal component analysis (PCA) or mixed linear models (MLM), differences in allele frequencies alone can produce significant signals in the absence of causal relationships (Stanton-Geddes et al., 2013). In addition, linkage disequilibrium (LD)-induced mismatches between markers and causal variants, phenotyping errors, and gene–environment interactions (G×E) further reduce the reproducibility of candidate gene associations across populations and environments (Table 1) (Zhu and Zhao, 2007).

When researchers attempt to mitigate these biases through stricter population structure correction, genome-wide background modeling, and multiple testing control, the analytical framework gradually converges toward that of GWAS (Baxter, 2020). This transition, in practice, diminishes the original advantages of the candidate gene approach—namely low cost and rapid validation—and explains its decline in the era of high-throughput genomics.

Table 1 Major sources of bias in candidate gene studies and their impacts on reproducibility

Bias type	Description	Impact on reproducibility
Publication bias	Positive findings are more likely to be published, while negative results are underreported	Accumulation of false associations and literature bias
Winner's curse	Significant effects often diminish in replication studies	Results are difficult to replicate in independent populations
Small sample size	Low detection power, making it difficult to identify loci with small effects	Insufficient statistical power, unstable results
Population stratification	Differences in allele frequency lead to spurious associations	Increased false positive rate, distorted signals across populations
LD mismatch	Mismatch between marker and causal locus	Incorrect interpretation of associations, difficulty in pinpointing causal genes
Phenotyping error	Noise masks the true effect and reduces signal strength	Inconsistencies across experimental results
Gene-environment interaction (G×E)	Effect size is unstable across environmental changes	Reduced reproducibility across environments

3.3 Practical contributions and limitations in plant breeding

Despite its gradual decline, the candidate gene approach has made landmark contributions to the identification and utilization of major-effect genes in plant breeding. A classic example is the Wx gene in rice, which controls amylose content and is closely associated with the biochemical pathway of starch synthesis. Allelic variation at this locus was rapidly identified and successfully applied in marker-assisted selection (MAS) to improve grain quality (Pflieger et al., 2001; Raj and Nadarajah, 2022). Another well-known case is the loss-of-function allele of BADH2, associated with fragrance traits in rice, which has been validated and widely utilized across multiple rice populations, becoming a hallmark example of candidate gene–based breeding. In crops such as cotton and wheat, candidate gene approaches have also been used to identify and deploy major genes conferring disease resistance and stress tolerance, contributing to crop improvement (Kushanov et al., 2021; Raj and Nadarajah, 2022).

However, agronomic traits such as yield and stress tolerance are typically governed by polygenic architectures with small-effect loci and gene–environment interactions (G×E), resulting in complex and dynamic genetic systems. The prior knowledge required for candidate gene selection is often insufficient to capture the full spectrum of allelic variation within regulatory networks, leading to low hit rates and limited explanatory power for complex traits (Pflieger et al., 2001; Zhu and Zhao, 2007). Moreover, differences in linkage disequilibrium (LD) patterns across ecotypes and breeding backgrounds further hinder the transferability and validation of candidate gene associations across populations (Stanton-Geddes et al., 2013).

Within modern genomic frameworks, the role of the candidate gene approach has shifted from a "discovery engine" to a "validation module". Specifically, it is now more suitable for prioritizing functional genes within genomic regions identified by GWAS or linkage mapping, followed by causal validation through eQTL analysis, fine mapping, and functional experiments. It also remains valuable for hypothesis-driven validation and targeted improvement, including rapid applications of gene editing technologies (Ibrahim et al., 2020). This transition has led to a complementary and synergistic relationship between candidate gene strategies, GWAS, and genomic selection (GS): GWAS enables large-scale, hypothesis-light discovery, while candidate gene approaches provide targeted functional validation, together advancing modern crop genetic improvement.

4 From Traditional Approaches to GWAS: A Paradigm Transition

Traditional linkage analysis and candidate gene approaches have achieved important progress within controlled populations and limited hypothesis-driven frameworks. However, their methodological boundaries have gradually become evident. Limitations in mapping resolution, statistical power, and cross-population generalizability are not isolated issues, but rather point to a common underlying challenge: under polygenic architectures characterized by numerous small-effect loci and complex population structures, local or hypothesis-driven analytical frameworks are insufficient to capture the global landscape of genetic variation. This realization has directly driven a paradigm shift toward systematic, genome-wide scanning approaches.

4.1 Innovations and limitations of GWAS (Table 2)

Genome-wide association studies (GWAS) represent a major paradigm shift in quantitative genetics. They move beyond the strong reliance on prior hypotheses inherent in linkage analysis and candidate gene strategies, instead adopting a genome-wide scanning framework with weak locus-specific assumptions. In large populations, GWAS systematically evaluates statistical associations between phenotypes and thousands to millions of single nucleotide polymorphisms (SNPs) (Tibbs Cortes et al., 2021; Ashwath et al., 2023; Bashir et al., 2024).

Table 2 Comparison among linkage mapping, candidate gene approaches, and genome-wide association studies (GWAS)

Dimension	Linkage mapping	Candidate gene approach	GWAS
Research paradigm	Genome-wide scan within linkage map, no locus-specific prior	Strong prior hypothesis based on known genes/pathways	Genome-wide scan, weak locus-specific prior
Study objective	Mapping QTL regions	Testing specific candidate genes	Genome-wide identification of trait-associated variants
Marker density	Low to moderate marker density	Low (targeted regions)	High-density SNP markers
Mapping resolution	Centimorgan-level resolution	Gene/locus-level resolution	Kilobase-level resolution
Search space	Linkage map scale	Preselected genes/regions	Whole genome
Population type	Biparental populations (F_2, RIL, DH, etc.)	Natural or selected populations	Diverse natural or breeding panels
Population structure	Simple or controlled	May involve population structure	Often complex, with stratification and relatedness
Primary statistic	LOD scores (linkage evidence)	t-test, regression, or χ^2 statistics	p-values or $-\log_{10}(p)$
Multiple testing burden	Relatively low	Relatively low	Substantial
Significance threshold	Permutation-based or empirical thresholds	Standard statistical thresholds	Bonferroni, FDR, or empirical thresholds
Main challenges	Low resolution, broad intervals	Dependence on prior hypotheses, risk of missing true genes	Population structure, LD, multiple testing
Common corrections	Limited correction strategies	Limited correction strategies	PCA, MLM/GRM, QC, etc.
Typical outcomes	Broad QTL intervals	Evidence for candidate genes	High-resolution association signals
Reproducibility	Dependent on population design	Dependent on hypotheses and validation	Dependent on sample size and models
Typical applications	Controlled systems with clear genetic backgrounds	Functional validation and hypothesis testing	Genome-wide analysis of complex traits

Note: Although linkage mapping and candidate gene approaches are both considered traditional strategies, they differ substantially in statistical framework and reliance on prior hypotheses

This transition has been enabled by advances in high-throughput genotyping technologies, including SNP arrays and next-generation sequencing, which allow high-density genotyping and genotype imputation. Within a linkage disequilibrium (LD) framework, these developments have increased mapping resolution to the kilobase (kb) level and, in some cases, to individual genes (Alqudah et al., 2020; Ashwath et al., 2023). Standardized quality control procedures (e.g., marker missingness thresholds and allele frequency filters), together with anchoring to reference genomes, have further improved comparability and reproducibility across platforms and populations (Bashir et al., 2024; Chang-Brahim et al., 2024).

From a statistical perspective, key advances in GWAS include the explicit modeling of population structure and relatedness, as well as the introduction of multiple error control strategies under large-scale multiple testing scenarios (Chang-Brahim et al., 2024; Nandi et al., 2024). For example, principal component analysis (PCA) is used to correct for population stratification, while mixed linear models (MLM) combined with genomic relationship matrices (GRM) account for relatedness and background genetic covariance, forming the current standard analytical framework (Tibbs Cortes et al., 2021; Susmitha et al., 2023). It should be noted that traditional linkage analysis achieves effective genome-wide error control through permutation testing, whereas in GWAS, the

choice of significance thresholds remains subject to methodological trade-offs, such as the conservativeness of Bonferroni correction and the instability of empirical thresholds.

Despite its advantages in resolution and genome-wide coverage, GWAS also has notable limitations. First, GWAS relies heavily on LD structure, meaning that significant signals often reflect LD blocks rather than causal variants, increasing uncertainty in interpretation. Second, GWAS has limited power to detect rare variants and structural variants, potentially underestimating their contributions. Third, residual confounding from population structure and cryptic relatedness may persist even after PCA and mixed-model correction. In addition, multiple testing correction involves a trade-off between controlling false positives and maintaining statistical power, and no universally optimal solution exists. Therefore, GWAS findings typically require validation through linkage analysis, fine mapping, and functional studies, highlighting the complementary nature of different approaches.

4.2 Implications for plant breeding

The emergence of GWAS has opened a new era for dissecting the genetic basis of complex traits in plants, with significant implications for modern breeding. Unlike traditional linkage mapping and candidate gene approaches, GWAS leverages natural populations and historical recombination events to achieve high-resolution mapping of quantitative trait nucleotides (QTNs), supported by large sample sizes and high-density genotyping. This enables the identification of robust and reproducible association signals for key agronomic traits such as yield, stress tolerance, and quality (Alqudah et al., 2020; Tibbs Cortes et al., 2021; Ashwath et al., 2023).

By integrating GWAS with multi-omics data, including expression quantitative trait loci (eQTL), transcriptomics, and metabolomics, researchers can accelerate the prioritization and validation of candidate genes and functional variants. These advances also provide reliable molecular markers for marker-assisted selection (MAS) and genomic selection (GS) (Bashir et al., 2024; Chang-Brahim et al., 2024). In crops such as rice, maize, and wheat, GWAS has been widely applied to dissect yield and drought-related traits, with validated loci directly translated into breeding markers (He et al., 2017; Nandi et al., 2024).

The integration of GWAS with variance component frameworks further enables seamless linkage with genomic selection (GS). Genome-wide markers can be used to construct genomic relationship matrices (GRM), which are then incorporated into G-BLUP or Bayesian prediction models to estimate genomic estimated breeding values (GEBVs), supporting cross-environment prediction and selection (He et al., 2017; Tibbs Cortes et al., 2021). GWAS results can also be used as weighted priors or feature subsets to improve prediction accuracy and guide optimal crossing strategies, thereby accelerating genetic gain (Susmitha et al., 2023).

Therefore, traditional linkage analysis, candidate gene approaches, and GWAS should not be viewed as mutually exclusive, but rather as complementary methods operating at different scales and under different modeling assumptions. Together, they form a multi-layered analytical framework for the genetic dissection of complex traits.

5 Discussion

The interpretation of GWAS findings has been strongly influenced by earlier frameworks from linkage analysis and candidate gene studies. The traditional “peak–interval–candidate” reasoning paradigm persists, where researchers often treat the sentinel SNP (i.e., the most statistically significant marker within an associated region) or nearby genes as causal variants. This practice overlooks allelic heterogeneity within linkage disequilibrium (LD) blocks, distal regulatory effects, and polygenic contributions (Gallagher and Chen-Plotkin, 2018; Tam et al., 2019; Uffelmann et al., 2021). Such a “linkage-driven mindset” leads to an overemphasis on single genes, while underestimating the roles of regulatory variation, structural variants, and chromatin architecture. Lessons from the candidate gene era further caution that functional annotation and statistical significance are not equivalent to causality; excessive reliance on prior knowledge may introduce publication bias and limit reproducibility across populations (Gallagher and Chen-Plotkin, 2018; Cano-Gamez and Trynka, 2020). Therefore, GWAS interpretation should be upgraded to an “evidence triangulation” framework: starting from credible sets, integrating eQTL/metabolite QTL colocalization, haplotype structure, cross-population replication, and functional validation.

In both human and crop studies, complex traits face shared challenges: high polygenicity, difficulty in detecting small-effect loci, susceptibility to spurious associations due to population structure and LD heterogeneity, and reduced transferability caused by environmental variation and phenotypic noise (Tam et al., 2019; Tibbs Cortes et al., 2021). Cross-population or cross-ecotype generalization is particularly challenging: polygenic risk scores (PRS) often show reduced predictive performance across ancestral groups, and GWAS signals in crops frequently fail to replicate across ecological or breeding backgrounds (Marigorta et al., 2018; Uffelmann et al., 2021; Abdellaoui et al., 2023). Rare variants and structural variants remain underrepresented in both domains, contributing to what is often referred to as the "genetic dark matter".

Addressing these challenges requires a coordinated evolution of sample size, population design, and statistical methodology. While large sample sizes improve statistical power, their marginal benefit diminishes without well-designed populations (e.g., NAM, MAGIC, multi-environment trials) and advanced analytical frameworks such as mixed linear models, Bayesian fine-mapping, and cross-cohort meta-analysis (Marigorta et al., 2018; Tibbs Cortes et al., 2021). Best practices therefore include incorporating power and bias simulations at the study design stage, standardizing quality control procedures and data dictionaries (i.e., unified definitions, coding schemes, and metadata standards), adopting reproducible analytical pipelines, and sharing summary statistics to ensure continuity across the "discovery–replication–validation" process (Uffelmann et al., 2021; Abdellaoui et al., 2023).

Traditional approaches are not obsolete but remain valuable in specific contexts. When targeting loci with large effects or rare alleles, strategies such as extreme phenotype sampling, selective genotyping, near-isogenic lines, and bulked segregant analysis (BSA) can enable rapid and high-confidence localization (Marigorta et al., 2018; Tibbs Cortes et al., 2021). Advances in high-coverage sequencing and long-read technologies also facilitate the resolution of complex structural variants in pedigrees or closely related materials, providing directional evidence for refining GWAS candidate regions.

Future directions point toward the integration of GWAS with post-GWAS methodologies. Bayesian fine-mapping, combined with functional priors, enables the identification of causal variants; polygenic risk scores (PRS) and genomic selection (GS) are converging conceptually, both relying on GRM-based modeling and cross-population calibration; and machine learning is advancing multi-omics integration, feature compression, and nonlinear modeling, enabling a closed-loop framework of "discovery-prediction-intervention." The joint evolution of these approaches in human and crop research holds promise for achieving precision intervention and controllable genetic improvement.

6 Conclusions

The evolution of statistical genetics—from linkage analysis, Haseman–Elston regression and variance component methods, to candidate gene studies and ultimately GWAS—reflects a transition from "locally hypothesis-driven" approaches to "genome-wide, hypothesis-light exploration". This progression represents not only methodological advancement but also a deepening integration of theory and practice. Traditional approaches are not obsolete stages of history, but foundational components of modern frameworks: the use of recombination and segregation in pedigrees, as well as the modeling of relatedness and identity-by-descent (IBD), directly informed the development of genomic relationship matrices (GRM) and mixed linear models (MLM); similarly, multiple testing correction and population structure adjustment have been systematically incorporated into GWAS.

A clear understanding of the statistical assumptions and limitations of these methods is essential for the proper interpretation of GWAS results. Assumptions underlying segregation and recombination in pedigree-based analysis, the prior-driven selection logic in candidate gene studies, and the treatment of linkage disequilibrium (LD) and allele frequency differences all reappear in modified forms within GWAS. Ignoring population structure or the effective number of independent tests can lead to inflated significance, while equating sentinel SNPs directly with causal variants risks overinterpretation. Therefore, an ideal analytical pathway requires the integration of statistical evidence, functional validation, and experimental confirmation, along with explicit quantification and reporting of uncertainty.

In plant breeding, these insights translate into practical design principles: incorporating power simulations and sample size planning at the study design stage; prioritizing populations that increase recombination and allelic diversity (e.g., NAM, MAGIC, and multi-environment trials); and controlling structural effects through standardized quality control and mixed-model frameworks. Within the "discovery–replication–validation" cycle, establishing cross-population transferability and predictive performance is essential for translating statistical signals into reliable tools for selection.

Authors' contributions

Fang Xuanjun and Wu Weiren conducted this study, including literature review, data analysis, and the drafting and revision of the manuscript. Both authors have read and approved the final version of the manuscript.

Acknowledgements

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

References

Abdellaoui A., Yengo L., Verweij K., and Visscher P., 2023, 15 years of GWAS discovery: Realizing the promise, American Journal of Human Genetics, 110(2): 179-194.

Alqudah A.M., Sallam A., Baenziger P.S., and Börner A., 2020, GWAS: fast-forwarding gene identification and characterization in temperate cereals: lessons from barley-a review, J. Adv. Res., 22: 119-135.

Ashwath M., Lavale S., Santhoshkumar A., Mohapatra S., Bhardwaj A., Dash U., Shiran K., Samantara K., and Wani S., 2023, Genome-wide association studies: an intuitive solution for SNP identification and gene mapping in trees, Funct. Integr. Genomics, 23(4): 297.

Bashir L., Mehmood A., Manzoor S., Thendral U., Yadav J., Saha S., Yadav M., Meena D., and Padder U., 2024, A comprehensive review on GWAS: basic concepts and role in agriculture, Int. J. Agric. Ext. Soc. Dev., 7(10): 302-314.

Baxter I., 2020, We aren't good at picking candidate genes, and it's slowing us down, Curr. Opin. Plant Biol., 54: 57-60.

Cano-Gamez E., and Trynka G., 2020, From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases, Front. Genet., 11: 424.

Chang-Brahim I., Koppensteiner L., Beltrame L., Bodner G., Saranti A., Salzinger J., Fanta-Jende P., Sulzbachner C., Bruckmüller F., Trognitz F., Samad-Zamini M., Zechner E., Holzinger A., and Molin E., 2024, Reviewing the essential roles of remote phenotyping, GWAS and explainable AI in practical marker-assisted selection for drought-tolerant winter wheat breeding, Front. Plant Sci., 15: 1319938.

Chen G.B., 2014, Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression, Front. Genet., 5: 107.

Chen G.B., 2016, On the reconciliation of missing heritability for genome-wide association studies, Eur. J. Hum. Genet., 24(12): 1810-1816.

Feingold E., 2002, Regression-based quantitative-trait-locus mapping in the 21st century, Am. J. Hum. Genet., 71(2): 217-222.

Gallagher M.D., and Chen-Plotkin A.S., 2018, The post-GWAS era: From association to function, Am. J. Hum. Genet., 102(5): 717-730.

He J., Meng S., Zhao T., Xing G., Yang S., Li Y., Guan R., Lu J., Wang Y., Xia Q., Yang B., and Gai J., 2017, An innovative procedure of genome-wide association analysis fits studies on germplasm population and plant breeding, Theor. Appl. Genet., 130(12): 2327-2343.

Ibrahim A., Zhang L., Niyitanga S., Afzal M., Xu Y., Zhang L., Zhang L., and Qi J., 2020, Principles and approaches of association mapping in plant breeding, Trop. Plant Biol., 13(1): 212-224.

Jeong M., Pazokitoroudi A., Liu Z., and Sankararaman S., 2024, Scalable summary-statistics-based heritability estimation method with individual genotype level accuracy, Genome Res., 34(7): 1286-1293.

Kushanov F., Turaev O., Ernazarova D., Gapparov B., Oripova B., Kudratova M., Rafieva F., Khalikov K., Erjigitov D., Khidirov M., Kholova M., Khusenov N., Amanboyeva R., Saha S., Yu J., and Abdurakhmonov I., 2021, Genetic diversity, QTL mapping, and marker-assisted selection technology in cotton (*Gossypium* spp.), Front. Plant Sci., 12: 779386.

Li C., Li Y., Bradbury P., Wu X., Shi Y., Song Y., Zhang D., Rodgers-Melnick E., Buckler E., Zhang Z., Li Y., and Wang T., 2015, Construction of high-quality recombination maps with low-coverage genomic sequencing for joint linkage analysis in maize, BMC Biol., 13: 78.

Liu H., and Chen G., 2022, A novel genomic prediction method combining randomized Haseman-Elston regression with a modified algorithm for Proven and Young for large genomic data, Crop J., 10(2): 550-554.

Marigorta U.M., Rodríguez J.A., Gibson G., and Navarro A., 2018, Replicability and prediction: Lessons and challenges from GWAS, Trends Genet., 34(7): 504-517.

Mei B., and Wang Z., 2016, An efficient method to handle the 'large *p*, small *n*'problem for genomewide association studies using Haseman-Elston regression, J. Genet., 95 (4): 847-852.

Meng L., Li H., Zhang L., and Wang J., 2015, QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations, Crop J., 3(3): 269-283.

Nandi S., Varotariya K., Luhana S., Kyada A., Saha A., Roy N., Sharma N., and Rambabu D., 2024, GWAS for identification of genomic regions and candidate genes in vegetable crops, Funct. Integr. Genomics, 24(6): 203.

Pasaniuc B., and Price A.L., 2016, Dissecting the genetics of complex traits using summary association statistics, Nat. Rev. Genet., 18(2): 117-127.

Pflieger S., Lefebvre V., and Causse M., 2001, The candidate gene approach in plant genetics: A review, Mol. Breed., 7(4): 275-291.

Qu P., Shi J., Chen T., Chen K., Shen C., Wang J., Zhao X., Ye G., Xu J., and Zhang L., 2020, Construction and integration of genetic linkage maps from three multi-parent advanced generation inter-cross populations in rice, Rice, 13(1): 13.

Raj S., and Nadarajah K., 2022, QTL and candidate genes: Techniques and advancement in abiotic stress resistance breeding of major cereals, Int. J. Mol. Sci., 24(1): 6.

Sebastiani P., Timofeev N., Dworkis D., Perls T.T., and Steinberg M.H., 2009, Genome-wide association studies and the genetic dissection of complex traits, Am. J. Hematol., 84(8): 504-515.

Sham P.C., and Purcell S., 2001, Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs, Am. J. Hum. Genet., 68(6): 1527-1532.

Sofer T., 2017, Confidence intervals for heritability via Haseman-Elston regression, Stat. Appl. Genet. Mol. Biol., 16(4): 259-273.

Stanton-Geddes J., Paape T., Epstein B., Briskine R., Yoder J., Mudge J., Bharti A., Farmer A., Zhou P., Denny R., May G., Erlandson S., Yakub M., Sugawara M., Sadowsky M., Young N., and Tiffin P., 2013, Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*, PLoS ONE, 8 (6): e65688.

Susmitha P., Kumar P., Yadav P., Sahoo S., Kaur G., Pandey M., Singh V., Tseng T., and Gangurde S., 2023, Genome-wide association study as a powerful tool for dissecting competitive traits in legumes, Front. Plant Sci., 14: 1123631.

Tam V., Patel N., Turcotte M., Bossé Y., Paré G., and Meyre D., 2019, Benefits and limitations of genome-wide association studies, Nat. Rev. Genet., 20(8): 467-484.

Taniguti C., Taniguti L., Amadeu R., Lau J., De Siqueira Gesteira G., De Paula Oliveira T., et al., 2022, Developing best practices for genotyping-by-sequencing analysis in the construction of linkage maps, GigaScience, 12(1): giad092.

Tibbs Cortes L., Zhang Z., and Yu J., 2021, Status and prospects of genome-wide association studies in plants, The plant genome, 14(1): e20077.

Uffelmann E., Huang Q.Q., Munung N.S., De Vries J., Okada Y., Martin A.R., Martin H.C., Lappalainen T., and Posthuma D., 2021, Genome-wide association studies, Nat. Rev. Methods Primers, 1(1): 1-21.

Wang T., and Elston R.C., 2005, Two-level Haseman-Elston regression for general pedigree data analysis, Genet. Epidemiol., 29(1): 12-22.

Wang X., Wang J., Xia X., Xu X., Li L., Cao S., Hao Y., and Zhang L., 2024, Effect of genotyping errors on linkage map construction based on repeated chip analysis of two recombinant inbred line populations in wheat (*Triticum aestivum* L.), BMC Plant Biol., 24(1): 306.

Watanabe K., Stringer S., Frei O., Mirkov U.G., de Leeuw C.A., Polderman T.J., et al., 2019, A global overview of pleiotropy and genetic architecture in complex traits, Nat. Genet., 51(9): 1339-1348.

Xu T., Qi G., Zhu J., Xu H., and Chen G., 2021, Subsampling technique to estimate variance component for UK-Biobank traits, Front. Genet., 12: 612045.

Xu Y., Li P., Yang Z., and Xu C., 2017, Genetic mapping of quantitative trait loci in crops, Crop J., 5(2): 175-184.

Zhang L., Li H., and Wang J., 2015, Linkage analysis and map construction in genetic populations of clonal F_1 and double cross, G3 (Bethesda), 5(3): 427-439.

Zhang Y., Jia Z., and Dunwell J.M., 2019, The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits, Front. Plant Sci., 10: 100.

Zhang Y., Wang M., Li Z., Yang X., Li K., Xie A., Dong F., Wang S.H., Yan J.B., and Liu J., 2024, An overview of detecting gene-trait associations by integrating GWAS summary statistics and eQTLs, Sci. China Life Sci., 67 (6): 1133-1154.

Zheng C., Boer M., and van Eeuwijk F., 2019, Construction of genetic linkage maps in multiparental populations, Genetics, 212(4): 1031-1044.

Zhou X., 2017, A unified framework for variance component estimation with summary statistics in genome-wide association studies, Ann. Appl. Stat., 11(4): 2027-2051.

Zhu M., and Zhao S., 2007, Candidate gene identification approach: Progress and challenges, Int. J. Biol. Sci., 3(7): 420-427.

Fang X.J., Wu W.R., and Tang J.L. (eds.), 2001, Crop DNA marker-assisted breeding, Science Press, Beijing, China, pp. 1-84.

Appendix 1 Standardized terminology and symbols in statistical genetics

Terminology in English	Symbol/Abbreviation	Brief definition	Common misuse notes
LOD score	LOD	$\log_{10}\{L(\theta)/L(0.5)\}$, quantifies evidence for linkage vs. no linkage	LOD ≠ $-\log_{10}(p)$ (only approximately related)
Identity by descent	IBD	Probability that alleles are inherited from a common ancestor	IBD ≠ "inbreeding"
Identity by state	IBS	Alleles identical in sequence but not necessarily from a common ancestor	IBS ≠ IBD
Principal components analysis	PCA	Dimensionality reduction of genotype matrix to capture population structure	PCA alone is insufficient to account for close relatedness
Genomic relationship matrix	GRM	Kinship matrix (K) constructed from genome-wide markers	Requires standardized genotype coding
Mixed linear model	MLM	Framework including fixed and random effects (often with GRM)	Variance components and λ_{GC} should be reported
False discovery rate	FDR	Expected proportion of false positives (e.g., BH procedure)	FDR ≠ FWER
Bonferroni correction	-	Controls FWER by α/m	Overly conservative under strong LD
Effective tests	m_{eff}	Number of independent tests accounting for LD	Should be estimated via spectral decomposition
Narrow-sense heritability	h^2	$\sigma_g^2/(\sigma_g^2+\sigma_e^2)$	May be biased by environmental stratification
Linkage disequilibrium	LD	Non-random association between alleles	Report r^2 and genomic distance
Quantitative trait locus/quantitative trait nucleotide	QTL/QTN	Locus/variant affecting quantitative traits	QTL ≠ causal variant
Permutation test	-	Permutes phenotype or genotype to derive significance thresholds and control FWER	Must preserve study design (e.g., stratification)
Genomic control	λ_{GC}	Median(χ^2) / 0.456 to assess inflation	Diagnostic only, not sufficient correction
Recombinant inbred lines	RILs	Lines derived from repeated selfing and recombination fixation	-
Backcross population	BC	Population generated by backcrossing to a parent	-
Doubled haploids	DH	Completely homozygous lines produced by chromosome doubling	-
Nested association mapping	NAM	Multi-family design with diverse parents crossed to a common parent	-
Multiparent advanced generation inter-cross	MAGIC	Multi-parent intercross populations with high recombination density	-
Genomic selection	GS	Prediction of breeding values using genome-wide markers	Complementary to GWAS
Best linear unbiased prediction	BLUP	Prediction of individual effects under MLM	-
Bulked segregant analysis	BSA	Pooling extreme phenotypes to compare allele frequencies	Coarse-mapping tool

Appendix 2 Simplified relationships and key computational considerations of common statistical measures (for conceptual understanding)

To facilitate readers' understanding of core concepts in statistical genetics and the quantitative relationships among different methods discussed in the main text, this appendix summarizes simplified formulas and intuitive derivations of several commonly used statistical measures. These expressions are intended to illustrate the internal connections among linkage analysis, variance component models, and GWAS statistics, rather than to provide

complete or rigorous mathematical proofs. The derivations are based on standard assumptions, and their scope and limitations have been discussed in the main text. Readers are encouraged to consult original references and consider specific data characteristics in practical applications.

S2-1: The relationship between LOD scores, genetic distance, and chi-square statistics (conceptual level)

Definition:

The LOD score is defined as:

$$LOD=\log_{10}\{L(\vartheta)/L(0.5)\}$$

Where, $L(\vartheta)$ and $L(0.5)$ denote the likelihood under a given recombination rate and the null hypothesis of no linkage, respectively.

Approximate relationship:

Under standard regularity conditions, the LOD score can be approximately related to the chi-square statistic:

$$\chi_{df}^2 \approx 2\ln(10)\cdot LOD,$$

The degrees of freedom (df) depend on the specific model settings.

Conversion of genetic distance:

Common functions for converting recombination rate to genetic distance include:

Haldane mapping function: $d=-\frac{1}{2}\ln(1-2\theta)$ (in cM);

Kosambi mapping function: $d=\frac{1}{4}\ln\{(1+2\theta)/(1-2\theta)\}$

Note:

Significance thresholds for LOD scores should generally be determined using empirical methods such as permutation testing to control the family-wise error rate (FWER) at the genome-wide level.

S2-2: Core concepts and statistical interpretation of Haseman–Elston (HE) regression

Classical form:

HE regression models the squared phenotypic difference of sibling pairs:

$$S=(y_1-y_2)^2$$

as a function of their IBD sharing proportion π:

$$S=\alpha-\beta\pi+\varepsilon_S$$

Where, β is proportional to the additive genetic variance at the locus or genomic region of interest.

Modified form:

An alternative formulation uses the cross-product:

$$C=(y_1-\bar{y})(y_2-\bar{y})$$

Regression on π:

$$C=\gamma+\delta\pi+\eta$$

Under linkage, $\delta > 0$ is typically expected.

Note:

Estimation errors in IBD sharing and sparse marker density may reduce statistical power. In practice, sliding-window or local smoothing strategies can be applied to mitigate these effects.

S2-3: Interpretation of genetic variance and heritability in mixed linear models

Model formulation:

The mixed linear model is typically expressed as:

$$y=X\beta+Zu+\epsilon,$$

Where,

$$u\sim N(0, \sigma_g^2 K), \epsilon\sim N(0, \sigma_e^2 I).$$

Heritability estimation:

Narrow-sense heritability is defined as:

$$h^2=\frac{\sigma_g^2}{(\sigma_g^2+\sigma_e^2)},$$

and is commonly estimated using restricted maximum likelihood (REML).

GRM construction (conceptual):

The genomic relationship matrix K is typically constructed from SNP genotypes that are centered and scaled according to allele frequencies.

Note:

Heritability estimates should be reported with standard errors, and it should be specified whether environmental stratification or batch effects are included in the model.

S2-4: Handling significance thresholds under multiple testing

Bonferroni correction:

$$\alpha^* = \alpha/m.$$

Benjamini–Hochberg (BH) FDR control:

For ordered p-values $P_{(i)}$, identify the largest k such that:

$$P_{(k)}\leq k\alpha/m$$

and declare the first k tests significant.

Effective number of tests:

Under linkage disequilibrium (LD), the effective number of independent tests (m_{eff}) can be estimated using spectral decomposition.

Note:

Under strong LD, replacing the nominal number of tests m with m_{eff} is often more appropriate.

S2-5: Diagnostic interpretation of genomic control and QQ plots

Genomic control factor:

λ_{GC} = median (χ^2) / 0.456 (1 df).

Diagnostic considerations:

QQ plots should ideally include 95% confidence bands. After appropriate correction, λ_{GC} is expected to be close to 1.

S2-6: Intuitive understanding of statistical power and non-central parameters (quantitative traits)

Single-SNP case:

$$NCP \approx n \bullet 2p(1-p)\beta^2/\sigma^2 .$$

Under LD between marker and causal variant:

$$NCP_{tag} \approx r^2 \times NCP_{causal} .$$

Note:
Statistical power is highly sensitive to sample size and LD strength (r^2). Interpretation of association results should explicitly consider LD between markers and underlying causal variants.

S2-7: Workflow for obtaining empirical significance thresholds based on permutation tests (overview)

Typical steps include:
(1) Fix genotypes and randomly permute phenotypes (or permute within strata);
(2) Record the maximum test statistic across the genome for each permutation;
(3) Use the 1-α quantile of the empirical distribution as the significance threshold;
(4) Typically, at least 1 000 permutations are required for stable estimation.

S2-8: Implementation considerations for PCA-based correction

After centering and allele-frequency scaling of the genotype matrix, singular value decomposition (SVD) or eigenvalue decomposition can be applied to G or GG^T. The top k principal components are then included as covariates in the model.

Because closely related individuals can distort principal component structure, it is advisable to remove highly related samples prior to PCA.

S2-9: Integrative framework of linkage and association analysis (conceptual illustration of NAM/MAGIC)

NAM and MAGIC populations increase recombination events and allelic diversity, combining advantages of both linkage and association mapping. Statistically, they often employ joint linkage mapping and association testing, while absorbing family background effects through random effects or hierarchical structures. This enables more robust effect estimation across multiple genetic backgrounds and results in narrower confidence intervals.