

evaluation and practice, we propose a standardized workflow of “training-validation-freezing-external evaluation,” emphasizing a multi-dimensional assessment based on relative  $R^2$ /AUC, calibration slope, and decision-curve net benefit. This framework is further complemented by ancestry-stratified and LD-perturbation sensitivity analyses, small-sample recalibration in target populations, and cost-benefit evaluation, supported by benchmark datasets, reference panels, and transparent reporting standards, with the aim of promoting the development of PRS/PGS as a transferable, interpretable, and equitable tool for applications in human health and crop improvement.

## 1 Fundamental Framework and Methodological Evolution of PRS/PGS

From a statistical inference perspective, the construction of PRS/PGS can be understood as a process that translates population-level association signals into individual-level predictive quantities. Specifically, the marginal effect estimates obtained from GWAS are not directly suitable for prediction; instead, they must be re-estimated under appropriate modeling assumptions that account for linkage disequilibrium structure and effect size distributions. This typically involves shrinkage and aggregation procedures that yield more stable and generalizable effect representations. These regularized effects are then projected onto individual genotype profiles to quantify genetic risk at the individual level. In this sense, PRS can be viewed as a model-dependent predictive function, jointly determined by effect estimation, regularization strategies, and genotype encoding, reflecting a continuous inferential pathway from association signal extraction to individual risk prediction.

Differences among methods fundamentally arise from distinct modeling assumptions regarding effect size distributions, linkage disequilibrium (LD) structure, and sparsity, thereby implying different statistical targets (estimands). Under this framework, the evolution of PRS methodology can be understood as a progression from “independent locus approximation” to “LD-aware modeling,” and further toward the integration of functional and ancestry information.

### 1.1 Classical clumping and thresholding (C+T)

Clumping and thresholding (C+T) is one of the earliest and most widely used approaches for constructing PRS/PGS (Sima et al., 2024) (Figure 1). This method begins with single-marker GWAS effect estimates, ranks candidate variants by statistical significance ( $p$ -values), and performs LD clumping using predefined window sizes and  $r^2$  thresholds to retain representative “sentinel” SNPs. Individual scores are then calculated via linear aggregation:

$$PRS_i = \sum_{m=1}^M \beta_m x_{im}$$

where  $\beta_m$  denotes the marginal effect size of the  $m$ -th SNP, and  $x_{im}$  represents the genotype of individual  $i$ . From a statistical modeling perspective, the C+T approach can be interpreted as a sparse estimation strategy that performs variable selection through hard thresholding. Under this framework, only variants exceeding a predefined significance threshold are retained, implicitly approximating the genetic architecture of a trait as being driven by a limited number of loci with relatively large effects. While this formulation simplifies the model structure, it also entails a substantial reduction of correlation information among variants. In practice, achieving a balance between model simplicity and predictive performance typically requires systematic exploration across a range of significance thresholds and linkage disequilibrium parameters, with model selection guided by validation data to identify an appropriate parameter configuration.

C+T offers advantages including simplicity, low computational cost, direct compatibility with GWAS summary statistics, and strong interpretability, making it a useful baseline method or rapid screening tool for large-scale traits (Wang et al., 2023). However, its “hard” LD pruning discards potentially informative variants and prevents optimal weighting within LD blocks. Moreover, its sensitivity to threshold selection and reference panels often leads to poor transferability across populations (Jayasinghe et al., 2024; Kachuri et al., 2024). Fundamentally, this reflects substantial shifts in the estimand when LD structure is ignored in cross-population settings.