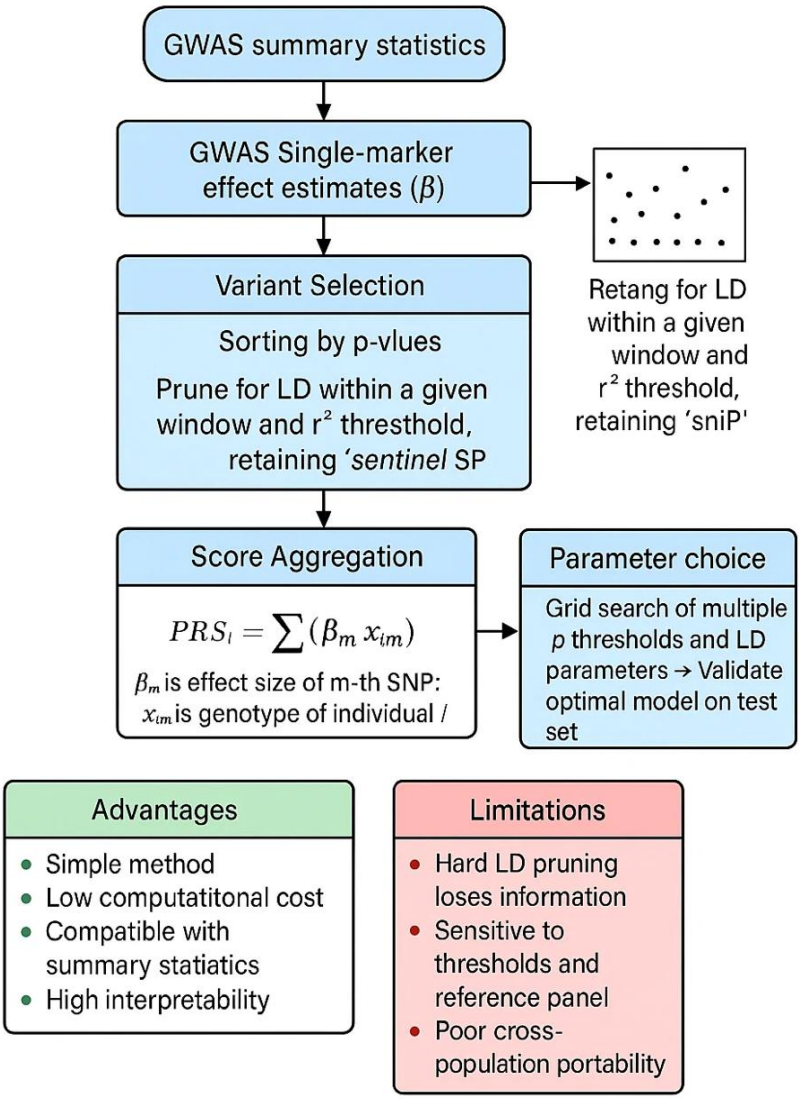

Figure 1 Workflow of the clumping and thresholding (C+T) approach for constructing polygenic risk scores (PRS)

Note: This figure illustrates the standard workflow of the clumping and thresholding (C+T) method for constructing polygenic risk scores (PRS). Starting from GWAS summary statistics, single-marker effect estimates (β) are obtained and used for variant selection. SNPs are first ranked by statistical significance (p-values), followed by linkage disequilibrium (LD)-based clumping within a specified genomic window and r^2threshold, retaining representative "sentinel" variants while removing correlated markers. The selected variants are then aggregated into an individual-level PRS using a linear scoring function, where SNP effect sizes serve as weights and individual genotypes as predictors. Model parameters, including p-value thresholds and LD pruning criteria, are typically optimized via grid search in a validation dataset

The C+T approach is computationally efficient, interpretable, and compatible with GWAS summary statistics, making it a widely used baseline method. However, its reliance on hard LD pruning may discard informative variants and lead to suboptimal weighting within LD blocks. In addition, the method is sensitive to parameter choices and LD reference panels, which limits its portability across populations.

1.2 LD-aware and bayesian methods

To overcome the limitations of C+T in handling LD, methods that explicitly model LD structure have been developed, including LDpred, LDpred2, and PRS-CS (Weissbrod et al., 2022). These approaches incorporate an LD matrix $R$from a reference panel and jointly estimate SNP effects via de-correlation and shrinkage, aiming to obtain the posterior expectation:

$$\mathbb{E}(\beta \mid \hat{\beta}, R)$$