Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Trait-specific benefits of association-based marker preselection for genomic prediction in slash pine

Xianyin Ding
xianyinding@gmail.com

Chinese Academy of Forestry
Yue Zhang
Zhejiang Yalin Biotechnology Co., Ltd
Qifu Luan
Chinese Academy of Forestry
Huijin Fan
Southwest Forestry University
Zifeng Tan
Chinese Academy of Forestry
Jingmin Jiang
Chinese Academy of Forestry

Research Article

Keywords: slash pine, oleoresin and naval stores, trait-class-specific architectures, genomic selection, GWAS prior, machine learning

Posted Date: October 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10956385/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Slash pine (*Pinus elliottii* Engelm. var. *elliottii*) is a leading resin-producing conifer and multipurpose industrial tree species, valued as a renewable source of rosin and turpentine for the naval-stores and bio-based chemical industries, and of timber and pulp; yet simultaneous improvement of growth, wood quality, oleoresin yield, and resin chemistry is constrained by long generation intervals and late-expressed traits. Using 240 open-pollinated trees genotyped with 53,229 transcriptome-derived SNPs, we characterized marker associations for 12 traits and evaluated genomic selection (GS) with GWAS, SNP ranking, QTN discovery, and model fitting restricted to training folds. Training-fold GWAS showed differences in association patterns among trait classes: growth traits showed the strongest, most recurrent associations, wood-quality traits moderate signals, and oleoresin traits weaker, less stable associations. Annotation of significant, large-effect, and recurrent QTN-tagged transcripts converged on nucleotide and RNA processing, respiratory energy and cell-wall precursor supply, and lipid-terpenoid metabolism, providing hypotheses for growth, wood formation, and resin physiology. Using fold-contained GWAS outputs as marker priors, model-averaged predictive ability increased by 11.9% for growth, more than five-fold the 2.2% gain for wood quality, but did not improve oleoresin traits. The selected configurations for diameter at breast height and height yielded PA values of 0.493 and 0.365, respectively, no single GS model dominated across traits, and fixed large-effect or recurrent QTN covariates rarely surpassed optimized genomic baselines without fixed QTN effects, suggesting limited additional predictive value of these covariates under the tested configurations. GS-selected elite sets retained 80.3–95.1% of the same-cycle phenotype-BLUP benchmark gain; assuming 15-year conventional and 7-year GS cycles yielded a projected 1.7-2.0-fold annualized benchmark advantage. These results support further evaluation of trait-specific, GWAS-informed prediction in slash pine. Model selection reused the cross-validation results, and independent family, site, and generation validation is needed; realized breeding gains were not measured.

1. Introduction

Conifer breeding seeks simultaneous gains in productivity, wood quality, and oleoresin yield across populations that remain in the field for decades (Niu et al., 2022; Plomion et al., 2011). Unlike annual crops, forest trees combine delayed trait expression, demanding field-test logistics, pronounced environmental heterogeneity, and long reproductive cycles, a combination that renders conventional recurrent selection both slow and expensive (Gyllenstrand et al., 2007; Isik, 2014). Phenotypic evaluation and BLUP-based estimation of breeding values remain the backbone of tree improvement because they exploit field-trial data to separate genetic from environmental effects (Piepho et al., 2008). However, phenotype-based selection alone is insufficient to achieve the shorter breeding cycles now demanded by the forestry sector, particularly for traits with low heritability, late expression, or requirements for destructive or chemical assays (Grattapaglia, 2022; Grattapaglia and Resende, 2011; Isik, 2014). Conifer- and forest-tree-specific reviews have recently re-examined these foundations in the context of climate adaptation and shortened cycles (Sharma et al., 2024; Zhang et al., 2025).

Marker-assisted selection was once expected to accelerate forest breeding by tracking markers linked to large-effect QTLs, but its realized gains have proved modest for the highly polygenic traits that dominate forestry (Muranty et al., 2014). Because most economically important tree traits are controlled by many loci of small to moderate effect, individual QTLs explain little phenotypic variance and selection on a handful of markers captures only a fraction of genetic merit. Genomic selection (GS) circumvents this limitation by using genome-wide markers to predict genomic estimated breeding values (GEBVs) without having to resolve every causal variant (Meuwissen et al., 2001). The approach is especially attractive in conifers, where prediction at an early stage can compress the interval between crossing, evaluation, and parental recycling (Grattapaglia, 2022; Grattapaglia et al., 2018; Resende Jr. et al., 2012).

The accuracy of GS is governed by trait heritability, training-population size, the extent of marker-QTL linkage disequilibrium, the relatedness between training and selection candidates, marker density, and the assumed distribution of marker effects (Daetwyler et al., 2008; Wientjes et al., 2013; Jannink et al., 2010). GBLUP is robust and computationally efficient when many loci contribute small additive effects, whereas Bayesian variable-selection models can be advantageous when a subset of loci carry larger effects (Habier et al., 2011; Perez and de los Campos, 2014). Kernel and machine-learning methods, such as RKHS regression, support vector regression, tree ensembles, gradient boosting, and neural networks, can further capture nonlinear signal. However, their use requires careful validation in the small training populations typical of forest trees (Bellot et al., 2018; Gianola and van Kaam, 2008; Yin et al., 2020). More broadly, evidence across plant breeding shows that marker density, training design, and model class can all affect prediction performance. This supports explicit, case-specific benchmarking rather than reliance on a single default model (Bernardo and Yu, 2007; Crossa et al., 2017; Crossa et al., 2014; Heffner et al., 2009; Heslot et al., 2012; Jannink et al., 2010; Lorenz et al., 2011).

GWAS priors offer a principled way to sharpen GS: markers with stronger fold-specific associations can define compact, trait-relevant marker sets, and recurrent or large-effect QTNs can enter the model as fixed effects. Across crops, aquaculture species, and forest trees, GWAS-assisted marker preselection or fixed-effect QTN modelling has improved prediction in some settings (Chen et al., 2023; Luo et al., 2021; Spindel et al., 2016; Zhang et al., 2014). Yet the same strategy can fail when GWAS signals are unstable, when associated markers are merely LD proxies, when the background genomic model already absorbs the signal, or when too many fixed effects overwhelm a small training set. A more insidious risk is information leakage: if GWAS is run on individuals that later appear in the test set, marker ranking and QTN selection import test-set information and inflate apparent accuracy. A valid evaluation must therefore confine GWAS and all marker selection to each training fold. Accordingly, we partitioned GWAS evidence into SNP-wise significance, marker PVE, and fold recurrence. These classes were treated as training-fold marker priors rather than proof of causality, especially because our transcriptome-derived SNPs were not anchored to chromosome-scale coordinates. This fold-contained discipline mirrors the established association-mapping principle that population structure, relatedness, and marker testing must be controlled within the analysed sample (Yu et al., 2006; Price et al., 2006; Kang et al., 2008; Zhou and Stephens, 2012, 2014).

Slash pine is one of the most widely planted conifers in southern China and a mainstay of the country's naval-stores industry, valued for its rapid growth, broad adaptability, timber and pulpwood production, and high oleoresin yield—the renewable feedstock for rosin and turpentine (Diao et al., 2019; Ding et al., 2024). A previous transcriptome-based association study in this same breeding population delivered SNP resources for growth, wood-quality, and oleoresin traits, providing an ideal foundation for integrating GWAS with GS (Ding et al., 2022). Because the breeding objective is intrinsically multi-trait, elite individuals

must combine vigorous growth, sound wood mechanics, high resin yield, and a favourable terpene and resin-acid profile, we set out to convert these genomic and phenotypic resources into a fold-contained GS pipeline. Specifically, we aimed to (i) benchmark eight GS models across 12 marker densities under random versus GWAS p-value-preselected SNP sets, (ii) test whether PVE- and recurrence-frequency-based QTN covariates improve prediction once markers are prioritized within each training fold, and (iii) compare phenotype-BLUP and GS-based selection on a common phenotype-BLUP benchmark together with slash-pine-specific annualized-gain assumptions.

2. Materials and methods

2.1 Plant material and field trial

The breeding population traces to a slash pine clonal seed orchard assembled from 300 superior clones selected in provenance trials spanning Guangdong, Fujian, Jiangxi, Jiangsu, Hunan, Hubei, and Zhejiang, following the breeding resource described by Ding et al. (2022). Scions of the selected trees were grafted onto two-year-old rootstocks in 1976. Open-pollinated half-sib seed was collected in 1992, sown in 1993, and established as a family trial near Hangzhou, China, in 1994. The trial used a randomized complete block design with six blocks and six-tree row plots at 2 m × 3 m spacing, at a site with a mean annual temperature of approximately 17.0°C and annual precipitation of approximately 1480 mm. Stands were thinned after 10 years, and no oleoresin tapping had been performed before the present phenotyping.

Plant material collection and field investigations were conducted in accordance with relevant institutional, national, and international guidelines and legislation. Permission to access the experimental site and collect samples was granted by Changle Forest Farm, which co-manages the genetic resources with the Research Institute of Subtropical Forestry, Chinese Academy of Forestry. The plant materials were identified as Pinus elliottii Engelm. var. elliottii by Qifu Luan. Voucher specimens representing the materials examined in this study are deposited in the herbarium of the Research Institute of Subtropical Forestry, Chinese Academy of Forestry, under voucher accession number(s) [RISF-PEE-2019-001 to RISF-PEE-2019-240].

2.2 Phenotypic measurements

Phenotyping was conducted in July 2019 across four blocks of the trial. Within the 33 represented families, the three trees per plot with a diameter at breast height (DBH) close to the family mean were selected, and after excluding trees with severe physical damage, 240 individuals were retained for analysis. Twelve traits were measured (Data S1). Growth traits comprised DBH, tree height (Ht), crown width (Crown), and annual ring width (ARW). DBH was measured at 1.3 m with a diameter tape, height with a telescopic levelling pole, and crown width as the mean of the north-south and east-west diameters. ARW and the resistance-amplitude wood traits were obtained with an IML RESI PD500 Resistograph and processed in PDToolsPro (Isik and Li, 2003).

Wood-quality traits comprised modulus of elasticity (MOE), acoustic mean propagation velocity (AMP), and Pilodyn penetration (Pd). Wood cores were taken at breast height with a 5 mm increment borer, and wood density was determined by the maximum moisture-content method. Stress-wave velocity was recorded with a Hitman ST300, and MOE was derived from wood density and stress-wave velocity following established nondestructive wood-evaluation principles (Ross, 2015). Pilodyn penetration was measured on the north and south sides of the stem, and the average was retained after repeated readings whenever paired values differed by more than 2 mm.

Oleoresin traits comprised oleoresin yield (OY) together with alpha-pinene, beta-pinene, abietic acid, and levopimaric acid. OY was quantified by fixing a 15 mL tube to a borehole on the sunlit side of the trunk for 24 h and weighing the collected resin. Chemical components were profiled by GC-MS on an Agilent 6890/5975B platform: resin was dissolved in ethanol with tetramethylammonium hydroxide, separated under a 60–280°C temperature program, and identified against the NIST08 library and published spectra. The major monoterpene and resin-acid components were retained for GS, consistent with recent slash pine oleoresin phenotyping and functional-genomic work (Ding et al., 2025a; Ding et al., 2025b).

2.3 Transcriptome-based SNP discovery and genotype processing

Secondary xylem was sampled from all 240 trees under RNase-free conditions after removal of bark and phloem, snap-frozen in liquid nitrogen, and stored at -80°C. Total RNA was extracted with a polysaccharide-polyphenol kit and assessed on NanoDrop, Qubit, and Agilent 2100 instruments; high-quality RNA was used for cDNA library construction and Illumina sequencing. Reads were aligned to a third-generation, full-length slash pine reference transcriptome (Diao et al., 2019) with STAR (Dobin et al., 2013). SNPs were called with the GATK pipeline (McKenna et al., 2010), filtered in PLINK to remove markers with minor allele frequency below 0.05 or call rate below 80% (Purcell et al., 2007), and imputed with Beagle (Browning et al., 2018). The resulting HMP genotype matrix comprised 53,229 transcriptome-derived SNPs across 240 individuals (Data S1). Because these SNPs derive from transcriptome assemblies and are not anchored to physical chromosomes, all marker positions were treated as transcript-local coordinates, and downstream interpretation focused on allele-substitution classes, fold-contained association evidence, transcript- and gene-level candidate prioritization, and functional annotation.

2.4 Heritability estimation

Phenotype-BLUP breeding values were estimated under the following linear mixed model:

$$y = Xb + Za + e$$

1

where y is the phenotype vector, b contains fixed effects, a contains individual additive genetic effects, X and Z are incidence matrices, and e is the residual vector. Genomic heritability was estimated with all markers under GBLUP. The genomic relationship matrix (G) was constructed from standardized markers:

$$G = \frac{MM\prime}{m}$$

2

where M is the centered and standardized marker matrix, $M\prime$ denotes the transpose of M, and m is the number of markers. The mixed model partitioned the phenotypic variance into a genomic additive component and a residual component. Individual-tree genomic heritability was calculated as:

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

3

where σ_g^2 is the genomic additive variance and σ_e^2 is the residual variance. The mixed-model and GBLUP variance-component implementation followed genome-assisted prediction practice for quantitative traits (Covarrubias-Pazaran, 2016), and the resulting GBLUP heritability estimates, used to scale PA into accuracy (AC), are reported in Table S1.

2.5 Training-fold GWAS and SNP preselection

The validation design and analysis flow are summarized in Fig. 1. The core scheme comprised 10 repeats of 10-fold splitting, yielding 100 held-out tests; within each outer fold, 90% of trees formed the training set and 10% were held out. Genome-wide association analysis was run independently within each training set for every trait using the BLINK (Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway) model implemented in GAPIT v3.0 under R. To control confounding by population structure and relatedness, fold-specific principal components and a marker-derived kinship matrix were included as covariates in the GAPIT analysis, and the validation trees were excluded from all GWAS inputs. Only training-set GWAS outputs were used for SNP ranking, QTN-evidence classification, and QTN-covariate construction. For descriptive visualization, training-fold SNP summaries were condensed per trait by the minimum *P* value, maximum PVE, and recurrence frequency across folds, and a continuous circular Manhattan profile was drawn after ordering transcript-derived SNPs by marker group and transcript-local position, without assigning chromosome blocks. Three operational evidence classes were defined: SNPs meeting a nominal association threshold (SigSNP, -log10(P) > 4), markers meeting a PVE threshold (PVE-QTN, maximum PVE > = 0.05), and recurrent SNPs (RecSNP, detected in at least 50% of outer folds); their overlap was summarized in an UpSet plot. These labels identify candidate markers, not confirmed causal QTNs or genome-wide significance after correction for multiple marker tests. All evidence classes were then merged with allele metadata and transcript functional annotations, and genes tagged by nominally associated, high-PVE, and recurrent loci were tested for GO and KEGG pathway enrichment against the full annotated transcript-derived marker background using one-sided hypergeometric tests with Benjamini-Hochberg FDR correction (Ashburner et al., 2000; Benjamini and Hochberg, 1995; Kanehisa et al., 2023; Kanehisa and Goto, 2000). Circular Manhattan and QQ diagnostics were generated in R with CMplot from the fold-contained GWAS summary tables. Publication panels were refined from the same tables to preserve continuous transcript ordering, trait-specific colours, compact legends, and the source-specific QTN-evidence layering required for a chromosome-unanchored transcriptome-SNP panel.

To extract biologically meaningful information from chromosome-unanchored SNPs, candidate loci were further summarized at the transcript and gene level. For each trait-gene record, an evidence-weighted priority score combined bounded association strength, marker PVE, fold-recurrence percentage, and the number of supporting evidence classes, with weights of 0.40, 0.25, 0.20, and 0.15, respectively; association strength was capped at -log10(P) = 10 and marker PVE at 0.25 so that no single component dominated. The score served only to rank transcript-level candidates for interpretation, not as a posterior probability of causality. Enrichment plots displayed -log10(P value) on the x-axis, evidence source as point colour, and gene number as point size, with GO terms or KEGG pathways shared across evidence sources placed on a common row. Local KEGG map HTML files were parsed to recover KO-button and candidate-gene annotations, and a KEGG-informed candidate pathway model was drawn to summarize candidate modules rather than to infer causal chromosome intervals. Trait-gene and cross-trait candidate summaries are provided in Table S2, parsed local KEGG-map evidence in Data S3, and GO/KEGG enrichment and KO-mapping outputs in Table S4.

2.6 Genomic prediction models

Eight prediction models were evaluated. GBLUP used a linear genomic-relationship kernel built from standardized markers (Clark and van der Werf, 2013). RKHS regression used a Gaussian kernel within the kernel-regression framework for genomic prediction (de los Campos et al., 2010; Gianola and van Kaam, 2008). Support vector regression (SVR) was implemented in its linear form (Cortes and Vapnik, 1995). Extra Trees and Random Forest were fitted as tree ensembles with marker subsampling and minimum-leaf-size regularization (Breiman, 2001; Geurts et al., 2006). XGBoost was trained with a squared-error objective, histogram-based tree construction, shallow trees, shrinkage, column subsampling, and L1/L2 regularization (Chen and Guestrin, 2016). The MLP neural network (MLP-NN) used a single hidden layer with train-only dimensionality reduction whenever the marker count exceeded 3000 (Bellot et al., 2018). BayesC was fitted within the BGLR Bayesian-regression framework (Habier et al., 2011; Perez and de los Campos, 2014). The marker-kernel and ridge-regression components are further consistent with rrBLUP-style genomic-prediction implementations widely used in plant breeding (Endelman, 2011).

2.7 PVE and GWAS-recurrence-frequency QTN fixed-effect strategies

After the genomic-prediction benchmark, each trait's selected model and preselected SNP density served as the background marker set for QTN-strategy testing. The baseline strategy (None_QTN in tables and figures) retained this background without any QTN covariates. The PVE strategy added covariates from cumulative fold-specific PVE intervals of (0.01, 0.05), (0.01, 0.10), (0.01, 0.15), (0.01, 0.20), and (0.01, 0.25). The frequency strategy exploited nested training-only GWAS: within each outer training set, 10 inner GWAS replicates identified SNPs with -log10(P) > 3, from which recurrence frequency was

computed, and cumulative frequency sets of [50%,60%], [50%,70%], [50%,80%], [50%,90%], and [50%,100%] were tested. QTN SNPs were removed from the background marker matrix before modelling to avoid collinearity; they were then fitted as explicit covariates in the mixed and Bayesian models and appended as prioritized covariate features in the machine-learning models. Throughout, fixed-QTN denotes QTN-covariate modelling across all model classes; in machine-learning models, prioritized features do not represent fixed effects in the mixed-model sense.

2.8 Prediction metrics, statistical summaries, and model selection

For each held-out test fold, prediction ability (PA) was calculated as the Pearson correlation between test-set genomic breeding values and observed test-set phenotypes:

$$PA \; = \; cor(GEBV_{test},\, y_{test})$$

4

Heritability-scaled prediction accuracy (AC) was calculated by dividing PA by h, where h denotes the square root of h²:

$$AC \; = \; \frac{PA}{h}$$

5

PA was the primary criterion for model comparison. Candidate model and SNP-density configurations were ranked by mean cross-validated PA across the 100 held-out tests, with AC and residual AIC used only as secondary diagnostics to break effective ties. Because the same held-out folds informed both comparison and selection of candidate configurations, reported performance of the selected configurations may be optimistic and is not an unbiased estimate from an additional independent outer validation. AC is an approximate heritability-scaled diagnostic, not a directly observed correlation with true breeding values, and does not propagate uncertainty in the heritability estimates. For the QTN strategies, a non-baseline covariate strategy was retained only when it raised mean PA above the no-QTN baseline under the same trait, model, SNP density, and fold structure. Exploratory comparisons against GBLUP used fold-level paired resampling with false-discovery-rate adjustment; because repeated cross-validation folds are not fully independent, these annotations are interpreted descriptively alongside PA effect sizes.

2.9 Breeding values, multi-trait index, and genetic gain

Traditional selection used phenotype-BLUP estimated breeding values (EBVs), whereas GS selection used genomic estimated breeding values (GEBVs) from each trait's selected model, SNP set, and retained QTN strategy. The operational selection index combined DBH and Ht (growth), MOE and AMP (wood quality), and OY, alpha-pinene, beta-pinene, abietic acid, and levopimaric acid (oleoresin yield and composition). Crown, ARW, and Pd were analysed throughout the GS pipeline but excluded from the final multi-trait elite-selection index.

Because the target traits are measured on different scales, each trait-specific EBV or GEBV was standardized across candidate trees before the index was formed. Two ranking indices were then constructed in parallel: a phenotype-BLUP index for traditional selection and a GS index for genomic selection.

$$I_P\,(i) \; = \; \frac{1}{q}\,\Sigma_{\,t}\, z\,[EBV_{PBLUP}(\,i,t\,)]$$

6

$$I_{GS}\,(i) \; = \; \frac{1}{q}\,\Sigma_{\,t}\, z\,[GEBV_{GS}(\,i,t\,)]$$

7

where q is the number of target traits and z denotes standardization within a trait. For each Top N from 1 to 30, candidates were ranked separately by the phenotype-BLUP index and the GS index. To compare the two selected sets on a common scale, selected-set index gain was evaluated using the phenotype-BLUP index as the benchmark for both routes. Thus, phenotype-selected trees and GS-selected trees were both measured against the same candidate-population mean phenotype-BLUP index:

$$\Delta\, G_P\,(N) \; = \; mean\,(I_P \mid \; S_P\,(N)) \; - \; mean(I_P \mid C)$$

8

$$\Delta\, G_{GS}\,(N) \; = \; mean\,(I_P \mid \; S_{GS}\,(N)) \; - \; mean(I_P \mid C)$$

9

where $S_P\,(N)$ is the Top N set selected by phenotype-BLUP ranking, $S_{GS}\,(N)$ is the Top N set selected by GS ranking, and C is the full candidate population. The same phenotype-BLUP benchmark was also used for annualized gain comparison, preventing route-specific EBV and GEBV shrinkage scales from being compared directly. Annualized gain was calculated by dividing genetic gain by the corresponding breeding-cycle length:

$$\Delta\, G_{year}\,(N) \; = \; \frac{\Delta\, G\,(N)}{L}$$

10

The traditional slash pine breeding cycle was set at 15 years. Assuming GS saves approximately 8 years of progeny testing for growth, wood-quality, and oleoresin traits gives a 7-year GS cycle, and the annualized GS advantage was defined as:

$$R_{GS}(N) = \frac{\Delta G_{GS}(N)}{7} / \frac{\Delta G_{P}(N)}{15}$$

11

Trait-level expected responses from the breeder's equation for Top 1 to Top 30 were retained as supplementary context. These used the selection intensity implied by the selected proportion N/240, the trait-specific prediction accuracy, and the genomic additive standard deviation; the primary elite-selection result, however, remained the selected-set index comparison described above. These index differences and breeder's-equation projections are not realized genetic gains measured in progeny.

3. Results

3.1 Training-fold GWAS signals and candidate QTN evidence

The genotype panel comprised 53,229 transcriptome-derived SNPs distributed across 17,828 transcript marker groups. Biallelic substitutions were strongly transition-biased, with 33,708 transitions and 19,521 transversions retained, yielding a Ti/Tv ratio of 1.74 (Fig. 2A). The most frequent substitution classes were C/T, G/A, A/G, and T/C, consistent with the well-documented excess of transitions in expressed-sequence-derived markers. Training-fold GWAS yielded complete marker-prior summaries for all 12 traits across the 10-repeat, 10-fold design (Table S1). Every trait retained the full set of 53,229 SNP markers and 100 successful outer folds, each with 216 training and 24 held-out trees. Trait-class circular Manhattan profiles (Fig. 2B-D) summarized the relative concentration of signal among growth, wood-quality, and oleoresin traits. Genomic-control lambda values were close to unity, ranging from 0.962 (abietic acid) to 1.127 (Pd), and the shared-coordinate QQ overlay (Fig. 2E) showed differing tail deviations across traits. These descriptive diagnostics do not establish the calibration of fold-aggregated association statistics. Merging the evidence classes across folds yielded 862 SigSNP, 5,181 PVE-QTN, and 385 RecSNP trait-SNP records, with their unique and shared support resolved by the UpSet analysis (Fig. 2F; Table S2). Although the panel was not chromosome-anchored, overlap among SigSNP, PVE-QTN, and RecSNP highlighted transcript-level candidates for follow-up; recurrence across overlapping training folds is not independent replication.

3.2 Functional annotation suggested three candidate biological modules

The evidence-weighted candidate-transcript priority map (Fig. 3) re-projected these pathway-level signals onto individual genes, clarifying how the SigSNP, PVE-QTN, and RecSNP evidence classes converge at the transcript level and identifying the candidate transcripts with the strongest multi-source support. Functional annotation of the QTN-tagged transcripts then placed these prioritized genes into three candidate biological modules (Fig. S1, Table S4 and Fig. 4), providing hypotheses linking marker evidence to trait biology. GO and KEGG enrichment (Fig. S1; Table S4) consistently highlighted nucleotide and RNA-processing terms, together with respiratory metabolism, cell-wall-precursor supply, and lipid-metabolic categories, suggesting potential links to growth, wood formation, and resin biosynthesis; nominal enrichment and biochemical context should be distinguished from FDR-supported enrichment. These recurring signals were distilled into the KEGG-informed pathway model (Fig. 4), including (i) a nucleotide-pool, polymerase, and repair module linked to cell proliferation, transcriptional activity, and growth or wood-quality variation, (ii) a respiratory-energy and cell-wall-precursor module connecting oxidative phosphorylation, pentose/glucuronate metabolism, amino- and nucleotide-sugar metabolism, Golgi transport, and secondary-wall deposition, and (iii) a BCAA, CoA, lipid-remodelling, and terpenoid-resin module relevant to oleoresin yield, pinene composition, and resin-acid chemistry.

3.3 Highest-ranked GS model and marker density differed among traits

The eight-model benchmark revealed that no single GS model dominated across traits (Table 1 and Fig. 5). Among growth traits, XGBoost achieved the highest mean PA for both DBH and Ht (0.493 at 0.3K SNPs and 0.365 at 3K SNPs, respectively), whereas Crown was best predicted by Random Forest at 10K SNPs and ARW by Extra Trees at 3K SNPs. These cross-validation-selected configurations indicate differences in predictive performance within this dataset. Model rankings alone do not establish nonlinear genetic effects or epistasis, and independent validation is required before deployment.

Wood-quality traits followed a different pattern. MOE was best predicted by GBLUP using all markers, consistent with useful predictive information in genome-wide realized relationships, whereas AMP was best predicted by SVR at 40K SNPs and Pd by BayesC at 7K SNPs. Thus, even within a single trait class, the highest-ranked configuration differed; this comparison does not by itself identify the underlying genetic architecture.

Table 1
Cross-validation-selected GS model and SNP density for each trait.

Trait	Trait class	Selected GS model	Selected SNP set	Prediction ability	Accuracy	AIC
DBH	Growth	XGBoost_GPU	300	0.493	0.597	129.02
Ht	Growth	XGBoost_GPU	3000	0.365	0.458	111.96
Crown	Growth	RandomForest	10000	0.141	0.738	68.18
ARW	Growth	ExtraTrees	3000	0.349	0.442	56.34
MOE	Wood quality	GBLUP	All	0.157	0.639	79.98
AMP	Wood quality	SVR	40000	0.206	0.646	377.45
Pd	Wood quality	BayesC	7000	0.392	0.445	390.20
OY	Oleoresin	BayesC	100	0.087	0.865	185.82
Alpha_pinene	Oleoresin	GBLUP	500	0.088	0.881	138.42
Beta_pinene	Oleoresin	XGBoost_GPU	300	0.139	0.834	159.45
Abietic_acid	Oleoresin	RKHS	3000	0.091	0.911	113.38
Levopimaric_acid	Oleoresin	MLP_NN	20000	0.160	0.646	136.26

Notes: Selected GS model indicates the configuration with the highest mean PA within the cross-validated benchmarking design for each trait. Selected SNP set indicates the marker density and SNP-selection scheme used by that model, Marker-selection scheme is not shown in this table; Preselection and Random in Fig. 5 denote GWAS p-value-ranked selection and random SNP sampling, respectively. AIC is reported as a residual diagnostic and was not treated as a primary criterion for comparing fundamentally different model classes.

Oleoresin traits showed generally lower PA, in line with their lower GBLUP heritabilities (Table S1) although the low heritabilities and small sample size limit comparisons of heritability-scaled AC across traits. Their optimal models were again trait-specific: BayesC for OY, GBLUP for alpha-pinene, XGBoost for beta-pinene, RKHS for abietic acid, and MLP-NN for levopimaric acid. GWAS p-value preselection was most beneficial for growth traits, where mean PA across model-density combinations rose from 0.242 under random selection to 0.271 under preselection (+ 11.9%) and mean AC from 0.431 to 0.467 (+ 8.3%). Wood-quality traits gained more modestly (PA 0.214 to 0.219, + 2.2%; AC essentially stable at 0.498), whereas oleoresin traits showed no meaningful change (PA 0.101 under both schemes; AC stable at 0.748). The trait-class PA/AC summary is given in Table S1, the model-wise PA comparison relative to GBLUP in Fig. 6, and the corresponding AC and AIC trends in Figures S2 and S3.

3.4 PVE and frequency-defined QTN covariates added little over a selected no-QTN baseline

The PVE and GWAS-recurrence-frequency QTN-covariate strategies were tested after fixing each trait's background model and SNP density. The retained-strategy summary covers all 12 traits (Table 2 and Fig. 7), with PA responses to PVE and frequency QTNs shown in Figs. 7A and 7B. The no-QTN baseline was retained for nine of the 12 traits, indicating that adding QTN covariates did not improve cross-validated PA for most traits. Frequency-based QTNs produced only marginal gains, for Crown (Freq_[50%,100%], ΔPA = 0.0123), beta-pinene (Freq_[50%,80%], ΔPA = 0.0015; 10.43 fixed QTNs on average), and levopimaric acid (Freq_[50%,70%], ΔPA = 0.0072; 11.57 fixed QTNs on average). Strategy-specific AC and AIC diagnostics are provided in Figures S4-S7.

Table 2
Retained QTN strategy and representative GWAS-derived candidate loci across all 12 traits.

Trait	Selected GS model	Selected SNP set	Retained QTN strategy	Retained rule	Representative SNP ID	Allele change	MAF	Minimum P	Max PVE	Gene_ID
DBH	XGBoost_GPU	300	None_QTN baseline	None_QTN	SC338111_130	C/T	0.354	2.01e-07	0.240	c33811
Ht	XGBoost_GPU	3000	None_QTN baseline	None_QTN	SC156637_456	G/A	0.185	1.42e-07	0.148	c15663
Crown	RandomForest	10000	Frequency	Freq_[50%,100%]	SC168373_838	A/G	0.073	1.61e-06	0.145	c16837
ARW	ExtraTrees	3000	None_QTN baseline	None_QTN	SC328549_600	C/T	0.248	3.30e-08	0.186	c32854
MOE	GBLUP	All	None_QTN baseline	None_QTN	SC177028_1357	T/C	0.381	7.84e-08	0.259	c17702
AMP	SVR	40000	None_QTN baseline	None_QTN	SC158308_1556	G/A	0.098	3.46e-06	0.170	c15830
Pd	BayesC	7000	None_QTN baseline	None_QTN	SC147198_465	C/A	0.369	4.19e-07	0.141	c14719
OY	BayesC	100	None_QTN baseline	None_QTN	SC162207_1227	G/T	0.127	3.28e-06	0.115	c16220
Alpha_pinene	GBLUP	500	None_QTN baseline	None_QTN	SC169704_441	C/T	0.448	8.31e-07	0.122	c16970
Beta_pinene	XGBoost_GPU	300	Frequency	Freq_[50%,80%]	SC360402_1117	C/T	0.394	5.44e-07	0.108	c36040
Abietic_acid	RKHS	3000	None_QTN baseline	None_QTN	SC169425_1560	G/A	0.206	3.73e-07	0.245	c16942
Levopimaric_acid	MLP_NN	20000	Frequency	Freq_[50%,70%]	SC311368_794	C/T	0.050	3.62e-07	0.262	c31136

Notes: Selected GS model and Selected SNP set refer to the trait-specific background model selected before QTN-covariate testing. None_QTN baseline der selected background model and SNP density without additional QTN covariates. Retained QTN strategy indicates whether the strategy retained on the basis validated PA retained the no-QTN baseline or a recurrence-frequency-defined QTN covariate set. PA changes relative to the no-QTN baseline are summarized and Fig. 7. Representative SNP ID is the top fold-contained GWAS candidate selected per trait for display. Allele change is shown as reference/alternate alle allele frequency. Minimum P is the smallest fold-contained GWAS P value observed for that SNP, Max PVE is the maximum PVE reported across folds, Frequ supplementary records is the percentage of outer folds in which the SNP passed P < 0.001. Full QTN and functional-annotation records are provided in Table

The PVE strategy rarely improved prediction. For many traits, PVE thresholds admitted thousands of SNP covariates, a number large relative to the 216 training trees per fold, which frequently lowered PA or inflated residual AIC. This pattern favours a conservative reading: GWAS information supported marker ranking and compact background SNP sets for some traits, but the inclusion of QTN covariates required trait-specific validation and was not automatically beneficial.

3.5 GS-selected trees retained benchmark index gain with a projected annualized advantage

The final elite-selection index used the nine target traits: DBH, Ht, MOE, AMP, OY, alpha-pinene, beta-pinene, abietic acid, and levopimaric acid. Traditional selection ranked trees by a phenotype-BLUP EBV index, whereas GS selection ranked them by a GS index derived from the trait-specific selected models and QTN strategies. Crown, ARW, and Pd were not absent from the GS analysis; they were simply excluded from the final operational index in line with the target breeding objective.

The primary comparison applied the trees actually chosen by each ranking system but evaluated both selected sets on the same phenotype-BLUP index benchmark (Table S3 and Fig. 8). At Top 1, the phenotype-selected and GS-selected trees were different individuals, with benchmark gains of 1.038 and 0.971, respectively. At Top 10, phenotype selection reached 0.887 and GS selection 0.721, with 40.0% overlap; at Top 30, the values were 0.656 and 0.601, with 73.3% overlap. The trees chosen by each route at representative thresholds and the full Top 1–30 selections are provided in Table S3. Across Top 1 to Top 30, GS-selected sets retained 80.3–95.1% of the same-cycle phenotype-BLUP benchmark gain. GS therefore did not surpass phenotype-BLUP selection within the same cycle on this benchmark, but it identified partly distinct individuals and retained most of the benchmark index gain. Earlier selection was assumed in the breeding-cycle scenario and was not tested in young trees.

Annualization reshaped the practical interpretation (Fig. 8). Under the slash pine cycle assumptions of 15 years for phenotype-based selection and 7 years for GS, the common-benchmark annualized GS/phenotype ratio was 2.00 at Top 1, 1.74 at Top 10, and 1.96 at Top 30, spanning 1.72–2.04 across Top 1 to Top 30. The projected advantage therefore depends on the assumed 8-year reduction in breeding-cycle length while retaining the estimated benchmark index gain. The study did not measure realized progeny response, verify a 7-year operational cycle, or estimate the contribution of individual functional modules to the index gain. The modules in Fig. 4 remain hypotheses for follow-up rather than mechanistic explanations of selection response.

4. Discussion

4.1 Trait-specific model performance and limits to architectural inference

Our results offer no support for a single, universally superior GS model. Instead, the top-ranked model varied by trait, consistent with prediction models capturing different aspects of trait-associated variation (de los Campos et al., 2013; Meuwissen et al., 2001; Perez and de los Campos, 2014). GBLUP excelled for MOE and alpha-pinene, consistent with predictive information in realized relatedness and distributed marker effects, whereas BayesC was selected for Pd and OY, although these rankings do not establish that the traits are controlled by fewer large-effect loci. Tree-based methods won for several growth traits, XGBoost for DBH and Ht, Random Forest for Crown, and Extra Trees for ARW, but predictive superiority alone does not distinguish nonlinear genetic effects from relatedness, regularization, or sampling variation. Because each fold held only 216 training trees and many model-density combinations were benchmarked, these should be read as cross-validation-selected candidates rather than externally validated universal optima. RKHS and MLP-NN were retained for abietic acid and levopimaric acid, respectively, but warrant caution, since small training populations can leave flexible nonlinear models sensitive to noise (Bellot et al., 2018; Yin et al., 2020). Recent data-fusion frameworks for machine-learning-based GS likewise emphasize that gains from richer feature sets are realized only when the prediction architecture is matched to the trait-specific signal distribution (Wu et al., 2025).

Trait-dependent model behaviour of this kind has been reported across conifers and other forest trees. Across forest-tree studies, prediction performance has varied among models, traits, and validation settings (Lenz et al., 2017; Thistlethwaite et al., 2017; Calleja-Rodriguez et al., 2020). Other forest-tree studies likewise show that marker density and training-set composition can reorder models (Resende et al., 2012; Gamal El-Dien et al., 2015; Lenz et al., 2017; Thistlethwaite et al., 2017; Beaulieu et al., 2020).

4.2 GWAS preselection provides ranking prior rather than label

GWAS p-value preselection improved mean PA for growth and, to a lesser extent, wood-quality traits. This is consistent with the view that training-fold association signals can enrich marker sets for loci or LD blocks tied to the target phenotype, as reported in rice, aquaculture, Eucalyptus, and Norway spruce (Chen et al., 2023; Luo et al., 2021; Spindel et al., 2016; Zhang et al., 2014). The benefit of preselection nonetheless depends on training-population size, LD persistence, marker-QTL phase, and the relatedness between training and selection candidates (Daetwyler et al., 2008; Wientjes et al., 2013). In our population, oleoresin traits differed little between random and preselected markers, possibly reflecting lower heritability and less stable marginal associations for several resin chemical traits (Ding et al., 2023; Ding et al., 2022). Because the panel was transcriptome-derived, the association evidence also supplied candidate-gene and pathway context, but it should be read as a marker-prior and annotation signal rather than chromosome-scale causal localization. A parallel study in another slash pine breeding population recently demonstrated the same benefit of GWAS-guided preselection for pulpwood-related traits (Wu et al., 2026), providing additional support for evaluating the approach within this species, without establishing universal benefit.

The transcript-level prioritization adds a useful but deliberately conservative interpretation layer. Because the marker groups derive from expressed transcripts rather than chromosome-anchored scaffolds, shared candidate genes and enriched pathways indicate recurrent marker-prior and annotation support, not resolved causal intervals. The KEGG-informed diagrams accordingly summarize working biological modules: nucleotide and RNA-processing pathways plausibly linked to growth and wood-quality formation, and CoA, lipid-remodelling, and secondary-metabolism pathways plausibly linked to oleoresin yield and resin chemistry. These are testable hypotheses for follow-up expression, metabolite, and targeted-genotyping studies, not substitutes for independent functional validation.

Restricting marker discovery to training folds is important here. Ranking markers before cross-validation can transfer information from validation individuals into SNP selection and inflate accuracy; feature selection, marker ranking, dimensionality reduction, and model tuning must therefore be nested within the training partition whenever prediction is evaluated (Runcie and Cheng, 2019; Schulz-Streeck et al., 2011). Generating SNP rankings separately in each training fold prevents held-out target phenotypes from directly informing that fold's marker discovery. However, configuration selection using the same cross-validation results and relatedness between training and validation trees can still affect estimated performance and its transfer to deployment populations. Recent feature-selection benchmarking in maize has likewise shown that the choice of CV scheme alters apparent prediction accuracy more strongly than the choice of model class (Du et al., 2026), underscoring that leakage control is methodological, not optional.

4.3 Limited additional value of QTN covariates over selected baselines

That the no-QTN baseline prevailed for most traits is biologically sensible. GWAS p-values quantify marginal marker-trait association, whereas GS exploits multivariate marker patterns and realized relationships, so a significant marker may add little once the background marker matrix is in place (de los Campos et al., 2013; Meuwissen et al., 2001; VanRaden, 2008). Recurrent hits can also reflect LD blocks or family structure rather than independent causal variants, a familiar issue when training and selection candidates are related (Habier et al., 2007; Wientjes et al., 2013). And when PVE thresholds admitted many covariates, the set grew too large for the training size and raised overfitting risk, which is precisely why shrinkage, Bayesian variable selection, and strict validation remain central to genomic prediction (Habier et al., 2011; Perez and de los Campos, 2014; Runcie and Cheng, 2019). The GO and KEGG results thus provide functional context for QTN-tagged genes without proving causal pathways; they are most valuable as biologically coherent hypotheses connecting marker-prior evidence to metabolism, transport, signalling, and secondary-metabolism annotations.

These findings do not contradict reports of benefits from large-effect loci or GWAS-informed priors. Their usefulness may depend on whether QTN effects are stable, estimable within the training set, and complementary to the background genomic model (Chen et al., 2023; Merrick et al., 2021; Sarinelli et al., 2019). In this slash pine population, frequency-defined QTN covariates yielded only marginal gains for Crown, beta-pinene, and levopimaric acid and were unhelpful for most traits. The practical recommendation is to validate QTN covariates trait by trait and to default to the no-QTN baseline unless training-fold validation supports the QTN model.

4.4 Common-benchmark gain and the breeding-cycle reduction

The final comparison deliberately evaluated the trees actually chosen by each route, in keeping with the selection-index principle that breeding decisions should be judged on aggregate merit (Hazel, 1943). Phenotype selection ranked candidates by an index of phenotype-BLUP EBVs and GS by an index of GEBVs from trait-specific model and SNP configurations, but both selected sets were scored on the same phenotype-BLUP benchmark. This common-benchmark design avoids comparing gains across different EBV and GEBV shrinkage scales and provides a descriptive comparison of whether GS-selected trees retain estimated phenotype-BLUP index merit.

On this common benchmark, GS-selected sets had slightly lower same-cycle gain than phenotype-BLUP sets across Top 1 to Top 30, retaining 80.3–95.1% of the benchmark gain. This is expected, because the benchmark is itself built from phenotype-BLUP EBVs and therefore favours the phenotype-BLUP ranking. The operative question is whether GS retains enough benchmark merit to justify far earlier selection, and under the 15-year versus 7-year cycle assumptions used here, the calculation yielded a projected 1.72-2.04-fold annualized benchmark advantage across Top 1 to Top 30.

The annualized view matters most for an industrial tree crop such as slash pine, whose economic value is realized through the sustained supply of oleoresin, rosin, turpentine, and wood across successive deployment cycles. Forest-tree GS studies repeatedly stress that the largest practical gains come from shortening the generation interval and progeny-testing phase rather than from raising within-cycle accuracy alone (Grattapaglia and Resende, 2011; Grattapaglia et al., 2018; Isik, 2014; Lebedev et al., 2020). Here the traditional phenotype-based cycle was set at 15 years and GS was assumed to save roughly 8 years of growth, wood-quality, and oleoresin progeny testing, giving a 7-year cycle. This should be read as a breeding-scheme scenario, not as evidence that field testing becomes unnecessary: GS can bring forward candidate prioritization and parental recycling, but field testing remains essential for model updating, deployment validation, and monitoring age-age and genotype-by-environment stability (Beaulieu et al., 2014; Calleja-Rodriguez et al., 2020).

The scope of inference is limited by the 240 trees from 33 open-pollinated families at one site and phenotyping occasion. Selecting trees with DBH close to the family mean may restrict phenotypic variation and limit representation of the wider breeding population. Random individual-level folds can place relatives in both training and validation sets, so the reported PA may partly reflect within-population relatedness and does not establish prediction across families, sites, ages, or generations. Transcriptome-derived markers sample expressed regions and do not provide chromosome-wide coverage or chromosome-anchored causal intervals.

5. Conclusions

In summary, we evaluated a GS framework with training-fold marker discovery for slash pine, a multipurpose industrial conifer prized for its oleoresin-derived rosin and turpentine and for its wood, that unites transcriptome-derived SNPs, training-fold GWAS priors, multi-model benchmarking, QTN-covariate testing, functional interpretation, and multi-trait elite-tree selection. Training-fold GWAS preselection improved prediction chiefly for growth and wood-quality traits, while oleoresin prediction was largely unaffected. Fixed QTN covariates added value only for a small subset of traits, suggesting limited incremental predictive value beyond the selected marker backgrounds under the tested configurations. Functional annotation of QTN-tagged transcripts converged on three candidate modules—nucleotide processing, respiratory and cell-wall metabolism, and lipid-resin biochemistry—that furnish testable hypotheses for growth, wood formation, and resin chemistry. On a common phenotype-BLUP benchmark, GS-selected sets retained 80.3–95.1% of same-cycle benchmark index gain and showed a projected 1.72–2.04-fold annualized advantage under assumed 15-year conventional and 7-year GS cycles. These results support further evaluation of GWAS-informed marker prioritization and trait-specific prediction in slash pine. Their scope is limited by the small, related, single-site sample, reuse of cross-validation results for configuration selection, and reliance on an internal index benchmark and assumed cycle lengths. Independent family-structured, longitudinal, multi-site, and progeny validation is required before claims of operational gain or transfer to other tree crops.

Abbreviations

GWAS	Genome-wide association study
GS	Genomic selection
QTL	Quantitative trait locus
GBLUP	Genomic best linear unbiased prediction
PA	Prediction ability
PVE	Proportion of phenotypic variance explained
GEBVs	Genomic estimated breeding values
AIC	Akaike information criterion
LD	Linkage disequilibrium
FDR	False discovery rate
DBH	Diameter at breast height
AC	Heritability-scaled prediction accuracy
AMP	Acoustic mean propagation velocity
ARW	Annual ring width
BCAA	Branched-chain amino acid
BGLR	Bayesian Generalized Linear Regression
BLINK	Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway
BLUP	Best linear unbiased prediction
CoA	Coenzyme A
Crown	Crown width
CV	Cross-validation
EBVs	Estimated breeding values
GAPIT	Genome Association and Prediction Integrated Tool
GATK	Genome Analysis Toolkit
GC-MS	Gas chromatography-mass spectrometry
GO	Gene Ontology
HMP	HapMap format
Ht	Tree height
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
MAF	Minor allele frequency
MLP-NN	Multilayer perceptron neural network
MOE	Modulus of elasticity
OY	Oleoresin yield
Pd	Pilodyn penetration
PVE-QTN	Candidate quantitative trait nucleotide meeting the PVE threshold
QC	Quality control
QQ	Quantile-quantile
QTN	Quantitative trait nucleotide; used here as a candidate-marker label
RecSNP	Recurrent single-nucleotide polymorphism
RKHS	Reproducing kernel Hilbert space
SigSNP	Single-nucleotide polymorphism meeting the nominal association threshold
SNP	Single-nucleotide polymorphism
STAR	Spliced Transcripts Alignment to a Reference

SVR	Support vector regression
Ti/Tv	Transition-to-transversion ratio
XGBoost	Extreme gradient boosting

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The raw RNA sequencing data for the 240 slash pine individuals analysed in this study were previously reported by Ding et al. (2022) and are available in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA793059 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA793059), with run accessions SRR17382109–SRR17382348. The core input files supporting the analyses in this study are available as Data S1–S3 in Zenodo at https://doi.org/10.5281/zenodo.20768384. Consolidated supplementary tables are provided as Tables S1–S4, and supplementary diagnostic figures are provided as Figures S1–S7.

Conflict of interest

The authors declare that they have no conflict of interest.

Funding

This work was supported by the National Key Research and Development Program of China, Ministry of Science and Technology of the People's Republic of China (2022YFD2200204-2), and the Fundamental Research Funds for the Central Non-profit Research Institution of the Chinese Academy of Forestry (CAFYBB2025ZA011-1).

Authors' contributions

Conceptualization: Qifu Luan and Xianyin Ding. Methodology: Qifu Luan, Xianyin Ding, and Huijin Fan. Software and formal analysis: Xianyin Ding. Investigation and data curation: Xianyin Ding, Yue Zhang, Huijin Fan, Zifeng Tan, and Jingmin Jiang. Resources: Qifu Luan and Jingmin Jiang. Visualization: Xianyin Ding. Writing—original draft: Xianyin Ding. Writing—review and editing: Xianyin Ding, Qifu Luan, Yue Zhang, Huijin Fan, Zifeng Tan and Jingmin Jiang. Supervision and project administration: Qifu Luan and Jingmin Jiang.

Acknowledgements

The authors thank Wuchan Zhongda Changle Forest Farm Co., Ltd., Yuhang District, Hangzhou, Zhejiang Province, China, for providing valuable support during sample collection and storage.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9. 10.1038/75556.
2. Beaulieu J, Doerksen TK, MacKay J, Rainville A, Bousquet J. Genomic selection accuracies within and between environments and small breeding groups in white spruce. BMC Genomics. 2014;15:1048. 10.1186/1471-2164-15-1048.
3. Beaulieu J, Nadeau S, Ding C, Celedon JM, Azaiez A, Ritland C, Laverdière J, Deslauriers M, Adams G, Fullarton M, Bohlmann J, Lenz P, Bousquet J. Genomic selection for resistance to spruce budworm in white spruce and relationships with growth and wood quality traits. Evol Appl. 2020;13(10):2704–22. 10.1111/eva.13076.
4. Bellot P, de los Campos G, Perez-Enciso M. 2018. Can deep learning improve genomic prediction of complex human traits? Genetics, 210, 809–19. 10.1534/genetics.118.301298

5. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Ser B: Stat Methodol. 1995;57(1):289–300. 10.1111/j.2517-6161.1995.tb02031.x.
6. Bernardo R, Yu J. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 2007;47(3):1082–90. 10.2135/cropsci2006.11.0690.
7. Breiman L. Random forests. Mach Learn. 2001;45:5–32. 10.1023/A:1010933404324.
8. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 2018;103:338–48. 10.1016/j.ajhg.2018.07.015.
9. Calleja-Rodriguez A, Chen Z, Suontama M, Pan J, Wu HX. Genomic predictions with nonadditive effects improved estimates of additive effects and predictions of total genetic values in Pinus sylvestris. Front Plant Sci. 2021;12:666820. 10.3389/fpls.2021.666820.
10. Calleja-Rodriguez A, Pan J, Funda T, Chen Z, Baison J, Isik F, Abrahamsson S, Wu HX. Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. BMC Genomics. 2020;21:796. 10.1186/s12864-020-07188-4.
11. Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, pp. 785–794. 10.1145/2939672.2939785
12. Chen Z-Q, Baison J, Pan J, Karlsson B, Andersson B, Westin J, García-Gil MR, Wu HX. Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC Genomics. 2018;19:946. 10.1186/s12864-018-5256-y.
13. Chen Z-Q, Klingberg A, Hallingback HR, Wu HX. Preselection of QTL markers enhances accuracy of genomic selection in Norway spruce. BMC Genomics. 2023;24:147. 10.1186/s12864-023-09250-3.
14. Clark SA, van der Werf J. Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values. Volume 1019. Totowa, NJ: Humana; 2013. 10.1007/978-1-62703-447-0_13.
15. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. 10.1007/BF00994018.
16. Covarrubias-Pazaran G. Genome-assisted prediction of quantitative traits using the R package sommer. PLoS ONE. 2016;11:e0156744. 10.1371/journal.pone.0156744.
17. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–75. 10.1016/j.tplants.2017.08.011.
18. Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity. 2014;112(1):48–60. 10.1038/hdy.2013.16.
19. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE. 2008;3:e3395. 10.1371/journal.pone.0003395.
20. de los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res. 2010;92:295–308. 10.1017/S0016672310000285.
21. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193:327–45. 10.1534/genetics.112.143313.
22. Diao S, Ding X, Luan Q, Jiang J. A complete transcriptional landscape analysis of *Pinus elliottii* Engelm. using third-generation sequencing and comparative analysis in the *Pinus* phylogeny. Forests. 2019;10:942. 10.3390/f10110942.
23. Ding X, Diao S, Luan Q, Jiang J. Genetic analysis and elite tree selection of the main resin components of slash pine. Front Plant Sci. 2023;14:1079952. 10.3389/fpls.2023.1079952.
24. Ding X, Diao S, Luan Q, Wu HX, Zhang Y, Jiang J. A transcriptome-based association study of growth, wood quality, and oleoresin traits in a slash pine breeding population. PLoS Genet. 2022;18:e1010017. 10.1371/journal.pgen.1010017.
25. Ding X, Diao S, Zhang Y, Luan Q, Li Y, Jiang J, Wu HX. Insights into *PeERF168* gene in slash pine terpene biosynthesis: integrating high-throughput phenotyping, GWAS, and transgenic studies. Int J Biol Macromol. 2025a;300:139728. 10.1016/j.ijbiomac.2025.139728.
26. Ding X, Pelser PB, Xu C, Porth I, Cui M, El-Kassaby YA, Diao S, Luan Q, Li Y. Leveraging close-range UAV phenotyping and GWAS for enhanced understanding of slash pine growth dynamics. Inform Process Agric. 2025b;12:550–64. 10.1016/j.inpa.2025.07.002.
27. Ding X, Zhang Y, Sun J, Tan Z, Huang Q, Diao S, Wu Y, Luan Q, Jiang J. Genetic selection for growth, wood quality and resin traits of potential Slash pine for multiple industrial uses. Forestry Res. 2024;4:e023. 10.48130/forres-0024-0020.
28. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. 10.1093/bioinformatics/bts635.
29. Du Y, Jia N, Wang Y, Li R, Lu Y, Würschum T, Zhu X, Liu W. Benchmarking feature selection methods and prediction models for flowering time prediction in maize. Int J Mol Sci. 2026;27(1635). 10.3390/ijms27041635.
30. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011;4(3):250–5. 10.3835/plantgenome2011.08.0024.
31. Gamal El-Dien O, Ratcliffe B, Klapste J, Chen C, Porth I, El-Kassaby YA. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics. 2015;16:370. 10.1186/s12864-015-1597-y.
32. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63:3–42. 10.1007/s10994-006-6226-1.
33. Gianola D, van Kaam JBCHM. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;178:2289–303. 10.1534/genetics.107.084285.

34. Goddard ME, Hayes BJ. Genomic selection. J Anim Breed Genet. 2007;124:323–30. 10.1111/j.1439-0388.2007.00702.x.
35. Grattapaglia D. Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. Forests. 2022;13:1554. 10.3390/f13101554.
36. Grattapaglia D, Resende MDV. Genomic selection in forest tree breeding. Tree Genet Genomes. 2011;7:241–55. 10.1007/s11295-010-0328-4.
37. Grattapaglia D, Silva-Junior OB, Resende RT, Cappa EP, Müller BSF, Tan B, Isik F, Ratcliffe B, El-Kassaby YA. Quantitative genetics and genomics converge to accelerate forest tree breeding. Front Plant Sci. 2018;9:1693. 10.3389/fpls.2018.01693.
38. Gyllenstrand N, Clapham D, Kallman T, Lagercrantz U. A Norway spruce FLOWERING LOCUS T homolog is implicated in control of growth rhythm in conifers. Plant Physiol. 2007;144:248–57. 10.1104/pp.107.095802.
39. Habier D, Fernando RL, Dekkers JCM. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 2007;177:2389–97. 10.1534/genetics.107.081190.
40. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics. 2011;12:186. 10.1186/1471-2105-12-186.
41. Hazel LN. The genetic basis for constructing selection indexes. Genetics. 1943;28:476–90. 10.1093/genetics/28.6.476.
42. Heffner EL, Sorrells ME, Jannink J-L. Genomic selection for crop improvement. Crop Sci. 2009;49(1):1–12. 10.2135/cropsci2008.08.0512.
43. Heslot N, Yang H-P, Sorrells ME, Jannink J-L. Genomic selection in plant breeding: a comparison of models. Crop Sci. 2012;52(1):146–60. 10.2135/cropsci2011.06.0297.
44. Isik F. Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forest. 2014;45:379–401. 10.1007/s11056-014-9422-z.
45. Isik F, Li B. Rapid assessment of wood density of live trees using the Resistograph for selection in tree improvement programs. Can J For Res. 2003;33(12):2426–35. 10.1139/x03-176.
46. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics. 2010;9(2):166–77. 10.1093/bfgp/elq001.
47. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587–92. 10.1093/nar/gkac963.
48. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28(1):27–30. 10.1093/nar/28.1.27.
49. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–23. 10.1534/genetics.107.080101.
50. Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA. Genomic selection for forest tree improvement: methods, achievements and perspectives. Forests. 2020;11:1190. 10.3390/f11111190.
51. Lenz PRN, Beaulieu J, Mansfield SD, Clement S, Desponts M, Bousquet J. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC Genomics. 2017;18:335. 10.1186/s12864-017-3715-5.
52. Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L. Genomic selection in plant breeding. Adv Agron. 2011;110:77–123. 10.1016/B978-0-12-385531-2.00002-5.
53. Luo Z, Yu Y, Xiang J, Li F. Genomic selection using a subset of SNPs identified by genome-wide association analysis for disease resistance traits in aquaculture species. Aquaculture. 2021;539:736620. 10.1016/j.aquaculture.2021.736620.
54. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303. 10.1101/gr.107524.110.
55. Merrick LF, Burke AB, Chen X, Carter AH. Breeding with major and minor genes: genomic selection for quantitative disease resistance. Front Plant Sci. 2021;12:713667. 10.3389/fpls.2021.713667.
56. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29. 10.1093/genetics/157.4.1819.
57. Muranty H, Jorge V, Bastien C, Lepoittevin C, Bouffier L, Sanchez L. Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. Tree Genet Genomes. 2014;10:1491–510. 10.1007/s11295-014-0790-5.
58. Niu S, Li J, Bo W, Yang W, Zuccolo A, Giacomello S, Chen X, Han F, Yang J, Song Y, Nie Y, Zhou B, Wang P, Zuo Q, Zhang H, Ma J, Wang J, Wang L, Zhu Q, Zhao H, Liu Z, Zhang X, Liu T, Pei S, Li Z, Hu Y, Yang Y, Li W, Zan Y, Zhou L, Lin J, Yuan T, Li W, Li Y, Wei H, Wu HX. The Chinese pine genome and methylome unveil key features of conifer evolution. Cell. 2022;185(1):204–e217214. 10.1016/j.cell.2021.12.006.
59. P?rez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198:483–95. 10.1534/genetics.114.164442.
60. Piepho HP, Mohring J, Melchinger AE, Buchse A. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica. 2008;161:209–28. 10.1007/s10681-007-9449-8.
61. Plomion C, Bousquet J, Kole C, editors. Genetics, Genomics and Breeding of Conifers. Boca Raton: CRC; 2011. 10.1201/b11075.
62. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9. 10.1038/ng1847.
63. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75. 10.1086/519795.

64. Resende MFR, Muñoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M. Accuracy of genomic selection methods in a standard data set of loblolly pine. Genetics. 2012;190:1503–10. 10.1534/genetics.111.137026.
65. Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, Pappas GJ, Kilian A, Grattapaglia D. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol. 2012;194(1):116–28. 10.1111/j.1469-8137.2011.04038.x.
66. Ross RJ, editor. 2015. Nondestructive Evaluation of Wood: Second Edition. General Technical Report FPL-GTR-238. USDA Forest Service, Forest Products Laboratory, Madison, WI. 10.2737/FPL-GTR-238
67. Runcie DE, Cheng H. 2019. Pitfalls and remedies for cross validation with multi-trait genomic prediction methods. G3: Genes, Genomes, Genetics, 9, 3727–41. 10.1534/g3.119.400598
68. Sarinelli JM, Murphy JP, Tyagi P, Holland JB, Johnson JW, Mergoum M, Mason RE, Babar A, Harrison S, Sutton R, Griffey CA, Brown-Guedira G. Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. Theor Appl Genet. 2019;132:1247–61. 10.1007/s00122-019-03276-6.
69. Schulz-Streeck T, Ogutu JO, Piepho HP. 2011. Pre-selection of markers for genomic selection. BMC Proceedings 5, S12. 10.1186/1753-6561-5-S3-S12
70. Sharma U, Sankhyan HP, Kumari A, Thakur S, Thakur L, Mehta D, Sharma S, Sharma S, Sankhyan N. Genomic selection: a revolutionary approach for forest tree improvement in the wake of climate change. Euphytica. 2024;220(9). 10.1007/s10681-023-03263-5.
71. Spindel JE, Begum H, Akdemir D, Collard B, Redona E, Jannink JL, McCouch S. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity. 2016;116:395–408. 10.1038/hdy.2015.113.
72. Tan B, Grattapaglia D, Martins GS, Ferreira KZ, Sundberg B, Ingvarsson PK. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. BMC Plant Biol. 2017;17:110. 10.1186/s12870-017-1059-6.
73. Thistlethwaite FR, Ratcliffe B, Klapste J, Porth I, Chen C, Stoehr MU, El-Kassaby YA. Genomic prediction accuracies in space and time for height and wood density of Douglas-fir using exome capture as the genotyping platform. BMC Genomics. 2017;18:930. 10.1186/s12864-017-4258-5.
74. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23. 10.3168/jds.2007-0980.
75. Wientjes YCJ, Veerkamp RF, Calus MPL. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics. 2013;193:621–31. 10.1534/genetics.112.146290.
76. Wu H, Jin S, Xiang C, Tang J, Xian J, Zhang J, Zhao J, Feng X, Jiang D, Wu Y, Ding Y. GPS: harnessing data fusion strategies to improve the accuracy of machine learning-based genomic and phenotypic selection. Plant Commun. 2025;6(101416). 10.1016/j.xplc.2025.101416.
77. Wu Y, Ding X, Diao S, Huang Q, Shang G, Tan Z, Wu S, Hua X, He C, Luan Q, Chen Z-Q, Wu HX. Integrating GWAS-guided markers preselection with genomic selection enhances prediction of pulpwood-related traits in slash pine (Pinus elliottii Englem). BMC Plant Biol. 2026;911410.1186/s12870-026-09114-4.
78. Yin L, Zhang H, Zhou X, Yuan X, Zhao S, Li X, Liu X. KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. Genome Biol. 2020;21:146. 10.1186/s13059-020-02052-w.
79. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38(2):203–8. 10.1038/ng1702.
80. Zhang T, Sun X, Li J, Cui K. Application and development prospect of genomic selection breeding in coniferous trees. Planta. 2025;262(133). 10.1007/s00425-025-04849-0.
81. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, Li J, Simianer H. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS ONE. 2014;9:e93017. 10.1371/journal.pone.0093017.
82. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4. 10.1038/ng.2310.
83. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11(4):407–9. 10.1038/nmeth.2848.

Figures

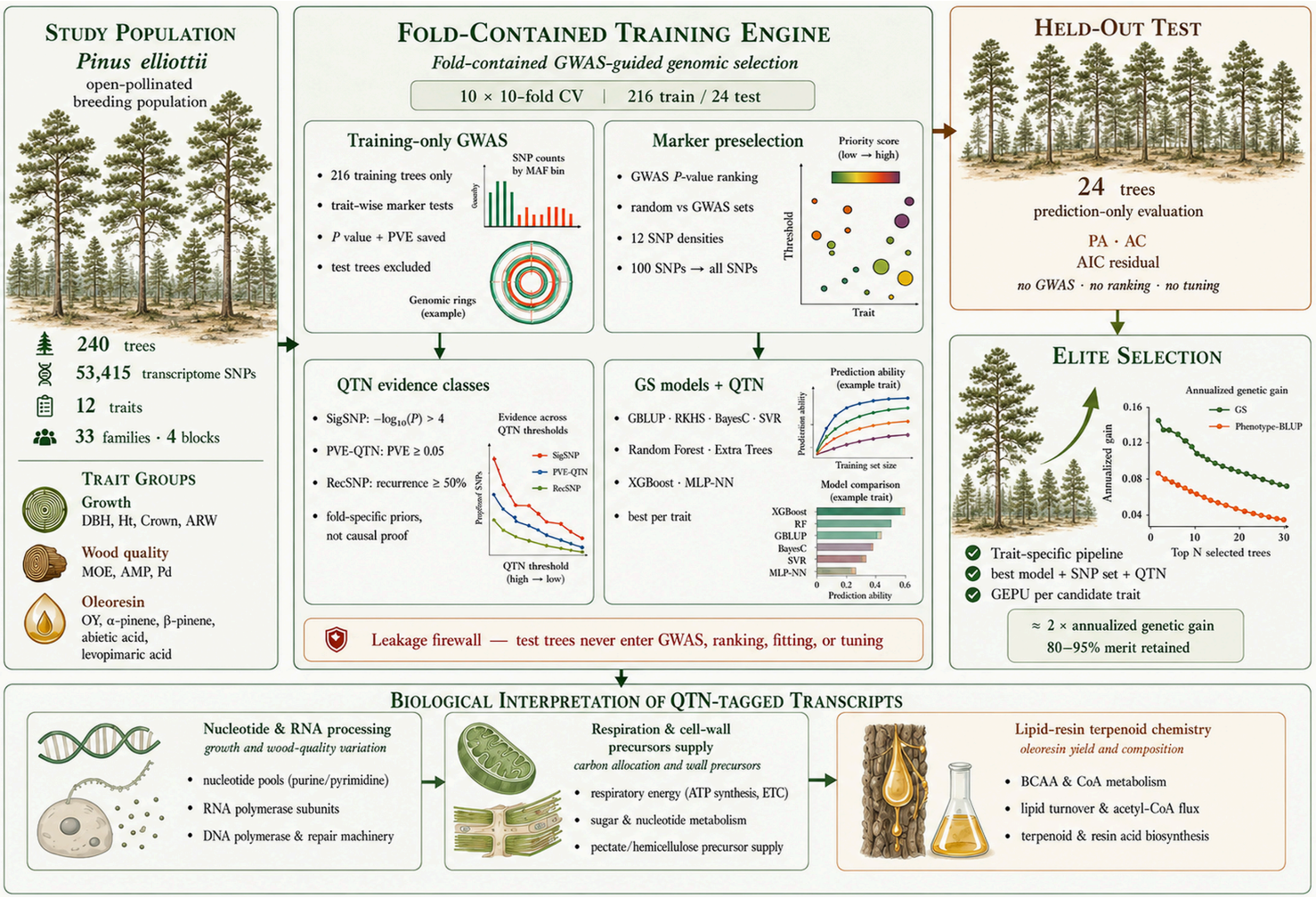


Figure 1

Fold-contained, GWAS-informed genomic selection and elite-tree selection workflow. The schematic summarizes the analytical pipeline connecting training-fold GWAS with genomic selection and multi-trait elite-tree selection. Within each outer fold, GWAS, SNP ranking, QTN detection, and model fitting used the training set. Held-out phenotypes were used for prediction evaluation, and the resulting cross-validation summaries also informed configuration selection. Fold-contained GWAS outputs were used to define nominally associated SNPs, high-PVE candidate QTNs, and recurrent SNPs, and these evidence classes were integrated with transcript functional annotations to propose candidate biological modules. Multiple SNP densities and eight GS models were benchmarked under random and GWAS-prioritized marker selection schemes, after which PVE- and recurrence-frequency-based QTN-covariate strategies were tested. Trait-specific cross-validation-selected pipelines were then used to predict GEBVs and to compare GS-based elite-tree selection with phenotype-BLUP selection on a common phenotype-BLUP index benchmark.

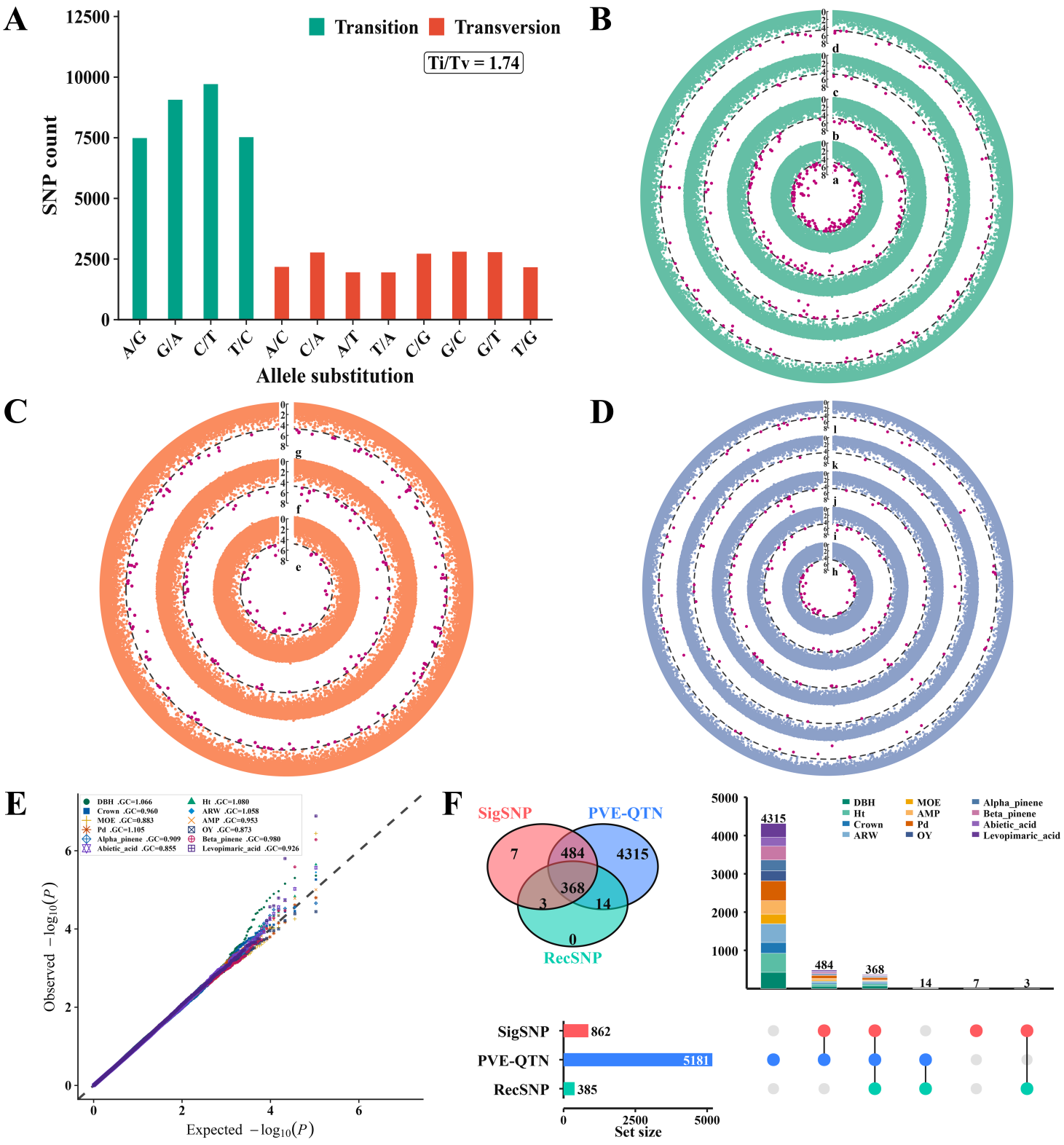


Figure 2

Genome-wide association landscape of fold-contained GWAS summaries. (A) SNP substitution spectrum of the transcriptome-derived marker panel: bars show the 12 directional allele-substitution classes retained after filtering, with fill colour distinguishing transitions from transversions; the Ti/Tv ratio summarizes the overall transition excess characteristic of expressed-sequence markers. (B–D) Trait-class circular Manhattan (CMplot) profiles for (B) growth traits (DBH, Ht, Crown, ARW), (C) wood-quality traits (MOE, AMP, Pd), and (D) oleoresin traits (OY, alpha-pinene, beta-pinene, abietic acid, levopimaric acid). Because the transcriptome-derived SNPs are not anchored to physical chromosomes, they are ordered as a single continuous marker sequence; each ring is one trait, radial height encodes −log10(P), the dark-grey dashed circle marks *P* = 1e-4 (−log10(P) = 4), and red points denote SNPs meeting the nominal association threshold. (E) Shared-coordinate QQ plot of fold-contained GWAS summaries across the 12 traits; all trait-specific curves are plotted in a single coordinate system against the bold null-expectation line, providing a descriptive comparison of the plotted distributions. (F) Overlap of GWAS-derived QTN evidence classes shown by an UpSet plot. Set-size bars count trait-SNP records supported by nominally associated SNPs (SigSNP, −log10(P) > 4), high-

PVE candidate markers (PVE-QTN, maximum PVE ≥ 0.05), or recurrent fold evidence (RecSNP, detection frequency ≥ 50%). Left horizontal bars show total set sizes, top vertical bars show intersection sizes, connected dots indicate the evidence-class combination, and stacked colours denote trait contributions. The 368 records shared by SigSNP, PVE-QTN, and RecSNP have support from all three operational evidence classes and were prioritized for downstream annotation and pathway interpretation.

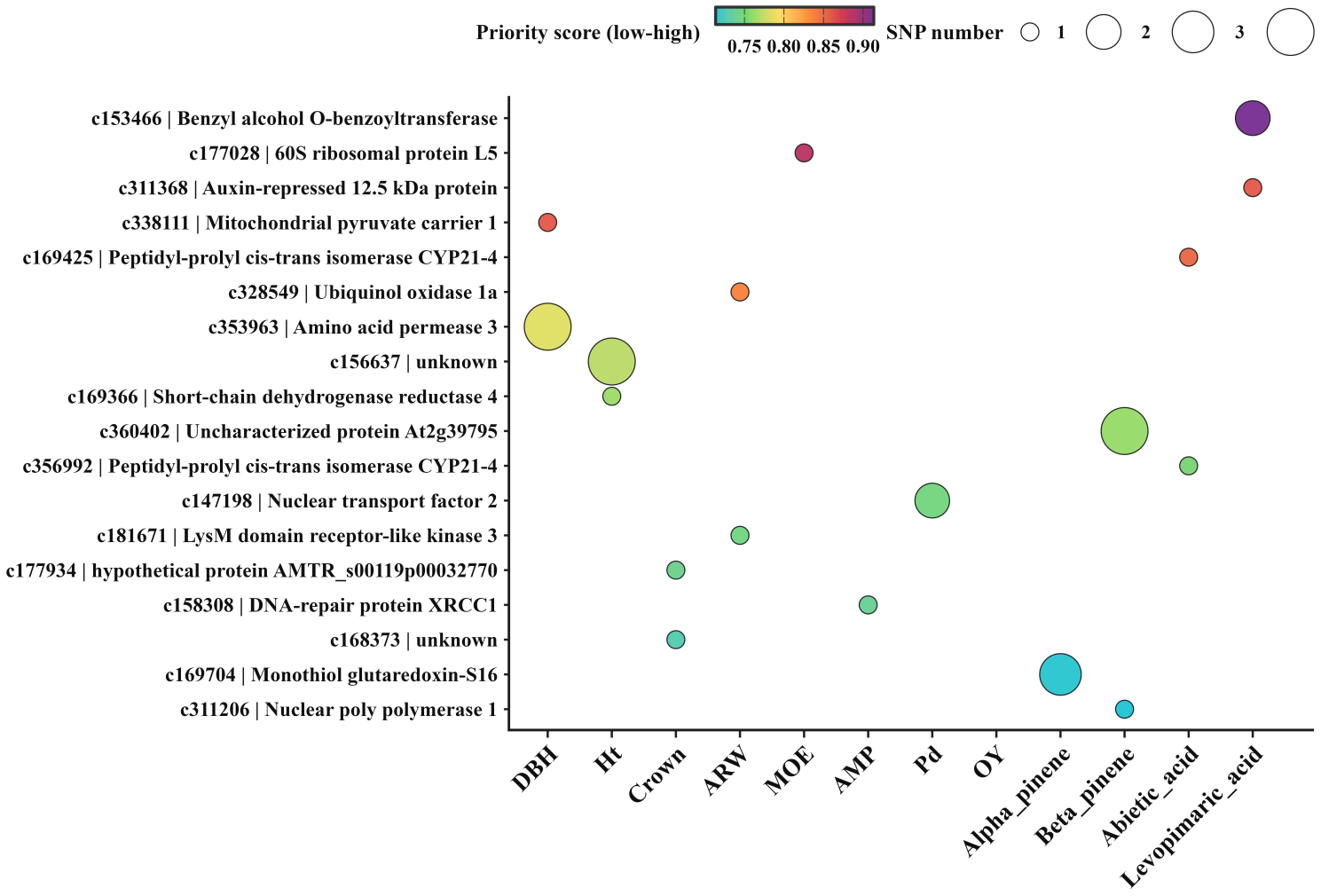


Figure 3

Candidate transcript prioritization of GWAS-derived QTN evidence. Evidence-weighted candidate-transcript priority map. Rows list gene identifiers with brief functional annotations drawn from the union of SigSNP, PVE-QTN, and RecSNP evidence; trait columns indicate where each candidate was implicated. Point size encodes the number of candidate SNPs per transcript-trait record, and fill colour encodes the evidence-weighted priority score.

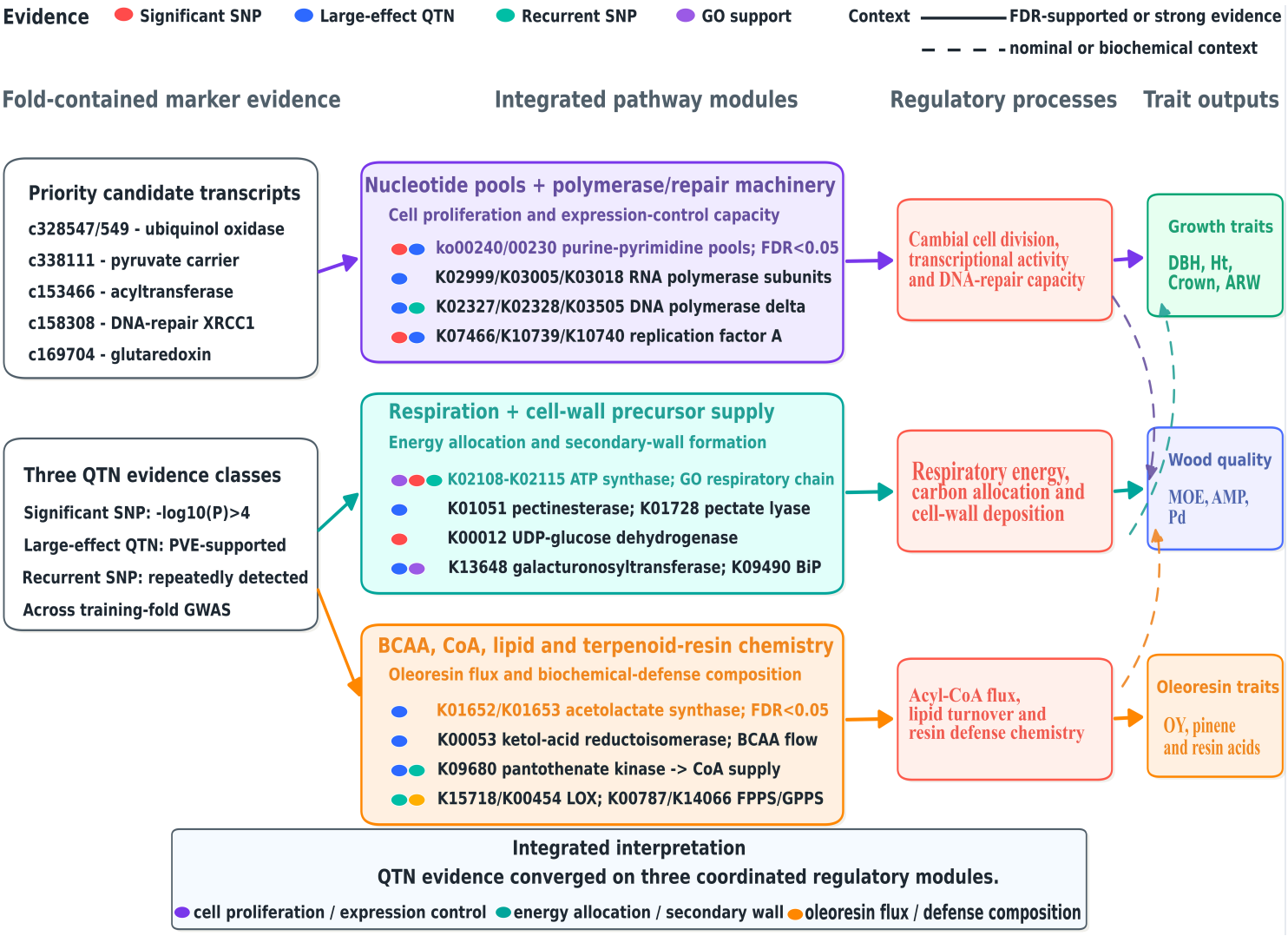


Figure 4

KEGG-informed pathway interpretation of GWAS-derived QTN evidence. This model integrates the three fold-contained marker-evidence classes with transcript functional annotation to summarize candidate biological routes potentially relevant to growth, wood-quality, and oleoresin traits. The KEGG-informed candidate pathway model derived from QTN-tagged transcript annotations contains three biologically interpretable modules: nucleotide pools with polymerase and repair machinery, respiratory energy with cell-wall precursor supply, and BCAA/CoA/lipid/terpenoid-resin chemistry. Solid edges indicate FDR-supported or strong evidence, whereas contextual edges indicate nominal enrichment or biochemical relevance. The model is intended as a hypothesis-generating annotation framework; the displayed connections are not experimentally established regulatory interactions or proof of causal pathways.

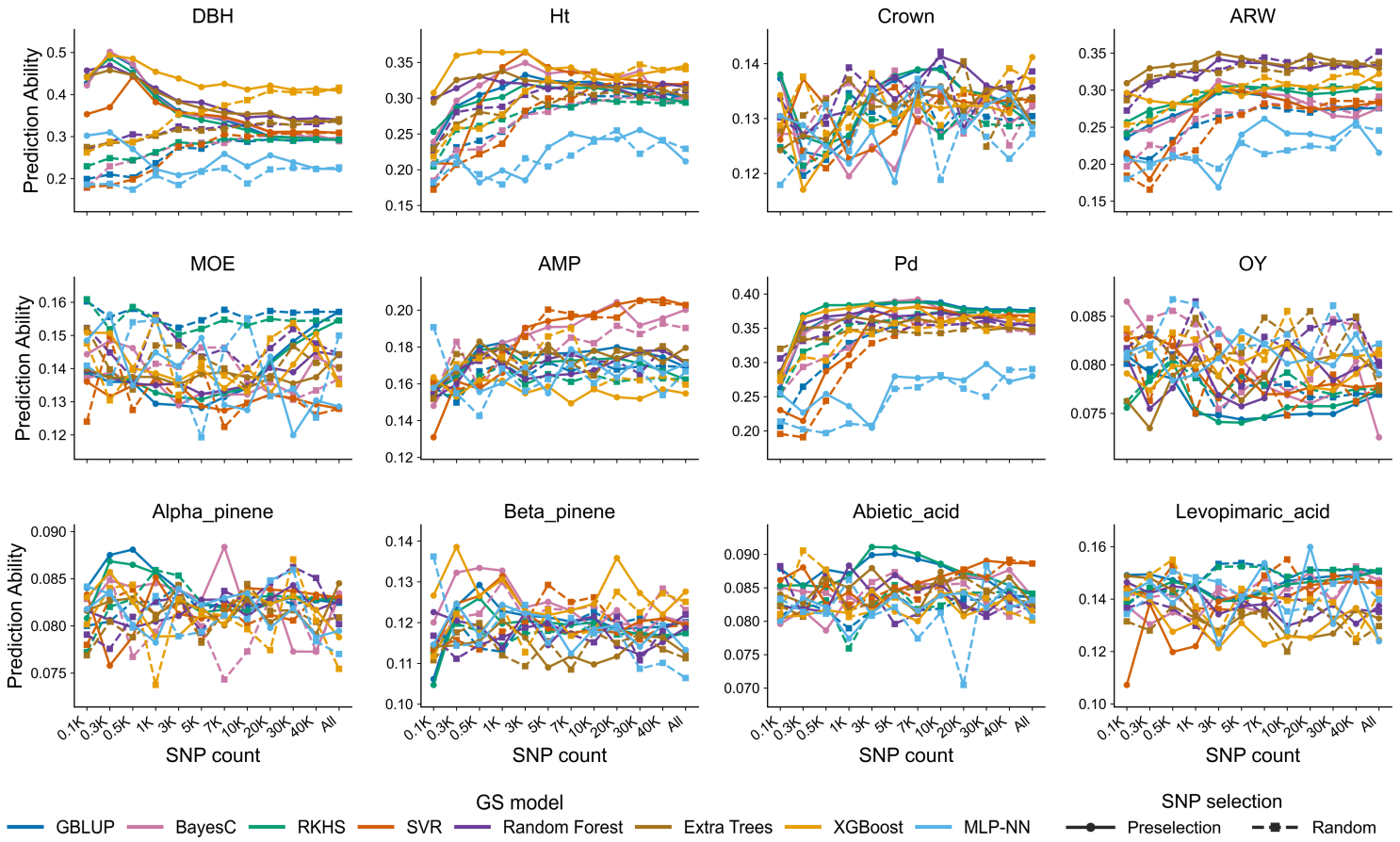


Figure 5

Prediction ability across SNP densities, GS models, and marker-selection schemes. Lines show mean prediction ability for each model across SNP-density levels, with solid lines for GWAS p-value preselection and dashed lines for random selection; colour denotes the model and each panel a trait. Prediction ability is the Pearson correlation between predicted GEBVs and observed test-set phenotypes.

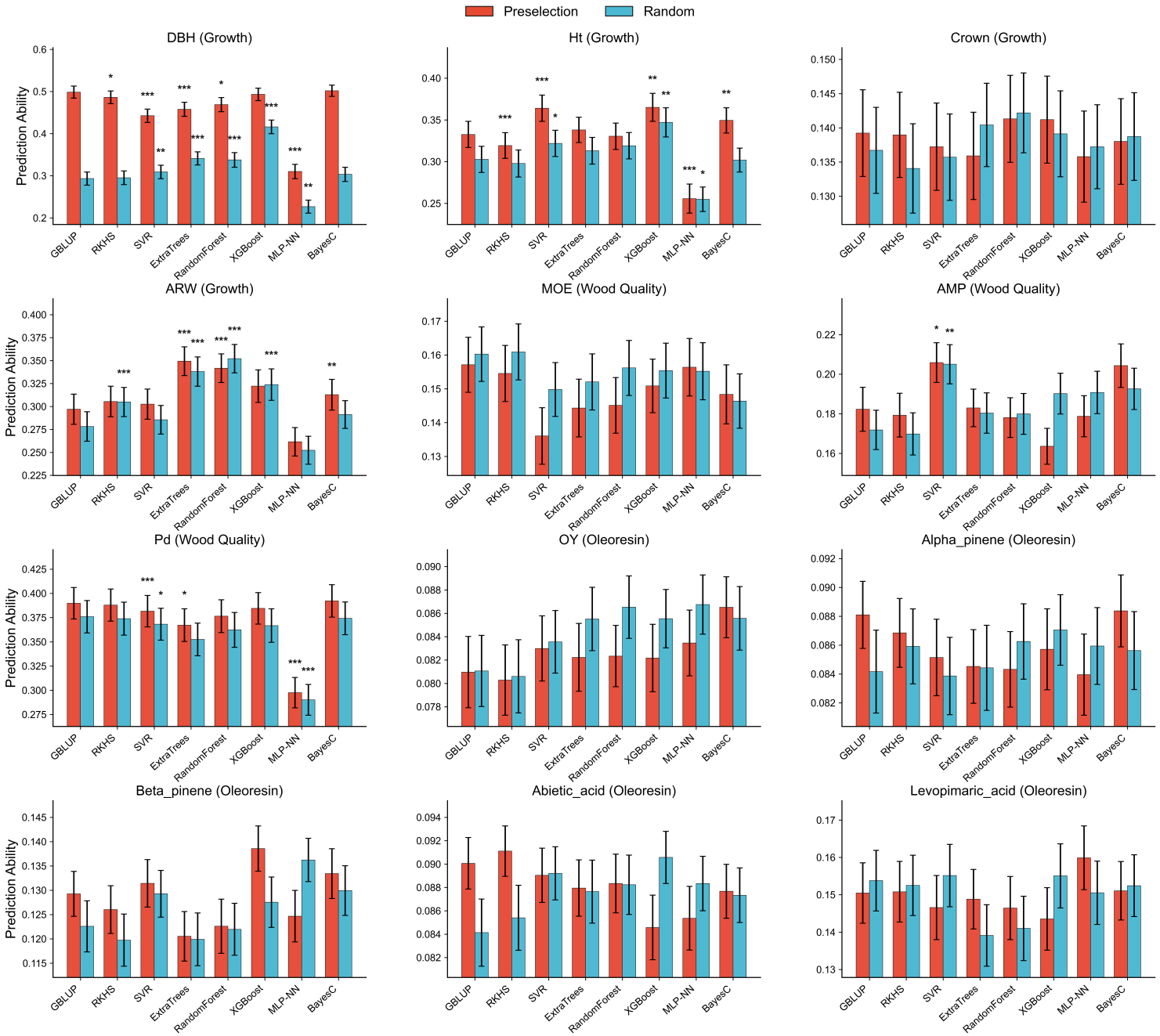


Figure 6

Prediction ability of alternative GS models relative to GBLUP. Bars show mean prediction ability across cross-validation folds for each model, grouped by trait, with error bars indicating fold-to-fold variation. Asterisks, where shown, mark exploratory paired comparisons with GBLUP after false-discovery-rate adjustment; because repeated cross-validation folds are not fully independent, these annotations are interpreted descriptively alongside PA effect sizes.

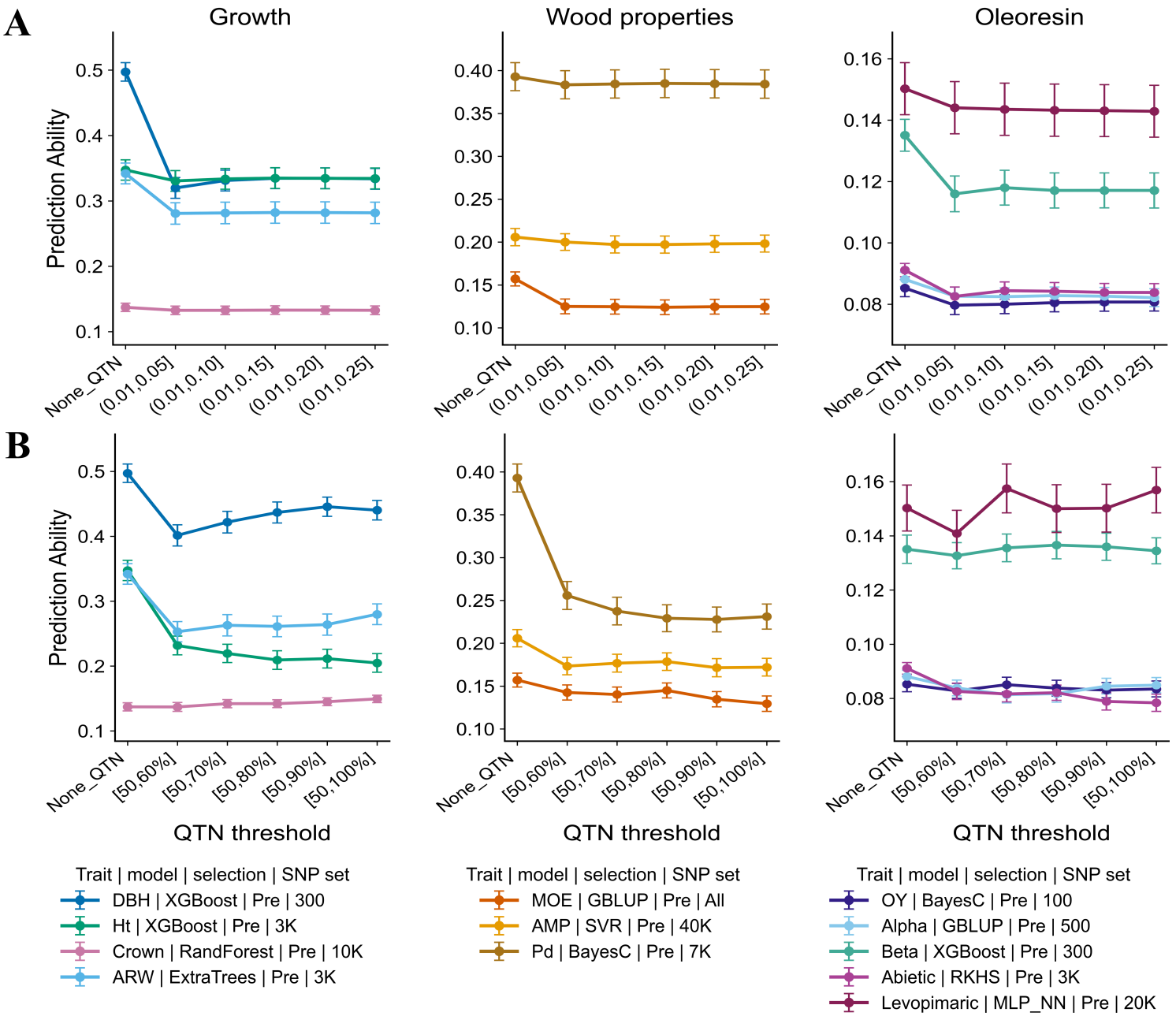


Figure 7

Effects of PVE-defined and fold-recurrence-frequency QTN-covariate strategies on prediction ability. (A) Cumulative PVE-defined QTN strategies and (B) cumulative fold-recurrence-frequency QTN strategies. Lines show mean prediction ability across folds for each trait and rule, and error bars show fold-to-fold variation. None_QTN is the per-trait no-QTN baseline, sharing the same selected GS model and SNP-density background as the QTN strategies but adding no QTN covariates.

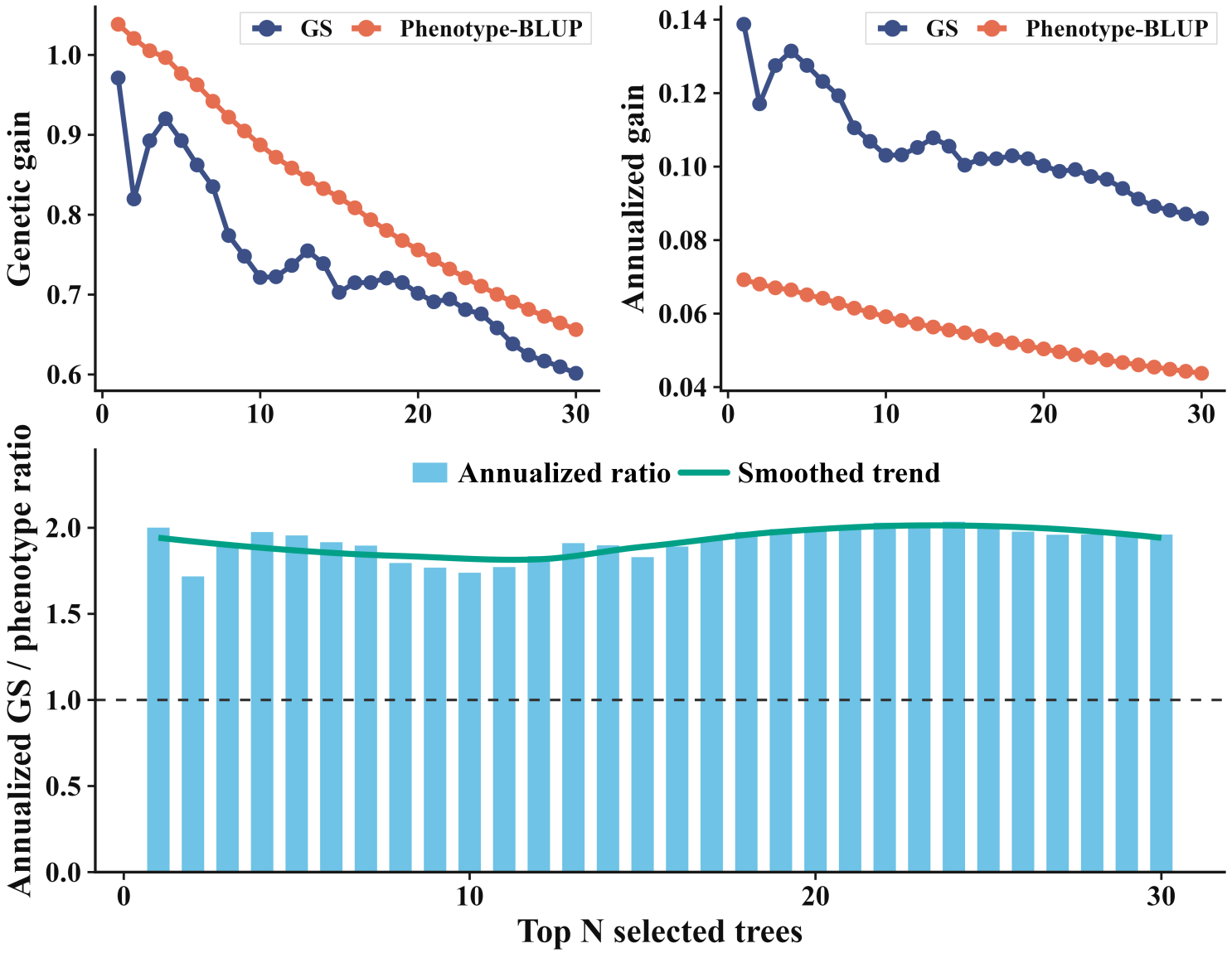


Figure 8

Benchmark index gain and projected annualized advantage of GS-based elite-tree selection from Top 1 to Top 30 under a common phenotype-BLUP benchmark. Upper panels show same-cycle and annualized benchmark index gain for phenotype-BLUP and GS selection. The lower panel shows the annualized GS-to-phenotype ratio as bars with a smoothed trend; values above 1 indicate a projected annualized GS advantage under the assumed 15-year traditional and 7-year GS cycles. These are index-based scenario estimates, not observed responses in a subsequent generation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesS1CoreGWASGSQC.xlsx
- SupplementaryTablesS3EliteSelectionGeneticGain.xlsx
- SupplementaryTablesS4GOKEGGFunctionalAnnotation.xlsx
- SupplementaryTablesS2QTNEvidenceandCandidates.xlsx
- SupplementaryFiguresS1toS7UPDATED.zip
- DataS1CorePhenotypeGenotypeCVSplits.zip
- DataS3FunctionalAnnotationGOKEGGInputs.zip
- DataS2GWASQTNEvidenceandVisualizationInputs.zip
- Supplementarymaterials.docx