

Genetic architecture and genomic prediction of stem forking across two generations of loblolly pine (*Pinus taeda* L.)

Pinar Guner

North Carolina State University

M. Nasir Shalizi

North Carolina State University

Trevor D. Walker

North Carolina State University

Fikret Isik

fisik@ncsu.edu

North Carolina State University

Research Article

Keywords: Threshold trait, QTL mapping, GWAS, single-step GBLUP, genomic prediction, cross-validation

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11243546/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Genetic architecture and genomic prediction of stem forking across two generations of loblolly pine (*Pinus taeda* L.)

Pinar Guner, M. Nasir Shalizi, Trevor D. Walker, Fikret Isik*

Cooperative Tree Improvement Program, Department of Forestry and Environmental Resources,
North Carolina State University, Raleigh, NC 27695, USA

*Corresponding author email: fisik@ncsu.edu

ORCID:

Pinar Guner: 0000-0002-2739-0614

M. Nasir Shalizi: 0000-0002-7773-741X

Trevor D. Walker: 0000-0002-7073-7898

Fikret Isik: 0000-0003-2833-7919

Abstract

Stem forking, assessed as a binary trait, is an important defect in *Pinus taeda* (loblolly pine) that reduces timber quality and product value. We investigated its genetic architecture and prediction accuracy using data from a two-generation breeding population evaluated in multi-environment trials. Genetic parameters were estimated using animal and family models with conditional penalized quasi-likelihood (PQL) and marginal PQL. We conducted genome-wide association and family-based linkage mapping to identify QTL. Individual-tree heritability on the latent scale was low (0.09-0.13), whereas family-mean heritability was high (0.93), indicating significant additive genetic control at the family mean level, but substantial environmental effect at the tree level. The probit PQL models produced slightly higher heritability estimates, whereas logistic and marginal models produced similar estimates. Association and linkage analyses identified a significant lineage-specific QTL on linkage group 3 but fitting it as a fixed covariate did not improve prediction accuracy. Across-generation family-level predictive ability was moderate (0.40). Forking had modest additive genetic correlations with tree height (0.36) and stem diameter (0.24). These results indicated that the underlying (unobserved) inheritance of stem forking in *P. taeda* is polygenic, while its expression among individual trees is strongly influenced by environmental and developmental factors. Response to genomic selection at the family level will be effective in reducing stem forking, while prediction of outcome for individual trees remains challenging.

Keywords: Threshold trait; QTL mapping; GWAS; single-step GBLUP; genomic prediction; cross-validation

Introduction

Pinus taeda L. (loblolly pine) is the most economically important forest tree species in the United States, with nearly one billion seedlings planted annually (McKeand 2019). Its rapid growth, broad adaptability, and suitability for large-scale plantation forestry have made it a foundation of the timber and pulpwood industries in the southern United States (Allen et al. 2005; Vance et al. 2010). Decades of selective breeding have improved growth, stem straightness, and resistance to fusiform rust, substantially increasing plantation productivity and economic value (Isik and McKeand 2019; McKeand et al. 2021). Nevertheless, stem forks and ramicorn branches remain persistent defects that limit the value of genetically improved plantations.

A stem fork occurs when the main stem divides into two or more codominant stems (**Supplementary Fig. 1**), whereas a ramicorn branch is a large branch with an unusually steep angle (Stovall et al. 2011). These defects introduce large knots, irregular grain, and poor stem form, reducing the proportion of trees suitable for high-value solid-wood products (Franklin and Callaham 1970; Espinoza 2009). They may also increase susceptibility to wind and ice damage (Jones and Wells 1969; Amateis and Burkhart 1996). Reducing the incidence of these defects is therefore important for improving both the resilience and economic value of *P. taeda* plantations.

Environmental factors such as frost damage or mechanical injury are well-documented triggers of forking (Franklin and Callaham 1970; Espinoza 2009; Walker et al. 2024). Genetic factors also contribute to stem forking, although their effects are generally more subtle (Xiong et al. 2010; Cumbie et al. 2012). Provenance studies of *P. taeda* have revealed variation in forking incidence among populations, with warmer coastal sources generally exhibiting higher forking rates than cold-hardy sources, especially when planted in colder environments where they are not adapted (Farjat et al. 2017; Walker et al. 2024; Trachta et al. 2025). However, previous studies often evaluated relatively few genetic entries or used experimental designs and analytical methods that limited accurate estimation of variance components. Consequently, the genetic architecture of forking, and the extent to which genetic susceptibility can be predicted across breeding generations, remains poorly understood.

Analyzing binary traits in plant and animal breeding, such as stem forking in *P. taeda*, presents statistical challenges. Generalized linear mixed models (GLMMs) provide a natural framework for analyzing binary traits because they accommodate non-Gaussian response distributions while accounting for fixed and random effects. Effects are estimated on the scale of the link function, while predictions can be transformed to the probability scale for biological interpretation. Many widely used GLMM implementations rely on penalized quasi-likelihood (PQL), also referred to as Schall's method or joint maximization (Breslow and Clayton 1993; Butler et al. 2023). PQL can severely underestimate variance components when trait incidence is below 10% or above 90%, or when animal models, also known as individual-tree models in forest genetics, are fitted to data in which each genetic individual has only one observation (Gilmour 2018; Butler et al.

2023). Incorrect GLMM specifications and inappropriate data structures can lead to misleading inferences from GLMM analyses (Bolker et al. 2009).

To address the shortcomings of PQL for animal models, Gilmour et al. (1985) proposed a marginal PQL formulation based on the expectation-maximization algorithm. The method retains the link function but calculates the working probabilities and weights from the fixed-effects component rather than from the full conditional predictor (Gilmour 2018). The recent implementation of marginal PQL in Echidna software provides a new opportunity to evaluate its performance for binary traits in large breeding populations (A. Gilmour, personal communication, June 2026). Here, we apply marginal PQL to stem forking in *P. taeda* and compare the resulting genetic parameter estimates with those obtained using conventional GLMM approaches.

Advances in DNA-marker technology and statistical modeling also provide opportunities to characterize the genetic architecture of stem forking and reduce its incidence through recurrent selection. Genomic selection is particularly well suited to polygenic traits (Montesinos-López et al. 2014; Bermann et al. 2021), because it can predict breeding values of offspring without phenotypes across generations (Isik et al. 2026). In contrast, genome-wide association studies (GWAS) can identify large-effect loci across pedigrees, but low allele frequencies, relatedness, and rapid linkage-disequilibrium decay in conifer populations may limit their power (Neale and Savolainen 2004; Parchman et al. 2012). Family-based linkage mapping provides a complementary approach for detecting major-effect quantitative trait loci (QTL) within informative pedigrees. A major-effect QTL for stem forking was previously identified in a single large full-sib family of *P. taeda* (Xiong et al. 2016), but its allele-substitution effect has not been validated in other pedigrees or breeding generations.

The novelty of this study lies in integrating a new binary-trait modeling approach with complementary genomic analyses in a large breeding population (19,165 trees) from two successive generations. Specifically, we combine regular and marginal PQL, population-wide association analysis, family-based linkage mapping, and cross-generation genomic prediction to determine whether stem forking is governed primarily by polygenic effects, major-effect loci, or a combination of both. This integrated design also allows us to test whether a QTL detected in one pedigree is transferable across the broader breeding population and whether explicitly accounting for that locus improves genomic prediction.

The objectives of this study were to (1) compare genetic-parameter estimates obtained using alternative models for a binary trait, (2) determine whether major-effect loci can be identified using GWAS and supported by within-family linkage analysis, and (3) evaluate genomic prediction across generations, including whether fitting a significant marker as a fixed effect improves predictive performance.

Materials and Methods

Plant material and phenotypic data

A two-generation population of *P. taeda* was used in this study. The generation 1 (G1) population was developed in the early 2000s and comprised 51 full-sib families produced from controlled crosses among 21 elite parents in three disconnected diallels, together with seven pollen-mix families treated as half-sib families in the pedigree-based analyses as common checks. In total, 58 families were derived from 24 parents. Approximately 2,400 progeny were clonally propagated through rooted cuttings and planted across eight field sites in the southern United States using alpha-cyclic incomplete block designs. Each clone was represented by up to one genetically identical copy, or ramet, per site, resulting in 1 to 8 ramets per clone across sites, with a median of approximately six ramets (Shalizi and Isik 2019). A total of 15,289 G1 trees were phenotyped at age six years after planting in 2017.

The generation 2 (G2) population was created in 2017 by selecting 73 clones from G1 and mating them to produce 67 full-sib families. In 2021, seedling progeny tests were established across four field sites using alpha-cyclic incomplete block designs. A total of 3,876 G2 seedlings were phenotyped at age four years after planting in 2025.

Stem forks and ramicorn branches were recorded and combined into a single binary phenotype. Trees exhibiting either a stem fork or a ramicorn branch were classified as affected (1), whereas trees exhibiting neither defect were classified as unaffected (0). Hereafter, this combined phenotype is referred to as forking. Tree height and diameter at breast height (DBH) were also assessed in the two populations. Across G1 and G2, phenotypic records were available for a total of 19,165 trees.

Genotyping and SNP filtering

Needle tissue was collected from G1 field trial clones in 2018 and from G2 progeny in 2021. Tissue samples were placed into individually labeled coin envelopes containing silica desiccant and allowed to dry at room temperature. Genomic DNA extractions were performed by the Forest Genomics Laboratory at the University of Florida.

Individuals were genotyped using the Pita50K Axiom™ SNP array (Applied Biosystems™), a medium-density platform developed for *P. taeda* (Caballero et al. 2021). The array comprises 46,439 high-quality SNP and indel markers distributed across the genome. All genotyping was performed by the North American Genomics Laboratory using Axiom™ Genotyping Services (Thermo Fisher Scientific).

Standard genotype calling and sample quality control were applied before downstream analyses. Markers were filtered to remove loci with call rate below 0.95, monomorphic loci, and loci with minor allele frequency (MAF) below 0.01. Samples with a call rate below 0.90 were excluded. After filtering, the final dataset used for analyses contained 15,848 SNPs. Missing genotypes

were imputed using the Wright method, assuming the Hardy-Weinberg equilibrium. A subset of 4,340 SNPs was anchored to the consensus linkage map (Lauer and Isik 2021) and was used for GWAS and linkage-group-based visualization and interpretation of association signals (**Supplementary Fig. 2**). Following all quality control steps, 2,067 genotyped trees from G1 and 1,936 genotyped trees from G2 were available for downstream analyses.

Statistical analyses

Genetic relationships and population structure

The additive genetic relationship matrix, $\mathbf{A}$, was constructed from pedigree records using Henderson's recursive path method (Henderson 1976). The inverse of this matrix was used to model the covariance among trees and to estimate the additive genetic variance among trees. For genotyped individuals, the genomic relationship matrix, $\mathbf{G}$, was calculated using VanRaden's allele-frequency method (VanRaden 2008):

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2 \sum p_i(1 - p_i)} \quad (1)$$

where $\mathbf{M}$ is an $n \times m$ matrix of genotype calls coded as 0 (homozygous major), 1 (heterozygous), or 2 (homozygous minor) for n individuals and m SNP markers, and $\mathbf{P}$ is an $n \times m$ matrix containing twice the minor allele frequency at each marker locus. The denominator is the sum of the expected genotype variances, $2p(1 - p)$, across loci.

Because not all trees were genotyped, pedigree- and marker-based relationship information was blended into the $\mathbf{H}$ matrix, in the single-step genomic BLUP (ssGBLUP) framework (Misztal et al. 2014). To construct $\mathbf{H}$, the pedigree relationship matrix $\mathbf{A}$ was partitioned into non-genotyped ($\mathbf{A}_{11}$) and genotyped ($\mathbf{A}_{22}$) individuals as follows.

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad (2)$$

The inverse of the single-step relationship matrix was then formed as

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \lambda(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) \end{bmatrix} \quad (3)$$

where λ is the weight factor. To improve numerical stability and reduce scale differences between pedigree- and marker-based relationships, $\mathbf{G}$ was tuned for compatibility with $\mathbf{A}_{22}$ and blended using a factor $\lambda = 0.90$ (VanRaden 2008; Christensen 2012). Relationship matrices were calculated using the AGHmatrix R package v3.0.1 (Amadeu et al. 2023).

To evaluate population structure in G1, principal component analysis (PCA) was conducted using genomic marker information. The first three principal components were retained based on the scree plot and included as covariates in GWAS to account for population structure (**Fig. 1**).

Single-step generalized linear mixed models

Generalized linear mixed models (GLMMs) were used to estimate variance components under both individual-tree (animal) and family (GCA) models using G1 individuals only. The G2 individuals were excluded from these analyses and were used exclusively in a cross-generation prediction analysis described later.

Binary forking observations were analyzed using GLMMs fitted by penalized quasi-likelihood (PQL). The general model in matrix format is

$$\mathbf{w} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (4)$$

where $\mathbf{w}$ is the working variable, $\mathbf{b}$ is the vector of fixed effects, including the intercept and site effects, $\mathbf{u}$ is the vector of additive genetic effects (tree effect), and $\mathbf{e}$ is the residual term. $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices, linking observations to levels of the fixed and random effects, respectively. The corresponding linear predictor $\boldsymbol{\eta}$ is

$$\boldsymbol{\eta} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} \quad (5)$$

Conditional GLMMs were fitted using logit and probit link functions in ASReml-R (Butler et al. 2023). Variance components were estimated on the latent, or liability, scale. Because the scale of the underlying liability is not identifiable from binary observations, the residual variance must be fixed to establish the scale. Under the logit link, the residual variance on the liability scale was fixed at $\pi^2/3 = 3.289$, corresponding to the variance of the standard logistic distribution. For the probit model, the residual variance was fixed at 1, corresponding to the variance of the standard normal distribution assumed for the underlying liability. The probit model therefore corresponds directly to the classical normally distributed liability-threshold model used in quantitative genetics (Falconer and Mackay 1996).

Penalized quasi-likelihood (PQL) estimation can produce biased fixed-effect and variance component estimates in GLMMs, particularly when sample size for random-effect levels is smaller than 10 (Breslow and Clayton 1993). To assess the sensitivity of variance component estimates to the standard conditional PQL formulation, we also fitted the marginal option (mPQL) implemented in Echidna as suggested by Arthur Gilmour (personal communication, June 2026). This marginal formulation retains the mixed-model structure but calculates the working probabilities and weights from the fixed-effects component rather than from the full conditional predictor including random effects. Thus, it should be interpreted as an alternative approximation within the GLMM/PQL framework, following the approach of Gilmour et al. (1985), rather than as a full likelihood or standard REML analysis of the observed binary responses. The working variable is defined as follows (Gilmour 2018).

$$\mathbf{w} = \boldsymbol{\eta} + \frac{y - \mu}{d} \quad (6)$$

where $\boldsymbol{\eta} = \mathbf{X}\mathbf{b}$ is the fixed-effect linear predictor on the logit scale, $\mu = \exp(\boldsymbol{\eta})/(1 + \exp(\boldsymbol{\eta}))$ is the corresponding predicted probability, and d is the derivative used to rescale the observed-scale residual to the working scale. Thus $(y - \mu)$ is the raw residual on the observed response scale, where y is the binary observation and μ is its expected probability. Dividing this residual by d rescales it to the linear-predictor, or working, scale. The marginal option was implemented in Echidna (Gilmour 2018), which is freely available for non-commercial use.

Individual-tree (animal) models

Individual-tree (animal) models were fitted using the conditional logit, conditional probit, and marginal PQL formulations described above. In these models, additive genetic effects were assumed to have expectations of $E(\mathbf{u}) = 0$ and variance $Var(\mathbf{u}) = \mathbf{H}\sigma_u^2$, where $\mathbf{H}$ is the genetic relationship matrix combining pedigree and genomic relationship information and σ_u^2 is the additive genetic variance.

Family (GCA) models

Family parental-effect models were fitted using the same conditional logit, conditional probit, and marginal PQL formulations. An overlaid design matrix was used to fit parental effects because, in conifer breeding populations, a parent can contribute as either female or male, as in a diallel mating design (Isik et al. 2017). In these models, the variance explained by the parental effect is a quarter of the additive genetic variance, which corresponds to a general combining ability (GCA) component. Compared with the individual-tree model, the number of observations per parental level was substantially larger, which can improve the stability of variance component estimation for binary traits. Because the family model included parental effects but no progeny-level genetic effect, it did not explicitly account for clonal replication. Ramets of the same clone were therefore treated as separate progeny sharing the same parents rather than as repeated observations of one genotype.

Validation of stem forking using ssGBLUP

Across-generation predictive ability was evaluated by comparing genomic estimated family-mean predictions from ssGBLUP models with observed family means in the G2 population. Phenotypes of the G2 population were masked during model fitting so that they did not contribute to the predictions. Observed family incidence was calculated as the proportion of forked trees within each family. G2 families contained between 12 and 37 genotyped and phenotyped trees.

Two ssGBLUP models with a logit link were trained using the G1 population, while the G2 population was used for validation. The models were fitted with and without the significant marker SNP505839 as a fixed covariate to evaluate whether including the marker improved

prediction. Both models included additive genetic effects based on the hybrid relationship matrix (**H**) and were fitted using ASReml-R (Butler et al. 2023). Predicted probabilities for G2 trees were obtained from the fitted linear predictor using the inverse-logit transformation.

Family-level predictive ability was calculated as the Pearson correlation between observed G2 family forking incidence and family-mean predicted probabilities. This statistic was interpreted as predictive ability rather than prediction accuracy because predictions were compared with observed phenotypic values rather than true breeding values (Isik et al. 2026).

Heritability calculations

For individual-tree models, narrow-sense (h_i^2) heritability for models with the logit link function was

$$h_{i_log}^2 = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3} \quad (7)$$

where σ_u^2 is the additive genetic variance. For the probit link function, the latent residual variance fixed at $\sigma_e^2 = 1$ was used. Under the marginal formulation, random-effect variances are absorbed into the liability-scale residual, so $\pi^2/3$ was used as the total phenotypic variance (Gilmour 2018).

$$h_{i_m}^2 = \frac{\sigma_u^2}{\pi^2/3} \quad (8)$$

For family models, the parental effect explains one-quarter of the additive genetic variance (Isik et al. 2017). Individual-tree narrow-sense heritability was therefore

$$h_i^2 = \frac{4\sigma_{GCA}^2}{2\sigma_{GCA}^2 + \sigma_e^2} \quad (9)$$

where $4\sigma_{GCA}^2$ is the additive genetic variance and $2\sigma_{GCA}^2$ is the contribution of the two parents to the phenotypic variance, σ_e^2 is the residual variance and varies according to the link function.

Family-mean heritability was calculated as follows.

$$h_f^2 = \frac{\sigma_{GCA}^2}{\left(\frac{p}{p-1}\right)\sigma_{GCA}^2 + \frac{\sigma_e^2}{(p-1)n}} \quad (10)$$

where p is the number of parents and n is the harmonic mean number of progeny per family. In the G1 population, $p = 24$, giving $p/(p-1) = 1.043$, and $n = 43$.

Multivariate ssGBLUP model

To estimate additive genetic correlations among forking, tree height, and diameter at breast height, a multivariate single-step genomic BLUP (ssGBLUP) model was fitted in Echidna version 2.08 (Gilmour 2018). Forking was modeled as a binary trait using a binomial distribution

with a logit link function, whereas tree height and diameter were modeled as continuous traits assuming a normal distribution. The multivariate model was specified as (Isik et al. 2017):

$$\mathbf{Y}_{nd} = (\mathbf{I}_d \otimes \mathbf{X}_{n \times (f+1)}) \mathbf{b}_{(f+1)d \times 1} + (\mathbf{I}_d \otimes \mathbf{Z}_{n \times r}) \mathbf{u}_{rd \times 1} + \mathbf{e}_{nd} \quad (11)$$

where $\mathbf{Y}_{nd}$ is the $nd \times 1$ vector of observations for $d = 3$ traits (forking, height, and diameter) and n is the number of observations. $\mathbf{I}_d$ is the identity matrix with dimension $d \times d$, $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices for fixed and additive genetic effects, respectively, and $\otimes$ denotes the Kronecker product. $f + 1$ is the number of fixed-effect coefficients (including the intercept), and r is the number of additive genetic effects. The vectors $\mathbf{b}$, $\mathbf{u}$, and $\mathbf{e}$ represent the fixed effects, additive genetic effects, and residuals, respectively. The additive genetic effects were assumed to follow

$$\mathbf{u} \sim N(0, \mathbf{G}_a \otimes \mathbf{H})$$

where $\mathbf{H}$ is the hybrid genetic relationship matrix and $\mathbf{G}_a$ is the $d \times d$ additive genetic variance-covariance matrix among forking, height, and diameter, defined as

$$\mathbf{G}_a = \begin{bmatrix} \sigma_F^2 & \sigma_{FH} & \sigma_{FD} \\ \sigma_{HF} & \sigma_H^2 & \sigma_{HD} \\ \sigma_{DF} & \sigma_{DH} & \sigma_D^2 \end{bmatrix} \quad (12)$$

where F , H , and D denote forking, tree height, and diameter at breast height, respectively. The diagonal elements are the trait-specific additive genetic variances, and the off-diagonal elements are the additive genetic covariances between trait pairs. Working-scale residuals were assumed to follow

$$\mathbf{e} \sim N(0, \mathbf{R} \otimes \mathbf{I}_n) \quad (13)$$

where $\mathbf{R}$ is the residual variance-covariance matrix among traits and $\mathbf{I}_n$ is an $n \times n$ identity matrix. The *vpredict* function in Echidna was used to obtain additive genetic correlations with approximate standard errors (Gilmour 2018).

Genome-wide association analysis

Genome-wide association analysis in the G1 population was conducted in two stages. In the first stage, spatially adjusted clone means for forking were estimated using a GLMM with a logit link. Clone identity and test site were fitted as fixed effects, while row and column positions within each trial were fitted as random effects to account for spatial heterogeneity. The model was fitted using ASReml-R (Butler et al. 2023). Treating clone effects as fixed avoided introducing shrinkage before marker effects and polygenic background were evaluated in the second stage (Holland and Piepho 2024). Adjusted clone means were estimated on the logit scale and

transformed to predicted probabilities using the inverse-logit function. These probabilities were used as phenotypic values in the second-stage GWAS.

In the second stage, marker-trait associations were evaluated in GAPIT3 using the multiple-locus mixed model (MLMM), Fixed and Random Model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) methods (Tang et al. 2016; Wang and Zhang 2021). Population structure was controlled using the first three principal components, and marker-derived kinship was incorporated according to the requirements of each method. Using three association methods allowed the consistency of marker associations to be evaluated under different approaches to controlling population structure, relatedness, and marker cofactors. Genome-wide significance was evaluated using a Bonferroni threshold at $\alpha = 0.05$. The Benjamini-Hochberg procedure with $q \leq 0.10$ was also applied as a secondary false discovery rate criterion. Quantile-quantile plots were used to evaluate model calibration, and Manhattan plots were used to visualize marker-trait associations.

QTL mapping in large families

QTL mapping was performed using 256 progeny from six full-sib families sharing N11210 as a common parent. Clone-level fixed-effect estimates for forking were used as pseudo-phenotypes. Marker genotypes were coded as A, H, and B, and genotype probabilities were calculated using **calc.genoprob** in the R/qtl package (Broman and Sen 2009). A genome-wide single-QTL scan was conducted using Haley-Knott regression implemented in **scanone**, with mate family included as a fixed covariate. The genome-wide 5% significance threshold was determined using 1,000 permutation tests. A 95% Bayesian credible interval was estimated for the significant QTL peak using **bayesint**.

Lineage-specific genomic prediction

Following GWAS and QTL mapping, descriptive summaries of forking incidence by SNP505839 genotype were examined across G1 pedigrees to explore whether the marker association differed among genetic backgrounds. These summaries suggested that the association was concentrated among descendants of N11210. SNP505839 was therefore evaluated using a separate binomial ssGBLUP model with a logit link fitted only to G1 phenotypic records. The model included test site and a fixed indicator for N11210-descendant status. SNP dosage was fitted as a fixed, lineage-specific allele-substitution covariate among descendants of N11210. Row and column effects nested within test site and the polygenic additive genetic effect based on the H matrix were fitted as random effects. The marker contribution to the linear predictor was defined as follows.

$$m_i = \begin{cases} \beta x_i, & \text{if tree } i \text{ is a descendant of N11210,} \\ 0, & \text{otherwise,} \end{cases} \quad (14)$$

where m_i is the marker contribution to the linear predictor, x_i is the SNP dosage coded as 0, 1, or 2 for TT, TC, and CC, respectively, and β is the allele substitution effect. The lineage-specific allele-substitution effect was evaluated using a 1-df Wald test. Because the N11210 lineage was identified through descriptive examination of the same G1 data, this analysis was considered post hoc, and the resulting Wald test was interpreted as exploratory. The odds ratio was calculated as $\exp(\beta)$, and the 95% confidence interval was calculated as $\exp(\beta \pm 1.96 SE)$. The model was fitted using ASReml-R (Butler et al. 2023). For visualization, fitted probabilities were averaged across ramets within each G1 clone before being summarized by SNP505839 genotype.

Evaluation of predictive performance

Individual-tree discrimination in G2 was quantified using the area under the receiver operating characteristic curve (AUC), calculated from predicted probabilities and observed binary outcomes with the pROC package in R (Robin et al. 2011). AUC represents the probability that a randomly selected forked tree receives a higher predicted probability than a randomly selected non-forked tree. Family-level predictive ability was quantified as the Pearson correlation between observed G2 family forking incidence and family-mean predicted probabilities. As a secondary analysis, AUC was calculated separately within each of the 23 marker-segregating G2 families containing both phenotypic classes.

Genomic localization of significant SNPs

To infer the genomic position of the significant SNP associated with forking, we aligned SNP flanking sequences to the reference assembly using local BLAST. Flanking sequences were queried against the *P. taeda* Pita v2.01 genome FASTA from the TreeGenes FTP repository, which provides scaffold-level genomic context. The reference assembly remains fragmented at the scaffold level; nevertheless, it enabled scaffold anchoring of marker sequences for future annotation and candidate-gene follow-up (Zimin et al. 2017).

Local databases were constructed from the downloaded genome FASTA using *makeblastdb* (NCBI BLAST), and queries were aligned with *blastn* using default scoring parameters. For each SNP, the top-scoring alignments (lowest E-value and highest bit score) were used to assign scaffold identifiers and approximate genomic coordinates. These scaffold placements were used as scaffold anchors for future annotation and validation.

Results

Forking incidence and genetic parameters

Overall forking incidence across the G1 and G2 populations was 43.5%. Forking incidence varied substantially among test sites, ranging from 14% to 64% across the eight G1 sites and from 25% to 72% across the four G2 sites.

Individual-tree narrow-sense heritability estimates ranged from 0.09 to 0.13 across six models (**Table 1**). Estimates from individual-tree models were slightly lower than those from family models and had lower standard errors. Probit models produced slightly higher heritability estimates than the standard logit and marginal logit models, which were similar. Family-mean heritability was 0.93 across models.

Predicted breeding values on the probability scale from ssGBLUP individual-tree models with different link functions were nearly perfectly correlated (**Fig. 2**). Forking showed a moderate positive additive genetic correlation with tree height ($r_a = 0.36$) and stem diameter ($r_a = 0.24$), while height and stem diameter had a strong correlation of 0.81 (**Table 2**).

The ssGBLUP model trained using the G1 population was used to predict forking incidence in the G2 population. At the family level, the correlation between observed family mean forking incidence and genomic estimated breeding values of families was 0.40 (**Fig. 3**).

Genome-wide association analysis

PCA revealed three broad genetic groups corresponding to the three disconnected diallel groups in G1 (**Fig. 1**). Marker SNP505839 on linkage group 3 was significantly ($P \leq 9.62 \times 10^{-11}$) associated with forking in all three GAPIT3 models and passed both the Bonferroni and Benjamini-Hochberg false discovery rate criteria (**Fig. 4**). This marker had a minor allele frequency of 0.08 and was selected for downstream evaluation. Among descendants of N11210, each additional copy of the C allele at the marker locus was associated with 45% higher odds of forking (**Fig. 5**). The estimated odds ratio was 1.45, with a 95% confidence interval of 1.17-1.80.

QTL mapping in large families

Joint linkage mapping of 256 progeny from six families sharing N11210 as a common parent detected a significant QTL associated with forking on linkage group 3 (**Fig. 6**). The maximum LOD score was approximately 4.5 and exceeded the genome-wide significance threshold of 3.9. The QTL region overlapped the linkage-group interval containing SNP505839, which was also identified in the GWAS. The 95% Bayesian credible interval surrounding the QTL peak spanned approximately 3.2 cM, providing a relatively narrow interval for further validation across the N11210 lineage.

Effect of fitting the QTL-linked marker

Following the identification of SNP505839 through GWAS and QTL mapping, the marker was included as a fixed covariate in the ssGBLUP model. Including SNP505839 did not improve family-level predictive ability compared with the model without the marker (**Fig. 3**).

Individual-tree prediction was weak in G2, with an overall AUC of 0.55, where 0.50 indicates no better discrimination than chance. AUC was also calculated within 23 marker-segregating families (**Fig. 7**). Under ssGBLUP, family-specific AUC values ranged from 0.45 to 0.84, with a mean of 0.59. Adding SNP505839 did not consistently improve prediction: mean AUC

decreased to 0.58, increasing in 8 families and decreasing in 13, and remaining unchanged. The predictive power in families with N11210 ancestry did not consistently improve when including the marker as a fixed effect.

Scaffold localization

BLAST analysis of the flanking sequence for the significant SNP505839 marker (5'-TGAAGAAGAAGAGGAAAAAGAATCTGTGGTTTTCTC/TCCATAGGGGTTTCCAC-3') identified its location within the fragmented *P. taeda* genome assembly (Pita v2.01). The SNP aligned with 100% sequence identity to scaffold APFE030399899.1 (6,574 bp), part of BioProject PRJNA174450, with a low E-value (2×10^{-12}). This alignment anchored the marker to a specific genomic scaffold but did not yield functional annotation or gene-level information due to the incomplete nature of the *P. taeda* genome assembly.

Discussion

This study combined PQL-based GLMMs with logit and probit link functions, QTL detection, and single-step genomic prediction to evaluate stem forking across two related generations of *P. taeda* (loblolly pine) breeding populations. This integrated approach allowed us to assess both the genetic basis of forking and the genetic evaluation for selection. The results showed that at the individual tree level, the trait outcome (forking or no forking) is mostly determined by environmental factors, while at the family-mean level additive genetic effects were significant, reaching a family-mean heritability of 0.93. Similar low individual-tree heritability and high family-mean heritability were reported in the species (Xiong et al. 2010; Cumbie et al. 2012). Significant family differences in stem forking have also been reported in several pine species (Ladrach and Lambeth 1991; Moura et al. 1998; Codesido and Fernández-López 2008). Our results extend this evidence across two-generation large breeding populations and support the use of family-based evaluation for selection against forking.

In this study, forking incidence varied considerably among sites in both generations, consistent with an important environmental contribution to trait expression. In a large clonally replicated study of loblolly pine, forking at the tree level was attributed to frost injury, leader breakage or dieback, and variation in soil nutrient status (Xiong et al. 2010). Multiple flushing and loss of apical dominance can also cause lateral shoots to outgrow the terminal shoot and produce a fork-like stem form (Cline and Harrington 2007). Stovall et al. (2011) reported higher forking incidence in fertilized plots than in control plots in a replicated clonal trial of *P. taeda*. These environmental and developmental effects explain the limited value of a single tree's phenotype as an indicator of its genetic tendency to fork.

A methodological contribution of this study was the evaluation of a marginal PQL for stem forking to assess a potential bias (Gilmour et al. 1985). The bias is especially relevant to individual-tree (animal) models because each tree has only one binary observation (Gilmour 2018; Butler et al. 2023). Studies of binary traits in livestock have shown that alternative models

can yield different heritability estimates while producing similar predictions or rankings (Martinez et al. 2005; Saborío-Montero et al. 2018). Applying the marginal PQL procedure in this study provided an assessment of the sensitivity of heritability estimates to the estimation method. The marginal PQL did not increase heritability estimates relative to the standard logit model, while estimates from the probit model were slightly higher. The large sample size, information contributed by pedigree, and a high incidence of stem forking may have reduced the potential for PQL bias. Nevertheless, predicted probabilities from the three model formulations were almost perfectly correlated, indicating that model choice had little effect on the ranking of trees or families.

A further contribution of this study was the identification of a QTL supported by both association and linkage analyses within the N11210 pedigree. Among N11210 descendants, each additional copy of the C allele at SNP505839 was associated with higher log-odds of forking under the additive allele-substitution model. Previous QTL mapping in *P. taeda* also identified loci affecting stem forking within a clonally replicated full-sib family (Xiong et al. 2016). Our results provide additional evidence that QTL for forking can be identified within individual pedigrees. Because the association was detected only among N11210 descendants, the SNP505839 marker should be considered a candidate marker pending further validation.

We tested the practical value of the SNP505839 marker by including it as a fixed-effect covariate in a single-step genomic prediction model. Incorporating the marker did not consistently improve predictive ability, demonstrating that a QTL does not necessarily provide additional predictive value at the population level. Lauer et al. (2021) also reported a large-effect, family-specific QTL for tree height in this population with inconsistent results. Differences in marker-QTL linkage phase, rapid decay of linkage disequilibrium, and low allele frequency of relevant alleles could limit the transferability of marker associations across pedigrees and generations. Single-step GBLUP is efficient for selection of complex traits, including traits following the threshold model of quantitative genetics, since the method traces all loci with large or small effects across the genome using dense marker coverage, pedigree, and phenotype (Misztal et al. 2020).

Modest positive additive genetic correlations of stem forking with tree height and stem diameter suggested that selection for growth alone could increase forking liability in *P. taeda*. Cumbie et al. (2012) also reported a positive genetic correlation between forking and growth in a multi-environment study of loblolly pine. Walker et al. (2024) found that families with higher breeding values for forking subsequently had more stem defects and ice-storm damage, along with lower predicted sawtimber potential. These findings support evaluating forking alongside growth, stem form, and disease resistance in a multi-trait selection framework.

These findings have direct implications for *P. taeda* breeding. Although stem forking is mostly influenced by environmental factors at the individual-tree level, differences among family means are partly under additive genetic control, indicating that family-level truncation selection based

on genomic selection would be effective to reduce the overall frequency of stem forking in breeding populations.

Author contributions

PG conducted the data analysis and prepared the initial draft of the manuscript. FI guided the overall data analyses and contributed writing the first draft. FI, TDW and MNS contributed to the study design. TDW and MNS helped with data analyses, produced figures and edited the manuscript. All authors contributed to and approved the final version.

Acknowledgements

The authors thank the members of the North Carolina State University Cooperative Tree Improvement Program for establishing and evaluating the genetic field trials and for providing financial and in-kind support.

Competing interests

The authors declare no competing interests.

Funding

This research was funded by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) project “Genomic Selection in Forest Trees: Beyond the Proof of Concept” (project director: Fikret Isik, Award number: 2019–67013-29169) and North Carolina State University Cooperative Tree Improvement Program.

Data Archiving Statement

The phenotype, pedigree, inverse of genetic relationship matrices, and SNP marker data used in analyses are archived on DRYAD repository.

References

- Amadeu RR, Garcia AAF, Munoz PR, Ferrão LFV (2023) AGHmatrix: genetic relationship matrices in R. *Bioinformatics* 39(7):btad445. [https://doi: 10.1093/bioinformatics/btad445](https://doi.org/10.1093/bioinformatics/btad445).
- Amateis RL, Burkhart HE (1996) Impact of heavy glaze in a loblolly pine spacing trial. *Southern Journal of Applied Forestry* 20:151–155. <https://doi.org/10.1093/sjaf/20.3.151>
- Bermann M, Legarra A, Hollifield MK, et al (2021) Validation of single-step GBLUP genomic predictions from threshold models using the linear regression method: An application in chicken mortality. *J Anim Breed Genet* 138:4–13. <https://doi.org/10.1111/jbgs.12507>

- Bolker BM, Brooks ME, Clark CJ, et al (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24:127–135. <https://doi.org/10.1016/j.tree.2008.10.008>
- Breslow NE, Clayton DG (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88:9–25. <https://doi.org/10.1080/01621459.1993.10594284>
- Broman KW, Sen S (2009) *A Guide to QTL Mapping with R/qlt*. Springer Science & Business Media
- Butler BJ D, Cullis B, Gilmour A, et al (2023) *ASReml-R Reference Manual Version 4.2*
- Caballero M, Lauer E, Bennett J, et al (2021) Toward genomic selection in *Pinus taeda*: Integrating resources to support array design in a complex conifer genome. *Appl Plant Sci* 9:e11439. <https://doi.org/10.1002/aps3.11439>
- Christensen OF (2012) Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet Sel Evol* 44:37. <https://doi.org/10.1186/1297-9686-44-37>
- Cline M, Harrington C (2007) Apical dominance and apical control in multiple flushing of temperate woody species. *Canadian Journal of Forest Research* 37: 74-83. <https://doi.org/10.1139/x06-218>
- Codesido V, Fernández-López J (2008) Juvenile genetic parameter estimates for vigour, stem form, branching habit and survival in three radiata pine (*Pinus radiata* D.Don) progeny tests in Galicia, NW Spain. *Eur J Forest Res* 127:315–325. <https://doi.org/10.1007/s10342-008-0207-9>
- Cumbie WP, Isik F, McKeand SE (2012) Genetic improvement of sawtimber potential in loblolly pine. *Forest Science* 58:168–177. <https://doi.org/10.5849/forsci.09-060>
- Espinoza JA (2009) Genetic and nutritional effects on stem sinuosity in loblolly pine. PhD dissertation 102
- Falconer DS, Mackay FCT (1996) *Introduction to Quantitative Genetics*. pp 464–464
- Farjat AE, Chamblee AK, Isik F, et al (2017) Variation among loblolly pine seed sources across diverse environments in the southeastern United States. *Forest Science* 63:39–48. <https://doi.org/10.5849/forsci.15-107>
- Franklin EC, Callaham RZ (1970) Multinodality, branching, and forking in lodgepole pine (*Pinus contorta* var. *murrayana* Engelm.). *Silvae Genet* 19:180–184
- Gilmour A (2018) Echidna mixed model software. In: *Proceedings of the World Congress Applied to Livestock Genetics*

- Gilmour AR, Anderson RD, RAE AL (1985) The analysis of binomial data by a generalized linear mixed model. *Biometrika* 72:593–599. <https://doi.org/10.1093/biomet/72.3.593>
- Henderson CR (1976) A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics* 32:69–83. <https://doi.org/10.2307/2529339>
- Holland JB, Piepho H-P (2024) Don't BLUP Twice. *G3 Genes|Genomes|Genetics* Volume 14:
- Isik F, Holland J, Maltecca C (2017) *Genetic Data Analysis for Plant and Animal Breeding*. Springer International Publishing, 400 p.
- Isik F, McKeand SE (2019) Fourth cycle breeding and testing strategy for *Pinus taeda* in the NC State University Cooperative Tree Improvement Program. *Tree Genetics & Genomes* 15:1–12. <https://doi.org/10.1007/s11295-019-1377-y>
- Isik F, Shalizi MN, Walker TD (2026) Genomic selection validated across two generations of loblolly pine breeding. *G3 Genes|Genomes|Genetics* jkag122. <https://doi.org/10.1093/g3journal/jkag122>
- Jones EP, Wells OO (1969) Ice damage in a Georgia planting of loblolly pine from different seed sources. Res Note SE-126 Asheville, NC: US Department of Agriculture, Forest Service, Southeastern Forest Experiment Station 4 p 126:
- Ladrach WE, Lambeth C (1991) Growth and heritability estimates for a seven-year-old open-pollinated *Pinus patula* progeny test in Colombia. *Silvae genet* 40:169–173
- Lauer E, Holland J, Isik F (2021) Prediction ability of genome-wide markers in *Pinus taeda* L. within and between population is affected by relatedness to the training population and trait genetic architecture. *G3 (Bethesda)* 12:jkab405. <https://doi.org/10.1093/g3journal/jkab405>
- Lauer E, Isik F (2021) Major QTL confer race-nonspecific resistance in the co-evolved *Cronartium quercuum* f. *sp. fusiforme*–*Pinus taeda* pathosystem. *Heredity* 127:288–299. <https://doi.org/10.1038/s41437-021-00451-8>
- Lee Allen H, Fox TR, Campbell RG (2005) What is Ahead for Intensive Pine Plantation Silviculture in the South? *Southern Journal of Applied Forestry* 29:62–69. <https://doi.org/10.1093/sjaf/29.2.62>
- Martinez GE, Koch RM, Cundiff LV, et al (2005) Genetic parameters for stayability, stayability at calving, and stayability at weaning to specified ages for Hereford cows. *J Anim Sci* 83:2033–2042. <https://doi.org/10.2527/2005.8392033x>
- McKeand SE (2019) The Evolution of a Seedling Market for Genetically Improved Loblolly Pine in the Southern United States. *Journal of Forestry* 117:293–301. <https://doi.org/10.1093/jofore/fvz006>

- McKeand SE, Payn KG, Heine AJ, Abt RC (2021) Economic Significance of Continued Improvement of Loblolly Pine Genetics and Its Efficient Deployment to Landowners in the Southern United States. *Journal of Forestry* 119:62–72. <https://doi.org/10.1093/jofore/fvaa044>
- Misztal I, Legarra A, Aguilar I (2014) Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science* 97:3943–3952. <https://doi.org/10.3168/jds.2013-7752>
- Misztal I, Legarra A, Aguilar I, others (2020) Current status of genomic evaluation and single-step methods. *Animal Frontiers*
- Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, et al (2014) Threshold Models for Genome-Enabled Prediction of Ordinal Categorical Traits in Plant Breeding. *G3 (Bethesda)* 5:291–300. <https://doi.org/10.1534/g3.114.016188>
- Moura VPG, Dvorak WS, Hodge GR (1998) Provenance and family variation of *Pinus oocarpa* grown in the Brazilian cerrado. *Forest Ecology and Management* 109:315–322. [https://doi.org/10.1016/S0378-1127\(98\)00265-5](https://doi.org/10.1016/S0378-1127(98)00265-5)
- Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. *Trends in Plant Science* 9:325–330. <https://doi.org/10.1016/j.tplants.2004.05.006>
- Parchman TL, Gompert Z, Mudge J, et al (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology* 21:2991–3005. <https://doi.org/10.1111/j.1365-294X.2012.05513.x>
- Robin X, Turck N, Hainard A, et al (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. <https://doi.org/10.1186/1471-2105-12-77>
- Saborío-Montero A, Vargas-Leitón B, Romero-Zúñiga JJ, Camacho-Sandoval J (2018) Additive genetic and heterosis effects for milk fever in a population of Jersey, Holstein × Jersey, and Holstein cattle under grazing conditions. *Journal of Dairy Science* 101:9128–9134. <https://doi.org/10.3168/jds.2017-14234>
- Shalizi MN, Isik F (2019) Genetic parameter estimates and G×E interaction in a large cloned population of *Pinus taeda* L. *Tree Genetics & Genomes* 15:46. <https://doi.org/10.1007/s11295-019-1352-7>
- Stovall JP, Carlson CA, Seiler JR, et al (2011) Growth and stem quality responses to fertilizer application by 21 loblolly pine clones in the Virginia Piedmont. *Forest Ecology and Management* 261:362–372. <https://doi.org/10.1016/j.foreco.2010.10.018>
- Tang Y, Liu X, Wang J, et al (2016) GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *The Plant Genome* 9:plantgenome2015.11.0120. <https://doi.org/10.3835/plantgenome2015.11.0120>

- Trachta M, McKeand SE, Walker TD (2025) A comparison of non-improved provenances and improved checklots from a *Pinus taeda* breeding program. *Tree Genetics & Genomes* 21:24. <https://doi.org/10.1007/s11295-025-01708-7>
- Vance ED, Maguire DA, Zalesny RSJ (2010) Research strategies for increasing productivity of intensively managed forest plantations. *Journal of Forestry* 108: 183-192 108:183–192
- VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>
- Walker TD, Maynor JA, Isik F, et al (2024) Stem Defect Rates and Ice Storm Damage for Families of *Pinus taeda* from Coastal and Piedmont Provenances Planted on a North Carolina Piedmont Site. *for sci* 70:311–322. <https://doi.org/10.1093/forsci/fxae016>
- Wang J, Zhang Z (2021) GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction
- Xiong J, Isik F, McKeand S, Whetten R (2010) Genetic Variation of Stem Forking in Loblolly Pine. *Forest Science* 56:429–436. <https://doi.org/10.1093/forestscience/56.5.429>
- Xiong JS, McKeand SE, Isik F, et al (2016) Quantitative trait loci influencing forking defects in an outbred pedigree of loblolly pine. *BMC Genetics* 17:138. <https://doi.org/10.1186/s12863-016-0446-6>
- Zimin AV, Stevens KA, Crepeau MW, et al (2017) An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience* 6:giw016. <https://doi.org/10.1093/gigascience/giw016>

Tables and Figures

Table 1 Variance components, individual-tree narrow-sense heritability, and family-mean heritability from ssGBLUP models fitted using logit, probit, and marginal-logit link functions. Estimates are followed by approximate standard errors where available.

A) Individual-tree (animal) model

Estimates	Logit	Probit	Marginal
Additive variance	0.315 ± 0.0447	0.117 ± 0.0164	0.299 ± 0.0430
Residual variance	3.289 ± NA	1.000 ± NA	3.289 ± NA
Phenotypic variance	3.604 ± 0.0447	1.117 ± 0.0164	3.289 ± NA
Individual-tree heritability	0.09 ± 0.011	0.10 ± 0.013	0.09 ± 0.013

B) Family (GCA) model

Estimates	Logit	Probit	Marginal
Additive variance	0.379 ± 0.1103	0.138 ± 0.0400	0.365 ± 0.1056
Residual variance	$3.289 \pm \text{NA}$	$1.000 \pm \text{NA}$	$3.289 \pm \text{NA}$
Phenotypic variance	3.668 ± 0.0276	1.069 ± 0.0200	$3.289 \pm \text{NA}$
Individual-tree heritability	0.11 ± 0.030	0.13 ± 0.035	0.11 ± 0.032
Family mean heritability	0.93 ± 0.009	0.93 ± 0.007	0.93 ± 0.009

Table 2 Additive genetic variances (diagonal), additive genetic correlations (above the diagonal), and genetic covariances (below the diagonal) among stem forking, height, and diameter at breast height (DBH) estimated from the multivariate ssGBLUP model. Forking was modeled using a binomial distribution with a logit link, whereas height and DBH were modeled using Gaussian distributions with identity links. Estimates were followed by approximate standard errors.

	Forking	Height	DBH
Forking	0.374 ± 0.048	0.36 ± 0.072	0.24 ± 0.072
Height	0.299 ± 0.064	1.890 ± 0.169	0.81 ± 0.021
DBH	0.055 ± 0.017	0.422 ± 0.042	0.144 ± 0.012

Fig. 1 Population structure analysis of the G1. (A) Heatmap of realized additive genomic relationships among 2,067 individuals, estimated using 4,340 SNPs. Individuals are grouped and ordered by family along both axes. Darker blue squares along the diagonal indicate within-family relatedness (~ 0.5 for full-sibs), while lighter off-diagonal blocks reflect relatedness among families (~ 0.25 for half-sibs or less). (B) Principal component analysis (PCA) shows the first two principal components (PC1 and PC2), which explain 5% and 4% of the genetic variation, respectively. Individuals are grouped based on pairwise genetic distances. (C) Neighbor-joining tree based on the genomic relationship matrix, revealing three major genetic clusters. (D) Scree plot showing the variance explained by the first ten PCs; the optimal number of PCs retained ($k = 3$) is marked in red based on the elbow criterion.

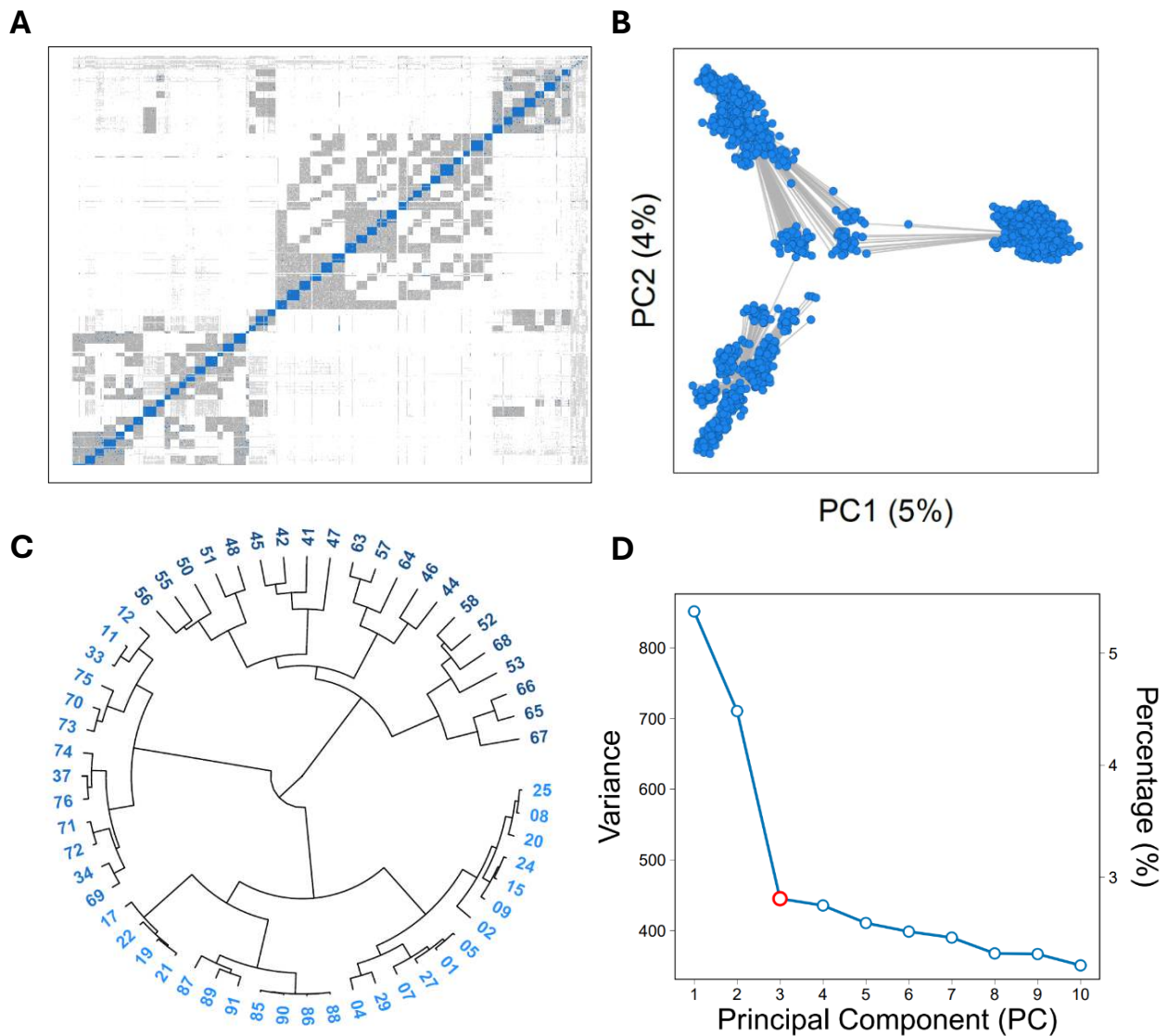


Fig. 2 Predicted breeding values (probability of forking) obtained from individual-tree ssGBLUP models fitted with logit, probit, and marginal formulations. Histograms in diagonal panels show the distributions of predicted probability values for individual trees, lower-triangle panels show pairwise scatterplots among model formulations, and upper-triangle panels show Pearson correlations (r). The near-unity correlations indicate that predicted probabilities were highly consistent among the three individual-tree model formulations.

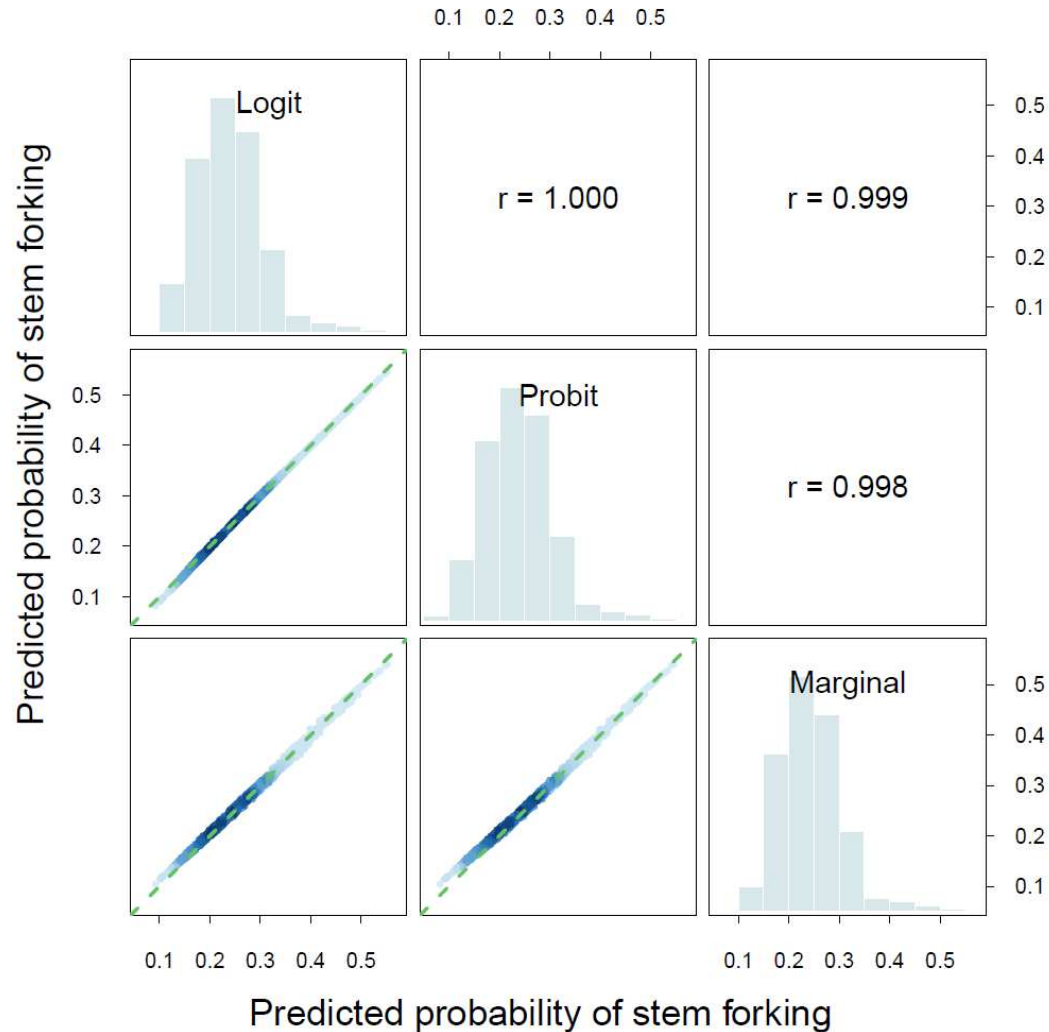


Fig. 3 Validation of G2 family-level prediction for stem forking. Observed forking incidence in G2 full-sib families, represented by 12 to 37 genotyped and phenotyped trees per family, was compared with predicted family-mean forking probabilities from models trained using G1 phenotypes and genotypes, while G2 phenotypes were masked. Each point represents one G2 full-sib family. The upper panel shows ssGBLUP predictions, and the lower panel shows ssGBLUP with SNP505839 fitted as a fixed effect. Solid lines show simple linear regression fit, and Pearson correlations (r) are shown within panels. Family-level predictive ability was modest, and including SNP505839 did not improve prediction relative to ssGBLUP alone.

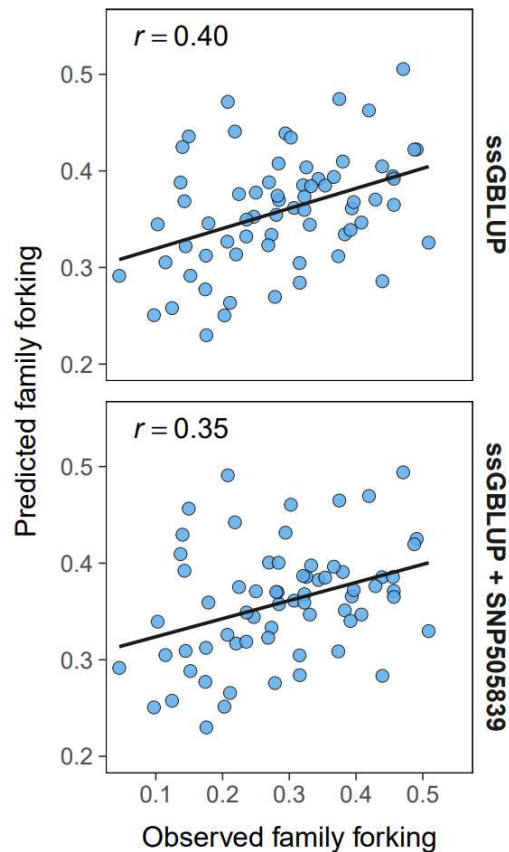


Fig. 4 Genome-wide association analysis of stem forking using the FarmCPU method, with a comparative evaluation of multi-locus models. Manhattan plot displaying the $-\log_{10}$ p-values of SNP associations across 12 linkage groups. The solid horizontal line denotes the Benjamini-Hochberg false discovery rate threshold ($q = 0.10$), and the dashed line indicates the Bonferroni-corrected genome-wide significance threshold ($\alpha = 0.05$). Quantile-quantile plot comparing observed versus expected $-\log_{10}(p)$ values for Fixed and random model Circulating Probability Unification (FarmCPU; blue), Multi-Locus Mixed Model (MLMM; orange), and Bayesian Information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK; green). Deviation from the null expectation is primarily confined to the extreme tail, consistent with a limited number of true association signals after accounting for relatedness and population structure.

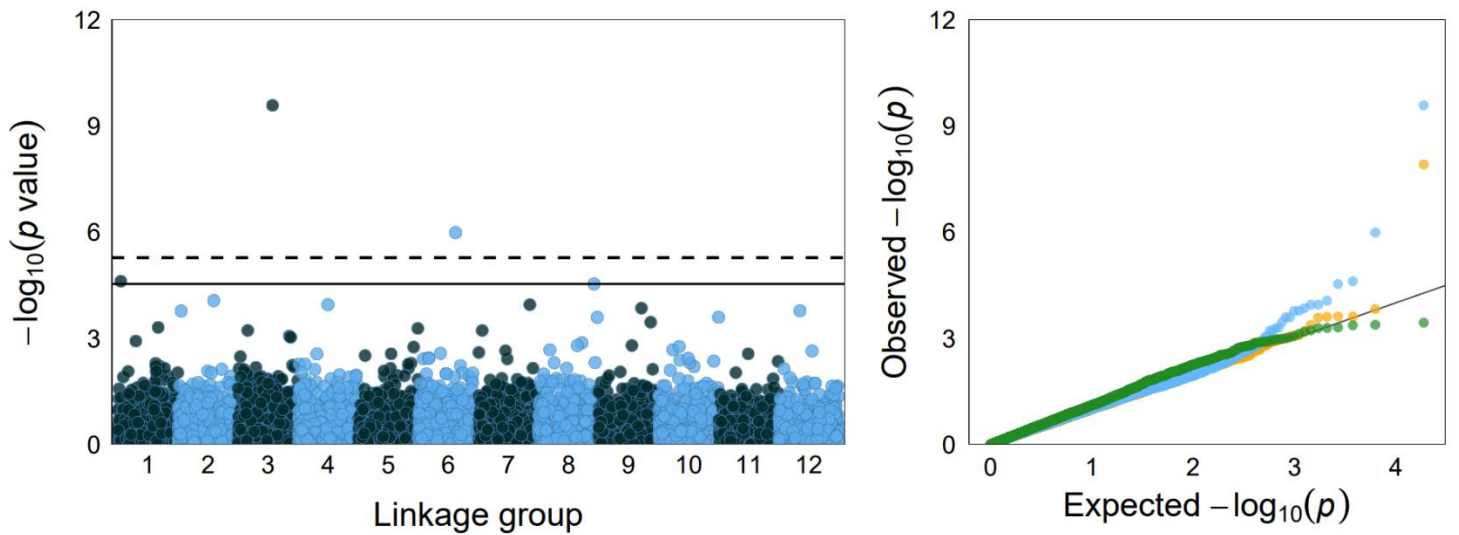


Fig. 5 Clone-level predicted probabilities of forking for marker SNP505839 within the tree N11210 descendants in the G1 population. Violin plots show predictions from an allele-substitution model using additive dosage coding, TT = 0, TC = 1, and CC = 2, shown in blue. Boxplots summarize the distribution of clone-level predicted probabilities within each genotype class, and red points indicate genotype-class means. The number of clones in each genotype class is shown below the violins. The additive allele-substitution effect was significant.

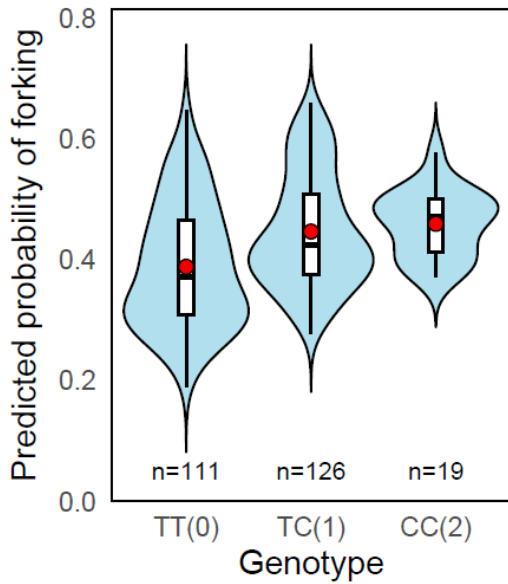


Fig. 6 QTL mapping of stem forking in 256 progeny from six full-sib families sharing N11210 as a common parent. The genome-wide scan was conducted using Haley-Knott regression. The dashed horizontal line is the genome-wide 5% significance threshold determined from 1,000 permutation tests ($\text{LOD} = 3.9$). Each point represents the LOD score at a marker position across the 12 linkage groups. A significant QTL peak was detected on linkage group 3.

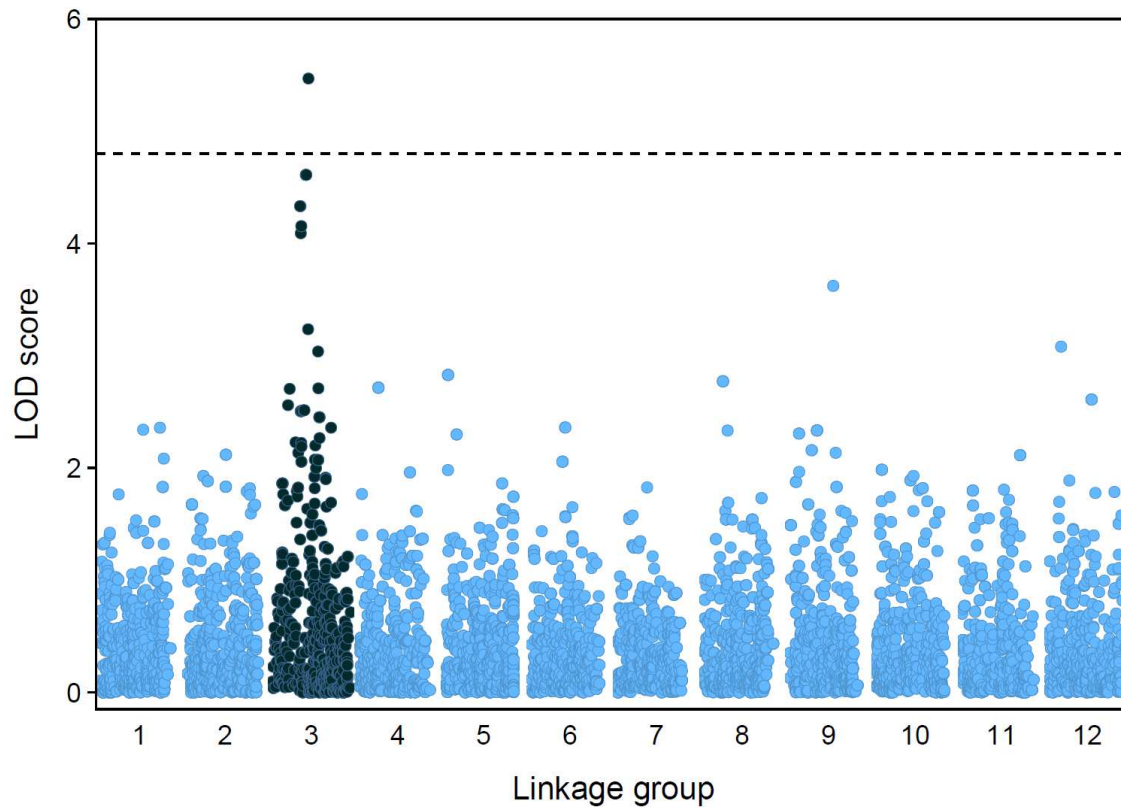


Fig. 7 Within-family prediction discrimination for stem forking in the G2 validation population. Models used ssGBLUP with and without the marker SNP505839 fitted as a fixed effect. Bars show the area under the receiver operating characteristic curve (AUC) for classifying forked and non-forked trees within each full-sib family. The dashed horizontal line indicates $AUC = 0.5$, corresponding to no discrimination. Families with N11210 ancestry are denoted with an asterisk.

