Received: 10 October 2024 | Accepted: 23 January 2025

DOI: 10.1002/tpg2.70005

The Plant Genome

ORIGINAL ARTICLE

Development of a 101.6K liquid-phased probe for GWAS and genomic selection in pine wilt disease-resistance breeding in Masson pine

Jingyi Zhu[1,2] | **Qinghua Liu**[1,3] | **Shu Diao**[1,4] | **Zhichun Zhou**[1,4] | **Yangdong Wang**[1,4] | **Xianyin Ding**[1,4] | **Mingyue Cao**[5] | **Dinghui Luo**[6]

[1]Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou, China

[2]College of Landscape Architecture, Nanjing Forestry University, Nanjing, China

[3]State Key Laboratory of Tree Genetics and Breeding, Chinese Academy of Forestry, Beijing, China

[4]Zhejiang Provincial Key Laboratory of Tree Breeding, Hangzhou, China

[5]Higentec Co., Ltd, Changsha, China

[6]Linhai Natural Resources and Planning Bureau, Linhai, China

Correspondence
Qinghua Liu and Shu Diao, Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou, China.
Email: liuqh@caf.ac.cn and diaoshu0802@163.com

Assigned to Associate Editor Zhonghu He.

Funding information
Biological Breeding-National Science and Technology Major Project, Grant/Award Number: 2022ZD0401603; Zhejiang Science and Technology Major Program on Agricultural, New Variety Breeding, Department of Science and Technology of Zhejiang Province, People's Republic of China, Grant/Award Number: 2021C02070-5-2; Fundamental Research

Abstract

Masson pine (*Pinus massoniana* Lamb.), indigenous to southern China, faces serious threats from pine wilt disease (PWD). Several natural genotypes have survived PWD outbreaks. Conducting genetic breeding with these resistant genotypes holds promise for enhancing resistance to PWD in Masson pine at its source. We conducted a genome-wide association study (GWAS) and genomic selection (GS) on 1013 Masson pine seedlings from 72 half-sib families to advance disease-resistance breeding. A set of efficient 101.6K liquid-phased probes was developed for single-nucleotide polymorphisms (SNPs) genotyping through target sequencing. PWD inoculation experiments were then performed to obtain phenotypic data for these populations. Our analysis reveals that the targeted sequencing data successfully divided the experimental population into three subpopulations consistent with the provenance, verifying the reliability of the liquid-phased probe. A total of 548 SNPs were considerably associated with disease-resistance traits using four GWAS

Abbreviations: AUDPC, area under the disease progress curve; BA, Bayesian model based on the scaled-t distribution; BV, breeding value; cGPS, genotyping using the pinpoint sequencing of liquid-captured targets; CTAB, cetyl trimethyl ammonium bromide; CV, cross-validation; DNNGP, deep neural network-based method for genomic prediction; FarmCPU, fixed and random model circulating probability unification; Fst, fixation index; gBLUP, genomic best linear unbiased prediction; GEBV, genomic estimated breeding value; GS, genomic selection; GWAS, genome-wide association study; He, expected heterozygosity; Het, heterozygosity; Ho, observed heterozygosity; LD, linkage disequilibrium; LMM, logistic mixed model; MAF, minor allele frequency; MLM, mixed linear model; PA, predictive ability; PC, prediction accuracy; PCA, principal component analysis; PVE, phenotypic variance explanation; PWD, pine wilt disease; PWN, pine wood nematode; *Q*–*Q* plot, quantile–quantile plot; RNA-seq, RNA sequencing; rrBLUP, ridge regression best linear unbiased prediction; SNP, single-nucleotide polymorphism; SUPER, settlement of MLM under progressively exclusive relationship; WGS, whole-genome sequencing.

Funds for the Central Non-profit Research Institute of CAFs., Grant/Award Number: CAFYBB2021ZG001-4

algorithms. Among them, 283 were located on or linked to 169 genes, including common plant disease resistance-related protein families such as NBS-LRR and AP2/ERF. The DNNGP (deep neural network-based method for genomic prediction) model demonstrated superior performance in GS, achieving a maximum predictive accuracy of 0.71. The accuracy of disease resistance predictions reached 90% for the top 20% of the testing population ordered by resistance genomic estimated breeding value. This study establishes a foundational framework for advancing research on disease-resistant genes in *P. massoniana* and offers preliminary evidence supporting the feasibility of utilizing GS for the early identification of disease-resistant individuals.

Plain Language Summary

Pine wilt disease is a serious problem for Masson pine, an important economic tree in China. The disease threatens the health of these trees, making it important to find ways to protect them. One solution is to breed Masson pine trees with disease-resistant genes. However, the pine tree genome is very large and difficult to fully study, making it hard to find genes that could make the trees resistant. To address this, we use a special probe, which is a tool that helps gather important genetic information more efficiently. By combining this genetic information with experiments that test which tree has resistance to the disease, we can identify genes that help protect the trees. Additionally, genomic selection can help select resistant trees more quickly, speeding up the process of breeding resistant trees. This research is a key step in understanding and preventing pine wilt disease in Masson pine.

1 | INTRODUCTION

Pine wilt disease (PWD), caused by the pine wood nematode (PWN, *Bursaphelenchus xylophilus*), represents a great threat to forest ecosystems (Mamiya & Kiyohara, 1972). PWD was first discovered in the United States in 1979 and subsequently spread to Europe and Asia over the following decades (Mota et al., 1999). In China, Masson pine (*Pinus massoniana* Lamb.)—a major timber and resin-producing species extensively distributed across 17 subtropical provinces and of considerable economic value—is among the primary targets of PWD. Upon infestation, Masson pine exhibits acute symptoms of dehydration, discoloration, and eventual death upon infestation (Futai, 2013), leading to profound losses in forest ecology and forestry economics. Although the isolation of infected trees (M. Li et al., 2022) and trunk injection (D. Li et al., 2023) have been widely applied to prevent PWD, it continues to spread. Some natural individuals of Masson pine have survived PWD outbreaks (Zhao & Li, 2008), suggesting the presence of wild genotypes with resistance to the disease. Leveraging these wild-resistant genotypes through genetic breeding presents an effective approach to preventing PWD at its source. However, the development of PWD-resistant varieties through traditional breeding methods is a time-consuming process.

Molecular breeding is an effective strategy to expedite and optimize the development of disease-resistant varieties (F. Liu et al., 2024). Identifying genetic loci and key genes associated with PWD resistance is essential, as single-nucleotide polymorphism (SNP) loci can help in early selection, while key genes represent potential targets for enhancing resistance through the application of transgenic or gene-editing technologies. Transcriptome studies have revealed that the molecular mechanisms involved in disease resistance are highly complex, including the pathways of plant hormone signaling, secondary metabolism, oxidative stress responses, plant defense responses, and resistance reactions (Modesto et al., 2022). However, the key regulatory genes that govern these defense pathways remain unidentified. Genome-wide association study (GWAS) has emerged as a powerful tool for dissecting the genetic architecture of disease-resistance traits in plants. This method has successfully identified genetic loci and resistance genes for common crop diseases such as wheat rust (Vikas et al., 2022) and wheat powdery mildew (Du et al., 2021). GWAS is particularly effective in identifying genetic loci and genes associated with PWD resistance, making it a

crucial tool for advancing the understanding and improvement of resistance traits. In addition to GWAS, genomic selection (GS) is also a promising breeding approach that leverages high-density markers spanning the entire genome to accelerate breeding cycles, particularly in forestry (Diao et al., 2016). This technique has been successfully applied in trees such as *Eucalyptus* (Cappa et al., 2019) and *Pinus* (Isik et al., 2016) for predicting growth and wood traits. Thus, it is an ideal strategy for developing Masson pine varieties resistant to PWD.

A primary challenge when applying GWAS and GS for breeding PWD-resistant varieties is genotyping in pine species, given their enormous and uncharacterized genomes. Recent advances in genotyping strategies, such as the development of liquid- or solid-phased probes, have shown great promise in addressing these challenges for large populations. SNP-based solid arrays are effectively adopted in genotyping for pine species and have been employed in Maritime pine (*Pinus pinaster* Aiton) (Plomion et al., 2015), Loblolly pine (*Pinus taeda* L.) (Caballero et al., 2021; De La Torre et al., 2018), four European pine species (Perry et al., 2020), six tropical pine species (Jackson et al., 2021), and Scots pine (*Pinus sylvestris* L.) (Kastally et al., 2022). Despite their effectiveness, SNP-based solid arrays are associated with certain limitations, such as high costs, an inability to incorporate new target SNPs, and only genotyping the designed SNPs. Liquid-phased probes offer a more cost-effective and flexible solution for genotyping across a wide range of SNPs, as they can simultaneously capture multiple loci within the genome. For example, J. J. Liu et al. (2016) and Neves et al. (2013) conducted exome-targeted sequencing of Maritime pine using liquid-phased probes. Diao et al. (2024) developed a 51K SNP array using a published SNP set, which was applied to targeted capture sequencing in pine species such as Slash pine (*Pinus elliottii* Engelm.), Loblolly pine, and Caribbean pine (*Pinus caribaea* Morelet). Scholars employed genotyping using the pinpoint sequencing of liquid-captured targets (cGPS) as a new type of liquid-phased probe to successfully achieve SNP-targeted sequencing (Meng et al., 2024). Therefore, target sequencing with liquid-phased probes appears to be the most suitable approach for the genetic characterization of Masson pine due to its flexibility, efficiency, and cost-effectiveness.

In this study, leveraging whole-genome sequencing (WGS) data of Masson pine and integrating previous RNA sequencing (RNA-seq) analyses, we designed a set of evenly distributed SNP markers across the entire genome to construct a 101.6K liquid-phased probe array, enabling cost-effective batch sequencing of a large number of samples. Within this framework, we conducted inoculation experiments and targeted sequencing on a total of 1013 Masson pine genotypes to evaluate their disease-resistance capabilities. Using high-density SNP genotyping and phenotype data, our study aims to (1) validate the reliability of the liquid-phased probe array through population structure analysis; (2) identify SNPs associated with PWD resistance via GWAS and explore resistance candidate genes by identifying significant SNPs and RNA-seq; and (3) assess the potential of GS prediction models for breeding disease-resistant Masson pine. This study facilitates further research on disease-resistant genes in Masson pine, as well as gene function studies and cultivar breeding.

Core Ideas

- It is urgent to explore resistance genes of pine wilt disease (PWD) for variety improvement.
- The developed 101.6K liquid-phased probe can obtain large quantities of Masson Pine genotype data economically.
- With genome-wide association study (GWAS), a candidate disease-resistant single-nucleotide polymorphism (SNP) set was constructed for gene mining and GS model optimization.
- A candidate resistance gene set was constructed to aid in the study of the pathogenesis of PWD.
- We preliminarily tested and demonstrated the feasibility of using GS for early identification of PWD resistance.

2 | MATERIALS AND METHODS

2.1 | Plant materials and resistance evaluation

We performed WGS on 30 Masson pine individuals naturally distributed across seven southern provinces of China. A huge number of wild Masson pines exhibiting natural resistance to PWD were collected and subsequently planted in the Linhai seed orchard (28°51′ N, 121°07′ E) for further propagation. Following free pollination, a total of 72 unrelated Masson pines from nine provinces were selected as maternal trees for breeding. A total of 1013 half-sib family seedlings were collected and planted in Hangzhou and Linhai as training populations for inoculation experiments and targeted sequencing. Each family contained 10 or more individuals, thereby representing the genetic resources associated with disease resistance in Masson pine across southern China. The practical application of the GS model was quantified using an additional 49 Masson pine individuals, whose disease resistance was confirmed through three rounds of inoculation experiments. There was no kinship between the 49 samples and the training population. Detailed information is reported in Table S1.

For the PWN inoculation experiment, the highly pathogenic and proliferative strain "Guangde 3B" was

FIGURE 1 Severity levels of pine wilt disease (PWD) in Masson pines. The progression of PWD across a single genotype. The severity level, ranging from 0 to 5, is depicted sequentially from left to right.

selected for mixing with strains isolated from dead Masson pines in Anhui and Zhejiang provinces. Previous experiments confirmed the strong pathogenic ability of the mixed population. The nematodes were subsequently cultured in wheat medium supplemented with *Botrytis cinerea*. Artificial inoculation was performed in July 2023 (35˚C) using a method that simulates beetle feeding on Masson pine, as described previously (Q. Liu et al., 2017). Observations were conducted weekly after inoculation. Plants were classified into six distinct levels based on the severity of disease symptoms: level 0: healthy plants with no visible symptoms; level 1: minor wilting and bending of young shoots; level 2: approximately 25% of leaves exhibiting pale yellow discoloration and wilting; level 3: around 50% of leaves showing pale discoloration and browning, indicative of PWD infection; level 4: approximately 75% of leaves turning pale, 50% of leaves showing browning; and level 5: all leaves displaying yellow–brown discoloration, resulting in complete plant mortality (Figure 1).

The deviation value was employed as an index of the family resistance capability (Xu & Tadao, 2006) and is calculated as:

$$\mathrm{Dv} = (S_{\mathrm{F}} - \bar{S})/\sigma,$$

where S_{F} is the survival rate of the family, $\bar{S}$ is the overall survival rate, and σ is the overall standard deviation. Based on the ratio of the family deviation value to the overall deviation value, the family resistance ability is classified into five levels, divided by the boundaries -1.5σ, -0.5σ, 1σ, and 1.5σ.

The area under the disease progress curve (AUDPC) (Shaner, 1977) was used as an index of the individual genotype's disease resistance capability and is calculated as:

$$\mathrm{AUDPC} = \sum_{i=1}^{n} \left[\frac{\mathrm{SB}_{i+1} + \mathrm{SB}_i}{2}\right] \left[t_{i+1} - t_i\right],$$

where SB_i is the severity level of the disease at the ith observation, with values of 0, …, 5; t_i is the time of the ith observation; and n is the total number of evaluations.

To thoroughly investigate the genetic structure of disease resistance using GWAS and identify optimal phenotypes to enhance prediction accuracy (PC) in GS, we utilized a range of computational methods to generate multiple sets of phenotypic data (Table S2). Y1 represents the final disease resistance performance of genotypes after infection with PWN, based on the survival status at the end of the 12th week, with a value of 1 indicating survival and a value of 0 indicating death. Y1 is the direct observation at the end of the 12th week and is treated as a binary variable. Y2 and Y3 represent the AUDPC values at the end of the 8th and 12th weeks, respectively, standardized as continuous numerical variables. Y4 and Y5 are derived from Y2 and Y3, respectively, weighted by corresponding family resistance indices, enhancing the genetic basis of the family lineage on individual disease resistance performance.

2.2 | Design and synthesis of the liquid-phased probe

To obtain SNP markers across the entire genome and identify candidate regions for liquid-phased probe design, we performed WGS on 30 Masson pine samples. Genomic DNA was extracted from leaf tissue using the cetyl trimethyl ammonium bromide (CTAB) method. WGS libraries were prepared following the MGIEasy Fast (MGI) enzyme digestion library protocol. Sequencing was conducted using the DNBseq platform (MGI, China). Raw sequencing data underwent quality control using fastp 0.23.2 (Z. Chen et al., 2018) with default parameters, followed by alignment to the *P. tabuliformis* reference genome (Niu et al., 2021) using Sentieon V2020 (Freed et al., 2017). The alignment results were sorted and deduplicated using Samtools 1.17 (H. Li et al., 2009). SNP detection was then performed using Sentieon V2020. SNPs and indels were annotated using ANNOVAR (K. Wang et al., 2010), which identified the genomic regions and types of mutations at the variant sites. The SNP data obtained from resequencing underwent two rounds of filtering. The first round of

filtering included the removal of SNPs on contigs, SNPs located within 10 bp of indels, and SNPs with a miss rate < 0.1, minor allele frequency (MAF) > 0.05, and heterozygosity (Het) < 0.6. Moreover, SNPs with a sequencing depth between 10× and 40× were retained. Following this, SNPs with 100 bp upstream and 100 bp downstream were further screened according to liquid-phased probe design principles, with the following criteria: (1) 20%–80% GC content; (2) no repetitive sequence regions; (3) no N bases; and (4) whole-genome copy number between 1 and 3. The filtered SNPs were then used for the subsequent probe design.

We selected SNPs for liquid-phased probe design in terms of three characteristics. First, we explored SNPs for candidate genes of PWD resistance. Three batches of differential gene expression analysis were performed using inoculation experiments and RNA-seq of Masson pine with different resistance levels (unpublished). RNA-seq results using the following selection criteria: p_{adj} (adjusted p-value) < 0.05 and |log2FoldChange| > 1. Candidate gene sequences were aligned to the *P. tabuliformis* reference genome, retaining genes with a sequence alignment rate > 70% and sequence similarity > 90%. Second, SNP genotyping was performed using the liquid-phased probe from other pine species. We conducted the targeted capture sequencing of 10 Masson pine samples using the 51 K liquid-phased probe designed for *P. elliottii* and *P. taeda* (Diao et al., 2024). We selected high-stability probes based on the following criteria: the probes had a single copy in the *P. tabuliformis* genome, exhibited a matching similarity greater than 80%, and achieved a 100% detection rate across all 10 Masson pine individuals. Third, SNPs were selected based on the genomic physical distribution. We used PLINK 1.9 (Purcell et al., 2007) to perform linkage disequilibrium (LD) analysis on the SNPs filtered in the first round. SNPs were filtered using the LD parameter R^2 ≥ 0.2 to remove highly linked sites. SNPs with MAF ≥ 0.1 were prioritized. Through the above steps, probes were considered for uniform distribution across the genome. Finally, SNPs from the three sources were combined and deduplicated for the final probe design (Figure 2).

We selected 12 Masson pine samples for cGPS to validate the efficiency of the 101.6K liquid-phased probe array, with technical replicates performed on four of these samples. Genomic DNA was extracted from fresh leaf tissue using the CTAB magnetic bead method. High-quality, intact DNA (1.8 < OD260/280 < 2.0, OD260/230 > 2.0) was selected for targeted capture sequencing. The sequencing procedures followed the methods described by previous studies (Diao et al., 2024; Meng et al., 2024), including library construction, hybrid capture, and targeted sequencing. Raw data quality control was performed using fastp with default parameters, and alignment was conducted using BWA 0.7.17 (Li & Durbin, 2009) against the reference genome. Variant analysis of the sequencing results was performed using GATK 4.2.6.1 (McKenna et al., 2010) to obtain genotype calls for the targeted SNPs and the surrounding sequences. Finally, quality control measures for missing data (miss < 0.1) and MAF (>0.05) were performed using PLINK.

2.3 | Genetic structure and population structure analysis

The fixation index (Fst) and nucleotide diversity (π) were calculated for each subgroup using Vcftools (Danecek et al., 2011). The MAF, observed heterozygosity (Ho), and expected heterozygosity (He) were computed for the populations using PLINK. The subpopulation structure of the Masson pine populations was analyzed using the model-based clustering algorithm in Admixture 1.22 (Alexander & Lange, 2011). Individuals were allocated to distinct subpopulations according to their familial lineage. The clustering results were visualized as bar plots using the Admixture Q Matrix Viz function in TBtools (C. Chen et al., 2020). To construct the phylogenetic tree based on individual adjacency, the phylogenetic matrix was generated using vcf2phylip 2.0 (Olatoye et al., 2019). The resulting matrix was imported into MEGA11 (Tamura et al., 2021) for the construction of the phylogenetic tree, employing the bootstrap method with 5000 replicates and the P-distance for phylogenetic testing. The phylogenetic tree was subsequently optimized using the online tool ITOL (Letunić & Bork, 2021) (https://itol.embl.de/). Population principal component analysis (PCA) was performed using GCTA (Yang et al., 2011). Kinship analysis was performed using TASSEL 5 (Bradbury et al., 2007), and the results of the PCA and kinship analysis were visualized using the R (R Core Team) package ggplot2.

2.4 | GWAS and candidate gene identification

Four models were used for GWAS. The settlement of MLM under progressively exclusive relationship (SUPER) (Q. Wang et al., 2014) and fixed and random model circulating probability unification (FarmCPU) (X. Liu et al., 2016) methods were employed to comprehensively identify potential SNPs in GAPIT 3.0 (J. Wang & Zhang, 2021). IIIVmrMLM (Li et al., 2022) implements a compressed variance component mixed model approach (3VmrMLM) to estimate additive and dominance effects, as well as their interactions with environmental and epistatic effects, while controlling for all possible polygenic backgrounds. This approach was adopted for the identification of significant and recommended loci. The logistic mixed model (LMM) was employed to test genetic associations using the GENESIS package (Gogarten et al., 2019). Principal components calculated by PC-AiR

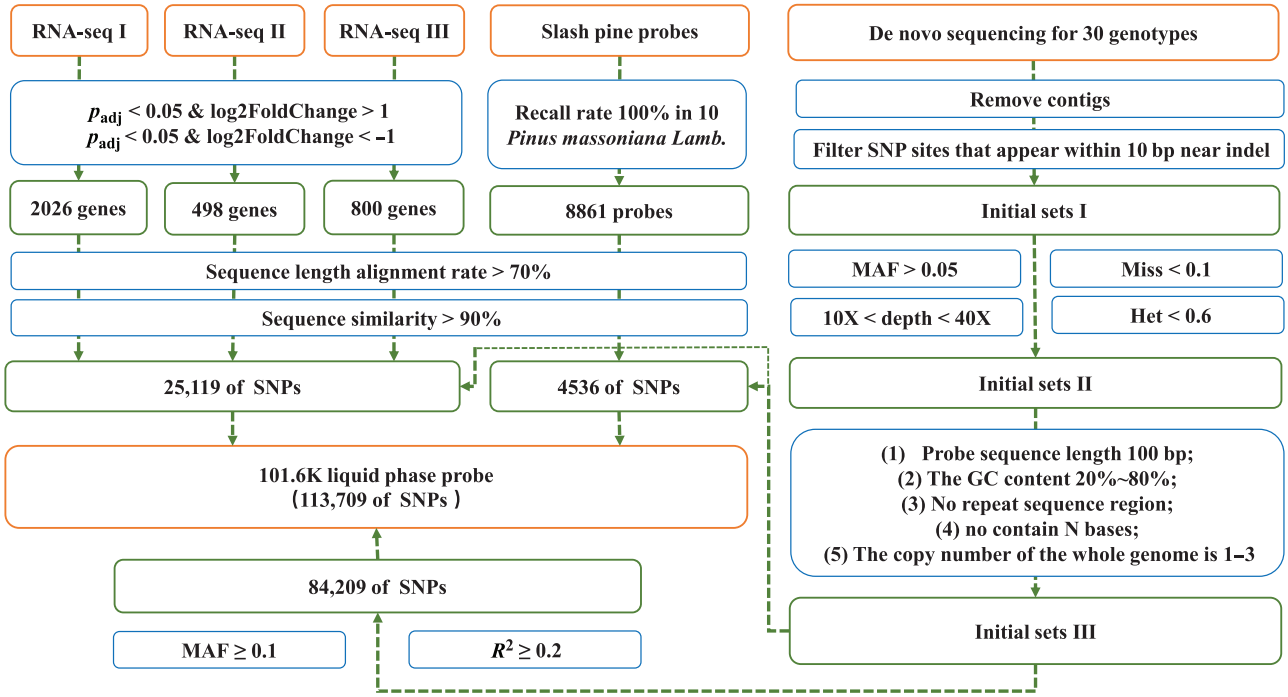

FIGURE 2 Schematic workflow of the liquid-phased probe array design. The orange boxes represent the input and output data, the blue boxes represent the filtering process, and the green boxes represent the intermediate data. A total of 113,709 single-nucleotide polymorphisms (SNPs) were added to the liquid-phased probe array design. MAF, minor allele frequency.

were used as fixed effect covariates to account for known and unknown relatedness in the LMM (Conomos et al., 2015). The kinship matrix (or genetic relationship matrix) estimated from PC-Relate was used as a random effect to explain phenotypic correlations due to population structure (Conomos et al., 2016). The aforementioned models used the *F*-test and Bonferroni correction for multiple testing. Quantile–Quantile (*Q*–*Q*) plots were employed to validate the presence of false positives. A mixed linear model (MLM) was utilized to estimate the heritability (h^2) for each phenotype.

The LD decay distance was calculated using PopLDdecay (C. Zhang et al., 2018). The gene functional annotation of genes was conducted using the online tool eggnog-mapper (http://eggnog-mapper.embl.de/) (Cantalapiedra et al., 2021). LDBlockShow was employed for the LD calculations and visualizations between SNPs (Dong et al., 2020). Haplotype analysis of SNPs was conducted using the R package geneHapR (R. Zhang et al., 2023). SNPs linked to functional genes were extracted separately and subjected to correlation analysis with phenotype using the R package SKAT (Lee et al., 2012). The transcriptome and haplotype differential results were visualized using the R package ggplot2.

2.5 | GS and resistance prediction

Four models were used to predict the genomic estimated breeding value (GEBV) for resistance to PWD. Genomic best linear unbiased prediction (gBLUP) was implemented using the GAPIT package. The gBLUP model is described as:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mu + \epsilon$$

where $\mathbf{y}$ is the vector of phenotypic values; $\mathbf{X}$ is the design matrix for fixed effects; β is the fixed effect parameters; $\mathbf{Z}$ is the genotype matrix; μ is the random genetic effects; and ϵ is the residual term. The ridge regression best linear unbiased prediction (rrBLUP) model was derived from the rrBLUP package in R (Endelman, 2011) and is described as:

$$\mathbf{y} = \mathbf{Xb} + \sum_{i=1}^{M} \mathbf{z}_i g_i + e,$$

where $\mathbf{y}$ is the vector of phenotypic values; $\mathbf{X}$ is the matrix of fixed effect coefficients; $\mathbf{b}$ is the vector of fixed effects; $\mathbf{z}_i$ is the vector of genotypes for the *i*th marker; g_i is the effect of the *i*th marker; and *e* is the residual error. The Bayesian model based on the scaled-t distribution (BA) was derived using the BGLR package in R (Pérez & de Los Campos, 2014), with 3000 iterations performed and 500 iterations discarded as burn-in. The DNNGP (deep neural network-based method for genomic prediction) model (K. Wang et al., 2023), selected as the nonlinear model, is based on deep neural networks and can capture complex nonadditive effects.

To ensure accuracy, we employed fivefold cross-validation (CV), randomly dividing the phenotype and genotype datasets

of each sample into five equally sized components. The model was trained on four folds to predict the GEBVs for the remaining fold with 100 repetitions. We assessed the predictive ability (PA) by calculating the Pearson correlation coefficient (ρ^2) between the observed values (Y) and model-predicted GEBVs and evaluated PC using PA/$\sqrt{h^2}$. To investigate the predictive accuracy of the linear models across different numbers of markers, we randomly selected six subsets of SNPs uniformly distributed throughout the genome, with sizes of 1K, 5K, 10K, 50K, 100K, and 200K markers. To determine the optimal predictive accuracy of DNNGP, the genotype data were processed into four PCA sets with sizes of 10, 50, 100, and 500. To enhance model PC, the breeding value (BV) for phenotypes Y1 (with no manual processing) and Y4 (superior performance in GWAS) were computed using GCTA and incorporated into the analysis (Merrick et al., 2022). We controlled the model by selecting significant SNPs as fixed factors. Statistical analysis was conducted using SPSS 27 (IBM, SPSS Inc.), and single-factor analysis of variance was performed on the computed results. Post hoc multiple comparisons were conducted using Tukey's honestly significant difference to assess significant differences in the PA. The best model and settings were used to perform 100 repetitions of disease resistance predictions for Masson pine individuals outside the modeling population. The practical application of the model was assessed by comparing the predicted and actual resistances.

3 | RESULTS AND ANALYSIS

3.1 | Masson pine evaluation

Due to differences in the physiological age of the seedlings planted in Hangzhou and Linhai, the results from each location are presented separately. In the Hangzhou nursery, a total of 725 1-year-old seedlings were evaluated. PWD symptoms began to manifest in the second week and peaked by the sixth week. Mortality symptoms appeared from the fourth week, with an overall mortality rate of 41.16% by the 12th week. In contrast, the Linhai nursery contained 288 3-year-old seedlings. Symptoms in this group also began to appear in the second week but peaked later (in the eighth week). Mortality symptoms were observed from the fourth week, resulting in a higher overall mortality rate of 64.21% by the 12th week (Figure 3). We identified five highly resistant families and 11 resistant families (Table S3), respectively, from these two locations. The marked difference in disease incidence between the Linhai and Hangzhou nurseries is likely attributable to variations in physiological age. To minimize errors associated with age differences, we implemented max–min normalization on the phenotype data and included the planting location as a covariate in subsequent analyses.

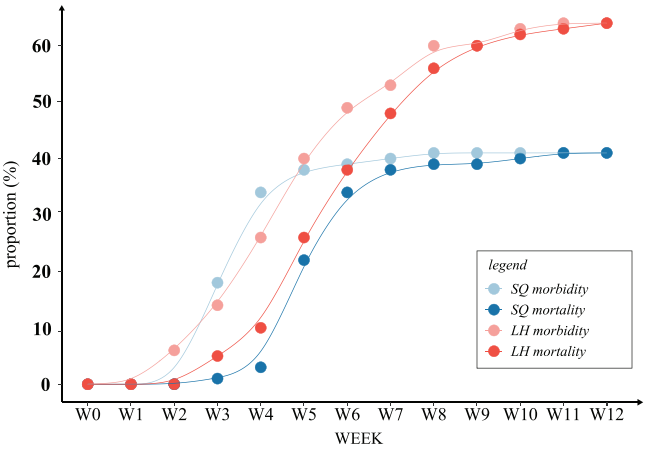

FIGURE 3 Morbidity and mortality of the pine wilt disease (PWD) inoculation experiment. Red and blue denote the 2- and 1-year-old populations, respectively. The light- and dark-colored curves illustrate the temporal variation in population morbidity and mortality, respectively.

3.2 | 101.6K Liquid-phased probe

WGS yielded an average GC content of 38.19% and an average sequencing depth of 24.43×, with an alignment rate of 98.34% compared to the reference genome. These metrics indicate high-quality sequencing and confirm a close genetic relationship between Masson pine and Chinese pine. By leveraging a substantial number of SNP markers derived from WGS for probe development (Figure S1), the Masson pine 101.6K liquid-phased probe was developed successfully. This probe comprises 101,597 capture probes, targeting 113,709 SNPs of interest (Table 1), with 29,501 SNPs located on candidate genes implicated in disease resistance by RNA-seq. The average distance between SNPs is 214,461 bp. Compared to the annotation information of the reference genome, there are 57,213 SNPs (50.32%) distributed on genes, which can be directly linked to 15,505 functional genes. In addition, 56,496 SNPs are distributed in intergenic regions, effectively meeting the requirement for uniform genome coverage by the probe (Figure S2, Table S4). Based on these markers, a total of 101,597 capture probes were synthesized for the targeted regions.

The targeted sequencing of 12 samples detected an average of 102,204 SNPs, with average detection, Het, and missing rates of 98.82%, 32.07%, and 0.67%, respectively, and an average depth of 22.0771×. All sequencing indexes met the expected design standards. Technical replicates of four samples exhibited an average genotype concordance rate of 97.83%, indicating low technical error. Following successful probe validation, targeted sequencing was performed on 1013 Masson pine individuals. After hard filtering with GATK, a total of 553,699,380 SNPs were obtained. Quality control was then applied for missing data and minimum allele frequencies,

TABLE 1 Single-nucleotide polymorphism (SNP) annotation for whole-genome sequencing (WGS) and the 101.6K liquid-phased probe.

Category	Number of SNPs	
	WGS	101.6K liquid-phased probe
Exonic	4,302,161	32,176
UTR3	799,374	4179
UTR5	639,065	3490
UTR5; UTR3	11,912	113
Splicing	16,095	34
Intronic	105,417,648	16,184
Upstream	3,517,626	1093
Downstream	3,293,808	1365
Upstream/ downstream	305,644	101
Intergenic	1,504,206,721	54,974
Stop gain	102,325	0
Stop loss	5570	0
Synonymous	884,413	0
Non-synonymous	1,859,502	0
Total	1,643,703,150	113,709

resulting in 572,079 high-quality SNPs for the subsequent analysis.

3.3 | Population and genetic structure analysis

The CV errors for the number of preselected subgroups of 2, 3, 4, 5, and 6 were 0.51305, 0.50756, 0.50300, 0.49923, and 0.49511, respectively, and no inflection point was observed. This indicates that the calculation results do not provide the optimal number of subgroups. Based on the results determined from Admixture, we categorized the population into three subpopulations: western source (pop W), central source (pop C), and eastern source (pop E) (Figure 4A,C). This corresponds to known geographic sources. Subpopulations within each family were consistent, with 190, 511, and 307 genotypes allocated to each subpopulation, respectively. The genotype distribution across provinces is shown in Figure 4B, indicating that diseased and healthy individuals are present in each province.

To further validate the reliability of the subpopulation classification, we constructed a neighbor-joining tree based on individual data to assess the genetic relationships among individuals (Figure 4D). The results show that pop W has a relatively distant genetic relationship to the other two subpopulations, while pop C and pop E are more closely related. PCA conducted in PLINK confirmed these findings, agreeing with the subpopulation classification (Figure 4E).

We analyzed the π and Fst values among subpopulations, using the top 1% of values as the population characteristics (Figure 4F). The Fst values between pop C and pop E, pop E and pop W, and pop C and pop W were 0.226, 0.287, and 0.351, respectively. This indicates that significant differentiation is present among the three subpopulations, with pop W exhibiting the greatest level of differentiation. All subpopulations displayed high genetic richness. In particular, the π values of pop C, pop E, pop W, and the training population were 2.08×10^{-3}, 2.21×10^{-3}, 2.16×10^{-3}, and 2.17×10^{-3}, respectively, reflecting a substantial representation of Masson pine genetic resources in southern China.

To assess the potential of the liquid-phased probe array in evaluating population genetic variation, we calculated the MAF, Ho, and He (Figure 4G). The coefficient of variation in Ho (0.63) was the highest, followed closely by MAF (0.61), while He (0.44) exhibited the least variation. The average MAF value was 0.22, suggesting the presence of relatively common minor alleles and reflecting a high degree of genomic variation and genetic diversity. The average Ho (0.30) was close to the average He (0.31), indicating that after the probe design and manual screening, the allele distribution in the population was relatively uniform. The kinship analysis results are presented in Figure S3. Numerous small blocks are located along the diagonal, proving the presence of a large number of closely related groups in the material. This accurately represents the kinship of our 72 half-sib family materials, further confirming the reliability of the liquid phase probe.

3.4 | Genome-wide association study

Among the four GWAS models, the logistic and IIIVmrMLM models were the most effective, determining 474 and 380 significant SNPs, respectively. The SUPER and FarmCPU models only identified 33 and 11 significant SNPs, respectively. Numerous overlapping sites were observed between phenotypes. The numbers of overlapping and nonoverlapping sites in the logistic model were 195 and 279, respectively (41.14% overlap rate); those in the IIIVmrMLM model were 253 and 127, respectively (33.42% overlap rate); those in the SUPER model were 16 and 17, respectively (48.48% overlap rate); and those in the FarmCPU model were 5 and 6, respectively (45.45% overlap rate) (Table 2, Table S5, Figure S4). After filtration, we identified 547 significant associations between SNPs and five disease-resistance traits (Figure 5). All significant SNPs were extracted to calculate the heritability of the trait, determined as 0.26–0.32. These significant SNPs accounted for 1‰ of the total but contributed to 40.00%–44.84% of the heritability. The high contribution of these loci to the heritability further verifies their value in studying disease resistance traits.

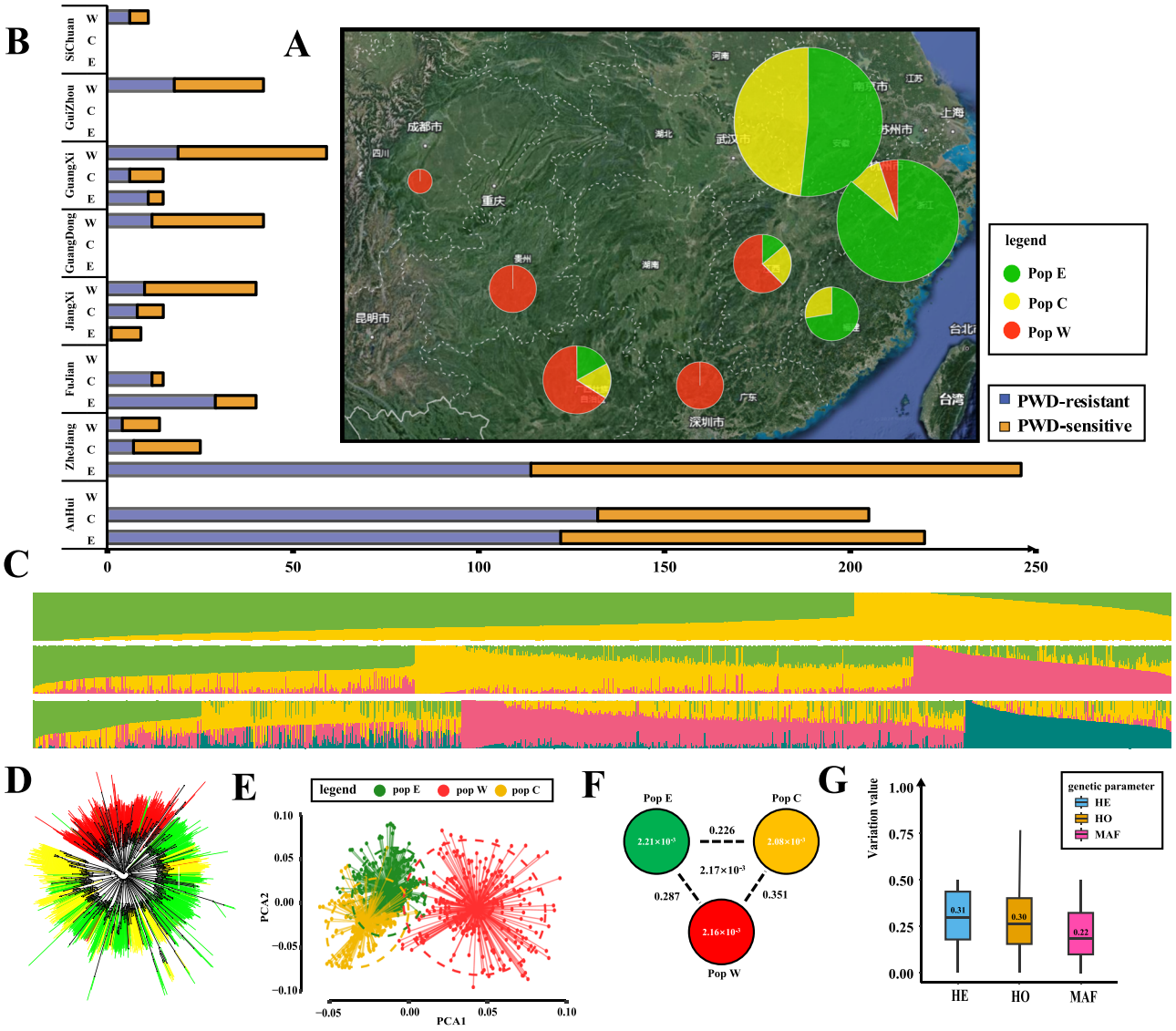

FIGURE 4 Population and genetic structure analysis of Masson pine samples. (A) The dotted lines on the map represent the different provinces of China. The pie chart size represents the proportion of total genotypes in the province and the pie chart color ratio represents the proportion of genotypes in the three subgroups in the province. (B) Comparison of the number of genotypes and number of disease-resistant and disease-susceptible individuals across different provinces and subpopulations. (C) Admixture results of different subpopulations, from top to bottom, for $K = 2$, $K = 3$, and $K = 4$, respectively. (D) Rootless evolutionary trees based on individuals. (E) Principal component analysis. (F) The numbers within and between subpopulations represent π and fixation index (Fst), respectively. (G) Population genetic parameters. He, expected heterozygosity; Ho, observed heterozygosity; MAF, minor allele frequency; PCA, principal component analysis; PWD, pine wilt disease.

SNPs captured by the same liquid phase probe are represented as SNP clusters with strong internal correlation. Thus, directly summing the phenotypic variance explanation (PVE) of significant SNPs will lead to data distortion due to LD. Based on the results of the IIIVmrMLM model, we selected the PVE of the SNPs with the smallest *p*-value in each cluster to represent the whole cluster. By superposition, the PVE of these significant SNPs in traits Y1–Y5 totaled 46.6%–48.7%. Among these, Y4 exhibited the highest number of 275 SNP associations (Figure 5). For the analysis, models incorporated location and subpopulation differences as covariates to control the false positive rate.

3.5 | Identification of candidate genes

To construct a disease-resistance candidate gene set, we first calculated LD decay distances based on the WGS data. Given the large number of SNPs, we used PopLDdecay to compute the LD decay separately for each chromosome and then plotted the average values (Figure 6A). The distance

TABLE 2 Significant single-nucleotide polymorphisms (SNPs), annotated genes, and heritability of different phenotypes.

Mode	Type	Y1	Y2	Y3	Y4	Y5	All
FarmCPU	SNPs	NA	3	3	2	3	6
	Genes	NA	0	1	1	0	2
SUPER	SNPs	NA	4	4	8	17	17
	Genes	NA	3	2	3	8	5
Logistic (GENESIS)	SNPs	NA	33	11	275	155	279
	Genes	NA	8	3	54	31	55
IIIVmrMLM(SIG)	SNPs	66	74	80	77	83	253
	Genes	33	28	29	32	42	114
MLM	h^2	0.66	0.65	0.67	0.73	0.73	NA

Abbreviations: FarmCPU, fixed and random model circulating probability unification; h^2, heritability; MLM, mixed linear model; NA, not available; SUPER, settlement of MLM under progressively exclusive relationship.

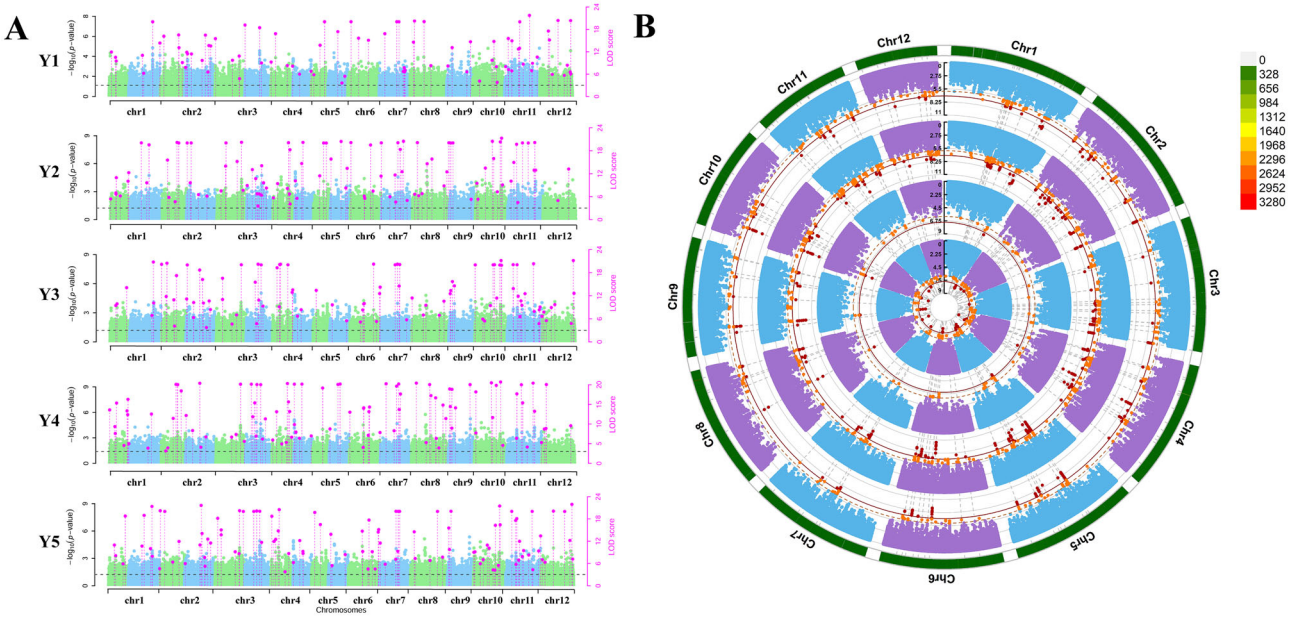

FIGURE 5 Manhattan plot for the IIIVmrMLM and logistic models. (A) Left side shows the original calculated *p*-values and the right side shows the logarithm of odds (LOD) value of the significant and suggested sites. (B) Y2–Y5, from the inner circle to the outer circle, respectively. The dashed yellow and red lines are the subsignificant interval 10^{-6} and the significant interval 10^{-7}, respectively.

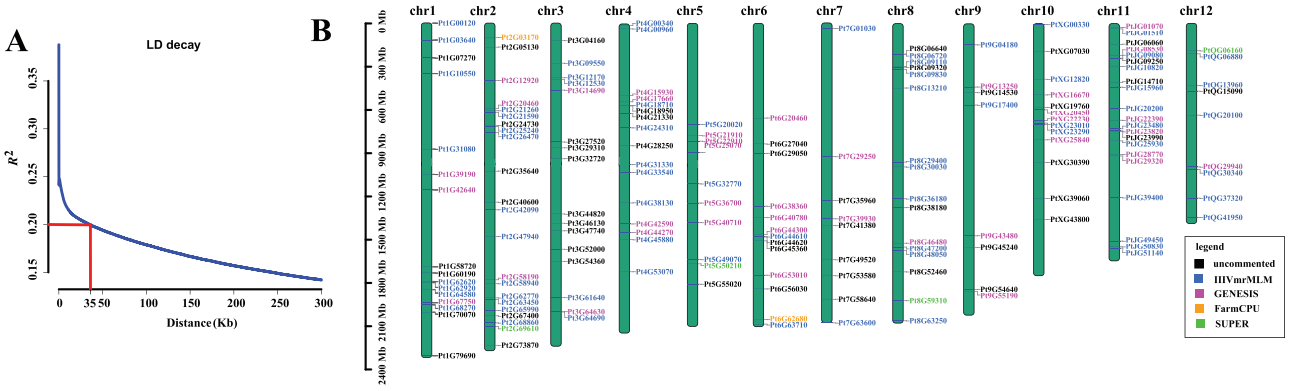

FIGURE 6 Linkage disequilibrium (LD) decay of *Pinus massoniana* and distribution of annotated genes on chromosomes. (A) Blue represents the decay of LD with distance for genetic markers, while red indicates the distance corresponding to half of the maximum LD. (B) Different colors represent the model sources of the annotated genes. FarmCPU, fixed and random model circulating probability unification; SUPER, settlement of MLM under progressively exclusive relationship.

TABLE 3 Computational information statistics of genes of interest.

Gene ID	Chr	Start(bp)	End(bp)	RNA-Seq	GWAS	Haplotype	SKAT
Pt1G67750	1	1980803189	1980928781	0.0057	2.41E-08	2.00E-06	0.0158
Pt3G14690	3	464627216	464631540	2.76E-10	3.46E-09	7.20E-08	0.0535
Pt4G15930	4	495664162	495669395	0.0055	7.40E-08	8.30E-04	0.0282
Pt5G21910	5	729207729	729208771	1.08E-07	4.09E-10	2.80E-07	0.0383
Pt5G40710	5	1424344879	1424372646	0.0014	9.28E-09	2.30E-05	0.0049

Abbreviation: GWAS, genome-wide association study.

corresponding to half of the maximum linkage was used as the average linkage distance. Based on the reference genome annotation files, we searched the upstream and downstream regions of significant SNPs. A total of 169 functional genes fell within these linkage ranges and were used to construct the disease-resistance candidate gene set (Table 2, Table S5). We extracted the protein sequences for online annotation, obtaining annotation information for 118 genes (Table S6) and their relative positions on the chromosomes (Figure 6B). Combining these results with the earlier RNA-seq studies, we identified 51 genes that were significant in GWAS and differential gene analysis. These genes were prioritized as key candidate disease-resistance genes.

Here, we describe five genes that showed significant results in GWAS, RNA-seq, and the haplotype analysis. The five genes are involved in the LD with surrounding SNPs and are associated with known disease-resistant gene families (Table 3). These gene-linked SNP clusters were extracted separately and were observed to be significantly correlated with phenotypes. The gene Pt1G67750 (Figures 7A,F,K) encodes products that include protein domains such as LRRNT_2, LRR_8, and Pkinase, which may be associated with serine or threonine kinase activities. The gene Pt3G14690 (Figures 7B,G,L) produces products containing lipoxygenase and PLAT protein domains, which may influence various physiological processes, including plant growth and development, pest resistance, and aging by participating in the biosynthesis of oxylipins, metal ion binding, and linoleic acid metabolism. The gene Pt4G15930 (Figures 7C,H,M) includes protein domains such as LRRNT_2, LRR_1, LRR_6, and LRR_8, and belongs to the same LRR family as Pt1G67750, which is involved in the regulation of protein kinase and serine/threonine kinase activities. The gene Pt5G21910 (Figures 7D,I,N) encodes products containing the PLAT protein domain and is annotated as Embryo-Specific Protein 3 (ATS3), which is involved in various abiotic stress responses and the cell membrane structure. The gene Pt5G40710 (Figures 7E,J,O) encodes products with EF-hand_1, EF-hand_5, and EF-hand_6 protein domains. These protein domains may be involved in plant–pathogen interactions by influencing calcium ion binding on the membrane.

3.6 | GS for PWD resistance

We randomly divided 1013 samples into training and testing populations to perform fivefold CV, evaluating the predictive abilities of different models (gBLUP, rrBLUP, and BA), multiple numbers of SNP (1K, 5K, 10K, 50K, 100K, and 200K), and different phenotypes (Y1, Y4, Y1_BV, and Y4_BV). Each combination was computed using 100 repetitions (Figure 8A). The results indicate that the PA of the models improved to varying extents as the number of SNPs increased. Specifically, as the number of SNPs expanded from 1000 to 200,000, the average PC improved from 0.04 to 0.15. When the number of SNPs reached 100K, the increase in accuracy plateaued, indicating that the predictive performance had essentially reached a bottleneck. When using the raw phenotypic data Y1 and Y4, the predictive abilities of the three models did not differ significantly. However, when BV was employed, gBLUP demonstrated significantly superior PA compared to the other models, showing considerable potential. Thus, gBLUP was selected for the subsequent optimization.

We selected 50 markers from the significant SNPs as fixed effects to control the models. The results showed that the predictive accuracy of Y1_BV was consistently higher than Y1 across all SNP quantities and achieved the best prediction performance at 100K ($PA_{max} = 0.56$, $\overline{PA} = 0.44$, $PC_{max} = 0.69$, and $\overline{PC} = 0.54$) (Figure 8B). When significant SNPs were included as fixed effects, the required training population size for achieving maximum predictive accuracy decreased from 100 to 50K. The results for the Y4 group were similar to those for Y1, with Y4_BV reaching an optimal prediction performance at 100K ($PA_{max} = 0.46$, $\overline{PA} = 0.33$, $PC_{max} = 0.54$, and $\overline{PC} = 0.39$), and the inclusion of fixed effects halved the training population size required to achieve maximum predictive accuracy (Figure 8C). When the nonlinear model DNNGP was used, phenotype Y1_BV and genotype PCA 50 were selected, and the prediction ability reached the upper limit after 1000 iterations ($PA_{max} = 0.58$, $\overline{PA} = 0.50$, $PC_{max} = 0.71$, and $\overline{PC} = 0.62$) (Figure 8D). A total of 49 Masson pine individuals with known resistance (real R & real S) were selected for genomic prediction. The mean of the BVs from 100 repetitions of resistance predictions (pred R and pred S) was used as

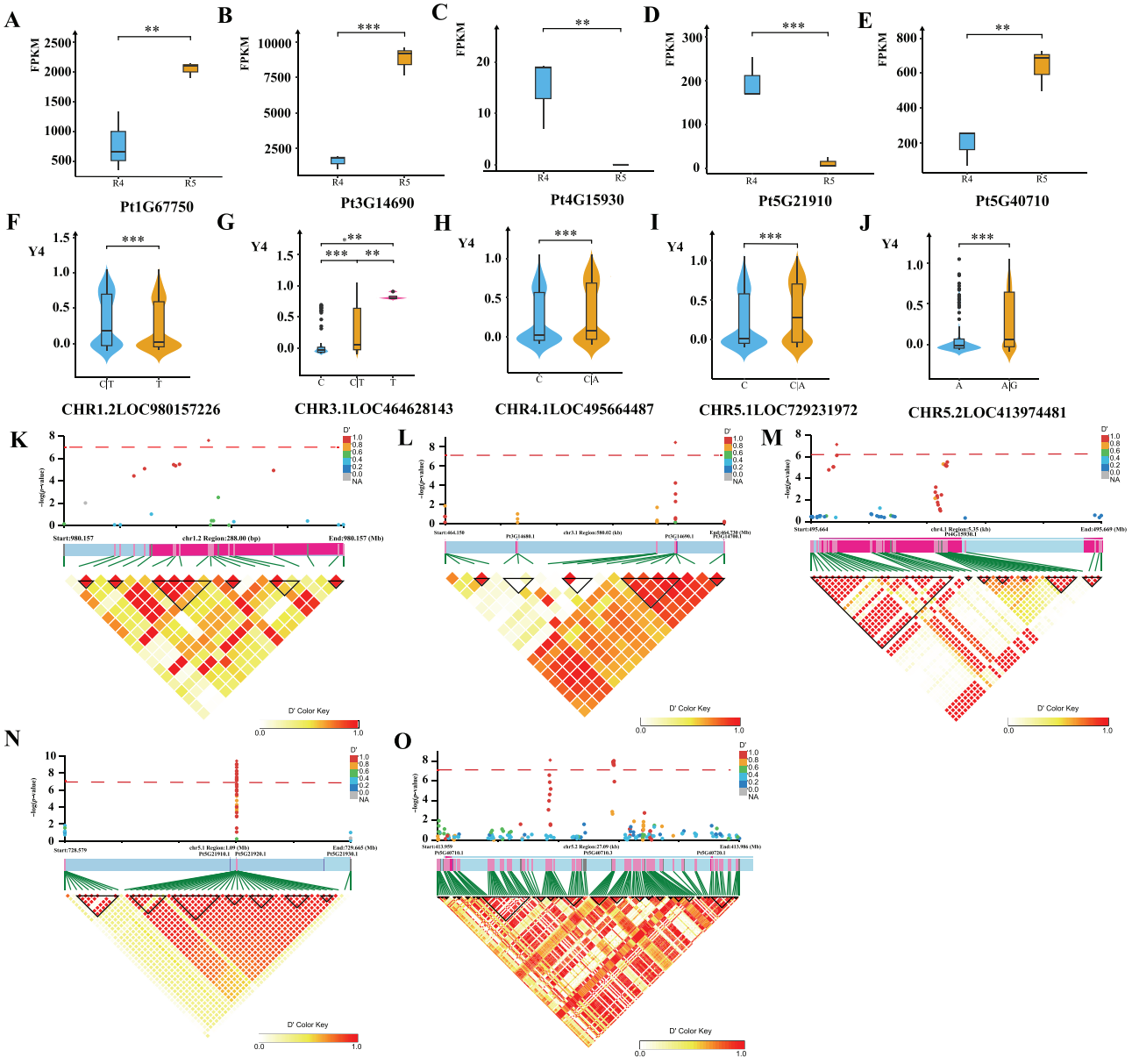

FIGURE 7 Five genes linked to disease resistance. (A–E) RNA-seq (fragments per kilobase of exon model per million mapped fragments [FPKM] value) was used to detect the gene expression of different haplotypes in different inoculation periods. Data are average values, and *, **, and *** indicate $p < 0.05$, 0.01, and 0.001, respectively. (F–J) Violin map of phenotypes based on single-nucleotide polymorphism (SNP) haplotypes. The center line represents the median, the box limits are the upper quartile and the lower quartile, and the whisker represents the data range. Significance analysis was performed using double-tail *T*-tests. (K–O) linkage disequilibrium (LD) block + genome-wide association study (GWAS) + annotation of interested genes. The scatter plots represent the *p*-value distribution of the SNPs, with the colors indicating the correlation degree. The gene annotation plot in the center shows the relative positions of genes and SNPs. The bottom diagram is the LD heatmap.

a ranking to evaluate the model's performance under optimal conditions (phenotype Y1_BV and 100K SNPs). A sample was recorded as "correct" if it was real R and pred R, or real S and pred S. Based on the order of the predicted BVs, different testing population proportions were selected, and the PC was the proportion of "correct" samples. The disease resistance PC reached 80% and 90% for the top 20% of the testing population (ordered by resistance GEBV) in gBLUP and DNNGP, respectively. As the observation range expanded, the PC gradually decreased, but the prediction accuracies for the top 40% of the testing population were still above 70% (Figure 8E).

4 | DISCUSSION

The slow progress in breeding resistance to PWD in Masson pine is primarily due to long generation times and extended breeding cycles. With the rapid spread of PWD, there is an

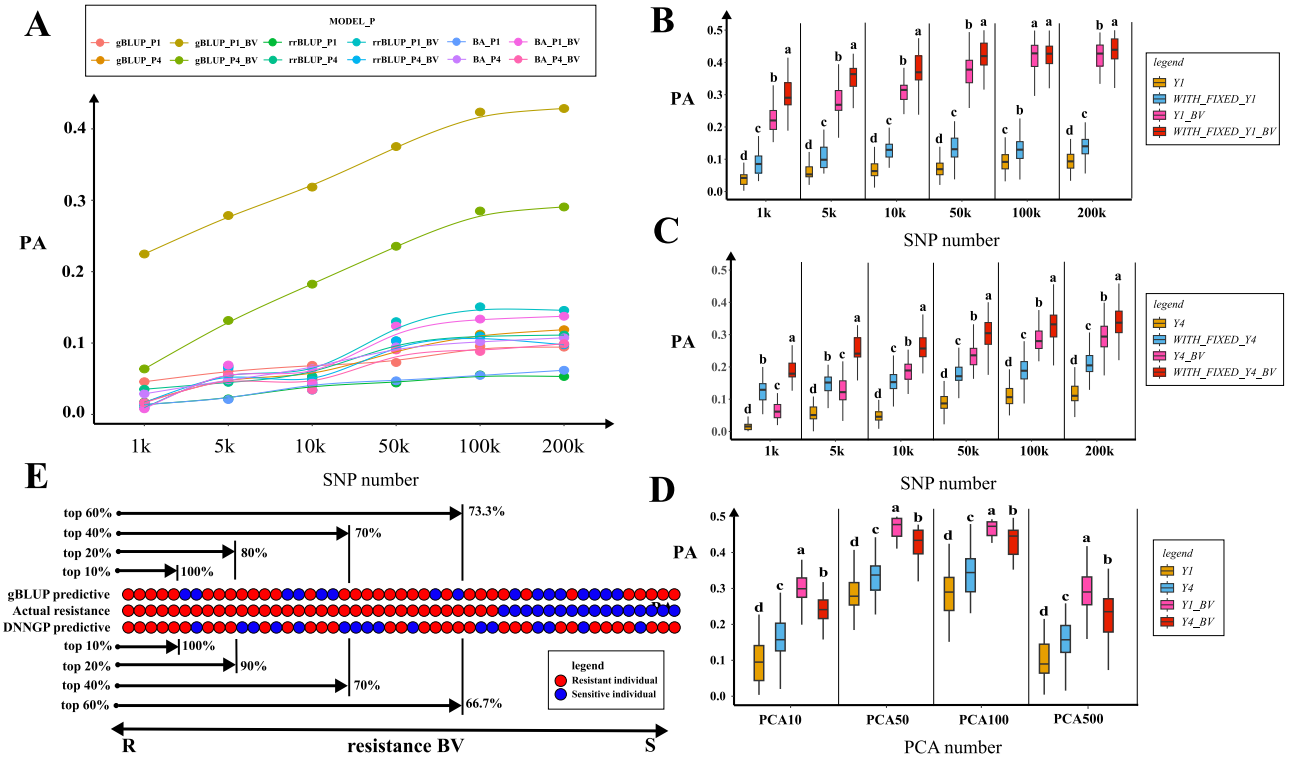

FIGURE 8 Variation in prediction ability under different conditions. (A) Different colored curves represent a combination of distinct models and phenotypes, illustrating how the predictive performance varies with changes in the number of single-nucleotide polymorphisms (SNPs). (B) Predictive ability of Y1 and related phenotypes under different SNP numbers. (C) Predictive ability of Y4 and related phenotypes under different SNP numbers. (D) Predictive ability of deep neural network-based method for genomic prediction (DNNGP) under different principal component analysis (PCA) numbers. (E) Comparison between the actual and predicted disease resistance in the validation population. On the left, higher predicted breeding values correspond to stronger resistance (R), and on the right, lower predicted breeding values indicate weaker resistance (S). Red and blue denote the actual disease resistance ability. BV, breeding value; gBLUP, genomic best linear unbiased prediction; PA, predictive ability.

urgent need to adopt genomic breeding strategies, including whole-genome association analysis and genomic selection, to accelerate resistance breeding efforts. To meet the demand for extensive genotyping, it is essential to develop rapid and cost-effective genetic typing tools.

4.1 | Benefits of the liquid-phased probe

The liquid-phased probe can capture all SNP sites within the design interval, enabling the determination of a large number of linked site groups. Compared to the single-point outputs with solid-phased chips (Kastally et al., 2022; Liu et al., 2016; Neves et al., 2013), the advantage of using these clusters for analysis lies in reducing the risk of errors and omissions and facilitating the presentation of the analytical results. Although such tight linkage in clusters may cause inflation in the *Q–Q* plots and PVE, excluding false positive markers by linkage ensures the reliability of the results. Given the vast genome of Masson pine, the clusters provided by the liquid-phased probe can offer more information and are more conducive to the targeted gene discovery task. Compared to mainstream solid-phased probes, the liquid-phased probe offers higher cost-effectiveness and greater flexibility, allowing for the addition or removal of panel SNPs as required. This enables us to fully utilize the analytical results to optimize and upgrade the probe.

To maximize coverage within a limited number of probes, the probes were distributed evenly rather than solely on the genes, with approximately half targeting genes (57,213) and the other half targeting intergenic regions (56,496). This strategy enhances the probe's versatility and aids other studies related to Masson pine. Probes in intergenic regions provide substantial information, as pseudo-genes within these regions can still have regulatory functions. Pseudo-genes are reported to integrate into upstream regulatory elements, intronic or UTR regions, and exonic regions, influencing gene expression through DNA sequence exchanges (Yadav et al., 2023). Genes encoding cyclin-dependent protein kinases have been identified as pseudo-genes in spruce (*Picea asperat*) (Kvarnheden et al., 1998). Numerous pseudo-genes have been reported in disease-resistant genes (R genes) in rice (Luo et al., 2012). Gene families such as BTB/POZ proteins, terpene synthases, chalcone synthases, and cytochrome P450 proteins are also

relatively enriched in pseudo-genes in rice (Zou et al., 2009). These gene families have been widely identified in studies related to PWD resistance in pine (Zhao & Li, 2008).

The quality of the reference genome assembly significantly impacts the efficiency of the liquid-phased probe capture. Initial studies on the whole exome capture of *P. taeda* and *P. elliottii* indicate that the capture efficiency of probes overlapping multiple exons decreases without a reference genome (Neves et al., 2014). Liu et al. (2016) designed oligonucleotide probes targeting 199,723 exons extracted from the *P. taeda* v1.0 reference genome, achieving an average mapping rate of 67% on sequencing data. Diao et al. (2024) designed liquid-phased probes using an improved assembly version (v2.01) of the *P. taeda* genome (Zimin et al., 2017), demonstrating improved efficiency with a mapping rate of up to 83.27%. Given the close phylogenetic relationship between *P. massoniana* and *P. tabuliformis*, we selected the published genome of Chinese pine as the reference genome. A total of 113,709 SNPs were designed, with an average of 102,204 capture SNPs per sample and an average probe capture efficiency of 89.88%. This improvement in the probe mapping success rate further validates the reliability of the Chinese pine reference genome. It is anticipated that future releases of more high-quality genomes and annotation information will contribute to enhancing the success and efficiency of probe design.

4.2 | Suitable phenotypic and model facilitate GWAS

Common quantitative traits such as tree height and diameter at breast height are easily obtained through direct measurement methods, providing continuous and normally distributed phenotype data. These data can quantify the effects of numerous minor alleles on correlated traits. However, converting qualitative disease resistance traits into continuous numerical variables that fully reflect genotype disease resistance is a complicated task. Due to the large size of woody plants, using the average incidence of clonal individuals to evaluate disease resistance in Masson pine presents challenges when assessing genotypic resistance. Therefore, developing a method that effectively reflects genotype-specific disease cycles and differences in disease resistance between genotypes is crucial. The AUDPC, which reflects the time of disease onset and the rate of disease progression, has been applied in crop disease-resistance phenotyping (Binalf et al., 2024; de Carvalho Paulino et al., 2021). We computed deviation values on a pedigree basis and performed secondary assignment to highlight the genetic basis of disease resistance within families. Deviations were weighted to harmonize numerical values. After data normalization, we obtained resistance phenotype values following a bimodal normal distribution, effectively reflecting inter-individual differences in resistance (Figure S5). To assess resistance to PWD in Masson pine, we adopted a phenotypic calculation method that considers the individual performance and familial genetic background, aiming to provide a reference for disease resistance research in other forest trees.

IIIVmrMLM, a novel tool for detecting nucleotides associated with quantitative traits, has been applied in fine-mapping studies of flax resistance to wilt disease, identifying 18 nucleotides associated with quantitative traits, with one major quantitative trait locus directly explaining 20%–48% of the phenotypic variation (Cloutier et al., 2024). Our analysis yielded 253 significant SNPs, collectively PVE from 0.466 to 0.487. The model offers fast computation speed and minimal memory usage, enhancing the analysis experience. However, results for non-recommended loci are not included in the output files, posing challenges to the presentation of the results. The LMM in the GENESIS package, developed for association analysis targeting disease-resistance traits, has been widely used in the medical field (Pan et al., 2022) and has recently been applied to crop disease resistance locus screening (Bhattarai et al., 2022, 2023). We obtained a total of 279 significant SNPs annotated to 55 candidate genes, demonstrating the high suitability of our disease-resistance phenotype calculation approach to mainstream GWAS analysis software. The substantial number of SNPs exhibits a significant phenotypic contribution. However, the PVE at individual loci remains below 5%, indicating the absence of genes with high effect sizes. This further corroborates the notion that PWD is a multifaceted outcome arising from the intricate interplay between numerous genes and pathogens, suggesting that the efficacy of disease resistance breeding—which relies predominantly on the enhancement of a limited number of genes—may be considerably constrained.

4.3 | Genetic loci of PWD resistance

WGS and LD decay analyses in Masson pine reveal an average linkage distance of 35 kb. Compared to crops, *P. massoniana* exhibits longer LD decay distances, which facilitates probe site coverage across larger gene sections. Based on the reference genome of *P. tabulaeformis*, the whole-genome length of the 12 chromosomes is 24,405,604,838 bp. Combined with the number (101.6K) and length (100 bp) of probes and the average LD (35 kb) of Masson pine, the theoretical maximum capture range is 7,112,000,000 bp, accounting for 29.14% of the whole genome. Using the average LD as a standard, a total of 169 genes were located on or linked to significant SNPs. A total of 118 of these genes were subsequently annotated, including common plant disease resistance-related protein families such as NBS-LRR (nucleotide-binding site-leucine-rich repeat), as well as families involved in cellular waste metabolism and detoxification,

including MatE, domains related to immune response and transmembrane transport such as TIR and PLAT domains, and the AP2/ERF transcription factor family.

In this study, we identified seven candidate disease-resistance genes belonging to the *LRR* gene family. NBS-LRR is the largest gene family associated with disease resistance in plants. These resistance proteins can directly or indirectly recognize pathogen effectors and may also identify invasive effector molecules with the assistance of host cells (Yuan et al., 2021). They mediate resistance to a variety of pathogens, including bacteria, fungi, oomycetes, viruses, and nematodes (H. Zhang et al., 2022). Recent research has shown that enhancing plant disease resistance can be achieved by targeting specific members of this gene family. For example, the *Pm21* gene encoding a CC-NBS-LRR protein confers resistance to wheat powdery mildew (Xing et al., 2018), providing a reference for improving resistance in Masson pine through candidate gene validation. NLR proteins with an amino-terminal Toll/interleukin-1 receptor (TIR) domain activate defense responses through the NADase activity of the TIR domain, playing a crucial role in pattern-triggered immunity (PTI), which is key in enhancing plant defenses during PTI (Tian et al., 2021). Our results indicate that the downstream protein of Pt1G00120 contains a TIR domain and exhibits significant association in association and transcriptome analyses, suggesting its potential involvement in the resistance response to PWD invasion. The PtJG01510 gene located on chromosome 11 of Masson pine is annotated to include an AP2 domain. As an important transcription factor family in plants, AP2/ERF includes many genes that encode proteins involved in controlling disease-resistance pathways (Gutterson & Reuber, 2004). They regulate various developmental and stress response pathways through mechanisms such as transcriptional and posttranscriptional regulation (Gibbs et al., 2015; Phukan et al., 2017).

The disease development process of PWD involves extensive transport and the information exchange of membrane-associated substances. Therefore, membrane-bound recognition proteins may play a critical role in plant disease resistance (Futai, 2013). The PLAT domain appears in different membrane- or lipid-associated proteins as either single or repeated domains (Shin et al., 2004) and is involved in protein–protein and protein–membrane interactions (Hu & Barr, 2005). The identified Pt5G21910 and Pt3G14690 both have PLAT domains in their translated protein structures. Although research on plant PLAT domain proteins is limited, there is some evidence linking them to non-biological stress tolerance in plants (Hyun et al., 2015). The MatE family, as transmembrane transport proteins, drives the absorption of heavy metals, transporting them to the leaves where they are ultimately sequestered in vacuoles or cell walls (Rascio & Navari-Izzo, 2011). Family members are directly or indirectly involved in disease resistance, aluminum detoxification, toxic metal efflux, secondary metabolite transport, and plant hormone transport. The Pt1G62620 gene identified in this study encodes a protein belonging to the MatE family and may participate in the disease resistance process. Specific functions of some MatE genes have been characterized in *Arabidopsis*; for example, *ADS1* regulates plant disease resistance by encoding a MatE transporter (Sun et al., 2011).

4.4 | GS potential in disease-resistant improvement

Among all models, DNNGP, a nonlinear model based on machine learning, achieved the best prediction effect, with average and maximum prediction accuracies of 0.62 and 0.71, respectively. These accuracies were slightly higher than those of gBLUP, determined as the optimal linear model. This may be the result of successfully capturing the nonadditive effect. In addition to the high PC, this model also has a faster calculation speed compared with the conventional linear model, demonstrating the great potential of machine learning for GS applications.

We also found that under the same conditions, using the calculated BVs as phenotypic data can significantly improve the prediction ability of the gBLUP model. This may be because the GCTA and gBLUP calculations are based on the G-matrix. When calculating the BV, GCTA first calculates the G-matrix and then calculates the variance of the phenotypic data according to the G-matrix to obtain the BV of each sample. Including the BV in the gBLUP model as the fitting result of the G-matrix can reduce the divergence of the model and improve the PC. Under scenarios with an adequate number of SNPs, the ρ2max under the gBLUP model can reach 0.56. Compared to other disease resistance predictions, such as Wen's study on soybean resistant to white mold disease with GS prediction accuracies of 0.62–0.64 (Zhang et al., 2018), the PC in this study is also at a high level.

When conducting GS analysis with large sample sizes, challenges such as extensive memory space requirements and lengthy computation times must be addressed. We found that incorporating significant SNPs as fixed factors in gBLUP can help the model achieve the desired PC earlier, a strategy supported by multiple studies (Anilkumar et al., 2023; Z. Chen et al., 2023; Kim et al., 2022). By adding 50 significant SNPs in GWAS as fixed factors, the required number of SNPs to achieve optimal PC decreased from 100 to 50K, a 50% reduction. This implies that by simplifying covariate settings, it is possible to reduce the number of required SNPs, thereby decreasing the computational memory and time required, which aids in improving computational efficiency. Therefore, conducting GWAS analysis before GS is necessary, as the significant SNPs identified by GWAS have practical value in

subsequent GS analysis. We also tested the impact of the number of significant SNP fixed factors on the model accuracy and found that increasing their number did not raise the upper limit of the model accuracy nor decrease the required number of SNPs to achieve the desired PC. Instead, it exacerbated linear dependencies among factors, making model fitting more challenging. In extreme cases, an excessive number of fixed factors can lead to matrix singularity, rendering the model uncomputable. Therefore, the number of fixed factors should be moderated appropriately to maintain model stability and effectiveness.

Genomic selection has proven effective in screening for disease-resistant Masson pine germplasm, demonstrating considerable practical application potential. We applied the constructed GS model to predict resistance in 49 samples that were unrelated to the modeling population. The PC reached 80% when focusing on the top 20% of the samples. Although the PC gradually declined as the observation range increased, it consistently remained above 70%. This demonstrates that GS can support early resistance selection to some extent, providing reliable predictions even under substantial selection pressure.

5 | CONCLUSION

In this study, 1013 seedlings of Masson pine were inoculated with PWNs, and five groups of phenotype data were computed. To acquire extensive genome-wide SNP data, a Masson pine 101.6K liquid-phased probe was designed and developed based on resequencing data and transcriptome analysis data from 30 Masson pine samples. The probe targeted a total of 113,709 SNPs, which are relatively evenly distributed across the whole genome. Targeted sequencing of 1013 Masson pine seedlings by the liquid-phased probe was conducted to obtain a large amount of high-quality SNP information. The experimental population was effectively divided into three subgroups according to geographical characteristics through population structure analysis. The phenotypic and genotypic data of the 1013 Masson pine seedlings were used for GWAS and as a training population for GS. For GWAS, we employed the IIIVmrMLM, logistic, SUPER, and FarmCPU models to identify a resistance candidate SNP set comprising 548 significant SNPs. In addition, we annotated a resistance candidate gene set containing 169 functional genes. This set includes NBS-LRR, AP2/ERF, and other important protein families and transcription factors related to disease resistance. Four models were used for GS analysis: gBLUP, rrBLUP, BA, and DNNGP. Among these, DNNGP demonstrated the highest PC, with a maximum accuracy of 0.71. The accuracy of the actual disease resistance prediction was 90% in the top 20% of the resistance samples of the testing populations. The liquid-phased probe can economically realize batch sequencing of a large number of samples, thus meeting the genomic research needs of a large amount of Masson pine genetic information, improving our understanding of the genetic basis of Masson pine resistance to PWD, and providing an economical and effective method for the genotyping of large pine genomes. The disease-resistant candidate SNP and gene sets identified through GWAS can facilitate studies on the pathogenesis of PWD. The GS model was developed to expedite the breeding process for new resistant varieties of Masson pine through early selection.

AUTHOR CONTRIBUTIONS

Jingyi Zhu: Data curation; formal analysis; investigation; methodology; software; writing—original draft. **Qinghua Liu**: Conceptualization; project administration; resources; supervision; writing—review and editing. **Shu Diao**: Methodology; supervision; writing—review and editing. **Zhichun Zhou**: Supervision. **Yangdong Wang**: Supervision. **Xianyin Ding**: Methodology; writing—review and editing. **Mingyue Cao**: Methodology; resources. **Dinghui Luo**: Data curation; investigation.

ACKNOWLEDGMENTS

We are grateful to thank all of the researchers and the students in forest tree genetics and breeding research group of the Research Institute of Subtropical Forestry, Chinese Academy of Forestry. This research was funded by the Science and Technology Innovation 2030 Program, Ministry of Science and Technology, People's Republic of China (2022ZD0401603), Zhejiang Science and Technology Major Program on Agricultural, New Variety Breeding, Department of Science and Technology of Zhejiang Province, People's Republic of China (2021C02070-5-2), and the Fundamental Research Funds for the Central Non-profit Research Institute of CAFs. (grant no. CAFYBB2021ZG001-4).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The 101.6K liquid-phased probe set is freely available and can be found on GitHub (https://github.com/nbsx510/Masson-Pine-101.6K-probe). More data will be made available on request.

ORCID

Jingyi Zhu https://orcid.org/0009-0001-0120-0253
Mingyue Cao https://orcid.org/0000-0001-5924-6038

REFERENCES

Alexander, D. H., & Lange, K. (2011). Enhancements to the admixture algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*, Article 246. https://doi.org/10.1186/1471-2105-12-246

Anilkumar, C., Muhammed Azharudheen, T. P., Sah, R. P., Sunitha, N. C., Devanna, B. N., Marndi, B. C., & Patra, B. C. (2023). Gene based markers improve precision of genome-wide association studies and accuracy of genomic predictions in rice breeding. *Heredity*, *130*(5), 335–345. https://doi.org/10.1038/s41437-023-00599-5

Bhattarai, G., Olaoye, D., Mou, B., Correll, J. C., & Shi, A. (2022). Mapping and selection of downy mildew resistance in spinach cv. Whale by low coverage whole genome sequencing. *Frontiers in Plant Science*, *13*, 1012923. https://doi.org/10.3389/fpls.2022.1012923

Bhattarai, G., Shi, A., Mou, B., & Correll, J. C. (2023). Skim resequencing finely maps the downy mildew resistance loci rpf2 and rpf3 in spinach cultivars whale and lazio. *Horticulture Research*, *10*, uhad076. https://doi.org/10.1093/hr/uhad076

Binalf, L., Shifa, H., & Tadesse, W. (2024). Association mapping of septoria tritici blotch resistance in bread wheat in bale and arsi highlands, ethiopia. *Heliyon*, *10*(6), e32265. https://doi.org/10.1016/j.heliyon.2024.e32265

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). Tassel: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Caballero, M., Lauer, E., Bennett, J., Zaman, S., McEvoy, S. L., Acosta, J. J., Jackson, C., Townsend, L., Eckert, A. J., Whetten, R., Loopstra, C. A., Holliday, J. A., Mandal, M., Wegrzyn, J. L., & Isik, F. (2021). Toward genomic selection in *Pinus taeda*: Integrating resources to support array design in a complex conifer genome. *Applications in Plant Sciences*, *9*(6), 11439. https://doi.org/10.1002/aps3.11439

Cantalapiedra, C. P., Hernández-Plaza, A., Letunić, I., Bork, P., & Huerta-Cepas, J. (2021). Eggnog-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Cappa, E. P., de Lima, B. M., Junior, O. B. S., Garcia, C. C., Mansfield, S. D., & Grattapaglia, D. (2019). Improving genomic prediction of growth and wood traits in eucalyptus using phenotypes from non-genotyped trees by single-step gblup. *Plant Science*, *284*, 9–15. https://doi.org/10.1016/j.plantsci.2019.03.017

Chen, C., Chen, H., Zhang, Y., Thomas, H., Frank, M., He, Y., & Xia, R. (2020). Tbtools: An integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant*, *13*(8), 1194–1202. https://doi.org/10.1016/j.molp.2020.06.009

Chen, Z., Baison, J., Jin, P., Karlsson, B., Andersson, B., Westin, J., Gil, M. R. G., & Wu, H. X. (2018). 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*, *19*(1), Article 946. https://doi.org/10.1186/s12864-018-5256-y

Chen, Z., Klingberg, A., Hallingbäck, H. R., & Wu, H. (2023). Preselection of qtl markers enhances accuracy of genomic selection in norway spruce. *BMC Genomics*, *24*(1), Article 147. https://doi.org/10.1186/s12864-023-09250-3

Cloutier, S., Edwards, T., Zheng, C., Booker, H., Islam, T., Nabetani, K., Kutcher, H. R., Molina, O. E., & You, F. M. (2024). Fine-mapping of a major locus for fusarium wilt resistance in flax (*Linum usitatissimum* L.). *Theoretical and Applied Genetics*, *137*(1), Article 27. https://doi.org/10.1007/s00122-023-04528-2

Conomos, M. P., Miller, M. B., & Thornton, T. A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, *39*(4), 276–293. https://doi.org/10.1002/gepi.21896

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, *98*(1), 127–148. https://doi.org/10.1016/j.ajhg.2015.11.022

Danecek, P., Auton, A., Abecasis, G. R., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and vcftools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Carvalho Paulino, J. F., de Almeida, C. P., Bueno, C. J., Song, Q., Neto, R. F., Carbonell, S. A. M., Chiorato, A. F., & Benchimol-Reis, L. L. (2021). Genome-wide association study reveals genomic regions associated with fusarium wilt resistance in common bean. *Genes*, *12*(5), 765. https://doi.org/10.3390/genes12050765

De La Torre, A. R., Puiu, D., Crepeau, M. W., Stevens, K., Salzberg, S. L., Langley, C. H., & Neale, D. B. (2018). Genomic architecture of complex traits in loblolly pine. *New Phytologist*, *221*(4), 1789–1801. https://doi.org/10.1111/nph.15535

Diao, S., Ding, X., Luan, Q., Chen, Z., Wu, H. X., Li, X., Zhang, Y., Sun, J., Wu, Y., Zou, L. H., & Jiang, J. (2024). Development of 51 k liquid-phased probe array for loblolly and slash pines and its application to gwas of slash pine breeding population. *Industrial Crops and Products*, *216*, 118777. https://doi.org/10.1016/j.indcrop.2024.118777

Diao, S., Hou, Y., Xie, Y., & Sun, X. (2016). Age trends of genetic parameters, early selection and family by site interactions for growth traits in *Larix kaempferi* open-pollinated families. *BMC Genomic Data*, *17*(1), Article 104. https://doi.org/10.1186/s12863-016-0400-7

Dong, S., He, W., Ji, J., Zhang, C., Guo, Y., & Yang, T. L. (2020). LDBlockShow: A fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Briefings in Bioinformatics*, *22*(4), bbaa227. https://doi.org/10.1093/bib/bbaa227

Du, X., Xu, W., Peng, C., Li, C., Zhang, Y., & Hu, L. (2021). Identification and validation of a novel locus, qpm-3bl, for adult plant resistance to powdery mildew in wheat using multilocus gwas. *BMC Plant Biology*, *21*(1), Article 357. https://doi.org/10.1186/s12870-021-03093-4

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrBLUP. *The Plant Genome*, *4*(3), 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Freed, D., Aldana, R., Ja, W., & Js, E. (2017). The sentieon genomics tools—A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv*. https://doi.org/10.1101/115717

Futai, K. (2013). Pine wood nematode, *Bursaphelenchus xylophilus*. *Annual Review of Phytopathology*, *51*, 61–83. https://doi.org/10.1146/annurev-phyto-081211-172910

Gibbs, D. J., Conde, J. V., Berckhan, S., Prasad, G., Mendiondo, G. M., & Holdsworth, M. J. (2015). Group vii ethylene response factors coordinate oxygen and nitric oxide signal transduction and stress responses in plants. *Plant Physiology*, *169*(1), 23–31. https://doi.org/10.1104/pp.15.00338

Gogarten, S. M., Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., Rice, K., & Conomos, M. P. (2019). Genetic association testing using the genesis r/bioconductor package. *Bioinformatics*, *35*(24), 5346–5348. https://doi.org/10.1093/bioinformatics/btz567

Gutterson, N., & Reuber, T. L. (2004). Regulation of disease resistance pathways by ap2/erf transcription factors. *Current Opinion in Plant Biology*, *7*(4), 465–471. https://doi.org/10.1016/j.pbi.2004.04.007

Hu, J., & Barr, M. M. (2005). Atp-2 interacts with the plat domain of lov-1 and is involved in caenorhabditis elegans polycystin signaling. *Molecular Biology of the Cell*, *16*(2), 458–469. https://doi.org/10.1091/mbc.e04-09-0851

Hyun, T. K., Albacete, A., van der Graaff, E., Eom, S. H., Großkinsky, D. K., Böhm, H., Janschek, U., Rim, Y., Ali, W. W., Kim, S. Y., & Roitsch, T. (2015). The arabidopsis plat domain protein1 promotes abiotic stress tolerance and growth in tobacco. *Transgenic Research*, *24*(4), 651–663. https://doi.org/10.1007/s11248-015-9868-6

Isik, F., Bartholomé, J., Farjat, A. E., Chancerel, É., Raffin, A., Sánchez, L., Plomion, C., & Bouffier, L. (2016). Genomic selection in maritime pine. *Plant Science*, *242*, 108–119. https://doi.org/10.1016/j.plantsci.2015.08.006

Jackson, C., Christie, N., Reynolds, S. M., Marais, C., Tii-Kuzu, Y., Caballero, M., Kampman, T., Visser, E. A., Naidoo, S., Kain, D., Whetten, R., Isik, F., Wegrzyn, J., Hodge, G. R., Acosta, J. J., & Myburg, A. A. (2021). A genome-wide SNP genotyping resource for tropical pine tree species. *Molecular Ecology Resources*, *22*(2), 695–710. https://doi.org/10.1111/1755-0998.13484

Kastally, C., Niskanen, A. K., Perry, A., Kujala, S., Avia, K., Cervantes, S., Haapanen, M., Kesälahti, R., Kumpula, T. A., Mattila, T. M., Ojeda, D. I., Tyrmi, J., Wachowiak, W., Cavers, S., Kärkkäinen, K., Savolainen, O., & Pyhäjärvi, T. (2022). Taming the massive genome of scots pine with pisy50k, a new genotyping array for conifer research. *The Plant Journal*, *109*(5), 1337–1350. https://doi.org/10.1111/tpj.15628

Kim, G. W., Hong, J., Lee, H., Kwon, J., Kim, D., & Kang, B. C. (2022). Genomic selection with fixed-effect markers improves the prediction accuracy for capsaicinoid contents in *capsicum annuum*. *Horticulture Research*, *9*, uhac204. https://doi.org/10.1093/hr/uhac204

Kvarnheden, A., Albert, V. A., & Engström, P. (1998). Molecular evolution of cdc2 pseudogenes in spruce (*picea*). *Plant Molecular Biology*, *36*(5), 767–774. https://doi.org/10.1023/a:1005901413475

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., & Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, *91*(2), 224–237. https://doi.org/10.1016/j.ajhg.2012.06.007

Letunić, I., & Bork, P. (2021). Interactive tree of life (itol) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(1), 293–296. https://doi.org/10.1093/nar/gkab301

Li, D., Li, Y., Wang, X., Zhang, W., Wen, X., Liu, Z., Feng, Y., & Zhang, X. (2023). Engineered pine endophytic bacillus toyonensis with nematocidal and colonization abilities for pine wilt disease control. *Frontiers in Microbiology*, *14*, 1240984. https://doi.org/10.3389/fmicb.2023.1240984

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, R. E., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. R., & Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, M., Li, H., Ding, X., Wang, L., Wang, X., & Chen, F. (2022). The detection of pine wilt disease: A literature review. *International Journal of Molecular Sciences*, *23*(18), 10797. https://doi.org/10.3390/ijms231810797

Li, M., Zhang, Y., Xiang, Y. T., Liu, M., & Zhang, Y. (2022). Iiivmrmlm: The r and c++ tools associated with 3vmrmlm, a comprehensive gwas method for dissecting quantitative traits. *Molecular Plant*, *15*(8), 1251–1253. https://doi.org/10.1016/j.molp.2022.06.002

Liu, F., Baye, W., Zhao, K., Tang, S., Xie, Q., & Xie, P. (2024). Unravelling sorghum functional genomics and molecular breeding: Past achievements and future prospects. *Journal of Genetics and Genomics*. https://doi.org/10.1016/j.jgg.2024.07.016

Liu, J. J., Schoettle, A. W., Sniezko, R. A., Sturrock, R. N., Zamany, A., Williams, H., Ha, A., Chan, D., Danchok, B., Savin, D. P., & Kegley, A. (2016). Genetic mapping of *Pinus flexilis* major gene (cr4) for resistance to white pine blister rust using transcriptome-based SNP genotyping. *BMC Genomics*, *17*(1), Article 735. https://doi.org/10.1186/s12864-016-3079-2

Liu, Q., Wei, Y., Xu, L., Hao, Y., Chen, X., & Zhou, Z. (2017). Transcriptomic profiling reveals differentially expressed genes associated with pine wood nematode resistance in masson pine (*Pinus massoniana* Lamb.). *Scientific Reports*, *43*(6), 995–1008. https://doi.org/10.1038/s41598-017-04944-7

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genetics*, *12*(3), 1005767. https://doi.org/10.1371/journal.pgen.1005767

Luo, S., Zhang, Y., Hu, Q., Chen, J., Li, K., Chen, L., Liu, H., Wang, W., & Kuang, H. (2012). Dynamic nucleotide-binding site and leucine-rich repeat-encoding genes in the grass family. *Plant Physiology*, *159*(1), 197–210. https://doi.org/10.1104/pp.111.192062

Mamiya, Y., & Kiyohara, T. (1972). Description of *Bursaphelenchus lignicolus* n. Sp. (Nematoda: Aphelenchoididae) from pine wood and histopathology of nematode-infested trees. *Nematologica*, *18*, 120–124.

McKenna, A., Hanna, M. G., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. J., & DePristo, M. A. (2010). The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meng, Y., Zhang, W., Cheng, Y., Wu, Y., Wu, H., He, M., Chen, S., Man, C., Gao, H., Du, L., Chen, Q., & Wang, F. (2024). Development and verification of a 10k liquid chip for hainan black goat based on genotyping by pinpoint sequencing of liquid captured targets. *BMC Genomic Data*, *25*(1), Article 44. https://doi.org/10.1186/s12863-024-01228-8

Merrick, L. F., Herr, A. W., Sandhu, K. S., Lozada, D. N., & Carter, A. H. (2022). Optimizing plant breeding programs for genomic selection. *Agronomy*, *12*(3), 714. https://doi.org/10.3390/agronomy12030714

Modesto, I., Mendes, A., Carrasquinho, I., & Miguel, C. M. G. (2022). Molecular defense response of pine trees (*Pinus* spp.) To the parasitic nematode *Bursaphelenchus xylophilus*. *Cells*, *11*(20), 3208. https://doi.org/10.3390/cells11203208

Mota, M., Braasch, H., Bravo, M. A., Penas, A. C., Burgermeister, W., Metge, K., & Sousa, E. (1999). First report of *Bursaphelenchus xylophilus* in Portugal and in Europe. *Nematology*, *1*, 727–734. https://doi.org/10.1163/156854199508757

Neves, L. G., Davis, J. M., Barbazuk, W. B., & Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. *The Plant Journal*, *75*(1), 146–156. https://doi.org/10.1111/tpj.12193

Neves, L. G., Davis, J. M., Barbazuk, W. B., & Kirst, M. (2014). A high-density gene map of loblolly pine (*Pinus taedal*.) Based on exome sequence capture genotyping. *G3: Genes, Genomes, Genetics*, *4*(1), 29–37. https://doi.org/10.1534/g3.113.008714

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., … Wu, H. X. (2021). The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell*, *185*(1), 204–217. https://doi.org/10.1016/j.cell.2021.12.006

Olatoye, M. O., Clark, L. V., Wang, J., Yang, X., Yamada, T., Sacks, E. J., & Lipka, A. E. (2019). Evaluation of genomic selection and marker-assisted selection in miscanthus and energycane. *Molecular Breeding*, *39*, Article 171. https://doi.org/10.1007/s11032-019-1081-5

Pan, Y., Sun, X., Mi, X., Huang, Z., Hsu, Y. Y., Hixson, J. E., Munzy, D., Metcalf, G., Franceschini, N., Tin, A., Köttgen, A., Francis, M., Brody, J. A., Kestenbaum, B., Sitlani, C. M., Mychaleckyj, J. C., Kramer, H., Lange, L. A., Guo, X., … Kelly, T. N. (2022). Whole-exome sequencing study identifies four novel gene loci associated with diabetic kidney disease. *Human Molecular Genetics*, *32*(6), 1048–1060. https://doi.org/10.1093/hmg/ddac290

Pérez, P., & de Los Campos, G. (2014). Genome-wide regression and prediction with the bglr statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Perry, A., Wachowiak, W., Downing, A., Talbot, R., & Cavers, S. (2020). Development of a single nucleotide polymorphism array for population genomic studies in four European pine species. *Molecular Ecology Resources*, *20*(6), 1697–1705. https://doi.org/10.1111/1755-0998.13223

Phukan, U. J., Jeena, G. S., Tripathi, V., & Shukla, R. (2017). Regulation of apetala2/ethylene response factors in plants. *Frontiers in Plant Science*, *8*, Article 150. https://doi.org/10.3389/fpls.2017.00150

Plomion, C., Bartholomé, J., Lesur, I., Boury, C., Quilón, I. R., Lagraulet, H., Ehrenmann, F., Bouffier, L., Gion, J. M., Grivet, D., de Miguel, M., de María, N., Cervera, M. T., Bagnoli, F., Isik, F., Vendramin, G. G., & Martínez, S. C. G. (2015). High-density snp assay development for genetic analysis in maritime pine (*Pinus pinaster*). *Molecular Ecology Resources*, *16*(2), 574–587. https://doi.org/10.1111/1755-0998.12464

Purcell, S., Neale, B. M., Brown, K. T., Thomas, L., Ferreira, M., Bender, D. B., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rascio, N., & Navari-Izzo, F. (2011). Heavy metal hyperaccumulating plants: How and why do they do it? And what makes them so interesting? *Plant Science*, *180*(2), 169–181. https://doi.org/10.1016/j.plantsci.2010.08.016

Shaner, G. (1977). The effect of nitrogen fertilization on the expression of slow-mildewing resistance in knox wheat. *Phytopathology*, *77*, 1051. https://doi.org/10.1094/phyto-67-1051

Shin, R., An, J. M., Park, C. J., Kim, Y. J., Joo, S., Kim, W. T., & Paek, K. H. (2004). *Capsicum annuum* tobacco mosaic virus-induced clone 1 expression perturbation alters the plant's response to ethylene and interferes with the redox homeostasis. *Plant Physiology*, *135*(1), 561–573. https://doi.org/10.1104/pp.103.035436

Sun, X., Gilroy, E. M., Chini, A., Nurmberg, P. L., Hein, I., Lacomme, C., Birch, P. R. J., Hussain, A., Yun, B. W., & Loake, G. J. (2011). Ads1 encodes a mate-transporter that negatively regulates plant disease resistance. *New Phytologist*, *192*(2), 471–482. https://doi.org/10.1111/j.1469-8137.2011.03820.x

Tamura, K., Stecher, G., & Kumar, S. (2021). Mega11: Molecular evolutionary genetics analysis version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/molbev/msab120

Tian, H., Wu, Z., Chen, S., Ao, K., Huang, W., Yaghmaiean, H., Sun, T., Xu, F., Zhang, Y., Wang, S., Li, X., & Zhang, Y. (2021). Activation of tir signalling boosts pattern-triggered immunity. *Nature*, *598*(7881), 500–503. https://doi.org/10.1038/s41586-021-03987-1

Vikas, V. K., Pradhan, A. K., Budhlakoti, N., Mishra, D. C., Chandra, T., Bhardwaj, S., Kumar, S., Sivasamy, M., Jayaprakash, P., Rani, N., Shajitha, P., Peter, J., Geetha, M., Mir, R. R., Singh, K., & Kumar, S. (2022). Multi-locus genome-wide association studies (ml-gwas) reveal novel genomic regions associated with seedling and adult plant stage leaf rust resistance in bread wheat (*Triticum aestivum* L.). *Heredity*, *128*(6), 434–449. https://doi.org/10.1038/s41437-022-00525-1

Wang, J., & Zhang, Z. (2021). Gapit version 3: Boosting power and accuracy for genomic association and prediction. *Genomics, Proteomics & Bioinformatics*, *19*(4), 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, K., Abid, M., Rasheed, A., Crossa, J., Hearne, S., & Li, H. (2023). Dnngp, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*, *16*(1), 279–293. https://doi.org/10.1016/j.molp.2022.11.004

Wang, K., Li, M., & Hákonarson, H. (2010). Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., & Zhang, Z. (2014). A super powerful method for genome wide association study. *PLoS One*, *9*(9), 107684. https://doi.org/10.1371/journal.pone.0107684

Xing, L., Hu, P., Liu, J., Witek, K., Zhou, S., Xu, J., Zhou, W., Gao, L., Huang, Z., Zhang, R., Wang, X., Chen, P., Wang, H., Jones, J. D. G., Karafiátová, M., Vrána, J., Bartoš, J., Doležel, J., Tian, Y., … Cao, A. (2018). *Pm21* from *Haynaldia villosa* encodes a CC-NBS-LRR protein conferring powdery mildew resistance in wheat. *Molecular Plant*, *11*(6), 874–878. https://doi.org/10.1016/j.molp.2018.02.013

Xu, L., & Tadao, T. (2006). Selection and evaluation of pine wood nematode resistant candidate for *Pinus massoniana*. *Anhui Agricultural Sciences*, *39*(2), 8–10. (In Chinese). https://doi.org/10.13989/j.cnki.0517-6611.2006.17.059

Yadav, S., Kalwan, G., Meena, S., Gill, S. S., Yadava, Y. K., Gaikwad, K., & Jain, P. (2023). Unravelling the due importance of pseudogenes and their resurrection in plants. *Plant Physiology and Biochemistry*, *203*, 108062. https://doi.org/10.1016/j.plaphy.2023.108062

Yang, J., Lee, S., Goddard, M. E., & Visscher, P. M. (2011). Gcta: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yuan, C., Li, C., Zhao, X., Cheng, Y., Wang, J., Mou, Y., Sun, Q., & Shan, S. (2021). Genome-wide identification and characterization of hsp90-rar1-sgt1-complex members from arachis genomes and their responses to biotic and abiotic stresses. *Frontiers in Genetics*, *12*, 689669. https://doi.org/10.3389/fgene.2021.689669

Zhang, C., Dong, S., Xu, J., He, W., & Yang, T. L. (2018). Popldecay: A fast and effective tool for linkage disequilibrium decay analysis

based on variant call format files. *Bioinformatics*, *35*(10), 1786–1788. https://doi.org/10.1093/bioinformatics/bty875

Zhang, H., Zi, Y., Liu, Z., Sun, Y., Li, X., Wu, J., Zhou, G., & Wan, Y. (2022). The cassava nbs-lrr genes confer resistance to cassava bacterial blight. *Frontiers in Plant Science*, *13*, 790140. https://doi.org/10.3389/fpls.2022.790140

Zhang, R., Jia, G., & Diao, X. (2023). Genehapr: An r package for gene haplotypic statistics and visualization. *BMC Bioinformatics*, *24*(1), Article 199. https://doi.org/10.1186/s12859-023-05318-9

Zhang, W., Tan, R., Zhang, S., Collins, P. J., Yuan, J., Du, W., Gu, C., Ou, S., Song, Q., An, Y., Boyse, J. F., Chilvers, M. I., & Wang, D. (2018). Integrating GWAS and gene expression data for functional characterization of resistance to white mould in soya bean. *Plant Biotechnology Journal*, *16*(11), 1825–1835. https://doi.org/10.1111/pbi.12918

Zhao, B. G., & Li, R. G. (2008). The role of bacteria associated with the pine wood nematode in pathogenicity and toxin-production related to pine wilt. In B. Zhao, K. Futai, J. Sutherland, & Y. Takeuchi (Eds.), *Pine wilt disease* (pp. 250–259) Springer.

Zimin, A. V., Stevens, K., Crepeau, M. W., Puiu, D., Wegrzyn, J. L., Yorke, J. A., Langley, C. H., Neale, D. B., & Salzberg, S. L. (2017). An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience*, *6*(1), giw016. https://doi.org/10.1093/gigascience/giw016

Zou, C., Lehti-Shiu, M. D., Nissen, F. T., Prakash, T., Buell, C. R., & Shiu, S. H. (2009). Evolutionary and expression signatures of pseudogenes in arabidopsis and rice. *Plant Physiology*, *151*(1), 3–15. https://doi.org/10.1104/pp.109.140632

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhu, J., Liu, Q., Diao, S., Zhou, Z., Wang, Y., Ding, X., Cao, M., & Luo, D. (2025). Development of a 101.6K liquid-phased probe for GWAS and genomic selection in pine wilt disease-resistance breeding in Masson pine. *The Plant Genome*, *18*, e70005. https://doi.org/10.1002/tpg2.70005