

RESEARCH

Open Access



Development and application of a genotyping by target sequencing single-nucleotide polymorphism array panel in *Salix suchowensis*

Yu Han^{1†}, Shaobo Gu^{1†}, Mingjia Zhu², Wei Liu¹, Landi Feng¹, Tongming Yin³, Xuemeng Gao¹, Yanjun Zan^{4,5}, Rui Huang⁶, Yan Ji^{4,5*} and Jianquan Liu^{2*}

Abstract

Salix suchowensis is an important species of *Salix*, known for its rapid growth property and wide application in environmental construction, ecological restoration, wicker production, and biomass energy production. Due to its significance as a sustainable biological resource, *S. suchowensis* has been the centre of intensive breeding. However, rapid improvement of growth and biomass has been hindered by a lack of genomic resources. To address this limitation, we designed a liquid-phase probe array by genotyping by target sequencing technology. Using whole-genome resequencing data, a total of 39,076 SNPs were selected for the array panel, consisting of trait-associated SNPs, intragenic SNPs, and intergenic SNPs. This panel was validated by genotyping 550 new samples, demonstrating high call rates and effective capture of the population structure. Genome-wide association analysis identified 72 SNPs associated with plant height and ground diameter. Additionally, the array panel shows a high potential for genomic selection, with high prediction accuracy for various traits. These results highlight the efficiency of this panel in capturing genomic variations that are highly valuable for future genetic research and breeding applications.

Keywords Liquid-phase probe array, Genome-wide association study, Genomic selection, Plant height, Ground diameter

[†]Yu Han and Shaobo Gu contributed equally.

*Correspondence:

Yan Ji

jijian@caas.cn

Jianquan Liu

liujq@nwpb.ac.cn

¹College of Life Sciences, Sichuan University, Chengdu, China

²State Key Laboratory of Herbage Innovation and Grassland Agro-Ecosystem, College of Ecology, Lanzhou University, Lanzhou, China

³The Key Laboratory of Tree Genetics and Biotechnology of Jiangsu Province and Education, Department of China, Nanjing Forestry University, Nanjing, China

⁴Department of Plant Physiology, Umeå Plant Science Center, Integrated Science Lab, Umeå University, Umeå, Sweden

⁵Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao, China

⁶Trine University, Angola, IN, USA

Introduction

Willows (*Salix*) are dioecious, catkin-bearing woody plants, together with poplars (*Populus*), which comprise the *Salicaceae* family. Willows consist of more than 300 species, including shrubs and trees, which are widely distributed in diverse ecological niches, particularly in temperate and arctic zones. Asia is considered the centre of origin for *Salix*, with China alone harbouring around 270 species [1, 2]. The physiological adaptations and ecological resilience of willows make them ideal candidates for ecological conservation and restoration projects across various climatic zones [3, 4]. Moreover, willows have a significant commercial impact, especially in wood production. Their rapid growth and short generation period



make them an ideal renewable biological resource for biomass production, with additional economic value in manufacturing, medicinal use, and ornamental cultivation [3–6]. Although willows have a long breeding history, genetic improvement has been limited by a lack of genomic resources and a poor understanding of the genetic basis of growth traits [7–12]. Recently, genomics-informed breeding strategies have been widely applied to a variety of crops [13–16]. Adopting this approach in willow breeding could facilitate rapid genetic improvement of key growth traits.

Salix suchowensis is a diploid ($2n=12x=38$) species with the first high-quality reference genome in this genus and around 30,000 protein-coding genes were annotated in the assembled 356 Mb genome [7, 11]. This species grows naturally in rivers and stream-sides of northern China. Growth traits are key agronomic characteristics that directly affect biomass production and are primary targets for genetic improvement [17]. This makes it an ideal model for functional genomics and genetic studies in willows. Previous studies have identified quantitative trait loci (QTLs) underlying growth traits in *S. suchowensis* using amplified fragment length polymorphism markers and genotype by sequencing [18]. However, there is an urgent need for a high-throughput, cost-effective, time-saving, and accurate genotyping method for large-scale population research in *S. suchowensis*. Targeting by sequencing is a promising method to provide high-resolution single-nucleotide polymorphisms (SNPs), which can be used for genome-wide association studies (GWAS), genomic selection (GS), and population genetic studies [19–23]. The up-to-date technology is genotyping by target sequencing (GBTS), which enhances flexibility and scalability since the designed probes can be easily updated and expanded [24]. The GBTS SNP panel has been successfully applied in various species, including rice, maize, and soybean, proving its efficiency in genetic research [24, 25]. Thus, the development of a GBTS SNP panel for *S. suchowensis* would be a valuable resource for genetic research and breeding programs.

In this research, we designed a 40k GBTS SNP panel for *S. suchowensis* based on whole-genome resequencing data. This SNP array panel was designed to capture the genetic diversity for genetic research and breeding applications. We validated this SNP panel by genotyping 550 new samples and conducting GWAS to identify candidate genes associated with growth traits. Finally, we assessed the prediction accuracy of the SNP panel for plant height and ground diameter in *S. suchowensis*.

Materials and methods

Selection of 40k SNPs from WGS

A total of 4,531,211 SNPs obtained from whole-genome resequencing in our previous study were used. All SNPs

are bi-allelic variants with a call rate above 90% and a minor allele frequency (MAF) greater than 0.05. In 2023, 600 rooted cuttings from 8 full-sib families were planted in Leshan (LS; N29.55°, E103.77°), Pengzhou (PZ; N30.99°, E103.96°), and Yibin (YB; N28.75°, E104.64°). Phenotypic data for plant height (PH) and ground diameter (GD) were collected from all three locations, and transcriptome data were gathered in LS to perform GWAS and expression quantitative trait loci (eQTL) mapping. In total, 25 SNPs associated with growth traits and 938 SNPs associated with gene expression were prioritised in the SNP panel design. All genetic data, phenotype data, and research results from our previous study were reused in this study [26].

Moreover, we categorised these SNPs based on their relative genomic position with gene annotation. Considering the potential significant influence of SNPs located at the start and stop codons, such SNPs were included in the SNP panel design (239 at the start codon and 82 at the stop codon).

Then, 3,235,172 SNPs located within gene regions (including gene body, 3 kb flanking region of upstream and downstream) were further filtered to ensure at least one SNP per gene, thereby maximising the coverage of regions with annotated genes. For each gene, SNPs inside the gene region were used to calculate pairwise R^2 values. After that, SNPs with the top 10 largest number of tagged SNPs ($R^2 > 0.6$) were retained. The threshold we selected under the consideration that these SNPs could best represent the genetic variations inside corresponding genes. Next, the SNP with the highest MAF was chosen for each gene, for its high occurrence rate in this population.

Furthermore, for the 1,296,039 SNPs located outside the gene region, linkage disequilibrium (LD) filtering was performed to obtain representative SNPs. Using PLINK with a 200 kb window, a 10 kb stepsize, and an R^2 threshold of > 0.1 , we identified 124,192 SNPs. These SNPs were subsequently down-sampled to 12,419 SNPs [27].

Based on the preliminary selection of SNPs, we designed the liquid-phase probe sequences located within the upstream and downstream regions of the SNP loci. The probe sequences were designed using Novogene's SNP array algorithm based on thermodynamic stability principles to ensure high capture efficiency. The probe length was set between 80–120 bp with an average of approximately 100 bp, while maintaining GC content within 20%–70%. The number of SNPs in probe capture regions was restricted to fewer than 5. The SNPs of the panel that do not meet the standards will be filtered. This step ensured high capture efficiency of this SNP panel. Finally, 39,076 SNPs—nearly 40k—were retained as the final SNP panel.

Plant materials

After designing the 40k SNP panel, we randomly selected two samples from each of the previously researched samples, resulting in a total of 16 samples, to validate the quality of the SNP panel. These samples were collected from an experimental field in 2023, which contained 600 samples. The youngest, most recent leaves from the tips of the main stems were collected and immediately frozen in liquid nitrogen for DNA extraction.

In this study, 550 additional samples from 8 new full-sib families (subfamily 1: 95, subfamily 2: 66, subfamily 3: 103, subfamily 4: 61, subfamily 5: 59, subfamily 6: 56, subfamily 7: 75, subfamily 8: 35) of *S. suchowensis* were selected to perform GBTS and genetic investigation. The rooted cuttings with the same growth were planted at two locations, Yibin (YB; N28.75°, E104.64°) and Pengzhou (PZ; N30.99°, E103.96°), in Sichuan province, China, in 2024. The field soil of each location was smoothed to be equal for all soil compositions. After being planted, we immediately watered all cuttings evenly to ensure their establishment. During the growth period, weeds were removed manually, and no additional fertilization was applied. Fifteen days after planting, two to three fresh leaves were collected from each sample to serve as material for DNA extraction. The collected leaves were immediately frozen in liquid nitrogen to preserve DNA integrity. Phenotypic data collection began 30 days after planting, at a point when plants exhibit measurable differences in growth traits. At this stage, we take measurements of plant height (PH) and ground diameter (GD) at 15-day intervals at both YB and PZ locations until the end of the growing season (a total of five measuring point). PH was measured as the vertical distance from the soil surface to the tip of the main stem (apical meristem). GD was determined as the diameter of the stem at approximately 5 cm above the soil surface. These two traits measured the vertical and horizontal growth of the plant, capturing major traits related to biomass accumulation. The phenotypic value used for analysis was the growth measurement at each point.

Validation of the 40k SNP panel

All samples were sequenced using GBTS, with an average of 2 Gb raw reads per sample. The raw data were processed using Fastp to filter out low-quality reads and adapters [28]. Clean data were then aligned to the reference genome of *S. suchowensis* using BWA-MEM [29]. The aligned data were processed with SAMBAMBA to sort and remove duplicate reads. Genome coverage was calculated using SAMtools [30, 31]. Genetic variants were detected using BCFtools [32]. The genotyping data were filtered with the following parameters: QUAL > 30, F_MISSING < 0.5, DP > 2, and "-m2 -M2". SNPs located at the designed positions were extracted for further

analysis, including the calculation of the sample call rate and site call rate. The sample call rate is defined as the average SNP call rate across all SNPs for a given sample, while the site call rate refers to the SNP call rate across all samples. For further analysis, SNPs with MAF < 0.05 were excluded using PLINK [27].

Principal component analysis

The population structure of the 600 *S. suchowensis* individuals was evaluated using principal component analysis (PCA) based on the whole-genomic SNPs and the 40k selected SNPs. PCA was performed using the GCTA software [33]. For whole-genomic SNP data, we select all SNPs from the previous study as input to perform PCA analysis. For the 40k selected SNPs data, we extracted these SNPs from whole-genomic SNPs as input to perform PCA analysis. The PCA results were visualised by the plot function of the R language.

Genetic analysis

The kinship heritability for each trait was estimated by fitting a mixed linear model using GCTA software [33].

$$y = \mu + Zu + e \quad (1)$$

Here, y is the phenotype value, μ is the intercept of the model. Z is the corresponding design matrix that satisfies $ZZ^T/m = G$, where G is the kinship matrix, m is the SNP number. Therefore, the distribution of u is $N(0, I\sigma_u^2)$. e is the residual with $e \sim N(0, I\sigma_e^2)$. The kinship heritability was calculated as $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$. For comparison, the genomic kinship matrix G was calculated based on the whole-genome SNPs and the 40k selected SNPs separately. Phenotypic data were collected from previous study. After planting 550 new samples in YB and PZ, PH and GD were collected to perform a genome-wide association study (GWAS). GWAS was performed using GEMMA with a mixed linear model [34]. A Bonferroni-corrected genome-wide significance threshold (0.05 divided by the number of SNPs) was used for genome-wide significance. The significantly associated SNPs were annotated using the *S. suchowensis* genome annotation file. Genes within a 10 kb flanking region upstream and downstream of the associated SNPs were extracted for gene ontology (GO) enrichment analysis using the clusterProfiler package in R [35].

Genomic selection

To assess the quality of this SNP panel in GS, the Genomic Best Linear Unbiased Prediction (GBLUP) method was employed. This method, described in model 1, integrates additive genetic effects into a mixed linear

model framework, enabling the estimation of genomic estimated breeding value (GEBV) based on SNP-derived relationships. This model was implemented in the hglm package in R [36]. The G matrix was calculated based on the 40k selected SNPs. To evaluate the accuracy of genomic selection, five-fold cross-validation was used to split the dataset into training and testing sets. For each fold, GEBVs were calculated for individuals in the testing set based on the G matrix and model parameters derived from the corresponding training set. The correlation

between the calculated and predicted breeding values across the five test sets was used to assess the accuracy of genomic selection.

Results

Probe design of *S. suchowensis*

A total of 4,531,211 high-quality SNPs from 600 *S. suchowensis* F1 individuals were obtained for probe design. Four selections were applied to select the SNP panel (Fig. 1): (1) SNPs associated with growth traits

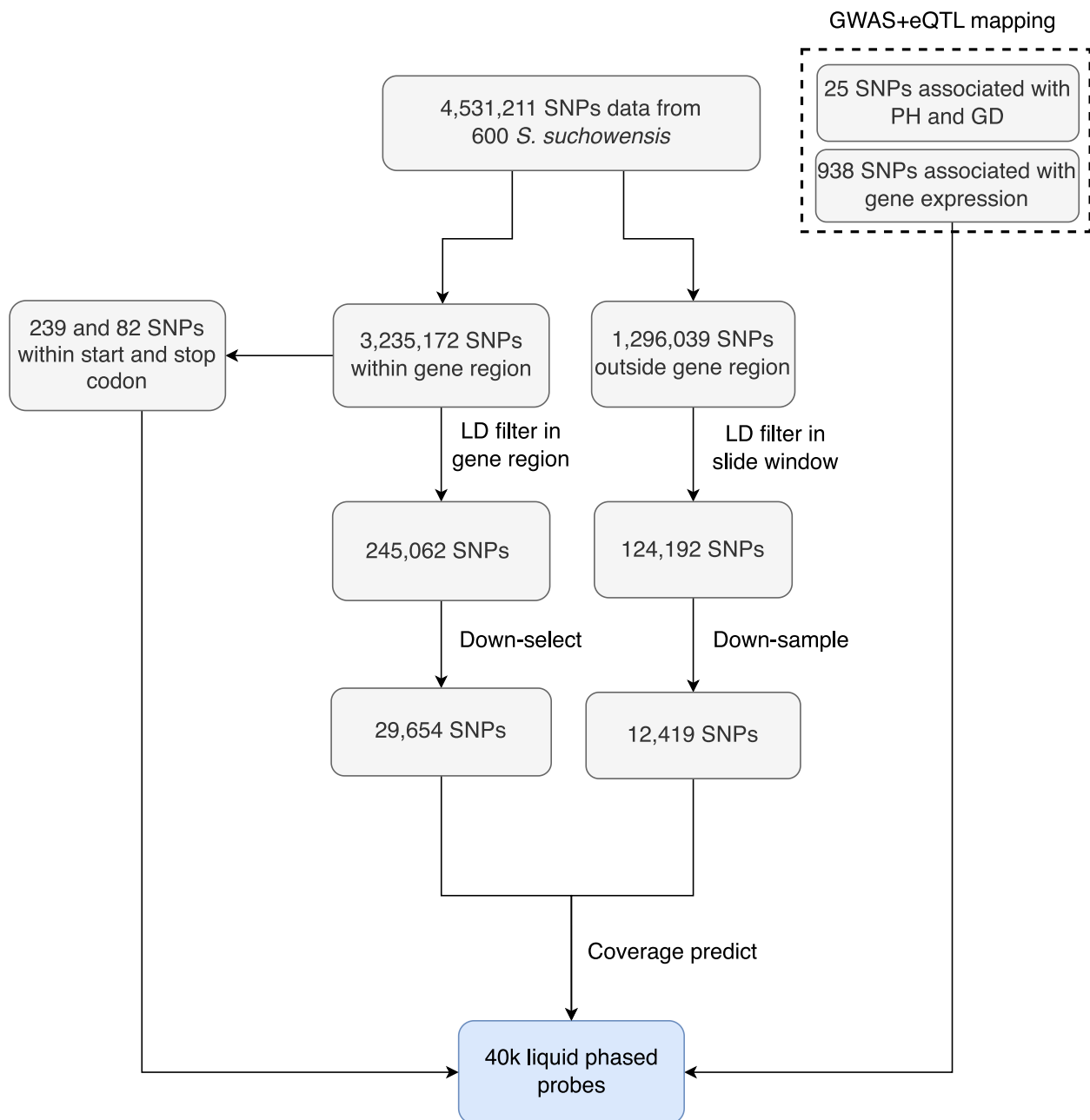


Fig. 1 Schematic representation of the probe design workflow. Each box denotes the number of SNPs after each selection step. The abbreviated selection steps can be found in the materials and methods

(25 SNPs) and molecular traits (938 SNPs) identified through GWAS and eQTL mapping were prioritised for their functional relevance. (2) SNPs located in the start and stop codons (239 and 82 SNPs, respectively) were included due to their potential role in gene regulation. (3) At least one SNP per gene within the gene body were selected to ensure maximum gene space coverage. (4) SNPs were evenly distributed in intergenic regions to account for the even genome distribution of SNP panel. Finally, after filtering SNPs by their phychemical properties (see materials and methods), we retained 39,076 probes for the final SNP panel (Supplementary Table 1), which were evenly distributed across the genome (Fig. 2A). The number of SNPs per chromosome is highly correlated with chromosome length (Pearson correlation coefficient = 0.96, p -value = 1.07×10^{-10}), with the highest number of SNPs on chromosome 1 (4,401 SNPs) and the

lowest number on chromosome 9 (938 SNPs) (Fig. 2B). With MAF higher than 0.05 as the selection criterion, the median MAF of the 40k SNP panel was 0.43 (Fig. 2C), which ensured the selected SNP panel SNPs were widespread in the population. These SNPs are located in different genomic regions, including the gene body and intergenic region (Fig. 2D). Following our selection criteria, most of the selected SNPs are distributed within the gene structure, with the most SNPs located in the exon region (11,728 SNPs).

40k SNPs captured the population structure of *S. suchowensis*

To evaluate whether the 40k selected SNPs could capture genetic relatedness and diversity of the *S. suchowensis* population, we compared the population structure of the 600 *S. suchowensis* captured by whole-genomic SNPs and

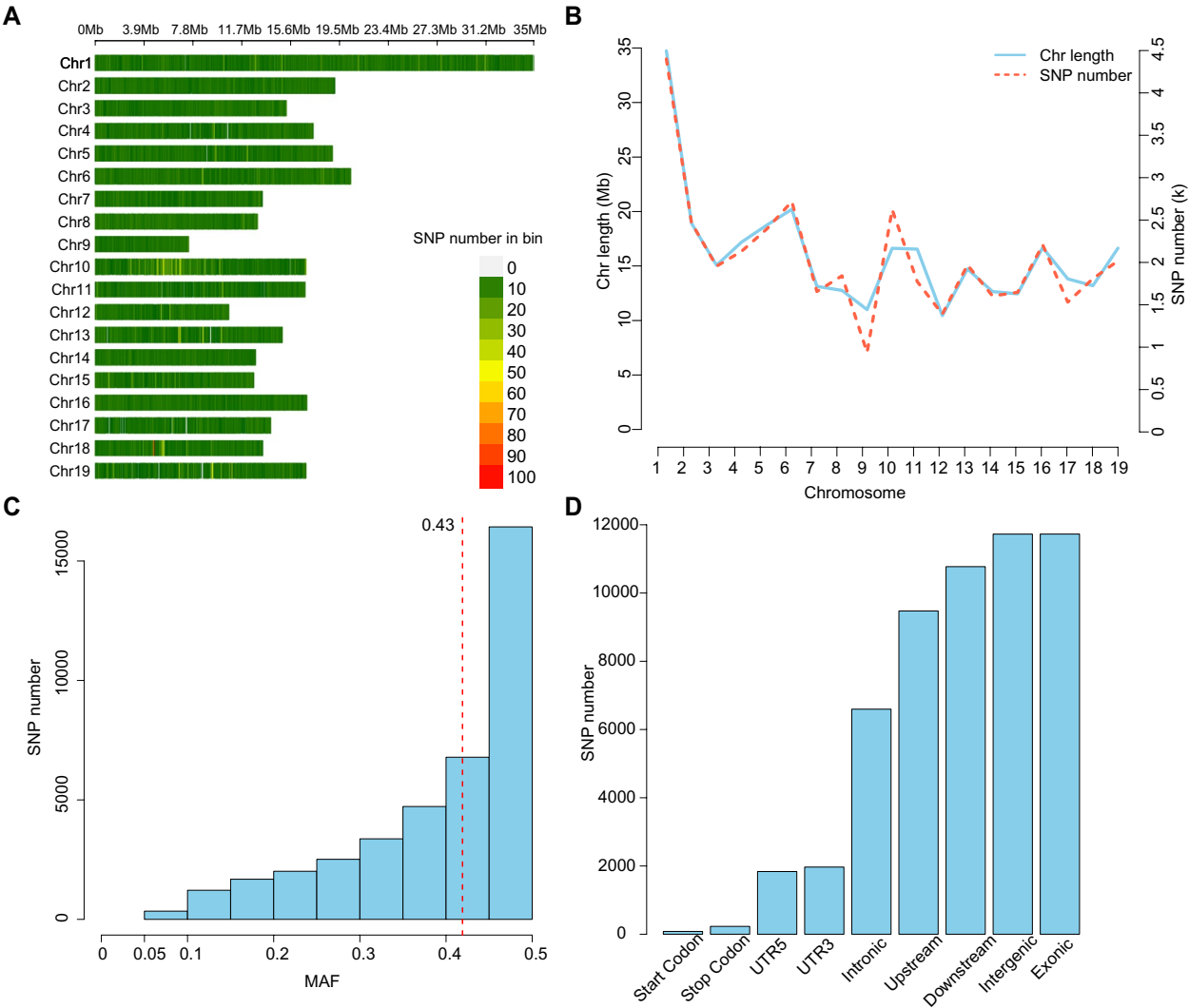


Fig. 2 Characteristics of the 40k probes. **A** The distribution of SNPs by chromosome. Each bar represents a 50 kb genome bin, colored by the number of SNPs. **B** The plot of chromosome length (colored in blue) and the number of SNPs per chromosome (colored in red). **C** MAF of the SNP panel. **D** Distribution of SNPs with different genomic features

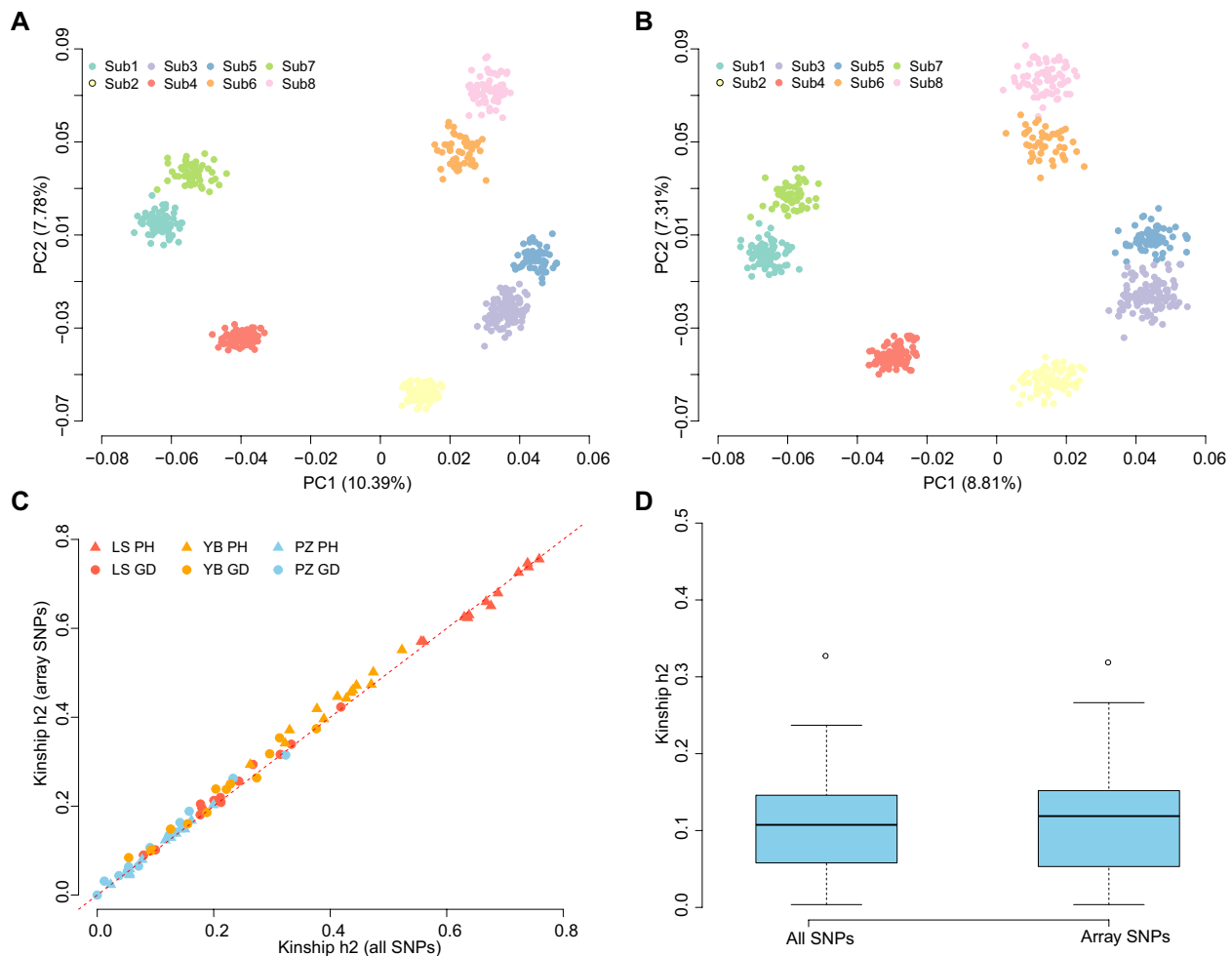


Fig. 3 The 40k SNP panel captured the population structure of *S. suchowensis*. **A** and **B** PCA plot of the 600 *S. suchowensis* individuals based on the whole-genomic SNPs and the 40k selected SNPs. Each color represents a family. **C** The estimated kinship heritability of growth traits was based on the whole-genomic SNPs and 40k selected SNPs. Different colors and point shapes were used to indicate the source of the phenotype data from the previous study. For example, LS PH means the plant height phenotype measured in Leshan city. The red line represents the diagonal line. **D** The boxplot of the heritability of all traits in Pengzhou

the 40k selected SNPs. PCA analysis indicated a similar population structure in each family (Fig. 3A, B). In the PCA plots, clear separations along the first and second principal components reveal distinct genetic signals for each family, with clusters displaying consistent internal distribution and minimal overlap. Consistent with that, the 600 individuals from eight families, individuals can be clearly separated into eight groups in the PCA plot, both by the whole-genomic SNPs and 40k selected SNPs. This suggests that the selected SNPs are highly efficient in capturing genomic relationships among the studied individuals.

Furthermore, we compared the estimated kinship heritability of measured growth traits (PH and GD at YB, PZ, and LS) to assess the quality of selected SNPs in genetic research. This comparative analysis was performed across multiple environments and traits, ensuring a comprehensive evaluation under a wide range of scenarios.

Similarly, the estimated heritability of growth traits from the whole-genome sequenced SNPs and 40k selected SNPs was highly correlated (Pearson correlation coefficient = 0.99, p -value = 2.82×10^{-93} , Fig. 3C). For instance, at Pengzhou, the difference in heritability median between whole-genome sequenced SNPs and 40k selected SNPs was just 5.68×10^{-3} . Such subtle differences highlighted the high quality of the 40k SNP panel in capturing the genetic variance of studied traits, making it an ideal approach for future genetic research and breeding applications.

Validation of the probe efficiency

Here, we selected two individuals from each family to validate the capture ability of synthetic DNA probes. As illustrated in Fig. 4A, the SNP panel showed a high call rate across all selected samples, with a minimal call rate of 96.95%, and the combined SNP call rate reached 99.91%.

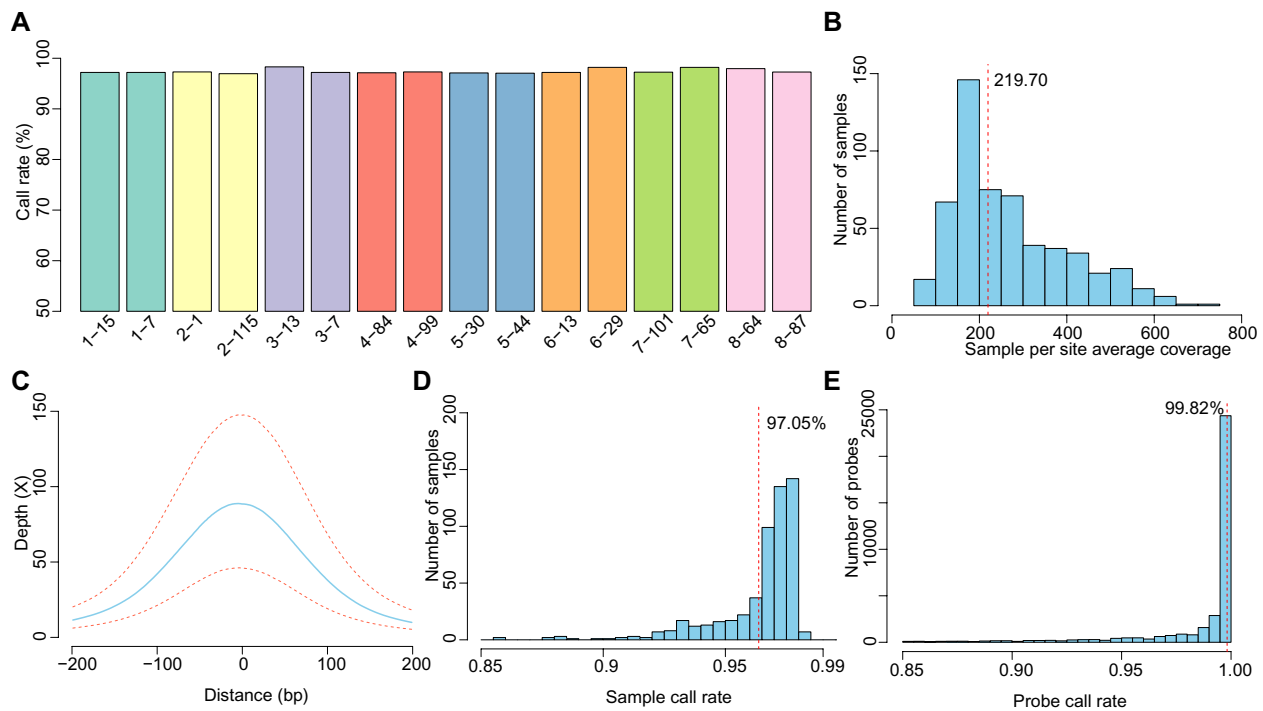


Fig. 4 Validation of the probe efficiency. **A** The call rate of the 40k SNP panel in 16 selected samples. The color of each bar represents the family, which is consistent with Fig. 3 **(B)** The coverage of the 40k SNP panel in 550 new samples. The red line indicates the median value. **C** The distribution of median sequence depth of the 40k SNP panel in 550 new samples across 400bp base (blue). Distance 0 represents the selected snp position, upstream was shown as minus, and downstream was shown as positive number. The red line above the blue line indicates the top 25% depth of each base, and the red line below the blue line indicates the bottom 25% depth of each base. **D** The sample call rate of the 40k SNP panel in 550 new samples. The red line indicates the median value. **E** The site call rate of each 40k SNP panel. The red line indicates the median value

Then we applied this SNP panel in a breeding population with 550 samples from eight additional families by performing GBTS. With ~2 Gb raw sequencing reads per individual, 50% of samples showed an average coverage of 219.70X at targeted positions (Fig. 4B). The high sequencing, with minimal fluctuations in coverage across targeted probe regions, ensured robust SNP calling for downstream analysis. For each probe, the sequence depth showed a normal distribution, with a median depth of 88.49X, and gradually decreased towards flanking region (Fig. 4C). These results demonstrated that the designed probes could bind DNA with high accuracy and specificity. After SNP calling, a high call rate (median 97.05%) across all samples was achieved (Fig. 4D). Moreover, for designed probes, 50% of probes showed a high call rate across all sites (median 99.82%) (Fig. 4E). Altogether, these results indicated a high efficiency of the designed SNP panel.

Applying the SNP panel in GWAS to identify candidate genes for growth traits

To explore the implementation ability of this SNP panel in dissecting the genetic architecture of growth traits, we measured the phenotype of PH and GD on 550 samples at 15-day intervals in YB and PZ from one month after

planting (Supplementary Table 2), and performed GWAS analysis. A total of 72 SNPs were significantly associated with growth traits (Fig. 5A-D, Supplementary Table 3). There were 63 SNPs distributed inside the body of 50 genes, and nine SNPs were located at intergenic regions (Fig. 5E). We noticed that there were three SNPs located in the stop codon, which may introduce gene expression change and further influence the growth traits. These results highlighted the significance of the SNP panel in dissecting the genetic basis of growth traits. For instance, SNP 4:16980574 (G → A) is located in the stop codon of *IMY05_004G0170200*. *IMY05_004G0170200* is annotated as a vacuolar-sorting receptor, which recognises specific protein sorting signals. Such variation may affect plant growth by mediating the sorting and trafficking of proteins to vacuoles [37]. Within 10 kb flanking region of all associated SNPs, 169 genes were detected. GO enrichment analysis indicated that these genes were enriched in the biological process of cell division, cell development, and growth (Fig. 5F), suggesting they were promising candidates for future validation.

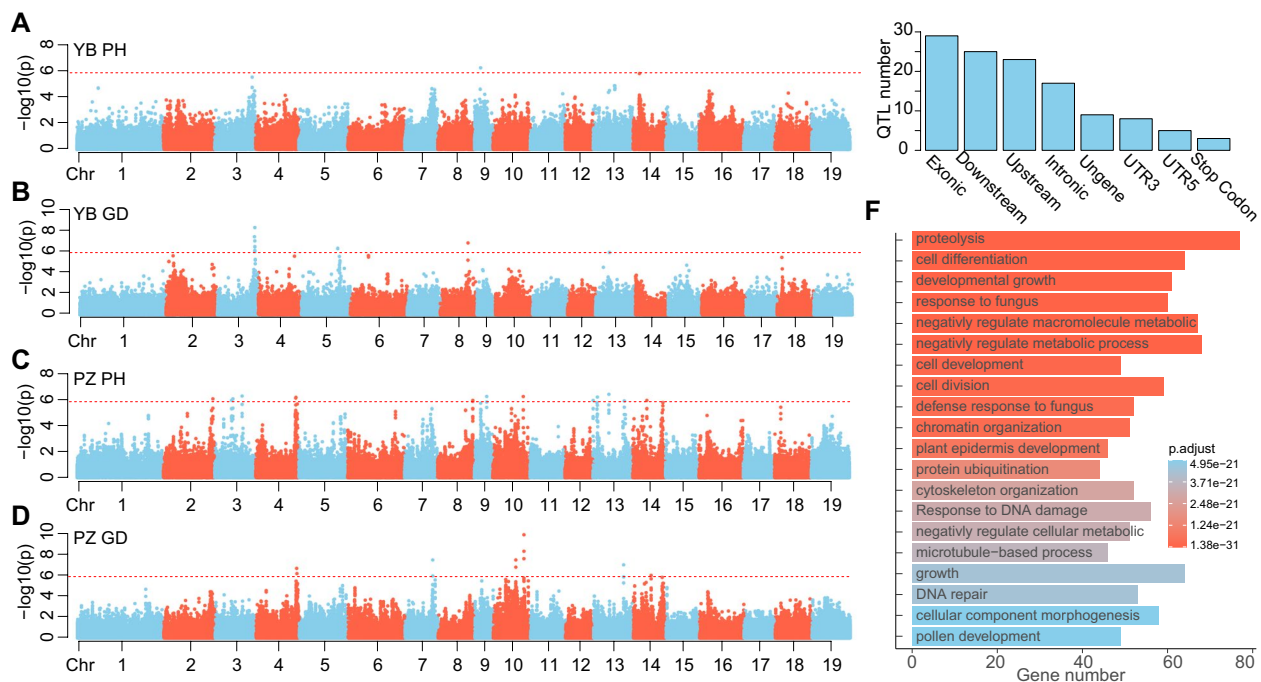


Fig. 5 GWAS of growth traits identified candidate genes. **A–D** The overlapping Manhattan plot of GWAS for the five measured plant heights (PH) and ground diameter (GD) in Yibin (YB) and Pengzhou (PZ). The red line represents the Bonferroni corrected genome-wide significance threshold. **E** The distribution of significantly associated SNPs in gene and intergenic regions. **F** GO enrichment analysis of genes within 10 kb flanking region of significantly associated SNPs

The designed SNP panel enables accurate genomic selection in *S. suchowensis*

GS is an efficient approach in rapid genetic improvement of livestock, crop and forest trees. It could significantly improve breeding efficiency by selecting the elite individual without measuring phenotype. After using five-fold cross-validation to split the training and testing sets, the Pearson correlation of the calculated and predicted breeding value in five test sets was visualised. As shown in Fig. 6, all the traits measured at different plant locations and times showed a high prediction accuracy, with a median of 0.90.

Discussion

The design and performance of the SNP panel

Common strategy for SNP panel design typically focuses on evenly distributing SNPs across the genome, aiming at maximizing the genome coverage [19]. In this study, we adopted a strategy that prioritizes the biological function of SNPs, while maintaining even distribution across the genome. As described above, these selection criteria consider candidate SNPs that may influence growth traits and gene expression, SNPs that may have an effect on gene translation, the most representative SNPs located within gene body, and SNPs that fill gaps of the selected SNPs in terms of genome coverage. Ultimately, 39,076 SNPs were selected. The consistent population structure and high kinship heritability captured by both

whole-genome SNPs and the SNP panel indicate a high quality of the designed SNP panel. The SNP panel's performance was further validated by genotyping 550 samples of *S. suchowensis*, yielding a high call rate across all samples (median 97.05%) and all sites (median 99.82%). These high call rates demonstrate the efficiency of the SNP panel in genotyping. However, the SNP panel and probe sequence that were designed for GBTS may not be suitable for other sequencing platforms. Despite this, when compared to solid-phase gene chips, the GBTS SNP panel enhances the flexibility of probe design and provides more comprehensive information [24, 38], making it a valuable tool for genotyping applications. For instance, additional SNPs can be easily incorporated into probe designs for further population studies. Moreover, the target sequencing can capture a wide range of genomic variants in the probe region, offering many kinds of molecular marker data, such as SNPs, SSRs, and InDels, and providing a much larger number of markers than that obtained from multiplex PCR [24].

Application of SNP panel in genetic research and breeding

As genetic research in willow has not progressed as rapidly as in other crops, and the first high-quality reference genome for the model species *S. suchowensis* was only published five years ago, the molecular mechanisms and genetic architecture of *S. suchowensis* remain unclear. The 40k SNP panel designed in this study provides a valuable

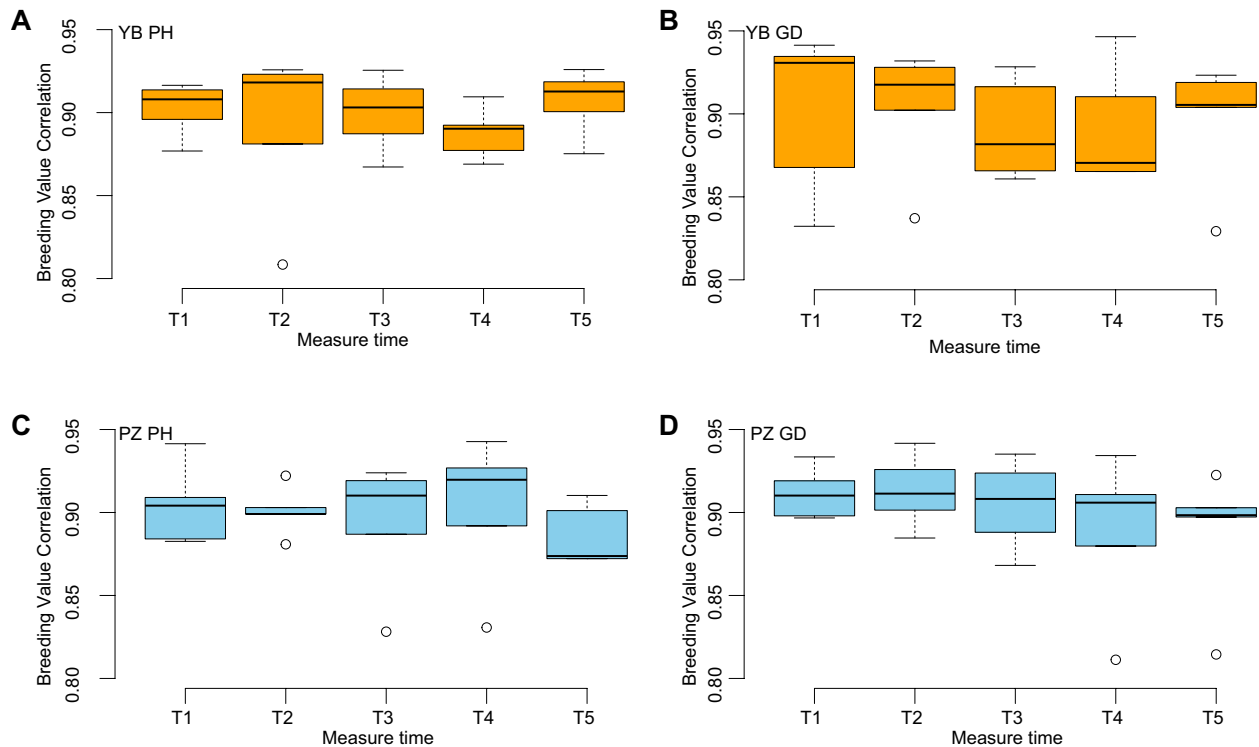


Fig. 6 The accuracy of genomic selection in *S. suchowensis*. **A–D** The boxplot of correlation of calculated and predicted breeding value in five test sets for plant height (PH) and ground diameter (GD) in Yibin (YB) and Pengzhou (PZ). Each measurement point from the first to the fifth is represented as T1 to T5, respectively

resource for advancing genetic research in this species. The SNP panel was successfully applied in GWAS to identify QTL associated with growth traits. A total of 72 SNPs significantly associated with growth traits were identified, and genes around these SNPs were enriched in biological processes such as cell division, cell development, and growth. These findings lay the foundation for unravelling the genetic architecture of growth traits in *S. suchowensis*. In future research, in addition to experimentally verifying these 72 SNPs to confirm their biological functions, it is also necessary to introduce multiple environmental factors to discover the effect of the environment on the phenotypic variation and consider using continuously updated and iterated GWAS models and software to obtain more accurate results.

The Omnigenic model is proposed for the genetic mechanism of complex traits, which describes that complex traits are not only directly affected by a few core genes from the core pathways, but also indirectly regulated by a large number of peripheral genes [39, 40]. Core genes are usually located at the central nodes of biological networks and play a key regulatory role in phenotypes, while peripheral genes regulate the topological structure of the network by influencing or being affected by core genes. These genes collectively shape the variation of phenotype. In the GWAS analysis results, the genetic structure of the ground diameter and plant height traits

embody this model in *S. suchowensis*. Although some genes near SNP loci are functionally candidate genes for regulating growth and development, there are still some SNP loci significantly related to traits distributed in inter-genic regions or other functional genes. These core genes may have regulatory relationships with a large number of peripheral genes, influencing the core growth pathway through a complex gene network. However, the limited sample size and trait phenotype in this study may not provide sufficient statistical power to detect the minor yet important associations.

In future research, implementing this SNP panel could enable large-scale SNP genotyping at a lower cost, facilitating the genetic dissection of growth traits. The high accuracy of genomic selection noted in this study, with a median correlation of 0.90, further supports the feasibility of using this SNP panel for GS, underscoring its potential in future breeding applications. Although GS is still in its early stages for *S. suchowensis*, it has emerged as a key technology for breeding improvement [41–43], addressing the challenges of long breeding cycles. With ongoing updates to SNP panels, the accumulation of genotype–phenotype data, and enhanced plant environments, prediction models can be further refined [44, 45]. This progress is expected to improve GS accuracy and accelerate the genetic enhancement of *S. suchowensis*.

Conclusion

We designed a 40k SNP panel for *S. suchowensis* by including trait-associated SNPs, gene body SNPs, intra-genic SNPs, and intergenic SNPs, which are evenly distributed across the genome. We performed a comprehensive evaluation analysis and revealed its high quality in capturing population structure. The designed SNP panel was successfully applied in GWAS to identify candidate genes for growth traits. Furthermore, the SNP panel enabled accurate genomic selection, demonstrating its potential for future genetic research and breeding applications. Overall, this SNP panel provides a valuable resource for *S. suchowensis*.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-026-12690-2>.

Supplementary Material 1.

Acknowledgements

We are grateful to the members of the Key Laboratory of Tree Genetics and Biotechnology of Jiangsu Province of Nanjing Forestry University for their help.

Authors' contributions

Conceptualization, Y.H., S.G.; methodology, Y.H., S.G., Y.Z., Y.J.; formal analysis, Y.H., W.L.; investigation, Y.H., S.G., W.L., M.Z., L.F., X.G.; data collection and curation, Y.H., M.Z., W.L., L.F.; manuscript writing, Y.H., S.G., T.Y., Y.J., J.L.; manuscript revising, Y.H., S.G., R.H.; supervision, Y.J., J.L.

Funding

This research was funded by the National Key Research and Development Program of China (2021YFD2200202).

Data availability

The genotype files generated by GBTS SNP array can be accessed through the <https://doi.org/10.6084/m9.figshare.29365658>. The raw sequencing data are available in the China National Center for Bioinformation PRJCA041895.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 June 2025 / Accepted: 19 February 2026

Published online: 02 March 2026

References

- Newsholme C. Willows: the genus *Salix*. Timber Press; 2003.
- Argus GW. Infrageneric classification of *Salix* (Salicaceae) in the New World. Syst Bot Monogr. 1997;52:1–121. <https://doi.org/10.2307/25096638>.
- Kuzovkina YA, Quigley MF. Willows beyond wetlands: uses of *salix* L. species for environmental projects. Water Air Soil Pollut. 2005;162:183–204. <https://doi.org/10.1007/s11270-005-6272-5>.
- Elowson S. Willow as a vegetation filter for cleaning of polluted drainage water from agricultural land. Biomass Bioenergy. 1999;16:281–90.
- Desborough MJR, Keeling DM. The aspirin story – from willow to wonder drug. Br J Haematol. 2017;177:674–83. <https://doi.org/10.1111/bjh.14520>.
- Christersson L, Sennnerby-Forsse L, Zsuffa L. The role and significance of woody biomass plantations in Swedish agriculture. For Chron. 1993;69:687–93.
- Wei S, Yang Y, Yin T. The chromosome-scale assembly of the willow genome provides insight into Salicaceae genome evolution. Hortic Res. 2020;7:45. <https://doi.org/10.1038/s41438-020-0268-6>.
- Zhang J, Yuan H, Li Y, Chen Y, Liu G, Ye M, et al. Genome sequencing and phylogenetic analysis of allotetraploid *Salix matsudana* Koidz. Hortic Res. 2020;7:1. <https://doi.org/10.1038/s41438-020-00424-8>.
- He L, Jia KH, Zhang RG, Wang Y, Shi TL, Li ZC, et al. Chromosome-scale assembly of the genome of *Salix dunnii* reveals a male-heterogametic sex determination system on chromosome 7. Mol Ecol Resour. 2021;21. <https://doi.org/10.1111/1755-0998.13362>.
- Hu N, Sanderson BJ, Guo M, Feng G, Gambhir D, Hale H, et al. Evolution of a ZW sex chromosome system in willows. Nat Commun. 2023;14:7144. <https://doi.org/10.1038/s41467-023-42880-5>.
- Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, et al. The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res. 2014;24:1274–7.
- Hou J, Ye N, Zhang D, Chen Y, Fang L, Dai X, et al. Different autosomes evolved into sex chromosomes in the sister genera of *Salix* and *Populus*. Sci Rep. 2015;5. <https://doi.org/10.1038/srep09076>.
- Varshney RK, Bohra A, Yu J, Graner A, Zhang Q, Sorrells ME. Designing future crops: genomics-assisted breeding comes of age. Trends Plant Sci. 2021;26:631–49.
- Wang N, Li H, Huang S. Rational redomestication for future agriculture. Annu Rev Plant Biol. 2025;76:637–62.
- Alemu A, Åstrand J, Montesinos-López OA, y Sánchez JL, Fernández-González J, Tadesse W, et al. Genomic selection in plant breeding: key factors shaping two decades of progress. Mol Plant. 2024;17:552–78.
- Basu U, Parida SK. Restructuring plant types for developing tailor-made crops. Plant Biotechnol J. 2021;21:1106–22.
- He L, Liao S, Applequist WL, Chen S-P. The valid publication of *Salix suchowensis* (Salicaceae). PhytoKeys. 2019;131:27–35.
- Chen L, Liu L, Yang G, Li X, Dai X, Xue L, et al. Expression quantitative trait locus of wood formation-related genes in *Salix suchowensis*. Int J Mol Sci. 2023;25:247. <https://doi.org/10.3390/ijms25010247>.
- Thomson MJ. High-throughput SNP genotyping to accelerate crop improvement. Plant Breed Biotechnol. 2014;2:195–212.
- Balagué-Dobón L, Cáceres A, González JR. Fully exploiting SNP arrays: a systematic review on the tools to extract underlying genomic structure. Brief Bioinform. 2022;23:bbac043.
- Kim K-W, Nawade B, Nam J, Chu S-H, Ha J, Park Y-J. Development of an inclusive 580K SNP array and its application for genomic selection and genome-wide association studies in rice. Front Plant Sci. 2022;13:1036177.
- Freudl R. Signal peptides for recombinant protein secretion in bacterial expression systems. Microb Cell Factories. 2018;17:1–10. <https://doi.org/10.1186/s12934-018-0901-3>.
- Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault AM, et al. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. Syst Biol. 2018;68:594–606.
- Guo Z, Yang Q, Huang F, Zheng H, Sang Z, Xu Y, et al. Development of high-resolution multiple-SNP arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip. Plant Commun. 2021;2:100230.
- Lee CY, Cheon K-S, Shin Y, Oh H, Jeong Y-M, Jang H, et al. Development and application of a target capture sequencing SNP-genotyping platform in rice. Genes. 2022;13:794.
- Liu W, Gu S, Zhu M, Han Y, Dai X, Wu H, et al. Genome and transcriptome analysis provide insights into the genetics and breeding of *Salix suchowensis* growth traits. New For. 2025;56:1–21.
- Chen ZL, Meng JM, Cao Y, Yin JL, Fang RQ, Fan SB, et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. Nat Commun. 2019;10:3404. <https://doi.org/10.1038/s41467-019-11337-z>.
- Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023;2:e107.

29. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv.1303.3997v2. 2013. <https://doi.org/10.48550/arXiv.1303.3997>.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
31. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015;31:2032–4.
32. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2020;10:giab008.
33. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88:76–82. <https://doi.org/10.1016/j.ajhg.2010.11.011>.
34. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 2012;44:821–4. <https://doi.org/10.1038/ng.2310>.
35. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb)*. 2021;2:100141.
36. Alam M, Rönneberg L, Shen X. Fitting conditional and simultaneous autoregressive spatial models in hglm. *R J*. 2015;7:5.
37. Masclaux FG, Galaud JP, Pont-Lezica R. The riddle of the plant vacuolar sorting receptors. *Protoplasma*. 2005;226:103–8. <https://doi.org/10.1007/s00709-005-0117-3>.
38. Dalma-Weiszhausz DD, Warrington JA, Tanimoto EY, Miyada CG. The affymetrix GeneChip platform: an overview. *Methods Enzymol*. 2006;410:3–28.
39. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169:1177–86. <https://doi.org/10.1016/j.cell.2017.05.038>.
40. Mathieson I. The omnigenic model and polygenic prediction of complex traits. *Am J Hum Genet*. 2021;108:1558–63. <https://doi.org/10.1016/j.ajhg.2021.07.003>.
41. Resende MDV, Resende MFR, Sansaloni CP, Petrolini CD, Missiaggia AA, Aguiar AM, et al. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol*. 2012;194:116–28. <https://doi.org/10.1111/j.1469-8137.2011.04038.x>.
42. Resende MFR, Muñoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM, et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics*. 2012;190:1503–10. <https://doi.org/10.1534/genetics.111.137026>.
43. Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, et al. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol*. 2012;193:617–24. <https://doi.org/10.1111/j.1469-8137.2011.03895.x>.
44. Lenz PRN, Nadeau S, Mottet M, Perron M, Isabel N, Beaulieu J, et al. Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. *Evol Appl*. 2020;13:76–94. <https://doi.org/10.1111/eva.12823>.
45. Grattapaglia D. Status and Perspectives of Genomic Selection in Forest Tree Breeding. In: Varshney RK, Roorkiwal M, Sorrells ME, editors. *Genomic Selection for Crop Improvement: New Molecular Breeding Strategies for Crop Improvement*. Cham: Springer International Publishing; 2017. p. 199–249. https://doi.org/10.1007/978-3-319-63170-7_9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.