Construction of a SNP fingerprinting of exotic pine germplasm resources in China based on 51K liquid-phased probes

Yadi Wu
Research Institute of Subtropical Forestry, Chinese Academy of Forestry
Shu Diao
Research Institute of Subtropical Forestry, Chinese Academy of Forestry
Xianyin Ding
Research Institute of Subtropical Forestry, Chinese Academy of Forestry
Qinyun Huang
Research Institute of Subtropical Forestry, Chinese Academy of Forestry
Qifu Luan
qifu.luan@caf.ac.cn

Research Institute of Subtropical Forestry, Chinese Academy of Forestry

Research Article

Keywords: DNA fingerprinting, pine variety identification, liquid-phased probes, Pinus elliottii, Pinus taeda

Posted Date: February 12th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3910803/v1

Additional Declarations: No competing interests reported.

Construction of a SNP fingerprinting of exotic pine germplasm resources in China based on 51K liquid-phased probes.

Yadi Wu[1,2], Shu Diao[1], Xianyin Ding[1], Qinyun Huang[1] and Qifu Luan[1,*]

* These authors contributed equally to this work

[1] Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou 311400, China

[2] College of Forestry, Nanjing Forestry University, Nanjing 210037, China

Word Count: 4337 words (main body) with 6 figures and 3 tables

Article submitted to: BMC Plant Biology

Corresponding author: Qifu Luan

Tel.: **+86-13588372689**, **E-mail:** qifu.luan@caf.ac.cn

Abstract

【Background】 Slash pine (*Pinus elliottii*), loblolly pine (*P. taeda*), caribbean pine (*P. caribaea*) and their hybrid pine are the major timber species in China, which have been introduced from North America for 100 years. Distinguishing them is challenging due to the nutritional organs similarity and the combined characteristics inherited from their parent species in hybrid pines. **【Results】** In this study, we aimed at constructing a set of DNA fingerprint of 38 pine varieties including four kinds of pines species for accurately identifying them. The genotypes of 38 pines were captured by 51K liquid-phased probes developed by our team and 5,60,567 SNPs were genotyped by the following next-generation sequencing. A total of 344 core SNPs were obtained through the screening of the minor allele frequency, miss rate, heterozygosity rate and other parameter conditions. Additionally, employing the Random Forest model in conjunction with PCA led to the retention of 28 SNPs demonstrating identification capabilities for the pine varieties. **【Conclusions】** The DNA fingerprint of the 38 pine varieties was successfully established using the set of 28 SNPs, which serve as a valuable reference for identifying the exotic pine varieties, managing germplasm , and conducting genetic diversity analysis.

Key words: DNA fingerprinting, pine variety identification, liquid-phased probes, *Pinus elliottii*, *Pinus taeda*

Abbreviations: PEE: *Pinus elliottii*; PTA: *P. taeda*; PEC: *P. elliottii* × *P. caribaea*; PAC: *P. taeda*× *P. caribaea*; SNP: single nucleotide polymorphism; SSR: simple sequence repeats; AFLP: amplified fragment length polymorphism; RAPD: random amplified polymorphic DNA; RFLP: restriction fragment length polymorphism; MAS: molecular marker-assisted selection; MAF: minor allele

frequency; PIC: polymorphism information content; LD: linked disequilibrium; hdw: Hardy-Weinberg; HeThe expected heterozygosity; Ho: observed heterozygosity.

1.INTRODUCTION

Slash pine (*Pinus elliottii*), loblolly pine (*P. taeda*) and caribbean pine (*P. caribaea*) have been cultivated in China for over a hundred years since their introduction. They are characterized by rapid growth, wide adaptability and high-yield oleoresin, which make them become important timber[1, 2] and resin-producing[3] species in southern region of China. There are some complementary traits between slash pine and caribbean pine as well as loblolly pine and caribbean pine[4]. Their interspecific hybrid offspring , *P. elliottii* × *P. caribaea* and *P. taeda*× *P. caribaea*, have shown superior growth characteristics to their parents in the tropical and subtropical regions of China[5] and have become important afforestation species in the suitable areas. The demand for seedlings of exotic pines, especially for slash pine, is very strong because of the extensive planting area. Slash pine covers an area of 1.5 million hectares, ranking among Chinese 10 largest artificial forest. The seedling systems have been established in nearly 10 provinces, which have selected and bred nearly 100 varieties of afforestation. However, it is tough to differentiate these pine varieties even species only relying on traditional morphological identification as result of their similar nutritional organs, and their hybrid pine phenotypes combined the characteristics of both parents[6]. The growth cycle of forest trees is long, and the efficiency of breeding will be greatly reduced if the breeding is carried out after the seedlings are mature[7]. Indistinguishable species and long-term breeding cycle both make carry out molecular marker-assisted selection (MAS) in pines necessary. The molecular markers developed in recent years can distinguish the genetic characteristics of the species from the level of genetic material

because of their strong stability and high feasibility [8, 9].It has great significance for the identification of forest tree varieties, genetic diversity analysis, and germplasm research by using molecular markers.

DNA fingerprinting is a technical tool that enables to efficiently identify plant and animal lineages. This technique was first proposed by a British scientist Jeffreys[10]. DNA fingerprinting refers to a region of extremely high polymorphism in the human genome, where DNA bands formed after enzymatic cleavage of DNA can distinguish different human bodies, like to a person's fingerprint, with individual uniqueness[9]. The limitations of using phenotypes to identify varieties can be overcome by using DNA fingerprinting and not only the costs of identification can be saved, but also the utilization efficiency of germplasm resources can be significantly improved. The key steps to construct an accurate and stable DNA fingerprinting includes choosing the suitable molecular marker, genotyping, screening out a set of core marker loci combination and establish the DNA fingerprinting. There are several types of molecular markers were widely applied in recent years, including amplified fragment length polymorphism (AFLP) markers[11], random amplified polymorphic DNA (RAPD) markers [12, 13], restriction fragment length polymorphism (RFLP) markers[14], simple sequence repeats (SSR) markers[15, 16] and single nucleotide polymorphism (SNP) markers. Among these markers, SSR and SNP markers are the most efficient. In 2005, the International Union for the Protection of New Varieties of Plants (UPOV) identified SSR and SNP markers as the main methods for constructing DNA fingerprinting databases [17]. In the past research, SSR markers were often used to construct DNA fingerprinting [18, 19]. A single SSR marker has a series of advantages, such as more information content, fewer required loci, good repeatability, and most of them are co-dominant markers [20, 21], so it is widely used in variety identification and analysis. However, SSR markers also have certain limitations, such as small number of loci, complex operation, high cost, and low throughput

[22-24]. In contrast, SNP markers have obvious advantages: (1) SNP is the smallest unit of genetic variation in the genome, which is easier to integrate data and simpler in data statistics and analysis; (2) Low mutation frequency and higher genetic stability; (3) High sequencing throughput and low cost [25, 26], which make it become the most suitable molecular marker for MAS. In recent years, SNP markers have been widely used in the construction of DNA fingerprints of a variety of plants[27, 28], such as rice(*Oryza sativa* L.) [9], corn(*Zea mays* L.) [29], brassica napus(*Brassica napus* L.) [30], camellia camellia(*Camellia oleifera* Abel) [8], tea tree(*Camellia sinensis*) [31] and so on. In forestry trees, SNPs are commonly used for development of molecular markers, whole-genome association analysis and construction of high-density linkage genetic maps[32-34].

In the progress of MAS in pine trees, the difficulties lie not only in selecting appropriate molecular markers but also in efficiently and cost-effectively conducting genotyping. The cost of genotyping slash pines and loblolly pines is always very high because of the huge genome[35]. Both resequencing and RNA-seq cost much whereas SNP-based arrays are less laborious. Genotyping by target sequencing (GTBS) is theoretically an ideal genotyping strategy and the cost of targeted sequencing based on in-solution probes is relatively lower [36, 37], but this also not suitable for whole-exome sequencing or more boarder targeted sequencing in pine species because the price is still high. Therefore, our research group (Forest Germplasm Resources Research Group, Institute of Subtropical Forestry, Chinese Academy of Forestry) have designed the 51K liquid-phased probes (unpublished) for genotyping a large number of samples for slash pine and loblolly pine, which provide technical support for our study.

In this study, the 51K liquid-phase probe was used to genotype 38 pine varieties from 4 exotic pines. The genotyped SNP markers were filtered by the parameters of minor allele frequency, miss rate,

heterozygosity rate and the polymorphism information content (PIC) to get the core combination of SNPs that can efficiently identify the germplasm of exotic pines. Employing the Random Forest model in conjunction with PCA to simplify the core SNPs and constructing DNA fingerprints, which can be used as a reference for identifying foreign pines as a good breeder, germplasm management, and genetic diversity analysis.

2 Materials and methods

2.1. Materials and DNA extraction

A total of 38 samples were selected to construct DNA fingerprinting, including 29 accessions of *Pinus elliottii* (PEE), 2 accessions of *Pinus taeda* (PTA), 2 accessions of hybrid of *P. elliottii* × *P. caribbaea* (PEC) and 5 accessions of hybrid of *P. taeda*× *P. caribaea* (PAC). All materials were collected in Changle Forest Farm in Zhejiang Province, Hangzhou city. The samples in this study all have excellent traits in growth, wood property or turpentine production (Table1). In May 2023, the young needles were collected from the trees and refrigerated in ice bags before DNA was extracted.

Table 1．Basic information of germplasm resources of 38 pine trees

Species	Number of sample	Sampling site (Changle Forest Farm)	Female source	Excellent trait
Pinus elliottii (PEE)	2	PEE 2nd seed orchard (2012)	PEE breeding population	Fast-growing
	5	PEE seed orchard of elite lines (2006)	PEE elite family	
	4	PEE high-turpentine	PEE high-turpentine	High-turpentine

		family test forest (2015)	orchard from American	production
Pinus taeda （PTA）	2	PTA 2nd seed orchard (2012)	PTA breeding population	Fast-growing, excellent timber
P. taeda × P. caribaea （PAC）	5	Hybrid pine family test forest (2007)	PTA clone family 31062	Fast-growing, excellent timber
		Hybrid pine family test forest (2002)	PTA clone family 31041	
P. elliottii × P. caribaea （PEC）	2	Hybrid pine conle test forest (2009)	PTA clone family IV-60	Fast-growing, excellent timber

M5 HiPer Universal DNA Mini Kit (Mei5 Biotechnology, Beijing) was used to extract the DNA of the sample needles. The genome target region were captured by 51K liquid-phased probes designed by our lab for subsequently next-generating sequencing. The specific experimental steps are as follows:(1) The DNA was segmented and the end of the DNA fragment was repaired by Pre-PCR to amplify the library; (2) The probe hybridizes with the target region; (3) Hybridization probes were captured by magnetic beads using streptomycin affinity labeling; (4) The enriched target fragment was eluted and amplified by Post-PCR after capture;(5) Second generation sequencing was performed on the target region.

2.2. Data quality control and genotyping

The captured DNA region were sequenced by Illumina Xten high-throughput Sequencing platform. The raw data were stored in FASTQ format. The adapters and low quality data were filtered by FastQC software[38] to obtain the clean reads. Subsequently, single nucleotide polymorphism (SNP) detection is mainly implemented using GATK software toolkit (https://gatk.broadinstitute.org/hc/en-us). According to the localization results of clean Reads in the reference genome, samtools was used to remove duplications (Mark Duplication), GATK was used to perform local realignment and base

recalibration. To ensure the accuracy of SNP detected, GATK was used to detect and filter single nucleotide polymorphisms, and the final SNP loci was obtained. The mutation results are displayed and saved in vcf files.

2.3. Population genetic structure and phylogenetic analysis

Admixture software was used to analyze the optimal cluster number of samples, and MEGA 11.0 [39] software was used to construct a Neighbor-Joining phylogenetic tree to analyze the genetic structure of the population, then combining the result of principal component analysis (PCA) to analyze the kinship between the test samples.

2.4. Core SNP selection and functional annotation

The raw genotyped SNPs were screened in the following criteria: SNP loci in the Raw data were selected, and the screening criteria were as follows:(1)biallelic sites; (2) minor allele frequency (MAF) > 0.3; (3) miss rate=0; (4) conform to the Hardy-Weinberg equilibrium; (5) linkage disequilibrium (LD) value > 0.2; (6) PIC>0.35. The polymorphism information content (PIC) value was calculated according to minor allele frequency:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=n+1}^{n} 2 p_i^2 p_j^2$$

p_i and p_j are the i th and j th allele frequencies, respectively, and n is the number of alleles; (7) heterozygosity rate<0.25 [8, 9, 30, 40, 41]. The core SNPs were obtained after the screening of the above steps. VCFtools [42] and BCFtools [43] were used to perform the filtration.

Based on the functional annotation of SNPs in the 51K liquid-phased probes, the GO functional annotation information of core SNPs was extracted and visualized in R software.

2.5. SNP assessment and construction of DNA fingerprinting

Random forest is used to further simplify the number of core SNPs, aiming to identify as many samples as possible with a small number of SNPs, and it is implemented by the R package "Random Forest". The Mean Decrease Accuracy index generated by random forest for each SNP was sorted in descending order. SNP loci with the highest Mean Decrease Accuracy index were remained [44].

The software pcadapt v4.3.2 (https://bcm-uga.github.io/pcadapt/articles/pcadapt.html) were used to carry out the PCA analysis, which is used to check the efficiency of identification by different SNPs combination [8]. PCA calculations are performed by sequentially increasing the number of SNPs. The process is halted when clear separation among the samples is achieved on the PCA.

3.RESULTS

3.1. Genetic structure and phylogenetic relationship of 38 pine varieties

A total of 5,60,567 SNP loci were genotyped for 38 pine varieties by target sequencing of 51 K liquid phased probes. The genotyping rate was 94.61%. A total of 183,849 SNPs with a minor allele frequency (MAF) < 0.01 and a miss rate > 0.8 were retained as as the primary dataset for this experiment. The cluster situation of 38 varieties were estimated using ADMIXTURE software based on 1,83,849 SNPs loci. The cross-validation error value (CV error) is minimized when the value of K is 2, indicating that 38 varieties were divided into two clusters (Fig. 1A). All 29 PEE samples and their hybrids(PEC) were in the cluster I and all PTA samples and their hybrids were in the cluster II. The phylogenetic analysis shows the samle results (Fig.1B), which is consistent with the results of the ADMIXTURE software.

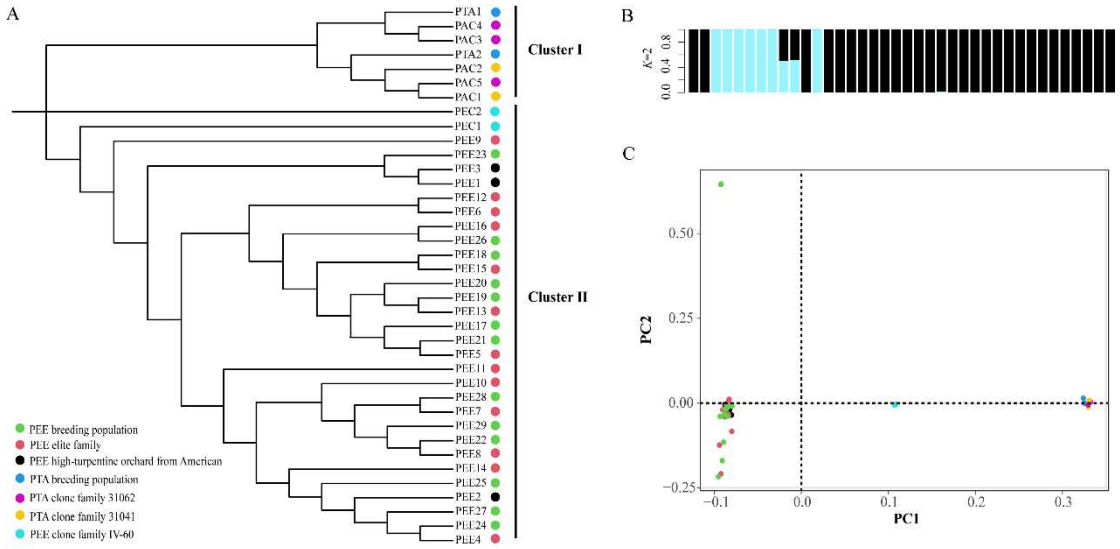

Fig.1. Phylogenetic and population genetic analyses of 38 pine varieties. **A:** The phylogenetic analysis ,the different circle represent the different female resources of the samples. **B:** Genetic structure analysis of 38 pines at *K*=2. **C:** Principal component analysis, the colors of the circles were consistent with the phylogenetic tree.

There is no obvious correlation between the genetic relationship and their collection origin. PEE samples come from 3 different origin so that the genetic distance among samples of PEE is relatively far. The PEC varieties containing the genes of Caribbean pines as the male parent have a independent position in Cluster I.As the same, the loblolly pine hibrids (PAC) share the genes of Caribbean and loblolly pines as the parents,so their positions in Cluster II are a little far away frome their female PTA.

3.2. Screening of core SNPs and functional annotation

Based on the MAF, miss rate, linked disequilibrium (LD) value, heterozygosity, polymorphism information content (PIC) and other indicators of SNPs, a rigorous screening process (Table 2) was formulated to screen high-quality core SNPs. A total of 344 SNPs were finally screened, which were regarded as the representative loci of the 38 pine germplasm resources as well as candidate loci for the construction of the DNA fingerprinting profile.

The MAF of 344 core SNPs ranged from 0.3026 to 0.5, with an average of 0.3524(Fig.2A). The PIC values ranged from 0.4221 to 0.5, with an average value of 0.4507(Fig.2B). The nucleotide polymorphisms (Π) value ranges from 8.55439E-07 to 1.95789E-06, with an average value of 9.21476E-07((Fig.2C); The expected heterozygosity (He) of 344 SNPs in 38 samples was 0.5433, and the observed heterozygosity (Ho) was between 0.3692 and 0.9826, with an average of 0.7811 (Fig.2D).

Table 2. The setting parameters for screening core SNPs

Screening Step	The number of remaining SNPs
Raw data	1,83,849
MAF>0.3，miss rate=0	22,593
hdw	13,931
ld>0.2	8,502
PIC>0.35	8,502
He<0.25	344

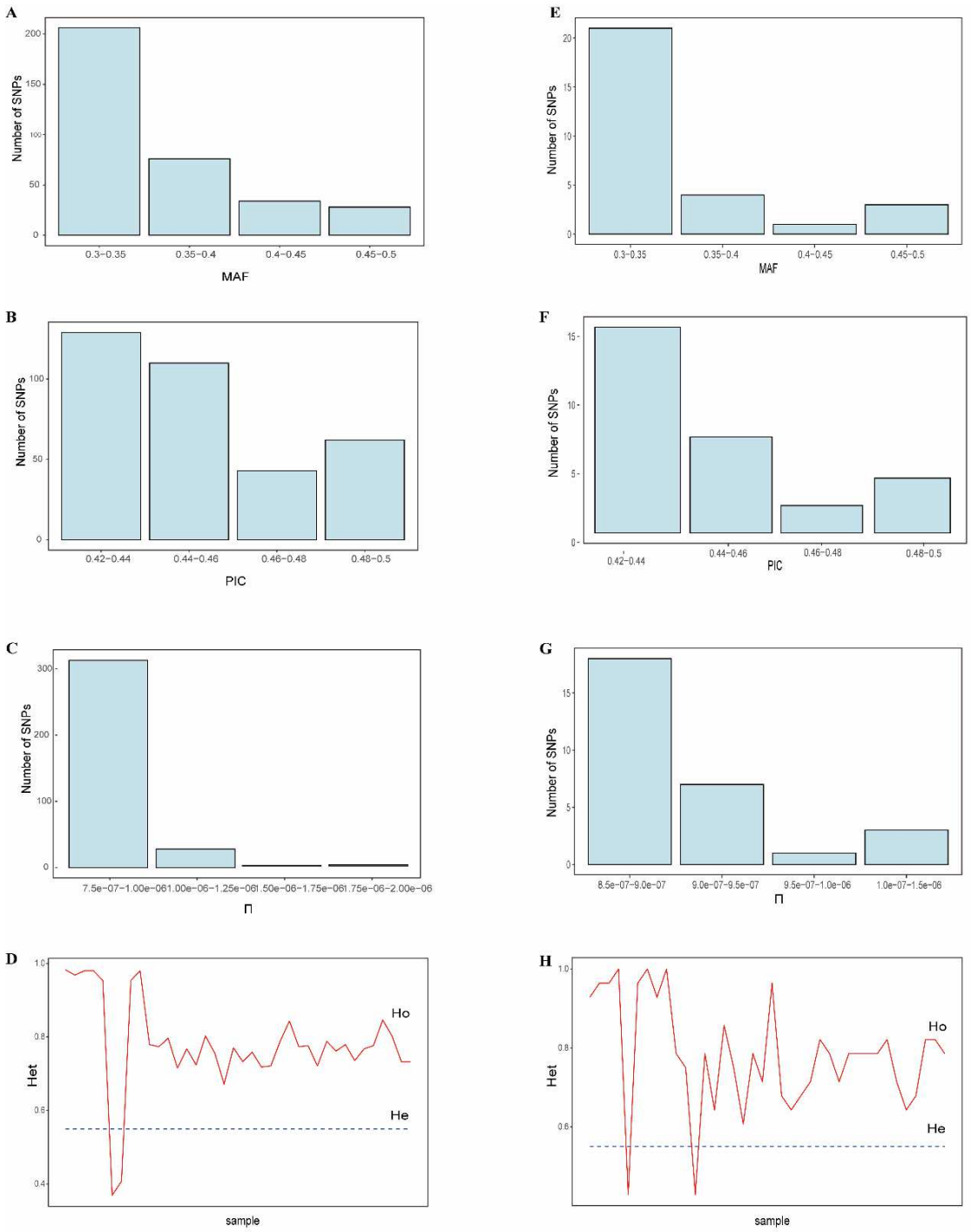

Fig.2. Genetic parameter information of 344 SNPs and 28 SNPs. **A-D:** MAF (A), PIC (B), Π(C), observed heterozygosity and expand heterozygosity(D) of 344 SNPs; **E-H:** MAF (E), PIC (F), Π(G), observed heterozygosity and expand heterozygosity(H) of 28 SNPs.

To clarify the genes where the core SNPs loci are located and the functions of these genes, the

annotation information of 344 core SNPs in this screening was counted based on the 51K liquid-phased probes(Fig.3). The results showed that 192 SNPs are located within genes with functional annotations, participating in various fundamental metabolic activities. The genes associated with the 119 SNPs are related to molecular functions, such as the activity of protease or protein kinase, and most of the SNPs are related to the binding of important substances such as DNA and ATP. The genes associated with the 32 SNPs are related to cell components, with 16 of them were related to cell membrane synthesis. The genes associated with 22 SNPs are related to physiological processes and participate in important physiological processes such as lipid metabolism and cell proliferation.

The gene that we found have the functions of active energy, metabolism and physiological, which could be analyzed by correlation analysis in a large population, and the identification of the good seed resources could be further carried out at the molecular-functional level. It has a great significance in elucidating the trait characteristics of different elite germplasm resources. The annotation information of these genes shows a lower association with cell components, which possibly due to the lack of annotation information of related genes.

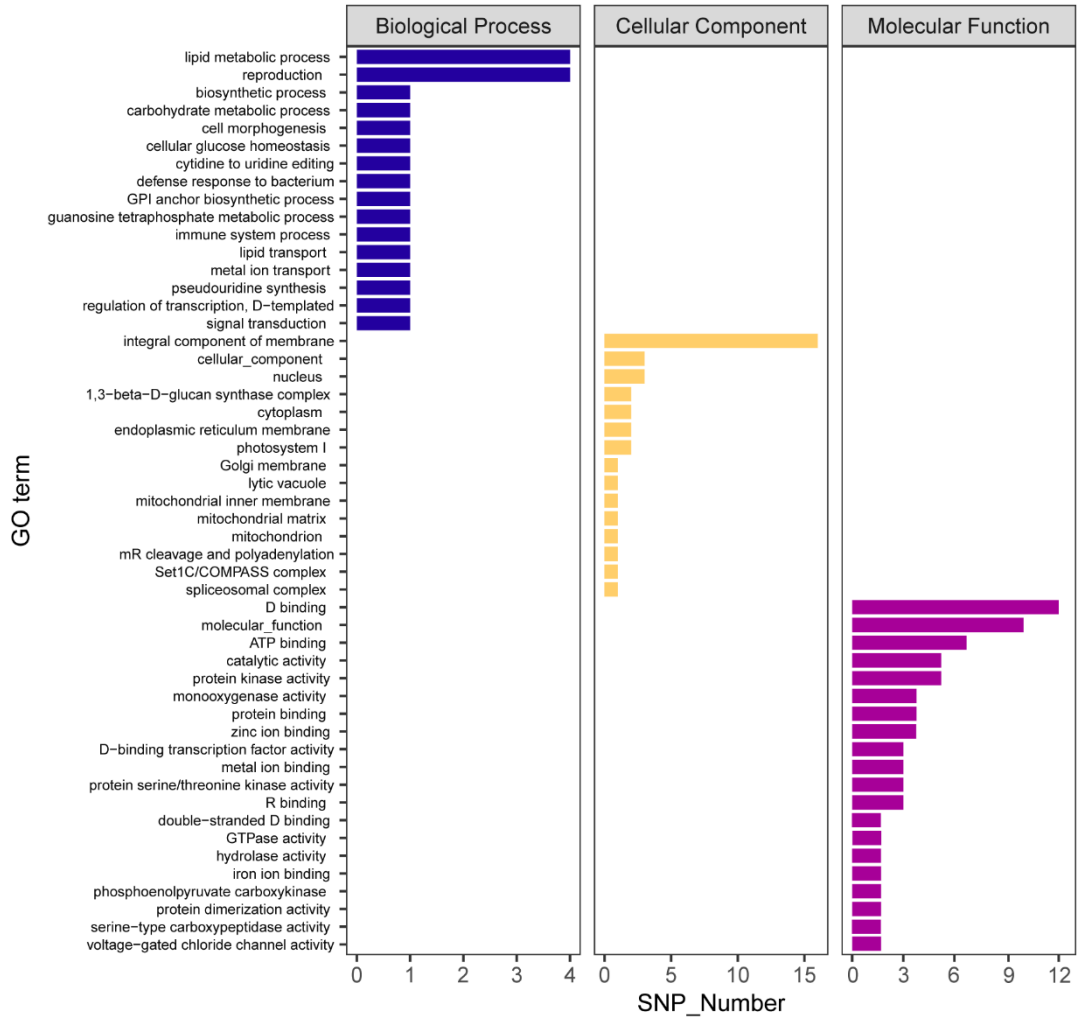

Fig.3. Annotation of core SNPs functions based on 51K liquid-phase probes.

3.3. Evaluation of core SNPs and construction of DNA fingerprinting

To improve the efficiency of variety identification, the core SNP loci need to be streamlined, aiming to identify a larger number of individuals with the fewest loci as much as possible. The Mean Decrease Accuracy index was generated for each SNP in the Random Forest model. Then, according to the value of the Mean Decrease Accuracy index, we sort SNPs in descending order, and preserve the SNP at the top of the order. Combining with the results of PCA to screen SNPs, a total of 28 loci were finally selected as representative loci for the 38 samples.

To evaluate whether the selected SNPs could effectively detect population diversity, principal component analysis and phylogenetic analysis were performed using 344 core SNPs and 28 SNPs for fingerprint construction, respectively (Fig.4, Fig.5). The results of PCA showed that the clustering effect of 344 SNPs and 28 SNPs was consistent with the clustering effect of the 1,83,849 SNPs (Figure 1C). With the streamlining of the number of SNPs, the proportion of differential loci among samples expanded. The final samples were significantly different in 28 SNP loci, which not only preserved the species characteristics, but also distinguished different individuals in the same species. According to the phylogenetic tree, four pine species cluster independly based on 344 core SNPs.The core SNPs in this experiment can effectively represent the genetic diversity of the test population. Finally, DNA fingerprinting was constructed using 28 SNP loci (Table 3, Fig.6).

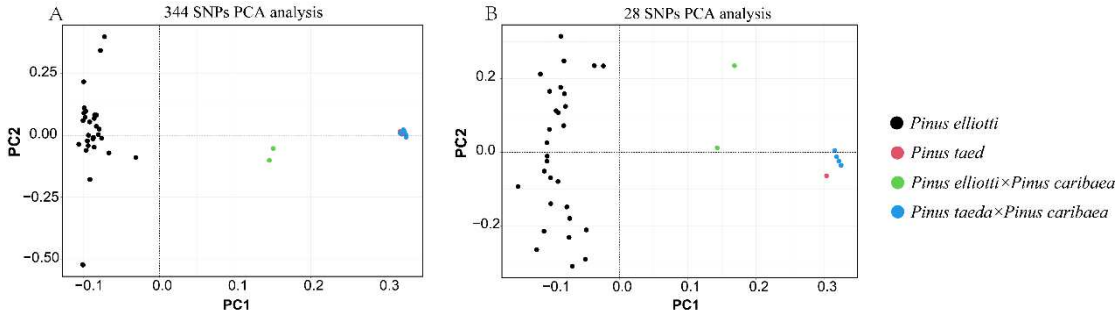

Figure.4. Principal component analysis based on 344 SNPs(**A**) and 28 SNPs(**B**). The black circle represents PEE samples. The red circle represents PTA samples. The green circle represents PEC samples. The blue circle represents PAC samples.

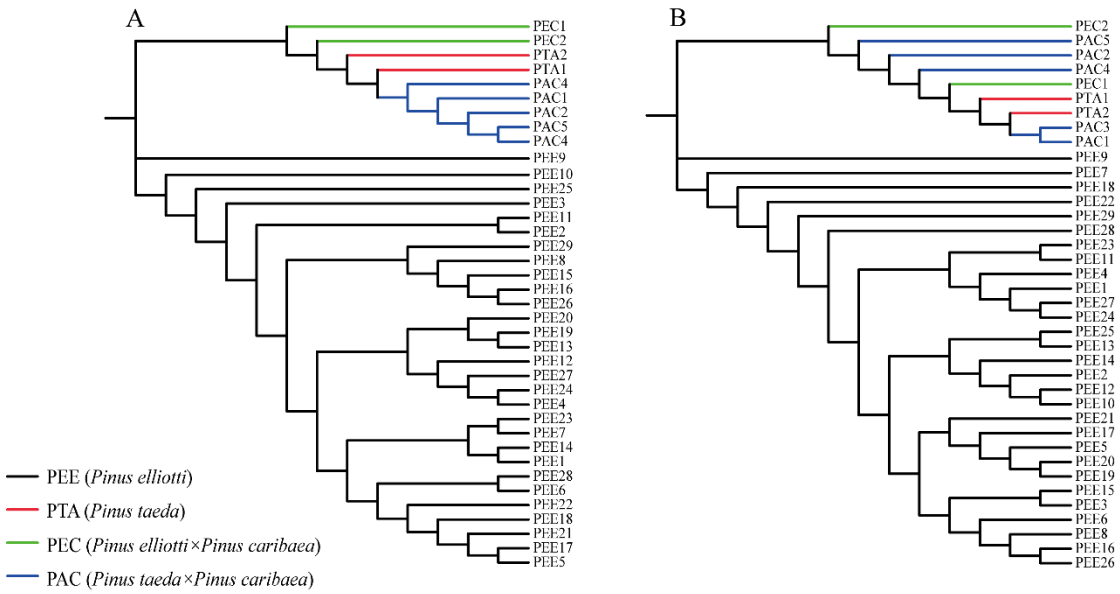

Fig.5. Phylogenetic relationship of 38 pines were analyzed based on 344 SNPs (A) and 28 SNPs (B). The black line represents PEE samples. The red line represents PTA samples. The green line represents PEC samples. The blue line represents PAC samples.

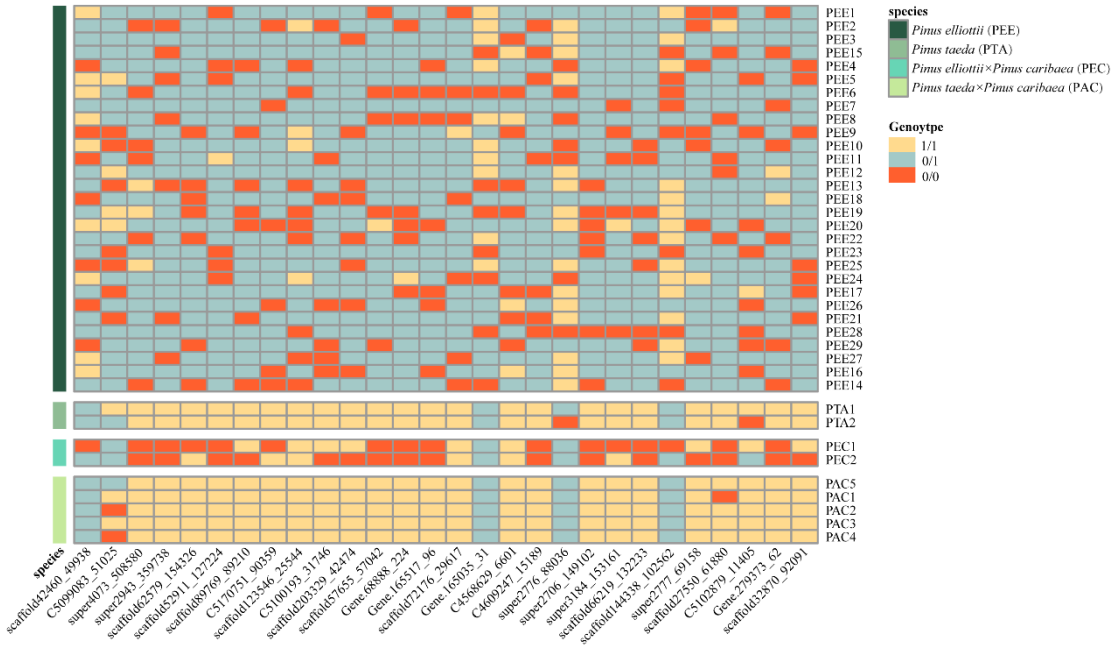

Fig.6. DNA fingerprinting of 38 pine germplasm resources. Genotype is 1/1 represent that it is homozygote of major alle gene, and genotype is 0/0 represent that it is homozygote of minor alley gene, and genotype is 0/1 represent that it is heterozygote. The different color in the left represents different exotic pines species.

The MAF of the 28 SNPs ranged from 0.302632 to 0.5, with an average of 0.314474 (Fig.2E). The PIC value ranged from 0.422092 to 0.5, with an average value of 0.444796 (Fig.2F). Π values ranged from 8.55439E-07 to 1.01333E-06, with an average value of 9.01E-07 (Fig.2G). The He of 38 samples at 28 loci was 0.55, and the Ho ranged from 0.428571 to 1, with an average value of 0.781955 (Fig.2H). The average value of MAF and PIC of 28 SNPs were lower than those of 344 core SNPs, but these values were still in the range of 0.3-0.5, indicating that the 28 SNPs we screened were still highly polymorphic.

Table 3. 28 SNP markers after simplification

Number	CHROM	POS	Allele A	Allele B	MAF	PIC	Π
1	super2706	1491092	A	T	0.302632	0.42200	8.55E-07
2	super2776	88036	G	A	0.5	0.5	1.01E-06
3	super2777	69158	G	C	0.342105	0.45013	9.12E-07
4	super2943	3597382	A	G	0.302632	0.42200	8.55E-07
5	super3184	1531611	G	C	0.315789	0.43213	8.76E-07
6	super4073	5085800	A	G	0.368421	0.46537	9.43E-07
7	scaffold2755	6188000	G	A	0.302632	0.42200	8.55E-07
8	scaffold3287	9209100	C	T	0.315789	0.43213	8.76E-07

9	scaffold424690	49938	C	T	0.342105	0.450138		9.12E-07
10	scaffold529114	12722	A	G	0.315789	0.432133		8.76E-07
11	scaffold576555	57042	T	C	0.302632	0.422092		8.55E-07
12	scaffold62579	15432	C	T	0.315789	0.432133		8.76E-07
13	scaffold66219	13223	C	T	0.302632	0.422092		8.55E-07
14	scaffold72176	29617	C	T	0.355263	0.458102		9.28E-07
15	scaffold89769	89210	T	C	0.315789	0.432133		8.76E-07
16	scaffold123546	25544	T	C	0.460526	0.496884		1.01E-06
17	scaffold144338	10256	A	G	0.486842	0.499654		1.01E-06
18	scaffold203329	42474	G	C	0.328947	0.441482		8.95E-07
19	C4568629	6601	G	A	0.447368	0.494476		1.00E-06
20	C4609247	15189	A	G	0.302632	0.422092		8.55E-07
21	C5099083	51025	G	C	0.302632	0.422092		8.55E-07
22	C5100193	31746	G	A	0.315789	0.432133		8.76E-07
23	C5102879	11405	T	C	0.328947	0.441482		8.95E-07

24	C5170751	90359	G	A	0.302632	0.422092	8.55E-07
25	Gene.165035	31	A	G	0.368421	0.465374	9.43E-07
26	Gene.165517	96	T	C	0.302632	0.422092	8.55E-07
27	Gene.279373	62	G	C	0.355263	0.458102	9.28E-07
28	Gene.68888	224	T	C	0.328947	0.441482	8.95E-07
Average					0.343985	0.343985	9.01E-07

4.DISCUSSION

DNA fingerprinting is a technique that can effectively identify the genotypes of animals and plants, which is mainly used in the studies of follow aspects: (1) genotype identification; (2) genetic diversity, population structure and genetic relatedness; (3) genome constitution of hybrids and polyploids. In recent years, with the development of sequencing technology, DNA fingerprinting based on SNP molecular markers has been widely used in crops and horticultural plants genotype identification and population genetic diversity analysis[45].

In this study, a total of 38 samples of exotic pine varieties were selected, including PEE (*Pinus elliottii*), PTA (*Pinus taeda*), PEC (*P. elliottii* × *P. caribaea*) and PAC (*P. taeda*× *P. caribaea*). SNPs was captured by 51K liquid-phased probes. After screening and evaluation, 344 core SNPs were obtained. Finally, DNA fingerprints were constructed for these exotic pines.

4.1. Construction of DNA fingerprinting for the exotic pines

In China, the exotic pines refer to several kinds of pine introduced from the United States in the

1930s, including slash pine, loblolly pine, Caribbean pine and their hybrids. These pine species are so similar in phenotype that they are almost indistinguishable at seedling stage. After adulthood, the tree can be distinguished by its stem canopy, branch angle and other traits, but the discrimination efficiency is still very low. As the exotic pines are widely planted in China, their interspecific hybrids have become important afforestation tree species, and these hybrid pines are difficult to distinguish by nutritional organs. This has caused a great obstacle to the efficient management of germplasm resources.

The conifer trees have huge genomes, and the three exotic pines in this research have a genome of over 20GB [46, 47]. Construction of a cost and stable DNA fingerprinting is challenging. Genome-resequencing is too cost, and transcriptome-sequencing or Specific-Locus Amplified Fragment Sequencing (SLAF) could not get a stable results [35]. Genotyping by target sequencing (GTBS) is theoretically an ideal genotyping strategy and the cost of targeted sequencing based on in-solution probes is relatively lower, which is an optimal method for the construction of a huge genome DNA fingerprinting [37, 48]. In this study, a total of 344 core SNPs representing the exotic pine samples were selected by strict screening of genetic parameters such as MAF, miss rate, heterozygosity rate and PIC of SNPs. Then, a total of 28 SNPs were selected to construct DNA fingerprint through further simplification and evaluation, which can completely distinguish 38 pine samples.

According to the values of genetic parameters, the nucleotide polymorphism and genetic diversity of the samples can be performed, whether using 28 SNPs or 344 SNPs. There are obvious differences between the DNA fingerprints of different species. Addition, the same species shows a high degree of consistency and differences with other species in some specific sites, these sites are the key to quickly distinguish them.

4.2. Genetic relation of exotic pine germplasm in China

The exotic pines have been cultivated for a hundred years since they were introduction from America [49, 50]. They are widely planted in subtropical and tropical areas in China currently and provide a rich source of variation for pine germplasm breeding. Understanding the genetic relationships between varieties at the genomic level is necessary and helpful to improve breeding efficiency of exotic pine germplasm.

In this study, the 51K liquid-phase probe was used to capture the SNP molecular markers of 38 pine samples, including PEE, PTA, PEC and PAC. The principal component analysis, phylogenetic analysis and genetic population analysis were carried out to clarify the genetic structure and kinship of samples. Then, we found that the results of these analysis are basically consistent. Two sub population can be divided from the pine population. PEE and PEC cluster together while PTA and PAC cluster together. The reason for this cluster result is mainly due to the difference in their parents. The kinship among the PEE samples and the genetic distance between PEE and PEC are distant. This result may be related to the genetic background of slash pine. We selected a large number of slash pine samples involving three different sources, and this result suggests that the genetic background of slash pine is more complex than that of other three pine species.

In the selection of parents in cross breeding, a close genetic relationship of the hybrid parents is not conducive to the selection of elite new varieties. On the contrary, if the genetic relationship of the hybrid parents is far away and the genotypic differences are large, the genetic variation of the offspring is more extensive. Therefore, varieties between different groups or clusters can be selected as parent materials for hybridization, so as to cultivate excellent hybrid offspring[51]. In this study, the genetic relationship between PEE samples that clustered into multiple branches can be selected as hybrid

parents in the future. As the hybrid progeny of slash pine and Caribbean pine, a long genetic distance was shown between the two PEC samples.Combining with the excellent growth trait of PEC samples, we speculate that their mother tree has strong genetic diversity. Addition, all PTA and PAC samples perform an excellent growth trait, and they were cluster in same group, indicating that PTA samples also can be chosen as the hybrid parent in the future.

The results of our study provide a favorable reference and basis for better utilization of pines germplasm resources in China, and this will be of great significance in the selection of parental lines in future breeding programs of pines and for the effective utilization of heterosis in the breeding work.

5. CONCLUSIONS

In this study, 38 pines were used as test samples, including PEE (*Pinus elliottii*), PTA (*P. taeda*), PEC (*P. elliottii* × *P. caribaea*) and PAC (*P. taeda*× *P. elliottii*). SNPs were captured by 51K liquid-phase probes of slash pine and loblolly pine. The minor allele frequency, miss rate, PIC and heterozygous rate of SNPs were strictly screened, and 344 core SNPs that can initially identify pine elite strains were obtained. Subsequently, 28 SNPs were selected as the optimal loci' combination, and they could identify all the materials used in this study. However, the identification and analysis effect of this set of SNPs on other exotic pine germplasm still need to be analyzed according to the actual test results.

Declarations

Ethics approval and consent to participate

The authors confirm that all the experimental methods complied with the IUCN Policy Statement

on Research Involving Species at Risk of Extinction and the Convention on the Trade in Endangered Species of Wild Fauna and Flora.

The pine samples used in this research are collected at Changle Forest Farm, Zhejiang, China, and the seeds are freely available for scientifc research. Permission to collect the samples was obtained from Changle Forest Farm and Research Institute of Subtropical Forestry, Chinese Academy of Forestry for only research purposes.

Consent for publication

Not applicable.

Availability of data and materials

The data have been requested for a Bioproject in the NCBI database (National Center for Biotechnology Information (nih.gov)), and specific information on the biological samples has been submitted with the accession number PRJNA1073171.BioSample submission: SUB14206957. The datasets presented in this study can be found in online repositories. The direct web link can be found below: https://zenodo.org/records/10615722

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the Science and Technology Innovation 2030 Program, Ministry of Science and Technology, People's Republic of China (2023ZD0405902) and the National Key Research and Development Program of Ministry of Science and Technology of the People’s Republic of China (2022YFD2200023-2).

Authors' contributions

L.Q.: Conceptualization, Supervision, Project administration, Funding acquisition **W.Y.:** Investigation, Resources, Data Curation, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization **D.S.:** Writing - Review & Editing **D. X.:** Investigation, Resources, Methodology **H.Q.:** Investigation, Resources. All authors commented on the paper. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

We thank the China Shengshidai (Beijing) Biotech Co., Ltd. for their technical assistance in DNA capture sequencing based on 51K liquid-phased probes.

Authors' information

First Author: Wu Yadi

Corresponding author: Luan Qifu

Other Author: Diao Shu, Ding Xianyin, Huang Qinyun.

REFERENCES

1. Ding X, Tao X, Diao S, Luan Q, Jiang J: **Evaluation of basic density of standing trees of *Pinus elliottii* by Pilodyn and Resistograph**. *Journal of Nanjing Forestry University(Natural Sciences Edition)* 2020, **44**(03):142-148.
2. Luan Q, Ding X, Xu Y, Xu L, Jiang J: **Study on the Application of Principal component regression Analysis in early selection of loblolly Pine**. *Journal of Anhui Agricultural University* 2019, **46**(02):264-269.
3. Ding X, Li Y, Zhang Y, Diao S, Luan Q, Jiang J: **Genetic analysis and elite tree selection of the main resin components of slash pine**. *Frontiers in Plant Science* 2023, **14**:1079952.
4. Luan Q: **Cross breeding and genetic Analysis of several Pine trees**. *Doctor.* Chinese Academy of Forestry; 2011.
5. Situ R, Zhong d, Ye W, Situ W, Mao J: **Establishment of near Infrared Prediction Model for Total content of 16 flavonoids in *Pinus elliottii*× *PinusCaribaea***. *Forestry and Environmental Science* 2021, **37**(04):1-6.
6. Yuan X, Li Z, Xiong L, Song S, Zheng X, Tang Z, Yuan Z, Li L: **Effective identification of varieties by nucleotide polymorphisms and its application for essentially derived variety identification in rice**. *BMC bioinformatics* 2022, **23**(1):1-16.

7. KANG X: **Research progress of forest genetics and tree breeding**. *Journal of Nanjing Forestry University* 2020, **44**(3):1.
8. Lin P, WANG K, Yao X, Ren H: **Construction of Molecular ID Card of main varieties of Camellia oleifera based on Transcriptome SNP**. *Journal of Integrative Agriculture* 2023, **56**(02):217-235.
9. Ma X, Cui D, Han B, Jiao C, Han L: **Genomic DNA fingerprinting Identification of Rice Germplasm Resources**. *Journal of Plant Genetic Resources* 2023, **24**(04):1106-1113.
10. Jeffreys AJ, Wilson V, Thein SL: **Hypervariable 'minisatellite'regions in human DNA**. *Nature* 1985, **314**(6006):67-73.
11. Vuylsteke M, Peleman JD, Van Eijk MJ: **AFLP technology for DNA fingerprinting**. *Nature protocols* 2007, **2**(6):1387-1398.
12. Darine T, Allagui M, Rouaissi M, Boudabbous A: **Pathogenicity and RAPD analysis of Phytophthora nicotianae pathogenic to pepper in Tunisia**. *Physiological and Molecular Plant Pathology* 2007, **70**(4-6):142-148.
13. Micheli MR, Bova R, Pascale E, D'Ambrosio E: **Reproducible DNA fingerprinting with the random amplified polymorphic DNA (RAPD) method**. *Nucleic Acids Research* 1994, **22**(10):1921.
14. Lefebvre V, Palloix A, Rives M: **Nuclear RFLP between pepper cultivars (Capsicum annuum L.)**. *Euphytica* 1993, **71**(3):189-199.
15. Jamil S, Shahzad R, Yasmeen E, Rahman SU, Younas M, Iqbal MZ: **DNA fingerprinting of Pakistani maize hybrids and parental lines using simple sequence repeat markers**. *Pak J Bot* 2020, **52**(6):2133-2145.
16. Wang Z, Pan Y, Luo J, You Q, Xu L, Zhang H, Que Y: **SSR-based genetic identity of sugarcane clones and its potential application in breeding and variety extension**. *Sugar Tech* 2020, **22**:367-378.
17. UPOV: **Guidelines for DNA profiling: molecular marker selection and database construction (BMT guidelines)**. 2010.
18. Wen M, Wang H, Xia Z, Zou M, Lu C, Wang W: **Developmenrt of EST-SSR and genomic-SSR markers to assess genetic diversity in Jatropha Curcas L**. *BMC research notes* 2010, **3**(1):1-8.
19. Zietkiewicz E, Rafalski A, Labuda D: **Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification**. *Genomics* 1994, **20**(2):176-183.
20. Lanka CL, Ram M, Krishna SM: **DNA Fingerprinting of Crops and Its Significance in Crop Improvement**. *International Journal of Plant & Soil Science* 2023, **35**(16):232-242.
21. Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P: **Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences**. *American journal of botany* 2012, **99**(2):193-208.
22. Jenkins S, Gibson N: **High-throughput SNP genotyping**. *Comparative and Functional Genomics* 2002, **3**(1):57-66.
23. Kim S, Misra A: **SNP genotyping: technologies and biomedical applications**. *Annu Rev Biomed Eng* 2007, **9**:289-320.
24. Mahabir A, Motilal LA, Gopaulchan D, Ramkissoon S, Sankar A, Umaharan P: **Development of a core SNP panel for cacao (*Theobroma cacao L.*) identity analysis**. *Genome* 2020, **63**(2):103-114.
25. Montes JM, Technow F, Martin M, Becker K: **Genetic diversity in Jatropha curcas L. assessed**

with SSR and SNP markers. *Diversity* 2014, **6**(3):551-566.

26. Xia W, Luo T, Zhang W, Mason AS, Huang D, Huang X, Tang W, Dou Y, Zhang C, Xiao Y: **Development of high-density SNP markers and their application in evaluating genetic diversity and population structure in Elaeis guineensis**. *Frontiers in plant science* 2019, **10**:130.
27. McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y, Cho YG, Huang N, Ishii T, Blair M: **Microsatellite marker development, mapping and applications in rice genetics and breeding**. *Oryza: from molecule to plant* 1997:89-99.
28. Nagaraju J, Kathirvel M, Kumar RR, Siddiq E, Hasnain SE: **Genetic analysis of traditional and evolved Basmati and non-Basmati rice varieties by using fluorescence-based ISSR-PCR and SSR markers**. *Proceedings of the National Academy of Sciences* 2002, **99**(9):5836-5841.
29. Tian H-L, Wang F-G, Zhao J-R, Yi H-M, Wang L, Wang R, Yang Y, Song W: **Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties**. *Molecular Breeding* 2015, **35**:1-11.
30. Wang S, Huang Y, Liang C, Wang J, Yang Q: **Construction of fingerprints of *Brassica napus* based on whole Genome SNP**. *Chinese Journal of Oil Crops Sciences* 2022, **44**(05):966-972.
31. Fan X, Yu W, Cai C, Lin Y, Wang Z, Fang w, Zhang J, Ye N: **Construction of Molecular ID Card of Tea Variety Resources based on SNP Marker**. *Journal of Integrative Agriculture* 2021, **54**(08):1751-1772.
32. Bentley N, Grauke L, Klein P: **Genotyping by sequencing (GBS) and SNP marker analysis of diverse accessions of pecan (*Carya illinoinensis*)**. *Tree Genetics & Genomes* 2019, **15**:1-17.
33. Chancerel E, Lepoittevin C, Le Provost G, Lin Y-C, Jaramillo-Correa JP, Eckert AJ, Wegrzyn JL, Zelenika D, Boland A, Frigerio J-M: **Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine**. *BMC genomics* 2011, **12**(1):1-14.
34. De La Torre AR, Puiu D, Crepeau MW, Stevens K, Salzberg SL, Langley CH, Neale DB: **Genomic architecture of complex traits in loblolly pine**. *New Phytologist* 2019, **221**(4):1789-1801.
35. Ding X, Diao S, Luan Q, Wu HX, Zhang Y, Jiang J: **A transcriptome-based association study of growth, wood quality, and oleoresin traits in a slash pine breeding population**. *PLoS genetics* 2022, **18**(2):e1010017.
36. de Oliveira Junkes CF, de Araújo Júnior AT, de Lima JC, de Costa F, Füller T, de Almeida MR, Neis FA, da Silva Rodrigues-Correa KC, Fett JP, Fett-Neto AG: **Resin tapping transcriptome in adult slash pine (*Pinus elliottii var. elliottii*)**. *Industrial Crops and Products* 2019, **139**:111545.
37. Guo Z, Yang Q, Huang F, Zheng H, Sang Z, Xu Y, Zhang C, Wu K, Tao J, Prasanna BM: **Development of high-resolution multiple-SNP arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip**. *Plant communications* 2021, **2**(6).
38. Lo C-C, Chain PS: **Rapid evaluation and quality control of next generation sequencing data with FaQCs**. *BMC bioinformatics* 2014, **15**(1):1-8.
39. Tamura K, Stecher G, Kumar S: **MEGA11: molecular evolutionary genetics analysis version 11**. *Molecular biology and evolution* 2021, **38**(7):3022-3027.
40. Wang Y, Lv H, Xiang X, Yang A, Feng Q, Dai P, Li Y, Jiang X, Liu G, Zhang X: **Construction of a SNP fingerprinting database and population genetic analysis of cigar tobacco germplasm resources in China**. *Frontiers in plant science* 2021, **12**:618133.

41. Yang Y, Lyu M, Liu J, Wu J, Wang Q, Xie T, Li H, Chen R, Sun D, Yang Y: **Construction of an SNP fingerprinting database and population genetic analysis of 329 cauliflower cultivars**. *BMC Plant Biology* 2022, **22**(1):1-11.
42. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**(15):2156-2158.
43. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM: **Twelve years of SAMtools and BCFtools**. *Gigascience* 2021, **10**(2):giab008.
44. Liu Y, Qin X, He S, Li W, Wang Y, Jia Z, Liu M: **Screening SNP Loci with High Information content by combining PCA with Random Forest-- Application in breed Identification of Sheep**. *Computer Engineering and Application* 2018, **54**(16):235-240.
45. Nybom H, Weising K, Rotter B: **DNA fingerprinting in botany: past, present, future**. *Investigative genetics* 2014, **5**(1):1-35.
46. Ahuja MR, Neale DB: **Evolution of genome size in conifers**. *Silvae genetica* 2005, **54**(3):126-137.
47. Prunier J, Verta JP, MacKay JJ: **Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function**. *New Phytologist* 2016, **209**(1):44-62.
48. Chen H, Xie W, He H, Yu H, Chen W, Li J, Yu R, Yao Y, Zhang W, He Y: **A high-density SNP genotyping array for rice biology and molecular breeding**. *Molecular plant* 2014, **7**(3):541-553.
49. Zhang Y, Ding X, Luan Q, Jiang J, Diao S: **Identification and Tissue-Specific Expression Analysis of CYP720B Subfamily Genes in Slash Pine and Loblolly Pine**. *Forests* 2022, **13**(2):283.
50. Zhang Y, Luan Q, Jiang J, Li Y: **Prediction and utilization of malondialdehyde in exotic pine under drought stress using near-infrared spectroscopy**. *Frontiers in Plant Science* 2021, **12**:735275.
51. Gao Y, Xie Y, Yang Z, Chen L, Wu T: **Genetic relationship among different tea plum varieties based on flowering difference and ISSR-PCR analysis**. *NON-WOOD FOREST RESEARCH* 2022, **40**(04):209-217.