Core set construction and association analysis of *Pinus massoniana* from Guangdong province in southern China using SLAF-seq

Qingsong Bai¹,², Yanling Cai¹,², Boxiang He¹,², Wanchuan Liu³, Qingyou Pan³ & Qian Zhang¹,²

Germplasm resource collection and utilization are important in forestry species breeding. High-throughput sequencing technologies have been playing increasing roles in forestry breeding. In this study, specific-locus amplified fragment sequencing (SLAF-seq) was employed to analyze 149 masson pine (*Pinus massoniana*) accessions collected from Guangdong in China. A large number of 471,660 SNPs in the total collection were identified from 599,164 polymorphic SLAF tags. Population structure analysis showed that 149 masson pines could not be obviously divided into subpopulations. Two core sets, containing 29 masson pine accessions for increasing resin and wood yield respectively, were obtained from the total collection. Phenotypic analyses of five traits showed abundant variations, 25 suggestive and 9 significant SNPs were associated with the resin-yielding capacity (RYC’) and volume of wood (VW) using EMMAX and FaST-LMM; 22 suggestive and 11 significant SNPs were associated with RYC’ and VW using mrMLM and FASTmrMLM. Moreover, a large number of associated SNPs were detected in trait HT, DBH, RW and RYC using mrMLM, FASTmrMLM, FASTmrEMMA and ISIS EM-BLASSO. The core germplasm sets would be a valuable resource for masson pine improvement and breeding. In addition, the associated SNP markers would be meaningful for masson pine resource selection.

Masson pine (*Pinus massoniana*) is a native species that grows throughout central and southern China. Besides its wide uses in the wood, pulp and paper industries, this species has long been employed as the main source of resin, a hydrocarbon secretion of many plants that is widely used to produce resin and turpentine for the chemical industry¹. Masson pine is the most important resin tapping tree species in China and should thus be preserved². However, due to its high commercial value, this species has been subjected to over-exploitation during past decades, leading to a gradual decrease in genetic resources³. Protection and sustainable use of the preserved masson pine resource are urgent problems for researchers.

Genetic structure and diversity analyses could help to scientifically simplify the resources. Various types of molecular markers, including RAPD, SRAP, SSR, and ISSR, have been used to estimate genetic relationships and genetic distances in masson pine⁴–⁹. Single nucleotide polymorphisms (SNPs) have been widely reported in recent years because they are the most abundant and stable type of genetic marker in most genomes¹⁰. Deep sequencing technology has been rapidly developed to exploit these advantages and has enabled the high-throughput identification of SNPs¹¹–¹³, albeit with the disadvantage of becoming cost-prohibitive when the population is large. The genomes of conifer trees such as *Pinus taeda* are complex and fairly long¹⁴. To reduce time and labor costs, reduced-representation genome sequencing has been widely used in plant genome sequencing¹⁴. Considering that whole-genome deep sequencing is still expensive and usually unnecessary¹⁵, several simplified and cost-effective methods for SNP discovery and high-throughput genotyping have been developed, such as reduced representation library (RRL) sequencing¹⁶, restriction-site associated DNA sequencing (RAD)¹⁶,¹⁷, and
two-enzyme genotyping by sequencing (GBS)\(^\text{18}\). In recent years, a new strategy for de novo SNP discovery and genotyping of large populations, referred to as specific-locus amplified fragment sequencing (SLAF-seq), has been employed\(^\text{19}\). SLAF-seq is a high throughput, highly fast, highly efficient and cost-effective method for developing large-scale SNP and InDel markers\(^\text{19}\). By using enzyme digestion techniques, an SLAF-seq library containing specific size fragments of DNA can be obtained. Then, we could identify a polymorphic specific SNP locus from all of the accessions through software alignment. This high-resolution method has been tested on many organisms, including crape myrtle\(^\text{20}\), cucumber\(^\text{21}\), rapeseed\(^\text{22}\), sesame\(^\text{23}\) and soybean\(^\text{24}\). Moreover, this method has been widely used in GWAS for important traits\(^\text{20,25,26}\), as well as in the development of core germplasm\(^\text{27}\).

To better understand the genetic relationship and the genetic architecture of wood and resin yield traits of the \(P.\ massoniana\) accessions in Guangdong province, we conduct a genome-wide SNP discovery based on the SLAF-seq method. The identified SNPs were used to examine the masson pine population structure. Then, we selected a core set of masson pine germplasm resources for improving resin-yielding capacity (RYC) and volume of wood (VW). Finally, a genome-wide association study (GWAS) strategy was used to identify the SNP locus associated with growth, wood and resin yield traits. The results would be of great value for masson pine selection and breeding.

Results

**Sequencing quality statistics.** By SLAF-seq, 1759.00 M reads were obtained from this experiment. The average Q3 value was 92.78%, and the average GC content was 37.95% (see Supplementary Table S1). A large number of 3,232,864 SLAF tags were identified throughout the masson pine genome. The average sequencing depth of the tags was 16.98 \(\times\) (see Supplementary Table S2). Subsequently, a total of 599,164 polymorphic SLAF tags containing 2,774,976 SNPs were developed for the 149 samples that were used for further analysis. After filtering out the invalid SNPs, 471,660 SNPs were remained among the 149 masson pine accessions.

**Population structure and linkage disequilibrium analysis.** We applied clustering analysis to the samples using ADMIXTURE software (Fig. 1a). This method has been used with large sample sizes, exhibiting a strong capability to assign individuals into populations. The estimated membership fractions of the 149 accessions for different values of K ranged from 1 to 10, and the maximum likelihood revealed by the population structure
showed an optimum value of 1 (K = 1; Fig. 1b), indicating that the masson pines in Guangdong could not be
categorized into different subpopulations. It is important to use population-based methods to separate accessions
from mixed populations into unstructured subpopulations, as this allows for association analyses between
phenotypes and molecular bands to be conducted in homogenous subpopulations39. Population analysis indicated
that these masson pines were not excessively separated and could be used for association analysis. We also used
the structure and fastStructure to calculate appropriate K value (see Supplementary Fig. S1). The results showed
that the highest delta K value was obtained when K of the masson pine population was 2; the highest marginal
likelihood was obtained when the K value was 6. The geographical distributions of the masson pines were also
not consistent with the population structure in the two methods. Some more discussions should be added in the
population structure analysis of masson pines in Guangdong.

Linkage disequilibrium (LD) is the non-random association of alleles at different loci and may indicate the
genetic forces that structure the genome39. Investigations of genetic diversity and LD are prerequisites for associ-
| Accession | Location | VW  | RYC  | Core set for wood | Core set for resin |
|-----------|----------|-----|------|-------------------|-------------------|
| GW54      | DQ       | 0.42| 342.56 | ●                  | ○                  |
| GW111     | DQ       | 0.76| 302.02 | ●                  | ○                  |
| GW9       | GZ       | 0.28| 177.93 | ●                  | ○                  |
| GW24      | GZ       | 0.45| 235.84 | ●                  | ○                  |
| GW29      | GZ       | 0.31| 193.82 | ●                  | ○                  |
| GW31      | GZ       | 0.3 | 206.39 | ●                  | ○                  |
| GW45      | GZ       | 0.3 | 220.82 | ●                  | ○                  |
| GW30      | GZ       | 0.26| 213.42 | ●                  | ○                  |
| GW51      | GZ       | 0.26| 184.64 | ●                  | ○                  |
| GW106     | GZ       | 0.22| 192.44 | ●                  | ○                  |
| GW2       | XY       | 0.3 | 252.16 | ●                  | ○                  |
| GW4       | XY       | 0.32| 202.59 | ●                  | ○                  |
| GW7       | XY       | 0.34| 225.14 | ●                  | ○                  |
| GW8       | XY       | 0.42| 248.18 | ●                  | ○                  |
| GW20      | XY       | 0.29| 237.42 | ●                  | ○                  |
| GW23      | XY       | 0.35| 260.06 | ●                  | ○                  |
| GW32      | XY       | 0.31| 186.59 | ●                  | ○                  |
| GW36      | XY       | 0.32| 204.42 | ●                  | ○                  |
| GW48      | XY       | 0.29| 201.89 | ●                  | ○                  |
| GW71      | XY       | 0.37| 237.71 | ●                  | ○                  |
| GW72      | XY       | 0.44| 238.97 | ●                  | ○                  |
| GW78      | XY       | 0.44| 301.81 | ●                  | ○                  |
| GW85      | XY       | 0.32| 180.46 | ●                  | ○                  |
| GW93      | XY       | 0.33| 254.26 | ●                  | ○                  |
| GW116     | XY       | 0.38| 191.65 | ●                  | ○                  |
| GW118     | XY       | 0.39| 255.66 | ●                  | ○                  |
| GW27      | XY       | 0.26| 184.92 | ●                  | ○                  |
| GW50      | XY       | 0.28| 218.07 | ●                  | ○                  |
| GW87      | XY       | 0.42| 186.79 | ●                  | ○                  |
| GW103     | XY       | 0.39| 168.39 | ●                  | ○                  |
| GW109     | XY       | 0.3 | 164.84 | ●                  | ○                  |
| GW142     | XY       | 0.33| 176.77 | ●                  | ○                  |
| GW91      | YN       | 0.5 | 192.2  | ●                  | ○                  |
| GW52      | YN       | 0.32| 170.98 | ●                  | ○                  |

**Table 1.** Phenotypes and categories of core sets for wood and resin.

**Figure 2.** The PCA plots of core germplasm sets for resin and wood. (a) The PCA plot of the core set for resin. (b) The PCA plot of the core set for wood.
the SNPs developed by the EMMAX method completely overlapped with the SNPs developed by the FaST-LMM method irrespective of trait RYC’ or VW (Table 2). Eight SNPs (Marker643442, Marker650102, Marker530780, Marker297054, Marker279561, Marker210060, Marker526082, Marker582947) that significantly associated with RYC’ were simultaneously developed by EMMAX and FaST-LMM methods, which indicated that those SNPs were very valuable and significant in breeding. However, no SNPs were developed by MLM methods in all the traits.

In this study, we also used the multi-locus methods mrMLM, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO, pKWmEB and pLARmEB in mrMLM.GUI version 3.2 to identify associated SNPs. The result showed that 11 SNPs and 11 SNPs were associated with trait RYC’ and VW using mrMLM method, including 8 significant SNPs (Marker124737, Marker174624, Marker482425, Marker279561, Marker526082, Marker582947) that significantly associated with RYC’ were simultaneously developed by EMMAX and FaST-LMM methods, which indicated that those SNPs were very valuable and significant in breeding. However, no SNPs were developed by MLM methods in all the traits.

In this study, we also used the multi-locus methods mrMLM, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO, pKWmEB and pLARmEB in mrMLM.GUI version 3.2 to identify associated SNPs. The result showed that 11 SNPs and 11 SNPs were associated with trait RYC’ and VW using mrMLM method, including 8 significant SNPs (Marker124737, Marker174624, Marker482425, Marker279561, Marker526082, Marker582947) that significantly associated with RYC’ were simultaneously developed by EMMAX and FaST-LMM methods, which indicated that those SNPs were very valuable and significant in breeding. However, no SNPs were developed by MLM methods in all the traits.

Figure 3. The manhattan plots and Q–Q plots of traits RYC’ and VW using EMMAX and FaST-LMM. Each dot in the Manhattan plot represents one SNP. The horizontal dotted red and blue lines indicate the suggestive and significant thresholds.
Table 2. RYC’ and VW associated SNPs using EMMAX and FAST-LMM. Note: ○ means that the SNP could be developed by that model.

| Trait | Marker | Position | Alleles | EMMAX P-value | FaST-LMM P-value |
|-------|--------|----------|---------|---------------|------------------|
| RYC’  | Marker643442 | 96       | T/C     | 3.12E-09      | 3.12E-11         |
|       | Marker650102 | 85       | A/G     | 2.22E-10      | 1.15E-10         |
|       | Marker530780 | 30       | G/T     | 2.06E-10      | 1.70E-10         |
|       | Marker297054 | 111      | T/G     | 1.82E-08      | 7.07E-10         |
|       | Marker279561 | 86       | G/A     | 1.30E-09      | 1.04E-09         |
|       | Marker210060 | 123      | C/T     | 8.90E-09      | 4.87E-09         |
|       | Marker26082 | 61       | C/T     | 1.28E-08      | 1.24E-08         |
|       | Marker582947 | 199      | G/A     | 1.52E-08      | 1.51E-08         |
|       | Marker231539 | 255      | T/G     | 2.37E-08      | 1.63E-08         |
|       | Marker441368 | 48       | C/T     | 2.76E-08      | 2.63E-08         |
|       | Marker437366 | 105      | C/T     | 5.29E-08      | 5.23E-08         |
|       | Marker507605 | 70       | C/A     | 8.88E-08      | 7.84E-08         |
|       | Marker411469 | 183      | T/C     | 1.58E-07      | 5.96E-08         |
|       | Marker591100 | 241      | G/C     | 1.57E-07      | 9.35E-08         |
|       | Marker394053 | 201      | G/A     | 1.57E-07      | 8.36E-08         |
|       | Marker1560613 | 202     | T/C     | 1.03E-07      |                  |
|       | Marker373316 | 73       | C/T     | 1.20E-07      |                  |
|       | Marker475399 | 42       | C/T     | 1.38E-07      |                  |
|       | Marker607165 | 200      | T/C     | 1.68E-07      |                  |
|       | Marker611740 | 8        | G/T     | 1.91E-07      |                  |
|       | Marker335582 | 178      | T/C     | 1.04E-07      |                  |
|       | Marker613704 | 194      | C/T     | 1.47E-07      |                  |
|       | Marker179037 | 157      | G/C     | 1.23E-07      |                  |
|       | Marker103424 | 115      | C/T     | 1.31E-07      |                  |
|       | Marker188304 | 115      | G/A     | 1.71E-07      |                  |

Figure 4. The manhattan plots and Q–Q plots of traits RYC’ and VW using mrMLM. The horizontal dotted red lines indicate the suggestive thresholds.
have provided research workers novel insight into \textit{P. massoniana} FusA super family, pepsin_retropepsin_like super family, ribokinase_pfkB_like super family and rve super family. Table S5). The genes were involved in RNase_H_like super family, RT_like super family, RVT_2 super family, SNPs, 26 SLAF sequences were directly located on the conserved domain of functional genes (see Supplementary database, most results were located on the non-coding regions in genomic DNA. Among all the associated database on the website. We screened out the SNPs located SLAF sequences. After BLASTN analysis with pub-

development of next-generation sequencing, sequencing technologies, such as GBS, RAD-seq, and SLAF-seq, development of SNP markers using SLAF-seq can currently meet the needs of GWAS in masson pine19. In this study, 

SLAFs for SNP detection. The genetic structure of diverse masson pine accessions was estimated using 471,660 149 masson pine accessions collected from different regions of Guangdong in China were employed to develop 

13. Traditional breeding methods are usually inefficient for forestry species. In recent years, genomic data 

ers that facilitate selection of trees with high yields of resin and wood will have a major impact on masson pine breeding10. With the development of next-generation sequencing, sequencing technologies, such as GBS, RAD-seq, and SLAF-seq, are now available for the identification of abundant SNPs in a wide range of plant species17,22. Therefore, development of SNP markers using SLAF-seq can currently meet the needs of GWAS in masson pine19. In this study, 

149 masson pine accessions collected from different regions of Guangdong in China were employed to develop SLAFs for SNP detection. The genetic structure of diverse masson pine accessions was estimated using 471,660 SNPs. Long-term selection gain of forestry trees requires large numbers of resources with genetic variability. Therefore, the examination of the population structure and genetic diversity are both important for the a breeding program13. In this study, we received different population numbers using cross-validation, delta K analysis and fastStructure analysis. The cross-validation support the result of K ≥ 3) with trait RYC', VW, HT, DBH and RW by at least two methods (see Supplementary Table S4).

**Discussion**

It takes a long time to evaluate the growth and economic traits in a conventional breeding program, and markers that facilitate selection of trees with high yields of resin and wood will have a major impact on masson pine breeding10. Traditional breeding methods are usually inefficient for forestry species. In recent years, genomic data have provided research workers novel insight into \textit{P. massoniana} genetic diversity and evolution30,31,32. Until now, there is no available public \textit{P. massoniana} genome database on the website. We screened out the SNPs located SLAF sequences. After BLASTN analysis with public database, most results were located on the non-coding regions in genomic DNA. Among all the associated SNPs, 26 SLAF sequences were directly located on the conserved domain of functional genes (see Supplementary Table S5). The genes were involved in RNase_H_like super family, RT_like super family, RVT_2 super family, FusA super family, pepsin_retropepsin_like super family, ribokinase_pfkB_like super family and rve super family.

**Table 3. RYC’ and VW associated SNPs using mrMLM.**

| Trait | Marker | Position | Alleles | mrMLM | P-value | FASTmrMLM | P-value |
|-------|--------|----------|---------|--------|---------|-----------|---------|
| RYC'  | Marker174624 | 234 | G/T | | 1.14E-10 | | 1.14E-10 |
|       | Marker482425 | 206 | G/T | | 3.06E-10 | | 3.06E-10 |
|       | Marker279561 | 86 | G/A | | 1.57E-09 | | 1.57E-09 |
|       | Marker370341 | 258 | G/T | | 2.46E-09 | | 2.46E-09 |
|       | Marker504006 | 117 | G/G | | 4.04E-09 | | 4.04E-09 |
|       | Marker124737 | 43 | G/T | | 5.79E-09 | | 5.79E-09 |
|       | Marker271387 | 80 | A/G | | 6.79E-09 | | 6.79E-09 |
|       | Marker283415 | 234 | G/T | | 1.42E-08 | | 1.42E-08 |
|       | Marker248105 | 192 | G/T | | 8.77E-08 | | 8.77E-08 |
|       | Marker278935 | 7 | G/T | | 1.01E-07 | | 1.01E-07 |
|       | Marker278935 | 256 | A/C | | 1.31E-07 | | 1.31E-07 |
| VW    | Marker218315 | 18 | G/A | | 5.16E-10 | | 5.16E-10 |
|       | Marker163256 | 68 | G/A | | 8.3E-09 | | 8.3E-09 |
|       | Marker164392 | 258 | T/C | | 2.12E-08 | | 2.12E-08 |
|       | Marker103247 | 160 | T/C | | 2.55E-08 | | 2.55E-08 |
|       | Marker147979 | 68 | T/G | | 2.55E-08 | | 2.55E-08 |
|       | Marker167966 | 196 | A/G | | 3.74E-08 | | 3.74E-08 |
|       | Marker417759 | 73 | G/A | | 4.83E-08 | | 4.83E-08 |
|       | Marker211769 | 144 | G/A | | 1.13E-07 | | 1.13E-07 |
|       | Marker271496 | 74 | C/G | | 1.13E-07 | | 1.13E-07 |
|       | Marker308123 | 258 | T/G | | 1.52E-07 | | 1.52E-07 |
|       | Marker490183 | 52 | C/T | | 1.99E-07 | | 1.99E-07 |

**Gene identification of associated SNPs.** Until now, there is no available public \textit{P. massoniana} genome database on the website. We screened out the SNPs located SLAF sequences. After BLASTN analysis with public database, most results were located on the non-coding regions in genomic DNA. Among all the associated SNPs, 26 SLAF sequences were directly located on the conserved domain of functional genes (see Supplementary Table S5). The genes were involved in RNase_H_like super family, RT_like super family, RVT_2 super family, FusA super family, pepsin_retropepsin_like super family, ribokinase_pfkB_like super family and rve super family.
related. Guangdong is surrounded by numerous mountains and has an independent geographical environment. Thus, the gene exchanges of masson pines may be limited in the whole province. Furthermore, masson pine has a long period of cultivation history; the breeding work and provenance tests started in Guangdong province in the last century. The plus families of masson pine were planted across Guangdong province. The intermixed relationships among some masson pines collected from different regions may be induced by the cultivation history. By using clustering analysis, the masson pines were divided into different subgroups by the genetic distance, which meant that there are great differences among these germplasm resources. In future study, masson pines derived from other regions should be collected and compared to the current resources.

Molecular markers have been employed to develop core germplasm sets in multiple tree species, e.g. western white pine\textsuperscript{34}, olive\textsuperscript{28,35}, litchi\textsuperscript{36}, pear\textsuperscript{37}, and Chinese fir\textsuperscript{38} have been examined using SNPs developed by reduced-representation genome sequencing. A core set percentage of 20~30\% of the total collection was once suggested at a general scale of the population\textsuperscript{27}. The fixed size of the core set depends on the purpose of the study, and different kinds of plants require different sampling percentages\textsuperscript{39}. Long-term selection gain requires genetic

Figure 5. The manhattan plots and Q–Q plots of traits HT, DBH, RW and VW using mrMLM. The horizontal dotted red lines indicate the suggestive thresholds.
variability; thus, it is important to examine not only population structure but also genetic diversity. Across the 149 masson pine accessions examined in this study, we observed a mean genetic distance of 0.232, with a range from 0.008 to 0.292. Furthermore, genomic characterization revealed high genetic diversity within the 149 masson pine accessions; therefore, we decided to identify a core germplasm set to improve masson pine breeding efficiency. It is important and meaningful to select a fully representative germplasm set from a large masson pine collection. In this study, the core sets of wood and resin showed higher genetic distances than the total collection. In addition, the core set of wood showed a high level of genetic gain expectation (41.78%) for trait VW; the core set of resin showed a high level of genetic gain expectation (40.75%) for trait RYC. The core germplasm sets, for the purpose of improving resin and wood yield, were scientifically simplified resources that would be useful for masson pine breeding.

The GWAS analyses of complex traits in forestry conifer trees, especially conifer trees with large genomes, require an enormous density of SNP markers. The decay of LD over physical distances in a population determines the density of the marker coverage needed to perform a GWAS. The faster LD decays, the more markers are likely needed in GWAS analysis for complex traits. LD estimates in this study based on the specific length sequences indicated a very fast decay. Excavation of favorable markers is necessary for improving masson pine breeding efficiency using molecular assisted selection (MAS). GWAS offers increased opportunities for detecting susceptible loci for complex traits. Masson pine is an economic tree species for resin and wood. In the breeding project of masson pine, both resin and wood yields are important breeding targets. Therefore, discovering SNPs related to resin and wood producing capacity is important for improving masson pine breeding efficiency.

In the present study, we focused on the GWAS of quantitative traits, including growth traits and the resin and wood yield in masson pine. The phenotypes of complex traits often result from the combined actions of multiple genes and environmental factors, all of which can easily lead to lost heritability. Therefore, only those traits with high heritability can be stably detected. The traits in masson pines, especially RYC and VW, have been demonstrated to have high heritability. Furthermore, more extensive linkage disequilibrium has been found in conifer trees. In our study, the number of SNPs identified from 149 masson pine germplasm resources is large enough, and GWAS can be feasible in masson pine even though the genome may be generally large. RYC and VW are important traits for representing masson pine producing capacity and economic value and have an important value in breeding. In this study, five traits (HT, DBH, RW, VW, and RYC) were selected for GWAS analyses. All of the traits showed large phenotypic variation, supporting the suitability of GWAS for these traits. Thus, we presented GWAS analyses of these important traits in masson pine. In our study, the suggestive SNPs associated with traits RYC’ and VW were different from the SNPs identified using EMMAX and FaST-LMM. Only one common SNP (Marker279561) was developed in trait RYC’ and VW using these methods, which meant that different types of GWAS methods can provide complementary results with each other and provide us with more sufficient results. Moreover, no SNPs were developed in trait HT, DBH, RW and RYC using EMMAX and FaST-LMM, while a large number of SNPs were detected using multi-locus methods in mrMLM. The multi-locus GWAS methods in mrMLM.GUI provide more possibilities in detecting associated SNPs. Hence, a group of various types of GWAS methods should be applied in future studies.

High correlations between these traits were identified, and strong positive correlations existed among the traits DBH, RW, VW, and RYC (see Supplementary Table S6). The SNPs developed in trait RW were totally detected in trait RYC which meant that these SNPs have significance in the selection of high resin yield masson pines. However, the other trait did not show correlation ships, indicating that it is also necessary to develop additional SNPs at higher levels in the future. In recent years, MAS and genome selection (GS) have been the most popular methods in plant breeding. GWAS and GS can each compensate for the other’s deficiencies, and both approaches are likely to be useful in conifer breeding. The developed SNP markers in GWAS can be directly used for both MAS and GS, and both approaches are likely to be useful in conifer breeding. Genotyping based on reduced-representation genome sequencing (RRGS) has become popular in a wide range of plant species. The various types of RRGS methods, among which SLAF-seq is also widely used, have overcome the cost problem and have simplified the problem of identifying a large number of DNA markers in conifer species with large genomes as well as the large number of samples in the scientific research of forestry breeding. *P. massoniana* has not been completely genome sequenced. By using BLASTN with the public database and conserved domain search, several SNPs were located on the conserved domain in some unusual genes. The other SNPs were mainly distributed on the noncoding region of genome DNA. Further annotations and functional analysis of those SNPs are necessary. Future studies of masson pine should not merely focus on RRGS methods, a variety of methods such as exon capturing and comparative transcriptome sequencing should be also considered for detecting SNPs and functional genes. The SNPs developed from exon-seq and RNA-seq are usually distributed on the transcript sequences and has been successfully used in conifer species.

**Conclusion**

In this study, SLAF-seq technology was used to develop 471,660 filtered SNPs from 149 *P. massoniana* accessions in Guangdong. The population structure and genetic relationship analyses of these masson pines showed a chaotic genetic relationship but various genetic distances. We obtained core germplasm sets including 29 masson pine accessions for increasing wood and resin production, respectively. Multiple methods were used in GWAS of five traits and the results provided us different associated SNPs. The application of various GWAS methods can enrich the number of associated SNPs. The core germplasm resources and identified SNPs have meaningful application values in *P. massoniana* selection and breeding.
Materials and Methods

Experimental materials. A total of 149 masson pine accessions were selected for obtaining SNP markers (see Supplementary Table S1). The masson pines were collected from Boluo (BL), Chaoan (CA), Deqing (DQ), Dongyuan (DY), Gaozhou (GZ), Lianzhou (LZ), Xinyi (XY), Yingde (YD), and Yunan (YN) in Guangdong province in southern China; in the latitude 21°55′N – 23°87′N, longitude 110°47′E – 114°41′E, and at elevations from 35 m to 458 m. Those lines were planted in a masson pine seed orchard in 1989 by the grafting method. For each accession, 0.5 g of clean conifer needles was selected from each accession for further DNA extraction.

DNA extraction and SLAF-seq. Total masson pine genomic DNA was extracted using the DP320 DNA secure Plant Kit (TIANGEN China); the quality and quantity of DNA were then inspected using 0.8% gel electrophoresis. The quantified DNA was diluted to 20 μg·mL⁻¹ and was stored at −20°C before use. The masson pine genomic DNA was analyzed according to the SLAF-seq method19. To obtain evenly distributed SLAF tags and to avoid repetitive SLAF tags for maximum SLAF-seq efficiency, simulated restriction enzyme digestion was carried out in silicon. Sequencing libraries of each accession were constructed through digestion with the restriction enzymes EcoRV and Scal to obtain the SLAF tags, and *Oryza sativa* genome DNA was used as a control to assess the normal rate of enzyme digestion. A single nucleotide (A) overhang was added to the digested fragments using dATP at 37°C, and then duplex tag-labeled sequencing adapters were ligated to the A-tailed DNA with T4 DNA ligase. The PCR products were purified and pooled. The pooled samples were separated via electrophoresis on a 2% agarose gel. Fragments with indices and adapters from 264 to 414 bp were excised and purified. Finally, the purified gel product was sequenced using the Illumina HiSeq2500 system (Illumina, Inc., San Diego, CA, USA) at the Biomarker Technologies Corporation in Beijing.

Genotyping and quality control. After sequencing, reads with double ends were compared with similar sequences that could be labeled as candidate SLAFs to proceed with the next step. The SLAF tags were defined as the group with the most samples. The samples with the most tags were used as references, and GATK and SAMTOOLS were employed for SNP calling65,66. SNPs were removed if the integrity < 0.8 and minor allele frequency (MAF) ≤ 0.05. After these steps, the remaining SNPs were developed to calculate genetic structure, and the relationships were retained for genome-wide association study (GWAS).

Structure, phylogenetic and genetic kinship among accessions. SNPs were used to calculate pairwise kinship relationships among the 149 accessions by using SPAGeDi software62. Negative kinship values between two accessions indicate a poorer relationship than expected, and this was corrected to 063. ADMIIXTURE was employed to investigate population structure based on the maximum-likelihood method64. The predefined K, which indicates the number of groups in a population, varied from 1 to 10 in ADMIIXTURE models. Cross-validation, delta K and marginal likelihood against K were used to select the most probable value of K65. A phylogenetic tree based on the neighbor-joining method was constructed in MEGA 6.0 using the developed SNPs63. A PCA with Cluster software was used to cluster the masson pine population65. Genetic distance and population structure were used to develop an initial core germplasm set by CoreHunter software65. The results combined with phenotypic data VW and RYC were used to confirm the final core set.

Phenotypic data collection and analysis. Phenotype data, including height (HT), diameter at breast height (DBH), resin weight (RW), volume of wood (VW) and resin-yielding capacity (RYC), of 122 lines from 605 clone individuals were measured in 2010. RYC’ data were collected and calculated from 69 plus trees. Firstly, we collected the phenotype data of individuals. Then, the average value of individuals from the same accession was used as the final phenotype data. Resins were collected on sunny days from July to October using the narrow face system as described by COPPEN and HONE1. Trees were sampled once per day by removal of a sliver of wood from the stem without the application of a stimulant. VW and RYC were calculated by the formulas given below.

The VW of an individual was calculated as follows43:

\[
VW = 6.2341803 \times 10^{-5} \times DBH^{1.8551497} \times HT^{0.95682492},
\]

where: VW is the volume of wood from an individual tree; DBH is the diameter at breast height in meters, and HT is the height of the tree in meters.

The RYC of an individual was calculated as follows43:

\[
RYC = \frac{Wt}{(D \times Wd/C)},
\]

where: RYC is the resin-yielding capacity of an individual tree; Wt is the total weight of collected resin of a tree; D is the cutting time for resin tapping per tree; Wd is the total width of the narrow tapping face; and C is the circumference of the trunk where the bark was cut.

Association analysis. The GWAS analysis was performed by multiple methods, namely, the Mixed Linear Model (MLM) in TASSLE software67, Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) in FaST-LMM software68, Efficient Mixed-Model Association eXpedited (EMMAX) in EMMAX software61 and six methods, including multi-locus random effect mixed linear model (mrMLM)62, fast Multi-locus random effect mixed linear model (FASTmrMLM)63, fast multi-locus random-SNP-effect EMMAX (FASTmrEMMA)64, iterative modified-set sure independence screening Expectation-Maximization-Bayesian least absolute shrinkage and selection operator (ISIS EM-BLASSO)65, polygenic-background-control-based Kruskal-Wallis test with empirical Bayes (pKWmEB)66 and polygenic-background-control-based least angle regression plus empirical Bayes
Gene identification of associated SNPs. We found the SLAF sequences that the suggestive SNPs located on and used the DNA sequences as queries to conduct BLASTn with the public database. Meanwhile, the SLAF sequences were used to make a conserved domain database analysis using NCBI's Conserved Domain Database.

Data Availability
All of the data generated or analyzed during this study are included in this published article (and its Supplementary Information files).

References
1. Coppen, J. J. W. & Hone, G. A. Gun Naval Stores: Turpentine And Rosin From Pine Resin Ch. Non-wood forest products 2, 62 (Natural Resources Institute, Food and Agriculture Organization of the United Nations, 1995).
2. Wang, Z., Calderon, M. M. & Carandang, M. G. Effects of resin tapping on optimal rotation age of pine plantation. Journal of Forest Economics 11, 245–260, https://doi.org/10.1016/j.jfe.2005.10.001 (2006).
3. Shi, K. S., Li, Z. Y., Lin, F. M. & Zheng, R. The Development Of China's Forestry: Review And Prospects 110 (The Environmental Science Press of China, 1998).
4. Bai, T.-D., Xu, L.-A., Xu, M. & Wang, Z.-R. Characterization of masson pine (Pinus massoniana Lamb.) microsatellite DNA by 454 genome shotgun sequencing. Tree Genet. Genome. 10, 429–437, https://doi.org/10.1007/s11295-013-0684-y (2014).
5. Chen, W., Cao, M., Wang, Y., Zhou, Z. & Xu, L.-A. A genetic linkage map of Pinus massoniana based on SSR, SSR and ESTP markers. Silvae Genet. 63, 1–9, https://doi.org/10.1515/sig-2014-0001 (2014).
6. Feng, Y. H., Yang, Z.-Q., Wang, L., Luo, Q. F. & Li, H. G. Development and characterization of SSR markers from Pinus massoniana and their transferability to P. elliottii, P. caribaea and P. yunnanensis. Gen. Mol. Res. 13, 1508–1513, https://doi.org/10.4238/2014.March.12.2 (2014).
7. Feng, S. L., Li, Q. F., Li, D., Wang, Z. F. & Wang, D. P. Genetic diversity of Pinus massoniana revealed by RAPD markers. Silvae Genet. 52, 60–63, https://doi.org/10.1007/s10722-004-5730-z (2003).
8. Zhang, Y., Chu, D.-Y., Jin, G.-Q. & Zhou, Z.-C. Relationship between hybrid performance and genetic distances among Pinus massoniana clones based on ISSR molecular marker. For. Res. 23, 215–220, https://doi.org/10.3724/SP.J.1238.2010.00474 (2010).
9. Zhang, Y., Yang, Q., Zhou, Z. & Jin, G. Divergence among masson pine parents revealed by geographical origins and SSR markers and their relationships with progeny performance. New Forests 44, 341–355, https://doi.org/10.1007/s11056-012-9340-x (2013).
10. Liu, J. et al. An improved allele-specific PCR primer design method for SNP marker analysis and its application. Plant Methods 8, 34, https://doi.org/10.1186/1746-4811-8-34 (2012).
11. Davey, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12, 499–510, https://doi.org/10.1038/nrg3012 (2011).
12. Jiménez-Geonet, J. M. Next generation quantitative genetics in plants. Frontiers in Plant Science 2, 77, https://doi.org/10.3389/fpls.2011.00077 (2011).
13. Uchiyama, K. et al. Demonstration of genome-wide association studies for identifying markers for wood property and male strobili traits in Cryptomeria japonica. PLoS One 8, e79866, https://doi.org/10.1371/journal.pone.0079866 (2013).
14. Neale, D. B. et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15, R93, https://doi.org/10.1186/gb-2014-15-3-r93 (2014).
15. Van Tassell, C. P. et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods 5, 474–478, https://doi.org/10.1038/nmeth.1105 (2008).
16. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 248–252, https://doi.org/10.1101/gr.56871.2007 (2007).
17. Zhou, L. et al. Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. Sci. Rep. 5, 9350, https://doi.org/10.1038/srep09350 (2015).
18. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J.-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7, e32253, https://doi.org/10.1371/journal.pone.0032253 (2012).
19. Sun, X. et al. SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. PLoS One 8, e58700, https://doi.org/10.1371/journal.pone.0058700 (2013).
20. Ye, Y. et al. Identification and validation of SNP markers linked to dwarf traits using SLAF-seq technology in Lagerstroemia. PLoS One 11, e0158970, https://doi.org/10.1371/journal.pone.0158970 (2016).
21. Liang, D. et al. QTL mapping by SLAF-seq and expression analysis of candidate genes for aphid resistance in cucumber. Frontiers in Plant Science 7, 1000, https://doi.org/10.3389/fpls.2016.01000 (2016).
22. Zhou, Q. et al. Genome-wide SNP markers based on SLAF-seq uncover breeding traces in rapeseed (Brassica napus L.). Frontiers in Plant Science 8, 648, https://doi.org/10.3389/fpls.2017.00648 (2017).
23. Mei, H. et al. High-density genetic map construction and gene mapping of basal branching habit and flowers per leaf axil in sesame. Frontiers in Plant Science 8, 636, https://doi.org/10.3389/fpls.2017.00636 (2017).
24. Zhang, D. et al. High-density genetic mapping identifies new major loci for tolerance to low-phosphorus stress in soybean. Frontiers in Plant Science 7, 572, https://doi.org/10.3389/fpls.2016.00572 (2016).
25. Su, J. et al. Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in upland cotton. BMC Genomics 17, 687, https://doi.org/10.1186/s12864-016-2875-z (2016).
26. Xie, D. et al. Genome-wide association study identifying candidate genes influencing important agronomic traits of flix (Lunum usitatissimum L.) using SLAF-seq. Frontiers in Plant Science 7, 2232, https://doi.org/10.3389/fpls.2017.02232 (2018).
27. Su, W. et al. Genome-wide assessment of population structure and genetic diversity and development of a core germplasm set for sweet potato based on specific length amplified fragment (SLAF) sequencing. PLoS One 12, e0172066, https://doi.org/10.1371/journal.pone.0172066 (2017).
66. Ren, W.-L., Wen, Y.-J., Dunwell, J. M. & Zhang, Y.-M. pKWmEB: integration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* **120**, 208–218, https://doi.org/10.1038/s41437-017-0007-4 (2018).

67. Zhang, J. *et al.* pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* **118**, 517–524, https://doi.org/10.1038/hdy.2017.8 (2017).

68. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *NAR* **45**, D200–D203, https://doi.org/10.1093/nar/gkw1129 (2017).

**Acknowledgements**

This work was jointly supported by the Planning Subject of Guangdong Science and Technology Project (No. 2017B020232001), National Natural Science Foundation of China (No. 31800553), Forestry Science and Technology Innovation Program of Guangdong Province (No. 2017KJCX019) and ‘the 13th Five-Year-Plan’ in National Science and Technology for the Rural Development in China (No. 2017YFD0600303-3). We also thank for the Open Fund of Guangdong Key Laboratory for Innovative Development and Utilization of Forest Plant Germplasm, South China Agricultural University.

**Author Contributions**

Qingsong Bai carried out the experiments, data analyses, drafted the manuscript and participated in the project design; Qian Zhang chiefly designed the project, supervised the research and reviewed the manuscript; Yanling Cai and Boxiang He participated in the project design and data analyses, Wanchuan Liu and Qingyou Pan collected the phenotypic data; All authors have read and approved the manuscript.

**Additional Information**

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-49737-2.

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019