A Transcriptomic Variation Map Provides Insights into The Genetic Basis of Pinus massoniana Evolution and Association of Oleoresin Yield

CURRENT STATUS: UNDER REVIEW

Qinghua Liu
Research Institute of Subtropical Forestry Chinese Academy of Forestry

Yini Xie
Research Institute of Subtropical Forestry Chinese Academy of Forestry

Bin Liu
Research Institute of Subtropical Forestry Chinese Academy of Forestry

Huanhuan Yin
Sichuan Agricultural University

Zhichun Zhou
zczhou_risf@163.com
Corresponding Author
ORCiD: https://orcid.org/0000-0001-7085-3509

Zhongping Feng
Laoshan Forest Farm of Chunan County

Yadong Chen
Biomarker Technologies Corporation

DOI:
10.21203/rs.3.rs-19278/v1

SUBJECT AREAS
Plant Physiology and Morphology  Plant Molecular Biology and Genetics

KEYWORDS
SNP, genetic diversity, population structure, geographic origin, associative transcriptomics, Pinus massoniana
Abstract

Background: Masson pine (Pinus massoniana Lamb.), the dominant native coniferous species in South China, is commercially important for timber and oleoresin. However, knowledge of the genetic variability of the masson pine germplasm is still scarce. Here, the genetic diversity and population structure of masson pine germplasm were assessed using 204 wild accessions from 10 main distribution regions by 94,194 core SNPs obtained from transcriptome sequencing data.

Results: The average expected heterozygosity was 0.2724, implying abundant genetic diversity within masson pine germplasm. Analysis of molecular variance (AMOVA) revealed that 3.29% variation sourced from the genetic differentiation. Structure analysis identified two geographically distinct groups. Discriminant analysis of principal components (DAPC) showed one geographically distinct group was divided two clusters furtherly. Sichuan and Chongqing provenance is the geographical origin, and diffuses outward along two lines. Oleoresin yield might be reflected by the evolution, and exhibits two different changing trends between two diffusing line. It may be associated with the genes of chitinase, CYP720B, Cytochrome P450, ABC transporter and AP2/ERF based on SNPs and expression.

Conclusions: SNP markers by transcriptome sequencing have a strong power in evaluating genetic diversity within species and the genetic control of objective trait. The function of these genes will be verified and the strong associated genes with oleoresin yield will be used in the improvement of oleoresin yield by early genotype selection or genetic engineering.

Background

As a dominant native tree species, masson pine (Pinus massoniana Lamb.) is commercially important conifer for timber and oleoresin in China. The natural distribution extends from 21°41′N to 33°56′N and 102°10′E to 123°14′E with the planting area of 2 million hectares [1]. The provinces of Guangdong, Guangxi, Hu’nan, Sichuan, Chongqing, Guizhou, Zhejiang, Fujian and Jiangxi are the main natural distribution regions of masson pine in China [2]. Because masson pine has the characteristics of fast growth and tolerance to barren soil, it is often considered as a pioneer species for afforestation in bare mountain fields. Genetic diversity is critical for long-term survival species, which made the
species adapt to various abiotic and biotic stress and avoid extinction [3]. Large genetic variation can be observed among and within the natural populations based on provenance or family analysis for most trees species in growth, terpenoid, resistance, etc. [4-6]. To learn the genetic variation on main economic traits of masson pine, large-scale provenance experimentations have been carried in China since 1978. Two whole native range provenance trails and many partial native range provenance trails were built in China [2], which provided good materials to reveal the interplay and significance of several evolutionary forces causing phenotype diversity, and formulate gene conservation strategy capturing the natural genetic diversity within species. A classical geographical variation pattern in latitude was found for diameter at breast height (DBH) in masson pine [7].

As a secondary substances of masson pine, oleoresin is an important natural source in chemical industry [8, 9], defending against insects and disease [10, 11]. It also can be exploited as advanced liquid biofuels [12]. Significant genetic variation in oleoresin yield was also observed among different families of masson pine, which ranged from 14.12 and 50.55 g per day [13, 14] also reported that variation in oleoresin yield was heritable in P. taeda and could be increased 1.5- to 2.4-fold in one generation through selection.

Molecular markers are very useful in identifying germplasm, assessing biodiversity and describing the geographic patterns of genetic variation. Single nucleotide polymorphisms (SNPs) are commonly used in genetic study. Taking advantage of next generation sequencing (NGS) technologies, we could rapidly develop millions of SNPs for crops with low cost [9]. The high-throughput SNPs have been successfully used to evaluate genetic diversity, to infer population structure [15, 16] and kinships [17]. As an important forest tree in South China, high density SNPs map is essential for genetic innovation and traits improvements in masson pine’s future breeding. However, there are no reports about the whole genome sequences and developing SNP markers to study the genetic diversity and structure for masson pine. To date, only partial masson pine germplasms have been analyzed using random amplification polymorphic DNA (RAPD) [18], inter-simple sequence repeat (ISSR) [19], simple sequence repeat (SSR) [20] and inter-retrotransposon amplified polymorphism (IRAP) [21].

Associative transcriptomics has largely contributed to identify sequence polymorphisms and
transcript abundance linked to phenotypic variation, especially for non-model species [22, 23]. In addition, high quality full length transcripts are critical and useful for functional assays and understanding of genetic diversity [24]. In this work, we firstly constructed high quality transcript reference sequences through combination of full-length transcriptome and NGS-based unigenes. Then RNA-Seq of 204 representative accessions was adopted for de novo SNP discovery to generate a genome-wide variation map. The aims of our study were to 1) assess the genetic diversity, population structure and geographic origin of masson pine; 2) reveal the genes associated with oleoresin yield. The results would be useful for managing this species and expounding the forming mechanism of high-yielding oleoresin.

Results
Sequencing and variation discovery

The research and breeding program of masson pine have been hampered by lacking of high quality genome sequences, because of its extremely large and complex genome presumed by close species loblolly pine (P. taeda) [25]. To overcome this obstacle, we constructed a high quality full length transcripts data set from secondary xylem transcriptome using PacBio Single Molecule, Real-Time (SMRT) Sequencing platform. A total of 81,837 high-quality and non-redundant full-length transcripts were obtained from 18 Gb PacBio subreads. To explore the origins and patterns of genetic diversity, we also designed the population transcriptome experiments for 204 geographically diverse masson pine genotypes, which collected from main habitats in China. Totaling of 341,714 non-redundant unigenes were assembled. After combination of full length transcripts and unigenes, 423,288 non-redundant transcripts considered as reference sequences for further analysis (Additional file 1: Table S1). On average, 85.02% of the reads for each sample were successfully mapped to the reference sequences, suggesting the high completeness of reference transcripts (Additional file 2: Table S2). A total of 1,326,230 single nucleotide polymorphisms (SNPs) and 153,459 insertions/deletions (InDels) were detected from transcriptomes of 204 genotypes using GATK packages [26] with an average SNP density of 3.13 per transcript (Additional file 3: Table S3). Among these SNPs, 94,194 core SNPs with minor allele frequency (MAF) ≥ 0.05 and missing genotype calls <5% were remained for further
analysis, occupying 7.1 % of total set. These core SNPs included 23,864 (25.33%) non-synonymous (nSNPs) (Additional file 4: Table S4). This transcriptome variation map will benefit core germplasm identification, genetic variation research and artificial breeding.

Genetic diversity of Masson pine

The genetic diversity among *P. massoniana* germplasms from mainly distribution regions was investigated based on 94,194 SNPs. The observed heterozygosity (*H*₀) value was obviously lower than expected heterozygosity (*H*ₑ) value for each population, ranging from 0.2211 (Guangxi) to 0.2358 (Sichuan and Chongqing). The *H*ₑ values were similar among the different populations, and ranged from 0.3011 (Jiangxi) to 0.3124 (Sichuan and Chongqing) (Table 1). The values of inbreeding coefficient (F index) ranged from 0.2242 (Sichuan and Chongqing) to 0.2714 (Guangxi), with a value overall the population equal to 0.2731, indicating SNPs in Sichuan and Chongqing population have highest polymorphic. Putative differences among the nine populations were tested by AMOVA based on 94,194 SNPs (Table 2). The results showed that the differentiation among populations explained by 3.29% of the total variance. Only 0.01% of the variation was found among different subpopulations, suggesting the closed kindship within different subpopulations. In summary, our variant data set provides a comprehensive overview of genomic diversity at variously populational scales, and represents a rich source of genetic information for exploitation by both the academic and agricultural research communities.

Construction of *P. massoniana* core germplasm

The allelic diversity among *P. massoniana* accessions can be maximized by SNP markers. The redundancy curve showed that the masson pine allelic diversity could be represented by more than 40 core germplasms. These 40 representative genotypes which accounted for only 20% collection could represent more than 90.7% of the allelic diversity (Fig. 1 and Table 3). Therefore, the minimum size of the core germplasm could constructed by these 40 representative accessions, including 9 accessions from Zhejiang, 9 accessions from Guizhou and 7 accessions from Sichuan and Chongqing (Additional file 5: Table S5). To our knowledge, this is first comprehensive core germplasms
identification based on high density SNPs map in large population scale, which is valuable for *P. massoniana* breeding practices.

Population structure of *P. massoniana* germplasm

To further understand the evolutionary history of masson pine, we used ADMIXTURE [27] to estimate ancestry proportions for each accession. Genetic assignment analysis showed an optimal value of $K = 2$, which clearly separated the accessions of Chongqing and Sichuan from other wild genotypes (Additional file 6: Fig. S1). The first group which mainly included the clones from Chongqing and Sichuan provinces has high level signal of inter population admixture (Fig.2A). For $K = 3$, two new subpopulations from Central South China and Southeast China, respectively, arose from the accessions out of Chongqing and Sichuan. Notably, the Group I which included the clones from Chongqing and Sichuan provinces also showed the high levels of admixture. The Group II which contained major clones from Central South China including Guizhou, Guangxi, Guangdong and Hunan, showed introgression signal from Group I. It is possibly contributed by natural hybridizations through animal or wind after separation. The clones from southeast of China, including Fujian, Jiangxi, Zhejiang and Anhui provinces, were assigned into the Group III. Interestingly, Group III kept their homogeneous genetic background probably due to their geographical isolation that blocks interspecific hybridization (Fig. 2B). As expected, the population structure of masson pine genotypes is consistent with their geographical distributions.

A discriminant analysis of principal components (DAPC) revealed three genetic clusters driving the partitioning of diversity within our panel (Fig. 2C). Cluster I comprised only accessions from Sichuan and Chongqing (94.4%) (Additional file 7: Table S6); Cluster II included mainly clones from Jiangxi (100%), Fujian (100%), Zhejiang (100%) and Anhui (100%), which contributed 88.3% to Cluster II. Cluster III included major accessions from Guizhou (100%), Guangxi (96.6%), Guangdong (61.1%) and Hunan (88.2%). The three genetic clusters have obviously geographical isolation. Cluster I consist of accessions mainly lived in the West China. The accessions of Cluster II were mainly distributed in the Southeast China. Cluster III included accessions obtained from Central South China. The result of DAPC analysis was consistent with population structure analysis with $K=3$. 
We further estimated the genetic diversity of different clusters. The genome-wide nucleotide diversity ($\pi$) of Cluster I ($2.91 \times 10^{-2}$) was higher than those of Cluster II ($2.77 \times 10^{-2}$) and Cluster III ($2.83 \times 10^{-2}$), exhibited the highest diversity level. This result is also supported by $H_\text{e}$ value, which revealed the sequence diversity based on heterzygous sites (Table 4). The Nei’s genetic distance showed the values ranging from 0.135 (Cluster II vs. Cluster III) to 0.303 (Cluster I vs. Cluster II), while the pairwise $F_{\text{st}}$ ranged from 0.024 (Cluster II vs. Cluster III) to 0.110 (Cluster I vs. Cluster II) and the Nei’s and $F_{\text{st}}$ genetic distance of Cluster I vs. Cluster II was higher than that of Cluster I vs. Cluster III (Additional file 8: Table S7). These observations suggest that masson pine of sichuang basin has been maintained high genetic diversity, and has larger differentiation with that of Southeast China.

Geographical origin and diffusion of *P. massoniana* germplasm

To further elucidate the evolution map and spread pathway, we examined the phylogeny of 204 masson pine genotypes by building a neighbor joining phylogenetic tree (Fig. 2D). Meanwhile, *P. taeda* was assigned as outgroup of a maximum likelihood tree to identify the earliest diverged population, considering as progenitors of modern *P. massoniana* (Additional file 9: Fig. S2). The phylogenetic tree showed that the genotypes from Sichuan Basin (Sichuang, Chongqing) was closest to *P. taeda* and followed by other clades, suggesting that Sichuan Basin is the geographic origin of masson pine. Sichuan Basin was one of glacial refugees for many species at the last Pleistocene glaciations [28, 29], which may rescue the masson pine from extinction event. The masson pine gradually migrated to Guizhou plateau after end of glacial epoch and gradually adapted the plateau habitat. The genotypes of Hunan formed a subclade from a branch of Guizhou clade and followed by other masson pine genotypes of Central South China (Guangdong and Guangxi) and Southeast China (Jiangxi, Fujian, Zhejiang and Anhui). Notably, these genotypes clearly split into two subclades according their geographical distribution. This observation allowed us to propose a hypothesis of two orientation spreading lines in masson pine evolution map (Fig. 3A). One migration line is from Sichuan and Chongqing to Guizhou, to Hunan and then spread into Guangdong and Guangxi. The other line is from Sichuan and Chongqing to Guizhou, to Hunan, and then spread into Jiangxi, Fujian, Anhui and
Zhejiang. This hypothesis strongly supported by population structure evidence (Fig. 3B). The population differentiation is significantly larger between Sichuang/Chongqing and Guizhou \( (F_{st} = 0.13) \) than those of other populations, implying the strong genome variation for new natural adaption when firstly transfer the habitat from basin to plateau. The neucleotide diversity is slightly higher in progenitors of Sichuang/Chongqing population (Fig. 3A). Signals of introgression were detected between the populations for two dissemination lines by the TreeMix program. The hybridization signal from Sichuan Basin population to Guangdong/Guangxi population was detected (Fig.3C). The introgression might be mainly mediated by human activity.

Associative transcriptomics with oleoresin yield
The oleoresin yield in xylem varied substantially in the 204 clones of masson pine, its levels varying from 0.00 to 6.07 g·cm\(^{-1}\)·d\(^{-1}\) (Additional file 10: Table S8). The oleoresin yield appeared to be positive skew distribution among the accession (Fig. 4A). The oleoresin yield of accessions from Sichuan basin and Guizhou is significantly lower than that of Hunan accessions (Fig. 4B). In Central South China spreading path, oleoresin yield is slightly reducing when masson pine spreaded into Guangdong and Guangxi, but is still higher than that of accessions from Sichuan basin and Guizhou. In Southeast China spreading path, the oleoresin yield significantly increased when masson pine spreaded into the Southeast China, especial for Anhui, Zhejiang and Jiangxi.

Associative transcriptomics analysis identified 121 SNPs from 109 transcripts that were significantly associated with oleoresin yield at the significance level of \( P < 10^{-6} \) (Fig.4C, Additional file 11: Fig. S3 and Additional file 12: Table S9). The most significant SNP (c51955_f1p3_1546, \( R^2 = 0.51, P = 3.74E-19 \)) is localized in the transcript annotated as chitinase class I (Table 5). The mutated SNP took place the upstream of coding region, but the expression of the transcript (c51955_f1p3_1546) was not significantly correlated with oleoresin yield.

The family of CYP720B belonging to cytochrome P450 monooxygenases (P450), are an important enzyme involved in the biosynthesis of diterpene resin acids as the main content of oleoresin [30]. One CYP720B (c19795_f1p0_1763) and one Cytochrome P450 (c9591_f1p0_1663) were found
sequence associated with the oleoresin yield (9.60E-07, 1.85E-07). The mutated SNP of CYP720B led to non-synonymous mutations with the transition of codon CTC to TTC. The mutated SNP from Cytochrome P450 (c9591_f1p0_1663) belong to synonymous mutation. The expression of these two transcripts (c19795_f1p0_1763, c9591_f1p0_1663) was significantly correlated with oleoresin yield (P = 3.61E-08, P = 2.13E-08). The result of qRT-PCR using the high- and low-yielding oleoresin accessions showed the higher expression level for these two transcripts in high-yielding oleoresin masson pines (Fig.4D).

The sequences of AP2 domain transcription factor and ABC transporter were associated with oleoresin yield in P. teada [14]. In this study, two SNPs from AP2/Ethylene-responsive transcription factors (ERFs) (c24091_f1p1_1286, c8825_f1p0_1733) and one SNP from ABC transporter (c189021.graph_c0) was also found significantly associated with oleoresin yield in masson pine. The SNP from AP2/ERF (c8825_f1p0_1733) caused non-synonymous coding and the coding amino acid changing from cystine to arginine.

In addition, one SNP from the transcript of tubulin alpha chain (c20772_f1p4_1467) was significantly associated with oleoresin yield in sequence (P = 8.73E-16) and expression level (P = 4.83E-08) simultaneously. The SNP caused non-synonymous mutation with different codon CTC and TTC.

However, the function of tubulin alpha chain is unclear during the biosynthesis of oleoresin.

Discussion

The SNP markers have been used to evaluate diversity within many species, such as Populus trichocarpa [31], Vitis vinifera [32], Ginkgo biloba [33]. SNPs called from transcriptome sequencing is a more efficient strategy for characterizing diversity in non-model or massive-genome species, since the sequences are detected on the coding regions rather than the whole genome. In this study, 94,194 SNPs obtained by transcriptome sequencing were used to investigate the diversity of masson pine from 10 provinces or municipality. H_0 values (of approximately 0.22) were slightly lower than H_e values (0.30 across populations), suggesting that frequent inbreeding events happened within the populations (Table 1). Either H_e or H_0 can be used to assess genetic variation, but the H_0 value is often influenced by the level of inbreeding within population. Therefore, H_e is commonly more applied
for comparing genetic diversity among different species or population within the same species [3].
Masson pine has continuous distributions in larger native region of China. Hamrick et al. [34] found that the average expected heterozygosity within populations of tree species with widespread distributions was 0.228 using allozyme analyses. Huang and Zhang [35] reported that the $H_e$ value was 0.27 by isozyme analysis on six natural population of masson pine in Guizhou province. The genetic diversity of five population of masson pine in Fujian province was assessed and found that the average $H_e$ was 0.22 [36]. In this study, higher genetic variability detected might be attributed to the wider sampling regions, which almost over whole native region of masson pine. Similar consequence was observed for natural populations of Scots pine (P. sylvestris) [37, 38]. In addition, the difference of genetic diversity assessed in these studies could be also caused by the different marker types, sampling locations and sizes [39].
Both structure analysis and DAPC obviously separated Sichuan and Chongqing samples from the others (Fig. 2A, C). This differentiation was also in agreement with the results of $F_{st}$ and Nei’s genetic distance values (Additional file 9: Table S7), which revealed the germplasm from Sichuan and Chongqing had the highest values for $F_{st}$ and Nei’s genetic distance by DAPC, respectively. Although the structure analysis showed the minimum cross-validation error at $K = 2$, the cross-validation error at $K = 3$ was only slightly higher than that at $K = 2$. When $K = 2$, most of the germplasm from the other provinces not including Sichuan and Chongqing, were grouped one cluster (Fig. 2A). However, this cluster were divided into two groups at $K = 3$, which was strongly correspondence with the clusters of DAPC, despite with minor differences in the member of each cluster. The differentiation between Cluster II and Cluster III was relatively small with the $F_{st}$ and Nei’s genetic distance values of 0.024 and 0.135 respectively, suggesting the masson pine germplasm from South-central and Southeast China have a closer relationship than that from Southwest China. In addition, compared to the other two clusters, Cluster I composed of Sichuan and Chongqing germplasm located at the Southwest China had the highest genetic diversity with the $H_e$ of 0.318.
Climate is one of main effecting factors for adaptive evolution of forest trees [40, 41].
Hemisphere, warm subtropical and temperate climates with rich gymnosperms in the Eocene turned into cold and strong seasonal climates from the Oligocene onwards over the Cenozoic in the middle-latitude and high-latitude landmass, especially in the Quaternary with large-scale ice cover and glaciations [42–44]. Many tree species were extinction suffer severe cold during this period. However, some tree species better adapted cooler conditions sustained. In southern China, complex topography made numerous temperate forests survived the last glacial maximum at the "refugia" [45], such as Ginkgo biloba, Metasequoia glyptostroboides, Glyptostrobus pensilis and Liriodendron chinese, of which these species are still alive up to now in China. Sichuan Basin including central and eastern parts of Sichuan province and Chongqing Municipality was surrounded by Tibet Plateau, Dabashan Mountain, Wushan Mountain and the Yunnan-Guizhou Plateau (1000–3000 m a.s.l). But elevation of the bottom of Sichuan Basin was only ranging from 250 m to 750 m. Although the glaciations were also occurred in Sichuan Basin during the Quaternary, the cooler climate could not lead the plants and animals species extinct, such as surviving Stegodon-panda (Ailuropoda fauna). Therefore, the results suggested that Sichuan Basin is one main refugium for many species [46]. These species were expanding to lower elevations in the glacial periods, and reteating the refugia at higher elevations during the interglacial stages [45, 47, 48].

To explore the evolutionary history of masson pine, we use loblolly pine (P. taeda) as the reference. It expounded that masson pine from Sichuan Basin was the geographic origin. This is concordant with previous reported by Qin [46] through observing the characteristics of the needles. However, structure analysis showed two geographically distinct groups and DAPC identified three clusters in this study, which suggested the genes have been changing to adapt the habitat.

Masson pine spread to Guizhou province firstly from Sichuan Basin. Guizhou province is neighboring with Guangxi and Hunan provinces, but masson pine only spread to Hunan from Guizhou, subsequently. It might that Yunnan-Guizhou Plateau was the barrier hindering the spread of masson pine from Guizhou to Guangxi. Although Guangxi provenances were not highly distinguished using structure, DAPC and cladogram with other several provenances, the difference was significant between Guangxi and Guangdong provenances and the other provenances for growth traits. The
growth of Guangxi and Guangdong provenances was faster, which is related to the thermal resources of origins [7].

The breeding can be accelerated by selecting the genes related to target trait. NpABC1 was reported the first transporter involved in the secretion of terpenoids in soybean [49]. In conifers, oleoresin is transported from living cells to resin ducts and flowing upon the wounding in stems after stem suffering abiotic stimuli [50]. Westbrook et al. [14] found that SNPs located in ABC transporters were associated with oleoresin yield and inferred ABC transporters participate in oleoresin transportation. In this study, the results of SNPs also indicate that ABC transporters were significantly associated with oleoresin yield, which suggested that ABC transporters may play an important role in regulating the oleoresin yield through changing the sequences.

Chitinase play a key role in modifying cell wall structure. Zhong et al. [51] found the mutant of Chitinase (elp1) would lead lignin to be ectopically deposited in the stem of Arabidopsis, and walls of the lignified cells were not thickened. The function of chitinase might affect the transport rate of oleoresin from living cells to resin duct.

Most of Cytochrome P450s have been reported involving in the progress of secondary metabolism [52]. The CYP720B gene family of Cytochrome P450s is specific for conifer and can catalyze consecutive oxidation steps in the biosynthesis pathway of various diterpene resin acids as the main component of oleoresin [30]. We found that one SNP in CYP720B and one SNP in Cytochrome P450 were significantly associated with oleoresin yield, and the SNP in CYP720B led to nonsynonymous mutation. Therefore, the SNP in CYP720B was inferred have an important influence in determining oleoresin yield by changing sequence and expression level.

Ethylene can induce the biosynthesis and the formation of traumatic resin ducts in many conifers [53]. AP2/ERF transcription factors are involved in the regulation of ethylene-responsive gene expression in the ethylene signaling pathway during abiotic stress. Over expression of OsERE BP1 belonging to ERF family would cause increased expression of genes relating to lipid metabolism in rice [54]. In P. taeda, one SNP in AP2 domain transcription factor was also associated with oleoresin yield [14], which is verified by our results in masson pine.
Conclusion
It is important to understand the genetic architecture for improving the oleoresin yield in genetic
breeding process. Although the genome of masson pine has not been sequenced, we obtain
satisfactory result on genetic diversity, population structure and trait-gene association based on
94,194 SNPs using the full-length transcriptome as reference. Masson pine is clearly differentiated
two groups and Sichuan and Chongqing provenance is the geographical origin, then masson pine
diffuses outward along two lines. Oleoresin yield exhibits two different changing trends between two
diffusing line, and associated with the genes of chitinase, CYP720B, Cytochrome P450, ABC
transporter and AP2/ERF, among which some genes was also confirmed in other conifer. The function
of these genes will be verified in further work.

Methods
Sample collection
A clonal test of masson pine, located at Laoshan Forest Farm in the western of Zhejiang province,
China (119°02′ E,29°33′ N; altitude, 152 m above sea level), was used for this study. This trial
included 400 clones (genotypes) obtained from 10 provinces or municipality. In the 1980's, a national
technical cooperation group for masson pine was established in China, and the scions for these clones
were identified and provided by the every provincial technical cooperation group authorized by the
local government. Robust shoots as scions from wild trees were collected from upper crown of
masson pine in April 1985. Subsequently, the scions were grafted onto 2-year-old local seedlings of
masson pine using pith-cambium pairing grafting method carried by Research Institute of Subtropical
Forestry, Chinese Academy of Forestry, China. In next year, the clonal trail was established using
these grafted seedlings, and have a completely randomized design with 10 repetitions and 2.0 m ×
3.0 m spacing between trees. Experimental research on plants, including collection of plant material
have complied with institutional, national guidelines of China as well as international guidelines. Field
studies were conducted in accordance with local legislation. The authors comply with the Convention
on the Trade in Endangered Species of Wild Fauna and Flora.
In this study, 204 healthy clones of masson pine from 10 provinces or municipality have been
selected (Additional file 13: Table 10). Before 1997, Chongqing was a city of Sichuan province, we collected scions from Sichuan and Chongqing both belonging to Sichuan Basin. Therefore, the analysis would be carried considering the germplasm from Chongqing and Sichuan as a population.

Oleoresin yield was measured according Liu’s method [55] from May to October in 2017 and 2018. The oleoresin yield of each tree was calculated as the yield of the individual per day per cm streak length in grams. Simultaneously, 5 mm deep fresh secondary xylem tissues adjoining cambium were harvested from sample trees after removing the bark, phloem. These samples were put into liquid nitrogen immediately in field and then stored at -80 °C for RNA extraction. These experiments were under taken in Research Institute of Subtropical Forestry, Chinese Academy of Forestry, China.

RNA extraction and PacBio-based sequencing

Total RNA from each sample was separately extracted and evaluated according to Liu’s method [56]. Briefly, total RNA from each sample was extracted using the Plant RNA kit RN38 EASYSpin plus (Aidlab Biotech, Beijing, China). The concentration and integrity of the total RNA was detected using an Ultraspec TM 2100 Pro UV/visible spectrophotometer, and an Agilent 2100 Bioanalyzer. High-quality RNA samples were used to constructing cDNA libraries. One microgram RNA for each sample was equally pooled together and full-length cDNA was synthesized using the SMARTer™ PCR cDNA Synthesis Kit. The size of the full-length cDNAs were selected using Blue Pippin (SageScience, Beverly, MA, USA) and for three libraries of differently sized cDNA (1–2 kb, 2–3 kb, and > 3 kb) were build. Then, the size distribution of cDNA was quantified using Qubit fluorometer and the quality of three libraries was assessed using the Agilent Bioanalyzer 2100Bioanalyzer. Subsequently, SMRT sequencing was carried by the Pacific Biosciences RS II platform in Biomarker Technologies Corporation.

Next-generation sequencing

The mRNA was obtained from high-quality total RNA for each sample using magnetic beads enrichment procedure. Fragmentation Buffer was used to fragment mRNA randomly. The first- and second-strand cDNA were synthesized, respectively. All of cDNAs were purified using AMPure XP beads. After end repairing, adding A and adaptor ligation, the fragment size of purified cDNA was
selected using AMPure XP beads. Then, cDNA fragments were enriched by PCR amplification and the quality of cDNA library for each sample was assessed using Qubitfluorometer and the Agilent Bioanalyzer 2100 Bioanalyzer. Finally, the qualified cDNA library was paired-end sequenced on the Illumina HiSeq™ 2000 sequencing platform.

Quality control of RNAseq data

Low quality reads were filtered based on the following four rules: 1) If one end of a pair-end read had > 5% "N" bases, then the pair-end read was removed; 2) For each pair-end read, if one of them had an average base-quality less than 20, then they were removed; 3) For each read, we trimmed its 3’ bases if their quality scores are less than 13. The trimming was stopped at the base with quality score ≥ 13. After trimming, if the number of remaining bases was less than 40, then the pair-end reads were removed; 4) Duplicates of pair-end reads were removed. Clean data were then used to call both SNPs and insertions/deletions (InDels).

SNP and InDel Calling

Filtered reads were then mapped to the reference sequences (full-length transcriptome) using the BWA-MEM algorithm of the Burrows Wheeler Aligner. SNPs were called using Haplotype Calller in GATK across the 204 samples of masson pine. Finally, low-quality SNPs (QUAL < 30, MQ < 40.0, FS > 60.0 and QD < 2.0) were removed. InDels were called using the same pipeline as SNP calling. Raw InDels were filtered to reduce the false positives by GTAK Variant Filtration with the parameters: FS > 200, QD < 2, Read Pos Rank Sum < -20.0.

Genetic diversity analysis

The genetic parameters of observed heterozygosity (H_0), expected heterozygosity (H_e), the minor allele frequency (MAF) and inbreeding coefficient (F) were estimated by PLINK software (version 1.9). Variation among populations, among clones within populations and within clones was calculated via analysis of molecular variance (AMOVA) using Arlequin software (version 3.5.2).

Phylogenetic analyses

For phylogenetic tree, the genome of loblolly pine was downloaded firstly from NCBI database (SRX4454630). Then the genome of loblolly pine was aligned with the full-length transcriptome
sequences. Subsequently, SNPs were called from genome of loblolly pine. Then, the phylogenetic tree visualizing and editing assigning were performed using ITOL (http://itol.embl.de/). The divergence time between masson pine and loblolly pine was obtained by the online Time Tree software (http://timetree.org/). Finally, the divergence time for each germplasm was calculate and visualized by the Mcmctree package in Paml [57].

Population structure analyses

ADMIXTURE software [27] was used to visualize population genetic structure. The preset ancestral population numbers was ranging from K = 1 to K = 10. The most likely number of ancestral genetic groups was determined by the minimum K value on the cross-validation curve.

Discriminant analysis of principal components (DAPC)

For DAPC, genetic data were first transformed into uncorrelated components using PCA. The number of genetic clusters was then defined using k-means, a clustering algorithm that looks for the value of K that maximizes the variation between groups. The Bayesian Information Criterion (BIC) was calculated for K = 1–40 and the K value with the lowest BIC was selected as the optimal number of clusters. A discriminant analysis was then performed on the first 120 principal components using the function DAPC to efficiently describe the genetic clusters.

Identification of genes associated with oleoresion yield

The association between SNPs and oleoresin yield was carried by the mixed linear model (MLM) using TASSEL [58]. The P-value corresponding to each association was calculated, and the association was significant with the P-value ≤ 1.06E-5, which was estimated using 1/n named Bonferroni correction (n is the number of SNP markers).

Quantitative RT-PCR analysis

Ten high- and low-oleoresin yielding RNA samples were used in qRT-PCR. The primer pairs (Additional file 13: Table S10) for seven genes of chitinase, tubulin alpha chain, AP2/ERF, ABC transporter, CYP720 and cytochrome P450 design and the cDNA was amplification according to Liu’s method [56], and the expression levels of genes were calculated with the 2^{-\Delta\Delta Ct} method [59]. Elongation factor 1-alpha (EF 1- alpha) was used to normalize the transcript profiles.
Abbreviations
AMOVA: Analysis of molecular variance; DAPC: Discriminant analysis of principal components;
SNP: Single nucleotide polymorphisms; NGS: Next generation sequencing; ERFs: Ethylene-responsive transcription factors; AP2: APETALA2 domain transcription factor; RT-PCR: Quantitative reverse transcription-polymerase chain reaction.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
All data generated or analyzed during this study are included in this published article (and its additional information files). Sequencing data is available through NCBI database (SAMN13566835).

Competing interests
The authors declare that they have no competing interests.

Funding
This study was supported by the (Grant No. 2017YFD0600300) grant. The funding agency had no role in the design of the study and collection, analysis, and interpretation of data.

Authors’ contributions
LQH contributed to the design of the work; the acquisition, analysis, and interpretation of data and the writing of the manuscript. XYN contributed to the acquisition, and analysis of data. LB contributed to the acquisition and analysis of data. YHH contributed to the acquisition and analysis of data. FZP contributed to the plant material collection. CYD contributed to the acquisition and analysis of data. ZZC contributed to the plant material collection, the conception and design of the work. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the China National Key R&D Program during the 13th Five-year Plan Period (Grant No.
2017YFD0600300) and Project supported by Major Special Project on Breeding New Forest Trees in Zhejiang Province of China during the Thirteenth Five-plan Period (2016C02056-4). We greatly thank Dr. Hengfu Yin for his suggestion in manuscript writing.

References

1. Zhou ZC, Li JM, Jin GQ. Predicting the impact of climate warming on growth of masson pine using provenance regional trials. Forest Research. 2000;13: 295-300.

2. Qin GF, Zhou ZC. Germplasm resources of Chinese masson pine. Beijing: China Forestry Publishing House. 2012.

3. White TL, Adams, WT, Neale DB. Within-population variation-genetic diversity, mating systems and stand structure. In: Forest Genetics. Cambridge: Oxfordshire CABI. 2007; p149-185.

4. Manninen AM, Tarhanen S, Vuorinen M, Kainulainen P. Comparing the variation of needle and wood terpenoids in Scots pine provenances. Journal of Chemical Ecology. 2002; 28:211-228.

5. Lesser M, Parker Genetic variation in *Picea glauca* for growth and phenological traits from provenance tests in Ontario. Silvae Genetica. 2004; 53:141-148.

6. Trias OQ, Paehler J. Genetic variation in growth and blister-rust resistance in a *Pinus strobus* x *wallichiana* hybrid population. Silvae Genetica. 2004; 53:33-41.

7. Liu QH, Jin GQ, Zhang R, Chu DY, Qin GF, Zhou ZC. Provenance variation for growth, stem-form and wood density and provenance division of masson pine at 24 years old. Scientia Silvae Sinicae. 2009; 45: 55-60.

8. Kelkar VM, Geils BW, Becker DR, Overby ST, Neary DG. How to recover more value from small pine trees: essential oils and resins. Biomass Bioenerg. 2006;30: 316–320.

9. da Silva Rodrigues-Correa KC, Lima JC, Fett-Neto AG. Oleoresins from pine: production and industrial uses. In: Ramawat KG, Merillon JM, editor. Natural Products. Berlin: Springer. 2013; p. 4037–4060.

0. Martin DM, Bohlmann J. Molecular biochemistry and genomics of terpenoid defenses in conifers.
Rec. Adv. Phytochem. 2005; 39: 29-

1. Miller B, Madilao LL, Ralph S, Bohlmann J. Insect-induced conifer defense: White pine weevil and methyl jasmonate induce traumatic resinosis, de novo formed volatile emissions, and accumulation of terpenoid synthase and putative octadecanoid pathway transcript in Sitka spruce. Plant Physiol. 2005; 137: 369-

2. Harvey BG, Wright ME, Quintana RL. High-density renewable fuels based on the selective dimerization of pinenes. Energy & Fuels 2010; 24: 267–273.

3. Zeng LH, Zhang Q, He BX, Lian HM, Cai YL, Wang YS. Age Trends in genetic parameters for growth and resin-yielding capacity in masson pine. Silvae Genetica. 2013; 62: 7-18.

4. Westbrook JW, Resende MF Jr, Munoz P, Walker AR, Wegryn JL, Nelson CD, Neale DB, Kirst M, Huber DA, Gezan SA. Association genetics of oleoresin flow in loblolly pine: discovering genes and predicting phenotype for improved resistance to bark beetles and bioenergy potential. New Phytol. 2013; 199: 89-100.

5. Mercati F, De Lorenzis G, Brancadoro L, Lupini A, Abenavoli MR, Barbagallo MG, Di Lorenzo R, Scienza A, Sunseri F. High-throughput 8K SNP array to assess genetic variability of the main grapevine cultivars from Sicily. Tree Genet Genomes. 2016; 12:59.

6. Strillacci MG, Cozzi MC, Gorla E, Mosca F, Schiavini F, Roman-Ponce SI, Lopez FR, Schiavone A, Marzoni M, Cerolini S. Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. Animal. 2017; 11:737-45.

7. Sunseri F, Lupini A, Mauceri A, De Lorenzis G, Araniti F, Brancadoro L, Dattola A, Gullo G, Zappia R, Mercati F. Single nucleotide polymorphism profiles reveal an admixture genetic structure of grapevine germplasm from Calabria, Italy, uncovering its key role for the diversification of cultivars in the Mediterranean Basin. Aust J Grape Wine R. 2018; 24:345–359.

8. Peng SL, Li QF, Li D, Wang ZF, Wang DP. Genetic diversity of Pinus massoniana revealed by RAPD markers. Silvae Genetica. 2003; 52: 60-
9. Li ZH, Chen Y, Zhang DL, Yang MH, Jiang Y, Ding GJ, Tan XF. ISSR analysis of genetic diversity of *Pinnus massoniana* on Gupeng and Langshui nature populations in Guangxi. Chin Agric Sci Bull. 2009; 16:116-120.

10. Zhang Y, Yang Q, Zhou ZC, Jin GQ. Divergence among Masson pine parents revealed by geographical origins and SSR markers and their relationships with progeny performance. New Forests. 2013; 44:341-355.

11. Fan F, Cui B, Zhang T, Ding GJ, Wen XP. LTR-retrotransposon activation, IRAP marker development and its potential in genetic diversity assessment of masson pine (*Pinus massoniana*). Tree Genetics & Genomes. 2014; 10: 213-222.

12. Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, Hattori C, Werner P, Bancroft I. Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nature Biotechnology. 2012; 30:798-802.

13. Chen X, Liu X, Zhu S, Tang S, Mei S, Chen J, Li S, Liu M, Gu Y, Dai Q, et al. Transcriptome-referenced association study of clove shape traits in garlic. DNA Research. 2018; 25: 587-596.

14. Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, Zhu Y, Ji A, Zhang B, Hu S, et al. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. Plant J. 2015; 82: 951–961.

15. Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marcais G, Roberts M, Wegrzyn JL, de Jong PJ, et al. Sequencing and assembly of the 22-gb loblolly pine genome. Genetics. 2014; 196: 875-890.

16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997 25: 3389-3402.

17. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated
individuals. Genome research. 2009; 19: 1655-1664.

8. Ying LX, Zhang TT, Chiu CA, Chen TY, Luo SJ, Chen XY, Shen ZH. The phylogeography of Fagus hayatae (Fagaceae): Genetic isolation among populations. Ecology and Evolution. 2016; 6: 2805-16.

9. Song W, Cao LJ, Li BY, Gong YJ, Hoffmann AA, Wei SJ. Multiple refugia from penultimate glaciations in East Asia demonstrated by phylogeography and ecological modelling of an insect pest. BMC evolutionary biology. 2018; 18: 152.

10. Martin DM, Faldt J, Bohlmann J. Functional characterization of nine Norway spruce TPS genes and evolution of gymnosperm terpene synthases of the TPS-d Plant Physiol. 2004; 135: 1908-1927.

11. Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann M, Birol I. SNP discovery in black cottonwood (Populus trichocarpa) by population transcriptome resequencing. Molecular Ecology Resources. 2015; 11:81-92.

12. Lorenzis GD, Mercati F, Bergamini C, Cardone MF, Lupini A, Maucerri A, Caputo AR, Loredana Abbate L, Barbagallo MG, Antonacci D, et al. SNP genotyping elucidates the genetic diversity of Magna Graecia grapevine germplasm and its historical origin and dissemination. BMC Plant Biology. 2019; 19:7.

13. Wu Y, Zhou Q, Huang S, Wang G, Xu L. SNP development and diversity analysis for Ginkgo biloba based on transcriptome sequencing. Trees. 2019; 33: 587-597.

14. Hamrick JL, Godt MJW, Sherman-Broyles SL. Factors influencing levels of genetic diversity in woody plant species. New Forests. 1992; 6:95-124.

15. Huang QQ, Zhang YX. Study on the genetic structure in Pinus massoniana (Masson Pine) populations. Silvae Genetica. 2000; 49:190-195.

16. Wu RJ. Genetic diversity of protein in alcohol for pinus massoniana from natural population. Scientia Silvae Sinicae. 2002; 38: 160-165.
7. Bernhardsson C, Floran V, Ganea SL, Garcia-Gil. Present genetic structure is congruent with common origin of distant Scots pine populations in its Romanian distribution. Forest Ecol Manag. 2016; 361:131-141

8. Toth EG, Vendramin GG, Bagnoli F, Cseke K, Hohn M. High genetic diversity and distinct origin of recently fragmented Scots pine (Pinus sylvestris) populations along the Carpathians and the Pannonian Basin. Tree Genetics & Genomes. 2017; 13(2):47.

9. Syring JV, Tennessen JA, Jennings TN, Wegryn J, Scelfo-Dalbey C, Cronn R. Targeted capture sequencing in white bark pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome. Front Plant Sci. 2016; 7: 484.

10. Aitken SN, Yeaman S, Holliday JA, Wang TL, Curtis-McLane S. Adaptation, migration or extirpation: climate change outcomes for tree populations. Evolutionary Applications. 2008; 1: 95-

11. Alberto F, Aitken S, Alia R, Gonzalez-Martinez SC, Hanninen H, Kremer A, Lefevre F, Lenormand T, Yeaman S, Whetten R, et al. Evolutionary response to climate change-evidence from tree populations. Global Change Biology. 2013; 19: 1645-

12. Ivany LC, Patterson WP, Lohmann KC. Cooler winters as a possible cause of mass extinctions at the Eocene/Oligocene boundary. Nature. 2000; 407:887-890.

13. Dupont-Nivet G, Krijgsman W, Langereis CG, Abels HA, Dai S, Fang XM. Tibetan plateau aridification linked to global cooling at the Eocene-Oligocene transition. Nature. 2007; 445:635-638.

14. Eldrett JS, Greenwood DR, Harding IC, Huber M. Increased seasonality through the Eocene to Oligocene transition in northern high latitudes. Nature. 2009; 459:969-973.

15. Tian S, Jordi L, Wang HW, Ge Song, Zhang ZY. Molecular evidence for glacial expansion and interglacial retreat during Quaternary climatic changes in a montane temperate pine (Pinus kwangtungensis Chun ex Tsiang) in southern China. Plant Systematics & Evolution. 2010;
284:219-229.

6. Qin GF. Geographical Origin and Evolution of Masson Pine. Forest Research 2002; 15:406-412.

7. Qian H, Ricklefs RE. Large-scale processes and the Asian bias in species diversity of temperate plants. Nature. 2000; 407:180-182.

8. Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S. Hemisphere-scale differences in conifer evolutionary dynamics. Proc Natl Acad Sci USA. 2012; 109:16217-16221.

9. Jasinski M, Stukkens Y, Degand H, Purnelle B, Marchand-Brynaert J, Boutry M. A plant plasma membrane ATP binding cassette-type transporter is involved in antifungal terpenoid secretion. Plant Cell. 2001; 13: 1095-

10. McKay SAB. Insect attack and wounding induce traumatic resin duct development and gene expression of (-)-Pinene synthase in Sitka spruce. Plant Physiology. 2003; 133:368-378.

11. Zhong R. Ectopic Deposition of Lignin in the Pith of Stems of Two Arabidopsis Mutants. Plant Physiol. 2000; 123:59-70.

12. Schuler MA, Werck-Reichhart D. Functional genomics of P450s. Annu Rev Plant Biol. 2003; 54: 629-667.

13. Schmidt A, Nagel R, Krekling T, Christiansen E, Gershenzon J, Krokene P. Induction of isoprenyl diphosphate synthases, plant hormones and defense signalling genes correlates with traumatic resin duct formation in Norway spruce (Picea abies). Plant Molecular Biology. 2011; 77: 577-

14. Jisha V, Lavanya D, Jyothilakshmi V, Axel M, Saivishnupriya K, Rajeshwari R. Overexpression of an AP2/ERF type transcription factor OsEREBP1 confers biotic and abiotic stress tolerance in rice. PloS one. 2015; 10: e0127831.

15. Liu QH, Zhou ZZ, Fan HH, Liu YR. Genetic variation and correlation among resin yield, growth, and morphologic traits of Pinus massoniana. Silvae Genet. 2013; 62: 38-44.

16. Liu Q, Wei Y, Xu L, Hao Y, Chen X, Zhou Z. Transcriptomic profiling reveals differentially expressed genes associated with pine wood nematode resistance in masson pine (Pinus
massoniana). Scientific Reports. 2017; 7:4693.

7. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*. 2007; 24:1586-1591.

8. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Yogesh R, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007; 23:2633-2635.

9. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{ΔΔCt}$ method. Methods. 2001; 25: 402-408.

**Supplementary Files Legend**

**Additional file**

**Additional file 1**

Table S1. The statistical results after combination of full length transcripts and unigenes of 204 accessions of masson pine (XLSX 11 kb)

**Additional file 2**

Table S2. The numbers of reads mapped to reference sequences for 204 accessions of masson pine (XLS 41kb)

**Additional file 3**

Table S3. The SNPs and InDels obtained using transcriptomes of 204 accessions (XLS 33kb)

**Additional file 4**

Table S4. The effect of SNP mutation for each unigene (XLSX 3,642 kb)

**Additional file 5**

Table S5. Core accessions with the 90.7% and 95% allele proportion (XLS 27kb)

**Additional file 6**

Fig. S1. Cross-validation error rate for each K value (PDF 135kb)

**Additional file 7**

Table S6. Three genetic clusters identified by DAPC based on SNP markers (XLS 27kb)

**Additional file 8**

Table S7. Nei’s and $F_{st}$ genetic distance calculated on three clusters inferred by DAPC (DOCX 33kb)
**Additional file 9:**

**Fig. S2.** Phylogenetic relationships among masson pine population and *P. taeda* as outgroup (TIFF 589kb)

**Additional file 10:**

**Table S8.** Oleoresin yield in xylem of 204 clones of masson pine (XLS 29 kb)

**Additional file 11:**

**Fig. S3.** Quantile-quantile plots result from the transcriptome-based association study data for oleoresin yield (TIFF 55kb)

**Additional file 12:**

**Table S9.** SNP markers significantly associated with variation in oleoresin yield in masson pine (XLS 79 kb)

**Additional file 13:**

**Table S10.** Origin of the 204 clones of masson pine (XLS 26 kb)

**Additional file 14:**

**Table S11.** Primers were designed from the sequences of transcriptome library in masson pine by using Primer Premier 3.0 (DOCX 16 kb)

Tables
Due to technical limitations, Tables 1 - 5 are only available for download from the Supplementary Files section.

Figures
Figure 1

Core collection size with the proportion of alleles captured by SNP markers in masson pine
Characterization of the masson pine population. A and B, population structure and corresponding groups of 204 masson pines. GI, Group I; GII, Group II; GIII, Group III. The image was created in Google Earth. C, two-dimension (DAPC) discriminant analysis of principal component scatter plot. All genotypes were grouped in three clusters. Cluster I: Sichuan and Chongqing; Cluster II: Jiangxi, Fujian, Zhejiang and Anhui; Cluster III: Guizhou, Guangxi, Guangdong and Hunan. D, Phylogenetic tree of the 204 masson pine based on the 94,194 SNPs. SC&CQ, Sichuan and Chongqing. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion
whatever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 3

The spread pathway of masson pine from the SC&CQ origin. A, Two orientation spreading lines in masson pine evolution map. B, population structure of masson pine in two spread pathway, respectively. C, Signal of introgression among different masson pine populations detected by the TreeMix program. SC&CQ, Sichuan and Chongqing; GZ, Guizhou; HN, Hunan; GX, Guangxi; GD, Guangdong; JX, Jiangxi; FJ, Fujian; AH, Anhui; ZJ, Zhejiang.
Discover the candidate genes associated with oleoresin yield by associative transcriptomics in masson pine. A, Distribution of oleoresin yield among 204 masson pines. B, The variation of oleoresin yield among 9 populations. C, Manhattan plots result from the transcriptome-based association study data for oleoresin yield. a, Chitinase class I (c51955_f1p3_1546); b, Tubulin alpha chain (c20772_f1p4_1467); c, ABC transporter (c189021.graph_c0); c, AP2/ERF (c24091_f1p1_1286); c, AP2/ERF (c8825_f1p0_1733); e, CYP720B (c19795_f1p0_1763); f, Cytochrome P450 (c9591_f1p0_1663). D, Quantitative RT-PCR (qRT-PCR) validation of candidate genes associated with oleoresin yield.

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

Additional file 1 Table S1.xlsx
Additional file 8 Table S7.docx
Additional file 11 Fig. S3.pdf
Additional file 3 Table S3.xls
Additional file 2 Table S2.xls
Additional file 6 Fig. S1.tif
Additional file 14 Table S11.docx
Additional file 5 Table S5.xls
Additional file 13 Table S10.xlsx
Additional file 7 Table S6.xls
Additional file 10 Table S8.xls
Additional file 9 Fig. S2.tif
Additional file 12 Table S9.xls
Additional file 4 Table S4.xlsx
Tables.docx