Draft Genome of the Medicinal Tea Tree *Melaleuca Alternifolia* (Maiden & Betcher) Cheel (Myrtaceae)

Xiaoning Zhang (✉ sflzxn@163.com)

Guangxi Forestry Research Institute <https://orcid.org/0000-0003-2105-0560>

Silin Chen

Kunming Institute of Botany Chinese Academy of Sciences

Ye Zhang

Guangxi Forestry Research Institute

Yufei Xiao

Guangxi Forestry Research Institute

Yufeng Qin

Guangxi Forestry Research Institute

Qing Li

Kunming Institute of Botany Chinese Academy of Sciences

Li Liu

Kunming Institute of Botany Chinese Academy of Sciences

Buming Liu

Guangxi Key Laboratory of Traditional Chinese Medicine Quality Standards

Ling Chai

Guangxi Key Laboratory of Traditional Chinese Medicine Quality Standards

Hong Yang

Hubei University

Hailong Liu

Guangxi Forestry Research Institute <https://orcid.org/0000-0002-4983-8555>

Research Article

Keywords: Genome sequencing, *Melaleuca alternifolia*. Tea tree oil. Whole-genome shotgun sequencing (WGS)

Posted Date: February 15th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1293742/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background *Melaleuca alternifolia* (Maiden & Betcher) Cheel is a commercially important medicinal tea tree native to Australia. Tea tree oil (TTO), the essential oil distilled from its branches and leaves, has broad-spectrum germicidal activity and is highly valued in the pharmaceutical and cosmetic industries. Genome size is a key biology index, the study of which provides reference for genome engineering and sequencing work; besides, it also provides direct data for evolution biology and so on.

Methods and results In the present study, the next-generation sequencing (NGS) was used to investigate the whole genome of *Melaleuca alternifolia*. 114 Gb high quality sequence data were obtained and assembled into 1,838,159 scaffolds with an N50 length of 1021 bp. The assembled genome size is about 595 Mb, twice of that predicted by flow cytometer (300 Mb) and k-mer analysis (345 Mb). Benchmarking Universal Single-Copy Orthologs analyses indicated that only 11.3% of the conserved single-copy genes were miss. Repetitive regions cover over 40.43% of the genome. *M. alternifolia* maybe a diploid with highly heterozygous (heterozygosity is 0.8%). A total of 44,369 protein-coding genes were predicted in the assembled genome and 32,909 genes of these were functionally annotated, of which 8,436 genes were annotated in KEGG. Additionally, 457,661 simple sequence repeats (SSRs) and 1,109 transcription factors (TFs) form 67 TF families were identified in the assembled genome.

Conclusion This draft genome may provide a reference for the deep sequencing strategies and be useful for future functional or comparative genomics analyses.

Introduction

The Myrtaceae family includes about 130 genera and 5,000 species, many of which accumulate essential oils in their foliage and branchlets [1]. *Melaleuca alternifolia*, native to Australia, is thought to be a diploid species with a chromosome number of $2n = 22$ and is the best-known member of the *Melaleuca* genus planted in China [2]. The leaves and branches of *M. alternifolia* produce tea tree oil (TTO), which has broad-spectrum germicidal, antioxidant, anti-inflammatory, and anticancer activities [3] [4], and is highly valued in the pharmaceutical, cosmetic, and food industries [5, 6]. Terpenoid compounds contained in TTO also have important adaptive roles in plant defense against pests and abiotic stress.

M. alternifolia were introduced into China in the 1980s. After decades of efforts, we have selected superior individuals with high-quality TTO, established a mature tissue culture technology system, and realized large-scale tissue culture production. TTO consists of over 100 components, and terpinen-4-ol accounts for about one-third of the TTO in the terpinen-4-ol chemotype, which plays a critical role in the antimicrobial activity [7, 8]. Improved productivity of TTO will yield significant profit income, and might be a major target of breeders at present and for a long time [9].

Next-generation sequencing (NGS) can be used to generate and assemble high-quality genomes, based on its high throughput, high resolution, high precision, and low cost. Short-read sequencing is usually performed to estimate genome size, heterozygosity, repetitive rate, and evaluate the simple sequencing

repeat (SSR) markers at the genome level, in advance of long-read sequencing due to its cost-effective [10].

M. alternifolia genome had been sequenced recently [11], as well as several other Myrtaceae trees such as *Eucalyptus pauciflora* [12], *Eucalyptus grandis* [13], *Metrosideros polymorpha* [14], and *Leptospermum scoparium* [15]. These genome sequences not only provide foundational information for genetic engineering of better trees, but also contribute to a deeper understanding of their genomic evolution, while supplying a reference for genome sequencing projects of other tree species. Here, we employed whole-genome sequencing (WGS) and de novo assembly to produce a draft genome of *M. alternifolia* (terpinen-4-ol chemotype with high terpinen-4-ol content more than 30% and low 1,8-cineole content less than 5%) with its in-vitro aseptic plantlets. The size of the *M. alternifolia* genome was estimated using both the k-mer frequency distribution and flow cytometry (FCM). Moreover, the heterozygous and duplication rates of *M. alternifolia* genome were deduced using the JELLYFISH tool [16]. Our data provides valuable information for precise and in-depth sequencing and pave the way for further investigation of the genes involved in terpinen-4-ol biosynthesis.

Materials And Methods

Plant materials and DNA extraction

Melaleuca alternifolia (terpinen-4-ol chemotype, with high terpinen-4-ol and low 1,8-cineole) plantlets of tissue culture were grown at $25 \pm 2^{\circ}\text{C}$ with 16 h/8 h (light/dark) photoperiod and 2,500 lx illumination intensity for 1 month. Leaf tissues of *M. alternifolia* were collected and ground in liquid nitrogen for DNA extraction. Genomic DNA was then extracted using the DNA extraction kit DP350 (Tiangen biotech LTD, Beijing, China).

Flow cytometry analysis

Twenty milligram plant tissues were collected and immediately chopped in 1 mL cold nuclei isolation buffer (45 mM $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$, 20 mM MOPS, 30 mM Na-Citrate, 1% (w/v) PVP 40, 0.2% (v/v) Triton X-100, 10 mM Na_2EDTA , 20 $\mu\text{L/mL}$ β -mercaptoethanol, pH 7.0). Homogenate was filtered using 42 μm nylon mesh, and nuclei were collected in new sample tube. Then 50 mg/mL DNA fluorochrome propidium iodide (PI) simultaneously with RNase at 50 mg/mL were added and gently mixed with samples. The nuclei were incubated on ice for 1 h with occasional shaking before analysis. Maize (*Zea mays*) B37 was selected as the internal reference, and the relative fluorescence of the stained nuclei was measured using BD FACScalibur (Becton Dickinson, NY, USA).

Data processing

Two genomic DNA libraries with an insert size of 270 bp and 500 bp were constructed, respectively. Paired-end sequencing of both libraries was performed using the Hiseq X-ten platform (Illumina, NY, USA). To reduce the effect of sequencing errors on assembly, all raw sequencing reads were then filtered with

the SOAPnuke1.5.5 software [17]. To estimate the size and heterozygosity of the *M. alternifolia* genome, the filtered reads were used for k-mer (K = 17) determination within subsequent assembly steps [16]. Moreover, the JELLYFISH tool was used to simulate the heterozygosity and repetition rates of 15-Gb and 60-Gb data sets, which were randomly extracted from the filtered data used to calculate the k-mer.

De novo assembly and quality evaluation

SOAPdenovo v2.04 software [18] with default parameters were used to assemble the filtered reads for each library (270 bp and 500 bp inserts). Post-assembly processing included removal of contaminating bacterial and viral DNA sequences by aligning all assembled sequences to viral and bacterial genome sequences. The genome sequences we used were obtained from previous local (Basic Local Alignment Search Tool) BLASTn alignments and NCBI (National Center for Biotechnology Information, <https://www.ncbi.nlm.nih.gov/>) upload filter. Aligned sequences which shared >90% identity and >200 bp were removed from the final assembly. BUSCO v5.0.0 [19] with the viridiplantae_odb10 dataset which consisting of 425 Benchmarking Universal Single-Copy Orthologs (BUSCOs) from 57 species were used to evaluate the assembly completeness of gene content. GC content analysis was then performed and five sequences of each GC cluster were randomly selected for nucleotide (nt) library comparison, using Megablast software and the parameter was set with ‘-evaluate 1e-5’.

Genome annotation

To annotate the genome, BALSTp searches were performed against with NCBI NR database using Diamond v0.9.14.115 [20] with parameter ‘-evaluate 1e-5’. Moreover, genes in the *M. alternifolia* genome were also annotated by KEGG Automatic Annotation Server (KAAS) [21]. To generate the species-specific transposable elements (TEs) and repeats library of *M. alternifolia*, RepeatModeler v2.0.1 [22] was used in our research. The soft-mask repetitive elements in the *M. alternifolia* genome were identified using RepeatMasker v4.0.7 [23] running with NCBI/RMBLAST v2.6.0+ as search engine, and gff3 file was then generated based on the masked genome using AUGUSTUS v3.3 [24].

SSRs and TFs identification

MISA-microsatellite searching tool [25] was used to identify the SSRs in the *M. alternifolia* genome. Transcription factors (TFs) of genome-wide were identified and classified by the software iTAK [26].

Results

Genome sequencing

The young shoots of the *Melaleuca alternifolia* superior individuals (Figure 1a) were used as explants donor of in vitro culture. Induced sterile buds was grown in proliferation medium for 1 month with about 2-3 cm shoot height and 0.5 cm leaf length. They were cut into about 2 cm stem segments then transferred into rooting medium for another 1 month when developing more than 10 roots per shoot (Figure 1b). Well-rooted plantlets were transplanted into greenhouse in a red mud cup for three months

(Figure 1c) then used for afforestation (Figure 1d). The fresh leaves of these well-rooted plantlets were used as test materials to extract DNA for genome sequencing (Figure 1b). We generated a total of 402 and 465 million reads from the paired-end sequencing of our two *M. alternifolia* genomic libraries. After filtering and corrections, we retained 355 and 408 million reads for each library, corresponding to 53 Gb and 61 Gb of clean data, high-quality sequences for the 270-bp and 500-bp genomic libraries, with the clean data rate of 88.3% and 88.4%, respectively. When combined, the two genomic libraries therefore include 114 Gb data (Supplement Table S1).

Genome size estimation and heterozygous simulation

A k-mer analysis was performed to estimate the genome size. We determined the k-mer distribution in our sequencing data with $K = 17$. The distribution showed two peaks: a major peak at depth = 37, and a minor peak at depth = 18, the height of the main peak (depth = 37) was about twice of the minor peak (depth = 18). The value of the main peak was also about twice that of the minor peak (Figure 2). The total number of 17-mers amounted to 12,767 Mb of sequence, resulting in a genome size estimation of 345 Mb when taking the major peak (depth = 37) and using the formula: Genome size = k-mer num/peak depth.

The presence of the minor peak suggests that the *M. alternifolia* genome is highly heterozygous. We therefore simulated the heterozygosity and repetition rate of the *M. alternifolia* genome with JELLYFISH software [16], when using a depth of 37 as the main peak, the duplication rate was 41.95% and the heterozygosity rate was 0.80%, which is considered highly heterozygous [27], the read coverage is about 42 (Supplement Table S2).

Fluorescence cell sorting of purified nuclei was used to estimate the DNA content of the *M. alternifolia* genome. *M. alternifolia* cell nuclei were stained with the fluorescent dye propidium iodide (PI), which could be uniformly embedded into DNA double chains, and then flow cytometry was performed to obtain the fluorescence signal. As the number of nuclei labeled with PI is proportional to the DNA content, the relative DNA content in the sample could be quantified based on the observed fluorescence signal intensity. The mean fluorescence intensity for the reference genome (maize inbred line B73) was 215.13, whereas that for *M. alternifolia* was only 28.31 (Supplement Table S3 & Figure S1). Based on the known genome size of maize B73 of 2.3 Gb [28], we therefore estimate that the genome size of *M. alternifolia* is approximately 300 Mb and hypothesize that multiple peaks may have been caused by endoreduplication.

Sequence assembly and GC content analysis

SOAPdenovo software was then used for genome de novo assembly and 63 mers were selected. A total of 1,838,159 scaffolds were assembled into a final genome sequence of 595 Mb with an N50 length of 1,021 bp (Table 1). This de novo assembly produced a sequence twice as large as the estimated genome size by k-mer analysis (345 Mb) and flow cytometry (300 Mb), which may indicate that the assembled genome is a highly heterozygous diploid.

We calculated the GC content of the *M. alternifolia* genome in 20-kbp sliding windows without overlap (Supplement Figure S2). Our data fell into three clusters. Clusters A and B correspond to *M. alternifolia*, and the much smaller cluster C is presumed to represent DNA contaminants from exogenous sources such as bacteria, fungi, or insects. Mean GC content was 41.9%, as shown in Supplement Figure S1. The genome survey data represent 192 coverage of the *M. alternifolia* genome, and the splicing coverage was 198% (Table 1), both indicators of the high quality of our *M. alternifolia* sequence. Splicing coverage refers to the proportion of the assembled genome sequence to the whole genome. A splicing coverage of >95% is generally considered to be a robust genome assembly. Occasionally, the coverage is >100%, which may be due to the existence of repeat regions in the genome [29].

Table 1
Summary for the de novo assembly of the *M. alternifolia* genome.

Statistical item	Contig	Scaffold
N10 length (bp)	15,009	28,042
Number of N10	2,291	1,151
N50 length (bp)	796	1,021
Number of N50	100,342	71,431
N90 length (bp)	118	119
Number of N90	1,362,130	1,294,949
Maximum length (bp)	136,294	332,689
Total size (bp)	588,879,542	595,305,086
Total number (≥100 bp)	1,900,016	1,838,159
GC content (%)	42.4	41.9
Depth of sequencing (×)	192	
Splicing coverage (%)	198	

Genome completeness evaluation

The level of genome completeness for the assembled sequences was evaluated by using BUSCO v5.0.0 (BUSCO, RRID:SCR_015008), which quantitatively assesses genome completeness using evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from viridiplantae_odb10 (viridiplantae odb, RRID:SCR_011980; <http://busco.ezlab.org/>). BUSCO analysis showed that separate 67.5% and 21.2% of the 425 expected plant genes were identified as complete and fragmented genes respectively, while 11.3% of genes were considered to be missing from the assembled *M. alternifolia* genome sequence (Supplement Table S4).

Repeat element annotation

The total repeat length of the *M. alternifolia* genome (total length 595,305,086 bp) was estimated to be 240,657,903 bp, around 40.43% (Supplement Figure S3). 354,647,183 bp of non-redundant repetitive elements were also identified, accounting for 59.57%. About 6.93% of the genome can be attributed to TEs: DNA transposons, retrotransposons such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeat (LTR); 1.78% of the repeats were predicted to be small RNAs, satellites, simple and low-complexity repeats (Supplement Figure S3 & Table S2), among which 191,238 simple sequence repeats were annotated, which will provide valuable genetic markers to assist *M. alternifolia* breeding programs. The GC content was 42.39% across the genome.

The final assembly contains 44,369 predicted protein-coding genes (Supplementary Files S1), around 74.2% (32,909 genes) were functionally annotated in NCBI nonredundant protein (NR) database (Supplementary Files S2). Further, 8436 genes were annotated in KEGG metabolic pathways (Supplementary Files S3).

Table 2
Major types of repeat elements identified in the *M. alternifolia* genome assembly.

Repeat class	No. of elements	Total length, bp	% of genome
LINE	23,544	6,784,958	1.14
SINE	1,090	93,904	0.02
LTR	116,833	30,788,551	5.17
DNA transposons	14,545	3,588,626	0.60
Unclassified	1,457,484	186,499,107	31.33
Rolling-circles	5,871	2354517	0.40
Small RNA	29,702	2,262,577	0.38
Satellites	127	38,343	0.01
Simple repeats	191,238	7,015,736	1.18
Low complexity	25,911	1,231,584	0.21

SSR identification

The scaffolds were utilized for SSR identification to ensure higher reliability. A total of 457,661 SSRs, including 255,359 (55.8%) mono-, 176,264 (38.51%) di-, 21,775 (4.76%) tri-, 2839 (0.62%) tetra-, 810 (0.18%) penta-, and 614 (0.13%) hexanucleotide repeats, were identified in 355,921 scaffolds (Supplement Figure S4). The mononucleotide repeat was overwhelming, and the 20 most frequent SSR types were exhibited in Supplement Figure S5. The A/T repeat was the most abundant type, accounting for 51.09% of all SSRs, followed by AG/CT (32.95%) and C/G (4.71%). Other dinucleotides and trinucleotides repeat types, such as AT/AT, AAG/CTT also made up an outstanding proportion.

TF Identification

There are 1109 transcription factors (TFs) coding genes from 67 TF families were identified and classified using the online tools Itak (http://itak.feilab.net/cgi-bin/itak/online_itak.cgi) in *M. alternifolia*. 1109 TFs belonged to 67 TF family were identified in *M. alternifolia* genome (Supplemental Files S4). Top 10 TFs including C2H2, MYB-related, bHLH, AP2/ERF-ERF, zn-clus, NAC, bZIP, C3H, WRKY, MYB accounts for 56.45% of the total TFs (Figure 3).

Discussion

The results of flow cytometer revealed that the genome size of *M. alternifolia* is approximately 300 Mb (Supplement Table S3), which is similar to the result obtained by k-mer analysis, i.e., 345 Mb. The genome size of our results are consistent with the previous research (356Mb) of Calvert's (2017). And similar to that of *Metrosideros polymorpha* at 347 Mb [14] and *Leptospermum scoparium* at 297 Mb [15], but less than *Eucalyptus pauciflora* at 595 Mb and *Eucalyptus grandis* at 640 Mb [12, 13], all of which belong to the Myrtaceae family. The genome size of *M. alternifolia* is considered "very small" according to the definition of genome size [30]. Such a small genome will facilitate genome research and molecular manipulation, and provide insights into genome size variation in tree genomes.

A GC content that is too high (>65%) or too low (<25%) may result in sequence bias during whole-genome sequencing, and may negatively affect genome assembly. The mean GC content of our result was 41.9%, quite consistent with the result of Voelker's, 42%. According to the GC content and depth analysis, two of three detected clusters were derived from *M. alternifolia*. However, the de novo-assembled genome was 595 Mb (Table 1), which is twice about the genome sizes estimated using flow cytometry (300 Mb) and k-mer frequency distribution analyze (345 Mb). This larger assembly length was also performed in *Metrosideros polymorpha* [31]. This notable larger assembly length in *M. alternifolia* may suggest more than one haplotype were assembled. Based on these observations, we speculate that this species may be a diploid with a highly heterozygous genome.

Genome heterozygosity reflects the reproductive lifestyle of the species, such as self-incompatibility or outcrossing, but may confound the quality of the genome assembly. Our k-mer analysis identified two depth peaks: the height of the main peak (depth = 37) was about twice that of the minor peak (depth = 17), suggesting that the *M. alternifolia* genome is highly heterozygous, and the heterozygosity determined using JELLYFISH was 0.8% (Supplement Table S2) smaller than that of *Eucalyptus grandis* (1.0%) [12] and *Eucalyptus pauciflora* (1.5%) [13]. However, the *M. alternifolia* genome is on the threshold for being considered a highly heterozygous genome with heterozygosity $\geq 0.8\%$ [27].

Flowers of *M. alternifolia* are hermaphroditic, protandrous, pollinated by insects and predominantly outcrossing by nature, with occasional geitonogamy (self-pollination among flowers within the inflorescence) [32, 33]. The high heterozygosity of the *M. alternifolia* genome is likely the result of outcrossing [34] and highly confounds genome sequencing and assembly [35].

Contig N50 and BUSCO (BUSCO, RRID: SCR 015008) scores are usually used to assess and compare genome assemblies. A higher N50 length suggests an assembled genome with fewer and larger contigs. The length of contig N50 assembled by SOAPdenovo of *M. alternifolia* genome is 1,021 bp, lower than that of the previous research of this species (8,778bp) [11]. May be related to its short-read assembly method. However, it is much higher than that of *Mikania cordata* (312~352 bp) which used the same SOAPdenovo software [10]. Genome completeness, estimated by BUSCO using the viridiplantae dataset, was 88.7%, indicating a relatively high genome completeness. However, these two methods of short-read assembly were much lower than Pacific Biosciences using MaSuRCA assembly. Repetitive regions cover over 40.43% of the genome, which is similar with that in *Eucalyptus pauciflora* (44.77%) and *Eucalyptus grandis* (41.22%) [12].

A total of 44,369 protein-coding genes were predicted in the *M. alternifolia* assembled genome, more than those encoded by *Metrosideros polymorpha* (39,305) [14], *Eucalyptus grandis* (36,376) [13] and *Leptospermum scoparium* (31,220) [15]. 1109 TFs were annotated in the whole genome. C2H2, Myb related, bHLH were the top 3 TFs, accounted for 17.5%, 14.2%, 12.7%, respectively. Analysis of SSRs was conventionally performed in genome survey, 457,661 SSRs have been identified, A/T, AG/CT, AAG/CTT are the most common motifs of mono-, di-, and trinucleotide, respectively. This result showed the same pattern with that in the *Camellia sinensis* [36]. The genomic SSRs identified here would benefit assessing the genetic diversity of *M. alternifolia*.

In conclusion, we report a draft genome sequencing of *M. alternifolia*. This cost-effective strategy based on short-read data helped us to reveal its basic genomic parameters with small genome size, mid GC content, high heterozygosity rate. Its genome annotation has also been characterized, including repeat element annotation, the SSR identification, and TFs prediction at the genome level. This *M. alternifolia* draft genome will be useful for functional and comparative genomics research in the future.

Conclusions

We have sequenced the *Melaleuca alternifolia* genome and obtained 129 Gb of sequencing data from two genomic DNA libraries (270 bp and 500 bp), resulting in 114 Gb of clean data after quality control (Supplement Table S1). A total of 1,838,159 scaffolds were assembled for a total sequence length of 595 Mb, with an N50 length of 1,021 bp (Table 1). We estimated the genome size to be 345 Mb by k-mer frequency distribution and 300 Mb with FCM (Supplement Table S3). These numbers are only half of the total assembled sequence of 595 Mb. GC content and average depth analysis further confirmed that *M. alternifolia* may be a highly heterozygous diploid.

The draft genome assembly of *M. alternifolia* lays the foundation for further genome engineering and sequencing work, which can be used to accelerate the molecular breeding of *M. alternifolia* with improved tea tree oil yield and quality.

Declarations

Funding

This study was funded by the Department of Human Resources and Social Security of Guangxi Zhuang Autonomous Region, China (GuiCaiSheHan [2018]112); Guangxi Science and Technology Project (Gui Ke AD18281083 & Gui Ke AB18221058); Guangxi Key Laboratory of Traditional Chinese Medicine Quality Standards Open Project [202002].

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Conceptualization, Hong Yang and Hailong Liu; Data curation, Xiaoning Zhang and Silin Chen; Formal analysis, Ye Zhang, Yufei Xiao and Yufeng Qin; Funding acquisition, Hailong Liu; Investigation, Hong Yang; Methodology, Ye Zhang, Yufei Xiao and Yufeng Qin; Resources, Qing Li, Buming Liu and Ling Chai; Supervision, Hailong Liu; Visualization, Qing Li, Buming Liu and Ling Chai; Writing-original draft, Xiaoning Zhang and Silin Chen; Writing-review and editing, Li Liu.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent to publish

The authors affirm to grant the Publisher an exclusive licence to publish the article or to transfer copyright of the article to the Publisher.

Data Availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center (Nucleic Acids Res 2021), China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number CRA004893 that are publicly accessible at (<https://ngdc.cncb.ac.cn/gsa>) or at (<https://ngdc.cncb.ac.cn/gsa/browse/CRA004893>).

Supplemental Information

Supplemental information for this article can be found in the submission system.

References

1. Wu ZY, Raven PH (2013) Flora of China. Science Press: Beijing, China & Missouri Botanical Garden Press: St. Louis, MO, USA 321
2. Zhang X, Liang G, Yan Y, Yu Y, Yang G, Yang T (2000) Rapid propagation and polyploid induction in *Melaleuca alternifolia*. J Southwest Agric Univ 22(6):507-9
3. Chidi F, Bouhoudan A, Khaddor M (2020) Antifungal Effect of the Tea Tree Essential Oil (*Melaleuca alternifolia*) Against *Penicillium griseofulvum* and *Penicillium verrucosum*. J King Saud Univ Sci 32(3):2041-2045
4. Yadav E, Kumar S, Mahant S, Khatkar S, Rao R (2016) Tea tree oil: a promising essential oil. Journal of Essential Oil Research 29(3):201-13
5. Redondo-Blanco S, Fernandez J, Lopez-Ibanez S, Miguelez EM, Villar CJ, Lombo F (2020) Plant Phytochemicals in Food Preservation: Antifungal Bioactivity: A Review. J Food Prot 83(1):163-71
6. Lee JY, Lee J, Ko SW, Son BC, Lee JH, Kim CS, et al (2020) Fabrication of Antibacterial Nanofibrous Membrane Infused with Essential Oil Extracted from Tea Tree for Packaging Applications. Polymers (Basel) 12(1):125
7. Felipe LO, Junior W, Araujo KC, Fabrino DL (2018) Lactoferrin, chitosan and *Melaleuca alternifolia*-natural products that show promise in candidiasis treatment. Braz J Microbiol 49(2):212-9
8. Sharifi-Rad J, Salehi B, Varoni EM, Sharopov F, Yousaf Z, Ayatollahi SA, et al (2017) Plants of the *Melaleuca* Genus as Antimicrobial Agents: From Farm to Pharmacy. Phytother Res 31(10):1475-94
9. Bustos-Segura C, Padovan A, Kainer D, Foley WJ, Külheim C (2017) Transcriptome analysis of terpene chemotypes of *Melaleuca alternifolia* across different tissues. Plant Cell Environ 40(10):2406-25
10. Hong Y, Huang X, Li C, Ruan X, Wang Z, Su Y, et al (2020) Genome Survey Sequencing of In Vivo Mother Plant and In Vitro Plantlets of *Mikania cordata*. Plants 9(12):1665
11. Calvert J, Baten A, Butler J, Barkla B, Shepherd M (2017) Terpene synthase genes in *Melaleuca alternifolia*: comparative analysis of lineage-specific subfamily variation within Myrtaceae. Plant Syst Evol 304(1):111-21
12. Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, et al (2020) The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. Gigascience 9(1):1-12
13. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al (2014) The genome of *Eucalyptus grandis*. Nature 510(7505):356-62
14. Izuno A, Hatakeyama M, Nishiyama T, Tamaki I, Shimizu-Inatsugi R, Sasaki R, et al (2016) Genome sequencing of *Metrosideros polymorpha* (Myrtaceae), a dominant species in various habitats in the Hawaiian Islands with remarkable phenotypic variations. J Plant Res 129(4):727-36
15. Thrimawithana AH, Jones D, Hilario E, Grierson E, Ngo HM, Liachko I, et al (2019). A whole genome assembly of *Leptospermum scoparium* (Myrtaceae) for mānuka research. N Z J Crop Hortic Sci 47(4):233-60

16. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764-70
17. Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al (2018) SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and prepro-cessing of high-throughput sequencing data. *Gigascience* 7:1-6
18. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1(1):18
19. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* (31):3210-2
20. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Meth* 12(1):59-60
21. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35(Web Server issue):W182-5
22. Flynna JM, Hubleyb R, Gouberta C, Rosenb J, Clarka AG, Feschottea C, et al (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* 1:9451-7
23. Tempel S (2012) Using and understanding RepeatMasker. *Methods Mol Biol* 859:29-51
24. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34:W435-9
25. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics (Oxf)* 33(16):2583-5
26. Zheng Y, Jiao C, Sun H, G.Rosli H, A.Pombo M, Zhang P, et al (2016) iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. *Mol Plant* 9(12):1667-70
27. Wu Y, Xiao F, Xu H, Zhang T, Jiang X (2014). Genome Survey in *Cinnamomum camphora* L. *Presl. J Plant Genet Resour* 15(1):149-52
28. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956):1112-5
29. Desai A, Marwah VS, Yadav A, Jha V, Dhaygude K, Bangar U, et al (2013) Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PLoS One* 8(4):e60204
30. Camillo J, Leao AP, Alves AA, Formighieri EF, Azevedo AL, Nunes JD, et al (2014) Reassessment of the Genome Size in *Elaeis guineensis* and *Elaeis oleifera*, and Its Interspecific Hybrid. *Genom Insights* 7:13-22
31. Izuno A, Hatakeyama M, Nishiyama T, Tamaki I, Shimizu-Inatsugi R, Sasaki R, et al (2016) Genome sequencing of *Metrosideros polymorpha* (Myrtaceae), a dominant species in various habitats in the

- Hawaiian Islands with remarkable phenotypic variations. J Plant Res 129(4):727-736
32. Baskorowati L, Moncur MW, Doran JC, Kanowski PJ (2010) Reproductive biology of *Melaleuca alternifolia* (Myrtaceae) 1. Floral biology. Aust J Bot 58:373-83
33. Baskorowati L, Moncur MW, Cunningham SA, Doran JC, Kanowski PJ (2010). Reproductive biology of *Melaleuca alternifolia* (Myrtaceae) 2. Incompatibility and pollen transfer in relation to the breeding system. Aust J Bot 58:384-91
34. Butcher P, JC B, GF M (1992) Patterns of Genetic Diversity and Nature of the Breeding System in *Melaleuca alternifolia* (Myrtaceae). Aust J Bot 40:365-75
35. Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, et al (2018) Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. Proc Natl Acad Sci U S A. 115(18):E4151-E8
36. Wei Y, Jing W, Hua D, Youxiang Z, Mingming Z, Dingjin. H (2013) Characteristic Analysis and Application of Microsatellites from EST Sequence of *Camellia sinensis*(in chinese). Hubei Agricultural Sciences 52(24):6178-81

Figures

Figure 1

Afforestation with tissue culture plantlets of *M. alternifolia*. Adult superior strains (a); Single bud of about 2 cm height grown in rooting medium for 1 month (b); Well-rooted plantlets were transplanted into greenhouse in a red mud cup for 3 months (c); *M. alternifolia* stands (d).

Figure 2

K-mer ($k = 17$) analysis for *M. alternifolia* genome size estimation. The x-axis indicates the coverage depth ($\times$); the y-axis indicates the frequency of k-mers at a given depth divided by the total frequency at all depths.

Figure 3

Top 20 TF families of *M. alternifolia* genome.

Supplementary Files

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

- [Supplementaryfiguresandtables.pdf](#)
- [supplementaryfiles.rar](#)