Chromosomal-level genome assembly of *Populus lasiocarpa*

Zhiqin Long[1#], Yupeng Sang[1#], Jiajun Feng[1], Tingting Shi[1], Xuming Dan[1], Yulin Zhang[1], Jianquan Liu[1], Jing Wang[1*]

[1]Key Laboratory for Bio-Resources and Eco-Environment, College of Life Science, Sichuan University, Chengdu, Sichuan, China

#These authors contributed equally

*Correspondence: wangjing2019@scu.edu.cn

Abstract

Populus lasiocarpa, commonly called the Chinese necklace poplar, is a species of poplar native to humid forests of China, and is known for its large leaves that may reach dimensions of 35 × 25 cm. In this study, we generated a high-quality chromosomal-level *de-novo* assembly and annotation of *P. lasiocarpa* with a genome size of 419.54 Mb and a gene number of 39,008, which provide an important data support for the conservation and utilization of wild germplasm resources of *P. lasiocarpa*.

Introduction

Rapid global climate change is posing a main threat to biodiversity. Therefore, revealing the evolutionary history of species, understanding the patterns of genetic diversity, exploring the genetic mechanism of adaptive evolution, and evaluating adaptive capacity of species in the facing with changing environment have laid a theoretical foundation for genetic rescue and informing conservation strategies. *Populus lasiocarpa*, a species with ring distribution of populations wraping around Sichuan Basin, is unique poplar germplasm resources in China. The geographical barriers and heterogeneous environments constituted by the numerous uplifted mountain surrounding the basin offer ideal materials for studying geographic isolation and genetic mechanisms of adaptive evolution. Notably, a high-quality reference assembly is a key for future related studies. Therefore, we provide a high-quality *P. lasiocarpa* genome which assembled into chromosome-level here.

Results

Genome assembly and annotation

We first performed *K-mer* analysis to determine *P. lasiocarpa* genome size and composition via Illumina sequencing data, revealing an estimated genome size of 451.2 Mb and a heterozygosity rate of 0.6%. To obtain a high-quality genome assembly, we generated 49.95 Gb of Nanopore long-read sequences (~119×), 36.37 Gb of Illumina short-read data (~87×), and 61.56 Gb of high-throughput chromosome conformation capture (Hi-C) data (~149x) for *P. lasiocarpa*. Using these sequencing data, a 419.54 Mb non-redundant assembly was obtained with the contig N50 size of 9.19 Mb and contig number of 105 after removing redundant sequences and potential contaminated sequences (Table 1). Based on Hi-C read pairs, the assembled contigs were further anchored to 19 pseudo-chromosomes with an average anchoring rate of 99.23% (Table 2). The completeness and accuracy of the assembled genome were validated using benchmarking universal single-copy orthologues (BUSCO) showed that 1,346 complete plant orthologues (97.89%) were recalled (Table 3). The assembly was further evaluated by mapping short reads to the genome, which revealed a mapping rate and single-base accuracy (Depth>=5×) of 97.83% and 99.99%, respectively. Collectively, these results reflected the high level of completeness and reliability of our *P. lasiocarpa* genome assembly.

Table 1. Statistics of assembled contigs for *P. lasiocarpa*.

Type	Length	Number
Total	419,540,624	105
Longest	31,312,713	-
N50	8,459,212	15
N60	6,067,086	20
N70	3,878,110	27
N80	1,970,620	37
N90	917,550	54

Table 2. Scaffolding of contigs based on Hi-C data.

Peudo-chromosome	Length（Contig number）
LG01	52,992,279 (9)
LG02	28,222,690 (3)
LG03	26,620,501 (7)
LG04	26,318,386 (3)
LG05	25,382,582 (6)
LG06	22,749,246 (4)
LG07	22,695,253 (7)
LG08	21,045,204 (3)
LG09	20,558,685 (7)
LG10	18,743,377 (4)
LG11	18,014,361 (3)
LG12	17,702,534 (7)
LG13	17,409,083 (5)
LG14	17,212,443 (8)
LG15	17,022,449 (4)
LG16	16,768,192 (8)
LG17	16,667,112 (2)
LG18	15,551,821 (8)
LG19	14,625,206 (1)
Total	416,301,404 (99)

Table 3. Statistics of BUSCO evaluation for genome assembly.

BUSCO	Number	Percentage
Complete BUSCOs (C)	1,346	97.89%
Complete and single-copy BUSCOs (S)	1,089	79.20%
Complete and duplicated BUSCOs (D)	257	18.69%
Fragmented BUSCOs (F)	8	0.58%
Missing BUSCOs (M)	21	1.53%
Total BUSCO groups searched	1,375	100.00%

Genome annotation

We subsequently annotated repetitive elements and protein-coding genes for the final genome assembly. We identified 40.20% of the genome as transposable elements (TEs),

which were categorized as long terminal repeat retrotransposons (LTR-RTs) (20.52%), LINE (0.28%) and DNA transposons (15.88%). LTRs formed the most abundant category of TEs, with LTR/*Copia* and LTR/*Gypsy* occupying 3.88% and 11.31%, respectively (Table 4).

A total of 39,008 protein-coding genes were annotated with high confidence using a comprehensive strategy that combined homology-based searches, transcriptome-based predictions, and ab initio prediction. The average length for total gene regions, coding sequence (CDS) and intron sequence are 3558.82, 1093.61 and 1930.11 bp, respectively (Table 5). We further evaluated the quality of gene prediction by BUSCO and found that 1,557 out of the 1,614 (96.47%) highly conserved core proteins in the Embryophyta lineage were present in our gene annotation, of which 1300 (80.54%) were single-copy genes and 257 (15.92%) were duplicated. For the remaining conserved genes, 20 (1.24%) had fragmented matches and 37 (2.29%) were missing.

Among the predicted protein-coding genes, 93.96% could be annotated through at least one of the following protein-related databases: Pfam (67.10%), NR (90.62%), Interproscan (87.05%), KEGG (28.76%), Swiss-Prot (69.63%), KOG (79.65%), COG (32.28%), TrEMBLE (91.92%) and GO databases (71.01%) (Table 6).

Table 4. Statistics of repeat sequences in *P. lasiocarpa* genome.

TE type	Total size (bp)	Percentage of genome (%)
Class I: Retrotransposon	87,353,143	20.80
LTR Retrotransposon	63,727,493	20.52
Copia	16,263,189	3.88
Gypsy	47,464,304	11.31
unkonwn	22,341,447	5.33
LINE	1,184,497	0.28
Class II: DNA transposon	66,604,631	15.88
Helitron	37,895,370	9.03
CACTA	7,305,033	1.74
Mutator	13,665,986	3.26
PIF_Harbinger	2,166,554	0.52
Tcl_Mariner	393,947	0.09

hAT	5,177,741	1.23
Unclassified	14,723,521	3.51
Total	168,681,295	40.2

Table 5. Statistics of protein-coding genes

Feature	number/length
Gene number	39,008
Average gene length (bp)	3558.82
Mean exons number per mRNA	4.86
Mean CDS number per mRNA	4.74
Average CDS length (bp)	1093.61
Average intron length (bp)	1930.11

Table 6. Statistics of protein-coding gene functional annotation

Database	Number	Percentage (%)
Pfam	26,176	67.10
Interproscan	33,957	87.05
KEGG	11,219	28.76
NR	35,348	90.62
Swiss-Prot	27,161	69.63
KOG	31,070	79.65
COG	12,590	32.28
Tremble	35,855	91.92
GO	27,698	71.01
Unannotated	2,355	6.04

Materials and Methods

Plant materials and genome sequencing

One wild *P. lasiocarpa* individual (109.74E, 30.18N) was selected to harvest fresh young leaves and stems for obtaining high-quality genome assembly. Genomic DNA

was extracted from fresh mature leaves using a DNeasy® Tissue Kit (QIAGEN). For the short-read sequencing, 150 bp paired-end libraries with an insert size of 350 bp were constructed and sequenced on Illumina HiSeq X Ten platform. For the long-read sequencing, libraries for Nanopore long reads sequencing were built using large (>20 kb) DNA fragments with the Ligation Sequencing Kit 1D (SQK-LSK109), and sequenced using the PromethION platform (Oxford Nanopore Technologies). For the Hi-C experiment, the libraries were constructed from about 3g of fresh and young leaves and prepared with DpnII restriction enzyme, followed by sequencing on the Illumina NovaSeq platform. To assist gene annotation, total RNAs from fresh young leaves and stems were extracted with CTAB procedure to prepare RNA-seq libraries, which were sequenced on Illumina HiSeq X Ten platform.

Genome assembly and quality assessment

The Illumina short reads were first used to estimate the genome size of *P. lasiocarpa* via a 17-bp k-mer frequency analysis with Jellyfish (v2.3.0)[1]. NextDenovo (v2.0-beta.1, https://github.com/Nextomics/NextDenovo) was then used for the preliminary sequence assembly based on the Nanopore long reads. The raw long reads were first error-corrected via NextCorrect with parameters "reads_cutoff=1k, seed_cutoff=30k", and then assembled via NextGraph with default parameters. To improve the quality of the assembly, corrected ONT long reads (three rounds) and cleaned Illumina short reads (four rounds) were used to polish assembly by Racon v1.3.1[2] and Nextpolish v1.0.5[3], separately. The redundant sequences were subsequently removed by using perge_haplotigs v1.1.1[4] and the obtained genome assemblies were checked for DNA contamination by searching against the NCBI non-redundant nucleotide database (Nt) using BLASTN, with an E-value cutoff of 1e-5. Then, BUSCO (v4.0.5, embryophyta_odb10 download at 16-Oct-2020)[5] with default settings was applied to the assessment of assembly integrity.

The draft assembly was further scaffolded using Hi-C reads. Briefly, the Hi-C reads were filtered by fastp v0.20.0[6] with same parameters described above. The clean reads were then aligned into the draft assembled sequences using bowtie2 v2.3.2[7] with

parameters '-end-to-end, -very-sensitive -L 30'. The mapped Hi-C reads were processed to obtain the valid reads pairs by HiC-Pro v2.11.4[8]. Scaffolds were anchored into 19 pseudo-chromosomes using LACHESIS[9] with parameters CLUSTER MIN RE SITES=100, CLUSTER MAX LINK DENSITY=2.5, CLUSTER NONINFORMATIVE RATIO=1.4, ORDER MIN N RES IN TRUNK=60, ORDER MIN N RES IN SHREDS=60, and then followed by manual correction.

Gene prediction and functional annotation

Before gene prediction, preliminary TE annotation was performed by EDTA v1.9.3[10] pipeline and TEs annotated as LTR/unknown were re-classified using TEsorter v1.2.5[11] . The EDTA- constructed TE libraries were applied to mask the whole genome sequences using RepeatMasker v4.10[12].

We integrated three strategies including homology-based prediction, transcriptome-based prediction, and *ab initio* prediction to predict the protein-coding genes. To perform homology-based prediction, we aligned the protein sequences from six species (*Populus trichocarpa*, *Populus euphratica*, *Salix brachista*, *Salix purpurea*, *Arabidopsis thaliana* and *Vitis vinifera*) to *P. lasiocarpa* genome by TBLASTN[13], and parsed the resultant alignments for homolog predictions using Genewise v2.4.1[14]. Furthermore, these assembled transcripts based on both genome-free and genome-guided using Trinity v2.8.4[15] were aligned to the genome assembly using PASA v2.4.1 to conduct transcriptome-based predictions. Augustus[16] with default parameters was used to incorporate the homology- and transcriptome-based gene models for *ab initio* gene prediction. In the end, all above gene models were integrated into a comprehensive gene set using EvidenceModeler v1.1.1[17] which was further updated for three rounds using PASA.

Gene functional annotation was conducted based on BLAST searches with 1e-5 E-value cutoff against four well-known public protein databases: the protein families database (Pfam), the NCBI non-redundant protein database (NR), the interproscan database, the KEGG database, the Swiss-Prot protein database, the Eukaryotic Orthologous Groups of proteins (KOG), the Translated European Molecular Biology

Laboratory (TrEMBL) database and the Gene Ontology (GO) database. The putative domains and GO terms of *P. lasiocarpa* genes were identified using the InterProScan program with default parameters.

References

1. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).
2. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737-746 (2017).
3. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253-2255 (2020).
4. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *Bmc Bioinformatics* **19**(2018).
5. Simao FA, Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
6. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884-890 (2018).
7. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-9 (2012).
8. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
9. Burton, J.N. *et al.* Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119-25 (2013).
10. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**, 275 (2019).
11. Zhang, R.-G. *et al.* TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* **9**, uhac017 (2022).
12. Tempel, S. Using and understanding RepeatMasker. *Methods Mol Biol* **859**, 29-51 (2012).
13. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1-9 (2009).
14. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome research* **14**, 988-995 (2004).
15. Haas, B.J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494-1512 (2013).
16. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* **33**, W465-W467 (2005).
17. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1-22 (2008).