A Chromosome-Level Genome Assembly of *Dendrobium Huoshanense* Using Long Reads and Hi-C Data

Bangxing Han[1,*,†], Yi Jing[2], Jun Dai[1], Tao Zheng[2,3,†], Fangli Gu[1], Qun Zhao[1], Fucheng Zhu[1], Xiangwen Song[1], Hui Deng[1], Peipei Wei[1], Cheng Song[1], Dong Liu[1], Xueping Jiang[1], Fang Wang[1], Yanjun Chen[1], Chuanbo Sun[1], Houjun Yao[1], Li Zhang[1], Naidong Chen[1], Shaotong Chen[1], Xiaoli Li[1], Yuan Wei[4], Zhen Ouyang[4], Hui Yan[5], Jiangjie Lu[6], Huizhong Wang[6], Lanping Guo[7], Lingdong Kong[8], Jing Zhao[9], Shaoping Li[9], Lifen Luo[10], Karsten Kristiansen[3], Zhan Feng[2], Silong Sun[2], Cunwu Chen[1,*], Zhen Yue[2,*], and Naifu Chen[1,*]

[1]Anhui Province Traditional Chinese Medicine Resource Protection and Sustainable Utilization Engineering Laboratory, West Anhui University, Liu'an, 230036, China

[2]BGI Genomics, BGI-Shenzhen, Shenzhen, 518083, China

[3]Department of Biology, University of Copenhagen

[4]School of Pharmacy, Jiangsu University, Zhenjiang, 212013, China

[5]School of Biotechnology, Jiangsu University of Science and Technology, Zhenjiang, 212032, China

[6]Zhejiang Provincial Key Laboratory for Genetic Improvement and Quality Control of Medicinal Plants, College of Life and Environmental Science, Hangzhou Normal University, Hangzhou, 311121, China

[7]State Key Laboratory Breeding Base of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, 100700, China

[8]School of Life Sciences, Nanjing University, Nanjing, 210093, China

[9]State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, 440499, Macao

[10]Shenzhen Maternity and Child Healthcare Hospital, Shenzhen

*Corresponding authors: E-mails: hanbx1978@sina.com; cunwuchen@126.com; yuezhen@genomics.cn; cnf505@126

Accepted: 7 October 2020

[†] These authors contributed equally to this work.

Abstract

Dendrobium huoshanense is used to treat various diseases in traditional Chinese medicine. Recent studies have identified active components. However, the lack of genomic data limits research on the biosynthesis and application of these therapeutic ingredients. To address this issue, we generated the first chromosome-level genome assembly and annotation of *D. huoshanense*. We integrated PacBio sequencing data, Illumina paired-end sequencing data, and Hi-C sequencing data to assemble a 1.285 Gb genome, with contig and scaffold N50 lengths of 598 kb and 71.79 Mb, respectively. We annotated 21,070 protein-coding genes and 0.96 Gb transposable elements, constituting 74.92% of the whole assembly. In addition, we identified 252 genes responsible for polysaccharide biosynthesis by Kyoto Encyclopedia of Genes and Genomes functional annotation. Our data provide a basis for further functional studies, particularly those focused on genes related to glycan biosynthesis and metabolism, and have implications for both conservation and medicine.

Key words: orchid, genome, de novo assembly, Hi-C assembly, annotation.

Significance

Polysaccharides and alkaloids were identified as the active ingredients for the therapeutic effects of *Dendrobium huoshanense*. But the biosynthetic pathways by which they are generated remain poorly understood. In this article, we report a chromosome-level genome and a set of high-quality genes of *Dendrobium huoshanense*. Furthermore, we identified 252 genes responsible for polysaccharide biosynthesis by Kyoto Encyclopedia of Genes and Genomes (KEGG) functional annotation. The construction of the genomic architecture of a medically important orchid will accelerate genomic and medical studies of this species and Orchidaceae.

Introduction

Dendrobium is the second largest genus in Orchidaceae (Chaudhary et al. 2012). The genus includes over 1,000 species widely distributed across tropical and subtropical regions of Asia and Oceania. In China, 74 species and two varieties of *Dendrobium* have been described (Xu et al. 2006). Among them, *Dendrobium huoshanense*, *Dendrobium officinale*, and *Dendrobium nobile* have established therapeutic value (Ng et al. 2012). *Dendrobium huoshanense* is one of the most valuable traditional Chinese herbal medicines. This variety grows exclusively in Huoshan County in western Anhui Province. The plant has been listed as threatened and endangered in the Convention on International Trade in Endangered Species of Wild Fauna and Flora (Jin et al. 2016). *Dendrobium huoshanense* harbors a variety of characteristic compounds (polysaccharides, terpenoids, stilbenoids, alkaloids, flavonoids, etc.) with functions in oxidation resistance, immunity, liver protection, and tumor suppression (Xu et al. 2013; Zha et al. 2014; Xie et al. 2016).

In traditional Chinese medicine, *D. huoshanense* had been used to treat various diseases as nourishing the stomach and nourishing Yin (Ng et al. 2012; Veronika et al. 2017). Recent studies have identified the active ingredients responsible for the therapeutic effects. Despite extensive evaluations of the chemical and pharmacological properties of *D. huoshanense* alkaloids (Li Juan et al. 2011), the biosynthetic pathways by which they are generated remain poorly understood. Analyses of the genome sequence and expressed sequence tags could improve our understanding of the biological mechanisms of action of natural active ingredients (Li Ying et al. 2010). The genomes of *D. officinale* and *D. catenatum* have recently been sequenced (Yan et al. 2015; Zhang et al. 2016). The assembled genomes are 1.35 Gb and 1.01 Gb, respectively and provide a basis for genome data mining aimed at elucidating biosynthetic pathways for medicinal polysaccharides and alkaloids. Although *D. huoshanense* is regarded as the species with the highest therapeutic value within the genus, the genome has not yet been published. In this article, we describe the plant material and full data sets used to assemble, annotate, and validate the *D. huoshanense* reference genome. First, PacBio sequencing data and Illumina whole-genome shotgun sequencing data were used for genome assembly. Second, we assessed the quality of the genome assembly by using data sets from the most recent version of BUSCO (Simao et al. 2015). Third, we annotated the assembled *D. huoshanense* genome. The assembled genome and annotation for *D. huoshanense* will accelerate research on this valuable plant.

Materials and Methods

Plant Material, DNA Extraction, and DNA Sequencing Library

Seeds of *D. huoshanense* collected from the *D. huoshanense* Conservation Center in Huoshan County were cultivated in sterile culture medium for 200 days. The leaves and root of a mature healthy plant were collected and stored at −80°C prior to DNA sequencing. The cetyltrimethyl ammonium bromide method was employed to isolate genomic DNA. The extracted DNA was used to construct a 20 kb PacBio library, a 500 bp Illumina paired-end library, and a 350 bp Hi-C library.

Whole-Genome Shotgun, Single-Molecular Real-Time, and Hi-C Sequencing

A 500 bp Illumina paired-end library was sequenced using the Illumina HiSeq X-Ten DNA Sequencer. The raw data were filtered using SOAPnuke1.5.6 (https://github.com/BGI-flexlab/SOAPnuke), with the following parameter settings to remove adaptor sequences and low-quality reads: -n 0.01 -l 20 -q 0.1 -i -Q 2 -G -M 2 -A 0.5 -d. A 20 kb PacBio library was constructed for sequencing using eight single-molecular real-time (SMRT) cells on the PacBio (sequel) platform. Raw reads were processed using the SMRT pipeline with a minimum read quality of 0.8. A 350 bp Hi-C sequencing library was constructed and sequenced using the MGISEQ-2000 DNA Sequencer. The clean data, after the removal of duplicates using Juicer (Durand et al. 2016), were processed using 3D-DNA (Dudchenko et al. 2017) for integration into the *D. huoshanense* genome assembly.

Genome Size Estimation and Genome Assembly

The genome size of *D. huoshanense* was estimated by a k-mer analysis of shotgun sequencing data using Jellyfish and KmerFreq v5.0 (Marçais and Kingsford 2011). The estimated genome size was 1.29 Gb (supplementary fig. 1, Supplementary Material online). The *D. huoshanense* genome

was assembled using PacBio long-read sequencing data, followed by the integration of Illumina paired-end data and Hi-C sequencing data. First, the PacBio sequencing data were de novo assembled into contigs using smartdenovo (Liu et al. 2020) after a correction process with Canu (Koren et al. 2017) with the following parameter settings: minReadLength > 3,000 and minOverlapLength > 500. Next, Pilon (Walker et al. 2014) was used to improve the accuracy of the genome assembly by integrating Illumina sequencing data. Then, purge_haplotigs (Roach et al. 2018) was used with the parameters contigcov -l 5 -m 80 -h 190 and purge -a 65 to identify and delete duplicate contigs resulting from heterozygosity in the plant material. Finally, Hi-C (Lieberman-Aiden et al. 2009; Durand et al. 2016; Dudchenko et al. 2017) technology was used to anchor primary contigs to pseudo-molecules and remove redundancy.

Gene and Transposable Element Annotation

A combination of de novo prediction and homology-based methods was used to predict protein-coding genes. De novo prediction was performed using Semi-HMM-based Nucleic Acid Parser (Johnson et al. 2008) and AUGUSTUS (Stanke et al. 2006) trained with de novo assembled transcripts collected from four organs of *D. huoshanense* (the root, stem, leaf, and flower). Homology-based methods were based on the detection of homologous gene sets of five species, *Arabidopsis thaliana*, *D. catenatum*, *Apostasia shenzhenica*, *Phalaenopsis equestris*, and *Oryza sativa* in the *D. huoshanense* genome. After masking repetitive elements using RepBase in the genome assembly, MAKER (Holt and Yandell 2011) was used to integrate the de novo and homology-based prediction results. Functional annotation of the gene set was performed using Blast v2.2.31 (Altschul et al. 1990) to compare the genes with eight protein databases, including SwissProt, TrEMBL, Kyoto Encyclopedia of Genes and Genomes (KEGG), InterPro, NR, KOG, and GO. Both de novo prediction and homology-based prediction methods were used to detect transposable elements (TEs). For de novo prediction, a repetitive sequence data set was constructed using RepeatModeler (Tarailo-Graovac and Chen 2009). Then, RepeatMasker was used to search this data set for TEs. In the homology-based method, RepeatMasker v4.0.7 and RepeatProteinMask v4.0.7 were used to identify TEs by aligning the genome assembly to RepBase v21.12 (Bao et al. 2015).

Assessment of Genome and Gene Quality

Benchmarking Universal Single-Copy Orthologs (Simao et al. 2015) (BUSCO v3) with a total of 1,375 ortholog groups from the Embryophyta Dataset was used to assess the completeness of the genome assembly and gene sets predicted.

Results and Discussion

Genome Assembly and Gene Annotation

We generated 135.63 Gb whole-genome shotgun sequencing data; 139.15 Gb single-molecule real-time sequencing data; and 179.51 Gb Hi-C sequencing data with 105-fold, 108-fold, and 139-fold coverage (supplementary table 1, Supplementary Material online). The final genome assembly was 1.285 Gb in length, with a Contig N50 of 598 kb and Scaffold N50 of 71.79 Mb (table 1). We aligned the filtered reads to the assembled genome sequence using Burrows-Wheeler-Alignment Tool (Li and Durbin 2009) and calculated the base number and percentage of bases with different frequency depths in the genome. The Guanine and Cytosine content and average sequencing depth were approximately 38% and 60×, respectively (supplementary fig. 2, Supplementary Material online). We also generated a sequencing depth plot and found that the percentage of sequences with a depth of less than 10 was lower than 5% (supplementary fig. 3, Supplementary Material online).

We predicted 21,070 genes, with an average mRNA length of 9,877 bp, an average Coding DNA Sequence length of 1,202 bp, and an average intron length of 2,206 bp. A total of 20,904 genes were functionally annotated, accounting for 99.21% of the predicted genes. We also identified 1,495 noncoding RNA genes from the assembly. Furthermore, we

Table 1
Summary of the genome assembly and annotation tables

Genome assembly	Estimated genome size	1.29 Gb	
	Guanine and Cytosine content	38%	
	N50 length (contig)	598 kb	
	Longest contig	6.11 Mb	
	Total length of contigs	1.28 Gb	
	N50 length (scaffold)	71.79 Mb	
	Longest scaffold	100.20 Mb	
	Total length of scaffolds	1.29 Gb	
Transposable elements	Annotation	Percent (%)	Total length
	DNA	5.56	71.48 Mb
	LINE	12.04	154.75 Mb
	SINE	0.01	131.45 kb
	LTR	65.53	842.36 Mb
	Other	0.00	9.35 kb
	Unknown	4.38	56.29 Mb
	Total	74.92	0.96 Gb
Protein-coding genes	Predicted genes	21,070	
	Average transcript length (bp)	9,877.52	
	Average coding sequence length (bp)	1,202.62	
	Average exon length (bp)	270.66	
	Average intron length (bp)	2206.22	
	Functionally annotated	20,904	

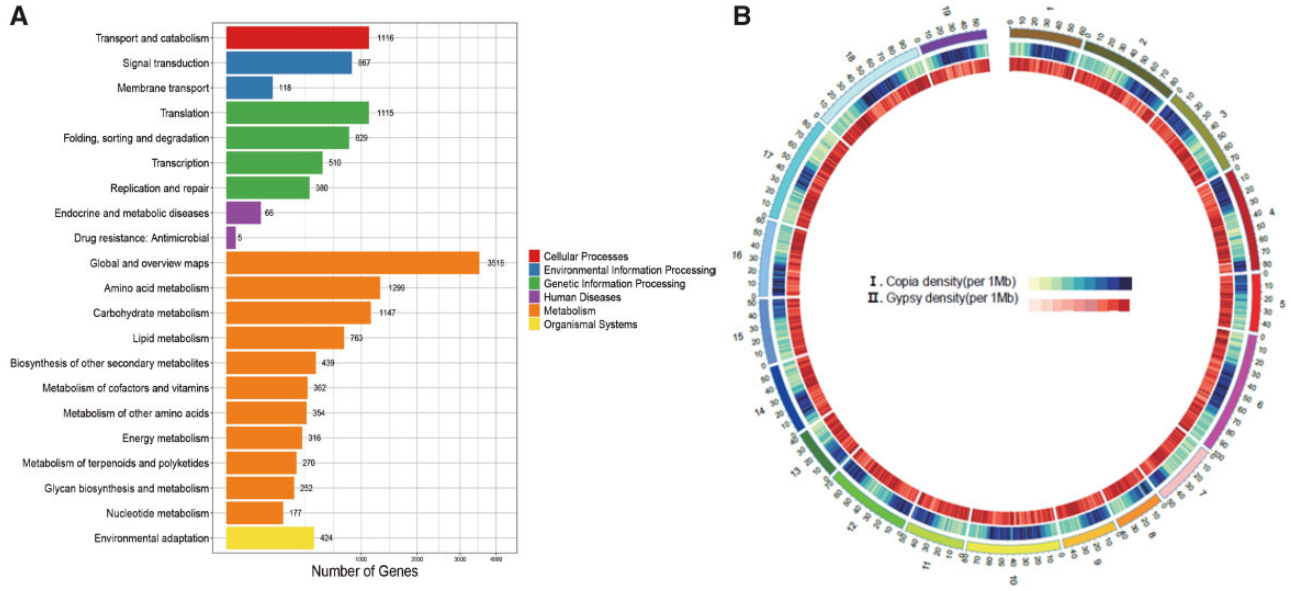

FIG. 1.—(*A*) Functional classification of *Dendrobium huoshanense* genes using the KEGG database. (*B*) *Copia* and *Gypsy* element distributions along the chromosomes of *D. huoshanense*. *Copia* and *Gypsy* densities represent the proportions of *Copia* and *Gypsy* elements within 1 Mb intervals.

functionally classified 14,552 (69.07%) *D. huoshanense* genes using KEGG. In particular, we identified 252 genes related to glycan biosynthesis and metabolism (fig. 1*A*). These findings are consistent with previous experimental studies indicating the importance of polysaccharides in *D. huoshanense* (Xu et al. 2013; Zha et al. 2014).

We assessed the *D. huoshanense* genome assembly and quality of annotated gene models using version 3 of Benchmarking Universal Single-Copy Orthologs (Simao et al. 2015) (BUSCO) with a total of 1,375 ortholog groups from the Embryophyta Dataset. The assembled genome and annotated genes had greater than 86.6% and 78.1% completeness (supplementary table 2, Supplementary Material online), indicating a relatively complete genome assembly and gene prediction.

Transposable Element and Long Terminal Repeat Distributions

We identified 1.02 Gb repetitive elements and 0.96 Gb TEs in the assembled *D. huoshanense* genome. The TEs are summarized in supplementary table 3, Supplementary Material online. In brief, 65.53% of the assembled genome was long terminal repeats, 85% of which belonged to subtypes *Copia* and *Gypsy*. *Gypsy* elements were distributed evenly along the chromosomes, while *Copia* showed a biased distribution. On seven chromosomes, *Copia* elements were concentrated at one end, whereas on the remaining 12 chromosomes, they were concentrated near the chromosome center (fig. 1*B*).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2017YFC1700701), Major Program of Increase or Decrease in the Central Government (2060302), Anhui Provincial Science Fund for Distinguished Young Scholars (1808085J17), China Agriculture Research System (CARS-21), and the University Synergy Innovation Program of Anhui Province (GXXT-2019-043, GXXT-2019-049).

Data Availability

The data have been deposited in the NCBI GenBank database under the BioProject Number PRJNA597621. It was also deposited in the CNGBdb with accession code CNP0000830.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11.

Chaudhary B, Chattopadhyay P, Verma N, Banerjee N. 2012. Understanding the phylomorphological implications of pollinia from Dendrobium (Orchidaceae). Am J Plant Sci. 03(06):816–828.

Dudchenko O, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356(6333):92–95.

Durand NC, et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3(1):95–98.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12(1):491.

Jin Q, et al. 2016. Metabolic analysis of medicinal *Dendrobium officinale* and *Dendrobium huoshanense* during different growth years. PLoS One 11(1):e0146607.

Johnson AD, et al. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24(24):2938–2939.
Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760.
Li J, Li S, Huang D, Zhao X, Cai G. 2011. Advances in the of resources, constituents and pharmacological effects of *Dendrobium officinale*. JKDS Review T. 29:74–79.
Li Y, et al. 2010. EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. BMC Genomics 11(1):268.
Lieberman-Aiden E, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6):764–770.
Ng TB, et al. 2012. Review of research on Dendrobium, a prized folk medicine. Appl Microbiol Biotechnol. 93(5):1795–1803.
Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19(1):460.
Liu H, Wu S, Li A, Ruan J. 2020. Ultra-fast de novo assembler using long noisy reads. Preprints 2020090207.
Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34(Web Server):W435–W439.
Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocol Bioinformatics Chapter 4: Unit 4.10.
Veronika C, Frederic B, Annelise L. 2017. Dendrobium: sources of active ingredients to treat age-related pathologies. Aging Dis. 8(6):827–849.
Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.
Xie S-Z, et al. 2016. Polysaccharide of *Dendrobium huoshanense* activates macrophages via toll-like receptor 4-mediated signaling pathways. Carbohydr Polym. 146:292–300.
Xu H, Wang Z, Ding X, Zhou K, Xu L. 2006. Differentiation of Dendrobium species used as "Huangcao Shihu" by rDNA ITS sequence analysis. Planta Med. 72(1):89–92.
Xu J, et al. 2013. Chemistry, bioactivity and quality control of Dendrobium, a commonly used tonic herb in traditional Chinese medicine. Phytochem Rev. 12(2):341–367.
Yan L, et al. 2015. The genome of *Dendrobium officinale* illuminates the biology of the important traditional Chinese orchid herb. Mol Plant. 8(6):922–934.
Zha X-Q, et al. 2014. Immunoregulatory activities of *Dendrobium huoshanense* polysaccharides in mouse intestine, spleen and liver. Int J Biol Macromol. 64:377–382.
Zhang G-Q, et al. 2016. The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. Sci Rep. 6:19029.