


Chromosome-scale assembly of the *Sparassis latifolia* genome obtained using long-read and Hi-C sequencing

Chi Yang ^{1,2}, Lu Ma,^{1,2} Donglai Xiao,^{1,2} Xiaoyu Liu,^{1,2} Xiaoling Jiang,^{1,2} Zhenghe Ying,^{1,2} and Yanquan Lin^{1,2,*}

¹Institute of Edible Mushroom, Fujian Academy of Agricultural Sciences, Fuzhou 350014, China

²National and Local Joint Engineering Research Center for Breeding & Cultivation of Featured Edible Mushroom, Fujian Academy of Agricultural Sciences, Fuzhou 350014, China

*Corresponding author: Institute of Edible Mushroom, Fujian Academy of Agricultural Sciences, Fuzhou 350014, China. Email: lyq-406@163.com

Abstract

Sparassis latifolia is a valuable edible mushroom cultivated in China. In 2018, our research group reported an incomplete and low-quality genome of *S. latifolia* obtained by Illumina HiSeq 2500 sequencing. These limitations in the available genome have constrained genetic and genomic studies in this mushroom resource. Herein, an updated draft genome sequence of *S. latifolia* was generated by Oxford Nanopore sequencing and the high-through chromosome conformation capture (Hi-C) technique. A total of 8.24 Gb of Oxford Nanopore long reads representing ~198.08X coverage of the *S. latifolia* genome were generated. Subsequently, a high-quality genome of 41.41 Mb, with scaffold and contig N50 sizes of 3.31 and 1.51 Mb, respectively, was assembled. Hi-C scaffolding of the genome resulted in 12 pseudochromosomes containing 93.56% of the bases in the assembled genome. Genome annotation further revealed that 17.47% of the genome was composed of repetitive sequences. In addition, 13,103 protein-coding genes were predicted, among which 98.72% were functionally annotated. BUSCO assay results further revealed that there were 92.07% complete BUSCOs. The improved chromosome-scale assembly and genome features described here will aid further molecular elucidation of various traits, breeding of *S. latifolia*, and evolutionary studies with related taxa.

Keywords: *Sparassis latifolia*; genome; Hi-C sequencing; Oxford Nanopore sequencing

Introduction

Sparassis latifolia Y. C. Dai et Z. Wang (*Sparassidaceae*, *Polyporales*, and *Agaricomycetes*), collections from Asian *Sparassis* (Dai et al. 2006) exhibit diverse biological and pharmacologic activities (Thi Nhu Ngoc et al. 2018; Uchida et al. 2019; Wang et al. 2019). *S. latifolia* is the commonly cultivated *Sparassis* species in China (Yang et al. 2017). To date, total fresh fruit production in Chinese factories is over 20 tons/d. Despite the significant economic and medical value of *S. latifolia*, its genetic information remains limited.

In 2018, our group reported that the *S. latifolia* has a size of 48.13 megabases (Mb) and 12,471 predicted genes (Xiao et al. 2018). Based on this genome sequence, we explored the mechanism of light response and primordia formation of *S. latifolia* (Xiao et al. 2017; Yang et al. 2019). The genome of *S. latifolia* was also deposited at the Joint Genome Institute (JGI, project Id: 1105659) and Genebank (PRJNA562364), consisting of 35.66 and 39.32 Mb genome lengths, respectively. *S. crispa* is another *Sparassis* species with a reported genome size of 39.0 Mb encoding for 13,157 predicted genes (Kiyama et al. 2018). Another report showed the *S. crispa* genome is 40.406 Mb in length, and contains 18,917 predicted contigs. They also revealed that the

complete mitochondrial genome of *S. crispa* is 139,253 bp long, containing 47 genes (Bashir et al. 2020). However, the assembly level of all these studies was under chromosome-scale and the contig number were still high (3848 and 184, respectively).

Oxford Nanopore sequencing (ONT) reads the nucleotide sequence by detecting changes in electrical current signals when a DNA molecule is forced through a biological nanopore (Chen et al. 2020a). Compared to the short reads generated by Illumina sequencing, the much longer reads produced by ONT span larger genome regions, resulting in more complete assemblies (Jain et al. 2016). ONT also significantly improves the assembly completeness as compared to the assembly generated using Illumina reads only (Murigneux et al. 2020). Besides, research on signal simulation of nanopore sequences is highly desirable for method developments of nanopore sequencing applications.

This study aimed to assemble a high-quality chromosome-scale reference genome of *S. latifolia* using ONT combined with high-through chromosome conformation capture (Hi-C) scaffolding. The improved reference genome will facilitate molecular breeding of *S. latifolia* and advance our understanding of its genetics and evolution.

Received: January 19, 2021. Accepted: May 10, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Materials and methods

DNA preparation and sequencing

The SP-C strain of *S. latifolia* was grown and maintained on potato dextrose agar (PDA) slants and preserved at the Institute of Edible Mushroom, Fujian Academy of Agricultural Sciences (Fuzhou, China). Genomic DNA was isolated from the mycelium of *S. latifolia* by the cetyl-trimethyl ammonium bromide method (Biel and Parrish 1986). The gDNA was then size-selected and sequenced on an Oxford Nanopore PromethION system by BioMarker Technology Co., Ltd. (Beijing, China). Libraries were prepared using the Ligation Sequencing Kit (Oxford Nanopore Technologies Inc., Oxford, UK). The prepared libraries were sequenced on the PromethION device. The reads were base-called with Albacore (v2.3.1) using the following options: barcoding (to enable demultiplexing) and disable_filtering (to include low-quality reads) (Sošić and Šikić 2017). The high-accuracy base-calling mode was used to base-call the signal in FAST5 files and outputted FASTQ files. Low-quality reads, reads with adapter sequences, and reads shorter than 2000 nt were filtered out before assembly. The reads quality and reads' statistics were analyzed by using NanoStat v1.2.1.

Genome assembly

The *S. latifolia* genome was assembled using NECAT v0.01 (Chen et al. 2020b), then polished by Pilon (Walker et al. 2014) with Illumina short reads, to further eliminate Indel and SNP (single nucleotide polymorphism) errors. BUSCO v3 assessment using single-copy orthologous genes was subsequently performed to confirm the quality of the assembled genome (Simão et al. 2015). The specific BUSCO gene set was fungi_odb9, which contains 290 conserved core genes of fungi.

Hi-C library construction and assembly of the chromosome

Genomic DNA was isolated from the mycelium of *S. latifolia*. Genomic DNA extraction, library preparation, and sequencing were carried out by Biomarker Technologies, Beijing, China. Hi-C sequencing libraries were constructed and their concentration and insert size detected using Qubit2.0 and Agilent 2100. Samples with high-quality nuclei were subjected to the Hi-C procedure (Yang et al. 2018). The chromatin was digested using Hind III restriction enzyme and ligated together *in situ* after biotinylation. DNA fragments were subsequently enriched via the interaction of biotin and blunt-end ligation and then subjected to Illumina HiSeq sequencing. Clean reads were mapped to the *S. latifolia* genome using BWA (Li and Durbin 2009) under its default parameters. Paired-end reads were separately mapped to the genome, followed by filtering out of dangling ends, self-annealing sequences, and dumped pairs (Roach et al. 2018). Valid paired-end reads of unique, mapped paired-end reads were collected using HiC-Pro (v2.10) (Servant et al. 2015). The order and direction of scaffolds/contigs were clustered into super scaffolds using LACHESIS (Burton et al. 2013), based on the relationships among valid reads.

Genome annotation

Identification and construction of the *de novo* repeat library were performed by LTR_FINDER v1.05 (Xu and Wang 2007), MITE-Hunter (Han and Wessler 2010), RepeatScout v1.0.5 (Price et al. 2005), and PILER-DF v2.4 (Edgar and Myers 2005). PASTEClassifier (Wicker et al. 2007) was used to classify the database then merged with Repbase's (Jurka et al. 2005) database to generate the final repeat sequence. The RepeatMasker v4.0.6 (Chen 2004) software

was finally used to search the known repeat sequences and map them onto the *de novo* repeat libraries. This step was done to identify novel repeat sequences based on the built repeat sequence database.

The combined use of *ab initio* prediction, homology-based prediction, and transcriptome-assisted prediction was used to identify protein-coding genes. For *ab initio* prediction, Augustus (v3.2.3) (Stanke et al. 2006), Geneid (v1.4) (Alioto et al. 2018), Genescan (v1.0) (Burge and Karlin 1997), GlimmerHMM (v3.04) (Majoros et al. 2004), and SNAP (v2013.11.29) (Korf 2004) software were employed under their default parameters. The GeMoMa v1.3.1 software (Jens et al. 2016) was used for homologous protein-based prediction. RNA-seq data were mapped to the sponge gourd genome using Hisat2 (v2.0.4) and Stringtie (v1.2.3) (Pertea et al. 2016) for transcriptome-based prediction. Amino acid sequences were predicted from the assemblies using TransDecoder (v2.0) (The Broad Institute, Cambridge, MA, USA). The results were integrated using the EVM (v1.1.1) (Haas et al. 2008) software to predict all genes.

The protein sequences were subsequently aligned to protein databases for gene annotation. The databases included gene ontology (GO) (Ashburner et al. 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (<http://www.genome.jp/kegg/>) (Ogata et al. 1999), InterPro (<https://www.ebi.ac.uk/interpro/>) (Jones et al. 2014), Swiss-Prot (<http://www.uniprot.org/>) (Boeckmann et al. 2003), and TrEMBL (<http://www.uniprot.org/>) (Boeckmann et al. 2003). Detection of reliable tRNA positions was accomplished by tRNAscan-SE (v2.0.3) (Lowe and Eddy 1997). Noncoding RNAs (ncRNAs) were predicted through an RFAM (v12.0) (Griffiths-Jones et al. 2005) database search using the Infernal software (v1.0) (Nawrocki and Eddy 2013) under its default parameters.

Comparative genomics analysis

Putative orthologous genes were constructed from two *S. latifolia* (SP-C strain in this study and CCMJ1100 strain in JGI) and one *S. crispa* (Kiyama et al. 2018). The OrthoMCL (Li et al. 2003) software was used to classify the protein sequences and analyze the gene families. The classification involved a statistical analysis of the gene families unique to each strain, the gene families shared by the strains, and single-copy gene families for each strain. The gene families were functionally annotated in the Pfam database and their Venn diagram constructed based on their statistical results. The PAML (Yang 2007) software was then used to calculate Ka/Ks ratios of gene pairs in single-copy gene families. Evolutionary trees based on single-copy gene families were constructed using the phyML 3.0 (Guindon et al. 2010) software to study the evolutionary relationships between species. The nucleotide sequences of the single-copy orthologous group from OrthoMCL clustering were connected to form a supergene. The maximum likelihood method was used to construct a phylogenetic tree. The genome sequence of *Wolfiporia cocos* (strain MD-104) were added as out tree in the phylogenetic tree, which was found to be closed to *S. latifolia* SPC in our previous research (Xiao et al. 2018). Comparisons of SP-C protein sequences with each reference genome were made through BLAST analysis (Altschul et al. 1997). Nucleic acid level crosstalk between the genomes pairs was then obtained based on the position of the homologous genes on the genome sequence and plotted using MCSanX (Wang et al. 2012).

Data availability

Genome assembly was submitted to NCBI under the BioProject accession number PRJNA686158. The genome accession is JAENRS000000000 and the version described in this study is version JAENRS010000000. The RNA-Seq data had been deposited in NCBI under accession GSE173822. Supplementary material is available at figshare: <https://doi.org/10.25387/g3.14450568>.

Results and discussion

Sequencing and assembly of the genome

The *S. latifolia* strain SP-C used for sequence was preserved at the Institute of Edible Mushroom, Fujian Academy of Agricultural Sciences (Fuzhou, China). The sequencing depth was 199.08X (Table 1), which yielded 8.86 Gb of genomic data. Removal of adapters yielded 8.24 Gb of clean data and 14.42 kb of subread N50. The assembled genome had 22 scaffolds, and N50 had significantly increased to 3.31 Mb, compared to 472 scaffolds with N50 of 0.46 Mb (Xiao et al. 2018). Assembly statistics are summarized in Supplementary Table S1. The primary contigs were further polished using the Pilon program (Walker et al. 2014) with Illumina short reads to improve accuracy. The post-correction genome size was 41,412,529 bp, with a contig N50 of 1,509,579 bp. The average GC content in the corrected genome was 51.51%.

Assessment of genomic integrity

Approximately 98.74% of the Illumina resequencing reads were mapped to the assembly (Supplementary Table S2). BUSCO assay results further revealed that there were 92.07% complete BUSCOs (Supplementary Table S3) which indicated that the assembly integrity was adequate.

Hi-C

The Hi-C approach efficiently uses high-throughput sequencing to determine the state of genome folding by measuring the contact frequency between loci pairs (Lieberman-Aiden et al. 2009). Nearly 39.5 million paired-end reads (11.8 Gb) were collected with a GC content of 53.12% and a Q20 ratio (the percentage of clean reads more than 20 bp) of 97.24% (Supplementary Table S4). Hi-C library quality was assessed based on the read mapping ratio and the content of invalid interaction pairs. A high-quality Hi-C library plays a vital role in increasing the final effective data volume, which directly reflects the quality of Hi-C library construction. Invalid interaction pairs mainly include self-circle ligation, dangling ends, re-ligation, and dumped pairs (Belton et al. 2012; Imakaev et al. 2012; Servant et al. 2015). The ratio of

mapped reads and valid interaction pairs was 90.83 and 86.67%, respectively (Supplementary Tables S5 and S6).

Hi-C assembly located 38,744,916 bp genomic sequences, accounting for 93.56% of the total sequence length on the 12 chromosomes. There were 39 corresponding sequences (contigs) accounting for 79.59% of the entire sequence. The sequences mapped to the chromosomes that determined the chromosomal order and direction were 38,744,916 bp long, accounting for 100% of the total length of the sequence. There were 39 corresponding sequences accounting for 100% of the sequence mapped to the chromosome. Detailed sequence distribution of each chromosome is outlined in Table 2.

For the Hi-C assembled chromosomes, the genome was cut into 20 kb bins with equal length. The number of Hi-C read pairs covered between any two bins was then used as the intensity signal of the interaction between the bins to construct a heat map. The heat map (Figure 1) revealed that the genome was divided into multiple chromosomes. The interaction intensity at the diagonal position within each group was higher than that of the off-diagonal position, indicating that the interaction between adjacent sequences (diagonal position) in the Hi-C assembly was high. However, there was a weak interaction signal between non-adjacent sequences (off-diagonal positions), which is consistent with the principle of Hi-C assisted genome assembly. This finding suggests that the genome assembly effect was good.

Genome annotation

The *S. latifolia* genome contained 7.23 Mb repetitive sequences that accounted for 17.47% of the genome (Supplementary Table S7), which was longer than obtained in our previous study (5.19 Mb, 10.79%) (Xiao et al. 2018). Five major types of repeats were detected: class I, class II, potential host gene, SSR, and unknown duplications. Among them, class I comprised the largest proportion (11.11%, total length of 4.60 Mb) followed by the novel repeats (4.54% total length of 1.88 Mb) of the genome.

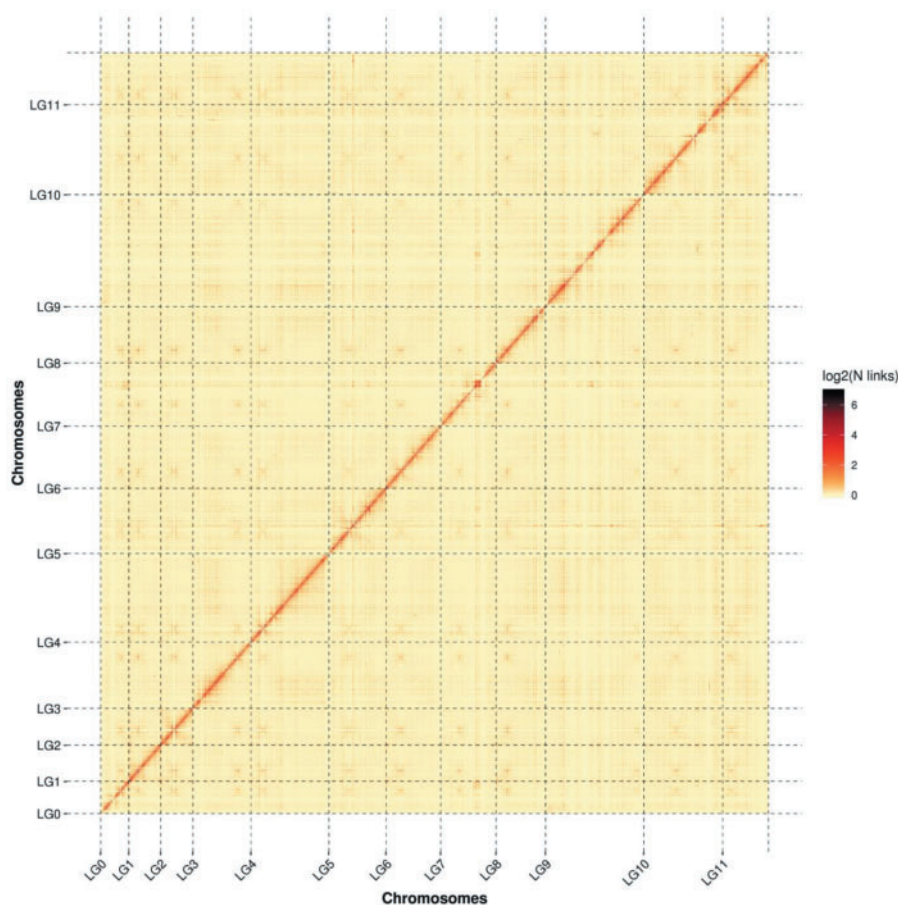
Annotation was done for 13,103 protein-coding genes in *S. latifolia* (Supplementary Table S8). Among them, 12,936 (98.72%) were supported by transcriptome data and homolog prediction. The number of protein-coding genes was higher than that of *S. latifolia* previously reported by our group (12,471) (Xiao et al. 2018) and JGI (Project Id: 1105659). However, it was lower than that of *S. crispa* (13,157) (Kiyama et al. 2018). As shown in Supplementary Table S9, the average length of the predicted genes was 1882.98 bp, while that of their coding sequences was 220.98 bp. In addition, there was an average of 6.2 exons per gene with a length of 238.51 bp per exon. The average intron length was 77.64 bp.

Table 1 Statistics of *Sparassis* taxa genome

Feature	<i>S. latifolia</i> (this study)	<i>S. latifolia</i> (Xiao et al. 2018)	<i>S. latifolia</i> (JGI)	<i>S. crispa</i> (Kiyama et al. 2018)	<i>S. crispa</i> (Bashir et al. 2020)	<i>S. latifolia</i> (Sampled in Larch tree)
Scaffold length (bp)	41,412,529	48,134,914	—	—	—	39,318,455
Scaffold number	22	472	184	—	—	—
Contig number	49	3,848	184	32	18,917	93
GC content (%)	51.51	51.43	—	51.4	—	51.4
Sequencing read coverage depth	199.08 X	601X	74.16x	—	—	—
Number of protein-coding genes	13,103	12,471	12,815	13,157	—	—
Length of genome assembly (Mb)	41.41	48.13	35.66	39	40.41	39.32
Mitochondrial genome (bp)	—	—	133,796	—	139,253	—
Sequencing technology	ONT	Illumina HiSeq 2500	PacBio	PacBio RSII	—	Oxford Nanopore GridION
Accession	PRJNA686158	PRJNA318565	1105659	PRJDB5582	—	PRJNA562364

Table 2 Summary of Hi-C-assisted assembly chromosome lengths

Group	Sequence number	Sequence length (bp)
Lachesis Group0	3	1,652,215
Lachesis Group1	2	1,841,158
Lachesis Group2	2	1,873,959
Lachesis Group3	2	3,372,947
Lachesis Group4	1	4,516,377
Lachesis Group5	3	3,305,915
Lachesis Group6	3	3,178,441
Lachesis Group7	5	3,208,453
Lachesis Group8	4	2,878,786
Lachesis Group9	7	5,687,837
Lachesis Group10	5	4,576,772
Lachesis Group11	2	2,652,056
Total sequences clustered (ratio %)	39 (79.59)	38,744,916 (93.56)
Total sequences ordered and oriented (ratio %)	39 (100)	38,744,916 (100)

**Figure 1** Intensity signal heat map of the Hi-C chromosome. Lachesis Group (LG) means chromosome.

Functional annotation further revealed that 5608, 3341, 7344, 5967, 11,187, 11,252, and 3497 genes were annotated to the KOG, GO, Pfam, Swissprot, TrEMBL, NR, and KEGG databases, respectively. In accord with this, 11,281 (86.09% of the total) genes had at least one hit to the public databases (Supplementary Table S10). In addition, 126 transfer RNAs, 75 ribosomal RNAs, and 36 other noncoding RNAs were identified in the *S. latifolia* genome (Supplementary Table S11). There were also 452 identified pseudogenes with premature stop codons or frameshift mutations (Supplementary Table S12). Nevertheless, this number was significantly higher than previously reported by our group (8 pseudogenes) (Xiao et al. 2018).

Comparison of the genomes of the *Sparassis* taxa

OrthoMCL (Li et al. 2003) was used to classify gene families with single and multiple copies from *Sparassis* taxa, resulting in 909 *S. latifolia* SPC strain-specific genes (Table 3). *S. latifolia* SPC had more common genes with *S. crispa* SCP (10,083) than with *S. latifolia* CCMJ1100 (8813) (Figure 2A). Moreover, phylogenetic analyses revealed that *S. latifolia* SPC was more closely related to *S. crispa* SCP (Figure 2B). Synteny and collinearity analysis between genomes was conducted using MCScanX (Wang et al. 2012) to further characterize the genomic differences between the newly sequenced *S. latifolia* SPC genomes and the other *Sparassis* taxa strains. *S. latifolia* SPC and *S. crispa* SCP genomes

Table 3 Statistics of comparison of the genomes of the *Sparassis* taxa

Species name	Total gene	Cluster gene	Total family	Unique Gene family	Unique gene
<i>S. latifolia</i> SPC	13,103	12,286	10,401	42	909
<i>S. latifolia</i> CCMJ1100	12,815	10,471	9,189	117	2,672
<i>S. crispa</i> SCP	13,157	12,103	10,309	48	1,164

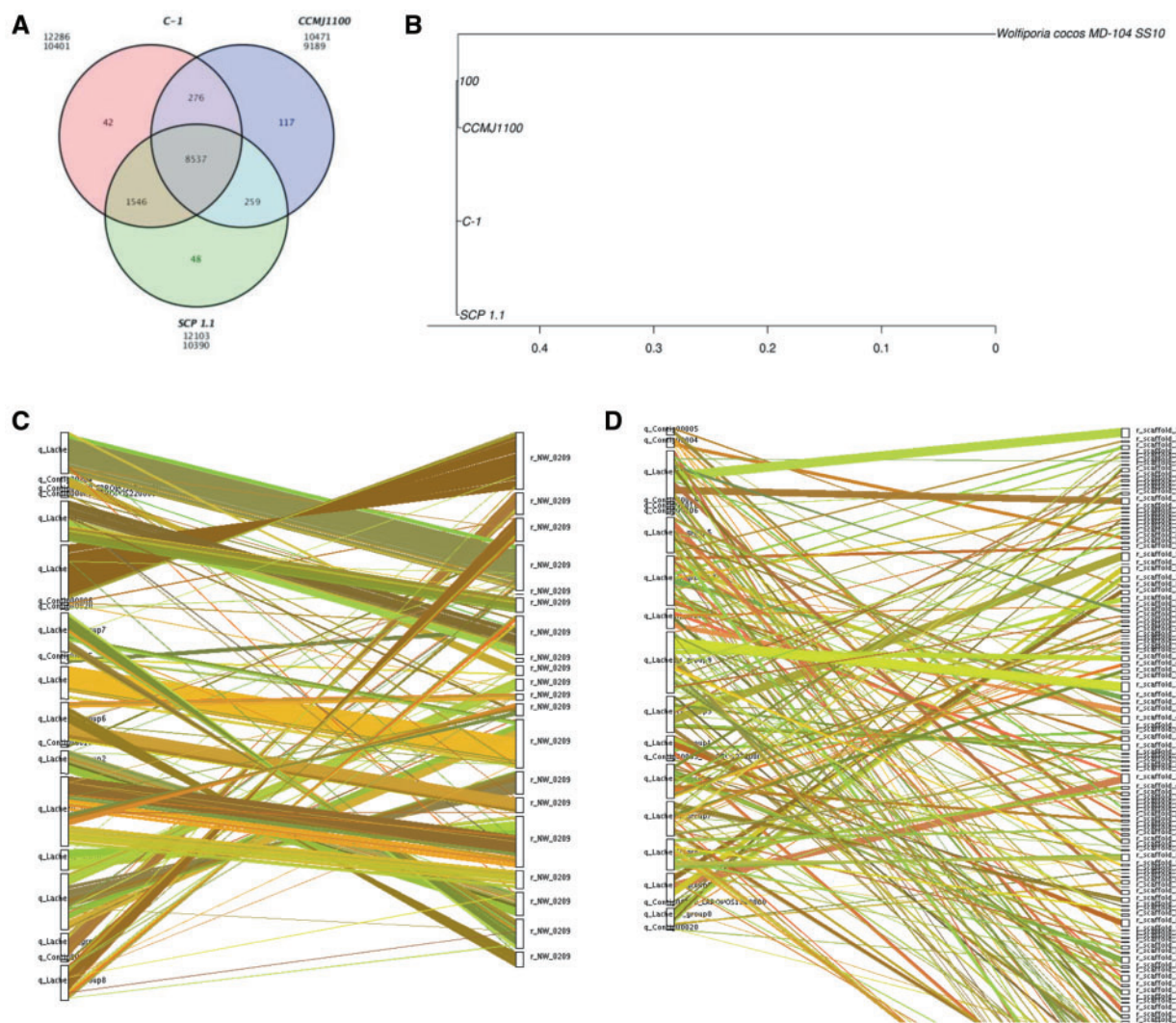


Figure 2 Comparison of the genomes of the *Sparassis* taxa. (A) Gene family annotation Venn diagram. OrthoMCL software was used to classify the protein sequences and analyze the gene families. The gene families were functionally annotated in the Pfam database. (B) Phylogenetic tree of *Sparassis* taxa. The single-copy genes from OrthoMCL clustering were linked into super genes after muscle comparison. Maximum likelihood method was used to construct a phylogenetic tree. The genome sequence of *W. cocos* (strain MD-104) was added as out tree. (C, D) Collinearity of *S. latifolia* SPC with *S. crispa* SCP and *S. latifolia* CCMJ1100. Comparisons of SP-C protein sequences with each reference genome were made through BLAST analysis. Nucleic acid level crosstalk between the genomes pairs was then obtained based on the position of the homologous genes on the genome sequence and plotted using MCSanX.

were found to be highly collinear (Figure 2, C and D). Collinearity analysis further confirmed that the assembled genome was of high quality.

Conclusions

Presented here is the chromosome level assembly of a genome from the *Sparassis* genera. A 41.41 Mb chromosome-level reference genome of *S. latifolia* was assembled, and its 13,103 protein-coding genes were annotated. The improved assembly and genome features described herein will aid further molecular

elucidation of various traits, breeding of *S. latifolia*, and evolutionary studies with related taxa.

Acknowledgments

Conceptualization: C.Y. and Y.L.; Formal analysis: L.M. and Z.Y.; Funding acquisition: C.Y., D.X., and Y.L.; Investigation: L.M., X.L., and X.J.; Methodology: C.Y., L.M., and D.X.; Project administration: L.M. and C.Y.; Software: C.Y.; Supervision: Y.L.; Validation: L.M. and D.X.; Writing—original draft: C.Y.; Writing—review and editing: C.Y. and D.X.

Funding

This work was supported by the Natural Science Foundation of Fujian province of China (2020J011378), the Special Fund for Scientific Research in the Public Interest of Fujian Province (2020R1035003 and 2020R1035005), and Free exploring scientific and technological innovation projects in Fujian Academy of Agricultural Sciences (ZYTS2020013).

Conflicts of interest

None declared.

Literature cited

- Alioto T, Blanco E, Parra G, Guigó R. 2018. Using geneid to identify genes. *Curr Protoc Bioinformatics*. 64:e56.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389–3402.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 25:25–29.
- Bashir KMI, Rheu KM, Kim M-S, Cho M-G. 2020. The complete mitochondrial genome of an edible mushroom, *Sparassis crispa*. *Mitochondrial DNA B Resour*. 5:862–863.
- Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, et al. 2012. Hi-c: a comprehensive technique to capture the conformation of genomes. *Methods*. 58:268–276.
- Biel SW, Parrish FW. 1986. Isolation of DNA from fungal mycelia and sclerotia without use of density gradient ultracentrifugation. *Anal Biochem*. 154:21–55.
- Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, et al. 2003. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*. 31:365–370.
- Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 268:78–94.
- Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, et al. 2013. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol*. 31:1119–1125.
- Chen N. 2004. Using repeatmasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 5:Chapter 4:Unit 4.10.
- Chen W, Zhang P, Song L, Yang J, Han C. 2020a. Simulation of nanopore sequencing signals based on bigru. *Sensors (Basel)*. 20:7244.
- Chen Y, Nie F, Xie S-Q, Zheng Y-F, Bray T, et al. 2020b. Fast and accurate assembly of nanopore reads via progressive error correction and adaptive read selection. *bioRxiv*. 2020. 2002.2001.930107.
- Dai YC, Wang Z, Binder M, Hibbett DS. 2006. Phylogeny and a new species of *sparassis* (polyporales, basidiomycota): evidence from mitochondrial atp6, nuclear rDNA and rpb2 genes. *Mycologia*. 98: 584–592.
- Edgar RC, Myers EW. 2005. Piler: identification and classification of genomic repeats. *Bioinformatics*. 21:i152–i158.
- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*. 33:D121–124.
- Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phylml 3.0. *Syst Biol*. 59:307–321.
- Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, et al. 2008. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biol*. 9:R7.
- Han Y, Wessler SR. 2010. Mite-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res*. 38:e199.
- Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, et al. 2012. Iterative correction of Hi-c data reveals hallmarks of chromosome organization. *Nat Methods*. 9:999–1003.
- Jain M, Olsen HE, Paten B, Akeson M. 2016. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 17:239.
- Jens K, Michael W, Erickson JL, Schattat MH, Jan G, et al. 2016. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res*. 44:e89.
- Jones P, Binns D, Chang HY, Fraser M, Li W, et al. 2014. Interproscan 5: genome-scale protein function classification. *Bioinformatics*. 30:1236–1240.
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110:462–467.
- Kiyama R, Furutani Y, Kawaguchi K, Nakanishi T. 2018. Genome sequence of the cauliflower mushroom *Sparassis crispa* (hanabiratake) and its association with beneficial usage. *Sci Rep*. 8:16053.
- Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics*. 5:59.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 25:1754–1760.
- Li L, Stoekert CJ, Jr, Roos DS. 2003. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13: 2178–2189.
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 326:289–293.
- Lowe TM, Eddy SR. 1997. Trnscan-se: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 25:955–964.
- Majoros WH, Pertea M, Salzberg SL. 2004. Tigrscan and glimmerhmm: two open source ab initio eukaryotic gene-finders. *Bioinformatics*. 20:2878–2879.
- Murigneux V, Rai SK, Furtado A, Bruxner TJC, Tian W, et al. 2020. Comparison of long-read methods for sequencing and assembly of a plant genome. *Gigascience*. 9:giaa146.
- Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 29:2933–2935.
- Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. 1999. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 27:29–34.
- Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with hisat, stringtie and ballgown. *Nat Protoc*. 11:1650–1667.
- Price AL, Jones NC, Pevzner PA. 2005. *De novo* identification of repeat families in large genomes. *Bioinformatics*. 21:i351–358.
- Roach MJ, Schmidt SA, Borneman AR. 2018. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*. 19:460.
- Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, et al. 2015. Hic-pro: an optimized and flexible pipeline for Hi-c data processing. *Genome Biol*. 16:259.
- Simão FA, Waterhouse RM, Panagiotis I, Kriventseva EV, Zdobnov EM. 2015. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31: 3210–3212.

- Šošić M, Šikic M. 2017. Edlib: A c/c++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*. 33:1394–1395.
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, et al. 2006. Augustus: *Ab initio* prediction of alternative transcripts. *Nucleic Acids Res*. 34:W435–W439.
- Thi Nhu Ngoc L, Oh YK, Lee YJ, Lee YC. 2018. Effects of *Sparassis crispa* in medical therapeutics: a systematic review and meta-analysis of randomized controlled trials. *Int J Mol Sci*. 19:1487.
- Uchida M, Horii N, Hasegawa N, Oyanagi E, Yano H, et al. 2019. *Sparassis crispa* intake improves the reduced lipopolysaccharide-induced TNF-alpha production that occurs upon exhaustive exercise in mice. *Nutrients*. 11:2049.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9:e112963.
- Wang Y, Tang H, Debarry JD, Tan X, Li J, et al. 2012. Mcscanx: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 40:e49.
- Wang Z, Liu J, Zhong X, Li J, Wang X, et al. 2019. Rapid characterization of chemical components in edible mushroom *Sparassis crispa* by UPLC-orbitrap MS analysis and potential inhibitory effects on allergic rhinitis. *Molecules*. 24:3014.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 8:973–982.
- Xiao D, Zhang D, Ma L, Wang H, Lin Y. 2017. Preliminary study on differentially expressed genes of *Sparassis latifolia* under light inducing. *Edible Fungi China*. 36:60–63.
- Xiao DL, Ma L, Yang C, Ying ZH, Jiang XL, et al. 2018. *De novo* sequencing of a *Sparassis latifolia* genome and its associated comparative analyses. *Can J Infect Dis Med Microbiol*. 2018: 1857170.
- Xu Z, Wang H. 2007. Ltr_finder: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. 35: W265–W268.
- Yang C, Ma L, Xiao D, Ying Z, Jiang X, et al. 2019. Integration of ATAC-seq and RNA-seq identifies key genes in light-induced primordia formation of *Sparassis latifolia*. *IJMS*. 21:185.
- Yang C, Ma L, Ying ZH, Jiang XL, Lin YQ. 2017. Sequence analysis and expression of a blue-light photoreceptor gene, *slwc-1* from the cauliflower mushroom *Sparassis latifolia*. *Curr Microbiol*. 74: 469–475.
- Yang X, Yue Y, Li H, Ding W, Chen G, et al. 2018. The chromosome-level quality genome provides insights into the evolution of the biosynthesis genes for aroma compounds of *osmanthus fragrans*. *Hortic Res*. 5:72–72.
- Yang Z. 2007. Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Communicating editor: M. Sachs