scientific data

OPEN

DATA DESCRIPTOR

Fungal photobiont and microbiome genome composition in the *Cladonia uncialis* tripartite symbiosis

Zhuang Dong[1,2,3,7], Meng Shan Sun[2,4,7], Yue Dong He[5], Li Zhou[2,4], Wei Xiang[1,2,3], XinYao Li[2,3], Peng Huang[2,3,6] ✉ & Jian Guo Zeng[2,3,6] ✉

As symbiotic complexes formed through the association of bacteria or algae with fungi, lichens exhibit exceptional adaptability to extreme environments and function as pioneer species in rocky habitat ecological succession. The absence of high quality chromosome-level genome has constrained investigations into lichen adaptive evolution, while functional contributions of symbiotic bacterial communities remain inadequately explored. This study presents the chromosome-level genome assembly of the mycobiont *Cladonia uncialis*, comprising 28 chromosomes with a total size of 43.49 Mb, generated through integrated PacBio HiFi and Hi-C methodologies. We characterized the symbiotic microbiota using integrated short and long-read sequencing and constructed 31 metagenome-assembled genomes. The community was dominated by *Ascomycota* (41.16%), *Proteobacteria* (17.61%), and *Bacteroidota* (14.20%). Long-read sequencing significantly enhanced detection sensitivity for low-abundance taxa. This study provides essential genomic resources and comprehensive profiles of the symbiotic microbiota, enabling mechanistic exploration of adaptive evolution within lichen symbiotic systems under extreme environmental conditions.

Background & Summary

Lichens represent unique symbiotic complexes primarily formed between fungi (mycobionts) and photosynthetic partners (photobionts, which can be algae, cyanobacteria, or both), along with a diverse community of other bacteria[1]. They exhibit remarkable ecological adaptability globally, particularly thriving in extreme habitats characterized by low temperatures, drought, and intense radiation. Lichens occupy irreplaceable niches and act as pioneer species, promoting the ecological succession of rock substrates[2–5]. Although prior work has uncovered key genetic adaptations in *Cladonia borealis* to Antarctic extremes[6], fundamental knowledge gaps persist: (1) The lack of high quality chromosome-scale lichen genomes hinders comprehensive investigation into structure-function relationships underlying evolutionary adaptation; (2) Prevailing binary symbiosis models, centered on mycobiont-photobiont interactions, neglect the functional significance of bacterial consortia in holobiont environmental resilience.

C. uncialis represents a characteristic species within the fruticose lichen genus *Cladonia*, exhibiting a creeping growth form with hollow, terete branches[7]. It typically colonizes humus-rich or sandy substrates and is widely distributed across high-latitude habitats such as coniferous forests and heathlands, with notable prevalence throughout the Northern Hemisphere[8]. This species is distinguished by its biosynthesis of lichen acids, a metabolic feature rarely documented in other lichen taxa[9]. Although genomic sequencing of its symbiotic mycobiont has been conducted using the Illumina platform[10], the assembly has impeded subsequent functional and comparative genomic investigations. Focused on *C. uncialis* collected from Mount Changbai, China, this study achieved the chromosome-level genome assembly of its algae and fungal partner combining PacBio Revio long-read sequencing and Hi-C scaffolding. GC content analysis showed a marked bimodal distribution

[1]College of Horticulture, Hunan Agricultural University, Changsha, 410128, Hunan, China. [2]Yuelushan Laboratory, Changsha, 410000, China. [3]Hunan Key Laboratory of Traditional Chinese Veterinary Medicine, Hunan Agricultural University, Changsha, 410128, Hunan, China. [4]Hunan Institute of Nuclear Agriculture Sciences and Chinese Herbal Medicines, Changsha, 410125, Hunan, China. [5]College of Bioscience and Biotechnology, Hunan Agricultural University, Changsha, 410128, Hunan, China. [6]College of Veterinary Medicine, Hunan Agricultural University, Changsha, 410128, Hunan, China. [7]These authors contributed equally: Zhuang Dong, Meng Shan Sun. ✉e-mail: huangpeng@hunau.edu.cn; zengjianguo@hunau.edu.cn

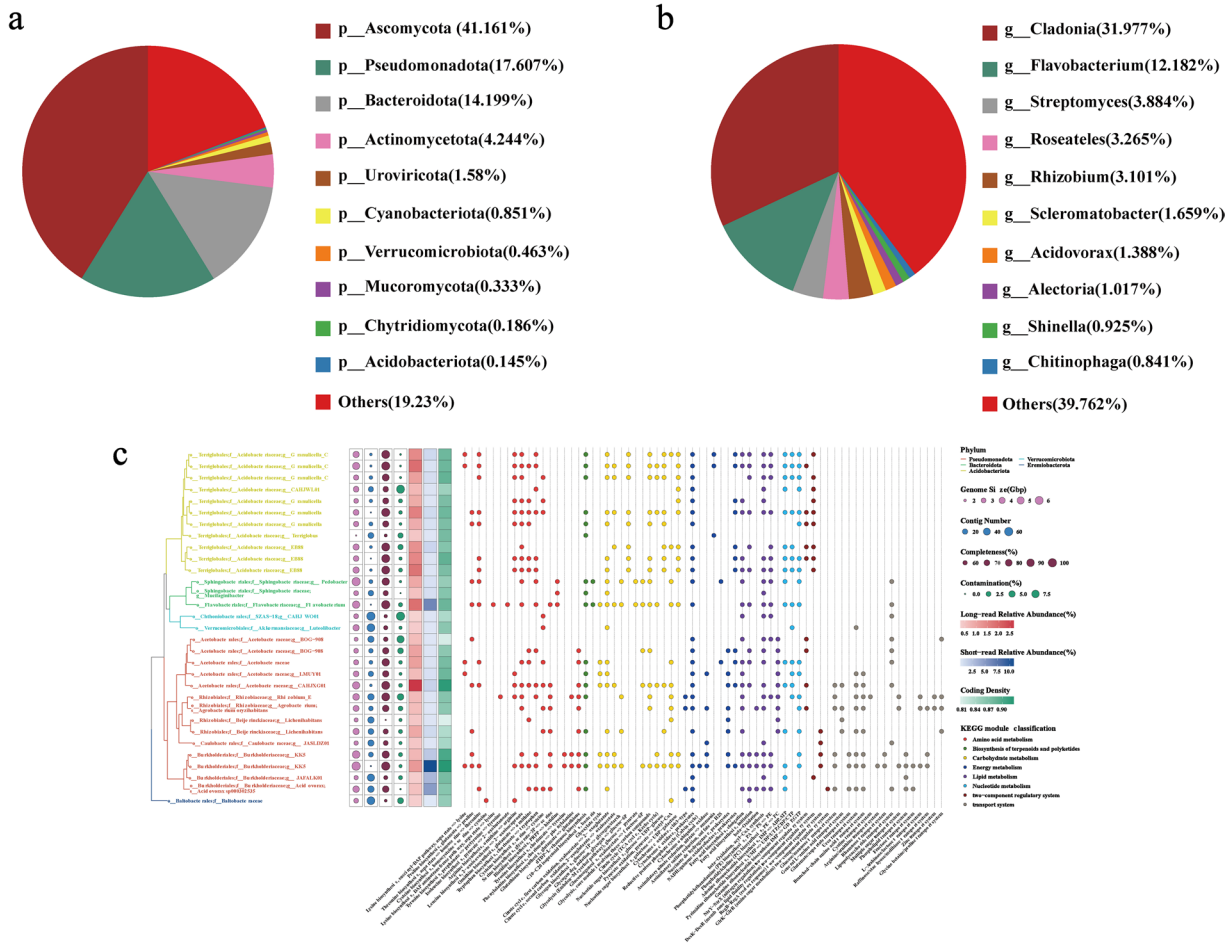

Fig. 1 Distribution of GC content and Hi-C interaction heatmap. (**a**) Density distribution profiles of GC content and sequencing depth. (**b**) Hi-C interaction heatmap of alga. (**c**) Hi-C interaction heatmap of fungi.

(Fig. 1a), featuring a major peak within the typical fungal genomic range and a minor peak of sequences with a distinct GC content. These non-homologous sequences in the minor peak are likely derived from the eukaryotic photosynthetic algal symbiont. This algal genomic fraction potentially encodes novel gene functions or biosynthetic clusters, thereby contributing to the distinctive ecological adaptation and metabolic diversity of *C. uncialis* and offering new clues for understanding metabolic synergy in this complex symbiotic system.

The photobiont genome assembly has a total size of 60.04 Mb, of which 43.79 Mb (72.9%) has been assembled into 18 chromosomes (Fig. 1b). Assessment with BUSCO (Benchmarking Universal Single-Copy Orthologs) revealed an incompleteness of 21.2%, indicating that a portion of the genome is missing from the assembly (Table 1). The initial fungal genome assembly (89 contigs, N50 = 1.66 Mbp) was significantly improved with Hi-C, increasing the contig N50 to 1.71 Mb. A total of 39.25 Mb (90.2%) of the 43.49 Mb genome was anchored into 28 chromosomes using 29 contigs (Figs. 1c, 2a), achieving a scaffold N50 of 1.71 Mb with contig length and number anchoring rates of 90.27% and 32.58%, respectively.

The interaction heatmap reveals distinct diagonal signals and pronounced inter-chromosomal interaction patterns. Among the 28 chromosomes, 27 (Chr1–Chr25 and Chr27–Chr28) are each assembled into a single contig, demonstrating exceptional continuity and accuracy, while only Chr26 was constructed from two contigs (Table 2). The strong intra-chromosomal interactions and well-defined chromosomal boundaries collectively confirm the achievement of a high-quality chromosome-level genome assembly, effectively mitigating the assembly fragmentation previously associated with short-read sequencing[10]. The observed inter-chromosomal interactions are non-random and likely represent biological signals, suggesting that these interaction hotspots may harbor regulatory networks involved in the coordinated expression of lichen acid biosynthetic gene clusters. Repetitive sequences constituted 15.85% (6,890,764 bp) of the assembled genome. The majority of these repeats were identified through a combination of methods: de novo prediction accounted for the largest proportion (14.05%), followed by long terminal repeat retrotransposons (LTRs; 3.03%) AUGUSTUS predicted 10,381 genes, with an average gene length of 1,996.5 bp, an average CDS length of 1,710.0 bp, and 3.26 exons per gene. GlimmHMM predicted 12,928 genes, averaging 1,581.5 bp in gene length, 1,400.8 bp in CDS length, and 2.41 exons per gene. Homology-based prediction using five reference species (*Bacidia gigantensis*, *Imshaugia aleurites*, etc.) yielded gene counts ranging from 4,464 to 11,391, with *B. gigantensis* predictions exhibiting the largest average intron length (284.2 bp).

BUSCOs	Fungi				algae	
	Assembly		Annotation		Assembly	
/	Proteins	Percentage (%)	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs	1653	96.9	1606	94.1	1187	78.2
Complete and single-copy BUSCOs	1403	82.2	1353	79.3	1154	76
Complete and duplicated BUSCOs	250	14.7	253	14.8	33	2.2
Fragmented BUSCOs	6	0.4	11	0.6	9	0.6
Missing BUSCOs	47	2.8	89	5.2	323	21.2
Total BUSCO groups searched	1706	100	1706	100	1519	100

Table 1. BUSCO assessment of annotation results.

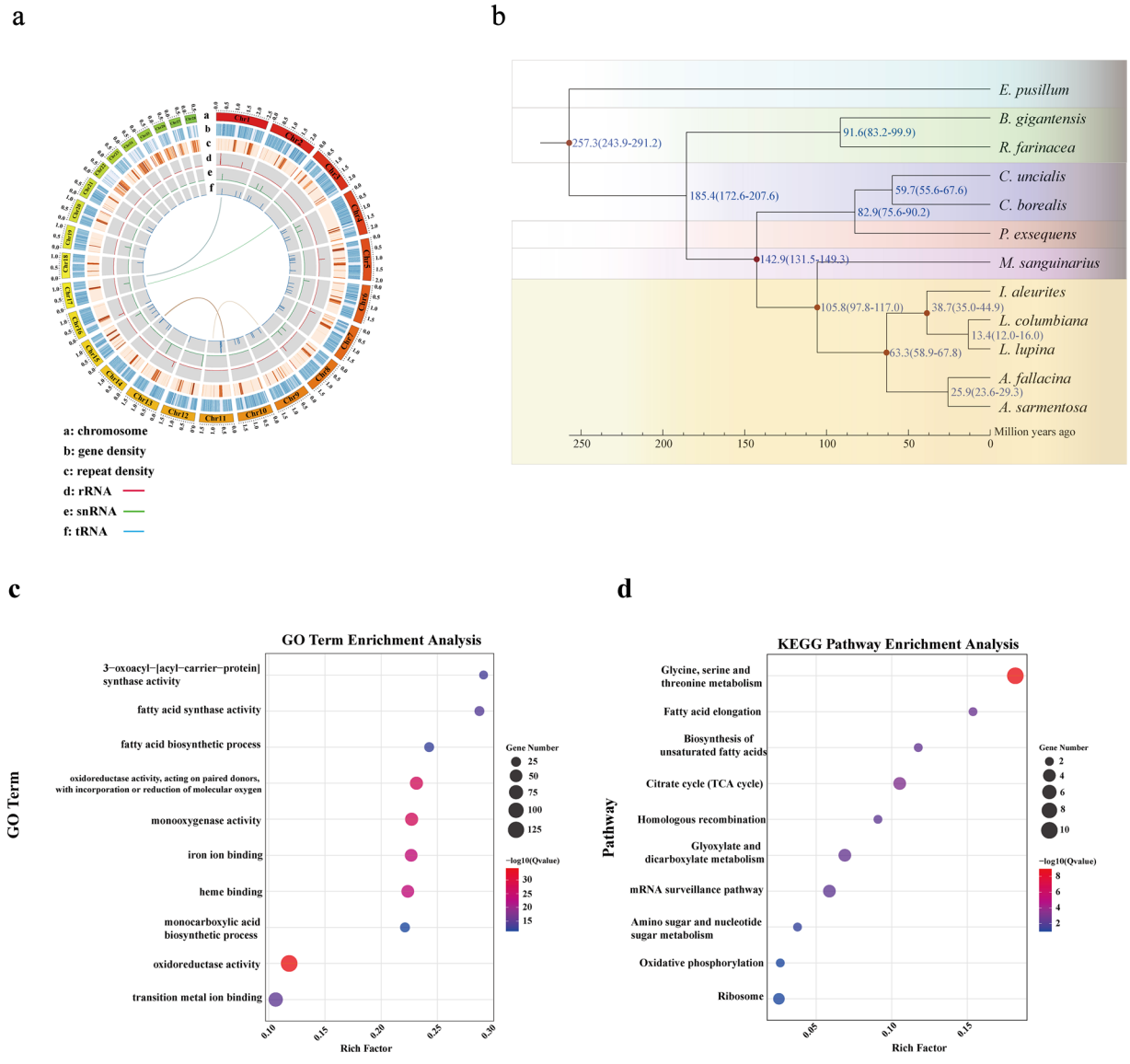

Fig. 2 Genomic evolution and the functional characteristics of expanded gene families. (**a**) Genomic features. (**b**) Estimation of the divergence time of 12 species of lichen formation. (**c**) Gene Ontology (GO) enrichment of expanded gene families. (**d**) KEGG pathway enrichment of expanded gene families.

The phylogenetic tree revealed that *C. uncialis* and *C. borealis* form a monophyletic clade, which diverged from *Cladoniaceae* around 59.7 million years ago and shares a common ancestor with *Puttea exsequens* dating to approximately 82.9 Mya (Fig. 2b). Genomic comparison further revealed substantial restructuring, including the expansion of 278 and contraction of 227 gene families (Fig. 2b), indicative of enhanced selective pressures. Functional annotation of expanded gene families revealed significant enrichment in oxidoreductase

Chromosome	Contig number	Sequence length (bp)
Chr1	1	2,535,146
Chr2	1	2,297,551
Chr3	1	2,238,088
Chr4	1	2,215,610
Chr5	1	2,051,164
Chr6	1	1,877,742
Chr7	1	1,761,947
Chr8	1	1,727,001
Chr9	1	1,711,767
Chr10	1	1,705,437
Chr11	1	1,662,487
Chr12	1	1,660,226
Chr13	1	1,659,478
Chr14	1	1,431,018
Chr15	1	1,345,410
Chr16	1	1,327,181
Chr17	1	1,243,289
Chr18	1	1,227,391
Chr19	1	1,124,758
Chr20	1	1,079,015
Chr21	1	960,098
Chr22	1	738,716
Chr23	1	708,339
Chr24	1	681,229
Chr25	1	650,128
Chr26	2	559,544
Chr27	1	539,818
Chr28	1	539,052

Table 2. Contig counts and total lengths per chromosome in the *C. uncialis* genome assembly.

activity (GO:0016491), fatty acid synthase activity (GO:0004312), and monocarboxylic/fatty acid biosynthesis (GO:0072330, GO:0006633) (Fig. 2c). Oxidoreductase activity represented the predominant functional category, while monocarboxylic acid biosynthesis was the most statistically significant pathway. KEGG (Kyoto Encyclopedia of Genes and Genomes) analysis identified four core metabolic modules: oxidative phosphorylation (ko00190), glycine/serine/threonine metabolism (ko00260), amino and nucleotide sugar metabolism (ko00520), and unsaturated fatty acid biosynthesis (ko01040) (Fig. 2d). The environmental adaptation of *C. uncialis* appears to be underpinned by enhanced functions in energy production, oxidative stress resistance, and membrane maintenance.

Trans.orf/RNAseq evidence supported 3,035 genes, characterized by an average gene length of 3,446.4 bp, an average exon length of 922.4 bp, and an average intron length of 69.7 bp. The MAKER pipeline generated an integrated set of 11,049 genes, with an average gene length of 2,049.4 bp, an average CDS length of 1,543.6 bp, and 3.09 exons per gene. Subsequent refinement using PASA improved gene structures, increasing the average gene length to 2,150.4 bp and the average exon number to 3.14 per gene (Table 3). Non-coding RNA annotation identified 56 tRNA genes, 10 rRNA genes, and 15 snRNA genes. The 10 rRNA genes comprised two 18S, two 28S, and six 5S rRNA genes; the 28S rRNA genes accounted for the majority (70.6%) of the total rRNA sequence length. 15 snRNA genes included eight CD-box, one HACA-box, and six splicing-related types. Neither miRNA nor scaRNA annotations were detected (Table 4). The annotation of transposable elements (TEs) revealed that repetitive sequences constitute 14.67% of the genome. Long terminal repeat (LTR) retrotransposons (8.59%) and elements of unknown classification (7.45%) were identified as the dominant components (Table 5). The prevalence of gene duplication was evidenced by gene family clustering, which identified multiple-copy orthologs (2,060 genes, 18.65%) as the most abundant group among the 11,045 genes analyzed. Among the 5,464 single nucleotide polymorphisms (SNPs) identified across the genome, heterozygous SNPs were predominant, constituting 4,754 (87.01%) of the total, while homozygous SNPs accounted for the remaining 710 (12.99%). Based on short-read sequencing technology, metagenomic sequencing was performed on the symbiotic microbial community of *C.uncialis*, resulting in the annotation of 325,124 genes (Table 6).

At the phylum level, *Ascomycota* (41.16%), *Pseudomonadota* (formerly *Proteobacteria*; 17.61%), and *Bacteroidota* (14.20%) constituted the dominant taxa (Fig. 3a), while at the genus level, the fungal symbiont *Cladonia* (31.98%) and bacterial symbiont *Flavobacterium* (12.18%) exhibited the highest relative abundances (Fig. 3b). Based on long-read metagenomic sequencing data, 31 metagenome-assembled genomes (MAGs) were successfully constructed. Among them, 2 MAGs were identified at the species level: semibin2_9 was identified as *Acidovorax* sp000302535, and semibin2_22 was identified as *Agrobacterium oryzihabitans*. Long-read

Gene set		Number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	AUGUSTUS	10381	1996.45	1709.98	3.26	521.22	125.61
/	GlimmHMM	12928	1581.5	1400.76	2.41	580.33	127.85
Homolog	B.gigantensis	9199	1611.31	1159.57	2.59	447.79	284.2
/	I.aleurites	10425	1514.71	1200.83	2.57	467.36	200
/	l.columbiana	11391	1392.29	1155.48	2.63	440.03	145.64
/	l.lupina	10816	1468.48	1202.69	2.67	450.66	159.28
/	M.sanguinarius	10687	1475.68	1193.69	2.58	463.55	179.03
/	S.pombe	4464	1129.41	885.23	2.49	355.4	163.79
trans.orf/RNAseq		3035	3446.38	1542.97	3.54	922.36	69.66
MAKER		11049	2049.38	1543.6	3.09	599.82	92.17
PASA		11,045	2,150.43	1,545.92	3.14	621.95	92.39

Table 3. The statistical results of gene prediction.

Type		Copy	Average length(bp)	Total length(bp)	% of genome
miRNA		0	0	0	0
tRNA		56	92.14286	5160	0.011865
rRNA	rRNA	10	1922.2	19222	0.044201
	18S	2	2471	4942	0.011364
	28S	2	6792	13584	0.031236
	5S	6	116	696	0.0016
snRNA	snRNA	15	134.66667	2020	0.004645
	CD-box	8	111.375	891	0.002049
	HACA-box	1	151	151	0.000347
	splicing	6	163	978	0.002249
	scaRNA	0	0	0	0

Table 4. Statistics of non-coding RNA annotation results.

Type	RepeatMasker TEs Length (Bp)	RepeatMasker TEs % in genome	RepeatProteinMask TEs Length (Bp)	Repeat Protein Mask TEs % in genome	De novo Length (Bp)	De novo % in genome	Combined TEs Length (Bp)	Combined TEs % % in genome
DNA	64930	0.15	22914	0.05	0	0	84170	0.19
LINE	42507	0.1	14676	0.03	0	0	55993	0.13
SINE	143	0	0	0	0	0	143	0
LTR	176470	0.41	215002	0.49	3559092	8.18	3734592	8.59
Other	19	0	0	0	0	0	19	0
Unknown	691	0	11505	0.03	3228725	7.42	3240784	7.45
Total TE	266331	0.61	263965	0.61	6109671	14.05	6380133	14.67

Table 5. The statistics of classification results of repeated sequences.

sequencing significantly improved the detection rate of low-abundance microbial communities, such as semibin2_31 (PacBio: 0.35% vs MGI: 0.01%) (Table 7). KEGG-based functional analysis showed that core metabolic pathways were significantly enriched in 31 MAGs (Fig. 3c). Among them, 25 MAGs contained prokaryotic cytochrome c oxidase, suggesting active oxidative phosphorylation. Biosynthesis and initiation phases (in 22 and 19 MAGs, respectively) involved acetyl-CoA carboxylase and fatty acid synthase, indicating functional machinery for producing malonyl-CoA and hexadecanoic acid. Concurrent activity of β-oxidation and acyl-CoA synthetases (23 MAGs) pointed to the use of external lipids for acetyl-CoA generation. This catabolism-anabolism link supported ATP synthesis via the TCA cycle and fatty acid elongation. The dTDP-L-rhamnose pathway (17 MAGs) was also enriched, potentially aiding synthesis of surface polysaccharides. Furthermore, a complete monocarboxylic acid cycle (19 MAGs) implied that symbiotic bacteria use this route to produce ATP and reducing equivalents ($NADH/FADH_2$).

Previous genomic studies on the lichen genus *Cladonia* have been largely limited to assemblies with low continuity, such as the contig-level assembly (N50 = 34.7 kb)[6,10], and the draft genomes of both the fungal and algal symbionts of *Cladonia grayi* have only reached scaffold level, with N50 values of 243.4 kb and 784.9 kb, respectively[11]. Compared to the previously published chromosome-level genomes of *Cladonia* lichens[12,13], this study assembled a larger genome of approximately 43.49 Mb, which not only expands the known genome size range

	Metagenomic short-read sequencing data	Genome non-redundant genes statistics	Genome predicted genes	Metagenomic long-read sequencing data	Metagenomic long-read sequencing assembly
Number	193330	325124	339946	1373353	12382
integrity:start	/	61943 (19.05%)	65302 (19.21%)	/	/
integrity:end	/	86895 (26.73%)	92435 (27.19%)	/	/
integrity:all	/	152921 (47.03%)	158623 (46.66%)	/	/
integrity:none	/	23365 (7.19%)	23586 (6.94%)	/	/
Total (bp)	302551460	157318623	162383268	21157642953	1164446647
Average (bp)	1564.95	483.87	477.67	15405.8	94043.5
Max (bp)	307566	18330	18330	50276	6562713
Min (bp)	500	102	102		8040
N50	2399	627	621	15614	185900
L50	23786	74948	77919	/	/
GC(%)	54.43	58.06	57.85	/	/

Table 6. Summary statistics of genome annotation and metagenomic assembly.

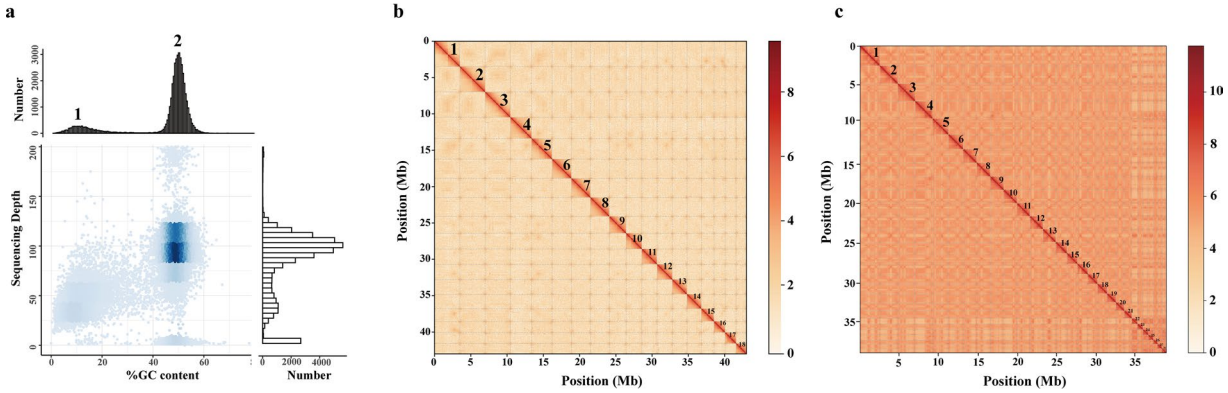

Fig. 3 Characteristics of the symbiotic microbiome. (**a,b**) Phylum and genus level taxonomic profiles from short-read sequencing. (**c**) KEGG-based annotation of metabolic pathways in 31 host-associated symbiotic bacterial MAGs.

for this genus but also suggests the presence of more complex genetic information. Furthermore, our assembly demonstrates superior continuity (Contig N50 = 1.71 Mb) and achieved a high chromosome anchoring rate of 98.66% using Hi-C technology, providing a more reliable foundation for precise gene localization and functional annotation. In contrast to previous studies focusing primarily on the fungal host, our approach integrated genomic and metagenomic strategies based on the contemporary understanding of lichens as symbiotic systems comprising fungi, algae, and specific bacterial communities. By employing long-read sequencing technology to overcome the challenges of assembling complex symbiotic systems, we first revealed the "core-satellite" structure and functional attributes of the symbiotic microbial community within this lichen system.

The relative abundances of microbial phyla reported in this study, notably *Proteobacteria* and *Bacteroidetes*, should be regarded as "technical abundances" derived from specific methodological conditions, rather than as absolute abundances in the original samples. As all samples were processed through a standardized and optimized pipeline, the systematic bias introduced by the methodology is consistent across samples. Consequently, observed differences in community structure are more likely to reflect true biological variation rather than technical artifacts[14]. We therefore conclude that the key findings of this study concerning the relative composition of distinct microbial communities are robust. To better address such methodological limitations and more accurately uncover *in situ* microbial community structure, *in situ* metagenomic and metatranscriptomic approaches are considered promising alternatives[15,16]. In future work, we plan to implement culture-independent or DNA extraction-free *in situ* techniques to achieve a more comprehensive and unbiased characterization of the complete microbial consortium within lichen symbiotic systems.

Methods

The *C. uncialis* sample was collected in December 2023 from the northern slope of Changbai Mountain (128°16′10″E, 42°20′29″N) in Antu County, Jilin Province, China, at an ambient temperature of approximately −25 °C. Immediately after collection, the sample was placed into a sterile cryovial and rapidly frozen in liquid nitrogen. During transport to the laboratory, the sample was maintained in a dry ice environment to ensure consistent low-temperature preservation. Upon arrival, it was promptly transferred to a −80 °C ultra-low temperature freezer for storage until DNA extraction. For subsequent analyses, 200.0 mg g, 1000.0 mg, and

Genome	PacBio Relative Abundance (%)	MGI Relative Abundance (%)	classification
complete.209	1.1896138	10.842532	g__*KK5*;s__
complete.223	1.8862169	6.2975926	g__*Flavobacterium*;s__
semibin2_9	0.5475069	4.8685894	g__*Acidovorax*;s__*Acidovorax* sp000302535
metabat2.631	2.662776	1.3988123	g__*CAHJXG01*;s__
semibin2_18	0.3564348	3.090497	g__*JAFALK01*;s__
metabat2.119	0.7072619	1.9875702	g__*KK5*;s__
metabat2.303	1.3217254	1.1432718	g__*EB88*;s__
semibin2_2	1.4296474	0.9588933	g__*Granulicella_C*;s__
complete.182	1.578299	0.7565463	g__*Granulicella*;s__
metabat2.445	1.8932424	0.343993	g__*Granulicella_C*;s__
semibin2_1	1.8060101	0.1447961	g__*EB88*;s__
metabat2.325	0.9324797	0.8311867	g__*Lichenihabitans*;s__
complete.80	1.4802755	0.1945076	g__*EB88*;s__
semibin2_4	1.102366	0.5638828	g__*Granulicella_C*;s__
semibin2_5	0.9265796	0.4168355	g__*Pedobacter*;s__
semibin2_7	1.0959816	0.2351909	f__*Acetobacteraceae*;g__;s__
metabat2.429	0.8632949	0.3744559	g__*Granulicella*;s__
semibin2_12	1.0399697	0.1012976	g__*BOG-908*;s__
metabat2.514	0.5316572	0.5424176	g__*Mucilaginibacter*;s__
semibin2_21	0.6360606	0.3871656	g__*LMUY01*;s__
semibin2_27	0.8913447	0.1226039	g__*Granulicella*;s__
semibin2_49	0.6838769	0.2035214	g__*CAHJWL01*;s__
semibin2_8	0.6492726	0.0969147	g__*Rhizobium_E*;s__
semibin2_22	0.621351	0.0854441	g__*Agrobacterium*;s__*Agrobacterium* oryzihabitans
semibin2_16	0.6205139	0.0635475	g__*BOG-908*;s__
metabat2.558	0.4439194	0.099185	g__*Lichenihabitans*;s__
semibin2_20	0.4454537	0.0713062	g__*Terriglobus*;s__
semibin2_14	0.4670585	0.0414418	g__*JASLDZ01*;s__
semibin2_17	0.385272	0.0756512	g__*Luteolibacter*;s__
semibin2_403	0.3013759	0.1168207	g__*CAHJWO01*;s__
semibin2_31	0.3490408	0.0140036	f__*Baltobacteraceae*;g__;s__

Table 7. PacBio and MGI metagenomic relative abundance by microbial classification.

1000.0 mg were allocated for short-read, long-read metagenomic sequencing, and long-read genomic sequencing, respectively.

The DNA extraction procedure utilized the following instruments and consumables. Tissue homogenization was performed using a FastPrep-24 instrument (MP Biomedicals, USA) with 0.5 mm zirconium beads (BioSpec Products, USA). Liquid handling steps were carried out using Eppendorf Research® plus manual pipettes (Eppendorf SE, Germany). A refrigerated centrifuge (Model: 5430 R, Eppendorf SE, Germany) equipped with F-35-6-30 and FA-45-24-11 rotors was used for centrifugation. Key centrifugation steps were conducted using sterile, nuclease-free 50 mL conical-bottom polypropylene tubes (CHIPRO). Subsequent purification steps employed the spin columns and collection tubes (2 mL) from the PowerSoil Pro Kit (QIAGEN, Hilden, Germany). Nucleic acid concentration was quantified using a Qubit 4.0 Fluorometer (Thermo Fisher Scientific, USA). The PowerSoil Pro Kit was acquired from QIAGEN (Hilden, Germany); absolute ethanol was purchased from Sinopharm Chemical Reagent Co., Ltd.; and ethylenediaminetetraacetic acid (EDTA) was obtained from Solarbio Science & Technology Co., Ltd.

DNA Extraction. Surface-associated matrices and debris were meticulously removed from lichen specimens using sterile scalpels. Under aseptic conditions, the samples were flash-frozen in liquid nitrogen and ground to a fine powder using a pre-chilled mortar and pestle. The powder was immediately homogenized in DNA/RNA Shield (Zymo Research) to stabilize nucleic acids. DNA extraction was performed using the PowerSoil Pro Kit (QIAGEN) with a key modification: insoluble polyvinylpolypyrrolidone (PVPP) was added to the Cell Lysis Solution at a final concentration of 5% (w/v) to adsorb polyphenolic compounds. Subsequent purification steps followed the protocol provided by the manufacturer, and all centrifugations were performed at 4 °C. Briefly, after cell lysis, an initial high-speed centrifugation step (5 min) was conducted to pellet debris. The supernatant was then subjected to ethanol precipitation (13,000 rpm, 10 min), followed by a wash step with ethanol (13,000 rpm, 10 min) to remove residual salts. Qualified DNA samples ($\geq$50 ng/μL, A_{260}/A_{280} ratio = 1.8–2.0) were used for library construction. Genomic DNA was fragmented to a target size of ~350 bp using a focused-ultrasonicator.

Size-selected fragments then underwent standard library preparation procedures, including end repair, adapter ligation, and PCR (Polymerase Chain Reaction) amplification.

Library construction. Short-read and long-read sequencing libraries were constructed using the BGISEQ PE150 and PacBio HiFi strategies, respectively. The BGISEQ PE150 library was prepared with 1 μg of starting DNA following the standard protocol of the MGI DNA Library Preparation Kit (Vazyme, Nanjing). The PacBio HiFi library was constructed with 15 μg DNA using the SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences) in strict accordance with the manufacturer's instructions. The DNA used for both short-read and long-read library construction was derived from the 'DNA Extraction' section. (Sequencing services were provided by Wuhan Frasergen Bioinformatics Co., Ltd.).

Genome sequencing. The genome was initially assembled using PacBio Sequel II long-read sequencing data, followed by iterative error correction with Pilon v1.22[17] based on 66.72 Gb of short-read data. A total of 4.74 Gb of Hi-C sequencing data was acquired. Hi-C clean reads were aligned to the preliminary assembly using BWA v0.7.17[18], followed by chromatin interaction-guided chromosome scaffolding with Juicer v1.6[19]. Read mapping rates were calculated using minimap2 v2.24[20], and coverage and depth distributions were analyzed using SAMtools v1.17[21]. Single-copy orthologous genes were identified using OrthoDB v11[22], and genome completeness was assessed using BUSCO v5.4.3[23] against the *Ascomycota* lineage dataset. Genome-wide SNPcalling was performed using GATK v4.3.0[24] following the Best Practices workflow (HaplotypeCaller + VariantFiltration), and homozygous/heterozygous site ratios were quantified. Tandem repeats were identified using TRF v4.09[25], and in parallel, transposable elements were annotated using RepeatMasker v4.1.5 and RepeatProteinMask[26]. tRNA genes were predicted using tRNAscan-SE v2.0.1[27], and non-coding RNAs (miRNAs and snRNAs) were characterized using Rfam covariance models with INFERNAL v1.1.4[28]. Additional long-read analyses included: NCBI BLAST+ v2.11.0[29] alignment against the Non-redundant Nucleotide (NT) database with the parameters-task megablast and -evalue 1e-5; prokaryotic gene prediction using Glimmer v3.02[30]; annotation of tRNA and rRNA with tRNAscan-SE v2.0.9[27] and RNAmmer v1.2[31]; prediction of secondary metabolite biosynthetic gene clusters using antiSMASH v7.1.0[32]; genome annotation (including CDS, rRNA, and tRNA) via Prokka v1.14.6[33]; and clustering of 16S rRNA genes at a 97% sequence identity threshold using CD-HIT-EST v4.8[34]. PASA v2.4.1 was employed with its default parameters to refine the gene structures[35]. A comparative genomic analysis was performed starting with the clustering of orthologous genes from all examined species using OrthoFinder (v 2.5.4)[36] with the parameters '-M msa -S diamond'. Single-copy orthologs common to all species were identified for phylogenetic inference. Each of these gene families was aligned using MUSCLE (v 3.8.31)[37] with '-maxiters 16', and the resulting alignments were concatenated into a supermatrix. A maximum likelihood phylogeny was constructed using RAxML (v 8.2.11)[38] with the command '-f a -x 12345 -p 12345 -# 100 -m PROTGAMMAAUTO'. Divergence times at key nodes were estimated by incorporating known divergence times from the TimeTree[39] database as calibration points, using r8s (v1. 81)[40] followed by the 'mcmctree' program in PAML (v 4. 10. 0)[41] with 'nsample = 10000'. Gene family expansion and contraction across the phylogeny were assessed using CAFÉ (v 4. 2. 1)[42] with the parameters '-p 0.05 -r 10000 -filter'. Finally, to detect positive selection driving species differentiation, a branch-site model was implemented via the 'codeml' program in PAML to analyze specific lineages. All software mentioned above was run using their default parameters unless otherwise specified.

Metagenomic sequencing. *Short-read metagenomic sequencing.* Metagenomic sequencing was performed on the MGI DNBSEQ-T7 platform. Raw reads were subjected to quality control and filtering using SOAPnuke (v 2.1.0)[43] with parameters '–lowQual = 20', '–nRate = 0.005', and '–qualRate = 0.5' to obtain high-quality clean data. Subsequent de novo assembly of the filtered reads was conducted with MEGAHIT (v 1.2.9)[44], which utilizes iterative k-mer optimization to generate contigs. Open reading frames (ORFs) were predicted from these contigs using MetaGeneMark (v3.38)[45] with default parameters, and genes shorter than 100 nucleotides were filtered out[46,47]. The predicted gene sets from all samples were then pooled to construct a non-redundant gene catalog using CD-HIT (v 4.8.1)[34] with sequence clustering thresholds set at 95% identity and 90% coverage; the longest sequence in each cluster was designated as the representative. For taxonomic annotation, the protein sequences of these representative unigenes were aligned against the bacterial subset of the NCBI non-redundant protein database (NR, version 20210213)[48] using DIAMOND (v2.0.9) blastp with an E-value cutoff of 1e-5. The top hit based on the highest alignment score was selected for both functional and taxonomic annotation, with species assignment being inferred from the corresponding NCBI taxonomy database[49].

Long-read metagenomic sequencing. Long-read metagenomic sequencing was conducted on the PacBio Revio platform, during which subreads (single-molecule effective inserts) were converted into high-fidelity reads (HiFi reads) using the CCS pipeline within the SMRT Link software package (v 13.0.0)[50]. The CCS analysis was performed with a minimum of 3 passes and a minimum predicted accuracy threshold of 0.99 (Q20). Metagenome-assembled genomes (MAGs) were reconstructed using a two-step strategy: initial assembly of raw sequencing data into contigs was performed with hifiasm-meta (v 0.2.1)[51], followed by binning implemented in the HiFi-MAG-Pipeline[52]. For taxonomic classification, the assembled contigs were aligned to the NCBI non-redundant nucleotide database (NT, version 20210214) using BLASTN from the NCBI BLAST+ package (v2.11.0+)[29] under the megablast task with an E-value threshold of 1e-5 and a maximum of three target sequences per query. Functional profiling of the resulting MAGs was carried out with EnrichM v0.6.5 for KEGG analysis. Specifically, MAG-encoded protein sequences were annotated by aligning them against the KO-integrated EnrichM v10 database (based on UniRef100) using Diamond v0.9.22 with default parameters[53]. The Diamond blastp search was conducted with the --sensitive mode. The best alignment result for each query

sequence was selected (--max-target-seqs 1), with thresholds set for E-value (<1e-5) and sequence coverage (>50%). Subsequently, KO functional modules within the MAGs were identified using the "classify" function.

Data Records

The dataset is available at the National Center for Biotechnology Information (NCBI) and the European Nucleotide Archive (ENA). The raw genomic sequencing data, short-read metagenomic data, and long-read metagenomic data have been deposited under BioProject accessions PRJNA1348763, PRJNA1348760, and PRJNA1348772, respectively. The specific read set accessions (SRA runs) corresponding to each library are as follows:

BioProject PRJNA1348763: All data in this project were generated from the PacBio HiFi library. The complete set of 39 individual SRA run accessions (SRR35859719–SRR35859753, SRR35878531–SRR35878534) can be viewed by accessing the BioProject page for SRP636986 on the NCBI website and navigating to the 'SRA' section.

BioProject PRJNA1348760: This project contains data generated from the BGI PE150 library. The run accessions SRR36204696 and SRR36204697 can be viewed by accessing the BioProject page for SRP649239.

BioProject PRJNA1348772: The data for this project were generated from the PacBio HiFi library. The run accession SRR35874174 can be viewed by accessing the BioProject page for SRP637105.

Additionally, the genome assembly and annotation are available under accession CDRNKR01. The 31 metagenome-assembled genomes (MAGs) reconstructed from long-read metagenomic data are available under accession CDRNLQ01.

Below are the direct access links and guidance for reviewers and readers to access the data:

NCBI Sequence Read Archive

https://identifiers.org/ncbi/insdc.sra:SRP636986 (2025)[54] (Raw Genomic Sequencing Data)

https://identifiers.org/ncbi/insdc.sra:SRP649239[55] (Short-Read Metagenomic Data)

https://identifiers.org/ncbi/insdc.sra:SRP637105 (2025)[56] (Long-Read Metagenomic Data)

European Nucleotide Archive

https://identifiers.org/ncbi/insdc.gca:GCA_977014525.1 (2025)[57] (Assembly and annotation of genome)

https://identifiers.org/ncbi/insdc.gca:GCA_977014875.1 (2025)[58] (Sequence of 31 MAGs).

Technical Validation

Genomic data evaluation and validation. The assembled sequence contains 96.9% complete conserved genes, among which 82.2% are single-copy and 14.7% are duplicated. Fragmented genes account for 0.4%, and the missing rate is 2.8%. The completeness of protein annotation reaches 94.1%, including 79.3% single-copy genes and 14.8% duplicated genes, with 0.6% fragmented genes and a 5.2% missing rate (Table 1). These results indicate that the genome assembly has achieved an excellent level of integrity, and the quality of protein annotation is highly reliable. After quality control, the short-read data retain 66.72 Gb of high-quality data, consisting of 444,790,912 reads with a Q30 value of 89.5% and a GC content of 48.8%. The long-read data amount to 4.74 Gb, including 338,057 reads with an average length of 14,011 bp and a GC content of 48.8%. Initial assembly was performed using long-read data, followed by iterative correction with Pilon based on short-read data. Of the 426,528,166 valid Hi-C read pairs, 375,189,794 (87.96%) were mapped to the genome assembly. Among these mapped pairs, 171,153,240 (45.62%) had mates located on different contigs or scaffolds. After applying a quality filter (mapQ $\geq$ 5), 164,018,571 interaction pairs (43.72% of mapped reads) were retained.

A total of 28 contigs longer than 100 kb, with a combined length of 39.22 Mb, were anchored into chromosomes by Hi-C scaffolding. This achieved an anchoring rate of 98.66% for the pre-assembly sequence length (39.75 Mb). All the above indicate that the assembly continuity has reached the chromosomal level. Verification of key functional elements detected 56 tRNAs, a complete rRNA operon (28S/18S/5.8S), and 15 snRNAs, demonstrating the integrity of basic genetic elements.

Metagenomic data evaluation and validation. Quality control processing of the 73.93 Gb of raw data derived from short-read sequencing of the symbiotic microbiota associated with *C. uncialis* samples yielded 48,755,046 high-quality reads (72.55 Gb), corresponding to a data retention rate of 98.13%. Following quality control, sequence metrics confirmed high data quality with Q20 and Q30 values of 99.19% and 97.34%, respectively.

De novo assembly generated 193,330 contigs totaling 302.55 Mb, averaging 1,564.95 bp per contig. Gene deduplication analysis yielded 325,124 non-redundant open reading frames, with a total length of 157.32 Mb and an average length of 483.87 bp. Analysis of structural integrity revealed that 47.03% were complete ORFs possessing both start and stop codons, 19.05% were partial ORFs containing only a start codon, 26.73% were partial ORFs containing only a stop codon, and 7.19% were incomplete ORFs lacking both codons. Continuity metrics showed an N50 of 627 bp, with maximum and minimum ORF lengths of 18.33 kb and 102 bp, respectively. The overall GC content was 58.06% (Table 6), consistent with the nucleotide composition characteristic of typical protein-coding regions. MetaGeneMark analysis identified 339,946 open reading frames (ORFs), encompassing a total length of 162.38 Mb and exhibiting an average length of 477.67 bp (Table 6). Collectively, these results indicate that the assembly derived from short-read sequencing exhibits a highly fragmented nature and structural incompleteness.

High-accuracy long-read sequencing yielded 1,373,353 high-quality reads, totaling 21.16 Gb. These reads exhibited an N50 length of 15,614 bp, an average length of 15,405.8 bp, and a maximum read length of 50,276 bp; corresponding quality metrics Q20 and Q30 reached 97.48% and 93.91%, respectively, meeting the high-accuracy criterion. Subsequent sequence assembly generated 12,382 contigs with a total assembly size of 1.16 Gb. Assembly continuity analysis revealed a contig N50 of 185,900 bp, an average contig length of

94,043.5 bp, a maximum contig length of 6.56 Mb, and the absence of sequence gaps (Table 6). This long-read sequencing data combines high accuracy with exceptional continuity, facilitating the reconstruction of microbial genomes. The relative abundances of bacterial phyla reported, particularly Proteobacteria and Bacteroidetes, should be considered "technical abundances" under the specific methodology, rather than absolute abundances in samples. As all samples were processed uniformly with this optimized protocol, we conclude that the core findings on relative changes in microbial community structures are robust and valid. Since methodological biases are consistent across samples, intergroup differences are more likely to reflect true biological variation rather than technical artifacts.

Data availability

All data generated and analysed in this study are publicly accessible. Raw sequencing reads can be retrieved from the NCBI BioProject database under accessions PRJNA1348763, PRJNA1348760, and PRJNA1348772. The genome assembly, annotation, and MAGs are available from the European Nucleotide Archive under accessions CDRNKR01 and CDRNLQ01, respectively. The eleven publicly available genomic datasets used for comparative analysis in this study were all sourced from the NCBI database. The accession numbers, listed in the order of analysis, are as follows: GCF_000464535.1 (*Endocarpon pusillum*), GCF_019456465.1 (*Bacidia gigantensis*), GCA_029948375.1 (*Ramalina farinacea*), GCA_018257855.2 (*Cladonia borealis*), GCA_022814085.1 (*Puttea exsequens*), GCA_022814195.1 (*Mycoblastus sanguinarius*), GCA_905337355.1 (*Imshaugia aleurites*), GCF_014066305.1 (*Letharia columbiana*), GCF_014066315.1 (*Letharia lupina*), GCA_905337325.1 (*Alectoria fallacina*), and GCA_904859925.1 (*Alectoria sarmentosa*). The involved data from the NCBI database cited in References[12,13] are available under accessions GCA_947623385.2 and GCA_963971245.1, respectively.

Code availability

No custom code was developed for this study. All data processing and analyses were performed following the standard protocols and guidelines provided by the respective bioinformatics tools. Detailed parameter settings employed for each tool are described in the Methods section.

Received: 11 July 2025; Accepted: 13 January 2026;
Published online: 30 January 2026

References

1. He, Z. & Naganuma, T. Chronicle of Research into Lichen-Associated Bacteria. *Microorganisms* **10**(11), 2111, https://doi.org/10.3390/microorganisms10112111 (2022).
2. Dal Grande, F. *et al*. Adaptive differentiation coincides with local bioclimatic conditions along an elevational cline in populations of a lichen-forming fungus. *BMC Evolutionary Biology* **17**(1), 93, https://doi.org/10.1186/s12862-017-0929-8 (2017).
3. Meeßen, J. *et al*. The Effect of High-Dose Ionizing Radiation on the Isolated Photobiont of the Astrobiological Model Lichen *Circinaria gyrosa*. *Astrobiology* **17**(2), 154–162, https://doi.org/10.1089/ast.2015.1453 (2017).
4. Parasyri, A. *et al*. Lichen as Micro-Ecosystem: Extremophilic Behavior with Astrobiotechnological Applications. *Astrobiology* **18**(12), 1528–1542, https://doi.org/10.1089/ast.2017.1789 (2018).
5. Adams, J. N. *et al*. The reference genome assembly of the bright cobblestone lichen, *Acarospora socialis*. *Journal of Heredity* **114**(6), 707–714, https://doi.org/10.1093/jhered/esad052 (2023).
6. Cho, M. *et al*. An Antarctic lichen isolate (*Cladonia borealis*) genome reveals potential adaptation to extreme environments. *Scientific Reports* **14**(1), 1342, https://doi.org/10.1038/s41598-024-51895-x (2024).
7. Phinney, N. H., Asplund, J. & Gauslaa, Y. The lichen cushion: A functional perspective of color and size of a dominant growth form on glacier forelands. *Fungal Biology* **126**(5), 375–384, https://doi.org/10.1016/j.funbio.2022.03.001 (2022).
8. Studzińska-Sroka, E. *et al*. *Cladonia uncialis* as a valuable raw material of biosynthetic compounds against clinical strains of bacteria and fungi. *Acta Biochimica Polonica* **66**(4), 597–603, https://doi.org/10.18388/abp.2019_2891 (2019).
9. Abdel-Hameed, M. *et al*. Putative identification of the usnic acid biosynthetic gene cluster by de novo whole-genome sequencing of a lichen-forming fungus. *Fungal Biology* **120**(3), 306–316, https://doi.org/10.1016/j.funbio.2015.10.009 (2016).
10. Bertrand, R. L., Abdel-Hameed, M. & Sorensen, J. L. Lichen biosynthetic gene clusters. Part I. Genome sequencing reveals a rich biosynthetic potential. *Journal of Natural Products* **81**(4), 723–731, https://doi.org/10.1021/acs.jnatprod.7b00769 (2018).
11. Armaleo, D. *et al*. The lichen symbiosis re-viewed through the genomes of *Cladonia grayi* and its algal partner *Asterochloris glomerata*. *BMC Genomics* **20**(1), 605, https://doi.org/10.1186/s12864-019-5629-x (2019).
12. *NCBI Genome* https://identifiers.org/ncbi/insdc.gca:GCA_947623385.2 (2024).
13. *NCBI Genome* https://identifiers.org/ncbi/insdc.gca:GCA_963971245.1 (2024).
14. Zhang, X. *et al*. Impacts of typical pharmaceuticals and personal care products on the performance and microbial community of a sponge-based moving bed biofilm reactor. *Bioresource Technology* **295**, 122298, https://doi.org/10.1016/j.biortech.2019.122298 (2020).
15. Emiola, A. & Oh, J. High throughput *in situ* metagenomic measurement of bacterial replication at ultra-low sequencing coverage. *Nature Communications* **9**(1), 4956, https://doi.org/10.1038/s41467-018-07240-8 (2018).
16. Tagirdzhanova, G. *et al*. Complexity of the lichen symbiosis revealed by metagenome and transcriptome analysis of *Xanthoria parietina*. *Current Biology* **35**(4), 799–817, https://doi.org/10.1016/j.cub.2024.12.041 (2025).
17. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**(11), e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
18. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*:1303.3997. https://doi.org/10.48550/arXiv.1303.3997 (2013).
19. Martínez, L. *et al*. Chromosome-Level Assembly and Annotation of the Genome of the Endangered Giant Patagonian Bumble Bee *Bombus dahlbomii*. *Genome Biology and Evolution* **16**, evae146, https://doi.org/10.1093/gbe/evae146 (2024).
20. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(17), 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
21. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
22. Zdobnov, E. M. *et al*. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **49**(D1), D389–D393, https://doi.org/10.1093/nar/gkaa1009 (2021).

23. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in Molecular Biology* **1962**, 227–245, https://doi.org/10.1007/978-1-4939-9173-0_14 (2019).
24. Zhou, Y. *et al*. A high-performance computational workflow to accelerate GATK SNP detection across a 25-genome dataset [J]. *BMC biology* **22**(1), 13, https://doi.org/10.1186/s12915-024-01820-5 (2024).
25. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**(2), 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
26. Bailly-Bechet, M., Haudry, A. & Lerat, E. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA* **5**(1), 13, https://doi.org/10.1186/1759-8753-5-13 (2014).
27. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**(5), 955–964, https://doi.org/10.1093/nar/25.5.955 (1997).
28. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**(22), 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
29. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**(1), 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
30. Delcher, A. L. *et al*. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**(6), 673–679, https://doi.org/10.1093/bioinformatics/btm009 (2007).
31. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**(9), 3100–3108, https://doi.org/10.1093/nar/gkm160 (2007).
32. Blin, K. *et al*. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research* **51**(W1), W46–W50, https://doi.org/10.1093/nar/gkad344 (2023).
33. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**(14), 2068–2069, https://doi.org/10.1093/bioinformatics/btu153 (2014).
34. Fu, L. *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**(24), 3150–3152, https://doi.org/10.1093/bioinformatics/bts565 (2012).
35. Valencia-Pesqueira, L. M. *et al*. Chromosome-level de novo genome assembly of wild, anoxia-tolerant crucian carp, *Carassius carassius*. *Scientific Data* **12**(1), 491, https://doi.org/10.1038/s41597-025-04813-3 (2025).
36. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**, 238, https://doi.org/10.1186/s13059-019-1832-y (2019).
37. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**(5), 1792–1797, https://doi.org/10.1093/nar/gkh340 (2004).
38. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9), 1312–1313, https://doi.org/10.1093/bioinformatics/btu033 (2014).
39. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23), 2971–2972, https://doi.org/10.1093/bioinformatics/btl505 (2006).
40. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**(2), 301–302, https://doi.org/10.1093/bioinformatics/19.2.301 (2003).
41. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**(5), 555–556, https://doi.org/10.1093/molbev/msm088 (1997).
42. De Bie, T., Cristianini, N., Demuth, J. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**(10), 1269–1271, https://doi.org/10.1093/bioinformatics/btl097 (2006).
43. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**(1), gix120, https://doi.org/10.1093/gigascience/gix120 (2018).
44. Li, D. *et al*. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**(10), 1674–1676, https://doi.org/10.1093/bioinformatics/btv033 (2015).
45. Karlsson, F. H. *et al*. Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature Communications* **3**, 1245, https://doi.org/10.1038/ncomms2266 (2012).
46. Qin, J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**(7285), 59–65, https://doi.org/10.1038/nature08821 (2010).
47. Zeller, G. *et al*. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10**(11), 766, https://doi.org/10.15252/msb.20145645 (2014).
48. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **33**(Database issue), D501–D504, https://doi.org/10.1093/nar/gki025 (2005).
49. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**(1), 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
50. Wenger, A. M. *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**(10), 1155–1162, https://doi.org/10.1038/s41587-019-0217-9 (2019).
51. Feng, X. *et al*. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods* **19**(6), 671–674, https://doi.org/10.1038/s41592-022-01478-3 (2022).
52. Portik, D. M. *et al*. Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods. *bioRxiv*, 2024.05.10.593587. https://doi.org/10.1101/2024.05.10.593587 (2024).
53. Gautam, A., Felderhoff, H., Bağci, C. & Huson, D. H. Using AnnoTree to get more assignments, faster, in DIAMOND+ MEGAN microbiome analysis. *Msystems* **7**(1), e01408-21, https://doi.org/10.1128/msystems.01408-21 (2022).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP636986 (2025).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP649239 (2025).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP637105 (2025).
57. *European Nucleotide Archive* https://identifiers.org/ncbi/insdc.gca:GCA_977014525.1 (2025).
58. *European Nucleotide Archive* https://identifiers.org/ncbi/insdc.gca:GCA_977014875.1 (2025).

Acknowledgements

This research project was funded by National Key R&D Program of China (2023YFD1301200).

Author contributions

Zhuang Dong: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. Peng Huang*: Conceptualization, Supervision, Funding acquisition, Writing – review & editing. Jian Guo Zeng*: Conceptualization, Supervision, Project administration, Writing – review & editing. Meng Shan Sun: Investigation, Writing–review & editing. Yue Dong He: Investigation. Wei Xiang: Investigation. Li Zhou: Formal analysis, Investigation. Xin Yao Li: Formal analysis.

Competing interests

All authors declare no competing interests, financial or non-financial, that are directly or indirectly related to the work submitted for publication.

Additional information

Correspondence and requests for materials should be addressed to P.H. or J.G.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.