scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosomal-level genome assembly of the autotetraploid *Anoectochilus roxburghii* (Jinxianlian, Orchidaceae)

Jing Fang[1,2,5], Mei Wu[3,5], Xinyi Zhang[1], Yang Chen[4], Lingjuan Tou[3], Jinjing Xu[3], Xiangjun Kong[3] ✉ & Yingxiong Qiu[1] ✉

Anoectochilus roxburghii **(Orchidaceae), commonly known as Jinxianlian, is a highly valued traditional Chinese herbal medicine. Here, we present a high-quality chromosome-level genome assembly of the cultivar 'Jinkang No.1'. Genome survey analyses suggest that the cultivar is an autotetraploid. The assembled genome spans 5.17 Gb, with a contig N50 of 24.90 Mb, and 93.4% (4.82 Gb) is anchored onto 80 pseudo-chromosomes across 20 homologous groups. Annotation of the genome assembly identifies 76.42% repetitive elements and 88,106 protein-coding genes, with 94.6% of these genes functionally annotated. *Ks* analysis of collinear gene pairs indicates a recent species-specific polyploidization event, likely resulting in autotetraploidization. This reference genome provides a valuable resource for functional genomics, evolutionary biology, and molecular breeding studies of *A. roxburghii* and its related species.**

Background & Summary

Anoectochilus roxburghii (Orchidaceae), commonly known as Jinxianlian, is a perennial herb widely distributed across the tropical and subtropical regions of Asia[1,2]. It is a highly valued traditional Chinese medicine that has been widely utilized as a tonic and health food in China for nearly a thousand years[2]. *A. roxburghii* is renowned for its unique pharmacological activities, including hepatoprotection, antihyperliposis, antihyperglycemia, and antiosteoporosis[3,4]. Empirical studies have identified kinsenoside, flavonoids and alkaloids as the primary bioactive constituents in *A. roxburghii* and its related species[5,6]. However, the content of these bioactive constituents varies greatly among different *Anoectochilus* species, leading to differences in therapeutic efficacy and quality[7]. Accurate identification of *Anoectochilus* species due to their highly similar morphological and anatomical characteristics[7]. Moreover, despite their medicinal importance, the molecular mechanisms underlying the biosynthesis of the major bioactive compounds remain largely unexplored[6,8]. A high-quality reference genome is essential for elucidating the genomic architecture and molecular pathways underlying bioactive compound synthesis, as well as for developing molecular markers to enable precise species identification within *Anoectochilus*. To date, no whole genome sequences of *Anoectochilus* species have been reported, representing a significant gap in both fundamental and applied research.

In this study, we present the first high-quality chromosome-level genome assembly of *A. roxburghii*, generated using a combination of the Pacific Biosciences (PacBio) high-fidelity (HiFi) long reads, high-throughput chromosome conformation capture (Hi-C) reads, and Illumina short-read sequencing reads (Table 1). The final assembly spans 5.17 Gb, with a contig N50 of 24.90 Mb (Table 2), representing not only the first genome for any *Anoectochilus* species, but also the largest orchid genome assembled to date[9–11]. A total of 4.82 Gb of contig sequences were anchored onto 80 pseudo-chromosomes across 20 homologous groups (Table 2). Genome

[1]State Key Laboratory of Plant Diversity and Specialty Crops, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan, Hubei, 430074, China. [2]University of Chinese Academy of Sciences, Beijing, 100049, China. [3]Provincial Key Laboratory of Characteristic Aquatic Vegetable Breeding and Cultivation, Jinhua Academy of Agricultural Sciences (Zhejiang Institute of Agricultural Machinery), Jinhua, Zhejiang, 321000, China. [4]Ningbo Academy of Agricultural Sciences, Ningbo, Zhejiang, 315040, China. [5]These authors contributed equally: Jing Fang, Mei Wu. ✉e-mail: junekx@163.com; qiuyingxiong@wbgcas.cn

libraries	Total data (Gb)	Sequence coverage (×)
Illumina reads	205	39.68
PacBio HiFi reads	292	56.53
Hi-C reads	668	129.31
Total	1,165	225.52

Table 1. Summary statistics of genome sequencing data of *A. roxburghii*.

Item	Number	Size
Genome assembly		
Total contigs	3,735	5,165,849,521 bp
Average contigs	—	1,383,092 bp
Contig L50/N50	80	24,895,964 bp
Total scaffolds	3,452	5,165,877,821 bp
Average scaffolds	—	1,496,488 bp
Scaffold L50/N50	36	60,986,586 bp
Pseudo-chromosomes	80	4,824,854,306 bp
GC content	—	39.10%
BUSCO completeness of assembly	—	96.00%
QV of assembly	—	35.4212
Homology SNPs	746	1.5e-05%
Genome annotation		
Repetitive sequences	—	3,947,667,842 bp
Protein-coding genes	88,106	1,022,708,498 bp
Genes in pseudo-chromosomes	77,368	992,816,206 bp
Average gene length	—	11,120 bp
Average exons per gene	4.85	—
Average exon length	—	264 bp
Average intron length	—	2,552 bp
miRNA	74,465	—
tRNA	36,211	—
rRNA	38,488	—
snRNA	4,325	—
BUSCO completeness of annotation	—	94.60%

Table 2. Summary statistics of assembly and annotation of the assembly genome.

annotation predicted approximately 3.95 Gb of repetitive sequences, 88,106 protein-coding genes and 153,489 non-coding RNA genes (Table 2). This reference genome provides a valuable resource for molecular marker development and future evolutionary, ecological, and functional genomic studies in *Anoectochilus*.

Methods

Sample collection and genome sequencing. For whole-genome sequencing, we followed the protocol of Piet *et al.*[12] to minimize endoreplication-induced variability and exclusively collected fresh young leaves from a tissue-cultured individual of the cultivar 'Jinkang No.1', cultivated at the Jinhua Academy of Agricultural Sciences in Jinhua City, Zhejiang Province, China. Genomic DNA was extracted from young leaves using the modified CTAB method[13]. A short-read library was first generated using the NEB Next® Ultra™ DNA Library Prep Kit for Illumina (NEB, USA), following the manufacturer's recommended protocol. Paired-end (PE) reads (150 bp) were sequenced on the Illumina NovaSeq 6000 platform, yielding 205 Gb (39.68×) of raw sequences (Table 1) for initial genome survey. Subsequently, a PacBio HiFi SMRTbell Library was constructed using the SMRTbell Express Template Prep Kit 2.0 (PacBio, CA, USA). Sequencing on the PacBio Sequel II platform generated approximately 292 Gb of high-fidelity (HiFi) reads, corresponding to an average genome coverage of 56.53× (Table 1). Additionally, a Hi-C library was prepared according to the standard protocol[14] and sequenced on the DNBSEQ-T7 platform (PE 150 bp). This yielded a total of 668 Gb (129.31×) of raw Hi-C data, which was subsequently used for chromosome-level scaffolding and assembly.

For transcriptome sequencing, four tissues (i.e. flowers, roots, stems, and leaves) were collected from the same individual for total RNA extraction using RNAprep Pure Plant Kit (Tiangen, Beijing). RNA-seq libraries were prepared using the TruSeq RNA Library Kit (Illumina, NEB) and sequenced on the Illumina NovaSeq 6000 platform. Each tissue sample generated an average of 6.5 Gb of clean reads, which were used for gene structure annotation (Table 3).

Sample	Raw Base (Gb)	Clean Base (Gb)	Effective (%)	Error (%)	Q30 (%)	GC (%)
Aro-R1	6.39	6.25	97.83	0.03	94.27	44.68
Aro-L1	6.62	6.49	98.12	0.02	94.73	45.27
Aro-S1	6.58	6.45	98.05	0.02	94.38	46.41
Aro-F1	6.62	6.46	97.65	0.03	94.12	44.88

Table 3. Summary of quality statistics for Illumina RNA-seq data.

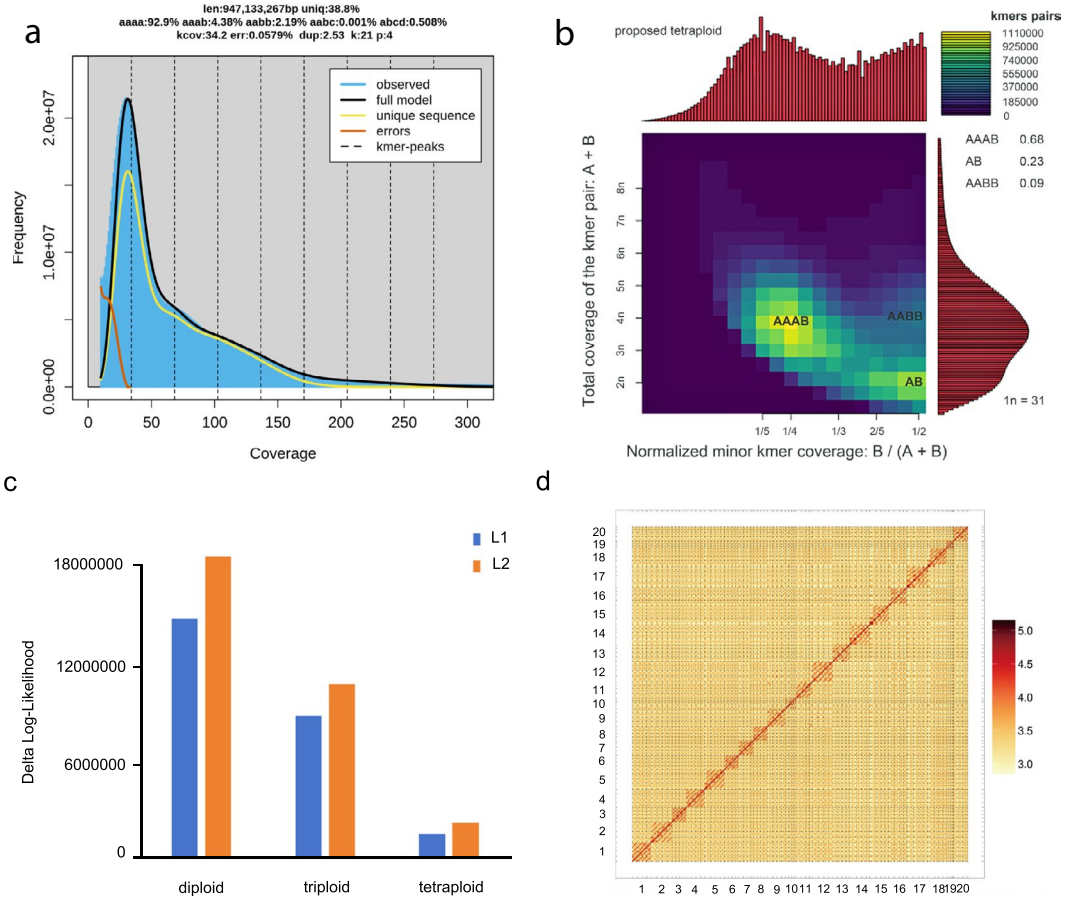

Fig. 1 Genome survey and assembly of *A. roxburghii* cultivar 'Jinkang No.1'. (**a**) The frequency distribution of the 21-mers. (**b**) Smudgeplots for ploidy estimation. (**c**) Bar plots of delta log-likelihood for all fixed models (diploid, triploid and tetraploid), with L1 and L2 representing two subsets of the Illumina genomic sequencing data. (**d**) Hi-C heatmap showing chromatin interaction intensity across assembled pseudo-chromosomes.

Genome survey and assembly. Raw Illumina sequencing reads were first processed using FASTP v0.12.4[15] to remove low quality reads and adapter sequences. After quality control, a total of 170 Gb of clean Illumina short reads were retained. *K*-mer frequency distributions were generated with KMC v3.0[16], and subsequently analyzed using GenomeScope v2.0 (https://github.com/tbenavi1/genomescope2.0). Based on *k*-mers analysis ($k = 21$), the estimated genome size of *A. roxburghii* was 3.79 Gb, with a heterozygosity rate of 2.57% (Fig. 1a). GenomeScope revealed predominantly homozygous *k*-mers (AAAA: 92.9%), while heterozygous patterns (AAAB: 4.38%; AABB: 2.19%) were consistent with expectations for an autotetraploid genome, where heterozygosity in an AAAB configuration is typically higher than in an AABB configuration[17] (Fig. 1a). To further assess genome ploidy and structure, SMUDGEPLOT v0.2.4 (https://github.com/KamilSJaron/smudgeplot) was employed. This analysis indicated a 1:4 minor allele ratio (B/[A + B]) and tetraploid coverage ($4n = 124$), with AAAB heterozygosity (68%) predominating over AABB (9%) (Fig. 1b). Additionally, genome ploidy was inferred using nQuire[18], which estimates ploidy levels by modeling allele frequency distributions from reads mapped to a reference genome or haplotypic contigs. The statistical analysis showed that the tetraploid model had the lowest delta log-likelihood, strongly supporting a tetraploid genome for the cultivar 'Jinkang No.1' (Fig. 1c). Collectively, these findings indicate that the genome of *A. roxburghii* 'Jinkang No.1' possesses a tetraploid genome, with a complex structure likely indicative of an autotetraploid origin.

Type	Number	Length (bp)	Percentage (%)
SINE	1,654	3,944,376	0.08
LINE	341,879	136,155,232	2.64
LTR	2,359,474	3,181,787,886	61.59
DNA	2,271,618	561,355,701	10.87
Unknown	143,630	46,899,890	0.91
Total	—	3,947,667,842	76.42

Table 4. Summary of repeat elements in the assembly genome. Note: Percentage indicate the proportion of bases in each class of repetitive sequences relative to the total genome length.

For long-read sequencing, PacBio CCS reads were filtered using CCS v.6.4.0 (https://github.com/PacificBiosciences/ccs) (minimum length = 50, min-passes = 3, min-rq = 0.99), yielding 17,398,615 high-quality reads. These reads were assembled by HIFIASM v0.16.1[19] with the "--n-hap 4" option, resulting in a 5.17 Gb contig-level assembly composed of 3,735 contigs and a contig N50 of 24.90 Mb. To construct chromosome-level scaffolds, Hi-C data were employed to cluster, order and orient the assembled contigs to different haplotype clusters using ALLHiC v.0.9.8[20] (--minREs 50 --maxlinkdensity 3 --NonInformativeRabio 2). Clean pair-end Hi-C reads were mapped to the contig assembly using BWA 0.7.17[21] with default parameters. The ALLHIC_prune function was used to remove inter-allelic and weak interaction signals from BAM files. Contigs with strong interaction signals were grouped into the same partition using a hierarchical linkage algorithm, and this process was repeated until homologous chromosomes were clearly separated into distinct partitions. Unanchored contigs were subsequently assigned to existing partitions based on the intensities of their chromosomal interaction signals. Within each partition, contigs were further optimized in terms of order and orientation. Chromosome-level scaffolding was finalized using the built-in command of "ALLHiC_build". Manual adjustments were carried out in JUICEBOX v.2.14.00[22], where contigs with strong mutual interactions were grouped into the same cluster. Within each cluster, assembly errors such as misjoin, translocation, and inversion were corrected, and the chromosome boundaries were refined based on the relative strength of intra- versus inter-chromosomal interactions. Ultimately, 20 homologous clusters were resolved, each containing four homologous chromosomes, which were sorted in descending order of length. The final assembly generated a high-quality genome of 5.17 Gb for *A. roxburghii*, with 93.4% of sequences (4.82 Gb) successfully anchored to 80 pseudo-chromosomes (Table 2; Fig. 1d).

Genome annotation and assessment. A combined approach of homology-based alignment and *de novo* prediction was used to identify repetitive sequences. For homolog-based detection, REPEATMASKER (http://www.repeatmasker.org/) was used to identify sequences matching known repeats in the Repbase database (http://www.girinst.org/repbase). For *de novo* prediction, repeat sequence libraries were constructed using LTR_FINDER v.1.07[23], REPEATSCOUT v.1.0.5[24] and REPEATMODELER v.2.0.3[25], all with default parameters. Tandem repeat sequences were detected by TRF v.4.09.1[26]. In total, 3.95 Gb repetitive sequences were predicted, accounting for 76.42% of the genome assembly (Table 2). Long terminal repeats (LTRs) were the predominant category, comprising 80.59% of all identified repetitive sequences (Table 4; Fig. 2a).

Gene structure annotation was performed using homology-based prediction, *ab initio* gene prediction, and transcriptome-assisted prediction with repeat-masked genome sequence. For homology-based prediction, protein-coding sequences from *Apostasia shenzhenica*[10] (GCA_002786265.1), *Dendrobium nobile*[27] (GCA_022539455.1), *Dendrobium chrysotoxum*[28] (GCA_019925795.1), *Phalaenopsis equestris*[29] (GCF_001263595.1), *Vanilla planifolia*[30] (GCA_016413895.1), and *Arabidopsis thaliana* (GCF_000001735.4)[31] were downloaded from NCBI website (https://www.ncbi.nlm.nih.gov) and aligned to the assembled genome using TBLASTN v.2.2.26 with a threshold of E-value ≤ 1e−5[32]. Matching proteins were then realigned to the genome for accurate spliced alignments using GENEWISE v2.4.1[33]. A*b initio* gene prediction was performed with AUGUSTUS v.3.4.0[34], GLIMMERHMM v.3.0.4[35], SNAP[36] and GENEID v.1.4.4[37]. For transcriptome-assisted gene prediction, RNA-seq reads from four tissues were aligned to the reference genome using HISAT v2.2.1[38] with default parameters. The alignment results were then used as input to STRINGTIE v2.2.1[39] for transcript assembly. All genes predicted by the above methods were integrated using EVIDENCEMODELER v1.1.1 (EVM)[40] to produce a non-redundant gene set, which were further refined using PASA v.2.5.1[41] to incorporate untranslated regions (UTRs) and alternative splicing information. In total, 88,106 protein coding genes were predicted in the *A. roxburghii* genome, with an average gene length of 11,120 bp (Table 2).

Gene function annotation were performed by aligning predicted protein sequences against several public databases using BLASTp (E-value ≤ 1e−5), including Swiss-Prot[42], NR (FAM v.35.0[43], KEGG v.102.0[44], GO v.2022-03-22[45] and InterPro v.88.0[46]. As a result, 85,429 genes (96.96% of all predicted genes) were successfully annotated with functional information (Table 5). Transfer RNAs (tRNAs) were identified using tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscanSE/). For ribosomal RNAs (rRNAs), sequences from closely related species were used as references and aligned using BLAST (E-value ≤ 1e−5). Other non-coding RNAs (ncRNAs), including microRNAs (miRNAs) and small nuclear RNAs (snRNAs), were identified by searching the Rfam database using infernal with default parameters (http://infernal.janelia.org/). In total, the genome of *A. roxburghii* contained 74,465 miRNAs, 36,211 tRNAs, 38,488 rRNAs, and 4,325 snRNAs were detected (Table 2). To assess the completeness of the predicted gene set, Benchmarking Universal Single-Copy Orthologs (BUSCO)[47] analysis

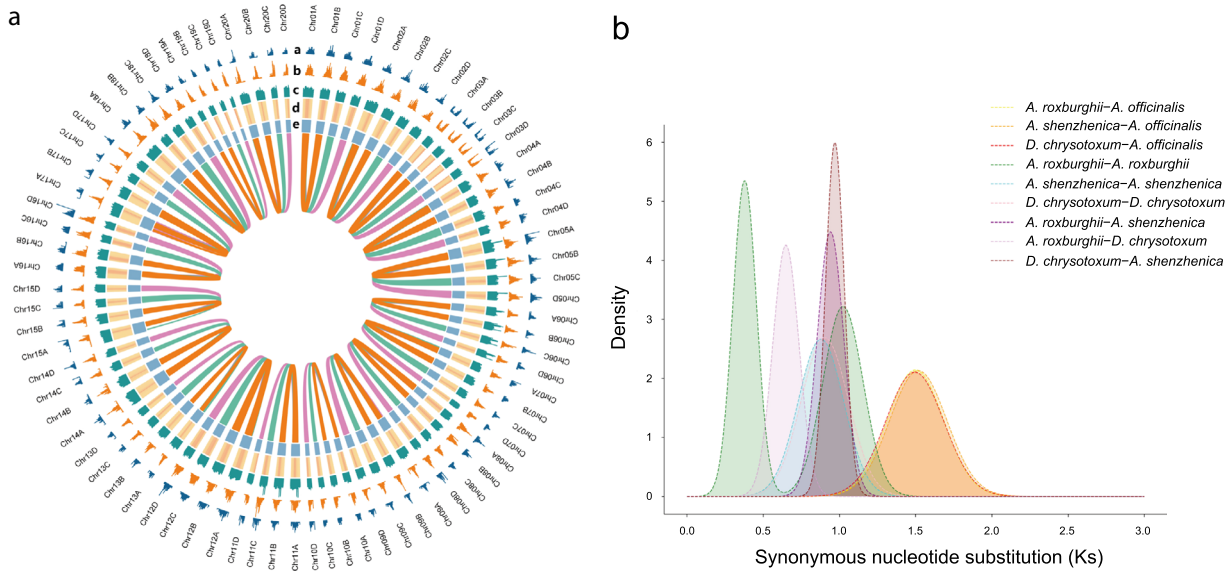

Fig. 2 Genomic features and evolutionary analysis of *A. roxburghii*. (**a**) Circos plot illustrating genomic features of *A. roxburghii*. From the outermost circle to the innermost tracks (**a**–**e**): indel density, SNP density, GC ratio, gene density, and intragenomic syntenic relationships. (**b**) Distribution of synonymous substitution (*Ks*) rates among collinear paralogous gene pairs in *A. roxburghii*, *A. officinalis*, *A. shenzhenica*, and *D. chrysotoxum*.

Type	Number	Percent (%)
Total	88,106	—
NR	83,060	94.27
Swissport	70,186	79.66
KEGG	67,033	76.08
InterPro	83,090	94.31
Pfam	67,050	76.10
GO	55,980	63.54
Annotated	85,429	96.96
Unannotated	2,677	3.04

Table 5. Summary statistics of gene functional annotation in the *A. roxburghii* genome. Note: Percentage indicates the proportion of genes annotated in each database relative to the total number of predicted genes in the genome.

was performed using the embryophyta_odb10 (1,614 BUSCOs), with 94.6% of conserved embryophyte genes successfully detected (Table 2).

Whole-genome duplication analysis. Based on the findings of Zhang *et al.*[10], WGDI v0.6.5[48] analysis of synonymous substitution rates (*Ks*) distributions revealed three distinct peaks (Fig. 2b) through intra- and interspecific comparisons among *A. roxburghii*, *A. shenzhenica*, *D. chrysotoxum*, and *Asparagus officinalis* (GCA_001876935.1). The *Ks* peaks at 0.7–1.1 and 1.5 correspond to two ancient whole-genome duplication (WGD) events that are broadly conserved across orchids and monocots, respectively[10,49]. Notably, a peak around 0.40 in *A. roxburghii* indicates a more recent species-specific polyploidization event, likely representing an autotetraploidization.

Data Records

The datasets generated in this study are deposited in public repositories to ensure accessibility and long-term preservation. All raw sequencing data are available under NCBI BioProject PRJNA1202387[50], comprising PacBio HiFi reads (SRR31827589–SRR31827593[51–55]) for primary genome assembly, Hi-C data (SRR31823870–SRR31823871[56,57]) for chromatin scaffolding, Illumina short reads (SRR31823823–SRR31823825[58–60]) for error correction, and RNA-seq data from four tissues (SRR31824678–SRR31824681[61–64]) for annotation. The final chromosome-scale genome assembly is available under GenBank accession GCA_051820535.1[65], and comprehensive annotation files including structural annotations in GFF3 format and genomic sequences in FASTA format are provided via Figshare[66].

Category	Metric	Value
Reads	Mapping rate (%)	99.71
Genome	Average sequencing depth (×)	38.80
	Coverage (%)	99.21
	Coverage at least 4 × (%)	98.83
	Coverage at least 10 × (%)	98.02
	Coverage at least 20 × (%)	90.98

Table 6. Summary statistics of the integrity of the assembly genome of *A. roxburghii*.

Technical Validation

The quality of the genome assembly was evaluated using three complementary approaches. First, genome completeness was assessed using BUSCO[47], which identified 96.0% of conserved embryophytes genes, indicating high completeness (Table 2). Second, assembly accuracy was measured using MERQURY[67], a *k*-mer-based method, which yielded a high-quality value (QV) of 35.42 (Table 2). Third, Illumina short reads were aligned to the assembled genome using BWA v0.7.17[21] to assess assembly integrity and sequencing uniformity. A total of 99.71% of reads were properly mapped, with 99.21% coverage across the genome (Table 6). Collectively, these results confirm that a genome assembly of high quality, accuracy, and completeness was successfully generated.

Data availability

All raw sequencing data generated in this study are available under NCBI BioProject PRJNA1202387[50], which includes PacBio HiFi reads (SRR31827589–SRR31827593[51–55]), Hi-C data (SRR31823870–SRR31823871[56,57]), Illumina short reads (SRR31823823–SRR31823825[58–60]), and RNA-seq data (SRR31824678–SRR31824681[61–64]). The final assembled genome and annotation files are both available from GenBank under accession number GCA_051820535.1[65] and the Figshare repository[66].

Code availability

No custom scripts or unpublished algorithms were employed during data processing. All analytical procedures in this study were carried out using established bioinformatics tools and pipelines in accordance with their official documentation and recommended protocols. Detailed descriptions of software versions and parameter settings have been provided in the Methods section. Unless otherwise stated, all parameters were applied at their default configurations as suggested by the tool developers.

Received: 7 January 2025; Accepted: 29 August 2025;
Published online: 08 October 2025

References

1. Zhang, Y., Li, Y., Chen, X., Meng, Z. & Guo, S. Combined metabolome and transcriptome analyses reveal the effects of mycorrhizal fungus *Ceratobasidium* sp. AR2 on the flavonoid accumulation in *Anoectochilus roxburghii* during different growth stages. *Int. J. Mol. Sci.* **21**, 564 (2020).
2. Ye, S., Shao, Q. & Zhang, A. *Anoectochilus roxburghii*: a review of its phytochemistry, pharmacology, and clinical applications. *J. Ethnopharmacol.* **209**, 184–202 (2017).
3. Qiu, Y. *et al.* Isolation, structural and bioactivities of polysaccharides from *Anoectochilus roxburghii* (Wall.) Lindl.: a review. *Int. J. Biol. Macromol.* **236**, 123883 (2023).
4. Zeng, B. *et al. Anoectochilus roxburghii* polysaccharide prevents carbon tetrachloride-induced liver injury in mice by metabolomic analysis. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **1152**, 122202 (2020).
5. Lu, L. *et al.* Advances in the therapeutic application and pharmacological properties of kinsenoside against inflammation and oxidative stress-induced disorders. *Front. Pharmacol.* **13**, 1009550 (2022).
6. Ye, B. *et al.* Beneficial effects of endophytic fungi from the *Anoectochilus* and *Ludisia* species on the growth and secondary metabolism of *Anoectochilus roxburghii*. *ACS Omega* **5**, 3487–3497 (2020).
7. Wu, Y. *et al.* Quantitative determination of multi-class bioactive constituents for quality assessment of ten *Anoectochilus*, four *Goodyera* and one *Ludisia* species in China. *Chin. Herb. Med.* **12**, 430–439 (2020).
8. Qi, C. *et al.* Kinsenoside: a promising bioactive compound from *Anoectochilus* species. *Curr. Med. Sci.* **38**, 11–18 (2018).
9. Li, M. *et al.* Genomes of leafy and leafless *Platanthera* orchids illuminate the evolution of mycoheterotrophy. *Nat. Plants* **8**, 373–388 (2022).
10. Zhang, G. *et al.* The *Apostasia* genome and the evolution of orchids. *Nature* **549**, 379–383 (2017).
11. Zhang, G. *et al.* The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci. Rep.* **6**, 19029 (2016).
12. Piet, Q. *et al.* A chromosome-level, haplotype-phased *Vanilla planifolia* genome highlights the challenge of partial endoreplication for accurate whole-genome assembly. *Plant Commun.* **3**, 100330 (2022).
13. Agbagwa, I. O., Datta, S., Patil, P. G., Singh, P. & Nadarajan, N. A protocol for high-quality genomic DNA extraction from legumes. *Genet. Mol. Res.* **11**, 4632–4639 (2012).
14. Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
15. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
16. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
17. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
18. Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S. & Burbano, H. A. nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* **19**, 122 (2018).
19. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

20. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
22. Robinson, J. T. *et al*. Juicebox.js provides a cloud-based visualization system for Hi-C Data. *Cell Syst.* **6**, 256–258.el (2018).
23. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic. Acids. Res.* **35**, W265–W268 (2007).
24. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
25. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
26. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic. Acids. Res.* **27**, 573–580 (1999).
27. Xu, Q. *et al*. Chromosome-scale assembly of the *Dendrobium nobile* genome provides insights into the molecular mechanism of the biosynthesis of the medicinal active ingredient of *Dendrobium*. *Front. Genet.* **13**, 844622 (2022).
28. Zhang, Y. *et al*. Chromosome-scale assembly of the *Dendrobium chrysotoxum* genome enhances the understanding of orchid evolution. *Hortic. Res.* **8**, 183 (2021).
29. Cai, J. *et al*. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat. Genet.* **47**, 65–72 (2015).
30. Hasing, T. *et al*. A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nat. Food* **1**, 811–819 (2020).
31. Reiser, L. *et al*. The Arabidopsis Information Resource in 2024. *Genetics* **227**, iyae027 (2024).
32. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41 (2006).
33. Madeira, F. *et al*. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic. Acids. Res.* **50**, W276–W279 (2022).
34. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic. Acids. Res.* **33**, W465–W467 (2005).
35. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
36. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
37. Alioto, T., Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr. Protoc. Bioinformatics* **64**, e56 (2018).
38. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
39. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
40. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
41. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic. Acids. Res.* **31**, 5654–5666 (2003).
42. Boutet, E. *et al*. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.* **1374**, 23–54 (2016).
43. Mistry, J. *et al*. Pfam: the protein families database in 2021. *Nucleic. Acids. Res.* **49**, D412–D419 (2021).
44. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic. Acids. Res.* **44**, D457–D462 (2016).
45. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic. Acids. Res.* **49**, D325–D334 (2021).
46. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic. Acids. Res.* **51**, D418–D427 (2023).
47. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
48. Sun, P. *et al*. WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).
49. Ruprecht, C. *et al*. Revisiting ancestral polyploidy in plants. *Sci. Adv.* **3**, e1603195 (2017).
50. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1202387 (2024).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31827589 (2024).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31827590 (2024).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31827591 (2024).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31827592 (2024).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31827593 (2024).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31823870 (2024).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31823871 (2024).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31823823 (2024).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31823824 (2024).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31823825 (2024).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31824678 (2024).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31824679 (2024).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31824680 (2024).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31824681 (2024).
65. *NCBI GeneBank* https://identifiers.org/ncbi/insdc.gca:GCA_051820535.1 (2025).
66. Fang, J. Genome assembly and annotation of Jinxianlian (*Anoectochilus roxburghii*). *Figshare* https://doi.org/10.6084/m9.figshare.28094054 (2024).
67. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

Acknowledgements

This work was supported by The Major Science and Technology Projects of Breeding New Varieties of Agriculture in Zhejiang Province (No. 2021C02074).

Author contributions

Y.X.Q. and X.J.K. designed and conceived the study. M.W., L.J.T. and J.J.X. collected the samples. J.F. performed the analyses. J.F. and X.Y.Z. wrote the draft. Y.X.Q., Y.C. and X.Y.Z. revised the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.K. or Y.Q.