scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The telomere-to-telomere chromosome-scale genome assembly of *Acremonium chrysogenum*

Chao Han[1,2,4] ✉, Yiping Zhang[2,4], Hengyu Liang[2], Minliang Chen[2], Jiahuang Li[3] ✉ & Zichun Hua[1] ✉

Acremonium chrysogenum **is a notable filamentous fungus recognized for its essential contribution to the pharmaceutical sector through the biosynthesis of cephalosporin C (CPC). CPC functions as a key intermediate in the biosynthesis of β-lactam antibiotics, which are employed to combat bacterial infections. This study successfully generated a telomere-to-telomere (T2T) chromosome-scale genome sequence for *A. chrysogenum*, combining BGI short reads, PacBio HiFi long reads, and Hi-C technology. This genome sequence contained eight complete chromosomes (29.00 Mb) and a circular mitochondrial genome (27.27 kb), featuring an N50 length of 3.87 Mb. Repetitive elements accounted for 9.65% of genomic content, and a total of 7,745 genes involved in protein coding were annotated. This well-assembled reference genome of *A. chrysogenum* serves as an important foundation for elucidating the biosynthetic pathway of cephalosporin C and for molecular breeding. Furthermore, it offers valuable insights into chromosome organization, genome evolution, and regulatory mechanisms, facilitating future advancements in antibiotic research and fungal biotechnology.**

Background & Summary

Acremonium chrysogenum, recently reclassified as *Hapsidospora chrysogena*, is a critical filamentous fungus in industrial production, best known as the natural producer of cephalosporin C, a key precursor in the synthesis of cephalosporin antibiotics. Antibiotics play an indispensable role in clinical therapy, with cephalosporins accounting for 20% of the antibiotic market[1]. Most antibiotics derive from microbial secondary metabolites and their derivatives[2]. In 1954, Professor Giuseppe Brotzu discovered a fungus capable of producing cephalosporin C[3]. Initially named *Cephalosporium acremonium*, it was later reclassified as *Acremonium chrysogenum* and, most recently in June 2023, renamed *Hapsidospora chrysogena*[4]. Since the name *A. chrysogenum* has been widely adopted in scientific literature, we continue to use this designation in the present study, in accordance with other researchers, to avoid any potential confusion. The discovery revealed the fungus's broad antibacterial properties against diverse bacterial types and marked a critical milestone in antibiotic research enabling large-scale production. Over time, *Acremonium chrysogenum* has emerged as a major producer of cephalosporin antibiotics. As the primary raw material for synthesizing 7-aminocephalosporanic acid (7-ACA), a crucial intermediate, its output and cost of cephalosporin C significantly affect the supply and pricing of cephalosporin antibiotics. Given their broad antimicrobial efficacy and relative safety for both humans and animals, cephalosporin antibiotics are widely used to treat bacterial infections, establishing them as a cornerstone of clinical anti-infection therapy[5].

In 2014, Terfehr and Dominik's team constructed the first draft genome assembly of *Acremonium chrysogenum*, generating 541 scaffolds (N50: 16.69 kb) based on second-generation sequencing data[6]. However, this assembly contained 1,189 sequence gaps, resulting in a highly fragmented genome structure. Although subsequent studies have systematically elucidated the cephalosporin C (CPC) biosynthetic pathway[7], the incomplete

[1]The State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, China. [2]Henan Joincare Biopharma Research Institute Co. Ltd, Jinyuan Street 8, Jiaozuo, 454000, People's Republic of China. [3]School of Biopharmacy, China Pharmaceutical University, Nanjing, 211198, China. [4]These authors contributed equally: Chao Han, Yiping Zhang. ✉e-mail: chaohansyn@gmail.com; lijiah@cpu.edu.cn; zchua@nju.edu.cn

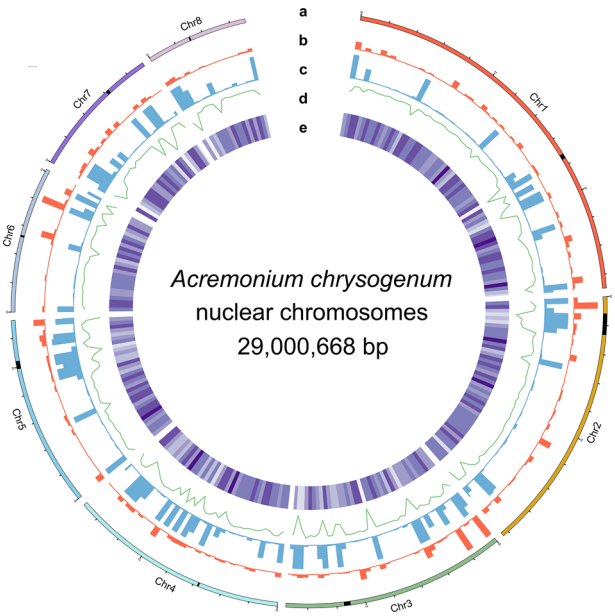

Fig. 1 Chromosomal distribution of different elements in *A. chrysogenum*. Displayed from the outermost to the innermost ring: (**a**) Chromosomes karyotype (The black area in the chromosome is the location of the centromere). (**b**) Copia denisity. (**c**) TE denisity. (**d**) GC denisity. (**e**) Gene density.

Feature	*A. chrysogenum* CGMCC3.3795	*A. chrysogenum* ATCC11550	*A. chrysogenum RNCM 408D*
Genome size	29.0 Mb	28.6 Mb	28.9 Mb
Number of Contigs	10	2,799	2,258
N50 of Contigs	3.8 Mb	149.5 kb	33.9 kb
Number of Scaffolds	8	541	2,258
N50 of Scaffolds	3.9 Mb	166.9 kb	33.9 kb
GC content of the genome (%)	55	55	55
Number of gaps	2	1,189	not provided
Assembly level	Chromosome	Scaffold	Contig

Table 1. Summary of the *A. chrysogenum* genome assembly comparisons.

genome assembly has significantly hindered fundamental research into the species' genetic regulatory mechanisms and secondary metabolite networks. To advance genomic studies of *Hapsidospora* species, the establishment of a high-resolution reference genome has become an urgent necessity. Compared to previous fragmented assemblies, this more complete and contiguous genome assembly provides a molecular foundation for investigating cephalosporin C biosynthesis while also being regarded as an important genetic asset for strain engineering and the development of efficient breeding strategies.

In the present work, we generated the telomere-to-telomere (T2T) chromosome-scale genome (Fig. 1) for *Acremonium chrysogenum*, achieved through the integration of cutting-edge sequencing methods such as BGI-generated short reads, PacBio HiFi reads, and Hi-C scaffolding. The final genome measured 29.00 Mb, featuring an N50 of 3.87 Mb, with the genome sequences organized into eight chromosomes. Annotation identified 7,745 protein-coding genes. Additionally, a circular mitochondrial genome of 27.73 kb in length was successfully assembled. The quality of our assembled genome was significantly improved compared to previous versions (Table 1)[6,8]. The presence of a high-quality, complete genome assembly is crucial for fundamental biological research. The genomic datas generated in this research not only establish a basis for subsequent studies into the molecular mechanisms of cephalosporin biosynthesis in *Acremonium chrysogenum*, but also pave the way for innovative applications in strain improvement, metabolic pathway optimization, and antibiotic development.

Methods

Sample collection and sequencing. *Collection and extraction of DNA samples.* The *A. chrysogenum* strain CGMCC3.3795 was routinely cultivated and maintained on slant culture medium. For genomic DNA extraction, hyphae grown for 7 days at 30 °C and 50% humidity were harvested from these slant cultures. These harvested hyphae were then inoculated into Gelrite Yeast Extract and Malt Extract (GYM) liquid medium

Sequencing	Raw bases (Gb)	Raw reads	N50 length (bp)	Application
HiFi	1.78	143,530	13,877	Assembly
Hi-C	52.09	173,638,290	2 × 150	Chromosome construction
BGI DNBSEQ-T7	3.50	23,363,760	2 × 150	Genome evaluation
RNA-seq	6.92	46,140,436	2 × 150	Genome annotation

Table 2. Summary of sequencing data of *A. chrysogenum* for telomere-to-telomere assembly and genome annotation.

K-mer	Number	Depth	Genome Size (Mb)	Heterozygous Ratio (%)	Duplication Ratio (%)
17	3,108,257,032	95	31.98	0.21	16.11

Table 3. Summary of the *A. chrysogenum* genome size estimation by 17-mer analysis.

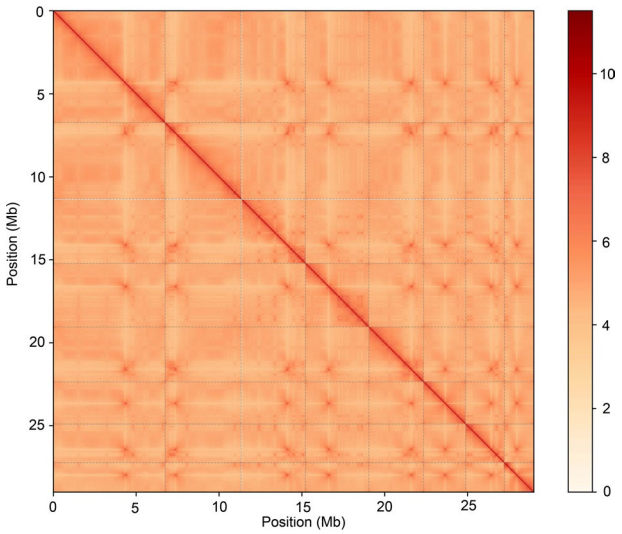

Fig. 2 A visualization of the genome-wide Hi-C interaction matrix at a resolution of 250 kb.

incubated with shaking (180 rpm) for 42 hours at 30 °C before collection. Genomic DNA used for sequencing was subsequently isolated from collected hyphae by applying an adapted CTAB protocol[9].

Illumina, PacBio and Hi-C sequencing. A combined sequencing strategy was used to build a high-quality genome, drawing on BGI short reads, PacBio long reads, and Hi-C methods. Short-read data were produced by constructing a 150 bp insert-size library performed using DNBSEQ-T7 (BGI, Shenzhen, China), yielding 2.33 million reads and 3.50 Gb of raw sequences (Table 2). Using the Template Prep Kit 2.0, a SMRTbell library was generated to acquire PacBio long-read data, which was sequenced via the Sequel II platform (Pacific Biosciences, Menlo Park, USA). This process yielded 1.78 Gb of HiFi reads with an N50 length of 13.88 kb (Table 2). The Hi-C protocol followed standard procedures, including formaldehyde crosslinking, DNA cleavage and biotinylation, fragment ligation, and purification. After random fragmentation of the DNA into 300–500 bp fragments, libraries were prepared and sequenced using the DNBSEQ-T7 platform, resulting in 52.09 Gb of sequencing output (Table 2)[10].

RNA sequencing and analysis. Hyphal RNA was extracted using TRIzol reagent (Invitrogen, CA, USA). Poly-T oligo-bound magnetic beads were used to purify mRNA from total RNA. Following mRNA extraction, RNA-seq libraries were constructed using the VAHTS Universal V6 Kit for MGI (Vazyme, Nanjing, China), with unique index codes assigned as per the manufacturer's protocol. These libraries were then sequenced on the DNBSEQ-T7 platform, producing 6.92 Gb of raw RNA-seq output (Table 2).

Genome sequence assembly. *Genome survey.* We used SOAPnuke software[11] (v2.1.0) to clean the raw short reads and obtain clean reads. The clean reads served to ascertain the entire genomic scale. A k-mer analysis (17-mer) was performed on the short-read data to generate a frequency distribution. Based on this analysis, the genome size was inferred to be roughly 31.98 Mb. Additionally, the content of repetitive elements was approximately 16.11%, while genomic heterozygosity measured about 0.21%, both determined using GCE software[12] (v1.0.2) (Table 3).

Repetitive sequence	Number/% in genome
Tandem repeats	0.71
DNA	0.63
LINE	0.22
SINE	0.00122
LTR	7.61
Unknown	0.88
Total	9.76%
Protein-coding genes	
the total number of genes	7,745
the average gene length (bp)	2,455.11
the average cds_length (bp)	1,570.14
the average exon_number of per gene	3.02
the average exon_length (bp)	748.7
the average intron_length (bp)	96.16
Function annotation	
InterPro	5,827 (75.24%)
GO	5,856 (75.61%)
KEGG_ALL	7,312 (94.41%)
KEGG_KO	3,383 (43.68%)
Swissprot	5,211 (67.28%)
TrEMBL	7,713 (99.59%)
NR	7,714 (99.60%)
Annotated	7,714 (99.60%)
Unannotated	31 (0.40%)

Table 4. Summary of repeat elements.

Genome assembly. Sequencing of SMRTbell libraries took place on the PacBio Sequel II platform, and HiFi reads resulted from ccs processing(v6.3.0; https://github.com/pacificbiosciences/unanimity), applying the option '-minPasses 3′. Highly precise HiFi reads, approximately 15 kilobases long and with over 99% accuracy, served as input for genome assembly using Flye[13] (v2.9) with default settings. To process the Hi-C sequencing data, Trimmomatic[14] (v0.39) was used for adapter trimming and quality filtering. Subsequently, the filtered reads were mapped to the 10 contigs using Juicer[15] (v1.6; https://github.com/aidenlab/juicer) to compute interaction frequencies. Next, 3D-DNA[16] (v180922) was employed to perform misjoin adjustments in two iterations (-r1) under default configuration. After contig orientation, Juicer was used to construct interaction matrices, and manual refinement was carried out using Juicebox Assembly Tools[16] (v1.11.08). These contigs were anchored onto eight pseudochromosomes (Fig. 2), consistent with previous karyotype analysis of *A. chrysogenum*, which confirmed the presence of eight chromosomes in this species[17,18]. In total, the assembled genome of *A. chrysogenum* CGMCC3.3795 is approximately 29.00 Mb achieving an N50 scaffold of 3.87 Mb (Table 1). These sequences were successfully arranged into eight pseudochromosomes.

The mitochondrial genome was assembled from short-read data using GetOrganelle software[19], with the mitochondrial genome sequence of *A. chrysogenum* ATCC11550 (accession number KF757229.1) serving as the reference. The resulting mitochondrial genome is 27.73 kb in size.

Genome annotation. Detection of repetitive sequences within the *A. chrysogenum* genome involved both de novo and homology-based approaches. A de novo repeat library was constructed using RepeatModeler[20] (v2.0.1) (http://www.repeatmasker.org/RepeatModeler/) for initial identification of repetitive sequences. Homology-based tools RepeatMasker[21] and RepeatProteinMask (v4.1.2), along with Repbase library[22,23] (https://www.girinst.org/repbase/), were subsequently applied to annotate characterized transposable elements (TEs). LTR_FINDER[24] (v1.0.7) was utilized to detect long terminal repeat (LTR) retrotransposons. The Tandem Repeat Finder (TRF) package[25] was applied to locate tandemly repeated regions, whereas RepeatMasker was utilized to uncover non-interspersed repetitive elements, such as simple sequences, satellite DNA, and regions of low complexity. Finally, repetitive element datasets obtained from the two methods were integrated into a unified library, and RepeatMasker was applied to annotate the overall repeat content. In total, 2.83 Mb of repetitive elements, comprising 9.76% of the *A. chrysogenum* genome, were identified. Of the identified repeats, LTR elements accounted for the largest portion, comprising 7.61% of the genome (Table 4).

Inferring architecture involved integrating ab initio, homology-based, along with transcriptome-based strategies. Before annotating, RepeatMasker masked *A. chrysogenum*'s repetitive regions, performing both soft and hard masking. For the ab initio-based approach, we adopted Augustus[26] (v3.4.0) and GlimmerHMM[27] (v3.0.4). For the homology-guided approach, protein datasets from *A. chrysogenum* ATCC11550, *Penicillium chrysogenum*, and *Aspergillus luchuensis* were obtained via the National Center for Biotechnology Information (NCBI) database and aligned with the *A. chrysogenum* CGMCC3.3795 genome using TBLASTN[28] (NCBI BLAST

Type		Copy	Total length(bp)	% of genome
miRNA		0	0	0.00
tRNA		111	9,969	0.03
rRNA	rRNA	28	18,559	0.06
	18S	3	5,385	0.02
	28S	3	10,617	0.04
	5S	22	2,557	0.01
snRNA	snRNA	19	2,381	0.01
	CD-box	12	1,288	0.00
	HACA-box	1	179	0.00
	splicing	6	914	0.00
	scaRNA	0	0	0.00

Table 5. Summary statistics of noncoding RNA.

Chr	upstream_start	upstream_end	downstream_start	downstream_end
Superscaffold1	1	135	6,741,395	6,741,520
Superscaffold2	1	238	4,610,212	4,610,372
Superscaffold3	1	127	3,877,092	3,877,216
Superscaffold4	1	120	3,821,661	3,821,780
Superscaffold5	1	134	3,309,648	3,309,767
Superscaffold6	1	133	2,539,188	2,539,301
Superscaffold7	1	122	2,332,562	2,332,690
Superscaffold8	1	120	1,768,644	1,768,666

Table 6. Summary of telomere information of the *A. chrysogenum* genome.

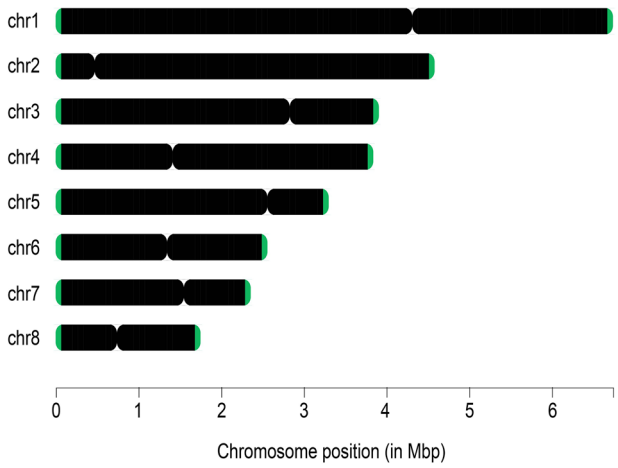

Fig. 3 Distribution map of centromeres and telomeres on the assembled *A. chrysogenum* genome. The telomeric sequences are represented by the green termini, while the central region of crossover corresponds to the centromeric sequence.

v2.11.0+). Subsequently, gene models were predicted through homology using Exonerate[29] (v2.4.0). For the transcriptome-based method, high-quality RNA-Seq data were first assembled into transcript sequences with Trinity[30] (v2.8.5), followed by gene structure annotation using PASA[31] (v2.4.1). MAKER[32] (v3.01.03) combined outcomes from all methods, yielding consensus models. This ultimately provided consistent, separate structures representing the gene architecture of *A. chrysogenum*.

Gene functions were predicted by identifying top hits through comparisons against several functional databases, including NCBI Non-Redundant (NR)[33], Gene Ontology (GO)[34,35], Translation of European Molecular Biology Laboratory (TrEMBL)[36], InterPro[37], Swiss-Prot[36], and the Kyoto Encyclopedia of Genes and Genomes (KEGG)[38], utilizing BLASTP[28] (NCBI BLAST v2.11.0+) setting the E-value to 1E-5. Our analysis identified

chr	start	end
Superscaffold1	4,304,000	4,389,000
Superscaffold2	291,000	660,000
Superscaffold3	2,762,000	2,876,000
Superscaffold4	1,393,000	1,423,000
Superscaffold5	2,469,000	2,608,000
Superscaffold6	1,300,000	1,335,000
Superscaffold7	1,533,000	1,587,000
Superscaffold8	753,000	781,000

Table 7. Summary of predicted centromere information of the *A. chrysogenum* genome.

Application	ReadNum	BaseCount(Gb)	ReadLength(bp)	Q20(%)	Q30(%)	GC Content(%)
survey	23,363,760	3.50	150;150	98.70	95.10	52.90
Hi-C	173,638,290	52.09	150;150	97.31	91.21	55.15

Table 8. Statistics of genomic of the BGI short-read sequencing data.

7,745 protein-coding genes, with 7,714 (99.60%) of these successfully assigned functional annotations in a minimum of one database (Table 4).

Identification of non-coding RNA genes. TRNA genes were detected using tRNAscan-SE[39] (v2.0.9) under default settings. For rRNA identification, RNAmmer[40] (v1.2) was employed. Infernal[41] (v1.1.4) was employed to detect miRNAs and snRNAs by aligning sequences to the Rfam database[42] (v14.1) using standard parameters. In total, the *A. chrysogenum* genome contains 111 tRNAs, 28 rRNAs, 19 snRNAs, and no miRNAs (Table 5).

Telomere and centromere identification. Telomere identification was carried out by detecting characteristic telomeric repeat sequences (CCCTAA at the 5′ end and TTAGGG at the 3′ end). Of the 16 expected telomeres (two per chromosome across eight chromosomes), 15 were successfully identified. HiFi telomeric reads were manually inspected, and the missing telomeric sequences on the final chromosome were patched accordingly (Table 6, Fig. 3). Centromere identification was performed using Tandem Repeats Finder[25] to detect clusters of candidate centromeric tandem repeats. These repeats were found to be continuous and occupied the majority of each chromosome. To determine precise centromere positions, the candidate centromeric tandem repeats were integrated with Hi-C interaction heatmaps (Table 7, Fig. 3).

Data Records

The genome assembly data is available on GenBank under accession number GCA_965197235[43].

Raw sequencing data can be retrieved from the Genome Sequence Archive (GSA) at NGDC under accession number CRA023113[44].

The genome annotation files, gene CDS, and protein data have been submitted to Figshare[45].

Technical Validation

DNA sample quality. We evaluated the extracted DNA's quality and quantity through a combination of methods: a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), a Qubit 3.0 Fluorometer with a dsDNA HS Assay Kit (Life Technologies, Carlsbad, CA, USA), and electrophoresis using 0.8% agarose gel.

Sequencing data assessment. Short-read data were processed using SOAPnuke[11] (v2.1.0), revealing a GC content of 52.9% and Q20/Q30 accuracy rates of 98.7% and 95.1%. The Hi-C dataset had a GC content of 55.15%, with corresponding Q20 and Q30 of 97.31% and 91.21% (Table 8).

Assessment of genome assembly quality. Multiple complementary methods were applied for assessing the *A. chrysogenum* assembly. As a primary step, BUSCO[46] (v5.2.2) was run with the fungi_odb10 database (758 genes) to gauge genome completeness, resulting in a robust 99.3% completeness score. Next, BGI short reads were mapped onto the assembled genome via BWA[47] (v0.7.17), resulting in a 99.99% coverage rate. Third, mapping of TGS long reads was performed with Minimap2[48] (v2.24), resulting in 100% genome coverage. Fourth, LTR_retriever[49] estimated assembly quality using the LTR Assembly Index (LAI), producing a 30.06 score, which meets the threshold for a golden reference genome. Fifth, Mequery[50] (v1.3) evaluated overall base-level quality and completeness, producing metrics of 53.51 QV and 98% completeness. The results collectively confirm the *A. chrysogenum* assembly's high accuracy, completeness, and stability.

Code availability

All experimental procedures were performed using standard, off-the-shelf tools, eliminating the need for custom scripting. Furthermore, every bioinformatics tool and pipeline employed is publicly accessible. Unless specific parameters are otherwise noted, the default settings recommended by their developers were consistently applied.

Received: 12 March 2025; Accepted: 17 July 2025;
Published online: 07 August 2025

References

1. Klein, E. Y. *et al*. Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *Proc. Natl. Acad. Sci.* **115**, E3463–E3470 (2018).
2. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
3. Newton, G. G. F. & Abraham, E. P. Cephalosporin C, a New Antibiotic containing Sulphur and D-α-Aminoadipic Acid. *Nature* **175**, 548–548 (1955).
4. Hou, L. W. *et al*. Redisposition of acremonium-like fungi in *Hypocreales*. *Stud. Mycol.* **105**, 23–203 (2023).
5. Brakhage, A. A. Molecular Regulation of β-Lactam Biosynthesis in Filamentous Fungi. *Microbiol. Mol. Biol. Rev.* **62**, 547–585 (1998).
6. Terfehr, D. *et al*. Genome Sequence and Annotation of *Acremonium chrysogenum*, Producer of the β-Lactam Antibiotic Cephalosporin C. *Genome Announc.* **2**, https://doi.org/10.1128/genomea.00948-14 (2014).
7. Schmitt, E. K., Hoff, B. & Kück, U. Regulation of Cephalosporin Biosynthesis. in *Molecular Biotechnolgy of Fungal beta-Lactam Antibiotics and Related Peptide Synthetases: −/−* (ed. Brakhage, A. A.) 1–43, https://doi.org/10.1007/b99256 (Springer, Berlin, Heidelberg, 2004).
8. Zhgun, A. A. Comparative Genomic Analysis Reveals Key Changes in the Genome of *Acremonium chrysogenum* That Occurred During Classical Strain Improvement for Production of Antibiotic Cephalosporin C. *Int. J. Mol. Sci.* **26**, 181 (2025).
9. Protocol for Extracting DNA from Plant Samples Using CTAB. *OPS Diagnostics LLC* https://opsdiagnostics.com/notes/protocols/ctab_protocol_for_plants.htm.
10. Rao, S. S. P. *et al*. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
11. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience* **7**, gix120 (2018).
12. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Preprint at https://doi.org/10.48550/arXiv.1308.2012 (2020).
13. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
14. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
15. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
16. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
17. Walz, M. & Kück, U. Polymorphic karyotypes in related *Acremonium* strains. *Curr. Genet.* **19**, 73–76 (1991).
18. Dumina, M. V., Zhgun, A. A., Domracheva, A. G., Novak, M. I. & El'darov, M. A. Chromosomal polymorphism of *Acremonium chrysogenum* strains producing cephalosporin C. *Russ. J. Genet.* **48**, 778–784 (2012).
19. Jin, J.-J. *et al*. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
20. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
21. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **25**, 4.10.1–4.10.14 (2009).
22. Jurka, J. Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
23. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
24. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
25. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
26. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
27. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
29. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
30. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).
31. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
32. Cantarel, B. L. *et al*. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
33. 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 Res.* **33**, D501–D504 (2005).
34. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
35. The Gene Ontology Consortium. *et al*. The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
36. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
37. Mitchell, A. *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, D213–D221 (2015).
38. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
39. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
40. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
41. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
42. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
43. *NCBI Genbank*, https://identifiers.org/ncbi/insdc.gca:GCA_965197235.1 (2025).
44. *NGDC Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA023113 (2025).
45. Han, C. First telomere-to-telomere genome assembly of *Acremonium chrysogenum*, Important producer of antibiotics. *figshare* https://doi.org/10.6084/m9.figshare.28440755.v1 (2025).

46. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 227–245, https://doi.org/10.1007/978-1-4939-9173-0_14 (Springer, New York, NY, 2019).
47. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 0 Bytes, https://doi.org/10.6084/M9.FIGSHARE.963153.V1 (2014).
48. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
49. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
50. 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 National Natural Science Foundation of China (32250016, 82130106); the Natural Science Foundation of Jiangsu Province (BK20243001, BG2024026, BE2023695), the Changzhou Municipal Department of Science and Technology (CE20246001, CJ20235009).

Author contributions

Z.H. supervised the research. C.H., Y.Z. and H.L. contributed to research design, C.H., Y.Z., H.L. and M.C. analyzed the data. C.H., Y.Z., J.L. and Z.H. drafted and revised the manuscript. All authors have reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.H., J.L. or Z.H.