BMC Genomic Data

DATA NOTE Open Access

A chromosome-level genome assembly of *Verticillium albo-atrum*, an dangerous quarantine pathogen known for causing verticillium wilt

Yang Chen[1†], Yunfang Chen[2†], Yanfeng Deng[2], Zhenchuan Mao[1], Yan Li[1], Jianlong Zhao[1], Guohua Chen[1*] and Jian Ling[1*]

Abstract

Objectives *Verticillium albo-atrum* is one of the most dangerous quarantine pathogen, which is a soil-borne pathogen known for causing verticillium wilt, a disease that affects a wide range of plants, including many economically important crops. However, the lack of high-quality genome resource has greatly limited the research of molecular and evolutionary mechanisms of *Verticillium albo-atrum*. The highly-quality genome of *Verticillium albo-atrum* provides a valuable resource for better understanding of the biological characteristics.

Data description We sequenced and assembled the genome of *Verticillium albo-atrum* using ONT long reads combined with DNBSEQ-T7 paired-end short reads and anchored 11 contigs into 8 chromosomes using Hi-C chromatin contact information, yielding a 35.95 Mb chromosome-level genome assembly with a N50 of 4.20 Mb. In addition, transcript-based annotation identified 9967 protein-coding genes, of which 84.17% were functionally annotated. BUSCO analysis demonstrated that this genome assembly has a high-level completeness of 96.47% gene coverage.

Keywords *Verticillium albo-atrum*, Genome assembly, Hi-C chromatin contact information

Objective

Verticillium albo-atrum is a considered as a quarantine pathogen which is a soil-borne pathogen that can infect a great number of important dicotyledonous plant species, such as potato, tomato and cotton [1]. The most typical characteristic of infected plants is the appearance of the change of the color of plant vascular bundles, leading to symptoms such as discoloration, wilting, plant stunting and eventual mortality [2]. It is widely distributed in the world, including the United States, Canada, Slovenia, The United Kingdom, Germany, Poland, Japan and other countries [3].

†Yang Chen and Yunfang Chen contributed equally to this work.

*Correspondence:
Guohua Chen
chenguohuacaas@126.com
Jian Ling
lingjian@caas.cn
[1]State Key Laboratory of Vegetable Biobreeding, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2]Comprehensive Technology Center of Suzhou Customs, Suzhou, China

Previous studies have indicated that *Verticillium albo-atrum* poses a serious threat to local agricultural production in many regions. For example, the production of hop has been caused by outbreaks of a highly virulent strain of *Verticillium albo-atrum*, which induces lethal wilt symptoms, with rapid plant withering [4]. In order to facilitate the pathogenic mechanism of *Verticillium albo-atrum*, the high-quality genome resource is of great importance. However, the genome information of *Verticillium albo-atrum* is still limited. Although genome assemblies of *Verticillium albo-atrum* have been reported (Genbank: GCA_024085555.1, GCA_050229555.1 and GCA_002851705.1), all of them are scaffold-level, which are sequenced employing the short reads of IIIumina sequencing technology. In this study, we successfully assembled the *Verticillium albo-atrum* chromosome-level genome using ONT long reads combined with DNBSEQ-T7 paired-end short reads and Hi-C chromatin contact information. The high-quality genome of *Verticillium albo-atrum* would provide a valuable resource into the molecular and evolutionary mechanisms of *Verticillium albo-atrum.*

Data description

The strain of *Verticillium albo-atrum* (DSM12231) was provided by Deutsche Sammlung von Mikroorganismen und Zellkulturen and was routinely grown on potato dextrose agar (PDA) medium at 25 °C. To process the data, the collected strain was harvested from Czapek-Dox medium set at 180 rpm and immediately frozen in liquid nitrogen. High-molecular-weight genomic DNA extraction was used CTAB method [5], followed by purification using a QIAGEN Genomic kit. RNA extraction was performed using a RNeasy Mini Kit (Qiagen) to obtain RNA-seq data [6]. A total of 19.0 Gb DNBSEQ-T7 paired-end short reads,10.3 Gb ONT long reads, 8.1 Gb Hi-C reads and 7.9 Gb paired-end Illumina reads were generated for the genome survey, assembly and annotation (Table 1, Data file 1 [7], Data set 1–4 [8]).

For second-generation sequencing, DNBSEQ library was constructed from the genomic DNA following the Plus DNA Library Prep Kit sequenced by DNBSEQ. For third-generation sequencing, construction of genomic DNA libraries was also performed using the Oxford Nanopore Technologies (ONT) sequencing kit (SQK-LSK110, EXP-NBD196). "k-mer analysis" is performed on DNBSEQ-T7 paired-end short reads using Jellyfish

Table 1 Overview of all figures, data files and tables

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Figure1	Hi-C contact heatmap of *Verticillium albo-atrum*	Image (.jpeg)	Zenodo(https://doi.org/10.5281/zenodo.16568166) [18]
Figure2	Genome features of *Verticillium albo-atrum*	Image (.jpeg)	Zenodo(https://doi.org/10.5281/zenodo.15770251) [34]
Data file1	Summary of library sequencing data	Word file (.docx)	Zenodo(https://doi.org/10.5281/zenodo.15770415) [7]
Data file2	Genome size estimate using GenomeScope	Word file (.docx)	Zenodo(https://doi.org/10.5281/zenodo.16606682) [8]
Data file3	Assembly statistics of the *Verticillium albo-atrum*	Word file (.docx)	Zenodo(https://doi.org/10.5281/zenodo.15770457) [13]
Data file4	Statistics of Hi-C data	Word file (.docx)	Zenodo(https://doi.org/10.5281/zenodo.15770947) [17]
Data set1	ONT sequence long reads of *Verticillium albo-atrum* genomic DNA	Fastq file (.fastq.gz)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set2	DNBSEQ-T7 paired-end short reads of *Verticillium albo-atrum* genomic DNA	Fastq file (.fastq.gz)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set3	Hi-C sequence reads of *Verticillium albo-atrum* genomic DNA	Fastq file (.fastq.gz)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set4	RNA-seq reads of *Verticillium albo-atrum*	Fastq file (. fastq.gz)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set5	Genome assembly data of *Verticillium albo-atrum*	Fasta file (.fasta)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set6	Genome annotation of *Verticillium albo-atrum*	GFF file (.gff)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set7	Coding DNA Sequence of *Verticillium albo-atrum*	Fasta file (. fasta)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set8	Noncoding RNA Sequence of *Verticillium albo-atrum*	Fasta file (. fasta)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Data set9	Secreted Protein Sequence of *Verticillium albo-atrum*	faa.file (.faa)	NGDC(https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832) [8]
Table S1	Stastics of secretory protein	Excel file (.xlsx)	Zenodo(https://doi.org/10.5281/zenodo.15770758) [38]
Table S2	Stastics of secondary metabolic genes (PKSNRPS)	Excel file (.xlsx)	Zenodo(https://doi.org/10.5281/zenodo.15770560) [39]
Supplementary file1	Genome assembly and annotation of *Verticillium albo-atrum* and *Verticillium dahliae*	Excel file (.xlsx).	Zenodo(https://doi.org/10.5281/zenodo.16343579) [40]

(v 2.2.3). When K value is 21, the peak of k-mer distribution curve is 24. The *Verticillium albo-atrum* genome size is estimated to be 34 Mb (Table 1, Data file 2 [9]). The ONT long reads were assembled using Canu (v 1.5) [10] and NextDenovo (v 2.5.0). To improve genome completeness and accuracy, we combined the two assembly results using QUICKMERGE (v 0.3) software. The merged genome were polished by three rounds of Racon using ONT long reads [11] and then were polished by three rounds of Pilon used by DNBSEQ-T7 short reads [12]. As shown in Data file 3 [13], the polished genome primary assembly resulted in 11 contigs, with a length of 35.95 Mb and an N50 length of 4.20 Mb (Table 1, Data file 3 [13], Data set 5–9 [8]). In addition, Hi-C library was constructed and sequenced using DNBSEQ-T7 platform. Clean reads of Hi-C sequences were obtained after quality control, which were processed and mapped to contigs by juicer [14]. Genome error correction and assembly were then performed by 3D-DNA process [15], and manual inspection and adjustment were finally performed by a juicer box [16]. By using Hi-C data, a total of 35.95 Mb assembly, which consists of 11 contigs, were anchored into 8 chromosomes (Table 1, Data file 4 [17], Fig. 1 [18]). We also have identified the telomeres by a python script TelomereSearch.py (https://github.com/jamiemcg/TelomereSearch). Among all eight chromosomes, we identified telomere sequences at both ends of four chromosomes (Chr1, Chr2, Chr6, and Chr7). Additionally, telomeric

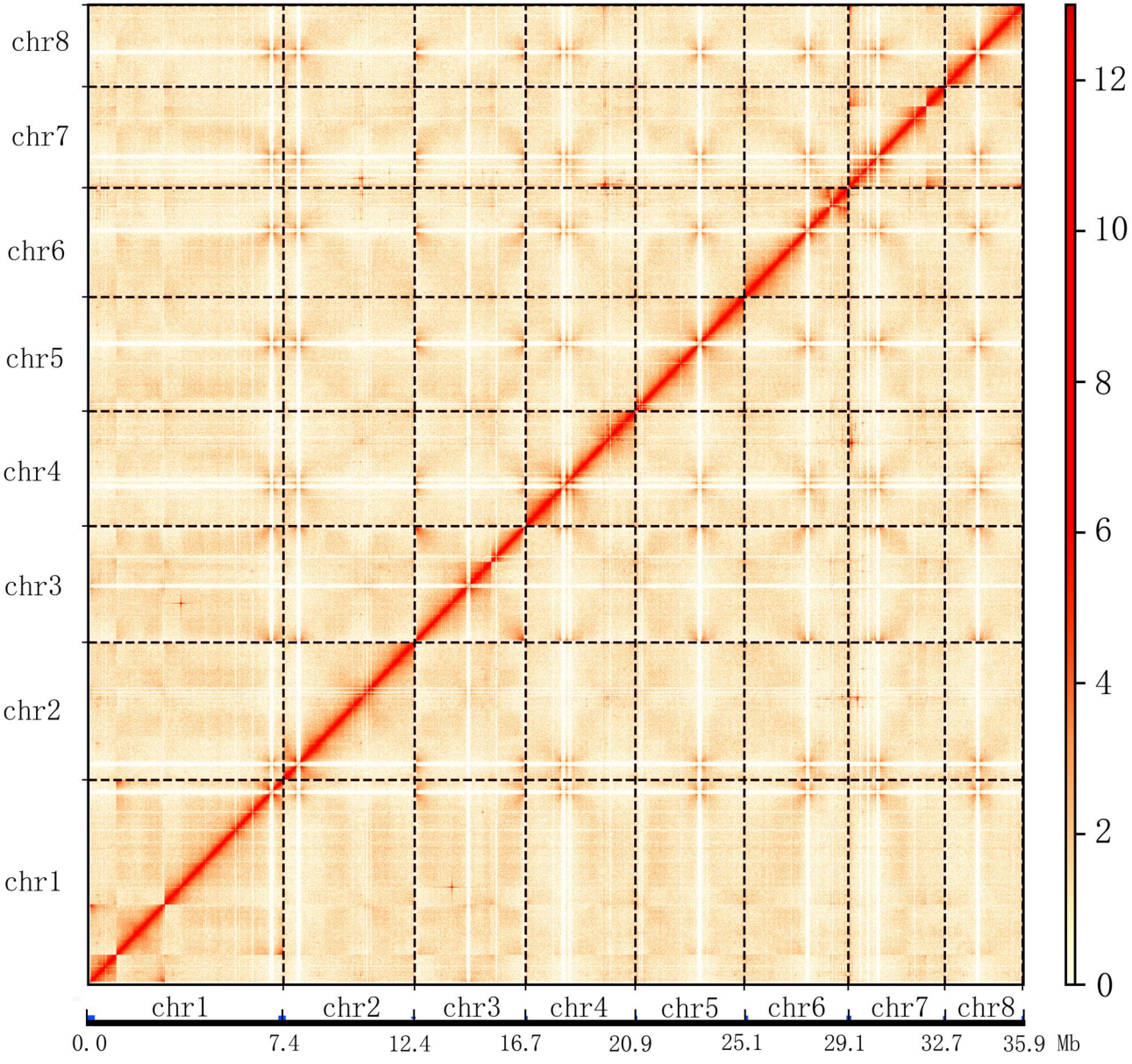

Fig. 1 Hi-C contact heatmap of *Verticillium albo-atrum*

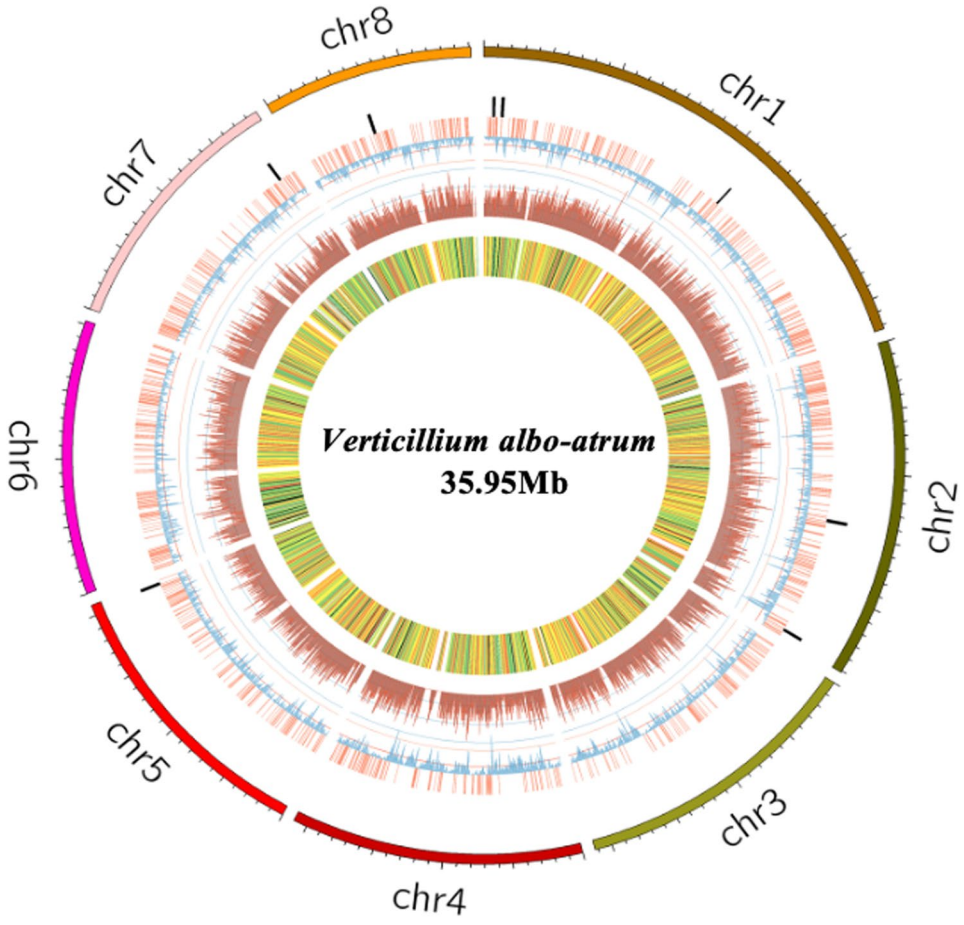

Fig. 2 Genome features of *Verticillium albo-atrum*

repeats were detected at the startings of three other four chromosomes (Chr3, Chr4, and Chr5), while telomeric repeats were detected at the end for Chr8.

A total of 1 μg RNA per sample was used for RNA sequencing library preparation. Sequencing libraries were generated using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England Bio-labs, Ipswich, ME, USA) following the manufacturer's recommendations. RNA-seq data performed on IIumina platform was used for gene annotation. We employed *de novo* and homology methods to identify the repeated sequences. In the process of homology, we used Repbase database (v 20.05) [19]. For *de novo* method, we utilized MITE-HUNTER [20], REPEATSCOUT, LTR-FINDER [21] and PILER-DF [22] to build the *de novo* repeat sequence library. We created the final repeat sequence from the *de novo* constructed database along with the Repbase database. Finally, we employed REPEATMASKER (v 4.0.9) [23] to identify and classify the repeats in *Verticillium albo-atrum.* The total size of the repeats was 3,537,903 bp, accounting for 9.84% of the *Verticillium albo-atrum* genome. For gene structure annotation, we used three methods for the prediction of protein-coding genes. Based on homologous genes, gene prediction was conducted using GEMOMA (v 1.3) [24]. The ab initio predictions were completed through the utilization of GENESCAN [25], AUGUSTUS (v 2.7) [26], GLIMMERHMM (v 3.0.4) [27], GENEID (v 1.4) [28] and SNAP [29]. The RNA-seq reads of *Verticillium albo-atrum* were subjected to assembly via HISAT (v 2.0.4) [30] and STRINGTIE (v 1.2.3) [31], subsequently facilitated by PASA (v 2.0.2) [32] for elucidating unique i-genes, which contributed to the gene prediction process. Finally, we amalgamated the outcomes employing EVIDENCEMODELER [33]. The predicted number of protein-encoding genes was 9967, of which 84.17% were functionally annotated. As illustrated in Fig. 2 (Table 1; Fig. 2 [34]), using 10-kb genomic windows to visualize repeat sequence and gene contents, our analysis revealed that chr4 exhibited the highest repeat content (12.84%), while chr1 showed the highest gene content (32.32%) and highest number of secret proteins (215). The BUSCO completeness of *Verticillium albo-atrum* was 96.47%. In addition, to annotate gene function, we submitted the protein sequence of *Verticillium albo-atrum* to eggnog (v2.1.12) and compared with some protein sequence

databases, such as KOG [35], Nr [36]and Nt [37]. We annotated 1,068 secretory protein-coding genes and 27 secondary metabolic genes (PKS/NRPS), with complete listings in Table S1 (Table 1, Table S1 [38]), and Table S2 (Table 1, Table S1 [39]).

Limitation

Despite the high quality of the current genome assembly of *Verticillium albo-atrum*, there is still room for achieving a higher quality genome of *Verticillium albo-atrum.* This can be accomplished by integrating different assembly outcomes from other assemblers, conducting fine revisions among other approaches. In addition, the genome annotation of *Verticillium albo-atrum* can also be further improved.

Abbreviations

ONT	Oxford Nanopore Technology
Hi-C	High-throughput chromosome conformation capture
RNA-seq	RNA sequencing
BUSCO	Benchmarking Universal Single-Copy Orthologs

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-025-01372-9.

The relative supplementary file has been deposited in Zenodo and is listed in Table 1 (Table 1, Supplementary file1, [40]).

Acknowledgements

Not applicable.

Authors' contributions

All authors reviewed the manuscript.

Funding

This research is funded by the National Key R&D Program of China (2023YFC2604900), Key Research and Development Program of Shaanxi (2021ZDLNY05-01), and the National Natural Science Foundation of China (grant numbers: 31571962, 32372508), China Agriculture Research System (CARS-23).

Data availability

The whole genome data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center,Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number GWHF1CX00000000.1, that is publicly athttps://ngdc.cncb.ac.cn/gsa. The raw sequence data of ***Verticillium albo-atrum***have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center (Nucleic Acids Res 2022), China National Center for Bioinformation/ Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA023962) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa.Other relative data information have been also deposited in Zenodo.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 10 April 2025 / Accepted: 24 September 2025
Published online: 21 October 2025

References

1. Pegg GF, Brady BL. Verticillium wilts[M] . Volume 151. CABI . 2002. p109–110.
2. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, et al. Comparative genomics yields. Insights into niche adaptation of plant vascular wilt pathogens. PloS Pathog. 2011;7:e1002137–1002155.
3. Usami T, Itoh M, Amemiya Y. Mating type gene MAT1-2-1 is common among Japanese isolates of verticillium dahliae. Physiol Mol Plant Pathol. 2008;73:133–7.
4. Radisek S, Jakse J, Javornik B. Genetic variability and virulence among verticillium albo-atrum isolates from hop. Eur J Plant Pathol. 2006;116(4):301–14.
5. Kuo P, Henderson IR, Lambing C. CTAB DNA extraction and. genotyping-by-sequencing to map meiotic crossovers in plants. Methods Mol Biol. 2022;2484:43–53.
6. Zhan G, Guo J, Tian Y, et al. High-throughput RNA sequencing reveals differences between the transcriptomes of the five spore forms of puccinia striiformis f. sp. tritici, the wheat Stripe rust pathogen. Stress Biology. 2023;3:29.
7. Chen Y. Summary of library sequencing data. Zenodo. 2025. https://doi.org/10.5281/zenodo.15770415.
8. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. Nucleic Acids Res. 2024;52(D1):D18–32. https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA032832.
9. Chen Y. Genome size estimate using genomescope. Zenodo. 2025. https://doi.org/10.5281/zenodo.16606682.
10. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, and., Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
11. Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate de Novo genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.
12. Walker BJ et al. Pilon: an integrated tool for comprehensive microbial. Variant Detect Genome Assembly Improvement. PLoS One. 2014;9:e112963.
13. Chen Y. Assembly statistics of the verticillium albo-atrum. Zenodo. 2025. https://doi.org/10.5281/zenodo.15770457.
14. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99–101.
15. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, et al. De Novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.
16. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click. System for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3:95–8.
17. Chen Y. Stastics of Hi-C data. Zenodo. 2025. https://doi.org/10.5281/zenodo.15770947.
18. Chen Y. Figure 1 Hi-C contact heatmap of verticillium albo-atrum.jpeg. Zenodo. 2025. https://doi.org/10.5281/zenodo.16568166.
19. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
20. Han Y, Wessler SR. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from. Genomic sequences. Nucleic Acids Res. 2010;38:e199.
21. Price AL, Jones NC, Pevzner PA. De Novo identification of repeat families in large genomes. Bioinformatics. 2005;21:i351–8.
22. Edgar RC, Myers EW. PILER: identification and. Classification of genomic repeats. Bioinformatics. 2005;21:i152–8.
23. Tarailo-Graovac M, Chen N. Using repeat masker to identify. Repetitive elements in genomic sequences. Curr Protoc Bioinf. 2009;25:4101–41014.
24. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, and., Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;44:e89.
25. Burge C, Karlin S. Prediction of complete gene structures in. human genomic. DNA J Mol Biol. 1997;268:78–94.
26. Stanke M, Waack S. Gene prediction with a hidden Markov. Model and a new intron submodel. Bioinformatics. 2003;19:ii215–25.

27. Majoros WH, Pertea M, Salzberg SL. TigrScan and. Glimmer HMM: two open source Ab initio eukaryotic. gene-finders. Bioinformatics. 2004;20:2878–9.
28. Blanco E, Parra G, Guigó R. Using Geneid to identify genes. Curr Protocols In Bioinformatics. 2007;18(1):4.3. 1-4.3.28.
29. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.
30. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced. Aligner with low memory requirements. Nat Methods. 2015;12:357–60.
31. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved. Reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
32. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Robin CR. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis[J]. BMC Genomics. 2006;7(1):32.
33. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation. Using evidence modeler and the program to assemble spliced alignments. Genome Biol. 2008;9:R7.
34. Chen Y. Genome features of verticillium albo-atrum. Zenodo. 2025. https://doi.org/10.5281/zenodo.15770251.
35. Galperin MY, Wolf YI, Makarova KS, Alvarez RV, Landsman D, Koonin EV. COG database update: focus on. microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021;49:D274–81.
36. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences. (RefSeq): A curated non-redundant sequence. Database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35:D61–5.
37. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res. 2004;32:D138–1141.
38. Chen Y. Stastics of secretory protein.xlsx. Zenodo. 2025. https://doi.org/10.5281/zenodo.15770758.
39. Chen Y. Stastics of secondary metabolic genes (PKSNRPS). Zenodo. 2025. https://doi.org/10.5281/zenodo.15770560.
40. Chen Y. Genome assembly and annotation of *Verticillium albo-atrum* and *Verticillium dahliae*. Zenodo. 2025. https://doi.org/10.5281/zenodo.16343579.

Publisher's Note