BMC Genomic Data

DATA NOTE Open Access

A high-quality chromosome-level genome assembly of *Antiaris toxicaria*

Weicheng Huang[1], Jiaxin Xiang[2], Yamei Ding[1,3], Wanzhen Liu[1,3], Ni Fang[1], Yongmei Xiong[4], Seping Dai[4*] and Hui Yu[1,3,5*]

Abstract

Objectives *Antiaris toxicaria* is a tall tree belonging to the Moraceae family, known for its medicinal value. Its latex contains various cardiac glycosides, which hold significant research and potential application value. However, the lack of genomic resources for *A. toxicaria* currently hinders molecular genetic studies on its medicinal components. For its effective conservation and elucidation of the distinctive genetic traits of and medical components, we present its chromosome-level genome assembly.

Data description Here, we assembled two haplotypes of *A. toxicaria*, including a 671.73-Mb HapA subgenome containing 27,213 genes and a 666.41-Mb HapB subgenome containing 28,840 genes. Their contig N50 sizes were 90.18 and 90.29 Mb, respectively. The transposable elements represented 61.15% and 64.13% of the total assembled genome in HapA and HapB subgenome, respectively. A total of 27,213 and 28,840 genes were predicted in the two haplotypes. Hopefully, this chromosome-level genome of *A. toxicaria* will provide a valuable resource to enhance understanding of the biosynthesis of medicinal compounds.

Keywords *Antiaris toxicaria*, Genome assembly, HiFi, ONT ultralong, Hi-C, Transcriptome

Objective

Antiaris is a genus in the Moraceae, all species are large trees with tall and straight trunk and plank-like roots. There are approximately seven species and three varieties worldwide, with only one species, *Antiaris toxicaria*, found in China, distributed in Guangdong, Hainan, Guangxi, and southern Yunnan. *A. toxicaria* is well known for its ornamental value, ecological function and cultural importance. Previous studies on the secondary metabolites of this species have indicated that its latex [1, 2] and seeds [3] are rich in cardiac glycosides. The sap of the leaves and branches of *A. toxicaria* contains highly toxic substances [4], with the main toxic components being cardiac glycosides such as α-antiarin [5], antioside [6, 7], and convallatoxin [8, 9], which have effects including enhancing heart function, inducing vomiting and diarrhea, and possessing anesthetic properties [10].

*Correspondence:
Seping Dai
gzifla_dsp@gz.gov.cn
Hui Yu
yuhui@scib.ac.cn
[1]Key Laboratory of National Forestry and Grassland Administration on Plant Conservation and Utilization in Southern China, South China Botanical Garden, The Chinese Academy of Sciences, Guangzhou 510650, China
[2]Department of Computer Science, Hong Kong Baptist University, Kowloon 999077, Hong Kong, China
[3]Guangdong Provincial Key Laboratory of Applied Botany, The Chinese Academy of Sciences, Guangzhou 510650, China
[4]Guangzhou Institute of Forestry and Landscape Architecture, Guanzhou 510405, China
[5]State Key Laboratory of Plant Diversity and Specialty Crops, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, Guangdong 510650, China

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Summary of library sequencing data	Word file (.docx)	Figshare, https://doi.org/10.6084/m9.figshare.28328342 [13]
Data file 2	K-mer-based estimation of genome characters	Image file (.jpg)	Figshare, https://doi.org/10.6084/m9.figshare.28328369 [14]
Data file 3	Hi-C interactive heatmap	Image file (.jpg)	Figshare, https://doi.org/10.6084/m9.figshare.28328372 [15]
Data file 4	Statistics of genome assembly and annotation	Word file (.docx)	Figshare, https://doi.org/10.6084/m9.figshare.28328378 [16]
Data file 5	Summary of gene functional annotation	Word file (.docx)	Figshare, https://doi.org/10.6084/m9.figshare.28328387 [17]
Data file 6	Gene function on Haplotype A	TXT file (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.28328396 [18]
Data file 7	Gene function on Haplotype B	TXT file (.txt)	Figshare, https://doi.org/10.6084/m9.figshare.28328405 [19]
Data file 8	Statistical results of the repetitive sequences	Word file (.docx)	Figshare, https://doi.org/10.6084/m9.figshare.28328408 [20]
Data file 9	Summary of noncoding RNA genes	Word file (.docx)	Figshare, https://doi.org/10.6084/m9.figshare.28328414 [21]
Data set 1	Illumina survey data of *A. toxicaria*	Fastq file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR32205349
Data set 2	PacBio HiFi reads of *A. toxicaria*	Bam file (.bam)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR32203223
Data set 3	ONT Ultra-long reads of *A. toxicaria*	Fastq file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR32203292
Data set 4	Hi-C reads of *A. toxicaria*	Fastq file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR32205131
Data set 5	Genome assembly data for HapA	Fasta file (.fasta)	Figshare, https://doi.org/10.6084/m9.figshare.28328498 [22]
Data set 6	Genome assembly data for HapB	Fasta file (.fasta)	Figshare, https://doi.org/10.6084/m9.figshare.28328528 [23]
Data set 7	Transcriptome data of *Antiaris toxicaria*	Fastq file (.fastq)	NCBI Sequence Read Archive, https://identifiers.org/ncbi/insdc.sra:SRR32202871
Data set 8	Gene prediction on HapA	GFF3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.28328429 [24]
Data set 9	Gene prediction on HapB	GFF3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.28328432 [25]
Data set 10	Transposable elements annotation on HapA	GFF3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.28328444 [26]
Data set 11	Transposable elements annotation on HapB	GFF3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.28328450 [27]
Data set 12	Noncoding RNA prediction on HapA	GFF3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.28328456 [28]
Data set 13	Noncoding RNA prediction on HapB	GFF3 file (.gff3)	Figshare, https://doi.org/10.6084/m9.figshare.28328459 [29]

Additionally, HPLC screening of *A. toxicaria* extracts revealed the presence of gallic acid, catechins, chlorogenic acid, caffeic acid, ellagic acid, epigallocatechin, rutin, isoquercitrin, quercitrin, quercetin and kaempferol [11]. Therefore, *A. toxicaria* holds significant research and commercial value. However, our understanding of the biosynthesis and regulatory mechanisms of secondary metabolites in *A. toxicaria* is limited, and further research is needed on the candidate genes and transcription factors involved in cardiac glycoside biosynthesis pathways.

In this study, we successfully assembled the *A. toxicaria* chromosome-level genome using high-fidelity (HiFi) reads and high-throughput chromosome conformation capture (Hi-C) sequencing technologies. This study reports the high-quality genome of *A. toxicaria*. We believe that this research will provide important resources for studying the biosynthetic mechanisms of this species.

Data description

A. toxicaria samples were obtained from the South China Botanical Garden (23.18°N, 113.36°E), Guangzhou, China. Fresh leaves of *A. toxicaria* were collected for PacBio HiFi, ONT ultralong, and Hi-C sequencing. A PCR-free SMRTBell library was constructed using high-quality purified long reading DNA for PacBio HiFi sequencing. The ONT PromethION sequencer was used to generate ONT ultralong reads. Hi-C libraries were constructed and sequenced using BGI platform. Stems, leaves, and seeds of *A. toxicaria* were frozen in liquid nitrogen and stored at – 80 °C for transcriptome analyses. All Illumina sequencing data were filtered to obtain clean data using the fastp v0.23.1 [12] for subsequent analysis. A total of 128.34 Gb (~ 191.06 × coverage) paired-end Illumina reads (Table 1; Data set 1), 32.7 Gb (~ 48.68 × coverage) PacBio HiFi long reads (Table 1; Data set 2), 16.71 Gb ONT Ultra-long reads (~ 24.87 × coverage) (Table 1; Data set 3), and Hi-C reads (~ 126.14×coverage) (Table 1; Data set 4) were generated for the genome survey, and assembly (Table 1; Data file 1).

Before genome assembly, we used the GCE (Genomic Charactor Estimator) v 1.0.2 [30] to assess the genome size based on Illumina short reads. The genome size of *A. toxicaria* was estimated to be approximately 729.84 Mb based on the assessment results when using kmer length of 17 bp, showing a high degree of repeat content (70.62%) and heterozygosity (0.57%) (Table 1; Data file 2).

The PacBio HiFi, ONT Ultra-long, and Hi-C data were assembled using Hifiasm [31] with the default parameters. Then, the Hi-C data was aligned to the HapA and HapB subgenomes, respectively, and classified as valid or invalid interaction pairs using the Juicer pipeline [32] and YaHS v1.2 [33]. Meanwhile, misassembled contigs were detected, corrected manually and oriented to chromosomes through Juicebox v1.11.08 [32]. The corrected ONT and PacBio HiFi reads were used to replace the gap region using TGS-GapCloser v1.2.1 [34], and then obtained the haplotype-resolved gap-free genome of *A. toxicaria*. Finally, the *A. toxicaria* genome was ultimately phased into two haplotypes, comprising a total of 26 pseudochromosomes, with HapA spanning approximately 671.73 Mb and featuring a contig N50 of 90.18 Mb (Table 1; Data files 3–4; Data set 5). Similarly, HapB spans around 666.41 Mb with a contig N50 of 90.29 Mb (Table 1; Data files 3–4; Data set 6). Moreover, the GC content of HapA was 35.65%, while that of HapB was 35.61% (Table 1; Data file 4).

The genome completeness was assessed by searching the gene content of the embryophyta_odb10 database (1,614 expected genes from the embryophyta) with BUSCO v4.1.2 [35], showed that, the proportions of complete BUSCOs (including single-copy and multi-copy) of these two haplotypes were 98.5% and 98.6%, respectively (Table 1; Data file 4). The quality of repetitive genomic regions was assessed using the LAI v3.2 program [36], which exhibited LAI values of 16.4 (HapA) and 14.72 (HapB) (Table 1; Data file 4). Then the per-base consensus accuracy (QV) was estimated with Merqury v1.365 [37] using PacBio HiFi long reads, resulting in QV values of 47.13 and 47.1 (Table 1; Data file 4). Short-reads and long-reads were mapped to the genome with BWA v0.7.13-r1126 [38] and Minimap2 v2.21 [39], and we found that the genome coverage of sequencing data exceeded 99% (Table 1; Data file 4).

Protein-coding genes was predicted using homology-based, transcriptome-based, and ab initio prediction methods. First, we used homologies as protein-based evidence for predicting gene sets using GeneWise v2.4.1 [40]. Transcriptome data were mapped using HISAT2 v2.1.0 [41] (Table 1; Data set 7). ab initio prediction using packages AUGUSTUS v3.4.0 [42], trained by the transcriptome data. To generate a comprehensive protein-coding gene set, we used the GETA v2.6.1 (Genome-wide Electronic Tool for Annotation) pipeline (https://github.com/chenlianfu/geta) to integrate annotations from all homology-based, transcriptome-based, and ab initio predictions. Then Functional annotation of the protein-coding genes was carried out by blast searches against databases, including the NCBI nr [43], SwissPort [44], KOG [45], eggNOG [46], Pfam [47], GO [48], and KEGG [49]. In total, we obtained 27,213 and 28,840 protein-coding genes of the HapA and HapB subgenomes, respectively (Table 1; Data file 4; Data sets 8–9). Moreover, 26,906 (98.87%) genes of the HapA subgenome and 26,360 (98.8%) genes of the HapB subgenome were supported by multiple functional databases (Table 1; Data files 5–7).

To identify Transposable elements (TEs), we used the pipeline of Extensive de-novo TE Annotator (EDTA) v2.1.0 [50], which combines both structural-based and homology-based predictions. For noncoding RNA prediction, the tRNA genes were predicted using tRNAscan-SE v2.0.6 [51]. Others, including miRNA, rRNA and snRNA genes, were detected by comparison with the Rfam database [52] using CMsearch v1.1.3 [53] with the default parameters. A total of 427.39 Mb of TEs were identified, accounting for 64.13% of the HapB subgenome, which was higher than the HapA subgenome (Table 1; Data file 8; Data sets 10–11). In addition, the long terminal repeat retrotransposons (LTRs) were the predominant repeats covering 55.63% (370.77 Mb) of the HapB subgenome, and the Copia and Gypsy-type LTRs were the largest LTR subfamilies, accounting for 15.89% (105.89 Mb) and 39.10% (260.58 Mb), respectively (Table 1; Data file 8; Data sets 10–11). Moreover, 456 tRNAs and 111 miRNAs were identified in the *A. toxicaria* subgenome (Table 1; Data file 9; Data sets 12–13). 1,637 and 1,182 rRNAs were identified in the HapA and HapB subgenomes, respectively (Table 1; Data file 9; Data sets 12–13).

Limitations

Genome and transcriptome data are available in this study, but there is a lack of proteome and metabolome data from different tissues, as well as multi-omics correlation analysis.

Abbreviations

HiFi	High fidelity
ONT	Oxford Nanopore Technology
Hi-C	High-throughput chromosome conformation capture
HapA	Haplotype A
HapB	Haplotype B
BUSCO	Benchmarking Universal Single-Copy Orthologs
LAI	LTR Assembly Index
QV	Consensus quality
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOG	Eukaryotic Orthologous Groups
Nr	Non-redundant
LTR	Long-Terminal Repeat
TE	Transposon Eleme
NCBI	National Center for Biotechnology Information

Acknowledgements

We thank the reviewers for their time, expertise, and helpful suggestions to improve our manuscript.

Authors' contributions
HY supervised the project. WCH, YMD, WZL and JXX conducted data analysis. WCH, NF collected samples. YMX and SPD edited the manuscript. All authors contributed to writing the manuscript.

Funding
This work was supported by National Key R & D Program of China (2023YFE0107400), Guangzhou Ecological Landscape Technology Collaborative Innovation Center (202206010058), Science and Technology Projects in Guangzhou (E33309), and the Guangdong Flagship Project of Basic and Applied Basic Research (2023B0303050001).

Data availability
The raw sequencing data for HiFi, Hi-C, RNA-seq, and ONT Ultra-long reads were submitted to NCBI Sequence Read Archive database under BioProject accession PRJNA1218215. The chromosomal-level genome assembly file were deposited in the Figshare database with DOIs 10.6084/m9.figshare.28328498 [22] and 10.6084/m9.figshare.28328528 [23]. Moreover, the gene structure, gene function, TE and non-coding RNA annatition files also have been deposited at the Figshare database [24–29].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 December 2024 / Accepted: 10 March 2025
Published online: 24 March 2025

References

1. Liu Q, Tang JS, Hu MJ, Liu J, Chen HF, Gao H, et al. Antiproliferative cardiac glycosides from the latex of *Antiaris toxicaria*. J Nat Prod. 2013;76(9):1771–80. https://doi.org/10.1021/np4005147.
2. Que DM, Gan YJ, Zeng YB, Dai HF, et al. Cytotoxic cardenolides from the latex of *Antiaris toxicaria*. J Trop Subtrop Bot. 2010;18:440–4.
3. Zuo WJ, DongWH, Zhao YX, Chen HQ, Mei WL, Dai HF. Two new strophanthidol cardenolides from the the seeds of *Antiaris toxicaria*. Phytochem Lett. 2013;6(1):1–4. https://doi.org/10.1016/j.phytol.2012.10.001.
4. Carter CA, Forney RW, Gray EA, Gehring AM, Schneider TL, Young DB, et al. Toxicarioside A. A new cardenolide isolated from *Antiaris toxicaria* latex-derived dart poison. Assignment of the tH- and tsC-NMR shifts for an antiarigenin aglycone. Tetrahedron. 1997;53(40):13557–66. https://doi.org/10.1016/S0040-4020(97)00895-8.
5. Tian DM, Qiao J, Bao YZ, Liu J, Zhang XK, Sun XL, et al. Design and synthesis of biotinylated cardiac glycosides for probing Nur77 protein inducting pathway. Bioorg Med Chem Lett. 2019;29(5):707–12. https://doi.org/10.1016/j.bmcl.2019.01.015.
6. Kopp B, Bauer WP, Bernkop-Schnurch A. Analysis of some Malaysian dart poisons. J Ethnopharmacol. 1992;36(1):57–62. https://doi.org/10.1016/0378-8741(92)90061-u.
7. Agrawal P, Akhade M, Laddha K, Narkhede S, Mirgal A, Salunke C. Quantification of Convallatoxin in *Antiaris toxicaria* leuschseeds by RP-HPLC. Anal Chem Lett. 2014;4(3):172–7. https://doi.org/10.1080/22297928.2014.925821.
8. Yang SY, Kim NH, Cho YS, Lee H, Kwon HJ. Convallatoxin, a dual inducer of autophagy and apoptosis, inhibits angiogenesis in vitro and in vivo. PLoS ONE. 2014;9(3):e91094. https://doi.org/10.1371/journal.pone.0091094.
9. Shi LS, Liao YR, Su MJ, Lee AS, Kuo PC, Damu AG, et al. Cardiac glycosides from *Antiaris toxicaria* with potent cardiotonic activity. J Nat Prod. 2010;73(7):1214–22. https://doi.org/10.1021/np9005212.
10. Mei WL, Gan YJ, Dai HF. Advances in studies on chemical constituents of *Antiaris toxicaria* and their Pharmacological activities. Tradit Chin Herb Drugs. 2008;39:151–4.
11. Subiono T, Tavip MA. Qualitative and quantitative phytochemicals of leaves, bark and roots of *Antiaris toxicaria* lesch., a promising natural medicinal plant and source of pesticides. Plant Sci Today. 2023;10(1):5–10. https://doi.org/10.14719/pst.1896.
12. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
13. Huang WC. Data files 1: Summary of library sequencing data. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328342.
14. Huang WC. Data files 2: K-mer-based Estimation of genome characters. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328369
15. Huang WC. Data files 3: Hi-C interactive heatmap. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328372
16. Huang WC. Data files 4: Statistics of genome assembly and annotation. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328378.
17. Huang WC. Data files 5: Summary of gene functional annotation. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328387.
18. Huang WC. Data files 6: gene function on haplotype A. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328396
19. Huang WC. Data files 7: gene function on haplotype B. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328405
20. Huang WC. Data files 8: Statistical results of the repetitive sequences. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328408
21. Huang WC. Data files 9: Summary of noncoding RNA genes. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328414.
22. Huang WC. Data set 5: Genome assembly data for Haplotype A. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328444.
23. Huang WC. Data set 6: Genome assembly data for Haplotype B. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328444
24. Huang WC. Data set 8: gene prediction on haplotype A. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328498
25. Huang WC. Data set 9: gene prediction on haplotype B. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328528
26. Huang WC. Data set 10: transposable elements annotation on haplotype A. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328444
27. Huang WC. Data set 11: transposable elements annotation on haplotype B. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328450
28. Huang WC. Data set 12: noncoding RNA prediction on haplotype A. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328456
29. Huang WC. Data set 13: noncoding RNA prediction on haplotype B. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28328459
30. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv preprint arXiv:1308.2012. https://doi.org/10.48550/arXiv.1308.2012
31. Feng X, Cheng H, Portik D, Li H. Metagenome assembly of high-fidelity long reads with hifiasm-meta. Nat Methods. 2022;19(6):671–4. https://doi.org/10.1038/s41592-022-01478-3.
32. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a One-Click system for analyzing Loop-Resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8. https://doi.org/10.1016/j.cels.2016.07.002.
33. Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39(1):btac808. https://doi.org/10.1093/bioinformatics/btac808.
34. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience. 2020;9(9):giaa094. https://doi.org/10.1093/gigascience/giaa094.
35. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.
36. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR assembly index (LAI). Nucleic Acids Res. 2018;46(32):e126. https://doi.org/10.1093/nar/gky730.
37. Rhie A, Walenz BP, Koren S, Phillippy AM, Merqury. Reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):245. https://doi.org/10.1186/s13059-020-02134-9.
38. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95. https://doi.org/10.1093/bioinformatics/btp698.
39. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.

40. Birney E, Clamp M, Durbin R. GeneWise and genomewise. Genome Res. 2004;14(5):988–95. https://doi.org/10.1101/gr.1865504.
41. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60. https://doi.org/10.1038/nmeth.3317.
42. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: Ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9. https://doi.org/10.1093/nar/gkl200.
43. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National center for biotechnology information in 2023. Nucleic Acids Res. 2023;51(D1):D29–38. https://doi.org/10.1093/nar/gkac1032.
44. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement trembl in 1999. Nucleic Acids Res. 1999;27(1):49–54. https://doi.org/10.1093/nar/27.1.49.
45. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41. https://doi.org/10.1186/1471-2105-4-41.
46. Hernandez-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR, et al. EggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Res. 2023;51(D1):D389–94. https://doi.org/10.1093/nar/gkac1022.
47. Mistry J, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913.
48. Ashburner M, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25:25–9. https://doi.org/10.1038/75556.
49. Kanehis M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
50. Su W, Ou S, Hufford MB, Peterson T. A tutorial of EDTA: extensive de Novo TE annotator. Methods Mol Biol. 2021;2250:55–67. https://doi.org/10.1007/978-1-0716-1134-0_4.
51. Chan PP, Lin BY, Mak AJ, Lowe TM. TRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021;49(16):9077–96. https://doi.org/10.1093/nar/gkab688.
52. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43(D1):D130–7. https://doi.org/10.1093/nar/gku1063.
53. Cui X, Lu Z, Wang S, Jing-Yan Wang J, Gao X, CMsearch. Simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. Bioinformatics. 2016;32(12):i332–40. https://doi.org/10.1093/bioinformatics/btw271.

Publisher's Note