scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the parasitoid wasp *Aenasius arizonensis*

Wan-Ying Dong[1,7], Tian-Yu Huang[2,7], Sheng-Yuan Zhao[3], Juan Zhang[4], Yang Lei[1,5], Jun Huang[1], Zhong-Shi Zhou[2,6] ✉ & Yao-Bin Lu[1,3] ✉

Aenasius arizonensis **is an important solitary endoparasitoid successfully used for biocontrol of cotton mealybug. However, lacking genomic resources has limited molecular-level investigations. Our exploration produced a superior genomic assembly of *A. arizonensis* from the chromosome level by combining MGISEQ short reads, Hi-C scaffolding, and PacBio Revio sequencing techniques. The genome measured 398.69 Mb, including a contig N50 of 4.73 Mb, a BUSCO completeness level of 97.07%, and a scaffold N50 of 35.96 Mb. Hi-C data were further utilized cluster and anchor 98.66% of the genome sequences into 11 chromosomes. Approximately, 165.90 Mb, representing about 41.61% of the genome, was identified as repeat elements. Non-coding sequence annotation identified 171 rRNAs, 117 small RNAs, 331 regulatory RNAs, and 872 tRNAs. Genome annotation reveals 11,727 protein-coding genes, with 10,842 (92.45%) genes functionally annotated. In summary, our chromosome-level genome assembly serves as a significant resource for advancing research on Encyrtidae parasitoids.**

Background & Summary

Aenasius arizonensis (Girault, 1915) (Hymenoptera: Encyrtidae) is an obligate endoparasitoid that affects cotton mealybug *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae), harmful invasive pests feeding on over 200 plant species, including field crops and horticultural plants[1]. Taxonomically, *A. arizonensis* is recognized as a junior synonym with *Aenasius bambawalei* (Hayat, 2009), a species presumably native to North America and recently recorded in India[2], Pakistan[3,4], China[5], Iran[6], Turkey[7], Israel[8], and Australia[9]. The parasitoid wasp *A. arizonensis* significantly suppresses the *P. solenopsis* population by parasitizing 2nd and 3rd instar nymphs and female adults under laboratory and field conditions[10]. The parasitization efficiency reached 78 80% in Turkey[11] and even up to 90–95% in India[12] and Pakistan[13]. The parasitoid wasp kills *P. solenopsis* directly and drastically lowers the species' fertility, fecundity, and survival rate[14]. Despite its ecological importance, key aspects of *A. arizonensis* biology remain poorly understood. These include its biological traits[10,11,15,16], chemical ecology[17,18], venom function[19,20], ant-mealybug-parasitoids interactions[21,22], and their responses to temperature[23,24] and insecticide stresses[25,26]. To date, genetic research on this wasp has been limited to transcriptomic studies and mitochondrial genome analyses[27–30]. Access to a high-quality, assembled, as well as annotated genomic assembly for *A. arizonensis* offers a critical foundation for exploring diverse biological processes, including host localization, venomics, gender characterization, and genetic evolution.

Herein, we produce a chromosomal assembly for *A. arizonensis* by integrating PacBio long-read, MGISEQ short-read, and high-throughput chromosome conformation capture (Hi-C) approaches. A genome of 398.69 Mb was produced with contig and scaffold N50 of 4.73 Mb and 35.96 Mb, respectively. Hi-C data underwent clustering and anchoring into 11 chromosomes. Repeat elements constitute a significant portion of the

[1]State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-Products, Key Laboratory of Biotechnology in Plant Protection of MOA of China and Zhejiang Province, Institute of Plant Protection and Microbiology, Zhejiang Academy of Agricultural Sciences, Hangzhou, 310021, China. [2]State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, 100193, China. [3]Institute of Bio-Interaction, Xianghu Laboratory, Hangzhou, 311258, China. [4]Zhejiang Institute of Landscape Plants and Flowers, Zhejiang Xiaoshan Institute of Cotton & Bast Fiber Crops, Hangzhou, 311251, China. [5]College of Advanced Agricultural Sciences, Zhejiang A&F University, Hangzhou, 311300, China. [6]National Nanfan Research Institute, Chinese Academy of Agricultural Sciences, Sanya, 572019, China. [7]These authors contributed equally: Wan-Ying Dong, Tian-Yu Huang. ✉e-mail: zhouzhongshi@caas.cn; luybcn@163.com

Sequencing strategy	Platform	Usage	Reads number	Raw data (Gb)	Clean data (Gb)
Short-reads	MGISEQ- T7	Genome survey	282,502,066	42.38	39.15
Long-reads	PacBio Revio	Genome assembly	784,970	—	16.48
Hi-C	MGISEQ- T7	Chromosome-level assembly	450,525,948	67.58	67.54
RNA-seq	MGISEQ-T7	Gene structure annotation	86,851,686	13.03	13.00

Table 1. Sequencing and methodologies employed to assemble the *Aenasius arizonensis* genome.

genome, accounting for 165.90 Mb (~41.61%) of the total assembly. Non-coding sequence annotation identified 171 rRNAs, 117 small RNAs, 331 regulatory RNAs, and 872 tRNAs. To functionally characterize the genome, we performed structural and functional annotation using transcriptome data from female *A. arizonensis*. A total of 11,727 protein-coding genes were identified, of which 92.45% were successfully annotated. As the first chromosome-level genome assembly within the genus *Aenasius*, this high-quality reference genome offers a valuable to advance our understanding of the biocontrol capabilities of *A. arizonensis*. Furthermore, it is a critical foundation for following the investigation of the genetics, evolution, and host-parasitoid interactions within Encyrtidae parasitoids.

Methods

Sample collection and rearing. *Aenasius arizonensis* specimens analyzed in this research were collected from *Hibiscus mutabilis* plants located in the suburban areas of Hangzhou, Zhejiang, China. These parasitoids were continuously cultured over 30 generations, using their natural host, the mealybug *P. solenopsis*, under standardized environmental conditions. The conditions for raising were kept at 27 ± 1 °C, 70% ± 5% relative humidity and a light-dark sequence of 14: 10 hours. The parasitoids underwent rearing in nylon net cages (50 × 50 × 60 cm). Sprouted potato tubers were also provided in the cage to feed mealybugs.

Library construction and sequencing. For genomic DNA extraction, twenty newly emerged female adults from the laboratory population were surface-sterilized and processed using the QIAGEN Genomic-Tip (Qiagen, Germany). The extracted DNA was purified using a Grandomics Genomic kit (GrandOmics, China), following standardized protocols provided by the manufacturers for routine sequencing applications. Total RNA was isolated from an additional twenty newly emerged females of the laboratory population with TRIzol (Invitrogen, USA), adhering to the manufacturer's directions. Following extraction, both DNA and RNA were assessed through multiple methods: 1% agarose gels were used to check for integrity, a NanoDrop^{TM} One UV-Vis spectrophotometer (Thermo Fisher, USA) was employed for measurement, and a Qubit® 4.0 Fluorometer (Invitrogen, USA) determined the concentration.

Genomic DNA was fragmented randomly for short-read sequencing, and libraries with insert sizes 200–500 bp were created employing an Agencourt AMPure XP-Medium Kit (Beckman, USA). These libraries underwent sequencing using the MGISEQ-T7 platform at GrandOmics (Wuhan, China), generating 42.38 Gb of raw data for genome survey analysis. The sequencing data exhibited high quality, with Q20 and Q30 rates of 99.21% and 96.91%, respectively, with an average coverage depth of 74.64 ×. After quality filtering with fastp v0.23.4[31], 39.15 Gb of clean data were retained, of which 37.94 Gb (96.91%) exceeding the Q30 quality threshold (Table 1).

An SMRTbell library was generated with an SMRTbell® Prep Kit 3.0 (Pacific Biosciences, USA) for PacBio HiFi long-reads. This library underwent sequencing using a PacBio Revio with the Revio Polymerase Kit. This generated 16.48 Gb of high-quality HiFi reads, providing 28.88 × coverage for genome assembly. The clean data exhibited an N50 length of 21,531 bp, a maximum read length of 61,351 bp, and an average length of 20,999.70 bp (Table 1).

To perform Hi-C sequencing, muscle tissue underwent exposure to 2% formaldehyde to stabilize DNA-protein interactions through crosslink. Chromatin was incubated with DpnII. The Hi-C samples underwent a series of steps, including biotin labeling, blunt-end ligation, and DNA purification. The final Hi-C library was sequenced on the MGISEQ-T7 platform, generating paired-end 150-bp reads to map spatial interactions within the chromosomes. A total of 67.58 Gb of Hi-C read data was generated, with 67.54 Gb retained after filtering using fastp v0.23.4[31]. Of these, 65.23 Gb (96.58%) exceeded the Q30 quality threshold (Table 1).

RNA sequencing (RNA-Seq) libraries were constructed utilizing a NEBNext® Ultra^{TM} RNA Library Prep Kit (NEB, USA) and sequenced using an MGISEQ-T7. RNA-Seq data (13.03 Gb) was generated, which was then utilized in whole-genome protein-coding gene forecasting.

Genome survey. Before assembly, the genome size as well as heterozygosity were projected using k-mer on MGI paired-ended raw reads. In summary, 42.38 Gb raw data underwent quality filtering with fastp v0.21.0[31] (parameters: -n 0 -f 5 -F 5 -t 5 -T 5 -q 20), resulting in 39.15 Gb of clean data (Table 1). Filtered reads underwent processing using KMC v3.2.1[32] (parameters: -k21 -ci1 -cs1000000) to generate k-21 mers frequency distribution and assess heterozygosity. Genome size was performed using the FindGSE program[33] and GenomeScope v1.0.0[34] under default configurations. The analysis revealed a genome size of 428.82 Mb and heterozygosity of 0.60%, determined from the k-mer depth distribution (Fig. 1).

Genome assembly. Contig assembly from PacBio HiFi reads was performed using hifiasm v0.19[35] (default parameters). To achieve high accuracy, the initial draft genome was further polished with NextPolish v1.2.4[36], incorporating short-read sequencing data generated from the MGISEQ-T7 platform. The assembled genome was aligned with all Illumina paired-end evaluation with the Burrows-Wheeler Aligner (BWA) v0.7.12-r1039[37], while

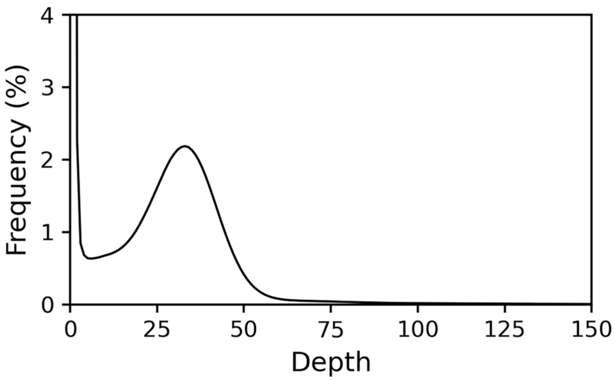

Fig. 1 The analysis of the 21-mer sequences within the *Aenasius arizonensis* genome. The X-axis illustrates k-mer depth, while the Y-axis denotes k-mer frequency at a given depth.

Genome features	Statistics
Draft genome size (bp)	406,665,476
Contig number	225
Contigs N50 size (bp)	4,728,094
Scaffold number	25
Scaffold N50 size (bp)	35,959,675
% of sequences anchored to chromosomes	98.66
Number of chromosomes	11
Total length of chromosomes (bp)	398,693,586
GC content (%)	35.42

Table 2. Summary of the assembled *Aenasius arizonensis* genome.

base accuracy was evaluated using SAMtools v1.4[38] and Bcftools v1.8.0[39]. The sequencing reads were assessed for alignment rate and genome coverage utilizing Minimap2 vr41[40] (parameters: -x map-hifi). A thorough contaminant screening was conducted to ensure the integrity and purity of the draft genome assembly, utilizing blast v2.9[41] to align the assembly against the NT Database. This process identified and eliminated 7 contaminant contigs. Additionally, similarity searches were conducted using Purge_Dups[42] (parameters: -f .9) to identify and discard redundant contigs, resulting in a final assembly. The initial assembly comprises 225 contigs, totaling 406.67 Mb in length, and features a contig N50 of 4.73 Mb (see Table 2).

Hi-C scaffolding. Raw Hi-C results underwent data analysis via Hi-C-Pro v2.8.1[43] and quality inspection procedures utilizing fastp v0.21.6[31]. Clean reads underwent alignment to the draft genome assembly with bowtie2 v2.3.2[44] (parameter: -end-to-end, -very-sensitive -L 30). Subsequently, uniquely mapped paired-end reads were processed through Hi-C-Pro v2.8.1 to filter out incorrect pairs, including self-cycle, dangling ends, re-ligations, and dumped sequences. A total of 99,670,120 valid interaction pairs were retained for scaffold correction. These pairs were utilized to cluster, order, and orient contigs onto chromosomes using LACHESIS[45] (parameters: CLUSTER_MIN_RE_SITES=100, CLUSTER_MAX_LINK_DENSITY=2.5, CLUSTER NONINFORMATIVE RATIO=1.4, ORDER MIN N RES IN TRUNK=60, ORDER MIN N RES IN SHREDS=60).

The total genome length measured 398,693,586 bp, with 393,330,049 bp (98.66%) successfully anchored to 11 chromosomes, ranging 22.97–45.46 Mb (Tables 2, 3). A genome-wide chromatin interaction Hi-C heatmap, generated using Python, revealing strong diagonal interaction signals, further validating chromosome-level assembly quality (Fig. 2). Additionally, the chromosomal landscape was visualized using the Advanced Circos tool integrated into TBtools[46] (Fig. 3).

Genomic repeat and non-coding RNA annotation. Repeats identified in the assembled genome were separated into two categories: tandem repeats (TRs) and transposable elements (TEs). The detection of TRs was performed with GMATA v2.2[47] (default settings) and Tandem Repeats Finder v4.07b[48] (parameters: 2 7 7 80 10 50 500 -f -d -h -r). TE were annotated utilizing a hybrid methodology that integrates homology based and *de novo* techniques. The *de novo* repeat approach used LTR-retriever[49] (default parameters) and RepeatModeler version open-1.0.11 (parameters: engine wublast). Predicted repeats were classified via TEclass v2.1.3[50], and merged with Repbase[51] entries to compile a species-specific TE library, retaining redundant sequences. RepeatMasker v1.331[52] was then employed to identify TE sequences through homology searches against the library. Collectively, 4.41 Mb of TRs and 154.60 Mb of TEs were annotated, constituting 41.61% of the genome assembly (Table 4).

Non-coding RNA (ncRNA) labeling was accomplished by mapping genomic sequences to the RFAM database[53] (http://rfam.xfam.org/) with Infernal v1.1.2[54] (default parameters). Transfer RNA (tRNA) was detected with tRNAscan-SE v2.0[55] (parameters:--thread 4 -E -I). Ribosome RNA (rRNA) and subunits were annotated with

Chromosome	Chromosome size (bp)	Contig number
1	45,460,858	33
2	43,488,874	27
3	42,485,375	23
4	38,061,147	23
5	35,958,775	10
6	35,786,642	11
7	34,695,690	22
8	33,884,628	22
9	31,547,313	17
10	28,993,959	15
11	22,966,788	9

Table 3. Overview of eleven assembled *Aenasius arizonensis* chromosomes.

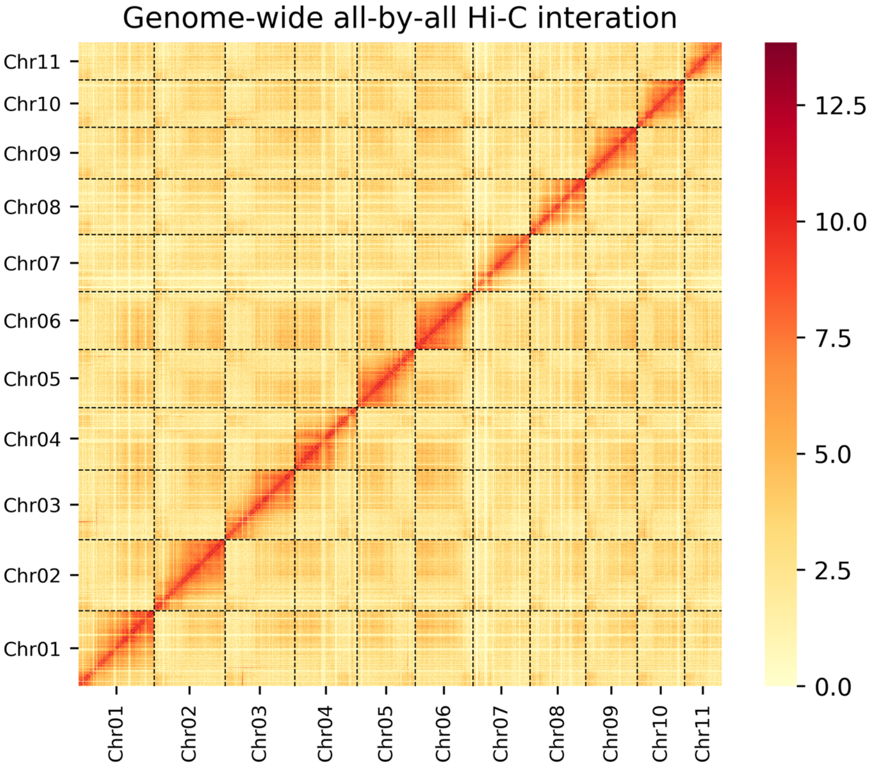

Fig. 2 Hi-C assembly heatmap of *Aenasius arizonensis*. The X- and Y-axes indicate the sequential order of bins along their respective chromosome groups. The interaction intensity of Hi-C associations is depicted using a color gradient, illustrated on the left, from yellow, representing low-intensity to red, representing high-intensity.

RNAmmer v1.2[56], which was configured with the settings "-S euk -m lsu, ssu, tus -gff". Based on Rfam databases, the *A. arizonensis* genome contains 171 rRNAs, 117 small RNAs, 331 regulatory RNAs, and 872 tRNAs (Table 5).

Gene modeling and functional predictions. Following repeat sequence masking, intact protein-coding gene models were predicted through an integrative pipeline combining three independent approaches: homology-based prediction, transcriptome-based prediction, and *de novo* prediction. For homology-based prediction, protein sets from five insects species including *Copidosoma floridanum*[57], *Eretmocerus hayati*, *Nasonia vitripennis*, *Eurytoma adleriae*, and *Ormyrus pomaceus* were retrieved from Insectbase 2.0[58] and aligned with GeMoMa v1.6.1[59] (default parameters). In order to make a prediction based on the transcriptome, high-quality RNA-seq data were aligned to the *A. arizonensis* genome assembly via STAR v2.7.3a[60]. This was then followed by the assembly of transcripts with Stringtie v1.3.4d[61]. Open reading frames (ORFs) underwent characterization utilizing PASA v2.3.3[62] to generate a training dataset. *De novo* gene models were constructed using Augustus v3.3.1[63] and GlimmerHMM v3.0.4[64]. Gene models from these methodologies were combined into an integrated set utilizing EVidenceModeler (EVM) v1.1.1 (default parameters), which was further retained with PASA v2.3.3 to annotate untranslated regions. The genome annotation revealed 11,727 genes that code for proteins (Table 6); these genes possessed a mean length of 17,936.89 bp and a mean length of 1,716.43 bp for their

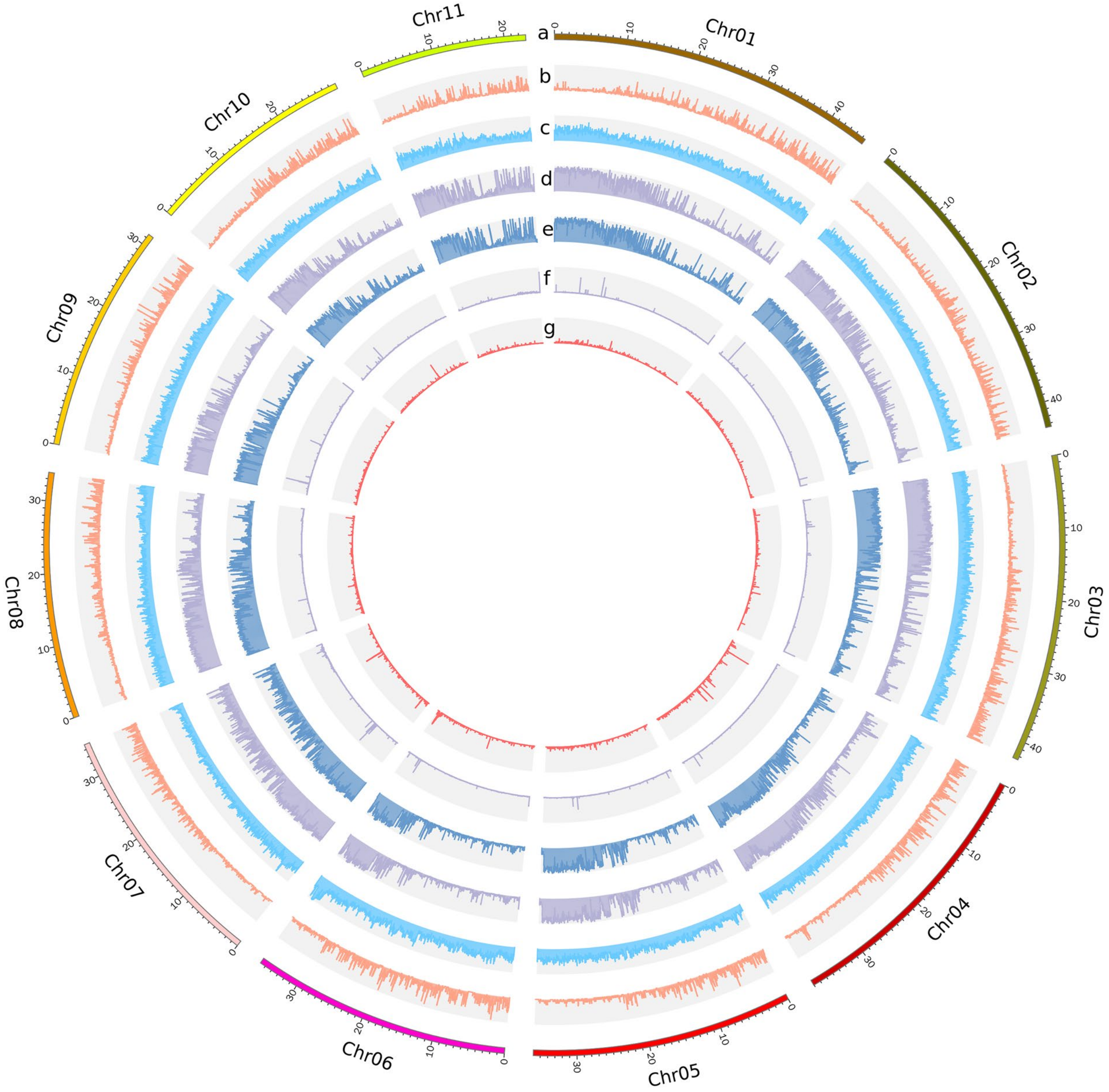

Fig. 3 Circos plot illustrating genome characteristics of *Aenasius arizonensis*, with all data represented in 50-kb genomic windows. (a) Chromosome the length (Mb); (b) gene density in each Mb (0–15) (c) GC abundance in each Mb (0%–100%); (d) repeat elements abundance (0%–100%); (e) transposable elements abundance (0%–100%); (f) tandem repeats abundance (0%–88.73%); (g) non-coding RNA abundance (0–19).

coding sequence (CDS). On average, 6.92 exons were found in each gene through structural analysis, with exon and intron lengths averaging 248.08 bp and 2,740.41 bp, respectively (Table 7).

Predicted protein-coding genes underwent functional characterization through alignment to five major databases: Kyoto Encyclopedia of Gene and Genomes (KEGG)[65], Eukaryotic Orthologous Groups of protein (KOG)[66], the National Center for Biotechnology Information (NCBI) non-redundant database (NR), Gene Ontology (GO)[67], and SwissProt[68]. Predicted functional domains and GO identities were defined using InterProScan[69] program (default parameters). Protein sequences integrated by EVM were compared against the mentioned databases using BLASTp (https://blast.ncbi.nlm.nih.gov/Blast.cgi), with a threshold E-value of 1e-5[62]. Consensus annotations from all five databases were integrated using EVM v1.1.1[62], resulting in successful functional annotation of 10,842 genes (92.45% of predicted genes) (Table 8).

Data Records

PacBio long-read sequences, Hi-C data, MGI short-read sequencing, and transcriptomic sequences can be accessed in NCBI Sequence Read Archive database (accession number PRJNA1178347)[70]. Specifically, genomic MGI sequencing data, PacBio sequel II long-read data, Hi-C sequencing results, and transcriptome sequences can be found in NCBI (accession number SRP541658)[71]. The Genbank accession number for the Whole Genome Shotgun project is JBISGU000000000[72]. Additionally, genome annotations can be found using Figshare: https://doi.org/10.6084/m9.figshare.27933360[73].

Type			Number of elements	Length of sequence (bp)	Percentage of sequence (%)
TEs	Class I: Retroelement	LINE	74,638	25,834,248	6.48
		LTR	83,767	37,258,922	9.35
		SINE	5,891	651,582	0.16
		Total	164,296	63,744,752	15.99
	Class II: DNA transposon	DNA	357,536	83,191,135	20.87
		RC	8,822	1,610,915	0.40
		MITE	20,666	6,048,250	1.52
		Total	387,024	90,850,30	22.79
	Total TEs		551,320	154,595,052	38.78
Tandem Repeats	SSR		79,905	999,375	0.25
	Tandem repeat elements		37,592	3,407,250	0.85
	Total		117,497	4,406,625	1.11
Simple repeats			1,914	200,178	0.05
Other			2,348	319,792	0.08
Unknown			28,907	6,367,098	1.60
Low complexity			91	11,573	0.00
Total repeats			702,077	165,900,318	41.61

Table 4. Overview of repetitive sequences within the *Aenasius arizonensis* genome.

Class	Type	Copy number	Average length (bp)	Total length (bp)	Percentage of sequence (%)
rRNA (171)	18 S	45	1,934.73	87,063	0.0218
	28 S	25	4,289.52	107,238	0.0269
	5.8 S	49	155.00	7,595	0.0019
	5 S	45	115.09	5,179	0.0013
snRNA (117)	snRNA	19	102.00	1,938	0.0005
	miRNA	43	82.53	3,549	0.0009
	Spliceosomal	45	158.16	7,117	0.0018
	Other	10	213.30	2,133	0.0005
Regulatory	cis-regulatory elements	331	47.01	15,560	0.0039
tRNA	tRNA	872	75.78	66,077	0.0166

Table 5. Overview of non-coding RNAs within the *Aenasius arizonensis* genome.

	Gene set	Total number of genes	Average gene length (bp)	Average CDS length (bp)	Average exons number per gene	Average exon length (bp)	Average intron length (bp)
Homoloy	*C. floridanum*	20,784	17,848.14	1,448.45	5.12	282.7	3,977.07
	E. hayati	45,650	17,463.63	1,266.15	3.66	346.08	6,092.69
	N. vitripennis	26,058	18,498.17	1,446.22	5.07	285.02	4,185.49
	E. adleriae	33,952	13,395.79	1,191.98	4.13	288.4	3,895.2
	O. pomaceus	33,602	12,987.71	1,206.77	3.93	307.0	4,019.69
	GeMoMa	33,390	11,282.02	1,076.59	3.38	318.19	4,281.75
Transcriptome	NGS RNA seq	16,291	23,141.57	3,543.85	7.89	449.42	2,846.25
	PASA	15,792	22,419.65	3,549.13	7.88	450.45	2,743.21
De novo	AUGUSTUS	12,881	18,318.39	1,679.79	6.93	242.36	2,805.41
	GlimmerHMM	28,064	12,719.03	823.41	4.5	182.88	3,396.29
Final	EVM	11,727	17,936.89	1,716.43	6.92	248.08	2,740.41

Table 6. Gene annotation results within the *Aenasius arizonensis* genome, generated through three different strategies.

Technical Validation

To ensure the quality of the genome assembly, three approaches were used to evaluate its accuracy and completeness. First, MGI short-read were aligned to the genome with BWA v0.7.12-r1039[37], achieving a 99.92% alignment rate and of 99.76% genome coverage. The genome exhibited heterozygous (0.002074%) and homozygous (0.000708%) nucleotide polymorphisms (SNPs), respectively, demonstrating elevated accuracy. Second, completeness and accuracy of core genes in the integrated genome were evaluated using the Core Eukaryotic Genes Mapping Approach (CEGMA, v2)[74]. Of 248 core eukaryotic genes (CEGs) identified in

Species	Total number of genes	Average transcript length (bp)	Average CDS length (bp)	Average exons number per gene	Average exon length (bp)	Average intron length (bp)
A.arizonensis	11,727	17,936.89	1,716.43	6.92	248.08	2,740.41
C. floridanum	11,908	17,376.98	1,591.67	6.12	260.0	3,081.95
E. hayati	23,927	4,843.5	1,262.94	4.11	307.52	1,152.5
N. vitripennis	13,571	11,806.81	1,617.06	6.07	266.19	2,007.91
E. adleriae	26,747	2,390.22	1,082.89	3.69	293.34	485.72
O. pomaceus	20,984	2,670.97	1,161.47	4.24	273.97	465.99

Table 7. Comparison analysis of protein-coding genes annotations between the *Aenasius arizonensis* genome and other parasitoid species.

Type		Number	Percent (%)
Annotation	NR	10,776	91.89
	KEGG	6,786	57.87
	KOG	7,622	65.00
	GO	7,059	60.19
	Swissprot	8,800	75.04
Total	Annotated	10,842	92.45
	Gene	11,727	—

Table 8. Gene functional annotation of *Aenasius arizonensis*.

Data	Complete gene (%)	Single-copied gene (%)	Duplicated gene (%)	Fragmented gene (%)	Missing gene (%)
Contig level assembly	96.89	92.75	4.14	0.42	2.68
Chromosome level assembly	97.07	95.68	1.39	0.44	2.49
All protein-coding gene	94.73	93.08	1.65	0.47	4.80
Functionally annotated protein-coding gene	94.73	93.08	1.65	0.47	4.80

Table 9. Assessment of genome assembly and protein-coding gene completeness using BUSCO analysis.

CEGMA, 246 (99.19%) were successfully assembled, with 239 CEGs (96.37%) being complete. According to these results, the genome assembly contains a nearly full complement of core genes. Third, genome completeness was quantified using Benchmarking Universal Single Copy Orthologs (BUSCO v4.0.5)[75] (parameters: -l endopterygota_odb10 -g genome). At the contig level, the completeness of single-copy genes reached 96.89%, with 92.75% being single-copy and 4.14% duplicated. At the chromosome level, the completeness of single-copy genes was 97.07%, consisting of 95.68% single-copy and 1.39% duplicated genes. Functional annotation further corroborated these results, showing that 94.73% of protein-coding genes were classified as complete, with 93.08% being single-copy and 1.65% duplicated (Table 9). Collectively, these analyses demonstrate increased accuracy and completeness.

Code availability

No custom scripts or code were used in this study.

Received: 9 December 2024; Accepted: 14 April 2025;
Published online: 17 May 2025

References

1. Fand, B. B. & Suroshe, S. S. The invasive mealybug *Phenacoccus solenopsis* Tinsley, a threat to tropical and subtropical agricultural and horticultural production systems-A review. *Crop Prot.* **69**, 34–43, https://doi.org/10.1016/j.cropro.2014.12.001 (2015).
2. Hayat, M. Description of a new species of *Aenasius* Walker (Hymenoptera: Encyrtidae), parasitoid of the mealybug, *Phenacoccus solenopsis* Tinsley (Homoptera: Pseudococcidae) in India. *Biosystematica* **3**, 21–26 (2009).
3. Ashfaq, M., Shah, G. S., Noor, A. R., Ansari, S. P. & Mansoor, S. Report of a parasitic wasp (Hymenoptera: Encyrtidae) parasitizing cotton mealybug (Hemiptera: Pseudococcidae) in Pakistan and use of PCR for estimating parasitism levels. *Biocontrol Sci. Techn.* **20**, 625–630, https://doi.org/10.1080/09583151003693535 (2010).
4. Bodlah, I., Ahmad, M., Nasir, M. F. & Naeem, M. Record of *Aenasius bambawalei* Hayat, 2009 (Hymenoptera: Encyrtidae), a parasitoid of *Phenacoccus solenopsis* (Sternorrhyncha: Pseudococcidae) from Punjab, Pakistan. *Pak. J. Zool.* **42**, 533–536 (2010).
5. Chen, H. Y., Cao, R. X. & Xu, Z. F. First record of *Aenasius bambawalei* Hayat (Hymenoptera: Encyrtidae) a parasitoid of the mealybug, *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae) from China. *J. Environ. Entomol.* **32**, 280–282, https://doi.org/10.3969/j.issn.1674-0858.2010.02.022 (2010).
6. Fallahzadeh, M., Japoshvili, G., Abdimaleki, R. & Saghaei, N. New records of Tetracneminae (Hymenoptera, Chalcidoidea, Encyrtidae) from Iran. *Turk. J. Zool.* **38**, 515–518, https://doi.org/10.3906/zoo-1309-28 (2014).

7. Çalişkan Keçe, A. F., Kahya, D., Hayat, M. & Ulusoy, M. R. A new record of a parasitoid (Hymenoptera: Encyrtidae) of an invasive mealybug *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae) from Turkey. *Türk. Biyo. Mücadele Derg.* **9**, 31–37, https://doi.org/10.31019/tbmd.436856 (2018).
8. Spodek, M. *et al.* The cotton mealybug, *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae) in Israel: pest status, host plants and natural enemies. *Phytoparasitica* **46**, 45–55, https://doi.org/10.1007/s12600-018-0642-1 (2018).
9. Spargo, G., Khan, M. & Byers, K. A parasitoid of solenopsis mealybug found at Emerald. *Aust Cottongrower.* **34**, 22–23 (2013).
10. Shera, P. S. & Karmakar, P. Effect of mating combinations on the host parasitisation and sex allocation in solitary endoparasitoid, *Aenasius arizonensis* (Hymenoptera: Encyrtidae). *Biocontrol Sci. Techn.* **28**, 49–61, https://doi.org/10.1080/09583157.2017.1413707 (2018).
11. Kahya, D., Çalişkan Keçe, A. F. & Ulusoy, M. R. Determining some biological parameters of *Aenasius arizonensis* (Girault) (Hymenoptera: Encyrtidae) on cotton mealybug and the rate of parasitism in field conditions. *Turk. J. Zool.* **45**, 416–424, https://doi.org/10.3906/zoo-2102-2 (2021).
12. Tanwar, R. K., Jeyakumar, P., Singh, A., Jafri, A. A. & Bambawale, O. M. Survey for cotton mealybug *Phenacoccus solenopsis* (Tinsley) and its natural enemies. *J. Environ. Biol.* **32**, 381–384 (2011).
13. Khuhro, S. N., Kalroo, A. M. & Mahmood, R. Present status of mealybug *Phenacoccus solenopsis* (Tinsley) on cotton and other plants in Sindh (Pakistan). World Cotton Research Conference-5; 2011 2011-11-07; Mumbai: Excel India, New Delhi; 2011. 268–271.
14. Shahzad, M. Q., Abdin, Z. U., Abbas, S. K., Tahir, M. & Hussain, F. Parasitic effects of solitary endoparasitoid, *Aenasius bambawalei* Hayat (Hymenoptera: Encyrtidae) on cotton mealybug, *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae). *Adv. Entomol.* **4**, 90–96, https://doi.org/10.4236/ae.2016.42010 (2016).
15. Li, P., Wu, T. D., Ren, Y. J., Xu, Z. F. & Zhou, Z. S. Developmental fitness of *Aenasius bambawalei* (Encyrtidae) reared on *Phenacoccus solenopsis* reared from different species of host plants. *Biocontrol Sci. Techn.* **30**, 1162–1167, https://doi.org/10.1080/09583157.2020.1776840 (2020).
16. Huang, J., Zhi, F., Zhang, J., Li, X. & Lu, Y. The parasitoid *Aenasius arizonensis* prefers its natural host but can parasitize a nonnatural host in the laboratory: an opportunity for control of a new invasive mealybug. *J. Pest Sci.* **95**, 593–604, https://doi.org/10.1007/s10340-021-01406-y (2022).
17. Li, Q. L. *et al.* Optimization of reverse chemical ecology method: false positive binding of *Aenasius bambawalei* odorant binding protein 1 caused by uncertain binding mechanism. *Insect Mol. Biol.* **27**, 305–318, https://doi.org/10.1111/imb.12372 (2018).
18. Xu, C. *et al.* Discovery of behaviorally active semiochemicals in *Aenasius bambawalei* using a reverse chemical ecology approach. *Pest Manag. Sci.* **77**, 2843–2853, https://doi.org/10.1002/ps.6319 (2021).
19. Abbas, S. K., Abdin, Z., Arif, M. J. & Jamil, A. Functional analysis of the venom of mealybug parasitoid *Aenasius bambawalei* (Hymenoptera: Encyrtidae). *Biologia* **69**, 1046–1050, https://doi.org/10.2478/s11756-014-0402-8 (2014).
20. Abbas, S. K., Abdin, Z. U., Arshad, M., Hussain, F. & Jamil, A. *In vitro* studies for the evaluation of insecticidal potential of the venom of endoparasitic wasp *Aenasius arizonensis* (Girault) (Hymenoptera, Encyrtidae). *Int. J. Pept. Res. Ther.* **27**, 47–54, https://doi.org/10.1007/s10989-020-10062-2 (2020).
21. Huang, J., Zhang, P. J., Zhang, J. & Tang, Y. Y. An ant-coccid mutualism affects the behavior of the parasitoid *Aenasius bambawalei*, but not that of the ghost ant *Tetramorium bicarinatum*. *Sci. Rep.* **7**, 5175, https://doi.org/10.1038/s41598-017-05442-6 (2017).
22. Xu, C., Li, Q. L., Qu, X. B., Chen, J. & Zhou, A. M. Ant-hemipteran association decreases parasitism of *Phenacoccus solenopsis* by endoparasitoid *Aenasius bambawalei*. *Ecol. Entomol.* **45**, 290–299, https://doi.org/10.1111/een.12797 (2019).
23. Zhang, J., Tang, Y. & Huang, J. The effects of temperature on the development, morphology, and fecundity of *Aenasius bambawalei* (=*Aenasius arizonensis*). *Insects* **12**, 833, https://doi.org/10.3390/insects12090833 (2021).
24. Thimmegowda, M. N., Sachin, S. S. & Sagar, D. Thermal tolerance mechanism of invasive cotton mealybug parasitoid, *Aenasius arizonensis* Girault (Hemiptera: Encyrtidae). *J. Asia-Pac. Entomol.* **27**, 102178, https://doi.org/10.1016/j.aspen.2023.102178 (2024).
25. Karmakar, P. & Shera, P. S. Lethal and sublethal effects of insecticides used in cotton crop on the mealybug endoparasitoid *Aenasius arizonensis*. *Int. J. Pest Manage.* **66**, 13–22, https://doi.org/10.1080/09670874.2018.1538544 (2018).
26. Shankarganesh, K., Ricupero, M. & Sabtharishi, S. Field evolved insecticide resistance in the cotton mealybug *Phenacoccus solenopsis* and its direct and indirect impacts on the endoparasitoid *Aenasius arizonensis*. *Sci. Rep.* **12**, 16764, https://doi.org/10.1038/s41598-022-20779-3 (2022).
27. Shaina, H., Abdin, Z. U., Webb, B. A., Arif, M. J. & Jamil, A. *De novo* sequencing and transcriptome analysis of venom glands of endoparasitoid *Aenasius arizonensis* (Girault) (=*Aenasius bambawalei* Hayat) (Hymenoptera, Encyrtidae). *Toxicon* **121**, 134–144, https://doi.org/10.1016/j.toxicon.2016.08.022 (2016).
28. Nie, X. P. *et al.* Antennal transcriptome and odorant binding protein expression profiles of an invasive mealybug and its parasitoid. *J. Appl. Entomol.* **142**, 149–161, https://doi.org/10.1111/jen.12417 (2017).
29. Ma, Y., Zheng, B., Zhu, J., Tang, P. & Chen, X. The mitochondrial genome of *Aenasius arizonensis* (Hymenoptera: Encyrtidae) with novel gene order. *Mitochondrial DNA Part B* **4**, 2023–2024, https://doi.org/10.1080/23802359.2019.1617052 (2019).
30. Zhang, J., Huang, J., Tang, Y. & Long, X. Transcriptome profile analysis of the accompanying migratory parasitic wasp *Aenasius bambawalei* (=*Aenasius arizonensis* girault) (Hymenoptera: Encyrtidae): Genes related to fertilization involved at different stage of ovary development. *Biocell* **46**, 195–205, https://doi.org/10.32604/biocell.2022.016563 (2022).
31. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
32. Deorowicz, S., Kokot, M., Grabowski, S. & Debudaj-Grabysz, A. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* **31**, 1569–1576, https://doi.org/10.1093/bioinformatics/btv022 (2015).
33. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. *findGSE*: estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies. *Bioinformatics* **34**, 550–557, https://doi.org/10.1093/bioinformatics/btx637 (2018).
34. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
35. 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, https://doi.org/10.1038/s41592-020-01056-5 (2021).
36. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
37. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595, https://doi.org/10.1093/bioinformatics/btp698 (2010).
38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
39. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039, https://doi.org/10.1093/bioinformatics/btx100 (2017).
40. Li, H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110, https://doi.org/10.1093/bioinformatics/btw152 (2016).
41. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238, https://doi.org/10.1186/1471-2105-13-238 (2012).
42. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).

43. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259, https://doi.org/10.1186/s13059-015-0831-x (2015).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
45. Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125, https://doi.org/10.1038/nbt.2727 (2013).
46. Chen, C. *et al.* TBtools: an integrative toolkit developed for interactive analysis of big biological data. *Mol. Plant* **13**, 1194–1202, https://doi.org/10.1016/j.molp.2020.06.009 (2020).
47. Wang, X. & Wang, L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* **7**, 1350, https://doi.org/10.3389/fpls.2016.01350 (2016).
48. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
49. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
50. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass-a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330, https://doi.org/10.1093/bioinformatics/btp084 (2009).
51. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
52. Bedell, J. A., Korf, I. & Gish, W. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16**, 1040–1041, https://doi.org/10.1093/bioinformatics/16.11.1040 (2000).
53. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124, https://doi.org/10.1093/nar/gki081 (2005).
54. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
55. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096, https://doi.org/10.1093/nar/gkab688 (2021).
56. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108, https://doi.org/10.1093/nar/gkm160 (2007).
57. Toga, K. *et al.* Long-read genome assembly of the Japanese parasitic wasp *Copidosoma floridanum* (Hymenoptera: Encyrtidae). *G3-Genes Genom. Genet.* **14**, e127, https://doi.org/10.1093/g3journal/jkae127 (2024).
58. Mei, Y. *et al.* InsectBase 2.0: a comprehensive gene resource for insects. *Nucleic Acids Res.* **50**, D1040–D1045, https://doi.org/10.1093/nar/gkab1090 (2022).
59. Keilwagen, J. *et al.* Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89, https://doi.org/10.1093/nar/gkw092 (2016).
60. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21, https://doi.org/10.1093/bioinformatics/bts635 (2013).
61. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
62. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
63. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
64. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
65. Ogata, H. *et al.* KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34, https://doi.org/10.1093/nar/27.1.29 (1999).
66. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269, https://doi.org/10.1093/nar/gku1223 (2015).
67. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
68. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27**, 49–54, https://doi.org/10.1093/nar/27.1.49 (1999).
69. Zdobnov, E. M. & Apweiler, R. InterProScan-an integration platform for the signature-recognition methods in *InterPro*. *Bioinformatics* **17**, 847–848, https://doi.org/10.1093/bioinformatics/17.9.847 (2001).
70. *NCBI Bioproject* http://identifiers.org/ncbi/bioproject:PRJNA1178347 (2024).
71. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP541658 (2024).
72. Dong, W. *Aenasius arizonensis* isolate DW-2024, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JBISGU000000000 (2024).
73. Huang, T. Genome annotation results. *figshare*. https://doi.org/10.6084/m9.figshare.27933360.v1 (2024).
74. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067, https://doi.org/10.1093/bioinformatics/btm071 (2007).
75. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* **1**, e323, https://doi.org/10.1002/cpz1.323 (2021).

Acknowledgements

This study was supported by the Zhejiang Provincial Natural Science Foundation of China (LQ23C140003) and the National Key Research and Development Program of China (2022YFC2601400), Postdoctoral Fellowship Program (Grade C) of China Postdoctoral Science Foundation (GZC20233066) and Fellowship from the China Postdoctoral Science Foundation (2023M743837).

Author contributions

Z.S.Z. and Y.B.L. conceived the study and directed the research. W.Y.D., J.Z. and Y.L. contribute to the materials for sequencing. W.Y.D. and T.Y.H. performed the experiments and analyzed the data. W.Y.D. drafted the manuscript. W.Y.D., T.Y.H., S.Y.Z. and J.H. revised the manuscript. All authors reviewed the final manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Z.-S.Z. or Y.-B.L.

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

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