Nucleic Acids Research, 2026, gkag1087
https://doi.org/10.1093/nar/gkag1087
Database issue

AnnelidaDB: a comprehensive platform of taxonomic and multi-omic resources across annelids

Yunlong Li[1], Haokun Liu[2], Maeva Perez[1], Yi-Xuan Li[1], Fan Hui[1], Zihui Shen[1], Yanan Sun[3], José M. Martín-Durán[4], Greg W. Rouse[5], Pei-Yuan Qian[6,7], Jianliang Xu[8], Xike Xie[2,*], Jian-Wen Qiu[1,*]

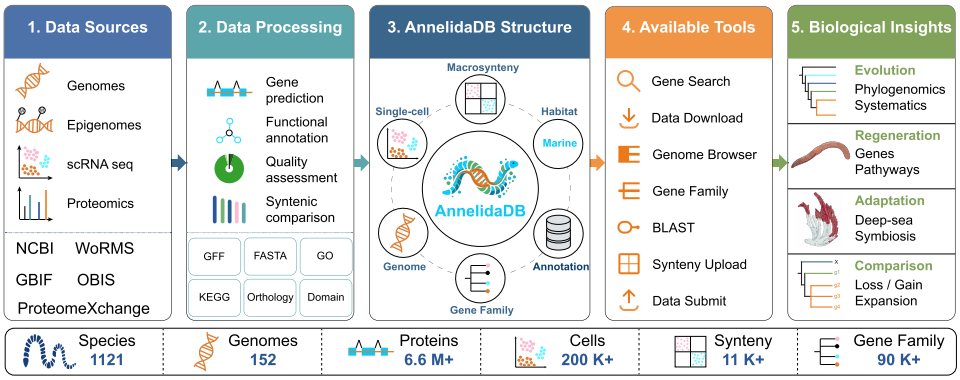
[1]Department of Biology, Hong Kong Baptist University, Hong Kong SAR, China
[2]Data Darkness Lab, School of Biomedical Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu 215000, China
[3]Laboratory of Marine Organism Taxonomy and Phylogeny, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China,
[4]School of Biological and Behavioural Sciences, Queen Mary University of London, London E1 4NS, United Kingdom
[5]Scripps Institution of Oceanography, University of California, San Diego 92093-0202, United States
[6]Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou, Guangdong 511458, China
[7]Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[8]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

*To whom correspondence should be addressed. Email: qiujw@hkbu.edu.hk
Correspondence may also be addressed to Xike Xie. Email: xkxie@ustc.edu.cn

Abstract

Annelids inhabit marine, freshwater, and terrestrial ecosystems and are important models for studying animal evolution, regeneration, and host–microbe interactions, yet their rapidly growing omics resources remain dispersed across repositories and publications. We developed AnnelidaDB (www.annelidadb.com), a species-centred database built on an updated phylogenetic framework that integrates genomes, mitochondrial genomes, transcriptomes, proteomes, epigenomes, single-cell datasets, host-associated symbiont genomes, and comparative synteny for Annelida. The current release covers 1121 species, including 152 species with nuclear genomes, 177 with mitochondrial genomes, 591 with transcriptome data, 5 with single-cell datasets comprising 209 802 cells, and 30 bacterial symbiont genomes. AnnelidaDB organizes these resources through 10 navigation directories and connects species, genome, gene, functional annotation, and gene-family records. Users can browse genomic assemblies, explore interactive single-cell atlases and 12 200 precomputed Oxford plots, run BLAST queries and search genes against the annelid-specific database, and submit chromosome-level data for syntenic analysis against 20 Annelida linkage groups. Sequences, annotations, and private results are available for download. As a comprehensive multi-omics resource dedicated to annelids, AnnelidaDB provides an accessible platform for comparative investigations of annelid genome evolution, adaptation, regeneration, and symbiosis.

Graphical abstract



Received: August 13, 2026. **Revised:** October 12, 2026. **Accepted:** October 13, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

Introduction

The phylum Annelida, comprising over 22 000 described species of segmented worms, represents one of the most diverse and ecologically significant lineages within the superphylum Lophotrochozoa (Spiralia) [1, 2]. Spanning marine, freshwater, and terrestrial ecosystems, annelids play indispensable roles as primary decomposers, critical links in global food webs, and reliable bioindicators of soil and aquatic health [3–5]. Beyond their ecological roles, annelids have emerged as powerful model organisms in evolutionary and developmental biology, offering insights into processes such as whole-body regeneration and habitat transitions [6–9]. The phylum also exhibits remarkable evolutionary plasticity, including extreme physiological adaptations [10, 11]. Notable examples are deep-sea hydrothermal vent tube worms (e.g. *Riftia pachyptila*), which have lost their digestive systems and rely entirely on obligate chemoautotrophic symbionts for energy and nutrition [12]. In addition, annelids display remarkable diversity in chromosome numbers and rearrangements, making them an appropriate system for studying karyotype evolution across deep time [9, 13].

Advances in high-throughput sequencing have driven a rapid expansion of annelid multi-omics data over the past decade [6, 14–18]. Despite this growth, genomic, transcriptomic, and single-cell datasets remain scattered across broad-scope public repositories, such as the National Center for Biotechnology Information (NCBI) [19], the European Nucleotide Archive (ENA) [20], Figshare, Dryad, and GitHub. The utility of these genomic assemblies is often hindered by missing gene feature files and inconsistent deposition across databases. As of 1 September 2026, only 10 of 236 genome assemblies hosted on NCBI and 3 of 251 assemblies in ENA had associated gene feature files (details listed on the About page, www.annelidadb.com/about). While other major animal phyla already benefit from well-curated, taxon-specific, and highly integrated genomics resources—such as MolluscDB for the phylum Mollusca [21, 22], Echinobase for the phylum Echinodermata [23], and Hymenoptera Genome Database for hymenopteran insect species [24]—a dedicated platform for annelids remains absent. Currently, only a limited number of resources have been integrated into databases not dedicated to annelids, such as some deep-sea annelids in the Deep Ocean Omics database (a multi-omics atlas for deep-ocean organisms, 14 deep-sea annelid assemblies, 12 of them with gene feature files) [25] and some earthworms in MetaInvert (Metagenomic Monitoring of Soil Communities, 8 annelid assemblies) [26]. General-purpose databases also lack annelid-specific curation, standardized functional annotations, and a unified user interface, creating a bottleneck for cross-species comparisons and multi-omics integration, particularly for researchers without specialized computational expertise.

A further complication is the persistence of an outdated phylogenetic framework in mainstream databases. For example, NCBI's Taxonomy Database reflects traditional, morphology-based classifications that treat groups such as Polychaeta and Clitellata (Oligochaeta and Hirudinea) as distinct, equal-ranking branches. This paraphyletic view conflicts with modern phylogenomic analyses, which support the placement of Clitellata within the polychaete radiation and classify the phylum into major clades such as Errantia and Sedentaria, as well as several early-branching lineages [27–29]. Reliance on outdated taxonomy obscures relationships supported by the cited phylogenomic studies and hampers phylogenetic interpretation in comparative studies.

Table 1. Summary statistics for AnnelidaDB

Data	Statistics
Species	1121
Eukaryotic genomes	152
Chromosome-level genomes	111
Genomes with gene models	150
Species with single-cell data	5
Single-cell datasets	9
Cells	209 802
Cell clusters	171
Mitochondrial genomes	177
Proteomic studies	90
Species with proteomic data	27
Species with epigenomic data	13
Epigenomic studies	26
Species with DNA sequencing	718
Species with RNA sequencing	591
Symbiont genomes	30
Oxford plots	12 200
Gene families	91 629
Protein sequences from gene models	6 689 710

(1) Protein sequences were counted as individual peptide records compiled from multiple gene-model predictions.
(2) Oxford synteny plots represent successful pairwise comparisons among the 111 chromosome-level assemblies and comparisons between these assemblies and the ancestral Annelida linkage groups (AnLGs).
(3) Gene families were delineated using evidence-based or source gene models for species with available RNA-seq data, and ANNEVO-derived predictions for the remaining species.

To address these issues, we developed AnnelidaDB (www.annelidadb.com), a comprehensive, integrated multi-omics database dedicated to Annelida. AnnelidaDB centralizes standardized biological data, currently encompassing genomic assemblies for 152 species, transcriptomes for 591 species, 177 mitochondrial genomes, single-cell RNA-sequencing data resolving 209 802 cells, alongside epigenomic and symbiont datasets (Table 1). By retaining source annotations and generating evidence-based and *ab initio* gene models, we compiled 6 689 710 genome-derived protein sequence records against 427 545 annelid proteins in NCBI proteins (frozen on 1 September 2026). This count combines records from different annotation sets and does not represent the number of distinct biological genes. To facilitate comparative studies, the platform features an interactive view across genomic assemblies, providing a literature-curated navigation hierarchy alongside the source classification and guiding users through a modern evolutionary framework. AnnelidaDB provides integrated analytical tools—including BLAST, JBrowse2, Gene Search, and macrosyntenic analysis. Finally, we implemented a community submission portal for researchers to contribute new datasets and corrections. By bridging multi-omics data with accessible, phylogenetically informed analysis, AnnelidaDB serves as a collaborative hub for the global annelid research community and allied fields.

Materials and methods

Data collection and inclusion criteria

AnnelidaDB integrates public sequence records, omics-study metadata, taxonomic information, and ecological occurrence data. Nuclear and mitochondrial genome assemblies and their source annotations were collected from the NCBI Nucleotide

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

database and NCBI Assembly resources, and other depositions with associated publications (see data sources on the About page, www.annelidadb.com/about/). DNA- and RNA-sequencing metadata were retrieved from NCBI Sequence Read Archive (SRA). Proteomics records were collated from ProteomeXchange and linked publications where available. Epigenomic and single-cell records were curated from public repositories and their associated studies. Host-associated symbiont assemblies were retained when the source record or publication provided an identifiable annelid host and sufficient assembly metadata.

Taxonomy normalization and ecological links

NCBI Taxonomy identifiers serve as identifiers to link species records with sequence and omics resources [30]. Where available, identifiers and cross-references provided by the source databases were used to establish links to corresponding records in the World Register of Marine Species (WoRMS) [31], Global Biodiversity Information Facility (GBIF; www.gbif.org), and Ocean Biodiversity Information System (OBIS; obis.org). The originating resources can be accessed through their application programming interfaces. The node-level relationship table has been deposited in the GitHub repository (github.com/ylify/AnnelidaDB). GBIF occurrence coordinates are rendered as an interactive distribution map, and available GBIF images link to source records. WoRMS and OBIS occurrence metadata are summarized to describe the habitats and sampled depth ranges, respectively. The integration of these identifiers is distinct from the literature-based relationships adopted for the taxonomic navigation hierarchy. Rather than reconstructing a phylogeny from primary sequence data, we adopted a reference topology compiled from established phylogenomic literature to display the taxonomic relationships [28, 29, 32]. The interface presents both the NCBI classification and an updated, phylogenetically informed view, which supports users in inspecting the source taxonomy while navigating a modern hierarchy. These external data remain attributable to their originating services and are not intended as exhaustive species distributions. Images displayed in the Species Rapid View section are sourced from the public domain, with source links and attributions provided on the About page (www.annelidadb.com/about/).

Genome processing, quality assessment, and annotation

For each nuclear genome, AnnelidaDB records assembly level, assembly size, GC content, scaffold N50, and predicted gene count. Genome and gene-set completeness were assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) v6.0 with default parameters [33] against the metazoa_odb12 and lophotrochozoa_odb12.2 lineage datasets [34]. Assembly composition and completeness were visualized using BlobToolKit-style plots [35]. Source gene models are retained when available (shown on the About page, www.annelidadb.com/about/). Where RNA-seq reads were publicly available, we performed evidence-based gene prediction. It included the RNA-seq alignment using STAR v2.7.10b [36], transcript reconstruction using StringTie v2.2.1 [37], *ab initio* prediction using AUGUTUS v3.4.0 [38], and homolog prediction using miniprot v0.12 from the publicly available annelid protein sequences [39]; then the three types of evidence were integrated using EvidenceModeler v2.1.0 [40], resulting in the final gene feature files. In addition, all genomes received a second, *ab initio* gene annotation using ANNEVO [41]. Transposable elements were predicted using RepeatModeler v2.0.1 and RepeatMasker v4.1.2.p1 using default parameters [42, 43]. Coding sequences and amino acids were extracted using gff3_files_to_proteins.pl from TransDecoder v5.5.0 [44]. The source codes are provided on the About page (www.annelidadb.com/about/).

Mitochondrial genomes were re-annotated using MitoZ v3.6 [45]. Web pages retain accession identifiers, genome length, and topology, display an annotated circular map, and provide tabular gene-feature summaries. FASTA files for the mitochondrial genome, CDSs, genes, proteins, ribosomal RNAs, and transfer RNAs can be selected for download.

Bacterial symbiont genomes were re-annotated and visualized using Bakta v1.9.2 [46]. Completeness and contamination were evaluated using CheckM2 v1.0.1 [47]. Taxonomic classifications were assigned using GTDB-tk v2.6.1 [48].

Protein-coding genes were functionally annotated using eggNOG-mapper v2.1.13 and InterProScan v5.75–106.0 [49–52]. The web interface displays descriptions, preferred gene names, Gene Ontology terms, COG categories, KEGG orthology/pathway fields, EC numbers, CAZy assignments, and protein-family or domain matches where available. InterProScan member-database results include Pfam, PANTHER, Gene3D, SUPERFAMILY, NCBIfam, PROSITE, and MobiDBLite. Gene, CDS, and protein sequences are linked to genomic coordinates and gene-family membership. Repeat summaries and locations are provided as separate downloadable tracks where available.

Multi-omics metadata and single-cell processing

Transcriptome records are organized by SRA run and include species, tissue, developmental stage, sequencing platform, library layout, spot count, and base count when present in the source record. Proteome and epigenome hubs use study-centred records that retain species, tissue, stage, experimental modality, method, publication, and year. The current release displays 4585 transcriptome runs from 420 studies, 90 proteomic studies, and 26 epigenomic studies (Table 1).

Single-cell datasets are organized by species, tissue, developmental stage, study, and platform. Each dataset page identifies the source of the displayed results. For author-provided processed objects, including RDS and h5ad files, we retain the original embeddings, cluster labels, and available expression and marker-gene results without recomputing these analyses. The reference genome assembly and annotation for these objects follow those used in the original study. Datasets available as expression matrices in GEO were reanalysed using Seurat v5.4.0 [53] and DoubletFinder v2.0.6 [54]. For datasets available only as raw sequencing reads, reads were mapped to the reference genome using CellRanger v7.2.0 [55]. Unless otherwise specified, analyses performed by AnnelidaDB use the reference genome assembly and annotation provided in the corresponding genome record. The relevant R code is provided for users to inspect and reuse. The interface provides an interactive 2D embedding, cluster selection, marker-gene tables, gene-expression overlays, and expression boxplots. Users can download UMAP coordinates, R objects, expression matrices, and marker-gene results where available. Selecting a gene from the single-cell interface shows the corresponding AnnelidaDB gene detail page, connecting cell-level patterns to sequence,

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

genomic location, functional annotation, and gene-family context.

Gene families, BLAST, and comparative synteny

Protein-coding genes were grouped into gene families linking representative genes, member genes, represented species, and functional annotations. The web release distinguishes evidence-based, ANNEVO-derived, and merged family sets. The longest isoform of each gene was selected from the corresponding gene feature files and analysed with OrthoFinder v2.5.4 under default parameters to infer orthogroups [56].

Custom BLAST databases for complete nucleotide and protein sequences derived from integrated annelid genomes were powered by the BLAST + executable suite v2.17.0, supporting blastn, blastp, blastx, tblastn, and tblastx searches [57].

For macrosynteny analysis, chromosome-scale gene coordinates and protein homology were used to associate annelid chromosomes with 20 AnLGs, consistent with previous research [9, 13]. The upload workflow accepts paired protein (.pep) and coordinate (.bed) files, excludes alternative isoforms and short scaffolds according to the preprocessing rules, compares query proteins with a database of 3858 AnLG-associated orthologs, retains supported best matches, and produces a karyotype plot and an Oxford plot (Fig. 4D). The prediction of AnLGs and pairwise comparisons followed previous studies [58, 59]. Briefly, protein sequences from five chromosome-level assemblies, including two non-annelid outgroups, were compared using DIAMOND blastp v2.1.8 for pairwise comparisons with parameters –evalue 1e-3 and –max_target_seqs 50 000 [60]. These included two spiralian outgroups—the scallop *Mizuhopecten yessoensis* (Mollusca) and the ribbon worm *Lineus longissimus* (Nemertea)—alongside three representative annelids: *Owenia fusiformis* (basal Annelida / Palaeoannelida), *Streblospio benedicti* (Sedentaria), and *Alitta virens* (Errantia). Only mutual best hits present in all five assemblies and located in significantly linked groups (identified using macrosyntR) were retained, yielding 20 AnLGs [61]. Each orthologous group used to define the AnLGs contains protein sequences from the five reference species used to construct the linkage-group framework. For comparisons between AnLGs and a chromosome-level assembly, genes hit by at least 4 of 5 genes in AnLG orthologs were considered conserved. Visualizations were generated with Matplotlib v3.10.9 [62]. BCnS labels follow the ancestral Bilateria–Cnidaria–Sponge linkage-group framework described and used by Schultz *et al.* [59]. Correspondences between AnLGs and BCnS groups were inferred by comparing BCnS annotations across the five reference species. These correspondences are defined at the linkage-group level and provide a connection to the broader metazoan framework. Because the current implementation does not provide gene-level BCnS assignments for the plotted proteins, a separate BCnS-labelled dotplot is not currently available. In total, 12 200 precomputed pairwise Oxford plots spanning 111 chromosome-level assemblies are available for browsing and download.

Database construction and web implementation

The AnnelidaDB database and web application were developed using Django v4.2 and Python v3.8, with the primary database hosted on Alibaba Cloud. All software dependencies were managed and installed via BioConda [63]. An SQLite database (db.sqlite3) was incorporated to manage species records. Genome assemblies and annotation tracks are visualized with JBrowse 2 [64].

The web interface is fully responsive and freely accessible on personal computers and mobile devices. We verified stability and cross-compatibility of AnnelidaDB on Google Chrome, Microsoft Edge, Firefox, and Apple Safari. An automated email notification system (annelidadb@foxmail.com) alerts users upon completion of time-consuming analytical jobs, manages user feedback, and coordinates submission of new data.

Database architecture and core modules

Ten navigation directories

AnnelidaDB presents content through 10 top-level directories: Taxonomy, Genome, Proteome, Epigenome, Mitochondria, Transcriptome, Single-cell, Synteny, Symbiont, and Tools (Fig. 1A). This structure separates major data types without breaking species- and gene-level connections. The home page summarizes available content and provides three entry points: an interactive phylogenetic overview, a taxonomic quick search, and a rapid species-card view (Fig. 1B–E). Expanding a phylogeny node exposes descendant groups, whereas selecting a terminal taxon opens its species record.

In the current release, the Taxonomy directory covers 1121 annelid species with at least one omic dataset. The Genome directory contains 152 species, of which 111 have chromosome-level assemblies, and 150 have at least one gene-model set. AnnelidaDB also indexes 6 689 710 genome-derived protein sequences, 177 mitogenomes, 591 transcriptomic resources, 209 802 cells, 30 symbiont genomes, and 12 200 precomputed oxford plots (Table 1). These figures provide a snapshot of database content rather than biological diversity.

Taxonomy, genome, and species-centred records

To resolve fragmentation across repositories, AnnelidaDB provides an integrated navigational architecture centred on the Taxonomy and Genome hubs. These hubs consolidate ecological, taxonomic, and molecular data into unified species and genome records, enabling users to benchmark assembly contiguity and annotation completeness side-by-side and to transition seamlessly to interactive genomic browsers.

The Taxonomy hub is designed as a cross-omics inventory (Fig. 2A). Each row reports class, order, family, species, and NCBI Taxonomy ID and indicates whether nuclear genome, mitochondrial genome, DNA sequencing, RNA sequencing, epigenomic, proteomic, single-cell, or symbiont resources are available. Users can switch between NCBI-derived classifications and updated phylogenetic relationships, filter the table, and move directly to a species record.

The Genome hub compares assembly and annotation metrics across species (Fig. 2B). In addition to assembly level, size, GC content, and scaffold N50, the table shows the BUSCO completeness at the genome level and gene level (i.e. the source/evidence-based and ANNEVO gene sets). This arrangement helps distinguish assembly contiguity from gene set and genome completeness before selecting a genome for downstream analysis.

Species records consolidate information that would otherwise require navigating across several repositories (Fig. 2C). A species overview reports taxonomic identifiers, classification, habitat, and depth information where available, and

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

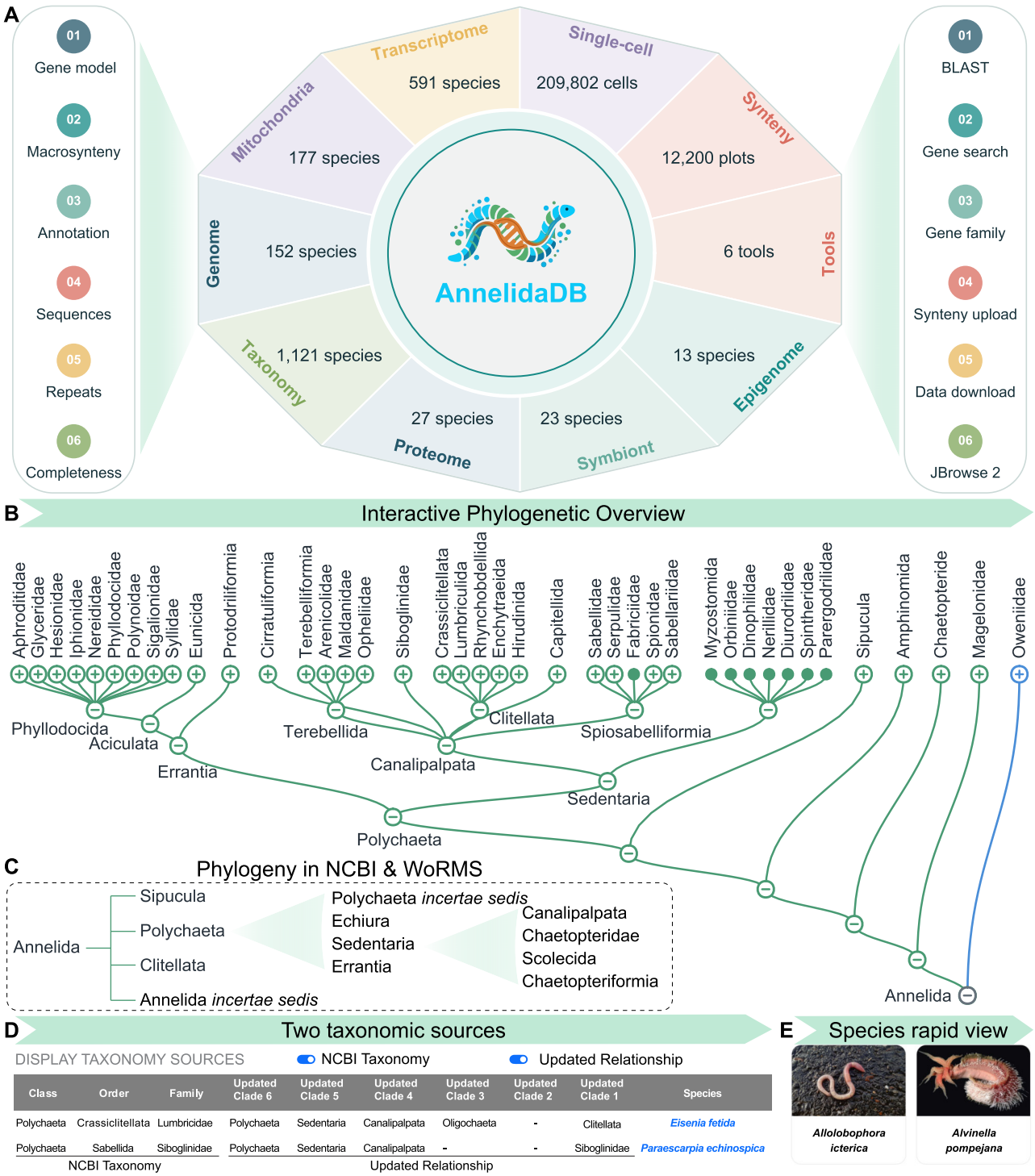


Figure 1. Overall architecture and taxonomic entry points of AnnelidaDB. (**A**) AnnelidaDB is organized into 10 top-level directories: Taxonomy, Genome, Proteome, Epigenome, Mitochondria, Transcriptome, Single-cell, Synteny, Symbiont, and Tools. Summary counts indicate the scale of each resource in the illustrated release. Genome records connect assembly statistics, gene models, macrosynteny, functional annotations, sequences, repeats, and completeness metrics. Analysis and access functions include BLAST, Gene Search, Gene Family, Synteny Upload, data download, and JBrowse 2. (**B**) Interactive phylogenetic overview of major annelid lineages; nodes can be expanded to inspect descendant taxa and species records. Note: Tree topology and clade relationships were synthesized from previous phylogenomic studies [28, 29, 32] as a literature-curated classification hierarchy, not reconstructed *de novo* from our dataset. (**C**) Simplified representation of classifications retrieved from NCBI Taxonomy and WoRMS. (**D**) Taxonomy-source controls allow users to compare the original NCBI classification with the updated phylogenetic relationships adopted by AnnelidaDB. The interactive table displays both systems alongside species-level records, exemplified by *Eisenia fetida* and *Paraescarpia echinospica*. (**E**) Rapid species-card view linking representative taxa to species-centred pages. Only taxa with genomic data are shown.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

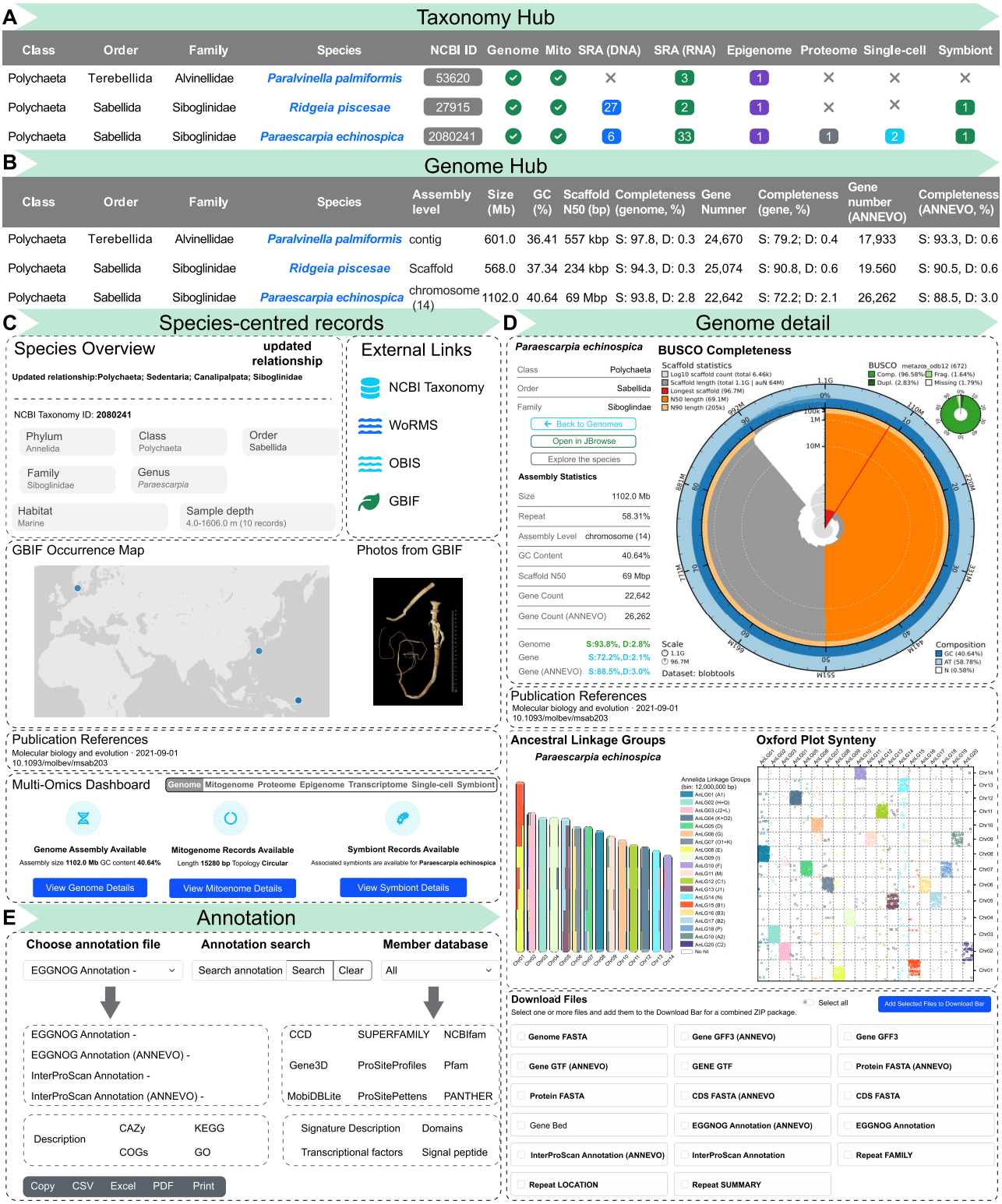


Figure 2. Taxonomy, genome, and species-centred exploration in AnnelidaDB. (**A**) Taxonomy hub showing taxonomic relationships, NCBI identifiers, and the presence of nuclear-genome, mitochondrial-genome, DNA-sequencing, RNA-sequencing, epigenome, proteome, single-cell, and symbiont records. (**B**) Genome hub comparing assembly level, size, GC content, scaffold N50, genome completeness, source gene models, and ANNEVO gene models. (**C**) Species-centred record integrating taxonomy, external links, a GBIF occurrence map and GBIF images, publications, and a multi-omics dashboard. (**D**) Genome detail page for *Paraescarpia echinospica*, including assembly statistics, BUSCO completeness, assembly composition, downloadable files, Annelida linkage-group assignments, and Oxford-plot synteny. (**E**) Searchable eggNOG and InterProScan annotation tables with filtering by member database and tabular export.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

provides links to the corresponding record on NCBI Taxonomy, WoRMS, OBIS, and GBIF. GBIF occurrence points and source-attributed images provide geographic and visual context. A multi-omics dashboard indicates which modules contain data for the species and redirects to the corresponding detail pages.

Genome detail pages combine assembly statistics, BUSCO summaries, assembly-composition graphics, publications, AnLG assignments (bar plots and Oxford plots), and file downloads (Fig. 2D). Users can open the assemblies directly in JBrowse 2 and move from genome-scale tracks to individual genes. The annotation panel supports full-text search across eggNOG and InterProScan results (Fig. 2E). It allows users to filter InterProScan member databases according to their preferences. Tables can be copied or exported in common formats, and gene IDs link to the gene detail view for the selected search results.

Transcriptome, proteome, epigenome, and mitogenome hubs

To overcome fragmentation of multi-omics datasets, AnnelidaDB establishes specialized, searchable metadata hubs and provides precise filtering by biological context (e.g. taxonomy, tissue, developmental stage).

The Transcriptome hub exposes cached SRA metadata for annelid RNA-seq runs (Fig. 3A). Users can filter by accession, platform, library layout, taxonomy, tissue, or developmental stage and can export selected records or filtered results. Links back to the source archive retain accession-level provenance.

The Proteome and Epigenome hubs organize heterogeneous experiments using common study-level metadata fields (Fig. 3B and C). Proteome records summarize tissue, stage, biological topic, and publication detail. Epigenome records retain tissue, stage, and modality, including DNA methylation and chromatin-assay datasets (ATAC-seq) where available.

The Mitogenome Hub lists species, genome length, topology, GenBank accession, and RefSeq accession (Fig. 3D). Detail pages display a circular genome map, summarize gene content, and visualize feature coordinates. Mitogenome data can be browsed and downloaded independently of nuclear-genome records.

Single-cell atlases

Single-cell RNA sequencing resolves cellular heterogeneity beyond bulk RNA-seq, which averages across tissues and masks rare or transient states [65]. This high resolution is indispensable for deciphering complex biological processes such as regeneration, segment formation, tissue specialization, and host-symbiont physiology in annelids [7, 15, 16, 66].

To bridge fragmented cell-type labels, marker sets, and processed objects across studies, the AnnelidaDB Single-cell directory provides an interactive atlas (Fig. 3E) that currently curates nine datasets from five species, encompassing 209 802 cells and 171 annotated clusters. By coupling dataset-level metadata with gene-aware visualizations, users can select clusters, inspect marker genes, project gene expression onto low-dimensional embeddings, and quantitatively compare expression distributions. Cross-linking these single-cell visual modalities and core gene records connects cell-type-specific expression patterns to genomic context and functional annotation, translating static observations into testable comparative hypotheses across diverse annelid tissues and lineages. Expression values and cell-type labels should be interpreted within the context of each dataset; the shared interface does not by itself establish quantitative comparability across studies or species. Users can download available processed objects (.rds or .h5ad) and related files for offline analysis and further comparison.

Host-associated symbionts

Symbiosis is a major evolutionary adaptation route by which several annelid lineages occupy chemically challenging habitats [67, 68]. Deep-sea siboglinids, e.g. derive nutrition from intracellular chemoautotrophic bacteria, and genomic, proteomic, and single-cell studies show that host physiology and microbial metabolism must be interpreted together [10, 12, 14, 16, 69]. Microbial genomes lose context without host identity, tissue information, or assembly quality. Linking each symbiont genome to a curated annelid host therefore creates the unit of comparison needed to investigate how symbioses evolve and support the ecological specialization of the host.

The Symbiont hub addresses this need (Fig. 3F). It currently connects 30 bacterial symbiont species to 23 annelid hosts and reports genome size, GC content, contig number, estimated completeness, contamination, and microbial taxonomy. Symbiont detail pages display assembly and annotation summaries, circular genome views, and selectable genome, protein, gene, and GFF3 downloads. Host identifiers preserve the connection between each microbial assembly and the corresponding annelid species record.

Database access and analytical tools

Interconnected gene, annotation, and gene-family records

Many questions in annelid biology begin with a named gene or phenotype and require distinguishing orthologs, paralogs, and lineage-specific changes across deeply diverged species. Integrating sequence similarity with gene models, functional annotations, genomic positions, and family membership is essential [3, 7].

AnnelidaDB treats genes, functional annotations, and gene families as connected objects rather than separate result tables (Fig. 4A). Gene Search accepts a gene name, description, or annotation term and can be restricted by species. A gene detail record reports the species, taxonomy ID, description, genomic interval, strand, protein and CDS sequences, gene-family membership, and available eggNOG and InterProScan annotations. From this page, users can open the genomic interval in JBrowse 2, inspect other family members, or return to the species and genome views.

The Gene Family directory lists family IDs, source sets, representative genes, species counts, and member counts. Selecting a gene family directs users to a detailed view of its constituent members and functional annotations. This bidirectional navigation supports workflows that start from a functionally described gene to explore its expression across annelids, or from a gene family to examine the annotation and genomic context of individual members.

BLAST search against AnnelidaDB reference databases

To leverage the expanded gene models and annotations, AnnelidaDB provides a BLAST + service (Fig. 4C) against custom, annelid-specific nucleotide and protein databases. All standard algorithms (blastn, blastp, blastx, tblastn, and

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

Figure 3. Multi-omics and host-associated symbiont resources. (**A**) Transcriptome hub with SRA run, species, tissue, developmental stage, platform, library layout, spot, and base metadata. (**B**) Proteome Hub summarizing species, tissue, stage, biological topic, and publication link. (**C**) Epigenome Hub organizing species, tissue, stage, assay type, and publication link. (**D**) Mitogenome hub and detail view showing accession, topology, a circular genome map, gene features, and selectable FASTA downloads. (**E**) Single-cell dataset record and interactive atlas with cluster-level visualization, marker genes, gene-expression overlays, boxplots, functional-annotation links, and data downloads. (**F**) Host-associated symbiont genome record with assembly quality, taxonomy, circular visualization, annotation summary, and downloadable files.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

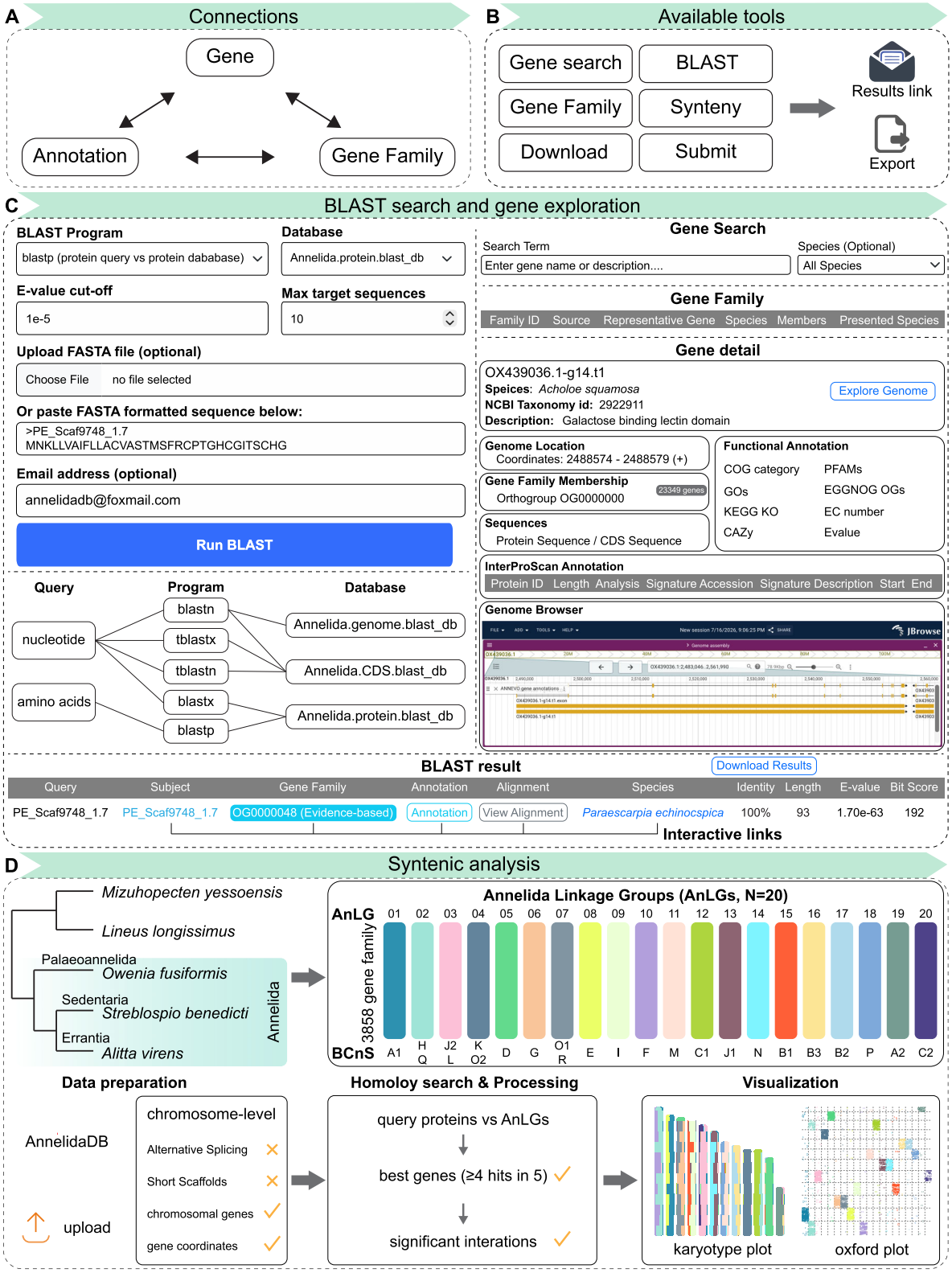


Figure 4. Gene-centred tools and comparative synteny workflows. (**A**) Bidirectional connections among genes, functional annotations, and gene families. (**B**) Main access and analysis tools, with downloadable outputs or persistent result links where applicable. (**C**) BLAST searches against AnnelidaDB sequence databases and gene-centred exploration. Nucleotide or amino-acid queries can be compared with genome, CDS, or protein databases using the appropriate BLAST program; hits link to alignments, annotations, gene families, species records, and genomic context in JBrowse 2. (**D**) Synteny workflow for chromosome-level user data. Protein and BED files are filtered, compared with the AnLG reference set, and visualized as a karyotype plot and Oxford plot. The upper panel summarizes the 20 AnLGs and their group-level correspondences to the Bilateria–Cnidaria–Sponge (BCnS) linkage-group framework.

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

tblastx) are supported with configurable parameters, including *E*-value thresholds and target limits. Users can submit queries via FASTA input and optionally provide an email address to receive job links. Output includes standard alignment metrics (identity, *E*-value, bit score) and an interactive alignment view. Hits matching AnnelidaDB genes link back to detailed functional annotations, gene families, and precise genomic neighbourhoods, enabling sequence-centred exploration, including for pre-publication data.

Synteny browsing and user-submitted comparisons

With the rapid accumulation of genomic resources, particularly chromosome-level genome assemblies, synteny and karyotype evolution have become central for interpreting the structural basis of evolutionary divergence and convergence. Conserved linkage groups coexist with lineage-specific rearrangement, including rearrangement bursts associated with the origin of non-marine clades [6, 9, 13]. Gene-level homology alone cannot distinguish chromosome fusion, fission, and translocation histories; these patterns emerge when homologous genes are compared in macrosynteny across species. AnnelidaDB provides an online platform to explore these patterns across annelid genomes.

The Synteny directory provides 12 200 precomputed Oxford plots spanning 111 species with chromosome-level assemblies. Selecting a species pair displays the distribution of homologous genes among chromosomes and enables retrieval of the plot. The available plots can be browsed individually in the Synteny module (https://www.annelidadb.com/synteny/). After users select 'Species 1', the available comparison options are displayed under 'Species 2'. Genome pages complement pairwise views with karyotype and Oxford plots that assign chromosomes to the 20 AnLGs.

For species not yet represented by a precomputed comparison, the Synteny Upload tool accepts a protein FASTA file and its BED-format coordinate file (Fig. 4D). The job maps supported query genes to the AnLG reference and returns a karyotype visualization, an Oxford plot, and intermediate results, placing newly assembled chromosome-level genomes into the same comparative framework without software installation. An optional email address provides a persistent result link when the computation is complete.

Download, export, and community submission

The value of a multi-omics resource depends not only on what data user can view and reuse, but also on how easily it can be audited, updated, and integrated. This is particularly important for Annelida, where new genomes are accumulating rapidly, yet datasets are dispersed across repositories, and assemblies may lack accessible gene-feature files. Selective downloads enable reproducible downstream analysis with explicit sequence and annotation provenance, while reviewed community submissions allow specialists to contribute new datasets and correct taxonomic or metadata errors while preserving curation control.

Genome, mitogenome, and symbiont pages show file-level checkboxes, and hub tables support record-level export. Available genome files include assembly FASTA, GFF3, GTF, CDS, protein, and gene FASTA files; standardized ANNEVO models, eggNOG annotations, InterProScan annotations, and repeat tracks are offered where available. The download bar persists across pages and packages the final selection as a ZIP archive.

AnnelidaDB is curated and maintained by Prof. Jian-Wen Qiu and Dr Yunlong Li at Hong Kong Baptist University. We target routine resource updates approximately every 3 months, with the timing and scope of individual updates guided by the team's research activities and feedback from users. AnnelidaDB welcomes community contributions of omics datasets across all data types represented in the database. Researchers can submit datasets for storage and integration through the submission portal (https://www.annelidadb.com/submit/). No specific database-ready file format is required at submission. Submitters are encouraged, but not required, to provide MD5 checksums calculated from the original files, allowing file integrity to be verified after transfer. Following submission, the AnnelidaDB team curates the data and converts or adapts the submitted files into formats compatible with the database. Submitted genome assemblies undergo a standardized processing and assessment workflow, including transposable-element prediction and completeness assessment. These analyses use consistent software versions within each database release, improving consistency across records and supporting downstream comparative analyses. We communicate directly with submitters by email to clarify dataset content and metadata, resolve integration issues, and ensure that their contributions are accurately represented. This process facilitates the incorporation of diverse community-generated resources and the continued expansion of AnnelidaDB.

Example: connecting multi-omics records for *Paraescarpia echinospica*

The deep-sea siboglinid *Paraescarpia echinospica* illustrates how AnnelidaDB supports exploration of existing resources relevant to host–symbiont biology [14, 69, 70]. Starting from the species record (NCBI Taxonomy ID 2080241), users can access the host nuclear genome, inspect assembly quality, and compare source and ANNEVO-derived gene sets using their separate completeness summaries. For a candidate host gene identified from a publication or functional annotation, users can inspect its sequence, annotation, and genomic location through the gene and genome views. The species record also links the plume and trophosome single-cell datasets, where users can examine gene-expression patterns and cluster annotations within each dataset. These expression views retain their dataset-specific interpretation. The same host record links the associated bacterial symbiont genome and its annotation, allowing users to retrieve microbial information alongside the relevant host resources. Source accessions, study metadata and publication links preserve the context of these records, while available sequence, annotation, and single-cell files can be downloaded for further investigation. This organization provides a connected route from a host species and candidate gene to dataset-specific expression views and a host-associated microbial record, reducing the need to locate the resources independently across multiple portals.

Future directions

To ensure its long-term utility and relevance to the scientific community, we have outlined a robust roadmap for the continuous development of AnnelidaDB. We are committed to

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

maintaining and regularly updating the database as new omics data become available. We aim to expand and integrate biological content, including in-depth analyses of gene family evolution, multi-dimensional gene expression patterns, and regulatory trends by systematically integrating proteomics, single-cell transcriptomics, and epigenomic modification data. We will add a module for organizing, version tracking, and comparing multiple genome assemblies for the same species. To enhance the usability, reproducibility, and analytical capabilities of the database, we will deploy additional built-in bioinformatics tools and deposit source code used in publications on AnnelidaDB. Planned features include interactive phylogenomic pipelines, advanced cross-species single-cell data integration and automated cell-type annotation, and online pipelines for structural and functional genome annotation. Finally, we will explore the integration of Large Language Models trained on AnnelidaDB to provide targeted predictions, data mining, and biological insights tailored explicitly to annelid biology, transforming AnnelidaDB from a static data repository into a dynamic, predictive resource.

Acknowledgements

We thank Carmen Liu and Yu Sheung Law for technical and logistical support.

Author contributions: Yunlong Li (Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Resources [lead], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Haokun Liu (Formal analysis [equal], Methodology [equal], Writing—review & editing [equal]), Maeva Perez (Formal analysis [equal], Methodology [equal], Writing—review & editing [equal]), Yi-Xuan Li (Formal analysis [equal], Methodology [equal], Validation [equal], Writing—review & editing [equal]), Fan Hui (Writing—review & editing [equal]), Zihui Shen (Writing—review & editing [equal]), Yanan Sun (Writing—review & editing [equal]), José M. Martín-Durán (Writing—review & editing [equal]), Greg W. Rouse (Writing—review & editing [equal]), Pei-Yuan Qian (Writing—review & editing [equal]), Jianliang Xu (Writing—review & editing [equal]), Xike Xie (Formal analysis [equal], Methodology [equal], Writing—review & editing [equal]), and Jian-WenQiu (Conceptualization, Funding acquisition, Validation, Writing—review & editing).

Conflict of interest

None declared.

Funding

This work was supported by the Research Grants Council of Hong Kong (Collaborative Research Fund [C2013-22GF] to J.-W.Q. and the General Research Fund [121010] to J.-W.Q.), the National Natural Science Foundation of China (62472400 to X.X.), and the Sanming Project of Medicine in Shenzhen (SZSM202411033 to X.X.).

Data availability

AnnelidaDB is freely available at www.annelidadb.com without login or registration. Genome, mitogenome, symbiont, annotation, single-cell, and comparative-genomics files can be downloaded through the download bar on the corresponding record page. Information about data sources (e.g. publication and accession number) and software employed or integrated in this database is provided on the About page (www.annelidadb.com/about/). Users can refer to the AnnelidaDB User Guide (www.annelidadb.com/user-guide/) for instructions on navigating the database, selecting datasets, and interpreting results. To support reproducibility and reuse, the AnnelidaDB source code, selected example files, and node-level tables are available through the project's GitHub repository (github.com/ylify/AnnelidaDB) and figshare (doi.org/10.6084/m9.figshare.34467162). These materials are intended to facilitate reproduction of the database workflows and adaptation by other research communities. The authors are committed to maintaining the resource at the current URL for at least 5 years post-publication, supported by dedicated, pre-paid cloud server hosting.

References

1. Rouse G, Pleijel F, Tilic E. *Annelida*. Oxford University Press. 2022. https://doi.org/10.1093/oso/9780199692309.001.0001
2. Magalhães WF, Hutchings P, Oceguera-Figueroa A *et al.* Segmented worms (Phylum Annelida): a celebration of twenty years of progress through Zootaxa and call for action on the taxonomic work that remains. *Zootaxa* 2021;4979:190–211. https://doi.org/10.11646/zootaxa.4979.1.18
3. Phillips HRP, Guerra CA, Bartz MLC *et al.* Global distribution of earthworm diversity. *Science* 2019;366:480–5. https://doi.org/10.1126/science.aax4851
4. Hasan MN, Reynolds JW, Narayanan SP *et al.* Global diversity and taxonomic overview of the phylum Annelida: an integrative review. *Zootaxa* 2026;5849:502–24. https://doi.org/10.11646/zootaxa.5849.4.3
5. Capa M, Hutchings P. Annelid diversity: historical overview and future perspectives. *Diversity* 2021;13:129. https://doi.org/10.3390/d13030129
6. Martín-Zamora FM, Liang Y, Guynes K *et al.* Annelid functional genomics reveal the origins of bilaterian life cycles. *Nature* 2023;615:105–10. https://doi.org/10.1038/s41586-022-05636-7
7. Shao Y, Wang X-B, Zhang J-J *et al.* Genome and single-cell RNA-sequencing of the earthworm *Eisenia andrei* identifies cellular mechanisms underlying regeneration. *Nat Commun* 2020;11:2656. https://doi.org/10.1038/s41467-020-16454-8
8. Struck TH, Golombek A, Weigert A *et al.* The evolution of annelids reveals two adaptive routes to the interstitial realm. *Curr Biol* 2015;25:1993–9. https://doi.org/10.1016/j.cub.2015.06.007
9. Vargas-Chávez C, Benítez-Álvarez L, Martínez-Redondo GI *et al.* An episodic burst of massive genomic rearrangements and the origin of non-marine annelids. *Nat Ecol Evol* 2025;9:1263–79. https://doi.org/10.1038/s41559-025-02728-1
10. Moggioli G, Panossian B, Sun Y *et al.* Distinct genomic routes underlie transitions to specialised symbiotic lifestyles in deep-sea annelid worms. *Nat Commun* 2023;14:2814. https://doi.org/10.1038/s41467-023-38521-6
11. Chipman AD. *Organismic Animal Biology: An Evolutionary Approach*. Oxford University Press, 2024. https://doi.org/10.1093/oso/9780192893581.001.0001
12. Markert S, Arndt C, Felbeck H *et al.* Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. *Science* 2007;315:247–50. https://doi.org/10.1126/science.1132913
13. Lewin TD, Liao IJ, Luo Y-J. Annelid comparative genomics and the evolution of massive lineage-specific genome rearrangement in bilaterians. *Mol Biol Evol* 2024;41:msae172.
14. Sun Y, Sun J, Yang Y *et al.* Genomic signatures supporting the symbiosis and formation of chitinous tube in the deep-sea

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

tubeworm *Paraescarpia echinospica*. *Mol Biol Evol* 2021;38:4116–34. https://doi.org/10.1093/molbev/msab203
15. Davies BE, Gonzalez P, Sur A *et al.* A chromosome-level assembly and functional genomic resources for the model annelid *Capitella teleta*. *Genome Biol Evol* 2026;18 evag041. https://doi.org/10.1093/gbe/evag041
16. Wang H, Xiao H, Feng B *et al.* Single-cell RNA-seq reveals distinct metabolic "microniches" and close host-symbiont interactions in deep-sea chemosynthetic tubeworm. *Sci Adv* 2024;10:eadn3053. https://doi.org/10.1126/sciadv.adn3053
17. Guynes K, Sarre LA, Carrillo-Baltodano AM *et al.* Annelid methylomes reveal ancestral developmental and aging-associated epigenetic erosion across Bilateria. *Genome Biol* 2024;25:204. https://doi.org/10.1186/s13059-024-03346-z
18. Stockinger AW, Adelmann L, Fahrenberger M *et al.* Molecular profiles, sources and lineage restrictions of stem cells in an annelid regeneration model. *Nat Commun* 2024;15:9882. https://doi.org/10.1038/s41467-024-54041-3
19. Sayers EW, Beck J, Bolton EE *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* 2025;53:D20–9. https://doi.org/10.1093/nar/gkae979
20. Burgin J, Ahamed A, Cummins C *et al.* The European Nucleotide Archive in 2022. *Nucleic Acids Res* 2023;51:D121–5. https://doi.org/10.1093/nar/gkac1051
21. Liu F, Li Y, Yu H *et al.* MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. *Nucleic Acids Res* 2021;49:D1556. https://doi.org/10.1093/nar/gkaa1166
22. Liu F, Cai B, Lian S *et al.* MolluscDB 2.0: a comprehensive functional and evolutionary genomics database for over 1400 molluscan species. *Nucleic Acids Res* 2025;53:D1075–86. https://doi.org/10.1093/nar/gkae1026
23. Kudtarkar P, Cameron RA. Echinobase: an expanding resource for echinoderm genomic information. *Database* 2017;2017:bax074.
24. Walsh AT, Triant DA, Le Tourneau JJ *et al.* Hymenoptera Genome Database: new genomes and annotation datasets for improved go enrichment and orthologue analyses. *Nucleic Acids Res* 2022;50:D1032–9. https://doi.org/10.1093/nar/gkab1018
25. She J, Qian P-Y, Wu L. DOO: integrated multi-omics resources for deep ocean organisms. *Nucleic Acids Res* 2026;54:D1031–45. https://doi.org/10.1093/nar/gkaf1096
26. Collins G, Schneider C, Boštjančić LL *et al.* The MetaInvert soil invertebrate genome resource provides insights into below-ground biodiversity and evolution. *Commun Biol* 2023;6:1241. https://doi.org/10.1038/s42003-023-05621-4
27. Weigert A, Bleidorn C. Current status of annelid phylogeny. *Org Divers Evol* 2016;16:345–62. https://doi.org/10.1007/s13127-016-0265-7
28. Struck TH, Paul C, Hill N *et al.* Phylogenomic analyses unravel annelid evolution. *Nature* 2011;471:95–8. https://doi.org/10.1038/nature09864
29. Andrade SCS, Novo M, Kawauchi GY *et al.* Articulating "archiannelids": phylogenomics and annelid relationships, with emphasis on meiofaunal taxa. *Mol Biol Evol* 2015;32:2860–75. https://doi.org/10.1093/molbev/msv157
30. Schoch CL, Ciufo S, Domrachev M *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* 2020;2020:baaa062.
31. Vandepitte L, Dekeyzer S, Decock W *et al.* The World Register of Marine Species (WoRMS) through the looking glass: insights from the data management team in light of the crystal anniversary of WoRMS. *Hydrobiologia* 2025;852:1–22. https://doi.org/10.1007/s10750-024-05688-y
32. Li Y, Kocot KM, Whelan NV *et al.* Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. *Zoologica Scripta* 2017;46:200–13. https://doi.org/10.1111/zsc.12201
33. Manni M, Berkeley MR, Seppey M *et al.* BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;38:4647–54. https://doi.org/10.1093/molbev/msab199
34. Tegenfeldt F, Kuznetsov D, Manni M *et al.* OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res* 2025;53:D516–22. https://doi.org/10.1093/nar/gkae987
35. Challis R, Richards E, Rajan J *et al.* BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)* 2020;10:1361–74. https://doi.org/10.1534/g3.119.400908
36. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635
37. Pertea M, Pertea GM, Antonescu CM *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;33:290–5. https://doi.org/10.1038/nbt.3122
38. Stanke M, Keller O, Gunduz I *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res* 2006;34:W435–9. https://doi.org/10.1093/nar/gkl200
39. Li H. Protein-to-genome alignment with miniprot. *Bioinformatics* 2023;39:btad014.
40. Haas BJ, Salzberg SL, Zhu W *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 2008;9:R7. https://doi.org/10.1186/gb-2008-9-1-r7
41. Zhang P, Xu T, Wang S *et al.* Highly accurate *ab initio* gene annotation with ANNEVO. *Nat Methods* 2026;23:740–8. https://doi.org/10.1038/s41592-026-03036-7
42. Flynn JM, Hubley R, Goubert C *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* 2020;117:9451–57. https://doi.org/10.1073/pnas.1921046117
43. Tempel S. Using and Understanding RepeatMasker. In: Bigot Y (ed.), *Mobile Genetic Elements. Methods in Molecular Biology*.Humana Press, 2012, 859, 29–51. https://doi.org/10.1007/978-1-61779-603-6 10.1007/978-1-61779-603-6_2
44. Haas BJ, Papanicolaou A, Yassour M *et al.* *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 2013;8:1494–512. https://doi.org/10.1038/nprot.2013.084
45. Meng G, Li Y, Yang C *et al.* MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic Acids Res* 2019;47:e63. https://doi.org/10.1093/nar/gkz173
46. Schwengers O, Jelonek L, Dieckmann MA *et al.* Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* 2021;7:000685.
47. Chklovski A, Parks DH, Woodcroft BJ *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;20:1203–12. https://doi.org/10.1038/s41592-023-01940-w
48. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;38:5315–6. https://doi.org/10.1093/bioinformatics/btac672
49. Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al.* eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 2021;38:5825–9. https://doi.org/10.1093/molbev/msab293
50. Jones P, Binns D, Chang H-Y *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;30:1236–40. https://doi.org/10.1093/bioinformatics/btu031
51. Huerta-Cepas J, Szklarczyk D, Heller D *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–14. https://doi.org/10.1093/nar/gky1085

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

52. Blum M, Andreeva A, Florentino LC *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res* 2025;53:D444–56. https://doi.org/10.1093/nar/gkae1082
53. Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal, and scalable single-cell analysis. *Nat Biotechnol* 2024;42:293–304. https://doi.org/10.1038/s41587-023-01767-y
54. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst* 2019;8:329–37. https://doi.org/10.1016/j.cels.2019.03.003
55. Zheng GXY, Terry JM, Belgrader P *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;8:14049. https://doi.org/10.1038/ncomms14049
56. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019;20:238. https://doi.org/10.1186/s13059-019-1832-y
57. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinf* 2009;10:421. https://doi.org/10.1186/1471-2105-10-421
58. Sigwart JD, Li Y, Chen Z *et al.* Still waters run deep in large-scale genome rearrangements of morphologically conservative Polyplacophora. *eLife* 2025;13:RP102542. https://doi.org/10.7554/eLife.102542
59. Schultz DT, Haddock SHD, Bredeson JV *et al.* Ancient gene linkages support ctenophores as sister to other animals. *Nature* 2023;618:110–7. https://doi.org/10.1038/s41586-023-05936-6
60. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;18:366–8. https://doi.org/10.1038/s41592-021-01101-x
61. El Hilali S, Copley RR. macrosyntR : drawing automatically ordered Oxford Grids from standard genomic files in R. bioRxiv, https://doi.org/10.1101/2023.01.26.525673, 27 January 2023, preprint: not peer reviewed.
62. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng* 2007;9:90–5. https://doi.org/10.1109/MCSE.2007.55
63. Grüning B, Dale R, Sjödin A *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;15:475–6. https://doi.org/10.1038/s41592-018-0046-7
64. Diesh C, Stevens GJ, Xie P *et al.* JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol* 2023;24:74. https://doi.org/10.1186/s13059-023-02914-z
65. Baysoy A, Bai Z, Satija R *et al.* The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;24:695–713. https://doi.org/10.1038/s41580-023-00615-w
66. Li J, Wang J, Zhang P *et al.* Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. *Nat Genet* 2022;54:1711–20. https://doi.org/10.1038/s41588-022-01197-7
67. Thornhill DJ, Fielman KT, Santos SR *et al.* Siboglinid-bacteria endosymbiosis. *Commun Integr Biol* 2008;1:163–6. https://doi.org/10.4161/cib.1.2.7108
68. Dubilier N, Blazejak A, Rühland C. Symbioses between Bacteria and Gutless Marine Oligochaetes. In: Overmann J (ed.), *Molecular Basis of Symbiosis*. Berlin, Heidelberg: Springer, 2025, 41, 251–75. https://doi.org/10.1007/3-540-28221-1 10.1007/3-540-28221-1_12
69. Yang Y, Sun J, Sun Y *et al.* Genomic, transcriptomic, and proteomic insights into the symbiosis of deep-sea tubeworm holobionts. *ISME J* 2020;14:135–50. https://doi.org/10.1038/s41396-019-0520-y
70. Sun Y, Liang Q, Sun J *et al.* The mitochondrial genome of the deep-sea tubeworm *Paraescarpia echinospica* (Siboglinidae, Annelida) and its phylogenetic implications. *Mitochondrial DNA Part B* 2018;3:131–2. https://doi.org/10.1080/23802359.2018.1424576

Downloaded from academic.oup.com/nar/advance-article/doi/10.1093/nar/gkag1087/9005756 by guest on 31 October 2026

Received: August 13, 2026. **Revised:** October 12, 2026. **Accepted:** October 13, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.