MATEdb2, a Collection of High-Quality Metazoan Proteomes across the Animal Tree of Life to Speed Up Phylogenomic Studies

Gemma I. Martínez-Redondo [1,*], Carlos Vargas-Chávez [1], Klara Eleftheriadi [1], Lisandra Benítez-Álvarez [1], Marçal Vázquez-Valls [1], Rosa Fernández [1,*]

[1]Metazoa Phylogenomics Lab, Biodiversity Program, Institute of Evolutionary Biology (CSIC-University Pompeu Fabra), 08003 Barcelona, Spain

*Corresponding authors: E-mails: gemma.martinez@ibe.upf-csic.es; rosa.fernandez@ibe.upf-csic.es.

Accepted: October 17, 2024

Abstract

Recent advances in high-throughput sequencing have exponentially increased the number of genomic data available for animals (Metazoa) in the last decades, with high-quality chromosome-level genomes being published almost daily. Nevertheless, generating a new genome is not an easy task due to the high cost of genome sequencing, the high complexity of assembly, and the lack of standardized protocols for genome annotation. The lack of consensus in the annotation and publication of genome files hinders research by making researchers lose time in reformatting the files for their purposes but can also reduce the quality of the genetic repertoire for an evolutionary study. Thus, the use of transcriptomes obtained using the same pipeline as a proxy for the genetic content of species remains a valuable resource that is easier to obtain, cheaper, and more comparable than genomes. In a previous study, we presented the Metazoan Assemblies from Transcriptomic Ensembles database (MATEdb), a repository of high-quality transcriptomic and genomic data for the two most diverse animal phyla, Arthropoda and Mollusca. Here, we present the newest version of MATEdb (MATEdb2) that overcomes some of the previous limitations of our database: (i) we include data from all animal phyla where public data are available, and (ii) we provide gene annotations extracted from the original GFF genome files using the same pipeline. In total, we provide proteomes inferred from high-quality transcriptomic or genomic data for almost 1,000 animal species, including the longest isoforms, all isoforms, and functional annotation based on sequence homology and protein language models, as well as the embedding representations of the sequences. We believe this new version of MATEdb will accelerate research on animal phylogenomics while saving thousands of hours of computational work in a plea for open, greener, and collaborative science.

Key words: comparative genomics, phylogenomics, genome analysis, proteomes, sequence databases.

Significance

The increasing availability of genomic sequences has advanced animal phylogenomic studies, but inconsistencies in data preprocessing and annotation hinder their full potential. Previous studies provided valuable resources, yet issues like outdated data and varying computational pipelines create challenges in data comparability and accuracy. MATEdb2 addresses these gaps by expanding taxonomic coverage to nearly 1,000 species and introducing a standardized pipeline for processing genome data, ensuring consistency and quality. The new database also includes advanced functional annotations using cutting-edge protein language models. These improvements make MATEdb2 a more reliable and comprehensive resource, significantly enhancing the accuracy and utility of comparative genomic studies across the Animal Tree of Life.

Introduction

In the midst of an explosion in the availability of genomic sequences, the advancement of phylogenomic, phylotranscriptomic, and comparative genomic studies in animals is hindered by the preprocessing and homogenization of the input data. With high-quality chromosome-level genomes being published almost daily in the last few years, we are gaining access to new biological knowledge that is helping to solve trickier scientific questions, such as the identity of the sister taxon to Bivalvia (Song et al. 2023) or the evolution of noncoding and repetitive regions (Osmanski et al. 2023). In addition, the use of transcriptomes as a proxy of a species proteome continues to be a main source of proteome data as a cheaper and easier alternative for phylogenetic inference (Zapata et al. 2014; Mongiardino Koch et al. 2018; Erséus et al. 2020, among others) and gene repertoire evolution (De Oliveira et al. 2016; Thoma et al. 2019; Fernández and Gabaldón 2020) in less-studied animals.

Together, these genomic and transcriptomic studies have provided a vast number of resources for a plethora of animals that cannot be directly used in phylogenomic studies before proper preprocessing. This is especially true for older data sets where data quality is much lower and can have a high impact on the results obtained. Moreover, the use of different computational pipelines for data processing makes data not comparable and prone to false positives and negatives. To some extent, the transcriptome assembly methodology used can impact the comparability among different data sets. For example, the number of "genes" inferred with Trinity for the subset of mollusk transcriptomes obtained from Krug et al. (2022) is significantly different—$P < 0.1$—to the ones we obtained (supplementary fig. S1, Supplementary Material online). Additionally, due to different nomenclature across files, the "ready-to-use" protein files provided in some genome sequencing projects cannot be easily matched with the other genome files for additional analyses. This mainly impacts research groups with lower computational resources or experience who cannot leverage the publicly available data into their research. To help alleviate these issues, we previously published the Metazoan Assemblies from Transcriptomic Ensembles database (MATEdb) containing high-quality transcriptome assemblies for 335 arthropods and mollusks (Fernández et al. 2022). Here, we present its second version, MATEdb2, that differs from the previous one in three main aspects: (i) we have increased the taxonomic sampling to all animal phyla with high-quality data publicly available and provide the first transcriptomic sequences for some animal taxa; (ii) we include a standardized pipeline for extracting protein sequences directly from GFF genomic files, rather than relying on precomputed protein files, facilitating replication and integration with the corresponding genomic data; and (iii) we provide the functional annotation of all proteins using a protein language-based new methodology that outperforms traditional methods (Barrios-Núñez et al. 2024). We hope that this newer version of MATEdb accelerates research on animal evolution by providing a wider taxonomic resource of high-quality proteomes across the Animal Tree of Life.

Results

Increased Taxonomic Coverage

The first version of MATEdb (Fernández et al. 2022) included high-quality data sets from 335 species of arthropods and mollusks, with special attention to lineage representation within each phylum. Here, we provide a newer version of MATEdb that expands the taxonomic representation across the Animal Tree of Life by incorporating a total of 970 species from virtually all animal phyla that have publicly available genomic or transcriptomic data, as well as some outgroup species relevant for understanding animal evolution. Taxon sampling tried to maximize the taxonomic representation within each phylum while considering the quality of the data. The number of proteomes per phylum included in MATEdb2 and the distribution of gene number are represented in Fig. 1a and b, and the complete list of species and their metadata is included in supplementary table S1, Supplementary Material online.

Improved Analytical Pipeline for Genomes

In the previous version of MATEdb (Fernández et al. 2022), we directly downloaded the coding DNA sequences (CDS) and proteome files from the public repositories in the case of genomes. However, a closer inspection of both files together with their corresponding genome sequence and annotation revealed incongruences between them that needed to be manually curated. Just looking at gene numbers, only 9 out of the 59 genomes we are keeping from the previous version of MATEdb had the same number of protein-coding genes (just considering the longest isoform) in the peptide and GFF files. In addition, most of the proteomes differed in more than 1,000 genes, with 15 having more than 10k of difference. This is caused by the lack of consensus in the annotation and publication of genome files, with some authors uploading modified versions of the protein sequences that do not map directly with the reported GFF and FASTA files, hindering the utility of those files for additional analyses. Another case we encountered when building the new version MATEdb2 was for the only chromosome-level ctenophore genome available back then (*Hormiphora californensis*), in which the proteins extracted directly from the GFF and FASTA files contained premature stop codons in virtually all the sequences, which made us discard this species. Moreover, even highly curated public databases can contain wrong or missing data, such as the case of *Apis mellifera* and *Anopheles gambiae*'s CDS file in UniProt (UniProt Consortium 2023) containing only a

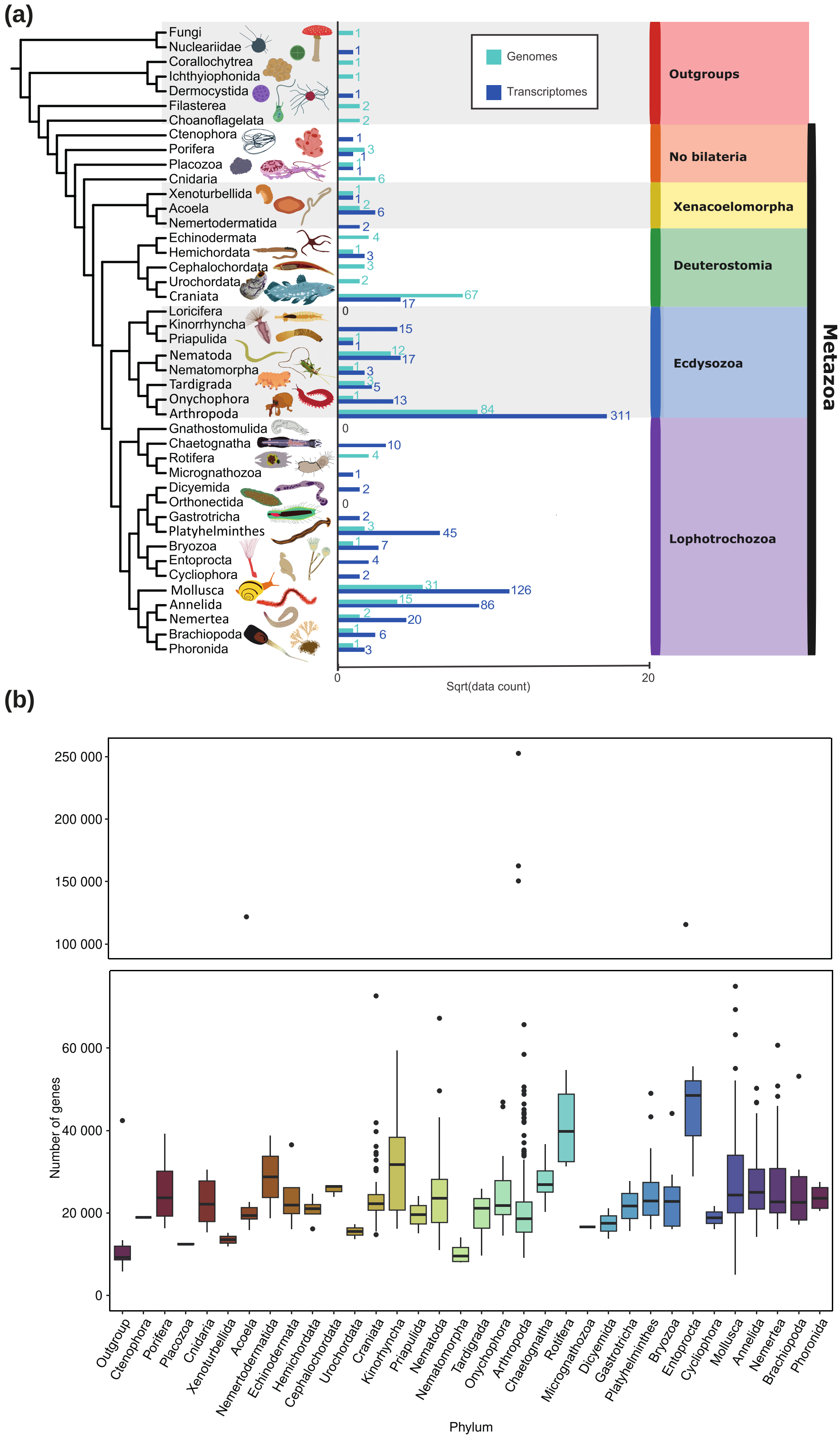

Fig. 1. Taxonomic representation of species included in MATEdb2. a) The number of data sets per phylum is separated by the type of data type: genomes and transcriptomes. The phylogenetic tree topology was manually constructed based on the latest phylogenomic studies or accepted topologies (Giribet and Edgecombe 2017; Lu et al. 2017; Laumer et al. 2019; Marlétaz et al. 2019; Khalturin et al. 2022). b) Distribution of the number of genes per proteome across phyla. All artwork was designed for this study by Gemma I. Martínez-Redondo.

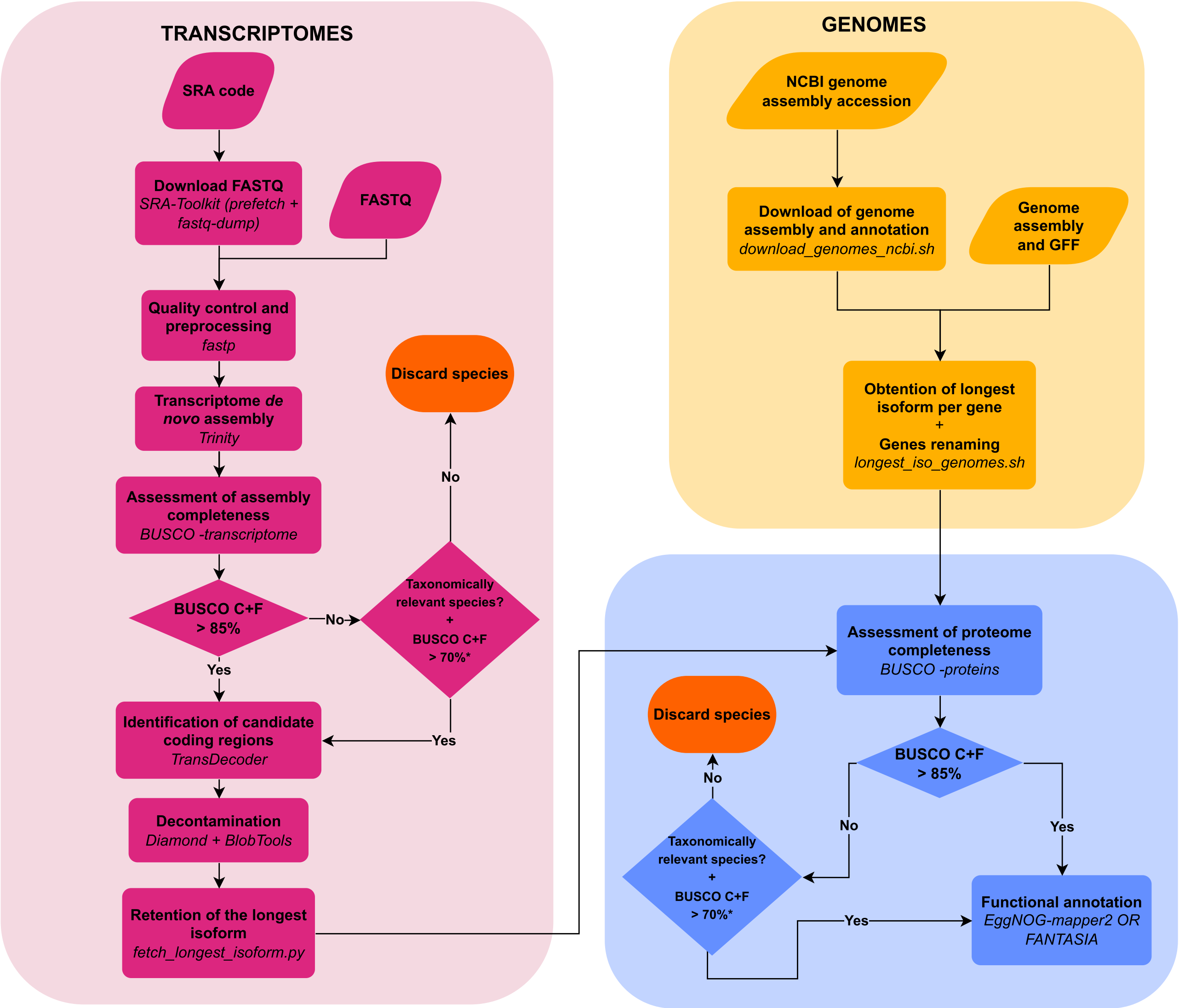
TRANSCRIPTOMES
SRA code
Download FASTQ
SRA-Toolkit (prefetch + fastq-dump)
FASTQ
Quality control and preprocessing
fastp
Transcriptome de novo assembly
Trinity
Assessment of assembly completeness
BUSCO -transcriptome
BUSCO C+F > 85%
No
Taxonomically relevant species? + BUSCO C+F > 70%*
No
Discard species
Yes
Yes
Identification of candidate coding regions
TransDecoder
Decontamination
Diamond + BlobTools
Retention of the longest isoform
fetch_longest_isoform.py
GENOMES
NCBI genome assembly accession
Download of genome assembly and annotation
download_genomes_ncbi.sh
Genome assembly and GFF
Obtention of longest isoform per gene + Genes renaming
longest_iso_genomes.sh
Assessment of proteome completeness
BUSCO -proteins
BUSCO C+F > 85%
No
Yes
Taxonomically relevant species? + BUSCO C+F > 70%*
No
Discard species
Yes
Functional annotation
EggNOG-mapper2 OR FANTASIA

Fig. 2. Pipeline followed to generate the MATEdb2 database. All steps differing from the MATEdb original pipeline are discussed in detail in the main text.

couple of sequences instead of the whole proteome. Therefore, we have included in the newer version of MATEdb a standardized pipeline for obtaining the CDS and protein files using directly the FASTA and GFF files of the corresponding genome (Fig. 2, in yellow).

Assessing Gene Completeness of Proteomes

Gene completeness was assessed using BUSCO in protein mode against the metazoa_odb10 reference set (except for the outgroup species, where eukaryota_odb10 was used). More than 75% of our species passed the threshold of 85% complete plus fragmented used in MATEdb (Fernández et al. 2022). The remaining 25% includes almost all representatives of tardigrades, annelids, nematodes, acoels, and some representatives of other phyla (see supplementary table S1, Supplementary Material online). As we want to maximize the taxon representation of animal lineages while keeping data sets of high quality, we lowered the threshold value to 70% in these cases, a value previously used in other studies. These values may represent biological features of the genomes of these lineages (Barreira et al. 2021) or just a lack of representation of the lineage in the BUSCO reference data sets. As an exception, after this new threshold, 8 animal and 2 outgroup transcriptome assemblies have been included with a slightly lower BUSCO score due to their taxonomic relevance (e.g. they were one of the only two representatives of their lineage, such as in the case of the Priapulida). A list of discarded data sets can also be found in supplementary table S2, Supplementary Material online.

Functional Annotation of the Gene Repertoire

In MATEdb (Fernández et al. 2022), ∼30% of all proteins remained unannotated when using traditional homology-based functional annotation. This percentage increases to ∼50% in this second version, which includes a broader representation of animal taxa. Phyla that are better annotated in the public databases (i.e. Craniata) have a higher proportion of annotated protein-coding genes when using homology-based methods. To overcome this limitation, we annotated the longest isoform gene list for each data set using the FANTASIA pipeline (https://github.com/MetazoaPhylogenomicsLab/FANTASIA). FANTASIA is a pipeline that allows the functional annotation of whole proteomes using GOPredSim (Littmann et al. 2021), a protein language-based method that transfers GO terms based on embedding similarity. In brief, embeddings are vectorized representations of protein sequences generated using protein language models, such as ProtT5 (Elnaggar et al. 2022), that consider protein sequences as sentences and apply natural language processing tools to extract information from them.

With FANTASIA, we address the homology-based taxonomic bias in functional annotation by assigning putative GO terms to virtually all proteins in the data set, regardless of their taxonomic group. Besides the GO terms predicted by FANTASIA, we also provide the raw per-protein ProtT5 embeddings, the same model that is being integrated into UniProt. While this approach offers advantages over traditional methods (Barrios-Núñez et al. 2024; Martínez-Redondo et al. 2024), it remains uncertain if any bias exists in the results produced by the protein language model depending on the data set it was trained on, and more research is needed to clarify this. More details about the pipeline, the method or the benchmarking and comparison with deep learning, homology-based methods, and another protein language model can be checked elsewhere (Barrios-Núñez et al. 2024; Martínez-Redondo et al. 2024).

Discussion

We presented here the second version of MATEdb (MATEdb2), with almost 1,000 animal species data. This newer version overcomes some of the previous restrictions of our database, including the restricted taxonomic representation of only arthropods and mollusks, and the use of previously preprocessed peptide and CDS files for genomes. Nevertheless, it is not devoid of limitations. The main limitation of this newer version is still the genome annotations. Even though we use an alternative approach that considers the incongruences found in some of the publicly preprocessed files, there may still be biases between the proteomes. These biases are caused by the heterogeneity of genome annotation methodologies, which can affect downstream analyses, such as ortholog inference (Weisman et al. 2022). These biases are typically ignored by phylogenomic studies that use publicly available preprocessed files. Nevertheless, correcting for this limitation by reannotating the genomes using the same methodology is computationally expensive and is still biased toward species where additional data that improves this annotation (e.g. RNA-seq) are available.

Materials and Methods

The analytical pipeline of MATEdb2 is shown in Fig. 2. In brief, the differences with the pipeline depicted in MATEdb (Fernández et al. 2022) are the following: (i) we included a standardized pipeline for obtaining the longest isoform from genomes (see below); (ii) for a few exceptions, we lowered the threshold used to consider a data set as high quality to 70% C + F (complete plus fragmented) BUSCO score (Manni et al. 2021), as the original 85% threshold was too restrictive when prioritizing a wide taxonomic sampling and the inclusion of biologically interesting species that are not widely studied; and (iii) we added per-gene functional annotation based on protein language embedding similarity.

Further details about the pipeline are shown in Fig. 2 and on the GitHub repository.

Compilation of Genomic Data

Genome assembly (FASTA) and annotation (GFF) files for each species were downloaded through NCBI Datasets (Sayers et al. 2023) or from the direct URL download link for genomes available in other repositories. The database source for each species is referenced in supplementary table S1, Supplementary Material online, while the bash script "download_genomes.sh" used to automatize the downloading of several files is included in the GitHub repository and the Singularity container (see Data Availability).

Once downloaded, we used AGAT (Dainat et al. 2023) to obtain the GFF containing only the longest isoforms which was then used to get the FASTA file with the longest protein sequence for each gene (and its corresponding CDS). In addition, we renamed the sequences to match the structure used in the transcriptomic part of the MATEdb2 pipeline and obtained a conversion file to keep track of the original names. These steps were performed using a custom bash script "longest_iso_genomes.sh," also included in the GitHub repository and container.

Functional Annotation of the Gene Repertoire

The longest isoform gene list for each data set was annotated with the homology-based software eggNOG-mapper v2 (Cantalapiedra et al. 2021) and the FANTASIA pipeline (https://github.com/MetazoaPhylogenomicsLab/FANTASIA).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Centro de Supercomputación de Galicia (CESGA) and the HPC Drago from the Centro Superior de Investigaciones Científicas for access to computer resources.

Author Contributions

This database results from the collaborative effort of lab members from the Metazoa Phylogenomics Lab to offer the scientific community the possibility to reuse some of the data generated for their projects. G.I.M.-R., C.V.-C., K.E., L.B.-Á., and M.V.-V. contributed assemblies to the data repository. G.I.M.-R. created the pipeline custom scripts for the genome data analyses and designed the MATEdb logo. K.E. created the Singularity container. C.V.-C. and R.F. contributed to the creation and management of the database. C.V.-C. created and curated the Github repository. R.F. provided resources and supervised the project. G.I.M.-R. wrote the first version of the manuscript. All authors revised and approved the final version of the manuscript.

Funding

G.I.M.-R. acknowledges the support of Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya and European Social Fund (ESF) "Investing in your future" (grant 2021 FI_B 00476). R.F. acknowledges support from the following sources of funding: Ramón y Cajal fellowship (grant agreement no. RYC2017-22492 funded by MCIN/AEI/10.13039/501100011033 and European Social Fund (ESF) "Investing in your future"), the Agencia Estatal de Investigación (project PID2019-108824GA-I00 funded by MCIN/AEI/10.13039/501100011033), the European Research Council (this project has received funding from the European Research Council [ERC] under the European's Union's Horizon 2020 research and innovation program [grant agreement no. 948281]), the Human Frontier Science Program (grant no. RGY0056/2022), and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (AGAUR 2021-SGR00420).

Data Availability

Scripts and commands in the pipeline and the Supplementary Material online (supplementary tables S1 and S2 and fig. S1, Supplementary Material online) can be found in the following repository: https://github.com/MetazoaPhylogenomicsLab/MATEdb2. For transcriptomes, the data repository contains (i) de novo transcriptome assemblies; (ii) their candidate coding regions within transcripts (both at the level of nucleotide and amino acid sequences); (iii) the coding regions filtered using their contamination profile (i.e. only metazoan content or eukaryote for outgroups); (iv) the longest isoforms of the amino acid candidate coding regions; (v) the gene content completeness score as assessed against the BUSCO reference sets; and (vi) orthology and protein language-based gene annotations, and per-protein ProtT5 embeddings. In the case of genomes, only files (iv), (v), and (vi) are provided in MATEdb2, together with a filtered version of file (iii) with just the longest CDS per gene. The database is hosted on our own server and will be there indefinitely. The database will be expanded as we incorporate new data sets from underrepresented lineages, such as nematodes, or as requested to be incorporated by the scientific community if resources allow it. Links for downloading can be found in the following file in the GitHub repository: https://github.com/MetazoaPhylogenomicsLab/MATEdb2/blob/main/linksforMATEdb2.txt. We provide a Singularity container for easy implementation of the tools used to generate the files in the database with the appropriate software versions along with their dependencies (https://cloud.sylabs.io/library/klarael.metazomics/matedb2/matedb2.sif). The software included are the following: SRA Toolkit v2.10.7 (https://github.com/ncbi/sra–tools); fastp v0.20.1 (https://github.com/OpenGene/fastp; Chen et al. 2018); Trinity v2.11.0 (Grabherr et al. 2011); BUSCO v5.3.2 (Manni et al. 2021); TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder); Diamond v2.0.8 (Buchfink et al. 2015); BlobTools v2.3.3 (Challis et al. 2020); NCBI Datasets v13.42.0; eggNOG-mapper v2.1.9 (Cantalapiedra et al. 2021); seqkit v2.1.0 (Shen et al. 2024); AGAT v0.9.1 (Dainat et al. 2023); as well as some custom scripts.

Literature Cited

Dainat J, Hereñú D, Murray KD, Davis E, Crouch K, LucileSol, Agostinho N, Pascal-Git, Zollman Z, Tayyrov. NBISweden/AGAT: AGAT-v1.2.0 (v1.2.0). Zenodo. 2023

Barreira SN, Nguyen AD, Fredriksen MT, Wolfsberg TG, Moreland RT, Baxevanis AD. AniProtDB: a collection of consistently generated metazoan proteomes for comparative genomics studies. Mol Biol Evol. 2021:38(10):4628–4633. https://doi.org/10.1093/molbev/msab165.

Barrios-Núñez I, Martínez-Redondo GI, Medina-Burgos P, Cases I, Fernández R, Rojas AM. Decoding functional proteome information in model organisms using protein language models. NAR Genom Bioinform. 2024:6(3):lqae078. https://doi.org/10.1093/nargab/lqae078.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015:12(1):59–60. https://doi.org/10.1038/nmeth.3176.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021:38(12):5825–5829. https://doi.org/10.1093/molbev/msab293.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. G3 (Bethesda). 2020:10(4):1361–1374. https://doi.org/10.1534/g3.119.400908.

Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018:34(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

De Oliveira AL, Wollesen T, Kristof A, Scherholz M, Redl E, Todt C, Bleidorn C, Wanninger A. Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. BMC Genomics. 2016:17(1):905. https://doi.org/10.1186/s12864-016-3080-9.

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022:44(10):7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.

Erséus C, Williams BW, Horn KM, Halanych KM, Santos SR, James SW, Creuzé des Châtelliers M, Anderson FE. Phylogenomic analyses reveal a palaeozoic radiation and support a freshwater origin for clitellate annelids. Zool Scr. 2020:49(5):614–640. https://doi.org/10.1111/zsc.12426.

Fernández R, Gabaldón T. Gene gain and loss across the metazoan tree of life. Nat Ecol Evol. 2020:4(4):524–533. https://doi.org/10.1038/s41559-019-1069-x.

Fernández R, Tonzo V, Simón Guerrero C, Lozano-Fernandez J, Martínez-Redondo GI, Balart-García P, Aristide L, Eleftheriadi K, Vargas-Chávez C. MATEdb, a data repository of high-quality metazoan transcriptome assemblies to accelerate phylogenomic studies. Peer Community J. 2022:2. Article e58. https://doi.org/10.24072/pcjournal.177.

Giribet G, Edgecombe GD. Current understanding of Ecdysozoa and its internal phylogenetic relationships. Integr Comp Biol. 2017:57(3):455–466. https://doi.org/10.1093/icb/icx072.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol. 2011:29(7):644–652. https://doi.org/10.1038/nbt.1883.

Khalturin K, Shunatova N, Shchenkov S, Sasakura Y, Kawamitsu M, Satoh N. Polyzoa is back: the effect of complete gene sets on the placement of Ectoprocta and Entoprocta. Sci Adv. 2022:8(26):eabo4400. https://doi.org/10.1126/sciadv.abo4400.

Krug PJ, Caplins SA, Algoso K, Thomas K, Valdés ÁA, Wade R, Wong NLWS, Eernisse DJ, Kocot KM. Phylogenomic resolution of the root of Panpulmonata, a hyperdiverse radiation of gastropods: new insight into the evolution of air breathing. Proc Biol Sci. 2022:289(1972):20211855. https://doi.org/10.1098/rspb.2021.1855.

Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Riesgo A, Andrade SCS, Sterrer W, Sørensen MV, Giribet G. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc Biol Sci. 2019:286(1906):20190831. https://doi.org/10.1098/rspb.2019.0831.

Littmann M, Heinzinger M, Dallago C, Olenyi T, Rost B. Embeddings from deep learning transfer GO annotations beyond homology. Sci Rep. 2021:11(1):1160. https://doi.org/10.1038/s41598-020-80786-0.

Lu TM, Kanda M, Satoh N, Furuya H. The phylogenetic position of dicyemid mesozoans offers insights into spiralian evolution. Zoological Lett. 2017:3:6. https://doi.org/10.1186/s40851-017-0068-5.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 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(10):4647–4654. https://doi.org/10.1093/molbev/msab199.

Marlétaz F, Peijnenburg KTCA, Goto T, Satoh N, Rokhsar DS. A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. Curr Biol. 2019:29(2):312–318.e3. https://doi.org/10.1016/j.cub.2018.11.042.

Martínez-Redondo GI, Barrios-Núñez I, Vázquez-Valls M, Rojas AM, Fernández R. Illuminating the functional landscape of the dark proteome across the animal tree of life through natural language processing models. bioRxiv 582465. https://doi.org/10.1101/2024.02.28.582465, May 29, 2024, preprint: not peer reviewed.

Mongiardino Koch N, Coppard SE, Lessios HA, Briggs DE, Mooi R, Rouse GW. A phylogenomic resolution of the sea urchin tree of life. BMC Evolutionary Biology. 2018:18(1). https://doi.org/10.1186/s12862-018-1300-4.

Osmanski AB, Paulat NS, Korstian J, Grimshaw JR, Halsey M, Sullivan KAM, Moreno-Santillán DD, Crookshanks C, Roberts J, Garcia C, et al. Insights into mammalian TE diversity through the curation of 248 genome assemblies. Science. 2023:380(6643):eabn1430. https://doi.org/10.1126/science.abn1430.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Farrell CM, Feldgarden M, Fine AM, Funk K, et al. Database resources of the national center for biotechnology information in 2023. Nucleic Acids Res. 2023:51(D1):D29–D38. https://doi.org/10.1093/nar/gkac1032.

Shen W, Sipos B, Zhao L. SeqKit2: a Swiss army knife for sequence and alignment processing. Imeta. 2024:3(3):e191. https://doi.org/10.1002/imt2.191.

Song H, Wang Y, Shao H, Li Z, Hu P, Yap-Chiongco MK, Shi P, Zhang T, Li C, Wang Y, et al. Scaphopoda is the sister taxon to Bivalvia: evidence of ancient incomplete lineage sorting. Proc Natl Acad Sci U S A. 2023:120(40):e2302361120. https://doi.org/10.1073/pnas.2302361120.

Thoma M, Missbach C, Jordan MD, Grosse-Wilde E, Newcomb RD, Hansson BS. Transcriptome surveys in silverfish suggest a multistep origin of the insect odorant receptor gene family. Front Ecol Evol. 2019:7:281. https://doi.org/10.3389/fevo.2019.00281.

UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023:51(D1):D523–D531. https://doi.org/10.1093/nar/gkac1052.

Weisman CM, Murray AW, Eddy SR. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. Curr Biol. 2022:32(12):2632–2639.e2. https://doi.org/10.1016/j.cub.2022.04.085.

Zapata F, Wilson NG, Howison M, Andrade SC, Jörger KM, Schrödl M, Goetz FE, Giribet G, Dunn CW. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. Proc Biol Sci. 2014:281(1794):20141739. https://doi.org/10.1098/rspb.2014.1739.

Associate editor: Christopher Wheat