Title: **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 (Spanish Research Council-University Pompeu Fabra). Passeig marítim de la Barceloneta 37-49, 08003 Barcelona (Spain)

[*] Corresponding authors: gemma.martinez@ibe.upf-csic.es, rosa.fernandez@ibe.upf-csic.es

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: (1) we include data from all animal phyla where public data is available, (2) we provide gene annotations from genomes obtained using the same pipeline. In total, we provide proteomes inferred from high-quality transcriptomic or genomic data for almost 1000 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.

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 non-coding 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 (Erséus et al., 2020; Mongiardino Koch et al., 2018; Zapata et al., 2014, among others) and gene repertoire evolution (De Oliveira et al., 2016; Fernández & Gabaldón, 2020; Thoma et al., 2019) 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 datasets 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. For instance, the transcriptome assembly methodology used can impact the comparability among different datasets (e.g. in our experience the number of 'genes' inferred with Trinity may vary up to one order of magnitude depending on the version), while the 'ready-to-use' protein files provided in some genome sequencing projects cannot be easily matched with the other genome files for additional analyses due to different nomenclature across files. 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: (1) 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; (2) we include a standardized pipeline for obtaining the protein sequences from genomes instead of adding the precomputed files available, making it easier to replicate; (3) we provide the functional annotation of all proteins using a 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.

Material and Methods

Increased taxonomic coverage

The first version of MATEdb (Fernández et al., 2022) included high-quality datasets 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 is represented in Figure 1, and the complete list of species and their metadata is included in Table S1.

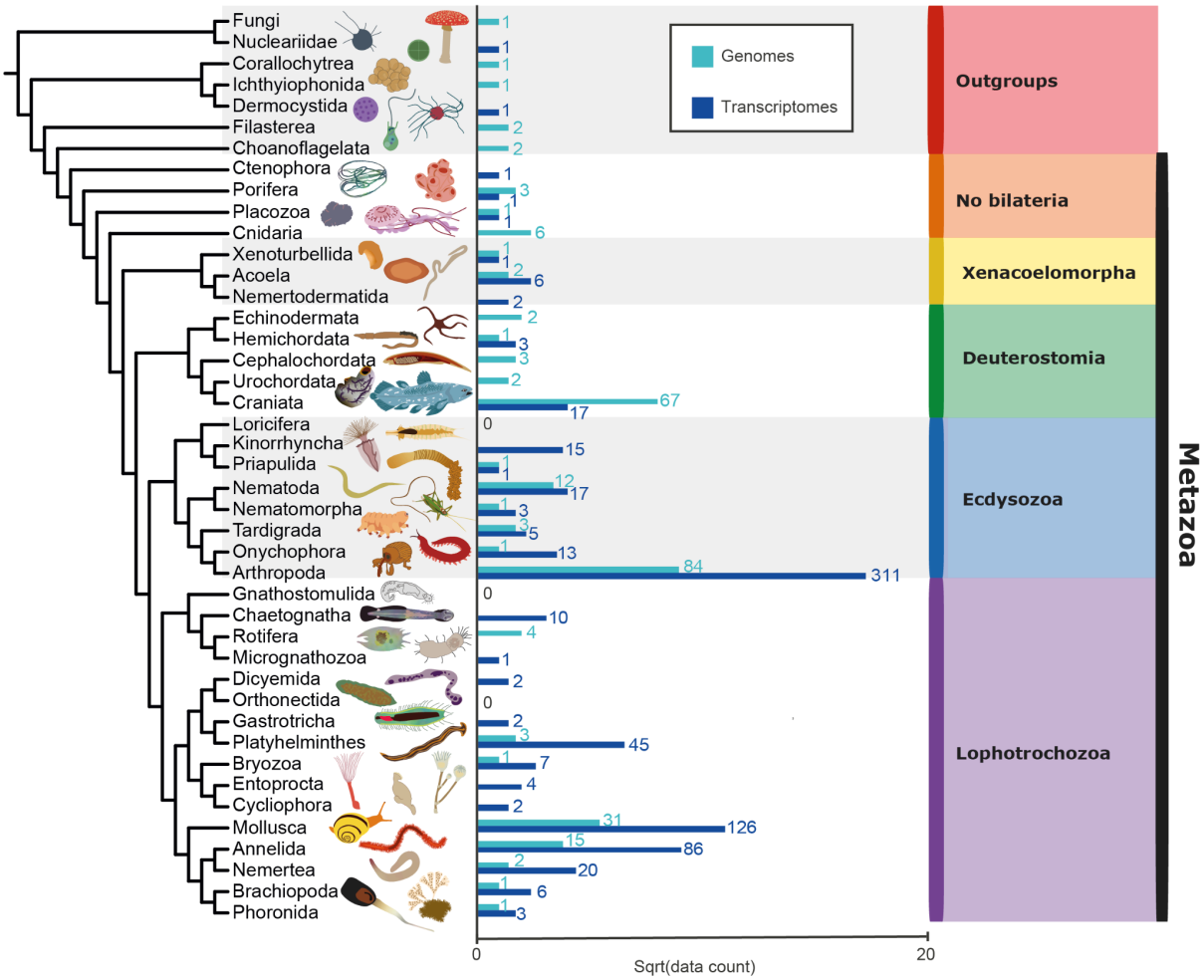

Figure 1. Taxonomic representation of species included in MATEdb2. The number of datasets per phylum is separated by the type of data (genomes and transcriptomes in light and dark blue, respectively).

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. 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 file, hindering the utili of those files for additional analyses. 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

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.

The analytical pipeline of MATEdb2 is shown in Figure 2. In brief, the differences with the pipeline depicted in MATEdb (Fernández et al., 2022) are the following: (1) we included a standardized pipeline for obtaining the longest isoform from genomes; (2) for a few exceptions, we lowered the threshold used to consider a dataset 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. Further details about the pipeline are shown in Figure 2.

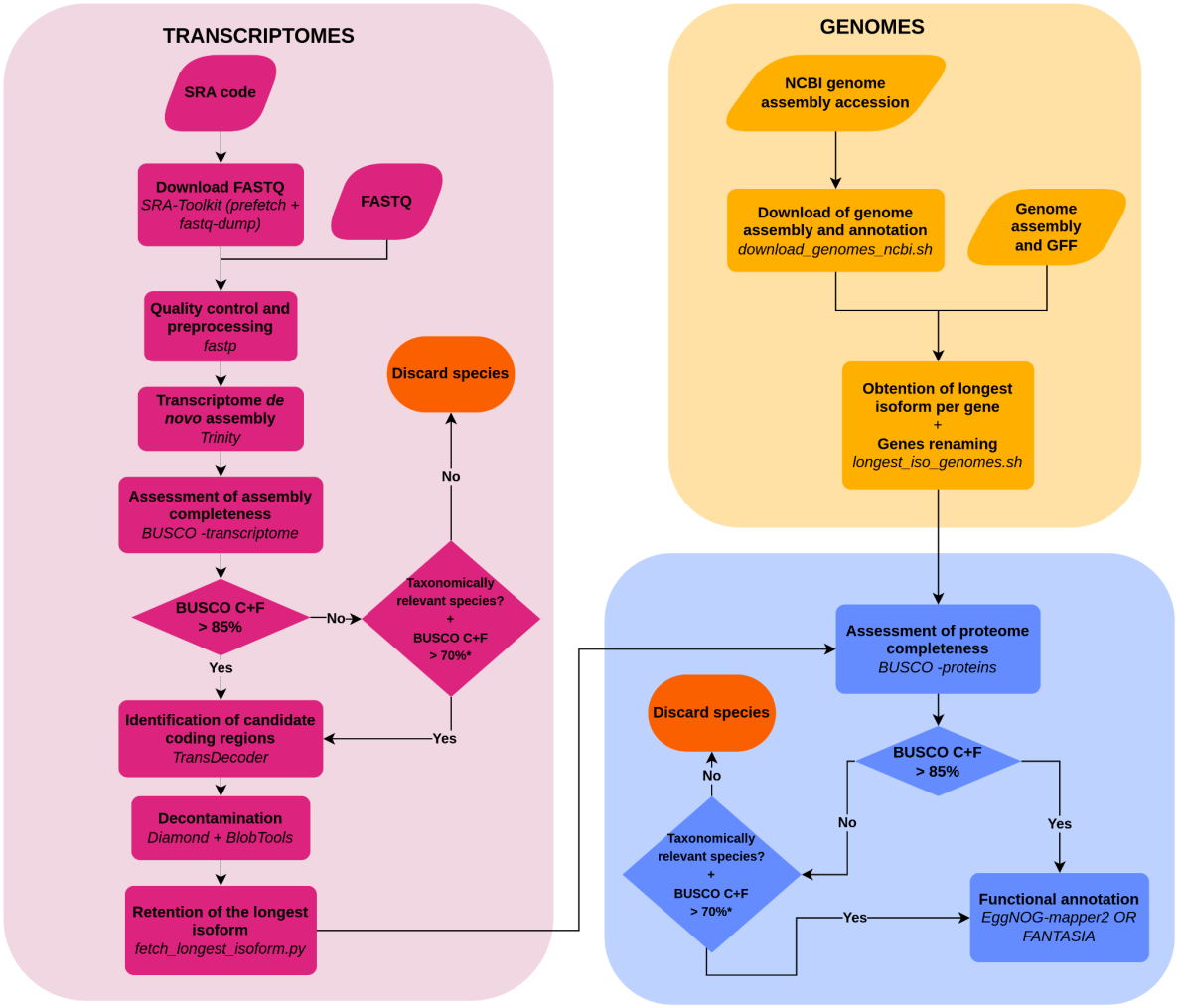

Figure 2. Pipeline followed to generate the MATEdb2 database. All steps differing from the MATEdb original pipeline are discussed in detail in the main text.

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 the Supplementary Material, Table S1, 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 (Creators Jacques Dainat1 Darío Hereñú Dr. K. D. Murray2 Ed Davis3 Kathryn Crouch4 LucileSol Nuno Agostinho5 pascal-git Zachary Zollman tayyrov Show affiliations 1. IRD 2. Max Planck Institut für Entwicklungsbiologie, Tübingen 3. Oregon State University 4. @VEuPathDB 5. European Bioinformatics Institute | EMBL-EBI, n.d.) 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.

Finally, 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 Table S1). As we want to maximize the taxon representation of animal lineages while keeping datasets of high quality, we lowered the threshold value to 70% in these cases, a value that has been previously used in other studies, as these values may indeed represent biological features of the genomes of these lineages (Barreira et al., 2021). 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 the only representatives of their lineage, such as in the case of the hagfish).

Functional annotation of the gene repertoire

The longest isoform gene list for each dataset was annotated with the homology-based software eggNOG-mapper v2 (Cantalapiedra et al., 2021) and the

FANTASIA pipeline (https://github.com/MetazoaPhylogenomicsLab/FANTASIA). FANTASIA is a pipeline that allows the 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. Here, besides the GO terms predicted by FANTASIA, we provide the raw per-protein ProtT5 embeddings. More details about the pipeline, the method or the benchmarking and comparison with homology-based methods can be checked in (Barrios-Núñez et al., 2024) and Martínez-Redondo et al. 2024.

Database availability

Scripts and commands

Scripts and commands in the pipeline and the supplementary Table S1 can be found in the following repository: https://github.com/MetazoaPhylogenomicsLab/MATEdb2

Files deposited in the repository

For transcriptomes, the data repository contains (1) de novo transcriptome assemblies, (2) their candidate coding regions within transcripts (both at the level of nucleotide and amino acid sequences), (3) the coding regions filtered using their contamination profile (ie, only metazoan content or eukaryote for outgroups), (4) the longest isoforms of the amino acid candidate coding regions, (5) the gene content completeness score as assessed against the BUSCO reference sets, and (6) orthology and protein language-based gene annotations, and per-protein ProtT5 embeddings. In the case of genomes, only files (4), (5), and (6) are provided in MATEdb2, together with a filtered version of the file (3) with just the longest CDS per gene.

Software availability

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 is the following: SRA Toolkit v2.10.7, fastp v0.20.1, Trinity v2.11.0, BUSCO v5.3.2,

TransDecoder v5.5.0, Diamond v2.0.8, BlobTools v2.3.3, NCBI datasets v13.42.0, eggNOG-mapper v2.1.9, seqkit v2.1.0, AGAT v0.9.1, as well as some custom scripts.

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. GIMR, CVC, LBA, MVV, and KE contributed assemblies to the data repository. GIMR created the pipeline custom scripts for the genome data analyses and designed the MATEdb logo. KE created the Singularity container. CVC and RF contributed to the creation and management of the database. CVC created and curated the Github repository. RF provided resources and supervised the project. GIMR wrote the first version of the manuscript. All authors revised and approved the final version of the manuscript.

Acknowledgments

GIMR acknowledges the support of Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya and ESF Investing in your future (grant 2021 FI_B 00476). RF 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 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 programme (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). We also thank Centro de Supercomputación de Galicia (CESGA) and the HPC Drago from the Centro Superior de Investigaciones Científicas for access to computer resources.

References

Barreira, S. N., Nguyen, A.-D., Fredriksen, M. T., Wolfsberg, T. G., Moreland, R. T., &

Baxevanis, A. D. (2021). AniProtDB: A Collection of Consistently Generated Metazoan Proteomes for Comparative Genomics Studies. *Molecular Biology and Evolution*, *38*(10), 4628–4633. https://doi.org/10.1093/molbev/msab165

Barrios-Núñez, I., Martínez-Redondo, G. I., Medina-Burgos, P., Cases, I., Fernández, R., & Rojas, A. M. (2024). Decoding proteome functional information in model organisms using protein language models. In *bioRxiv* (p. 2024.02.14.580341). https://doi.org/10.1101/2024.02.14.580341

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Creators Jacques Dainat1 Darío Hereñú Dr. K. D. Murray2 Ed Davis3 Kathryn Crouch4 LucileSol Nuno Agostinho5 pascal-git Zachary Zollman tayyrov Show affiliations 1. IRD 2. Max Planck Institut für Entwicklungsbiologie, Tübingen 3. Oregon State University 4. @VEuPathDB 5. European Bioinformatics Institute | EMBL-EBI. (n.d.). *NBISweden/AGAT: AGAT-v1.2.0*. https://doi.org/10.5281/zenodo.8178877

De Oliveira, A. L., Wollesen, T., Kristof, A., Scherholz, M., Redl, E., Todt, C., Bleidorn, C., & Wanninger, A. (2016). Comparative transcriptomics enlarges the toolkit of known developmental genes in mollusks. *BMC Genomics*, *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., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

Erséus, C., Williams, B. W., Horn, K. M., Halanych, K. M., Santos, S. R., James, S. W., Creuzé des Châtelliers, M., & Anderson, F. E. (2020). Phylogenomic analyses reveal a Palaeozoic radiation and support a freshwater origin for clitellate annelids. *Zoologica*

Scripta, *49*(5), 614–640. https://doi.org/10.1111/zsc.12426

Fernández, R., & Gabaldón, T. (2020). Gene gain and loss across the metazoan tree of life. *Nature Ecology & Evolution*, *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, G. I., Balart-García, P., Aristide, L., Eleftheriadi, K., & Vargas-Chávez, C. (2022). MATEdb, a data repository of high-quality metazoan transcriptome assemblies to accelerate phylogenomic studies. *Peer Community Journal*, *2*(e58). https://doi.org/10.24072/pcjournal.177

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

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Martínez-Redondo, G.I., Barrios-Núñez, I., Vázquez-Valls, M., Rojas, A.M., & Fernández, R. (2024). Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models.

Osmanski, A. B., Paulat, N. S., Korstian, J., Grimshaw, J. R., Halsey, M., Sullivan, K. A. M., Moreno-Santillán, D. D., Crookshanks, C., Roberts, J., Garcia, C., Johnson, M. G., Densmore, L. D., Stevens, R. D., Zoonomia Consortium†, Rosen, J., Storer, J. M., Hubley, R., Smit, A. F. A., Dávalos, L. M., … Ray, D. A. (2023). Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science*, *380*(6643), eabn1430. https://doi.org/10.1126/science.abn1430

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Farrell, C. M., Feldgarden, M., Fine, A. M., Funk, K., Hatcher, E., Kannan, S., Kelly, C., Kim, S., Klimke, W., Landrum, M. J., Lathrop, S., Lu, Z., Madden, T. L., … Sherry, S. T. (2023). Database resources of the National Center for Biotechnology Information in 2023.

Nucleic Acids Research, *51*(D1), D29–D38. https://doi.org/10.1093/nar/gkac1032

Song, H., Wang, Y., Shao, H., Li, Z., Hu, P., Yap-Chiongco, M. K., Shi, P., Zhang, T., Li, C., Wang, Y., Ma, P., Vinther, J., Wang, H., & Kocot, K. M. (2023). Scaphopoda is the sister taxon to Bivalvia: Evidence of ancient incomplete lineage sorting. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(40), e2302361120. https://doi.org/10.1073/pnas.2302361120

Thoma, M., Missbach, C., Jordan, M. D., Grosse-Wilde, E., Newcomb, R. D., & Hansson, B. S. (2019). Transcriptome surveys in silverfish suggest a multistep origin of the insect odorant receptor gene family. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00281

UniProt Consortium. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Zapata, F., Wilson, N. G., Howison, M., Andrade, S. C. S., Jörger, K. M., Schrödl, M., Goetz, F. E., Giribet, G., & Dunn, C. W. (2014). Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. *Proceedings. Biological Sciences / The Royal Society*, *281*(1794), 20141739. https://doi.org/10.1098/rspb.2014.1739