Title: Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models

Gemma I. Martínez-Redondo[1*], Israel Barrios-Núñez[2], Marçal Vázquez-Valls[1], Ana M. Rojas[2†*] & Rosa Fernández[1†*]

[1] Metazoa Phylogenomics Lab, Institute of Evolutionary Biology (CSIC-UPF), Passeig marítim de la Barceloneta 37-49, 08003 Barcelona, Spain.

[2] Computational Biology and Bioinformatics, Andalusian Center for Developmental Biology (CABD-CSIC), 41013 Sevilla, Spain.

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

[†] Contributed equally

Abstract

Understanding how coding genes and their functions evolve over time is a key aspect of evolutionary biology. Protein coding genes poorly understood or characterized at the functional level may be related to important evolutionary innovations, potentially leading to incomplete or inaccurate models of evolutionary change, and limiting the ability to identify conserved or lineage-specific features. Homology-based methodologies often fail to transfer functional annotations in a large fraction of the coding gene repertoire in non-model organisms. This is particularly relevant in animals, where a high number of their coding genes yield no functional annotation. Here, we leverage machine learning and natural language processing models to investigate functional annotation in the 'dark proteome' (defined as the unknown functional landscape') of ca. 1,000 gene repertoires of virtually all animal phyla, totaling ca. 23.2 million coding genes. Gene ontology annotations were transferred to virtually all genes, with the model ProtT5 outperforming both homology-based and other machine learning-based models. We then explored the 'dark proteome' of all animal phyla revealing an enrichment in functions related to immune response, viral infection, response to stimuli, development, or signaling, among others. Furthermore, we provide an open-access pipeline - FANTASIA - to implement and benchmark these methodologies in any dataset. Our results uncover the putative functions of poorly understood protein-coding genes across the Animal Tree of Life and contribute to a more comprehensive understanding of the molecular basis of animal evolution.

Introduction

We cannot fully understand the evolution and complexity of life without understanding the function of its components. Assigning a function to proteins is vital to study gene and genome evolution. Common methods for protein functional annotation include metabolic signaling pathway analysis, protein domain analysis and Gene Ontology (GO) analysis. Right after the first genomes were sequenced, the GO Consortium created a controlled common vocabulary -ontologies- to describe the function of gene products in any organism in a comparative manner[1]. These ontologies are directed acyclic graphs (DAGs) of GO terms grouped into 3 categories: biological process, molecular function, and cellular component. Thus, when we talk about functional annotation, we are referring to the assignment of a function -in the form of GO terms- to a gene (or, more properly, the protein and non-coding RNA molecules encoded by a gene), a gene family, or the whole gene repertoire of a species (its proteome). This is a key aspect of evolutionary biology that allows us to investigate the functional landscape of genomes and proteomes.

One widespread problem in non-model organisms is the need for a complete functional annotation of the gene repertoire, as a high number of genes remain unannotated and therefore 'hidden' from our sight (what we call here the 'dark proteome'). The underlying factor causing this lack of functional annotation is the lack of homology since traditional methods, such as eggNOG-mapper[2] or PANTHER[3] heavily rely on sequence homology for transference of function, constraining the results to sequences conserved to some degree. This issue is at the heart of our need for progress in functional annotation, since there is a great number of orphan proteins that are no longer identifiable by homology-based methods due to their particular amino acid arrangements (e.g. intrinsically disordered proteins) or evolutionary history (e.g. encoded by fast-evolving genes or originated *de novo* in a specific lineage or species). Some illustrative examples in animals are the case of water bears (phylum Tardigrada), with 41.1% of the genes in their genomes with no homology to any genes in the non-redundant database of the National Centre for Biotechnology Information[4], or the freshwater sponge *Ephydatia muelleri*, with 50% of its ca. 20k genes without homology to any other known genes with functional annotations in the same database[5]. This highlights the imperative necessity of novel methodologies able to unlock the 'hidden biology' behind these genes.

With the advent of initiatives such as the Earth Biogenome Project[6] or the European Reference Genome Atlas[7,8], where new genomes from non-model organisms are being sequenced and released daily, we need to leverage faster and scalable sequence-based functional prediction methods. In this line, embracing concepts from orthogonal disciplines, such as computer science, could alleviate the problem. This has been the case for 3D protein structure prediction by AlphaFold[9], a paradigm-shifting case of the use of artificial intelligence (AI) to address the protein structure prediction problem, a long-standing problem in biology. Functional annotation prediction methods have also suffered considerable development thanks to AI in the past few years. AI methods implemented range from deep learning[10] to natural language processing (NPL) tools that extract rules from the language of protein sequences in the form of vectorized embeddings (protein language models[11,12]). In addition, some methods have been created to predict the function of specific proteins, such as enzymes[13] or prokaryotic viral proteins[14], while others have been successfully used to predict non-GO term-based functions in a wider context[15,16].

Here, we addressed the issue of the lack of annotation of genes across metazoans by applying newly developed AI-based functional annotation methodologies to a dataset composed of 970 gene repertoires (i.e. proteomes) inferred from high-quality genomes and transcriptomes from virtually all animal phyla, totaling ca. 23.5 million genes. To do so, we compared the performance of 3 functional annotation methods: one based on homology[2], one based on deep learning and sequence similarity[10], and one based on protein language models[12] (with two models tested), using the parameters previously benchmarked in controlled functional datasets and empirically tested in model organisms by our team[17]. We show that protein language model-based annotations outperformed homology- and deep learning-based ones, and discuss the implications of our findings in furthering our understanding of the functional landscape of the 'dark proteome' across the Animal Tree of Life. Furthermore, we developed an open-access pipeline readily available to the scientific community (FANTASIA, Functional ANnoTAtion based on embedding space SImilArity) to functionally annotate any dataset following the best performing method, as discussed below. These methodologies set the road towards a deeper understanding of the functional landscape of the animal kingdom at an unprecedented level.

Results

GOPredSim with the protein language model ProtT5 annotates virtually all genes in all gene repertoires with high confidence

We functionally annotated 970 gene repertoires from high-quality genomes and transcriptomes from species representing virtually all animal phyla, totaling 23,184,398 genes (Fig. 1a). GO terms were predicted using 3 different methods: (1) eggNOG-mapper[2], based on function transfer from orthologous genes annotated in the eggNOG database; (2) DeepGOPlus[10], a deep-learning model combined with sequence similarity data; and (3) GOPredSim[12] that uses annotation transfer based on sequence embedding similarity with embeddings obtained using NLP-based models. Here, we tested and compared the protein language models ProtT5[11] and SeqVec[18]. The results of the proportion of annotated genes by each program are shown in Figure 1b (dark purple), with eggNOG-mapper failing to annotate almost half of all the genes in our dataset, unlike the other 3 methods. EggNOG-mapper annotates a higher proportion of genes in the phyla that are better annotated in the public databases (i.e. Craniata, Fig. 1a), while this bias is not seen in the other methods. The number of annotated genes after filtering the most general -i.e. not informative- GO terms (see Methods) is only slightly reduced (Figure 1b, light purple).

DeepGOPlus and GOPredSim methods estimate the reliability index (RI) for each GO term, which measures how much you can trust the result based on the Euclidean distance between the query embeddings and the target proteins in the lookup reference dataset (i.e. dataset of proteins with known GO annotations) from which GO terms are transferred. This reliability index equals 1 when the exact protein sequence is included in the lookup reference dataset. Even though this threshold is based on a small set of experimentally confirmed proteins from model organisms without considering a high diversity of organisms, we also applied this filter in our annotated proteomes to see how much information we lost (measured as the proportion of genes without GO annotation after filtering). In total, after this filter, the number of genes annotated by GOPredSim with both models was reduced, losing 11.32% for ProtT5 and 70.35% with SeqVec (Figure 1b, dark pink). Combining depth and RI-based filtering in DeepGOplus reduced the proportion of annotated genes by ca. 25% (Figure 1b, light pink). Per-phylum

proportion of annotated genes after filtering followed the same pattern as in eggNOG-mapper, with better-annotated phyla having a higher proportion (Supp. Figure 1a). Thus, in terms of the proportion of annotated genes, all 3 non-homology-based methods outperform eggNOG-mapper, with GOPredSim-ProtT5 being the one that annotates the highest number when applying the most restrictive filters.

Regarding the number of GO terms assigned per gene, DeepGOPlus, and eggNOG-mapper assigned overall higher numbers of GO terms, reaching maximum values higher than 4000 and 1200, respectively, for the biological process. Contrarily, the number of GO terms assigned to each gene by GOPredSim was less than 200 for the biological process. In all programs, the number of biological process GO terms was the highest of the three categories, followed by the molecular function and the cellular component, in this order. These patterns were constant regardless of the filtering method used, even though filtering by depth resulted in a slight reduction of the maximum and mean values (Supp. Figure 2).

To assess the quality of the non-homology-based annotations, we first compared the information content (IC) of their GO terms (see Methods). Our results showed that DeepGOPlus annotated genes with GO terms that were very general (i.e. less nformative) for all 3 GO terms categories (Fig 1c). Contrarily, eggNOG-mapper and GOPredSim annotations were more specific and informative in all categories, with GOPredSim values being slightly higher (Fig. 1c). There were no differences between both models of GOPredSim (SeqVec or ProtT5).

Next, we assessed how similar were the annotations of each of the methods against the eggNOG-mapper annotations, as eggNOG-mapper is considered the standard for functional annotation in non-model organisms. Our results showed that the annotations of all methods varied to some extent compared to those of eggNOG-mapper. Annotations obtained for biological process were less similar than the ones obtained for molecular function and cellular component for all comparisons (Fig. 1d), which is consistent with the higher diversity of biological process GO terms[19]. In addition, GOPredSim-ProtT5 semantic similarity to eggNOG-mapper annotation was comparable to GOPredSim-SeqVec for all 3 GO categories (Fig. 1d, light and dark blue, respectively, Supp. Fig. 3). Contrarily, DeepGOPlus annotations were more similar to the ones from eggNOG-mapper than any of the two GOPredSim models (Fig. 1d, purple). This similarity can be partly explained by the lower specificity of GO terms

annotated by DeepGOPlus and eggNOG-mapper, that decreased as semantic similarity got higher (Supp. Fig. 4a-f), something not observed with GOPredSim-ProtT5 (Supp. Fig. 4g-l).

In summary, GOPredSim-ProtT5 outperforms the other methods in terms of the proportion of annotated genes and GO term specificity and informativeness, and therefore we consider it as the best-performing model for GO annotation across animal phyla. Moreover, its information content and semantic similarity did not differ among phyla (Supp. Fig 1b,c).

GOPredSim resulted in a high proportion of GO terms with a reliability index of 1 for species whose annotation was already present in the lookup dataset. Thus, we next tested the performance of GOPredSim-ProtT5 in those species (Supp. File 1) when removing them from the lookup dataset. We analyzed the 11 species that contained more than 100 sequences in the lookup dataset (Fig. 1e). Comparing the GOPredSim-ProtT5 annotation with and without these species in the lookup dataset revealed significant differences in reliability index distribution (Supp. Fig. 5a) for species with a high proportion of proteins in the lookup dataset. However, the resultant annotations showed no difference in information (Supp. Fig. 5b) or semantic similarity (Fig. 1e). Semantic similarity between both annotations was very high, with less than 25% of proteins having a semantic similarity lower than 0.7 for biological process (Fig. 1e). Surprisingly, removing the species from the lookup dataset sometimes increased its reliability index while preserving high semantic similarity (Supp. Fig. 5c-e).

For the following analyses, we decided to proceed with the unfiltered annotated set of GOs inferred by GOPredSim-ProtT5. This decision was based on 4 factors: (1) reliability index thresholds were estimated based on a very reduced dataset strongly biased towards model organisms, and their application may impact considerably the amount of information we retrieve for non-model species, (2) reducing the reliability index values in model organisms by removing those species from the lookup dataset did not change the functional annotation, (3) information content of GO terms predicted by GOPredSim-ProtT5 is independent of their reliability index[17], and (4) we identified genes with GO terms with low reliability index aligning well with the results of GOPredSim-ProtT5 and showing experimentally confirmed functions. Firstly, the human pseudogene LOFF_G0001716 (or AL627230.1), found to be differentially expressed in a bone mechanical response experiment[20], was inferred to have a role in wound healing after annotation with GOPredSim-ProtT5 (GO:0042060), a term related to angiogenesis, a prerequisite for bone formation. Secondly, the gene CG41561 in *D. melanogaster*, known to be

downregulated in cohesin knockdown mutants[21], exhibited GO terms assigned by GOPredSim-ProtT5 related to viral response and immunity (GO:0039554, GO:0030683, GO:0039548 and GO:1900226). This suggests a potential viral of transposable element origin for this gene, with putative involvement in chromatin remodeling regulation given its association with MDA-5, IRF3, and NLRP3 in humans[22–24]. We therefore considered this evidence as supportive for not filtering considering the reliability index previously suggested[12].

Annotation of the longest isoform per gene for each species recovers virtually all GO terms from all isoforms within that gene

In evolutionary studies of non-model organisms, it is common practice to use the longest isoform of each gene as a proxy for the whole gene, particularly in phylogenomic studies aiming at species tree reconstruction or gene repertoire evolution, to reduce the redundancy of the sequences included in the dataset and have only one sequence representing each gene. This is commonly preferred to the use of the consensus sequence per gene as these are artificial constructs that do not have a real biological function. Still, it remains unclear whether ignoring shorter isoforms affects the prediction of gene function or not. To see how this annotation may change, we annotated the functions of a subset of 93 proteomes representing all animal phyla including all isoforms per gene and compared it with the annotation previously inferred using only the longest isoform. Our results show that GOPredSim-ProtT5, with and without filtering by reliability index, is the method where fewer differences are observed in the numbers of annotated genes when either considering all isoforms or just the longest, followed by eggNOG-mapper (Fig. 2a, first and third panel, respectively). Among all methods, DeepGOPlus performed the worst and unpredictably when applying the depth and reliability index filters, with a maximum of more than 25% genes annotated when using all isoforms and some species having more genes annotated using just the longest isoform than when including all isoforms (Fig. 2a, second panel). With the unfiltered datasets, all three methods DeepGOPlus, GOPredSim-ProtT5, and GOPredSim-SeqVec showed no differences in the number of genes annotated (Supp. Figure 6a-c).

We compared how similar the annotations were at the gene level when annotating only the longest isoform or all of them per gene. More than half of the genes had the same or very similar annotation for all 4 methods (Fig. 2b). The method that had the lowest similarity was GOPredSim-ProtT5, but it was still 0.5 or higher for 75% of the genes (Fig. 2b). In addition, the

patterns of information content were highly similar using either the longest (Fig. 1c) or all isoforms (Supp. Fig. 6d) when comparing the different methods. Overall, these results indicate that the annotations at the level of gene are very similar if we annotate only the longest isoform per gene or all of them for virtually all the methods tested, which will reduce computational efforts on one side and allow the exploration of proteomes reported in the literature that include only the longest isoform per gene on the other, as is common in phylogenomic studies of non-model organisms.

Revealing the 'hidden biology' in the dark proteome across the Animal Tree of Life

What functional information have we been unable to access before with the traditional homology-based methods? What are the hidden functions of the dark proteome in animals unrevealed so far? To answer these questions, we performed a GO enrichment analysis per phylum of the functions inferred by GOPredSim-ProtT5 for the genes not annotated by eggNOG-mapper (see Methods). While 34.43% of all GO terms were found to be phylum-specific, 48 GO terms were common to all phyla (Table 1, Supp. Figure 7). Common GO terms were mainly related to viral response and the immune system, functions that have been proven to be under adaptive evolution in animals[25–27]. For instance, one common GO term was GO:0090729 (toxin activity), revealing a set of candidate genes that may be relevant to understanding the defense or hunting mechanisms of animals, as toxins and resistance to them is another trait usually under strong selection[28]. As an example, we identified an undescribed tachylectin-5-like protein in a scorpion species annotated with this GO term (Figure 3b, see Methods). Tachylectin-5 is a protein from the hemolymph of horseshoe crabs involved in its innate immune response against pathogens[29], which makes their 'blue blood' indispensable for drug testing. Related proteins have been found to be expressed in the venom glands of some spiders[30], suggesting a putative toxin-like role in scorpions. Moreover, several transposable element proteins (TEs) are usually functionally annotated as viruses, due to their viral origin[31]. Thus, viral GO terms might also be related to TEs, which play an important role in genome function and evolution[31].

Among non-common GO terms, we found plant and fungi-specific functions (e.g. "Photosynthesis"). However, tracing back the taxonomic origin of the GO terms for our subset of 93 animals, we found that 78% out of the 64% matching genes were of animal origin.

Regardless, these cases might be considered as convergent changes at the functional level across distant taxonomic scales that do not need to be any methodological artifact.

Next, we summarized the enriched functions by assigning each GO term to an Alliance for Genomics Resources (AGR) GO Slim GO for a simplified visualization of their function (Fig. 3a). We found that for all animal phyla, most of the enriched biological processes functions newly unveiled were related to response to stimuli, followed by immune system and developmental processes, signaling, and the establishment of localization (Fig. 3a, left). Molecular function and cellular component functions revealed that these newly annotated proteins have mainly catalytic activity or bind other molecules as parts of protein complexes (Fig. 3a, center, and right, respectively). This suggests that genes not annotated by eggNOG-mapper are probably proteins that either catalyze some protein, RNA, or DNA chemical modification, or bind to other molecules, form macromolecular complexes, and are involved in the regulation of essential processes for animals, such as their interaction with the environment (internal or external, biotic or abiotic) and their development.

Phylum or lineage-specific enrichment results also show interesting hidden functions related to their unique biological characteristics. For instance, tardigrades, commonly named water bears or moss piglets, are known for their high resilience to environmental stressors, such as desiccation or high levels of radiation[4]. Tardigrades contain newly annotated genes related to the response to stressors such as UV-A, osmotic stress, or heat, which may help us to understand how these animals can survive in such extreme conditions (Fig. 4a). Another example is the enigmatic monotypic phylum Micrognathozoa (Fig. 4b). These microscopic animals are known for their complex pharyngeal apparatus[32]. Here, we recover functions enriched in pharyngeal pumping and digestive system regulation, together with touch and taste perception, which may help us to understand the origin of such a complex pharynx. In addition, no males have been ever described and micrognathozoans are assumed to be parthenogenetic[32]. Nevertheless, our results revealed GO terms related to sperm and egg fusion, suggesting that males may exist but have not been observed yet (probably due to their small size as occurs in rotifers), and thus they may have sexual reproduction. Another possibility, if we consider them asexual, is that the genes that are involved in these sexual functions in other animals have not been lost in this phylum and were co-opted for another mechanism that shares some similarities at the molecular level. The last example is the ctenophores (Fig. 4c), whose putative position in the Animal Tree of Life as the sister lineage to

the rest of animals[33] and their unique complex syncytial nervous system[34] suggests that neurons may have evolved twice in animals. Here, newly inferred functions previously unannotated exposed GO terms related to neuronal activity (i.e., regulation of neuronal synaptic plasticity, neuropeptide signaling, dendritic transport of messenger ribonucleoprotein complex, and regulation of potassium ion transport), providing new putative annotated candidate genes that may shed light into the evolutionary origin of the nervous system in ctenophores. GO enrichments of the 'hidden proteome' of each animal phyla are not discussed here due to space constraints, but we encourage the readers to inspect them in the Supplementary File 2, as we believe they can serve as a very valuable source of candidate genes to further explore with other methodologies for hypothesis-driven functional testing.

Finally, as the dynamic structural conformation of proteins is partly responsible for the 'darkness' of proteins[35], we also inferred the disorder of each protein and found that there was a significant association between structure and annotation (p-value < 2.2e-16), with unannotated proteins being either globular (ordered) or mostly disordered (Fig. 3c). For instance, gene CG11373 from *D. melanogaster* encodes a mostly disordered protein (Fig. 3d) and lacks GO term annotation in Flybase[36]. Through FANTASIA, we inferred its function to be related to meiotic cell cycle regulation (Table 2), which is consistent with its described expression in the male germline and testis somatic cells[36](Fig. 3e). This highlights the power of using protein language-based methods in a non-negligible proportion of functionally-relevant[37] proteins in eukaryotic genomes[38], intrinsically disordered proteins (IDPs), wherestructure/domain prediction methods fail as they remain unfolded[39] and homology-based methods may fail due to their high evolutionary rate[40]. This opens a promising venue for functional annotation in IDPs, even if putatively.

FANTASIA: one pipeline to annotate them all

The best-performing method for GO prediction as highlighted by this study and our previous in model organisms[17], GOPredSim-ProtT5, is publicly available as a web server to explore functional annotation of single sequences (https://embed.protein.properties/). However, its automatic application for whole proteomes has not been implemented yet, nor does it include the necessary steps for the necessary data preprocessing or output configuration. In this context, to facilitate the annotation of whole proteomes using protein language models, we have developed a pipeline called FANTASIA (Functional ANnoTAtion based on embedding space

SImilArity). FANTASIA automatically takes as input a proteome file (either the longest isoform or the full set of isoforms for all genes), removes identical sequences using CD-HIT[41,42] and sequences longer than 5000 amino acids (due to a length constraint in ProtT5), and executes GOPredSim-ProtT5 for all sequences (Fig. 5a). Then, it converts the standard GOPredSim output file to the input file format for topGO [43] to facilitate its application in a wider biological workflow. When manually installing the pipeline and not using the provided singularity container, it allows the user to select CPUs or GPUs depending on the computer architecture where the pipeline is executed, with CPUs selected by default. GPUs decrease the execution time (Fig. 5b), which facilitates faster execution for several proteomes, but it does not influence the memory required. Memory use by GOPredSim-ProtT5 is directly proportional to the number of sequences in the proteome, regardless of the use of GPUs (Fig. 5c), and can be calculated using the following formula: $y = 4.4 + 0.0021\ x$, where y is the memory in gigabytes and x the size of the proteome in terms of number of sequences.

Discussion

The lack of functional annotation of genes has affected the biological inferences we can make in research, relying only on the gene families whose function has already been experimentally tested. This is particularly relevant in non-model animal species, as half of the genes cannot be functionally annotated based on homology, as shown above. Therefore, functional annotation has become a great challenge for the ongoing and future worldwide genome sequencing projects and any comparative or evolutionary genomics study, where traditional approaches can only partially capture the new biological features of those organisms. While the information inferred using homology-based methods can still be relevant, we are missing an important part of the biology of those organisms that may help us understand their uniqueness.

New methods for protein structure prediction[9,44] can help alleviate the problem for proteins that fall within the classical structure-function paradigm, but their large-scale application to a large number of proteins is still limited, as well as the accuracy of their predictions for intrinsically disordered proteins and other proteins with dynamic structures[39]. This is why faster sequence-based functional inference is a good alternative for research questions involving a high number of species (proteomes) with diverse types of proteins. In a previous study, we assessed the performance of several AI-based functional annotation methods when inferring functions in model organisms and found that protein language models provided a good

alternative in terms of coverage, informativeness, and precision when compared to deep learning-based and profile-based tools[17]. Here, we extended that comparison to non-model organisms across the animal kingdom and assessed how AI-based methods behaved when compared to the widely used homology-based eggNOG-mapper. eggNOG-mapper is based on the assignment of input proteins to pre-computed orthologous gene families and the posterior transfer of GO terms from these gene families. This entails several problems: (1) it is assumed that genes that are orthologs perform the same function in different species. This is known as the 'ortholog conjecture' and there is an ongoing debate on its universal truthfulnes[45]; (2) it relies on the correct identification of homology, which is especially hard for remote homologs, highly divergent sequences, or organisms that are not well studied and results in a large number of proteins without functional annotation. Besides its limitations, given the lack of better alternatives, it has become the standard functional annotation method in the field of evolutionary and comparative genomics of non-model organisms.

Our results highlight how protein language-based methods can help illuminate the animal 'dark functional landscape', i.e. to infer the function of proteins not previously annotated using homology-based methods. We recovered functions for genes that remained unannotated before but, how much can we trust these functions? When comparing how similar the annotations obtained using protein-language models to the ones obtained in homology-based methods are, we saw that the similarity was around 0.5 of average. This raises the question of whether we are accurately predicting the function of those genes, given that the functions obtained with the homology-based methods have not been experimentally validated either and could also be artifacts of the methodology. Further functional experiments and analyses will help to elucidate this issue. In the meantime, we believe that this resource is extremely helpful to advance comparative and functional genomics studies in animals but encourage researchers to proceed with caution when interpreting GO terms in genes that are not functionally validated, as we would do in the case of homology-based methods as well.

Furthermore, all these methods, both AI and homology-based, are still dependent on the existing functions in the public databases. The two protein language models tested here, SeqVec and ProtT5, were trained using Uniref50, which contains information on not only animal species but also plants, fungi, other eukaryotes, bacteria, archaea, and viruses. Therefore, it is not surprising to see among our results some plant or fungi-specific functions, especially in the biological process category. This does not necessarily invalidate our results, as functional

capabilities transcend evolutionary relationships (as seen by convergence). Thus it is still possible that some proteins in such distant organisms have similar functions at the molecular level while participating in different physiological processes. In addition, these databases are still mostly focused on a limited set of organisms relevant to human activities creating a detrimental bias when identifying the possible protein functions of less studied organisms. To alleviate these problems, it is important to consider biological diversity when selecting a reference dataset and enhance the currently available methods. Newer protein language models trained on larger datasets are being published[46–48] as well as new methods based on existing models[49] that may provide more accurate inferences in the near future. This may allow us to dig deeper into the results obtained when annotating a whole proteome or even solving other relevant evolutionary problems like remote homology detection[47,50].

To test the applicability of protein language model-based functional annotation to identify the previously 'hidden' biology in animals, we created an automatic pipeline called FANTASIA (Functional ANnoTAtion based on embedding space SImilArity) that automatically assigns GO terms to a whole proteome sequence file. We used FANTASIA on the proteomes of almost 1000 animal species and identified in the unannotated dark proteome proteins putatively involved in functions related to the response to the environment, immunity, metabolism, development and reproduction, processes differing among animal phyla and whose further understanding may help us understand their morphological, physiological and ecological diversity. In addition, we discovered putative functions of relevance to understanding the biology of less studied animal phyla, such as tardigrade resilience, micrognathozoan reproduction, or ctenophore neuronal system. Given that most of the experiments in animals are performed on only a handful of model organisms, these bioinformatic functional predictions will provide a starting point to further explore the molecular basis behind these animal biological processes. In addition, functional annotation with FANTASIA enables us to assign putative functions to previously unannotated intrinsically disordered proteins that are experimentally harder to work with. Without doubt, genetic and evolutionary studies in animals will benefit from the more complete functional annotation of the gene repertoire provided by this methodology, allowing us to put a biological meaning -even if putatively- into genes that are evolutionary relevant (e.g. duplicated, lost, or fast-evolving) but have remained functionally 'hidden' until now[51,52], which opens the door to a much deeper understanding on the animal evolutionary chronicle. All in all, in the era of genomics, protein language models are promising alternatives to homology-based models for

inferring the functions of genes in non-model organisms. Because important biological functions can be found, even in the 'darkest' of genes, if one has a way to turn on the light.

Methods

Input datasets and processing

We applied the methodology described in MATEdb2[53] to obtain the proteome files for 961 animals and 9 outgroups containing the longest peptide sequence (or isoform) per gene (Supp. File 1). In brief, we retrieved all datasets from the public databases. Transcriptomes were assembled *de novo* to ensure homogeneity in data processing. Genome annotations were downloaded from their respective repositories. All final and intermediate files are publicly available in MATEdb2[53]. Next, we created a subset of 93 species (Supp. File 1) maximizing the representation across the Animal Tree of Life to obtain all the isoforms (proteins) per each gene to test whether there were differences between annotating all isoforms or only the longest.

We filtered all proteomes to remove sequences that were identical in sequence using CD-HIT[41,42] and proteins that were larger than 5000 amino acids. This is due to a constraint in the embedding length when using the ProtT5 model of GOPredSim. We used the filtered proteomes as input for all annotation methods even if the filter was not needed in all of them to make sure results were comparable.

In addition, we performed two extra filters: 1) a depth-based filtering for eggNOG-mapper and DeepGOPlus results, as GOPredSim automatically removes GO terms that have a depth in the GO DAG lower than 2 (i.e. the terms are very general, close to or at the root of the DAG); 2) a reliability-based filtering for GOPredSim and DeepGOPlus results. This filtering is based on an index that measures how much you can trust the result based on the Euclidean distance between the query embeddings and the target proteins in the lookup reference dataset from which GO terms are transferred. Thresholds for filtering were obtained from the source publications based on evaluations using CAFA3[54] (GOPredSim[12] thresholds of 0.35 for biological process, 0.28 for molecular function, and 0.29 for cellular component, and DeepGOPlus[10] thresholds of 0.22 for biological process, 0.18 for molecular function, and 0.27 for cellular component). The latter filter was only applied in the first analyses since we decided to proceed with the unfiltered (or only depth-filtered) datasets (see main text).

When exploring the effect of using all isoforms for functional annotation, we first annotated all isoform sequences and then obtained per-gene annotations by collapsing the annotation of each isoform to retrieve a list of unique GOs.

Information Content

To see how informative the resultant GO terms were given their position within the GO graph, we estimated their Information Content (IC) using a custom script (https://github.com/cbbio/Compute_IC) which is defined as[55]:

$$IC =- log2(p(t))$$

Where p(t), the probability of a term occurring in a dataset, is calculated as:

$$p(t) = 1 - \frac{'Offspring\ Count'}{'Offspring\ Count' + 'Ancestors\ Count'}$$

Offspring and ancestors count refer to the number of proteins that are assigned to that term given a reference Gene Product Association Data file (GAPD). In this case, the reference GAPD, or universe of proteins, combined all GAPD files for all species available.

Higher IC implies a more specific GO term while low IC means that the GO term is more general.

Semantic similarity

To assess how similar two different annotations are, we obtained the lists of genes that were commonly annotated by all methods for the representative subset of species representing all phyla (n=93) and calculated for each gene their semantic similarity between the GO terms inferred using each method and eggNOG-mapper independently for each of the 3 GO terms categories. To be more conservative in our analysis, we used the most filtered dataset for each method (i.e., filtering by reliability index in GOPredSim-ProtT5 and GOPredSim-SeqVec and by reliability index and depth in DeepGOPlus). We estimated the Wang semantic similarity[56] using pygosemsim (https://github.com/mojaie/pygosemsim). This similarity measure, unlike other

measures, is based on the GO DAG topology, making it more stable than others based on information content that rely on the available annotations. Semantic similarity values range from 0 to 1, with 0 meaning no similarity and 1 meaning (almost) identical.

GO enrichment for genes not annotated by eggNOG-mapper and visualization

We obtained the list of genes that were not annotated by eggNOG-mapper for each species and performed a GO enrichment analysis using topGO[43] against the whole set of proteins of that species to identify which functions were being missed by eggNOG-mapper. We used GO terms predicted by GOPredSim-ProtT5. TopGO, in contrast to other GO enrichment methods, is based on the GO DAG, allowing us to take relationships between GO terms into account.
From the GO enrichment per species, we obtained a list of GO terms enriched per phylum by combining the results per species. For that, in the case of GO terms that were enriched in at least one species, we combined the p-values using the ordmeta method[57].

We identified the list of GO terms that were commonly enriched in all animal phyla. Among all common GO terms (Table 1), we focused on toxin activity (GO:0090729) to assess the performance of the functional annotation. Among all animals with venom, we selected the scorpion *Centruroides sculpturatus* to look into the genes with this GO term that were not annotated by eggNOG-mapper. We used ToxinPred2[58] to further filter this list and identified CSCU1_DN0_c0_g23215_i1.p1 (XP_023242874.1, uncharacterized protein) as a putative homolog of tachylectin-5 (XP_023224331.1). We further assessed this by predicting their structure with ColabFold[44] and comparing them with the experimentally-determined structure of the horseshoe crab *Tachypleus tridentatus* tachylectin-5A (PDB 1JC9) using the PDB pairwise structure alignment tool.

To visualize the significantly enriched GO terms per phylum and GO category, we first calculated the Wang semantic similarity matrix for each dataset and compared several clustering algorithms using the simplifyEnrichment R package[59], and proceeded with the K-means algorithm (Supp. File 3). Then, to summarize the function of each cluster, we obtained their representative GO term by selecting the GO term with the lowest closeness value (the ‘farthest’) in the subset semantic similarity network for that cluster. Results were plotted in a TreeMap.
We then combined the results of all phyla by assigning each significant GO term to the GO slim AGR (Alliance of Genomics Resources) subset using the map-to-slim.py script available in

GOAtools[60]. Between 10 and 60% of GO terms per phylum were not assigned to a category and were further clustered as explained above. The GO representative terms of each of the 49 clusters were later manually assigned to the GO slim AGR classification and used to include unassigned GOs to the GO-GO category list. Final results were normalized per phylum (row) and visualized in a heatmap.

Disorder prediction of proteins

Average disorder (i.e. the average of the tendency for each amino acid to be in a region lacking a fixed structure) per protein was predicted for our dataset using a modified Python implementation of IUPred3[61] that takes a multi-FASTA as input file and calculates the average disorder per protein. Proteins were then classified according to their disorder in 4 classes previously described[62]: (1) globular (disorder <0.1); (2) nearly ordered (disorder 0.1-0.3); (3) partially disordered (disorder 0.3-0.7) and (4) mostly disordered (disorder >0.7). We performed a χ^2 test to assess the statistical significance of the association between disorder category and not being annotated by egg-NOG-mapper.

GO version

GO knowledgebase is regularly updated with new releases that remove obsolete GO terms or add new ones, which hinders comparison between analyses that use different versions[63]. Hence, we used the same version for all of the analyses whenever possible to improve comparability. For DeepGOPlus and GOPredSim annotation, which are dependent on the GO release, we used the release of March 22nd 2022. This same release was used for the GOslim subsets. For information content and semantic similarity analyses, we used the release of June 11th 2023.

Computational resources

GPU-based GO annotations were performed in an HPC cluster (Drago, Spanish National Research Council, CSIC) using nodes with an NVIDIA Ampere A100 graphic card with 512 GB and Dual Intel Xeon Gold 6248R 24C 3.0GHz processors. Only one species (*Panulirus ornatus*) was annotated using CPUs in another supercomputer (Centro de Supercomputación de Galicia, CESGA, Finisterrae III) due to the limitation in the maximum node memory. NVIDIA A100 GPU

nodes with 32 CPUs per node and a maximum memory of 250G and CPUs were used for testing the performance of FANTASIA in CESGA, Finisterrae III.

Data availability

All proteomes, EggNOG-mapper annotations and GOPredSim-ProtT5 per-protein embeddings and annotations are available in MATEdb2[53]. DeepGOPlus and GOPredSim-Seqvec annotations are provided in Zenodo[65-70], together with information content, semantic similarity and other results. The complete list of the files can be found on GitHub (https://github.com/MetazoaPhylogenomicsLab/Martinez_Redondo_et_al_2024_Dark_Proteome_Animal_Tree_Of_Life).

Code availability

FANTASIA pipeline and dependencies are included in a Singularity (v3.8.3+231-g9dceb4240) container available on GitHub (https://github.com/MetazoaPhylogenomicsLab/FANTASIA), together with an installation guide for its use on any computer (with or without GPUs). In addition, given the dependence of the software versions on the computational architecture used, and the amount of resources needed, we are currently working on its implementation as a Galaxy tool[64] to make it more easily incorporated into biological research.

Steps followed for the creation of several intermediate files not available on Zenodo, as well as additional scripts used for the analyses and creating figures are available on GitHub (https://github.com/MetazoaPhylogenomicsLab/Martinez_Redondo_et_al_2024_Dark_Proteome_Animal_Tree_Of_Life).

Acknowledgments

GIMR, AMR, and RF acknowledge funding and support from the LifeHUB/CSIC research network (PIE-202120E047). 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). AMR and IB acknowledge support from MDM-2016-0687 and PID2021-127503OB-I00 funded by MCIN. We also thank Centro de Supercomputación de Galicia and CSIC for access to computer resources (CESGA and DRAGO respectively).

Author contributions

GIMR and RF designed the study. GIMR performed the main analyses, created the singularity container, and designed the figures. GIMR and IB wrote the code and the pipeline to automate the methods. MVV assisted with some data analysis. GIMR, AMR and RF wrote the first version of the manuscript. RF and AMR provided resources, the rationale and concepts behind the work, and supervised the study. All authors revised and approved the final version of the manuscript.

References

1. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
2. Cantalapiedra, C. P., 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.* **38**, 5825–5829 (2021).
3. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022).
4. Hashimoto, T. *et al.* Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nat. Commun.* **7**, 12808 (2016).
5. Kenny, N. J. *et al.* Tracing animal genomic evolution with the chromosomal-level assembly

of the freshwater sponge Ephydatia muelleri. *Nat. Commun.* **11**, 3676 (2020).

6. Lewin, H. A. *et al.* The Earth BioGenome Project 2020: Starting the clock. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).
7. Mazzoni, C. J., Ciofi, C. & Waterhouse, R. M. Biodiversity: an atlas of European reference genomes. *Nature* **619**, 252 (2023).
8. Mc Cartney, A. M. *et al.* The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics. *bioRxiv* 2023.09.25.559365 (2023) doi:10.1101/2023.09.25.559365.
9. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
10. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).
11. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
12. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Rep.* **11**, 1160 (2021).
13. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
14. Flamholz, Z. N., Biller, S. J. & Kelly, L. Large language models improve annotation of prokaryotic viral proteins. *Nat Microbiol* **9**, 537–549 (2024).
15. Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–937 (2022).
16. Miller, D., Stern, A. & Burstein, D. Deciphering microbial gene function using natural language processing. *Nat. Commun.* **13**, 5731 (2022).
17. Barrios-Nunez, I. *et al.* Decoding proteome functional information in model organisms using protein language models. *bioRxiv* 2024.02.14.580341 (2024)

doi:10.1101/2024.02.14.580341.

18. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**, 723 (2019).
19. Thomas, P. D. The Gene Ontology and the Meaning of Biological Function. *Methods Mol. Biol.* **1446**, 15–24 (2017).
20. Zhang, C. *et al.* Mapping the Response of Human Osteocytes in Native Matrix to Mechanical Loading Using RNA Sequencing. *JBMR Plus* **7**, e10721 (2023).
21. MacPherson, R. A., Shankar, V., Anholt, R. R. H. & Mackay, T. F. C. Genetic and Genomic Analyses of Drosophila melanogaster Models of Chromatin Modification Disorders. *bioRxiv* (2023) doi:10.1101/2023.03.30.534923.
22. Clapes, T. *et al.* Chemotherapy-induced transposable elements activate MDA5 to enhance haematopoietic regeneration. *Nat. Cell Biol.* **23**, 704–717 (2021).
23. Platanitis, E. *et al.* Interferons reshape the 3D conformation and accessibility of macrophage chromatin. *iScience* **25**, 103840 (2022).
24. Menon, M. P. & Hua, K.-F. The Long Non-coding RNAs: Paramount Regulators of the NLRP3 Inflammasome. *Front. Immunol.* **11**, 569524 (2020).
25. Kober, K. M. & Pogson, G. H. Genome-wide signals of positive selection in strongylocentrotid sea urchins. *BMC Genomics* **18**, 555 (2017).
26. McTaggart, S. J., Obbard, D. J., Conlon, C. & Little, T. J. Immune genes undergo more adaptive evolution than non-immune system genes in Daphnia pulex. *BMC Evol. Biol.* **12**, 63 (2012).
27. Shultz, A. J. & Sackton, T. B. Immune genes are hotspots of shared positive selection across birds and mammals. *Elife* **8**, (2019).
28. van Thiel, J. *et al.* Convergent evolution of toxin resistance in animals. *Biol. Rev. Camb. Philos. Soc.* **97**, 1823–1843 (2022).
29. Kawabata, S. & Iwanaga, S. Role of lectins in the innate immunity of horseshoe crab. *Dev.*

Comp. Immunol. **23**, 391–400 (1999).

30. Carlson, D. E. & Hedin, M. Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes. *PLoS One* **12**, e0174102 (2017).
31. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu. Rev. Genet.* **54**, 539–561 (2020).
32. Kristensen, R. M. & Funch, P. Micrognathozoa: a new class with complicated jaws like those of Rotifera and Gnathostomulida. *J. Morphol.* **246**, 1–49 (2000).
33. Li, Y., Shen, X.-X., Evans, B., Dunn, C. W. & Rokas, A. Rooting the Animal Tree of Life. *Mol. Biol. Evol.* **38**, 4322–4333 (2021).
34. Burkhardt, P. *et al.* Syncytial nerve net in a ctenophore adds insights on the evolution of nervous systems. *Science* **380**, 293–297 (2023).
35. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15898–15903 (2015).
36. Öztürk-Çolak, A. *et al.* FlyBase: updates to the Drosophila genes and genomes database. *Genetics*, iyad211 (2024).
37. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* (2023).
38. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004).
39. Perrakis, A. & Sixma, T. K. AI revolutions in biology: The joys and perils of AlphaFold. *EMBO Rep.* **22**, e54046 (2021).
40. Trivedi, R. & Nagarajaram, H. A. Intrinsically Disordered Proteins: An Overview. *Int. J. Mol. Sci.* **23**, (2022).
41. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein

or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

42. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
43. Adrian Alexa, J. R. *topGO*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.TOPGO.
44. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
45. Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. & Dessimoz, C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.* **8**, e1002514 (2012).
46. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
47. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
48. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
49. Hamamsy, T. *et al.* Learning sequence, structure, and function representations of proteins with language models. *bioRxiv* (2023) doi:10.1101/2023.11.26.568742.
50. Kaminski, K., Ludwiczak, J., Pawlicki, K., Alva, V. & Dunin-Horkawicz, S. pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics* **39**, (2023).
51. Aristide, L. & Fernández, R. Genomic Insights into Mollusk Terrestrialization: Parallel and Convergent Gene Family Expansions as Key Facilitators in Out-of-the-Sea Transitions. *Genome Biol. Evol.* **15**, (2023).
52. Fernández, R. & Gabaldón, T. Gene gain and loss across the metazoan tree of life. *Nat Ecol Evol* **4**, 524–533 (2020).
53. Martinez-Redondo, G. I. *et al.* MATEdb2, a collection of high-quality metazoan proteomes

across the Animal Tree of Life to speed up phylogenomic studies. *bioRxiv* 2024.02.21.581367 (2024) doi:10.1101/2024.02.21.581367.

54. Zhou, N. *et al.* The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 244 (2019).
55. Louie, B., Bergen, S., Higdon, R. & Kolker, E. Quantifying protein function specificity in the gene ontology. *Stand. Genomic Sci.* **2**, 238–244 (2010).
56. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
57. Yoon, S., Baik, B., Park, T. & Nam, D. Powerful p-value combination methods to detect incomplete association. *Sci. Rep.* **11**, 6980 (2021).
58. Sharma, N., Naorem, L. D., Jain, S. & Raghava, G. P. S. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief. Bioinform.* **23**, (2022).
59. Gu, Z. & Hübschmann, D. simplifyEnrichment: A Bioconductor Package for Clustering and Visualizing Functional Enrichment Results. *Genomics Proteomics Bioinformatics* **21**, 190–202 (2023).
60. Klopfenstein, D. V. *et al.* GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
61. Erdős, G., Pajkos, M. & Dosztányi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic Acids Res.* **49**, W297–W303 (2021).
62. Avramov, M. *et al.* Identification of Intrinsically Disordered Proteins and Regions in a Non-Model Insect Species Ostrinia nubilalis (Hbn.). *Biomolecules* **12**, (2022).
63. Tomczak, A. *et al.* Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci. Rep.* **8**, 5115 (2018).
64. Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative

biomedical analyses: 2022 update. *Nucleic Acids Res.* **50**, W345–W351 (2022).

65. Martínez-Redondo, G.I. (2024). Data (part 1) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. Zenodo. https://doi.org/10.5281/zenodo.10714961

66. Martínez-Redondo, G.I. (2024). Data (part 2) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. Zenodo. https://doi.org/10.5281/zenodo.10717484

67. Martínez-Redondo, G.I. (2024). Data (part 3) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. Zenodo. https://doi.org/10.5281/zenodo.10717774

68. Martínez-Redondo, G.I. (2024). Data (part 4) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. Zenodo. https://doi.org/10.5281/zenodo.10717783

69. Martínez-Redondo, G.I. (2024). Data (part 5) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. Zenodo. https://doi.org/10.5281/zenodo.10717885

70. Martínez-Redondo, G.I. (2024). Data (part 6) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. Zenodo. https://doi.org/10.5281/zenodo.10717910

Tables

Table 1: GO terms significantly enriched in the ‘dark proteome’ that were common to all animal phyla.

GO term	Description	GO category
GO:0030435	sporulation resulting in formation of a cellular spore	Biological process
GO:0075733	intracellular transport of virus	Biological process
GO:0046718	viral entry into host cell	Biological process
GO:0046740	transport of virus in host, cell to cell	Biological process
GO:0032755	positive regulation of interleukin-6 production	Biological process
GO:0019064	fusion of virus membrane with host plasma membrane	Biological process
GO:0019058	viral life cycle	Biological process
GO:0019083	viral transcription	Biological process
GO:0019073	viral DNA genome packaging	Biological process
GO:0019076	viral release from host cell	Biological process
GO:0007218	neuropeptide signaling pathway	Biological process
GO:0039526	modulation by virus of host apoptotic process	Biological process
GO:0039502	suppression by virus of host type I interferon-mediated signaling pathway	Biological process
GO:0039503	suppression by virus of host innate immune response	Biological process
GO:0039693	viral DNA genome replication	Biological process
GO:0039644	suppression by virus of host NF-kappaB transcription factor activity	Biological process
GO:0016032	viral process	Biological process
GO:0006278	RNA-dependent DNA biosynthetic process	Biological process
GO:0031640	killing of cells of other organism	Biological process
GO:0015074	DNA integration	Biological process
GO:0003964	RNA-directed DNA polymerase activity	Molecular function
GO:0005179	hormone activity	Molecular function
GO:0039660	structural constituent of virion	Molecular function
GO:0030354	melanin-concentrating hormone activity	Molecular function
GO:0090729	toxin activity	Molecular function
GO:0031777	type 1 melanin-concentrating hormone receptor binding	Molecular function
GO:0019865	immunoglobulin binding	Molecular function

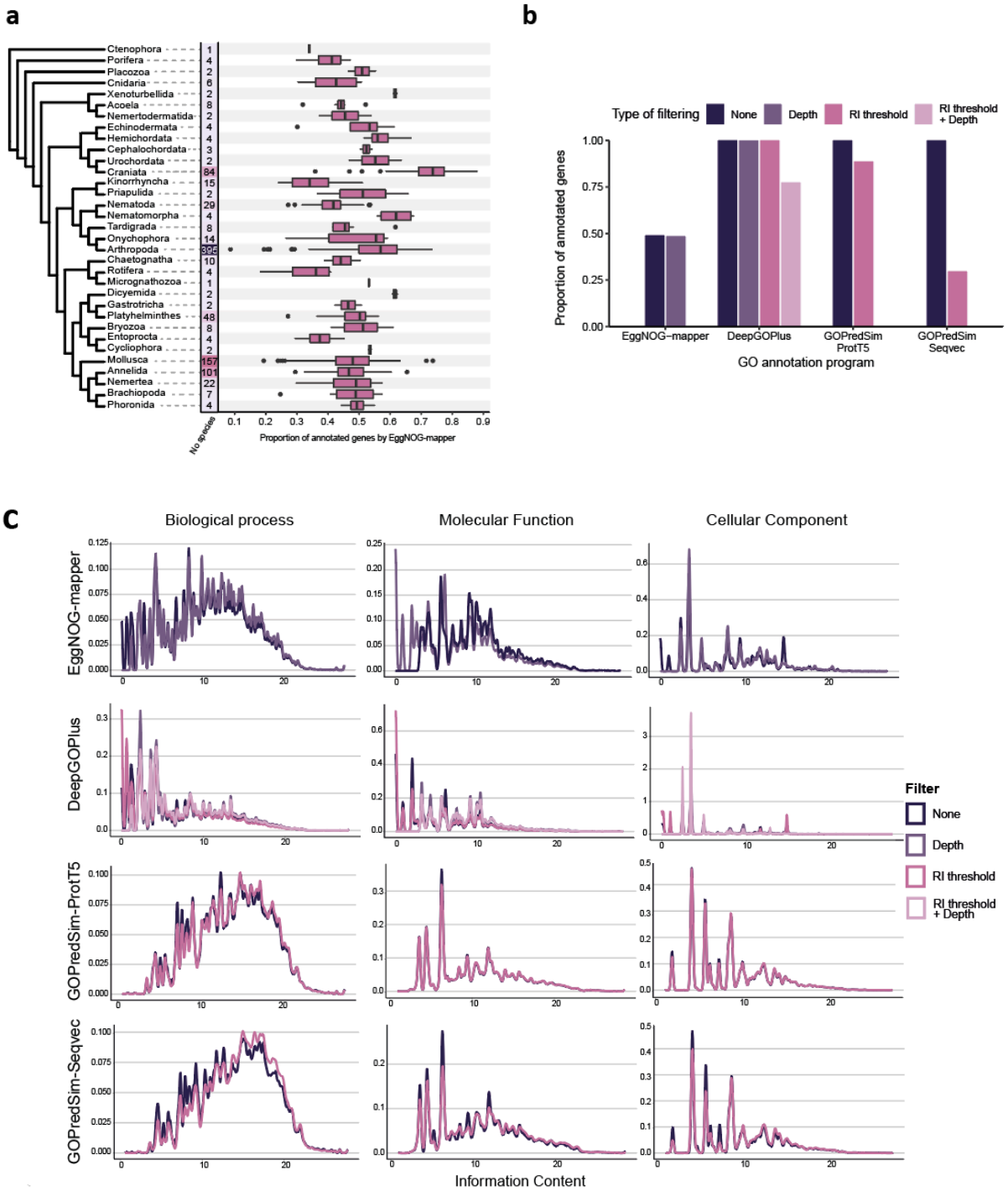
GO:0004867	serine-type endopeptidase inhibitor activity	Molecular function
GO:0044423	virion component	Cellular component
GO:0033655	host cell cytoplasm part	Cellular component

Table 2: FANTASIA GO terms for gene CG11373 in *D. melanogaster*.

GO term	Description	GO category
GO:0051321	meiotic cell cycle	Biological process
GO:0140538	negative regulation of conjugation with zygote	Biological process
GO:0030435	sporulation resulting in formation of a cellular spore	Biological process
GO:0043086	negative regulation of catalytic activity	Biological process
GO:0110046	signal transduction involved in cell cycle switching, mitotic to meiotic cell cycle	Biological process
GO:0030430	host cell cytoplasm	Cellular component
GO:0042025	host cell nucleus	Cellular component
GO:0030291	protein serine/threonine kinase inhibitor activity	Molecular function
GO:0005515	protein binding	Molecular function

Figures

Figure 1



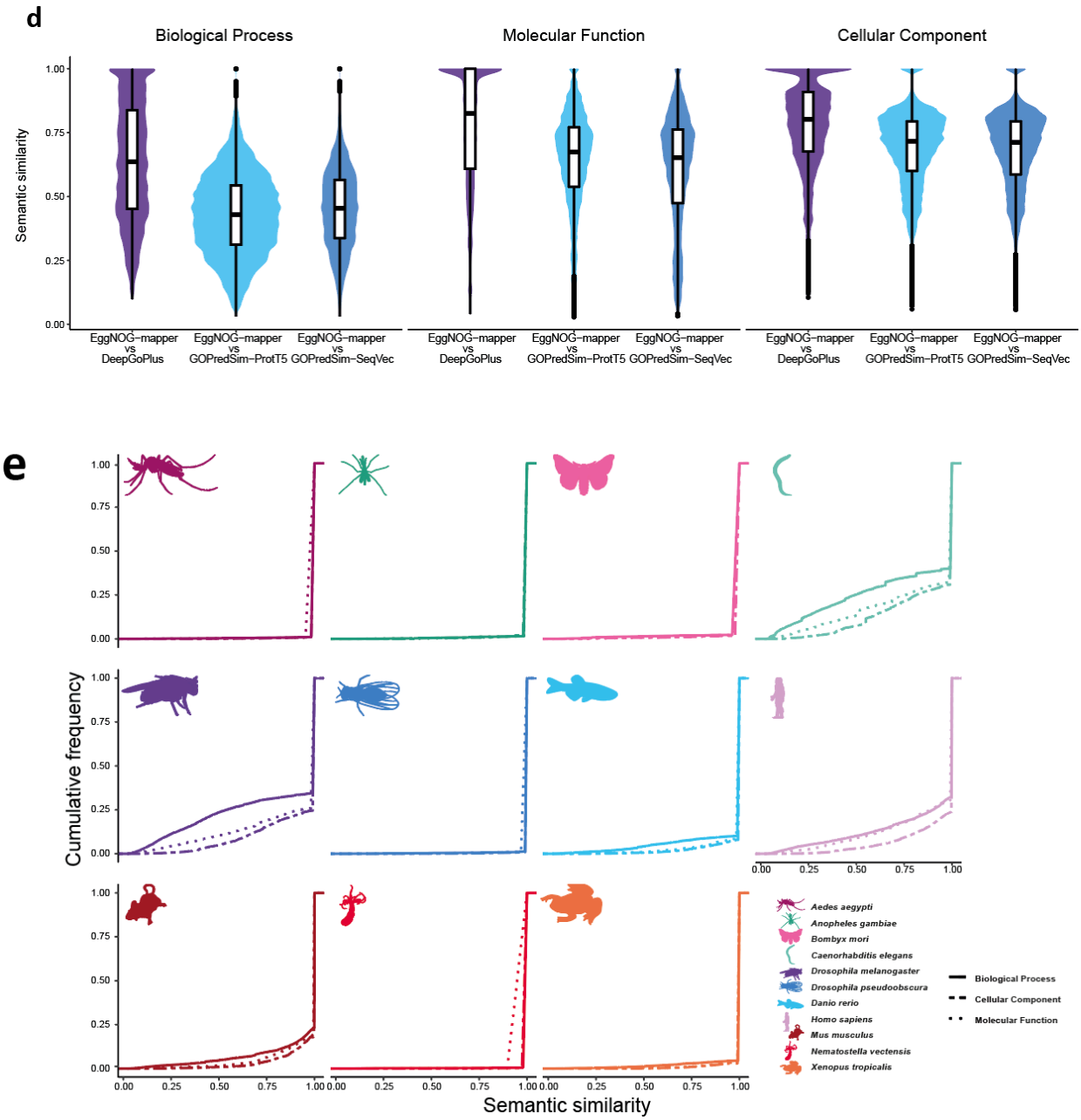


Fig. 1: Comparison of functional annotation methods across the Animal Tree of Life. a, Number of animal species per phylum included in this study (left) and the proportion of genes annotated by eggNOG-mapper per phylum (right). **b,** Proportion of genes annotated by eggNOG-mapper, DeepGOPlus, GOPredSim-ProtT5, and GOPredSim-SeqVec without applying any filter (dark purple), filtering by depth (light purple), filtering by reliability index (RI, pink) and both filters (light pink)**. c,** Comparison of the Information Content (IC) of the annotations made by the different methods, for every filtering and GO category. **d,** Semantic similarity of gene annotations from DeepGOPlus, GOPredSim-ProtT5, and GOPredSim-SeqVec compared to

eggNOG-mapper for all 3 GO categories. **e,** Semantic similarity of the GOPredSim-ProtT5 GO annotations obtained for 11 model organisms when removing their sequences from the GOPredSim lookup dataset compared to the original annotation.

Figure 2

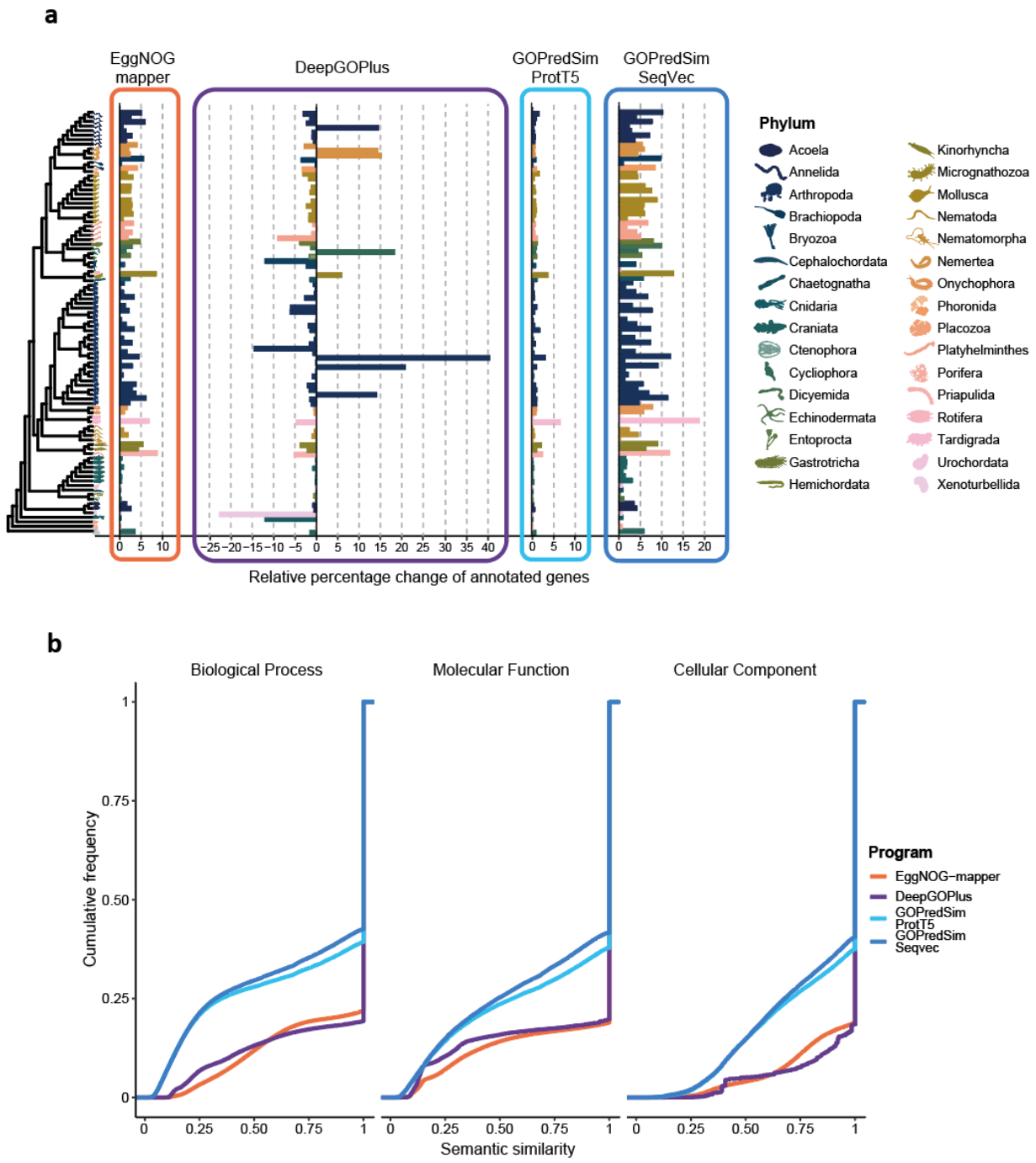
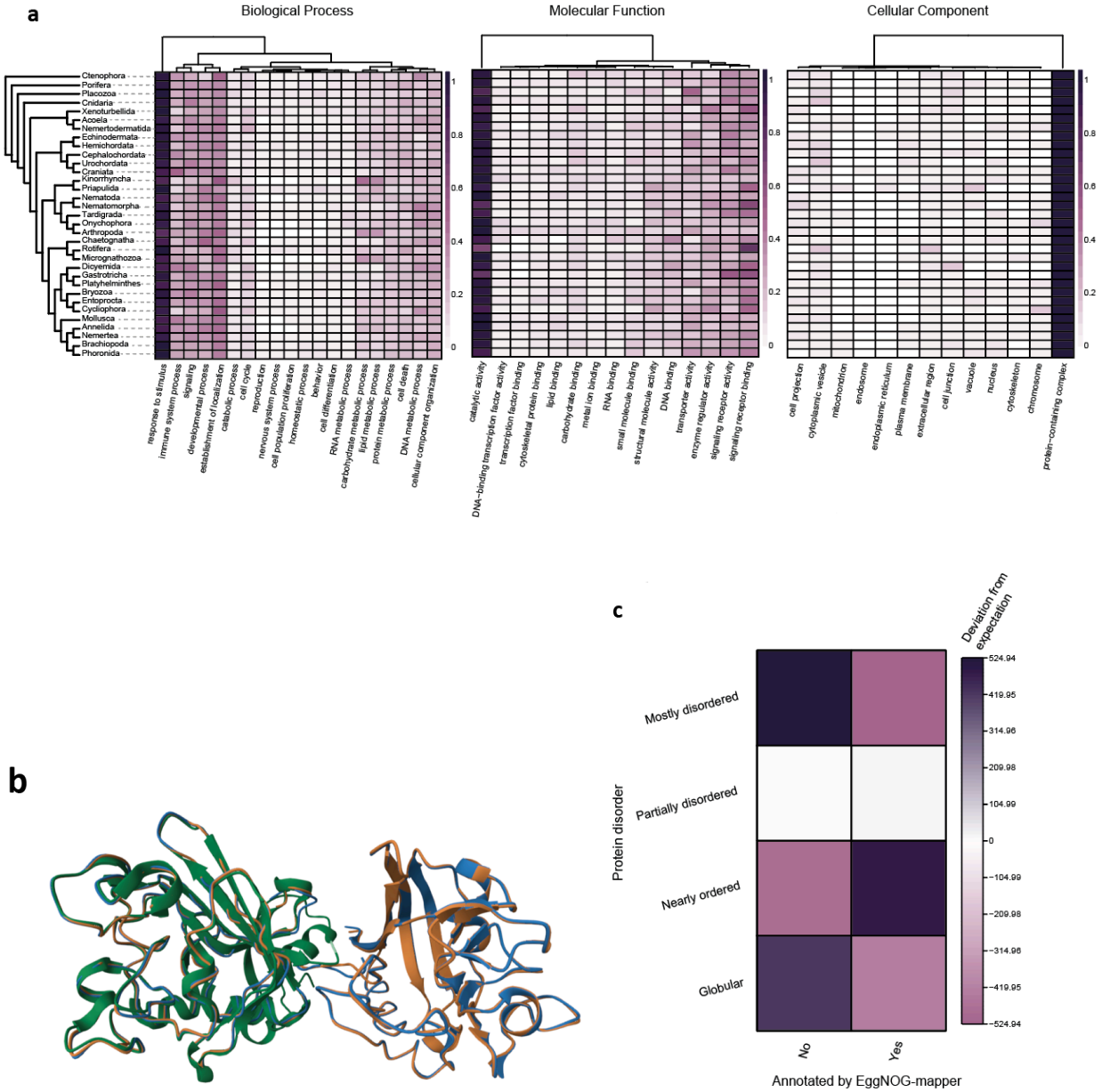


Fig. 2: Per-gene functional annotation differences when using all isoforms or just the longest isoform. a, Percentage change of genes annotated when using all isoforms compared to genes annotated using the longest isoform. Results are shown for the most filtered datasets, i.e. DeepGOPlus with depth and reliability index filters and GOPredSim-ProtT5 and

GOPredSim-SeqVec filtered by reliability index. **b,** Semantic similarity when comparing per-gene functional annotations obtained using all isoforms with the annotations obtained using the longest isoform**.**

Figure 3


a
Biological Process
Molecular Function
Cellular Component
Ctenophora
Porifera
Placozoa
Cnidaria
Xenoturbellida
Acoela
Nemertodermatida
Echinodermata
Hemichordata
Cephalochordata
Urochordata
Craniata
Kinorhyncha
Priapulida
Nematoda
Nematomorpha
Tardigrada
Onychophora
Arthropoda
Chaetognatha
Rotifera
Micrognathozoa
Dicyemida
Gastrotricha
Platyhelminthes
Bryozoa
Entoprocta
Cycliophora
Mollusca
Annelida
Nemertea
Brachiopoda
Phoronida
response to stimulus
immune system process
signaling
developmental process
establishment of localization
catabolic process
cell cycle
reproduction
nervous system process
cell population proliferation
homeostatic process
behavior
cell differentiation
RNA metabolic process
carbohydrate metabolic process
lipid metabolic process
protein metabolic process
cell death
DNA metabolic process
cellular component organization
catalytic activity
DNA-binding transcription factor activity
transcription factor binding
cytoskeletal protein binding
lipid binding
carbohydrate binding
metal ion binding
RNA binding
small molecule binding
structural molecule activity
DNA binding
transporter activity
enzyme regulator activity
signaling receptor activity
signaling receptor binding
cell projection
cytoplasmic vesicle
mitochondrion
endosome
endoplasmic reticulum
plasma membrane
extracellular region
cell junction
vacuole
nucleus
cytoskeleton
chromosome
protein-containing complex
1
0.8
0.6
0.4
0.2
0
b
c
Deviation from expectation
524.94
419.95
314.96
209.98
104.99
0
-104.99
-209.98
-314.96
-419.95
-524.94
Protein disorder
Mostly disordered
Partially disordered
Nearly ordered
Globular
No
Yes
Annotated by EggNOG-mapper

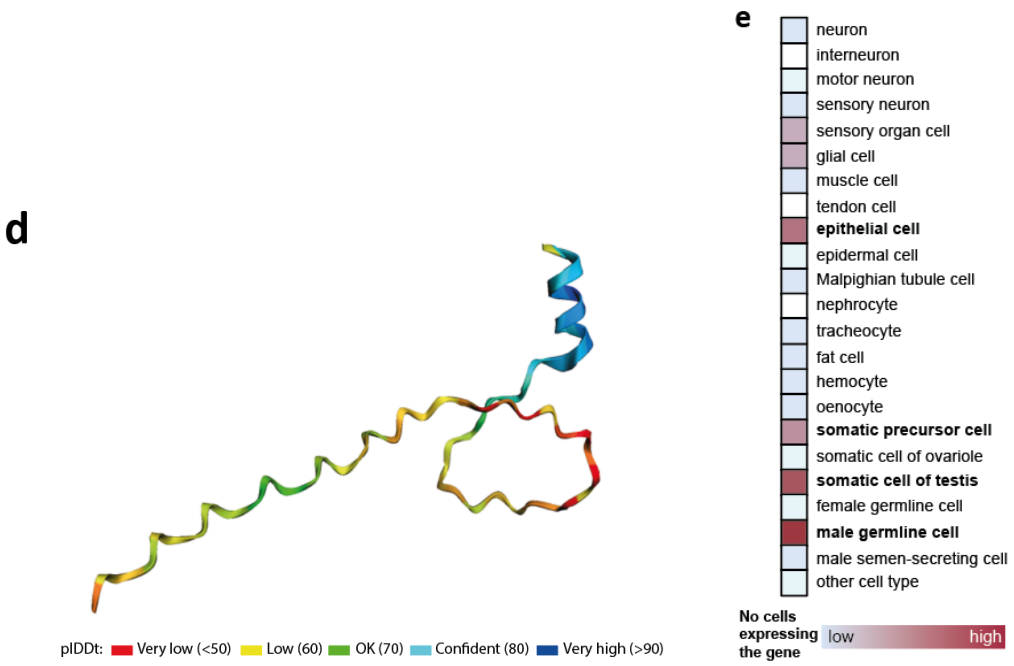
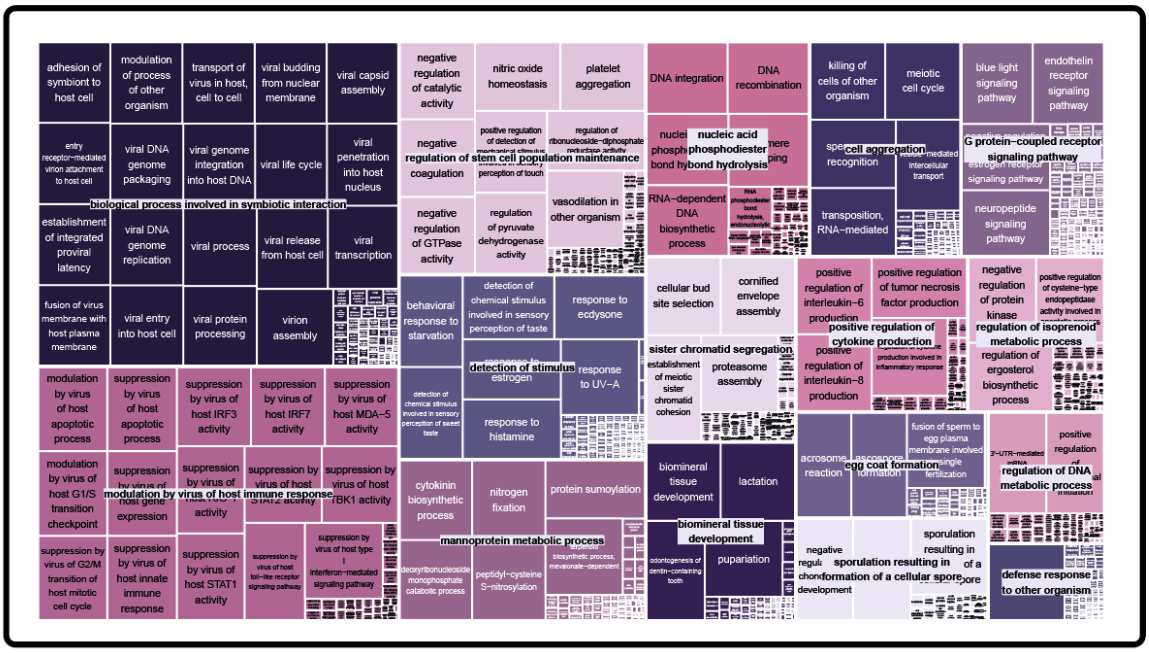


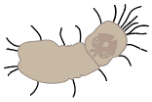
Fig. 3: Functional annotation of the dark proteome across the Animal Tree of Life. a, Per-phylum normalized number of enriched GO terms assigned to AGR GO slim terms. **b,** Structure alignment of *C. sculpturatus* uncharacterized protein XP_023242874.1 (in orange), tachylectin-5A-like XP_023224331.1 (in blue), and *T. tridentatus* 1JC9 (in green). RMSD against the uncharacterized protein is 0.44 and 0.53. respectively. **c,** Chi-square residuals showing the association between being annotated or not by eggNOG-mapper and the disorder of the protein. Colors less similar to white indicate a higher (more purple) or lower (more pink) number of newly annotated proteins than expected. **d,** ColabFold[44] structure prediction for gene CG11373 in *D. melanogaster* (unknown function). **e,** Fly Cell Atlas expression summary ribbons for gene CG11373 in *D. melanogaster* as shown in FlyBase.

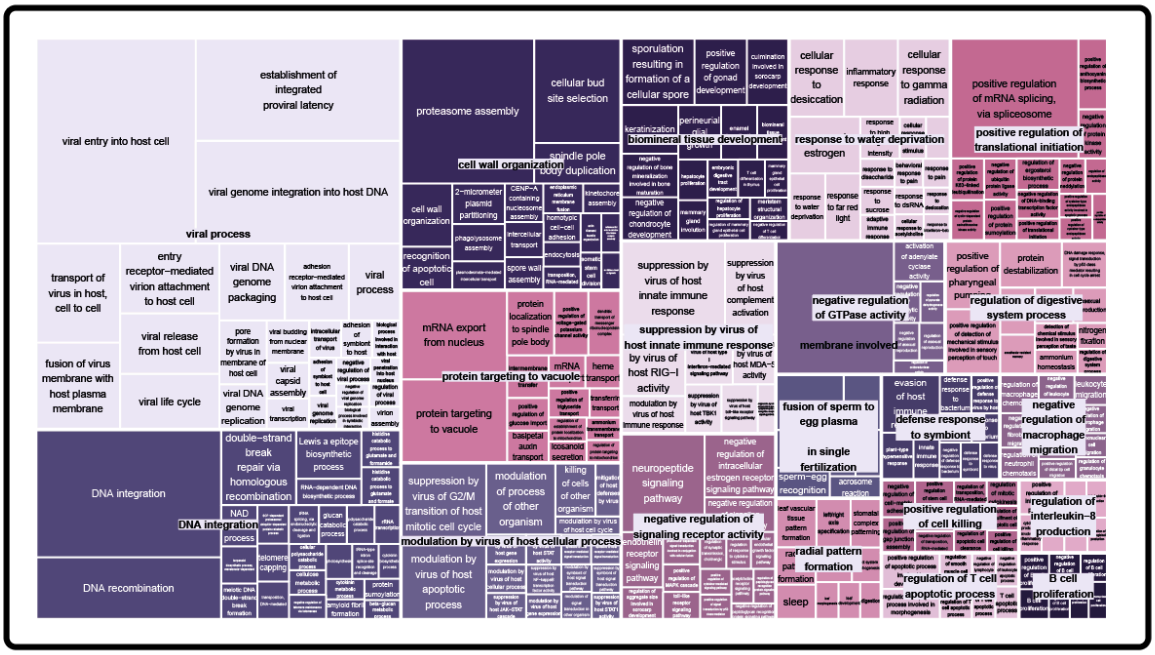
Figure 4

a Tardigrada



b Micrognathozoa





c Ctenophora

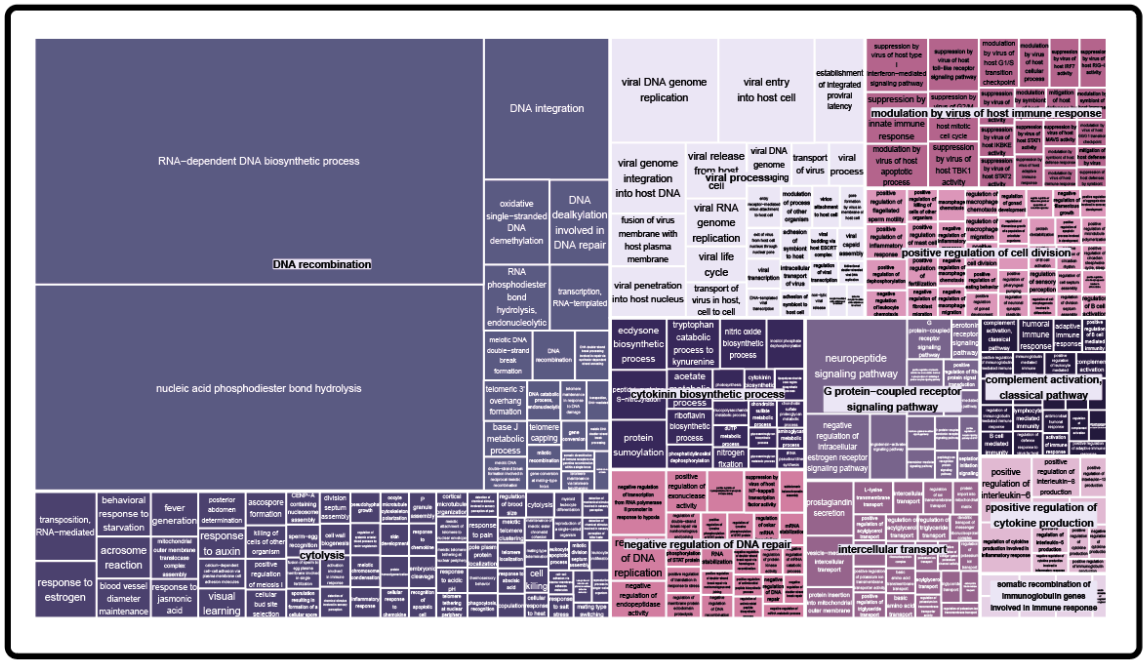


Fig. 4: Enriched biological process GO terms in some animal phyla. a, Tardigrada. **b,** Micrognathozoa**. c,** Ctenophora.

Figure 5

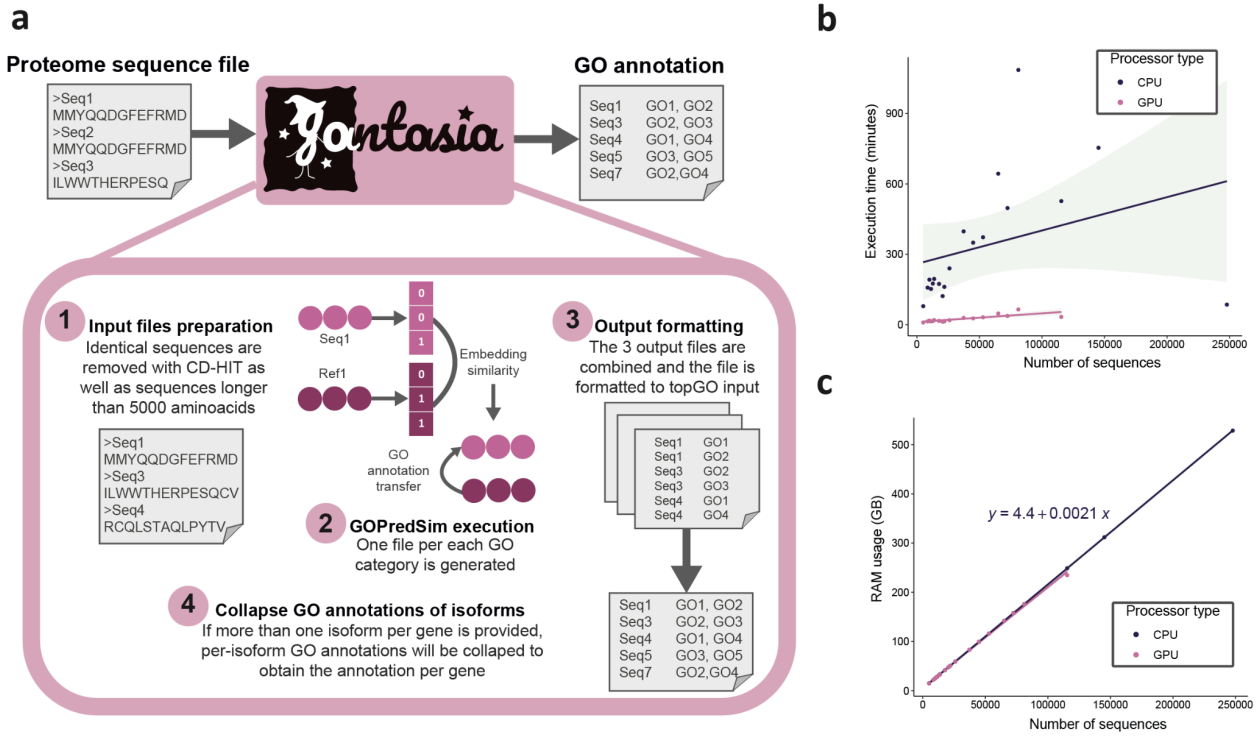
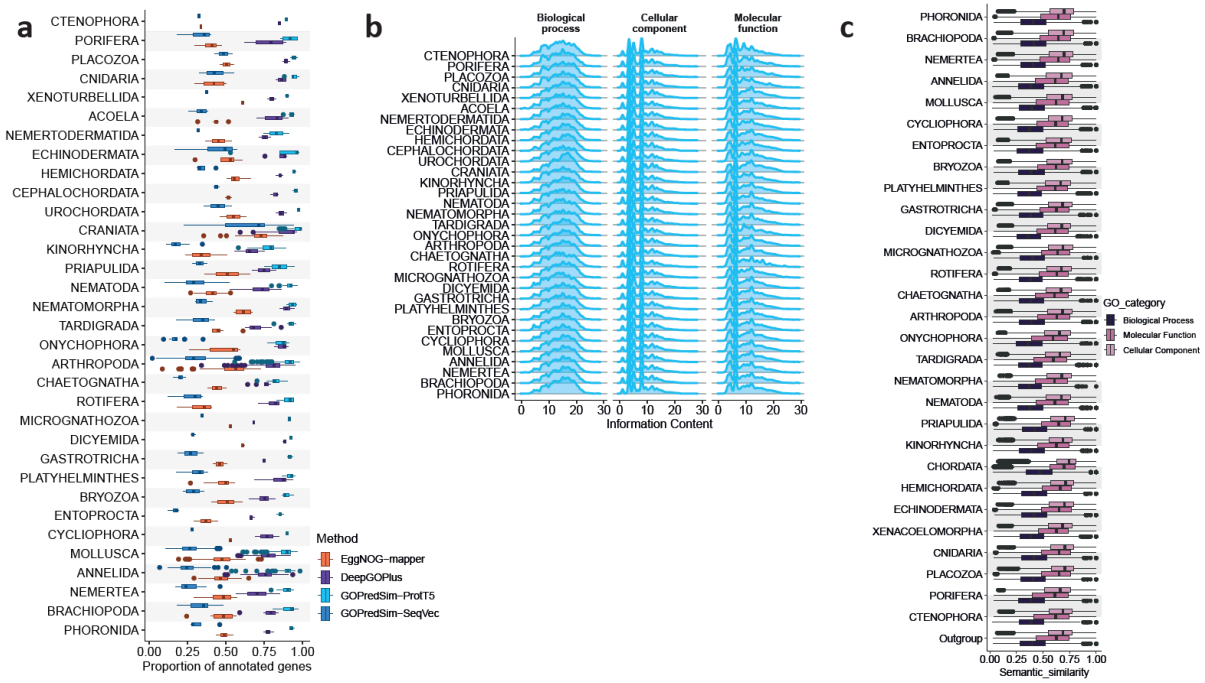


Fig. 5: Functional ANnoTAtion based on embedding space SImilArity (FANTASIA) pipeline. a, From an input peptide file (either including the longest isoform or all isoforms for all genes in a proteome), FANTASIA first removes 100% identical sequences and sequences with a length higher than 5000 amino acids. Then, it automatically creates the configuration file needed by GOPredSim and executes the software to obtain the list of GO terms assigned to each protein. **b,** Relationship between the size of the proteome (number of sequences) and the execution time using CPUs or GPUs**. c,** Relationship between the size of the proteome (number of sequences) and the memory employed when using CPUs or GPUs.

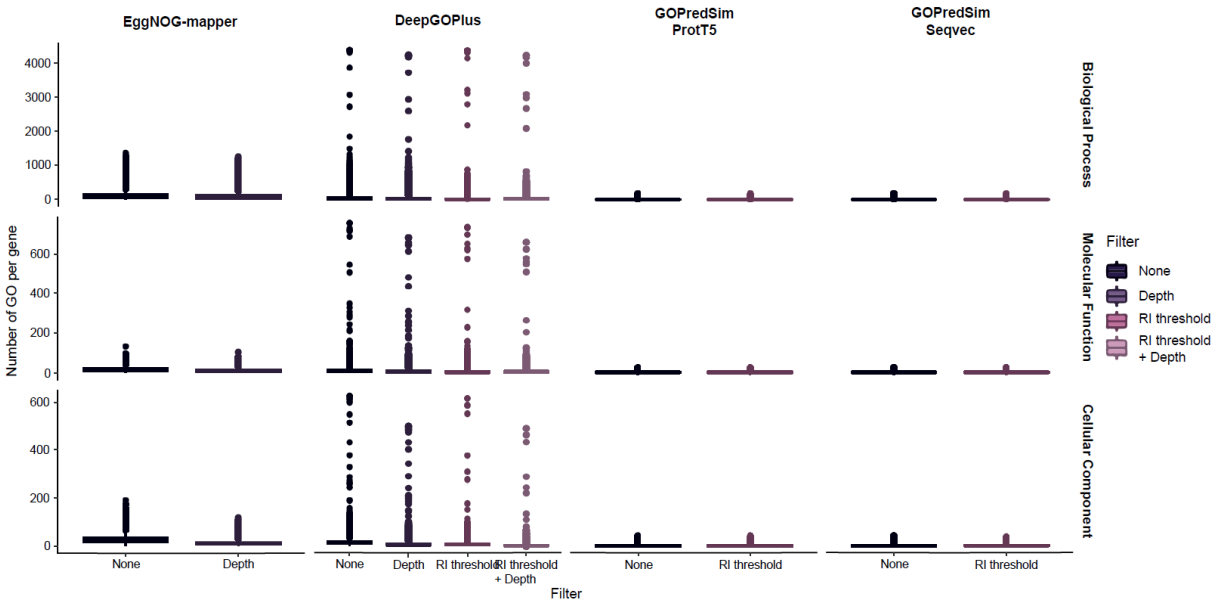
Supplementary figures

Suppl. Fig S1



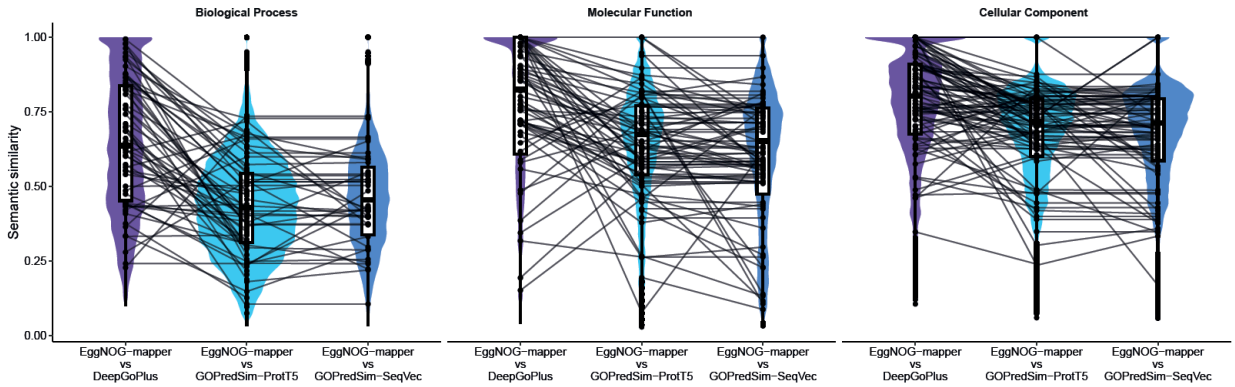
Supp. Fig. 1: Comparison of functional annotation methods per phylum. a, Proportion of annotated genes per phylum in all 4 methods tested. **b,** Information content per phylum for all 3 GO categories inferred by GOPredSim-ProtT5 for the reliability index filtered dataset. **c,** Semantic similarity between the GO terms inferred by EggNOG-mapper and the GO categories inferred by GOPredSim-ProtT5 for the reliability index filtered dataset per phylum for all 3 GO categories.

Suppl. Fig. S2



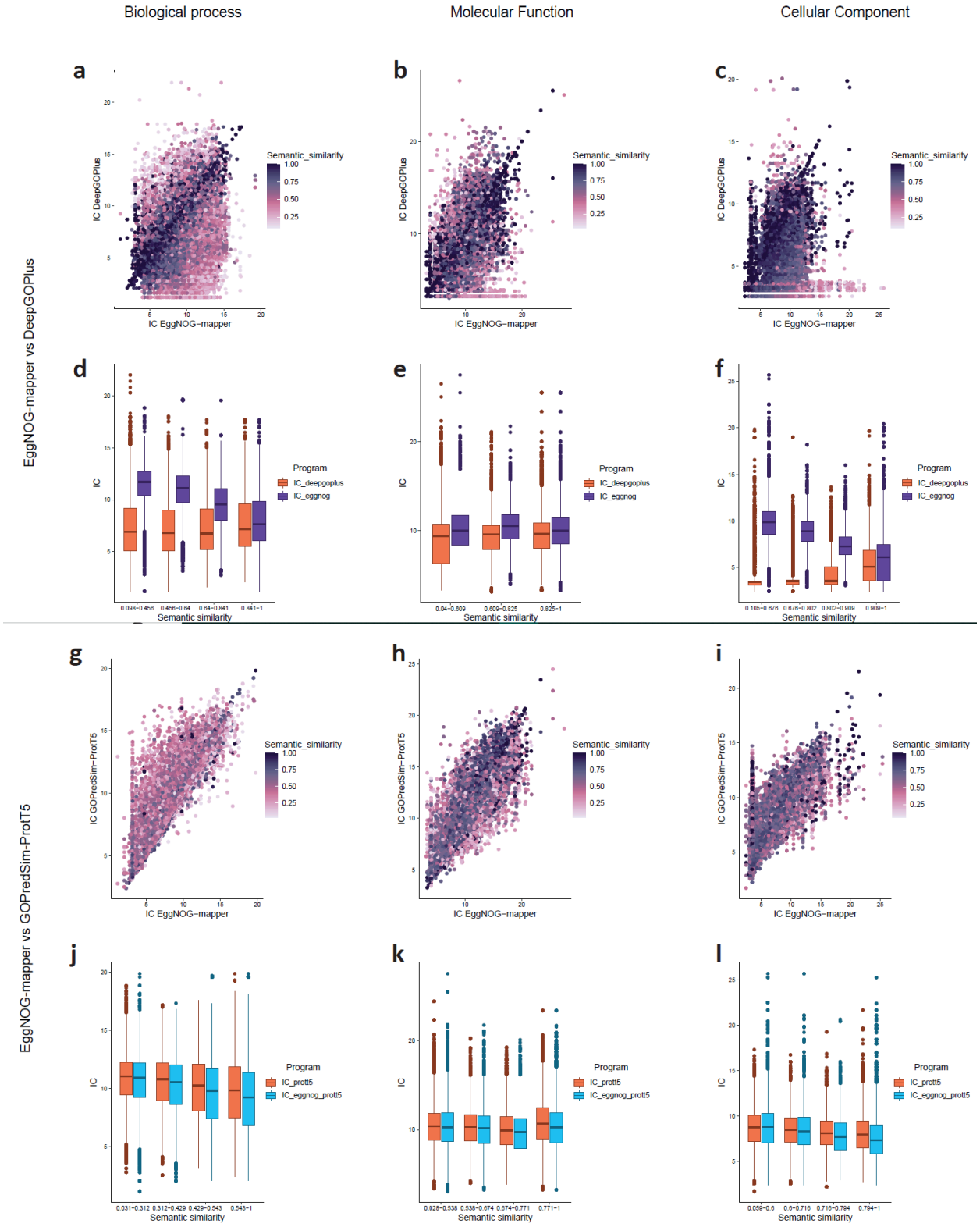
Supp. Fig. 2: Number of GO terms of each GO category per gene in all tested functional annotation methods.

Suppl. Fig. S3



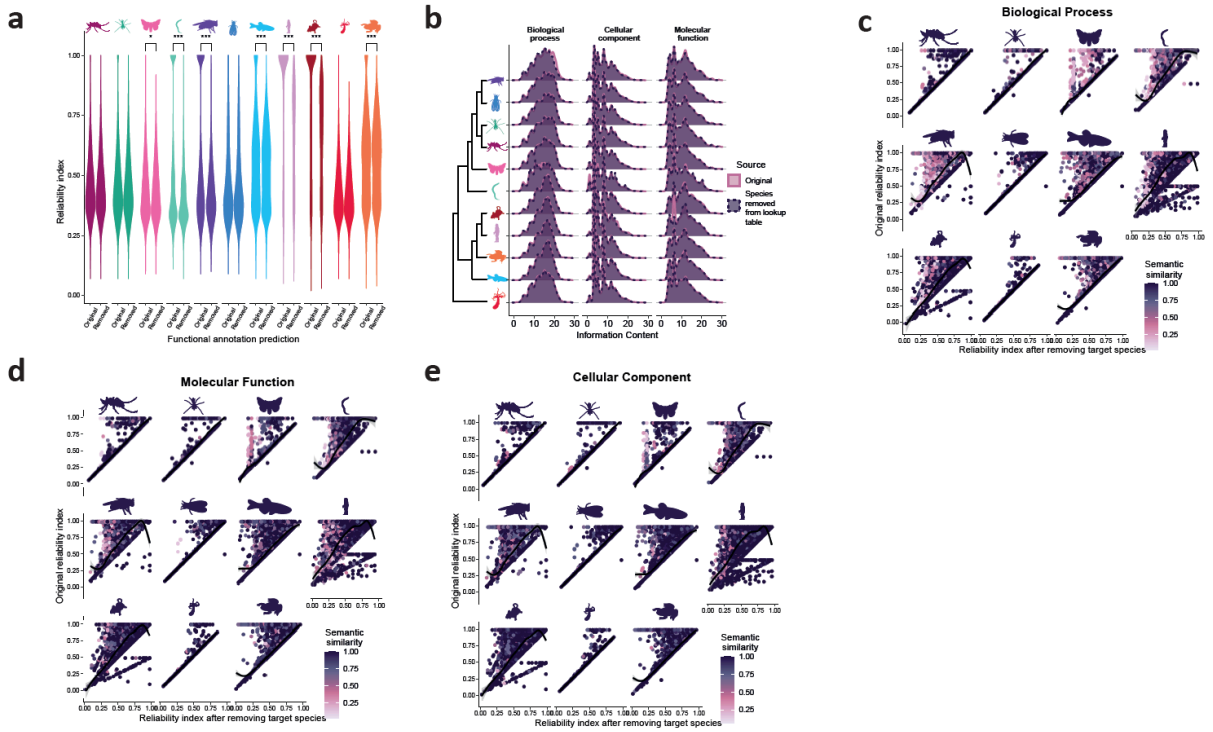
Supp. Fig. 3: Variation in the semantic similarity of the functional annotation using DeepGOPlus and GOPredSim (ProtT5 and SeqVec) against eggNOG-mapper for of a random subset of 100 genes.

Suppl. Fig. S4


Biological process
Molecular Function
Cellular Component
EggNOG-mapper vs DeepGOPlus
EggNOG-mapper vs GOPredSim-ProtT5
a
b
c
d
e
f
g
h
i
j
k
l
IC DeepGOPlus
IC EggNOG-mapper
IC GOPredSim-ProtT5
Semantic_similarity
1.00
0.75
0.50
0.25
IC
Semantic similarity
Program
IC_deepgoplus
IC_eggnog
IC_prott5
IC_eggnog_prott5

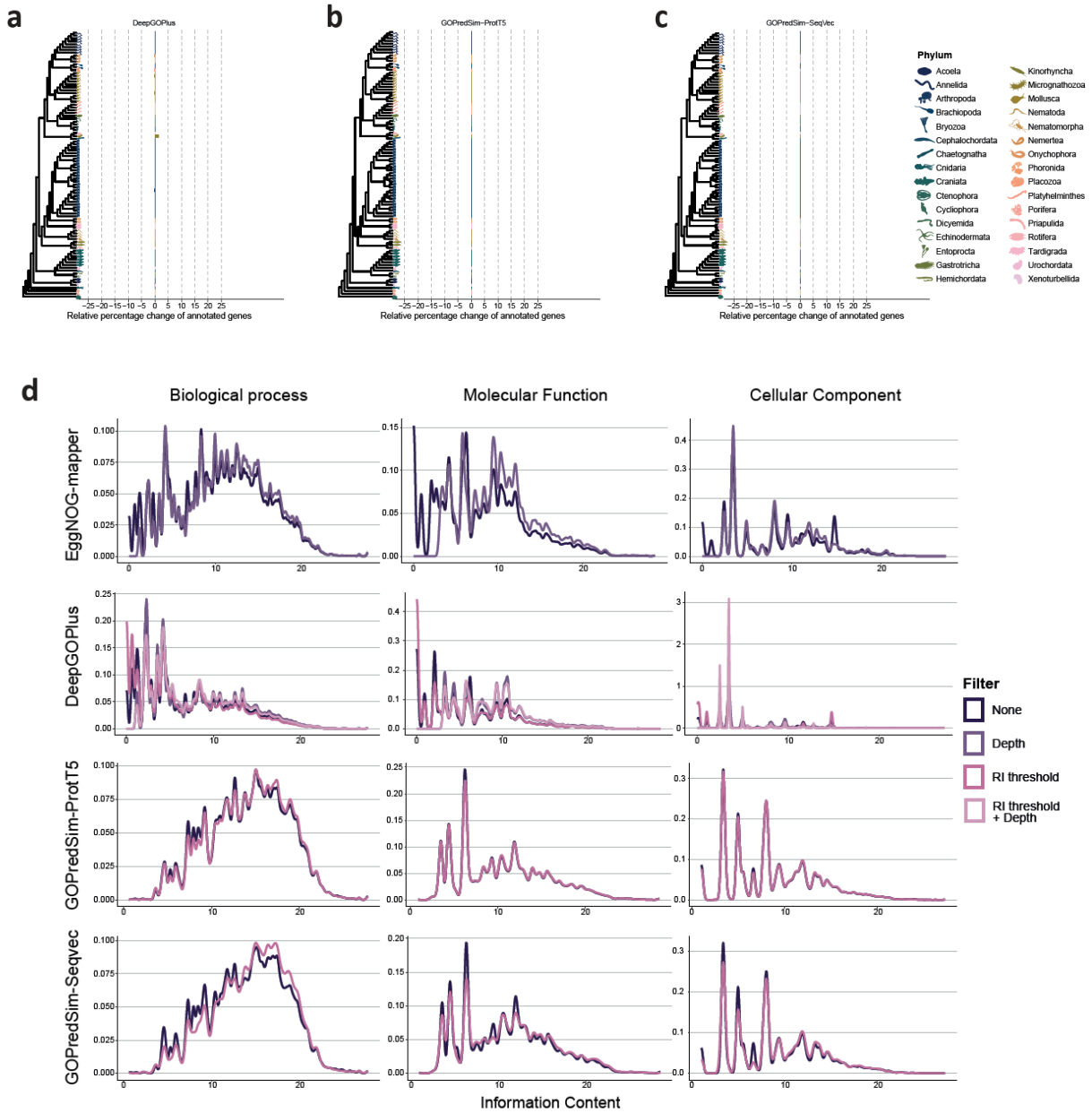
Supp. Fig. 4: Relationship between the difference of information content of GO terms inferred by the tested method and eggNOG-mapper and the semantic similarity between both annotations. a,d, Biological process GO category for the eggNOG-mapper and DeepGOPlus comparison. **b,e,** Molecular function GO category for the eggNOG-mapper and DeepGOPlus comparison. **c,f,** Cellular component GO category for the eggNOG-mapper and DeepGOPlus comparison. **g,j,** Biological process GO category for the eggNOG-mapper and GOPredSim-ProtT5 comparison. **h,k,** Molecular function GO category for the eggNOG-mapper and GOPredSim-ProtT5 comparison. **i,l,** Cellular component GO category for the eggNOG-mapper and GOPredSim-ProtT5 comparison.

Suppl. Fig. S5



Supp. Fig. 5: Comparison between GOPredSim-ProtT5 GO annotations obtained for 11 model organisms when removing their sequences from the GOPredSim lookup dataset compared to the original annotation. a, Reliability index. **b,** Information content. **c,** Biological process semantic similarity. **d,** Molecular function semantic similarity. **e,** Cellular component semantic similarity.

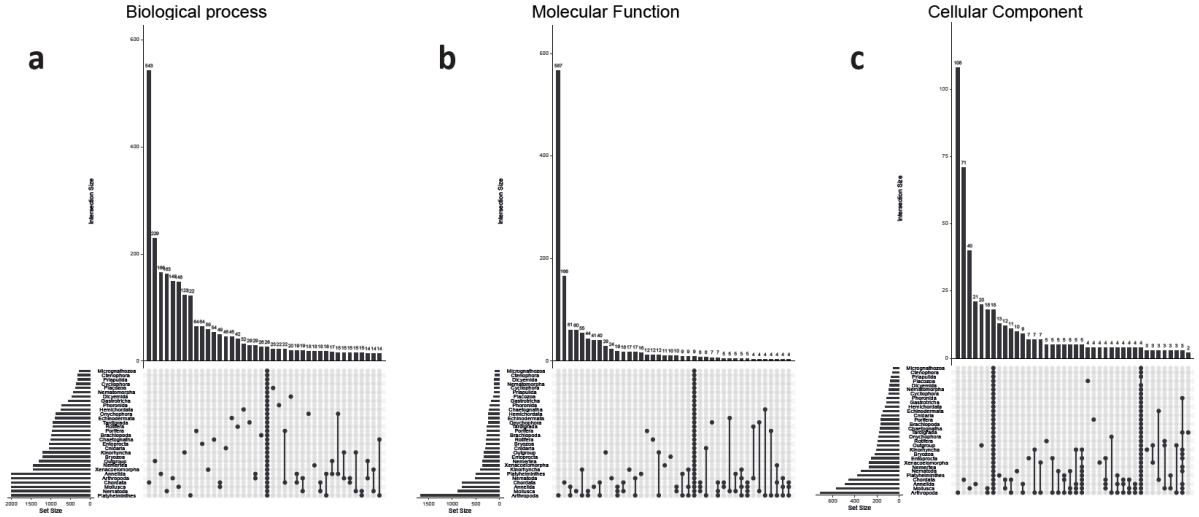
Suppl. Fig. S6



Supp. Fig. 6: Per-gene functional annotation differences when using all isoforms or just the longest isoform. a, Percentage change of genes annotated by DeepGOPlus when using all isoforms compared to genes annotated using the longest isoform (unfiltered dataset). **b,** Percentage change of genes annotated by GOPredSim-ProtT5 when using all isoforms compared to genes annotated using the longest isoform (unfiltered dataset). **c,** Percentage change of genes annotated by GOPredSim-SeqVec when using all isoforms compared to genes annotated using the longest isoform (unfiltered dataset). **d,** Comparison of the Information

Content (IC) of the per-gene isoform-based annotations made by the different methods, for every filtering and GO category.

Suppl. Fig. S7



Supp. Fig. 7: UpSet plot showing the GO terms intersection among all animal phyla. a, Biological process. **b,** Molecular function. **c,** Cellular component.