Cross-phyla protein annotation beyond sequence similarity through structural prediction and alignment

Fabian Ruperti[1, 2◕], Nikolaos Papadopoulos[1◕], Jacob Musser[1], Detlev Arendt[1,3*]

1 Developmental Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

2 Collaboration for joint Ph.D. degree between EMBL and Heidelberg University, Faculty of Biosciences 69117 Heidelberg, Germany.

3 Centre for Organismal Studies (COS), University of Heidelberg, 69120 Heidelberg, Germany

◕: These authors contributed equally to this work. *: arendt@embl.de

Abstract

Background: Annotating protein function is a major goal in molecular biology, yet experimentally determined knowledge is often limited to a few model organisms. In non-model species, the sequence-based prediction of gene orthology can be used to infer function, yet this approach loses its predictive power with longer evolutionary distances. Here we propose a pipeline for the functional annotation of proteins using structural similarity, exploiting the fact that protein structures are directly linked to function and can be more conserved than protein sequence.

Results: We propose a pipeline for the functional annotation of proteins via structural similarity (CoFFE) and use this to annotate the complete proteome of a sponge, an early-branching animal uniquely positioned for inferring the early history of animal cell types. CoFFE accurately identifies proteins with known homology in $> 90\%$ cases, and annotates an additional 50% of the proteome beyond sequence-based methods. Using this, we uncover new functions for sponge cell types, including extensive FGF, TGF and Ephrin signalling in sponge epithelia and redox metabolism and control in myopeptidocytes. Notably, we also annotate

genes that arose via horizontal gene transfer and likely participate in digesting algal and plant cell walls, which are specific to the enigmatic sponge mesocytes.

Conclusions: Our work demonstrates that structural similarity is more powerful to generate functional prediction and annotation than sequence alone, bridging longer evolutionary distances. We anticipate that with a growing number of experimental and predicted structures the approach outlined here will boost annotation efficiency especially for phyletically distant non-model species, and will facilitate discovery in numerous -omics datasets.

Keywords

Functional annotation, proteins, *Spongilla*, scRNA-seq, structural similarity, structure conservation, ortholog conjecture

Background

For many types of high-throughput experiments in molecular biology, knowing the protein function of the assayed genes is crucial to interpret the data. Since protein functional studies are limited to a few model species, amino acid sequence similarity has been used to infer the function of protein homologs. This approach exploits the observation that proteins with similar structures often have similar sequences, and that proteins with similar functions often have similar structures [1, 2]. However, homology detection over longer evolutionary distances remains challenging owing to the decay of protein similarity that abolishes evidence of historical continuity. This presents a severe bottleneck for inferring protein function across a wide expanse of the tree of life, particularly in distant organisms where many proteins fall in the "twilight zone", sharing a sequence identity only between 10-20% with proteins in characterised models [3, 4].

A way to venture more deeply into the twilight zone is to use structural similarity for functional annotation transfer, as structures are more conserved in evolution [5]. Until recently, this was not feasible since predicting protein structure from amino acid sequence required prior inference of sequence homology [6]. This has changed with the advent of AlphaFold [7], a deep learning AI system that can predict de novo protein structures with atomic resolution, together with novel approaches for identifying similar structures in large databases [8]. Entire proteomes can now be aligned to known structures, either experimentally solved or predicted by AlphaFold, and aligned to structures with characterised functions from model systems.

In this work, we predicted structures for the entire proteome of the freshwater sponge *Spongilla lacustris*, an early-branching animal, and aligned them against available structural databases. We show a 50% expansion of previous functional annotation, and revisit recent single-cell RNA-sequencing data [9] to discover additional aspects of sponge cell biology, such as extended cell signalling in pinacocytes, redox metabolism and control in myopeptidocytes, and polysaccharide digestion as a key function of the previously uncharacterized mesocytes.

Results

A protein structure-based pipeline enriches functional annotation for *Spongilla lacustris*

We created a structure-based pipeline for functional annotation transfer, which we refer to as **CoFFE** (**Co**lab**F**old-**F**oldSeek-**E**ggNOG). Instead of using amino acid sequences directly, we predict protein structures, align them against structural databases, and transfer the functional annotation of the best hits to the queries (for an overview, see Suppl. Fig. 1; details in Methods). As a test case, we chose to annotate proteins in the freshwater demosponge *Spongilla lacustris*, an early-branching animal with only about 20 cell types, making it a unique model for understanding the origins of specialised animal cells [9].

The *Spongilla lacustris* proteome contains 41943 predicted proteins, including protein isoforms. As a first step, we used the ColabFold [10] pipeline to predict three-dimensional structures of all 41943 predicted *Spongilla* proteins (all structures and metadata deposited to ModelArchive, see Methods). This represents the entirety of the *Spongilla* proteome except for 11 proteins, which were too long ($>$ 2900 amino acids) to be predicted by the available hardware. Confidence of predicted protein structures was assessed by calculating average pLDDT (predicted local distance difference) values. Notably, average pLDDT values for Spongilla protein structures were similar to those of more well-characterised animal models (Fig. 1A).

Next, we used the predicted protein structures as queries to search with FoldSeek [8] against AlphaFoldDB [11], SwissProt [12], and PDB [13]. After removing lower-quality matches we transferred the functional annotation of the best hits to the sponge proteins, producing annotations for slightly more than 60% of the proteome (25232 proteins). This represents an approximately 50% increase in annotated proteins compared to the state-of-the-art orthology pipeline EggNOG-mapper [14]. This moves Spongilla into a comparable level of functional annotation as *C. elegans* (70% at least described), one of the most intensively studied model species (Fig. 1B).

Exploration of the CoFFE results (Supplemental Note A) in terms of the distributions and correlation of different measures (pLDDT, structure/sequence bit score, perc. identity, etc.), as well as various protein characteristics (hydrophobicity, gravy index, polarity, etc.), did not reveal specific cause for concern and suggested little correlation between protein structure

prediction quality and chemical properties of the proteins.

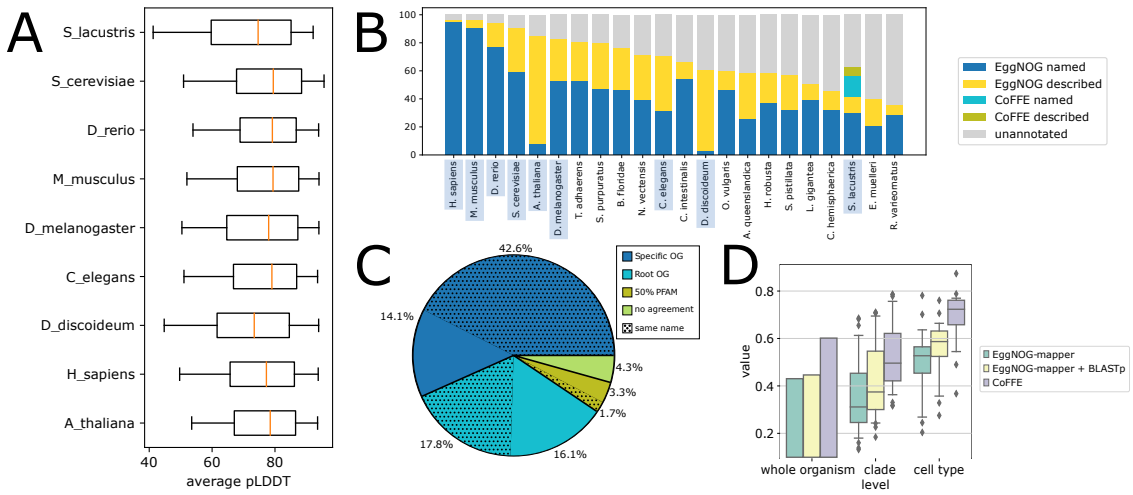


Fig 1. Structural prediction and alignment of the *Spongilla* proteome A) Distribution of average pLDDT per protein structure of predicted proteomes from common model species in comparison to *Spongilla lacustris*. B) Percentage of genes that receive functional annotation from EggNOG, and the enrichment of the *Spongilla lacustris* functional annotation by the CoFFE pipeline. Highlighted organisms appear in A). C) Overlap between EggNOG and CoFFE annotations. **Ortholog**: proteins identified as belonging to the same orthology group in the most recent common ancestor in the EggNOG database. textbfProtein fam.: proteins identified as belonging to the same eukaryote orthology group in the EggNOG database, indicating annotations represent homologs in the same gene family. 50% PFAM: half of the sequence-based PFAM domains are shared. No agreement: the CoFFE and EggNOG annotations differ. Subcategories with "same name" denote the fractions where EggNOG and CoFFE returned the same preferred name for a protein. D) Percent enrichment of *Spongilla* annotation at different levels with EggNOG-mapper (+ BLASTp) as the baseline.

CoFFE annotations agree with sequence-based annotation transfer

As a first step towards validating our pipeline, we compared CoFFE annotations to sequence-based annotations generated by EggNOG-mapper. This method annotates proteins by searching precomputed gene trees in the eggNOG orthology database and assigns query proteins to fine-grained orthology groups. A total of 16589 proteins were annotated by both

pipelines. Overall, 90.6% of proteins were annotated to homologs (Fig. 1C), including 56.7% orthologs and an additional 33.9% representing proteins in the same gene family. In the remaining cases, most (5% overall) shared PFAM domains in common [15], bringing the percentage of proteins with at least minimal agreement between CoFFE and EggNOG-mapper to ¿ 95%. These results indicate that CoFFE can generally be trusted to find true homologs or proteins related by shared domains.

We also compared our results to "legacy" annotations from the recently published Spongilla cell type atlas, which used BLASTp to supplement EggNOG-mapper annotations [9]. Compared to this combined sequence-based approach, CoFFE annotates more proteins organism-wide (60% to 40%). More importantly, CoFFE markedly improved the proportion of annotated cell type-specific markers (70%) compared with sequenced-based approaches (%56; Fig. 1D and Suppl. Note XX). This indicates CoFFE will be valuable for exploring cell type-specific functions in non-model organisms.

CoFFE expands annotation of signalling pathways in Sponge pinacocytes

Pinacocytes are contractile epithelial cells that line the sponge canal system, playing important roles in morphogenesis, barrier formation, and sponge whole-body contractions [16, 17]. Musser et al. showed pinacocytes upregulate genes involved in TGF/BMP and Wnt signalling, key pathways regulating development and morphogenesis in multicellular animals [18, 19]. Supporting this, CoFFE identified additional pinacocyte marker genes in each of these pathways.

In the FGF pathway, sequence-based annotations indicated expression of Fgf receptors in incurrent pinacocytes and apendopinacocytes, along with FGF mediator Frs and regulator Grb2, which are expressed in the entire family of pinacocyte cell types (Figure 2A). Extending this, CoFFE identified GAB1 and GAB2 (GBR2-associated-binding protein 1/2) as well as PTPN11/SHP2, which are necessary for signal transmission into the cell [20, 21]. Strikingly, CoFFE also detected the FGF ligand in *Spongilla*, which was not found using sequence-based approaches, revealing it is most highly expressed in incurrent pinacocytes, lophocytes, and sponge neuroid cells. Structural superposition of the protein with its best FoldSeek hit (UniProtID: P48804 , *Gallus gallus* FGF4) revealed an extensive alignment of large parts of the protein, with a RMSD of 0.89 over 543 atoms despite a sequence similarity of only 11.8% (Fig.

2B).

Musser *et al.* also described multiple genes involved in TGF-β signalling in the pinacocyte family, including Tgfbr1, Acvr, Smad and Smurf [22]. CoFFE extends the list of known actors in pinacocyte TGF signalling, adding important ligands such as INHBE, CHRDL1 or KCP.

Lastly, Ephrin and the ephrin receptor (Eph) are membrane-anchored signalling molecules that mediate communication between adjacent cells [23, 24]. Whereas multiple ephrin receptors were detected via sequence similarity, ephrin itself was only found in the *Spongilla* proteome using highly sensitive HMM profile searches [25, 26]. CoFFE annotates a *Spongilla* gene with differential expression in various pinacocytes as a homolog of *Caenorhabditis elegans* Efn-3 (UniProtID: Q19475). Although these proteins share only 22% sequence identity, the aligned structures achieve an RMSD of 1.59 over 580 atoms (Fig. 2C). A separate Hmmer search [27] using the ephrin Pfam profile (PF00812) picked up the same gene, suggesting CoFFE is at least as sensitive as curated HMM profile searches.

FGF, TGF-beta as well as Ephrin interestingly exhibit converging downstream signalling pathways, including PI3K/Akt, ERK/MAPK1, JNK or RhoA/ROCK, responsible for cell growth, differentiation, migration and cytoskeletal organisation [23, 28, 29]. Together, CoFFE and sequence-based methods identified the principal proteins involved in the downstream pathways, showing they are broadly expressed across most *Spongilla* cell types, consistent with their diverse functional roles. This example highlights that CoFFE is able to vastly extend the information about already hypothesised signalling pathways in cell types by annotating proteins showing extensively divergent sequences.

Redox metabolism and control in myopeptidocytes

The mesohyl of sponges is a collagenous, dynamic tissue forming large parts of the body between pinacocytes and the feeding choanocyte chambers [30]. Musser et al. identified five novel mesenchymal cell types [9] in *Spongilla*. Among them, myopeptidocytes are an abundant, largely uncharacterised cell type, forming long projections that contact other cells. Sequence-based annotations suggested the generation and degradation of hydrogen peroxide in myopeptidocytes, taking into account the expression of dual oxidase (Duox1), its maturation factor (DuoxA), and Catalase (Cat) [31]. Reduced states of metal ions, such as Cu^+, imported by copper transporters such as Slc31a2, are able to react with H_2O_2 (Fenton reaction) [32]

leading to the generation of hydroxyl radicals, prominent reactive oxygen species (ROS) in cells (Fig. 3), supported by the presence of Cyba [33]. However, further roles of ROS metabolism and function had so far remained unclear.

CoFFE identifies potential iron-based generation of H_2O_2 in myopeptidocytes by detecting Ferric-chelate reductase (FRRS1 and FRRS1L), which recycles Fe^{3+} to its reduced state. At the same time, CoFFE proposes downstream metabolism and function of the ROSs: Disulfide oxidoreductase (DsbB) as well as Flavin carrier protein (FLC) both play a role in oxidative protein folding [34, 35]. NmrA-like acts as a redox sensor in the cell [**?**]. Consistent with a possible redox regulation role, myopeptidocytes express a range of additional ROS metabolising and responsive proteins: Sulfiredoxin 1 (SRXN1) promotes resistance against oxidative stress damage [36] while AP endonuclease 1 (APEX1) protects against ROS induced DNA damage [37, 38]. We also identified a glutathione S-transferase member (HPGDS) as well as an additional member of methionine sulfoxide reductases (msrA), which are important enzymes involved in the repair of proteins damaged by oxidative stress [39, 40]. Finally, myopeptidocytes also express the LDL-receptor LRP2, which is known to bind apolipoprotein M (ApoM) [41]. Notably, CoFFE also detected ApoM in myopeptidocytes, suggesting a role in lipid metabolism. Consistent with this, Musser et al. documented round inclusions in myopeptidocytes that may represent lipid droplets. Lipid metabolism and redox control is tightly coupled in peroxisomes which are responsible for beta-oxidation of long-chain fatty acid [42].

Finding digestive roles for the uncharacterised mesocytes

Electron microscopy studies of the *Spongilla* feeding behaviour have demonstrated that prey endocytosis is initially carried out by pinacocytes and choanocytes. Food particles are passed on to digestive and phagocytic cells in the mesohyl, which are then thought to release undigested pellets into the excurrent canals [43]. These food particles often represent bacteria and algae, which are protected by polysaccharide and glycoprotein cell walls [44]. Breaking down cell walls would allow the sponge to extract additional energy and boost filter-feeding efficiency. Mesocytes are newly discovered sponge cell types and functionally uncharacterised medium-sized cells, whose name refers to their sparse location in the mesenchymal mesohyl [9]. Our new CoFFE annotations of previously unidentified mesocyte markers include genes such as expansin (yoaJ), glucan endo-1,3-beta-glucosidase (BG3), and spore cortex-lytic enzyme (sleB),

Annotation	Annotation origin	Taxonomy	Function
Aminohydrolase	legacy EggNOG	Bacteria	hydrolase acting on aminogroups
GMHA	legacy EggNOG	Bacteria	xanthan biosynthesis
Metallopeptidase M20	legacy EggNOG	Bacteria	metallopeptidase
yoaJ	refined by CoFFE	Bacteria	cell wall degradation
sleB	refined by CoFFE	Bacteria	hydrolase activity, cell wall organization
T5orf172	new annotation	Bacteria	hydrolase activity
BG3	new annotation	Streptophyta	cellulase activity
cellulase A family member	new annotation	Streptophyta	cellulase activity
Endochitinase	new annotation	Streptophyta	endochitinase
Chitinase class I	new annotation	Streptophyta	endochitinase

Table 1. Hydrolytic enzymes in *Spongilla* mesocytes.

all hydrolases that specifically degrade cell walls, cellulase, chitin, other polysaccharides [45–48] and proteins (Table 1). It is thus tempting to speculate that mesocytes represent cells specialised to digest polysaccharides that are otherwise difficult to hydrolyse.

Interestingly, enzymes such as cellulases are famously absent from the metazoan digestive repertoire [45]. CoFFE annotations of these mesocyte marker genes reveal many have most structural similarity to bacteria and plant/algal proteins, suggesting these may have arisen via horizontal gene transfer (HGT). In line with this, recent evidence from nematodes showed they expanded their dietary options via the HGT of a cellulase gene of eukaryotic origin [49]. Furthermore, a previous study analysing origins of genes in the marine demosponge Amphimedon queenslandica identified a high proportion of polysaccharide hydrolases and metallopeptidases originating from HGTs [50].

Analysis of the GC-content and codon usage of genes shown in Table 1 compared to the entire Spongilla transcriptome revealed no significant deviation for putative HGT genes (Figure 3A), consistent with the hypothesis that they are present in sponge genomes. To confirm this and explore the relative timing of HGT events, we searched for homologs of all proteins shown in Table 1 in openly available genomes of sponge species (Figure 3B). We detected every protein

in at least one other sponge. Corroborating previous studies, we identified hydrolases present in nearly all examined sponges. The only protein present in all analysed sponges is a metallopeptidase, similar to that detected in ctenophore comb jellies [51], indicating this transfer may have occurred prior to the last common ancestor of all animals. Strikingly, we found that three hydrolases were shared by only freshwater sponges, suggesting mesocyte specific functions may have expanded via HGT events after colonisation of new freshwater environments. Finally, the presence of predicted chitinases is intriguing, as chitin has been shown to be an important structural component of *Spongilla lacustris* [52]. This finding suggests another possible role for mesocytes as structural remodelers of the sponge skeleton made of polysaccharides.

Identifying well-folded, unannotated proteins in *Spongilla lacustris*

Using the CoFFE pipeline we annotated a total of 26344 out of 41943 predicted proteins. The remaining 15599 unannotated proteins were relatively short, and may represent incomplete fragments, untranslated sequences, or de novo lineage-specific genes. Notably, we found XX unannotated proteins with a pLDDT score greater than 70, indicating well-folded structures. Although many of these had poor FoldSeek alignments falling below our accepted bit score threshold, 316 unannotated and well-folded predicted proteins had no hit using FoldSeek. Manual inspection revealed that the overwhelming majority of these are predicted to be long helices, except for 35 non-helical structures. To ensure these hadn't somehow eluded sequence similarity searches, we used NCBI BLASTp to identify potential homologs in the NR database, even very remote ones (Suppl. Tabel 1). Seven of the sequences find no annotation nor get significant PFAM domain hits, but are broadly expressed in sponge cell types, presenting prime candidates for truly novel structures.

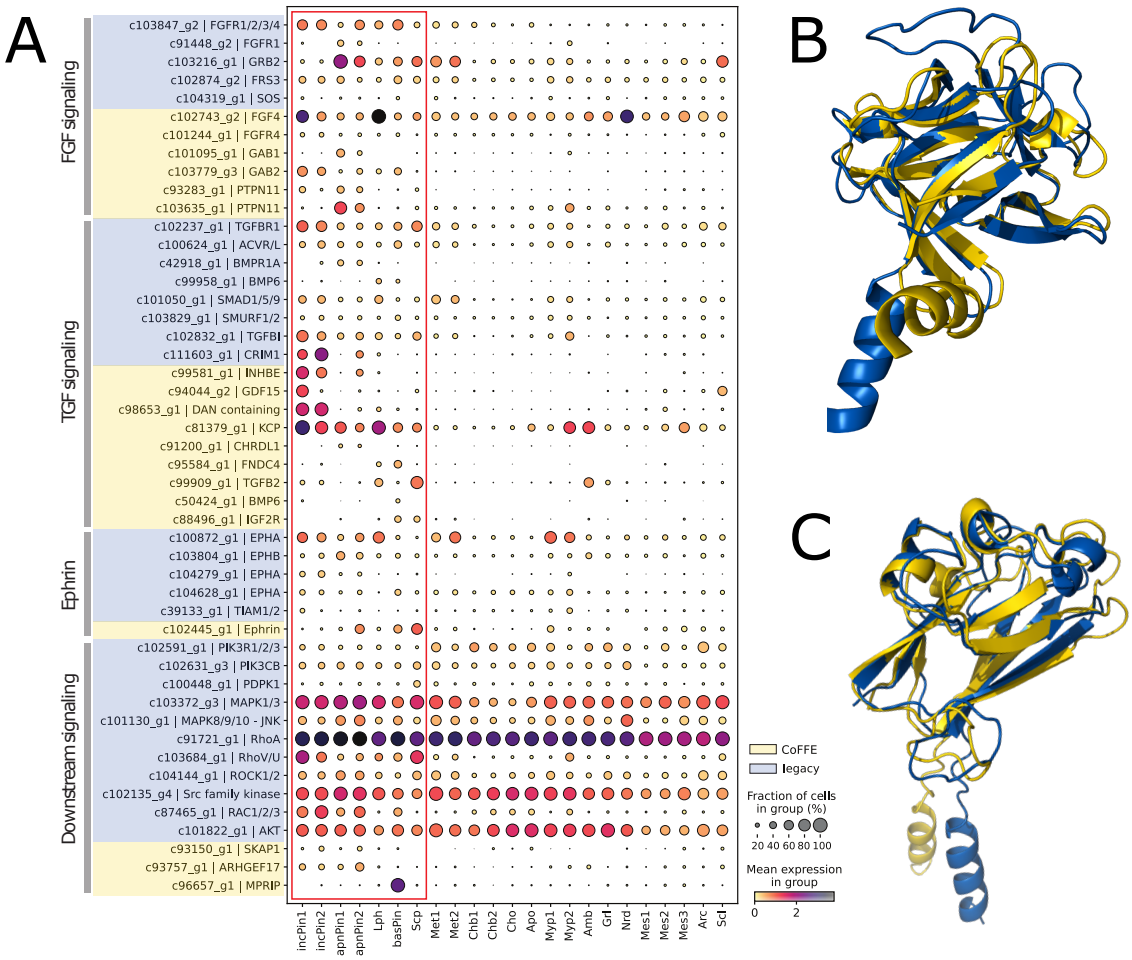


Fig 2. Signalling pathways in *Spongilla* pinacocytes A) Dotplot of pinacocyte signalling and effectors. Cell types of the pinacocyte family are highlighted by a red square. Genes on blue background are annotated by sequence-based methods ('legacy'). Genes on yellow background are annotated by CoFFE. B) Structural alignment of *Spongilla* FGF (blue, 61 - 230 aa) with *Gallus gallus* FGF4 (UniProtID: P48804) (yellow, 54 - 194 aa) (RMSD = 0.89 over 543 atoms). C) structural alignment of *Spongilla* ephrin (blue, 1 - 153 aa) with *C. elegans* efn-3 (UniProtID: Q19475) (yellow, 29 - 179 aa) (RMSD = 1.59 over 580 atoms). In both cases the structural similarity is apparent despite their low sequence identity of 11.8% and 22% respectively. Superpositions were created using the `super` command in PyMOL (V2.3.5) []. Cell type abbreviations: incPin - incurrent pinacocytes; apnPn - apendopinacocytes; lph - lophocytes; basPin - basopinacocytes; scp - sclerophorocytes; met - metabolocytes; chb - choanoblasts; cho - choanocytes; apo - apopylar cells; myp - myopeptidocytes; amb - amoebocytes; grl - granulocytes; nrd - neuroid cells; mes - mesocytes; arc - archaeocytes; scl - sclerocytes.

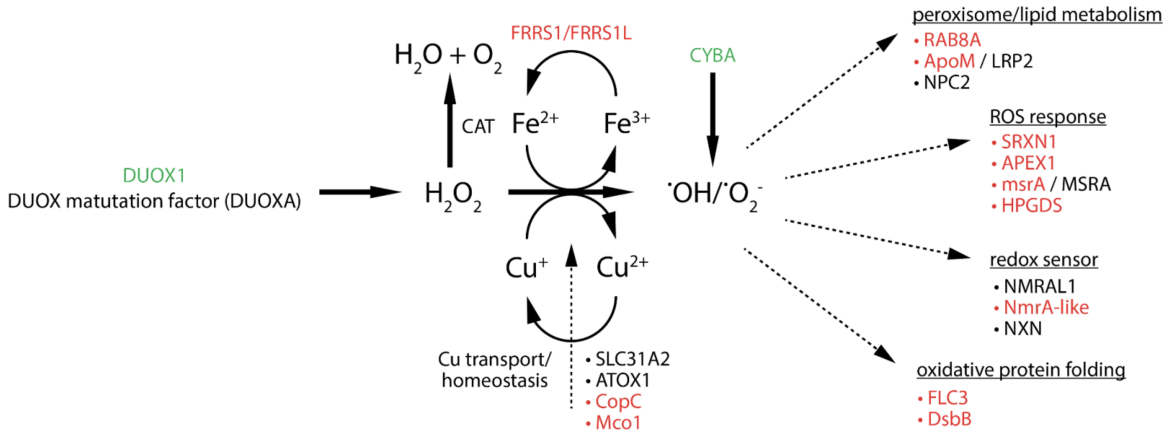


Fig 3. ROS metabolism and redox-control in Myopeptidocytes Myopeptidocytes differentially express multiple genes involved in redox control and ROS defence. Genes in black and green have been annotated using sequence based methods. Green proteins are known to interact. Genes in red have been identified using CoFFE.

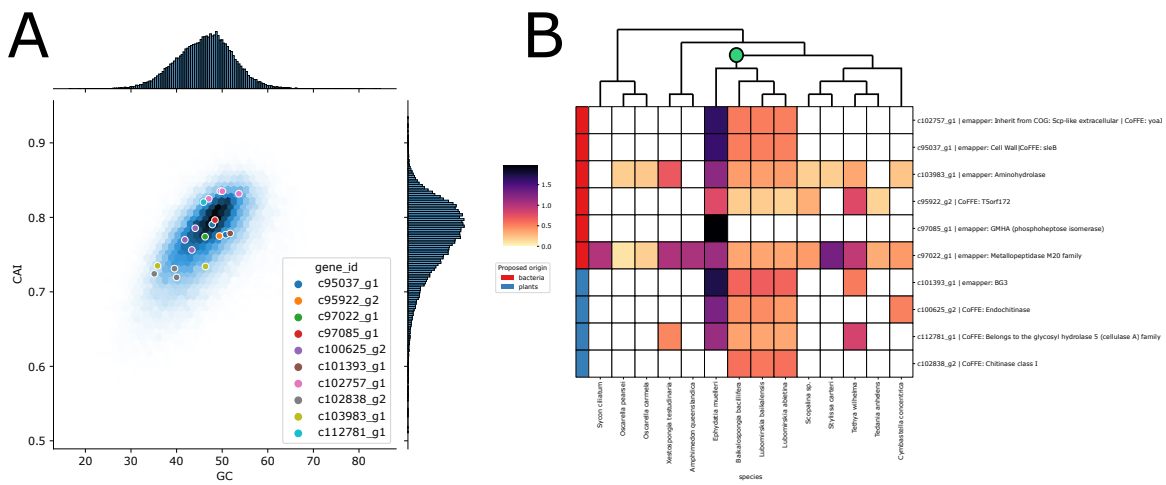


Fig 4. Horizontal gene transfer of mesocyte marker genes A) Comparison of GC-content and codon usage of mesocyte marker genes of non-metazoan origin with whole *Spongilla* transcriptome. B) Homology search of mesocyte marker genes in various sponge species. The green dot denotes the last common freshwater sponge ancestor.

Conclusions-Discussion

By predicting the protein structures of the entire proteome of the freshwater sponge *Spongilla lacustris* and aligning them against all currently available structures, we were able to increase the annotation of the proteome by approx. 50% (60% of the proteome in total, up from 40% when only using sequence information). For proteins with annotation from both sources, we found that in more than 90% of cases, sequence-based and structure-based annotation agreed at least to the protein superfamily level.

To explore the usefulness of structural similarity for functional annotation we revisited the annotation of marker genes previously identified from *Spongilla* scRNA-seq data. In the epithelial pinacocyte family, this significantly improved sequence-based annotation via the detection of key players in FGF, TGF-beta, as well as ephrin signalling. For the digestive peptidocyte family, we were able to infer complex modes of redox regulation as an important function of the myopeptidocytes. Finally, we detected polysaccharide and protein hydrolyzing enzymes of bacterial and plant/algal origin in the enigmatic mesocytes which strongly suggests that these digest cell walls of endocytosed prey. The mesocyte example also reveals cell types whose functional role is largely defined by genes acquired via HGTs from other pro- and eukaryotic organisms. This is remarkable both from a cell type evolutionary perspective and because such innovation had previously escaped our notice due to extensive sequence divergence following the HGT event.

The lack of reliable functional annotation has so far been a major bottleneck in the analysis of non-model species, in particular those separated from traditional models by large phylogenetic distances. Here, we demonstrate that by exploiting the evolutionary conservation of protein structure it is possible to dramatically improve protein functional annotations in non-model species. The pipeline described here can thus be used to query highly informative candidate genes from proteomics or single-cell -omics experiments. Although protein structural predictions for an entire proteome might be outside the technical capabilities of many labs due to the cost of procuring the necessary hardware, plans to expand the AlphaFold database [11] have been announced. Given the power of AlphaFold and the huge potential of expanded annotations, it is reasonable to expect that in the future protein sequences deposited in public databases will automatically receive predicted structures, paving the way for unique insights

into biological functions across the tree of life.

Materials and methods

In all boxplots, the box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers represent the $5 - 95\%$ percentiles.

Sequence-based annotation of the *Spongilla lacustris* proteome

Juvenile freshwater sponges (*Spongilla lacustris*), grown from gemmules, were used for bulk RNA isolation and sequencing. *De novo* transcriptome assembly with Trinity, returned 62180 putative isoforms, covering 95.2% of Metazoan BUSCOs [53]. To identify putative proteins, Transdecoder [54] (version 3.0.1) was used with a minimum open reading frame length of 70 amino acids, resulting in 41945 putative proteins. The longest putative protein per gene ID was kept. The resulting predicted proteome was annotated by EggNOG mapper [14, 55] (v2.1.7, default settings) via the website.

Legacy annotation

Musser *et al.* [9] used the putative proteins to create a phylome by constructing gene/protein trees for each protein (available online). The phylome information was used to refine the assignment of transcripts to genes. In some cases, 3' and 5' fragments of a gene were assigned to two different transcripts. These fragments were merged into the same merged gene name using the gene tree information. Functional annotations were supplemented by EggNOG mapper (v1) and `blastp` searches against human RefSeq (default parameters). This annotation was used in the original *Spongilla* scRNA sequencing publication and is present in the single-cell data. The legacy annotation was used for the single-cell data analysis, but not for the comparison between the CoFFE pipeline and the sequence-based annotation transfer.

Structure-based annotation of the *Spongilla lacustris* proteome

CoFFE and constituent tools

We designed a simple pipeline that goes from sequence to predicted structure to functional annotation (symbolically, **Co**lab**F**old-**F**oldSeek-**E**ggNOG, or CoFFE), and used it to annotate the *Spongilla lacustris* predicted proteome. We used ColabFold to predict structures and subsequently aligned the predicted structures against all currently available (solved and

predicted) protein structures using FoldSeek. Finally, we used structural similarity to transfer annotations from the best structural targets to their corresponding *Spongilla* protein queries. In the following we show an overview of the tools in use and a more detailed description of the CoFFE pipeline.

MMseqs2 [56] is a software suite for sequence-sequence, sequence-profile, and profile-profile search and clustering. It is orders of magnitude faster than BLAST at the same sensitivity and is widely adopted [14].

AlphaFold2 [7] is a neural network-based model that predicts protein three-dimensional structures from sequence, regularly achieving atomic accuracy even in cases where no similar structure is known. AlphaFold is widely considered to have revolutionised the field of structural bioinformatics, greatly outperforming the state of the art in the most recent iteration of the CASP challenge [57].

AlphaFold quantifies prediction confidence by pLDDT, the predicted local distance difference test on the $C\alpha$ atoms. Regions with pLDDT > 90 are modelled to high accuracy; regions with $70 <$ pLDDT < 90 should have a generally good backbone prediction; regions with $50 <$ pLDDT < 70 are low confidence and should be treated with caution; regions with pLDDT < 50 should not be interpreted and probably are disordered.

ColabFold [10] is a pipeline that combines fast homology searches via MMseqs2 [56] with AlphaFold2 [7] to predict protein structures 40 to 60 times faster than the original AlphaFold2.

FoldSeek [8] enables fast and sensitive comparison of large structure databases. FoldSeek's key innovation lies in the appropriate translation of structure states to a small alphabet, thus gaining access to all the heuristics of sequence search algorithms.

The CoFFE pipeline

A visual representation of the CoFFE pipeline can be found in supplement figure XX. To predict structures, we adapted the ColabFold pipeline as outlined in `https://colabfold.mmseqs.com/`

Multiple sequence alignment generation: We downloaded reference sequence databases (UniRef30, ColabFold DB) (databases.sh, adapted from setup_databases.sh) and calculated indices locally (create_index.sh). We were interested in homology detection at the limit of the twilight zone, so UniRef30, a 30% sequence identity clustered database based on UniRef100 [58],

was the adequate choice. Similarly, we indexed the *Spongilla* predicted proteome (spongilla_create_index.sh). We calculated MSAs for each *Spongilla* predicted protein using MMseqs2 [56] (align.sh, adapted from colabfold_search.sh).

Structure prediction: We predicted structures for all *Spongilla* predicted proteins using ColabFold [10] as a wrapper around AlphaFold2.

We split the MSAs in 32 batches and submitted each one to the EMBL cluster system (managed by `slurm` [59]); we used default arguments but added `--stop-at-score 85`. (milot.sh, M. Mirdita, personal communication). The calculations were done on NVIDIA A100 GPUs, on computers running CentOS Linux 7. We used GCC [60] version 10.2.0 and CUDA version 11.1.1-GCC-10.2.0 [61]. We processed the resulting PDB-formatted model files with Biopython's PDB module [62].

Structure search and annotation transfer: Structural search was conducted using FoldSeek. FoldSeek allows fast comparison of large structural databases. We downloaded PDB, SwissProt, and AlphaFold DB (fs_pdb.sh, fs_sp.sh, and fs_afdb.sh, respectively). For each *Spongilla* protein we kept the best-scoring AlphaFold2 model, and used them to construct a FoldSeek database fs_query.sh. These models were then used to search against the three structural databases. For each search we kept the FoldSeek hit with the highest corrected bit score in each database and aggregated the three result tables (AlphaFoldDB, PDB, SwissProt) into one. We imposed a bit score cutoff (1xe-05) on FoldSeek hits based on their bimodal distribution. Annotations of the best hits were gathered from either UniProt via its API [63] or through EggNOG mapper (v2.1.7, default settings) [14] by using the sequences of the best structural hits (pulled by UPIMAPI [64]). To facilitate downstream analysis, we extracted summary tables from each resource type. This procedure can be found in the corresponding notebook. A total of 1401 proteins received sequence annotation but their FoldSeek best hits were below the bit score cutoff.

Differential gene expression in single-cell transcriptomics data

We obtained the processed Seurat file from [9], and downloaded the lists of differentially expressed genes of clusters, cell types, and cell type clades from the supplemental material of the same publication (Supplemental Data S1 to S3; file science.abj2949_data_s1.xlsx; tabs "Diff. exp. 42 clusters", "Diff. Exp. cell types", "Cell type clade genes (OU tests"). The single-cell

data operates on the level of genes, so we transformed the sequence-derived and CoFFE annotations by merging isoform entries and keeping the entry with the best bit score.

We used the legacy annotation included in the file (phylome, EggNOG-mapper, and BLASTp-based), the sequence-derived annotation, and the CoFFE output to propose names for *Spongilla* proteins. We prioritized sequence-derived annotations (legacy annotation, EggNOG preferred name, EggNOG description) and fell back to CoFFE (CoFFE preferred name, CoFFE description) when there were none. We produced dotplots for the top 200 differentially expressed genes in each cell type and manually inspected them. We focused on the terminal (named) cell types; owing to the presence of continuously differentiating stem cells, the single-cell data contains many clusters of maturing or differentiating cells whose expression patterns do not distinguish them from their mature counterparts. Code and detailed explanations are available in the corresponding notebook.

Protein structure visualisation and superposition

In order to visualize predicted structures from the *Spongilla* proteome, we used PyMOL Version 2.3.5 [65]. Superposition with their respective best Foldseek hit was carried out using the `super` command. `super` creates sequence- independent superpositions and is more reliable for protein pairs with low sequence similarity.

Detection of ephrin orthologs in *Spognilla* using hmmer

To validate the ephrin ortholog detected by CoFFE in *Spongilla*, we recapitulated a previous effort to detect ephrins in different species using extensive hmmer protein profile search [26]. For this, we performed a hmmer search (v3.3.2) using the ephrin Pfam profile (PF00812) against the *Spongilla* proteome using default settings.

0.1 Detecting HGT in Spongilla

TO detect HGT we followed the proposal by Degnan [66] and sought to get multiple indications for HGT. In particular: phylogenetic evidence that the candidate gene is more closely related to foreign than to animal genes; genome data showing the candidate gene assembles into a contiguous stretch of DNA with neighboring genes unambiguously of animal origin (this requires, of course, the availability of a sequenced and assembled animal genome; the more

complete the assembly, the more confident the HGT identification); and gene sequence revealing metazoan-like compositional traits, including presence of introns, GC content and codon usage. Where possible, gene expression data showing active transcription of candidate genes in animal cell nuclei can enormously strengthen a case, and also addresses the issue of whether or not the HGT-acquired gene is active in its new genomic context.

We used `https://spongebase.net` to obtain transcriptomes and genomes for 13 sponge species, and obtained a 14th one directly from its repository (`https://bitbucket.org/molpalmuc/tethya_wilhelma-genome/`). We built sequence databases with MMseqs2 version 12-113e3 and searched against them (mmseqs easy-search) with the protein sequences of all isoforms of the HGT candidates. The resulting alignments were filtered to keep the best-scoring alignment per species per gene.

The table of best hits per genome was visualised in Python, and a sponge phylogeny was added manually based on [67] (also A. Riesgo, personal communication).

Acknowledgments

The authors thank Milot Mirdita and Martin Steinegger for discussions about ColabFold usage and technical support; Alexandros Pittis for his technical feedback; the Arendt lab for discussions; the ModelArchive team for their support in publishing the predicted structures; Ana Riesgo for sharing demosponge phylogenies. NP was funded by MSCA individual fellowship 101031984/DeCoDe Platy.

A Extracting quality indicators from the MAF pipeline

To get a better understanding of the relationship between the various metrics, we assessed the correlation between them (Fig. 5). We hoped to find a combination of measures that together would be a robust indicator of prediction quality for MAF.

Most obviously, query length, (uncorrected) structural bit score, and (uncorrected) sequence bit score correlate very strongly. Longer proteins will have more or longer aligned regions with their targets, and as the bit score is additive, will accumulate higher bit scores, regardless of whether the alignment happens in structure or sequence space. It seems these measures can be used interchangeably; out of principle, we would prefer to use query length, as it is the least derived.

A way to overcome the dependence on length is to normalize alignment scores by the length of the aligned region (*corrected bit score - (FS)*). This measure correlates strongly with the percent structure state identity. This is equivalent to percent sequence identity correlating strongly with the bit score in a sequence similarity search, and is not surprising at all. It seems that these measures can also be used interchangeably, but we would prefer to use percent structure identity, since it correlates less strongly with the other measures in the plot.

The rest of the measures (pLDDT, EggNOG bit score of FoldSeek best hit, percent aligned query, and input MSA size) don't correlate too strongly to each other. We decided to use them as-is.

Three things become apparent: first, higher pLDDT scores generally correlate with higher FoldSeek bit scores. This means that FoldSeek does a better job of identifying structural similarity for well-folded proteins. This seems intuitive: the atomic coordinates of well-folded proteins will be more deterministic than disordered proteins, so well-folded orthologs will overlap structurally over their entire length, while heavily disordered orthologs might only overlap structurally over very short patches.

Second, agreement of structural and sequence annotation correlates with higher bit score. This is also ok: the peptides that share the most specific orthogroup are the ones that are recognisably the same, regardless if we look at sequence or structure; peptides that at least share the root orthogroup are the ones with high sequence conservation but lower

The more confident the overlap between structure and sequence annotation, the higher the

average bit score is. This happens in all pLDDT buckets, but has a stronger effect the higher the average pLDDT is.

FoldSeek is more confident in identifying structural similarity for peptides whose structure annotation agrees with the sequence annotation to the most specific orthogroup. Peptides with root orthogroup overlap come in second, and peptides with PFAM domain overlap third. These categories correspond to structures being in the same protein family or superfamily, or at least sharing a couple of domains, indicating three broad levels of structural similarity.

Three, peptides without sequence annotation and peptides where structure and sequence annotation disagree don't have meaningful differences in their FoldSeek bit scores. This happens regardless of average pLDDT.

isoform	q len	#MSA	%seq. id.	bit score	plddt	BLASTp
c88419_g1_i1	154	462	0.337	580.0	85.01	CRN5
c104256_g1_i1	375	163	0.272	785.0	82.42	CRN5
c104256_g1_i2	299	184	0.267	781.0	90.35	CRN5
c104256_g1_i3	336	167	0.262	777.0	88.39	CRN5
c104256_g1_i4	332	158	0.263	796.0	90.26	CRN5
c104256_g1_i6	253	196	0.275	774.0	92.81	CRN5
c94352_g2_i1	127	1725	0.500	299.0	86.06	DD3-3
c89711_g1_i1	144	2424	0.441	380.0	86.73	DD3-3
c89711_g2_i1	75	2340	0.611	210.0	77.48	DD3-3
c112476_g1_i1	179	1450	0.352	434.0	83.19	DD3-3
c103626_g1_i1	670	3392	0.324	2224.0	91.33	DD3-3
c103626_g1_i2	670	3424	0.329	2258.0	91.18	DD3-3
c103626_g2_i1	667	3427	0.326	2215.0	90.01	DD3-3
c102584_g1_i1	578	3328	0.355	2091.0	81.55	DD3-3
c102584_g1_i2	187	1994	0.416	522.0	89.04	DD3-3
c101972_g1_i1	778	3481	0.378	2536.0	85.58	DD3-3
c101972_g1_i2	778	3481	0.378	2534.0	85.58	DD3-3
c101972_g1_i3	778	3481	0.378	2537.0	85.54	DD3-3
c104453_g2_i1	359	368	0.177	662.0	75.80	foll. epith. yolk prot. subunit
c104453_g2_i2	280	356	0.172	658.0	89.90	foll. epith. yolk prot. subunit
c104453_g2_i3	340	378	0.185	634.0	78.61	foll. epith. yolk prot. subunit
c104453_g2_i4	236	374	0.183	594.0	83.73	foll. epith. yolk prot. subunit
c104453_g3_i1	182	341	0.117	298.0	70.54	foll. epith. yolk prot. subunit
c98985_g1_i2	73	1852	0.138	45.0	70.12	XPO5 (Exportin-5)
c88005_g3_i1	130	3213	0.274	363.0	94.66	hemopexin repeat-containing protein
c70965_g1_i1	169	299	0.209	354.0	77.62	pinocchio
c103612_g1_i1	89	1253	0.152	46.0	73.88	kinesin-like protein KIF23
c102197_g1_i3	312	257	0.179	649.0	86.28	bamB
c100331_g2_i1	281	43509	0.219	821.0	82.62	hydralysin-2-like
c93555_g2_i2	77	439	0.109	80.0	87.77	?
c86044_g1_i1	87	7052	0.077	90.0	88.87	?
c83472_g1_i1	98	553	0.177	90.0	84.77	?
c108675_g1_i1	73	3174	0.224	51.0	78.17	?
c100701_g1_i2	69	1	0.123	61.0	81.77	?
c100429_g1_i1	79	43996	0.152	85.0	71.74	?

Table 2. High confidence structures with non-helical appearance. The last column contains a summary of the significant BLASTp results in the `nr` database. Question marks denote

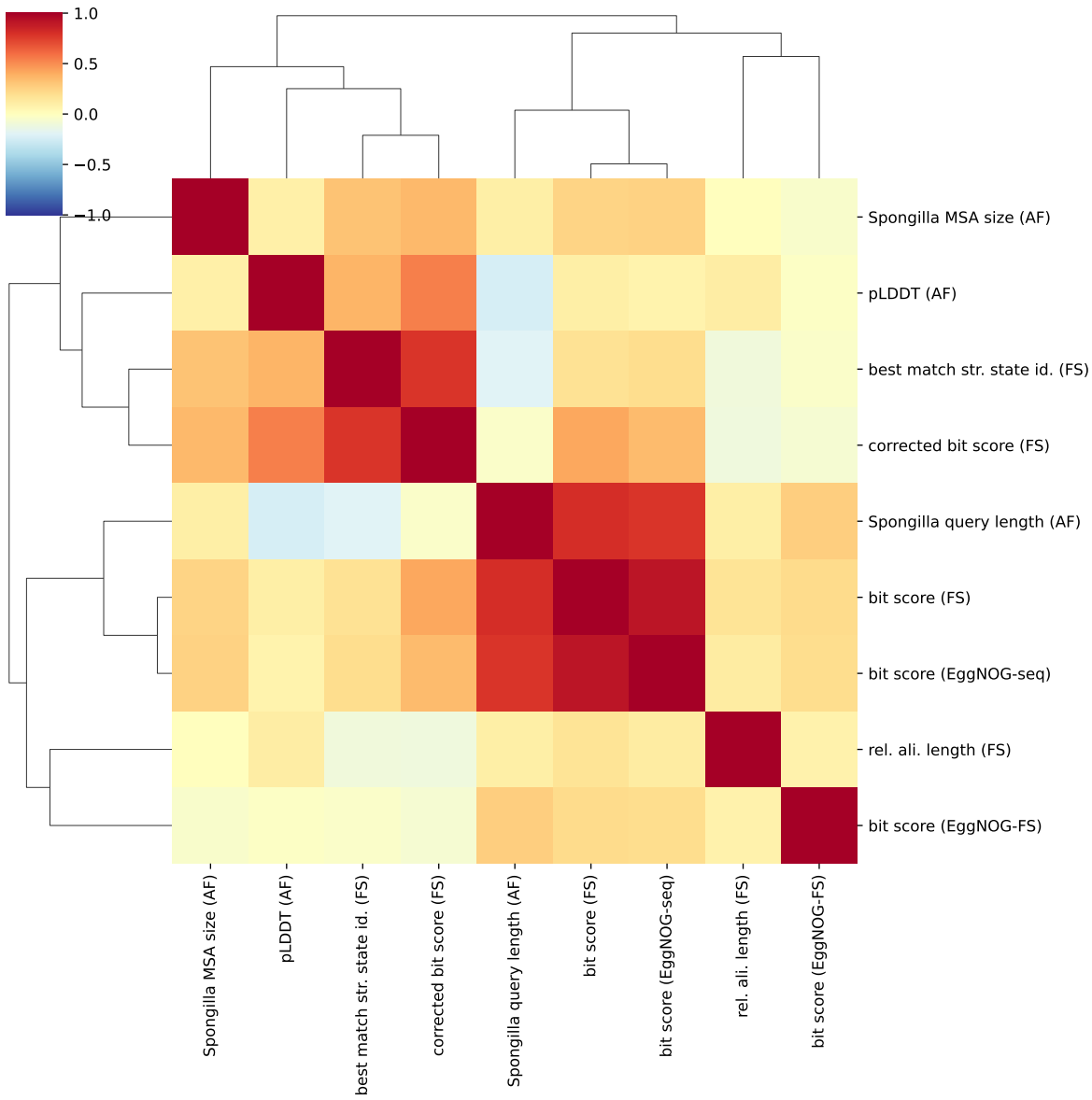


Fig 5. Pairwise correlation of various metrics. *Spongilla MSA size (AF)*: size of the input MSA used for AlphaFold. *plDDT (AF)*: average pLDDT of the predicted structure. *best match seq. id. (FS)*: structure state identity in the alignment between the query *corrected bit score (FS)*: the FoldSeek bit score divided by alignment length. *Spongilla query length (AF)*: the query length of the input peptide. *bit score (FS)*:
bit score of the FoldSeek alignment between the query *Spongilla* structure and the structure of the best FoldSeek target. *bit score (EggNOG-seq)*: bit score of the EggNOG alignment between the query *Spongilla* peptide sequence and the best target sequence. *rel. ali. length (FS)*: percentage of the query (*Spongilla*) structure aligned with the best target structure. *bit score (EggNOG-FS)*: bit score of the sequence alignment between the sequence of the best FoldSeek target and the EggNOG database.

References

1. Bradshaw RA, Neurath H, Walsh KA. Considerations of the concept of structural homology as applied to bovine carboxypeptidases A and B. Proceedings of the National Academy of Sciences. 1969;63(2):406–411.

2. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. The EMBO journal. 1986;5(4):823–826.

3. Rost B. Twilight zone of protein sequence alignments. Protein engineering. 1999;12(2):85–94.

4. Kabir MN, Wong L. EnsembleFam: towards more accurate protein family prediction in the twilight zone. BMC bioinformatics. 2022;23(1):1–20.

5. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins: Structure, Function, and Bioinformatics. 2009;77(3):499–508.

6. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research. 2018;46(W1):W296–W303.

7. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589.

8. van Kempen M, Kim S, Tumescheit C, Mirdita M, Söding J, Steinegger M. Foldseek: fast and accurate protein structure search. bioRxiv. 2022;.

9. Musser JM, Schippers KJ, Nickel M, Mizzon G, Kohn AB, Pape C, et al. Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. Science. 2021;374(6568):717–723.

10. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. bioRxiv. 2022;doi:10.1101/2021.08.15.456425.

11. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of

protein-sequence space with high-accuracy models. Nucleic acids research. 2022;50(D1):D439–D444.

12. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. In: Plant Bioinformatics. Springer; 2016. p. 23–54.

13. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, et al. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic acids research. 2021;49(D1):D437–D451.

14. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular biology and evolution. 2021;38(12):5825–5829.

15. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. Nucleic acids research. 2021;49(D1):D412–D419.

16. Elliott GR, Leys SP. Coordinated contractions effectively expel water from the aquiferous system of a freshwater sponge. Journal of Experimental Biology. 2007;210(21):3736–3748.

17. Nickel M, Scheer C, Hammel JU, Herzen J, Beckmann F. The contractile sponge epithelium sensu lato–body contraction of the demosponge Tethya wilhelma is mediated by the pinacoderm. Journal of Experimental Biology. 2011;214(10):1692–1698.

18. Adamska M, Degnan SM, Green KM, Adamski M, Craigie A, Larroux C, et al. Wnt and TGF-β expression in the sponge Amphimedon queenslandica and the origin of metazoan embryonic patterning. PloS one. 2007;2(10):e1031.

19. Schippers KJ, Nichols SA. Evidence of signaling and adhesion roles for β-catenin in the sponge Ephydatia muelleri. Molecular Biology and Evolution. 2018;35(6):1407–1421.

20. Li H, Tao C, Cai Z, Hertzler-Schaefer K, Collins TN, Wang F, et al. Frs2α and Shp2 signal independently of Gab to mediate FGF signaling in lens development. Journal of cell science. 2014;127(3):571–582.

21. Mossahebi-Mohammadi M, Quan M, Zhang JS, Li X. FGF signaling pathway: a key regulator of stem cell pluripotency. Frontiers in cell and developmental biology. 2020;8:79.

22. Massagué J. TGFβ signalling in context. Nature reviews Molecular cell biology. 2012;13(10):616–630.

23. Darling TK, Lamb TJ. Emerging roles for Eph receptors and ephrin ligands in immunity. Frontiers in immunology. 2019;10:1473.

24. Kania A, Klein R. Mechanisms of ephrin–Eph signalling in development, physiology and disease. Nature reviews Molecular cell biology. 2016;17(4):240–256.

25. Arcas A, Wilkinson DG, Nieto MÁ. The Evolutionary History of Ephs and Ephrins: Toward Multicellular Organisms. Molecular biology and evolution. 2020;37(2):379–394.

26. Krishnan A, Degnan BM, Degnan SM. The first identification of complete Eph-ephrin signalling in ctenophores and sponges reveals a role for neofunctionalization in the emergence of signalling domains. BMC evolutionary biology. 2019;19(1):1–17.

27. Eddy SR. Accelerated profile HMM searches. PLoS computational biology. 2011;7(10):e1002195.

28. Horbelt D, Denkis A, Knaus P. A portrait of Transforming Growth Factor β superfamily signalling: Background matters. The international journal of biochemistry & cell biology. 2012;44(3):469–474.

29. Turner N, Grose R. Fibroblast growth factor signalling: from development to cancer. Nature Reviews Cancer. 2010;10(2):116–129.

30. Bonasoro F, Wilkie IC, Bavestrello G, Cerrano C, Carnevali M. Dynamic structure of the mesohyl in the sponge Chondrosia reniformis (Porifera, Demospongiae). Zoomorphology. 2001;121(2):109–121.

31. Ameziane-El-Hassani R, Morand S, Boucher JL, Frapart YM, Apostolou D, Agnandji D, et al. Dual oxidase-2 has an intrinsic Ca2+-dependent H2O2-generating activity. Journal of Biological Chemistry. 2005;280(34):30046–30054.

32. Miller CJ, Rose AL, Waite TD. Importance of iron complexation for Fenton-mediated hydroxyl radical production at circumneutral pH. Frontiers in Marine Science. 2016;3:134.

33. Ueno N, Takeya R, Miyano K, Kikuchi H, Sumimoto H. The NADPH oxidase Nox3 constitutively produces superoxide in a p22phox-dependent manner: its regulation by oxidase organizers and activators. Journal of Biological Chemistry. 2005;280(24):23328–23339.

34. Protchenko O, Rodriguez-Suarez R, Androphy R, Bussey H, Philpott CC. A screen for genes of heme uptake identifies the FLC family required for import of FAD into the endoplasmic reticulum. Journal of Biological Chemistry. 2006;281(30):21445–21457.

35. Missiakas D, Georgopoulos C, Raina S. Identification and characterization of the Escherichia coli gene dsbB, whose product is involved in the formation of disulfide bonds in vivo. Proceedings of the National Academy of Sciences. 1993;90(15):7084–7088.

36. Chang TS, Jeong W, Woo HA, Lee SM, Park S, Rhee SG. Characterization of mammalian sulfiredoxin and its reactivation of hyperoxidized peroxiredoxin through reduction of cysteine sulfinic acid in the active site to cysteine. Journal of Biological Chemistry. 2004;279(49):50994–51001.

37. Pei DS, Jia PP, Luo JJ, Liu W, Strauss PR. AP endonuclease 1 (Apex1) influences brain development linking oxidative stress and DNA repair. Cell death & disease. 2019;10(5):1–14.

38. Hu Z, Ding X, Ji Y, Liu X, Ding Z. APEX1 protects against oxidative damage-induced cardiomyocyte apoptosis. Biocell. 2021;45(3):745.

39. Korbecki J, Baranowska-Bosiacka I, Gutowska I, Chlubek D. The effect of reactive oxygen species on the synthesis of prostanoids from arachidonic acid. J Physiol Pharmacol. 2013;64(4):409–421.

40. Boschi-Muller S, Azza S, Sanglier-Cianferani S, Talfournier F, Van Dorsselear A, Branlant G. A sulfenic acid enzyme intermediate is involved in the catalytic mechanism of peptide methionine sulfoxide reductase fromEscherichia coli. Journal of Biological Chemistry. 2000;275(46):35908–35913.

41. Faber K, Hvidberg V, Moestrup SK, Dahläck B, Nielsen LB. Megalin is a receptor for apolipoprotein M, and kidney-specific megalin-deficiency confers urinary excretion of apolipoprotein M. Molecular endocrinology. 2006;20(1):212–218.

42. Ayala A, Muñoz MF, Argüelles S. Lipid peroxidation: production, metabolism, and signaling mechanisms of malondialdehyde and 4-hydroxy-2-nonenal. Oxidative medicine and cellular longevity. 2014;2014.

43. Imsiecke G, Steffen R, Custodio M, Borojevic R, Müller WE. Formation of spicules by sclerocytes from the freshwater spongeEphydatia muelleri in short-term cultures in vitro. In Vitro Cellular & Developmental Biology-Animal. 1995;31(7):528–535.

44. Imam SH, Buchanan MJ, Shin HC, Snell WJ. The Chlamydomonas cell wall: characterization of the wall framework. The Journal of cell biology. 1985;101(4):1599–1607.

45. Davison A, Blaxter M. Ancient origin of glycosyl hydrolase family 9 cellulase genes. Molecular Biology and Evolution. 2005;22(5):1273–1284.

46. Kerff F, Amoroso A, Herman R, Sauvage E, Petrella S, Filée P, et al. Crystal structure and activity of Bacillus subtilis YoaJ (EXLX1), a bacterial expansin that promotes root colonization. Proceedings of the National Academy of Sciences. 2008;105(44):16876–16881.

47. Zavaliev R, Levy A, Gera A, Epel BL. Subcellular dynamics and role of Arabidopsis β-1, 3-glucanases in cell-to-cell movement of tobamoviruses. Molecular Plant-Microbe Interactions. 2013;26(9):1016–1030.

48. Makino S, Ito N, Inoue T, Miyata S, Moriyama R. A spore-lytic enzyme released from Bacillus cereus spores during germination. Microbiology. 1994;140(6):1403–1410.

49. Han Z, Sieriebriennikov B, Susoy V, Lo WS, Igreja C, Dong C, et al. Horizontally acquired cellulases assist the expansion of dietary range in Pristionchus nematodes. Molecular biology and evolution. 2022;.

50. Conaco C, Tsoulfas P, Sakarya O, Dolan A, Werren J, Kosik KS. Detection of prokaryotic genes in the Amphimedon queenslandica genome. PLoS One. 2016;11(3):e0151092.

51. Hernandez AM, Ryan JF. Horizontally transferred genes in the ctenophore Mnemiopsis leidyi. PeerJ. 2018;6:e5067.

52. Ehrlich H, Kaluzhnaya OV, Brunner E, Tsurkan MV, Ereskovsky A, Ilan M, et al. Identification and first insights into the structure and biosynthesis of chitin from the freshwater sponge Spongilla lacustris. Journal of Structural Biology. 2013;183(3):474–483.

53. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–3212.

54. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols. 2013;8(8):1494–1512.

55. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic acids research. 2019;47(D1):D309–D314.

56. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology. 2017;35(11):1026–1028.

57. Pereira J, Simpkin AJ, Hartmann MD, Rigden DJ, Keegan RM, Lupas AN. High-accuracy protein structure prediction in CASP14. Proteins: Structure, Function, and Bioinformatics. 2021;89(12):1687–1699.

58. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23(10):1282–1288.

59. Yoo AB, Jette MA, Grondona M. Slurm: Simple linux utility for resource management. In: Workshop on job scheduling strategies for parallel processing. Springer; 2003. p. 44–60.

60. GCC, the GNU Compiler Collection; 2020. https://web.archive.org/web/20220528080831/https://gcc.gnu.org/.

61. NVIDIA, Vingelmann P, Fitzek FHP. CUDA, release: 11.1.1; 2020. Available from: https://docs.nvidia.com/cuda/archive/11.1.1/.

62. Hamelryck T, Manderick B. PDB file parser and structure class implemented in Python. Bioinformatics. 2003;19(17):2308–2310.

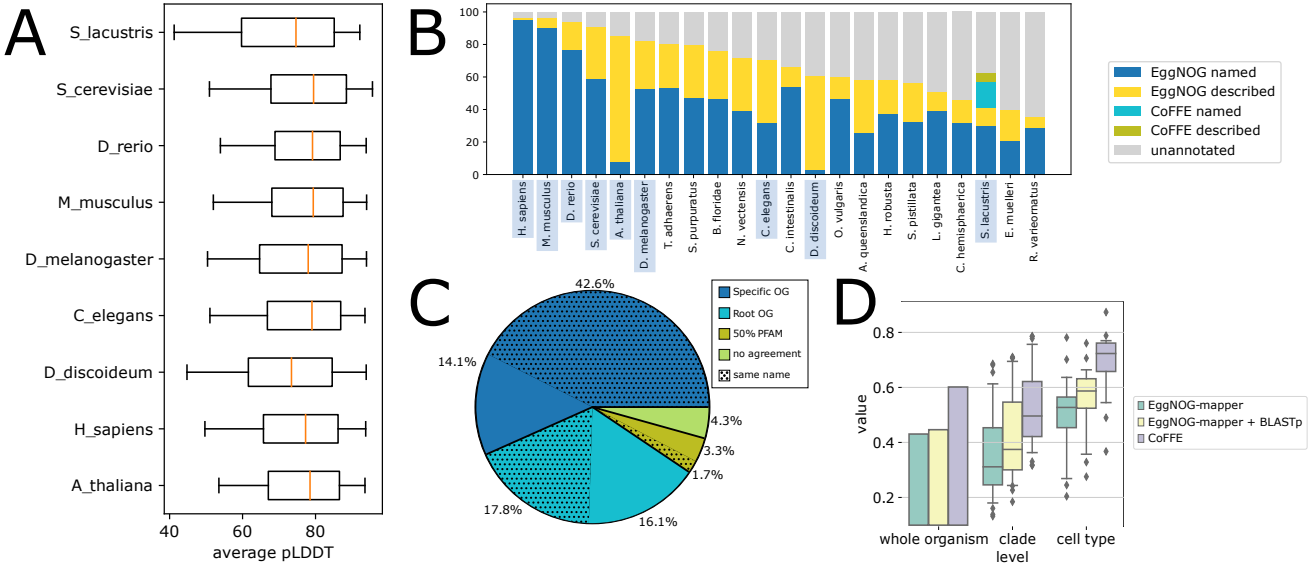
63. consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic acids research. 2021;49(D1):D480–D489.

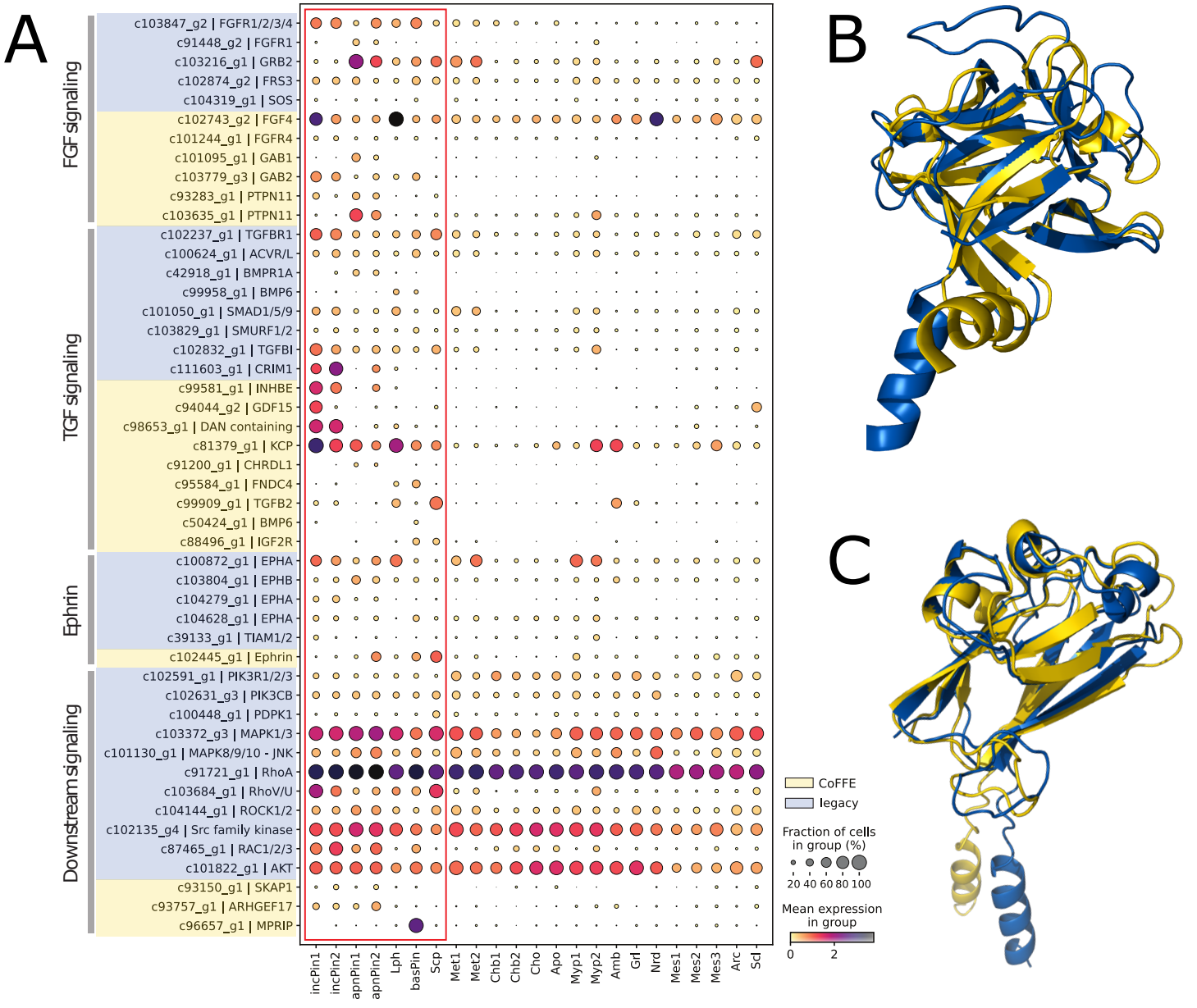
64. Sequeira JC, Rocha M, Alves MM, Salvador AF. UPIMAPI, reCOGnizer and KEGGCharter: Bioinformatics tools for functional annotation and visualization of (meta)-omics datasets. Computational and Structural Biotechnology Journal. 2022;20:1798–1810.

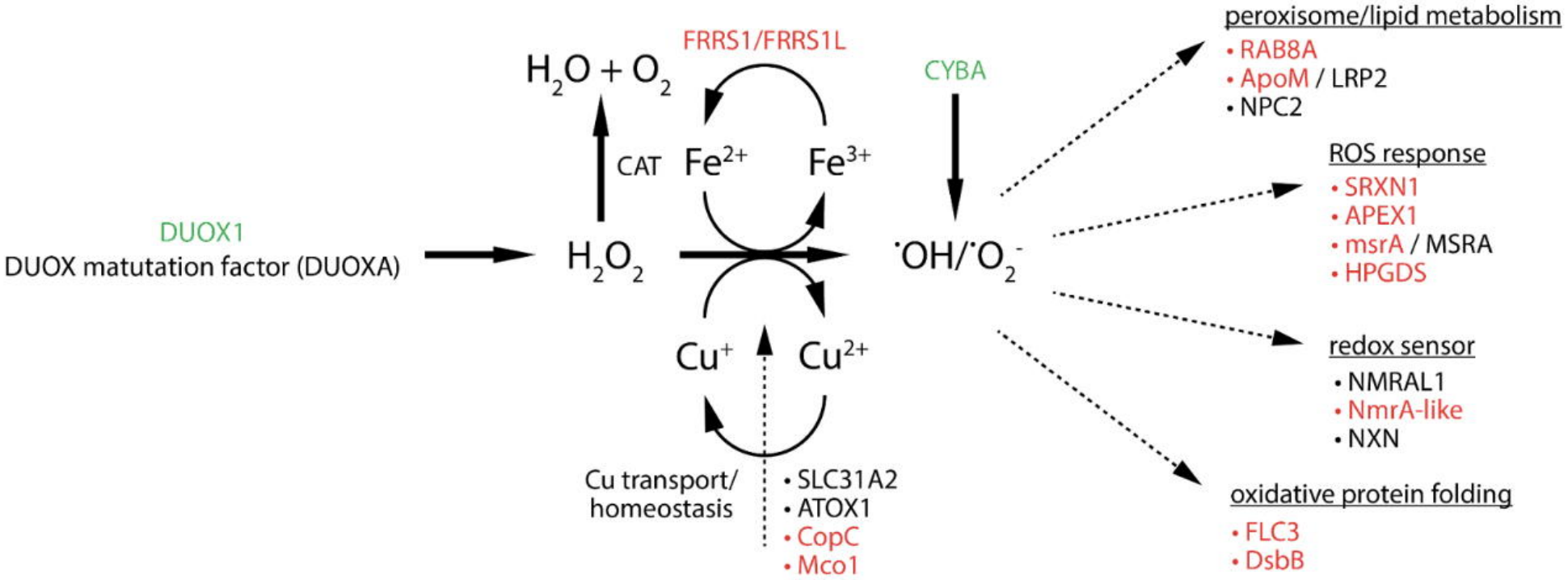
65. The PyMOL Molecular Graphics System; 2022. http://web.archive.org/web/20220624180204/https://pymol.org/.

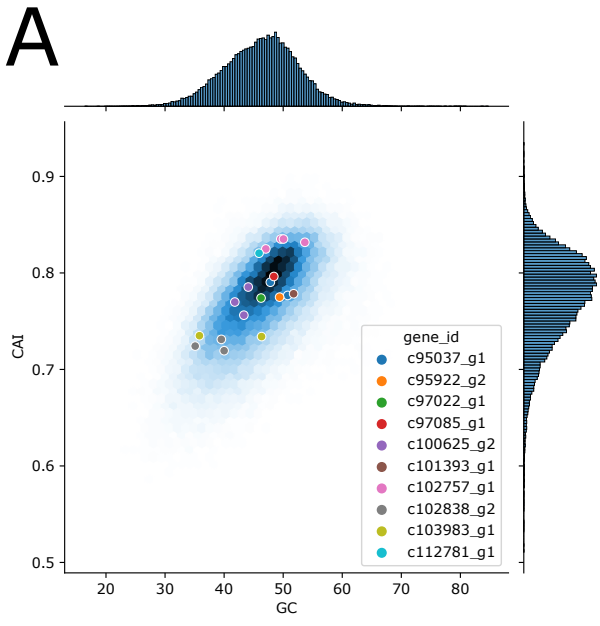
66. Degnan SM. Think laterally: horizontal gene transfer from symbiotic microbes may extend the phenotype of marine sessile hosts. Frontiers in microbiology. 2014;5:638.

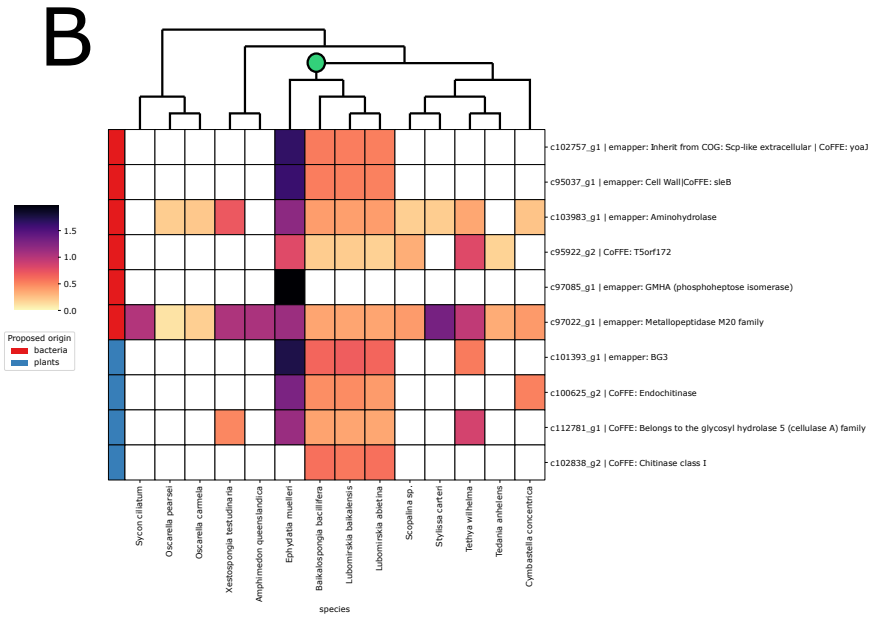
67. Plese B, Kenny NJ, Rossi ME, Cárdenas P, Schuster A, Taboada S, et al. Mitochondrial evolution in the Demospongiae (Porifera): Phylogeny, divergence time, and genome biology. Molecular Phylogenetics and Evolution. 2021;155:107011.


A
S_lacustris
S_cerevisiae
D_rerio
M_musculus
D_melanogaster
C_elegans
D_discoideum
H_sapiens
A_thaliana
40
60
80
average pLDDT
B
100
80
60
40
20
0
H. sapiens
M. musculus
D. rerio
S. cerevisiae
A. thaliana
D. melanogaster
T. adhaerens
S. purpuratus
B. floridae
N. vectensis
C. elegans
C. intestinalis
D. discoideum
O. vulgaris
A. queensland ica
H. robusta
S. pistillata
L. gigantea
C. hemisphaerica
S. lacustris
E. muelleri
R. varieomatus
EggNOG named
EggNOG described
CoFFE named
CoFFE described
unannotated
C
42.6%
14.1%
4.3%
3.3%
1.7%
17.8%
16.1%
Specific OG
Root OG
50% PFAM
no agreement
same name
D
0.8
0.6
0.4
0.2
value
whole organism
clade level
cell type
EggNOG-mapper
EggNOG-mapper + BLASTp
CoFFE


A
B
C
FGF signaling
TGF signaling
Ephrin
Downstream signaling
c103847_g2 | FGFR1/2/3/4
c91448_g2 | FGFR1
c103216_g1 | GRB2
c102874_g2 | FRS3
c104319_g1 | SOS
c102743_g2 | FGF4
c101244_g1 | FGFR4
c101095_g1 | GAB1
c103779_g3 | GAB2
c93283_g1 | PTPN11
c103635_g1 | PTPN11
c102237_g1 | TGFBR1
c100624_g1 | ACVR/L
c42918_g1 | BMPR1A
c99958_g1 | BMP6
c101050_g1 | SMAD1/5/9
c103829_g1 | SMURF1/2
c102832_g1 | TGFBI
c111603_g1 | CRIM1
c99581_g1 | INHBE
c94044_g2 | GDF15
c98653_g1 | DAN containing
c81379_g1 | KCP
c91200_g1 | CHRDL1
c95584_g1 | FNDC4
c99909_g1 | TGFB2
c50424_g1 | BMP6
c88496_g1 | IGF2R
c100872_g1 | EPHA
c103804_g1 | EPHB
c104279_g1 | EPHA
c104628_g1 | EPHA
c39133_g1 | TIAM1/2
c102445_g1 | Ephrin
c102591_g1 | PIK3R1/2/3
c102631_g3 | PIK3CB
c100448_g1 | PDPK1
c103372_g3 | MAPK1/3
c101130_g1 | MAPK8/9/10 - JNK
c91721_g1 | RhoA
c103684_g1 | RhoV/U
c104144_g1 | ROCK1/2
c102135_g4 | Src family kinase
c87465_g1 | RAC1/2/3
c101822_g1 | AKT
c93150_g1 | SKAP1
c93757_g1 | ARHGEF17
c96657_g1 | MPRIP
incPin1
incPin2
apnPin1
apnPin2
Lph
basPin
Scp
Met1
Met2
Chb1
Chb2
Cho
Apo
Myp1
Myp2
Amb
Grl
Nrd
Mes1
Mes2
Mes3
Arc
Scl
CoFFE
legacy
Fraction of cells in group (%)
20 40 60 80 100
Mean expression in group
0
2


DUOX1
DUOX matutation factor (DUOXA)
H_2O_2
CAT
$H_2O + O_2$
FRRS1/FRRS1L
Fe^{2+}
Fe^{3+}
Cu^+
Cu^{2+}
CYBA
$^{\cdot}OH/^{\cdot}O_2^-$
Cu transport/
homeostasis
• SLC31A2
• ATOX1
• CopC
• Mco1
peroxisome/lipid metabolism
• RAB8A
• ApoM / LRP2
• NPC2
ROS response
• SRXN1
• APEX1
• msrA / MSRA
• HPGDS
redox sensor
• NMRAL1
• NmrA-like
• NXN
oxidative protein folding
• FLC3
• DsbB


A
CAI
GC
0.9
0.8
0.7
0.6
0.5
20
30
40
50
60
70
80
gene_id
c95037_g1
c95922_g2
c97022_g1
c97085_g1
c100625_g2
c101393_g1
c102757_g1
c102838_g2
c103983_g1
c112781_g1


B
1.5
1.0
0.5
0.0
Proposed origin
bacteria
plants
c102757_g1 | emapper: Inherit from COG: Scp-like extracellular | CoFFE: yoaJ
c95037_g1 | emapper: Cell Wall|CoFFE: sleB
c103983_g1 | emapper: Aminohydrolase
c95922_g2 | CoFFE: T5orf172
c97085_g1 | emapper: GMHA (phosphoheptose isomerase)
c97022_g1 | emapper: Metallopeptidase M20 family
c101393_g1 | emapper: BG3
c100625_g2 | CoFFE: Endochitinase
c112781_g1 | CoFFE: Belongs to the glycosyl hydrolase 5 (cellulase A) family
c102838_g2 | CoFFE: Chitinase class I
Sycon ciliatum
Oscarella pearsei
Oscarella carmela
Xestospongia testudinaria
Amphimedon queensland ica
Ephydatia muelleri
Baikalospongia bacillifera
Lubomirskia baikalensis
Lubomirskia abietina
Scopalina sp.
Stylissa carteri
Tethya wilhelma
Tedania anhelens
Cymbastella concentrica
species