UNIVERSITAT DE BARCELONA

Assembling the largest 18S ribosomal RNA curated data set for non-bilaterian animals

Javier Arañó
Department of Genetics, Microbiology and Statistics / Faculty of Biology / University of Barcelona

05/08/2024

TFM
Master in Genetics and Genomics

Jesus Lozano Fernandez

Mattia Giacomelli

UNIVERSITAT DE BARCELONA

Assembling the largest 18S ribosomal RNA curated data set for non-bilaterian animals

Department of Genetics, Microbiology and Statistics/
Faculty of Biology / University of Barcelona

Master in Genetics and Genomics
05/08/2024

	Supervisor 1	Supervisor 2
Javier Arañó Ansola	Dr. Jesus Lozano	Dr. Mattia Giacomelli

ABSTRACT

The small subunit ribosomal RNA (SSU rRNA), named 18S in eukaryotes, is a gene universally present in all branches of the Tree of Life that has been instrumental to solve the most ancient relationships. Furthermore, it serves as a molecular identifier of biodiversity in environmental DNA studies. Despite its significance, specialised 18S databases still contain errors and struggle to keep pace with the rapidly changing eukaryotic taxonomy and the influx of novel diversity. This challenge hinders the assembly of reliable reference phylogenetic trees needed to identify novel 18S sequences obtained through metabarcoding studies.

As part of a larger project aimed at creating reference databases for eukaryotic 18S (EukRef), this work focuses on curating the 18S sequences of non-bilaterian animals (sponges, ctenophores, cnidarians and placozoans) in the Protist Ribosomal Reference Database (PR2). Through this curation process, we generated the largest backbone phylogenetic tree for non-bilaterians, essential for the taxonomic identification of 18S sequences within this group. Furthermore, we reaffirmed that 18S is a suitable phylogenetic marker for animals, as it can recover most well-known clades to the phylum level. Lastly, we confirmed that incorporating 18S secondary structure information into sequence alignment positively impacts the topology of the inferred animal Tree of Life.

SUSTAINABLE DEVELOPMENT GOALS (SDG)

In light of the global biodiversity crisis our planet is currently facing, expanding our knowledge of Earth's ecosystems and their inhabitants is crucial for their conservation. This project will contribute to enhancing our understanding of non-bilaterian animals and identifying known and unknown biodiversity within this group. Non-bilaterians are a diverse group of organisms that constitute a significant portion of the biomass in aquatic ecosystems and play multiple ecological roles. Therefore, in the final term, this research will clearly contribute to those SDG aimed at fighting biodiversity loss: Life Below Water (SDG 14), concretely impacting goal 14.2 “Sustainably manage and protect marine and coastal ecosystems to avoid significant adverse impacts, including by strengthening their resilience, and take action for their restoration in order to achieve healthy and productive oceans”; and Life on Land (SDG 15), especially goal 15.5 “Take urgent and significant action to reduce the degradation of natural habitats, halt the loss of biodiversity and protect and prevent the extinction of threatened species”.

Furthermore, the collaboration between evolutionary biologists, bioinformatics and other disciplines present in this work embodies the essence of Partnerships for the Goals (SDG 17), recognizing that solving complex challenges requires collective effort. Thus, it is evident that this research not only contributes to the Sustainable Development Goals but also demonstrates the potential of science to address these global objectives.

ABBREVIATIONS

SSU rRNA (Small subunit ribosomal RNA)
ToL (Tree of Life)
Long-Branch Attraction (LBA)
Maximum Likelihood (ML)

SUPRA-PHYLA TAXONOMIC GLOSSARY

Ambulacraria: supraphyletic group inside *Deuterostomia* composed of *Hemychordata* and *Echinodermata*.

***Lophotrochozoa*:** Its members are characterized by the lophophore, a feeding structure consisting of a ciliated crown of tentacles surrounding a mouth, and the developmental stage of the trochophore larva. It is composed of Gastrotricha, *Platyhelminthes, Bryozoa, Brachiopoda, Nemertea, Mollusca, Annelida, Entoprocta, Cycliophora, Rotifera* and *Gnathotomulida*

Ecdysozoa: Clade of molting animals consisting of *Kinhoryncha, Priapulida, Loricifera, Nematoda, Nematomorpha, Tardigrada, Onychophora* and *Arthropoda*

Protostomia: clade composed of the two supraphyletic groups *Ecdysozoa* and *Lophotrochozoa*

***Deuterostomia*:** clade characterized by lineages that form the anus before the mouth during embryonic development. It is composed of *Chordata*, *Hemychordata* and *Echinodermata.*

Bilateria: Clade of animals with bilateral symmetry consisting of three supraphylum clades (*Protostomia, Deuterostomia* and *Xenacoelomorpha*)

TABLE OF CONTENTS

1. INTRODUCTION 1
1.1 18S rRNA, an attribute of Earth's life 1
1.2 Metabarcoding survey of 18S and the uncover of new diversity 3
1.3 The relevance of curated 18S sequence databases 4
1.4 Past and present of 18S RNA in the context of reconstructing animal evolution . 5
2. OBJECTIVES 8
3. MATERIALS AND METHODS 8
3.1 PR2 database curation 8
3.2 Data and phylogenetic inference 9
3.3 Structural alignments 9
3. RESULTS 11
3.1 PR2 database curation 11
3.2 Metazoa phylogeny based on 18S sequences and sensitivity analyses 14
3.3 Addition of 18S secondary structure information 17
4. DISCUSSION 19
4.1 Curating PR2 database and generating a backbone tree for non-bilaterians 19
4.2 The reliability of 18S-based animal phylogeny 21
4.3 The effect of adding secondary structure to 18S-based phylogeny of animals ... 24
5. CONCLUSIONS 26

1. INTRODUCTION

1.1 18S rRNA, an attribute of Earth's life

The small subunit ribosomal RNA (SSU rRNA), 18S rRNA in eukaryotes or 16S in prokaryotes, is a non-protein coding, structural RNA found in all cellular organisms. SSU rRNA forms the structural core of the small subunit of the ribosome. It interacts with the large subunit ribosomal RNA (LSU) and ribosomal proteins, and together with transfer RNAs decodes messenger RNAs into amino acids, providing peptidyl transferase activity to form peptide bonds between adjacent amino acids during translation [1-3]. Due to this role in the vital process of translation, SSU rRNA's sequence and structure have been highly constrained during evolution [1]. Such constraints have been so strong that a solid hypothesis of homology for each site in the SSU rRNA sequence can be established among organisms across the entire Tree of Life (ToL) [1, 4].

SSU rRNA is the most conserved of all existing ribosomal RNAs and among the slowest evolving locus throughout living organisms, and hence, it has been very useful for examining ancient evolutionary events dating back to the Precambrian [4, 5]. In 1977, Carl Woese and George Fox were the first to employ this sequence to build a molecular phylogeny for the entire tree of life, arguing that it existed in all cellular life, had evolved slowly enough to be comparable across all life, and it was readily isolated from a variety of different unicellular organisms [6]. This work led to the description of the Archaea domain and the organisation of life in three separate domains (Eukarya, Eubacteria & Archaebacteria). Besides the reasons mentioned by Woese & Fox, SSU rRNA possesses other properties that contribute to its utility as a phylogenetic marker: 1) the presence of multiple extremely conserved regions flanking more variable regions within the SSU rRNA sequence allows the construction of nearly universal primers, facilitating sequencing efforts from previously unstudied groups; 2) the typical presence of many tandemly repeated copies of SSU in each nuclear eukaryote genome make it easy to get PCR product amplification; 3) the pattern of concerted evolution that occurs among repeated copies of an individual organism reduces individual polymorphism (In other words, each copy of an rRNA array is usually very similar to the other copies within individual genomes). These features facilitate the analysis of rDNA by direct or

environmental RNA or DNA sequencing, making SSU rRNA one of the most frequently sequenced genes [4, 5].

Furthermore, it must be noted that SSU rRNA, like all ribosomal RNAs, is a structural RNA, working as a functional unit rather than serving as a template to construct a protein. SSU rRNA secondary and even tertiary structures have been deciphered in multiple taxa (Fig. 1), enabling the identification of structural homology within studied organisms [3, 7]. Previous research has demonstrated that the simultaneous use of RNA sequences and their individual secondary structure increases the robustness and accuracy of phylogenetic analyses [8]. Thus, SSU's secondary and tertiary structures can be treated as a phylogenetically informative character when working with SSU as has been done in recent papers [8, 9].

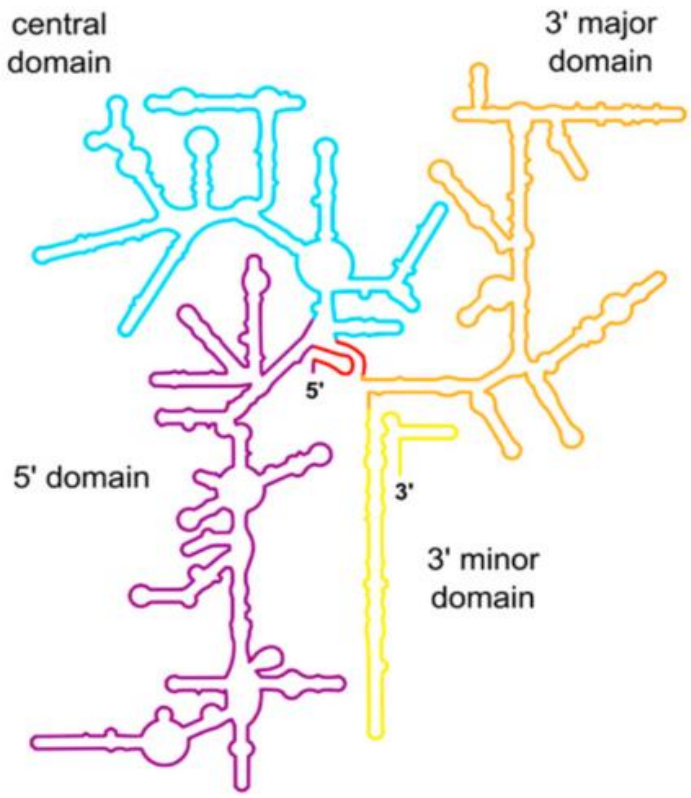


Figure 1. Secondary structure of the human 18S rRNA. The 18S rRNA folds into 45 helices that distribute into four structural domains distinguished here by different colours. Adapted from [66].

1.2 Metabarcoding survey of 18S and the uncover of new diversity

Multiple worldwide oceanographic expeditions have sampled marine samples across the oceans [10-13]. Based on the study of these collected samples, recent reports suggest that marine environments have not yet been fully sampled, implying the potential existence of unknown living lineages, which could potentially help to resolve the ToL [14]. As a case book study, the analysis of metagenomic samples led to the genome assembly of a new archaeal group, the superphylum Asgard, including the closest relatives of eukaryotes that share many of their genomic features. This finding reshaped our understanding of the evolution of life on Earth [15].

DNA metabarcoding is a PCR-based barcoding technique that allows us to specifically sequence well-known molecular markers, such as the 18S ribosomal gene, to discover hidden diversity [16]. Similarly, environmental DNA samples can be sequenced using metagenomic approaches, in which no phylogenetic marker is specifically targeted, but instead an environmental sample of bulk DNA or RNA is sheared into fragments which will be sequenced typically using a high-throughput sequencing approaches [17]. These techniques enable massive multispecies (or higher taxon) identification using the total and typically degraded DNA from an environmental sample (e.g. soil, water, faeces, etc.) [18]. Since Pace et al. applied the first universal primers for determining partial SSU rRNA sequences from bulk cellular RNA in 1985 [19], the popularity of barcoding-like studies steadily increased [1]. Metabarcoding employing various molecular markers (e.g. cytochrome *c* oxidase subunit I or 18S ribosomal RNA) has been successfully used to analyse the diversity of animals (*Metazoa*) inhabiting multiple environments [20-22]. For example, the study of a short fragment of the 18S in several European coastal sampling sites allowed to discover an uncharacterized clade of tunicates [14]. For all environmental DNA surveys, regardless of how sequences are acquired and how many are obtained, a common crucial step of "phylogenetic placement" should be performed to properly identify and approximate the taxonomy of the sequences in the sample. In this step, environmental sequences are aligned to a set of reference sequences about which taxonomic information is known. In the resulting phylogenetic tree, known as the

‘backbone tree’, the location of the new sequences on the tree in relation to the reference sequences reveals their approximate identity. Note that this does not represent a phylogenetic inference, because in phylogenetic placement the reference tree is kept fixed, implying that the environmental sequences are not inserted as new branches into the tree but rather “mapped” onto its branches. [23]. Phylogenetic placement for environmental DNA data requires a high-quality, up-to-date, source of curated sequences that can be used as a reliable backbone tree. However, this robust backbone is absent for most clades in the animal kingdom, therefore, the application of phylogenetic placement techniques are hampered by the absence of a *bona fide* 18S animal tree of life.

1.3 The relevance of curated 18S sequence databases

Reference databases of ribosomal DNA bring together sequences from known isolates as well as environmental sequence datasets [24]. While specialised SSU rRNA databases have improved significantly in recent years, they still contain errors and struggle to keep pace with the rapidly changing eukaryotic taxonomy and the influx of novel diversity [24]. Incorrect sequence annotation of SSU rRNA public databases hinders the assembling of a high-quality backbone tree needed for the accurate characterization of lineage diversity in environmental samples using phylogenetic placement techniques. Therefore, the optimization of ribosomal sequence databases is a crucial step toward discovering new taxa that can lead to the resolution of the ToL.

To address the errors present in databases, particularly for unicellular eukaryotic lineages (protists), the EukRef initiative was born. This community effort of phylogenetic curation aims to improve the taxonomic information associated with 18S rRNA sequences and create better reference databases for metabarcoding studies [24]. The EukRef community developed a pipeline for database curation. Following this approach, a set of sequences from the database is aligned with a phylogenetically accurate reference alignment. A phylogenetic tree is then inferred from this alignment and used to identify discrepancies such as long branches, which may be potential artifacts, or sequences that fall outside of its predicted taxon. Following the removal of these problematic data, a new alignment and tree are constructed with the remaining sequences [24].

The two main databases for eukaryotic ribosomal DNA sequences are SILVA [25], a general database that also includes Bacteria and Archaea ribosomal DNA, and the Protist Ribosomal Reference Database (PR2), which mostly focuses on protists but also includes metazoans, land plants, macroscopic fungi, and eukaryotic organelles [26]. PR2 database possesses an extensive set of metazoan 18S sequences that have not been properly curated yet. All these molecules have been annotated using the taxonomy assigned in the National Center for Biotechnology Information GenBank (NCBI) [27] database entries without further examination. To date, more than 135,000 sequences inside the PR2 database originally annotated as *Metazoa* can not be securely assigned to a specific clade, or even assure that they belong to an animal. EukRef, now integrated with PR2, is actively involved in the curation of this database. As part of the EukRef community, we are responsible for curating the non-bilaterian *Metazoa* sequences present in PR2. Through this process, we aim to generate a backbone 18S phylogenetic tree for non-bilaterian animals. This endeavor may lead to the correction of errors, discovery of new taxa, and enhance our understanding of the relationships at the base of the animal ToL.

1.4 Past and present of 18S RNA in the context of reconstructing animal evolution

In the year 2005, Science magazine included the "resolution of the Tree of Life" among the most important gaps in scientific knowledge at that time. Almost 20 years later, despite the analysis of massive amounts of both morphological and molecular data, many nodes of the tree of Life (ToL) remain unresolved. Even the evolutionary relationships within our kingdom, the *Metazoa*, remain controversial, in particular in determining its root and early splits [28]. In the last two decades, multigene analysis and phylogenomics methods have taken over. Nevertheless, animal phylogeny is not totally solved yet, and there are still challenging nodes that are unstable depending on the methodology used to infer the tree. For example, multiple large-scale phylogenomic studies have not managed yet to provide a consensus on determining which animal group is sister to the rest, sponges (*Porifera-sister*) or comb jellies (*Ctenophora-sister*) [29] (Fig. 2). This is not a trivial debate, as resolving relationships among extant lineages at the base of the animal tree is essential to understand the evolution of complex animal traits, such as the nervous

systems, mesoderm or muscles, as well as the interpretation of genome architecture and gene content [28, 30]. While sponges are relatively simple with no nerve cells, muscles or digestive system, ctenophores are sophisticated marine animals having all these complex systems. The classic scenario in which sponges were the first lineage to branch from the tree is consistent with a single origin of the nervous, muscular, and digestive system [31]. If ctenophores are the most distantly related, these traits could have evolved independently at least twice, in ctenophores and in the ancestor of Bilateria and cnidarians, in a complex pattern of convergent evolution (Fig. 2). Alternatively, they may have evolved in the common ancestor of animals and been lost in sponges and placozoans.

In the past, 18S rRNA had capital importance in our current knowledge of animal phylogeny. Before the rise of molecular approaches to infer phylogenies, most animal groups were grouped by superficial similarity or based on shared morphological characters. In many instances, largely based on decrees by authority figures, often without formal algorithms for evolutionary analyses of data [32]. A revolutionary change took place in 1988 when Field et al.'s paper "Molecular phylogeny of the animal kingdom" was published [33], displaying a phylogeny of animals based on the 18S. This paper ushered in an era of molecular systematics for higher-taxonomic level animal phylogeny, using nucleotide sequences to study evolutionary relationships within major animal phyla. In the following years, several groundbreaking hypotheses regarding animal phylogeny were formulated, which either contrasted or corroborated classical morphological views. Some novel hypotheses mostly based on initial molecular findings include, but are not limited to: *Lophotrochozoa* [34], *Ecdysozoa* [35], *Cnidaria* as sister to *Bilateria* [36], *Platyhelminthes* polyphyly [37], and *Ambulacraria* [38]. All of these findings have been subsequently confirmed by other sources of evidence and are now generally accepted [39].

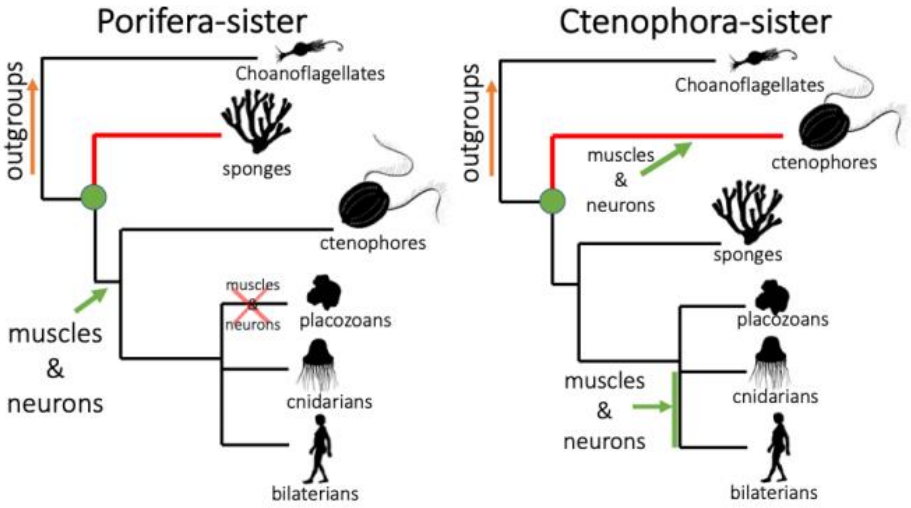


Figure 2. Main hypotheses on the early diversification of animals, with possible implications on the origin of the nervous or muscular system. It should be noted that although the relationships of placozoans with cnidarians and bilaterians are not resolved, they do not have massive implications of

Despite advancements in phylogenomics, the resolution of the root of the animal tree remains elusive. The 18S ribosomal RNA gene presents an opportunity to identify hidden diversity that can help mitigate phylogenetic artifacts that commonly affect phylogenomic datasets. Additionally, 18S still possesses sufficient phylogenetic signal to properly distinguish Metazoa species at higher taxonomic levels (e.g. phyla and class), and is one of the most sequenced genes in environmental DNA studies [20]. Furthermore, we can assess whether the secondary structure of 18S allows the recovery of the accepted *Metazoa* phylogeny, which has not been tested yet to the author's knowledge. We have joined forces as members of the EukRef and PR2 initiatives, committing ourselves to the curation of non-Bilateria animal sequences in the PR2 database. This work will provide a valuable resource for researchers to map their metabarcoding projects, enabling the characterization of 18S evolution in these animals and the discovery of previously unknown or undescribed animal lineages near the root of the tree.

2. OBJECTIVES

- **Curate the sequences in PR2 database annotated as non-bilaterian *Metazoa* following the EukRef pipeline, and generate in the process a backbone phylogenetic tree for these groups of animals**

- **Compare the reliability of 18S-based animal phylogeny to the current accepted animal relations according to phylogenomic approaches**

- **Analyse how the incorporation of secondary structure data to the 18S alignment affects the topology of 18S-based animal phylogeny**

3. MATERIALS AND METHODS

3.1 PR2 database curation

For PR2 database curation, a 18S sequence dataset containing 161 species representing all described non-bilaterian animal major lineages (i.e. Class level) was used as a reference alignment. After building this reference alignment, all non-bilaterian animal 18S sequences were retrieved from the PR2 database, yielding a total of 5,568 sequences. These sequences were divided into eight smaller, more manageable groups (*Placozoa*, *Ctenophora*, *Porifera I* (Hexactinellida+Demospongia), *Porifera II* (Homoscleromorpha+Calcarea), *Cnidaria I* (Schyphozoa), *Cnidaria II* (Hydrozoa), *Cnidaria* III (Anthozoa), and *Cnidaria* IV (Cubozoa)) and aligned with the previously selected reference sequences using SSU-align software [1]. Each subgroup alignment was trimmed of uninformative sites using ClipKIT [40]. After trimming, a Maximum Likelihood (ML) phylogenetic tree was inferred from each subgroup alignment with IQ-TREE [41] using the GTR+G4+I model. The resulting phylogenetic trees were used to identify discrepancies such as long branches, which may indicate potential artifacts, or sequences that fall outside their predicted taxon. All problematic sequences were flagged as potential errors and removed from the alignments. Subsequently, all subgroup alignments without the flagged sequences were unified into one general alignment. From

this general alignment, new rounds of tree inference, sequence flagging, and removal were conducted iteratively until no more problematic sequences could be identified.

3.2 Data and phylogenetic inference

529 complete or nearly complete 18S sequences from almost all animal classes were retrieved from NCBI (ANNEX 1). These sequences were aligned using the multiple sequence aligner software MAFFT with default settings [42]. The average size of the sequences was 1,830 bp, and once aligned, the size of this untrimmed version of the matrix was 7,658 sites. It is known that highly divergent sites in the alignment can negatively impact phylogenetic inference, therefore all constructed alignments were trimmed using ClipKit with smart-gap trimming mode. ClipKit aims to identify and retain parsimony-informative sites, which are known to be phylogenetically informative [40], and rendered a trimmed matrix of 5,291 positions. Subsequently, IQ-TREE was employed to determine the dataset's best-fitting (optimal) and less-fitting (suboptimal) substitution models. The same software was used to construct a ML tree using four different substitution models: the one identified by IQ-Tree as the best-fitting model for our data (TIM2+F+R10); the most complex substitution model for nucleotide matrices (GTR+G4+I); and JC69+R10, and HKY85+F+R10, with the latter two being simpler models with worst fit, according to ModelFinder. All phylogenetic analyses were performed in the Hercules computer cluster hosted at the University of Barcelona. The trimmed and untrimmed version of all matrices, as well as the resulting phylogenetic trees can be found at

https://drive.google.com/drive/folders/1KSj5TtfcB2mecWuX4F45EDdpaCUA8p1g?usp=drive_link.

3.3 Structural alignments

To add structural information to the 18S alignment, two different strategies were tested as outlined in figure 3. First, the individual secondary structure of 410 sequences from the original dataset was obtained from the specialized database RNAcentral [43]. Using

this sequence-structure dataset, a structurally-aware alignment was constructed using ClustalW [44] as implemented in 4SALE [45]. 4SALE uses a 12-letter translation table to encode the sequence-structure information of each individual taxa into a new sequence that is based on a 12-letter alphabet, known as pseudoprotein sequence. Pseudoprotein sequences are automatically aligned using a 12×12 scoring matrix. Second, the original dataset was aligned using SSU-align. This open-source software uses previously generated covariance model for specific phylogenetic ranges to create structure-aware alignments of SSU rRNA guided by a consensus 18S secondary structure [1]. In this case, a default covariance model for eukaryotic 18S included in SSU-align was employed to perform the guided alignment.

After trimming with ClipKit, a ML tree was reconstructed from both alignments using IqTree. For the SSU-align alignment tree, the GTR+G4+I model was used, and for the 4SALE alignment tree, the GTR20+G4+I model was employed. PhyKIT [52] was employed to obtain informative metrics about the generated ML trees and alignments. The support for all phylogenetic trees in this study was assessed using ultrafast bootstrap as implemented in IQ-TREE. FigTree [46] was used to visualize tree topology.

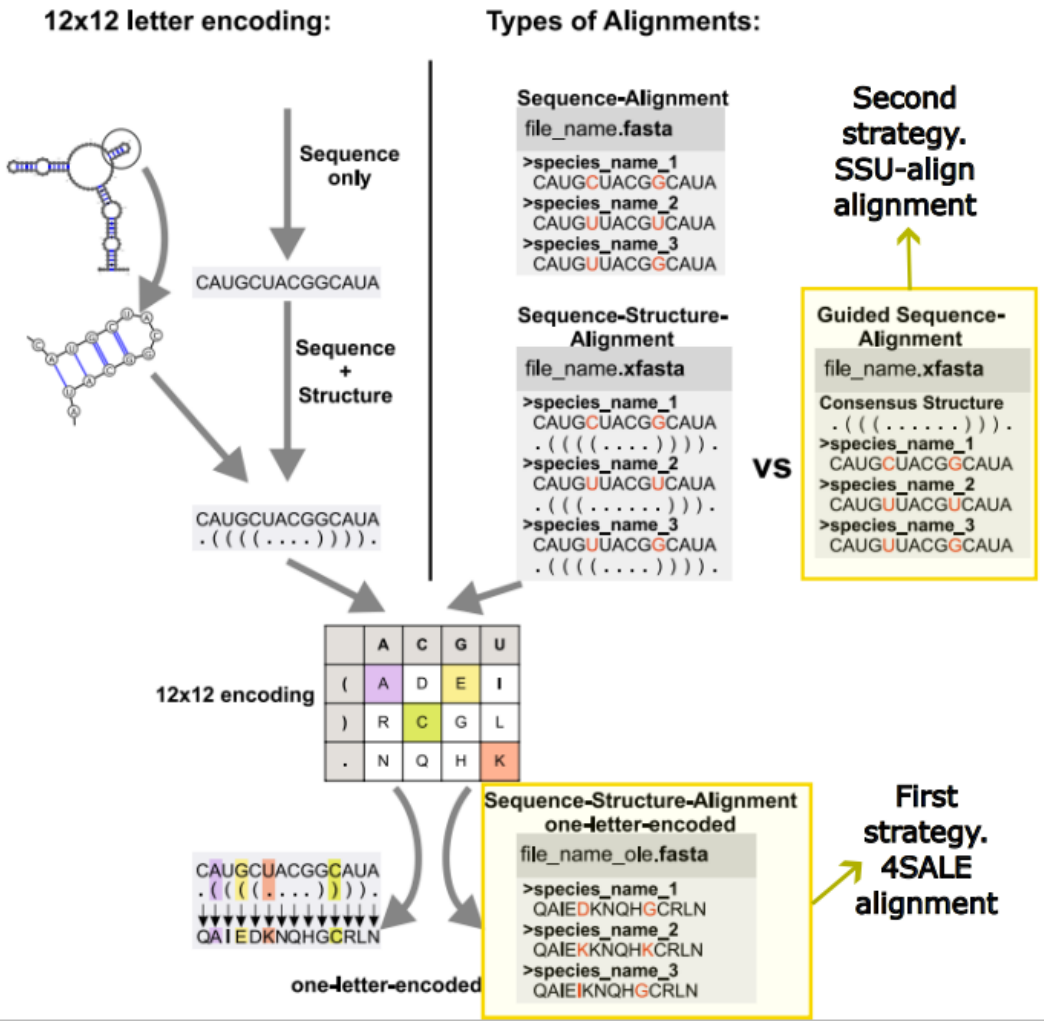


Figure 3. Left: Encoding of sequence-structure information as implemented in 4SALE. The figure shows the process of turning an RNA sequence with its individual secondary structure in the bracket-dot-bracket notation to a one-letter-encoded pseudoprotein that will be used for the alignment. Right: Different alignments are shown. They differ in terms of informational content (exemplarily highlighted in red). In The guided-sequence alignments, such as the SSU-align approach, the alignment is guided only by a consensus structure. Adapted from [8].
The type of alignments used in each one of the employes strategies strategies are highlighted.

3. RESULTS

3.1 PR2 database curation

After applying our curation pipeline to all 18S-sequences annotated as non-bilaterian animals in the PR2 database, 174 sequences (3.125%) were flagged as problematic or misidentified as summarized in figure 4. However, multiple reasons can justify the

flagging of a sequence (Fig. 5). In some cases, the flagged sequences were indeed grouped within a well-described animal clade that was not present in the database. This is the case of the 55 sequences in PR2 belonging to the cnidarian class *Staurozoa*, which were all incorrectly annotated as *Schyphozoa*, breaking the monophyly of the latter. Also, the same case applied to four putative *Hydrozoa* sequences that in many trees fell in a separated clade inside *Cnidaria* that have already been described as the class *Polypodiozoa*. In other cases, a sequence was flagged because it fell outside its annotated clade. This problem ranged from sequences placed within its predicted phyla but in a different class, to sequences clustered with the non-metazoa outgroup. Between these two extremes, we found sequences falling in other non-bilateria phyla, and others grouped with bilaterians.

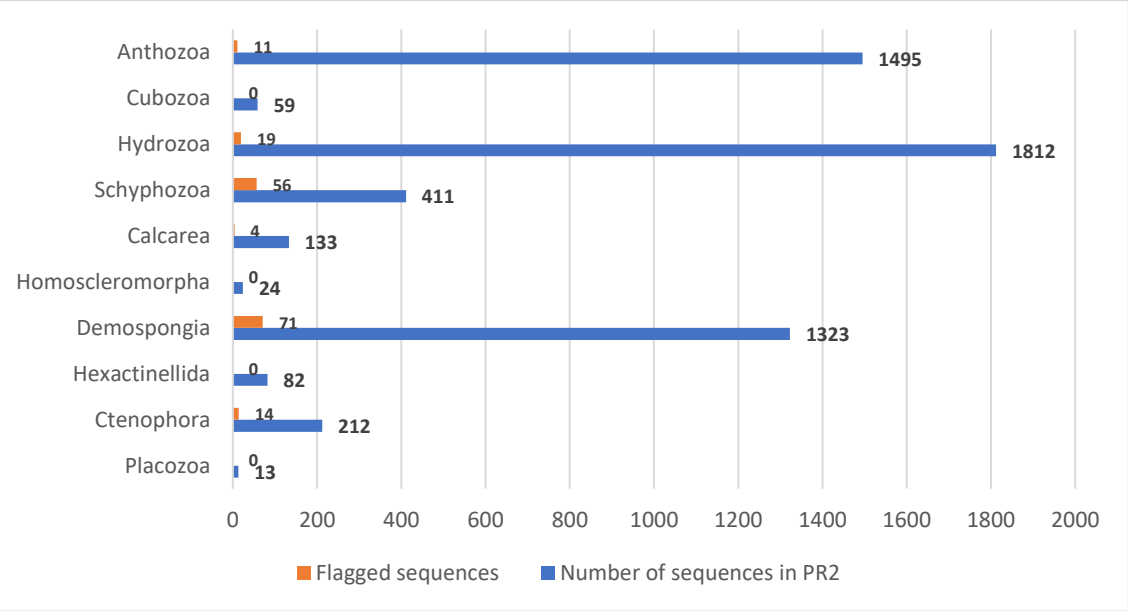


Figure 4. Number of current sequences of each non-bilaterian *Metazoa* class in the PR2 database compared to the number of sequences flagged during the curation. process.

Additionally, 31 flagged sequences were found in odd positions in the phylogeny. This is the case of a small clade of putative *Demospongia* which repetitively fell at the base of *Metazoa* separated from the other *Porifera* clades and ten supposed *Hydrozoa* sequences that were placed in a clade sister to all other *Cnidaria*. Something similar happened with 14 *Ctenophora* sequences that are placed in diverse positions at the base of its phyla, away from the crown ctenophores, and we conservatively removed them. After erasing all problematic sequences, the final aligment contained 5632 taxa without the outgroups. The output of this curation process has been a ML tree, made of a combination of the

reference alignment and the PR2 sequences that survived the curation process. This final tree recovered all non-bilaterian clades to the class level except *Polypodiozoa*, which in this case clustered inside *Anthozoa* (Fig. 6).

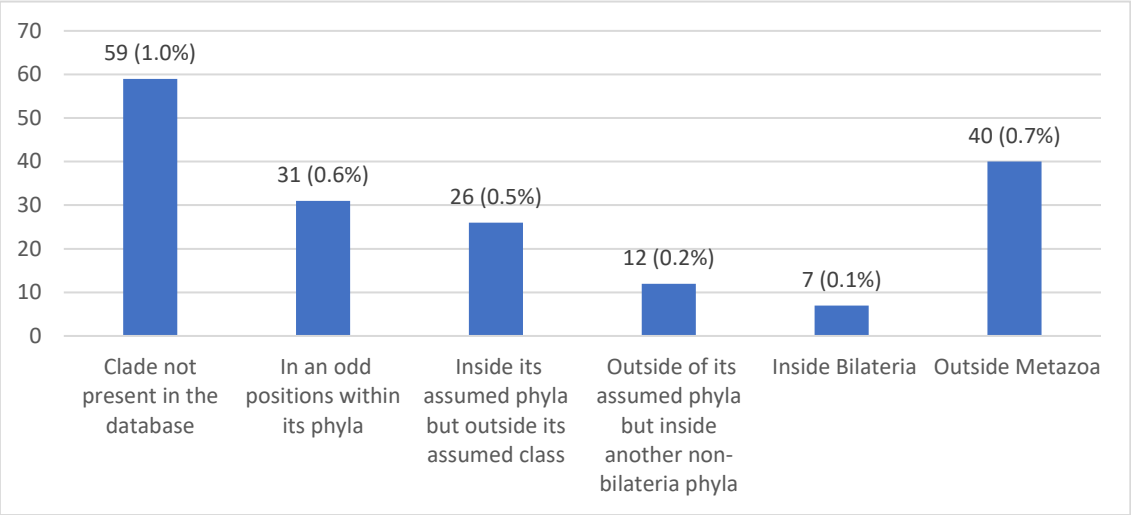


Figure 5. Reason for flagging each problematic sequence during the curation process of non-bilaterian *Metazoa* taxa in the PR2 database (percentage over total non-bilaterian animal sequences in PR2).

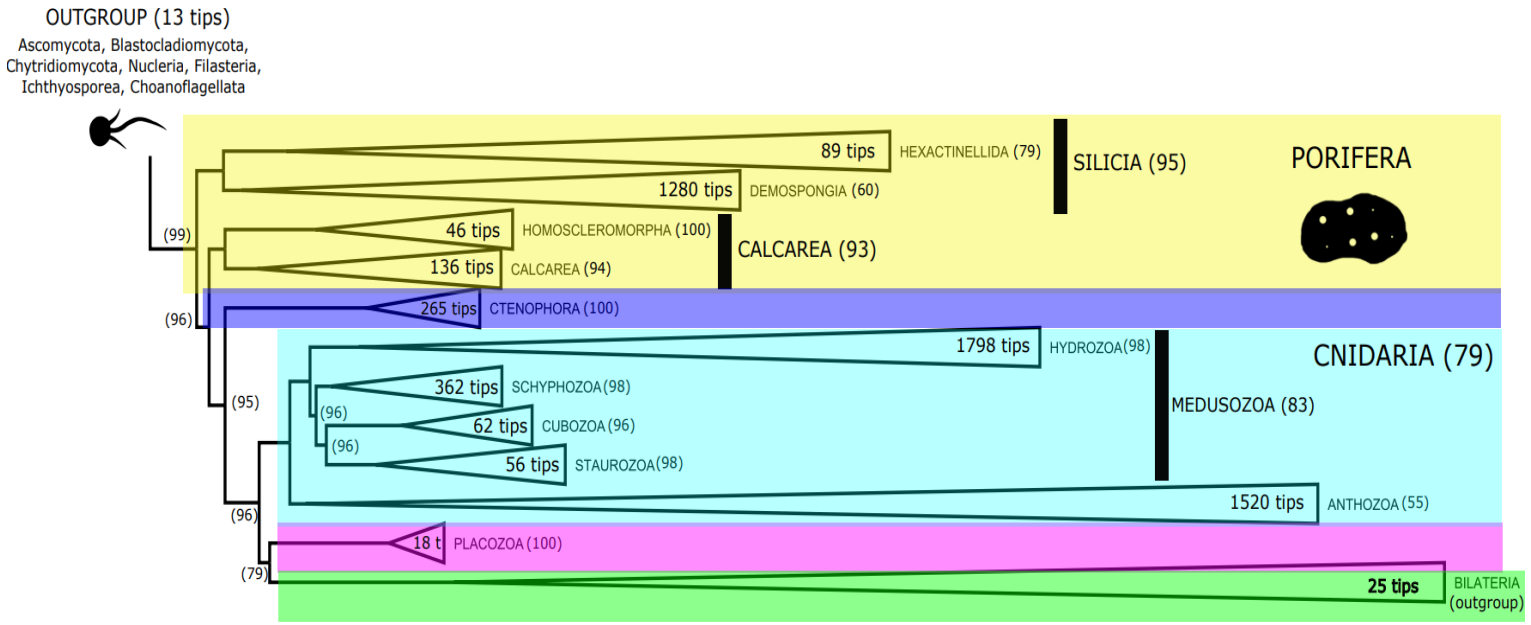


Figure 6. Collapsed ML phylogenetic tree made with all the 18S sequences in the PR2 database belonging to non-bilaterian animals after erasing all problematic sequences (bipartition support assessed with ultrafast bootstrap). The number of sequences of each clade can be seen inside each collapsed node. The outgroups are written from the farthest to the closest to *Metazoa*.

3.2 Metazoa phylogeny based on 18S sequences and sensitivity analyses

All four different ML trees inferred from the same alignment using different substitution models yielded similar topologies, retrieving animals into a single clade and recovering many accepted animal monophyletic groups at the phyla and supra-phylum levels. Nevertheless, the internal relationships between major animal phyla show some inconsistencies, especially within bilaterian animals (Fig. 7).

non-Bilateria

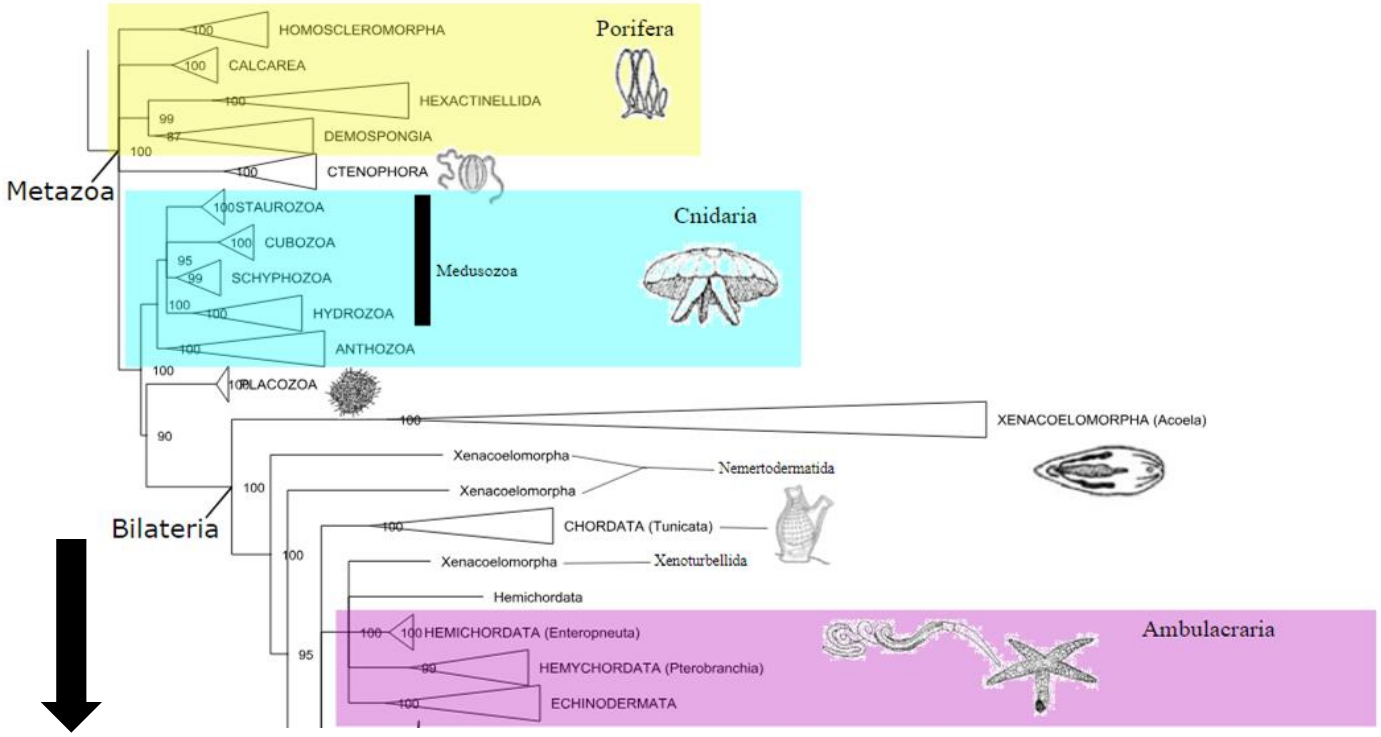
In all of the results, *Porifera* was recovered as a paraphyletic clade. *Calcarea and Silicia* sponges appeared splitting earlier from the rest of animals, with the latter sister to a monophyletic *Ctenophora*, although this last sister group relationship was poorly-supported and disappears when low-support nodes are collapsed (<85). Ctenophora, Cnidaria and Placozoa were invariably recovered with high support (<90) as monophyletic phyla, independently from the model employed. Within Cnidaria, all classes were recovered with almost full support, and containing a major split between Medusozoa classes (Scyphozoa, Staurozoa, Cubozoa & Hydrozoa) and corals (Anthozoa). All models displayed a sister group relationship between *Placozoa* and all bilaterian animals, but it was only highly supported in the GTR tree. Furthermore, among non-bilaterian animals, all class-level clades were recovered and highly supported with all employed substitution models.

Bilateria

We recovered *Bilateria* and *Protostomia* with almost full support in all tested conditions. *Xenacoelomorpha* worms were the earliest-splitting *Bilateria* in all inferred trees, but never as a monophyletic group because *Nemertodermatida* and *Xenoturbellida* split as the successive sister groups of the rest of bilaterians and *Xenoturbella bocki* was always placed within *Ambulacraria*. However, none of the inferred phylogenies recovered *Deuterostomia* or even the monophyly of the phylum *Chordata.* In all cases, *Tunicata* formed a clade separated from the other chordates.. Between the chordate clades, an *Ambulacraria,* containing the well-supported phylum *Hemichordata* and *Echinodermata*, was present in all trees, although containing the *Xenoturbellida* worm.

With all models, *Ecdysozoa* monophyly is disrupted because it contains the non-molting *Chaetognatha* worms, and other non-Ecdysozoan long-branched taxa. On the contrary, *Lophotrochozoa* is not fully recovered because some long branches are attracted towards *Ecdysozoa*. Anyways, a clade consisting mostly of *Ecdysozoa* and *Lophotrochozoa* are always recovered with high support. Most phyla within Ecdysozoa are recovered as monophyletic, except for *Onychophora* and *Arthropoda*, which contain long-branched taxa that are attracted towards other ecdyosozoan and lophotrocozoan long-branched groups. For example, the myriapod (Pauropoda) species breaks *Arthropoda* monophyly by being attracted to cephalopoda.

The internal relations within *Lophtrochozoa* are particularly unresolved, being prone to change when different models are applied to construct the tree. Only *Gnathostomulida, Platyhelminthes, Entoprocta, Rotifera, and Cycliophora* were recovered as stable monophyletic clades across all trees. All substitution models found *Mollusca* as a polyphyletic group divided into three clades: one clustering *Monoplacophora* and *Polyplacophora* mollusks, a second displaying monophyletic *Gasteropoda*, and another one with lower support grouping the remaining *Mollusca* classes. *Annelida* is only found in the GTR inferred tree, in the other cases the monophyly of the group is broken by a few *Polychaeta* species that fall in independent clades inside *Lophotrochozoa.*



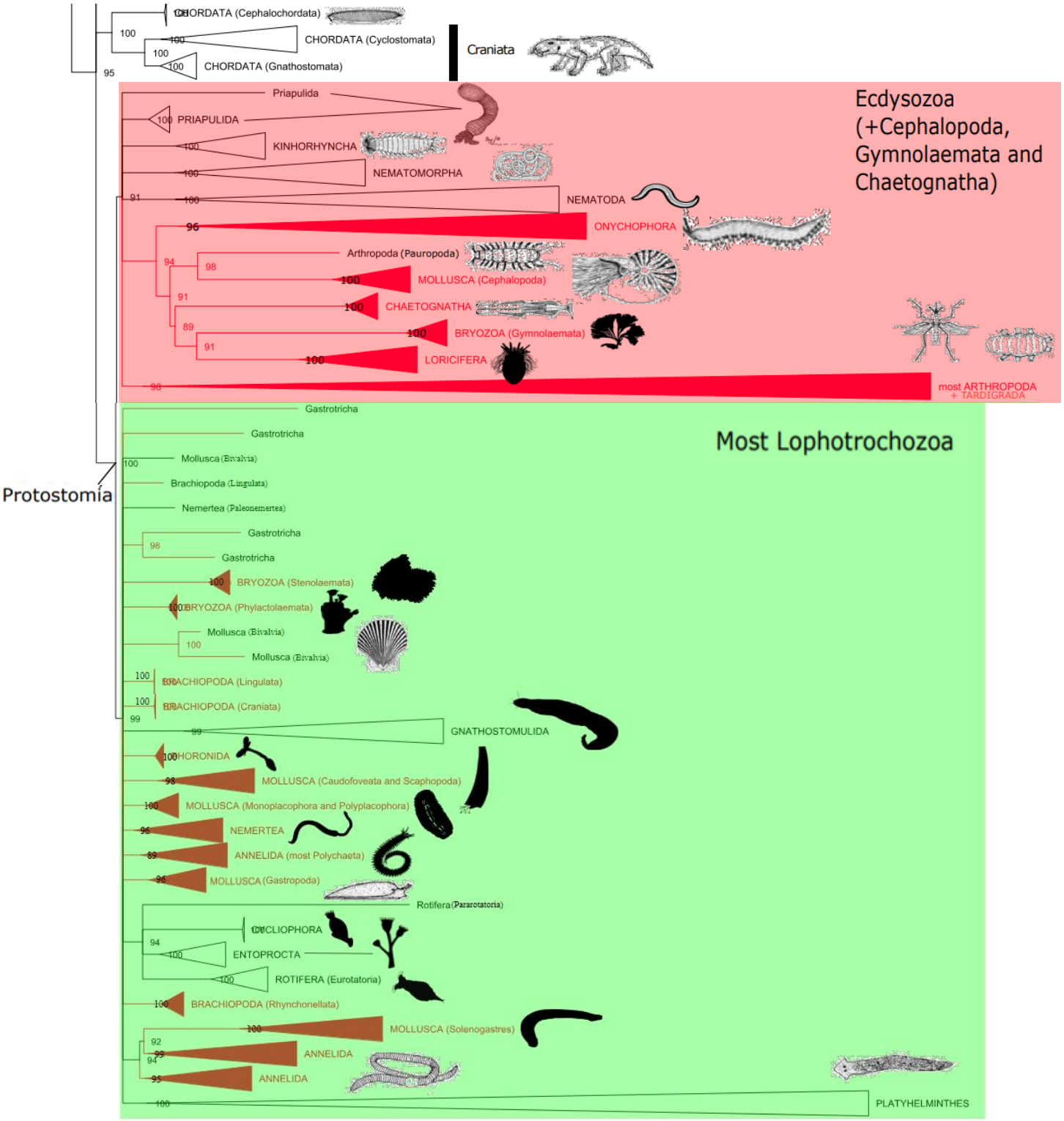
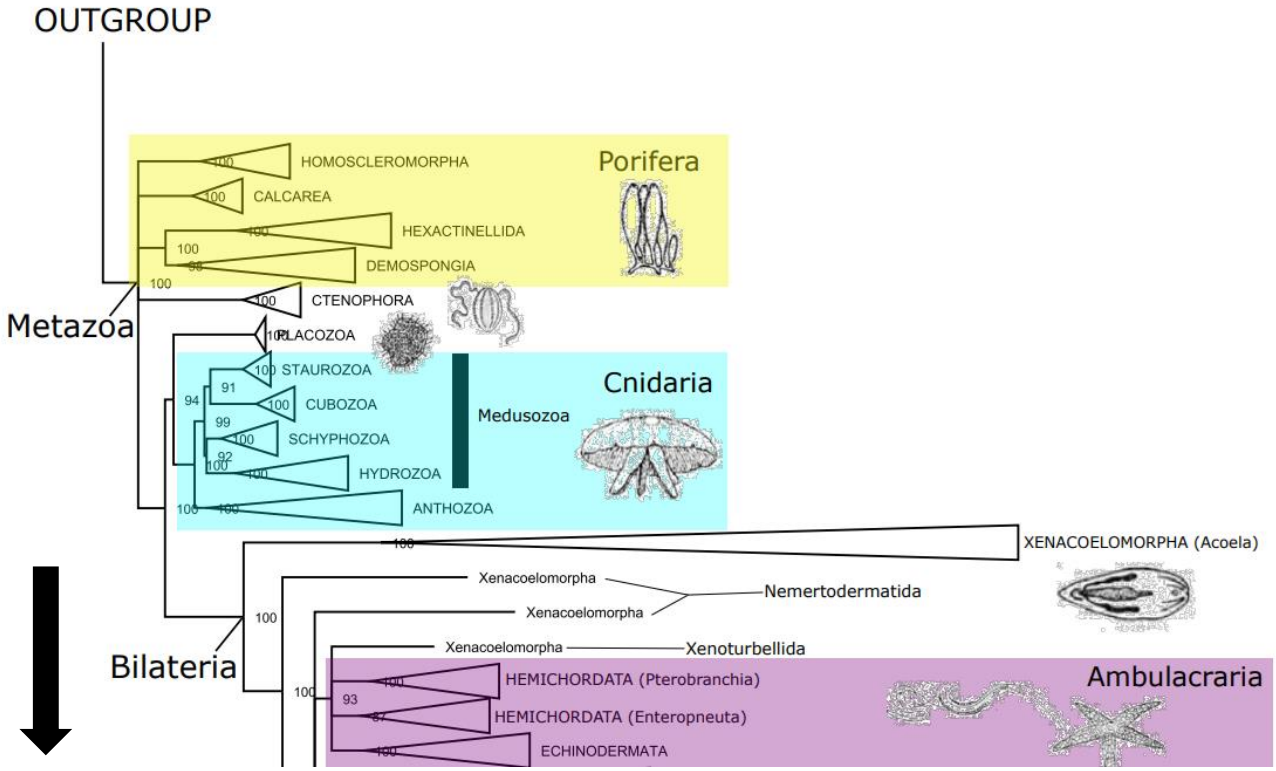


Figure 7. 529 taxa ML tree, calculated from a nearly complete 18S Mafft alignment using GTR+G4+I as substitution model. The unstable nodes when a different substitution model was used are marked in red. The nodes with a bipartition support under 85 have been collapsed.

3.3 Addition of 18S secondary structure information

The two methods used to add the secondary structure information of the alignment yielded divergent results. On the one hand, the pseudoprotein alignment generated with 4SALE resulted in a ML animal tree that could not improve the support for any recognized animal phyla compared to the sequence-only GTR tree. *Entoprocta*, *Rotifera*, *Placozoa*, and *Cycliophora* are the only recognized animal phyla that were recovered as monophyletic. In general, this methodology displayed a topology highly divergent when compared to all sequence-only trees and from the most accepted animal phylogeny.

On the other hand, the alignment generated with the guided sequence-structure alignment SSU-align displayed a ML tree consistent with our sequence-only results, but with two key changes (Fig. 8). It placed *Cnidaria* and *Placozoa* inside a highly supported clade sister to *Bilateria*, and it recovered with high support the monophyly of *Chordata* sister to *Protostomia*. Nevertheless, this alignment method did not solve the long-branching taxa clustered in the middle of *Ecdysozoa* that was observed in the sequence-only trees, nor improve the relationships within *Protostomia* phyla



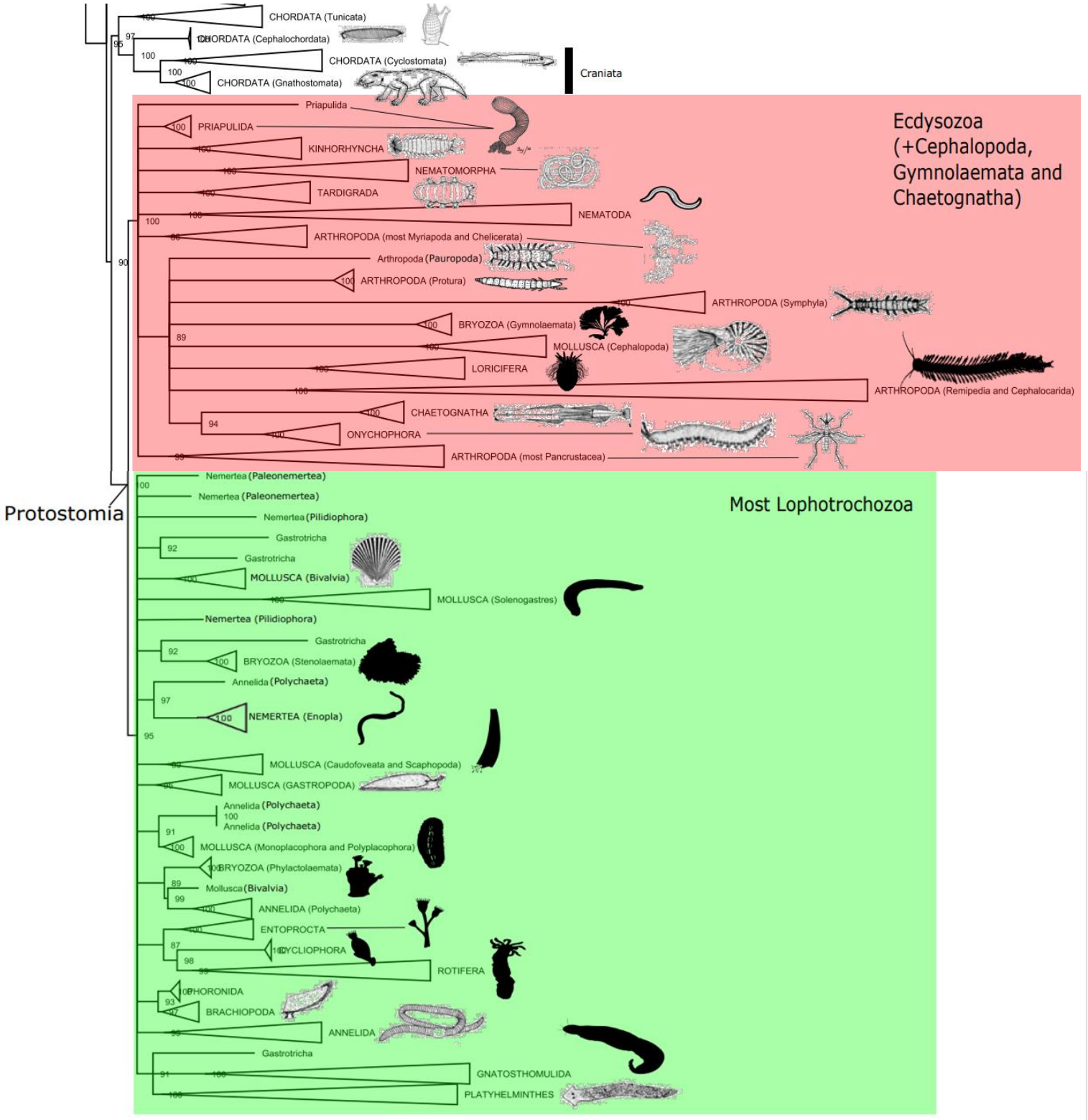


Figure 8. 529 taxa ML tree, calculated from a nearly complete 18S SSU-ali (structure-aware) alignment using GTR+G4+I as substitution model. The nodes with a bipartition support under 85 have been collapsed.

To determine which changes in the alignment could have caused the differences in topology that we observed between the sequence-only (Mafft aligned) and sequence-structure (SSU-align) trees, alignments were compared. As can be seen in Table 1, both untrimmed versions displayed similar alignment lengths, with the sequence-structure alignment being less than 100 nucleotides longer, but the level of similarity between measure by column score was low (around 9% of columns were fully identical). Additionally, both alignments presented a huge number of gaps, and at least one gap in every column. However, the main variation between the two kinds of alignments came after trimming with ClipKit. Only four sites were trimmed from the sequence-structure alignment and up to 2,367 sites were trimmed from the sequence-only alignment. This uneven proportion of trimmed sites caused the trimmed alignments to be almost completely different from one another, with a column score of only 0.0004.

Table 1. Comparison of SSU-align sequence-structure alignment and the Mafft sequence-only alignment.
Longer alignments are associated with strong phylogenetic signal.
Longer alignments when excluding sites with gaps are associated with strong phylogenetic signal.
Column score is a metric calculated by summing the coincident columns between two alignments over all columns in an alignment. Values range from 0 to 1 and higher values indicate more similar alignments [52].

	SSU-align alignment		**Mafft alignment**	
	Trimmed	**Not trimmed**	**Trimmed**	**Not trimmed**
Alignment length	7,723	7,727	5,291	7,658
Alignment length no gaps	0	0	0	0
Column Score	0.0004	0.0938	0.0004	0.0938

4. DISCUSSION

4.1 Curating PR2 database and generating a backbone tree for non-bilaterians

As stated in the introduction of this work, widely used reference DNA databases can still contain errors that, if not properly corrected, might hinder metabarcoding studies creating flaws in our biodiversity knowledge. Among the sequences identified as non-bilaterian

animals in the PR2 database, multiple contamination issues and miss-annotations were discovered in our test phylogenetic trees. If these problems were found at class and phyla levels, one could be concerned about the situation at shallower phylogenetic levels such as order or family. Thus, deeper curation in the main DNA sequence repositories should be encouraged for all groups in the ToL. Unsurprisingly, sponges were the phylum with the most contamination issues, with dozens of sequences annotated as *Porifera* which clustered with unicellular eukaryotes or bilaterian animals. Sponges are known to support diverse microbial and macrofaunal communities [47], which can cause contamination during DNA isolation process. Another issue in the PR2 database concerns taxonomic annotation, as for example *Staurozoa* and *Polypodiozoa* were absent classes belonging to Cnidaria, which are well described cnidarian groups since the early 2000s [48]. As a consequence, all sequences belonging to these two classes were wrongly annotated as *Schyphozoa* an *Hydrozoa* respectively.

Even more remarkable are the PR2 sequences that formed undescribed clades in the tree, such as the clades of early-branching *Porifera* and *Cnidaria,* and the multiple sequences that fell outside of all known ctenophores. These oddly placed sequences come from oceanic environmental samples from different independent studies and tend to be short and incomplete, most of them between 500 and 900 nucleotides. Despite some of these sequences being placed with high support, it is still early to hypothesise the existence of new clades among non-bilaterian animals based only on these problematic sequences, as they might just represent artifacts. However, an eye must be kept on them because they may constitute undescribed animal lineages. Future efforts should be focused on retrieving full-length 18S sequences and corroborating their sequence identity and taxonomy.

The final output of the PR2 database curation was a strong and well-supported ML phylogenetic tree containing thousands of 18S sequences spanning all non-bilaterian animals present in the database. This curated dataset will serve as a *bona fide* backbone phylogenetic tree in which novel 18S sequences from metabarcoding studies can be mapped using phylogenetic placement techniques. López-Escardó et al. (2018) already generated a phylogenetically curated metazoan 18S rRNA reference dataset to identify

barcoded sequences taxonomically [14]. However, this backbone tree was focused on the more taxon-rich bilaterian animals, with more than half of the entries in the dataset constituted by arthropods. It also contained 203 *Porifera*, 991 *Cnidaria*, 20 *Ctenophora* and 4 *Placozoa*, but these are far fewer than the 1,551 *Porifera*, 3,798 *Cnidaria*, 265 *Ctenophora* and 18 *Placozoa* sequences included in our backbone dataset. Therefore, used alone or even as an extension of López-Escardó and colaborator's dataset, it will be a useful tool to identify novel 18S metabarcodes that could belong to undescribed taxa among non-bilaterian animals.

4.2 The reliability of 18S-based animal phylogeny

The main reason for the abandonment of 18S-based phylogenies in favour of phylogenomics is the need for more data to overcome stochastic error. Stochastic error arises from the use of a small number of molecular characters (such as a single gene) that contain weak phylogenetic signal. With few nucleotides or amino acids in a data set, random substitutions into the same characters (homoplasies that mislead the number of substitutions inferred) may predominate over informative sites at certain branches of the tree [49]. Phylogenomic scale datasets drastically reduces stochastic errors in comparison to 18S locus alone, which rarely contains more than 2000 nucleotides. However, after stochastic error is minimised using larger datasets, systematic error becomes the main source of problem. This kind of error mainly stems from using incorrect model assumptions [49], such as using substitution models that cannot properly describe the heterogeneous process of evolution (i.e. genes and sites evolving at different pace). In theory, systematic error should be solvable by improving the models of evolution. However, it can be argued that due to the high complexity and heterogeneity of evolutionary processes, this methodological error will never be fully overcomed. Thus, it is interesting to identify which clades of the 18S-based animal phylogeny are most susceptible to change when using different substitution models and compare them to the current phylogenomic perspective on animal phylogeny. Thereby, determining the extent to which the divergence between these two approaches (few reliable data versus massive heterogeneous data) is caused by systematic rather than stochastic error.

Nowadays, most clades in the animal ToL are well-supported and apparently resolved using phylogenomic approaches. The current consensus of animal phylogeny broadly supports *Metazoa* monophyly and that all living animals belong to one of five monophyletic groups: *Porifera*, *Ctenophora, Cnidaria*, *Placozoa,* or *Bilateria* [49–51]. In the trees inferred with our 18S dataset, these hypotheses are also stable and well-supported, except for *Porifera,* which appears as a paraphyletic group. The paraphyly of sponges has been historically supported by SSU data [39] and was even recovered in a phylogenomic study conducted by Sperling et al. (2009), but only considering a relatively small number of nuclear genes [52]. Thus, *Porifera* paraphyly can be a good example of an artifact caused by a stochastic error that can be solved when large phylogenomic datasets are applied. Moreover, all trees inferred with our dataset supported *Placozoa* as a sister group to bilaterians. This relationship has been supported before by 18S data [36]. However, this is not concordant with the current phylogenomic consensus, which suggests that cnidarians are the sister group to *Bilateria* [49, 50], or sister to Placozoa [53].

Among bilaterians, *Deuterostomia,* and *Protostomia* are well-accepted clades in phylogenomics analyses [49-51], but the first one is never recovered with the 18S dataset used here. While *Deuterostomia* is a traditionally accepted clade, support for this grouping is weak compared to the support for *Protostomia* [51], and even using phylogenomic data have been proposed that the monophyly of *Deuterostomia* might derive from a systematic error [54]. Nevertheless, other 18S-based phylogenies have been able to recover a monophyletic *Deuterostomia* [55]. Regarding the monophyly of the phylum *Chordata,* major phylogenomic studies agree on a monophyletic *Chordata* with the two subphylums *Tunicata* and *Craniata* having a sister group relation [51]. However, in past 18S-based animal phylogenies, the support for the monophyly of *Chordata* is weak, mainly due to the variable placement of *Tunicata.* Among other positions, tunicates have been found as the basal deuterostome lineage, as in our dataset [39]. Hence, chordates paraphyly may be another case of stochastic error alleviated with genomic-scale data.

Within *Protostomia* there are three supported clades: *Chaetognatha*, *Ecdysozoa*, and *Lophotrocohozoa*. In all the 18S-based trees inferred in this study, these three major

groups are fully retrieved to a greater or lesser extent by the addition of long-branching protostome taxa within *Ecdysozoa.* This probably artifactual clade that contains all, or part of the following taxa: *Cephalopoda*, *Chaetognatha*, *Cephalocarida, Remipedia, Onychophora*, and *Pauropoda* myriapods, seems a clear case of Long-branch attraction (LBA). This phylogenetic artifact appears when rapidly evolving lineages are incorrectly inferred as closely related because they have undergone multiple molecular substitutions, and not because they are related by descent [56]. Furthermore, the studies on animal phylogeny inferred from combined 18S-28S data conducted by Mallat et al. [57, 58] also recovered the same clustering of long-branching taxa near the base of Arthropoda.

Despite the growing number of genomic and transcriptomic data available in public repositories, there are still some major areas of controversy inside the animal ToL among phylogenomic studies. These concerns the root of the animal tree, the placement of *Xenacoelomorpha* within *Bilateria*, and the exact relationships within *Protostomia* [49-51]. Regarding the root of the tree, 18S phylogenetic studies strongly support the *Porifera* sister hypothesis [34, 39, 55, 59]. In concordance, early-splitting *Porifera* groups remain topologically stable across all models tested with our dataset. Both phylogenomic and 18S phylogenetic analyses agree that three groups of marine worms, *Xenoturbellida* and the two related acoelomorph flatworm groups *Acoela* and *Nemertodermatida*, do not cluster with other *Platyhelminthes* [39, 49]. Previous 18S-based phylogenies, as well as the 18S trees generated here, fail to recover a monophyletic *Xenoacoelomorpha*, and in most cases, *Xenoturbellida* is placed among *Ambulacraria* [32].

On the relations inside *Protostomia*, current phylogenomic studies agree on the phyla belonging to *Ecdysozoa* and *Lophotrochozoa*, but the inner relationships are less clear [49]. Three sub-groups within *Ecdysozoa* seem to have a strong support in phylogenomic studies, *Nematoida* (*Nematoda* + *Nematomorpha*), *Scalidophora* (*Kinorhyncha* + *Priapulida* + *Loricifera*) and *Panarthropoda* (*Arthropoda* + *Tardigrada* + *Onychophora*) [49, 50]. None of these clades were supported in our phylogeny and they are not recovered in most 18S-based animal phylogenies [55]. *Lophotrochozoa* remains the major clade with the poorest internal resolution in rRNA-based phylogenies [60, 61]. This difficulty is especially evident within *Annelida* and *Mollusca*, both of which have several separate subgroups. Finally, *Chaetognatha*, or 'arrow worms', have been linked by different

phylogenomic studies to *Ecdysozoa*, *Lophotrochozoa*, or as a sister group to both [49]. Some other 18S phylogenies have already reported chaetognathe worms as part of *Ecdysozoa* [55, 62], as in the case of this study. Nevertheless, these previous studies acknowledge that LBA could be involved in this phylogenetic position.

4.3 The effect of adding secondary structure to 18S-based phylogeny of animals

We tested two methods to construct a structurally aware 18S alignment based on our dataset: a) convert the sequence-structure information of the entire alignment into a 12-letter alphabet pseudoprotein sequence, as implemented in 4SALE, and b) use SSU-align to build a guided sequence-structure alignment of our dataset based on a covariance model from eukaryotic 18S. The first method, using 4SALE, can be considered as unsuccessful due to its failure to recover most of the uncontested clades, such as phylum groups, in the animal kingdom. This result is surprising because the experiment followed the same methodology that Rapp and Wolff (2024) successfully used to construct an 18S rDNA sequence-structure phylogeny of eukaryotes [8]. In their study, Rapp and Wolf concluded that all tested sequence-structure approaches showed improvements compared to the respective sequence-only approaches, obtaining higher bootstrap support and recovering sister relationships between groups much more comparable to results obtained by multigene analyses. Although this methodology has not been applied to *Metazoa* phylogenies, it has successfully inferred phylogenies for many other groups of eukaryotes [63-65], consistently yielding accurate and robust trees. Therefore, we believe that the surprising results of our analysis using the 4SALE methodology might be due to some methodological error from our side. We plan to study in depth the causes of these discrepancies.

The second methodology based on a guided-structure alignment, using SSU-align, displayed more coherent results. The ML tree obtained from this alignment recovered almost the same clades present in the sequence-only GTR+G4+I tree. The two key changes in topology that the addition of secondary structure information caused (*Chordata* monophyly and a sister group relationship between *Cnidaria* and *Bilateria*) brought the 18S-based animal phylogeny closer to the current consensus based on

phylogenomic analyses. We are unaware of previous work constructing a full *Metazoa* ML phylogenetic tree employing just the 18S sequence-structure alignment. However, Mallatt et al (2010), performed a similar analysis using a sequence-structure manual alignment of combined SSU-LSU data from 371 taxa [58]. Despite using more nucleotides and GTR+G4 as substitution model, Mallat's sequence-structure phylogeny is less concordant with phylogenomics consensus than the results presented here, at least when observing the relations at phyla and supraphyla levels. Monophyletic *Chordata* and *Deuterostomia* were neither obtained due to *Tunicata* falling among *Lophotrochozoa* phyla. Furthermore, the relations among non-bilaterian phyla they found are much less clear than in our case, but that may be attributed to our more extensive taxon sampling on this area of the animal tree.

Addressing the characteristics of the two alignments –sequence only (Mafft) versus guided sequence-structure (SSU-align)– might facilitate finding the drivers of the changes in topology in the resulting ML trees. Both alignments are long (for an 18S alignment) and gappy, but that could be expected acknowledging that the aligned sequences are very divergent. One of the main differences between the two alignments is the divergent effect that smart-gap trimming implemented in ClipKit has in each of them. The smart-gap trimming method analyzes the distribution of gaps across the alignment to identify an optimal threshold that minimises excessive trimming, particularly in highly divergent sequences. The smart-gap approach implemented in the trimming software dynamically determines a threshold based on the gap distribution, meaning that alignments with more extensive or clustered gaps will result in greater trimming [40]. The sequence-only alignment indeed seems to present a higher concentration of gaps in certain regions compared to the guided sequence-structure alignment. The greater distribution of gaps in the sequence-structure alignment could explain, at least in part, the few changes in topology that are observed from the phylogenies resulting from both alignments.

5. CONCLUSIONS

- This thesis contributes to the community effort to create a proper database for 18S rRNA. In this process, we generated a backbone phylogenetic tree containing 5632 tips for non-bilaterian animals, which can serve as a tool for the taxonomic identification of 18S environmental sequences. These results will enhance our understanding of known biodiversity, uncover hidden diversity, and hopefully improve animal phylogenies.

- Despite being composed of fewer than 2000 nucleotides, the 18S gene successfully recovers most of the accepted clades in animal phylogeny to the phylum level. The main discrepancies between our sequence-only 18S-based animal phylogeny and the phylogenomic consensus include: paraphyletic *Porifera*, *Placozoa* sister to *Bilateria*, the polyphyly of *Deuterostomia*, *Chordata*, and *Xenoambulacraria.* Additionally, major protostome groups are truncated in our 18S phylogeny due to LBA issues.

- The guided sequence-structure alignment generated by SSU-align produced a metazoan phylogeny that is more congruent with the phylogenomic consensus than the phylogeny derived from the sequence-only alignment. This highlights that the incorporation of secondary structure information may improve 18S-based phylogenies.

BIBLIOGRAPHY

1- Nawrocki, E. P. (2009). *Structural RNA homology search and alignment using covariance models*. Washington University in St. Louis

2- Noller, H. F. (1991). Ribosomal RNA and translation. *Annual review of biochemistry*, *60*(1), 191-227.

3- 23- Noller, H. F. (2005). RNA structure: reading the ribosome. *Science*, *309*(5740), 1508-1514.

4- Hillis, D. M., & Dixon, M. T. (1991). Ribosomal DNA: Molecular Evolution and Phylogenetic Inference. *The Quarterly Review of Biology*, *66*(4), 411–453.

5- Machida, R. J., & Knowlton, N. (2012). PCR primers for metazoan nuclear 18S and 28S ribosomal DNA sequences.

6- Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088-5090.

7- Xie, Q., Lin, J., Qin, Y., Zhou, J., & Bu, W. (2011). Structural diversity of eukaryotic 18S rRNA and its impact on alignment and phylogenetic reconstruction. *Protein & cell*, *2*(2), 161-170.

8- Rapp, E., & Wolf, M. (2024). 18S rDNA sequence-structure phylogeny of the eukaryotes simultaneously inferred from sequences and their individual secondary structures. *BMC Research Notes*, *17*(1), 124.

9- Lis, J. A. (2023). Molecular Apomorphies in the Secondary and Tertiary Structures of Length-Variable Regions (LVRs) of 18S rRNA Shed Light on the Systematic Position of the Family Thaumastellidae (Hemiptera: Heteroptera: Pentatomoidea). *International Journal of Molecular Sciences*, *24*(9), 7758

10- Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., ... & de Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, *18*(8), 428-445.

11- Duarte, C. M. (2015). Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition.

12- Planes, S., & Allemand, D. (2023). Insights and achievements from the Tara Pacific expedition. *Nature communications*, *14*(1), 3131.

13- Anderson, R. F. (2020). GEOTRACES: Accelerating research on the marine biogeochemical cycles of trace elements and their isotopes. *Annual Review of Marine Science*, *12*(1), 49-85.

14- López-Escardó, D., Paps, J., De Vargas, C., Massana, R., Ruiz-Trillo, I., & Del Campo, J. (2018). Metabarcoding analysis on European coastal samples reveals new molecular metazoan diversity. *Scientific reports*, *8*(1), 9106.

15- Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., ... & Ettema, T. J. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173-179.

16- Hajibabaei, M., Singer, G. A., Hebert, P. D., & Hickey, D. A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *TRENDS in Genetics*, *23*(4), 167-172.

17- S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. (2005) Comparative metagenomics of microbial communities. Science, 308:554–557.

18- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular ecology*, *21*(8), 2045-2050.

19- N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen. Analyzing natural microbial populations by rRNA sequences. ASM News, 51:4–12, 1985

20- Clarke, L. J., Suter, L., Deagle, B. E., Polanowski, A. M., Terauds, A., Johnstone, G. J., & Stark, J. S. (2021). Environmental DNA metabarcoding for monitoring metazoan biodiversity in Antarctic nearshore ecosystems. *PeerJ*, *9*, e12458.

21- Di Capua, I., Piredda, R., Mazzocchi, M. G., & Zingone, A. (2021). Metazoan diversity and seasonality through eDNA metabarcoding at a Mediterranean long-term ecological research site. *ICES Journal of Marine Science*, *78*(9), 3303-3316.

22- Capra, E., Giannico, R., Montagna, M., Turri, F., Cremonesi, P., Strozzi, F., ... & Pizzi, F. (2016). A new primer set for DNA metabarcoding of soil Metazoa. *European Journal of Soil Biology*, *77*, 53-59.

23- Czech, L., Stamatakis, A., Dunthorn, M., & Barbera, P. (2022). Metagenomic analysis using phylogenetic placement—a review of the first decade. Frontiers in Bioinformatics, 2, 871393.

24- Del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L. F., Nenarokov, S., Massana, R., ... & Wegener Parfrey, L. (2018). EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS biology*, *16*(9), e2005849.

25- Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, *35*(21), 7188-7196.
26- Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... & Christen, R. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic acids research*, *41*(D1), D597-D604.
27- Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., ... & Yaschenko, E. (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, *36*(suppl_1), D13-D21.
28- Steenwyk, J., & King, N. (2024). From Genes to Genomes: Opportunities, Challenges, and a Roadmap for Synteny-based Phylogenomics.
29- Li, Y., Shen, X. X., Evans, B., Dunn, C. W., & Rokas, A. (2021). Rooting the animal tree of life. *Molecular Biology and Evolution*, *38*(10), 4322-4333.
30- Whelan, N. V., Kocot, K. M., Moroz, L. L., & Halanych, K. M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences*, *112*(18), 5773-5778.
31- Jékely, G., & Budd, G. E. (2021). Animal phylogeny: resolving the slugfest of ctenophores, sponges and acoels?. *Current Biology*, *31*(4), R202-R204.
32- Halanych, K. M. (2016). How our view of animal phylogeny was reshaped by molecular approaches: lessons learned. *Organisms Diversity & Evolution*, *16*, 319-328.
33- Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., Raff, E. C., Pace, N. R., & Raff, R. A. (1988). Molecular phylogeny of the animal kingdom. *Science (New York, N.Y.)*, *239*(4841 Pt
34- Halanych, K. M. (1996). Convergence in the feeding apparatuses of lophophorates and pterobranch hemichordates revealed by 18S rDNA: an interpretation. *The Biological Bulletin*, *190*(1), 1-5.
35- Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., & Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, *387*(6632), 489-493.
36- Collins, A. G. (1998). Evaluating multiple alternative hypotheses for the origin of Bilateria: an analysis of 18S rRNA molecular evidence. *Proceedings of the National Academy of Sciences*, *95*(26), 15458-15463.
37- Ruiz-Trillo, I., Riutort, M., Littlewood, D. T. J., Herniou, E. A., & Baguna, J. (1999). Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science*, *283*(5409), 1919-1923.
38- Halanych, K. M. (1995). The phylogenetic position of the pterobranch hemichordates based on 18S rDNA sequence data. *Molecular phylogenetics and Evolution*, *4*(1), 72-76.
39- Halanych, K. M. (2004). The new view of animal phylogeny. *Annu. Rev. Ecol. Evol. Syst.*, *35*, 229-256.
40- Steenwyk, J. L., Buida III, T. J., Li, Y., Shen, X. X., & Rokas, A. (2020). ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS biology*, *18*(12), e3001007.
41- Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268-274.
42- Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*, *9*(4), 286-298.
43- RNAcentral Consortium. (2021). RNAcentral 2021: Secondary structure integration, improved sequence search, and new member databases. *Nucleic Acids Research*, *49*(D1), D212-D220.
44- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. ClustalW and ClustalX Version 2.0. Bioinformatics. 2007.
45- Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M. 4SALE—a tool for synchronous RNA sequence and secondary structure alignment and editing. BMC Bioinformatics. 2006.
46- Rambaut, A. (n.d.). FigTree v1.4.4. Retrieved from http://tree.bio.ed.ac.uk/software/figtree/
47- Bell, J. J. (2008). The functional roles of marine sponges. *Estuarine, coastal and shelf science*, *79*(3), 341-353.
48- Miranda, L. S., Mills, C. E., Hirano, Y. M., Collins, A. G., & Marques, A. C. (2018). A review of the global diversity and natural history of stalked jellyfishes (Cnidaria, Staurozoa). *Marine Biodiversity*, *48*, 1695-1714.
49- Telford, M. J., Budd, G. E., & Philippe, H. (2015). Phylogenomic insights into animal evolution. *Current Biology*, *25*(19), R876-R887.
50- Dunn, C. W., Giribet, G., Edgecombe, G. D., & Hejnol, A. (2014). Animal phylogeny and its evolutionary implications. *Annual review of ecology, evolution, and systematics*, *45*(1), 371-395.

51- Giribet, G. (2016). New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. *Organisms Diversity & Evolution*, *16*, 419-426.
52- Sperling, E. A., Peterson, K. J., & Pisani, D. (2009). Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Molecular biology and evolution*, *26*(10), 2261-2274.
53- Laumer, C. E., Gruber-Vodicka, H., Hadfield, M. G., Pearse, V. B., Riesgo, A., Marioni, J. C., & Giribet, G. (2018). Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. *Elife*, *7*, e36278.
54- Kapli, P., Natsidis, P., Leite, D. J., Fursman, M., Jeffrie, N., Rahman, I. A., ... & Telford, M. J. (2021). Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria. *Science advances*, *7*(12), eabe2741.
55- Peterson, K. J., & Eernisse, D. J. (2001). Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. *Evolution & development*, *3*(3), 170-205.
55- Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS biology*, *9*(3), e1000602.
57- Mallatt, J., Craig, C. W., & Yoder, M. J. (2012). Nearly complete rRNA genes from 371 Animalia: updated structure-based alignment and detailed phylogenetic analysis. *Molecular phylogenetics and evolution*, *64*(3), 603-617.
58- Mallatt, J., Craig, C.W., Yoder, M.J., 2010. Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. Mol. Phylogenet. Evol. 55, 1–17.
59- Kim, J., Kim, W., & Cunningham, C. W. (1999). A new perspective on lower metazoan relationships from 18S rDNA sequences. *Molecular Biology and Evolution*, *16*(3), 423-427.
60- Bourlat, S.J., Nielsen, C., Economou, A.D., Telford, M.J., 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. Mol. Phylogenet.
61- Struck, T.H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K.M., 2007. Annelid phylogeny and the status of Sipuncula and Echiura. BMC Evol. Biol. 2007 (7), 57.
62- Halanych, K. M. (1996). Testing hypotheses of chaetognath origins: long branches revealed by 18S ribosomal DNA. *Systematic Biology*, *45*(2), 223-246.
63- Buchheim, M. A., Müller, T., & Wolf, M. (2017). 18S rDNA sequence-structure phylogeny of the Chlorophyceae with special emphasis on the Sphaeropleales. *Plant Gene*, *10*, 45-50.
64- Borges, A. R., Engstler, M., & Wolf, M. (2021). 18S rRNA gene sequence-structure phylogeny of the Trypanosomatida (Kinetoplastea, Euglenozoa) with special reference to Trypanosoma. *European Journal of Protistology*, *81*, 125824.
65- Rackevei, A. S., Karnkowska, A., & Wolf, M. (2023). 18 S r DNA sequence–structure phylogeny of the Euglenophyceae (Euglenozoa, Euglenida). *Journal of Eukaryotic Microbiology*, *70*(2), e12959.
66- Aubert, M., O'Donohue, M. F., Lebaron, S., & Gleizes, P. E. (2018). Pre-ribosomal RNA processing in human cells: from mechanisms to congenital diseases. *Biomolecules*, *8*(4), 123.