ESCI upf.
School of International Studies

# MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

## Gene families and the origins of complexity: A comparative genomics approach to better understand multicellularity


| | |
|---|---|
| **AUTOR/A:** Blai Crespo Selma | **NIA:** 106760 |
| **GRAU:** Bachelor Degree in Bioinformatics | |
| **CURS ACADÈMIC:** 3r | |
| **DATA:** 18 de Juny de 2024 | |
| **TUTOR/S:** Marta Alvarez-Presas | |

## FULL DE RESUM DEL TREBALL DE FI DE GRAU DEL BDBI (ESCI-UPF)

| | |
|---|---|
| **TÍTOL DEL PROJECTE:** Gene families and the origins of complexity: A comparative genomics approach to better understand multicellularity | |
| **AUTOR/A:** Blai Crespo Selma | **NIA:** 106760 |
| **CURS ACADÈMIC:** 3r | |
| **DATA:** 18 de Juny de 2024 | |
| **TUTOR/S:** Marta Alvarez-Presas | |

**PARAULES CLAU (mínim 3)**

- Català: Unicel·lularitat Holozoa, factors de transcripció, genòmica comparada, gens homeobox, filogènia, LIM, Sox
- Castellà: Unicelularidad Holozoa, Factores de Transcripción, Genómica Comparada, Genes Homeobox, Filogenia, LIM, Sox
- Anglès: Unicellular Holozoa, Transcription Factors, Comparative genomics, Homeobox genes, Phylogeny, LIM, Sox

**RESUM DEL PROJECTE (extensió màxima: 100 paraules per llengua)**


- Català: L'aparició d'organismes pluricel·lulars complexos planteja infinites preguntes en l'àmbit de la biologia evolutiva, especialment pel que fa als complicats mecanismes reguladors subjacents a la transició d'organismes unicel·lulars a entitats multicel·lulars complexes. Entre els factors fonamentals que impulsen l'evolució de la vida pluricel·lular, cal destacar l'aparició i diversificació de famílies de gens, com els factors de transcripció homeobox, que són clau en el desenvolupament i l'evolució dels animals. En la nostra recerca sobre la trajectòria evolutiva cap a la multicel·lularitat, ens centrem en l'anàlisi d'algunes famílies de gens homeobox utilitzant un bon repositori de genomes i transcriptomes complets al nostre abast, i fem una exhaustiva anàlisi genòmica comparativa mitjançant eines bioinformàtiques especialitzades, en un conjunt de dades que inclou una amplia mostra de taxons d'holozoa, cosa que encara no s'ha fet.

- Castellà: La aparición de organismos multicelulares complejos plantea infinitas preguntas dentro del ámbito de la biología evolutiva, particularmente en relación con los complicados mecanismos reguladores que subyacen a la transición de organismos unicelulares a entidades multicelulares complejas. Entre los factores fundamentales que impulsan la evolución de la vida multicelular, cabe destacar la aparición y diversificación de familias de genes, como los factores de transcripción

homeobox, que son clave en el desarrollo y la evolución animal. En nuestra investigación sobre la trayectoria evolutiva hacia la multicelularidad, nos centramos en el análisis de algunas familias de genes homeobox utilizando un buen repositorio de genomas y transcriptomas completos a nuestra disposición, y realizamos un análisis genómico comparativo exhaustivo utilizando herramientas bioinformáticas especializadas, sobre un conjunto de datos que incluye un amplio muestreo de taxones de Holozoa, algo que aún no se ha hecho.

- Anglès: The emergence of complex multicellular organisms raises infinite questions within the realm of evolutionary biology, particularly regarding the complicated regulatory mechanisms underlying the transition from unicellular organisms to complex multicellular entities. Among the fundamental factors driving the evolution of multicellular life, it is worth highlighting the emergence and diversification of gene families, such as homeobox transcription factors (TF), which are key in animal development and evolution. In our research on the evolutionary trajectory towards multicellularity, we focus on the analysis of some homeobox gene families using a good repository of complete genomes and transcriptomes at our disposal, and we conduct an exhaustive comparative genomic analysis using specialized bioinformatics tools, on a dataset that includes a broad taxon sampling of Holozoa, something that has not been done yet.

ESCI upf. School of International Studies

*Bachelor's Degree in Bioinformatics (UPF-UPC-UB)*

*Final Degree Project*

# Gene families and the origins of complexity: A comparative genomics approach to better understand multicellularity

Blai Crespo Selma[1]

Scientific director: Marta Álvarez-Presas[1]

[1] IBE, Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Pg. Maritim de la Barceloneta, 37, Ciutat Vella, 08003 Barcelona

**Abstract**

**Motivation:** The emergence of complex multicellular organisms raises infinite questions within the realm of evolutionary biology, particularly regarding the complicated regulatory mechanisms underlying the transition from unicellular organisms to complex multicellular entities. Among the fundamental factors driving the evolution of multicellular life, it is worth highlighting the emergence and diversification of gene families, such as homeobox transcription factors (TF), which are key in animal development and evolution. In our research on the evolutionary trajectory towards multicellularity, we focus on the analysis of some homeobox gene families using a good repository of complete genomes and transcriptomes at our disposal.

**Results:** We conduct an exhaustive comparative genomic analysis using specialized bioinformatics tools, on a dataset that includes a broad taxon sampling of Holozoa, something that has not been done yet. Through this approach, we have been able to find some proof of gene families like Sox or LIM in these taxa, even though they were before categorized as animal-specific. Consequently, this work conducted on other crucial gene families might help connect early-branching unicellular animal relatives to their multicellular animal counterparts.

**Supplementary information:** Supplementary data are available at this GitHub link, https://github.com/Blaicrespo/FINAL-DEGREE-PROJECT

## 1 Introduction

### From unicellularity to multicellularity

The transition from unicellular to multicellular life represents one of the most profound evolutionary leaps in the history of biology. This transition has occurred repeatedly across various eukaryotic lineages, resulting in the emergence of complex multicellular organisms such as animals (Metazoa) [1,2], fungi or algae [3-5]. Understanding how multicellular animals evolved from their unicellular ancestors is a central and challenging question in biology. Animals comprise a diverse group of multicellular organisms characterized by spatial and functional cell differentiation [1,2]. The evolution of multicellularity within the animal kingdom has led to an unparalleled diversity of body plans and developmental processes, far exceeding those found in other lineages. Central to the development of multicellular animals is the intricate regulation of gene expression, involving numerous key genes, signaling pathways, and molecular mechanisms [6,7]. To understand the origins of animal multicellularity, it is imperative to consider the phylogenetic relationships between the earliest branching animals and their closest living unicellular relatives [1,6,8]. The different hypotheses formulated so far suggest that the Last Unicellular Common Ancestor of animals (LUCA) possessed the genetic toolkit necessary for diverse development and reproduction pathways. Understanding the Last

Common Ancestor of animals (LCA) and LUCA, and exploring the evolutionary trajectory between them is crucial in piecing together the evolutionary events leading to animal multicellularity. The diversity of developmental processes observed across different animal phyla highlights the importance of investigating extant unicellular relatives, some of which exhibiting multiple developmental modes [6, 9]. Notably, robust evolutionary relationships have been established in Holozoa, between animals and their closest unicellular relatives, mainly choanoflagellates, filastereans, corallochytreans, and ichthyosporeans, [10, 11, 12, 13] (Figure 1). Members of this clade exhibit at least one temporary "multicellular" state also observed in Metazoa [14].

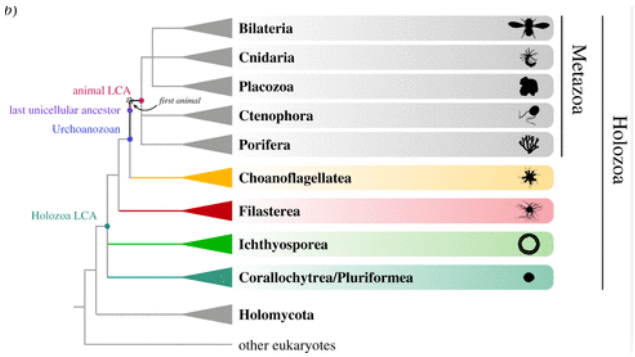


Figure 1: Metazoa and their unicellular relatives. Bilateria, Cnidaria, Placozoa, Ctenophora and Porifera indicate the major clades of Metazoa. Choanoflagellatea, Filasterea, Ichthyosporea and Corallochytrea/Pluriformea indicate the closest unicellular relatives of animals. Image from Ros-Rocher et al. [1].

In order to establish these evolutionary relationships between unicellular and multicellular species, it is essential to study specific transcription factors and basal gene families. Transcription factors (TF) are proteins that bind to specific DNA sequences to regulate gene expression, playing a crucial role in controlling cellular functions and development. They are pivotal in the emergence of multicellularity, highlighting their evolutionary dynamics and contribution to regulatory landscapes. Some of the genes we will discuss encode TFs themselves, and therefore help in some of their functions. We are particularly focusing on the homeobox gene families.

Homeobox genes are broadly categorized into two major classes: TALE (Three Amino Acid Loop Extension) and non-TALE [15, 16]. The TALE class is characterized by a three amino acid loop extension in their homeodomain, which influences their DNA-binding properties and interaction with other proteins. Non-TALE homeobox genes, on the other hand, lack this extension but are equally significant in their roles. The remarkable conservation of both types of homeobox genes across evolutionary time underscores their fundamental importance in developmental processes [17, 18]. Homeobox genes encode TFs that play crucial roles in regulating patterns of anatomical development (morphogenesis) in multicellular organisms. They are involved in the formation of various body structures during early embryonic development by controlling the expression of target genes that dictate cell differentiation and organ formation. This regulatory capacity makes homeobox genes essential for the proper spatial and temporal development of tissues and organs.

Studying these genes offers profound insights into how complex structures have evolved and diversified among different species. By comparing these genes across various species, we can unravel new information about their evolutionary relationships and identify their last common ancestor (LCA). Understanding the functions and evolutionary history of homeobox genes will provide significant contributions to our knowledge of multicellularity and evolutionary biology [19].

**Evolutionary Dynamics of Gene Families: From Unicellular to Multicellular Lineages**

The exploration of unicellular organisms has offered valuable insights into the early stages of multicellular evolution, particularly through studies of animal relatives like choanoflagellates and the filasterean *Capsaspora owczarzaki* [20, 21]. Research on their genomic landscapes and studies on the expansion of TALE and non-TALE homeobox genes and the evolution of TFs provide compelling evidence of genetic innovations driving the transition to multicellularity [22].

Some of the homeobox genes previously mentioned are or function as TFs themselves. Key research by de Mendoza [23] and Ferrier [24] explore

transcription factors and comparative studies, such as those by Joo [25] and de Mendoza and Sebé-Pedrós [26], reveal shared evolutionary patterns across diverse eukaryotic lineages. Advances in predictive modeling by Lambert [27] and insights from Bobola and Sagerström [28] deepen our understanding of TF evolution and functionality.

Recent genomic sequencing advancements, including work by de Mendoza and Ruiz-Trillo [29], uncover patterns of evolutionary conservation across multicellular and unicellular organisms, providing critical insights into the regulatory toolkit driving multicellular lineages. Notably, studies on key TFs like LIM homeobox genes and TALE homeobox gene families [30, 31] shed light on their central roles in developmental processes, offering invaluable insights into regulatory networks governing multicellular development.

**Gene Families Crucial for Multicellularity: Exploration and Analysis**

If we want to know more about our unicellular ancestors, investigating these gene families is essential to understanding the transition from unicellular to multicellular, given that they are very well conserved. Therefore, homeobox genes, especially Hox and ParaHox genes, which are animal-specific [32], are critical in multicellular development. These genes encode TFs that regulate other genes, essential for developmental processes [33, 34]. It has been suggested that Hox and ParaHox gene clusters existed in the LCA of animals, but they have not been found in basal lineages, so their phylogenies are poorly resolved. [35]

Srivastava et al. [36] highlighted the early evolution of LIM homeobox genes, important for cell fate and patterning. Grau-Bové et al. [22] explored genomic innovations in unicellular ancestors, revealing advanced regulatory networks predating multicellularity.

Sox family genes, vital in vertebrate sex determination and reproduction, have ancient homologs in unicellular relatives, indicating their fundamental regulatory roles [37, 38, 39]. Marshall Graves [38] detailed interactions between Sox genes, highlighting regulatory complexity aiding multicellularity linked to unicellularity. And some other studies tracing gene evolution and discussing some comparative genomics and synteny-based phylogenomics tools. [40]

Other homeobox gene families that could be relevant, and that we use for our analyses are POU [41], CERS-class [42], HNF-class [43], SIX [44], and CUT-class [45].

By examining and analyzing these genes and other crucial gene families in various unicellular relatives like choanoflagellates, filastereans, corallochytreans, and ichthyosporeans, we aim to uncover the genetic basis of the transition from unicellularity to multicellularity.

## 1.1. Objectives

The objective of this work is to generate new data and gain knowledge that will illuminate the evolution of transcription factors and homeobox gene families, and drive progress in the larger project investigating the origin of animal regulation. The main focus will be to study the evolutionary trajectory of transcriptional regulation and metabolism genes of unicellular relatives by taking advantage of the wealth of genomes available from these taxa in the lab, while also trying to generate an available pipeline with bioinformatics tools for the analysis of this data. We will focus on a few gene families from the homeobox group, for which there is no comparative genomics work at this level, and we hope to gain as much new information as possible which will contribute to a detailed understanding of evolution, genomics and gene regulation in the animals' single-celled relatives. This will bring us closer to answering the question of how the change from unicellularity to multicellularity happened.

# 2 Materials & methods

### 2.1. Unicellular Organisms of interest

In our study, we focus on a diverse array of unicellular organisms representing various lineages closely related to animals (Figure 2), although we also include various taxa of the most basal animal lineages, like ctenophores, cnidarians, poriferans and

placozoans, to gather additional information for comparative purposes. [47, 48] These organisms are key to understanding the evolutionary transition from unicellularity to multicellularity and provide valuable insights into the ancestral characteristics of early animal lineages. Specifically, we are working with choanoflagellates, filastereans, corallochytreans, and ichthyosporeans, which encompass a broad spectrum of unicellular eukaryotes.

### 2.2. Comparative genomics

We conducted our study on a selected group of gene families using a comprehensive proteome database encompassing 169 different proteomes from various taxa across Holozoa (see Table S1 in supplementary material), including some fungi that serve as an outgroup. Our primary focus is on the unicellular holozoans, our taxa of interest. To be able to do that, we wrote custom scripts (see link to GitHub for further information), in order to conduct comparative genomics analyses and further comprehend these gene families. See Figure 3 for the complete pipeline.

**Sequence search and phylogenetic reconstruction**

- Step 1: Input Acquisition and sequence deduplication

The first step is to obtain a relevant query file for the gene family under study. This file should represent key taxa families to ensure a comprehensive analysis. This involves querying public databases such as NCBI [49], InterPro [50], Uniprot [51] Pfam [52] or ENSEMBL [53], to obtain filtered sequences for the gene of interest across different species which can be relevant. Then the pipeline begins by requesting the user to provide the gene name and the name of the input sequences file obtained from the public database in FASTA format [54]. This step ensures specificity and relevance to the research question, laying a solid foundation for subsequent analyses. After acquiring the input, the script removes duplicate sequences from the dataset fetched from the public databses. Duplicate sequences can skew analyses and increase computational costs. Utilizing SeqKit [55], a versatile tool for sequence manipulation, the pipeline identifies and eliminates duplicate sequences.

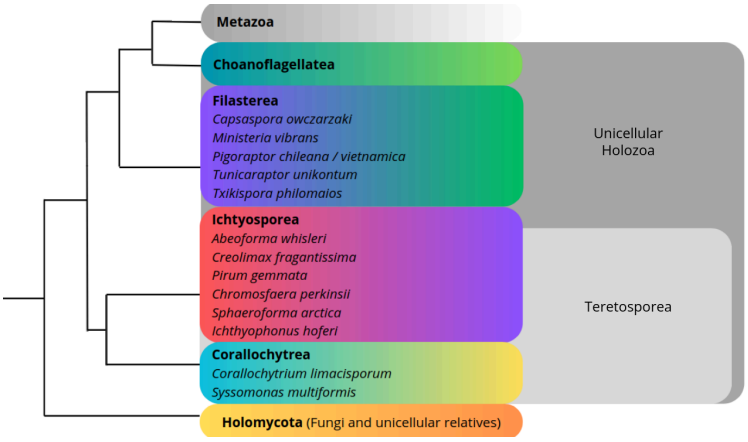


***Figure 2**:* Schematic phylogenetic tree depicting the evolutionary relationships of the four unicellular holozoan lineages with the Metazoa and Holomycota clades. Phylogenetic relationships are based on [46].

- Step 2: Sequence Alignment and Trimming

Then with a deduplicated dataset, the pipeline performs multiple sequence alignment (MSA) [56], using MAFFT [57], with the following parameters "*--maxiterate 1000 --globalpair --op 2.15 --leavegappyregion --reorder*". Post-alignment, the pipeline employs TrimAl [58], setting the gap threshold to 50%, to refine the alignment. TrimAl automatically identifies and removes poorly aligned regions, gaps, and ambiguously aligned segments, enhancing the alignment's quality and accuracy. This step mitigates noise and improves the reliability of downstream analyses.

- Step 3: HMM Database Construction

Upon obtaining a high-quality alignment, the pipeline constructs a Hidden Markov Model (HMM) [59] database using the HMMER suite [60]. HMMs are probabilistic models used for sequence analysis and classification. Building an HMM database from the aligned query sequences encapsulates conserved features and patterns, facilitating efficient sequence searching and classification.

- Step 4: HMM Search

The pipeline conducts an HMM search against a long collection of 169 predicted proteomes, using the previously constructed HMM database. HMMER suite facilitates this search, identifying protein sequences within the proteomes that exhibit significant similarity to the input gene. The script asks the user to define some parameters, such as the E-value threshold, which determines the cutoff for statistical significance in the results and can be changed depending on what is needed to guide the search to ensure significant matches. This step helps identifying orthologous sequences, which is meant as genes in different species that originated from a common ancestor, across diverse organisms, elucidating evolutionary relationships and functional conservation which helps narrowing down the search

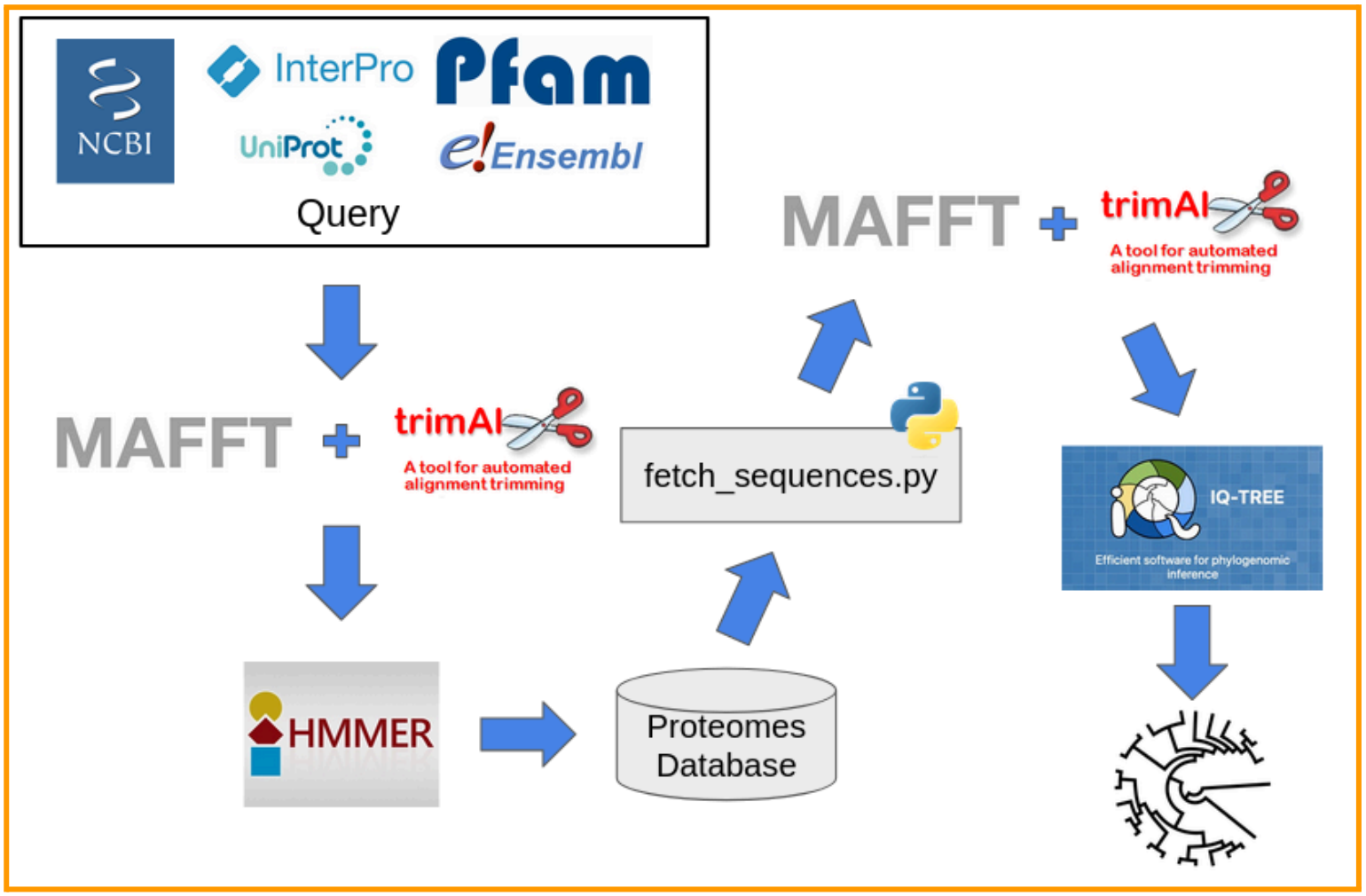


***Figure 3***: Schematic draw of the pipeline used in the comparative genomics analysis from step 1 to step 8. Order followed by arrows and finished with the resulting phylogenetic tree

for the following steps [61-63].

- Step 5: Post-processing of Search Results

Following the HMM search, the pipeline processes the search results to extract relevant information and to organize it for further analysis. This includes parsing output files, generating lists of hits, and cleaning up unnecessary information. Post-processing ensures the integrity and usability of the search results, facilitating subsequent analyses such as sequence retrieval and functional annotation.

- Step 6: Sequence Fetching

With organized search results, the pipeline fetches sequences corresponding to the identified hits from the proteomes. Another custom Python script retrieves the sequences from the original proteomes based on the hit lists generated in the previous step.

Sequence fetching enables us to obtain sequences of interest for further investigation.

- Step 7: Phylogenetic Analysis

After fetching the sequences, the pipeline performs another round of alignment using MAFFT, followed by trimming with TrimAl. Subsequently, a phylogenetic analysis is conducted to infer evolutionary relationships. Using IQ-TREE [64], the pipeline infers a Maximum Likelihood (ML) phylogenetic tree based on the aligned sequences. The parameters used are the following: "*iqtree -s FileAfterTrimming -m TESTMERGE -rcluster 10 -nt AUTO*" in order to let IQ-TREE find the best fit model, and the best number of CPU threads used for the data given. For a deeper analysis and validation of certain gene families, we performed bootstrapping using the following command: "*iqtree -s FileAfterTrimming -m TESTMERGE -rcluster 10 -bnni 10000 -nt AUTO"* . This tree provides insights into the evolutionary relatedness and divergence patterns among the analyzed sequences, contributing to our understanding of evolutionary processes. [65, 66]

- Step 8: Cleanup

The pipeline concludes with cleanup tasks to organize output files and remove temporary files generated during the analysis. This includes organizing output directories, removing redundant files, and ensuring the final results' integrity. Cleanup streamlines data management and facilitates reproducibility, ensuring the analysis remains well-organized and easily accessible for future reference.

The script finishes here (see Figure 3), but there is one step more done in the comparative genomics analysis work.

- Step 9: Motif search

Once the phylogeny is constructed, the next step is to analyze the unicellular organisms of interest depicted in the phylogenetic tree. Beginning by checking if these organisms possess the specific domains of interest, and discard any that do not. After completing this step, we use another custom script to remove unwanted sequences, then generate a new phylogenetic tree with the updated dataset. This process can be repeated as many times as necessary, with each iteration requiring a comparison to the previous results to ensure accuracy and consistency.

**Genomic and transcriptomic analysis**

After identifying which taxa possess the gene of interest in the earlier phase of the study, we proceed with a detailed analysis of the genomes and transcriptomes of the selected species. This step involves examining the genetic and transcriptional landscapes to gain deeper insights into the gene's regulation.

The first step in our detailed analysis involves uploading all the protein sequences that we identified as having the domains of interest into Geneious 7.1 [67]. Each sequence is carefully tagged and aligned to assess the conservation of these domains across different taxa. This alignment process is crucial for understanding the evolutionary stability and functional importance of the domains.

Following this, we upload all the available genomes of the unicellular holozoans into Geneious to map the domains within these genomes using the software's built-in mapping options. However, this approach did not yield satisfactory results due to inefficiencies in handling the large datasets and the complexity of the genomes involved.

To address these challenges, we developed a custom script utilizing Diamond BLAST [68]. This script creates a targeted database containing only the genomes relevant to our study, excluding unnecessary data and streamlining the analysis. Detailed instructions and the script are available on our GitHub repository for further reference.

Next, we select a protein sequence from any of the unicellular organisms where we identified the gene in the previous phase, preferably with a well-conserved domain, to serve as a query for a BLAST search [69] against the newly created genome database, using the following command: "*diamond blastp -d {db_path} -q {query_protein} -o*

*{blast_output} -k 15 --outfmt 6 -p 0*" . This BLAST search, done as part of the script, generates pairwise alignments, which are crucial for pinpointing the exact locations of the domains within the genomes, and gets an output file of only the best 15 hits in a tabular format table easily understandable.

Using the results from the pairwise alignments, we conduct a thorough search within the genomes to identify the presence and structure of the domains. If a sequence matches perfectly without any intervening sequences, it indicates the absence of introns in that region. Conversely, the presence of intervening sequences signifies the presence of introns, which we then would tag for further analysis. [70]

By systematically tagging introns and conserved domains, we can create a comprehensive map of the gene's regulatory elements and structural features across different unicellular holozoan species. This detailed mapping provides invaluable insights into the gene's evolution, function, and regulation, laying the groundwork for subsequent functional and comparative analyses.

After this, the next step of the analysis began by utilizing RNA-seq data that had already been generated for other projects in the lab. This provided a valuable resource for exploring gene expression across various organisms. Leveraging these results, we aimed to determine the expression patterns of specific genes of interest by examining their presence in the transcriptomes of different species, mainly focusing on the unicellular holozoans that were available.

To achieve this, we developed a series of scripts designed to search for the protein sequences corresponding to our target genes within the available transcriptomic data. This approach allowed us to identify whether these genes were actively expressed in the organisms under study, and also see how many copies or isoforms were of each gene. If a gene's protein sequence was detected in a transcriptome, it indicated that the gene was expressed in that particular organism. Conversely, the absence of the protein sequence suggested that the gene was not expressed.

## 3 Results and discussion

### Non-TALE homeobox genes present in unicellulars holozoans

To evaluate the evolutionary history of non-TALE homeobox genes, we examined the presence or absence of various gene families across a proteome database encompassing different organisms from all taxa. Our findings indicate that all the analyzed genes are present in Metazoa, prompting us to investigate whether these genes originated before the last common ancestor of animals or not.

To test these findings (see Figure 4), we inferred phylogenetic trees for all gene families. Notably, the LIM gene family was identified in Filasterea, Ichthyosporea and Corallochytrea. More specifically, the LIM + homeodomain was found in *Corallochytrium limacisporum (Clim),* both Indian and Hawaiian strains*, Syssomonas multiformis (Smult), Abeoforma whisleri (Awhis), Amoebidium parasiticum (Apara), Chromosfaera perkinsii (Cperk), Ichthyophonus hoferi (Ihofe), Pirum gemmata (Pgemm), Capsaspora owczarzaki (Cowcz), Pigoraptor chileana (Pchil), Pigoraptor vietnamica (Pviet), and Txikispora philomaios (Tphil).* In the case of the POU gene, it was detected in the choanoflagellate *Mylnosiga fluctuans (Mfluc)*, and in the corallochytrea *Corallochytrium limacisporum (Clim)*, both Indian and Hawaiian strains. Then, for Sox, we could find presence of the gene family in all unicellular holozoans, except for *Awhis* and *Pgemm*. (see Table S2 in supplementary material for a full dataset of the results). In the case


***Figure 4***: Presence of gene families across taxa. Each cell in the table represents the presence (denoted by a green square) of specific gene families within different taxa.

of Parahox genes and Hox genes we didn't find any presence of these families in any of the unicellular holozoans and therefore conclude that they are animal-specific. Then CERS-class gene family was found in the choanoflagellates *Codosiga hollandica (Choll), Didymoeca costata (Dcost), Mylnosiga fluctuans (Mfluc), Salpingoeca helianthica (Sheli), Salpingoeca macrocollata (Smacr), and Salpingoeca punica (Spuni)*. Also, in the filastereans *Capsaspora owczarzaki (Cowcz), Ministeria vibrans (Mvibr), Txikispora philomaios (Tphil)*, and the ichthyosporeans *Abeoforma whisleri (Awhis), Amoebidium parasiticum (Apara),* and *Creolimax fragrantissima (Cfrag)*. HNF-class and CUT-class gene families are not present either in any of the unicellular holozoans studied and therefore are probably animal-specific (Figure 4). The SIX gene from the SINE-class gene family is present in the choanoflagellate *Choanoeca flexa (Cflex)*, and in the ichthyosporeans *Creolimax fragrantissima (Cfrag), Chromosfaera perkinsii (Cperk), Pirum gemmata (Pgemm),* and *Sphaeroforma arctica (Sarct)*. And finally, the ALX gene from the PRD-class gene family is present in the corallochytrea *Syssomonas multiformis (Smult)*, and in the ichthyosporeans *Amoebidium parasiticum (Apara), Chromosfaera perkinsii (Cperk), Ichthyophonus hoferi (Ihofe)*, and *Pirum gemmata (Pgemm)*.

**Phylogenetic Analysis and Evolution of LIM Gene family**

LIM homeobox had been classified as an animal-specific non-TALE gene family [36], but was later found in unicellular holozoans [22]. We did a complete analysis of this particular gene family, in order to create a complete phylogeny and also a genomic analysis to get closer to know its genomic regulation and evolutionary history.

As previously mentioned, the LIM+homeodomain is present in Filasterea, Ichthyosporea, and Corallochytrea (see Figure 5). Within this context, we observe a distinct clade that includes all unicellular holozoans possessing the LIM-associated homeobox gene. This clade can be further divided based on the structural composition of the LIM and homeodomain regions. Specifically, we can categorize them into two distinct groups:

The first group comprises *Chromosphaera perkinsii (Cperk), Abeoforma whisleri (Awhis), Pirum gemmata (Pgemm), Txikispora philomaios (Tphil)*, and *Capsaspora owczarzaki (Cowcz)*. These organisms each have a single LIM domain in conjunction with a homeodomain (Figure 5).

In contrast, the second group (Figure 5) includes *Amoebidium parasiticum (Apara), Ichthyophonus hoferi (Ihofe), Corallochytrium limacisporum* from both Hawaiian (*ClimH*) and Indian (*ClimI*) strains, *Syssomonas multiformis (Smult), Pigoraptor chileana (Pchil)*, and *Pigoraptor vietnamica (Pviet*). These species are characterized by the presence of two LIM domains along with a homeodomain.

Additionally, we identified another clade of unicellular holozoans comprising protein sequences that were not included in the first analysis. This exclusion was due to these sequences possessing only one LIM domain without the accompanying homeodomain, indicating a structural divergence that precluded their inclusion in the main phylogenetic assessment.

Subsequently, we attempted to classify other copies or subfamilies of LIM homeodomains with the help of our phylogenetic analysis. These classifications allowed us to confidently determine which LIM genes are not present in unicellular holozoans. These subfamilies include *LIM Lhx3/4, LIM Lhx1/5, LIM Lhx6/8, LIM homeobox Awh,* and *LIM Lhx2/9*. Our findings indicate that these specific LIM subfamilies are exclusively found in Metazoa, suggesting that their emergence or evolution occurred after the last common ancestor of animals.

These results provide significant insights into the evolutionary trajectory of the LIM+homeodomain gene family. The distinction between unicellular and multicellular organisms based on the presence and structural composition of LIM domains underscores the complexity and diversification of gene families in the context of multicellular evolution. Importantly, even though most of these LIM subfamilies evolved exclusively in Metazoa after the divergence of unicellular and multicellular lineages, the presence of LIM+homeodomain genes in unicellular holozoans prior to this divergence is notable. This

finding emphasizes that the foundational elements of these genes existed before the evolution of multicellular organisms, playing a significant role in the early evolutionary stages. The subsequent evolution of these genes in Metazoa highlights their pivotal role in the development and complexity of multicellular organisms.

This evolutionary distinction underscores the importance of these genes in contributing to the intricate processes that underlie multicellular development. The presence of foundational LIM+homeodomain genes in unicellular ancestors suggests that these genes provided a pre-existing framework upon which further complexity could be built as organisms transitioned to multicellularity.

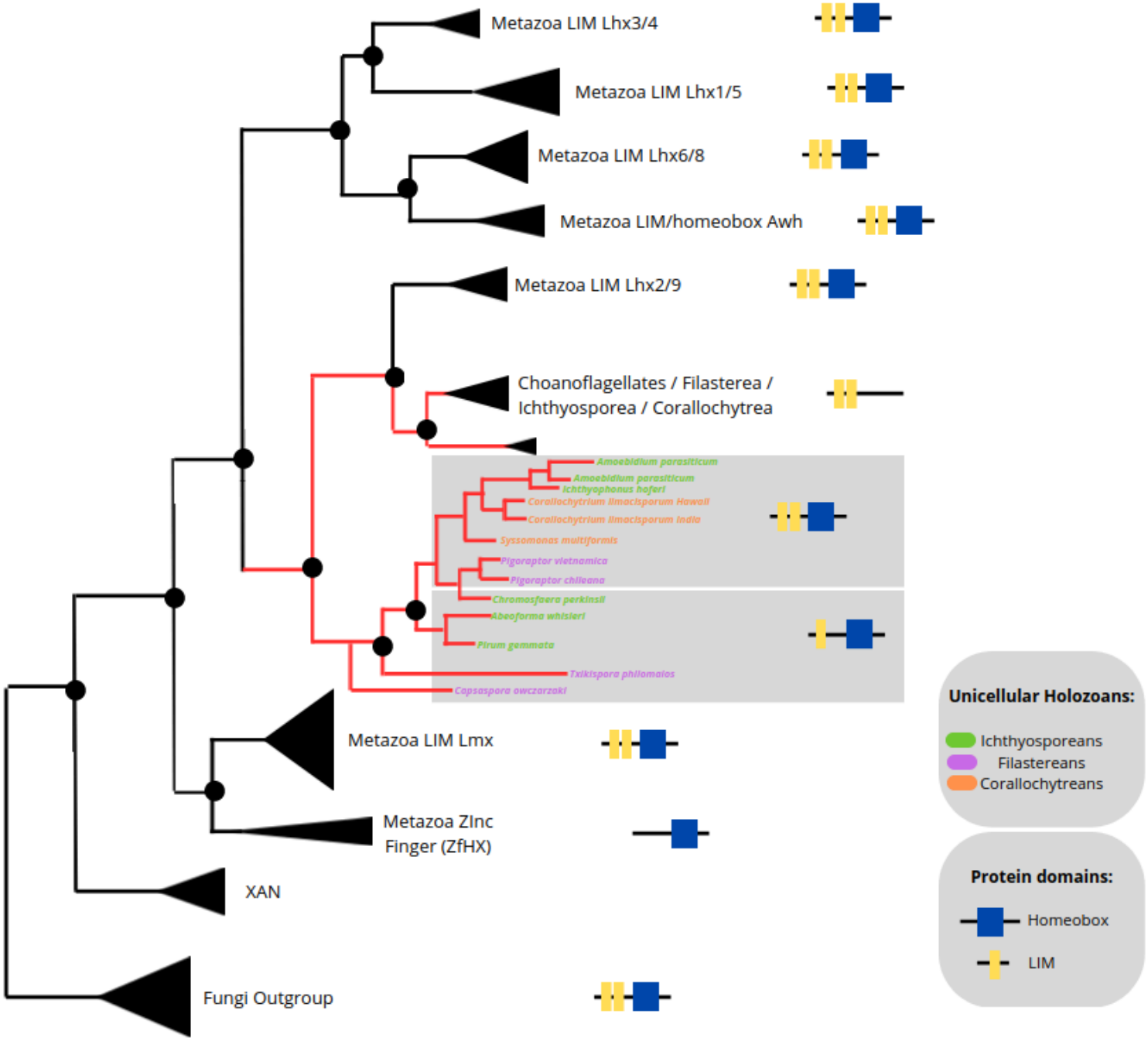


***Figure 5:*** Phylogenetic tree of LIM Homeobox Transcription Factors. Unicellular Holozoa LIM and all copies that exist in Metazoa. Mapped protein domain architectures in blue and yellow for all clades. Supported by Bootstrap.

**Functional Roles of LIM in Metazoa**

In Metazoa, LIM homeobox genes are known to interact with a variety of other proteins, contributing to their functional diversity. It is plausible that the absence of detectable LIM genes in choanoflagellates could be linked to the absence of these interacting partners. If the interacting proteins are not present or have significantly different forms in choanoflagellates, the LIM gene might not be functional or necessary. The variability and evolution of these interacting partners could also explain why LIM genes are not detected; the absence of co-evolved partners could lead to the gene's loss or functional replacement by other proteins. LIM homeobox genes play critical roles in the development and differentiation of tissues in Metazoa, including functions in the nervous system, muscle development, and organogenesis.

Additionally, we found no introns in the genomes of unicellular organisms. This absence of introns has significant implications for gene regulation and expression. Introns are known to play roles in the regulation of gene expression, alternative splicing, and the generation of protein diversity in multicellular organisms. Their absence in unicellular genomes suggests a simpler gene structure and potentially a more straightforward gene expression mechanism, and given that they have more compacted genomes it is not surprising that there are no introns.

**Loss of LIM in choanoflagellates**

Our analysis indicates an intriguing pattern: the apparent absence of LIM homeobox genes in choanoflagellates,which is the closest sister group to animals (see Figure 1), despite their presence in other unicellular holozoans and Metazoa. This observation prompts several hypotheses regarding the evolutionary and functional dynamics of LIM genes in choanoflagellates and their potential interactions with other proteins in Metazoa.

The seeming absence of LIM genes in choanoflagellates could be attributed to various evolutionary scenarios. One possibility is that LIM genes were indeed present in the common ancestor of choanoflagellates and other holozoans but were subsequently lost in the choanoflagellate lineage. Gene loss is a common evolutionary event and could occur due to redundancy or the acquisition of alternative regulatory mechanisms that render the gene unnecessary [71, 72]. Another consideration is that LIM genes might exist in choanoflagellates but have diverged significantly, making them difficult to detect using current genomic tools. This could result from extensive sequence variation or alterations in gene structure that obscure their identification. [73]

Another hypothesis is that choanoflagellates may possess a divergent form of the LIM genes that is not recognizable through conventional sequence homology searches. These genes might have undergone significant modifications, resulting in functional but structurally distinct versions. Such divergence could be driven by unique selective pressures in choanoflagellates, leading to the evolution of a specialized version of the LIM gene family adapted to their specific cellular and ecological contexts, but this possibility needs further investigation.

The complexity and multicellularity of Metazoa likely necessitate such regulatory genes to coordinate the intricate processes of development. In choanoflagellates, which are unicellular or form simple colonies, the regulatory demands are significantly lower. Consequently, the specific functions of LIM genes in Metazoa might not be required in choanoflagellates, leading to the gene's absence. Alternatively, these regulatory roles could be fulfilled by different proteins in choanoflagellates, reflecting an evolutionary divergence in gene function. This is the same case, for other unicellular groups, but LIM is present, so choanoflagellates might differ from them.

**Evolutionary Insights into the Sox Gene Family in Unicellular Holozoans**

Previous studies have suggested that the Sox gene family is animal-specific, with no known presence in unicellular holozoans [74-75]. However, our comprehensive analysis challenges this view. By examining our proteome database, we identified the Sox domain together with the HMG-box domain across all unicellular holozoan taxa. This finding

confirms that the Sox gene family existed prior to the LCA of animals. The detailed results (see Figure 6) show the presence of the Sox gene in all unicellular holozoans, except for *Abeoforma whisleri (Awhis)* and *Pirum gemmata (Pgemm)*.

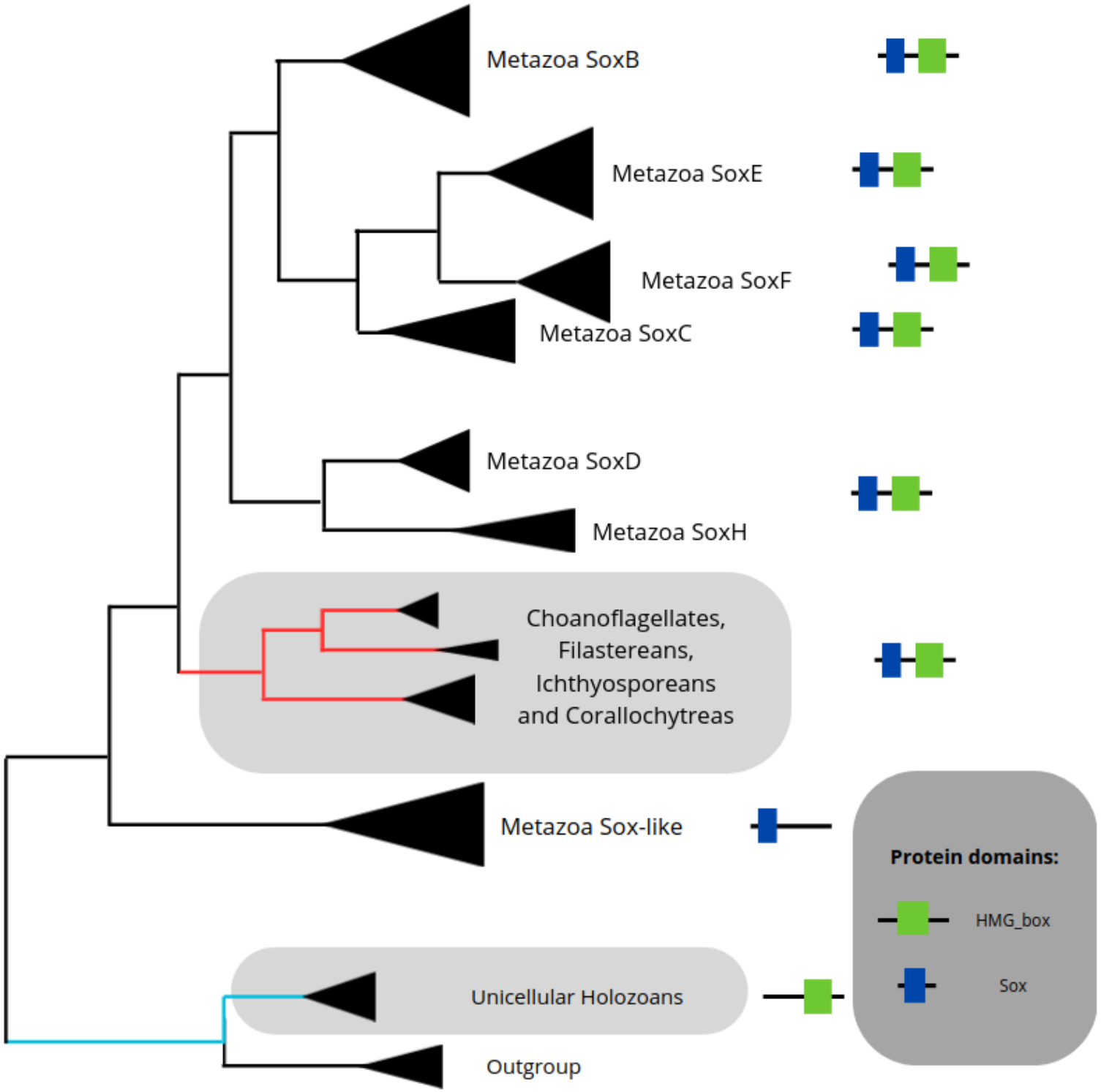


***Figure 6***: Reduced and schematic phylogenetic tree of unicellular Sox and the different copies of the Sox gene family in metazoa, with mapped protein domain architectures.

The phylogenetic analysis (Figure 6) reveals several critical insights into the evolutionary history of the Sox gene family. Initially, a clade of Sox-like genes that lacked the HMG-box domain was observed, indicating an early divergence or functional differentiation before the appearance of true Sox genes. Following this clade, the unicellular holozoans appear, suggesting that the presence of Sox genes in these organisms predates the emergence of animals. Within Metazoa, the phylogeny shows the emergence of multiple clades representing distinct Sox subfamilies. These include SoxH, SoxD, SoxC, SoxF, SoxE, and SoxB.

Each of these subfamilies contains further subdivisions with specific gene copies, reflecting the diversification and specialization of Sox genes in different lineages. [76-79] The presence of multiple Sox subfamilies and their respective copies within Metazoa suggests a significant expansion and functional diversification following the emergence of multicellular organisms. This expansion likely facilitated the complex regulatory networks necessary for the development and differentiation of various tissues and organs in multicellular animals. The Sox genes in unicellular holozoans, possessing both the Sox domain and HMG-box domain, indicate that the foundational elements of these genes were already in place before the transition to multicellularity. This early presence highlights the potential roles that these genes could have played in cellular regulation and interaction even in unicellular contexts. The subsequent diversification in Metazoa underscores the importance of these genes in the evolution of complex multicellular structures and crucial roles in development, stem cell maintenance, sex determination, nervous system function, muscle development and immune response.

**Summary of non-TALE Homeobox Gene Evolution Based on Phylogenetic analysis**

Our phylogenetic analysis summary (Figure 7) reveals critical insights into the evolutionary trajectory of various non-TALE homeobox gene families across Holozoa. This integrated summary and discussion highlight the evolutionary origins, lineage-specific retention, and potential functional implications of these genes.

**POU Gene Family**

POU domain proteins are TFs that contribute to a wide range of biological processes essential for animal development, tissue function, and physiological regulation. They appear to have originated before the last common ancestor of Holozoa, evidenced by their presence in all taxa except Filasterea. This suggests that POU genes were likely lost in the filasterean lineage. The consistent presence of POU genes across other taxa underscores their essential role in early Holozoan evolution. The loss in Filasterea could indicate a lineage-specific adaptation or redundancy reduction, potentially compensated by other regulatory mechanisms. (for more information see Figure S2 in supplementary material).

**Parahox / Hox gene family**

Our analysis indicates that Parahox and Hox gene families emerged after the last common ancestor of unicellular Holozoa but is present in all Metazoan taxa. This pattern suggests that Parahox genes were instrumental in the evolution of animals, contributing to the complexity and diversity observed in Metazoan body plans. Their absence in unicellular lineages underscores their specific association with multicellularity and the regulatory demands of more complex organisms (see Figure S3).

In animals, Hox and Parahox genes play crucial roles in the regulation of development by controlling the identity and differentiation of various body regions and structures along the anterior-posterior axis. These genes are responsible for specifying the formation of limbs, organs, and other tissues in a spatially and temporally coordinated manner. The precise regulation of gene expression by Hox and Parahox genes enables the development of complex body plans and specialized functions that are characteristic of multicellular organisms. In contrast, unicellular organisms do not require such intricate regulatory systems, as their single-celled structure

does not necessitate the same level of spatial and developmental control. Therefore, the sophisticated functions provided by Hox and Parahox genes are not essential for the life processes of unicellular entities.

**CERS-Class Gene Family**

The CERS-class gene family is present from the early stages of Holozoan evolution but is notably absent or altered in Corallochytrea. This loss or modification in Corallochytrea suggests that the CERS-class genes either underwent significant evolutionary changes or were replaced by other functionally equivalent genes. The retention of these genes in other lineages points to their fundamental role in early cellular processes that are fundamental to cellular physiology and organismal development, important since before the last common Holozoa ancestor (see Figure S4).

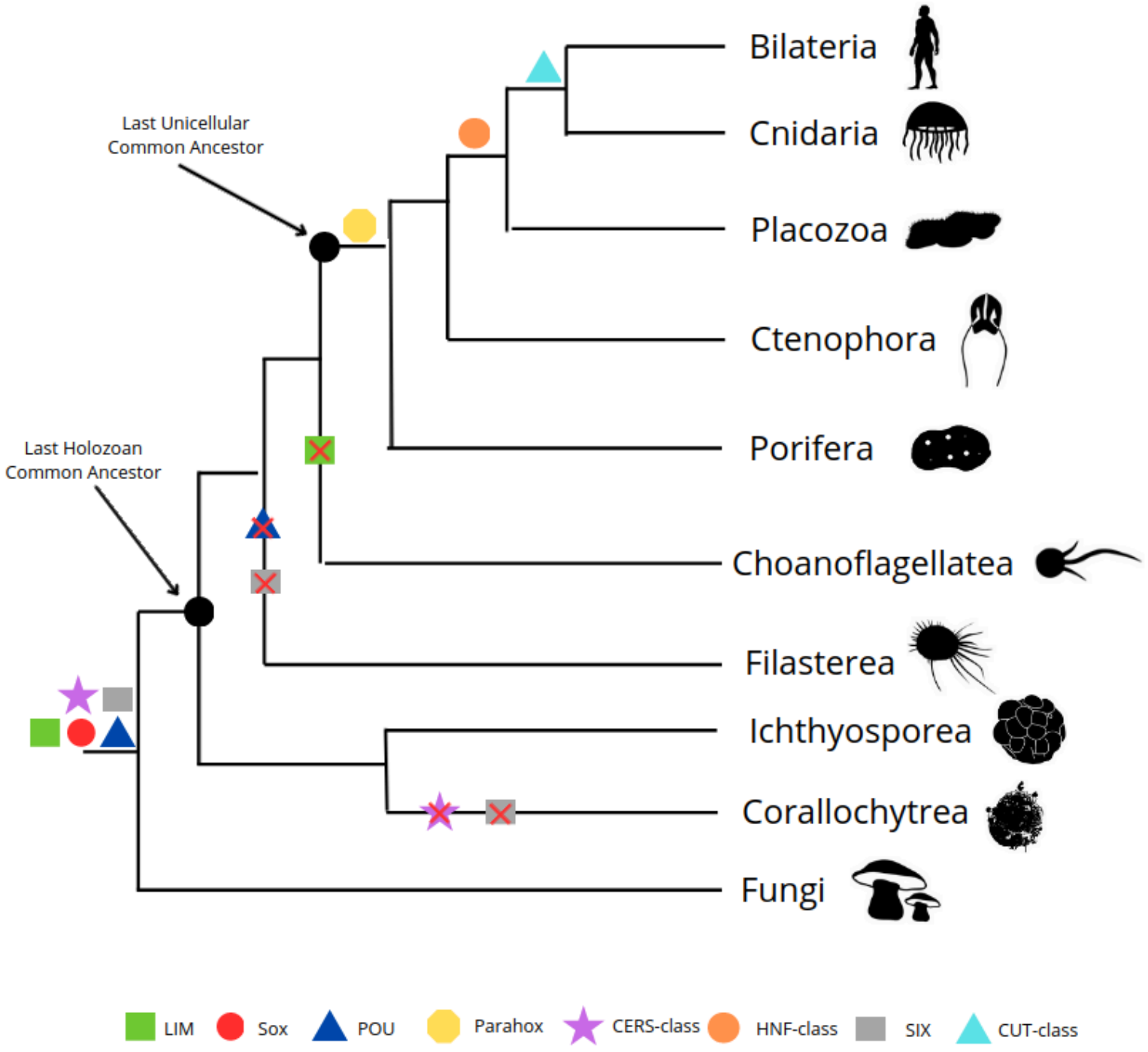

***Figure 7***: Schematic phylogenetic tree with all taxa. Colored shapes show when different genes first appeared during evolution. Red crosses over the shapes show possible gene loss.

**HNF-Class Gene Family**

The HNF-class gene family is exclusive to Metazoa, with presence limited to Placozoa, Cnidaria, and Bilateria (see Figure S5). This restricted distribution implies that HNF genes are linked to specific multicellular functions absent in early lineages of Metazoa, such as sponges or ctenophores. Their presence in Placozoa, Cnidaria, and Bilateria highlights their crucial role in regulatory networks governing the development and differentiation of these more complex organisms.

**SINE-Class Gene Family**

The SIX gene from the SINE-class gene family predates the LUCA of Holozoa but appears to be lost in Corallochytrea and Filasterea (see Figure S6). The absence of SIX genes in these lineages might indicate either a gene loss or significant functional divergence. The presence of SIX genes in other lineages suggests their importance in early Holozoan regulatory networks, and their loss could reflect unique evolutionary pressures in Corallochytrea and Filasterea.

**CUT-Class Gene Family**

The CUT-class gene family is the gene that most recently appeared, being exclusive to the clade comprising Cnidaria and Bilateria (see Figure S7). This suggests that CUT genes are associated with the increased complexity of these lineages, contributing to sophisticated body plans and regulatory mechanisms. The absence of CUT genes in simpler Metazoan lineages and unicellular Holozoans emphasizes their specialized role in the evolution of complex multicellular organisms.

Interestingly, we found that there are not many copies of this gene even in animals where it is present. This limited copy number could be due to the relatively recent emergence of the CUT-class gene family, which has not had sufficient time to diversify or undergo extensive duplication. As a result, the CUT genes may still be in the early stages of their evolutionary trajectory, gradually acquiring more specialized roles as multicellular complexity continues to evolve.

## 4. Conclusions

Overall, our study underscores the importance of detailed phylogenetic analysis in understanding the evolutionary history of gene families. By identifying specific structural variations and their corresponding phylogenetic clades, we can better comprehend the origins and evolutionary pathways that have led to the present diversity of gene families in both unicellular and multicellular organisms.

Non-TALE homeobox gene families in holozoans reveal diverse evolutionary paths, shedding light on their crucial roles in the transition from unicellularity to multicellularity in animals. The presence of LIM and Sox genes in unicellular holozoans challenges previous assumptions, suggesting early regulatory roles before multicellular complexity. Conversely, Parahox and Hox genes, absent in unicellular lineages but ubiquitous in Metazoa, highlight their emergence with multicellularity, crucial for body plan development. These insights deepen our understanding of genetic foundations shaping animal evolution and underscore the intricate interplay between regulatory genes and organismal complexity across holozoans.

Furthermore, this study has brought us closer to having a comprehensive pipeline for the analysis of gene family evolution. By enhancing bioinformatic tools and automating the analytical process, we have made it more accessible and efficient for researchers to conduct similar studies. These improvements in the pipeline facilitate faster and more accurate phylogenetic analyses, promoting further discoveries in evolutionary biology and advancing our understanding of the molecular mechanisms driving the diversity of life.

## 5. Future work

Future research directions could involve extending the current analysis to encompass other non-TALE homeobox genes, delving deeper into their evolutionary histories and functional roles across holozoans. Each gene family presents unique evolutionary patterns that could provide further insights into the genetic basis of multicellularity in animals. Additionally, exploring TALE homeobox genes would complement this study, offering a comprehensive comparison of different homeobox gene families.

Furthermore, there is potential for enhanced investigations through deeper transcriptomic and genomic analyses. These approaches could uncover additional layers of gene regulation and evolutionary dynamics, potentially including dn/ds or positive selection analyses to elucidate adaptive processes within these gene families.

Moreover, streamlining the analytical pipeline into a fully automated framework would significantly enhance accessibility and efficiency for researchers. As a bioinformatician, automating the entire process, not just individual components, would simplify complex analyses and accelerate discoveries in evolutionary biology and molecular genetics. This automation would facilitate broader utilization of these methods, promoting collaborative research efforts and advancing our understanding of gene family evolution across diverse biological contexts.

## Acknowledgements

First of all I would like to express my appreciation to my research supervisor, Marta Álvarez-Presas, for all the patience, the exceptional help and the infinite support she has given me throughout the duration of the project. I would also like to extend my gratitude to the entire group of people from the MulticellGenome Lab, for welcoming me with open arms since day one, and letting me work alongside them which has made me learn immeasurable things and it's something I will never forget. Finally, I would like to also thank my friends and family for their support and encouragement to me for years.

## References

1. Ros-Rocher, N., Pérez-Posada, A., Leger, M. M., & Ruiz-Trillo, I. (2021). The origin of animals: an ancestral reconstruction of the unicellular-to-multicellular transition. *Open biology*, *11*(2), 200359. https://doi.org/10.1098/rsob.200359
2. Brunet, T., & King, N. (2017). The Origin of Animal Multicellularity and Cell Differentiation. *Developmental cell*, *43*(2), 124–140. https://doi.org/10.1016/j.devcel.2017.09.016
3. Heaton, L.L.M., Jones, N.S. & Fricker, M.D. A mechanistic explanation of the transition to simple multicellularity in fungi. Nat Commun 11, 2594 (2020). https://doi.org/10.1038/s41467-020-16072-4
4. Karl J Niklas, Stuart A Newman, The many roads to and from multicellularity, *Journal of Experimental Botany*, Volume 71, Issue 11, 11 June 2020, Pages 3247–3253, https://doi.org/10.1093/jxb/erz547
5. Umen J. G. (2014). Green algae and the origins of multicellularity in the plant kingdom. *Cold Spring Harbor perspectives in biology*, *6*(11), a016170. https://doi.org/10.1101/cshperspect.a016170
6. Ruiz-Trillo, I., & de Mendoza, A. (2020). Towards understanding the origin of animal development. *Development (Cambridge, England)*, *147*(23), dev192575. https://doi.org/10.1242/dev.192575
7. Rokas A. (2008). The molecular origins of multicellular transitions. *Current opinion in genetics & development*, *18*(6), 472–478. https://doi.org/10.1016/j.gde.2008.09.004
8. Unicellular Relatives of Animals (Aleksandra Kożyczkowska , Iñaki Ruiz-Trillo and Elena Casacuberta)
9. Cavalier-Smith T. (2017). Origin of animal multicellularity: precursors, causes, consequences-the choanoflagellate/sponge transition, neurogenesis and the Cambrian explosion. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *372*(1713), 20150476. https://doi.org/10.1098/rstb.2015.0476
10. Hoffmeyer, T. T., & Burkhardt, P. (2016). Choanoflagellate models - Monosiga brevicollis and Salpingoeca rosetta. *Current opinion in genetics & development*, *39*, 42–47. https://doi.org/10.1016/j.gde.2016.05.016
11. Suga, H., & Ruiz-Trillo, I. (2013). Development of ichthyosporeans sheds light on the origin of

metazoan multicellularity. *Developmental biology*, *377*(1), 284–292. https://doi.org/10.1016/j.ydbio.2013.01.009
12. de Mendoza, A., Suga, H., Permanyer, J., Irimia, M., & Ruiz-Trillo, I. (2015). Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife*, *4*, e08904. https://doi.org/10.7554/eLife.08904
13. Kożyczkowska, A., Najle, S. R., Ocaña-Pallarès, E., Aresté, C., Shabardina, V., Ara, P. S., Ruiz-Trillo, I., & Casacuberta, E. (2021). Stable transfection in protist Corallochytrium limacisporum identifies novel cellular features among unicellular animals relatives. *Current biology : CB*, *31*(18), 4104–4110.e5. https://doi.org/10.1016/j.cub.2021.06.061
14. Sebé-Pedrós, A., Degnan, B. M., & Ruiz-Trillo, I. (2017). The origin of Metazoa: a unicellular perspective. *Nature reviews. Genetics*, *18*(8), 498–512. https://doi.org/10.1038/nrg.2017.21
15. Bobola, N., & Sagerström, C. G. (2024). TALE transcription factors: Cofactors no more. *Seminars in cell & developmental biology*, *152-153*, 76–84. https://doi.org/10.1016/j.semcdb.2022.11.015
16. Merabet, S., & Galliot, B. (2015). The TALE face of Hox proteins in animal evolution. Frontiers in genetics, 6, 267. https://doi.org/10.3389/fgene.2015.00267
17. Ferrier, David. (2016). Evolution of Homeobox Gene Clusters in Animals: The Giga-Cluster and Primary vs. Secondary Clustering. Frontiers in Ecology and Evolution. 4. 10.3389/fevo.2016.00036.
18. *Mishra, H., & Saran, S. (2015). Classification and expression analyses of homeobox genes from Dictyostelium discoideum. Journal of biosciences, 40(2), 241–255. https://doi.org/10.1007/s12038-015-9519-3*
*19.* The TALE face of Hox proteins in animal evolution (Samir Merabet, and Brigitte Gallio*)*
20. King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K. J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., … Rokhsar, D. (2008). The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature*, *451*(7180), 783–788. https://doi.org/10.1038/nature06617
21. Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, J. J., Sabidó, E., Gómez-Skarmeta, J. L., Di Croce, L., & Ruiz-Trillo, I. (2016). The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. *Cell*, *165*(5), 1224–1237. https://doi.org/10.1016/j.cell.2016.03.034
22. Grau-Bové, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T. A., & Ruiz-Trillo, I. (2017). Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife*, *6*, e26036. https://doi.org/10.7554/eLife.26036
23. de Mendoza, A., Sebé-Pedrós, A., Šestak, M. S., Matejcic, M., Torruella, G., Domazet-Loso, T., & Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), E4858–E4866. https://doi.org/10.1073/pnas.1311818110
24. Ferrier, D. E., & Minguillón, C. (2003). Evolution of the Hox/ParaHox gene clusters. *The International journal of developmental biology*, *47*(7-8), 605–611.
25. Joo, S., Wang, M. H., Lui, G., Lee, J., Barnas, A., Kim, E., Sudek, S., Worden, A. Z., & Lee, J. H. (2018). Common ancestry of heterodimerizing TALE homeobox transcription factors across Metazoa and Archaeplastida. *BMC biology*, *16*(1), 136. https://doi.org/10.1186/s12915-018-0605-5
26. de Mendoza, A., & Sebé-Pedrós, A. (2019). Origin and evolution of eukaryotic transcription factors. *Current opinion in genetics & development*, *58-59*, 25–32. https://doi.org/10.1016/j.gde.2019.07.010
27. Lambert, S. A., Yang, A. W. H., Sasse, A., Cowley, G., Albu, M., Caddick, M. X., Morris, Q. D., Weirauch, M. T., & Hughes, T. R. (2019). Similarity regression predicts evolution of transcription factor sequence specificity. *Nature genetics*, *51*(6), 981–989. https://doi.org/10.1038/s41588-019-0411-1
28. Bobola, N., & Sagerström, C. G. (2024). TALE transcription factors: Cofactors no more. Seminars in cell & developmental biology, 152-153, 76–84. https://doi.org/10.1016/j.semcdb.2022.11.015
29. de Mendoza, A., & Ruiz-Trillo, I. (2011). The mysterious evolutionary origin for the GNE gene and the root of bilateria. Molecular biology and evolution, 28(11), 2987–2991. https://doi.org/10.1093/molbev/msr142
30. Velyvis A, Qin J. LIM Domain and Its Binding to Target Proteins. In: Madame Curie Bioscience

Database [Internet]. Austin (TX): Landes Bioscience; 2000-2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6372/

31. Morino, Y., Hashimoto, N., & Wada, H. (2017). Expansion of TALE homeobox genes and the evolution of spiralian development. *Nature ecology & evolution*, *1*(12), 1942–1949. https://doi.org/10.1038/s41559-017-0351-z
32. Holland P. W. (2013). Evolution of homeobox genes. *Wiley interdisciplinary reviews. Developmental biology*, *2*(1), 31–45. https://doi.org/10.1002/wdev.78
33. Bürglin, T. (2001). Homeobox. In Encyclopedia of Genetics.
34. Holland, P.W., Booth, H.A.F. & Bruford, E.A. Classification and nomenclature of all human homeobox genes . *BMC Biol* **5**, 47 (2007). https://doi.org/10.1186/1741-7007-5-47
35. Mendivil-Ramos, Olivia & Barker, Daniel & Ferrier, David. (2012). Ghost Loci Imply Hox and ParaHox Existence in the Last Common Ancestor of Animals. Current biology : CB. 22. 1951-6. 10.1016/j.cub.2012.08.023.
36. Srivastava M, Larroux C, Lu DR, Mohanty K, Chapman J, Degnan BM, Rokhsar DS. Early evolution of the LIM homeobox gene family. BMC Biol. 2010 Jan 18;8:4. doi: 10.1186/1741-7007-8-4. PMID: 20082688; PMCID: PMC2828406.
37. Arumugam, Anitha & Senthilkumaran, Balasubramanian. (2021). Role of sox family genes in teleostean reproduction-an overview. Reproduction and Breeding. 1. 22-31. 10.1016/j.repbre.2021.02.004.
38. Graves J. A. (1998). Interactions between SRY and SOX genes in mammalian sex determination. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *20*(3), 264–269. https://doi.org/10.1002/
39. Jiang, L., Bi, D., Ding, H., Wu, X., Zhu, R., Zeng, J., Yang, X., & Kan, X. (2019). Systematic Identification and Evolution Analysis of *Sox* Genes in *Coturnix japonica* Based on Comparative Genomics. *Genes*, *10*(4), 314. https://doi.org/10.3390/genes10040314
40. Steenwyk, J.; King, N. From Genes to Genomes: Opportunities, Challenges, and a Roadmap for Synteny-based Phylogenomics. Preprints 2023, 2023090495.https://doi.org/10.20944/preprints202309.0495.v2
41. Gold DA, Gates RD, Jacobs DK. The early expansion and evolutionary dynamics of POU class genes. Mol Biol Evol. 2014 Dec;31(12):3136-47. doi: 10.1093/molbev/msu243. Epub 2014 Sep 25. PMID: 25261405; PMCID: PMC4245813.
42. Zhang, M., Li, Z., Liu, Y., Ding, X., Wang, Y., & Fan, S. (2023). The ceramide synthase (CERS/LASS) family: Functions involved in cancer progression. *Cellular oncology (Dordrecht)*, *46*(4), 825–845. https://doi.org/10.1007/s13402-023-00798-6
43. Shimeld S. M. (1997). Characterisation of amphioxus HNF-3 genes: conserved expression in the notochord and floor plate. *Developmental biology*, *183*(1), 74–85. https://doi.org/10.1006/dbio.1996.8481
44. Meurer, L., Ferdman, L., Belcher, B., & Camarata, T. (2021). The SIX Family of Transcription Factors: Common Themes Integrating Developmental and Cancer Biology. *Frontiers in cell and developmental biology*, *9*, 707854.https://doi.org/10.3389/fcell.2021.707854
45. Bürglin, T. R., & Cassata, G. (2002). Loss and gain of domains during evolution of cut superclass homeobox genes. *The International journal of developmental biology*, *46*(1), 115–123.
46. Ruiz-Trillo, I., Kin, K., & Casacuberta, E. (2023). The Origin of Metazoan Multicellularity: A Potential Microbial Black Swan Event. *Annual review of microbiology*, *77*, 499–516. https://doi.org/10.1146/annurev-micro-032421-120023
47. Medina, M., Collins, A. G., Silberman, J. D., & Sogin, M. L. (2001). Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(17), 9707–9712.https://doi.org/10.1073/pnas.171316998
48. Technau, U., & Steele, R. E. (2011). Evolutionary crossroads in developmental biology: Cnidaria. *Development (Cambridge, England)*, *138*(8), 1447–1458. https://doi.org/10.1242/dev.048959
49. Database resources of the National Center for Biotechnology Information in 2023. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Farrell CM, Feldgarden M, Fine AM, Funk K, Hatcher E, Kannan S, Kelly C, Kim S, Klimke W, Landrum MJ, Lathrop S, Lu Z, Madden TL, Malheiro A, Marchler-Bauer A, Murphy TD, Phan L, Pujar S, Rangwala SH, Schneider VA, Tse T, Wang J, Ye J, Trawick BW, Pruitt KD, Sherry ST. Sayers EW, et al. Nucleic

Acids Res. 2023 Jan 6;51(D1):D29-D38. doi: 10.1093/nar/gkac1032. Nucleic Acids Res. 2023. PMID: 36370100 Free PMC article.

50. Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, Julian Gough, Daniel H Haft, Ivica Letunić, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Christine A Orengo, Arun P Pandurangan, Catherine Rivoire, Christian J A Sigrist, Ian Sillitoe, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Cathy H Wu, Alex Bateman, InterPro in 2022, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, Pages D418–D427, https://doi.org/10.1093/nar/gkac993
51. Alex Bateman,Maria-Jesus Martin,Sandra Orchard,Michele Magrane, and et.al. Ahmad. Uniprot: The universal protein knowledgebase in 2023. Nucleic Acids Research, 51(D1): D523–D531, nov 21 2022. [Online; accessed 2023-06-13].
52. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. Nucleic acids research, 49(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913
53. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M. An overview of Ensembl. Genome Res. 2004 May;14(5):925-8. doi: 10.1101/gr.1860604. Epub 2004 Apr 12. PMID: 15078858; PMCID: PMC479121.
54. DavidJ.Lipman and William R. Rapid and sensitive protein similarity searches. Science, 227(4693):1435–1441, mar 22 1985. [Online; accessed 2023-06-13].
55. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One. 2016 Oct 5;11(10):e0163962. doi: 10.1371/journal.pone.0163962. PMID: 27706213; PMCID: PMC5051824.
56. Chao, J., Tang, F., & Xu, L. (2022). Developments in Algorithms for Sequence Alignment: A Review. Biomolecules, 12(4), 546. https://doi.org/10.3390/biom12040546
57. Kazutaka Katoh, Kazuharu Misawa, Kei‐ichi Kuma, Takashi Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Research, Volume 30, Issue 14, 15 July 2002, Pages 3059–3066, https://doi.org/10.1093/nar/gkf436
58. Salvador Capella-Gutiérrez, José M. Silla-Martínez, Toni Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics, Volume 25, Issue 15, August 2009, Pages 1972–1973, https://doi.org/10.1093/bioinformatics/btp348
59. Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2009 Sep;10(6):402-15. doi: 10.2174/138920209789177575. PMID: 20190955; PMCID: PMC2766791.
60. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W29-37. doi: 10.1093/nar/gkr367. Epub 2011 May 18. PMID: 21593126; PMCID: PMC3125773.
61. Pearson W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, *Chapter 3*, 3.1.1–3.1.8. https://doi.org/10.1002/0471250953.bi0301s42
62. Qian, B., & Goldstein, R. A. (2003). Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins*, *52*(3), 446–453. https://doi.org/10.1002/prot.10373
63. Margelevičius, M., Venclovas, Č. Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. BMC Bioinformatics 11, 89 (2010). https://doi.org/10.1186/1471-2105-11-89
64. 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. https://doi.org/10.1093/molbev/msu300
65. Gregory, T.R. Understanding Evolutionary Trees. *Evo Edu Outreach* **1**, 121–137 (2008). https://doi.org/10.1007/s12052-008-0035-x
66. Nagy LG, Merényi Z, Hegedüs B, Bálint B. Novel phylogenetic methods are needed for understanding gene function in the era of

mega-scale genome sequencing. Nucleic Acids Res. 2020 Mar 18;48(5):2209-2219. doi: 10.1093/nar/gkz1241. PMID: 31943056; PMCID: PMC7049691.
67. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012 Jun 15;28(12):1647-9. doi: 10.1093/bioinformatics/bts199. Epub 2012 Apr 27. PMID: 22543367; PMCID: PMC3371832.
68. Buchfink, B., Reuter, K. & Drost, HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021). https://doi.org/10.1038/s41592-021-01101-x
69. McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, *32*(Web Server issue), W20–W25. https://doi.org/10.1093/nar/gkh435
70. Manyuan Long, Michael Deutsch, Intron—exon structures of eukaryotic model organisms, *Nucleic Acids Research*, Volume 27, Issue 15, 1 August 1999, Pages 3219–3228, https://doi.org/10.1093/nar/27.15.3219
71. Jiménez-Marín, B., Rakijas, J.B., Tyagi, A. et al. Gene loss during a transition to multicellularity. Sci Rep 13, 5268 (2023). https://doi.org/10.1038/s41598-023-29742-2
72. Albalat, R., Cañestro, C. Evolution by gene loss. Nat Rev Genet 17, 379–391 (2016). https://doi.org/10.1038/nrg.2016.39
73. Otsuka, Jinya. 2023. "The Evolutionary Theory Along the Phylogenetic Tree of Unicellular Organisms". Annual Research & Review in Biology 38 (9):39-49. https://doi.org/10.9734/arrb/2023/v38i930605.
74. Koopman, P., Schepers, G., Brenner, S., & Venkatesh, B. (2004). Origin and diversity of the SOX transcription factor gene family: genome-wide analysis in Fugu rubripes. Gene, 328, 177–186. https://doi.org/10.1016/j.gene.2003.12.008
75. Jiang L, Bi D, Ding H, Wu X, Zhu R, Zeng J, Yang X, Kan X. Systematic Identification and Evolution Analysis of Sox Genes in Coturnix japonica Based on Comparative Genomics. Genes (Basel). 2019 Apr 22;10(4):314. doi: 10.3390/genes10040314. PMID: 31013663; PMCID: PMC6523956.
76. Sreenivasan, R., Gonen, N., & Sinclair, A. (2022). SOX Genes and Their Role in Disorders of Sex Development. Sexual development : genetics, molecular biology, evolution, endocrinology, embryology, and pathology of sex determination and differentiation, 16(2-3), 80–91. https://doi.org/10.1159/000524453
77. Kim, K., Kim, I. K., Yang, J. M., Lee, E., Koh, B. I., Song, S., Park, J., Lee, S., Choi, C., Kim, J. W., Kubota, Y., Koh, G. Y., & Kim, I. (2016). SoxF Transcription Factors Are Positive Feedback Regulators of VEGF Signaling. *Circulation research*, *119*(7), 839–852. https://doi.org/10.1161/CIRCRESAHA.116.308483
78. Schock, E. N., & LaBonne, C. (2020). Sorting Sox: Diverse Roles for Sox Transcription Factors During Neural Crest and Craniofacial Development. *Frontiers in physiology*, *11*, 606889. https://doi.org/10.3389/fphys.2020.606889
79. Penzo-Méndez AI. Critical roles for SoxC transcription factors in development and cancer. Int J Biochem Cell Biol. 2010 Mar;42(3):425-8. doi: 10.1016/j.biocel.2009.07.018. Epub 2009 Aug 3. PMID: 19651233; PMCID: PMC2862366.