

Unravelling the Origins and Evolution of the Animal Kingdom using Genomics

Cristina Guijarro

A thesis submitted for the degree of Doctor of Philosophy

Department of Biological Sciences

University of Essex

Date of submission January 2020

ABSTRACT

There are ~35 classified phyla/sub-phyla within the Animal Kingdom; some of which have unresolved relationships. The advent of genomics has made it possible to study new aspects of animal evolution, including comparative genomics (e.g., gene loss/gain, non-coding regions, synteny, etc), gene family evolution, and their evolutionary relationships using genome-wide data.

No study to date has compared all the wealth of genomic data available to understand the evolution of the Animal Kingdom. Using a core bioinformatics pipeline and dataset to infer Homology Groups (HGs), the losses and novelties of these HGs were proven integral to the diversification of the animal kingdom. The same core pipeline was used to extract homeobox gene HGs, a key family used to understand origin and diversification in animals. Gene trees were inferred from the core dataset HGs to determine the evolution of a gene family iconic in the study of animal body plans. Conserved animal genes were also mined using the same pipeline and dataset. Animal phylogenomics is one of the most controversial areas in modern evolutionary science. Whilst many new methods have been developed, no study to date has tried to assess the impact of gene age in the reconstruction of evolutionary trees.

The phyla with the largest count of HG losses also had the highest counts of HG novelties. Not all of these were strictly *de novo*, but the numbers suggest a re-manufacturing of the genetic material from the genes reduced to those that were more recently diverged.

A comprehensive classification of all the diversity of animal homeobox genes is lacking. The gene trees showed complex patterns, with similar homeobox expansions between more distant species, and interlapping homeobox families.

The highly conserved HGs recovered, for the animal phylogenies, well-established relationships between some phyla using maximum likelihood and Bayesian inference methods. Ctenophora was consistently recovered as sister to all other animals, and interesting relationships between ecdysozoans and lophotrochozoans. However, it was proven that it takes more than a highly conserved set of genes to infer a stable and correct phylogeny.

Each of the additional methods used to extend the core bioinformatics pipeline revealed a pattern of correlation, particularly among the fast evolving species, such as platyhelminthes,

nematodes and tardigrades. These HG losers, and gainers also had lineage specific homeobox clades, and caused artefactual problems in the phylogenies.

ACKNOWLEDGEMENTS

Firstly I need to thank my patient and hard-working supervisor Dr Jordi Paps-Montserrat. I am sincerely appreciative of the time he has dedicated to improving my writing skills, how I approach problems and the never ending patience he showed, even remotely from across the country. I gleaned not just knowledge but theoretical abilities that will get me through my career in the world of science. I have been so lucky to have his guidance as both a mentor and a friend, the Paps-Lab will always be a part of my life.

I need also to thank the rest of my PhD board members: Dr Antonio Marco and Dr Nicolae Radu Zabet, they provided me with excellent advice, insightful comments, confidence and new perspectives.

I am appreciative of Professor Peter Holland FRS at the University of Oxford for his expertise in publishing papers in this field, his comments and his help in publishing the 2nd chapter of this thesis. Without his guidance, it may have taken another year to publish.

Lastly, the biggest thanks goes to my family, without which I may not have survived the writing of this thesis. I thank my mum Caroline Clarke, for providing me with a roof over my head and encouraging me when I was hit with some hardships. I thank my sisters: Ana Crane and Nieves Guijarro for their mental support and proof reading my chapters for grammatical and typing errors. Most importantly, I thank my daughter Luna, she was born in the second year of my PhD, shortly after my confirmation board. She has made the whole PhD adventure a greater challenge, but she has also made the reward so much greater.

TABLE OF CONTENTS

ABSTRACT.....	2
ACKNOWLEDGEMENTS.....	4
TABLE OF CONTENTS.....	5
LIST OF FIGURES.....	7
LIST OF ABBREVIATIONS AND ACRONYMS.....	14
1 INTRODUCTION.....	15
1.2 COMPARATIVE GENOMICS OF BETTER KNOWN ANIMAL LINEAGES.....	21
1.3 GENE FAMILY EVOLUTION.....	33
1.4 METAZOA PHYLOGENY AND METHODOLOGY.....	42
1.5 THESIS AIMS.....	50
1.6 MOTIVATIONS AND RESEARCH QUESTIONS.....	53
1.7 THE BIOINFORMATICS PIPELINES.....	54
2 USE IT OR LOSE IT: WIDESPREAD PATTERNS OF GENE LOSS IN THE EVOLUTION OF ANIMAL GENOMES.....	55
2.1 SUMMARY.....	55
2.2 BACKGROUND.....	56
2.3 DISCUSSION & RESULTS.....	57
2.4 MATERIAL & METHODS.....	69
3 A NEXUS OF GENE AND MORPHOLOGY: HOMEBOX PROTEIN EVOLUTION IN ANIMALS USING GENE TREES AND AN EXTENSIVE TAXON SAMPLING.....	71
3.1 SUMMARY.....	71
3.2 BACKGROUND.....	73
3.3 RESULTS & DISCUSSION.....	75
3.4 MATERIALS & METHODS.....	103

4 RESAMPLING CORE/CONSERVED WHOLE GENOME LEVEL HOMOMOLOGY GROUPS TO INFER THE PHYLOGENY OF METAZOA.....	105
4.1 SUMMARY.....	105
4.2 BACKGROUND.....	107
4.3 RESULTS & DISCUSSION.....	110
METHODS & MATERIALS.....	128
5 DISCUSSION/CONCLUSIONS.....	131
5.1 DIVERSIFICATION OF SPECIFIC ANIMAL LINEAGES BY LOSS AND GAIN OF HOMOMOLOGY GROUPS	131
5.2 EVOLUTION OF BODY PLANS IN ANIMAL DIVERSIFICATION.....	133
5.3 IMPORTANT ANIMAL RELATIONSHIPS IN THE EMERGENCE OF ANIMAL PHYLA.....	135
5.4 IMPLICATIONS AND RECOMMENDATIONS TO FUTURE RESEARCH.....	136
5.5 FINAL WORDS.....	138
6 REFERENCES.....	139
7 APPENDICES.....	151

LIST OF FIGURES

Figure 1.1 *Phylogeny of the relationships within the Opisthokonta super group dividing into the two key lineages Holomycota and Holozoa in which animals placed in. Phylogeny as described in recent literature (Torruella et al., 2012; Paps et al., 2013).*

Figure 1.2 *Tree depicting the possible relationships of the ~35 Phyla in the Animal Kingdom. Faded taxon labels are some of the least well supported placements. This representation has been compiled from multiple literature sources (Laumer et al., 2015; Giribet, 2016a; Kocot, 2016; Kocot et al., 2017; Laumer et al., 2019).*

Figure 1.3 *Summary of novel and lost gene families in the emergence of Metazoa. Numbers with (+) denote novel gene families and numbers with (-) denote gene family losses. Highlighted are the lineages of interest in this thesis, results from (Paps & Holland, 2018).*

Figure 1.4 *A ANTP-Hox-like cladogram including previously unexplored lophotrochozoan homeobox genes. In mauve are well classified homeobox genes in the animal kingdom, in orange are the lesser classified and understood lophotrochozoan ANTP families, as adapted from (Somorjai et al., 2018).*

Figure 1.5 *Various conflicting topologies as cladograms for the internal relationships in the animal kingdom as summarised in recent research, with circles denoting ancestral nodes. (A) In 2008 this topology was considered very robust with one of the largest taxon samplings to that point (Dunn et al., 2018). (B) A very frequently agreed upon phylogeny, despite many unresolved nodes displayed as polytomies (Giribet, 2016b). (C) A most recent phylogeny to date using all the available animal phyla possible (Laumer et al., 2019). (D) A differing phylogeny produced in the same year as (C) in attempt to resolve the current polytomies in lophotrochozoans seen in (B) (Marlétaz et al., 2019).*

Figure 2.1 *Flow representation of the bioinformatics pipeline and each key step for the analysis/generation of novel and lost HGs.*

Figure 2.2 *Phylogenetic map for each of the animal genomes used in the bioinformatics analysis. Each of the main clades labelled for clarity, these are the clades used to infer HGs. As noted in the materials and methods, this is a consensus phylogeny taken from multiple literature sources.*

Figure 2.3 *BUSCO analysis results using the 303 eukaryote genes dataset. The threshold for each of the animal genomes was to have less than 15% missing genes. Where a single taxon had more than 15% genes missing, it was accepted if sharing an analysed clade with a genome meeting the completeness criterion.*

Figure 2.4 *Receiver operating characteristic (ROC) curve for the HGs in this study matching the eukaryote BUSCO gene sets. ROC produces a graphic that illustrates the diagnostic ability of a classifier system by plotting the true positive rate against the false positive rate. Graphs above the diagonal indicate a high proportion of true positives vs false positives. To quantify the amount of misassignment in our HG, we compared our clustering against the eukaryote and metazoan gene sets of BUSCO. BUSCO sets were mined from OrthoDB, which is based in Best-Reciprocal-Hit (BRH) BLAST. In contrast, our pipeline combines BRH plus Markov Clustering. BUSCO datasets contain single copy orthologs present in at least 90% of the species. The species sampling used to define the BUSCO gene sets are different to ours. The percentage of BUSCO genes misclassified in our pipeline was quantified. For the 303 orthology groups of the BUSCO eukaryotic dataset, the error rate of the assignments is 0.085%. We have an additional 80 eukaryotes that BUSCO do not have, 4 OrthoDB markers diverged beyond sequence similarity recognition, divided into two HGs according to major lineages. This is likely due to our extended dataset.*

Figure 2.5 *ROC curve for the HGs in this study matching the metazoan BUSCO gene sets. Similarly to the eukaryote test, to quantify the amount of misassignment in our HG, we compared our clustering against the metazoan gene sets of BUSCO. For the 978 orthology groups of the metazoans BUSCO, the error rate is 0.25%. We have an additional 38 animals that BUSCO do not have, 17 OrthoDB markers diverged beyond sequence similarity recognition, divided into two HGs according to major lineages. This is likely due to our extended Supplementary Data et with less bias towards vertebrates and a larger selection of non-arthropod protostomes and non-vertebrate deuterostomes.*

Figure 2.6 Reconstruction of ancestral genomic gains and losses in the Animal Kingdom.

Evolutionary relationships of the major groups included in this study (Halanych, 2004; Laumer et al., 2015; Kocot, 2016). Different categories of HG are indicated in each node, from top to bottom, Novel HG (+), Core Novel HG (++), Lost HG (-), and Core Lost HG (--). Organism outlines from phylopic.org and from the author (submitted to phylopic).

Figure 2.7 Levels of gene gains and losses at phylum level. Heatmap normalised by row displaying the amount of gene gains (green at highest numbers, blue at lower numbers) and loss (pink at highest loss, blue at fewer losses) for the animal phyla in this study.

Figure 2.8 Most abundantly lost and gained molecular functions (GOs). (A) Heatmap for core novel (++) GO molecular functions. Scale corresponds to percentage (%) of each molecular function in each core novel (++) HG per clade, calculated over the total spread of GO molecular functions. (B) Loss within a molecular function is indicated by filled blue circle (not necessarily loss of entire GO category). While different clades (columns) may have gained or lost the same functions, the actual HG gained or lost may be different. GO gained or lost in a clade refer to a subset of HG that perform that function, not all the HG associated with it.

Figure 2.9 Visual description for each of the HG types analysed in the bioinformatics pipeline.

“Loss” out group HG absence is optional. “Novel” in-group absence is optional. These allow some flexibility to account for incompleteness in genomes as detailed in Figure 2.3.

Figure 3.1 Gene tree consisting ~8000 homeodomain proteins. Coloured by superclass and rooted at a plant specific homeobox clade (HD-ZIP).

Figure 3.2 ROC analysis of the homeobox classification based on known HomeoDB classifications against same in-house database species. The % error rate for classification was found to be 0.87%. Numbers 2, 4 & 6 are cutpoints at which the test shows an abnormality.

Figure 3.3 The largest homeobox gene tree is dominated by PRD class genes with outgroups from HomeoDB. For this, and the following gene trees: each colour represents a different homeobox class. The light blue leaf labels are proteins that were not assigned a homeobox family, but were similar enough to cluster in the same gene tree. The diverging clade of rotifer

genes homologous to the Hdx family of POU class had Ultra-fast-bootstrap (UFBS) values of 100%.

Figure 3.4 One of the HOXL gene trees (ANTP class) with outgroups from HomeoDB. Hox9-13(15) appears to be well conserved throughout all the bilaterian animal clades, with lophotrochozoan genes closely related to vertebrate genes.

Figure 3.5 One of the HOXL HG ANTP class genes with outgroups from HomeoDB. Hox1 appears to be well conserved throughout all the bilaterian animal clades, but with some distinct phylum grouped clades. It is a large homeobox family with a couple of paralogues in each non-lophotrochozoan species.

Figure 3.6 One of the HOXL HG ANTP class genes with outgroups from HomeoDB. Hox3 originates in the stem of bilaterians. There has been some loss in some lophotrochozoan lineages, such as in some Mollusca and the brachiopod *Lingula anatina* here. The Hox3 in *Crassostrea gigas* and *Lottia gigantea* diverged slightly outside of the clade shown here.

Figure 3.7 Barx family one of the NKL ANTP class gene trees with outgroups from HomeoDB.

Figure 3.8 Barx family one of the NKL ANTP class HG genes with outgroups from HomeoDB. There has been significant expansions in cnidarians of the Barx.

Figure 3.9 Barhl family one of the NKL ANTP class HG genes with outgroups from HomeoDB. There is a clean monophyly of this homeobox family which is seen in every animal lineage.

Figure 3.10 The En family of ANTP NKL.

Figure 3.11 LIM class HG genes with outgroups from HomeoDB. This LIM class displays a divergence of the LIM families: Lhx1-8 and Lmx. Lmx is present in all the animal phyla, whilst Lhx expanded in bilaterians.

Figure 3.12 TALE-class. These genes have been classified as Pknox/Meis because it was too difficult to distinguish between them, although Meis is a Metazoan specific homeobox and Pknox evolved prior to the origin of animals.

Figure 3.13 CERS-class homeobox is dispersed through all the animal phyla and evolved prior to the origin of animals. There is a duplication event seen in vertebrates, but the rest of the animals have remained low copy number with distinct protostome and deuterostome clades with the exception of urochordates, which is likely an erroneous placement in the gene tree.

Figure 3.14 The ZF-class of homeobox is classified by having zinc-finger domains as well as the homeodomain (Bürglin & Affolter, 2016). It can have any number of additional domains and varying motifs, and for this reason it can be seen dispersed in paraphyletic clades across the gene tree.

Figure 3.15 The PROS-class has just the one homeobox family: Prox. This homeobox is dispersed throughout the animal phyla.

Figure 3.16 The HNF-class Hmbox family is a monophyletic clade. It has a lophotrochozoan specific divergence, with duplications seen in chordates, molluscs and annelids.

Figure 3.17 Presence of each homeobox gene family (gene families limited by HomeoDB classified genes, hence heavy bias towards vertebrates), as seen in each animal species, grouped by phyla. The named homeobox families and classifications only include the homeobox genes that have been identified in HomeoDB. Any novel or unclassified homeobox genes that have been found in this thesis, or that were already identified as homeobox genes but uncharacterised beyond that recognition have been collated under the class Other, family unassigned.

Figure 3.18 The gene tree pipeline including classification and identification of the homeobox HGs, described further in materials & methods.

Figure 4.1 A summarised consensus for the most commonly found animal phyla positions from the ML methods for all the protein sets described in methods & materials: HG60, HG90, HG101 and BUSCO293.

Figure 4.2 ML phylogeny of the HG60+ dataset with LG+R8 model. Additional protein data from non-genome sequences such as priapulids were used as a consensus to fill animal phylum gaps in HG60+ and HG90+. Non-animal outgroups collapsed.

Figure 4.3 ML phylogeny of the HG90+ dataset with LG+F+R7 model. There is a misplacement of outgroups between vertebrates and cephalochordates and elsewhere in the animal clade

are very erroneous and artefactual, displayed as collapsed triangle, and as seen with the *Bacillariophyta*.

Figure 4.4 ML phylogeny of the HG101 dataset using LG+F+R7 models. Every node has high UFBS values >98%. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

Figure 4.5 ML phylogeny of the BUSCO293 dataset using 293/303 BUSCO orthologues with LG+F+G model. Recovered a clade for Ctenophora and Porifera last common ancestor as first splitting extant lineage. This particular topology has lower (~70% UFBS) support for those nodes and a clade shared with all the fast evolving species. This phylogeny also recovers the shared non-bilaterian clade for Placozoa and Cnidaria, diverging before the emergence of bilaterians. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

Figure 4.6 ML phylogeny of the BUSCO293 dataset using 293/303 BUSCO orthologues with LG+C20+F model. Here a clade for Ctenophora and Porifera last common ancestor as first splitting extant lineage was recovered, and a shared clade for Placozoa and Cnidaria last common ancestor diverging before bilaterians emerged. There is poor support (UFBS <50%) for the unusual recovery of the three key bilaterian clades, but high internal support for each phyla. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

Figure 4.7 ML of HG60 dataset using LG+R10 model. Urochordates are unexpectedly recovered as first bilaterians, but with low UFBS support (72%) when compared to the rest of the nodes. Rotifera, Orthonectida and Platyhelminthe have high internal supporting relationships between them (100% UFBS), but the divergence of their last common ancestor has lower support, and this appears to be an artefact of LBA. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

Figure 4.8 ML of HG90 dataset using LG+R8 model. The phylogeny recovered here using a different dataset, but different model to HG60 in Figure 4.7 is identical. The UFBS values are higher. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

Figure 4.9 ML of the HG90 dataset using LG+F+R7 models. Whilst using the same dataset as in Figure 4.8, but different models, the phylogeny recovered here is more plausible. with urochordates recovered with vertebrates. The unexpected divergence prior to the cephalochordate clade is not well supported, and this is likely an artefact. Furthermore there is still the LBA issue occurring with the "fast evolving species". The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

Figure 4.10 Using the same dataset as HG101 in Figure 4.4, and Bayesian Inference (BI) with CAT+GTR model. There is high support for traditional clades in the 3 key bilaterian lineages: Ecdysozoa, Lophotrochozoa and Deuterostomia, with Platyhelminthes and Orthonectida are placed among lophotrochozoans with high bootstrap supports (98/100). There is a polytomy between the ambulacrarians and the chordates, both diverging from the last common bilaterian ancestor. Node points in green indicate posterior probabilities > 0.9. Using PhyloBayes 4.1: 2 Pb chains were run in parallel, with 2112 trees each. Maxdiff 0.98 and meandiff 0.09. The remaining outgroups have been collapsed to direct focus to the animals only.

Figure 4.11 Bayesian inference of the BUSCO293 dataset. Node points in green indicate posterior probabilities > 0.9. Using PhyloBayes 4.1: 2 Pb chains were run in parallel, with 1535 trees each. Maxdiff 0.98 and meandiff 0.08. The remaining outgroups have been collapsed to direct focus to the animals only.

Figure 4.12 The pipeline schema for each of the datasets and trees. Each step is introduced to reduce systematic errors in the resulting phylogenies. "auto model" refers to the parameter MFP + MERGE in IQTREE which uses ModelFinder (Kalyaanamoorthy et al., 2017) to choose an appropriate model.

Figure 5.1 A summary of all the results chapters. The most poorly supported nodes in chapter 4 in red, homeobox expansions for the origins of homeobox proteins from chapter 3 in green, size equals number of homeobox genes expanded and blue and red triangle sizes to match HG gains and losses from the biggest HG novel and loss counts.

LIST OF ABBREVIATIONS AND ACRONYMS

BI:	Bayesian inference
BLAST:	Basic local alignment search tool
BRH:	Best-reciprocal hit
HG:	Homology group
HMMs:	Hidden Markov models
LBA:	Long branch attraction
LCMA:	Last common metazoan ancestor
LCOA:	Last common opisthokont ancestor
MCL:	Markov-chain-clustering
ML:	Maximum likelihood
NS:	Nervous system
UFBS:	Ultra-fast bootstrap support

1 INTRODUCTION

The Animal Kingdom reveals intricate evolution, with various changes at molecular level in the genome, producing remarkable ecological and historical diversity. A well-supported evolutionary structure is necessary for comparative studies (Halanych, 2004). As well as supporting traditional hypotheses based on morphology, some of these new findings using molecular markers can challenge them too, being both controversial and leading to scepticism (Nosenko *et al.*, 2013). We are about to enter an introduction into the knowledge surrounding the animal tree of life because this is a fundamental topic having confounded scientists for a very long time.

1.1.1 EVOLUTIONARY RELATIONSHIPS IN ANIMALS

Eukaryotes are extremely diverse with many evolutionary adaptations to feeding, reproducing, and surviving. They are characterised by the possession of membrane-bound organelles, unlike prokaryotes (Burki, 2014). Animals are a part of the opisthokonts, a lineage whose relationships are supported by both molecular as well as morphological phylogenies. This super group divides into two key lineages: Holozoa, which contains Ichthyosporea, Filasterea, Choanoflagellata and Metazoa, and Holomycota, which contains Fungi and Nucleariida. These relationships are displayed in Figure 1.1. Hypotheses suggest that the last common opisthokont ancestor (LCOA) contained a stock of chitin synthases (Torruella *et al.*, 2015).

The animal tree of life as we know it today is a result of over a century of phylogenetic research based on anatomical and developmental features and, more recently, molecular data (Halanych, 2004). Metazoa are multicellular organisms, they can be classified into a paraphyletic group of animals: sponges, ctenophorans, cnidarians and placozoans, which emerged before bilaterians. The rest of animals are found within the Bilateria, which can be split into Deuterostomia, Lophotrochozoa and Ecdysozoa (Halanych, 2004). The monophyly of metazoans is well supported with a single origin of obligated multicellularity and other shared features as shown in Table 1.1 (Dunn *et al.*, 2014).

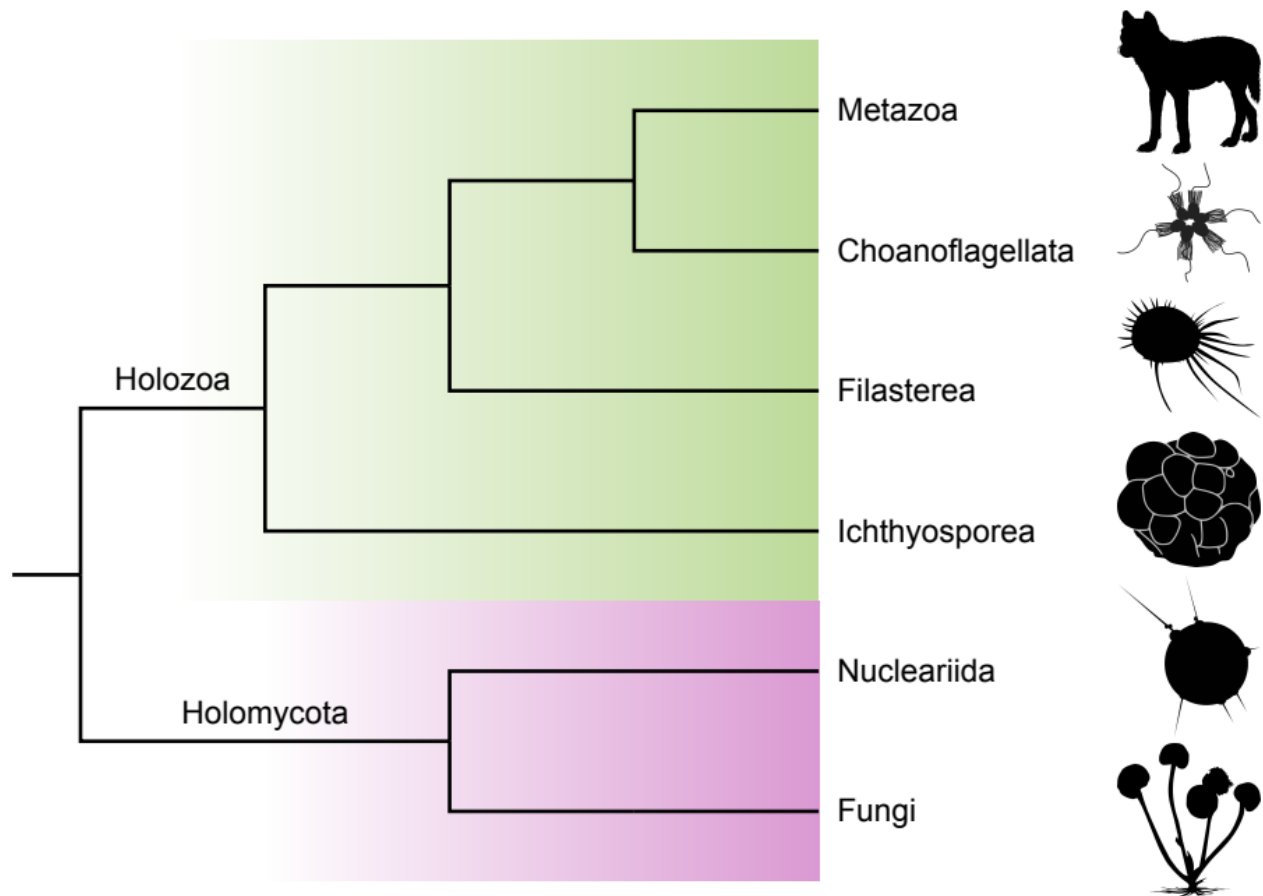


Figure 1.1 Phylogeny of the relationships within the Opisthokonta super group dividing into the two key lineages Holomycota and Holozoa in which animals placed in. Phylogeny as described in recent literature (Torruella *et al.*, 2012; Paps *et al.*, 2013).

Traditionally Porifera have been placed as the sister group to all animals (among the non-bilaterians, see Figure 1.2 & 1.3). Traditionally, all the animal clades except Porifera and Ctenophora have Hox and ParaHox genes (Dunn *et al.*, 2014; Giribet, 2016a). Recent studies have supported that calcareous sponges do have a ParaHox gene predating the emergence of poriferans (Fortunato *et al.*, 2014; Pastrana *et al.*, 2019). Cnidaria, Bilateria and Ctenophora all have nervous systems (NS), whilst Porifera and Placozoa do not. However, the homology of the elements of the NS further confuses the placement between clades, Ctenophora might have convergently evolved a NS, and remains outside Parahoxoa and within Eumetazoa (those with nerves) (Dunn *et al.*, 2014; Jékely *et al.*, 2015). Considering the relationships between animal phyla is important in unravelling the history and evolution of Metazoa. A general and possible summary of animal groups is shown in Figure 1.2.

Table 1.1 Key synapomorphies characterising each of the clades within this research topic, based on (Halanych, 2004; Dunn et al., 2014).

Clade	Synapomorphies
Opisthokonta	Chitin, with a single posterior flagellum
Metazoa	Collagen with mitochondrial gene reduction, oogenesis and spermatogenesis, with polar bodies
Bilateria	Bilateral symmetry with mesoderm, cephalization and both longitudinal and circular muscle structure
Deuterostomia	Enterocoely for embryo development leading to archimeric regionalization (prosome, mesosome, metasome), contain pharynx with ciliated gill slits
Lophotrochozoa	Triploblastic embryo development, spiral cleavage, protostome features, trochophore (“wheel”) larvae, the clade combines the trochophores and lophophores
Ecdysozoa	Chitinous external covering, all undergo ecdysis of this covering

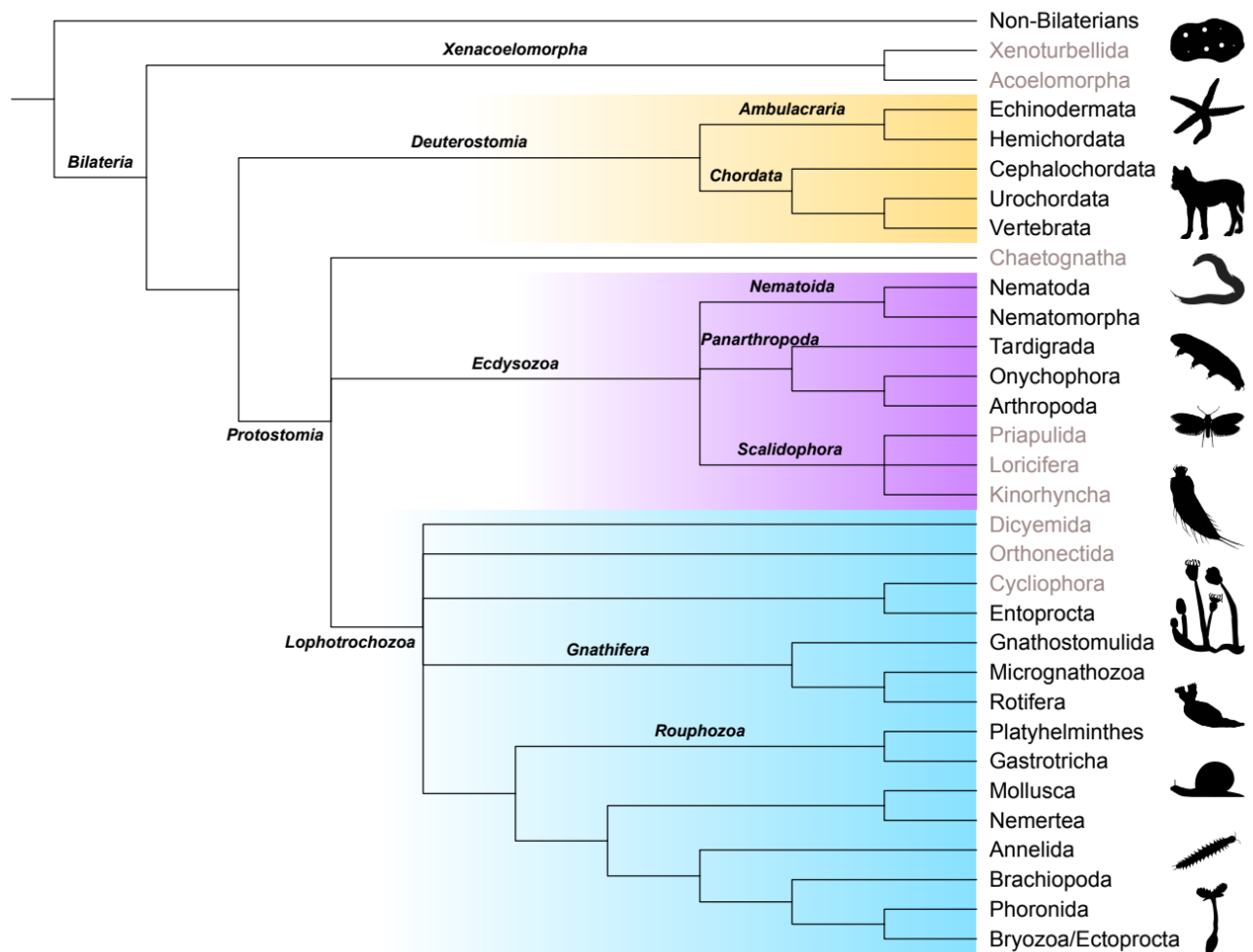


Figure 1.2 Tree depicting the possible relationships of the ~35 Phyla in the Animal Kingdom. Faded taxon labels are some of the least well supported placements. This representation has been compiled from multiple literature sources (Laumer et al., 2015; Giribet, 2016a; Kocot, 2016; Kocot et al., 2017; Laumer et al., 2019).

1.1.2 MOTIVATION FOR ANIMAL EVOLUTION

The origin of animals involved the transition from single-celled heterotrophic eukaryotes to the appearance of the first multicellular animal (Mills & Canfield, 2014). The further evolution of the animals as we know them today was the consequence of the radiation of the different phyla from this origin. It has been hypothesised that the lifestyles of animals was limited by the low oxygen levels in the Neo-proterozoic period. Coinciding with the Cambrian period, ~524-514 million years ago, and radiations of major animal lineages there were carbon isotope fluctuations as well as changes in oxygen levels atmospherically and oceanically (He *et al.*, 2019). Where the oxygen levels were low, it was expected that oxidative metabolism was inefficient and not enough to sustain complex animal life. As the levels of oxygen increased, the ability for aerobic life to develop rose, and so animals were able to further evolve (Mills & Canfield, 2014; He *et al.*, 2019). Recent studies show that some demosponges can survive in less than 4% of the current atmospheric oxygen levels. These are levels that are expected to have been around since before even the Ediacaran period, and therefore the increase in level of oxygen available probably has no real impact on the initial multicellularity of animals, but on the explosion in diversity (Knoll & Sperling, 2014).

Aside from the early increases in oxygen levels, further diversification within the Cambrian period is postulated to have occurred through dietary changes, evolving from osmotrophic organisms to carnivores. These were generally larger animals feeding on smaller animals, with much higher metabolic and oxygen requirements (Knoll & Sperling, 2014).

There is a constant battle between evidence from molecular data and fossil records supporting different phylogeny hypotheses and evolutionary histories, with molecular data placing animals or the metazoan radiation long before the appearance of the first fossil sponge (Dos Reis *et al.*, 2015). The molecular data supports how the metabolic requirements of the early animals adapted to the conditions of the environment (Mentel *et al.*, 2014).

1.1.3 THE ERA OF GENOMICS

In the last decade or so, phylogenomics has made massive progress in answering many evolution-based questions, this comes with the advancement in technology available, particularly with Next-Generation Sequencing able to quickly and relatively cheaply sequence whole genomes and transcriptomes, providing huge quantities of data. The only concerns holding back further

advancement in this era with huge datasets is the limitations in computational, man-power (Chan & Ragan, 2013), and the imbalance of available genomes across taxon disparity. Table 1.2 displays the number of assembled genomes per animal phylum published to date, as can be clearly seen, many diverse animal lineages are missing and there has been a biased focus on chordates, arthropods and nematodes.

The increase in the wealth of data, even early into the genomics era, has revealed shared features likely present in the last common metazoan ancestor (LCMA) (Brunet & King, 2017; Seb e-Pedr os et al., 2017). As the wealth of data increases, per every new genome sequenced and annotated, so does the knowledge on key metazoan lineage specific genomic expansions and contractions. Each of these fragments of knowledge collectively form an evolutionary tale on the diversification and the origin of multicellular animals.

With putative relationships between taxa in place (see Figure 1.2), this wealth of genomic data can be used to perform comparative genomics studies, such as the loss and the gain of novel genes in the origin and diversification of animals.

Table 1.2 *Assembled genomes that are publicly available up to 2017 as adapted from Simakov & Kawashima (2017).*

Major animal lineage	Phylum	Number of species genomes
Non-bilaterian	Porifera	1
	Ctenophora	10
	Placozoa	1
	Cnidaria	2
Basal-bilaterian	Acoelomorpha	0
	Xenoturbellida	0
Deuterostomia	Chordata	291
	Hemichordata	2
	Echinodermata	7
Ecdysozoa	Arthropoda	239
	Tardigrada	1
	Onychophora	0
	Kinorhyncha	0
	Priapulida	0
	Loricifera	0
	Nematomorpha	0

	Nematoda	81
Lophotrochozoa	Rotifera	2
	Orthonectida	1
	Chaetognatha	0
	Dicyemida	0
	Acanthocephala	0
	Gastrotricha	0
	Gnathostomulida	0
	Nemertea	0
	Platyhelminthes	29
	Phoronida	0
	Byrozoa/Ectoprocta	0
	Entoprocta	0
	Cycliophora	0
	Brachiopoda	1
	Mollusca	10
Annelida	2	

1.2 COMPARATIVE GENOMICS OF BETTER KNOWN ANIMAL LINEAGES

1.2.1 ANIMAL ORIGINS AND ANCESTORS

When discussing the beginnings of animals, we first need to look at the ancestors of animals. This involves in-depth analyses of the organisms closest to the transition from unicellular to multicellular. Combined with animal multicellularity are the specialised cell types unique to animals. Various studies have shown that the unicellular ancestors of animals already had some of the gene repertoire necessary to develop the functions for multicellularity. Choanoflagellates are one group of organisms that were expected traditionally to be the closest relatives of animals due to their similarity to the sponge choanocyte. Although this has been debated in a comparison of transcriptomes (Sebé-Pedrós *et al.*, 2017; Sogabe *et al.*, 2019). Otherwise, beyond morphological resemblance, relationships have been inferred with molecular data (Sebé-Pedrós *et al.*, 2017). Other close extant relatives of animals are the filastereans, such as *Capsaspora owczarzaki*, sister to the choanoflagellate and animal clade (Sebé-Pedrós *et al.*, 2017). These amoeboid organisms have genes that are essential to animal development (Sebé-Pedrós *et al.*, 2016). Comparing the regulatory genomes of animals against close non-animal eukaryotes determines which genes were present before animals, within the LCMA, and which ones are novel to the emergence of animals. This defining information begins to tell the tale of how animals came to be from their ancestors and more distant relatives.

The positions of the non-bilaterian animals and which was the first splitting animal is still uncertain. Opposing studies support either sponge (Philippe *et al.*, 2009; Pisani *et al.*, 2015; Whelan *et al.*, 2015a, 2015b; Feuda *et al.*, 2017; Simion *et al.*, 2017) or ctenophores (Dunn *et al.*, 2008; Ryan *et al.*, 2013; Mentel *et al.*, 2014; Moroz *et al.*, 2014) as the first splitting animal, and in some cases, even the position of placozoans has been debated (Laumer *et al.*, 2019). Genome analyses often place ctenophores as sister to all other metazoans (Ryan *et al.*, 2013; Moroz *et al.*, 2014), whilst the use of complex evolution models position sponges first (Pisani *et al.*, 2015; Whelan *et al.*, 2015a). A "sponges first" scenario offers a more parsimonious transition, for example in the case of NS. Either the LCMA had a NS which was lost in sponges and placozoans, or ctenophores evolved to have a nervous system independently and convergently to cnidarians and bilaterians (Jékely *et al.*, 2015). Comparative genomics is a key element to investigate the evolutionary history of animals.

In 2010, Srivastava *et al.* sequenced, assembled and annotated a demosponge genome from the phylum Porifera. For their analysis they hypothesised that the sponge was the first splitting lineage in the metazoan tree, sister to the Eumetazoa (the ctenophoran-cnidarian-bilaterian clade). They found 4670 ancestral gene families including the sponge and eumetazoa, with 27% of these being specific only to the metazoans. Around ~75% of this 27% arose by gene duplications within Metazoa, in particular transcription factor families such as homeodomains. The sequenced sponge had kinase domains from well-known metazoan families including EGFR, Met, DDR, ROR, Eph and Sevenless (Srivastava *et al.*, 2010).

Srivastava *et al.* (2010) assessed the key components of animal multicellularity: cell regulation, cell-adhesion, controlled cell death, gene regulation and signalling, innate immunity and allorecognition and specialisation of cells. With a small handful of non-metazoan genomes as an outgroup, 3 non-bilaterians, an arthropod, nematode, sea urchin and a human as diverse representatives, they defined a list of genomic synapomorphies for animals (Srivastava *et al.*, 2010). The putative orthologous genes were identified by reciprocal BLASTs (Basic local alignment search tool) against either the mouse, human or *Drosophila* genes, and ontologies taken from PANTHER hidden markov models (HMMs). Orthology was determined by phylogenetic trees using the neighbour joining method with 100 bootstraps (Srivastava *et al.*, 2010).

Whilst a diverse taxon sampling was used, it was still highly limited. For all the software used, only the software's default parameters were used, and these may not have been optimal for the analyses required. Srivastava *et al.* (2010) left open questions such as why there is such a diverse range in the morphological complexity of metazoans, and to what extent is this involved in the evolution of the animals, what determines the retention of some of these key multicellular components? With the presence of bilaterian transcription factor classes observed in most non-bilaterians, including sponges, is it the quantity rather than content that produces the diversity in animal body-plans?

Albalat and Cañestro (2016) proposed that the search of a comprehensive catalogue of gene loss in a diverse taxon sampling would be of great benefit to translational medicine as well as evolutionary biology. This is dependent on a reliable phylogenetic tree to provide an accurate gene loss map, limited by accurate annotations and phylogenetic inferences (Albalat & Cañestro, 2016). They reviewed the prevalence and importance of analysing gene loss in the evolution of animals, particularly within Eumetazoa. They collated studies which, upon sequencing various cnidarian genomes, found that there was a loss in complexity from the ancestral genome to the cnidarian

genome which was revealed in the loss of genes comparatively. This contradicted the traditional views that evolution leads to an increase in complexity and that evolution is spurred by novelty (Albalat & Cañestro, 2016).

In the quest to reconstruct the LCMA, Paps and Holland analysed the largest whole genome taxon sampling to date, comprising both animal outgroups and broad sampling of animal phyla. Using a phylogenetically aware pipeline, meaning the pipeline followed a set phylogeny for inference, they were able to establish a minimum protein-coding genome that would have been present in the last common metazoan ancestor (Paps & Holland, 2018). The reliability of their pipeline was determined by the correct clustering in homology groups of well-established gene classes, families and superfamilies. Twenty-five core gene novelties, new genes that are refractory to gene loss, were identified to be present in the emergence of metazoans, a large portion of these related to multicellularity such as transcription factors and cell-signalling pathways. Their results additionally showed elevated expansions of novel gene families in the emergence of major lineages Planulozoa (cnidarians, placozoans and bilaterians) and Bilateria, a summary of their numbers can be seen in Figure 1.3. The large and diverse selection of non-animal outgroups meant that these results were more reliable than a smaller and less diverse outgroup would have been, by reducing the false positives among novel genes (Paps & Holland, 2018).

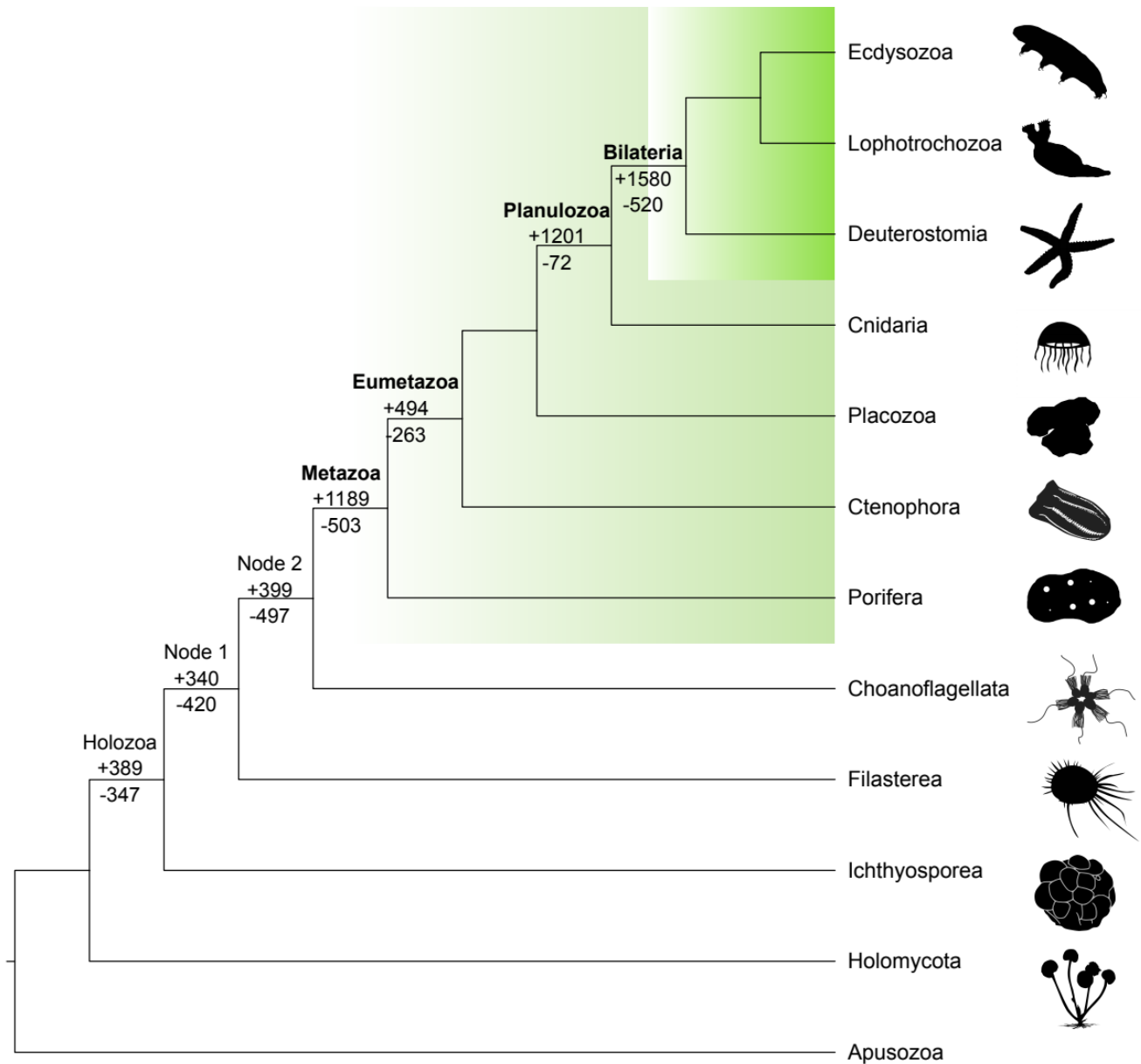


Figure 1.3 Summary of novel and lost gene families in the emergence of Metazoa. Numbers with (+) denote novel gene families and numbers with (-) denote gene family losses. Highlighted are the lineages of interest in this thesis, results from (Paps & Holland, 2018).

1.2.2 WHAT MAKES BILATERIANS STAND OUT?

Bilaterians diverge into 3 distinct lineages; Deuterostomia, Lophotrochozoa and Ecdysozoa, the two later forming the Protostomia (Figure 1.2 & 1.3). To understand the emergence of bilaterians, all 3 lineages are necessary for in depth analyses as well as those that are non-bilaterian. As noted by Simakov & Kawashima (2017), each new animal genome that has been sequenced, assembled and

annotated has revealed genomic novelties at both non-coding and coding levels (Simakov & Kawashima, 2017). Performing informative comparative genomics analysis requires a broad taxon sampling within animal phylogeny, so it is a given that the more diverse species sampled the more the resolution improves. Lophotrochozoa have received a little more attention in recent years with the sequencing of brachiopod, phoronid and nemertean genomes (Luo *et al.*, 2015, 2018), but the majority of in-depth comparative genomics analysis has covered model organisms such as humans and fruit flies in the deuterostome and ecdysozoan lineages respectively (Simakov & Kawashima, 2017).

Bilateria is one of the most well-supported animal lineages (Philippe *et al.*, 2005a). However, the emergence of the major animal lineages within the Bilateria clade still has some uncertainty, particularly surrounding the complex relationship with the first splitting bilaterians. According to molecular data, the clade Xenoacoelomorpha (Acoelomorpha as sister group to Xenoturbellida), has been placed as either the first splitting bilaterian and within Deuterostomia (Philippe *et al.*, 2011; Ruiz-Trillo & Paps, 2016), or a mix of both (Philippe *et al.*, 2011). Previously, Xenoacoelomorpha was also placed within the phylum Platyhelminthes as a simple flatworm member (Ruiz-Trillo & Paps, 2016). Platyhelminthes have also had a battle for a secure position in the animal phylogenetic structure, having been placed almost everywhere at some point (Kocot, 2016; Ruiz-Trillo & Paps, 2016). Given these conflicting placements, there are alternate theories about whether these animals derive from a simple ancestor, or are simplified (Ruiz-Trillo & Paps, 2016). At this point, comparative genomics can shed new light in this question.

LOPHOTROCHOZOA UNCERTAINTY

Simakov *et al.* (2013) looked at the evolution of Bilateria comparing 5 genomes from 3 lophotrochozoan phyla. They highlighted the poor level of taxon sampling within lophotrochozoans, with only Platyhelminthes having been thoroughly sequenced (mostly the parasitic members of the phylum), and how those well-studied genomes do not reveal a general model for evolution of lophotrochozoans. They assembled novel genomes for a limpet, leech and polychaete providing ~23,000-33,000 predicted protein coding genes for each genome (Simakov *et al.*, 2013). Through comparative genomics of 18 bilaterians, of which 5 were lophotrochozoas, and 4 non-bilaterian genomes, their results suggested that there were nearly ~9000 ancestral gene families in the last common ancestor of bilaterians with ~800 of those found in the 3 new lophotrochozoan genomes. The functions of these genes mainly included metabolic enzymes, epithelial sodium channels and G-protein-coupled receptor superfamilies. The ~9000 ancestral genes, with gene duplication events,

cover ~47-85% of bilaterian genes in modern genomes (Simakov *et al.*, 2013). Looking at gene loss, the 3 new genomes had retained ~94% of the bilaterian ancestor genes, much more than humans with 86% and even more so than the already sequenced platyhelminth genomes at a combined ~65% (Simakov *et al.*, 2013). Using BLAST methods, 231 putative lophotrochozoan-specific genes were extracted. HMMs showed that 188 of these actually had residual similarity to other bilaterian, non-lophotrochozoan gene families (Simakov *et al.*, 2013). The results of this paper provided clear evidence that comparative genomics using whole genomes from a range of taxa is essential to resolving the relationships between phyla. The poor retention of Bilateria ancestral genes in current model organisms, suggests that they may not be the most representative species, and possibly molluscs or annelids should be looked at more closely. Furthermore, whilst Simakov *et al.* (2013) increased the level of taxon sampling by including 3 new genomes for mollusc and annelid representation in the lophotrochozoans, increasing representative taxa from other phyla may provide an even more detailed picture, resolving uncertain relationships between them and providing better understanding for the evolution of these phyla.

The brachiopod genome, *Lingula anatina*, was first decoded in 2015, interestingly, these animals share many apparent anatomical characteristics with both deuterostomes and molluscs, such as hard tissue components seen in bone formation and shells (Luo *et al.*, 2015). The brachiopod genome has potential to bridge some gaps in the evolutionary pathway of lophotrochozoans alongside molluscs. A comparison in the genes expected to have a role in shell formation in molluscs shared with the *Lingula* genome showed that those were also shared with vertebrates, while the rest of the mollusc biomineralisation-related genes had evolved independently of brachiopods. *Lingula anatina* likely had independent expansions of the biomineralisation-related genes more involved in the calcium-phosphate based mineralisation that is seen in vertebrates (Luo *et al.*, 2015). Luo *et al.* (2018) furthered the analysis of Lophotrochozoa with the phoronid and nemertean genomes. Unfortunately, alongside many other genomes (Luo *et al.*, 2018), these were not all available at the time of the analyses within this thesis, and due to time constraints, could not be included. The authors, Luo *et al.* (2015), undertook comparative analysis on these newly sequenced genomes, and performed phylogenomic analyses with transcriptome markers for other lophotrochozoans. Due to unavailability of ectoprot and other closely related genomes, the lophotrochozoan relationships were still left uncertain. The genomes were compared with other animals, including non-lophotrochozoan bilaterians (31 in total including lophotrochozoan genomes). They found that lophotrochozoans and deuterostomes shared over 4000 genes not seen in ecdysozoans, over 1100 gene families novel to brachiopods, phoronids and nemerteans and ~7000 lophotrochozoan gene families. They found ~2800

gene families to be bilaterian specific, and exclusive of sponges and cnidarians. Their key finding showed that lophotrochozoans and deuterostomes had retained a bilaterian toolkit related to control of homeostasis and multicellularity (Luo *et al.*, 2018).

Discrepancies in results can easily be explained by how the homology groups or orthologies are defined as ancestral. For example, Simakov and collaborators defined ancestral genes to bilaterians as an orthologous group that contains 2 representatives from protostomes and deuterostomes and 2 from non-bilaterian metazoans (Simakov *et al.*, 2013). When Srivastava *et al.* (2010) studied the key components of animal multicellularity, they only required orthology groups to be present in two in-groups at least once and in an out-group for the gene families to be ancestral (Srivastava *et al.*, 2010) (Srivastava *et al.*, 2010). As could be expected, a stricter selection requiring more in-groups to have a gene family for the gene family to be ancestral would produce fewer ancestral orthology groups.

ECDYSOZOA - SO MUCH DATA, YET SO MUCH MISSING

Similarly to lophotrochozoans, the internal nodes within ecdysozoans are uncertain. Molecular and morphological data support a panarthropod clade, grouping onychophorans, tardigrades and arthropods. However, some molecular data tend to place tardigrades as sister to Nematoida (Nematoda and Nematomorpha) instead (Edgecombe, 2010). Genomic data has not been available for the other members of Ecdysozoa, although Scalidophora (Loricifera, Kinorhyncha, and Priapulida) is a traditionally accepted clade (Giribet & Edgecombe, 2017); although this has been questioned (Laumer *et al.*, 2019). The first animal genomes ever sequenced were those of ecdysozoans; the nematode *Caenorhabditis elegans* and the arthropod *Drosophila melanogaster*, known as model organisms, and since there has been a high level of sequencing data, almost equalling that of vertebrates, although not always made publicly available and nearly always transcriptome rather than genome data (Giribet & Edgecombe, 2017).

The sequencing of two tardigrade genomes garnered a lot of scientific interest given the apparently huge horizontal gene transfer they displayed. Much of this was disputed in *Hypsibius dujardini* as contamination, but a lower level of horizontal gene transfer was corroborated with another tardigrade genome *Ramazzottius varieornatus* (Arakawa, 2016; Hashimoto *et al.*, 2016; Yoshida *et al.*, 2017). Many uncharacterised novelties were found in the tardigrade species, particularly expansion of stress related gene families, and also a loss of gene pathways promoting stress damage, such as regulation of response to hypoxia. Using 27 other animal genomes on top of the 2 fully sequenced tardigrade genomes, and transcriptome data from 8 non-arthropod and non-

nematoid species, Yoshida *et al.* (2017) compared the tardigrade genomes to observe these tardigrade-specific genes. Orthofinder (Emms & Kelly, 2015) was utilised to cluster protein families among these animal species and KinFin to identify expansion and contraction of these protein clusters. They found that many animal-shared protein families were seen even more frequently in the tardigrades than other animals and 1486 clusters were tardigrade specific novel protein families. Tardigrade-specific protein families included; Wnt, Frizzled, and chibby (Yoshida *et al.*, 2017).

DEUTEROSTOMIA - GETTING THERE

Unlike lophotrochozoans, deuterostomes traditionally have a more stable and supported internal phylogeny (although not always; Philippe *et al.*, 2019), with hemichordates and echinoderms sister clades forming Ambulacraria, with ambulacrarians as sister to chordates. Pharyngeal gill slits unite ambulacrarians and chordates as a deuterostome ancestral trait, having been lost in amniotes and modern echinoderms only. Alongside this morphological support is the molecular support of shared Pax1/9 gene expression (Lowe *et al.*, 2015; Simakov *et al.*, 2015). With this well-supported phylogeny, Simakov *et al.* (2015) were able to distinguish around ~8700 homologous gene families in the last common deuterostome ancestor in comparing 30 animal genomes. Over 30 gene novelties were identified as specific to deuterostomes and no other metazoans, although approximately 12 of these were seen in microbes. An explanation for this weird finding comes in the forms of convergent loss along the other opisthokont lineages or horizontal gene transfer from marine microbes (Simakov *et al.*, 2015).

1.2.3 METHODS IN GENOME COMPARISON

Heavily related to phylogenetic studies are issues in orthology assignment. The use of paralogues instead of orthologues in phylogenetic reconstruction is argued to be a major source of phylogenetic conflict (Nosenko *et al.*, 2013). Torruella *et al.* (2012) identified that one way to avoid this potential pitfall was to use single-copy protein domains whose orthology is highly reliable. In selecting single-copy domains across taxa, paralogy problems were prevented and later down the line other systematic errors with molecular-level models were also reduced (Torruella *et al.*, 2012). The approach was successful for the dataset, but does have limitations, as new genomes are sequenced, do these regions remain as well conserved within increased numbers of genomes? As new genomes are included, these conserved single-copy domains may prove to be less conserved than believed, excluding some more distinct species, or they may prove to have more than one copy in other species,

reducing the actual count of conserved sites that can be used to infer phylogeny using accurate models. We go on to discuss alternative strategies and algorithms.

ORTHOLOGY ASSIGNMENT

Orthologous genes are those related through a speciation event, whilst paralogous sequences have occurred through a gene duplication event. At this point, there are already discrepancies in these definitions, which can cause confusion among scientists. Orthology is a concept in evolution, which is subject to change as knowledge surrounding evolution gains greater depths. The different evolutionary pathways a gene may have descended through, be it a combination of duplication and speciation events, may obscure their relationships. Further evolutionary complications such as fusion and recombination events can also merge the lines between homology types (Gabaldón, 2008; Gabaldón & Koonin, 2013; Holland *et al.*, 2017). There are 'asymmetrical' modes of evolution and whole genome duplication in which it is difficult to identify the gene relationships. A similar situation also holds true if only part of the locus is duplicated, then the terms orthologous and paralogous do not fit (Holland *et al.*, 2017).

Three well known methods for orthology assignment include; pairwise methods which comprises of best-reciprocal hits (BRH) between sequence pairs in different species, gene tree inference in which a phylogeny of genes is analysed for duplication and loss events, and species overlap methods where species connected to two sister nodes have a measured overlap to determine speciation or duplication against parent node (Gabaldón, 2008).

ORTHOMCL, ORTHOFINDER AND TRIBEMCL

The identification of orthologues is thought to be a reliable instrument in gene annotation, whereby an orthologous gene is theoretically well-conserved in sequence and therefore also in function. Orthologues are then further utilised in comparative genomics in order to classify organisms on a whole genome scale. OrthoMCL is a tool designed specifically to identify orthologues in eukaryotes (Li *et al.*, 2003; Tabari & Su, 2017). The OrthoMCL works by reciprocally comparing sequence similarity for all proteins of interest against each other and themselves using BLAST, separately between and within species. Producing a similarity matrix, normalised by species using weighted edges to counteract the effects of recent paralogues, from the BLAST results and using Markov-chain clustering (MCL) to group similar proteins. The MCL step uses the e-value from the BLAST search to inform the clustering. Then the orthologues are inferred by a using a MySQL

database that assumes orthology based on the e-value. The idea is that the algorithm reduces the impact of orthologue misassignment caused by over-similar recent paralogues through identifying and weighting these paralogues. The issue however remains that eukaryotes exhibit a higher rate of duplication than other organisms, causing confusion between functional divergence and redundancy (Li *et al.*, 2003). OrthoMCL does make headway into this problem (Li *et al.*, 2003; Tabari & Su, 2017). PorthoMCL is another implementation of OrthoMCL with the same output, but on a computationally parallel scale to improve efficiency issues (Tabari & Su, 2017).

Both mentioned pipelines make use of MCL to group together similar proteins based on BLAST results using graph theory. Tribe-MCL avoids the problems associated with assigning orthology altogether, it simply aims to group protein families together. Protein families are an essential element in functional and structural genomics (Enright *et al.*, 2002). Using the e-values for pairwise BLAST alignments, a two-dimensional network graph is constructed in which the sequences are the nodes and the edge lengths are proportional to their e-values. The Tribe-MCL algorithm then defines groups of genes based on how tightly clustered the sequences are; the input parameter of inflation and granularity defines the tightness of the clusters (Enright *et al.*, 2002). Small adjustments have been shown to have little to no effect on the eventual clusters produced (Enright *et al.*, 2002). An important ability of the MCL algorithm to note is the ability to cluster even multidomain proteins in which sequence similarity may fit into more than one protein family, which it does by the iterative inflation parameter (like bootstrapping) to assign it to the correct and stronger cluster (Enright *et al.*, 2002).

OrthoFinder is another pipeline intended to identify orthologues, however, rather than determining orthologues and paralogues as separate entities, it aims to extract orthogroups without bias. Similarly to OrthoMCL and PorthoMCL, OrthoFinder ignores synteny to solve the problem, where synteny is not conserved over large evolutionary distances (Emms & Kelly, 2015). OrthoFinder follows the same initial BLAST-based plus MCL clustering approach as OrthoMCL. However, the use of BLAST e-value to cluster genes might produce biases based on gene length (e.g. longer genes tend to get higher e-values for the same ratio of identical similarity matches in shorter genes). Instead, OrthoFinder uses the BLAST bit-score normalised by gene length to increase accuracy and reduce biases within orthogroups. Each orthogroup is defined as having derived from a single ancestral sequence. In retaining these orthogroup relationships, much evolutionary information is retained. Orthofinder can additionally be used to infer orthologues by calculating gene trees and species trees together, and reconciling those (Emms & Kelly, 2015). This pipeline improves upon OrthoMCL by reducing some of the bias introduced by sequence length and divergent sequences. It does this by

utilising bit-scores instead of e-values, so that sequence length has no determination in the similarity statistics, otherwise shorter sequences may score lower than they should (Emms & Kelly, 2015). However, there is little evidence for or against to suggest that removal of these biases works for all datasets to more accurately cluster these protein groups, since it has only been benchmarked against the model OrthBench datasets. There could be misassignment issues caused by fragmented sequences in which normalisation has overcompensated for the sequence length bias.

In summary, both OrthoMCL and OrthoFinder attempt to define groups of orthology from the output produced by Tribe-MCL

CONCERNS IN ORTHOLOGY

The use of gene trees to determine orthology is probably the least popular method currently. It relies on reliable species trees, the process is a complicated multi-step backwards and forwards requiring prior knowledge about the species, genes, and sequences. Some software programs have attempted to bypass the need for a reliable species tree, but these have other issues, such as more reliable gene trees, and even presumptions on rooting the trees. All tree-based methods are also computationally expensive (Kuzniar *et al.*, 2008).

Graph-based methods using precomputed BRHs or BLAST pairwise sequence similarity data such as Tribe-MCL and OrthoMCL are much more popular. A hybrid between both tree inference and graph-based methods is suitable for genome-wide analyses but does not provide multi-level phylogenetic resolution in *de novo* orthogroups. Most orthology detection methods, particularly tree-based approaches, currently still suffer with artefacts from chimeric sequences, caused by common evolutionary events including; fusion, fission, shuffling, gain and loss of protein domains (Kuzniar *et al.*, 2008).

Another problematic issue in orthology detection involves the reconstruction of gene loss. The loss of a paralogues in some species may lead to out-paralogues being incorrectly assigned as orthologues. Out-paralogues refer to speciation events following duplication events. Tree-based methods are able to process out-paralogue information to an extent with multiple gene losses, but graph-based methods for orthology are unable to cope with that type of situation. The misassignment of these false orthologues can have compounding effects on the rest of the orthologue assignments and the resulting evolutionary inference. Graph-based methods are able to cope well in the case of a single gene loss (Kuzniar *et al.*, 2008). Major tree-based methods are sufficient for many evolutionary

scenarios, but not for other complex scenarios including horizontal gene transfer (HGT), or are computationally expensive (Kuzniar et al., 2008).

With the concerns in orthology assignment as listed, the best approach may be a compromise, to go back to basics, avoid distinguishing between orthologues and paralogues and look only at the homologues as a whole as with Tribe-MCL, avoiding the oversimplification of evolution with these terms (Gabaldón & Koonin, 2013). Keeping this in mind, homology groups in this thesis are defined simply as clusters of related and similar proteins with lesser resemblance to other homology groups; these likely descend from a single ancestral sequence. These might coincide with different traditional categories of gene classification, from gene superfamilies (e.g., Wnt ligands, homeobox genes) to protein families and all the intermediate levels.

1.3 GENE FAMILY EVOLUTION

Gene trees provide an insight into the evolution of gene superfamilies and the relationships between them. Gene trees are constructed by aligning the sequences of homologous genes across species and inferring a phylogenetic tree to find the evolutionary relationships (Nam & Nei, 2005). Then, taking into account the evolutionary relationships between the taxa analysed, the patterns of gene duplications and losses can be inferred, informing the definition of different gene families.

1.3.1 HOMEODOMAIN DIVERSITY

A key gene superfamily used in understanding the evolution of animals is the Homeobox genes. Homeobox genes are transcription factors that play essential roles in animal development, mostly by defining body axis, such as the anterior-posterior axis in bilaterians, but also other types of patterning ranging from the formation of digits, to the patterns of butterfly wings, or cancer (Holland, 2013). They are characterised by the presence of a homeodomain that forms three helices; the second and third connected by a short loop forming a helix-turn-helix structure. The topology of the third one is determined by the primary sequence, and provides the specificity to different DNA target sequences (Holland, 2013).

Homeobox genes are prevalent throughout eukaryotes, but the vast majority of diversification has been seen in the evolution of animals (Holland, 2013). Homeobox genes in animals can be described in 11 classes and these further into the gene families as summarised in Table 1.3. The increase in homeobox genes in different animal lineages might be driven by whole genome duplications as well as gene duplication events, and it is thought such increase leads to an increase in the complexity of animal body plans (Holland, 2013).

Table 1.3 A summary of the identified homeobox genes and pseudogenes in animals as classified in HomeoDB2 (Zhong *et al.*, 2008; Zhong & Holland, 2011).

Classes Subclass Families

		Cdx Evx Gbx Gsx Hox1 Hox2 Hox3 Hox4 Hox5 Hox6-8 Hox9-13(15)
	HOXL	Meox Mnx Pdx
ANTP		
	NKL	Abox Ankx Barhl Bari Barx Bsx Dbx Dlx Emx En Hhex Hlx Hx Lbx Lcx Msx Msxlx Nanog Nedx Nk1 Nk2.1 Nk2.2 Nk3 Nk4 Nk5/Hmx Nk6 Nk7

	Noto Ro Tlx Vax Ventx
PRD	Alx AprdA AprdB AprdC AprdD AprdE Argfx Arx CG11294 Dmbx Dprx Drgx Dux Esx Gsc Hbn Hesx Hopx Isx Leutx Mix Nobox Obox Otp Otx Pax2/5/8 Pax3/7 Pax4/6 Phox Pitx Prop Prrx Rax Repo Rhox Sebox Shox Tprx Uncx Vsx
LIM	Isl Lhx1/5 Lhx2/9 Lhx3/4 Lhx6/8 Lmx
POU	Hdx Pou1 Pou2 Pou3 Pou4 Pou5 Pou6
HNF	Ahnfx Hmbox Hnf1
SINE	Six1/2 Six3/6 Six4/5
TALE	Atale Irx Meis Mkx Pbx Pknox Tgif
CUT	Acut Cmp Cux Onecut Satb
PROS	Prox
ZF	Adnp Azfh Tshz Zeb Zfhx Zhx/Homez
CERS	Cers
Other	Ahbx Beetlebox Bix Cphx Crxos1 Gm5585 Gm7235 LOC647589 Muxa Muxb NANOGNB Sia unassigned

1.3.2 EARLY HOMEBOX GENES

EUKARYOTES

Homeobox genes are specific to eukaryotes and have not been observed in Archaea or bacteria (Laughon & Scott, 1984). These genes do however have a similar peptide motif structure to the helix-turn-helix proteins found in bacteria, suggesting evolution from a similar gene (Laughon & Scott, 1984). Homeobox genes are so diverse that it has been difficult to pinpoint the early events leading to their evolution. The TALE class of homeoboxes is observed in most eukaryotes, with sufficient sequence similarity to show a shared origin between plants and animals (Holland, 2013). Only the TALE class is found across eukaryotes, all others likely evolved in different lineages (Holland, 2013). Parasitic and intracellular eukaryotes have lost homeobox genes, probably as a consequence of their body plan simplification (Holland, 2013).

The largest class of homeoboxes is the ANTP class. ANTP is specific to animals with around 50 known gene families to date, of which most emerged in the last common bilaterian ancestor. ANTP is divided into two subclasses of homeobox gene; Hox/paraHox-related (HOXL) and NK/NK-related (NKL) (Holland, 2013). Previous studies have shown that the NKL subclass genes can be found in all metazoans (Holland, 2013; Ferrier, 2016; Paps, 2018). The explosion of the ANTP class of genes in the emergence of animals and bilaterians could be a huge clue to the evolution and diversification of animals.

HOX/PARAHOX IN SPONGES

As mentioned before, only the NKL class has been found in sponge genomes, but never ANTP class genes until a ParaHox gene was reported by Fortunato *et al.* (2014). The 'ghost-locus' is a hypothesis that the absence of HOXL in the sponge *Amphimedon* is actually a secondary loss; this is supported by a synteny analysis. With this theory as a basis, Fortunato *et al.* (2014) analysed the ANTP class of homeoboxes in calcareous sponges: *Sycon ciliatum* and *Leucosolenia complicata*. They were unable to detect any HOXL genes in the sponges, except a single *Cdx* in each sponge. Their findings confirmed the 'ghost-locus' hypothesis, suggesting that the HOXL subclass predates the emergence of sponges. They also recommended the necessity to analyse more sponge genomes from multiple sponge lineages to visualise the whole homeobox ancestral story (Fortunato *et al.*, 2014).

In 2019 Pastrana *et al.* (2019) disputed these new findings as artefactual. They claim that the methodology was flawed and titled their paper similarly as '*Sponges lack paraHox genes*'. Fortunato *et al.* (2014) did not provide strong evidence in their approach, the inference was not as trustworthy as had they used neighbour-joining, maximum likelihood (ML) and Bayesian methods to result in differing clade assignments for the genes. The *Cdx* clade containing the sponge and bilaterian genes was placed far and separately from the rest of the HOXL genes, suggesting a phylogenetic signal issue (Fortunato *et al.*, 2014; Pastrana *et al.*, 2019). Pastrana *et al.* (2019) reanalysed the datasets used by Fortunato *et al.*, with additional taxa to reduce the sensitivity to the artefactual issues using multiple and varying gene-tree methods and mathematical models. The sponge gene identified as *Cdx* by Fortunato *et al.*, was retrieved alongside *Ankx*, of the NKL subclass. The placement of these genes appear to be inflicted by long branch attraction (LBA) (Pastrana *et al.*, 2019). LBA is a frequent issue in both species and gene tree inferences, this will be discussed later in this thesis.

BILATERIA

As previously discussed, some hypotheses place *Xenacoelomorpha* at the base of Bilateria, so exploring the homeobox set of genes in these animals could be of importance when looking at the evolution of animal diversity and body plans. Brauchle *et al.* (2018) compared, identified and classified the homeobox genes in RNA sequence data of these animals alongside existing transcriptome data and found all 11 animal homeobox gene classes present in the last common ancestor of bilateria (LCAB). Depending on the animal phylogeny, if xenacoelomorphs were the first splitting bilaterian, it is expected that the last common bilaterian ancestor also had the full 11 classes too (Brauchle *et al.*, 2018). The homeodomain sequences were retrieved and identified from the data through BLASTp of the NCBI and HomeoDB2 databases (Zhong & Holland, 2011; Brauchle *et al.*, 2018) and considered as long as the top 10 BLASTp results were in a Homeobox gene family. ML methods were used for phylogeny inference of the genes in species and class subsets for the 60-amino acid homeodomains with PhyML. To corroborate findings, the tree inferences were repeated with whole homeobox sequences instead of just the homeodomains. Whilst *Xenacoelomorpha* retained all 11 homeobox classes, this phylum was missing some animal-specific homeobox families. Cnidarians, as sister to bilaterians, possess several HOXL families, known as *antHox*, or anterior Hox, whilst xenacoelomorphs appeared to have only 1, thought to be equivalent to *Hox1* in humans. This supports a scenario in which many of these gene family homologues have been lost in *Xenacoelomorpha* (Brauchle *et al.*, 2018). It should be noted that previous studies do not support this finding, and found 3 Hox and 2 paraHox genes in *Nemertodermatida* (grouped with Acoels) (Jiménez-Guri *et al.*, 2006). This is relevant because it provides an alternative possible evolutionary history of these gene family homologues and of the phylogenetic position of these animal clades.

1.3.3 PATTERNS OF HOMEBOX EVOLUTION IN OTHER ANIMALS

Nam and Nei (2005) looked at the pattern of homeobox evolution in bilateral animals. They ran a phylogenetic analysis for the 2,031 homeobox sequences they extracted from comparative analysis of 11 bilateral taxa, of which 7 were deuterostomes, 2 lophotrochozoans and 2 ecdysozoans. They aligned only the homeodomains because these were the most alignable algorithmically, optimally, and computationally. They used the neighbour-joining algorithm to construct the tree rather than maximum-parsimony or maximum-likelihood methods due to the sheer quantity of sequences to analyse. The genes were divided into pre-defined groups according to structure and HMM similarity,

then a tree was constructed for these pre-defined groups (Nam & Nei, 2005). Using the phylogenetic results to assess how different homeobox genes had been lost or gained across the 11 taxa and how they had mutated, they found that some of the changes had not noticeably affected species phenotypically, some losses had beneficial effects, such as morphological differentiation, and that the duplication events left a backup, meaning that some of the genes were neutral and not essential for fitness of the animal (Nam & Nei, 2005).

Paps *et al.* (2012) investigated the diversity of transcription factors other than the homeobox super families: Paired, Fox, and Tbx superclasses, and the Sox class. Pfam sequences and Pfam profiles for human, mouse, and *drosophila* were mined and used to BLASTp and HMMER (a software package used to identify homologous sequences) for results in *Danio rerio*, *Ciona intestinalis*, *Branchiostoma floridae* and *Strongylocentrotus purpuratus* (Paps *et al.*, 2012). The authors extricated the domains of all these sequences and submitted them to multiple sequence alignments. Phylogenetic analysis was performed using the LG-Gamma-Invariant evolutionary model rather than other more frequently used models because this model is based on a much larger and diverse matrix estimation database (Paps *et al.*, 2012). The phylogenetic gene trees were used to comparatively assess each of the transcription factor super classes. The results for the Paired/Pax genes were expected to reveal similar patterns to homeobox results for amphioxus since the two have overlapping features. All chordates except amphioxus had lost the gene *eyegone*. For the rest of the transcription factors, amphioxus had retained all of them, similarly to homeobox genes in other studies (Paps *et al.*, 2012).

On the surface, annelids appear to be morphologically simple. However, despite the standard head followed by segmented trunk and tail, the variance in the head and tail shapes, internal anatomy and number or size of segments allows for an unlimited level of diversity within the phylum. This morphological diversity can be linked directly to homeobox developmental genes, so uncovering the evolution of these genes, even at phylum level, provides an immeasurable wealth of information that can be translated in principle. A custom HMM search tool was generated to detect HomeoDB genes in a translated genome assembly of the earthworm: *Eisenia fetida* by Zwarycz *et al.*, (2016). Regions unaligned to the homeodomains in HomeoDB were removed and aligned regions were extracted and generated into FASTA sequences. The same methods were repeated for protein models of other annelids publicly available. BLASTp was used to extract any possible missing homeodomains for all species, given that HomeoDB contains no annelid specific sequences in the data. All homeodomains were aligned and a gene-tree inferred using RAxML (a ML method) for each homeobox class.

Identification of homeoboxes in the gene trees was done by a majority rule. Gene family level losses and gains were observed in the annelids as compared to other non-annelid animals. Results showed a striking expansion in NKL, PRD, and LIM homeodomains for the new earthworm genome, and a collaborative expansion in the HOXL subclass for annelids. Hox3 could not be identified as present in *Eisenia fetida*, despite a finding from Simakov *et al.* (2013) in *Helobdella robusta* (Simakov *et al.*, 2013; Zwarycz *et al.*, 2016) and was deciphered as a loss after divergence in the last common ancestor with *Helobdella robusta* and the Pb Hox gene was lost in the joint lineage of the annelids *Helobdella robusta* and *Eisenia fetida*. Earthworms have a distinct spiralian cleavage to other annelids, the differences in homeobox expansion may be linked to this (Zwarycz *et al.*, 2016).

Highlighted above is just a miniscule fragment of the complex evolution involved in animal developmental genes. The idea that homeoboxes are so highly conserved has led to an oversimplification of ideas; this has caused a lesser comparative interest in developmental genes between phyla. With the ever-increasing number and disparity of species genomes sequenced and annotated, the evolution of homeobox developmental genes has fallen behind a little. Gene loss and gain can differ by the thousands between species, this is a consequence of evolutionary events, and homeoboxes are included in this. Evolution is asymmetrical, with duplication events and differing rates of mutation. Paralogs can diverge from orthologues at different rates, one copy may remain more static as the genome evolves whilst another gene lineage attracts more duplication events and multiple random mutations (Holland *et al.*, 2017). This type of divergence is common in homeoboxes, and has been seen particularly in homeobox families of vertebrates following whole genome duplications (WGD). Lepidopterans are a group of insects that have evolved and developed unique systems for survival, coincidentally, whilst the Hox cluster is relatively conserved among insects, through tandem duplication, lepidopterans have the largest HOX cluster in any animals, overflowing with divergence (Holland *et al.*, 2017). The retention of the Hox cluster makes pinpointing the origin of the genes more simple (Holland, 2013; Holland *et al.*, 2017). Molluscs have a highly divergent and expanded set of TALE homeoboxes and novel PRD gene families that remain unidentified, due to inversions and translocations that can separate tandem-duplicated genes and leave them dispersed throughout genomes (Paps *et al.*, 2015). This pattern is not unique to molluscs and has been observed in different lineages of lophotrochozoans at different times (Holland *et al.*, 2017, Somorjai *et al.*, 2018). There is still a lot left to discover about the homeobox family evolution in lophotrochozoans and other non-vertebrate animals. Somorjai *et al.* (2018) located some expanded homeobox families in molluscs such as the TALE and PRD gene families, but diverse homeobox families were found through the homeobox gene classes. Figure 1.4 illustrates the importance of including all the animal phyla in the

homeobox gene trees when observing the evolutionary development in animal body-plans. Only half the tree constitutes homeobox genes that are well-documented, from a small sample of well-studied taxa, the other half are lesser known, or previously undocumented genes, with lesser represented lophotrochozoan taxa (Somorjai et al., 2018).



Figure 1.4 A ANTP-Hox-like cladogram including previously unexplored lophotrochozoan homeobox genes. In mauve are well classified homeobox genes in the animal kingdom, in orange are the lesser classified and understood lophotrochozoan ANTP families, as adapted from (Somorjai et al., 2018).

GENE TREE ISSUES AND RESOLUTIONS IN METHODOLOGY

These aforementioned studies applied gene-tree methods to resolve a pattern of evolution in the gene families of animals, however, the essential construction of the phylogenetic trees differed. Differing versions of bootstrapping and the different analyses or models formed may contribute to consistency issues in results. ML is generally considered a good evolutionary method to delve into ancestry, whilst neighbour-joining is not, and nor was it intended to be (Lemey *et al.*, 2009). Neighbour-joining simply calculates distances between sequences, which is quicker but less accurate and phylogenetically correct. The use of a more accurate and evolutionarily aware model as considered by Paps and collaborators (2015) can provide more phylogenetically plausible results. The number of taxa and direct taxon sampling used also seems to have an effect on the phylogenetic signal and artefactual issues such as LBA as directed by Pastrana *et al.* (2019), increasing the number of homeobox genes across a larger range of species will likely have a positive impact to reduce these types of artefacts particularly when using parsimony and ML methods.

THE IMPORTANCE OF HOMEBOX EVOLUTION

Animal body plans date back to the Cambrian period, when different locomotion styles for different substrates evolved. Bilaterians evolved sensory organs with a central nervous system and anterior brain, muscle and skeletal systems and a digestive system with an entrance and exit. Homeobox genes of the ANTP Class encode transcription factors patterning body plans. Their presence in all bilaterians supports that these anatomical features were present in the last common ancestor of extant bilaterians, rather than independently evolved in different bilaterian lineages (Holland, 2015).

There is an abundance of research surrounding the metazoan-specific ANTP class, particularly in relation to Hox clusters, but there are knowledge gaps in other homeobox clusters and gene families. As explained before, Hox genes predominantly expanded at the emergence of bilaterians; they are prevalent in chordates and insects. Aside from these Hox clusters, there are many more homeobox clusters unique to or conserved in other animal lineage genomes. Other homeobox clusters include ANTP-class genes such as Hox, ParaHox, NK, PRD-class genes, TALE-class with the Irx cluster and SINE-class with the SIX cluster (Garcia-Fernández, 2005; Takatori *et al.*, 2008; Zhong *et al.*, 2008; Mazza *et al.*, 2010; Gómez-Marín *et al.*, 2015; Simakov *et al.*, 2015; Ferrier, 2016). One of the questions further research may answer is about the uncertainty of the evolution of these clusters, whether the formation of these clusters is primary (conserved from ancestral clusters) or secondary

(re-assortment of genes into lineage-specific clusters) (Ferrier, 2016). Answering this question may fill the gaps of ancient homeobox origins and evolutionary pathways in major animal lineages, but is only possible through expanding the animal genomes compared.

1.4 METAZOA PHYLOGENY AND METHODOLOGY

Phylogenomics has a great potential in resolving the early relationships between animals, which is essential in being able to study evolution and the transition from unicellular organisms to multicellular organisms. The current issue now lies not just in the availability of sequenced data, but in the analysis of it to infer phylogeny.

1.4.1 RELATIONSHIPS IN THE ANIMAL TREE OF LIFE

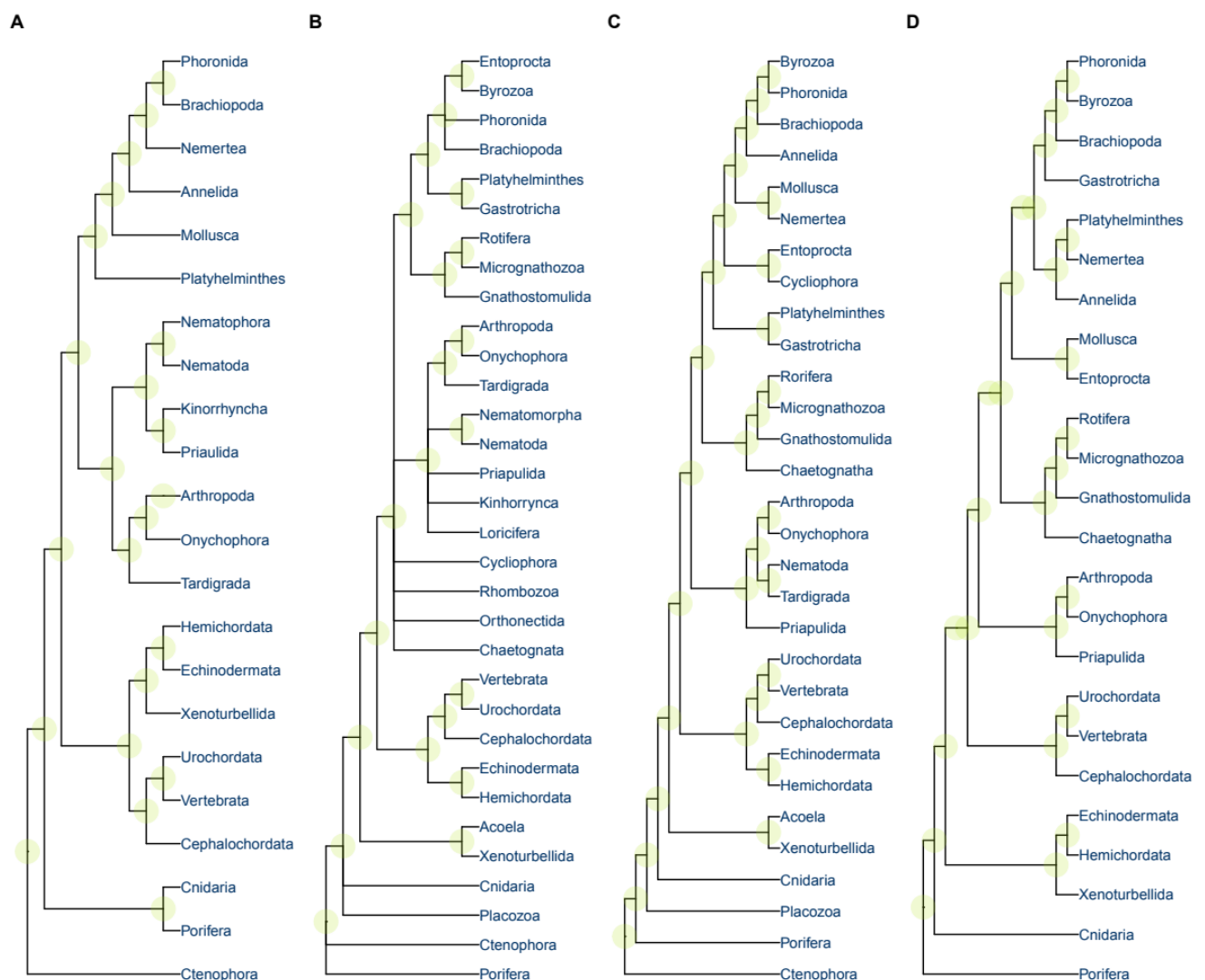


Figure 1.5 Various conflicting topologies as cladograms for the internal relationships in the animal kingdom as summarised in recent research, with circles denoting ancestral nodes. (A) In 2008 this topology was considered very robust with one of the largest taxon samplings to

that point (Dunn *et al.*, 2018). (B) A very frequently agreed upon phylogeny, despite many unresolved nodes displayed as polytomies (Giribet, 2016b). (C) A most recent phylogeny to date using all the available animal phyla possible (Laumer *et al.*, 2019). (D) A differing phylogeny produced in the same year as (C) in attempt to resolve the current polytomies in lophotrochozoans seen in (B) (Marlétaz *et al.*, 2019).

One of the most recent animal phylogenies, by Laumer *et al.* (2019), investigates the relationships between animals using genome-scale data from all phyla (except Orthonectida in which a genome was not available at the time). Discrepancies in phylogenetic interpretations can be attributed largely to the taxon sampling, mostly missing key taxa. Laumer *et al.* (2019) inferred various phylogenies using site-heterogeneous CAT + GTR models to confirm and establish major clades (see Figure 1.5). Through their various phylogenies, they concluded that Lophophorata (Entoprocta, Brachiopoda and Phoronida) (Figure 1.2) are monophyletic and Gnathifera are sister to Chaetognatha, leaving open questions about the nature of the first lophotrochozoan. Controversially they also found clade support for Cnidaria and Placozoa with Ctenophora remaining as sister group to all other Metazoa (Laumer *et al.*, 2018, 2019). Their key points to take away from their research included the importance of orthologue assignment and matrix construction. They constructed a super matrix of 201 opisthokonts, in which 1034 orthologues were selected for being present in at least 100 of those species. This selection was necessary to limit cross contamination, misassigned indexes and horizontal gene transfers. From this super matrix, various sub-matrices were extricated for ML analyses. Each submatrix was pruned and reduced to remove rogue taxa and orthologues (Laumer *et al.*, 2019). The results were very varied topologies. Unanswered questions included: Porifera or Ctenophora as sister to all other animals? Are bilaterians or Placozoa sister to cnidarians? The monophyly of Panarthropoda and Lophotrochozoa and defining the internal relationships of lophotrochozoans.

Prior to the Laumer *et al.* (2019) investigation, Jékely *et al.* (2015) dug into the big 'which animal is sister to all other animals?' question. Nervous systems always enter the discussion, with the shortest path of evolution placing sponges first, and molecular data often supporting ctenophores as the first splitting animal phylum. The molecular data used in phylogenies has been determined to alter the position of ctenophores according to how distant the non-animal out-groups are that are included. The more distant the out-group, the larger the effect of LBA (Nosenko *et al.*, 2013; Jékely *et al.*, 2015). Furthermore, the use of more metazoan taxa with expressed sequence tag (EST) data put ctenophores before sponge. However, removing the more distant out-groups in this setting, in

Bayesian analyses, places sponges before ctenophores (Pisani *et al.*, 2015). A similar problem exists in Bilateria, with the origin of complex eyes (Gehring, 2014). Pisani *et al.* (2015) further disputed that ctenophores were sister to all other animals and disagreed that genomic data does in fact not support that hypothesis when using a more 'correct' methodology. They proposed that Ctenophora is placed incorrectly as first splitting animal only when a phylogeny is incorrectly rooted using inappropriate species as an out-group (Pisani *et al.*, 2015). In contrast, Dunn *et al.* (2008) and Hejnol *et al.* (2009) support ctenophores first (Dunn *et al.*, 2008; Hejnol *et al.*, 2009). A further conflict in the non-bilateral animals came about when Philippe *et al.* (2009) proposed that sponges are monophyletic and not paraphyletic, which diminishes the view that the LCMA had a sponge-like body plan and that ctenophores are sister to cnidarians as a monophyletic clade (Philippe *et al.*, 2009).

When phylogenomic analysis contradicts traditional hypotheses accepted based on morphological features or previously seen molecular results, it is the molecular data that is questioned and not the traditional answer. Commonly seen problems for poor confidence in the major animal relationships are caused by several potential sources of error, which have controversially placed ctenophorans as sister group to other animals rather than the traditional poriferans. These include LBA, too few taxa, poor alignments and biased taxon sampling (Whelan *et al.*, 2015a).

BILATERIAN RELATIONSHIPS IN MAJOR LINEAGES

Only ~25 years to the writing of this thesis, was the monophyly of Protostomia with Lophotrochozoa and Ecdysozoa established using molecular data rather than morphological or embryological knowledge. Previously molecular support was limited in the genomes available or a phylogeny based on a small RNA-subunit. LBA issues have placed platyhelminthes with nematodes where nematodes are included in analyses, but otherwise with the rest of the lophotrochozoans if nematodes are eliminated (Philippe *et al.*, 2005a). With careful species sampling and exclusion of the fast evolving nematodes, a monophyly of ecdysozoans and lophotrochozoans is usually recovered with high support (Philippe *et al.*, 2005a). Protostomia and Deuterostomia clades were recovered with further support in 2008 (see Figure 1.5) (Dunn *et al.*, 2008).

Using Bayesian inference (BI), Paps *et al.* (2009) were unable to recover the monophyly of ecdysozoans, or lophotrochozoans clade and a monophyletic deuterostome clade with convergence. However, with ML approaches, the monophyly of Protostomia with Lophotrochozoa and Ecdysozoa is recovered with high support. Deuterostomia was recovered with slightly less support (Paps *et al.*, 2009). *Xenoturbella*, with its unstable nature, is thought to cause the low support for Deuterostomia,

as it is placed as sister to a strong Ambulacraria in the ML analyses (Paps *et al.*, 2009), and as sister group to all other bilaterians in the BI analyses. Whilst the internal nodes within all three clades were uncertain with lacking congruence across the ML and BI methods, at least the general placement with exclusion to *Xenoturbella* within the major clades of each phyla was supported (Paps *et al.*, 2009).

INTERNAL NODES OF LOPHOTROCHOZOA

Lophotrochozoan topology has remained unstable through the literature in part because of the the sampling of different taxa; this is due to genome availability and the completeness of the genomes that are available. Until recently molecular data for rotifers (Dunn *et al.*, 2008) and other phyla, such as molluscs has been incomplete or of poor quality (Paps *et al.*, 2015). Annelids are usually placed with brachiopods and phoronids, and rarely placed with nemerteans (Luo *et al.*, 2015; Kocot *et al.*, 2017; Luo *et al.*, 2018; Laumer *et al.*, 2019). Molluscs are frequently placed as a monophyletic clade, sister to whichever clade carries both brachiopods and annelids (Luo *et al.*, 2015, 2018; Kocot *et al.*, 2017; Laumer *et al.*, 2019). The relationship between brachiopods, annelids and molluscs is also held up by paleontological evidence in which the mollusc spicules and annelid/brachiopod chaetae are seen as derived from a common ancestor in the form of distinct fossil ‘coelosclerites’ (Dunn *et al.*, 2008).

Aside from lophotrochozoan taxa in which taxon sampling has not been lacking, one issue has been fast-evolving lineages such as platyhelminthes. They appear to attract other fast-clock phyla such as gastrotrichs or rotifers due to LBA, forming a clade named 'Platyzoa', for which there is little morphological or developmental agreement (Dunn *et al.*, 2008). Platyhelminthes have also been placed as 'basal' lophotrochozoans or with Nemertea branching paraphyletically alongside a clade of spiralian instead (Paps *et al.*, 2009).

A very recent study divides Lophotrochozoa into three well-defined sub-clades (see Figure 1.5); Entoprocts with molluscs, a monophyletic brachiopod, phoronid, and ectoproct sub-clade, and a whole new clade including platyhelminthes with annelids and nemerteans (Marlétaz *et al.*, 2019). An additional new clade includes chaetognaths with gnathiferans, rejecting the platyzoan hypothesis. It is clear that the lophotrochozoan internal nodes are totally unresolved.

SMALL UNCERTAINTIES IN DEUTEROSTOMIA

The support for internal nodes of deuterostomes are not as high as one would expect, including the monophyletic status of Chordata (Paps *et al.*, 2009). Urochordates as sister to vertebrates has been well-supported by fossils, however early molecular data supported the cephalochordate-vertebrate

grouping (Philippe *et al.*, 2005; Delsuc *et al.*, 2006). A paper by Delsuc and collaborators (2006) showed instead urochordates as sister to vertebrates (Delsuc *et al.*, 2006). The earlier placement of urochordates as early chordates is believed to be caused by LBA (Delsuc *et al.*, 2006).

Previously associated with platyhelminthes, acoelomorph flatworms are also sometimes placed as deuterostomes with Xenoturbellida (Philippe *et al.*, 2011; Delsuc *et al.*, 2018), contrary to their other possible position as sister group to all other bilaterians (Nephrozoa hypothesis) (Cannon *et al.*, 2016). Similarly, to flatworms and urochordates, acoelomorphs have been shown to have extremely fast rates of evolution, once again causing LBA artefacts. Their position depends on sampling of taxa, genes and evolutionary model, placing them at times as sister to all other deuterostomes, or even as a distant sister group to Ambulacraria (Philippe *et al.*, 2011).

1.4.2 ISSUES AND RESOLUTIONS IN PHYLOGENY METHODS

The idea of using phylogenomics, large sets of markers, is to reduce systematic errors as much as possible. However, phylogenomics comes with its own artefacts. All statistical probability methods make assumptions about the process of evolution in sequences. So far, no single process of evolution fits every sequence. Assuming homogeneity in a single amino acid substitution and a stationary sequence may lead to compositional bias towards species with similar sequence motifs, which may make mathematical statistical sense, but biologically the case is much more complex. Elevated rates of evolution in some species compared to others leads to artefactually grouping these organisms together. Both these issues lead to LBA (Delsuc *et al.*, 2005).

OUTGROUP AND INGROUP SAMPLING

The combination of different models and outgroups has proved to result in different phylogenies. This has been seen in the case of Pisani *et al.* (2015) and the *Mnemiopsis leidyi* genome paper (Ryan *et al.*, 2013). Ryan *et al.* (2013) recovered Ctenophora splitting in the animal Tree of Life before Porifera, although Porifera-first topologies were recovered when the site-heterogenous CAT model was used with close outgroups (e.g., Choanoflagellata); the statistical support was low for both, so these topologies were dismissed (Ryan *et al.*, 2013). Pisani *et al.* (2015) repeated those analyses, with similar results, and then removed *Xenoturbella bocki* to find much higher confidence and convergence in Porifera-first. However, the results remained the same for ctenophore-first when the out-group used more further related opisthokonts (Pisani *et al.*, 2015). The first-splitting animal remains inconclusive,

but the knowledge that the impact model and taxon sampling has is clear and needs to be well considered.

GENE SAMPLING

Traditional molecular data based phylogenomic studies used Small ribosomal subunit RNA gene sequences, otherwise known as 18S rDNA in order to determine bilaterian relationships (Halanych, 2004). The use of this subunit, however, is not immune to the effects of LBA (Halanych, 2015), and not without limitations. To solve this issue other protein-coding genes have been used in phylogenomics studies. This also comes with issues. The inclusion of additional markers has led to asymmetric molecular matrices that may have too many phyla and too small a marker selection, or not enough phyla with too many markers. It is this problem causing concern in the internal nodes of the major bilaterian clades (Paps *et al.*, 2009). Although asymmetrical matrices are not the main causes of concern nowadays. Other problems include the ability to infer paralogy/orthology, the rate of evolution in molecular heterogeneity, and missing data (Delsuc *et al.*, 2005).

MORE ON ORTHOLOGY DETERMINATION

After assembly of the genomes from raw data, bioinformatics pipelines used for phylogenomics are confronted with various issues. Once genes have been predicted, gene homology across taxa needs to be determined, followed by multiple sequence alignment. Orthology determination is one of the trickiest bends in the pipeline and weeding out the paralogues. These methods are commonly gene tree based. Problem genes are removed and then the phylogeny is inferred (Whelan *et al.*, 2015a).

Torruella *et al.* (2015) minimised the orthology determination issue by analysing single-copy protein domains. Due to the nature of the transcriptome data from next generation sequencing, a screening process was also carried out to reduce the possibility of cross contamination, particularly with species that grow with bacteria (Torruella *et al.*, 2015). This study aimed to reconstruct the phylogeny of Opisthokonta, and their results showed convergent evolution between animals and fungi of specialised osmotrophic lifestyles. The results further agreed with current traditional hypotheses with regards to the relationships between Metazoa using their single-copy protein phylogeny approach. Paralogy and horizontal gene transfer were detected manually from inferred single marker trees (Torruella *et al.*, 2015). Any horizontal gene transfer unnoticed could lead to skewed results in the phylogenomic analysis. Horizontal gene transfer gives rise to unexpected homology relationships

and can particularly complicate an evolutionary scenario when looking at convergent evolution and orthology (Gabaldón & Koonin, 2013).

Similarly and previously to Torruella, Nosenko *et al.* (2013) realised that gene selection plays a role in the ability of analyses performed to unravel the evolutionary history of metazoans. Systematic errors constantly assumed in the evolutionary models, that would predictably bring about LBAs and other similar scenarios, were eliminated to reconstruct a reliable phylogeny of animals. Nosenko *et al.*, constructed 3 different phylogenies with support by carefully sampling the genes used in the analysis and the composition of species sampled particularly in the outgroup (Nosenko *et al.*, 2013). As the most distant outgroup, fungi were removed from the alignments to prevent LBA artefacts. (Nosenko *et al.*, 2013). Presence and absence in the gene selection also lead to a bias. Often genes that are lost in all but one species are excluded from analyses, and genes present in outgroups but lost internally are also eliminated, creating a downwards bias (Pisani *et al.*, 2015).

It was concluded that animal phylogenetic trees can be more stable using a realistic amino acid substitution model which accounts for both biochemical patterns and evolutionary rates, although a particular model was not identified (Lartillot & Philippe, 2004; Nosenko *et al.*, 2013; Pisani *et al.*, 2015; Kocot *et al.*, 2017). The careful selection of representatives of each of the animal lineages also makes a difference, such as ensuring that non-bilaterian taxa comprise a good poriferan representation (including calcareous and homoscleromorph sponges), and at least one ctenophoran be used. Finally, that newly sequenced and more genomes of the earlier non-bilaterian animals may go a long way in successfully resolving the relationships between these early splitting animals (Nosenko *et al.*, 2013).

1.4.3 TREE INFERENCE ALGORITHMS

Central steps in phylogenomic inference are the selection of homologues across species, and the reconstruction of the evolutionary changes between homologues to construct a tree depicting the evolutionary relationships. There are 3 approaches to phylogenetic reconstruction: genetic distance such as neighbour-joining algorithms, maximum parsimony which minimises the number of evolutionary steps, and probabilistic, which is a statistical calculation of probability using algorithms such as ML and BI (Delsuc *et al.*, 2005). ML and BI have become very popular with the elevation of phylogenomic datasets.

BAYESIAN INFERENCE (BI) AND MAXIMUM-LIKELIHOOD (ML)

Combining prior probabilities, Bayesian methods derive the distribution of trees according to their posterior probability (values which give the probability of the tree based on prior probability, likelihood function, and the data), using Bayes' mathematical formula. This approach relies on a very specific and sophisticated model, which can potentially lead to high statistical support for an incorrect tree (Delsuc *et al.*, 2005).

ML selects the tree that maximizes the probability of observing the data under a given model through optimizing model parameters and finding the highest peak/most probable observed in the parameter space, such as GTR+ Γ +I (GTR-generalised time model, Γ -gamma distribution, I-proportion of invariable sites). For phylogenetic inference this approach is thought to be more robust than BI (Delsuc *et al.*, 2005).

Both BI and ML methods have distinct advantages and disadvantages, and both methods are often used to infer phylogenies using the same data as a measure of support for that phylogeny (Whelan *et al.*, 2015; Laumer *et al.*, 2019). In certain phylogenies, BI has proven a bias towards LBA, using the same simple evolutionary conditions, whilst ML is more efficient and requires lesser internal branch lengths to recover the same true topology. Similarly, given the nature of BI, it is also much more sensitive to model violation, stemming as both an advantage to the BI method and disadvantage. ML is less susceptible to this violation, but it is also less flexible to the use of combined and specific models (Philippe *et al.*, 2005b; Kolaczkowski & Thornton, 2009).

1.4.4 EVOLUTIONARY MODELS

CAT, GAMMA DISTRIBUTION, RECODING

The evolutionary model used in phylogenomic analyses is key to discrepancies seen in the literature. This is because there is no single evolutionary model which accounts for every possible amino acid substitution.

When site specific amino acid substitutions are modelled, Porifera are sister to all other animals, whereas in the same dataset, without site specific amino acid substitutions ctenophores are preferred, with lower support (Feuda *et al.*, 2017). Data recoding can be used to reduce compositional heterogeneity and allow for a more flexible and data-specific evolutionary model to be adapted (Feuda

et al., 2017). Data-recoding means reducing the effect of saturation by binning the nucleotides or amino acids into fewer groups, such as 6-Dayhoff categories. This process of grouping amino-acids with similar physiochemical categories reduced the noise seen in both rapid and slow-evolving substitutions (Susko & Roger, 2007; Feuda *et al.*, 2017).

CAT-GTR+G (category mixture model with gamma distribution) has been found to be the most sophisticated model, although for individual datasets other models are actually preferred (Feuda *et al.*, 2017). Whilst data-recoding has been shown to have an impact on the first-splitting animal topology, it hasn't necessarily had an impact on other phylogenetic relationships, only proven in testing the eukaryote orthologues defined by BUSCO (Benchmarking Universal Single-Copy Orthologues) (Simão *et al.*, 2015; Laumer *et al.*, 2018).

1.4.5 OUR CURRENT LESS-THAN-STABLE PHYLOGENY

Key papers in the field all agree Metazoa are sister group to Choanoflagellata (Torruella *et al.*, 2012; Paps *et al.*, 2013; Suga *et al.*, 2013), and within the metazoan clade one of the first-splitting animals is, most often Porifera, and almost equally frequently, Ctenophora (Philippe *et al.*, 2009; Srivastava *et al.*, 2010; Ryan *et al.*, 2013; Moroz *et al.*, 2014; Whelan *et al.*, 2017). Cnidaria are nearly always sister to Bilateria. And Bilateria split to form Deuterostomia and Protostomia. The exact relationships between phyla are uncertain beyond this point with disagreements between the morphological and molecular data and the history of gene families contradicting these also. The future recommendations in each phylogenetic analysis to tip the scales in any one direction and resolve the animal tree of life is always careful taxon sampling, clever mathematical models, and strict gene selection.

1.5 THESIS AIMS

In recent literature there has been a plethora of data and research into the origin and relatives of animals, with more in-depth research into the latter. Results have differed between studies due to three main reasons: the diversity of taxa considered, the variation in gene families and their definitions, and the actual analysis performed for reconstruction of the phylogenetic relationships.

With the decreased costs of next generation sequencing there are now more genomes sequenced than ever and they are always increasing in number. Now that there is a greater diversity and number of genomes available, this provides opportunity to reconstruct relationships within the Animal

Kingdom with an improved taxonomic sampling, providing more accurate results with fewer analytical artefacts. High computer power is also growing and the availability and requirements needed to process the huge amounts of genomic data cheaper and more possible. This also means that more accurate, but also more computationally intensive analyses, can be used to unravel the origin of animals. Now we can take advantage of this wealth of information and produce a comparative genomics overview of each of the major animal clades and their internal nodes to phylum classification level.

This thesis aims to tackle the current limitations in modern evolutionary studies to improve our understanding of the evolution of the Animal Kingdom by using as diverse taxon sampling as is possible among in- and out-groups whilst utilising a powerful in-house-developed bioinformatics pipeline.

Within comparative genomics, the extraction of ancestrally conserved homology groups, lost genes and genes novel to different lineages can be used to reveal patterns in the evolution and the genomic make-up of the different LCAs of animal lineages. In the next chapter of this thesis (chapter 2) I used over 100 eukaryote whole genomes, with around 2.6 million canonical proteins to achieve this task, a very competitive sampling for whole animal genome comparative genomic analysis. Moreover, a robust gene homology assignment approach can be the base to conduct more reliable gene family evolution studies (e.g., the homeobox genes) as well as more refined phylogenomic analyses.

Homeobox genes superfamily show high diversification in animals. Our larger taxonomic sampling could provide further understanding of the role of these genes in the evolution of animals. Gene trees with a dispersed spectrum of samples would produce a more detailed history of phylogeny, continuing on from the research already mentioned. Gene trees mapping the relationships of gene families/superfamilies may also reveal a pattern in the evolution of animals. In this project, I will analyse the largest selection of homeobox genes (over 8000 in 59 animals) across a broad selection of animals to date (chapter 3); I will test if they will contribute to the complex evolutionary knowledge already obtained in the gain and loss of genes as it reveals a map of the various developmental body plans routes each animal lineage has taken.

The use of molecular data for phylogenetic inference is key to understanding the relationships between the animal lineages, with careful consideration of the appropriate model and the method used as necessary. Using the most appropriate model for each dataset and ML methods should meet the

criteria for a robust and well-resolved phylogenetic topology, with a BI supporting backup. In the final results chapter of this thesis (chapter 4), I take a subset of the HGs that are ancestrally well-conserved among all the animal and other eukaryote genomes, as defined in the first results chapter. These HGs need to meet certain criteria to ensure they are ancestral and well-conserved, such as being orthologous or as close to single-copy as possible, therefore reducing a high range of evolutionary rates leading to LBA or any other undesired artefacts. I produce several protein matrices using these HGs with which to infer a reliable phylogeny.

Exploring the diversity of animals used in this thesis by linking together three distinct evolutionary analyses will provide an integrated view of the lineage specific morphological, molecular and developmental evolution.

It is expected that an increase in the number of disparate animal genomes and outgroups will produce a more accurate phylogeny for the metazoan tree of life, whilst reducing analytical artefacts such as LBA. That the gene loss/gain will reveal patterns of gene repertoire specificity for individual Phyla and the genome of the LCMA. Finally, that taxon-dense gene trees for highly diverse gene families found within animals will also reveal a pattern of functionality and morphological evolution as well as on a molecular level.

1.6 MOTIVATIONS AND RESEARCH QUESTIONS

Animals are incredible sentient life forms, and genomics codes for the mysterious blueprints of these sentient beings. There will always be incentive to delve into the intricate relationships between animals, because humans are animals, and we are humans. As science progresses alongside advancements in technology, the wealth of genomic data is becoming greater, and the quantity of data will always be more than there are bioinformaticians to analyse this high throughput. Here I have focussed on the diversification of animal phyla, with genome comparisons and patterns of evolution. The applications of the findings here in this thesis, whilst at this time considered a robust taxon sampling, can be upscaled even further. There were three main components in this thesis:

- i. Diversification of specific animal lineages by loss and gain of homology groups: Comparative Genomics of 102 eukaryote genomes from 16 different animal phyla, and 43 non-animal species. This bioinformatics pipeline was developed to form a sturdy and in-depth description of the genetic toolkits contained in phyla defined genomes and evaluate the loss and gain of homologous protein groups within extant animal clades.
- ii. Evolution of body plans in animal diversification: Homeobox gene trees to analyse the evolution of homeobox-related body plans using the well conserved homeodomains in the 102 eukaryote genomes. Here I used the resulting HGs from the comparative genomics pipeline I developed and known annotations within the dataset to extract all the HGs containing homeobox proteins. Gene trees were inferred for over 8,000 homeobox proteins from the 102 eukaryotes in 30 HGs, with one large gene tree containing all 8000. There were distinct homeobox Classes and Families in each HG.
- iii. Important animal relationships in the emergence of animal phyla: A phylogeny of the 102 eukaryote genome dataset using conserved HGs, extracted from the bioinformatics comparative genomics pipeline to infer an animal tree of life for the 16 animal phyla.

1.7 THE BIOINFORMATICS PIPELINES

All three components of this thesis are based on HGs inferred using the same bioinformatics pipeline. The initialisation of this pipeline to carry the three topics makes linking the results together simple. The most important feature here is the definition of the HGs. With such a diverse and as substantial as was available in 2016 taxon sampling, using an expect-value strict BLAST all vs all means of gathering similarity and then MCL on those statistics produced a very healthy and well defined matrix of HGs.

With corroboration from ROC analysis against the BUSCO results in the dataset used here, I am confident that the HGs defined can be taken forward and applied to many other comparative genomic analyses. Some of the resulting HGs have already been mined from the database to be used by others in the scientific community. There is little limitation to the use of these HGs in use with other animal genome studies, aside from the phyla that were not in the dataset used. The pipeline can be applied with any set of organisms following the instructions in the GitHub repository, for example, plants, fungi, choanoflagellates, etc.

Provided there is some annotation in at least one species, HG sequences of a specific function, cellular component, biological process and/or protein family can be extracted for detailed analysis. This powerful database tool was used to extract all the homeobox protein families for all the genomes in the dataset, and to extract well-conserved genes in animals to infer the relationships between animals in a phylogenetic tree.

The bioinformatics pipeline developed here, with instructions and scripts publicly available on GitHub (see appendices) can be adapted to any number of whole genome species from any lineage with ease. The only limitations in this project are the currently available protein models used in phylogeny methods and the variety of animal genomes publicly available. In this thesis, the results produced are novel and open up unlimited scope for further research. Up to this date I have used an unrivalled and unprecedented dataset of whole genomes, covering as many phyla that were available at the time I started to write.

2 USE IT OR LOSE IT: WIDESPREAD PATTERNS OF GENE LOSS IN THE EVOLUTION OF ANIMAL GENOMES

2.1 SUMMARY

The Animal Kingdom has an astonishing diversity, the product of over 550 million years of animal evolution. The current wealth of genome sequence data offers an opportunity to better understand the genomic basis of this disparity. Here we analyse a sampling of 102 whole genomes including >2.6 million protein sequences. We infer major genomic patterns associated with the variety of animal forms from superphylum to phylum level. We show a remarkable amount of gene loss that occurred during the evolution of two major groups of bilaterian animals, Ecdysozoa and Deuterostomia, and further loss in several deuterostome lineages. Deuterostomes and Protostomes also show large genome novelties. At the phylum level flatworms, nematodes and tardigrades show the largest reduction of gene complement, alongside gene novelty. These findings paint a picture of the evolution within the Animal Kingdom in which reductive evolution at protein-coding level played a major role in shaping genome composition. This chapter has been peer reviewed and published in *Nature Ecology & Evolution* (Guijarro-Clarke et al., 2020).

2.2 BACKGROUND

The Metazoa encompass an astonishing diversity of body forms. More than 30 animal phyla have been defined, the evolutionary relationships between which are well understood in broad outline, although some are still a matter of debate (Egger *et al.*, 2015; Jékely *et al.*, 2015; Giribet, 2016a; Marlétaz *et al.*, 2019). Understanding how their genomes have evolved can help us to better comprehend the origin of this disparity and reconstruct their evolutionary history. The Metazoa comprise sponges, ctenophores, cnidarians, placozoans, and bilaterians, with most of animal diversity found in the last of these. The Bilateria can be split into three major groups or superphyla — Deuterostomia, Lophotrochozoa and Ecdysozoa (Halanych, 2004), the latter two forming the Protostomia.

Gene losses and gains play major roles in evolution. The gain of new functions via assembly of modules from older genes or emergence of *de novo* coding regions has been proposed to be important during major evolutionary transitions such as the origin of animals (Grau-Bové *et al.*, 2017; Paps, 2018; Richter *et al.*, 2018). Gene loss has been associated with loss of anatomical structures in evolution, consistent with views that evolution can lead to both increases and decreases in complexity (Lankester, 1879; Denoeud *et al.*, 2010; Tsai *et al.*, 2013). Previous studies have shown the prevalence of gains and losses of genes and protein domains in the dawning of different groups of animals (Denoeud *et al.*, 2010; Moore & Bornberg-Bauer, 2012; Zmasek & Godzik, 2012; Tsai *et al.*, 2013; Albalat & Cañestro, 2016). The increasing availability of genome data is giving new opportunities to investigate animal genome evolution. For example, recent analyses have shown the importance of using large taxon sampling and a range of outgroups to reconstruct the minimum protein coding genome present in the ancestor of a clade (Dunwell *et al.*, 2017; Paps & Holland, 2018; Richter *et al.*, 2018).

2.3 DISCUSSION & RESULTS

To analyse origins of genes during the early evolution of mammals and Metazoa, a bioinformatics pipeline was introduced that used state-of-the-art methods of homology assignment (Dunwell *et al.*, 2017; Paps & Holland, 2018). The approach focuses on protein-coding genes, but other genomic elements (non-coding RNA genes, regulatory regions, transposable elements, etc) most likely also contributed to the diversification of metazoans. Here we apply a new, more flexible and efficient, version of this pipeline to a large collection of metazoan genomes (59 genomes from 16 animal phyla) and a greatly expanded representation of outgroups (43 genomes) specifically to investigate patterns of gain and loss of genes across major lineages of bilaterian animals. The use of complete genome sequences is particularly key to determining gene loss, since its inference from incomplete data sources is problematic.

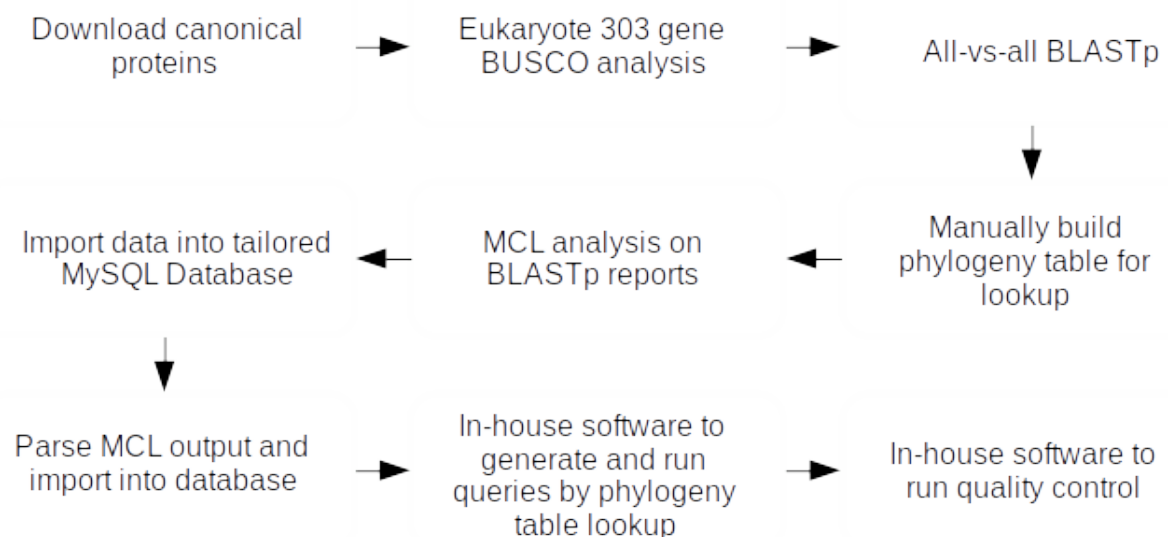


Figure 2.1 Flow representation of the bioinformatics pipeline and each key step for the analysis/generation of novel and lost HGs.

The pipeline developed is outlined in Material & Methods and Figure 2.1. Briefly, we assembled a dataset of 102 previously sequenced eukaryotic genomes (Figure 2.2), chosen for their phylogenetic position and quality as assessed using BUSCO (Kriventseva *et al.*, 2015; Simão *et al.*, 2015) (Figure 2.3). Over 2.6 million proteins were compared using a reciprocal BLASTp (Camacho *et al.*, 2009) (all-vs-all), and clustered with MCL (Enright *et al.*, 2002) into homology groups (HG). An HG is a group of protein-coding genes that differ from others consistently, independently of their

mechanism of origin (divergence, de novo origin etc). The extent of gene misassignment in HG clustering was assessed using metazoan and eukaryotic BUSCO gene sets for benchmarking, as well as performing receiver operating characteristic (ROC) analyses (Figure 2.4 and 2.5).

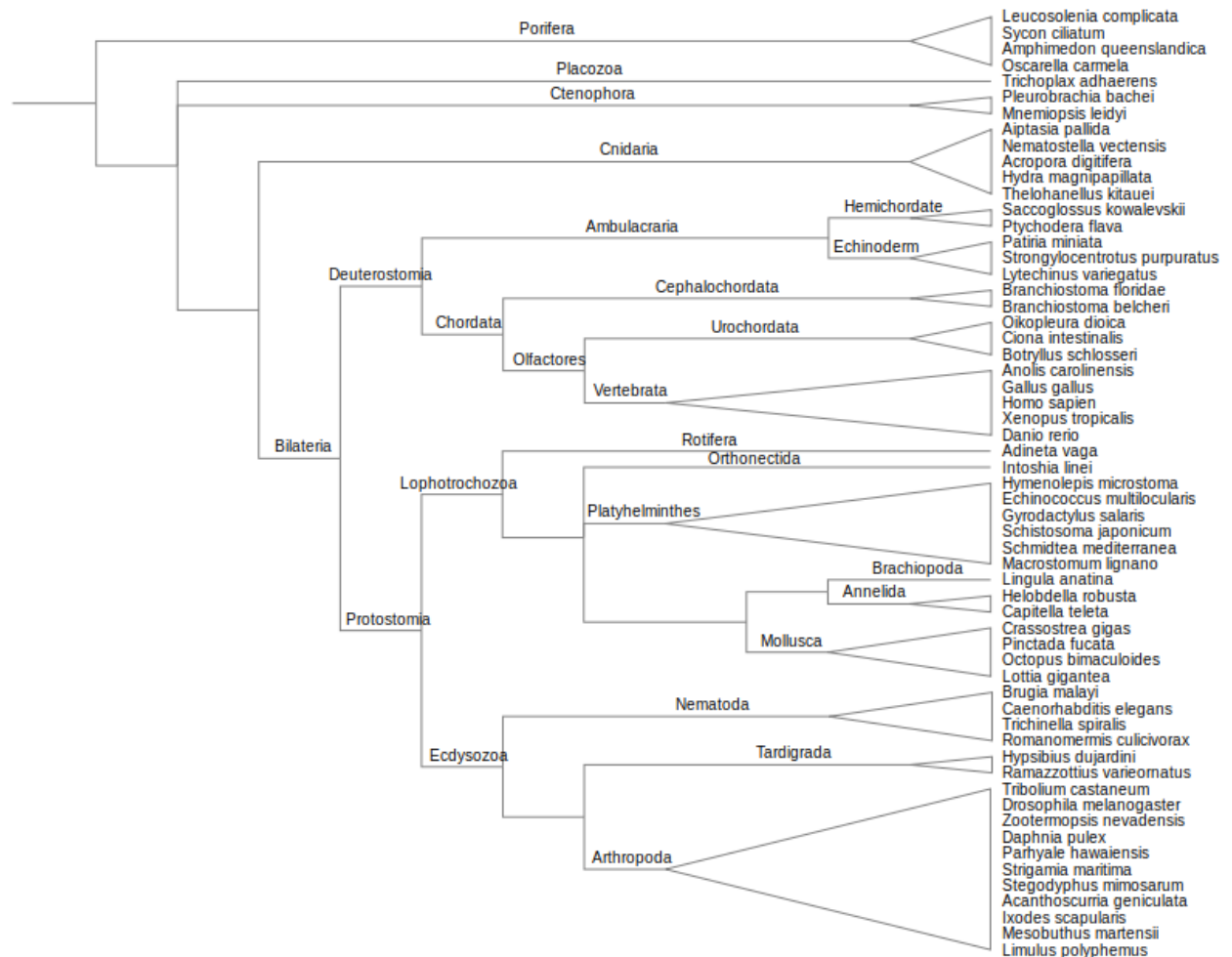


Figure 2.2 Phylogenetic map for each of the animal genomes used in the bioinformatics analysis. Each of the main clades labelled for clarity, these are the clades used to infer HGs. As noted in the materials and methods, this is a consensus phylogeny taken from multiple literature sources.

The HG were analysed in a MySQL database, which tabulates all species in the study classified following phylogenetic relationships (Figure 2.2). For each node of the phylogenetic tree, MySQL will find HG that are gained or lost by combining taxon presence/absence in each member of that clade. For example, an HG present only in species from the clade Vertebrata is considered a vertebrate novelty (Paps & Holland, 2018) (Figure 2.6). ‘Novel HG’ (denoted +) are sets of related genes that emerged in the stem lineage or last common ancestor (LCA) of an ingroup, ‘Core novel HG’ (++) are

Novel HGs highly retained in the ingroup (refractory to gene loss), ‘Lost HG’ (denoted -) specify HGs lost on the stem lineage of the clade (prior to the LCA), and ‘Core lost HG’ (--) are Lost HG that are highly retained in outgroup taxa. We propose that both categories of ‘core’ HG perform essential biological functions in the groups they are found, underpinning their preservation. However, the values of Core Novel HG are also affected by the number of genomes included in a clade (e.g. a clade with two genomes will display higher values of Core Novel HG than a clade with 10 representatives), and their evolutionary relatedness (clades composed by closely related genomes will show higher proportions of Core Novel HG. Core Novel HG for all nodes were further validated by BLASTp against the RefSeq database (Pruitt *et al.*, 2007). In the case of phyla with a single genome sampled (e.g. rotiferans, orthonectids, brachiopods), HG values may not be representative of the group; these values are not shown although these genomes are still important to infer HG categories in sister groups.

Figure 2.6 shows the numbers of Novel HG (+), Core novel HG (++), Lost HG (-) and Core lost HG (--) inferred for major evolutionary branches across the Bilateria. Values in the LCA of Metazoa are consistent with previous work (Paps & Holland, 2018), with minor discrepancies explained by expanded taxon sampling; for example, the number of metazoan Novel HG (+25) remains the same. Looking at the patterns of gene gain, Bilateria show the largest number of Novel HG (+1699) among the major animal clades indicating extensive origin of novel gene types. Bilaterians are characterised by a centralised nervous system with anterior elaboration (a ‘brain’). Consistent with this morphological change, our results reveal nervous system functions amongst novel bilaterian genes including genes encoding a diversity of neuropeptide receptors (e.g., orexin, neuropeptide FF and neuropeptide Y) and transcription factors (oligodendrocyte transcription factor 3, protein turtle and homeobox protein prospero). Despite the high levels of gene novelty, no Core novelties were detected suggesting genomic flexibility after the origin of Bilateria.

Further gene novelty is inferred in the evolutionary lineages leading to protostomes (+734) and deuterostomes (+280); within protostomes, lower novelty is detected on the stem lineages of lophotrochozoans (+60) and ecdysozoans (+97) nodes (Figure 2.6). Our sampling does not include representatives of Scalidophora (Priapulida, Loricifera, and Kinorhyncha) which are the sister group to the ecdysozoans sampled in this study. Low numbers of Core Novel HG are seen in the other major bilaterian clades (Figure 2.6 and Table 2.1). At the phylum level (Figure 2.7), particularly high levels of phylum-specific novelty are found in flatworms (+856), nematodes (+1187) and tardigrades (+945); the latter HG are all shared between two tardigrade genomes, including a version of *Hypsibius*

dujardini annotated excluding potential contaminations (Arakawa, 2016; Yoshida *et al.*, 2017). Many vertebrate novelties are related to immunity and signalling pathways. Figure 2.8 shows the most abundantly gained and lost molecular functions assigned by gene ontologies (Carbon *et al.*, 2017) (GOs) across all clades (Figure 2.8); however, we caution against over-interpretation of these since there is a bias in the quantity and quality of GO annotations between organisms. Core novel HG show GPCRs, receptors, and nucleic acid binding as some of the functions gained more often across clades (Figure 2.8A). Most clades show a broad spread of GO functions gained, while others concentrate gains in a few (e.g., echinoderms gained GPCRs and transporters, panarthropods gained transfer/carrier proteins and receptors).

Table 2.1 Novel Core HG for the major groups of bilaterian animals. HG column is a unique identifier to find the corresponding proteins in the same homologous set of proteins in the database, can also be found in the data provided in the appendix.

Group	HG	Protein	Molecular function	Uniprot ID
Protostomia	2367	<i>Drosophila melanogaster</i> : Rad-	GTP binding; GTPase	Q6NN22
		Gem/Kir	activity	
Ecdysozoa	10452	<i>D. melanogaster</i> :	Immunoglobulin-like	Q9VTG8
		Uncharacterized protein	domain	Q8SWY5
	10748	<i>Tribolium castaneum</i> :	Putative G-protein	D6WGI4
		Uncharacterized protein	coupled receptor	
Lophotrochozoa	2871	<i>D. melanogaster</i> : Morpheus	Regulation of embryonic	Q9V9X1
		(mey)	cell shape	
	6825	<i>Crassostrea gigas</i> :	Dopamine and orexin	K1QZE1
		Uncharacterized protein	receptors	K1QGB7
Deuterostomia	17100	<i>C. gigas</i> : Uncharacterized	Structural constituent of	V4B0P9
		protein	ribosome	
	8167	<i>Homo sapiens</i> : Sialidase-4;	Protein binding; Exo-	Q8WWR8
	12028	<i>H. sapiens</i> : Kremen protein 1	alpha-sialidase activity	
	13942	<i>H. sapiens</i> : Protein INCA1	Protein binding	Q96MU8
	14028	<i>H. sapiens</i> : 39S ribosomal	Cyclin binding	Q0VD86
protein L1, mitochondrial		RNA binding		
11983	<i>H. sapiens</i> : Glycoprotein	Hormone Activity	Q86YW7	
		hormone beta-5		



Figure 2.3 BUSCO analysis results using the 303 eukaryote genes dataset. The threshold for each of the animal genomes was to have less than 15% missing genes. Where a single taxon had more than 15% genes missing, it was accepted if sharing an analysed clade with a genome meeting the completeness criterion.

Gene loss shows a particularly interesting pattern. We deduce that very extensive gene loss, in excess of 1000 HG, occurred on the stem lineages of each of the three major bilaterian superphyla: Ecdysozoa (-4677), Lophotrochozoa (-1760) and Deuterostomia (-4231) (Figure 2.6). These values are in excess of the amount of gene novelty, suggesting that loss of genes or gene functions was important in shaping the distinctive biological characters of these clades. Similar patterns are not seen in the bilaterian node, where novelty is deduced to be more dominant in genome evolution (+1699 vs -745). The loss of bilaterian genes in the ecdysozoan lineage has been pointed out in previous studies (Simakov *et al.*, 2013; Luo *et al.*, 2018). The HG lost in ecdysozoans include several membrane proteins and signal transduction components; deuterostomes lost HGs include functions related to transmembrane proteins such as those that form gap junctions in invertebrates. Many phyla within these groups also show high degrees of HG loss, with many losses being of genes otherwise highly retained in outgroups (Core Lost HG): echinoderms (--680), urochordates (--845), nematodes (--401), tardigrades (--967), flatworms (--858), and annelids (--3179) (Figures 2.6 and 2.7). Occasional examples of convergent gene loss are detected, such as protein LEG1 homologue involved in multicellular development and small ubiquitin-related modifiers (SUMO), both lost in echinoderms and urochordates. Among molecular functions more often lost (Figure 2.8B) are transfer/carrier proteins, ribosomal proteins, and nucleic acid binding proteins; there are also differences between clades, with urochordates, ambulacrarians and tardigrades losing a genes with very diverse GO classifications.

Here we used a comprehensive taxon sampling together with comparative methods to infer the patterns of gene gains and losses of ancestral animal genomes. Our analyses support a major role of gene novelty in the origins of animals and bilaterians, consistent with origin of new biological characters, but in contrast we also deduced there was an exceptional amount of gene loss on the stem lineages of the major bilaterian supergroups: Ecdysozoa, Lophotrochozoa and Deuterostomia. Further gene loss occurred in the evolution of phyla within these groups, although in some cases loss seems largely balanced by novelty. The three animal phyla with the largest levels of gene loss - flatworms, nematodes, and tardigrades - also show remarkable levels of genomic novelty. This pattern could be explained by high gene turnover in the genome of their respective ancestors. Alternatively, it may be influenced by interaction between their biology and our methodology: these are 'fast-evolving' lineages, thus some of their genes may be highly divergent and have formed their own clusters. Gene loss has been suggested as an important force in the evolution of different groups of organisms, including in Metazoa and Fungi (Albalat & Cañestro, 2016). This study highlights the importance of

rich taxon sampling to understand the evolution of animals and sheds new light on the part that reductive evolution of gene complements has played in evolution of animal diversity.

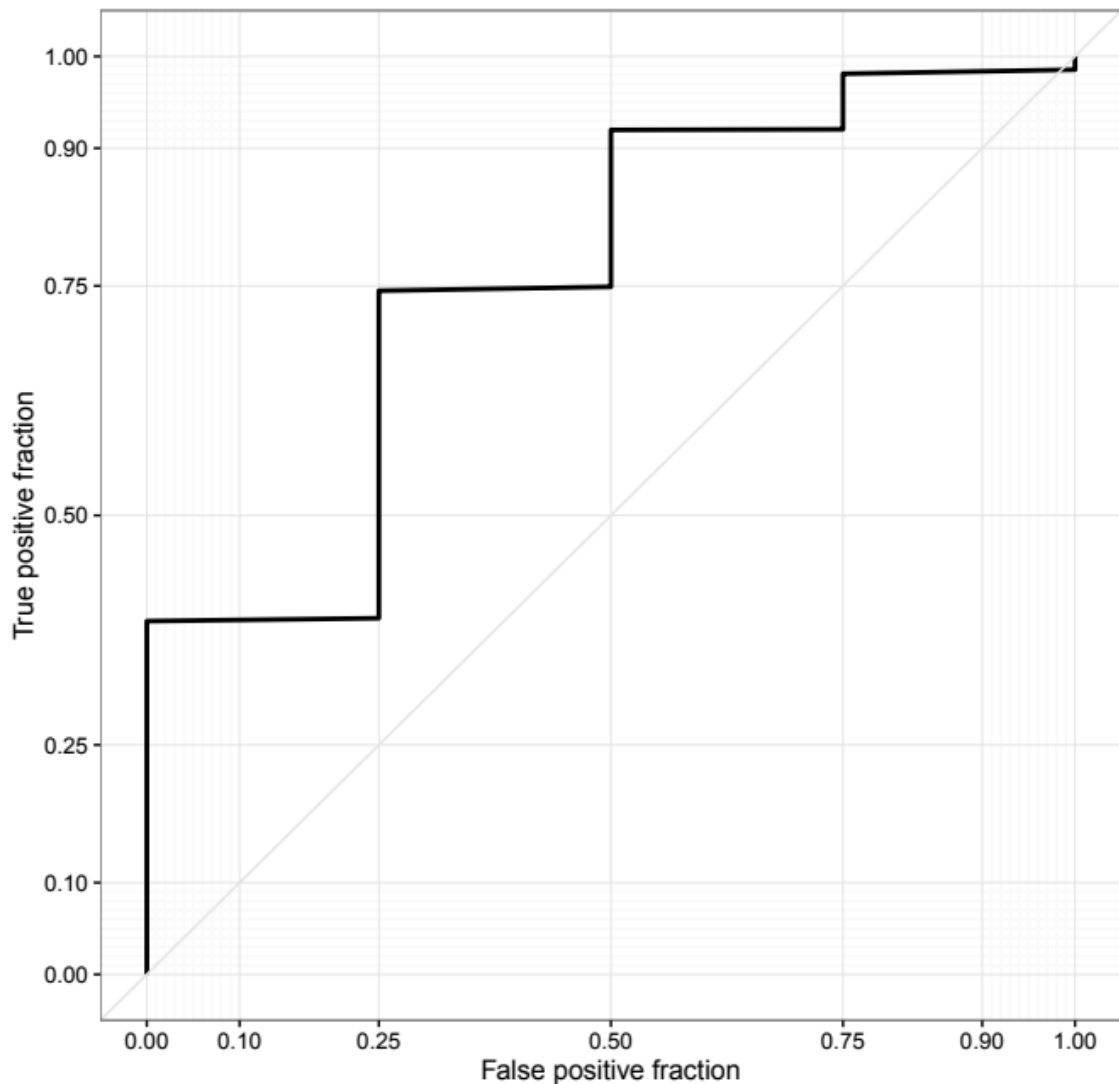


Figure 2.4 Receiver operating characteristic (ROC) curve for the HGs in this study matching the eukaryote BUSCO gene sets. ROC produces a graphic that illustrates the diagnostic ability of a classifier system by plotting the true positive rate against the false positive rate. Graphs above the diagonal indicate a high proportion of true positives vs false positives. To quantify the amount of misassignment in our HG, we compared our clustering against the eukaryote and metazoan gene sets of BUSCO. BUSCO sets were mined from OrthoDB, which is based in Best-Reciprocal-Hit (BRH) BLAST. In contrast, our pipeline combines BRH plus Markov Clustering. BUSCO datasets contain single copy orthologs present in at least 90% of the species. The species sampling used to define the BUSCO gene sets are different to ours. The percentage of BUSCO genes misclassified in our pipeline was quantified. For the 303 orthology groups of the BUSCO eukaryotic dataset, the error rate of the assignments is

0.085%. We have an additional 80 eukaryotes that BUSCO do not have, 4 OrthoDB markers diverged beyond sequence similarity recognition, divided into two HGs according to major lineages. This is likely due to our extended dataset.

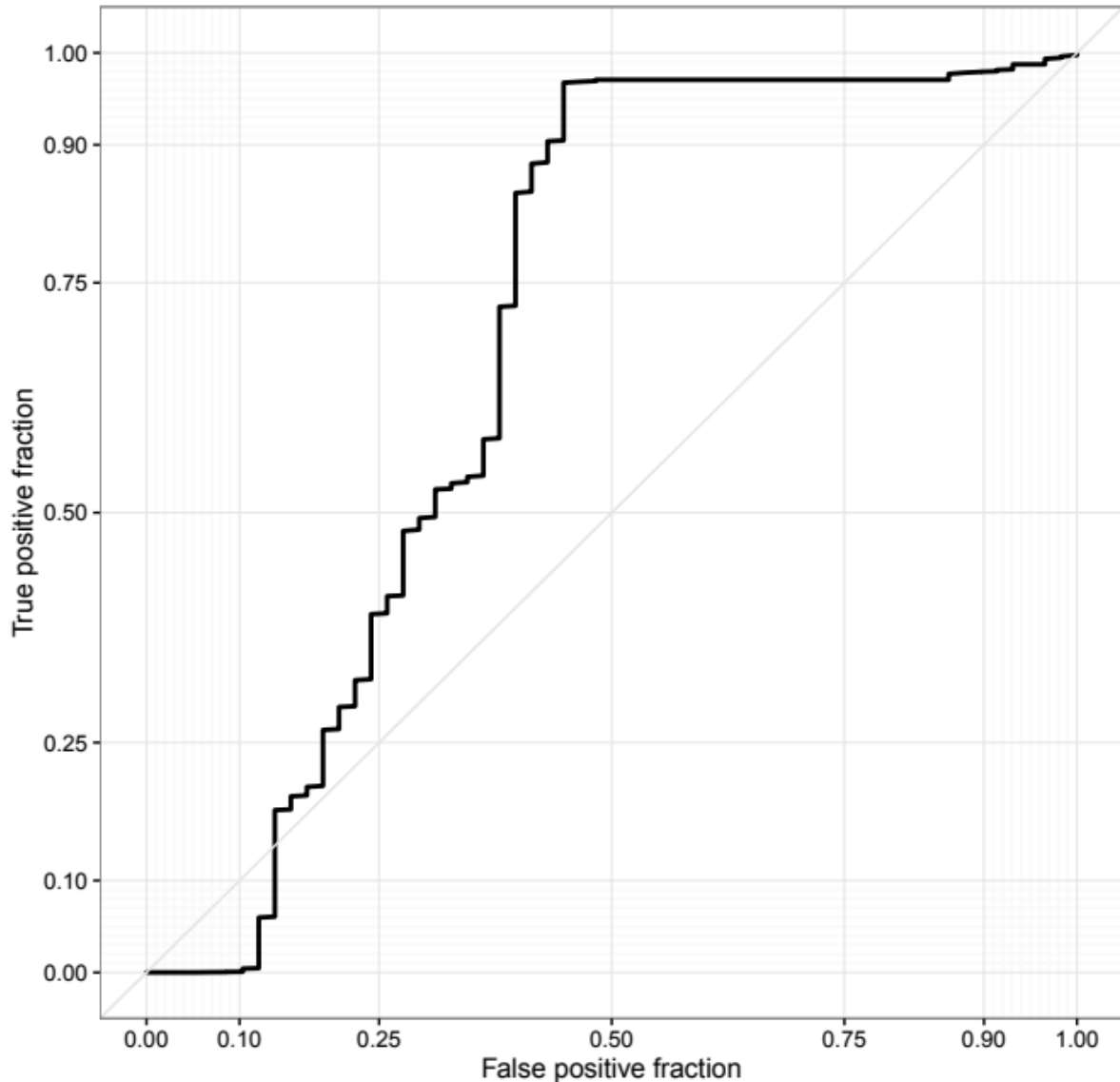


Figure 2.5 ROC curve for the HGs in this study matching the metazoan BUSCO gene sets.

Similarly to the eukaryote test, to quantify the amount of misassignment in our HG, we compared our clustering against the metazoan gene sets of BUSCO. For the 978 orthology groups of the metazoans BUSCO, the error rate is 0.25%. We have an additional 38 animals that BUSCO do not have, 17 OrthoDB markers diverged beyond sequence similarity recognition, divided into two HGs according to major lineages. This is likely due to our extended Supplementary Data et with less bias towards vertebrates and a larger selection of non-arthropod protostomes and non-vertebrate deuterostomes.

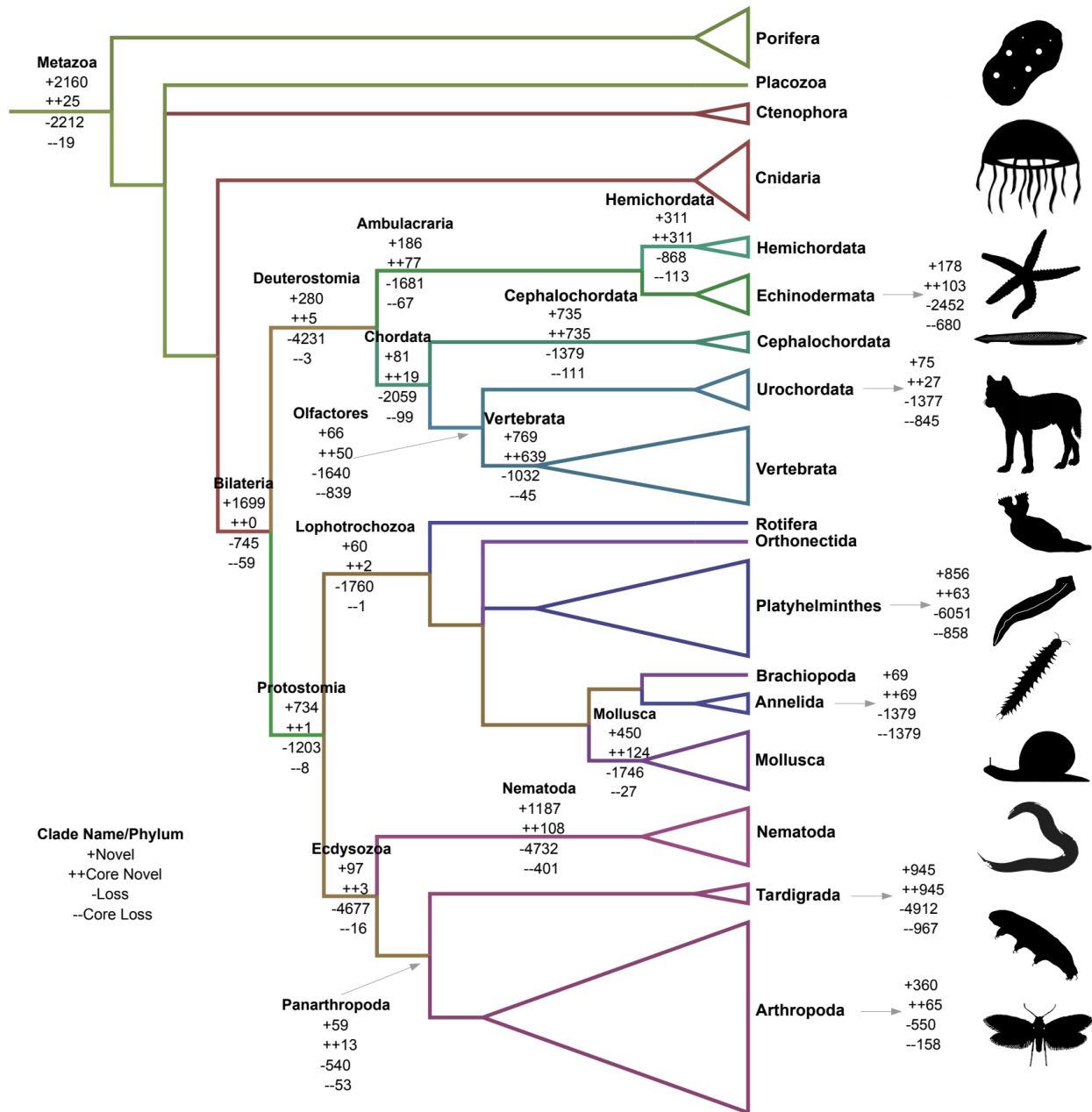


Figure 2.6 Reconstruction of ancestral genomic gains and losses in the Animal Kingdom.

Evolutionary relationships of the major groups included in this study (Halanych, 2004; Laumer et al., 2015; Kocot, 2016). Different categories of HG are indicated in each node, from top to bottom, Novel HG (+), Core Novel HG (++), Lost HG (-), and Core Lost HG (--). Organism outlines from phylopic.org and from the author (submitted to phylopic).

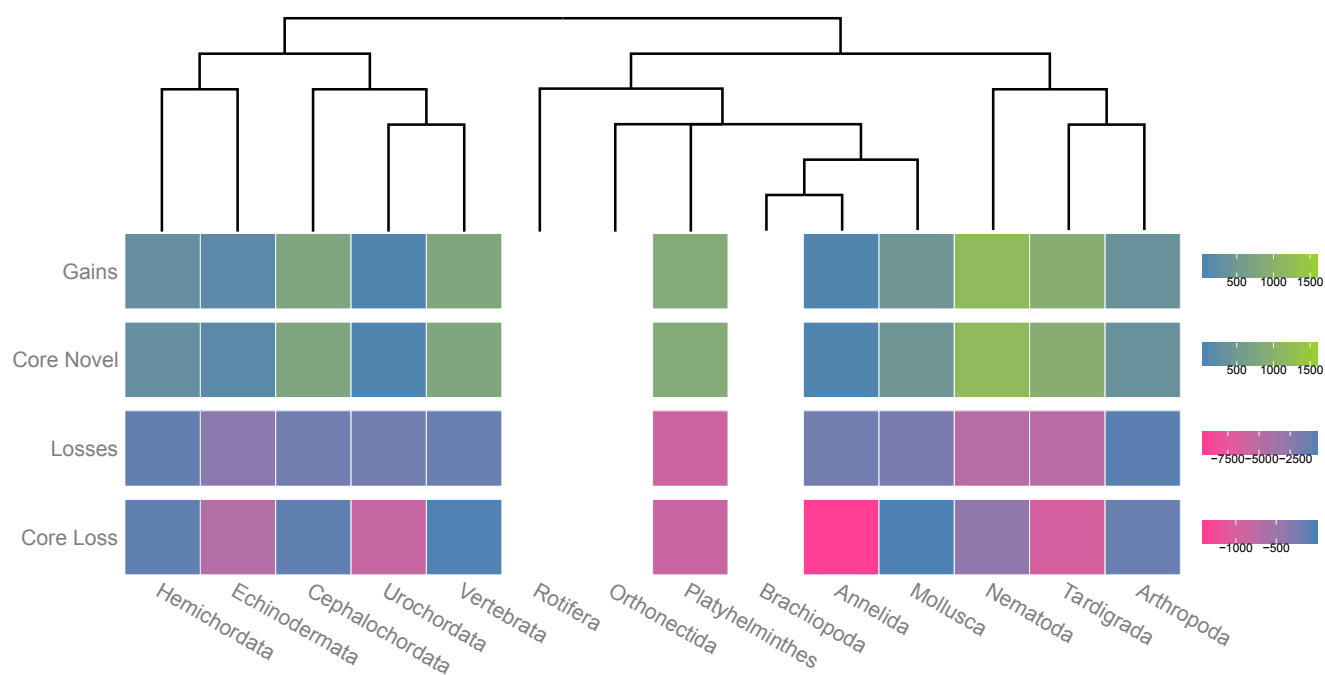


Figure 2.7 Levels of gene gains and losses at phylum level. Heatmap normalised by row displaying the amount of gene gains (green at highest numbers, blue at lower numbers) and loss (pink at highest loss, blue at fewer losses) for the animal phyla in this study.



Figure 2.8 Most abundantly lost and gained molecular functions (GOs). (A) Heatmap for core novel (++) GO molecular functions. Scale corresponds to percentage (%) of each molecular function in each core novel (++) HG per clade, calculated over the total spread of GO molecular functions. (B) Loss within a molecular function is indicated by filled blue circle (not necessarily loss of entire GO category). While different clades (columns) may have gained or lost the same functions, the actual HG gained or lost may be different. GO gained or lost in a clade refer to a subset of HG that perform that function, not all the HG associated with it.

2.4 MATERIAL & METHODS

GENOME COLLECTION

Canonical proteins from whole genomes were downloaded from 59 animal species, from ~16 phyla, and 43 non-animal eukaryotes. Genome annotation completeness was determined by BUSCO analysis (Simão *et al.*, 2015) using the eukaryote dataset of 303 orthologues. The cut-off criteria for genome quality was absence of more than 15% BUSCO orthologues in animals, unless the genome in question shared a phylum/subphylum with another high quality (>85% complete) genome.

COMPARATIVE GENOMICS

The selected genomes were compared using a reciprocal BLASTp (Camacho *et al.*, 2009) of all sequences against all sequences, with an e-value threshold of $\times 10^{-6}$. Markov Cluster Algorithm (Enright *et al.*, 2002) (MCL) was used to infer HGs from the BLASTp output, with default inflation parameter ($\mathbb{I}=2$). Gene ontologies (GOs) (Carbon *et al.*, 2017) were assigned to the different HGs using the Uniprot API for the sequences downloaded by Uniprot (Bateman *et al.*, 2017). Missing GOs were annotated using Interproscan (Jones *et al.*, 2014).

DEFINITION OF HOMOLOGY GROUPS (HG)

Following a consensus phylogeny, of most highly supported animal relationships, from well-known studies (Laumer *et al.*, 2015; Giribet, 2016; Kocot, 2016), the different types of HG (novel, core novel, etc; Figure 2.9) were inferred for the different clades through an in-house custom MySQL database. For the phyla with only two taxa, the definition of core novel HG and novel HG meet the same criteria, and so the HGs values are equal. The HG values for phyla represented by a single species (rotifers, orthonectids and brachiopods) are not comparable with other groups due to an excess of orphan genes, but they are useful to establish values for the other clades. For each type of HG, GOs were mined from Uniprot or obtained using InterProScan. Not all HGs were assigned GOs due to limited annotations in lesser studied phyla (e.g., tardigrades, orthonectids, rotifers). A reliability check was performed on the core novel HGs to assess their absence in the outgroups, using larger sampling. The RefSeq protein database (Pruitt *et al.*, 2007) was used, which has a comprehensive taxon and sequence sampling derived from studies ranging from single-gene analyses to transcriptomes and complete genomes. All the sequences from Novel Core HG were searched in RefSeq using BLASTp, expected cut-off value was $1e-6$ and identity >50% in the BLASTp parameters, and the option “–

negative_gilist” was used to exclude hits against the ingroup in the output files. Only in one case, Novel Core HG in cephalochordates, did two genes recover hits in other animals.

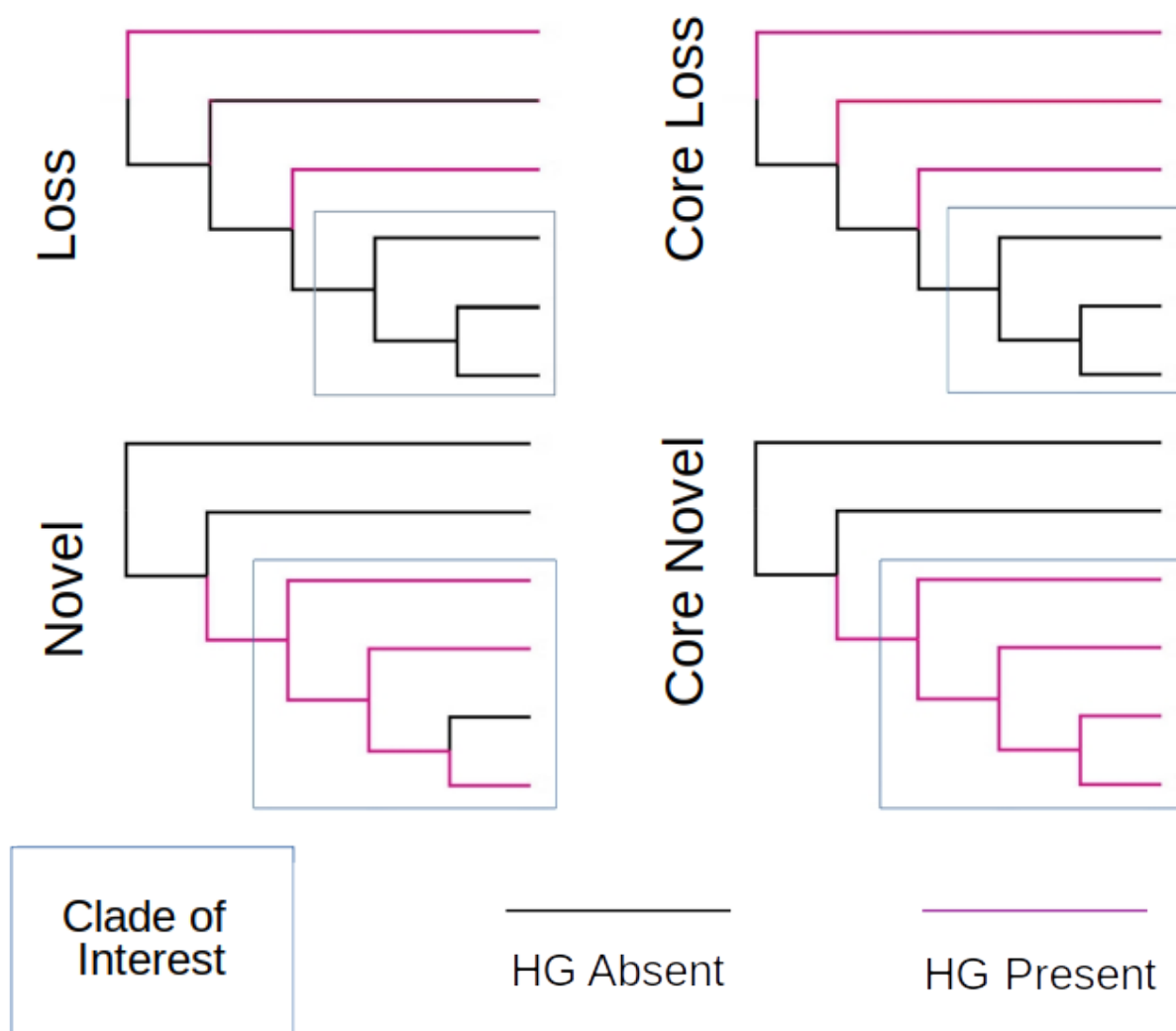


Figure 2.9 Visual description for each of the HG types analysed in the bioinformatics pipeline.

“Loss” out group HG absence is optional. “Novel” in-group absence is optional. These allow some flexibility to account for incompleteness in genomes as detailed in Figure 2.3.

3 A NEXUS OF GENE AND MORPHOLOGY: HOMEBOX PROTEIN EVOLUTION IN ANIMALS USING GENE TREES AND AN EXTENSIVE TAXON SAMPLING

3.1 SUMMARY

Homeobox genes comprise a superfamily of transcription factors that play a major role in the evolution of animal body plans. This makes them essential to understanding animal evolution, therefore it is essential to classify them. A major issue in homeobox classification is the gap in the diversity of animals represented. There is abundant homeobox knowledge surrounding key model organisms, such as in *Drosophila*, *Amphioxus* and vertebrates, but the research is lacking in the animal phyla that do not have model organism representatives. Here we use the comparative genomics pipeline and dataset (developed in chapter 2) with gene tree inference to collate ~8000 homeodomain proteins across 102 eukaryotes, of which 59 are animals and 43 non-animal out-groups. With a focus on animal homeoboxes, using representatives from 16 phyla, many of which do not have well annotated or classified homeobox genes at all, distinct clades of yet-to-be identified homeobox related genes have been observed. The gene trees show a representation bias in current classification towards chordates in previous known sequences. An expansion of distinct LIM-domain containing proteins unveils a consistent reshuffling of domains not previously observed within all the phyla except vertebrates and urochordates. Relationships between chordate specific and lophotrochozoan "unassigned" homeoboxes reveal that some superclasses may predate previous expectations and suggest patterns of whole homeobox gene class loss in the other phyla. HOX-like (HOXL) ANTP families show high support for diverging clades of HOXL ANTP genes in the annelid *Helobdella robusta* and urochordates. Lophotrochozoans have a large number of genes nested within Hox1 and Hox3 clades, related closely with deuterostome and arthropod genes. Homeobox evolution reveals a pattern of domain shuffling, domain combinations and constant expansion. Reduction of homeobox genes may be prompted by this reshuffling; repurposing the functions of neutral homeoboxes into

novel homeoboxes. There is likely to be an aspect of convergent evolution. This research into specific animal developmental genes can be used to describe the changing environment that spurred the divergence of each phylum. A more detailed analysis into the function and locations of these understudied homeobox families in invertebrates is necessary to uncover the importance of these evolutionary events in this history of metazoans.

3.2 BACKGROUND

A key feature in the evolution of animals is their morphological diversity and body plans. Key players in generating this disparity are transcription factors known as homeobox genes (Holland, 2013). Homeobox genes are prevalent throughout eukaryotes, but the vast majority of diversification has been seen in the evolution of animals. Homeobox genes in animals can be organised into 11-12 distinct superclasses (Holland, 2013; Bürglin & Affolter, 2016). Superclasses of homeodomain proteins in animals include: ANTP, PRD, POU, HNF, CUT, LIM, ZF, CERS, PROS, SIX/SO (SINE) and TALE (Holland, 2013; Bürglin & Affolter, 2016). Superclasses: ANTP, LIM, POU, SIX and TALE subclasses *Irx*, *Meis* and *Tgif* are metazoan specific (Sebé-Pedrós *et al.*, 2011; Paps & Holland, 2018).

An increase in homeobox genes in animals was driven by gene duplication events. It is suggested that an increase in the number of homeobox genes leads to an increase in animal complexity (Holland, 2013, 2015). Mutations in homeobox genes cause substantial alterations in animal morphology; this is due to the impact in developmental genes. This can amplify evolutionary leaps, for example the drastic change around 400 million years ago as animals transitioned from arthropods similar to extant crustaceans with multiple limbs on multiple body segments, to the extant insects we see today (Ronshaugen *et al.*, 2002; Chipman *et al.*, 2014; Bürglin & Affolter, 2016). Homeobox loss has also been observed to lead to these speedy evolutionary events, such as seen in unicellular eukaryotes (Paps *et al.*, 2012; Bürglin & Affolter, 2016).

Whilst homeobox gene inferences were limited within lophotrochozoan species prior to the last few years; genomic data for this group of animals is now more available than ever. Paps *et al.* (2015) analysed ~2000 homeobox gene sequences, including a lophotrochozoan majority. New lophotrochozoan-specific genes were described, believed to have a major impact in the morphology of the group, and well-known families were shown to be older than previously thought (Paps *et al.*, 2015; Barton-Owen *et al.*, 2018). For example, *Barx* and *Hopx* previously seen only in deuterostomes are also present in Lophotrochozoa, meaning that they originated at the stem of bilaterians. Variation is also seen in copy numbers, such as gene expansions in rotifers (Paps *et al.*, 2015). Research into animal homeoboxes has been dominated by bilaterian families, so novel and unclassified homeobox genes specific to non-chordate phyla have been largely neglected; there is a huge and unknown diversity of homeodomain proteins waiting to be explored (Paps *et al.*, 2015; Somorjai *et al.*, 2018).

Gene trees provide an insight into the evolution of gene superfamilies and the relationships between them (Nam & Nei, 2005). Here we use a comparative genomics pipeline with gene tree inference to collate ~8000 homeodomain proteins across 59 animals and 43 non-animal out-groups. Using this approach, we further unravel the current evolutionary narrative of animal morphological disparity considering homeobox expansion, loss, duplication and convergence.

3.3 RESULTS & DISCUSSION

Animal homeodomains were found divided across 30 out of 525,309 homology groups (HGs) in 102 whole animal genomes with over 2.6 million canonical proteins (see chapter 2). Homology groups in this study refer to proteins grouped by the Markov Cluster Algorithm (MCL) after an all-vs-all reciprocal BLASTp on the whole eukaryote genomes. 8,034 homeodomain containing proteins, including those downloaded from HomeoDB2 (Zhong *et al.*, 2008; Zhong & Holland, 2011), were collected, aligned using Multiple Sequence Alignment (MSA) methods, and trees were inferred using the Maximum Likelihood (ML) approach (Figure 3.1). The majority of the HGs clustered homeodomain superclasses, subclasses and some distinct families, such as ANTP/En, ANTP/Nk6 and HNF/Hmbox. The results are described per HG tree for easier reference.

3.3.1 CLASSIFICATION OF THE HOMEODOMAIN GENES

In the classification of the homeobox genes several factors are taken into account. The primary classification uses the BLASTp and InterProScan 5 similarity and domain signatures to assign a Homeobox class and family using signature databases as described in the methods and materials, and HomeoDB2. The secondary classification uses the protein trees to correct outlying annotations from the primary step, and fill in missing classifications where confidence, branch length and neighbouring leaves support an homeobox class and family. Finally, a ROC curve (Receiver Operating Characteristic curve) (Zhu *et al.*, 2010) is tested against HomeoDB against all the classifications to ensure that a high level of accuracy is achieved in the primary and secondary classifications. This is applied using a measure of sensitivity taken by comparison of the true positive rate against the false positive rate. In this instance, if the classification correctly matched the name assigned in a HomeoDB protein, it scores 1, if the classification did not match, it scored 0. The % false positive rate was 99.1% (Figure 3.2).

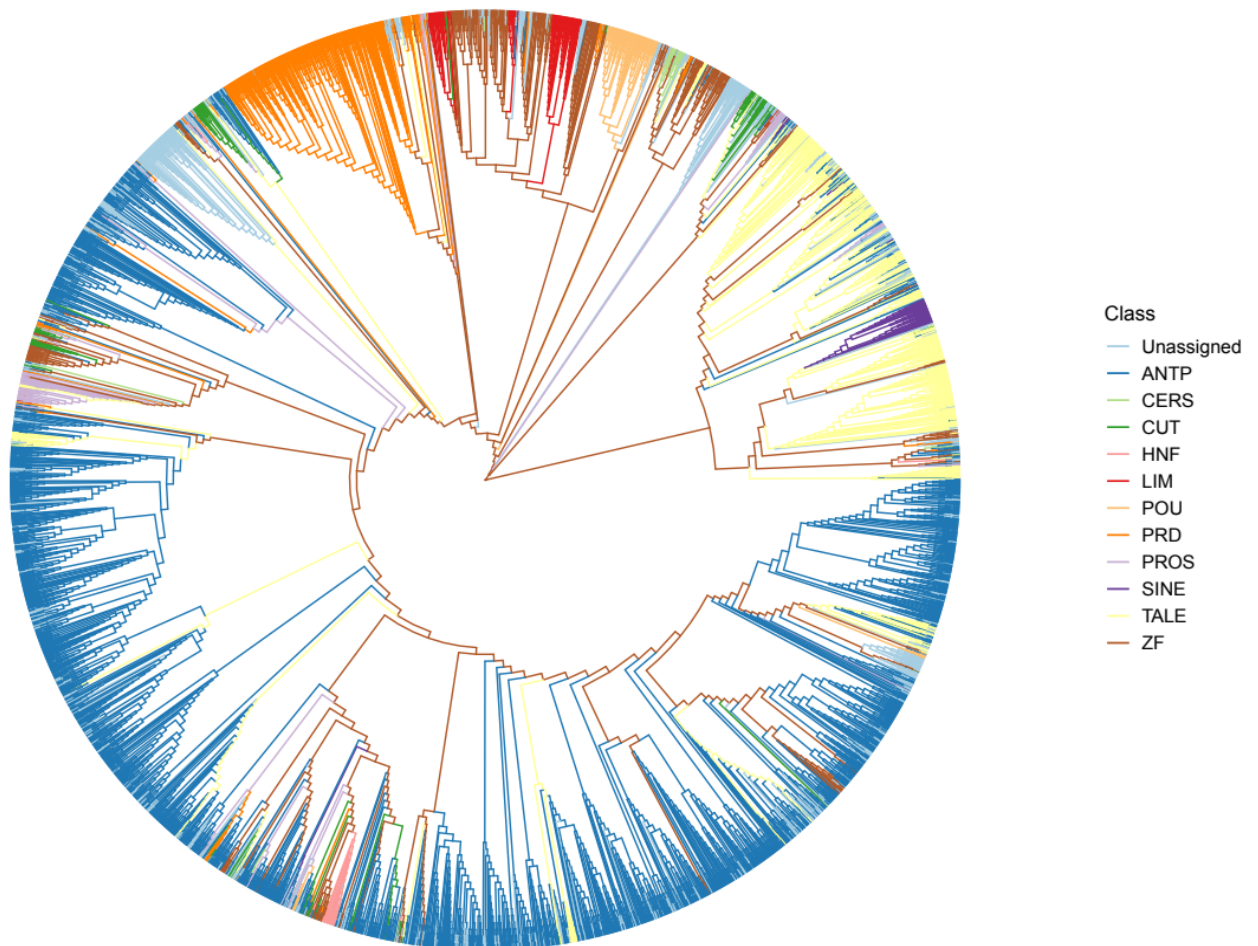


Figure 3.1 Gene tree consisting ~8000 homeodomain proteins. Coloured by superclass and rooted at a plant specific homeobox clade (HD-ZIP).

3.3.2 PROTEIN TREES

A tree gathering the 8034 sequences collated (Figure 3.1) did not recover the monophyly of some well-known superclasses, such as POU, PRD or LIM. This is most likely due to the short length of the homeodomain together with the large sequence sampling, increasing the levels of homoplasy across the alignment. Thus, further trees were inferred focusing on specific clusters (HGs) of homeobox genes.

PRD

The largest HG includes various PRD families (907/1416 proteins of this HG are PRD families) (Figure 3.3) among other superclasses of homeobox. The PRD superclass was divided into 3

paraphyletic clades (Figure 3.1). Shown in Table 3.1, the PRD superclass has evolved into lineage specific families in the Animal Kingdom, with many of these expanding after the divergence of the first splitting extant animal groups, pre-bilaterian (Porifera & Ctenophora).

POU

Interestingly, aside from the abundance of PRD families in the aforementioned HG, there appears to be a diverging clade of three genes, homologous to POU Hdx in the rotifer *Adineta vaga*, with high ultrafast bootstrap support, otherwise seen only in chordates (Figure 3.3).

ANTP

The next largest HG (Figure 3.4) is dominated by HOX-like (HOXL) ANTP families, it is a large homeobox family with several paralogues in each species. The gene tree shows high support for diverging clades of homeobox genes homologous to HOXL in the annelid *Helobdella robusta* and urochordates closely related to a vertebrate Hox9-13(15) clade. Every genome here has a Hox9-13(15) gene family. The appearance of these genes in sponges is artefactual and likely based on a HOXL signature similarity within the proteins sequences (Fonseca *et al.*, 2008). Within the same HG, lophotrochozoans have distinct clades within Hox1 and Hox3 (Figure 3.5 & 3.6), Hox3 is a smaller family in the lophotrochozoans, with some orthonectid, platyhelminthe and mollusc species having lost it altogether. Hox3 has been expressed in the mantle mesoderm and in the ventral ectoderm of early brachiopod gastrulae and later larvae (Schiemann *et al.*, 2017). The missing Hox3 in *Lingula anatina* is therefore surprising, especially given the completeness of the genome shown in the BUSCO analysis of chapter 2, which would suggest that it has not been lost in the genome assembly, and has simply not been detected in this pipeline due to lacking annotations. Within another large HG, Barx is seen in molluscs, exclusively within a deuterostome clade (Figure 3.7) in a different HG to the rest of the Barx family. The Barx family shows a reasonable expansion in cnidarians (Figure 3.8), which has been observed only very recently (Gold *et al.*, 2019). The Barhl family is another ANTP homeobox that has been well conserved across all the phyla, with a clear monophyletic clade sister to the deuterostome Barx family. There is a distinct divergence between deuterostome Barhl genes and ecdysozoan (Figure 3.9). Tardigrades, hemichordates and echinoderms each only have a single En ANTP family gene. This homeobox gene evolved in the last common ancestor of bilaterians and was not present prior, it is present in all the bilaterians and absent in ctenophores, poriferans, placozoans and cnidarians. Most bilaterian lineages have seen a duplication event and have more than one En homeobox gene (Figure 3.10). The presence of this homeobox family in every bilaterian is well

classified, previously labelled as part of the "SuperHox" cluster, it was expected to have originated in a bilaterian ancestor, here we see the same pattern (Ferrier, 2016).

LIM

In a LIM dominated HG there is a huge intermingling of LIM Lmx families with PRD Pax2/5/8, Muxb and ZF Tshz. LIM Lmx also has a highly supported branch which diverges greatly with representatives from all the animal phyla. These have been identified as LIM due to a collection of annotated sequences nested within the clade, and LIM domains detected in InterProScan v5 (Figure 3.11).

TALE

Poriferans and orthonectid (*Intosia linea*) share homology with a Meis/Pknox TALE clade alongside one annelid (*Capitella teleta*) and otherwise non-animal eukaryotes. Orthonectida Meis shares heritage among platyhelminthes. A distinct TALE clade, nested between Pbx and Meis has been lost in vertebrates and cephalochordates, but is in urochordates and protostomes, and there has been an unprecedented expansion of Meis family in invertebrates particularly, including non-bilaterians, lophotrochozoans and urochordates to name a few (Figure 3.12). LOC647589 (a human specific homeobox) is positioned as sister to amphioxus Atale TALE, a currently believed divergent amphioxus specific TALE homeobox (96% UFBS support).

CERS

Excluding only vertebrates, is a large diverging clade of Cers-like CERS proteins. It is separate from the other Cers CERS and detected by InterProScan5 to contain the Sphingosine N-acyltransferase Lag1/Lac1-like domain as well as Homeodomain. This clade is shared with non-animal eukaryotes (Figure 3.13). CERS cers is present in all animals except orthonectid *Intosia linea* and centipede *Strigamia maritima*. As the HGs become groups of fewer proteins, they contain the smaller known groups of homeodomains.

ZF

A few more notable observations include: tardigrades and *Daphnia pulex* have divergent ZF class genes alongside a single rotifer homeodomain. In this ZF HG, *Nematostella vectensis*, appears to have a homologous homeodomain, placing the origin of ZF as present in the ancestors of cnidarians

(Figure 3.14). This has not been seen before, ZF is usually associated among those homeoboxes that originated in the stem of bilaterians (Ryan *et al.*, 2006; Gold *et al.*, 2019).

CUT

Nematodes have an expansion of distinct Cmp family genes in the CUT superclass. These homeobox genes were clustered in a distinct and very small HG when compared to the rest. They do not appear to be artefactual because they have a well-supported relationship with HomeoDB classified nematode Cmp genes in the gene tree (>85% UFBS supports). Cmp and Satb both share a COMPASS domain, Cmp is likely an ecdysozoan specific, if not in this case a nematode specific homeobox; whilst Satb is vertebrate specific (Bürglin & Affolter, 2016). In this thesis, with the exclusion of the COMPASS domain, Satb and Cmp diverged beyond similarity recognition.

HNF & PROS

Superclasses PROS and HNF are much smaller than the other superclasses, but they also showed consistent monophyly across all the gene trees. Interestingly, whilst it has been expected that vertebrates underwent a duplication of this gene in a round of whole genome duplication after the divergence from cephalochordates, there may also be convergent duplication seen in the other bilaterian lineages (Figure 3.15). HNF also shows convergent duplications in Hmbox. Hmbox has been identified as a chordate-specific homeobox (Takatori *et al.*, 2008) (with the exception of a single HomeoDB classified Hmbox in nematodes), however, in the HNF HG gene tree, there is a highly supported (UFBS 94%) clade of protostome genes, homologous to chordate Hmbox genes, and in this thesis, classified as such (Figure 3.16).

The classified homeoboxes, (as seen in Figure 3.17), appear to be dominated by chordates with regard to presence (Figure 3.17). This is a reflection of missing annotations in non-chordate species, therefore this only accurately shows which of the listed families are present in each taxa. However the other animal phyla have a lot of unassigned homeodomains, these are seen in many large vertebrate exclusive clades across the gene trees (Figure 3.1).

HOMEBOX EVOLUTION

Evolution of the homeobox genes has likely been spurred by multiple event types. As seen in the data, homeobox families can be shared between animals at opposite ends of a spectrum, whilst at times missing from closer relatives. Phenotypical differences may not be noticeable despite losses of

homeoboxes, but the loss of some homeoboxes is probably linked to different adaptations, such as parasitism (Nam & Nei, 2005; Tsai *et al.*, 2013). For example, in tape worms, a possible adaptation for parasitism was the loss of 34 homeobox families in the bilaterian ancestor, some of which involved in neural development and others in through-gut specification (Tsai *et al.*, 2013). In cases where an ancient homeobox gene has been retained in a single taxon or phylum, it is probably neutral, where loss is not beneficial.

Table 3.1 The node origin of the homeobox families in relation to animals. Those listed under Metazoa may pre-date Metazoa. Basal2 refers to the node preceding the divergence of the first splitting extant animal.

Class	Family	Clade of origin	Class	Family	Clade of origin
ANTP	Abox	Basal2	OTHER	Muxb	Cephalochordata
ANTP	Ankx	Basal2	OTHER	Nanognb	Tetrapoda
ANTP	Barhl	Basal2	OTHER	Sia	Tetrapoda
ANTP	Bari	Basal2	OTHER	Unassigned	Basal2
ANTP	Barx	Metazoa	POU	Hdx	Basal2
ANTP	Bsx	Basal2	POU	Pou1	Metazoa
ANTP	Cdx	Basal2	POU	Pou2	Basal2
ANTP	Dbx	Basal2	POU	Pou3	Metazoa
ANTP	Dlx	Metazoa	POU	Pou4	Basal2
ANTP	Emx	Basal2	POU	Pou5	Tetrapoda
ANTP	En	Metazoa	POU	Pou6	Basal2
ANTP	Evx	Basal2	PRD	Alx	Metazoa
ANTP	Gbx	Basal2	PRD	Aprda	Cephalochordata
ANTP	Gsx	Basal2	PRD	Aprdb	Cephalochordata
ANTP	Hhex	Basal2	PRD	Aprdc	Cephalochordata
ANTP	Hlx	Basal2	PRD	Aprdd	Cephalochordata
ANTP	Hox1	Basal2	PRD	Aprde	Cephalochordata
ANTP	Hox2	Basal2	PRD	Argfx	Tetrapoda
ANTP	Hox3	Basal2	PRD	Arx	Basal2
ANTP	Hox4	Basal2	PRD	Cg11294	Mandibulata
ANTP	Hox5	Basal2	PRD	Dmbx	Basal2
ANTP	Hox6-8	Basal2	PRD	Dprx	Tetrapoda
ANTP	Hox9-13(15)	Metazoa	PRD	Drgx	Basal2
ANTP	Hx	Cephalochordata	PRD	Dux	Basal2
ANTP	Lbx	Metazoa	PRD	Esx	Tetrapoda
ANTP	Lcx	Cephalochordata	PRD	Gsc	Basal2
ANTP	Meox	Basal2	PRD	Hbn	Mandibulata
ANTP	Mnx	Basal2	PRD	Hesx	Basal2
ANTP	Msx	Basal2	PRD	Hopx	Basal2
ANTP	Msx1x	Basal2	PRD	Isx	Basal2
ANTP	Nanog	Basal2	PRD	Leutx	Tetrapoda
ANTP	Nedx	Basal2	PRD	Mix	Basal2
ANTP	Nk1	Basal2	PRD	Nobox	Basal2
ANTP	Nk2.1	Metazoa	PRD	Otp	Basal2
ANTP	Nk2.2	Metazoa	PRD	Otx	Basal2
ANTP	Nk3	Basal2	PRD	Pax2/5/8	Metazoa
ANTP	Nk4	Basal2	PRD	Pax3/7	Basal2

Class	Family	Clade of origin	Class	Family	Clade of origin
ANTP	Nk5/hmx	Basal2	PRD	Pax4/6	Basal2
ANTP	Nk6	Basal2	PRD	Phox	Basal2
ANTP	Nk7	Basal2	PRD	Pitx	Metazoa
ANTP	Noto	Basal2	PRD	Prop	Basal2
ANTP	Pdx	Basal2	PRD	Prrx	Basal2
ANTP	Ro	Basal2	PRD	Rax	Basal2
ANTP	Tlx	Basal2	PRD	Repo	Basal2
ANTP	Vax	Basal2	PRD	Rhox	Tetrapoda
ANTP	Ventx	Metazoa	PRD	Sebox	Tetrapoda
CERS	Cers	Metazoa	PRD	Shox	Basal2
CUT	Acut	Basal2	PRD	Tprx	Tetrapoda
CUT	Cmp	Metazoa	PRD	Uncx	Basal2
CUT	Cux	Basal2	PRD	Vsx	Basal2
CUT	Onecut	Basal2	PROS	Prox	Metazoa
CUT	Satb	Basal2	SINE	Six1/2	Metazoa
HNF	Ahnfx	Cephalochordata	SINE	Six3/6	Basal2
HNF	Hmbox	Basal2	SINE	Six4/5	Metazoa
HNF	Hnfl	Metazoa	TALE	Atale	Cephalochordata
LIM	Isl	Basal2	TALE	Irx	Metazoa
LIM	Lhx1/5	Metazoa	TALE	Meis	Basal2
LIM	Lhx2/9	Basal2	TALE	Mkx	Basal2
LIM	Lhx3/4	Basal2	TALE	Pbx	Basal2
LIM	Lhx6/8	Basal2	TALE	Pknox	Basal2
LIM	Lmx	Metazoa	TALE	Pknox/meis	Platyhelminthes
OTHER	Ahbx	Metazoa	TALE	Tgif	Basal2
OTHER	Beetlebox	Mandibulata	ZF	Adnp	Basal2
OTHER	Bix	Tetrapoda	ZF	Azfh	Basal2
OTHER	Cphx	Basal2	ZF	Tshz	Basal2
OTHER	Loc647589	Tetrapoda	ZF	Zeb	Basal2
OTHER	Muxa	Cephalochordata	ZF	Zfhx	Basal2
			ZF	Zhx/homez	Metazoa

There are many families of homeoboxes that appear to be fragmented across taxa, with a patchy distribution, that are otherwise lost in whole phyla. An example of this is the POU gene relative of Hdx seen only in chordates appearing to have a homologous protein in rotifer and a chelicerate. A potentially homologous to Hdx domain has been retained through the divergence of deuterostomes and protostomes to the lophotrochozoan rotifer. This is likely a case of homeodomain similarity, and not evidence that the rotifer or chelicerate have a Hdx gene. POU is a gene class linked to the emergence of animals, and despite a simple body plan in sponges, organs in other animals that are regulated by POU genes may share a deep common ancestry among the few sponge cell types (Gold

et al., 2014). Similarly, poriferans and orthonectid- *Intosia linea* share a TALE Meis/Pknox clade with one annelid (*Capitella teleta*) and otherwise non-animal eukaryotes. A distinct TALE clade, nested between Pbx and Meis has been lost in vertebrates and cephalochordates, but is in urochordates and protostomes.

Some of the most discernible unidentified homeobox clades among the gene trees show a reshuffling of domains. For example, the divergent LIM-domain-containing clade. It has been reported that 6/14 LIM homeoboxes were present in stem of Metazoans and the expansion of LIM homeoboxes in the first animal factored into the history of complex animal multicellularity (Koch *et al.*, 2012). López-Escardó *et al.* (2019) considered that POU predated the origin of animals, and only became more animal specific later in evolution. Their findings showed that novel animal genes in the emergence of animals were the result of duplications and rearranged protein domains (López-Escardó *et al.*, 2019). These combined-domain proteins, sitting close to known homeobox class proteins phylogenetically, but in a segregated clade may have been a rewiring of multiple homeobox or other related transcription factor genes. It is possible that the evolution of animal complexity may have more to do with the combinations and organisation/architecture of domains rather than gene number (Nam & Nei, 2005; Babushok *et al.*, 2007; Karaz *et al.*, 2016).

The Dbx family has 2 distinct clades, which could be expected due to suspected duplication events in the emergence of bilaterians (Karaz *et al.*, 2016). Being found in many divergent bilaterian lineages, the Dbx family likely was present in the last common ancestor of bilaterians, this hypothesis supports the gene tree inference in this research. Further biologically explainable results include how myriapods have retained many homeobox genes, that have been lost in other arthropods. Whilst found in the *Strigamia maritima* genome, Hmbox has been lost in all arthropods except in myriapods (Chipman *et al.*, 2014; Ferrier, 2016). However, there is a closely related gene in *Ixodes scapularis* splitting just before the HNF clade. The Hmbox may have been lost in crustaceans and hexapods, and individually lost in tardigrades, but retained in myriapods and ticks.

Lophotrochozoan patterns matched those obtained in other research (Paps & Riutort, 2012; Paps *et al.*, 2015; Morino *et al.*, 2017). We found at least one HNF family within each lophotrochozoan Phylum except Orthonectida, and Pou1 in each of the lophotrochozoan species analysed. This has been tentatively suspected before, although once thought to be non-bilaterian and deuterostome exclusive (Paps *et al.*, 2015). Such results suggest that most of the homeobox losses are not a one-time event, but in fact multiple losses in independent lineages. Alternatively, another theory is the effect of convergent evolution, with strictly conserved homeobox motifs (Fonseca *et al.*, 2008; Bürglin &

Affolter, 2016). This theory of convergence is supported by the unique clade separations seen in the tree branching from distinct homeobox classes (Figure 3.1) and also the divergence of some classes, such as LIM and ZF (Figure 3.11 and Figure 3.14).

More homeobox gene families were present either before or during the emergence of bilaterians than currently believed (Paps *et al.*, 2012). A nexus of homeobox evolution reveals a pattern of domain shuffling, combinations and constant expansion, although convergent evolution of some motifs seen in the superclasses and some subclasses/families of homeoboxes cannot be discounted. Homeobox reduction is prompted by the homeodomain shuffling, repurposing the functions of redundant or neutral homeoboxes. A more detailed analysis into the function and locations of these understudied homeobox families in under-represented animal phyla is necessary to uncover the importance of these evolutionary events in the history of metazoan body plans.

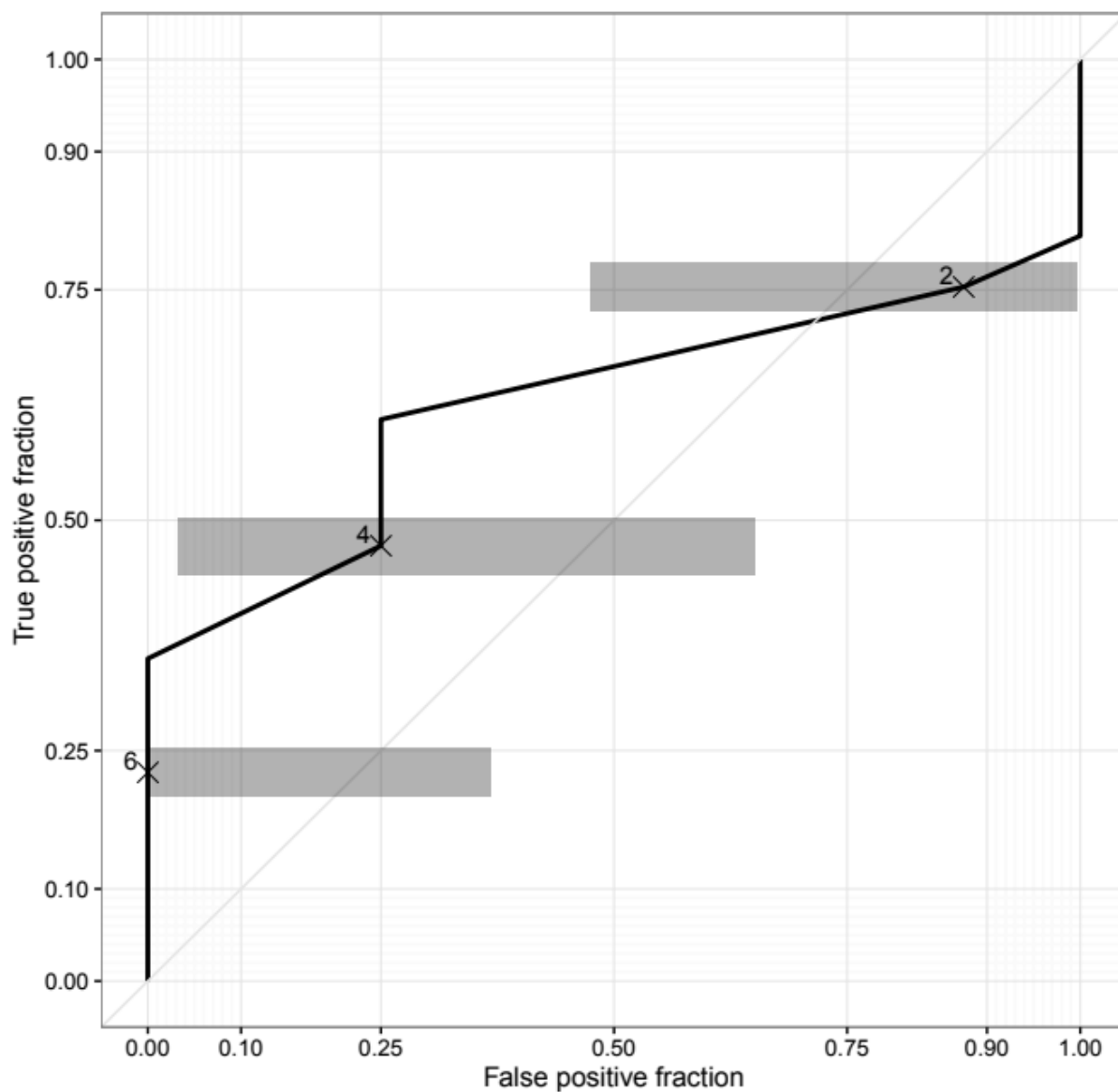


Figure 3.2 ROC analysis of the homeobox classification based on known HomeoDB classifications against same in-house database species. The % error rate for classification was found to be 0.87%. Numbers 2, 4 & 6 are cutpoints at which the test shows an abnormality.

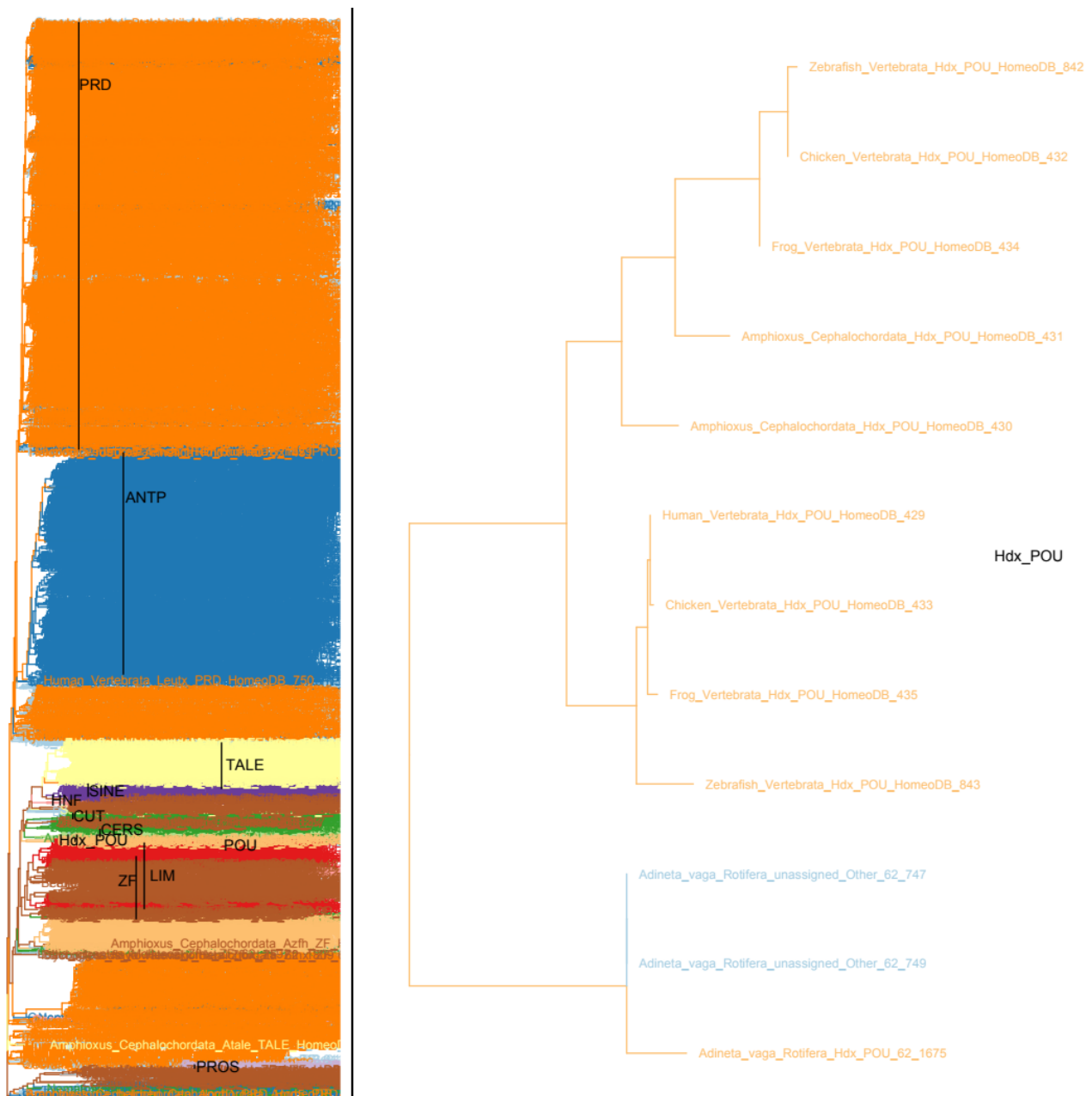


Figure 3.3 The largest homeobox gene tree is dominated by PRD class genes with outgroups from HomeoDB. For this, and the following gene trees: each colour represents a different homeobox class. The light blue leaf labels are proteins that were not assigned a homeobox family, but were similar enough to cluster in the same gene tree. The diverging clade of rotifer genes homologous to the Hdx family of POU class had Ultra-fast-bootstrap (UFBS) values of 100%.

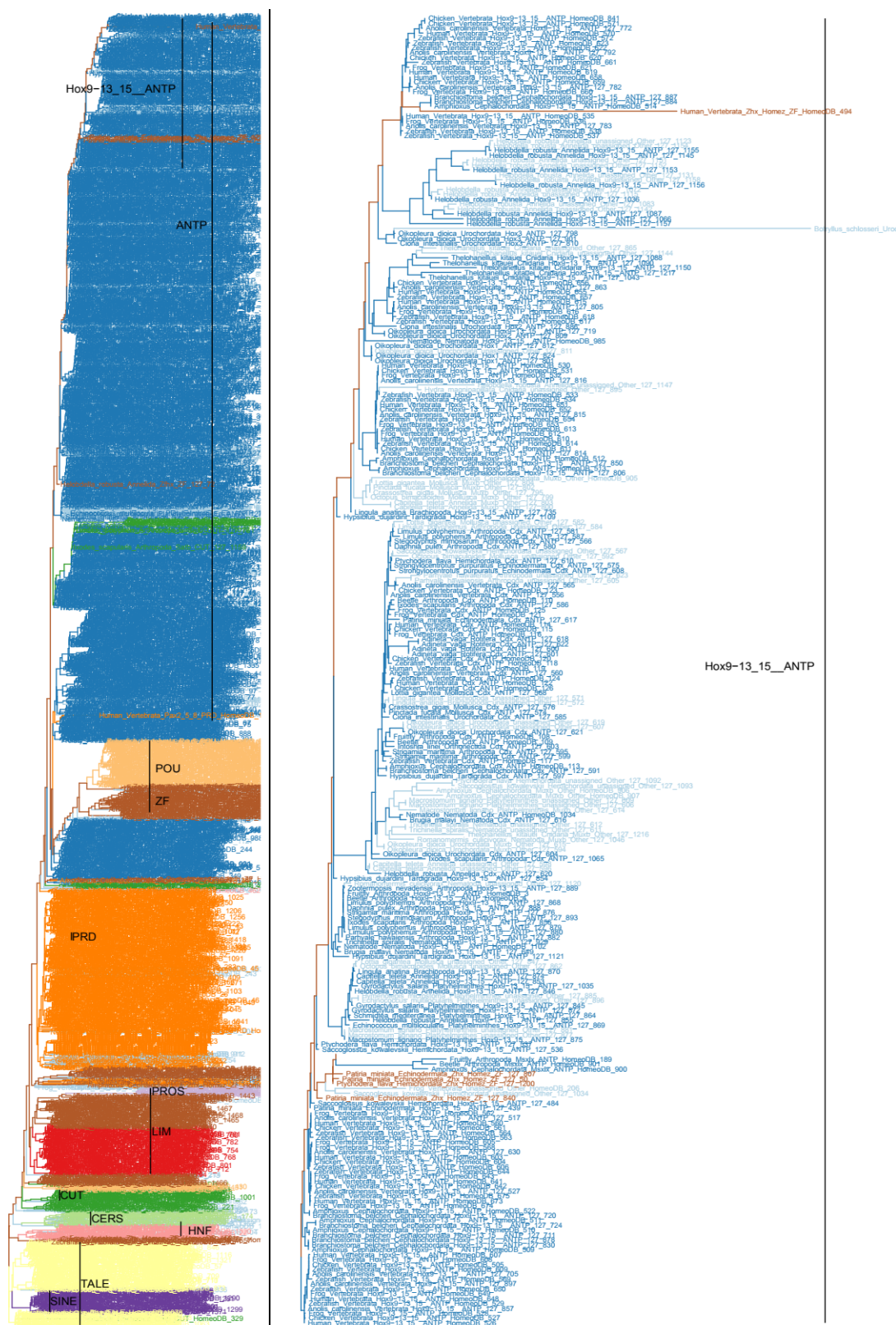


Figure 3.4 One of the HOXL gene trees (ANTP class) with outgroups from HomeoDB. *Hox9-13(15)* appears to be well conserved throughout all the bilaterian animal clades, with lophotrochozoan genes closely related to vertebrate genes.

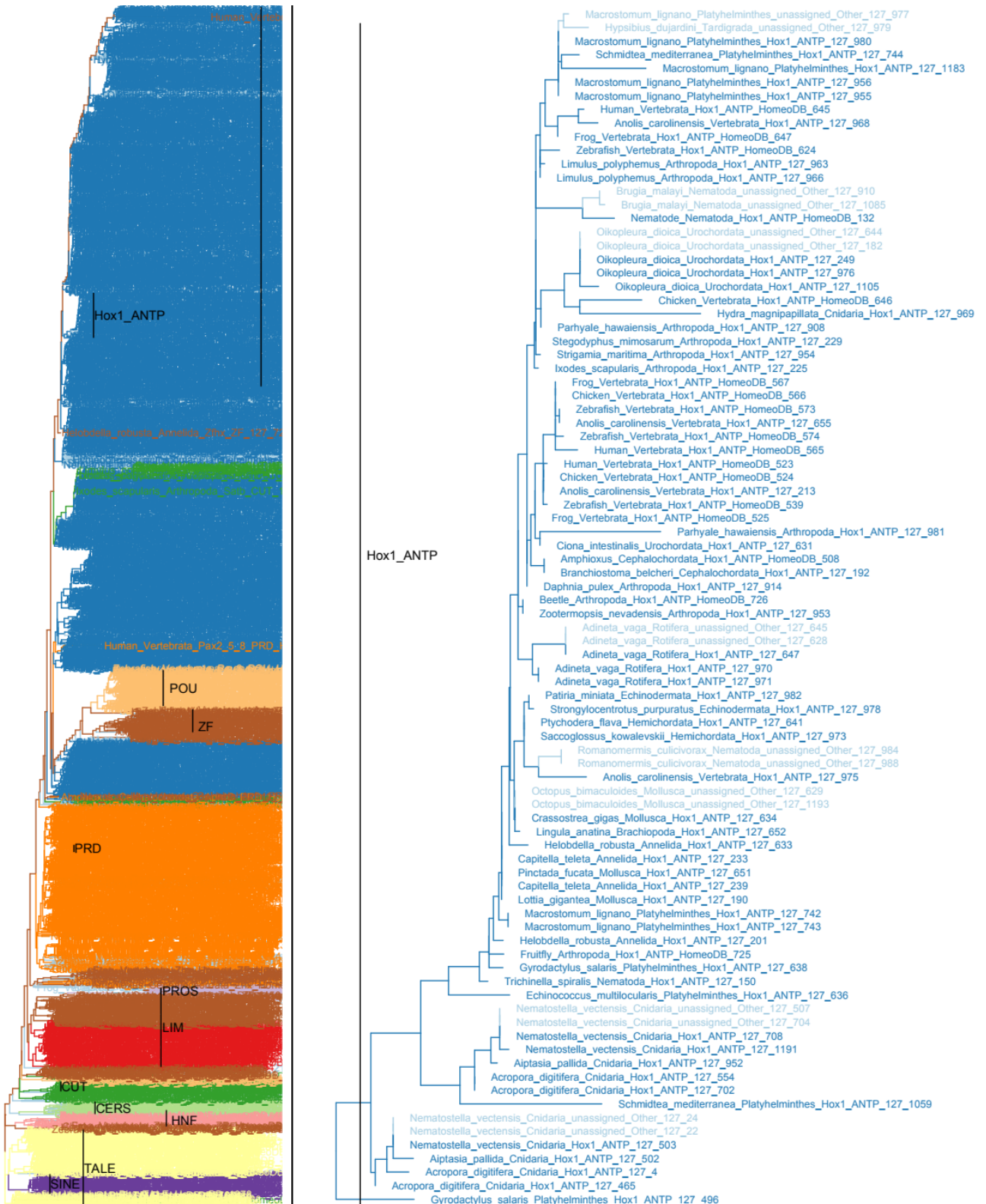


Figure 3.5 One of the HOXL HG ANTP class genes with outgroups from HomeoDB. *Hox1* appears to be well conserved throughout all the bilaterian animal clades, but with some distinct phylum grouped clades. It is a large homeobox family with a couple of paralogues in each non-lobophotrochozoan species.

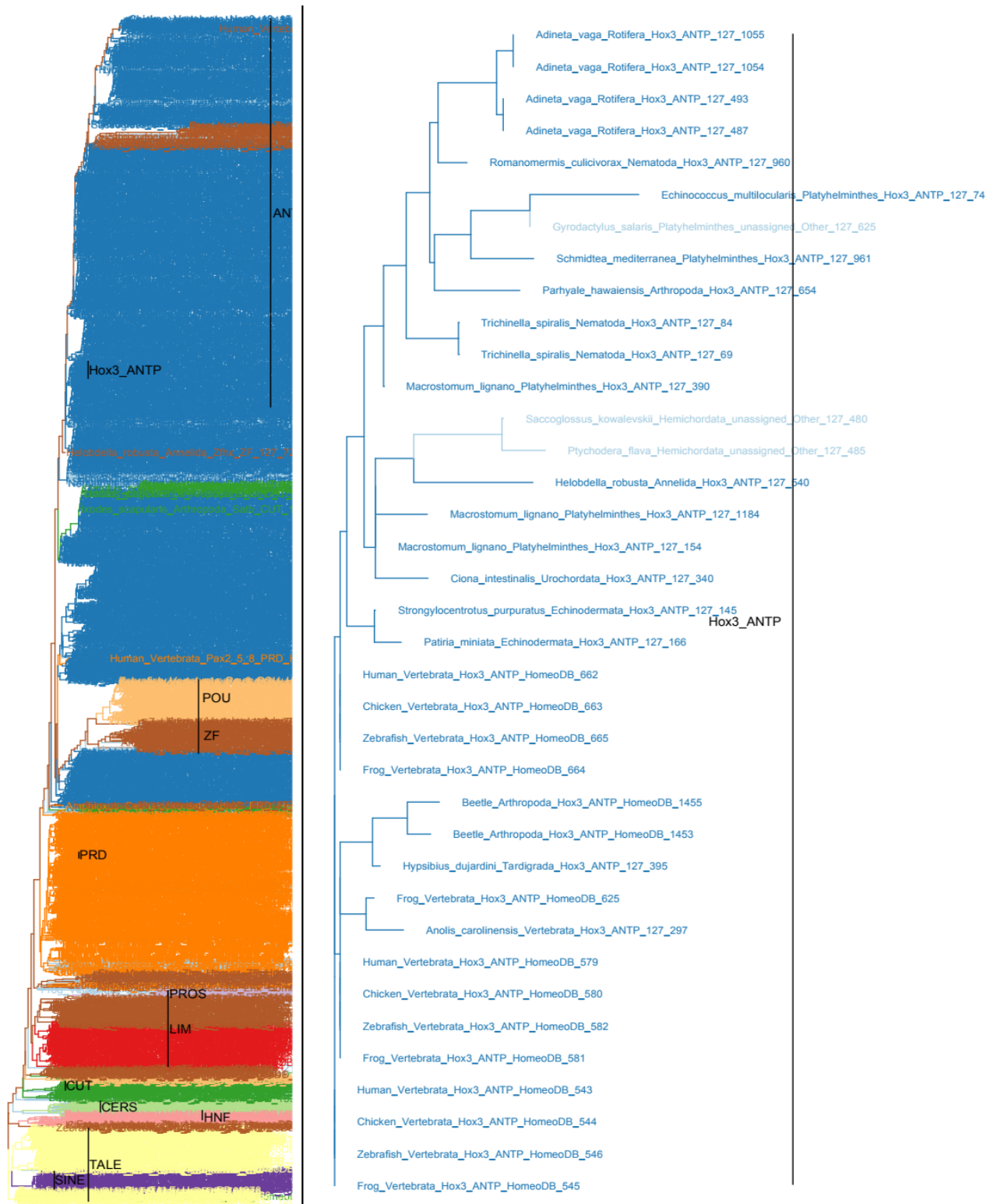


Figure 3.6 One of the HOXL HG ANTP class genes with outgroups from HomeoDB. *Hox3* originates in the stem of bilaterians. There has been some loss in some lophotrochozoan lineages, such as in some Mollusca and the brachiopod *Lingula anatina* here. The *Hox3* in *Crassostrea gigas* and *Lottia gigantea* diverged slightly outside of the clade shown here.

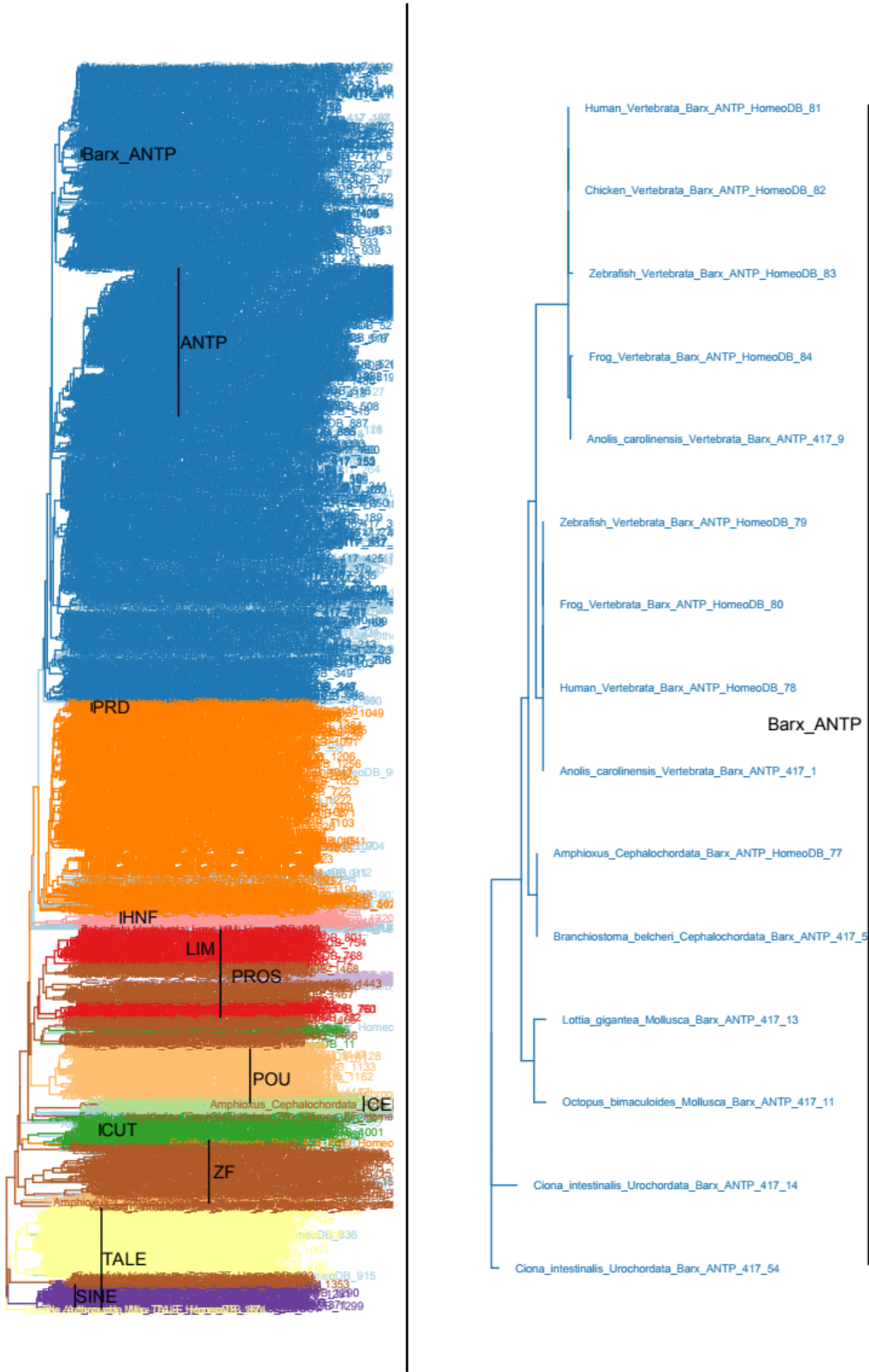


Figure 3.7 Barx family one of the NKL ANTP class gene trees with outgroups from HomeoDB.

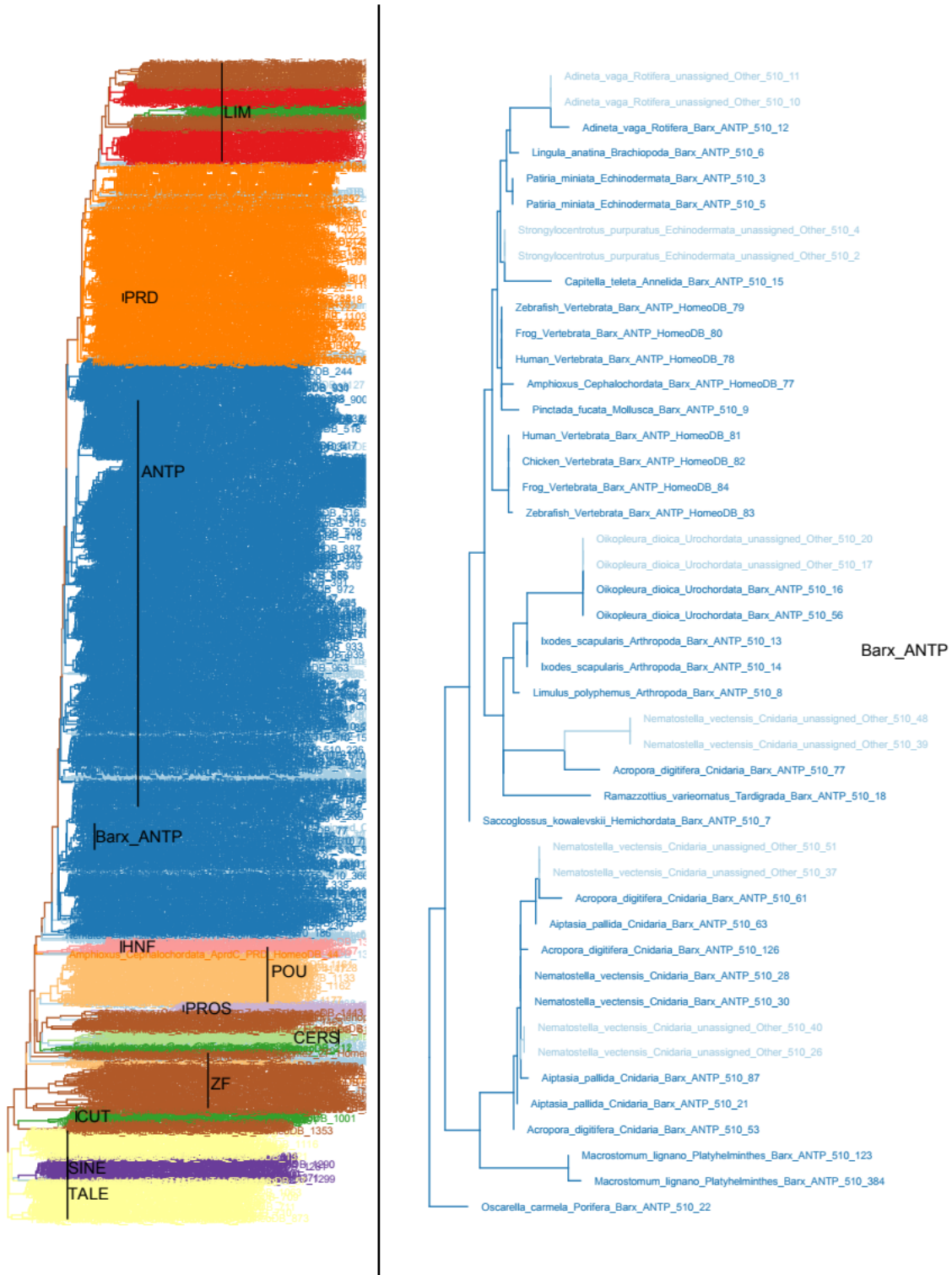


Figure 3.8 Barx family one of the NKL ANTP class HG genes with outgroups from HomeoDB. There has been significant expansions in cnidarians of the Barx.

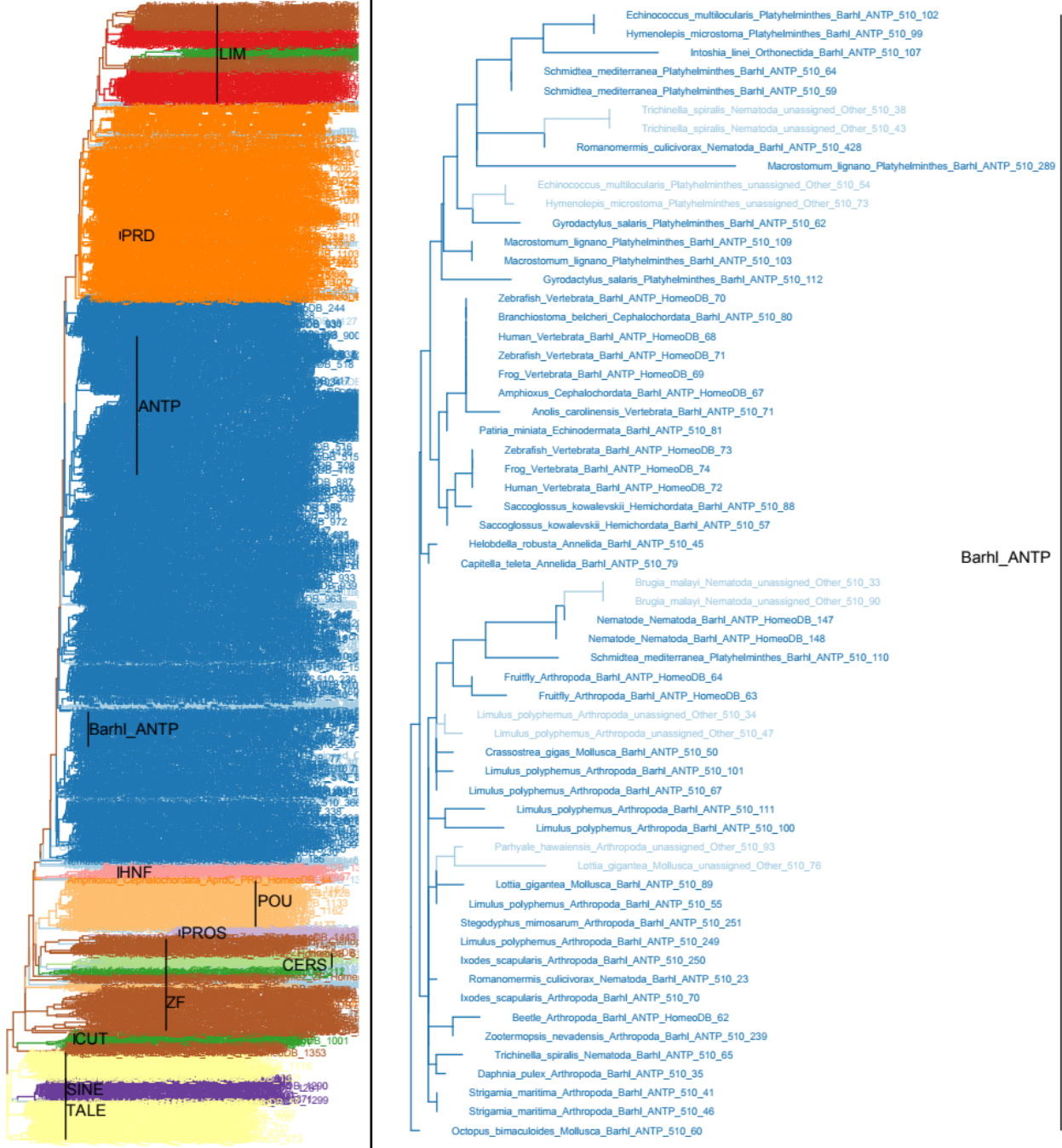


Figure 3.9 Barhl family one of the NKL ANTP class HG genes with outgroups from HomeoDB. There is a clean monophyly of this homeobox family which is seen in every animal lineage.

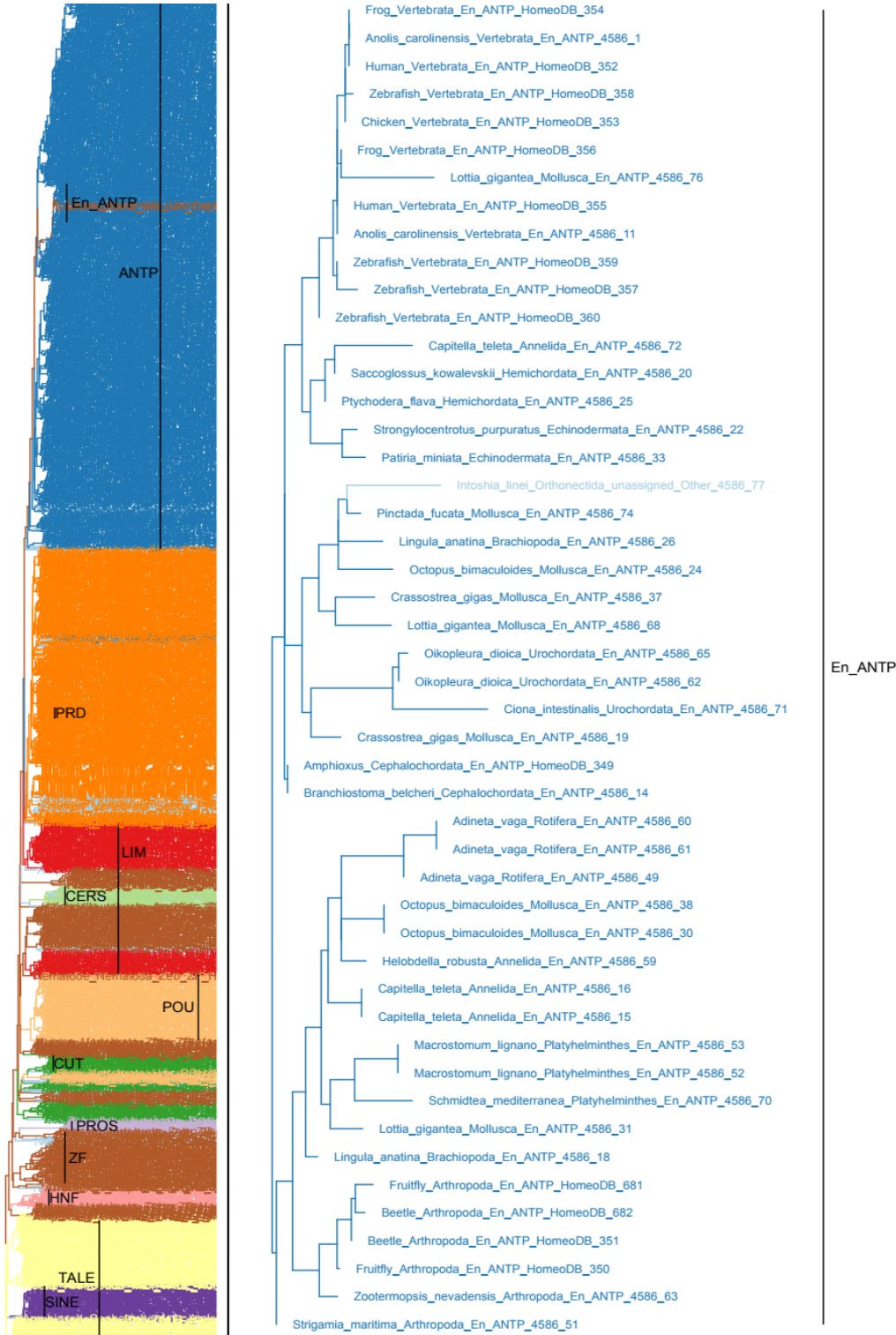


Figure 3.10 The En family of ANTP NKL.

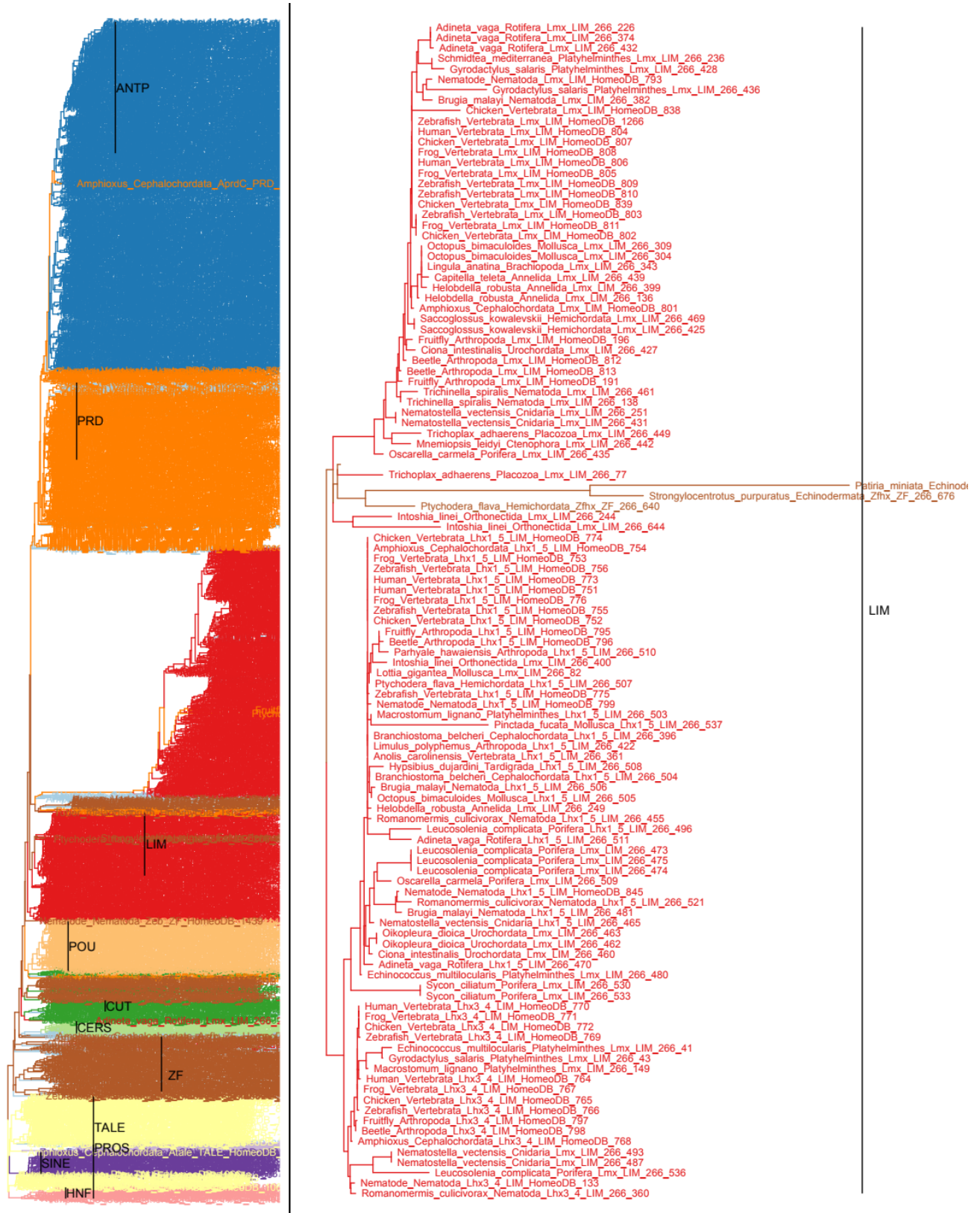


Figure 3.11 LIM class HG genes with outgroups from HomeoDB. This LIM class displays a divergence of the LIM families: *Lhx1-8* and *Lmx*. *Lmx* is present in all the animal phyla, whilst *Lhx* expanded in bilaterians.

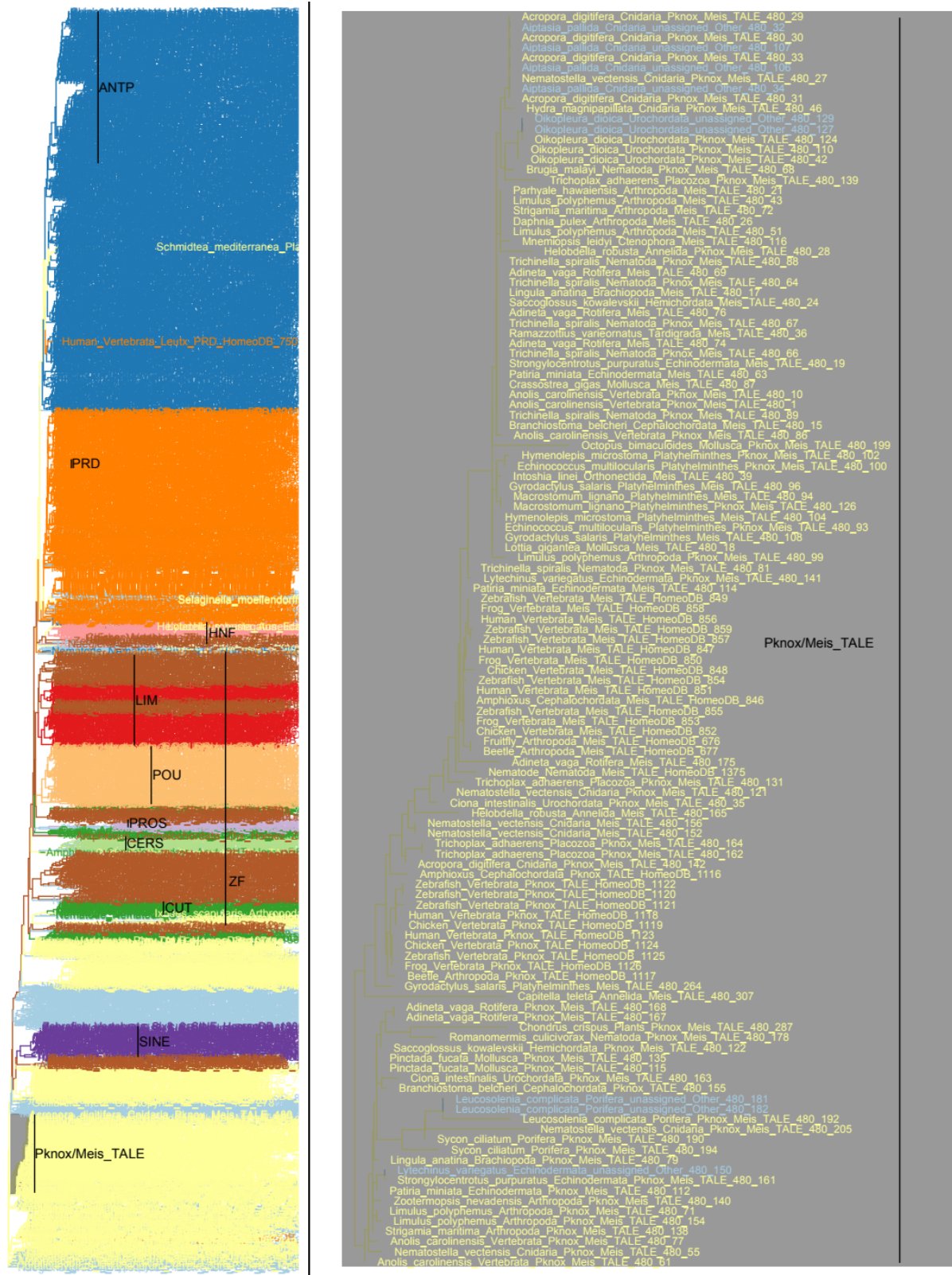


Figure 3.12 TALE-class. These genes have been classified as Pknnox/Meis because it was too difficult to distinguish between them, although Meis is a Metazoan specific homeobox and Pknnox evolved prior to the origin of animals.

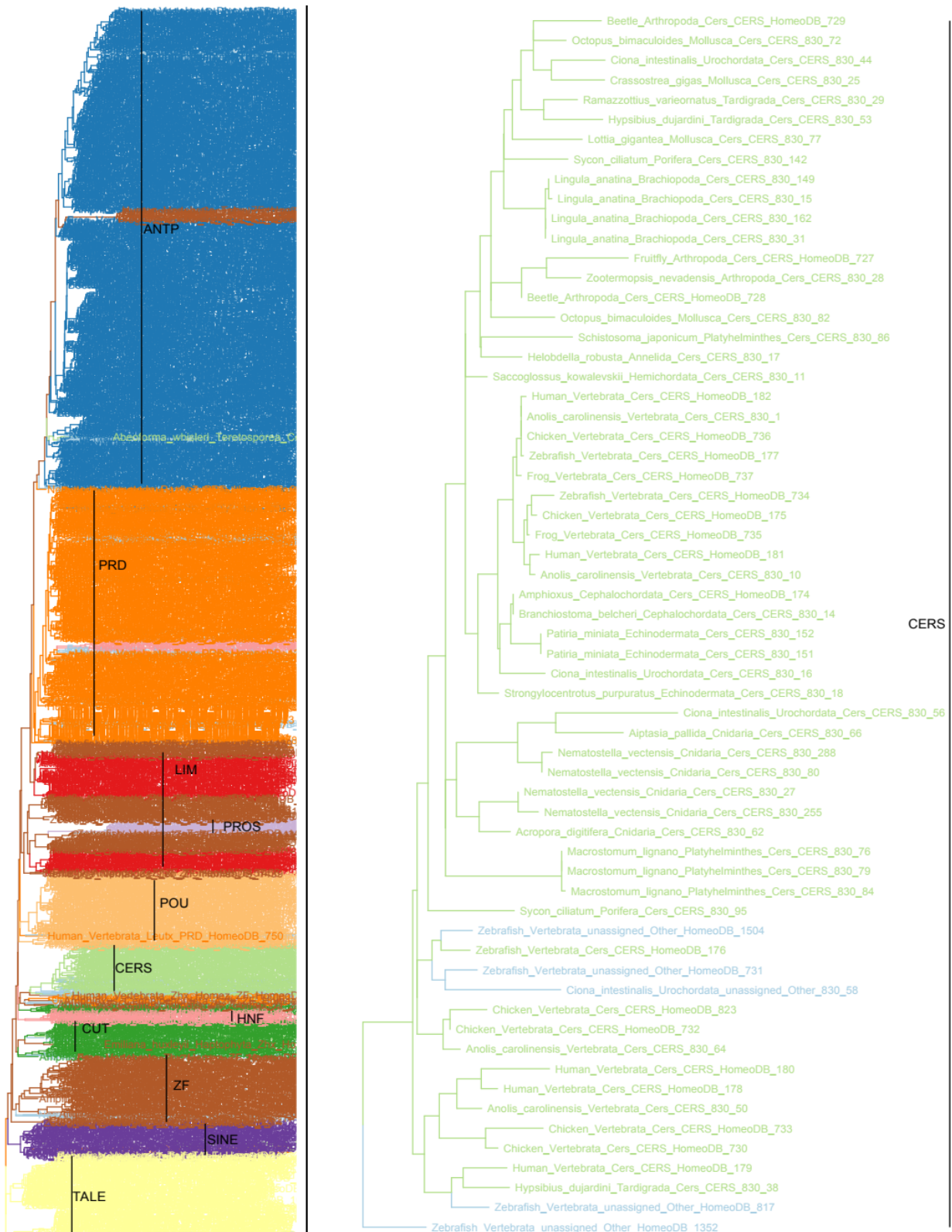


Figure 3.13 CERS-class homeobox is dispersed through all the animal phyla and evolved prior to the origin of animals. There is a duplication event seen in vertebrates, but the rest of the animals have remained low copy number with distinct protostome and deuterostome clades with the exception of urochordates, which is likely an erroneous placement in the gene tree.

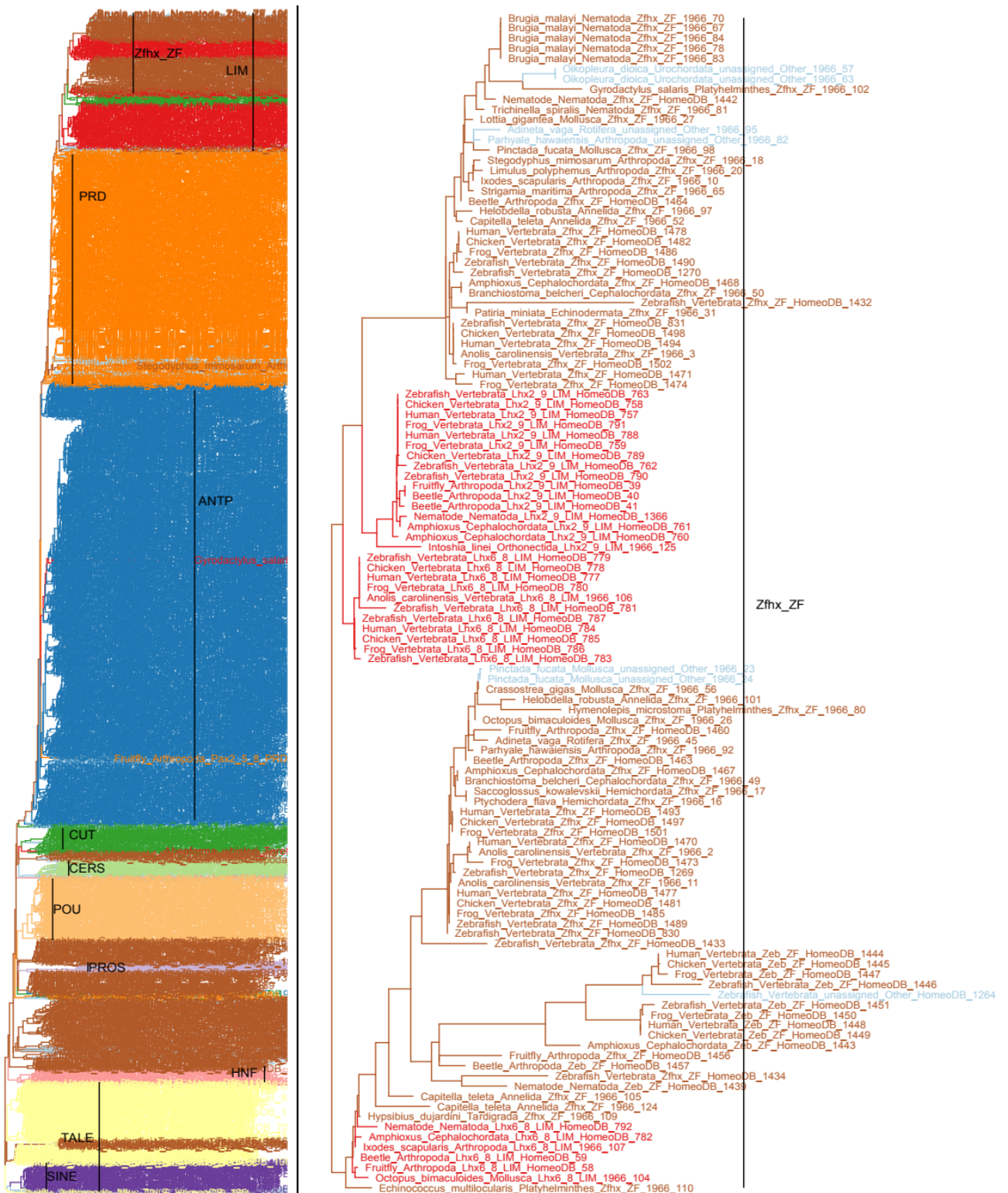


Figure 3.14 The ZF-class of homeobox is classified by having zinc-finger domains as well as the homeodomain (Bürglin & Affolter, 2016). It can have any number of additional domains and varying motifs, and for this reason it can be seen dispersed in paraphyletic clades across the gene tree.

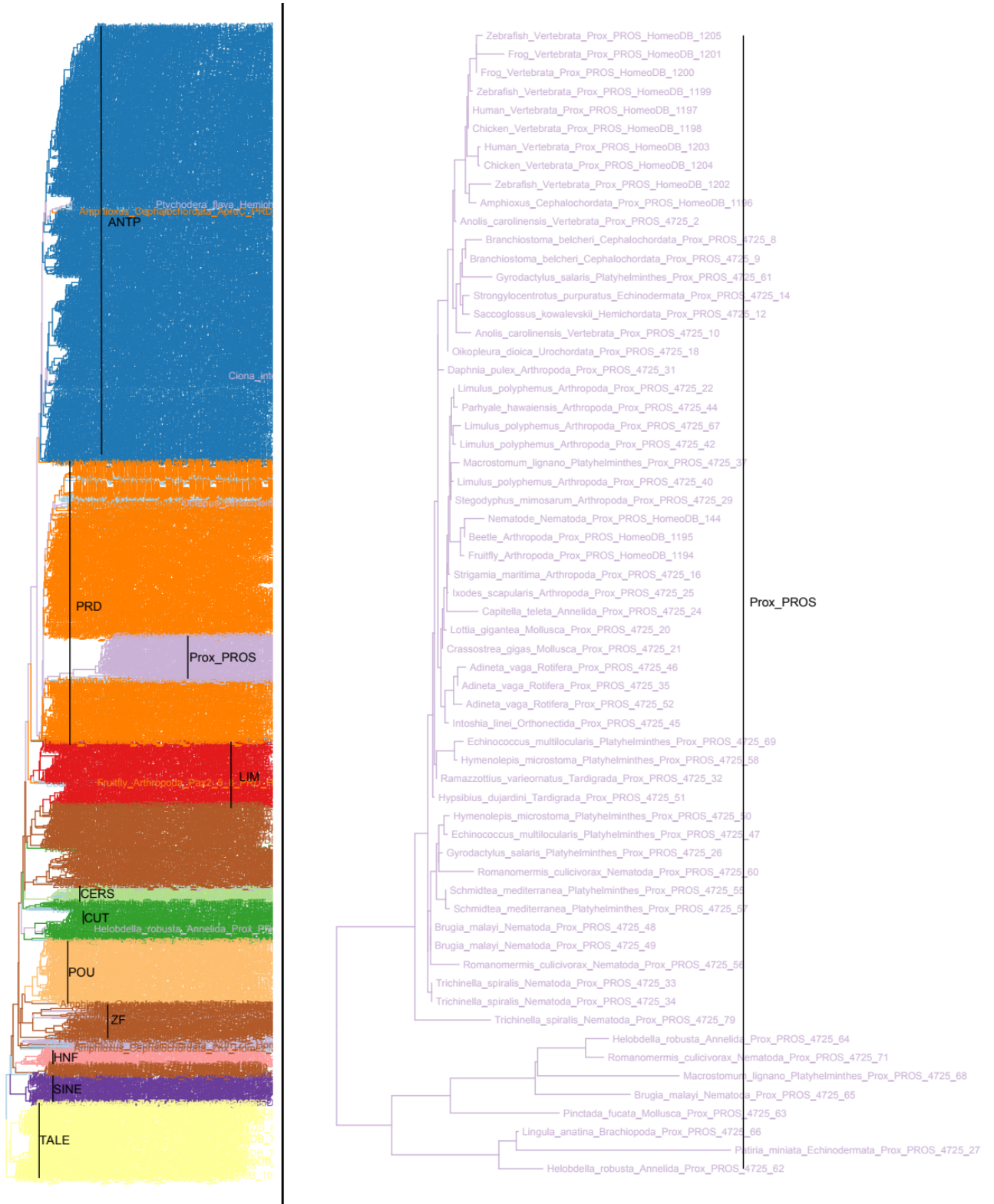


Figure 3.15 The PROS-class has just the one homeobox family: *Prox*. This homeobox is dispersed throughout the animal phyla.

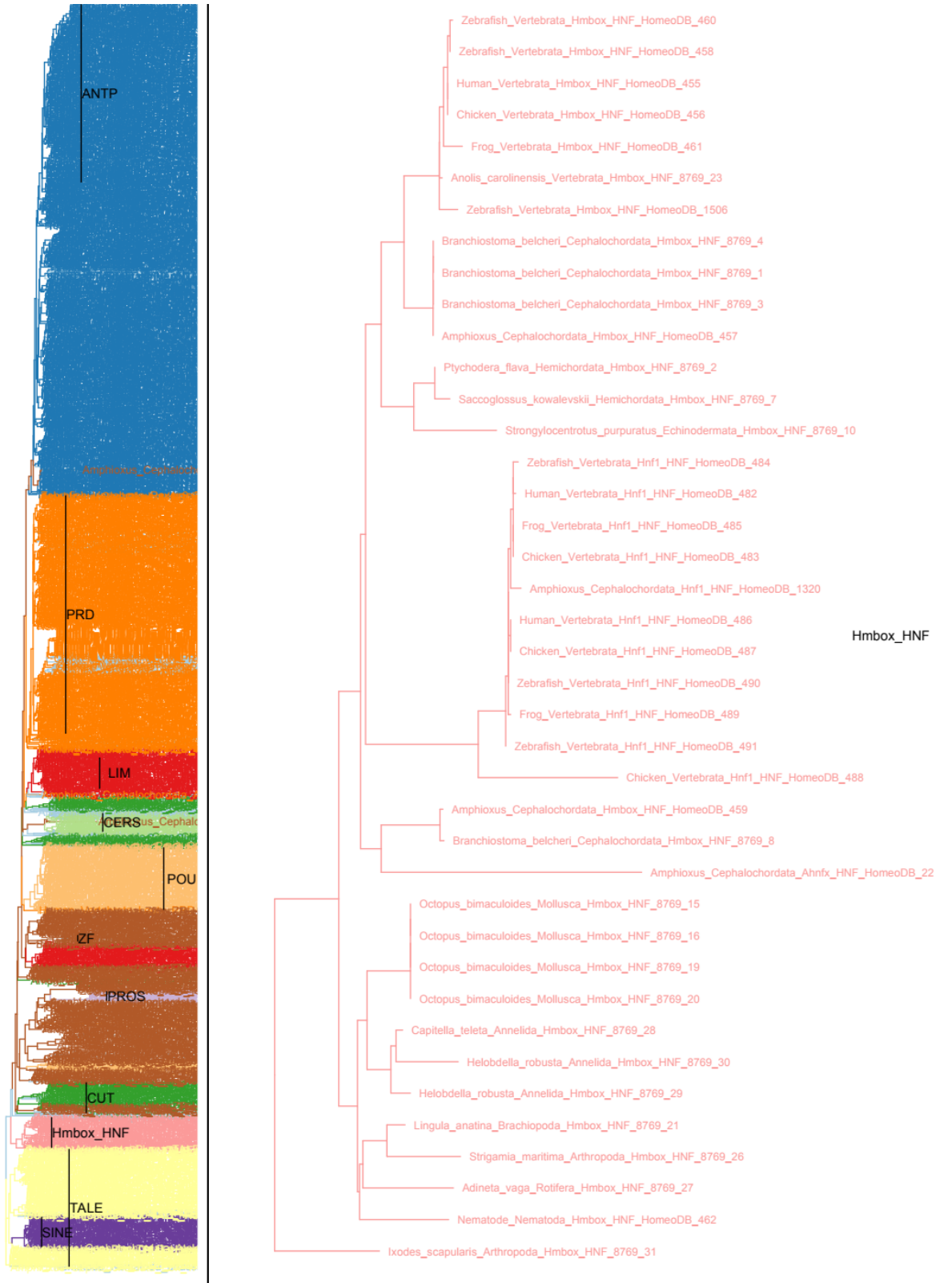


Figure 3.16 The HNF-class Hmbox family is a monophyletic clade. It has a lophotrochozoan specific divergence, with duplications seen in chordates, molluscs and annelids.

Figure 3.17 Presence of each homeobox gene family (gene families limited by HomeoDB classified genes, hence heavy bias towards vertebrates), as seen in each animal species, grouped by phyla. The named homeobox families and classifications only include the homeobox genes that have been identified in HomeoDB. Any novel or unclassified homeobox genes that have been found in this thesis, or that were already identified as homeobox genes but uncharacterised beyond that recognition have been collated under the class *Other*, family unassigned.

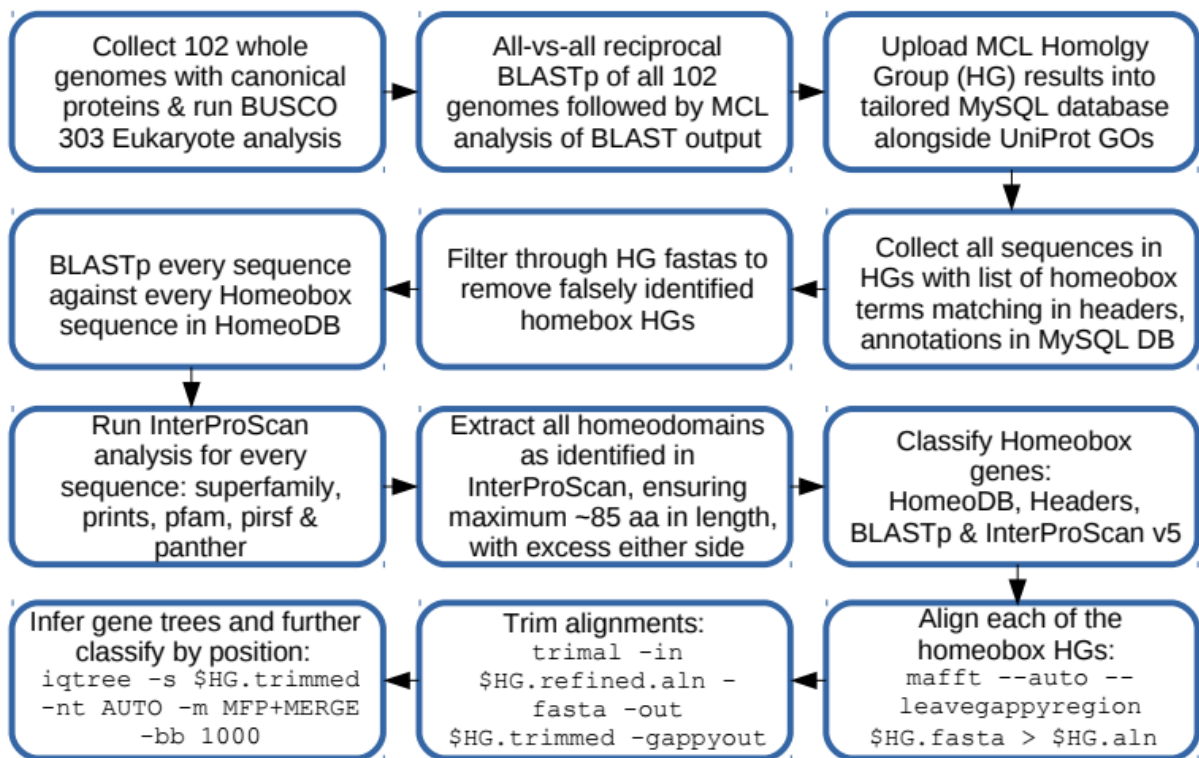


Figure 3.18 The gene tree pipeline including classification and identification of the homeobox HGs, described further in materials & methods.

3.4 MATERIALS & METHODS

COMPARATIVE GENOMICS

Whole genome proteins of 102 eukaryotes, of which 59 are animals were clustered into homology groups (HGs) by reciprocal BLASTp (Camacho *et al.*, 2009) with e-value = 10^{-6} cut off, and then MCL analysis (Enright *et al.*, 2002) with default $\tau=2$ parameters. The whole pipeline is simplified in Figure 3.18.

HOMEBOX DETECTION

Gene ontologies (GOs) were assigned to HGs using the Uniprot API for the sequences downloaded by Uniprot (Bateman *et al.*, 2017). Proteins, HGs, GOs and Panther protein profiles were uploaded to a MySQL database for automatic and accurate retrieval. All HGs with any Homeobox annotations (Panther families) or with classified homeodomain proteins (in FASTA format header descriptions) were downloaded from the in-house comparative genomics database as the whole protein FASTA files.

HOMEODOMAIN EXTRACTION

A Perl program was written to parse the analysis and extract only the homeobox domains (~60 amino acids, with 10 amino acids either side if available). Using InterProScan v5 (Jones *et al.*, 2014) with applications: Panther, PRINTS, Pfam, Superfamily and Pirsf, to select the homeodomain regions.

INITIAL HOMEBOX CLASSIFICATION

Each protein in the HG FASTA files was compared using BLASTp (Camacho *et al.*, 2009) against all the Homeobox Database (HomeoDB2) (Zhong *et al.*, 2008; Zhong & Holland, 2011) domains with e-value = 10^{-6} cut off. Each protein was also analysed using InterProScan 5 with default values to produce a tab separated file. A custom Perl script was developed to parse and evaluate the BLASTp and InterProScan results using the statistics alongside initial protein annotations and key motifs to classify each protein to Homeobox gene family level. Priority was given to >98% matches to the BLASTp results, otherwise the match with higher similarity was selected. In some cases,

particularly for the non-animals and lesser known Homeobox Classes, a clear classification could not be determined, and these were left as “Unknown”.

GENE TREES

For each HG, extracted domains were aligned with downloaded HomeoDB2 (Zhong & Holland, 2011) domain sequences, replacing the sequences from the corresponding sequences, (*Homo sapiens*, *Gallus gallus*, *Xenopus tropicalis*, *Danio rerio*, *Branchiostoma floridae*, *Drosophila melanogaster*, *Tribolium castaneum* and *Caenorhabditis elegans*) using MAFFT (Kato & Standley, 2013) with `--auto` and `--leavegappyregion` parameter and finally refined using MUSCLE `--refine` to improve the alignments where possible (Edgar, 2004). Alignments were trimmed using TrimAL (Capella-Gutiérrez *et al.*, 2009) with `--gappyout` parameter. Other trimming parameters were trialled, but these lost too much sequence in some cases. Trees were inferred using IQ-TREE v1.6.1 (Nguyen *et al.*, 2015) with automatic model finder and 1000 Ultra-fast bootstraps (Minh *et al.*, 2013). The HG sequences were also concatenated alongside the HomeoDB2 sequences to infer a single very large gene tree.

FURTHER HOMEBOX CLASSIFICATION BASED ON PHYLOGENY

In-house software was used to identify the "Unknown" homeobox proteins using the phylogenetic position in the gene tree by comparison to high confidence classified close ancestral/sister nodes. Some proteins remained unclassifiable, these were listed under "unassigned, Other". Classification accuracy was measured by comparing classification of known Homeobox identities from HomeoDB with same species proteins extracted from this pipeline. The homeobox classification accuracy was measured using ROC analysis (Figure 3.2) (W.Zhu *et al.*, 2010).

4 RESAMPLING CORE/CONSERVED WHOLE GENOME LEVEL HOMOLOGY GROUPS TO INFER THE PHYLOGENY OF METAZOA

4.1 SUMMARY

Animals are diverse, we can use homologous genes to infer the history of their relationships at an evolutionary level. Whilst the monophyly of the Metazoan tree of life has been well established, the evolutionary relationships of all the animal phyla are still unresolved. Phylogenomic analyses rely on protein evolutionary models, different tree inference algorithms and taxa selection. The determination of the position of individual phyla within the animal tree of life has become a battle field.

The current phylogenomic approaches, as discussed in chapter 1 have issues, we tackle these issues in this chapter. Here we attempt to infer a robust phylogeny of animals using a robust taxon sampling of whole sequenced genomes and a sophisticated pipeline to select well-conserved genes. This is likely the largest phylogeny using animal genomes to date. With strong statistical support in the ML and BI methods, for each of the models used, for all the datasets; we recovered Ctenophora or Ctenophora+Porifera as the first splitting extant animals and Placozoa+Cnidaria form a sister clade to all bilaterians (Planulozoa). The relationships between the phyla in the proposed Protostomia clade are even more complex. None of the models or conserved gene selections are able to account for the faster evolution seen in Platyhelminthes and Nematodes. In the unexpected event that the highly supported nodes are correct, they date back to earlier, vintage hypotheses that have only in recent years been rejected.

The genomic history in these findings are important to any animal research; if a model, algorithm and gene selection can be used to accurately infer the evolutionary relationships between animals, it can be applied in any animal phylogeny. Understanding the historical relationships between animal phyla and their super-groups is as essential as the question of the origin of life itself. This

research proposes additional hypotheses to be considered, poses more questions and begs for continual research to include more key phyla as in-groups and out-groups in future phylogenomic studies of animals. Furthermore, the experience acquired here has proven that it takes more than a conserved dataset to infer a true animal phylogeny and take into account species with fast evolved genomes.

4.2 BACKGROUND

Metazoa, an opisthokont Kingdom within eukaryotes are extremely diverse with many evolutionary adaptations to feeding and surviving. Eukaryotes are characterised by the possession of organelles (Burki, 2014). Animals are further characterised by their multicellularity and specialised cell types. The animal tree of life as we know it today is a result of decades of phylogenetic research based on anatomical, developmental features, and molecular data (Halanych, 2004; Dunn & Ryan, 2015).

Metazoa are multicellular animals; they can be split into bilateral animals; Deuterostomia, Lophotrochozoa and Ecdysozoa - and those before the emergence of bilateral animals; sponges, ctenophorans, cnidarians and placozoans. The monophyly of metazoans is well-supported with a single point of origin of multicellularity (Dunn *et al.*, 2014; Dunn & Ryan, 2015; Hejnol & Dunn, 2016). Traditionally Porifera is placed as the sister group to all the other animals (Philippe *et al.*, 2009; Pisani *et al.*, 2015; Whelan *et al.*, 2015b, 2015a; Feuda *et al.*, 2017; Simion *et al.*, 2017). Traditionally, it is believed that Hox genes are found in all the animal clades except Porifera and Ctenophora (Holland, 2013; Dunn *et al.*, 2014; Holland, 2015), although recently a putative ParaHox gene has been described in sponges (Fortunato *et al.*, 2015), this is still disputed however (Pastrana *et al.*, 2019).

4.2.1 OPEN ENDED QUESTIONS

As concluded by Laumer *et al.* (2019), there are still so many open ended questions that need answering, including; inferring the first splitting animal, the sister phylum to cnidarians and resolving the phylogenetic relationships behind the individual clades Panarthropoda and Lophotrochozoa (Laumer *et al.*, 2019). It is expected that an increase in the number and morphological diversity of animal genomes and out groups will produce a more accurate phylogeny for the metazoan tree of life, whilst reducing analytical artefacts such as long branch attraction. The selection criteria used to determine the homologous proteins between species will affect the outcome of the phylogenetic relationships between animal phyla. There are various different genes and morphological features that have been used in the past to determine relationships between animals, for example; Cnidaria, Bilateria and Ctenophora all have nervous systems/nerves, whilst Porifera and Placozoa do not - this traditionally places sponges as first splitting animals, unless ctenophores are believed to have

convergently evolved to have nerves, and remains outside Parahoxoa and within Eumetazoa (Dunn *et al.*, 2014; Holland, 2015, 2016). The dispute between Porifera-first (Philippe *et al.*, 2009; Pisani *et al.*, 2015; Whelan *et al.*, 2015b, 2015a; Feuda *et al.*, 2017; Simion *et al.*, 2017) and Ctenophora-first (Dunn *et al.*, 2008; Ryan *et al.*, 2013; Mentel *et al.*, 2014; Moroz *et al.*, 2014) is just one of many conflicts that need to be resolved.

As described in the first chapter, there are some methodological gaps in current phylogenies, causing concerns and inconclusive results, these include: orthology markers, molecular heterogeneity and missing data/gene loss.

4.2.2 ISSUES IN PHYLOGENIES

Orthology and paralogy is largely determined by a level of similarity, using BLAST (basic local alignment search tool) or RBH (reciprocal best hits), or computationally expensive tree inferences. Orthologous markers have been dominated/restricted by mitochondrial or ribosomal genes, particularly in invertebrates. As the number of available sequences and species has increased, these markers have become less useful in evolutionary inferences, and have actually been lost for some species altogether (Ballesteros & Hormiga, 2016). There is currently no one marker useful to all datasets, especially as the variation and quantity of data increases in this genomic era. Incorrectly inferred orthologues/paralogues can have drastic effects on phylogenomic analyses (Kocot *et al.*, 2017). For some genes, orthology markers are not possible to infer. This can be due to issues in pairwise/multiple alignment, phylogenetic inferences, and general technical limitations in the methods available to infer orthology (Luo *et al.*, 2018).

Heterogeneity is a molecular characteristic providing information in genetic variation. It is this difference between sequences which can be used to describe how one sequence has changed to become another. Heterogeneity and homogeneity together are necessary to infer evolutionary history between species. However, too much heterogeneity is a problem. This problem arises when species or groups of organisms evolve at very different rates. Fast-evolving species are frequently seen as 'unstable' and these are then eliminated from analyses to reduce the skewing of the phylogeny or artefacts such as long branch attraction (LBA), where the composition of amino acids between species are biased, grouping unrelated species together. A probable cause for this is the heavy weighting of GC content in a gene (Nosenko *et al.*, 2013; Kocot *et al.*, 2017). One recommended model to reduce the effect of compositional bias in heterogeneity is the use of the rate model 'Gamma' (Hejnol *et al.*,

2009). Another rate model is the "FreeRate model"; this model does not make any parametric statistical assumptions about the data, unlike most models, and it can be used in conjunction with Gamma. This model determines a rate specific to the dataset being used, and works the same way as Gamma, but without a parametric framework to begin with (Soubrier *et al.*, 2012). Aside from rate models, data-recoding can also be used to reduce the problem of heterogeneity, although this limits the type of model that can be used, and the model must take into account the specific data-recoding used (Feuda *et al.*, 2017). Data recoding is a method of placing amino acids into bins/categories to reduce the amount of erroneous heterogeneity that may bias the analyses (Susko & Roger, 2007).

Finally, missing data is a colossal obstacle when considering inclusive datasets to infer phylogenetic relationships. Smaller datasets, without missing data produces more robust phylogenies than larger datasets with missing data (Kocot *et al.*, 2017). Missing data accounts for incomplete genes, lost genes, and genes that have undergone a loss of amino acid sites. Missing data leads to LBA, by generating misleading positive signals, and causing misalignments in the sequences, decreasing the ability to resolve and limiting multiple substitution detection (Roure *et al.*, 2013). Recommendations to resolve this issue are most commonly to eradicate the taxa or the genes with missing data altogether, and to indicate in the research the level of missing data (Nosenko *et al.*, 2013; Roure *et al.*, 2013; Egger *et al.*, 2015; Whelan *et al.*, 2015a, 2015b; Cannon *et al.*, 2016; Kocot *et al.*, 2017). This is an unavoidable issue in some cases, for example isolated species, in which the dataset is limited (Roure *et al.*, 2013).

Here, with these issues in mind, we produce several comparable phylogenies using various dataset appropriate models, with well-conserved canonical genes, minimal missing data, and a large, varied taxon sampling to infer an accurate evolutionary relationship between animal phyla.

4.3 RESULTS & DISCUSSION

In brief (described further in Materials & Methods section), the genes for all the datasets were extracted from the comparative genomics pipeline developed in chapter 2. Canonical proteins from whole genomes ensure complete datasets. Conserved homology groups (HGs) selected from the MCL clustered, reciprocal BLAST+ ~2.6 million proteins were determined by presence in all the taxa available. Orthology assignment was reduced by selecting single or low copy genes across all the taxa, and heterogeneity bias reduced by using dataset specific and gamma rate models (R# and G noted in Table 4.1). Informative statistics for the tree inferences are shown in Table 4.1 and a summarised phylogeny in Figure 4.1. The summarised phylogeny as shown in Figure 4.1 is not to be accepted as is. Likely due to LBA and fast evolution, the Rotifera-Orthonectida-Platyhelminthes clade is sister to Ecdysozoa, whilst it is expected to be placed within the Lophotrochozoa clade (Kocot *et al.*, 2017). This will be discussed later in the results.

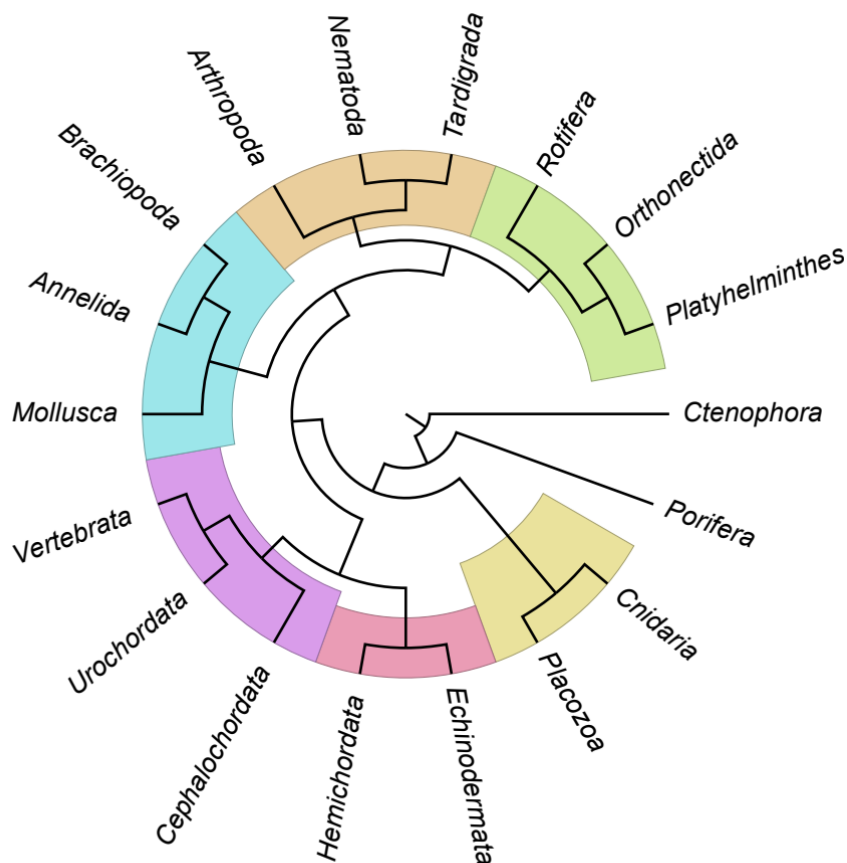


Figure 4.1 A summarised consensus for the most commonly found animal phyla positions from the ML methods for all the protein sets described in methods & materials: HG60, HG90, HG101 and BUSCO293.

The number of parsimony informative and constant sites in the two tree inferences which included additional, non-whole genome taxa (trees HG60+ and HG90+) were very low, and the trees likely to be very artefactual, so will not be included in this discussion in terms of inference results (Figure 4.2 & 4.3). HG101 (Figure 4.4) had the highest number of constant sites and homology groups (HGs) and the highest number of whole genome taxa. The shared HGs obtained in the HMM detection methods used by BUSCO 303 Eukaryote dataset contained the largest number of HGs and total sites (BUSCO293). However, there were many more gaps and many of the HGs were missing taxa. *Macrostomum lignano* had to be excluded from the analysis altogether due to gene loss in the case of the BUSCO obtained phylogenies (BUSCO293) (Figure 4.5 & 4.6).

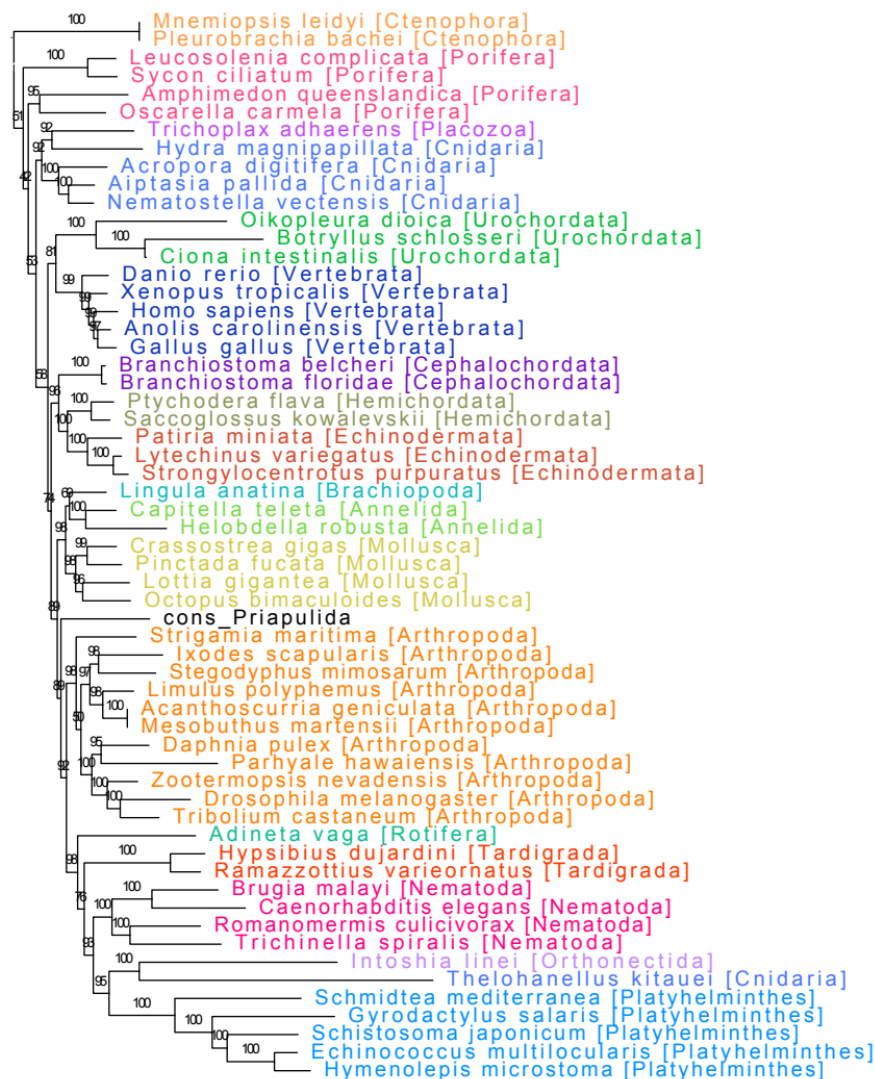


Figure 4.2 ML phylogeny of the HG60+ dataset with LG+R8 model. Additional protein data from non-genome sequences such as priapulids were used as a consensus to fill animal phylum gaps in HG60+ and HG90+. Non-animal outgroups collapsed.

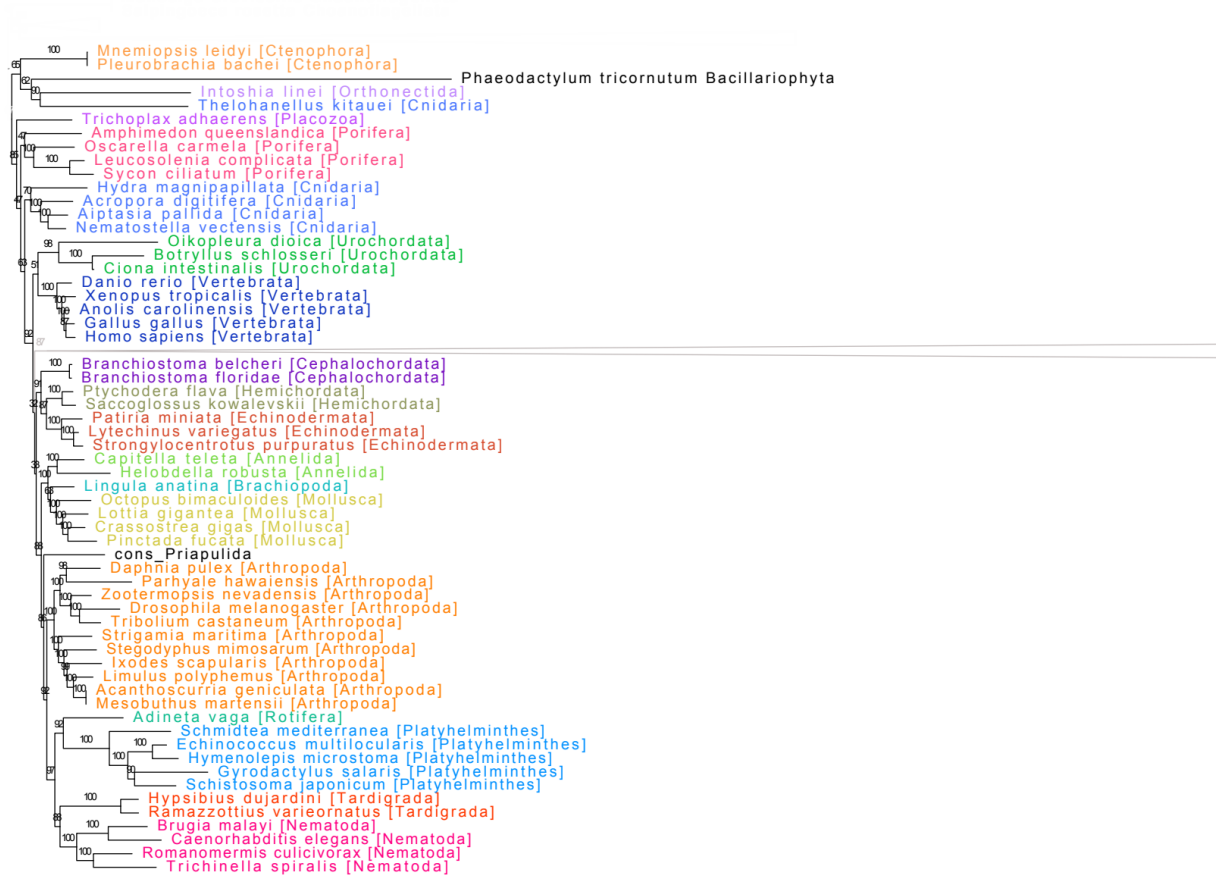


Figure 4.3 ML phylogeny of the HG90+ dataset with LG+F+R7 model. There is a misplacement of outgroups between vertebrates and cephalochordates and elsewhere in the animal clade are very erroneous and artefactual, displayed as collapsed triangle, and as seen with the Bacillariophyta.

Table 4.1 Amino acid based statistics prior to the tree inference methods. Trees HG60+ and HG90+ had no constant sites, whilst Tree HG101 had the most constant sites.

Tree ID	No. of Taxa	Parsimony informative sites/Total number of columns	Model used in IQ-Tree or PhyloBayes	No. of HGs	No. of constant sites
HG60 (Figure 4.7)	100	14,182 / 15,196	LG+R10	60	225
HG60+ (Figure 4.2)	107	3,489 / 3,689	LG+R8	60	0
HG90 (Figure 4.8)	100	21,134 / 22,258	LG+R8	90	206
HG90 (Figure 4.9)	100	9,562 / 10,356	LG+F+R7	90	145
HG90+ (Figure 4.3)	115	3,242 / 3,280	LG+F+R7	90	0
HG101 (Figure 4.4)	102	28,984 / 34,943	LG+F+R7	101	1,794
HG101 (Figure 4.10)	102	28,984 / 34,943	CAT+GTR(BI)	101	1,794
BUSCO293 (Figure 4.5)	101	43,069 / 46,759	LG+F+G	293	754
BUSCO293 (Figure 4.6)	101	43,069 / 46,759	LG+C20+F	293	754
BUSCO293 (Figure 4.11)	101	43,069 / 46,759	CAT+GTR(BI)	293	754

4.3.1 SISTER TO ALL OTHER ANIMALS: CTENOPHORA, PORIFERA OR (CTENOPHORA + PORIFERA)

There were high ultrafast bootstrap supporting (UFBS) values (99+/100) for Ctenophora as sister group to all other animals in all the phylogenies (Figures 4.4, 4.7, 4.8 & 4.9) except for BUSCO293 (Figures 4.5, 4.6 & 4.11). BUSCO293 dataset suggested an alternative hypothesis that Porifera are sister to Ctenophora in a joint clade that is sister to all other animals with high UFBS values. This Ctenophora-Porifera clade has been proposed, without certainty, in the genome paper of *Mnemiopsis leidyi* with both ML and BI support, although a small taxon sampling was used at the time (Ryan *et al.*, 2013). Ctenophora as sister to all other animals is a less popular hypothesis (Dunn *et al.*, 2008; Ryan *et al.*, 2013; Mentel *et al.*, 2014; Moroz *et al.*, 2014) than that of sponges as sister to all other animals (Philippe *et al.*, 2009; Pisani *et al.*, 2015; Whelan *et al.*, 2015b, 2015a; Feuda *et al.*, 2017; Simion *et al.*, 2017). The most common causes for concern are the different phylogenies inferred by ML and BI methods. BI is often called out for being biased, since it requires prior distribution input and probabilities, although it does have the ability to incorporate more complex substitution models. Ctenophora as sister to all other animals is more frequently recovered when ML

methods are used (Ryan *et al.*, 2013; Whelan *et al.*, 2015a; Laumer *et al.*, 2019), and sponges when BI in place (Whelan *et al.*, 2015a, 2015b, 2017).

The Ctenophora+Porifera sister hypothesis suggests that there was a genetic toolkit for a simple neural system in the ancestor to all extant animals that was lost by reduction in placozoans and sponges, and a more elaborate neural system evolved independently in ctenophores and eumetazoans (Ryan *et al.*, 2013; Ryan & Chiodin, 2015). This is also supported by the Ctenophora neural system differing more to the rest of the animals' neural systems (Moroz *et al.*, 2014). Many of the genes and patterns found in other animal neural systems are missing in ctenophores. Ctenophores have an independent neuromuscular transmitter and receptors to other animals, which supports an independently evolved nervous system (Moroz *et al.*, 2014). Further biological evidence for the Ctenophora+Porifera is also evidence for a Parahoxozoa clade (Ryan *et al.*, 2010; Laumer *et al.*, 2019), in which neither Ctenophores nor sponges have a parahox homeodomain complement (Ryan *et al.*, 2006, 2010, 2013, Dunn *et al.*, 2014, 2015; Pastrana *et al.*, 2019). There are additional biological evidences that support Ctenophora as the sister clade to all other animals, as well as evidence for Parahoxozoa. For example, a reduced and derived mitochondria in Ctenophores (Pett *et al.*, 2011; Kohn *et al.*, 2012), and a slow evolving mitochondrial genome in sponges and cnidarians (Wang & Lavrov, 2007; Halanych, 2015). It is plausible to suggest the mesodermal cell types in ctenophores evolved independently to those in Bilateria, with divergence ahead of Planulozoa (Ryan *et al.*, 2013). However, the ctenophores as sister to all other animals relationship is still disputed to be systematically erroneous. One reason for this is missing data in both sponges and ctenophores, as discussed by Dunn *et al.* (2015). A lot of information gleaned from the biology of both sponges and ctenophores has been studied in common with bilaterians with the implication that these non-bilaterians are primitive animals (Dunn *et al.*, 2015). Pisani *et al.* (2015) further elaborates on this basis for systematic errors in phylogenies and suggested that the analysis of gene content is potential for bias given that genes that have been lost in most species is necessary information (Pisani *et al.*, 2015). This bias is not a big issue in this study, since the dataset selected for inference was filtered to use HGs present in most species.

4.3.2 CNIDARIA + PLACOZOA - PARAHOXOA OR PLANULOZOA (PLACOZOA + EUMETAZOA)

The phylogenies supported two different hypotheses, both in support of Parahoxoa: Cnidarians as sister to Placozoa in a clade as sister to all bilaterians, or placozoans as sister to Planulozoa

(cnidarians + bilaterians clade). Only recently, a study by Laumer *et al.* (2018) looking at the relationships between the non bilaterian animals, investigated the use of genes with non-compositional bias/less compositional heterogeneity in phylogenomic inference with multiple placozoans. They found that cnidarians were the closest extant relative of the placozoan species. There are two possible theories used to explain this relationship: either placozoan ancestors had muscular, nervous and gastrointestinal systems, which were lost in extant placozoans or the bilaterian and cnidarian ancestor was placozoan-like, with cnidarians and bilaterians evolving those features independently (Laumer *et al.*, 2018); this was further supported in 2019, although equal support was seen for a distinct clade of Planulozoa, with Placozoa sister to a clade of bilaterians and cnidarians (Laumer *et al.*, 2019).

4.3.4 DEUTEROSTOMIA

We recovered a frequently agreed upon phylogeny for Deuterostomia, with the key lineage grouping the ambulacrarians (hemichordates and echinoderms) sister to chordates (Lowe *et al.*, 2015; Simakov *et al.*, 2015). Within the chordates, the most commonly recovered internal nodes grouped cephalochordates as sister to Olfactores (vertebrates and urochordates). Deuterostomes are well characterised by the homologous shared gill slits (Lowe *et al.*, 2015; Peterson & Eernisse, 2016).

AMBULACRARIA

In every tree inference from the analyses, Ambulacraria remains as a stable relationship between hemichordates and echinoderms with very supportive 100% ultra-fast bootstrap (UFBS) values. Ambulacraria is a key lineage in understanding the evolution of chordates and other deuterostomes (Cannon *et al.*, 2014; Hejnal & Lowe, 2014; Simakov *et al.*, 2015).

UROCHORDATA/CHORDATA/CEPHALOCHORDATA + VERTEBRATA

Although the internal positions of chordates are thought to be well understood, there are still some small conflicts. This can be seen in all the phylogenies except datasets HG101, HG101-1 and BUSCO293 (LG+C20+F) (Figures 4.4, 4.6 & 4.10). In the differing gene sets and proteins models used to infer the phylogenies in this study, very few of the positions including urochordates and cephalochordates concur. Where cephalochordates are recovered as sister group to vertebrates, the urochordate species can be placed erroneously with a clade of protostomes, or as a sister clade to the remaining deuterostomes. When Urochordates are recovered as sister group to vertebrates, a similar situation occurs with the cephalochordates. This has been observed in phylogenomic analyses prior to

this one (Marlétaz *et al.*, 2006; Delsuc *et al.*, 2008), and has been explained as an artefact of long branch attraction (LBA) (Philippe *et al.*, 2005b). Urochordates are much faster evolving animals (Paps *et al.*, 2012) than vertebrates and whilst molecular data supports cephalochordates as sister to vertebrates, the fossil record does not (Philippe *et al.*, 2005b, 2005a). Extant deuterostomes are so morphologically different, that close fossil relatives are indistinguishable between them (Delsuc *et al.*, 2006).

4.3.5 ECDYSOZOA

All of the tree inferences, in the datasets generated here, conflicted with the traditional monophyletic ecdysozoan hypotheses (Philippe *et al.*, 2005a; Paps *et al.*, 2009; Dunn *et al.*, 2014; Laumer *et al.*, 2015; Giribet & Edgecombe, 2017) with the clade groups arthropods, tardigrades, nematodes and other phylum. In the case of the datasets used here, Ecdysozoa has become a paraphyletic clade. Unexpectedly, in this scenario, platyhelminthes, Orthonectida and Rotifera have nested as sister clade to tardigrades and nematodes, forming a sister clade to the arthropods.

ARTHROPODA + NEMATODA + TARDIGRADA

Internal ecdysozoan relationships have never been fully resolved (Janssen *et al.*, 2014; Giribet & Edgecombe, 2017; Yoshida *et al.*, 2017). Some analyses place tardigrades as sister to arthropods, whilst others place tardigrades as sister to nematodes (Yoshida *et al.*, 2017). In the case here, tardigrades have been placed as sister to nematodes for every dataset and model. However, whilst the Ultra Fast Bootstrap Support (UFBS) values are high (>85%), they are not as comparably high as the rest of the nodes in any of the inferences (100%). Nematodes and Platyhelminthes have been seen grouped together in molecular data analysis previously (Simakov *et al.*, 2013), which has been thought to be due to having fewer informative residual indels than closer related phyla (Simakov *et al.*, 2013).

4.3.6 LOPHOTROCHOZOA

ROTIFERA + PLATYHELMINTHES + ORTHONECTIDA (PLATYZOA + ORTHONECTIDA)

Platyhelminthes have a higher rate of genomic turnover (as seen in chapter 2) than the other lophotrochozoan phyla; annelids, brachiopods and molluscs, attracting a divergent clade (including it's

closest living relatives; Orthonectida and Rotifera) to the base of the ecdysozoans alongside fast-evolving nematodes (Simakov *et al.*, 2013). Lophotrochozoans in general share some bilaterian core gene repertoires with deuterostomes that ecdysozoans and platyzoans (platyhelminthes and rotiferans) do not (Luo *et al.*, 2018). This insight can describe two possible scenarios: platyzoans and ecdysozoans independently lost these gene sets, or they shared an ancestor which had already undergone the reductive evolution - this being the less popular hypothesis (Struck *et al.*, 2014; Fröblius & Funch, 2017; Luo *et al.*, 2018). The trees (both inferred using BUSCO293 dataset) actually placed the lophotrochozoan phyla (molluscs, annelids and brachiopods) as sister to arthropods and more recently related to deuterostomes, whilst recovering a distinct clade for the platyzoans and the remaining ecdysozoans. The UFBS was lower than the average for these nodes, and the result is more than likely caused by missing representatives of the 303 orthologues within the latter taxa, causing a LBA effect. Orthonectids were once described as non-animal organisms which lacked specialised and differentiated cell types. Nowadays, it is agreed that they are "highly simplified" bilaterians who have undergone reductive evolution in adaptation to their parasitic way of life (Mikhailov *et al.*, 2016). These orthonectid characteristics have also had an impact in the grouping with platyhelminthes and rotifers in this way, leading to LBA. Orthonectids are usually recovered in a clade more closely related to annelids than platyhelminthes (Schiffer *et al.*, 2018). Without the other gnathiferan phyla included in this analysis (Chaetognatha, Gnathostomulida and Micrognathozoa), however, the results seen here, grouping rotifers close to platyhelminthes are reasonably supported (Fröblius & Funch, 2017). Platyhelminthes are often recovered as a paraphyletic clade within Lophotrochozoa (Kocot, 2016), but just as frequently as a monophyletic clade (Egger *et al.*, 2015). The BI on dataset HG101 shows strong support for many of the clades seen with the ML analysis. The `meandiff` (the average discrepancy between bipartitions) for two pb chains of 2112 trees each was 0.09 (Figure 4.10). The resulting phylogeny is increasingly promising. Ctenophora is still recovered as one of the extant first splitting animal phyla, followed by Porifera. The cnidarian-placozoan clade is present prior to bilaterian emergence. There is high support for traditional clades in the 3 key bilaterian lineages: Ecdysozoa, Lophotrochozoa and Deuterostomia, with Platyhelminthes and Orthonectida placed among lophotrochozoans with high bootstrap supports (98/100).

Both ML and BI methods struggled to place the cnidarian *Thelohanellus kitauei* with the other Cnidaria. The most significant observation is the placement of the rotifer, orthonectid and flatworms. BI places these within Lophotrochozoa, which is more frequently seen in other animal phylogenies. The relationship platyhelminthes, orthonectids and rotifers have with the other lophotrochozoan, or even protostome, phyla remains uncertain.

THE REST OF LOPHOTROCHOZOA: ANNELIDA + (MOLLUSCA + BRACHIOPODA)

Conventionally, Annelida, Mollusca and Brachiopod have always claimed status within a Lophotrochozoa super-group. This is well supported in all of the phylogenomic inferences recovered. More than one topology was recovered, describing the relationship between molluscs, annelids and brachiopods, equally frequently placing molluscs as sister to a monophyletic clade of annelids and brachiopods (Kocot *et al.*, 2017) or annelids as sister to brachiopods and molluscs. The usual consensus is to place annelids as the outer-most sister to the other two phyla (Paps *et al.*, 2009; Luo *et al.*, 2015; Kocot, 2016; Kocot *et al.*, 2017).

Overall there were small discrepancies in the results between inferences for the same taxon samples, the factors leading to these differences were the selection criteria for the genes and the inference models used. In order to avoid certain compositional bias, the datasets were selected to be as ancestrally conserved as possible, as mostly single copies across all the phyla. Additionally, the use of as many single copy HGs as possible reduced the impact orthologue determination may otherwise have. The models chosen were either dataset appropriate as suggested (Feuda *et al.*, 2017) for animal phylogenomic inferences, or automatically determined using ModelFinder (Kalyaanamoorthy *et al.*, 2017) implemented in IQ-TREE (Minh *et al.*, 2013; Nguyen *et al.*, 2015). Whilst none of the trees were wholly concurrent, together, where they agreed on relationships (Figure 4.1), they supported the key lineages, and described the more internal relationships as well. The taxon sampling was too disparate to resolve some of the internal nodes within Metazoa, and there are definitely important phyla missing; including these in future phylogenomic analyses will produce a more robust and reliable phylogeny, that will likely produce fewer erroneous and discrepant results.

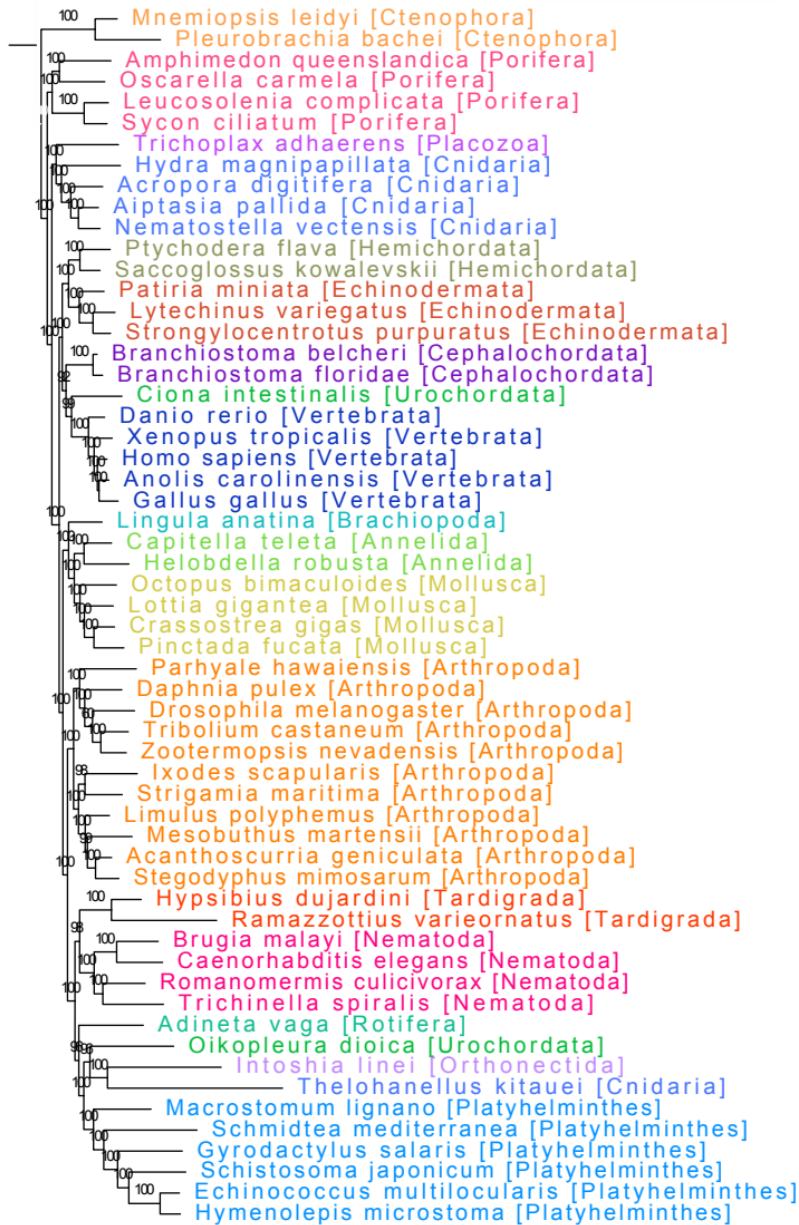


Figure 4.4 ML phylogeny of the HG101 dataset using LG+F+R7 models. Every node has high UFBS values >98%. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

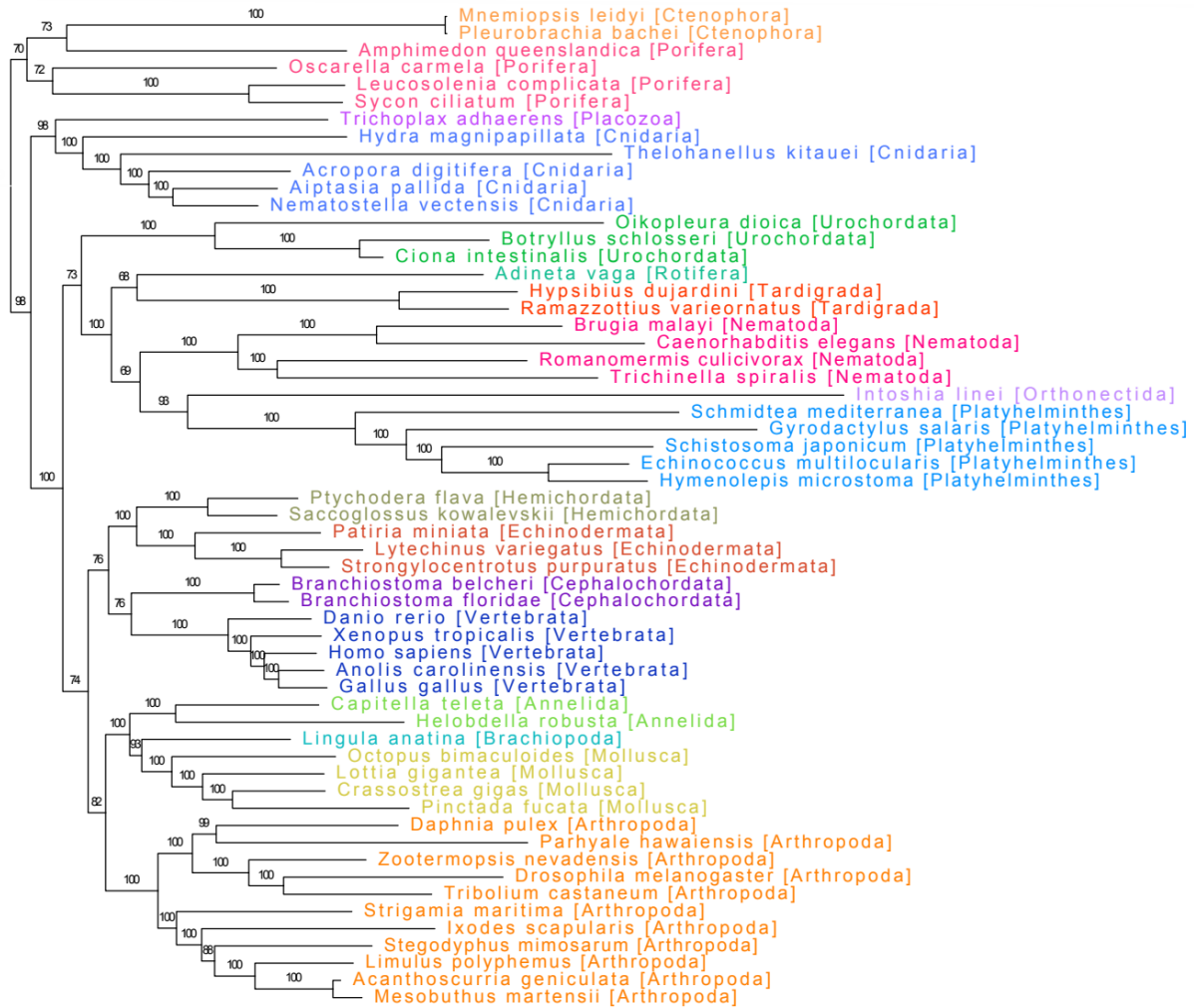


Figure 4.5 ML phylogeny of the BUSCO293 dataset using 293/303 BUSCO orthologues with LG+F+G model. Recovered a clade for Ctenophora and Porifera last common ancestor as first splitting extant lineage. This particular topology has lower (~70% UFBS) support for those nodes and a clade shared with all the fast evolving species. This phylogeny also recovers the shared non-bilaterian clade for Placozoa and Cnidaria, diverging before the emergence of bilaterians. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

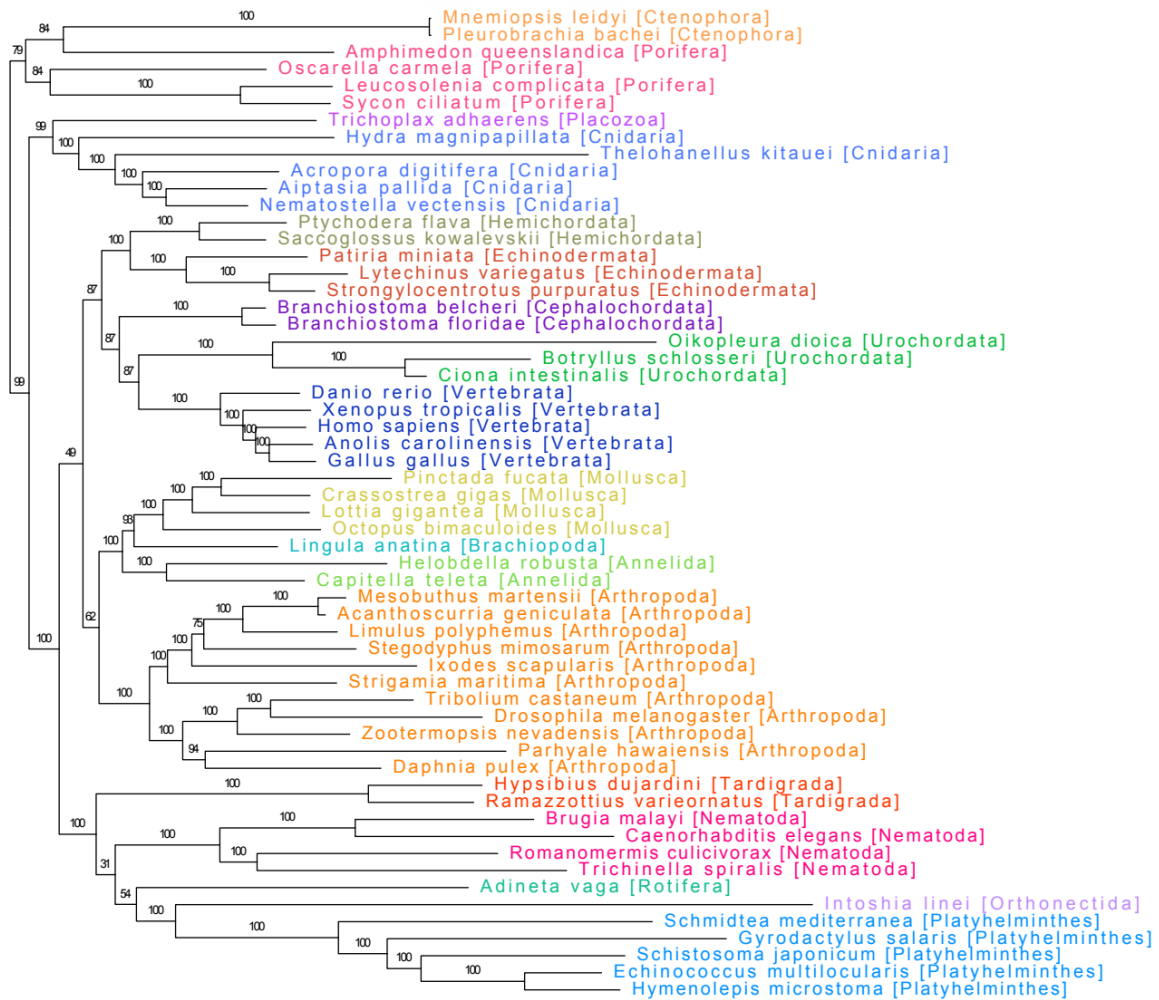


Figure 4.6 ML phylogeny of the BUSCO293 dataset using 293/303 BUSCO orthologues with LG+C20+F model. Here a clade for Ctenophora and Porifera last common ancestor as first splitting extant lineage was recovered, and a shared clade for Placozoa and Cnidaria last common ancestor diverging before bilaterians emerged. There is poor support (UFBS <50%) for the unusual recovery of the three key bilaterian clades, but high internal support for each phyla. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

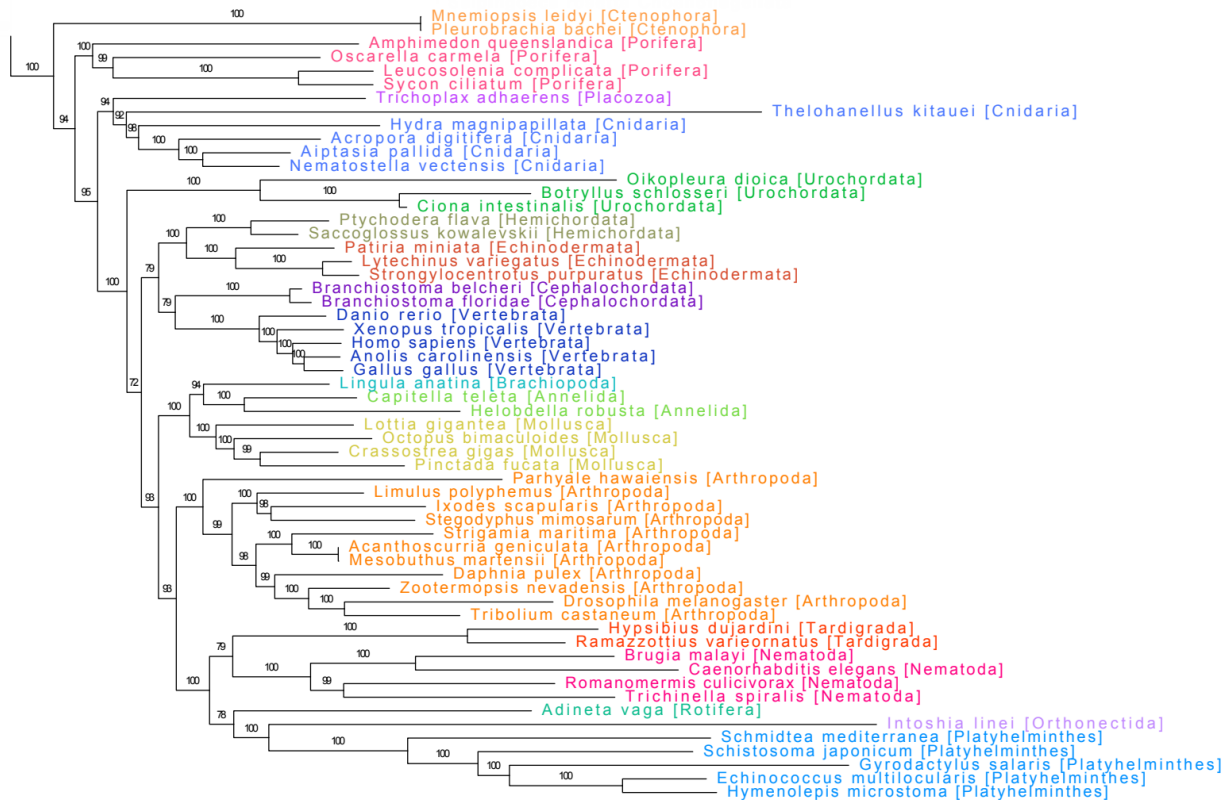


Figure 4.7 ML of HG60 dataset using LG+R10 model. Urochordates are unexpectedly recovered as first bilaterians, but with low UFBS support (72%) when compared to the rest of the nodes. Rotifera, Orthonectida and Platyhelminthe have high internal supporting relationships between them (100% UFBS), but the divergence of their last common ancestor has lower support, and this appears to be an artefact of LBA. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.



Figure 4.8 ML of HG90 dataset using LG+R8 model. The phylogeny recovered here using a different dataset, but different model to HG60 in Figure 4.7 is identical. The UFBS values are higher. The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

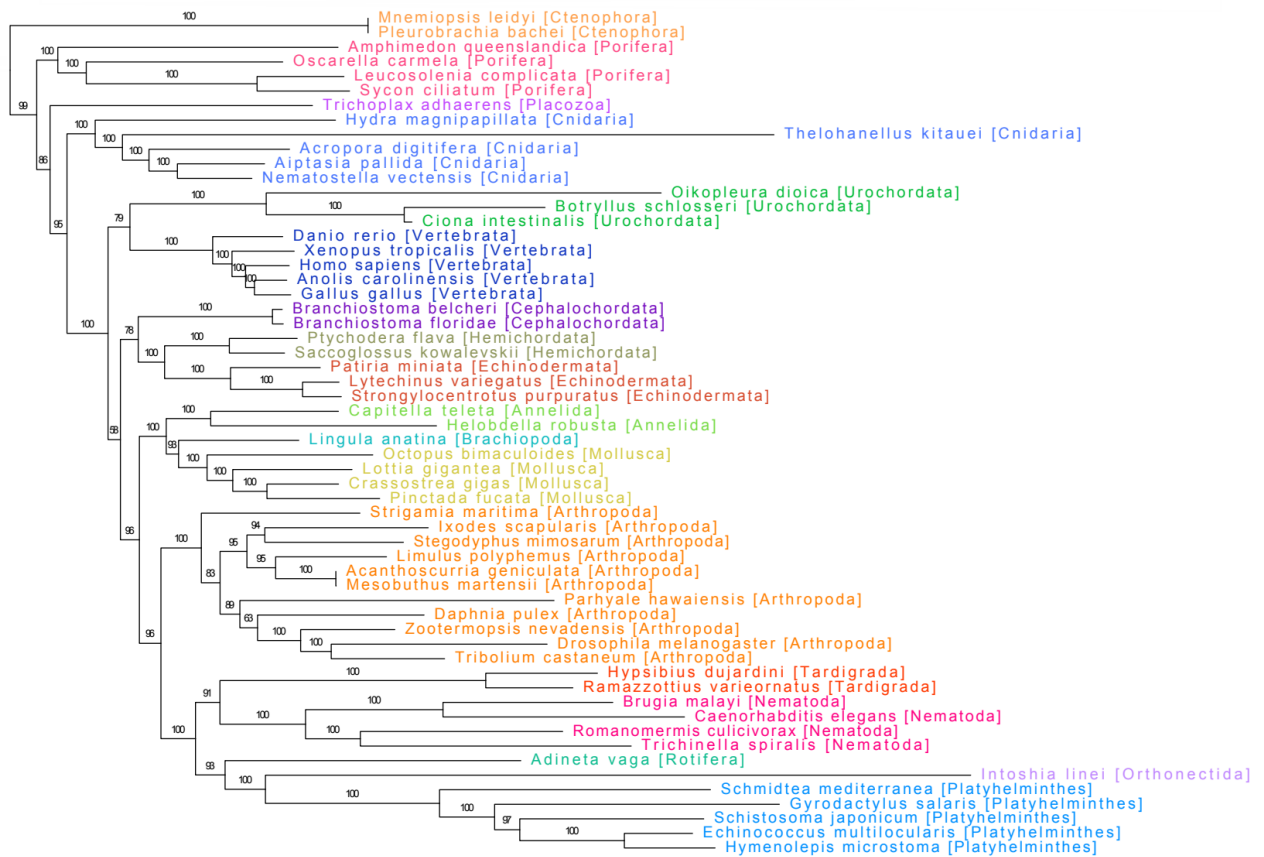


Figure 4.9 ML of the HG90 dataset using LG+F+R7 models. Whilst using the same dataset as in Figure 4.8, but different models, the phylogeny recovered here is more plausible. with urochordates recovered with vertebrates. The unexpected divergence prior to the cephalochordate clade is not well supported, and this is likely an artefact. Furthermore there is still the LBA issue occurring with the "fast evolving species". The non-animal out-group has been collapsed because the internal relationships inferred are irrelevant.

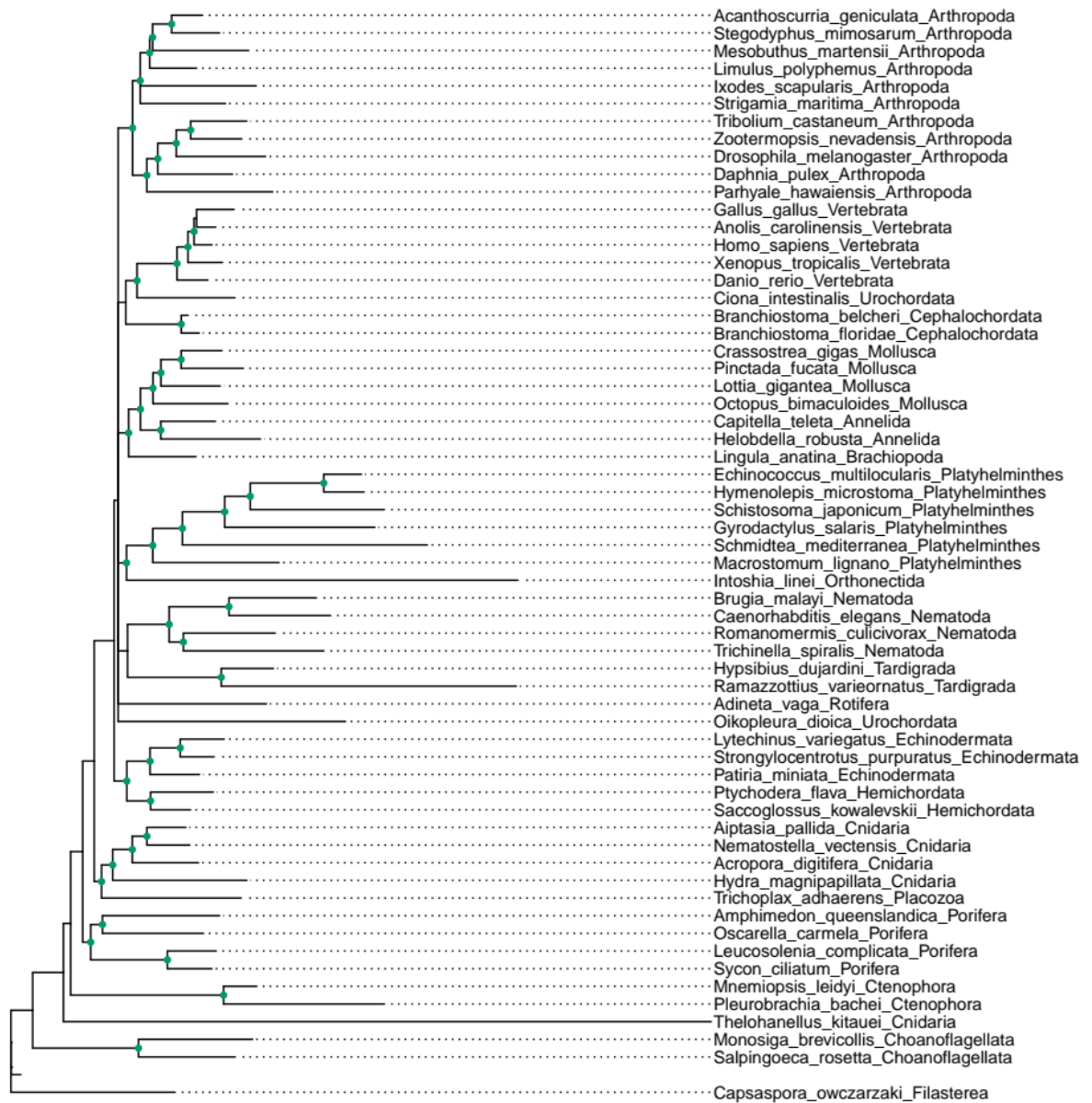


Figure 4.10 Using the same dataset as HG101 in Figure 4.4, and Bayesian Inference (BI) with CAT+GTR model. There is high support for traditional clades in the 3 key bilaterian lineages: Ecdysozoa, Lophotrochozoa and Deuterostomia, with Platyhelminthes and Orthonectida are placed among lophotrochozoans with high bootstrap supports (98/100). There is a polytomy between the ambulacrarians and the chordates, both diverging from the last common bilaterian ancestor. Node points in green indicate posterior probabilities > 0.9. Using PhyloBayes 4.1: 2 Pb chains were run in parallel, with 2112 trees each. Maxdiff 0.98 and meandiff 0.09. The remaining outgroups have been collapsed to direct focus to the animals only.

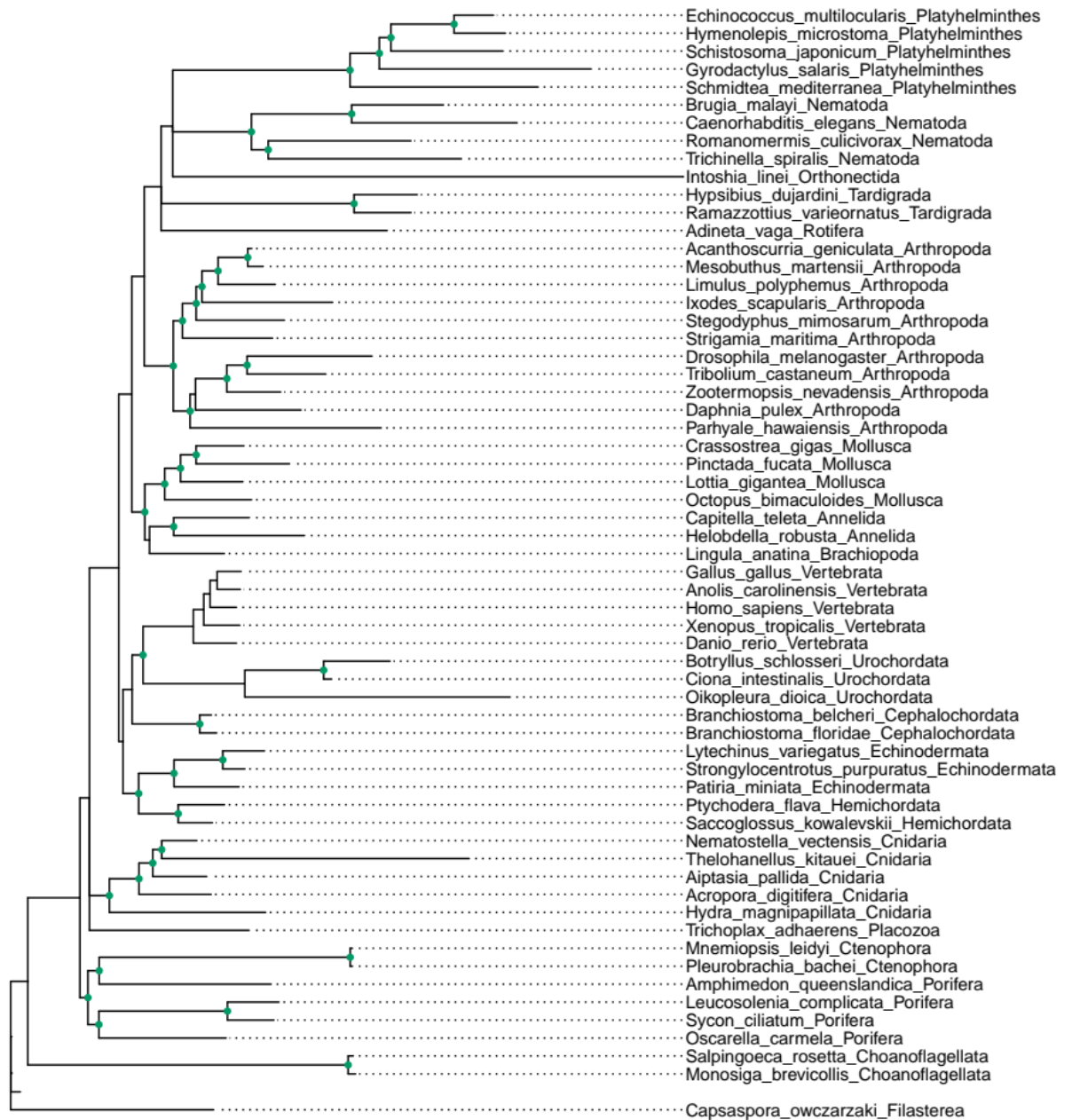


Figure 4.11 Bayesian inference of the BUSCO293 dataset. Node points in green indicate posterior probabilities > 0.9 . Using PhyloBayes 4.1: 2 Pb chains were run in parallel, with 1535 trees each. Maxdiff 0.98 and meandiff 0.08. The remaining outgroups have been collapsed to direct focus to the animals only.

The hot topic in 'tree of life' phylogenies is most frequently the discussion of systematic errors, causing artefacts. In this scenario we have made a tremendous effort in the direction of using more complete datasets by way of whole genomes (more than any other analyses at the time) (so less chance of missing data), selecting highly conserved genes (reducing the impact of lost or saturated genes/paralogues) and a large and varied taxon sampling (for a smooth transition of relationships between phyla with as few gaps as currently possible).

We need to look beyond the bias seen towards more traditionally accepted hypotheses, and welcome new hypotheses that show statistical promise and a biological explanation, even if that explanation is not the most direct path, evolution is complex after all, hence the conflict seen in the designation of the first (extant) divergent animal. The hypotheses presented in this study have minimal conflicts between them, but where they do differ, there is another study out there (Nosenko *et al.*, 2013; Jékely *et al.*, 2015; Giribet, 2016b; Feuda *et al.*, 2017; Kocot *et al.*, 2017; Laumer *et al.*, 2018, 2019) that corroborates similar relationships. The first splitting animal was most consistently Ctenophora, which has a plausible biological explanation: the mesodermal cell types evolved independently to those in Bilateria, with divergence ahead of Planulozoa.

Using similar methods to those used here, future phylogenomic inferences with more whole genome taxon samples covering all the known phyla missing in this study will reveal an improved and more detailed history of the relationships within the Animal Kingdom, further eliminating those well known systematic erroneous artefacts. The results provided resolution among some phyla, but the phylogeny as a whole proves that more than a well conserved dataset is required for total resolution, particularly within the fast evolving lineages such as platyhelminthes and nematodes.

METHODS & MATERIALS

Eight different phylogenies were inferred using different combinations of datasets (Table 4.1 and Figure 4.12). Extracting data from a published comparative genomics pipeline (Paps & Holland, 2018; Guijarro-Clarke *et al.*, 2020), in which 102 genomes were reciprocal BLASTp (Camacho *et al.*, 2009), on an all-vs-all basis with expect-value limitation of $1E^{-6}$. Markov Clustering analysis (MCL) (Enright *et al.*, 2002) clustered HGs using the default inflation parameter ($I=2$). Different selection criteria was used to determine each dataset used (Figure 4.12):

- Most retained single-copy (where possible) 60 HGs in at least 90/102 eukaryotes (tree HG60).
- Most retained single-copy 60 HGs in at least 90/102 eukaryote genomes with mixed taxon sequences of additional animal phyla (this mixing of sequences from multiple taxa is necessary where the genome is not available). Mixed taxa, with any representative available, sequences mined from NCBI non-redundant protein database using positive GI lists for each phyla not among the 102 genomes (HG60+).
- Most retained single-copy 90 HGs in at least 54/59 animal genomes and any of the 43 non-animal genome out-groups (HG90).
- Most retained single-copy 90 HGs in 59 animal genomes and mixed taxon sequences of additional animal phyla using BLAST similarity (HG90+).
- HGs retained in all the 102 eukaryote genomes regardless of number of copies or representatives in the organisms (HG101).
- HGs equivalent to the BUSCO 303 orthologues for eukaryotes (Simão *et al.*, 2015), reduced to 293 HGs due to majority missing representatives in the 102 genomes (BUSCO293).

Each of the datasets were aligned as HGs independently using MAFFT (-leavegappyregion) (Katoh & Standley, 2013), refined using MUSCLE (-refine) (Edgar, 2004) and trimmed using TrimAL (-gappyout) (Capella-Gutiérrez *et al.*, 2009). We implemented IQ-TREE (Nguyen *et al.*, 2015) to infer unrooted trees using -m MFP+MERGE to choose the most appropriate model automatically using ModelFinder (Kalyaanamoorthy *et al.*, 2017). Using branch lengths and trees rooted at the Choanoflagellata clade, orthologues were extracted from each tree where there were paralogues, together with the single copy proteins where available for each species. The HGs were

concatenated per species, retrimmed using TrimAL (`-gappyout`) to reduce the erroneous singleton sites. Tree HG90 (3.2) was additionally trimmed to reduce singletons even further. `-m MFP+MERGE` parameters were used to select the model for trees (tree HG60: LG+R10, and LG+R8, HG90: LG+R8, tree HG90 and HG90+: LG+F+R7). Close equivalent models to CAT+GTR in IQ-TREE were used for dataset HG101 and BUSCO293 (LG+F+R7 and LG+F+G). 1000 Ultrafast bootstrap approximation and SH-aLRT confidence values were also included in the IQ-TREE parameters (Minh *et al.*, 2013; Nguyen *et al.*, 2015). Further Bayesian analyses were performed on HG101 and BUSCO293 using PhyloBayes 4.1 `./pb` with `-cat -gtr` (Lartillot & Philippe, 2004).

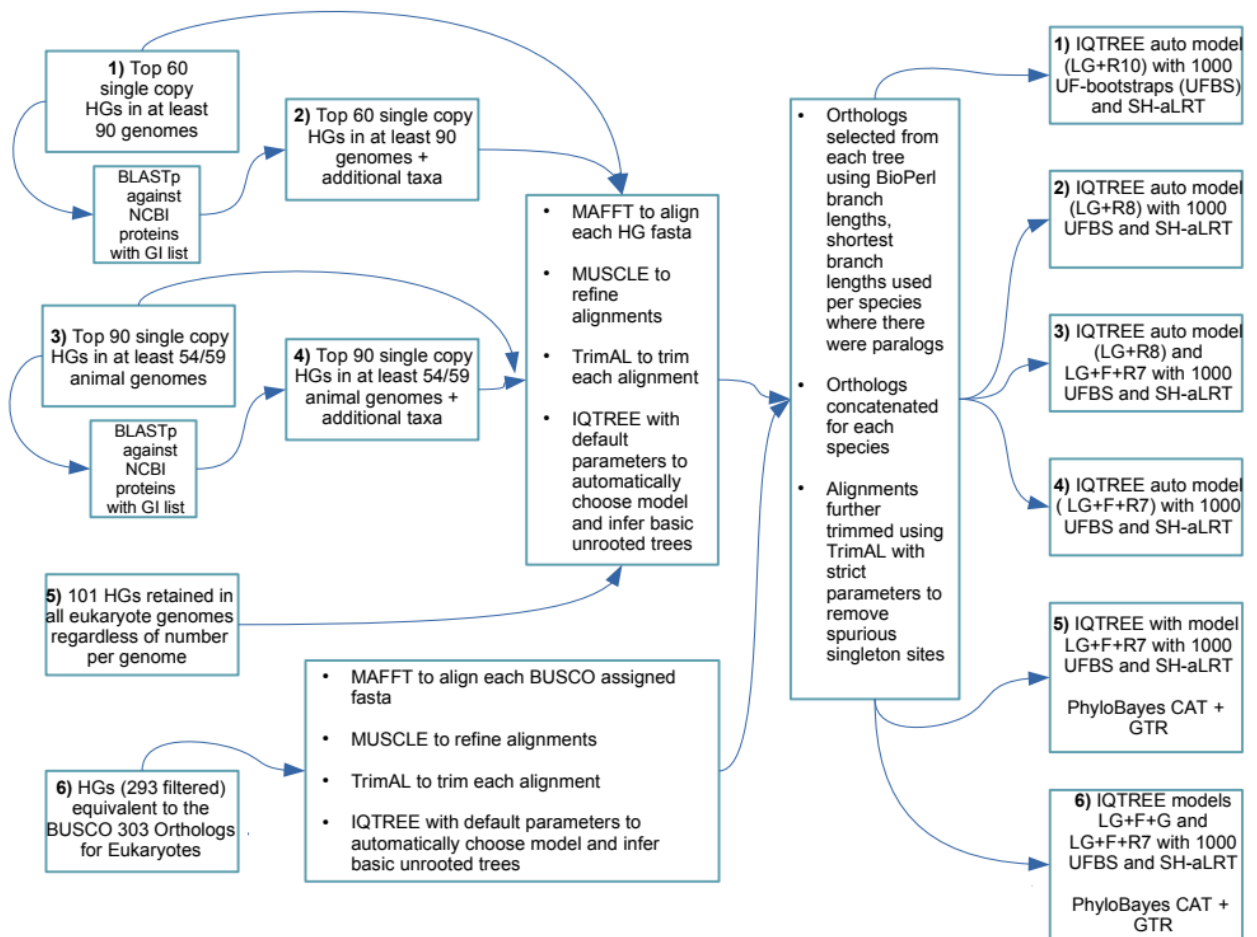


Figure 4.12 The pipeline schema for each of the datasets and trees. Each step is introduced to reduce systematic errors in the resulting phylogenies. "auto model" refers to the parameter *MFP + MERGE* in IQTREE which uses *ModelFinder* (Kalyanamoorthy et al., 2017) to choose an appropriate model.

5 DISCUSSION/CONCLUSIONS

5.1 DIVERSIFICATION OF SPECIFIC ANIMAL LINEAGES BY LOSS AND GAIN OF HOMOLOGY GROUPS

The role of gene gains, losses and recycling is central to understanding the evolution of the Animal Kingdom. We already knew there was a large role of gene novelty in the origin of animals. I further found that this was the case for Bilateria and these findings were consistent with the diversification seen in extant lineages. We do not know how exactly the three bilaterian supergroups (Deuterostomia, Ecdysozoa and Lophotrochozoa) diverged or which diverged first in the origin of Bilateria (Hejnol & Pang, 2016; Nielsen *et al.*, 2018). We follow the assumptions based on congruent phylogenies, and the losses and gains follow this traditional order. With or without this prior knowledge, we can at least deduce that many of the homologous novelties in the divergence of these major lineages are related to hormones and cellular interactions, hallmarks of animal multicellularity (Paps, 2018). After the huge burst in novelty seen at the origin of animals, the succeeding events generating the deuterostome and protostome groups showed similar, more specific bursts of novelty, fine-tuning their individual and specific toolkit in a changing environment, filled with competition and a necessity to be diverse to survive.

Further gene loss occurred in the evolution of phyla within these groups, although in some cases the loss is balanced by novelty, preventing a traditional reduction in the genome, in which it is simplified. Championing this genomic reduction, without simplification, the three animal phyla with the largest levels of gene loss - flatworms, nematodes, and tardigrades - also show striking levels of genomic novelty. There are at least two possible causes leading to these numbers. Either it is explained by a method of replacement. The ancestral genomes underwent a process of refurbishments, a high turnover in genomic content, and so genes were not lost as such, simply adapted beyond the point of similarity recognition, becoming a new set of homologous gene families with new or alternative functions (López-Escardó *et al.*, 2019). Or it may be that these particular phyla are evolving at faster rates than the rest, and thus, where there appears more novelty, and as an artefact, highly divergent, fast-clock genes have simply formed their own clusters, producing false losses and gains.

Conversely, excluding this pattern of simultaneous reduction and expansion, there are instances of unbalanced novelty in some lineages, and unprecedented loss in others. There are biological processes that would explain some of these, for example, heavy GC content throughout a genome leads to higher probability of mutation, a particularly fast evolving lineage. A lifestyle change from a free-living ancestral organism to that of parasitism (Hahn *et al.*, 2014). Duplication events. Horizontal gene transfer, retrotransposons. Random crossover (Paps & Holland, 2018; Richter *et al.*, 2018; Keeling & Burki, 2019; López-Escardó *et al.*, 2019).

These analyses show an unprecedented role for gene loss in the evolution of the Animal Kingdom. Gene loss has been observed frequently in eukaryote lineages (Suga *et al.*, 2013), but as an evolutionary function leading to divergent lineages in complex animals, the impact has been underestimated. The data here prove that gene loss is not just an evolutionary process to simplify animal lineages, but a process necessary in the turnover of genetic material ancestrally, leading to equally complex extant animals. Tardigrades are described as extremotolerant animals, they are able to withstand extreme environmental conditions. They have been shown to have undergone large losses of gene pathways which promote stress damage, whilst simultaneously expanding on metazoan specific damage-improvement related genes (Hashimoto *et al.*, 2016).

Until now, most comparative genomic studies surrounding loss and gain have been limited to smaller taxon sampling than here with fewer outgroups, the use of transcriptomes rather than genomes and a lack of diversity. The results unveiled here show the largest, most diverse comparative genomics study using whole genomes and a recognised pipeline methodology with proven outcome.

5.2 EVOLUTION OF BODY PLANS IN ANIMAL DIVERSIFICATION

Transcription factors are essential to determine the rate of expression of genes. The specific combination of transcription factors to developmental genes is what defines the morphological differences between animals. Many developmental genes and transcription factors are specific to animals, and there has been an abundance of research in this area, particularly in vertebrates (Fonseca *et al.*, 2008; Takatori *et al.*, 2008; Holland, 2013). Animal phyla have been organised in such a way that they are described by body-plan. In fact, it has become the norm to include a little about homeobox family presence as each animal genome is sequenced, particularly if it is not a well studied phyla. This however has its limitations. Homeobox analysis and expressions have been largely compared to those described in model organisms, and are most abundantly represented by vertebrates, *Drosophila melanogaster* and *Caenorhabditis elegans*. Furthermore, with a highly conserved homeodomain, homeobox genes are notoriously tricky to classify. Classification to date has relied on manual annotation and identifying the presence or absence of distinct and known motifs within the homeodomain, the number of domains in each homeobox or the type of additional homeodomains if any.

I uncovered an expansion of distinct LIM-domain containing proteins seen only in invertebrates (and exclusively in the vertebrate lizard *Anolis carolinensis*), which may be explained by a consistent reshuffling domains that have not been observed in other non-chordate phyla before (Koch *et al.*, 2012). The method used to extract homologous homeobox and homeobox-like proteins is comprehensive and includes uncharacterised proteins that have otherwise been ignored, and left unannotated. Using a combination of well known chordate specific homeodomain containing proteins from HomeoDB and a complete extraction of homeodomain containing proteins from homeobox HGs from non-chordate genomes, I revealed that some superclasses may predate previous suppositions, and there has been whole homeobox family loss in some phyla.

HOX-like (HOXL) ANTP families show high support for diverging clades of HOXL ANTP genes in the annelid *Helobdella robusta* and urochordates. Lophotrochozoans have a large number of genes nested in separate clades within HOXL, which are closely related to deuterostome and arthropod genes. The HOXL families have clean clades, which correlates to their important and essential function in Metazoa, they are well conserved, slow evolving genes (Halanych, 2004; Fonseca *et al.*, 2008; Bürglin & Affolter, 2016; Ferrier, 2016; Barton-Owen *et al.*, 2018).

The gene trees have, for most homeobox classes, distinct clades. The ZF class is the messy exception and shows a widespread of Zinc finger domains throughout homeobox families. This superclass, classified by the Zinc Finger domain, and associated domains frequently seen among them, is highly divergent and has undergone many small expansions in different directions (Takatori *et al.*, 2008). This polyphyletic tendency makes the ZF class extremely hard to classify, but still, as a class dispersed among all the animal phyla, the findings cannot be ignored. A reasonable explanation for the behaviour of the evolution in the ZF homeobox class is the tendency for zinc-fingers to attract and combine with other domains (Nam & Nei, 2005; Bürglin & Affolter, 2016; Ferrier, 2016). Given this likely pattern, I would recommend changing the way in which ZF class is classified, and perhaps it should be reclassified and looked at as more than one homeobox superclass instead, each ZF clade is seen dispersed across the three key bilaterian lineages: deuterostomes, ecdysozoans and lophotrochozoans.

Reduction of homeobox genes may be prompted by this reshuffling; repurposing the functions of neutral homeoboxes into novel homeoboxes. There is likely an aspect of convergent evolution, for example, the recombination of domains from other existing homeobox families, or duplications of the same domains in different animal lineages. Studying these complex genes involved in animal body plans, the functions, the convergent evolution and patterns of evolution in diverging families between the different animal phyla is necessary to understanding the events in metazoan evolution, and the division of phyla as they are defined today.

5.3 IMPORTANT ANIMAL RELATIONSHIPS IN THE EMERGENCE OF ANIMAL PHYLA

A robust evolutionary tree of animals is central to perform comparative evolutionary studies. We want to see accurate placements and relationships between individuals and groups that show a reasonable evolutionary path through ancestral genomes. The robustness of an evolutionary tree of animals is dependent on several factors, biological and methodological. The biggest problems causing artefacts and conflict in the interpretation of animal relationships and their shared ancestors goes beyond the animal tree of life, it is the simplistic statistical methods used. There is no single statistical solution that fits all the complexities in the biological events animals have undergone. Genes and species do not fit a constant rate molecular clock model and evolutionary rates can vary hugely (Whelan *et al.*, 2015a; Dos Reis *et al.*, 2016; Laumer *et al.*, 2018; Philippe *et al.*, 2019). An example of this is seen in the fast evolving flatworms. These exhibited a substantial turnover in genomic content in chapter 2, and a close relationship to another fast evolving clade, the nematodes, was inferred in chapter 4.

Another problem is the ability for scientists to escape traditional views. For a very long time, it has been believed that sponges were the first diverging extant animal, this was supported by a lacking homologous nervous system, and with other supporting evidence, is therefore the simplest and shortest route for the evolution of animals to have taken (Philippe *et al.*, 2009; Adamska, 2016; Feuda *et al.*, 2017). There is no shortage of molecular based comparative and phylogenomic studies as well as morphological deductions backing poriferans as the first animals. However, there is also no shortage vice versa for evidence backing ctenophorans as the first splitting animal (Ryan *et al.*, 2013; Moroz *et al.*, 2014; Halanych, 2015). Using four different datasets of slow evolving genes core to eukaryotes, animals and well known orthologues used by BUSCO, each result that was returned neglected to infer any poriferan species as first splitting animals. Ctenophorans were consistently first to diverge, and in some instances alongside sponges in a joint clade. Biological reasoning could agree with either hypothesis. I prefer the more complex biological pathway animal evolution has taken, with ctenophores diverging from the LCMA ahead of Porifera. Whilst the question has still not been resolved, and scientists may never agree, convergent loss across the animal kingdom is not out of the question (Fortunato *et al.*, 2015; Torruella *et al.*, 2015; Albalat & Cañestro, 2016).

5.4 IMPLICATIONS AND RECOMMENDATIONS TO FUTURE RESEARCH

The dataset used here used the broadest taxon sampling when it was started back in October 2016. There were several key phylogenetic positions in the animal kingdom without genome representatives. Since then, many more have been sequenced, assembled and annotated. These missing key lineages (in no particular order) include Onychophora, Xenoturbellida, Acoela, Nemertodermatida, Chaetognatha, Priapulida, Loricifera, Kinorhyncha, Dicyemeda, Cycliophora, Entoprocta, Gnathostomulida, Micrognathozoa, Gastrotricha, Nemertea, Nematomorpha, Phoronida and Bryozoa/Ectoprocta. Putative relationships of these are as shown in Figure 1.2, but as can be seen these animal groups are the least resolved among the animals. The genomes of some of these are now available, or will be available soon (Perea-Atienza *et al.*, 2015; Giribet, 2016a, 2016b), but the animal tree of life will become closer to a resolution only when the taxon sampling covers an even broader set of genomes, including all of these phyla (Giribet, 2016a).

The introduction of a single one of these additional missing animal genomes, once available to the scientific community, to the comparative genomics pipeline here could alter the results of lost and novel HGs drastically. Alternatively, the introduction of any one of these additional genomes to the dataset could also further support the results seen here. The same could be seen in the homeobox gene trees, where there are losses seen in some animal groups, these gaps could be filled. Homeobox families are likely to be expanded, as this would correlate with the morphological differences seen in each animal phyla body plan as classified (Gold *et al.*, 2014). Again, the inclusion of these missing phyla would go a long way to resolving the gaps in animal phylogeny and topologies. A smoother transition will be seen between species, there will be fewer erroneous positions causing LBA in terms of sudden jumps in seen in the fast evolving species (Simion *et al.*, 2017).

Expanding the genome dataset further to include more species per phyla, as well as more phyla will also cluster a stricter, and improved definition in HGs. Inflation could be reduced in the MCL parameters to decrease the granularity in this case, and increase the accuracy of homologous clustering because similarity groups based on e-values will have access to higher probabilities. The current set of HGs have already proven to provide a more inclusive dataset in chapter 4 for orthologues seen in all animals, where BUSCO revealed some gaps. This is because the BUSCO orthologue datasets are based on fewer species, which only include key model organisms. Further

work could incorporate the HGs analysed here in this thesis to provide signatures for HMMs, combining alternative supporting methods for comparisons on new animal whole genomes.

5.5 FINAL WORDS

Here in this thesis, I have provided an overview of the genome biology and evolution of the Animal Kingdom (see Figure 5.1 for summary of findings). Animals have been proven to be very dynamic at both genome level, and specific gene families. Evolution has no set plans, but follows the flow of constantly changing landscapes, adapting to fit niches. Future research should focus on increasing the taxon sampling with diversity in phyla and number of species, producing genomes of higher quality, and developing more robust methods to better understand the origin of these endless, intricately diverse life forms.

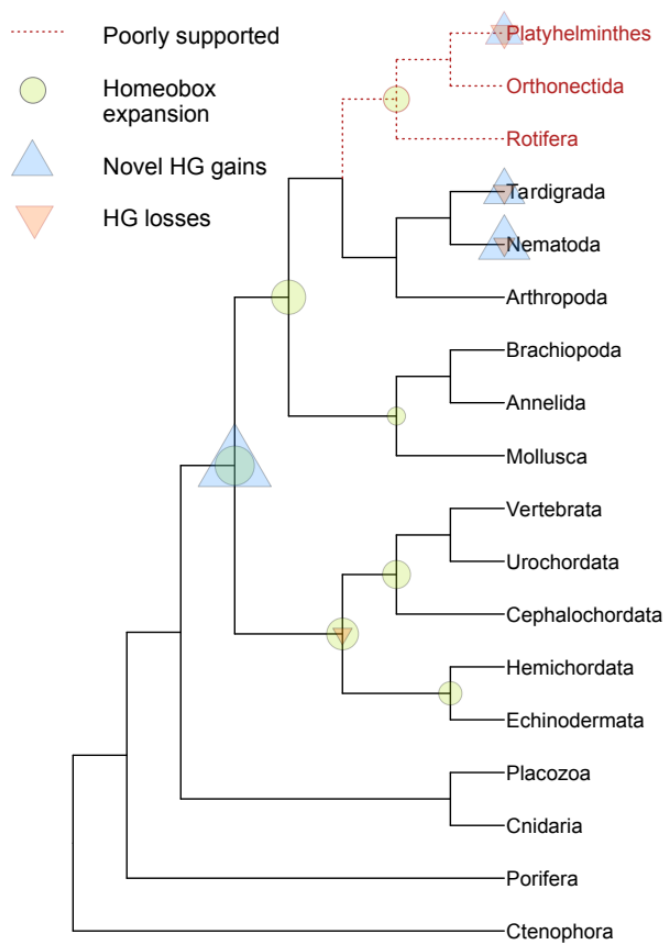


Figure 5.1 A summary of all the results chapters. The most poorly supported nodes in chapter 4 in red, homeobox expansions for the origins of homeobox proteins from chapter 3 in green, size equals number of homeobox genes expanded and blue and red triangle sizes to match HG gains and losses from the biggest HG novel and loss counts.

6 REFERENCES

1. Adamska M (2016) Sponges as models to study emergence of complex animals. *Current Opinion in Genetics and Development*, **39**, 21–28.
2. Albalat R, Cañestro C (2016) Evolution by gene loss. *Nature Reviews Genetics*, **17**, 379–391.
3. Arakawa K (2016) No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, E3057.
4. Babushok D V., Ostertag EM, Kazazian HH (2007) Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cellular and Molecular Life Sciences*, **64**, 542–554.
5. Ballesteros JA, Hormiga G (2016) A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Molecular Biology and Evolution*, **33**, 2117–2134.
6. Barton-Owen TB, Szabó R, Somorjai IML, Ferrier DEK (2018) A revised spiralian homeobox gene classification incorporating new polychaete transcriptomes reveals a diverse TALE class and a divergent hox gene. *Genome Biology and Evolution*, **10**, 2151–2167.
7. Bateman A, Martin MJ, O’Donovan C et al. (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, **45**, D158–D169.
8. Brauchle M, Bilican A, Eyer C et al. (2018) Xenacoelomorpha survey reveals that all 11 animal homeobox gene classes were present in the first bilaterians. *Genome Biology and Evolution*, **10**, 2205–2217.
9. Brunet T, King N (2017) The Origin of Animal Multicellularity and Cell Differentiation. *Developmental Cell*, **43**, 124–140.
10. Bürglin TR, Affolter M (2016) Homeodomain proteins: an update. *Chromosoma*, **125**, 497–521.
11. Burki F (2014) The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor Perspectives in Biology*, **6**, 1–19.
12. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: Architecture and applications. *BMC Bioinformatics*, **10**, 421.
13. Cannon JT, Kocot KM, Waits DS, Weese DA, Swalla BJ, Santos SR, Halanych KM (2014) Phylogenomic resolution of the hemichordate and echinoderm clade. *Current Biology*, **24**, 2827–2832.

14. Cannon JT, Vellutini BC, Smith J, Ronquist F, Jondelius U, Hejnol A (2016) Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, **530**, 89–93.
15. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
16. Carbon S, Dietze H, Lewis SE et al. (2017) Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research*, **45**, D331–D338.
17. Chan CX, Ragan MA (2013) Next-generation phylogenomics. *Biology Direct*, **8**, 1–6.
18. Chipman AD, Ferrier DEK, Brena C et al. (2014) The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. *PLoS Biology*, **12**.
19. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, **6**, 361–375.
20. Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**, 965–968.
21. Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H (2008) Additional molecular support for the new chordate phylogeny. *Genesis*, **46**, 592–604.
22. Delsuc F, Philippe H, Tsagkogeorga G et al. (2018) A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biology*, **16**.
23. Denoeud F, Henriot S, Mungpakdee S et al. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, **330**, 1381–1385.
24. Dunn CW, Ryan JF (2015) The evolution of animal genomes. *Current Opinion in Genetics and Development*, **35**, 25–32.
25. Dunn CW, Hejnol A, Matus DQ et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.
26. Dunn CW, Giribet G, Edgecombe GD, Hejnol A (2014) Animal Phylogeny and Its Evolutionary Implications. *Annual Review of Ecology, Evolution, and Systematics*, **45**, 371–395.
27. Dunn CW, Leys SP, Haddock SHD (2015) The hidden biology of sponges and ctenophores. *Trends in Ecology and Evolution*, **30**, 282–291.
28. Dunwell TL, Paps J, Holland PWH (2017) Novel and divergent genes in the evolution of placental mammals. *Proceedings of the Royal Society B: Biological Sciences*, **284**.
29. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

30. Edgecombe GD (2010) Arthropod phylogeny: An overview from the perspectives of morphology, molecular data and the fossil record. *Arthropod Structure and Development*, **39**, 74–87.
31. Egger B, Lapraz F, Tomiczek B et al. (2015) A transcriptomic-phylogenomic analysis of the evolutionary relationships of flatworms. *Current Biology*, **25**, 1347–1353.
32. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**, 1–14.
33. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**, 1575–1584.
34. Ferrier DEK (2016) Evolution of homeobox gene clusters in animals: The Giga-cluster and Primary vs. secondary clustering. *Frontiers in Ecology and Evolution*, **4**, 1–13.
35. Feuda R, Dohrmann M, Pett W et al. (2017) Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology*, **27**, 3864–3870.e4.
36. Fonseca NA, Vieira CP, Holland PWH, Vieira J (2008) Protein evolution of ANTP and PRD homeobox genes. *BMC Evolutionary Biology*, **8**, 200.
37. Fortunato SAV, Adamski M, Ramos OM, Leininger S, Liu J, Ferrier DEK, Adamska M (2014) Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature*, **514**, 620–623.
38. Fortunato SAV, Adamski M, Adamska M (2015) Comparative analyses of developmental transcription factor repertoires in sponges reveal unexpected complexity of the earliest animals. *Marine Genomics*, **24**, 121–129.
39. Fröblius AC, Funch P (2017) Rotiferan Hox genes give new insights into the evolution of metazoan bodyplans. *Nature Communications*, **8**.
40. Gabaldón T (2008) Large-scale assignment of orthology: back to phylogenetics? *Genome biology*, **9**, 235.
41. Gabaldón T, Koonin E V. (2013) Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, **14**, 360–366.
42. Garcia-Fernández J (2005) The genesis and evolution of homeobox gene clusters. *Nature Reviews Genetics*, **6**, 881–892.
43. Gehring WJ (2014) The evolution of vision. *Wiley Interdisciplinary Reviews: Developmental Biology*, **3**, 1–40.
44. Giribet G (2016a) Genomics and the animal tree of life: conflicts and future prospects. *Zoologica Scripta*, **45**, 14–21.

45. Giribet G (2016b) New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. *Organisms Diversity and Evolution*, **16**, 419–426.
46. Giribet G, Edgecombe GD (2017) Current understanding of Ecdysozoa and its internal phylogenetic relationships. *Integrative and Comparative Biology*, **57**, 455–466.
47. Gold DA, Gates RD, Jacobs DK (2014) The early expansion and evolutionary dynamics of POU class genes. *Molecular Biology and Evolution*, **31**, 3136–3147.
48. Gold DA, Katsuki T, Li Y et al. (2019) The genome of the jellyfish *Aurelia* and the evolution of animal complexity. *Nature Ecology and Evolution*, **3**, 96–104.
49. Gómez-Marín C, Tena JJ, Acemel RD et al. (2015) Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 7542–7547.
50. Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, Ruiz-Trillo I (2017) Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife*, **6**.
51. Guijarro-Clarke C, Holland PWH, Paps J (2020) Widespread patterns of gene loss in the evolution of the animal kingdom. *Nature Ecology and Evolution*, **4**, 519–523.
52. Hahn C, Fromm B, Bachmann L (2014) Comparative genomics of flatworms (Platyhelminthes) reveals shared genomic features of ecto- and endoparasitic neodermata. *Genome Biology and Evolution*, **6**, 1105–1117.
53. Halanych KM (2004) The new view of animal phylogeny. *Annual Review of Ecology, Evolution, and Systematics*, **35**, 229–256.
54. Halanych KM (2015) The ctenophore lineage is older than sponges? That cannot be right! or can it? *Journal of Experimental Biology*, **218**, 592–597.
55. Hashimoto T, Horikawa DD, Saito Y et al. (2016) Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nature Communications*, **7**, 12808.
56. He T, Zhu M, Mills BJW et al. (2019) Possible links between extreme oxygen perturbations and the Cambrian radiation of animals. *Nature Geoscience*, **12**, 468–474.
57. Hejnol A, Dunn CW (2016) Animal evolution: Are phyla real? *Current Biology*, **26**, R424–R426.
58. Hejnol A, Lowe CJ (2014) Animal evolution: Stiff or squishy notochord origins? *Current Biology*, **24**, R1131–R1133.
59. Hejnol A, Pang K (2016) Xenacoelomorpha's significance for understanding bilaterian evolution. *Current Opinion in Genetics and Development*, **39**, 48–54.

60. Hejnal A, Obst M, Stamatakis A et al. (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 4261–4270.
61. Holland PWH (2013) Evolution of homeobox genes. *Wiley Interdisciplinary Reviews: Developmental Biology*, **2**, 31–45.
62. Holland PWH (2015) Did homeobox gene duplications contribute to the Cambrian explosion? *Zoological Letters*, **1**, 1–8.
63. Holland ND (2016) Nervous systems and scenarios for the invertebrate-to-vertebrate transition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**.
64. Holland PWH, Marlétaz F, Maeso I, Dunwell TL, Paps J (2017) New genes from old: Asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372**.
65. Janssen R, Eriksson BJ, Tait NN, Budd GE (2014) Onychophoran Hox genes and the evolution of arthropod Hox gene expression. *Frontiers in Zoology*, **11**, 1–11.
66. Jékely G, Paps J, Nielsen C (2015) The phylogenetic position of ctenophores and the origin(s) of nervous systems. *EvoDevo*, **6**, 1.
67. Jiménez-Guri E, Paps J, García-Fernández J, Saló E (2006) Hox and ParaHox genes in Nemertodermatida, a basal bilaterian clade. *International Journal of Developmental Biology*, **50**, 675–679.
68. Jones P, Binns D, Chang HY et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
69. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermini LS (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.
70. Karaz S, Courgeon M, Lepetit H et al. (2016) Neuronal fate specification by the Dbx1 transcription factor is linked to the evolutionary acquisition of a novel functional domain. *EvoDevo*, **7**, 1–13.
71. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
72. Keeling PJ, Burki F (2019) Progress towards the Tree of Eukaryotes. *Current Biology*, **29**, R808–R817.
73. Knoll AH, Sperling EA (2014) Oxygen and animals in Earth history. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 3907–3908.

74. Koch BJ, Ryan JF, Baxevanis AD (2012) The diversification of the lim superclass at the base of the metazoa increased subcellular complexity and promoted multicellular specialization. *PLoS ONE*, **7**.
75. Kocot KM (2016) On 20 years of Lophotrochozoa. *Organisms Diversity and Evolution*, **16**, 329–343.
76. Kocot KM, Struck TH, Merkel J et al. (2017) Phylogenomics of lophotrochozoa with consideration of systematic error. *Systematic Biology*, **66**, 256–282.
77. Kohn AB, Citarella MR, Kocot KM, Bobkova Y V., Halanych KM, Moroz LL (2012) Rapid evolution of the compact and unusual mitochondrial genome in the ctenophore, *Pleurobrachia bachei*. *Molecular Phylogenetics and Evolution*, **63**, 203–207.
78. Kolaczkowski B, Thornton JW (2009) Long-branch attraction bias and inconsistency in bayesian phylogenetics. *PLoS ONE*, **4**.
79. Kriventseva E V., Tegenfeldt F, Petty TJ et al. (2015) OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, **43**, D250–D256.
80. Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JAM (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, **24**, 539–551.
81. Lankester ER (1879) *Degeneration: A Chapter in Darwinism*, Vol. 12. Macmillan and Company, 75 pp.
82. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, **21**, 1095–1109.
83. Laughon A, Scott MP (1984) Sequence of a *Drosophila* segmentation gene: Protein structure homology with DNA-binding proteins. *Nature*, **310**, 25–31.
84. Laumer CE, Bekkouche N, Kerbl A et al. (2015) Spiralian Phylogeny Informs the Evolution of Microscopic Lineages. *Current Biology*, **25**, 2000–2006.
85. Laumer CE, Gruber-Vodicka H, Hadfield MG, Pearse VB, Riesgo A, Marioni JC, Giribet G (2018) Support for a clade of placozoa and cnidaria in genes with minimal compositional bias. *eLife*, **7**, 1–19.
86. Laumer CE, Fernández R, Lemer S et al. (2019) Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B: Biological Sciences*, **286**, 20190831.
87. Lemey P, Salemi M, Vamdamme A (2009) *The Phylogenetic Handbook*. 723 pp.
88. Li L, Stoeckert CJJ, Roos DS (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research. *Genome Research*, **13**, 2178–2189.

89. López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I (2019) Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **374**.
90. Lowe CJ, Clarke DN, Medeiros DM, Rokhsar DS, Gerhart J (2015) The deuterostome context of chordate origins. *Nature*, **520**, 456–465.
91. Luo YJ, Takeuchi T, Koyanagi R et al. (2015) The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. *Nature Communications*, **6**, 1–10.
92. Luo YJ, Kanda M, Koyanagi R et al. (2018) Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads. *Nature Ecology and Evolution*, **2**, 141–151.
93. Marlétaz F, Martin E, Perez Y et al. (2006) *Chaetognath phylogenomics: a protostome with deuterostome-like development*, Vol. 16.
94. Marlétaz F, Peijnenburg KTCA, Goto T, Satoh N, Rokhsar DS (2019) A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. *Current Biology*, **29**, 312–318.e3.
95. Mazza ME, Pang K, Reitzel AM, Martindale MQ, Finnerty JR (2010) A conserved cluster of three PRD-class homeobox genes (homeobrain, rx and orthopedia) in the Cnidaria and Protostomia. *EvoDevo*, **1**, 1–15.
96. Mentel M, Röttger M, Leys S, Tielens AGM, Martin WF (2014) Of early animals, anaerobic mitochondria, and a modern sponge. *BioEssays*, **36**, 924–932.
97. Mikhailov K V., Slyusarev GS, Nikitin MA, Logacheva MD, Penin AA, Aleoshin V V., Panchin Y V. (2016) The Genome of *Intoshia linei* Affirms Orthonectids as Highly Simplified Spiralian. *Current Biology*, **26**, 1768–1774.
98. Mills DB, Canfield DE (2014) Oxygen and animal evolution: Did a rise of atmospheric oxygen ‘trigger’ the origin of animals? *BioEssays*, **36**, 1145–1155.
99. Minh BQ, Nguyen MAT, Von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution*, **30**, 1188–1195.
100. Moore AD, Bornberg-Bauer E (2012) The dynamics and evolutionary potential of domain loss and emergence. *Molecular Biology and Evolution*, **29**, 787–796.
101. Morino Y, Hashimoto N, Wada H (2017) Expansion of TALE homeobox genes and the evolution of spiralian development. *Nature Ecology and Evolution*, **1**, 1942–1949.
102. Moroz LL, Kocot KM, Citarella MR et al. (2014) The ctenophore genome and the evolutionary origins of neural systems. *Nature*, **510**, 109–114.

103. Nam J, Nei M (2005) Evolutionary change of the numbers of homeobox genes in bilateral animals. *Molecular Biology and Evolution*, **22**, 2386–2394.
104. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**, 268–274.
105. Nielsen C, Brunet T, Arendt D (2018) Evolution of the bilaterian mouth and anus. *Nature Ecology and Evolution*, **2**, 1358–1376.
106. Nosenko T, Schreiber F, Adamska M et al. (2013) Deep metazoan phylogeny: When different genes tell different stories. *Molecular Phylogenetics and Evolution*, **67**, 223–233.
107. Paps J (2018) What makes an animal? The molecular quest for the origin of the Animal Kingdom. *Integrative and Comparative Biology*, **58**, 654–665.
108. Paps J, Holland PWH (2018) Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nature Communications*, **9**, 1–8.
109. Paps J, Riutort M (2012) Molecular phylogeny of the phylum Gastrotricha: New data brings together molecules and morphology. *Molecular Phylogenetics and Evolution*, **63**, 208–212.
110. Paps J, Baguñ J, Riutort M (2009) Bilaterian phylogeny: A broad sampling of 13 nuclear genes provides a new lophotrochozoa phylogeny and supports a paraphyletic basal acoelomorpha. *Molecular Biology and Evolution*, **26**, 2397–2406.
111. Paps J, Holland PWH, Shimeld SM (2012) A genome-wide view of transcription factor gene diversity in chordate evolution: Less gene loss in amphioxus? *Briefings in Functional Genomics*, **11**, 177–186.
112. Paps J, Medina-Chacón LA, Marshall W, Suga H, Ruiz-Trillo I (2013) Molecular Phylogeny of Unikonts: New Insights into the Position of Apusomonads and Ancyromonads and the Internal Relationships of Opisthokonts. *Protist*, **164**, 2–12.
113. Paps J, Xu F, Zhang G, Holland PWH (2015) Reinforcing the egg-timer: Recruitment of novel Lophotrochozoa homeobox genes to early and late development in the Pacific oyster. *Genome Biology and Evolution*, **7**, 677–688.
114. Pastrana CC, DeBiasse MB, Ryan JF (2019) Sponges lack paraHox genes. *Genome Biology and Evolution*, **11**, 1250–1257.
115. Perea-Atienza E, Gavilán B, Chiodin M, Abril JF, Hoff KJ, Poustka AJ, Martínez P (2015) The nervous system of Xenacoelomorpha: A genomic perspective. *Journal of Experimental Biology*, **218**, 618–628.
116. Peterson KJ, Eernisse DJ (2016) The phylogeny, evolutionary developmental biology, and paleobiology of the Deuterostomia: 25 years of new techniques, new discoveries, and new ideas. *Organisms Diversity and Evolution*, **16**, 401–418.

117. Pett W, Ryan JF, Pang K, Mullikin JC, Martindale MQ, Baxevanis AD, Lavrov D V. (2011) Extreme mitochondrial evolution in the ctenophore *Mnemiopsis leidyi*: Insight from mtDNA and the nuclear genome. *Mitochondrial DNA*, **22**, 130–142.
118. Philippe H, Lartillot N, Brinkmann H (2005a) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and protostomia. *Molecular Biology and Evolution*, **22**, 1246–1253.
119. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005b) Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, **5**, 1–8.
120. Philippe H, Derelle R, Lopez P et al. (2009) Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Current Biology*, **19**, 706–712.
121. Philippe H, Brinkmann H, Copley RR et al. (2011) Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature*, **470**, 255–260.
122. Philippe H, Poustka AJ, Chiodin M et al. (2019) Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. *Current Biology*, **29**, 1818–1826.e6.
123. Pisani D, Pett W, Dohrmann M et al. (2015) Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 15402–15407.
124. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **35**, 61–65.
125. Dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z (2015) Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Current Biology*, **25**, 2939–2950.
126. Dos Reis M, Donoghue PCJ, Yang Z (2016) Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, **17**, 71–80.
127. Richter DJ, Fozouni P, Eisen MB, King N (2018) Gene family innovation, conservation and loss on the animal stem lineage. *eLife*, **7**, 1–3.
128. Ronshaugen M, McGinnis N, McGinnis W (2002) Hox protein mutation and macroevolution of the insect body plan. *Nature*, **415**, 914–917.
129. Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution*, **30**, 197–214.
130. Ruiz-Trillo I, Paps J (2016) Acoelomorpha: earliest branching bilaterians or deuterostomes? *Organisms Diversity and Evolution*, **16**, 391–399.

131. Ryan JF, Chiodin M (2015) Where is my mind? How sponges and placozoans may have lost neural cell types. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **370**, 20150059-.
132. Ryan JF, Burton PM, Mazza ME, Kwong GK, Mullikin JC, Finnerty JR (2006) The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: Evidence from the starlet sea anemone, *Nematostella vectensis*. *Genome Biology*, **7**, 677–688.
133. Ryan JF, Pang K, Mullikin JC, Martindale MQ, Baxeavanis AD (2010) The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *EvoDevo*, **1**, 1–18.
134. Ryan JF, Pang K, Schnitzler CE et al. (2013) The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science*, **342**, 1242592.
135. Schiemann SM, Martín-Durán JM, Børve A, Vellutini BC, Passamanek YJ, Hejnal A (2017) Clustered brachiopod Hox genes are not expressed collinearly and are associated with lophotrochozoan novelties. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, E1913–E1922.
136. Schiffer PH, Robertson HE, Telford MJ (2018) Orthonectids Are Highly Degenerate Annelid Worms. *Current Biology*, **28**, 1970–1974.e3.
137. Sebé-Pedrós A, De Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I (2011) Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. *Molecular Biology and Evolution*, **28**, 1241–1254.
138. Sebé-Pedrós A, Ballaré C, Parra-Acero H et al. (2016) The Dynamic Regulatory Genome of *Capsaspora* and the Origin of Animal Multicellularity. *Cell*, **165**, 1224–1237.
139. Sebé-Pedrós A, Degnan BM, Ruiz-Trillo I (2017) The origin of Metazoa: A unicellular perspective. *Nature Reviews Genetics*, **18**, 498–512.
140. Simakov O, Kawashima T (2017) Independent evolution of genomic characters during major metazoan transitions. *Developmental Biology*, **427**, 179–192.
141. Simakov O, Marletaz F, Cho SJ et al. (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature*, **493**, 526–531.
142. Simakov O, Kawashima T, Marlétaz F et al. (2015) Hemichordate genomes and deuterostome origins. *Nature*, **527**, 459–465.
143. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
144. Simion P, Philippe H, Baurain D et al. (2017) A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology*, **27**, 958–967.

145. Sogabe S, Hatleberg WL, Kocot KM et al. (2019) Pluripotency and the origin of animal multicellularity. *Nature*, **570**, 519–522.
146. Soubrier J, Steel M, Lee MSY, Der Sarkissian C, Guindon S, Ho SYW, Cooper A (2012) The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular Biology and Evolution*, **29**, 3345–3358.
147. Srivastava M, Simakov O, Chapman J et al. (2010) The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, **466**, 720–726.
148. Struck TH, Wey-Fabrizius AR, Golombek A et al. (2014) Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia. *Molecular Biology and Evolution*, **31**, 1833–1849.
149. Suga H, Chen Z, De Mendoza A et al. (2013) The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nature Communications*, **4**, 1–9.
150. Susko E, Roger AJ (2007) On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution*, **24**, 2139–2150.
151. Tabari E, Su Z (2017) PorthoMCL: Parallel orthology prediction using MCL for the realm of massive genome availability. *Big Data Analytics*, **2**, 1–5.
152. Takatori N, Butts T, Candiani S, Pestarino M, Ferrier DEK, Saiga H, Holland PWH (2008) Comprehensive survey and classification of homeobox genes in the genome of amphioxus, *Branchiostoma floridae*. *Development Genes and Evolution*, **218**, 579–590.
153. Torruella G, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I (2012) Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Molecular Biology and Evolution*, **29**, 531–544.
154. Torruella G, De Mendoza A, Grau-Bové X et al. (2015) Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology*, **25**, 2404–2410.
155. Tsai IJ, Zarowiecki M, Holroyd N et al. (2013) The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*, **496**, 57–63.
156. Wang X, Lavrov D V. (2007) Mitochondrial genome of the homoscleromorph *Oscarella carmela* (Porifera, Demospongiae) reveals unexpected complexity in the common ancestor of sponges and other animals. *Molecular Biology and Evolution*, **24**, 363–373.
157. Whelan N V., Kocot KM, Moroz LL, Halanych KM (2015a) Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 5773–5778.
158. Whelan N V., Kocot KM, Halanych KM (2015b) Employing Phylogenomics to Resolve the Relationships among Cnidarians, Ctenophores, Sponges, Placozoans, and Bilaterians. *Integrative and Comparative Biology*, **55**, 1084–1095.

159. Whelan N V., Kocot KM, Moroz TP et al. (2017) Ctenophore relationships and their placement as the sister group to all other animals. *Nature Ecology and Evolution*, **1**, 1737–1746.
160. Yoshida Y, Koutsovoulos G, Laetsch DR et al. (2017) *Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus*, Vol. 15. 1-40 pp.
161. Zhong YF, Holland PWH (2011) HomeoDB2: Functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution and Development*, **13**, 567–568.
162. Zhong YF, Butts T, Holland PWH (2008) HomeoDB: A database of homeobox gene diversity. *Evolution and Development*, **10**, 516–518.
163. Zhu W, Zeng N, Wang N (2010) Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS® implementations. *Northeast SAS Users Group 2010: Health Care and Life Sciences*, 1–9.
164. Zmasek CM, Godzik A (2012) This Déjà Vu Feeling-Analysis of Multidomain Protein Evolution in Eukaryotic Genomes. *PLoS Computational Biology*, **8**.
165. Zwarycz AS, Nossa CW, Putnam NH, Ryan JF (2016) Timing and scope of genomic expansion within annelida: Evidence from homeoboxes in the genome of the earthworm *eisenia fetida*. *Genome Biology and Evolution*, **8**, 271–281.

7 APPENDICES

7.1 GITHUB REPOSITORY FOR PIPELINE SCRIPTS

7.1.1 COMPARATIVE GENOMICS

<https://github.com/CristiGuijarro/ComparativeGenomics>

7.1.2 HOMEODOMAINS

<https://github.com/CristiGuijarro/Homeodomains>

7.2 MATERIALS SOURCES

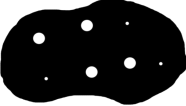







7.2.1 GENOME FILES










Species	Database Source
<i>Danio rerio</i>	Ensembl
<i>Xenopus tropicalis</i>	Uniprot
<i>Anolis carolinensis</i>	Uniprot
<i>Gallus gallus</i>	Uniprot
<i>Homo sapiens</i>	Ensembl
<i>Oikopleura dioica</i>	Uniprot
<i>Ciona intestinalis</i>	Ensembl
<i>Botryllus schlosseri</i>	http://botryllus.stanford.edu/botryllusgenome/download/
<i>Branchiostoma floridae</i>	Uniprot
<i>Branchiostoma belcheri</i>	http://genome.bucm.edu.cn/lancelet/download_data.php
<i>Patiria miniata</i>	http://www.echinobase.org/Echinobase/
<i>Strongylocentrotus purpuratus</i>	Uniprot
<i>Lytechinus variegatus</i>	http://www.echinobase.org/Echinobase/
<i>Saccoglossus kowalevskii</i>	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000003605.2_Skow_1.1
<i>Ptychodera flava</i>	https://groups.oist.jp/molgenu/hemichordate-genomes
<i>Hypsibius dujardini</i>	http://badger.bio.ed.ac.uk/H_dujardini/home/download large version
<i>Ramazzottius varieornatus</i>	http://www.ddbj.nig.ac.jp/whatsnew/wn160915-e.html
<i>Stegodyphus mimosarum</i>	Uniprot
<i>Acanthoscurria geniculata</i>	http://www.nature.com/ncomms/2014/140506/ncomms4765/full/ncomms4765.html#supplementary-information
<i>Ixodes scapularis</i>	Uniprot
<i>Mesobuthus martensii</i>	http://lifecenter.sgst.cn/main/en/scorpion.jsp
<i>Limulus polyphemus</i>	http://ryanlab.whitney.ufl.edu/genomes/Lpol/
<i>Daphnia pulex</i>	Uniprot
<i>Parhyale hawaiiensis</i>	https://s3-eu-west-1.amazonaws.com/phaw/phaw.3.0.genes.prot.fa


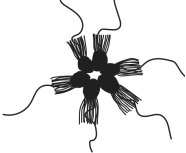
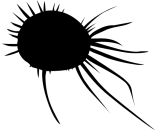
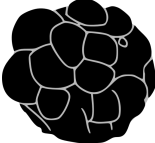
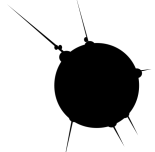
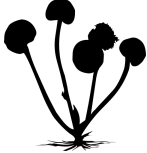
<i>Tribolium castaneum</i>	Uniprot
<i>Drosophila melanogaster</i>	EnsEMBL
<i>Zootermopsis nevadensis</i>	Uniprot
<i>Strigamia maritima</i>	EnsEMBL
<i>Brugia malayi</i>	Uniprot
<i>Caenorhabditis elegans</i>	EnsEMBL
<i>Trichinella spiralis</i>	Uniprot
<i>Romanomermis culicivorax</i>	http://nematodes.org/genomes/romanomermis_culicivorax/
<i>Helobdella robusta</i>	Uniprot
<i>Capitella teleta</i>	Uniprot
<i>Lingula anatina</i>	EnsEMBL
<i>Crassostrea gigas</i>	Uniprot
<i>Pinctada fucata</i>	http://marinegenomics.oist.jp/pearl/viewer/download?project_id=20
<i>Octopus bimaculoides</i>	Uniprot
<i>Lottia gigantea</i>	Uniprot
<i>Intoshia linei</i>	http://www.ncbi.nlm.nih.gov/genome/?term=Intoshia+linei%5Borgn%5D
<i>Hymenolepis microstoma</i>	Uniprot
<i>Echinococcus multilocularis</i>	Uniprot
<i>Gyrodactylus salaris</i>	http://invitro.titan.uio.no/gyrodactylus/downloads.html http://smedgd.stowers.org/downloads/
<i>Schmidtea mediterranea</i>	#MAKER_annotations_8211_Protein_FASTA_files http://parasite.wormbase.org/Macrostomum_lignano_prjna284736/Info/Index/
<i>Macrostomum lignano</i>	Uniprot
<i>Schistosoma japonicum</i>	Uniprot
<i>Adineta vaga</i>	http://www.genoscope.cns.fr/adineta/data/
<i>Aiptasia pallida</i>	http://aiptasia.reefgenomics.org/download/
<i>Nematostella vectensis</i>	Uniprot
<i>Acropora digitifera</i>	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF_000222465.1_Adig_1.1
<i>Hydra magnipapillata</i>	http://compagen.zoologie.uni-kiel.de/datasets.html
<i>Thelohanellus kitauei</i>	Uniprot
<i>Pleurobrachia bachei</i>	http://neurobase.rc.ufl.edu/pleurobrachia/download
<i>Mnemiopsis leidyi</i>	http://compagen.zoologie.uni-kiel.de/datasets.html
<i>Trichoplax adhaerens</i>	Uniprot
<i>Leucosolenia complicata</i>	http://datadryad.org/resource/doi:10.5061/dryad.tn0f3/4?show=full
<i>Sycon ciliatum</i>	http://datadryad.org/resource/doi:10.5061/dryad.tn0f3/3
<i>Amphimedon queenslandica</i>	http://amphimedon.qcloud.qcif.edu.au/downloads.html v2.1
<i>Oscarella carmela</i>	http://compagen.zoologie.uni-kiel.de/datasets.html
<i>Monosiga brevicollis</i>	Uniprot
<i>Salpingoeca rosetta</i>	Uniprot
<i>Capsaspora owczarzaki</i>	Uniprot
<i>Corallochytrium limacisporum</i>	Email Direct https://figshare.com/articles/
<i>Creolimax fragrantissima</i>	http://compagen.zoologie.uni-kiel.de/datasets.html Creolimax_fragrantissima_genome_data/1403592
<i>Abeoforma whisleri</i>	Email Direct
<i>Sphaeroforma arctica</i>	Uniprot
<i>Ichthyophonus nk52</i>	Email Direct
<i>Polysphondylium pallidum</i>	Uniprot
<i>Dictyostelium discoideum</i>	Uniprot
<i>Acanthamoeba castellanii</i>	Uniprot

<i>Entamoeba histolytica</i>	Uniprot
<i>Fonticula alba</i>	Uniprot
<i>Saccharomyces cerevisiae</i>	EnsEMBL
	http://mycor.nancy.inra.fr/IMGC/TuberGenome/download.php? select=fast
<i>Tuber melanosporum</i>	Uniprot
<i>Allomyces macrogynus</i>	Uniprot
<i>Spizellomyces punctatus</i>	Uniprot
<i>Thecamonas trahens</i>	Uniprot
	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/ GCF_000350225.1_ASM35022v2
<i>Chondrus crispus</i>	Uniprot
<i>Chlamydomonas reinhardtii</i>	Uniprot
	http://www.plantmorphogenesis.bio.titech.ac.jp/ ~algae_genome_project/klebsormidium/kf_download/ 160614_klebsormidium_v1.1_AA.fasta
<i>Klebsormidium flaccidum</i>	Uniprot
	http://cyanophora.rutgers.edu/cyanophora/ Cyanophora_paradoxa_MAKER_gene_predictions-022111-aa.fasta
<i>Cyanophora paradoxa</i>	Uniprot
<i>Volvox carteri</i>	Uniprot
<i>Ostreococcus tauri</i>	Uniprot
<i>Physcomitrella patens</i>	Uniprot
<i>Selaginella moellendorffii</i>	Uniprot
<i>Arabidopsis thaliana</i>	Uniprot
<i>Giardia lamblia</i>	Uniprot
	http://tritrypdb.org/common/downloads/Current_Release/ TruzziDm28c/fasta/data/
<i>Trypanosoma cruzi</i>	Uniprot
<i>Naegleria gruberi</i>	Uniprot
<i>Trichomonas vaginalis</i>	Uniprot
	http://genome.jgi.doe.gov/Guith1/Guith1.download.html
<i>Guillardia theta</i>	EnsEMBL
<i>Emiliana huxleyii</i>	EnsEMBL
<i>Toxoplasma gondii</i>	EnsEMBL
<i>Paramecium tetraurelia</i>	Uniprot
<i>Symbiodinium minutum</i>	Uniprot
	http://marinegenomics.oist.jp/symb/viewer/download?project_id=21
<i>Perkinsus marinus</i>	EnsEMBL
<i>Bigeloviella natans</i>	EnsEMBL
<i>Reticulomyxa filosa</i>	Uniprot
<i>Phaeodactylum tricorutum</i>	Uniprot
<i>Thalassiosira pseudonana</i>	Uniprot
	http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf? organism=Aplke1
<i>Aplanochytrium kerguelense</i>	Uniprot
<i>Phytophthora infestans</i>	Uniprot

7.2.2 PHYLOPIC SILHOUETTES

Phylopic.org silhouette	Source	Distribution
	http://phylopic.org/image/8b28733a-9eab-4510-9dcf-9235aaf44d14/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/27a08157-6943-4faf-9aa3-980249e5c376/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/59246275-9cc9-4adf-85c0-930b7f7b2633/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/26c0169f-b4a2-4871-8b30-e00db6e5958d/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/19f846e4-62f6-4081-9e52-b933792c5bcd/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/2e265034-1a52-4a5d-8b3c-2435079fa38b/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/77acd947-7660-4a34-8932-0b59d3cfe3fb/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/3042a73d-353d-4191-811f-9b12f57c958c/	Public Domain Dedication 1.0 license No Copyright

	http://phylopic.org/image/2dee030d-9f6d-4bab-87c4-c46869839b30/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/b81d728e-0e8c-4faf-8f40-edf999143f10/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/9da7781e-48eb-407d-a13f-d6de8954dde2/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/80639d7b-0856-45c1-ab44-7f075accda89/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/d6af4346-e56c-4f3d-84c7-fba921a293f1/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/b1b277bb-416a-4cdc-a07c-7e31d970b293/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/0fde2bb6-0472-4273-bf46-2d6073fa8fbc/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/09b58bf3-a79b-4740-a2c2-6c2940d8cd9a/	Creative Commons Attribution 3.0 Unported https://creativecommons.org/licenses/by/3.0/ Image by Mali'o Kodis, image from the Biodiversity Heritage Library No Copyright
	http://phylopic.org/image/cd62afdf-5b96-44fc-89b7-60d018cd4d5a/	Creative Commons Attribution 3.0 Unported https://creativecommons.org/licenses/by/3.0/ Image by Noah Schlottman, photo by Martin V. Sørensen No Copyright

	http://phylopic.org/image/e225d12f-acd5-4483-80c2-01a757bd4738/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/e5412511-0457-4887-bafb-0bd4bbc0809a/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/7ca79f27-b79b-4d24-8c22-c30a4c272749/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/dd39d950-11fb-4957-a320-51251ac34182/	Public Domain Mark 1.0 license No Copyright
	http://phylopic.org/image/1e40ed33-4524-452e-a482-af52e39b9c63/	Public Domain Dedication 1.0 license No Copyright
	http://phylopic.org/image/8cff2d66-6549-44d2-8304-d2dfecf53d78/	Public Domain Dedication 1.0 license No Copyright

7.3 SUPPLEMENTARY MATERIALS/DATA FILES

7.3.1 MYSQL DATABASE (ON USB)

Relational database containing 102 eukaryote genomes used in this thesis with the MCL analysis on the BLASTp results uploaded as homology groups. Holds information for each HG in terms of gene ontologies and protein families. All HG extractions and gain/loss inferences were queried against this database, which is necessary for many of the scripts in the GitHub repository to work and reproduce results.

7.3.2 ADDITIONAL GOS (ON USB)

Multiple comma separated value (csv) files containing specific GO analysis for each core and non-core HG inferred in chapter 2.

7.3.3 HOMEBOX CLASSIFICATION LOG FILE (ON USB)

Log file in csv file format listing annotations and FASTA accessions for each sequence extracted and classified in the homeobox analysis in chapter 3.

7.3.4 HOMEBOX GENE TREE FILES (NEWICK) (ON USB)

Newick format gene tree files with bootstraps and branch lengths for each homeobox HG.

7.3.5 ADDITIONAL HOMEBOX GENE TREE IMAGES (ON USB)

Each tree is a single HG. On the left side is the whole tree, on the right is the highlighted portion of the tree with homeodomains classified/identified in this thesis. These have not been included in the main thesis because they do not show particularly remarkable results.