UNIVERSITY OF LEEDS

An investigation into the contribution of gene remodeling to protein coding gene family evolution across the Metazoa

Peter Mulhair

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Faculty of Biological Sciences

School of Biology

June 2020

Acknowledgements

The work carried out over the course of this PhD has been an incredible journey of collaboration and support from a number of people. I am completely indebted to those who have helped me to achieve this degree.

My family has been a consistent source of inspiration and motivation my whole life. From early on my brothers and I were encouraged to pursue our interests to the best of our ability. They encouraged my move to another country to do this PhD and have always been there when I needed them the most.

My friends have been incredibly supportive, especially during the most difficult times of the past few years. Having a close knit network of amazing people has given me the confidence and ability to achieve my goals. Both in science and life, I know they will continue to help and support me.

My supervisor Mary is the main reason I am in this position today. Her undergraduate lectures on comparative biology and evolution sparked an interest that will influence my whole career. Her way of teaching, expertise, guidance, and kindness are unparalleled and I have learned so much from working with her. She has been the best mentor I could have wished for and I hope we can continue our close friendship and working relationship for a long time.

Finally, I wish to acknowledge a very good friend of mine and my undergraduate colleague, Robert Hickey. Without Rob to guide me and push me to get the best out of myself, I very much doubt I would be in this incredibly lucky position. I owe so much to him, and I hope he knows how much of an impact he has had on my career and my life.

Abstract

This thesis explores gene evolution throughout the history of Metazoa. This group of multicellular organisms represents a wide range of diversity, embodied by the number of species, the multitude of morphological and developmental traits, and the complexity within the genetic elements dictating these traits. Although a significant amount of research has been carried out on gene family emergence and expansion in animal genomes, comparatively little research has been published on how these genes are formed. Of specific interest here is the role of complex, reticulated mechanisms of gene evolution in forming new genes, these include processes such as gene fusion and fission - hereafter referred to as gene remodeling. Current methods of gene family prediction are not sensitive to these mechanisms of gene evolution. We apply both network and phylogenetic models to characterise the traits and role of gene remodeling across Metazoa and take a data focused approach to attempt to resolve remaining issues within the animal tree of life (AToL). In Chapter 2 we took a novel network approach to quantify the contribution of gene remodeling events to novel protein coding gene family evolution in the animal tree of life. Using graph theory we analysed the partial homology shared between a set of animal proteomes spanning most major clades, and we placed these gene remodeling events onto the species tree. In addition to this, in Chapter 3 we sought to assess the phylogenetic properties of these events and their ability to reconstruct AToL, ultimately our aim was to determine if gene fusions could be deployed to resolve contentious regions within AToL. As a consilient approach is most desirable in phylogeny reconstruction, in Chapter 4 we examine the potential for resolving AToL using a combination of data types, i.e. gene fusions and previously published phylogenomic datasets. Specifically, we examined potential issues in annotation of homology and orthology within previously published animal phylogenomic datasets and focussed on determining what impact inaccurate definitions of orthology have had on resolving difficult or contentious parts of AToL.

Table of contents

List of Tables

List of Figures

Abbreviations

AToL	The animal tree of life
BI	Bayesian inference
BIC	Bayesian Information Criterion
CHGs	Cluster of homologous genes
CI	Consistency index
DFS	Depth First Search
ESTs	Expressed sequence tags
GTR	General time reversible
HGT	Horizontal gene transfer
HMM	Hidden Markov models
HOGs	Hierarchical Orthologous groups
ILS	Incomplete lineage sorting
LBA	Long branch attraction
LCA	Last common ancestor
LRT	Likelihood ratio test
MCMC	Markov-chain Monte Carlo
ML	Maximum likelihood
MSA	Multiple sequence alignment
MA	Million years
MYA	Million years ago
ORF	Open reading frame
OTUs	Operational taxonomic units
PP	Posterior probability
RGCs	Rare genomic changes
SINEs	Short interspersed elements
SSN	Sequence similarity network
TE	Transposable element
UTR	Untranslated region
WGD	Whole genome duplication

Chapter 1: Introduction

1.1 Principles and practices of phylogenetic reconstruction

From Darwin's initial proposals for the use of trees to classify life in his seminal work in 1859, the history of systematics has developed substantially. Following the emergence of the modern synthesis, the field developed from basic classification into a sophisticated tool to understand the processes of molecular evolution, adaptation, and ultimately the diversity of life. This thesis will focus on the use of phylogenetics and comparative genomics to interpret the history and evolution of the protein coding content within animals and the patterns of speciation that led to the diversity we see today. To do so we will use phylogenetic trees, networks, and molecular sequence data in combination to uncover deep and complex patterns of evolution.

Modern phylogenies are essentially graphs that represent relationships; they consist of vertices (representing nodes on a tree) and edges (which represent branches). The leaves, often referred to as operational taxonomic units (OTUs) or terminal nodes, describe the taxa, or species, while the topology of the tree describes the relationships between these taxa. Trees can be either rooted or unrooted, where a root represents a point of origin from which one can trace the evolutionary history along the branches to the leaves. In this manner, rooted trees are a useful way to display directionality, as well as relationships between taxa. Theoretically at any point in time across a tree, a point of origin (e.g. for a given gene, morphological trait, or species) can be placed on a branch or node. Branch lengths can either represent rates of evolution (in an unconstrained tree), or time (in an ultrametric, or time tree), which gives insight into the rates of gain and loss or substitution within a gene or protein and timing for the origin of a species, respectively. In the latter case, trees are essential to provide the context to understand when and, in the case of fossil records, where the point of origin lies in earth's life history. Phylogenetic trees have been at the heart of evolutionary biology since the origins of the field, and even though they can be misleading in particular contexts (Doolittle 1999), they continue to be an integral part of our quest to understand the large and small-scale processes of how evolution underlies all life on earth.

1.1.1 Phylogenetic reconstruction using molecular data

Before the advent of protein and DNA sequencing, and more recently the influx of genomic data, the focus of phylogenetics/systematics was to use morphological data to describe species relationships. Nowadays, phylogenetics is applied in almost every branch of biology including: population genetics (Edwards 2009); pathogen evolution and epidemiology (Marra et al. 2003; Hadfield et al. 2018); cancer development and evolution (Somarelli et al. 2017; Salipante and Horwitz 2006); the identification of regulatory elements (Kellis et al. 2003; Lindblad-Toh et al. 2011); and, reconstructing ancestral genomes (Paten et al. 2008; Paps and Holland 2018).

Most phylogenetic analyses require some form of sequence data and a model which is used to estimate the probability of observing said data. This results in a statistical estimation of the relationships between the sequences which is represented on a tree (which is a specific instance of a network). Phylogenomics describes the application of genome-scale datasets involving large numbers of genes or proteins for phylogeny reconstruction. This has resulted in major advances in phylogenetic studies and systematics, however has also led to conflict between different datasets and data types and issues in reaching resolution due to statistical and time constraints. It is clear that many of the remaining phylogenetic issues, particularly within the animal phylogeny (see Section 1.2.2), may only be resolved by applying better models and by taking a consilient approach. Indeed, with the advent of more sequence data and larger phylogenomic datasets it is vital to remind ourselves that more data does not always mean more accurate inference, and data quality, accurate orthology inference, and appropriate model selection play an increasingly important role (Philippe and Roure 2011; Shen et al. 2017) (see Section 1.1.3 for more information).

1.1.2 Data types in phylogeny reconstruction

1.1.2.1 Discrete characters & morphology

Discrete character data, e.g. morphological traits, have a long history of use in systematics. Indeed, a large proportion of the tree of life, in particular the animal tree of life (AToL), was originally constructed through the use of morphological characters and many of those relationships still hold today. Molecular data provides a vast increase in the number of characters from which to infer a tree, and we see that traditional morphology and discrete characters are currently relied upon less for phylogeny reconstruction (Mooi and Gill 2010;

Lee and Palci 2015). Yet morphological data still has a key role to play as it (i) contains important phylogenetic information on the relationships between extinct and living species, and (ii), is essential for calibrating divergence times and rates of living species and in contentious parts of the tree (Wiens 2004). The issues with morphological characters for phylogenetic inference are well known (Scotland et al. 2003), for instance homology is based on assessment by experts, and convergent evolution is thought to be common. Nonetheless, arguments, based on unresolved branching orders that remain even after molecular phylogenetics analyses and the fact that there is no single source of data that contains all information required to resolve the tree of life, have cautioned that disregarding this type of data for phylogenetic reconstruction methods would overall decrease our power to elucidate species relationships (Wiens 2004; Jenner 2004).

Other types of discrete character data may also be used in conjunction with molecular data. For example, the presence or absence of genes in genomes has been suggested as a powerful approach (Fitz-Gibbon and House 1999; Lake and Rivera 2004; Ryan et al. 2013; Pisani et al. 2015; Tarver et al. 2018). Unlike standard phylogenomic approaches using concatenated amino acid sequence data for a large number of gene families, discrete data of gene presence and absence do not necessarily require accurate predictions of orthology. This is due to the fact that, in phylogenetic reconstruction using sequence data, incorrect ortholog assignment (i.e. the inclusion of paralogs within an orthologous gene family) introduces significantly noisy signal from the different patterns of substitution within paralogous genes, whereas alternatively using the patterns of presence and absence this issue is not present and thus does not have a major effect on the interpretation of the phylogenetic information. For the placement of deeply diverging lineages, such as those at the root of AToL, these traditional methods of large supermatrices of sequence data may be unable to fully resolve these branching orders. Instead, novel evolutionary signal may be gleaned from this discrete data type. Discrete character data has been applied to the contentious rooting of the animal tree, where patterns of presence and absence of both homologous and orthologous gene families in animals finds support for sponges as the primary emerging phylum (Pett et al. 2019). While this type of data may provide a novel approach to addressing remaining issues, more in depth analyses are required to measure the usefulness of gene presence/absence data, such as the incorporation of information about whole genome duplication events, varying rates of gain and loss in relation to gene family size, and the effects of different methods for homology and orthology inference. Further to this, it would be important to take into account the quality of the assembly and annotation of the genomes used, as the accurate patterns of gene presence and absence could be influenced by these factors.

1.1.2.2 Molecular data: nucleotide and amino acid data

The emergence of DNA sequencing technologies revolutionised the field of systematics. With an increase in methods to obtain molecular data, there was a significant increase in the size of datasets, and the types of different molecular data from which evolutionary information could be gleaned to infer a phylogenetic tree. Initial phylogenetic inference studies used nucleotide sequence data, often in the form of expressed sequence tags (ESTs), or mitochondrial DNA, due to their small size and the limited range of sequencing technologies (Dunn et al. 2008; Bourlat et al. 2006; Delsuc et al. 2006; Philippe et al. 2005). Later, nucleotide data from nuclear protein coding genes were applied due to the ease of inferring orthologous sequences and alignment construction across different species (Thomson et al. 2010). However, the use of nucleotide data can lead to issues in modeling the phylogenetic information, particularly when considering coding sequence data. The nucleotide positions within coding regions of the genome, especially the third position in a codon, are susceptible to saturation due to poorly constrained selection at some sites. This can lead to an underestimation of inferred rates of substitution which can have major implications on phylogenetic inference, namely leading to biases such as long branch attraction (LBA) (Yang and Rannala 2012). LBA is a well documented source of systematic error within phylogenetic studies, which derives from these patterns of unequal rates of evolution within different lineages in the tree (Felsenstein 1978). It manifests itself by incorrectly grouping lineages with long branches (i.e. fast rates of evolution) based on their shared rates of evolution rather than being closely related species.

Alternatively, nucleotide sequence data from non-coding regions has been found to be informative and in certain cases may contain more phylogenetic signal than coding sequences. For example, when comparing the usefulness of coding sequence data with intron sequences to resolve the phylogeny of Laurasiatheria (a clade of placental mammals), Chen et al. (2017) found that the non-coding intron data contained stronger, more congruent phylogenetic signal. It has now become common practice (particularly at deep timescales such as the root of the animal tree) to use amino acid sequence data removing some level of saturation and downstream phylogenetic bias from the data. This is perhaps counter-intuitive as the nucleotide sequence data is the primary source of information obtained from the sequencing of the genome, and as such some information may be lost in the translation to amino acid sequences (Townsend et al. 2008). Nonetheless, amino acid data has become the default especially for inferring deep diverging relationships (Laumer 2018), as is the case for many animal phylogenomic studies.

Amino acid re-coding has recently gained a reinvigorated interest (Morgan et al. 2013; Feuda et al. 2017; Laumer et al. 2019; Marlétaz et al. 2019; Philippe et al. 2019). This involves transforming the amino acid matrix into a smaller, higher order alphabet, known as Dayhoff categories (M. O. Dayhoff 1978) based on the physicochemical properties of amino acids. This reduces the 20 character states of amino acids to a smaller set, 6 in the case of the commonly used "Dayhoff-6" groupings (Susko and Roger 2007). The main reason for reducing the data matrix in this case is that it masks higher rates of substitution that occur within certain groups mitigating bias in the sequence data that may be caused by convergent patterns of amino acids as a result of chance rather than true evolutionary history. This bias is called compositional heterogeneity (Foster 2004), and may also be accounted for using sophisticated models of evolution such as those applied in P4 (Foster 2004) or PhyloBayes (Lartillot and Philippe 2004), which can account for different compositions between lineages (Node-discrete composition heterogeneity model implemented in P4) and between sites (CAT model implemented in PhyloBayes). The application of re-coding strategies has led to increased support for contentious regions in the animal phylogeny such as Porifera as the primary emerging animal lineage (Feuda et al. 2017) and sister grouping of Xenacoelomorpha with Ambulacraria (Philippe et al. 2019). Recently, the use of data re-coding to limit the effects of saturation within amino acid sequences has been contested (Hernandez and Ryan 2019). In particular, they found that non-recoding strategies outperformed 6-state amino acid re-coding in resolving sequence data containing compositional heterogeneity. Their findings suggest that while 6-state amino acid re-coding limits the effects of compositional heterogeneity, it also suffers from loss of phylogenetic signal due to the loss of a larger number of character states. Going forward, other re-coding strategies, such as schemes reducing amino acid data to 9, 12, 15 or 18 states may provide optimum trade-off between loss of signal and ability to account for compositional heterogeneity. Additionally, while models such as those applied in P4 (Foster 2004) contain a larger number of parameters and are thus computationally intensive, their application to unresolved regions may provide the necessary modeling of the data. This is particularly relevant at deep timescales where the effects of compositional heterogeneity on our ability to model sequence evolution may be profound.

1.1.2.3 Rare genomic changes as molecular markers

Rare genomic changes (RGCs) are a useful source of genomic characters to test phylogenetic hypotheses and they have been used as alternative phylogenetic markers to complement, for example, traditional protein coding data analyses (Rokas and Holland 2000; Bleidorn 2017). Examples of RGCs include gene fusions, mobile elements, microRNAs, introns, retroposon integration, and gene order rearrangements. RGCs are an attractive alternative to protein coding regions as they are often free from convergent site evolution (Rokas and Carroll 2008), heterogeneous rates of site substitution between lineages (Philippe and Lopez 2001), and functional constraints (Lee 1999). Another important feature of RGCs is how they are interpreted to change over time - primary sequence data is assumed to evolve, in general, in a clock-like manner where the number of changes within the sequence data are expected to follow a linear distribution. Alternatively, RGCs are thought to evolve in a non-clocklike manner, where the rare events generally occur in large periods of expansion, loss, or rearrangement, meaning modeling these characters is often difficult. However, this rate of evolution makes RGCs suitable for addressing difficult sets of relationships, for example those caused by short internal nodes that underwent rapid radiations and thus had little time for signal to accumulate, or long terminal branches where there has been enough time for signal to deteriorate (Boore 2006). There are a number of examples of difficult to resolve relationships with AToL (see Section 1.2.2.2), such as the root of the animal tree (where a large amount of time has passed to extant lineages making it difficult to extract phylogenetic signal for this region) and particularly at the branches subtending Bilateria (where short branches leading up to the clade were followed by large radiation events). These are prime cases where RGCs may provide key insight and complement protein coding sequence datasets. RGCs are thought to evolve in a non-clock-like manner, mainly due to the large genomic effects they cause, meaning they are rare events. Additionally, the rate at which these events occur means that they are less susceptible to homoplastic traits, such as convergent or parallel evolution and secondary loss (Rokas and Holland 2000). For example, microRNAs, which have been promoted as useful markers, lack any evidence of convergence (Tarver et al. 2013) and are thought to undergo low rates of secondary loss, with just a small number of microRNA families displaying the highest rates of loss (Tarver et al. 2018). Similarly, SINE retrotransposons are large scale mutations shown to have low rates of secondary loss and have been applied to phylogenetic issues in AToL (Shedlock and Okada 2000). Therefore, with respect to these characteristics RGCs are thought to be a useful set of markers of common descent.

Some examples of the application of RGCs to phylogeny reconstruction include: 51 retrotransposon events used to place parrots as the sister lineages to passerine birds (Suh et al. 2011); retroposon integrations using short interspersed elements (SINEs) providing support for the sister group of whales and hippos (Nikaido et al. 1999); and the signature sequence of Hox genes which supports the division of Protostomia into Lophotrochozoa and Ecdysozoa (de Rosa et al. 1999). More recently, microRNAs have emerged as a key set of molecular markers to address contentious regions in the animal phylogeny (Sperling et al. 2011; Campbell et al. 2011; Helm et al. 2012). Low rates of secondary loss, no evidence of convergent evolution (Tarver et al. 2013), and widespread distribution across the animal tree makes these regulatory genes ideal candidates to address remaining issues in AToL (Tarver et al. 2018).

1.1.2.4 Consilience in phylogeny reconstruction

For the remaining contentious regions within AToL it is clear that a singular approach or datatype may not be sufficient to reach resolution. Instead, the most compelling support for any relationship is based on consilience across different datasets including different data types. The topology of the mammal tree for example has been under consistent debate for quite some time (Meredith et al. 2011; McCormack et al. 2012; Song et al. 2012; O'Leary et al. 2013; Morgan et al. 2013). Recent work has combined protein coding sequences, microRNA presence/absence and sequences, and the reanalysis of previously published phylogenomic studies using better fitting models of evolution to address this question (Tarver et al. 2016). This multi-pronged approach provided robust support across multiple data types to place Atlantogenata as sister lineage to all other placental mammals; a topology supported by a previous study which applied heterogeneous models using amino acid data and Dayhoff re-coding (Morgan et al. 2013). This has shown that more sequence data alone does not necessarily mean greater accuracy and it is important that data and methodological issues should be considered. Multiple studies have shown the importance of limiting systematic errors within the data and the workflow to ensure accurate modeling of appropriate data. Some examples include ensuring accurate orthology inference (Siu-Ting et al. 2019), removal of cross-contaminants if transcriptome data is applied (Simion et al. 2017), removal of poorly aligned regions (Di Franco et al. 2019), removal of sites or genes that seriously violate model assumptions (Philippe and Roure 2011), and appropriate model selection (Morgan et al. 2013; Feuda et al. 2017). Finding a single, consistent AToL will require robust phylogenetic approaches as well as consilience across data types.

1.1.3 Key practices and pitfalls in phylogeny reconstruction

The workflow of a phylogenetic analysis can be complex and variable. The standard components of an analysis include taxon sampling and data collection, inference of homologous gene families, construction of multiple sequence alignments, model selection, and finally phylogenetic inference in a given statistical framework. Any of these points in the workflow are dependent on the phylogenetic question at hand, and selection of the optimum parameters is not an easy task. This section highlights some key practices and potential pitfalls in phylogeny reconstruction.

1.1.3.1 Taxon sampling

The generation of the initial dataset has obvious consequences for downstream results. It is important to sample taxa that cover the diversity of all clades surrounding your clade or branch of interest. Increased taxon sampling has been known for some time to be vital for accurate phylogenetic inference (Zwickl and Hillis 2002). Almost all phylogenetic inference methods produce unrooted trees, meaning that the direction of time across the tree cannot be interpreted. To create a rooted phylogeny, taxon sampling of a group of species outside the group of interest is necessary. This group is called the outgroup, while the species of interest are referred to as the ingroup. Outgroup selection is as important as ingroup selection as it places the root in the clade of interest and thus affects ingroup species relationships (Lyons-Weiler et al. 1998; Wilberg 2015). The effects of both extensive taxon sampling and outgroup selection on a phylogeny can be found in a number of animal phylogenomic studies analysing the root of the animal tree. Philippe et al. (2011) assessed the impact of reduced taxon sampling by comparing three studies, two with limited sampling of non-bilaterian species (Dunn et al. 2008; Schierwater et al. 2009) and one with greater taxon sampling (Philippe et al. 2009). The study with a larger number of non-bilaterian species (Philippe et al. 2009) found greater support for the monophyly of clades such as Porifera and Eumetazoa, compared to the other two studies which showed low support for these well defined clades. To assess the impact of taxon sampling on this highly supported topology, they reduced the taxon sampling of non-bilaterians from Philippe et al. (2009) to match the other to studies with low sampling and low support and reran the same phylogenetic methods as in the original study. They found drastically decreased support for well known clades such as Cnidaria, and no support for monophyletic clades of Porifera and

Eumetazoa, consistent with the idea that poor taxon sampling has significant implications on tree topology and support (Philippe et al. 2011). More recently, the monophyly of Characiforme fishes was recovered with full support in a study consisting of a larger sampling of species than other studies addressing the same topology, with sampling covering 23 out of the 24 clades within Characiforma (Betancur-R. et al. 2019). When they reduced the full dataset to obtain the same amount of sequence data but with less species, they failed to recover a consistent predominant branching order at the base of Characiforma, confirming the effects of taxon sampling on resolving this particular phylogeny (Betancur-R. et al. 2019).

Taxon sampling affects all downstream steps in phylogenetic analysis, including detection of orthologous sequences, alignment of homologous sites, and tree inference. The degree of sampling should be tailored to the evolutionary depth at hand, in order to capture the true evolutionary steps of each gene family under analysis (e.g. speciation, duplication, horizontal transfer, loss, parallel evolution). However, for certain regions of AToL where phylogenetic signal is weak, for example at ancient or short internal branches, even deep taxon sampling may not be enough to provide clarity on the true branching orders (Philippe et al. 1994). Additionally, including too many fast evolving outgroup species can greatly influence the ingroup topology. For example, resolving the topology at the root of the animal tree is particularly difficult owing to the fast evolving lineage of Ctenophora. The inclusion of outgroups that are too distantly related could result in systematic effects such as long branch attraction, potentially pulling the ctenophore lineage to the root of the animal tree (Pisani et al. 2015). In this case, it may be useful to run parallel phylogenetic analyses with different numbers of outgroup species to compare the effects of outgroup selection on the topology at the root of AToL (see Laumer et al. (2019)).

1.1.3.2 Ortholog assignment

Creating gene families consisting of orthologous genes in single copy is by many considered the standard for phylogenomic studies. In any case, differentiating between orthologous and paralogous genes is an essential step to obtain optimum signal from the dataset. In reality there are a number of methods and software to infer these relationships, and due to difficulty in creating the optimal possible set of orthologous families a significant amount of work has gone into standardising these methods (Altenhoff et al. 2016; Nichio et al. 2017). Some ortholog clustering tools are based on all-by-all alignment methods e.g. OMA (Altenhoff et al. 2019) and OrthoFinder (Emms and Kelly 2019), while some use predefined orthogroups and cluster using graph or tree based algorithms e.g. Orthograph (Petersen et al. 2017).

Although there are a number of methods and approaches to carry out this step, on which all evolutionary hypotheses are inferred, there still remains a number of concerning issues that are not addressed by the field. Perhaps the most well known issue is the innate difficulty in annotating true orthology. As ortholog/paralog annotation ideally requires a species tree for truly informed results (Szöllősi et al. 2015), this causes problems as the species tree is usually unknown, and is in fact the end goal. This can be confounded by the use of transcriptomic sequence data as a source of genes, which is still widely carried out (Simion et al. 2017; Philippe et al. 2019), that may cause genes that are not expressed in the given tissue at the time of sampling to be seen as absent from an organism for example. Additionally, it can be very challenging to distinguish between genes that have been in single copy since the last common ancestor (LCA) of the species versus those that were ancestrally in multi-copy but are in single copy in extant species due to subsequent lineage-specific loss. This means that a gene that looks like a single copy ortholog in current species may actually have been an ancestral paralog. In a study looking at the contentious phylogeny of Lissamphibia, inadvertent paralog selection has been found to confound phylogenetic inference, resulting in incorrect topologies with high support (Siu-Ting et al. 2019). Given inferred high rates of ancestral duplication and subsequent lineage loss throughout the animal tree (Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020), it is highly likely that these patterns are causing incongruent topologies with high support in studies addressing deep nodes in the animal phylogeny. Moreover, it is also possible that true orthologs have diverged to such a degree that they may be missed by the ortholog clustering tools, even in closely related species if taxon sampling is not dense enough. This effect of missing ortholog annotation was observed recently in flatworms when the ortholog detection methods were extended (Martín-Durán et al. 2017). Finally, each of the current clustering methods for finding homologous and orthologous genes are acting under the assumption that gene evolution is strictly tree-like (see Section 1.3.3 for more detail). We have known for quite some time that genes do not always evolve in this way, and more complex reticulate processes are possible (Haggerty et al. 2014). For example, gene families that emerged through mechanisms of gene fusion would be overlooked by current methods, thus removing potentially informative phylogenetic signal from the initial dataset. Orthology is a fundamental aspect of phylogenomics, one that is not trivial to define nor one that is often given adequate consideration. While some studies have assessed the impact of different methods of orthology annotation (Shen et al. 2018; Fernández et al. 2018; Altenhoff et al. 2019), more phylogenomic studies should consider the creation of orthologous groups as an important step to quality check the assumptions being made. While much consideration is given to appropriate taxon sampling (Philippe et al. 2011), accurate model selection (Morgan et al. 2013; Pisani et al. 2015; Feuda et al. 2017), and missing data (Roure et al. 2013; Streicher et al. 2016), measuring the effects of ortholog selection on the resulting phylogenetic tree may provide greater insight into difficult to resolve contentious parts of AToL.

1.1.3.3 Multiple sequence alignments

Alignment is the process of aligning positions that are thought to be homologous within related sequences, i.e. characters that have common ancestry (Kemena and Notredame 2009). The homologous sites within the resulting multiple sequence alignments (MSAs) provide the relevant evolutionary information from which the models of evolution will reconstruct the history of the genes and species at hand. Similar to ortholog assignment, methods and software to align sequences are numerous. Often within the workflow of a phylogenomic study, relatively little consideration is given to the alignment step, and the potential downstream effects that may occur if not carried out correctly. Potential reasons for this is the lack of consensus within the field as to the effects alignment, and particularly the filtering of MSAs, has on the resulting topology.

There are quite a large number of tools for sequence alignment, which employ a number of different algorithms. Muscle (Edgar 2004) and Mafft (Katoh et al. 2005) are both matrix based alignment tools that work in an iterative manner. MSAs are created and mapped to each other to compare alignment quality and inform subsequent distance calculations, until an optimal MSA is obtained. An alternative approach is implemented in HMMER which is based on probabilistic models called hidden Markov models (HMM) (Eddy 2008). These models are used to assign likelihoods to all possible combinations of alignment in order to find the most likely MSA.

Some studies into the effect of choice of alignment tool found that using more than one and comparing the resulting MSAs has little impact on the resulting topology (Dessimoz and Gil 2010), albeit with a small dataset. However, they did find that gaps within the MSA, which are commonly removed before phylogenetic inference, may contain signal useful for tree inference (Dessimoz and Gil 2010). The step of filtering alignments based on gapped or poorly aligned sequences was also assessed in a recent analysis on the mammal phylogeny (Ranwez and Chantret 2020). Filtering the regions/sites of uncertainty within an alignment is commonly carried out automatically on large phylogenomic datasets, either by completely removing them (using Gblock (Castresana 2000), trimAl (Capella-Gutiérrez et al. 2009), or BMGE (Criscuolo and Gribaldo 2010)) or by masking them (using HmmCleaner (Di Franco et al. 2019)). However, Ranwez and Chantret (2020) found that masking produced gene trees that were closer to the species topology and improved branch length estimates, when compared to alignments that were filtered by removing gapped or ambiguous sites. Provided care is given that the sequences that are aligned are truly homologous, they suggest that in large phylogenomic analysis, where manual inspection of MSAs is not possible, filtering by masking rather than removing ambiguous sites may provide a more robust resulting phylogeny.

1.1.3.4 Methods and models of phylogenetics

The aim of phylogenetic reconstruction is to use a set of parameters to estimate how the sequences in the alignment evolved over time, while minimising the effects of homoplasy. Methods for phylogenetic reconstruction can be distance based or character based (Yang and Rannala 2012). Distance based methods, such as those applied in neighbour joining (Saitou and Nei 1987), measure the distance between all sequence pairs and cluster them based on the distance matrix, resulting in a fully resolved phylogeny. Character based methods, such as maximum parsimony (MP), maximum likelihood (ML) or Bayesian inference (BI), compare all sequences in the alignment, considering one site at a time to estimate a score for each possible tree, thus resulting in a single tree (or a single consensus tree from a distribution of trees using BI) with the highest log parsimony or probability score. An evolutionary model informs on how the sequences evolved and consists of two components: a scheme (often referred to as the model) and a tree. The scheme is made up of an exchange rate matrix, which describes the probabilities of character state change, and a composition vector, which informs on the base frequencies within the sequence data. These schemes, or models, can be applied to nucleotide or amino acid sequences, and a number of different types have been described. Early models, such as the Jukes and Cantor (Jukes and Cantor 1969) or General time reversible (Tavare 1986) models, assumed that rates of change and composition are the same across the dataset. These models, known as homogeneous models, do not account for the biological reality that the rate of change for sites varies between lineages (rate heterogeneity), and that sites within a genome or between lineages may have restrictions or preferences for given nucleotides or amino acids (compositional heterogeneity). Homogeneous models use a single rate matrix and composition vector and may incorporate among site variation using a gamma distribution to describe the data (Yang 1996). While these models are useful for certain sequences or parts of a tree, but also due to their short running time, they miss out on relevant biological information that a more complex heterogeneous model may provide. The CAT model, implemented in PhyloBayes (Lartillot and Philippe 2004), is an infinite mixture model which is widely used to account for complexity in large datasets of deeply diverging sequences. PhyloBayes can model heterogeneity across sites, with the CAT model accounting for site-specific character parameters. This means that each column, or site, in the alignment has its own profile that is estimated from the sequence data. Therefore, instead of only a single set of parameters for the alignment, as in homogeneous models, a number of different averages can be applied across different sites in the data i.e. heterogeneous modelling.

As the levels of complexity in models of sequence evolution have increased, and as the size of datasets applied to phylogenetic studies has continued to grow, it has become paramount to use the best fitting models available. It has been shown that under fitting the data with a simple model is a major issue in generating incongruent topologies between studies (Feuda et al. 2017). The major issue with homogeneous models is that they are often too simplistic to account for complexity in biological data. This is particularly evident at deeper levels of taxon sampling, such as the root of the animal tree (~800 million years ago), where rates of change are higher over larger periods of time and between different lineages. Heterogeneous models have become commonplace in animal phylogenomic studies addressing relationships at deep timescales, the most common of which is CAT plus GTR, and often this is combined with data re-coding of amino acids to account for rate and compositional heterogeneity respectively (Simion et al. 2017; Feuda et al. 2017; Philippe et al. 2019; Laumer et al. 2019).

Maximum likelihood (ML) and Bayesian inference (BI) use these models to describe the data, and are hence known as model-based approaches, unlike maximum parsimony which does not use an explicit model. ML uses a likelihood function which is the probability of observing the data provided, given some hypothesis i.e. the topology and model of sequence evolution (Felsenstein 1981). This method allows sampling across all possible states (topologies) resulting in a single most likely tree with the highest likelihood score. Additionally, further support for each split in the tree can be obtained using a technique called bootstrapping (Felsenstein 1978). BI is closely related to ML, however, it differs in that the parameters are considered to be random variables, while ML considers them as fixed constants. This method estimates a tree by combining the likelihood of the tree with a prior probability of the tree given the model. This allows the generation of a posterior distribution, rather than the fixed likelihood value from ML, from which inferences of the topology may be made. Bayesian phylogenetic inference only became viable with the application of Markov-chain Monte Carlo (MCMC) algorithms (Metropolis et al. 1953; Hastings 1970). This allowed more rapid assessment of trees over parameter space, allowing movement from one tree to another and eventually visits trees in proportion to the posterior probabilities (Yang and Rannala 2012).

The choice of applying BI or ML approaches in phylogenomic studies is very much dependent on the phylogenetic question at hand, and thus the appropriate statistic. However, it is in the interpretation of results where BI has proven to be more advanced. The posterior probability (PP) of the tree generated in BI analysis is simply the probability of the tree given the data and the model. It is thus proportional to the prior probabilities multiplied by the weight of evidence provided by the data. For large datasets, as is often the case in deep animal phylogenomic studies, the posterior will be summarized on a single tree topology, amounting to almost certainty that the resulting tree is the most likely one, given the model and the data (Lartillot 2020). In likelihood analyses, it is not possible to calculate statistical inferences such as confidence intervals, and so interpretation of support is carried out using bootstrap analysis which can be difficult to interpret as it is not clear how large the bootstrap probability needs to be in order to statistically support a given branching order (Berry and Gascuel 1996; Susko 2010). However, it is important to note that Bayesian inference methods are sensitive to mode violations, with inappropriate, simplistic models resulting in overinflating posterior probabilities (Huelsenbeck and Rannala 2004).

1.1.3.5 Supertree and supermatrix approaches for tree inference

With the set of aligned sequences and the phylogenetic framework in which to run the analysis, there are two main approaches to infer the final species tree. First is the supertree approach, where each gene alignment is analysed separately to build gene trees which are then assembled using heuristic algorithms to create a single tree capturing the variation across the gene trees (Bininda-Emonds 2004). The advantage of this approach is that the resulting super tree is a step removed from the sequence data itself making it possible to summarise resulting trees that were obtained using different characters (Delsuc et al. 2005). It is also useful for studying the underlying differences between the individual subtrees, providing greater insight into the prevalence of potential artefacts within the sequence data and trees such as incomplete lineage sorting (ILS) or horizontal gene transfer (HGT). ILS is a significant source of gene tree and species tree discordance, which can result in positively misleading results or incongruence in the final species topology. It is caused when alleles are not perfectly segregated into all lineages and as a result may not coalesce or converge when they reach the common ancestral species. Thus when we trace their history back in time, the phylogenetic signal is lost and may result in incorrect species topologies.

A second approach for phylogenetic reconstruction is the supermatrix approach (de Queiroz and Gatesy 2007). In this method the gene alignments are concatenated into a single large matrix from which the final tree is inferred. This increased size of a single sequence alignment which contains more sites can give more appropriate model estimations across the whole alignment for site-heterogeneous mixture models such as the CAT-GTR model in Phylobayes (Lartillot and Philippe 2004). This may suppress other systematic biases in the sequence data such as long branch attraction (LBA) (Lartillot et al. 2007). However, this method ignores the different evolutionary dynamics between genes and may thus be more susceptible to underlying individual gene tree patterns such as ILS. If ILS is present in sufficient amount across the supermatrix this could lead to inconsistent or misleading topologies (Kubatko and Degnan 2007).

Which of these approaches to take when inferring the species phylogeny is dependent on a number of factors such as the evolutionary question along with the data at hand, and has a long history of division in the field over which approach is best (Bryant and Hahn 2020). When applying large phylogenomic datasets to a phylogenetic question, the application of coalescent models in a supertree approach is often not tractable. One possible alternative would be to use shortcut coalescence, however, the supermatrix approach has been shown to be superior to this method at deep evolutionary timescales (Gatesy and Springer 2014). Often, a combined approach using both frameworks, which has been applied when analysing the mammal phylogeny (Song et al. 2012) and the root of Lissamphibia (Siu-Ting et al. 2019), can provide a robust phylogenetic analysis if support for a given hypothesis is achieved across multiple analyses. For deep evolutionary questions such as the root of the animal tree supermatrix approaches remain most common, however, accounting for or at least measuring the possibility of gene discordance through a coalescence approach may prove useful in understanding the patterns of gene and species evolution in early metazoans.

1.2 Introduction to Animal Diversity and the Animal Tree of Life

1.2.1 Diversity in the Animal Kingdom

Animals consist of a diverse clade of organisms, comprising over 1.5 million described species comprising roughly 33 phyla (Zhang 2013). Actual numbers of biodiversity within animals, when including undescribed species, is much higher with a range of 15.3 - 163.2 million species, similar to that of fungi and protists. Diversity in animals occurs in a range of features such as development, morphology, body plan, behaviour, as well as molecular factors such as genome size and architecture, gene content and structure, and regulatory systems. Animal diversity and inter-relationships have captured scientists' imaginations for centuries, resulting in classifications being visited and revised a number of times (Aristotle et al 1862, Linnaeus 1758, Haeckel 1874a, (Cavalier-Smith 2007). Today we have a much broader view of animal evolution in terms of species richness and genome evolution. Based on our understanding of the animal tree we can map major transitions in phenotypic diversity. However, the molecular drivers of these transitions and the full set of phylogenetic relationships within the animal tree remain unresolved.

Animals emerged ~600-800 million years ago from a unicellular organism (dos Reis et al. 2015), that was either an undifferentiated ball of cells similar to extant choanoflagellates (Cavalier-Smith 2017; Sebé-Pedrós et al. 2017; Brunet and King 2017) or from a cell capable of existing in multiple cell states, similar to that of trans-differentiation in stem cells (Sogabe et al. 2019; Pozdnyakov et al. 2017). From their unicellular ancestral origin animals have diversified to cover a vast array of niches on our planet. The range of morphologies and niches we observe across the animal clade have been brought about through time by expansion, loss, reshuffling, regulatory shifts, and re-purposing of the content of animal genomes.

The sister group to animals, choanoflagellates, are single celled organisms with flagella, which together with animals form the clade known as Choanozoa (Cavalier-Smith and Chao 2003; Lang et al. 2002; Ruiz-Trillo et al. 2008). Sister to this is Filastera and outside of this clade is Teretosporea. Together with the clade consisting of fungi these five clades form the Opisthokonta. There are 5 major branches of life within the animal tree (Figure 1.1). The deepest diverging monophyletic clades are: Porifera (or sponges), Ctenophora (also known as comb jellies), Placozoa, and Cnidaria. Placozoa represent arguably the simplest morphology in animals, containing only 6 cell types and no organised tissue, nervous system or axial polarity (Eitel et al. 2018). The Cnidaria (jellyfish, anemones, and corals) display the most diversity in morphology, with a wide variety of complex trait evolution such as the multiple emergence of eyes (Picciani et al. 2018). Ctenophores, or comb jellies, comprise around 200 described species which radiated relatively recently in the clade (Simion et al. 2015). This group of organisms display a vast array of complex and unique traits including presence of nerve, muscle, and mesenchymal cells (Dunn et al. 2015).

The placement of the earliest emerging animal group remains controversial (Whelan et al. 2015; Pisani et al. 2015; Chang et al. 2015; Whelan et al. 2017; Feuda et al. 2017; Simion et al. 2017; Laumer et al. 2019). The major competing hypotheses support either sponges as the earliest animals, "sponge-sister" hypothesis, or the morphologically more complex comb jellies, "ctenophore-sister" (see section 1.2.2.2.1 for more detail). The Bilateria emerged later, ~550 million years ago (MYA), and represent a major transition in terms of morphological complexity, body plan diversity, innovation, and adaptability.

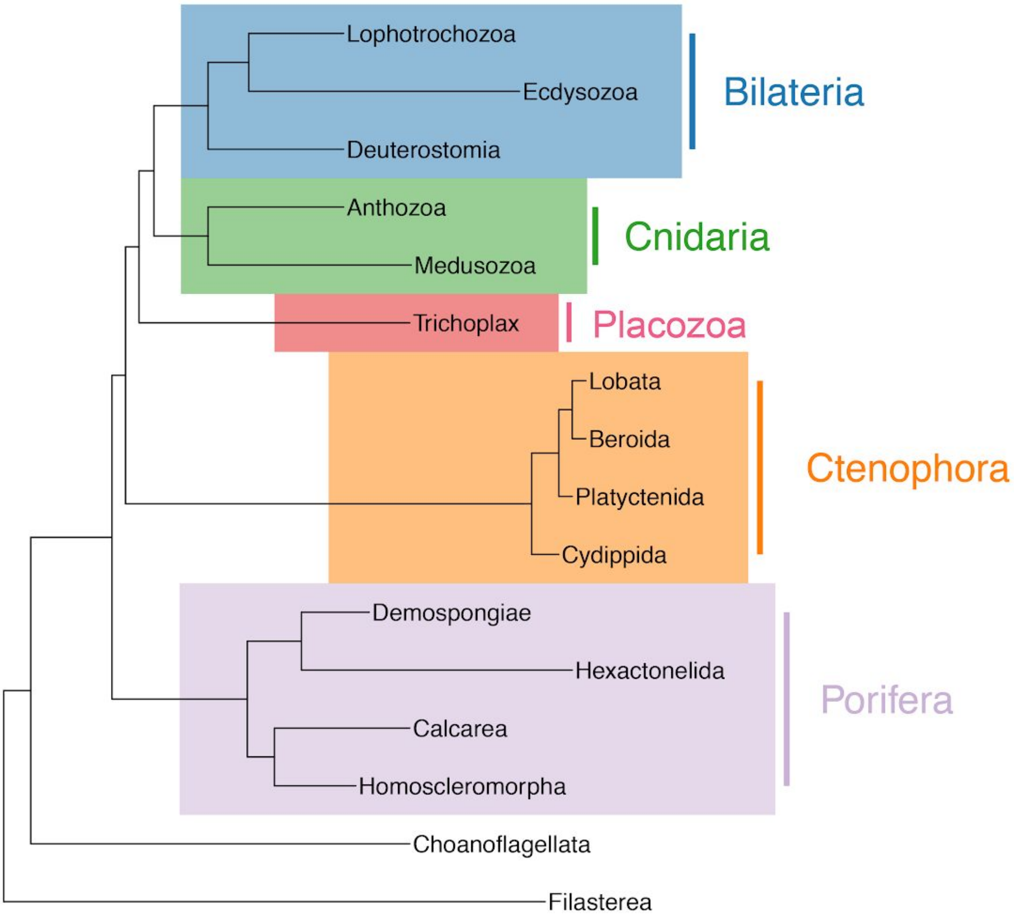


Figure 1.1. Major clades in AToL. Major clades sampled within the animal tree are colour coded, and selected outgroup lineages of closely related single cell organisms are left un-coloured for contrast.

The animal tree can be split into a number of major groups based on morphological traits alone: presence/absence of organised tissue or organ systems, body plan variation, and tissue layers (diploblastic: sponges, comb jellies, jellyfish, and coral, or triploblastic: Bilateria). Within the bilaterian clade, animals can be grouped based on the way in which these tissue layers form during gastrulation. Those in which the first opening (blastopore) becomes the anus are Deuterostomes, while those in which the blastopore becomes the mouth are Protostomes (although there are cases of deuterostomy like embryology in protostome organisms (Martín-Durán et al. 2016)). These traits also facilitate the further distinction between bilateral and non-bilateral organisms whereby most triploblasts have a through-gut with mouth and anus, while the diploblastic organisms have blind-ended guts (as found in Cnidaria and Ctenophora). Development provides further insight into the characterisation of these organisms through the types of early embryonic cleavage patterns i.e. radial cleavage, found in deuterostomes, versus spiral cleavage, most likely an ancestral state of Lophotrochozoa (or Spiralia) (Nielsen 2004; Lambert 2010; Hejnol 2010). These traits demonstrate the variety and diversity found throughout the animal tree, in terms of development and morphology. However, it also demonstrates the application of these morphological traits in understanding the evolution of animals. Indeed, many of these traits described over 100 years ago remain largely accurate in characterising and classifying animal phyla today.

1.2.2 Consensus and conflict in the Animal Tree of Life

With the advent of molecular biology and sequencing technologies, our understanding of the animal tree has reached a point of unsettled consensus in terms of the major phyla and internal relationships within the tree, and AToL is at the center of intense debate (Dohrmann and Wörheide 2013; Dunn et al. 2014; Telford et al. 2015; King and Rokas 2017; Laumer et al. 2019).

1.2.2.1 The phylogeny and evolutionary history of animals

As discussed earlier, there are 5 major lineages in AToL (Porifera, Ctenophora, Placozoa, Cnidaria, and Bilateria). The monophyly of these clades are well supported by both morphological and molecular data in most phylogenomic studies, and for a large number of the clades within these phyla we know the branching orders and phylogenetic relationships between the taxa (Figure 1.2). For example, the presence of the following morphological features: bilateral symmetry, a notochord and a dorsal nerve cord, and molting, define Bilateria, Chordata, and Ecdysozoa respectively.

Deuterostomia is one of the two major clades within Bilateria, and there is clarity in the internal relationships. Ambulacraria, a clade consisting of sister groups of Hemichordata (including acorn worms) and Echinodermata (e.g. starfish, sea urchins, sand dollars, and sea cucumbers), makes up the deuterostome phylum alongside Chordata, whose interior branching comprises Cephalochordata (lancelets) as sister group to Urochordata (sea squirts) and Vertebrata. However, the monophyly of Deuterostomia and the relationship between its two clades and the protostome clade has been questioned recently (Philippe et al. 2019). Protostomia is the sister clade to Deuterostomia and consists of Ecdysozoa and Lophotrochozoa (Figure 1.2). Ecdysozoa is comprised of Arthropoda, Nematoda (roundworms), Nematomorpha (horsehair worms), Tardigrada (water bears), Onychophora (velvet worms), and Priapulida (penis worms), amongst others. Its sister clade, Lophotrochozoa (or Spiralia), consists of an extremely diverse and species rich group of organisms whose relationships to one another have been notoriously difficult to resolve. Gnathifera is a group within Lophotrochozoa characterised by complex jaws, consisting of Gnathostomulida, Chaetognatha, and a sister group of Rotifera and Micrognathozoa (Bleidorn 2019). The monophyly of this clade, however, is often disputed owing to the unresolved placement of the enigmatic Chaetognatha (arrow worms), sometimes placed outside of Lophotrochozoa (Kocot et al. 2017; Marlétaz et al. 2019; Bleidorn 2019). The internal relationships between the remaining clades within Lophotrochozoa are still contentious.

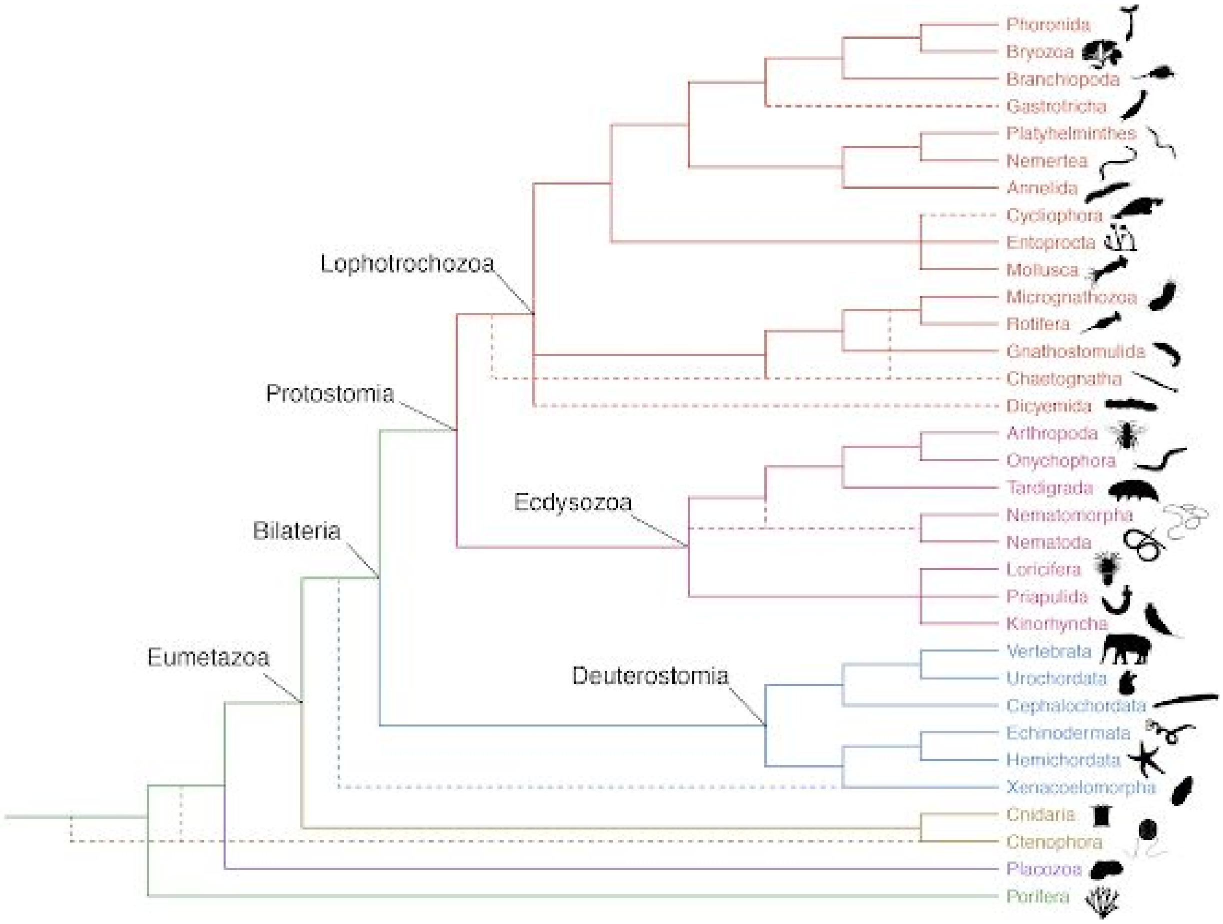


Figure 1.2. Current consensus in the animal phylogeny. Representation of our current understanding of the animal phylogeny adapted from Telford et al. (2015) (Figure used with permission of Publisher), with updated information from more recent studies. Major clade names are labelled and lineages are indicated by different colours. Dotted lines represent alternative topologies for regions in the tree which are unknown or unresolved.

1.2.2.2 Areas of major controversy in the phylogeny of AToL

1.2.2.2.1 Non-bilaterian animals & 'what came first?'

The placement of Porifera (sponges) or Ctenophora (comb jellies) as the earliest diverging animal lineage is perhaps the most intensely debated node in animal phylogenetics (Dohrmann and Wörheide 2013; Dunn et al. 2014; Telford et al. 2015; Dunn 2017; King and Rokas 2017; Whelan et al. 2015; Pisani et al. 2015; Chang et al. 2015; Whelan et al. 2017; Feuda et al. 2017; Simion et al. 2017; Pett et al. 2019; Laumer et al. 2019). The "sponge-sister" hypothesis, places Porifera as branching first with ctenophores grouping most often with Cnidaria and Bilateria. Alternatively, the “ctenophore-sister” hypothesis places ctenophores as earliest diverging with sponges branching second (Figure 1.2). Ctenophores are predatory, soft and gelatinous bodied animals that share complex morphological traits with both Cnidaria and Bilateria, including a nervous system and true muscle cells (Jager et al. 2011) (however, whether these traits and their underlying genetic components are homologous between these clades is still uncertain). Their placement at the root of AToL has considerable implications for our understanding of the evolution of animal phenotypes. Sponges represent relatively simple body plans with few cell types and their placement at the root of the animal phylogeny provides a simple explanation for the emergence of different cell types and indeed multicellularity. Ctenophores as the outgroup to all other animals suggests a more complex history and requires traits such as nervous systems and complicated systems of muscles across animals to have evolved a number of times independently, or to be lost in both sponges and placozoans (Jékely et al. 2015). The reasons for the uncertainty and conflict in these branching orders is manifold. Firstly, ctenophores represent a fast evolving clade with low taxonomic diversity. Long branches within the tree caused by fast evolving lineages cause systematic errors in phylogenetic reconstruction whereby the long branch leading to the ctenophore clade is attracted to the long branches of the outgroups, pulling the ctenophores to the root of the animal phylogeny (Philippe et al 2009). Subsequently, model misspecification, in particular the application of models that do not accommodate heterogeneity in rates of evolution across the data and the tree, leads to misrepresentation of the underlying molecular data which has major effects on the resulting species topology (Feuda et al. 2017).

1.2.2.2.2 Xenacoelomorpha - implications for the evolution of Bilateria

Xenacoelomorpha consist of marine worms for the most part, whose ultimate positioning within AToL has proven challenging to resolve. They have been placed either as the earliest diverging bilaterian clade (Ruiz-Trillo et al. 2004; Paps et al. 2009; Rouse et al. 2016; Cannon et al. 2016), or as sister to the Ambulacraria group (Bourlat et al. 2003; Bourlat et al. 2006; Philippe et al. 2011; Philippe et al. 2019) (Figure 1.2). These conflicting positions have major implications for interpreting and understanding how the Bilateria evolved, suggesting either a morphologically simplistic ancestral state for the phylum (i.e. the Nephrozoa hypothesis) or a secondary loss of complexity within the worm clade (i.e. the Xenambulacraria hypothesis). Xenacoelomorpha consists of two phyla, Xenoturbellida and Acoelomorpha (which includes two subgroups, Acoela and Nemertodermatida). Acoelomorphs were originally placed within Platyhelminthes in Protostomia due to similarities of morphological traits (Conway-Morris et al. 1985), however, the monophyly of Xenoacoelomorpha outside of Platyhelminthes is now well established. The subgroups within Xenacoelomorpha (particularly within Acoela) demonstrate a huge range of morphological variation in terms of size, development, and other complex traits such as nervous systems, digestive systems and reproductive organs (Achatz et al. 2013; Haszprunar 2016; Jondelius et al. 2019). The clade shares a number of morphological traits with Bilateria, including obvious ones such as bilateral symmetry and three germ layers with the mesoderm separating from the endomesoderm following gastrulation (Achatz et al. 2013; Hejnol 2015). However, they also share some key traits with Cnidaria, including the use of cilia for swimming or gliding (found in the planula stage of cnidarians), a similar cleavage pattern in early development found in all acoel species and a nemertodermatid species that is thought to be ancestral for Xenacoelomorpha that is similar to that of Cnidaria (Børve and Hejnol 2014), and finally the lack of any type of nephridia (an organ involved in excretion), also absent in cnidarians, distinguishes them from the rest of Bilateria which all possess this trait.

Analysis using molecular sequence data has struggled to confidently place this clade of enigmatic worms likely due to high rates of molecular evolution and heterogeneity in molecular composition observed in the Acoelomorpha clade. Confounded by undersampling of genomic data and application of unsuitable models of sequence evolution, this effect might cause the Xenacoelomorpha to pull to the root of Bilateria (Philippe et al. 2011). The resulting assumption that Xenacoelomorpha belongs nested within Deuterostomia, sister to the Ambulacraria, has some interesting evolutionary consequences. Namely, it suggests that traits which are found throughout Bilateria including the presence of a through gut, body cavity, and excretory organs were all lost in Xenacoelomorpha (Marlétaz 2019). Additionally, deuterostome traits such as endostylar tissue and gill slits would have been lost in the worms. The alternative hypothesis that Xenacoelomorpha are sister to Nephrozoa (all remaining bilaterian organisms) provides a more parsimonious view of the evolution of Bilateria, from a morphologically simple organism similar to the early stem bilaterians, to the evolution of a more complex type of organism. Similarly to the previous issues in placing the root of the animal tree, the placement of this unassuming group of worms has faced common issues in phylogenetic practices, resulting in huge obstacles in our understanding of the evolution of Bilateria.

1.2.2.2.3 Lophotrochozoa & Ecdysozoa - placements and monophyly of subclades within Protostomia

Lophotrochozoa is made up of 13 phyla (12 uncontroversially), however, the relationships between these clades are largely unresolved, and are often represented as a polytomy (Figure 1.2). Chaetognatha (arrow worms) represent an intriguing example of the uncertainty within Lophotrochozoa. With about 130 described species of this small, mostly plantonik, predator, this group of organisms have been placed within Deuterostomia based on embryonic and morphological characteristics (Marlétaz et al. 2006), but shared nervous system traits as well as more recent phylogenomic analysis using protein coding genes has confirmed them as part of Protostomia (Matus et al. 2006; Harzsch and Müller 2007; Kocot et al. 2017; Marlétaz et al. 2019). Where Chaetognatha lies in the protostome tree has been heavily debated, with implications for our understanding of the evolution of the clade. They have been placed as the sister lineage to the remaining Lophotrochozoa clade (Matus et al. 2006; Kocot et al. 2017) or within Lophotrochozoa affiliated with a clade called Gnathifera (either as the earliest branching lineage or within Gnathifera sister to a clade of Rotifera and Micrognathozoa) (Figure 1.2) (Kocot et al. 2017; Marlétaz et al. 2019; Bleidorn 2019). For the remaining clades within Lophotrochozoa, very few sister group relationships between any of them are shared across phylogenetic studies (Kocot et al. 2017; Marlétaz et al. 2019). Differences in taxon sampling, coding of characters, orthologous group selection, and model selection have all produced different topologies, resulting in uncertainty throughout the phylogeny (Bleidorn 2019).

Despite a large amount of genomic data, the certainty in the monophyly of and relationships within the constituent clades of Ecdysozoa also remain problematic. Major outstanding questions pertain to the lineages within Panarthropoda, the monophyly of Scalidophora, and the relationship between these two phyla and the remaining phylum of Nematoida and whether it is sister to Scalidophora forming the group Cycloneuralia or whether it is more closely related to Panarthropoda (Giribet and Edgecombe 2017) (Figure 1.2). The positioning of Tardigrada within Panarthropoda has yet to be resolved, some studies place them within Panarthropoda (either as sister to the remaining lineages, or sister to Arthropoda) (Pisani et al. 2013; Laumer et al. 2019) or outside of Panarthropoda as sister to the Nematoida (Hejnol et al. 2009; Borner et al. 2014; Laumer et al. 2015). The most recent study by Laumer et al. (2019) has found strong support for their placement within Panarthropoda using a Bayesian approach, however in the same study the tardigrades grouped with Nematoida using a maximum likelihood approach. Unlike the well sampled clades of Panarthropoda and Nematoida, which are most likely unresolved due to rogue taxa and fast evolving lineages, Scalidophora undoubtedly suffers from undersampling of species resulting in the ambiguity of this clade. The three lineages within this group, Loricifera, Priapulida (penis worms), and Kinorhyncha (mud dragons), have been placed in a number of positions, including as sister to Nematoida and Panarthropoda (in a study excluding Loricifera) (Borner et al. 2014), with Priapulida as sister to the remaining ecdysozoan groups (Hejnol et al. 2009), and most recently with this same branching order along with the placement of Loricifera sister to Nematoida (Laumer et al. 2019). All of these topologies have consequences for the existence of Cycloneuralia, a clade consisting of Scalidophora and Nematoida. For now the clearest consensus is perhaps monophyletic clades of Panarthropoda and Nematoida, in a polytomy with the remaining ecdysozoan lineages. The resolution of these deep splitting branches needs to be addressed with greater taxon sampling, especially of Loricifera, Priapulida, Kinorhyncha, and Nematomorpha. However, more appropriate models, that account for fast evolving lineages, should also be applied in order to reach consensus in this major group in the animal tree.

For some of the above unresolved regions of the AToL it is clear that the following features of the underlying data or methods have contributed: fast evolving or compositionally biased lineages/species, under sampling of key species, and poor model selection. However, inconsistencies and accuracy in gene family construction and lack of any appropriate model for a given part of a tree, could also be at the root of the alternative topologies.

1.2.3 Evolution of animal genomes - size, architecture, and complexity

1.2.3.1 Genome size & the C-value paradox

Within animals genome size varies by orders of magnitude from the smallest at 20Mb (20,000,000 base pairs), the nematode *Pratylenchus coffeae*, to the largest at 130Gb (130,000,000,000 base pairs), the lungfish *Protopterus aethiopicus* (Gregory et al. 2007). Genome size in animals is correlated with a number of molecular factors, such as gene retention, intron number and size, and number of mobile elements and repetitive regions, as well as other non-molecular factors, including cell size, development rate and complexity, and a negative correlation with cell division rate (Gregory 2002). The C-value is often used to represent genome size, given as the amount of DNA in a haploid genome, normally measured in picograms. We know from the C-value paradox (Gregory et al. 2007) that genome size does not predict complexity at either the genomic or organismal level (Figure 1.3A).

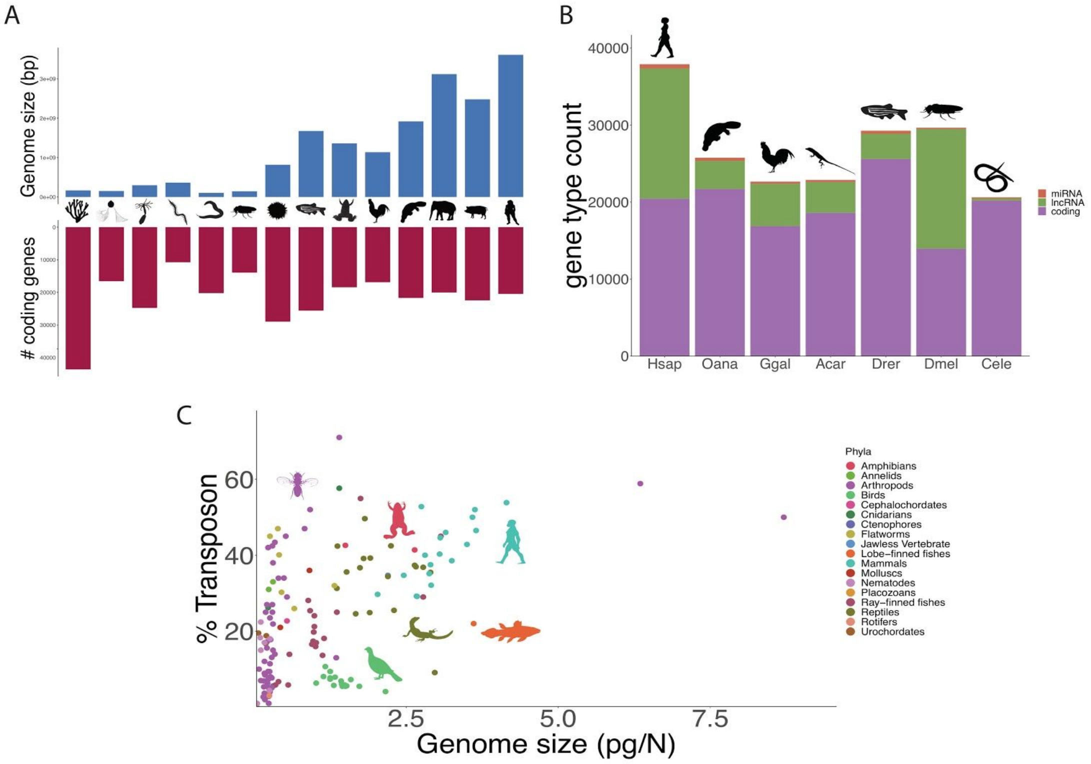


Figure 1.3. Coding and non-coding content of different animal lineages. Comparisons of : (A) genome size and protein coding gene count (B) the number of protein coding, and non-coding genes, and (C) genome size and percentage of genome that is transposable elements, across a diverse number of animals (adapted from Canapa et al. (2015)).

A major challenge in comparative genomics of the animal clade is the disproportionate sampling of sequenced genomes relative to described and undescribed species within a clade (Dunn and Ryan 2015). The variation we observe in the genomes of major animal groups or species include: gain and loss of gene families; gene duplication; whole genome duplication (WGD); emergence and modification of regulatory regions (such as microRNAs, long non-coding RNAs, and distal regulatory elements); changes in existing protein coding content (substitutions or rearrangements followed by sub- or neo-functionalization); gain of novel, or *de novo*, coding regions; gain and loss of transposable elements, and expansion in repetitive regions.

1.2.3.2 Evolution of protein coding content in animals

The genome of the last common ancestor of animals has been estimated to contain ~6,000 gene families, more than 1,000 of which were proposed to be unique to the Metazoa (Paps and Holland 2018; Richter et al. 2018). According to a recent study, the emergence of novel protein coding content varies across the animal tree, with increases of novelty found to correlate with nodes of major transition, such as the root of Metazoa, Planulozoa (Cnidaria + Bilateria), and Bilateria (Paps and Holland 2018). This suggests that protein coding content has an adaptive role at major transitions in phenotypic complexity. With predicted functions for these novel gene families in signalling, gene regulation, cell adhesion, and cell cycle these findings fit the hypothesis that the major transition events in animals, such as the emergence of multicellularity and evolution of complex body plans and new cell/tissue systems (e.g. nervous system, adaptive immunity and placentation), is correlated with an increase in novelty of gene content. Interestingly, and in contrast to the idea of animal specific novelty, comparisons of outgroup genomes such as the filasterean *Capsaspora owczarzaki* (Suga et al. 2013) and the choanoflagellate *Salpingoeca rosetta* (Fairclough et al. 2013) to animals has revealed that a large number of genes previously thought to be animal specific, or to play a role in traits unique to animals for example genes involved in signalling, cell adhesion, and development in animals, are in fact present in the pre-metazoan outgroup species (Richter et al. 2018). This suggests that the ancestral animal genome emerged through a complex pattern of old, new, rearranged, and repurposed gene content to result in the diversity within the animal tree we see today.

While a proportion of the protein coding genes that emerged at major transitions in animal evolution are conserved across all animal species, orphan/taxonomically restricted genes may help to further explain some variation in genome complexity between species. Orphan genes are generally young genes found in a single species or lineage, that share no homology to any other known gene (Tautz and Domazet-Lošo 2011). Not all orphan genes arise *de novo*, and they may emerge through processes such as duplication, rapid sequence divergence, and/or remodeling processes such as fusion/fission. Despite the lineage- or species- specificity of these genes, they are known to play a significant role in protein coding innovation (Van Oss and Carvunis 2019). The number of orphan genes can vary greatly across lineages, for example there are ~65,000 orphan genes predicted within Arthropoda (Wissler et al. 2013) whilst the estimates for other lineages such as primates, mouse and fruit fly are 270, 781, and ~200 respectively (Toll-Riera et al. 2009; Neme and Tautz 2013; Zhou et al. 2008; Chen et al. 2010).

Gene loss has also played a significant role in shaping the diversity of gene content across animal genomes. Recently, pervasive gene loss was found to occur across the animal tree in an unevenly distributed manner with lower rates of loss at deeper nodes and higher gene loss at shallower, lineage-specific nodes (Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020). These uneven rates of loss suggest that these processes are not stochastic and may provide genetic variation between lineages that could contribute to the wide diversity found across animal phyla (Fernández and Gabaldón 2020). Although gene loss can occur by random chance and be fixed by genetic drift, many examples demonstrate the adaptive role gene loss may play in the evolution of phenotypic traits (Albalat and Cañestro 2016). Indeed, these patterns of high gene loss often follow explosive and innovative periods in genome complexity (Wolf and Koonin 2013). An adaptive role for gene loss was recently found in cetaceans during the transition from land back to the ocean. This transition was accompanied by the loss of 85 genes some of which would be adaptive such as the reduced risk of thrombus formation during diving and oxidative stress–induced lung inflammation (Huelsmann et al. 2019). In humans, the non-functionalization of myosin, heavy chain 16 (MYH16) is thought to have been essential for the evolutionary origins of the species. The loss of this gene after the split from chimpanzees is thought to have occurred through adaptive selection leading to an increase in brain size and cranial capacity (Stedman et al. 2004). While these examples show how gene loss can be associated with phenotypic innovation, loss may also occur through neutral processes such as regressive evolution (Albalat and Cañestro 2016). A classic example of gene loss that has neutral effects on fitness is the recurrent loss of the l-gulonolactone oxidase (GLO) gene involved in vitamin C biosynthesis. This gene has been lost multiple times across vertebrates in response to a shift to diets rich in vitamin C (Moreau and Dabrowski 1998).

1.2.3.3 Transposable elements, repetitive regions & non-coding regions

The many structural and functional regions outside of protein coding genes also have a large impact on genome size and complexity. Transposable elements (TEs) for example, are mobile DNA fragments with the ability to rearrange the genome structure, and can introduce variation on which natural selection may act (Brandt et al. 2005; Volff 2006; Böhne et al. 2008; Oliver and Greene 2009). First discovered in maize almost 70 years ago (McClintock 1950), TEs are short nucleotide sequences with the ability to 'jump' to different regions in the genome, also capable of leaving a copy behind leading to enormously high element copy number, a pattern found particularly in eukaryotic genomes (Elliott and Gregory 2015). TEs can be divided into two major classes, based on their mode of transposition: (1) retrotransposons, which are mobilized through RNA intermediates that are reverse transcribed back into the genome, and (2) DNA transposons, that mobilized by a DNA intermediate (Bourque et al. 2018). However, there is an outstanding amount of diversity in TEs, as the two main classes may be subdivided further based on their mechanism of chromosomal integration (Bourque et al. 2018). In animals, while studies of TEs are biased largely toward vertebrate species, there are many striking examples of the role TEs play in driving innovation and their correlation with genome size. In insects, where genomes do not exceed <2 pg/N, a linear relationship has been found between genome size and percentage of TE in the genome, suggesting that increase in genome size in this group is directly correlated to expansions in transposons (Canapa et al. 2015). The proportion of the genome made up of TEs in vertebrates varies from ~5% in the pufferfish (*Tetraodon nigroviridis*) and the mallard (*Anas platyrhynchos*) to 55% and 54% in zebrafish (*Danio rerio*) and short-tailed opossum (*Monodelphis virginiana*) respectively (Canapa et al. 2015). This greater variation in the relative proportion of TE to genome size within and between vertebrates is most likely the result of WGDs and larger variability in gene loss and retention within vertebrate subphyla (Figure 1.3C). However, in general a positive correlation is found between the TEs and genome size (Chalopin et al. 2015).

Not only do TEs have the ability to rearrange whole genomes, the discovery of TE derived regulatory sequences, coding exons and other open reading frames, supports a model of ‘molecular domestication’ (Volff 2006; Piriyapongsa et al. 2007; Oliver and Greene 2009; Piacentini et al. 2014). The expansion of transposons and their subsequent co-option with coding content or regulatory elements can lead to the emergence of complex phenotypic traits (Chuong et al. 2016). For example, MER20 is a TE which contributed to the large-scale rewiring of gene regulatory networks involved in the emergence of pregnancy in placental mammals (Lynch et al. 2011). In addition, rapid proliferation of TEs within a genome have been found to correlate with species radiation events (Platt et al. 2014). It has been suggested that the rapid amplification of TEs may provide the opportunity for increased rates of variability within these TE derived functional elements which may be adaptive at times of environmental stress (Lanciano and Mirouze 2018; Horváth et al. 2017). In Drosophila, TE insertions involved in local adaptation (González et al. 2008), as well as TE insertion conferring adaptive response to temperate environments, where the frequency of TE inserts is correlated with environmental variables such as temperature and rainfall (González et al. 2010) has been shown, demonstrating the significant phenotypic and adaptive effects these set of mobile elements may confer.

Bursts in non-coding content evolution, some of which has the potential to increase levels of regulation across the genome, was necessary in large transitions in animal evolution such as the transition to multicellularity and the emergence of vertebrates (Heimberg et al. 2008). These non-coding elements, such as small non-coding RNA molecules, long non-coding RNAs, distal enhancers, and the cohesin-CTCF system, are essential to regulate cell-type specific and spatiotemporal expression, key for development and innovation of complex multicellular traits (Peter & Davidson 2015).

MicroRNAs are a class of small non-coding RNA that are abundant in most animal lineages and play a key role in development (Alberti and Cochella 2017) and the formation of cell types (Lim et al. 2005). These regulatory non-coding elements are thought to have emerged multiple times across eukaryotes, present in animals as well as in plants (Voinnet 2009), a single celled green algae (Molnár et al. 2007), a multicellular brown algae (Tarver et al. 2015), and in a social amoeba (Avesson et al. 2012). Across the animal tree there is variability in the presence and absence of miRNAs, the structure and length of pre-miRNAs, and the biogenesis pathways (Gaiti et al. 2017). For example, miRNAs and their associated biogenesis enzymes are present in all clades in the animal tree but are absent from Ctenophora and Placozoa (Gaiti et al. 2017). Interestingly, there are few to no miRNAs with shared primary sequence identity between bilaterian and non-bilaterian species. Additionally, the sponge *Amphimedon queenslandica* has miRNAs that are structurally more similar to plant miRNAs than bilaterian miRNAs (Gaiti et al. 2017). Recent research has found the presence of *bona fide* miRNAs, along with the complementary miRNA biogenesis machinery, in the unicellular holozoan clade Ichthyosporea (Bråte et al. 2018). These patterns all suggest that the miRNA complement in animals is older than previously thought and was not essential for the regulation or emergence of multicellularity (Gaiti et al. 2017). High rates of acquisition of these regulatory components in animals has been found to correlate with expansions in genome complexity and large leaps in phenotypic evolution (Heimberg et al. 2008). For example, at the base of vertebrates there were two WGD events which coincided with a major transition in the morphological complexity of animals. Heimberg et al. (2008) argue that, rather than the increase of genome complexity and protein coding content, it was the dramatic expansion of miRNA families at this point that was the driving force behind the evolution of morphological complexity in vertebrates. These patterns of miRNA innovation are also observed in invertebrate lineages. In the species-rich and morphologically diverse clade of Lepidoptera (butterflies and moths), the emergence of 11 novel miRNA families were found to have occured on the stem lineage of the clade, playing important regulatory roles in embryonic and wing patterning (Quah et al. 2015). It is clear that these regulatory elements play a significant role in the genomic and morphological diversity observed across animal species.

1.3 Molecular evolution of protein coding genes in animal genomes

1.3.1 Homology, orthology, paralogy & gene families

1.3.1.1 Homology

Homology (derived from the Greek words *homos*, same and *logos*, relation), is a difficult and fraught concept to define in molecular biology. Traditionally, in a strict tree-thinking sense, homologous genes are those that present significant similarity across the majority of their sequence length. Homology is used to represent shared traits between species that originated from a common ancestor. These homologous traits can be structural, developmental, or genetic. For example, the eye of a fish is homologous to the eye of a cow, but not to that of a squid, i.e., the trait shares common ancestry between cows and fish, but not between these two species and squid. Defining homologous clusters of genes is a crucial step in building phylogenetic trees to infer relationships between species (Felsenstein 2004), as well as for other comparative analyses. In essence homology represents the parts that make up the whole, or the modules that create the complex (Wagner 1996), and in this way we can use markers such as the eye, or a particular gene, to trace the evolutionary history of, and confer relationships between, these 'complexes' or species.

1.3.1.2 Orthology and Paralogy

Orthology is a type of homology that is used in the context of genetic material. Orthologs are homologous genes present in separate species that are derived from a speciation event only (Fitch 2000). Single gene orthologs, often seen as key markers for phylogenetic reconstruction, are orthologous genes that are present in only single copy in each species and are thus seen as unambiguous markers of species relationships (provided that there is no hidden paralogy). Paralogy refers to homologous genes derived from a single gene that was duplicated in an ancestral genome (Sonnhammer and Koonin 2002). However, there are cases where the duplication event was not ancestral but rather lineage specific, whereby after a speciation event orthologs in different species independently duplicate. Therefore, paralogs are split into two groups, 'outparalog' (cases where the ancestral ortholog duplicated before the speciation event) and 'inparalog' (where the duplication events happen independently on the lineage branches) (Sonnhammer and Koonin 2002). Inparalog, often called co-orthologs thus represent a possible source of information about the evolution of the species, and can be a useful for phylogenetic reconstruction (Sonnhammer and Östlund 2015).

1.3.1.3 Homology and homologous gene families

The inference of homology is a central component of molecular biology (including phylogenetics), providing the basis from which evolutionary history is inferred. This leaves us open to philosophical and methodological issues in our interpretation of homology and orthology (Lunter et al. 2008; Haggerty et al. 2014), and consequentially on our interpretation of phylogeny (Siu-Ting et al. 2019).

Traditional approaches in creating groups of homologous genes (or gene families) have interpreted gene and protein evolution as tree-like, even though it has long been known that sequence evolution does not always occur in a tree-like manner, e.g. processes such as horizontal gene transfer, gene fusion and fission, and domain shuffling contribute to the evolution of protein coding (and other) regions of the genome (Enright et al. 1999; Bapteste et al. 2005; Moore et al. 2008) (as discussed in more detail in Section 1.3.3). In fact, it has also long been held that these non-tree-like patterns are best represented using a network of homology (Sonnhammer and Kahn 1994; Park et al. 1997; Song et al. 2008; Bapteste et al. 2013; Haggerty et al. 2014). Nonetheless, a greater focus has been placed on defining strict gene families (often single copy orthologs) by clustering groups of homologous genes that represent full homology to one another (Enright et al. 2002; Li et al. 2003; Altenhoff et al. 2016; Emms and Kelly 2019). In doing so this carves the network of sequence similarity by cutting the edges representing partial homology between genes i.e. those found in fusion/fission genes (Tatusov et al. 1997). Additionally, analysing only strict orthologs does not avoid the ‘messiness’ of duplications, rather it ignores these issues and hides any potential duplication events that may have occurred and were subsequently lost (Dunn and Munro 2016; Siu-Ting et al. 2019). Phylogenetic analyses could undoubtedly benefit from taking a more holistic approach when constructing gene sets for phylogeny inference. Using the phylogenetic information contained within paralogous genes rather than discarding them for an assumed more informative set of single copy orthologous genes, could provide new insight into certain contentious regions within AToL. Additionally, accurately defining orthologs and paralogs will be an essential part of any phylogenomic workflow going forward to ensure that molecular datasets are enriched for signal (Guang et al. 2016). Finally, incorporating more complex patterns of gene evolution, such as gene fusion and fission, that go beyong standard boundaries of our understanding of homology could provide a novel set of phylogenetic markers.

1.3.1.3.1 Applying networks to uncover composite gene families

Graph theory is a section of mathematics which is aimed at representing pairwise relationships between objects, where a set of vertices (or nodes) are connected by a set of edges (or relations). This approach has been extremely useful in a number of biological contexts, such as studying ecological food webs (Krause et al. 2003), biodiversity measurements (Delgado-Baquerizo et al. 2020), phylogeographical patterns in viral outbreaks (Dudas et al. 2017), gene regulatory networks (Emmert-Streib et al. 2014), protein-protein interaction networks (Cafarelli et al. 2017), and sequence similarity networks (SSNs) (Song et al. 2008). SSNs are central to this thesis as they form the network base from which patterns of non-tree like gene evolution may be observed. These networks are constructed of nodes representing the genes, which are connected by edges which represent shared sequence homology (Figure 1.4A and B). Within the SSN it is possible to identify cliques, a subgraph in the larger network where each node in the subgraph shares homology to each other node. These cliques can be strict clusters of related genes, or they can show more complex patterns of homology (Figure 1.4B). The scenario of three distinct gene clusters showing homology between some but not all clusters represents the non-tree like patterns that are overlooked by standard gene family clustering approaches. From Figure 1.4B we can identify a pattern that is representative of a fusion gene cluster (Composite family, in red) is connected to two component (or parent) clusters (Parent family 1, in orange, and Parent family 2, in purple) that are otherwise unconnected. The structure of relationships represents a non-transitive pattern of similarity, where a cluster of genes connects two or more otherwise unconnected clusters. This structure can be described by alignments and phylogenetic trees (Figure 1.4C), where the parent families (orange and purple) align to the respective homologous regions in the composite gene, producing an N-rooted tree (where N is the number of roots, here N=2).

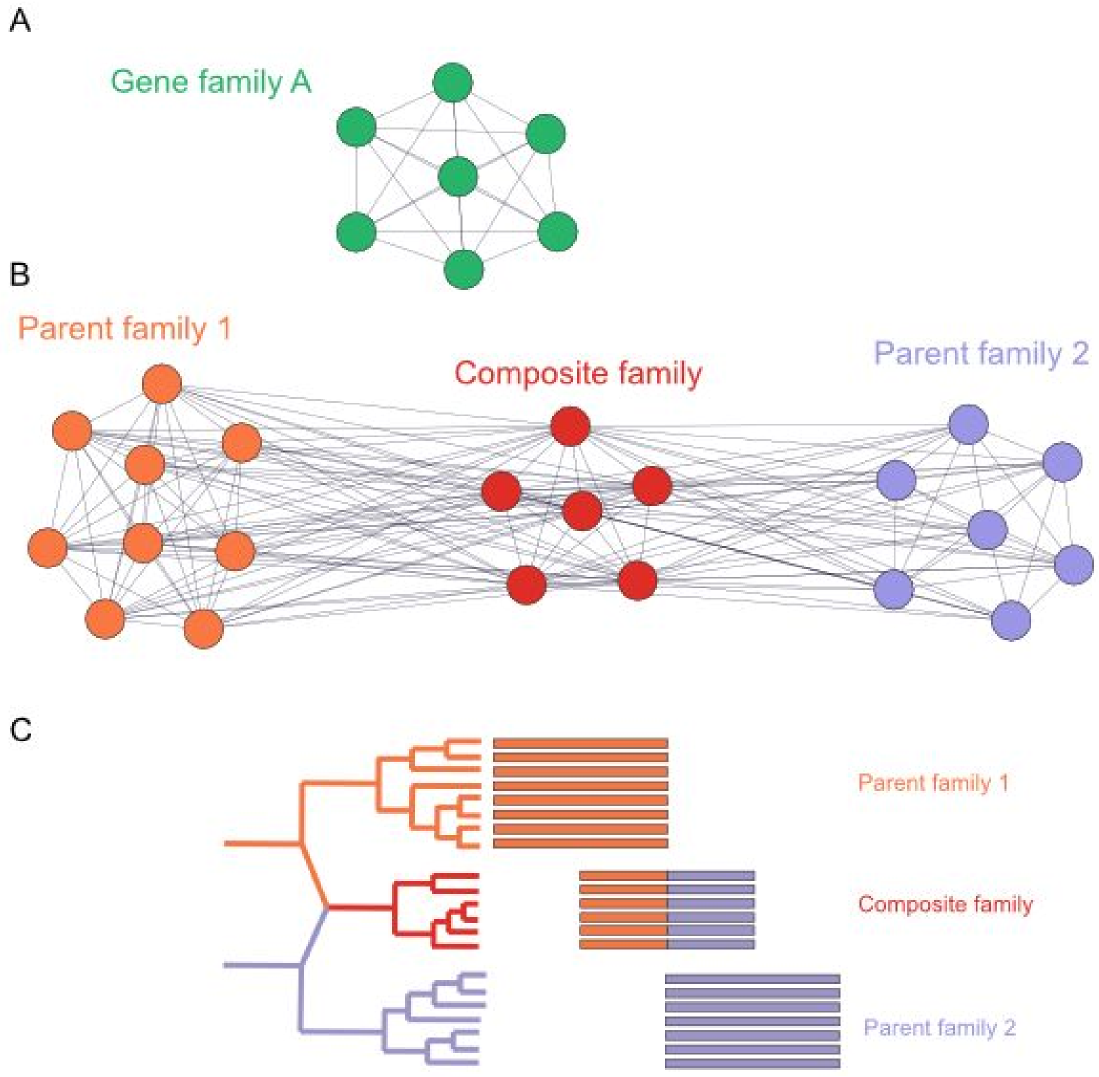


Figure 1.4. Representation of tree and non-tree like gene family evolution. (A) SSNs displaying a strict ortholog gene family (Gene family A, green) (B) a graphical depiction of a non-transitive structure representing a remodeling event with the Composite family gene cluster showing homology to both parent gene clusters, which do not share homology with each other. (C) The composite/component homology network is represented as a sequence alignment and phylogenetic tree with 2 roots.

1.3.2 Homoplasy

While homology is defined as the same trait found in different groups due to inheritance from a common ancestor, homoplasy is similarity that emerges from independent origins. Homoplasy can occur in a number of ways including convergence, parallelism and secondary loss or reversal. Convergent evolution is the process by which distantly related or unrelated species/sequences evolve the same or similar traits (Morris et al 2013), e.g., the presence of tetrapod-like hips bones in the blind cavefish *Cryptotora thamicola* (Flammang et al. 2016). Parallelism relates to the evolution of similar traits independently in closely related species/sequences. The independent *de novo* creation of antifreeze proteins in evolutionary distinct species of fish in the Antarctic and Arctic (Zhuang et al. 2019), demonstrates parallelism and its contribution to phenotypic innovation.

Convergence and parallelism are undoubtedly related processes, and their semantics have been debated; in contrast to the terms being separated, some suggest that they could both be considered under the umbrella term of “convergence” (Arendt and Reznick 2008; Scotland 2011). Often, convergent evolution is used to describe independent gains of phenotypic traits, while parallelism is used for the underlying homoplastic genetic mechanism by which the convergent phenotype evolved. The continuum between convergence and parallelism is one that provides fascinating insight into the way in which homoplasy may occur. Certain homoplastic traits may actually possess a deeper genetic mechanism (Wake et al. 2011). Traditionally, an example of convergent evolution of phenotypic traits has been the paired appendages in tetrapods and arthropods (e.g. the legs of a fly and the legs of a human). However, we now know that, while the trait did evolve independently, underlying the genetics and developmental evolution of this trait are a set of homologous gene clusters that share a common ancestor, i.e. the *Hox* genes (Shubin et al. 2009). This ancient set of homologous *Hox* genes control the initial outgrowths of the limbs from the body which then become patterned along the axis of the body (Shubin et al. 1997; Shubin et al. 2009). In this example, an underlying set of genes involved in body organisation and limb formation in all bilaterians were used in the independent evolution of this trait, suggesting that while the phenotypic trait is an example of convergent evolution, the genetic mechanisms are described in the context of parallel evolution as they use the same genetic components. Undoubtedly, these processes of homoplasy are not mutually exclusive and may occur in combination with, or independently of, each other.

The rate at which homoplasy occurs in animals, while difficult to quantify, is relatively unexplored with most studies analysing single cases of phenotypic convergence. However, as we know that homoplasy can occur at the sequence level (sometimes classified as parallelism), it is important to understand and account for the rate at which this occurs. It is well known that parallel evolution of nucleotide or amino acid sites within sequences occurs frequently (Rokas and Carroll 2008), and significantly hinders accurate phylogenetic reconstruction. These effects can be mitigated through the use of parameter rich models of sequence evolution (Yang 1994) and increased taxon sampling (Pollock et al. 2002). The rate at which homoplasy occurs in the context of gene family evolution, however, is still underexplored and underappreciated. In general, the rates of gene family evolution across animals have been measured without accounting for the possibility of parallel evolution, measuring gain along with secondary loss only (Fernández and Gabaldón 2020). A number of studies have found cases of parallel evolution of genes in animals formed through remodeling processes such as gene fusion (Lawn et al. 1997; Sayah et al. 2004; Brennan et al. 2008; Gaudry et al. 2014). Indeed, contrary to previous studies which suggest low rates of parallel emergence of multi-domain proteins (Gough 2005), one study found that between 25-70% of multi-domain proteins across eukaryotes have evolved from independent molecular events (Zmasek and Godzik 2012).

1.3.3 Mechanisms of gene and gene family evolution

1.3.3.1 Duplication and point mutation

Gene duplication is a major phenomenon driving novelty and adaptation in protein coding genes, providing a necessary source of genetic material with which natural selection can exert its effects (Conant and Wolfe 2008; Kaessmann 2010). Originally hypothesised by Susumu Ohno (Ohno 1970), evolution through duplication of genes, chromosomal regions, or whole genomes has been extensively characterised and established as central to genome and organism evolution (Lynch and Conery 2000; Van de Peer et al. 2009). Ohno's original hypothesis suggested a process whereby a duplicated region is freed up from the constraints of purifying selection allowing for a more rapid search of mutational space and potential for adaptation of a new function. This process, later denoted as neofunctionalization, is accompanied by other (likely more common) processes of gene duplication evolution known as subfunctionalization (Force et al. 1999) and pseudogenization. Subfunctionalisation describes the preservation of the ancestral function of a gene through complementary loss of subfunction between the original and duplicated genes through degenerative mutations (Lynch and Katju 2004). There are two distinct models for how subfunctionalization occurs (Conant and Wolfe 2008). The first is a neutral model whereby the duplicated and original gene gain mutations leading to the complementary loss of function, consequently preserving the multiple copies of the gene through purifying selection as together they are required for the ancestral function (Force et al. 1999). This model is one where neutral evolutionary processes are the main influence in gene and genome evolution (see Section 1.2.3). The other model describes an adaptive process whereby the ancestral gene has some conflict of function that cannot be optimised by natural selection, leading to duplication and adaptive mutations that separate the functions of the original and duplicate copies (Hittinger and Carroll 2007; Des Marais and Rausher 2008).

Neofunctionalization describes a process of adaptation through natural selection, whereby duplicated genes are free to mutate, opening the potential for the emergence of a new sequence with a novel function. An illustrative example of neofunctionalization in animals is found in leaf-eating monkeys, following the recent duplication of the pancreatic ribonuclease gene (Zhang et al. 2002). The duplicate copy of the ancestral pancreatic ribonuclease gene in the African leaf-eating monkey underwent adaptive evolution at particular sites within the protein to allow the organism to obtain nutrients from bacteria in the foregut. Incredibly, this duplication and subsequent adaptive evolution emerged independently in Asian leaf-eating monkeys, demonstrating not just the role of neofunctionalization in the generation of new genes and novel functions, but also the strength of convergent molecular evolution in the emergence of this genetic novelty.

Rates of gene duplication vary across the animal tree, with large bursts of duplications at deeper nodes in the tree and progressively lower rates towards the tips of the tree (Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020). This rate of gene duplication seems to have a converse relationship with gene loss in animal genomes, whereby gene loss is more prevalent towards these shallower nodes in the tree. Indeed this high rate of gene duplication and gene loss is thought to have played an important role in adaptive processes involved in major transitions in the animal tree, with divergence of duplicate genes and subsequent loss of alternative copies providing the necessary changes required for shifts in the genome diversity (Fernández and Gabaldón 2020; Albalat and Cañestro 2016). This alternative loss or retention of copied genes is thought to be most prevalent at certain parts in the animal tree, especially after whole genome duplication events (WGD). WGD is thought to be a response to escape the threat of extinction (Crow et al. 2006), and it subsequently leads to loss of specific genes which can result in different patterns of copy number and alternate duplicate retention, often driven by the dosage sensitivity of a gene (Makino and McLysaght 2010).

1.3.3.2 *De novo* gene genesis

De novo gene birth is the process whereby a new gene emerges from an ancestral non-genic region of the genome. This process represents a previously underestimated mechanism of new gene genesis that may produce a protein coding or regulatory RNA gene (Schmitz and Bornberg-Bauer 2017; Van Oss and Carvunis 2019). Contrary to the model of gene evolution through tinkering of previously established genes or gene regions (Ohno 1970; Jacob 1977), *de novo* gene birth represents a tree-like process of gene evolution from previously non-coding DNA. Exactly how and to what extent *de novo* gene genesis occurs across life is still relatively unknown. However, models by which the mechanism occurs have been proposed, and there are now several cases of the phenomenon, further supporting a major role of this process in the generation of genomic and phenotypic innovation (Tautz and Domazet-Lošo 2011).

Due to the difficulty in discovering ancient cases of *de novo* gene birth, most examples today are found in taxonomically-restricted genes (evolutionarily young genes which have emerged in a specific lineage or species genome). These young orphan genes are those that share no significant sequence homology to any other gene.

De novo genes can be characterised along a spectrum of definitions. *De novo* transcripts or protogenes (Carvunis et al. 2012) are intergenic transcripts that may not play a functional role, but are along the continuum between non-genic region and a fully functional *de novo* gene. In contrast a *de novo* gene is one that is functional by coding either for a protein or a functioning RNA (Heinen et al. 2009). This leads to questions about the mechanisms of how such processes occur.

In order for a *de novo* gene to emerge, a non-genic region must first obtain an open reading frame (ORF) and be transcribed. The order in which this occurs, “transcription first” or “ORF first”, is still debated as examples of both mechanisms are found across animals (Zhuang et al. 2019; Reinhardt et al. 2013). In animals, the emergence of these regions by transcription mediated processes shows tissue specificity. The transcription of these neutrally evolving parts of the genome is often facilitated by pervasive and promiscuous expression within the testes (White-Cooper and Davidson 2011). This gives credence to the “Out of Testis” hypothesis whereby the structure and levels of transcription within male gonads allows new genes to emerge stochastically at higher rates than other tissues (Levine et al. 2006; Kaessmann 2010; Wu et al. 2011; Luis Villanueva-Cañas et al. 2017). The overall rate of *de* *novo* gene birth varies greatly between lineages, and it's contribution to gene content is still uncertain due to difficulties in annotating and confirming legitimate *de novo* genes. However, in-depth analysis in a handful of examples have demonstrated the significant phenotypic impact of these events. For example, the recently described *Shj* gene that emerged *de novo* in mice is specifically expressed in females in the oviduct and is a key component of the transcriptional network at specific stages of the female reproductive cycle where knockout females were found to accelerate the time to the second litter and have increased the rates of infanticide (Xie et al. 2019).

1.3.3.3 Horizontal gene transfer

Tree-like schemes have a deep history in cladistics and evolutionary biology, but not all processes can be represented in such a linear manner. Whilst tree-like evolution describes descent from a single common ancestor, non tree-like evolution is the evolution of a component from multiple different sources, such as hybridisation between plants, introgression during the evolution of distinct hominin species, reticulate evolution of genes and genomes, and the horizontal transfer of genetic material. Horizontal gene transfer (HGT) or lateral gene transfer is a prime example of non-tree-like gene evolution, whereby genetic material is passed from one organism to another without it being a direct descendant of that organism. HGT in prokaryotes has long been known to be an important factor of their genome and pangenome evolution (Gogarten and Townsend 2005; McInerney et al. 2017). Cases of reticulate evolution by HGT in animals are more restricted, often associated with endosymbiosis or parasitism. However, recently there have been more discoveries of lateral transfer events across diverse phyla in animals, most often mediated between a prokaryotic source and the animal host. Examples of HGT into animal genomes have been described across a number of different systems in animals including in sponges (Conaco et al. 2016), ctenophores (Hernandez and Ryan 2018), cnidarians (Chapman et al. 2010; Baumgarten et al. 2015), arthropods (Kondo et al. 2002; Brinza et al. 2009; Nikoh et al. 2010; Husnik et al. 2013), rotifers (Gladyshev et al. 2008), tunicates (Sagane et al. 2010), and mammals (Mi et al. 2000). While the emergence of new genes in a lineage via HGT may be limited, it has shown to occur across a broad range of phyla to varying degrees.

1.3.3.4 Gene remodeling

Gene remodeling events constitute non-tree-like processes of gene evolution such as fusion, fission, domain shuffling, and exon shuffling. This form of reticulate evolution involves the physical shuffling of distinct genetic material resulting in novel sequence. Often associated with protein coding regions, this type of shuffling and rearrangement may also be used to combine coding regions with novel regulatory elements and UTRs (Katju et al. 2009; Zhou et al. 2008). If adaptive (fusion genes are found to be causal in some cancers (Demichelis et al. 2007; Agaram et al. 2015)), the combination of two or more independent genes to form a single transcription unit has the potential to rapidly give rise to new protein function and drive phenotypic innovation (Long et al. 2003). A classic example of new gene genesis through these recombinogenic processes is *jingwei* in *D. melanogaster* (Long and Langley 1993). This chimeric gene is derived from transposon copy of the Adh locus and a 5' end derived from another gene, *yande*, resulting in a novel phenotype whereby the protein obtains new specificity towards long-chain primary alcohols (Long and Langley 1993). Gene fusion may be DNA mediated, permanent remodeling within the genome through recombination or replication processes, or transcript mediated, where the adjacent genes are transcribed together via readthrough to produce a single chimeric transcript (Thomson et al. 2000; Parra et al. 2006). Underlying the majority of these recombinogenic processes is duplication, providing the raw material for remodeling while preserving the original content. As mentioned earlier, duplication can occur on a small scale, as with gene duplications, or larger scales, in segmental duplication and whole genome duplication. Gene duplication may facilitate remodeling in two ways (i) the parent genes undergo normal gene duplication, and the duplicate sequence as a whole is used during composite formation, (ii) alternatively, the parent gene partially duplicates, freeing up a portion of the sequence to be involved in a remodeling step with another full or partial gene sequence. Segmental duplications in hominoids have been found to provide the necessary material for the creation of composite genes which are transcribed (Bailey et al. 2002; She et al. 2004; Marques-Bonet et al. 2009) and translated (McCartney et al. 2019), some of which show signatures of positive selection, suggesting the evolution of a new function (Ciccarelli et al. 2005). Similarly, whole genome duplication events, which occurred several times throughout animal evolution, are thought to provide an influx of genetic material through which novel genes and functions may emerge. Thus, it is reasonable to believe that this may also loosen the selective constraints within the genome to facilitate higher rates of gene remodeling (Sémon and Wolfe 2007).

Transcript mediated fusions technically do not permanently alter the genome and form a novel gene, instead it involves the creation of a composite transcript following expression of two or more genes together as a single ORF. Often these genes are located adjacent to one another in the genome, with read-through providing the necessary machinery for the formation of a composite transcript with a potential novel protein product being formed (Thomson et al. 2000; Parra et al. 2006; Akiva et al. 2006; Denoeud et al. 2007; McCartney et al. 2019). However, it is possible for the fusion transcript to be reinserted into the genome via processes of retrotransposition. For example the fusion transcript *PIP5K1A* identified by Akiva et al. (2006) was later shown to be fixed within the genome of humans by retrotransposition, expressed in the testes, and underwent an intense period of positive selection suggesting acquisition of a novel function (Babushok et al. 2007).

The process of gene fission is linked with gene fusion in that the fission of a gene, which could otherwise be called partial gene duplication if the original copy is retained, can lead to the subsequent fusion of these smaller gene fragments with other genes (Kaessmann 2010). Gene fusion has been found to be more prevalent than gene fission, occurring three times more often than fission in yeast (Leonard and Richards 2012). Exon shuffling is the process by which two or more exons from different sources may be ectopically recombined or when exons are duplicated leading to a novel exon intron architecture within a gene (Long et al. 2003). Related to this, domain shuffling describes a similar process of recombination between and within proteins, where the structural and function components of the protein, called domains, may rearrange to form a new domain architecture. A significant amount of work has demonstrated the correlation between increased rates of domain rearrangements and an increase in protein and organism complexity in animals (Patthy 1996; Pawson and Nash 2003; Tordai et al. 2005; Vogel and Chothia 2006; Itoh et al. 2007; Kawashima et al. 2009; Bornberg-Bauer and Albà 2013; Thomas et al. 2020; Dohmen et al. 2020). Many examples of the role of domain rearrangements in driving functional complexity and innovation have been shown in animals, including the evolution of the extracellular matrix (Cromar et al. 2014), and the blood coagulation cascade (Patthy 1985), as well as key roles in signalling networks (Pawson 1995).

1.4 Thesis aims

(1) Investigate the contribution of gene remodeling to protein coding innovation in animals - Chapter 2.

In Chapter 2 we will explore reticulate patterns of gene evolution across the Metazoa by applying graph theory to quantify the rate and points in evolution in which it occurs. Given that current methods of homology assignment do not account for more complex patterns of gene evolution such as gene fusion and fission, we do not know how or at what rate these genes evolve. Here, we aim to describe the patterns of gene remodeling throughout animal evolution and describe their contribution to protein coding innovation in animal genomes.

(2) Assess the usefulness of composite genes as phylogenetic markers and test whether they can address contentious regions in AToL - Chapter 3.

In Chapter 3 we will assess the application of composite genes (i.e. fusion/fission genes) as novel phylogenetic markers to address contentious regions within AToL. Given the difficulty in resolving some long standing uncertainties in AToL using primary sequence data, different data types may provide a novel approach in helping to reach a consensus in the animal phylogeny. We categorize the properties of composite genes and ensure enrichment for teh characters whose evolutionary history provide the most useful signal to address the species phylogeny. In doing so we question the ability of composite genes to recapitulate known branching patterns in AToL, with the aim to address the contentious nodes that remain.

(3) Measure the impact of incorrect ortholog assignment on the placement of Xenacoelomorpha by reassessing previously published studies - Chapter 4.

Finally, in Chapter 4 we reassess previously published phylogenomic datasets by questioning the accuracy of the inference of orthology in large datasets and the effect it has on difficult to resolve topologies. Orthology assignment is a crucial but often overlooked part of phylogenomic studies, and the accuracy of current methods to define orthologous genes is seldom questioned. Using gene tree methods, we wished to uncover the prevalence of hidden parology within datasets constructed for phylogenetic reconstruction and, using the case studies of the difficult to place Xenacoelomorpha, test whether enriching for orthologous gene families would result in a different tree topology.

1.5 Result chapters contributions

Chapter 2 - Peter Mulhair designed experiments, carried out analyses, created images and wrote the manuscript. Raymon J Moran designed experiments and was involved in initial dataset construction. Mary J O’Connell conceived the initial study, designed experiments, and edited the manuscript.

Chapter 3 - Peter Mulhair designed experiments, carried out analyses, created images and wrote the manuscript. Mary J O’Connell conceived the initial study, designed experiments, and edited the manuscript.

Chapter 4 - Peter Mulhair conceived the initial study, designed experiments, carried out analyses, created images and wrote the manuscript. Mary J O’Connell conceived the initial study, designed experiments, and edited the manuscript.

Chapter 2: Gene remodeling is a major contributor to protein coding innovation in Metazoa

2.1 Introduction

Gene remodeling describes the physical rearrangement of genome content, by processes such as gene fusion and fission and exon and domain shuffling (see Section 1.3.3.4). While significant work has gone into the role of domain rearrangements in driving genomic and functional evolution, standard methods of gene family identification do not fully account for the reticulate nature of gene evolution. In this chapter we will focus on gene fusion and fission and will refer to both using the general term “gene remodeling”. Here the products of gene remodeling are referred to as “composite genes”. Composite genes represent a novel gene sequence comprising coding regions acquired by gene remodeling of “parent/component genes”. Within the animal phylogeny little is known about where, and at what rate gene remodeling occurs. Here, we apply a network approach to overcome the limits of gene family detection and uncover the set of composite genes present across 63 animal genomes from diverse phyla.

Major transitions in animal phenotype can be defined as events that radically change biological function and/or morphology (Nielsen 2008). Phenotypic transitions in the Metazoa have been well documented and include the emergence of multicellularity, the nervous system, and mineralized skeleton. However, the underlying genomic events that facilitated these changes are less well established. Recent work has made significant progress into establishing correlations between major phenotypic transitions and (i) the emergence of novel genes (Grau-Bové et al. 2017; Paps and Holland 2018; Richter et al. 2018; Fernández and Gabaldón 2020), and (ii) gene loss (Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020). However, the approaches used to define gene families in these and other studies have limited our view of gene evolution to strictly tree-like processes. As discussed in detail in Chapter 1, the definition of gene families has been delegated to software algorithms which search for local clusters of related sequences that share sequence homology within specific thresholds (Enright et al. 2002; Li et al. 2003; Altenhoff et al. 2016; Emms and Kelly 2019). While these standard approaches to gene families identification typically form subgraphs with high connectivity, composite gene families display more complex patterns within the network. A more pluralistic view of homology would allow us to take gene remodeling events into account. Using a sequence similarity network approach we could search for and quantify the set protein coding genes that shared partial homology to two or more separate parent gene families, thus allowing us to glean a more complete view of gene family evolution across Metazoa.

Poor genome assembly and gene misannotation are a concern for any comparative analysis, however, when studying composite genes this issue is even more pertinent. The impact of poor gene annotation, due to incorrect truncation of a protein coding gene or fusion of adjacent, separate genes, on misidentification of composite genes has been shown to be problematic in a number of diverse animal genomes (Nagy and Patthy 2011; Rödelsperger et al. 2019). Therefore, we attempted to reduce these effects of misannotation by testing whether a putative composite gene was expressed - thereby providing additional support that it is not an artifact of misassembly or misannotation.

Following the discovery and assessment of expression of our set of putative composite genes we sought to understand the evolution of gene remodeling throughout animal evolution. Gene gain and loss in animal genomes has been found to occur at unequal rates across the phylogeny. Additionally, the rates of gains and loss differ from each other at different parts of the tree, with gains thought to occur at higher rates at deeper nodes within the animal tree, even before the emergence of animals, and higher rates of loss closer to the tips of the tree (Fernández and Gabaldón 2020). With this in mind, we placed our composite clusters of homologous groups (CHGs) onto the animal tree in order to assess whether the rates of gain and loss are constant across the tree or, like standard gene families, they occur at bursts at particular nodes of the tree.

In this chapter we annotate and describe the evolution of new gene genesis by processes of gene remodeling. The emergence of new genes in animals has been described through many processes such as the emergence of ancestral novel gene families (Paps and Holland 2018; Dunwell et al. 2017), expansion and diversification through duplication and loss (Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020), de novo gene genesis (Toll-Riera et al. 2009), and processes of domain rearrangements (Dohmen et al. 2020; Thomas et al. 2020). However, there has not been a comprehensive study of the role of non-tree like gene evolution, such as gene remodeling by gene fusion and fission, across all phyla in AToL.

2.2 Materials and Methods

2.2.1 Dataset assembly, sequence similarity network construction and identification of putative remodeled genes

A dataset of approximately 1.2 million protein coding genes from a sample of 63 animal species representing all major clades within the animal tree was constructed (and average of 19,320 protein coding genes per species). Taxa were sampled to capture the known periods of major transition within animal evolution, and species representing all major nodes in the animal tree were included (Figure 2.1A). As the quality of data used was of particular importance in this study, given the potential for misidentification of composite genes, taxon sampling was guided by the quality of gene annotation of the available species genomes. To achieve this, we carried out two filtering steps of potential genomes. First, we searched for protein coding genes known to be present across all of Metazoa (412 in total) (Powell et al. 2012), ranking genomes as high quality (contain at least 70% of the genes) or low quality (less than 70% of conserved genes present in the genome). Next, a smaller set of 40 protein coding genes that are annotated as being present across all of life (Ciccarelli et al. 2006) were used as queries to search for their presence in the set of animal genomes. As this set of protein coding genes is more conserved, this allowed for a more strict filtering for quality of the genomes. All homology searching for the core set of metazoan and all of life protein coding genes was carried out using a reciprocal BLASTp approach (Altschul et al. 1990). Searching for the set of conserved genes in Metazoa and all of life within sampled genomes, ensured that genomes of high quality (deemed by the presence of these sets of conserved genes) were used in our analysis. See (Appendix 2.1 “metazoa63_AA.fasta”) for the final set of peptides from the resulting 63 animal genomes.

The second part of data assembly involved the construction of a time calibrated species tree. Node dates and species topology were obtained from TimeTree (Kumar et al. 2017), and contentious regions (such as the branching order at the root of the animal lineage) were resolved based on current literature on the animal phylogeny (Pisani et al. 2015; Simion et al. 2017; Feuda et al. 2017). Twelve of the species in our dataset were missing from the TimeTree database, and so to place their position and time of divergence, closely related species to these lineages were used as replacements. In most cases, sister species from the same genus were present, and a list of the closely related species used to replace them can be found in (Appendix 2.1 “timeTree_taxa_conversion.txt”). With other species, such as the case of *Ciona savignyi*, which was not present in TimeTree (Kumar et al. 2017), the divergence time between it and its sister lineage *Ciona intestinalis* was taken as 176 MYA from the literature (Delsuc et al. 2018).

The set of animal genomes to be included in our analyses were filtered based on annotation quality (see above). An all versus all BLASTp (Altschul et al. 1990) was carried out (E-value <= 1e-5, percent identity >= 30%). The statements of homology that were output from the BLASTp analysis of the approximately 1.2 million animal proteins were used to generate an SSN, using the cleanBlastp step in CompositeSearch (output from this stage is found in Appendix 2.1 “blast_allGenomes_Ray.out.cleanNetwork” SSN file). The software program CompositeSearch (Pathmanathan et al. 2018), which applies a modified Depth First Search (DFS) algorithm to annotate gene families followed by subsequent network searching to define composite genes and gene families, then takes this SSN as input and identifies composite gene clusters which are denoted by non-transitive triplet patterns in the SSN, whereby two or more previously unlinked clusters are connected (Figure 2.1B). We used an E-value cutoff of 1e-5, percent identity cutoff of 30%, and coverage threshold of 80%. This provides output files on both the gene families detected and the gene families annotated as composite. The composite gene families annotation file (blast_allGenomes_Ray_out_cleanNetwork.compositefamiliesinfo, Appendix 2.1) also provides information on the size of the composite gene families, the number of component families associated with the composite family, the size of the component gene families and the connectivity of the subgraph of composite genes within a family. The CompositeSearch software (Pathmanathan et al. 2018) also generates user friendly information on the composite gene families generated and general descriptions of the network, which are useful for interpretation of the data. Information such as the number of composite genes within a family, and the amount of overlap of the homologous regions between the distinct parent genes and the composite gene is all made readily available. As discussed previously the detection of composite genes may be prone to misidentification and false positives. Therefore, as an initial filtering step, a series of quality filters were applied to the putative composite gene families identified in the CompositeSearch analysis: firstly, singleton composite CHGs, i.e. those with only a single member in the CHG, were removed. This was to reduce the number of false positive composite genes within our dataset, as these singleton composites could not be ruled out as being annotation errors in the single species genome. In total this filtering step removed 48,640 CHGs out of the total of 77,085 CHGs. In addition, genes where the mapping of components to the composite was ambiguous or where the mapping of the components overlapped, were also removed (this removed a further 14,823), leaving a total of 13,632 remaining putative composite CHGs.

2.2.2 Determining if there is evidence for expression of unique breakpoints of composite genes using publicly available transcriptome data

To further reduce the potential for misannotation or misassembly errors contributing to the resulting dataset of remodeled genes we mined transcriptomic data to determine if the putative composite gene unique breakpoints were expressed (Figure 2.1C). A dataset of transcriptomes was assembled from RNA sequencing studies for 52/63 taxa in our dataset (RNAseq_dataTable.xlsx, Appendix 2.2), with at least one representative species from each of our major clades (i.e. non-bilaterian lineages, Ecdysozoa, Lophotrochozoa, and Deuterostomia). The remaining 11/63 taxa did not have RNA sequencing datasets available at the time of this study, these were Nomascus leucogenys, Sorex araneus, Procavia capensis, Echinops telfairi, Choloepus hoffmanni, Sarcophilus harrisii, Ficedula albicollis, Ciona savignyi, Capitella teleta, Helobdella robusta, Lottia gigantea. RNA seq reads were then mapped to all putative composite gene breakpoints using bowtie2 (Langmead and Salzberg 2012). This was carried out on an individual gene-by-gene basis for all putative composite genes in each taxa

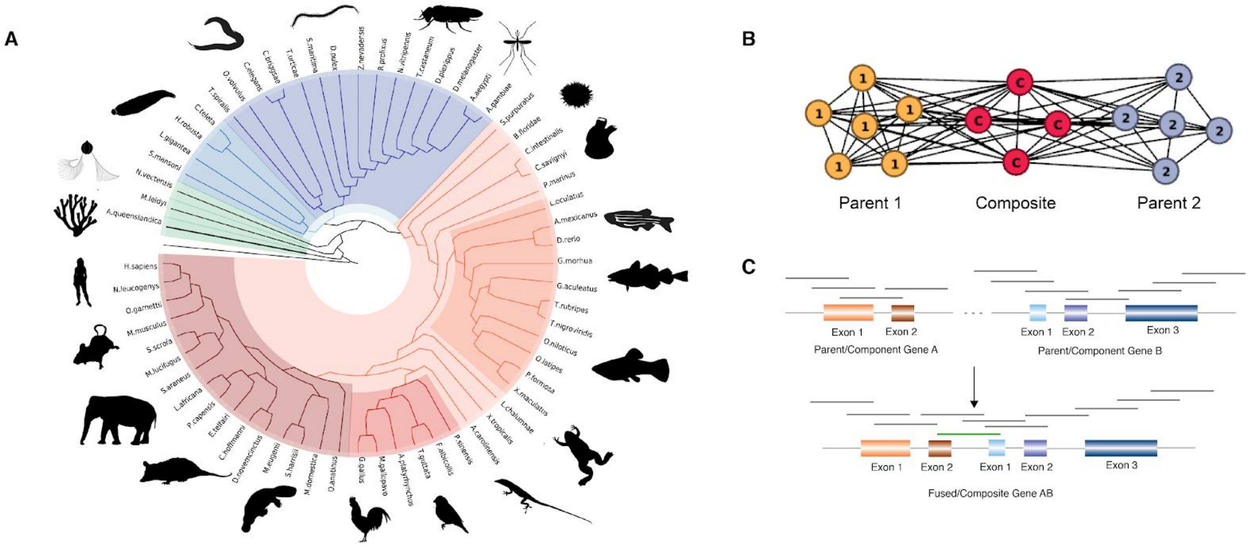


Figure 2.1. **Summary of data assembly and search for composite genes (A)** Distribution and phylogeny of species used in this study, coloured by major clades: Bilateria (pink); Lophotrochozoa (blue); Ecdysozoa (purple); Actinopterygii (orange); Aves (red); and

Mammalia (maroon). **(B)** Example of network subgraph representing non-transitive pattern of similarity, where the composite cluster (red) connects the two otherwise unconnected parent/component clusters (orange and purple). **(C)** Approach taken to validate the expression of the composite gene by mapping RNA reads to the unique composite breakpoint. Cartoon illustrates that if a read is found to cover the breakpoint, the composite gene is annotated as being expressed.

The breakpoint of the composite gene is the junction which joins the previously separated parent/component genes and is therefore unique to that composite gene. RNA seq reads mapping to this unique region provides support for the expression across this breakpoint for a given composite gene. To reduce the amount of data and run time of the RNA seq mapping, the number of composite genes which RNA seq reads were mapped to were reduced, taking representative composite genes from each CHG. The selection of a specific composite gene or genes for a given CHG was based on the phylogenetic distribution of the composite genes in the CHG and the quality of the transcriptome data for the species the genes are present in. Confirmation of expression of the representative gene(s) was extrapolated to confirm the expression and existence of the remaining composite genes in the CHG. The representative taxa for each of the Bilaterian clades included humans (Deuterostomia) and fruit fly (Protostomia), based on the amount of transcriptome data available for these species. If the best representative species was not present in a given composite family for RNAseq mapping, the next best representative species was taken and so on. Additionally, if a given CHG had two species in a single clade, the species with the more robust RNA seq data (i.e. if it had multiple tissues or timepoints sampled) were used as representative for that clade. As a result of these filters, we tested 70,662 composite genes (representing 12,048 composite CHGs), from 51 species, providing support for 7,774 composite CHG as having evidence for evidence of expression.

2.2.3 Mapping composite CHGs onto species tree

Based on current consensus in the animal tree for the 63 species sampled in this study (Pisani et al. 2015; Simion et al. 2017; Feuda et al. 2017) we reconstructed the gain/loss history of the composite CHGs, and used a constrained timetree (Kumar et al. 2017) to determine their rates of gain and loss. The distribution of gains and losses of composite CHGs across the tree was measured using a stochastic character mapping approach, implemented in RevBayes (Höhna et al. 2016). Setting the root to zero and using an Mk model with unequal transition rates, allowing a character to be lost and gained a number of times at different rates across the tree, we measured the gain and loss of each composite CHG. For each CHG we ran two mcmc chains for 5,000 generations each, allowing us to measure the precise timing of gain along each branch stochastically. This was carried out twice, first using the composite CHGs for our dataset of 63 animal genomes, and then adding outgroups, including a dataset of 63 animals genomes plus the choanoflagellate *Monosiga brevicollis* and a dataset of 63 animal genomes plus five outgroup taxa sampled deeper in the opisthokont tree, including two choanoflagellate species (*Monosiga brevicollis* and *Salpingoeca rosetta*), a filasteran (*Capsaspora owczarzaki*), an ichthyosporean (*Sphaeroforma arctica*), and a yeast (*Saccharomyces cerevisiae*). This allowed us to assess the effects of outgroup selection on the placements and numbers of gain and loss of composite CHGs across the animal tree.

2.2.4 Functional assessment of composite CHG

Each of our composite CHGs were functionally annotated using the KOG database (Galperin et al. 2015). The KOG database consists of four main functional categories which can split into specific KOG annotations representing more detailed functional descriptions. As the assignment of function in this way is defined by sequence homology, we took only the broadest descriptions of function i.e. four functional categories, for further analysis. We used RPS-BLAST (Altschul et al. 1990) (E-value 1e-05) to search our set of CHGs against the database and assigned KOG letters to each one. These were then parsed to describe the functional classifications at their broadest level for each CHG. We are aware, however, that assigning functions to composite genes in this way is problematic, as they still share sequence homology to their component genes, which may have a different function to the composite. For each of the 62 nodes in the species tree, we used the results from the phylogenetic mapping analysis (Section 2.2.3), to annotate the functional categories that emerged across the tree.

2.2.5 Role of remodeling in new gene genesis across Metazoa

Previous studies have identified novel gene families that have emerged during animal evolution (Paps and Holland 2018; Altenhoff et al. 2019), and the values vary depending on sampling and the specifics of gene family identification. We took the OMA (Altenhoff et al. 2019) dataset of Hierarchical Orthologous groups (HOGs) for our species and annotated each gene family according to their patterns of gain and loss. Placement of ‘novel’ genes using the OMA database allowed for a direct comparison of the contribution of different mechanisms to gene genesis. The OMA dataset consists of 45,612 novel genes in animals, and for each one we annotated it as being formed by remodeling or processes other than remodeling based on their presence or absence in our dataset of composite genes.

2.3 Results

2.3.1 Sequence similarity networks uncover a large set of remodeled genes that are confirmed with expression data

The analysis of the SSN from our dataset of protein coding genes revealed a significant proportion of gene clusters that are composites of other gene clusters. From our dataset of ~1.2 million genes, 322,093 were identified as composite. Once we removed singleton composites (totalling 48,640 composite CHGs) and applied quality filters describe in the materials and methods based on the structure of shared homology between composite and parent genes, we ended up with a set of 157,206 putative composite genes, corresponding to 13,632 putative composite gene families referred to throughout as clusters of homologous genes (CHGs). Only one composite CHG was present in all species across the animal tree (Figure 2.2). This composite cluster consisted of the DnaJ (Hsp40) gene, which produces a chaperone protein involved in binding to and folding of unfolded proteins (Caplan et al. 1993). The protein consists of three annotated pfam domains, the N-terminal DnaJ, the C-terminal DnaJ_C and the cysteine rich DnaJ_CXXCXGXG domain.

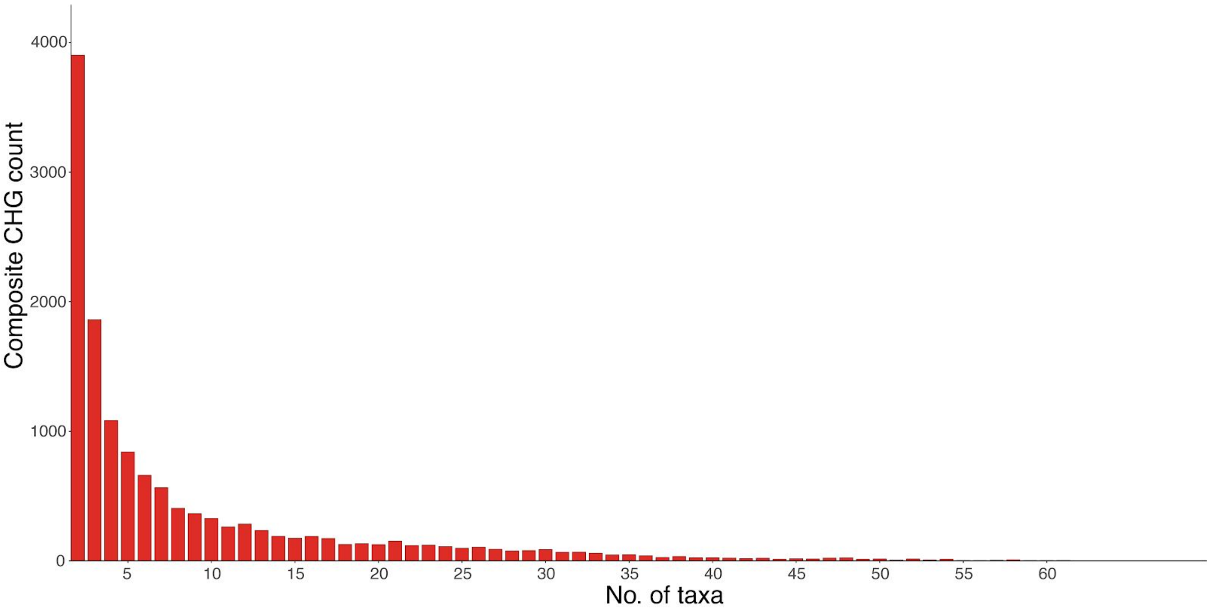


Figure 2.2. **Species counts in composite dataset.** Distribution of the number of species present in a given composite CHG, starting from two species up to the maximum of 63 species.

To further reduce the potential for misannotation or misassembly errors contributing to this signal we sought support for the expression of the unique composite breakpoints, the sequence location where the parent genes meet in the composite, from transcriptomic data. Transcriptome data was available for 12,048 (88%) of the composite CHGs, and of these 7,774 CHGs (65%) had at least one composite gene confirmed as transcribed as evidenced by at least one RNA read mapping across the breakpoint. Given that a particular composite gene may not be expressed at the time point sampled or in the tissue sampled, this proportion of "confirmed composites" is most likely a gross underestimate. Indeed, studies in zebrafish and mice have found that 26-34% (Wang et al. 2016) and 43% (Li et al. 2017) of RNA seq reads mapped to the respective genomes. Our results demonstrate that composite formation by remodeling is contributing to novel transcribed coding regions in animals.

2.3.2 Gene remodeling is prevalent across Metazoa, particularly at nodes of major phenotypic transition

Based on the current consensus for the topology of the animal tree (Pisani et al. 2015; Simion et al. 2017; Feuda et al. 2017) we reconstructed the gain/loss history of the composite CHGs for the 63 species sampled in this study, and then used a timetree (Kumar et al. 2017) to determine the patterns and rates of gain and loss. Firstly, the distribution of gene remodeling events is not linear across species sampled (Figure 2.3). For example, within Deuterostomia there were collectively 1,484 composite CHGs gained as compared to 712 composite CHGs gained in Protostomia and 34 composite CHGs in the three non-bilateria sampled. Additionally, the pattern of distribution of gene remodeling events suggests that a large proportion may be clade specific, e.g. we identify 299, 87 and 563 composite CHGs unique to Caenorhabditis, Culicidae and Euteleostomi respectively. Out of a total of 13,632 composite CHGs, 2,285 (17%) had a single node of origin meaning the vast majority were formed by multiple independent events - 55% of which were solely, independently gained in different non-sister species. The 2,285 composite CHGs of single origin are overrepresented at nodes of major transition in animal evolution (Figure 2.3), some of the largest number of single emergences of composite CHGs occur at internal nodes defining the following groups (in brackets are the raw number of composite CHGs at each node): Euteleostomi (563), Clupeocephala (92), Vertebrata (83), and Teleostei (69). Interestingly, a large number of gene remodeling events mapped to Nematoda (146) and in particular, the Caenorhabditis clade (299).

Heterogeneity in the raw number of composite CHG gains across the tree indicate bursts of remodeling events at the branches of major phenotypic evolution, particularly at Euteleostomi, Nematoda, and Caenorhabditis, all of which show a higher than average number of remodeling events (average of 41 ± 84 SD composite CHG gains observed across the whole phylogeny) (Figure 2.3). While the nodes with higher number of remodeling events are internal nodes, loss of the composite CHGs are more prevalent at shallower parts of the tree for example at Xenarthra (141 CHG losses), Dasyuromorphia (87 CHGs), Neognathae (81 CHGs). This higher number of losses is particularly evident within individual lineages, notably *Sorex araneus* (557 CHGs), *Choloepus hoffmanni* (489 CHGs), *Echinops telfairi* (464 CHGs), *Ornithorhynchus anatinus* (440 CHGs). The patterns of loss also vary across the tree, with the highest rates of loss found in lineages within the Euteleostomi. Although high numbers of remodeling events are observed in the Nematoda and Caenorhabditis clades, unlike patterns of high subsequent loss in Euteleostomi, far lower rates of loss are found.

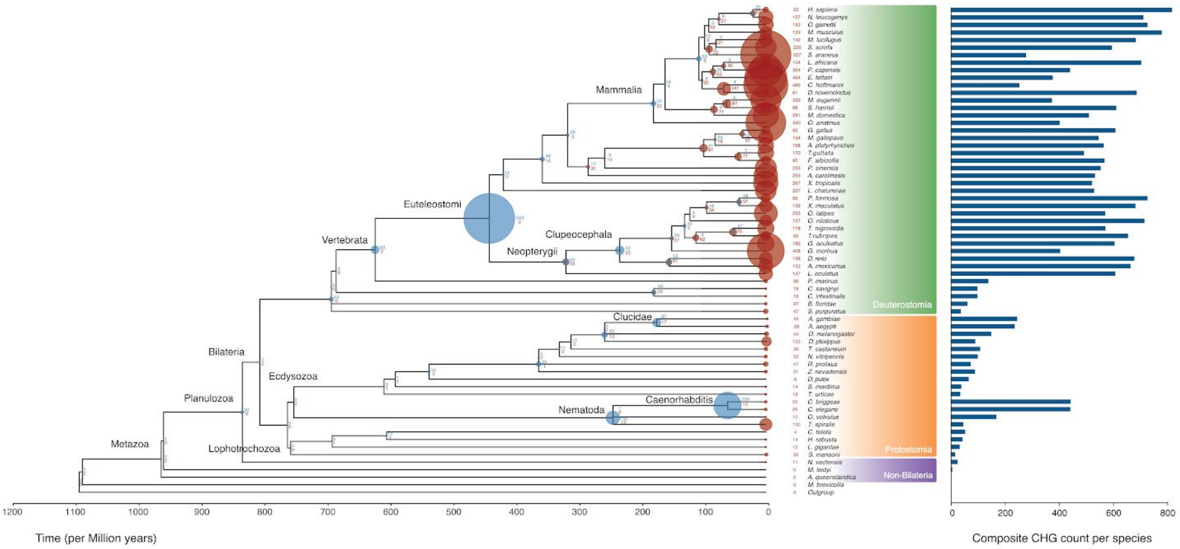


Figure 2.3. Gain and loss patterns for the 2,285 composite CHGs of single origin. The phylogeny with divergence times is shown for all species in the analysis, with certain clade names annotated. Gains are shown in blue and losses in red, and the size of the point on the node is proportional to the amount of gain/loss at that node. The histogram (right in blue) shows the total number of composite CHGs per taxon.

2.3.3 Composite CHGs gained at major transition nodes are predominantly signalling proteins

Functional annotation of the composite CHGs indicated that composite genes are more often annotated as carrying out cellular processes and signalling functions when compared to other functions such as information storage and processing or metabolism (at same level of KOG) (Figure 2.4). Others have also proposed a central role for cellular processes and signalling in major transition events throughout animal evolution (Paps and Holland 2018). Our analysis places gene remodeling as a central mechanism of new gene genesis generating bursts of novel gene families coincident with some of these major transition nodes.

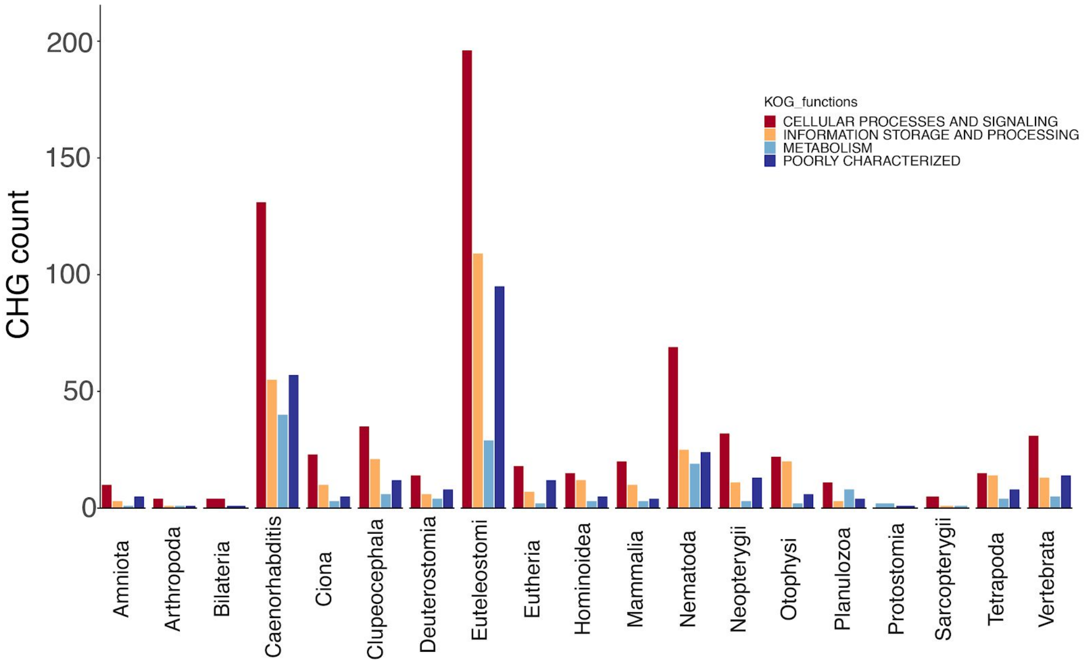


Figure 2.4. Functional categories of composite CHGs at representative major transition nodes within the animal tree. The major KOG categories are represented by cellular processes and signalling (red), information storage and procession (yellow), metabolism (light blue), and poorly characterised (dark blue). For each of the major transition clades in our species tree, the CHGs placed at emerging at those points in the tree are annotated by their KOG category functions.

2.3.4 Contribution of gene remodeling to existing dataset of "novel genes" reveals widespread contribution to protein coding novelty across Metazoa

The OMA dataset containing the 63 species in our dataset consists of 45,612 novel genes in animals (Altenhoff et al. 2015). The novel genes were annotated by parsing the Hierarchical Orthologous groups (HOGs) for our species and placing the evolutionary history (i.e. gained, duplicated, lost, or retained) of each gene on the species tree. For each node we compared the number of novel genes proposed by OMA (Altenhoff et al. 2019) to the number of novel composite genes proposed from our analysis (Figure 2.5). In general, the ratio of novel remodeled genes we identify per node to novel genes from OMA (Altenhoff et al. 2019) is similar across the tree. Certain internal nodes have higher numbers of novel genes formed by both processes, for example (clade name followed by total number of novel genes, % of novel genes that are remodeled at each node), Caenorhabditis (538, 37%), Euteleostomia

(3674, 30%), Bilateria (3313, 37%), and, Neopterygii (3026, 32%). Overall, the percentage of OMA classified novel genes that were formed through a remodeling event per node varies across the tree from 24% to 53%, with a mean of 36% (Figure 2.5). Clades where we see a greater than average contribution of remodeling to novel gene genesis include: Lophotrochozoa (49% of novel genes are remodeled); Annelida (48%), Nematoda (51%); Percomorphaceae (46%); Primates (45%), and Hominoidea (42%). Our findings provide further support for the role of new gene genesis by remodeling throughout the animal tree, ranging from 24 - 53% of all novel genes at a given node emerging through this process.

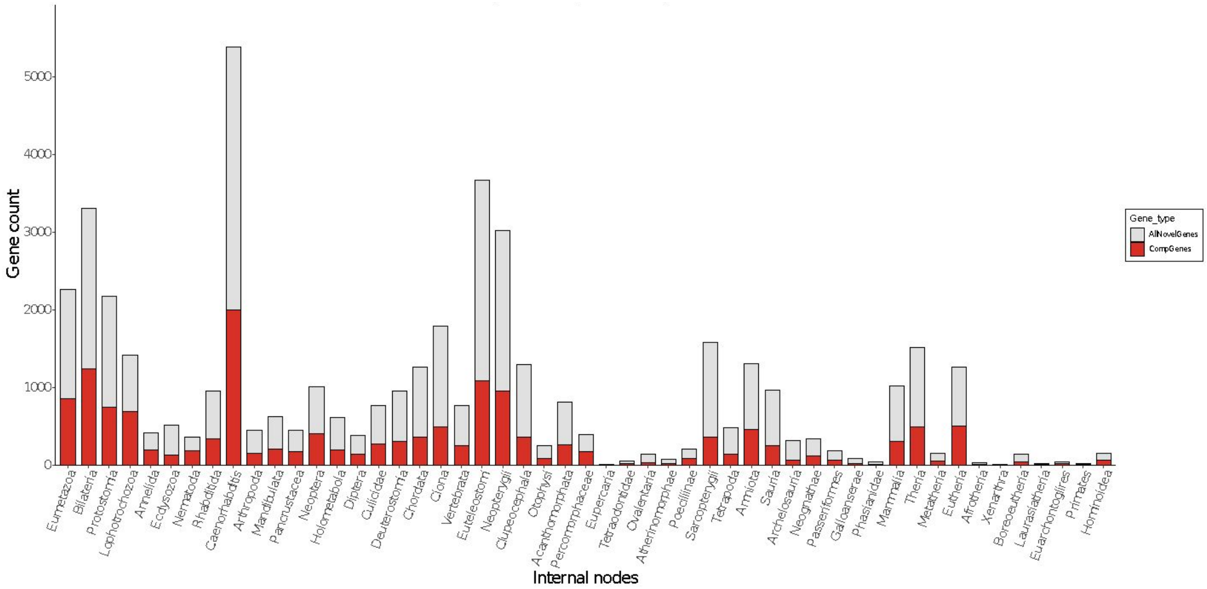


Figure 2.5. Proportion of novel genes emerging through remodeling processes and other processes. Each bar in the plot represents a node in the species tree. The bars are stacked and represent the number of 'novel' genes that emerged at that node by gene remodeling (red) or by processes other than remodeling (grey).

2.3.5 Composite genes in Metazoa tend to be larger, have more domain types, and are more likely to be formed by parents that are themselves composites

We have identified 13,632 composite CHGs and 40,217 parent CHGs (genes that contribute sequence to the composite gene). Of the 13,632 remodeled CHGs, 10,855 (80%) were formed from the most simple scenario of sequence contributions from just two parent genes. However, 11,805/13,632 (87%) composite CHGs were themselves parents (Figure 2.6A), illustrating the complex nature of gene remodeling, whereby a composite gene may itself be remodeled to produce a new composite gene and so on (Figure 2.6D). Indeed, we find that parents that are themselves composite are, on average, used in the formation of 17 composite genes as compared to an average of 10 events for parents that are not themselves the product of a remodeling event.

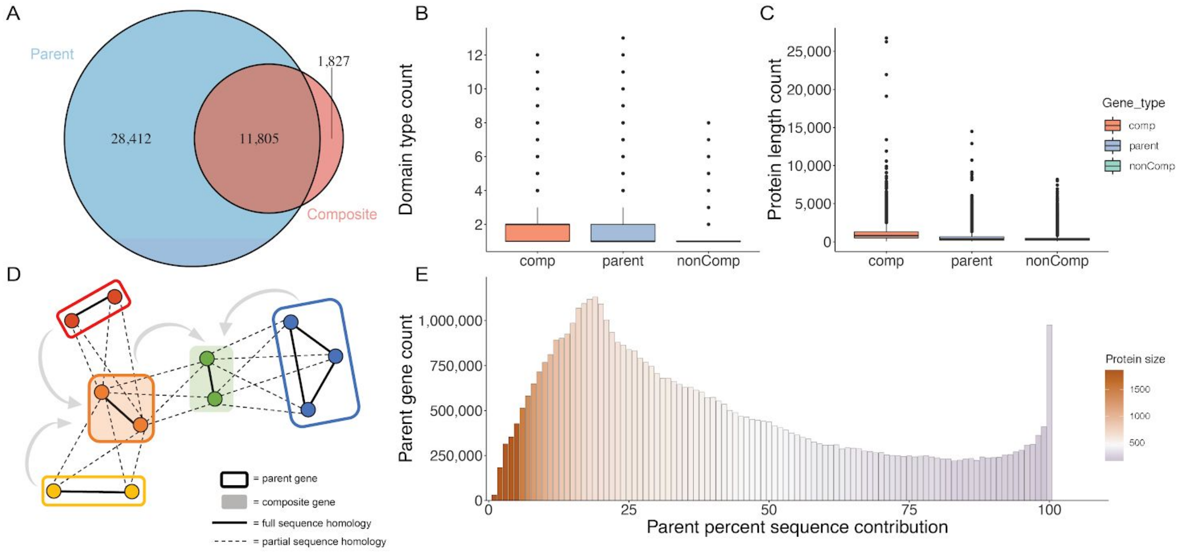


Figure 2.6. Characteristics of composite genes demonstrate the complex nature of gene remodeling and the role of domain modularity in producing complex proteins. (A) Venn diagram displaying the number of parent CHGs (blue) and composite CHGs (red) in our dataset, with the overlap representing parent CHGs which are themselves composite. **(B)** Boxplot plots showing distribution of domain type counts in the composite CHGs (red), parent CHGs (blue), and non-composite associated genes (green). **(C)** Boxplot plots showing distribution of gene lengths in the composite CHGs (red), parent CHGs (blue), and non-composite associated genes (green). **(D)** Model demonstrating the nested nature of gene remodeling, whereby a composite gene (orange) formed from distinct parent genes (red and yellow), may itself be used in a separate remodeling step in the formation of a new composite gene (green). **(E)** Distribution of the parent genes showing the proportion of the parent gene sequence that is provided to the composite gene during the remodeling process. The colour gradient indicates gene size, from the smaller genes (in the region of 0-500 amino acids) in blue to the larger genes (up to and including genes of 26,150 amino acids) in brown.

Next we categorised the ~1.2 million protein coding genes in our initial dataset based on whether they were composite genes, parent genes, or non-composite associated genes (i.e. neither composite nor parent) in order to compare protein length and domain architecture across these different categories. Each of our three datasets were annotated from the Pfam database using domain-specific hidden markov models (El-Gebali et al. 2019) (Table 2.1). We observe that composite genes have more domain types than either parent genes (P = 1.60e-148, T-test) or genes that are neither parents nor composite i.e. non-composite associated genes (P = 0, T-test) (Figure 2.6B). Additionally, remodeled genes tend to produce larger proteins than either of the parent (P = 0, T-test) or non-composite associated genes (P = 0, T-test) (Figure 2.6C).

	Composite	**All parents**	**Non-composite parents**	**Non-composite associated**
No. CHGs with annotated domains (full dataset numbers)	12,778 (13,362)	34,929 (40,215)	26,362 (28,412)	39,017 (54,632)
Total no. of domains present	27,348	61,727	38,943	48,560
No. unique domain types	3,720	4,203	3,952	6,159
Mean domain length (AAs)	125	118	116	131

Table 2.1. Domain information for composite, parent, non-composite parents, and non-composite associated genes. Rows in the table are as follows: the number of CHGs in the given dataset that have annotated pfam domains, along with the actual number of CHGs in the dataset; the total number of domains present across all CHGs in each dataset; the number unique of domains present in each dataset; the mean number of amino acids for the domains across each dataset.

Following this, we annotated the proportion of the parent gene that was used during the remodeling process to create the composite gene. We found a tendency for ~20% or 100% of the parent gene to be passed onto the composite gene, suggesting some evolutionary or structural constraint on what portion of the parent gene is used in the remodeled process (Figure 2.6E). One such constraint could be the modularity of the protein in question.

Proteins are made up of independent units called domains, which are thought to be the structural and functional units of selection for a protein. Indeed, the average length of the parent sequences used during a remodeling event is 100 amino acids, which is the average length of a domain in an animal protein. To investigate if the process of remodeling may be governed by specific domain features of a protein we analysed the differences in domain content between composite and parent proteins to see if there were any patterns of domain retention (domains consistently passed from parent gene to composite gene) during the process of remodeling across our dataset. Comparing domain types between parent and composite genes, we found no outlier domain(/s) overrepresented in all gene remodeling events. However, the domains WD40 and zf-C2H2 are present and retained at a higher rate than any other domain (WD40 is present in 391 parents and is retained in remodeled genes 58% of the time and zf-C2H2 is present in 339 parents and retained in remodeled genes 58% of the time) (Figure 2.7). These domains are known to be promiscuous (Basu et al. 2008), and are prevalent across all eukaryotes, playing key roles in a wide array of biological functions, including multi-protein complex assembly, DNA binding, and facilitating protein-protein interactions (Stirnimann et al. 2010; Fedotova et al. 2017). These types of functional domain may play an important role in facilitating new gene genesis by remodeling, by providing the novel sequence opportunities to stabilise and form new composite genes.

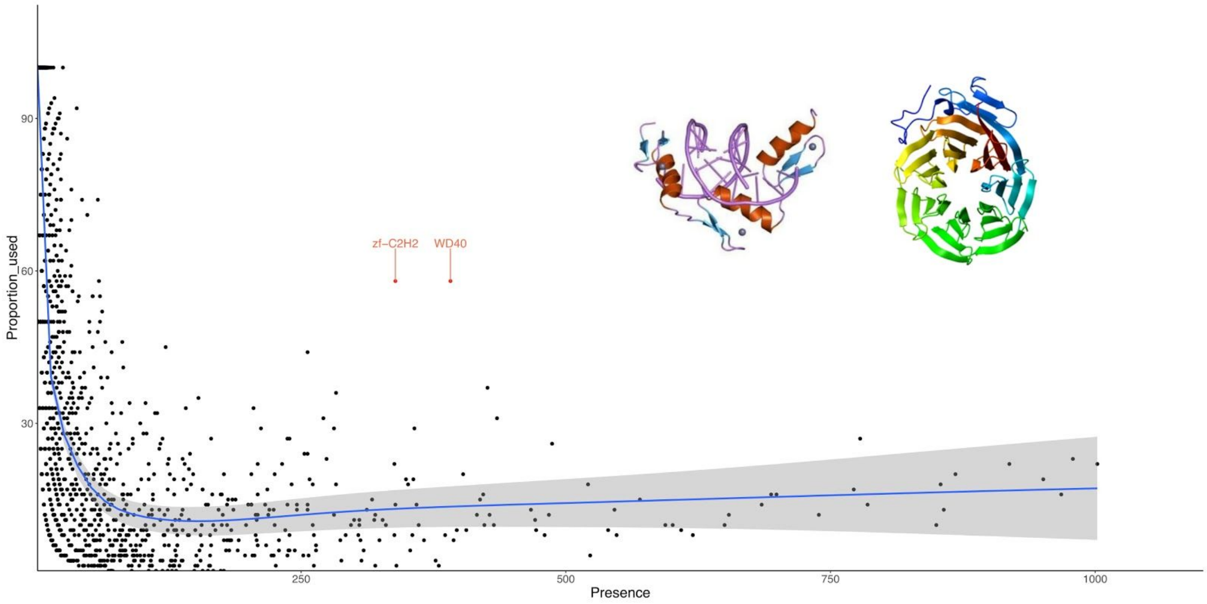


Figure 2.7. Domain usage during gene remodeling. This plot shows each domain (represented as points) in our dataset of parent CHGs. The x-axis shows the number of

times the domains are found across the full dataset of parents CHGs, and this is plotted against the proportion of times the domains are passed to a composite gene during a remodeling step (y-axis). Highlighted in red are two obvious outliers, the WD40 domain and the zf-C2H2 domain.

2.3.6 Rates of composite gene gain and loss

After mapping the number composite CHGs gained and lost at each node, we next measured the rate of gain and loss across the time calibrated tree. This allowed us to calculate the rate of gain and loss at each branch per million years, and test whether the actual rate of gain and loss of composite genes is clocklike or not. The average rate of gain across the tree is 0.27 per million years, versus a loss rate of 0.32 per million years. When we plot the rates of gain and loss for each branch onto the species tree, we find that certain branches display greater rates of change than others (Figure 2.8). The branch leading to the Caenorhabditis clade displays the highest rate of composite CHG gain, with 4.97 gains per million years. Other branches, such as the shallow branch leading to Hominoidea (~53 MYA), and the deeper branch leading to Euteleostomi (~500 MYA) display rates of composite gain above the average plus the standard deviation observed across the whole phylogeny (0.27 ± 1 composite gains per million years).

Additionally, gain and loss display different distributions in their rates across the tree. We find gains of composite CHGs distributed across deep internal and shallower nodes. Some branches with notably higher rates of gain include the branch leading to Euteleostomi, the branch leading to Caenorhabditis, and the branch leading to Hominidae (Figure 2.8A). Contrastingly, higher rates of composite CHG loss tend to be found nearer the tips of the tree, with few internal branches displaying a high rate of loss. Branches with rates of composite loss higher than the average plus standard deviation (0.32 ± 1 composite losses per million years) are found at the branches leading to Hominoidea (~53 MYA), Xenarthra (~66 MYA), Passeriformes (~67 MYA), Tetraodontidae (~52 MYA), Dasyuromorphia (~40 MYA), and Phasianidae (~42 MYA). The rates of loss also vary across the phylogeny, as we observe much higher rates of loss within the deuterostome branches (combined rate of 118.62 losses per million years across all nodes) versus the protostome branches (combined rate of 2.67 losses per million years across all nodes) (Figure 2.8B). The rates of gain in contrast, while higher in Deuterostomia, are more evenly distributed across different clades within the tree, with a gain rate of 10.37 per million years across all deuterostome nodes and a gain rate of 6.47 per million years across all protostome nodes (Figure 2.8A).

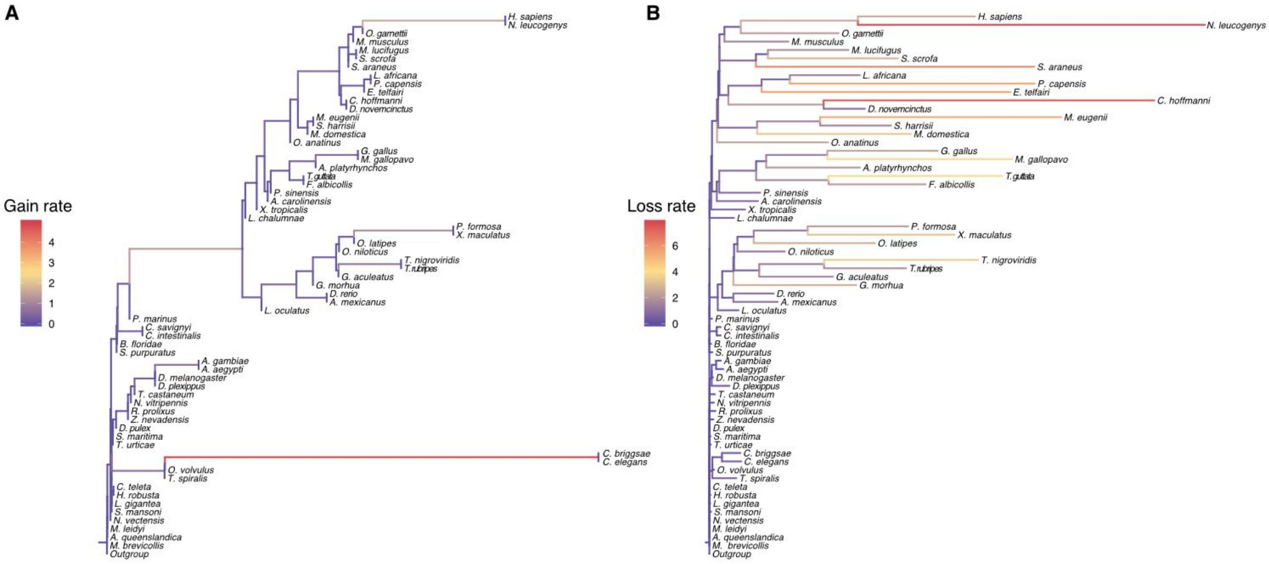


Figure 2.8. Rates of gain and loss of composite CHGs across ATol. (A) displays the rates of composite CHG gain across the phylogeny of 63 species in our dataset. Tree branch lengths represent the number of CHGs gained or lost per million years, and are coloured along a gradient based on higher (red) or lower (purple) rates of gain or loss. **(B)** Rates of composite CHG loss across the phylogeny, with branch lengths and colour gradient correlating to losses per million years.

This suggests that the rates of gain and lost are distributed unevenly across different branches on the tree. Indeed, when we plot the rate of gain and loss per node, ordered from oldest to youngest nodes, we see different rates of gain and loss across time (Figure 2.9). There are large increases in emergence of composite genes at internal nodes (Figure 2.9, blue line), particularly at Euteleostomi, Nematoda, Clupeocephala, and Culicidae. This is followed by subsequent increases in rates of gain at shallower, younger nodes such as Hominoidea and Poeciliinae where we see rates of gain in the region of 1.14-1.63/million years. A notable rise in the rate of gains is also found on the branch leading to Caenorhabditis, where there were 299 composite CHGs gained in this relatively young clade that diverged ~60 MYA. Alternatively, loss rates are greater in younger nodes (Figure 2.9, red line). It is important to note that tip branches are not present in this plot. The highest rates of loss are found at the tip branches, suggesting increased species-specific losses of composite CHGs.

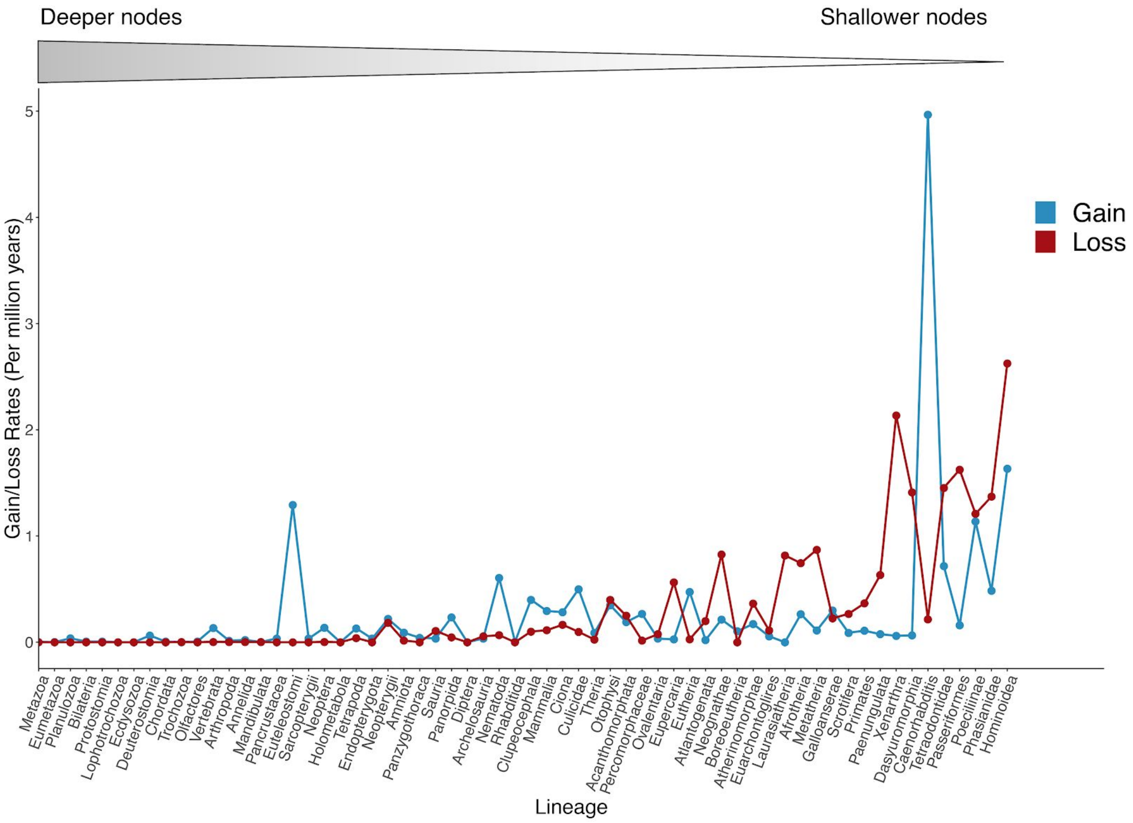


Figure 2.9. Rates of gain and loss of composite CHGs across nodes in the tree. Each node on the tree is represented on the x-axis and are ordered by their distance from the tips of the tree (older nodes on the left, younger nodes on the right). The rate of gain (blue) and loss (red) per million years is plotted for each node on the tree. Tip lineages are excluded from this plot.

The node with the largest number of emergent composite CHGs in the animal tree is Euteleostomi, with 563 CHGs of single origin (i.e. not found independently anywhere else in the tree). To assess the impact of secondary loss on this set, we annotated all branches with subsequent CHG loss and measured the rate of loss across all 563 composite CHGs which ranged from 0-60% of the overall set of losses of CHGs (Figure 2.10). The majority of losses in this clade are species-specific, with the common shrew (*Sorex araneus*), the platypus (*Ornithorhynchus anatinus*), and the two-toed sloth (*Choloepus hoffmanni*) showing the greatest number of losses. However, while the rates of loss are high in certain branches within the Euteleostomi, the average loss of CHGs across all branches in the clade is 15% indicating high level of retention of novel genes emerging by gene remodeling at this point in the tree.

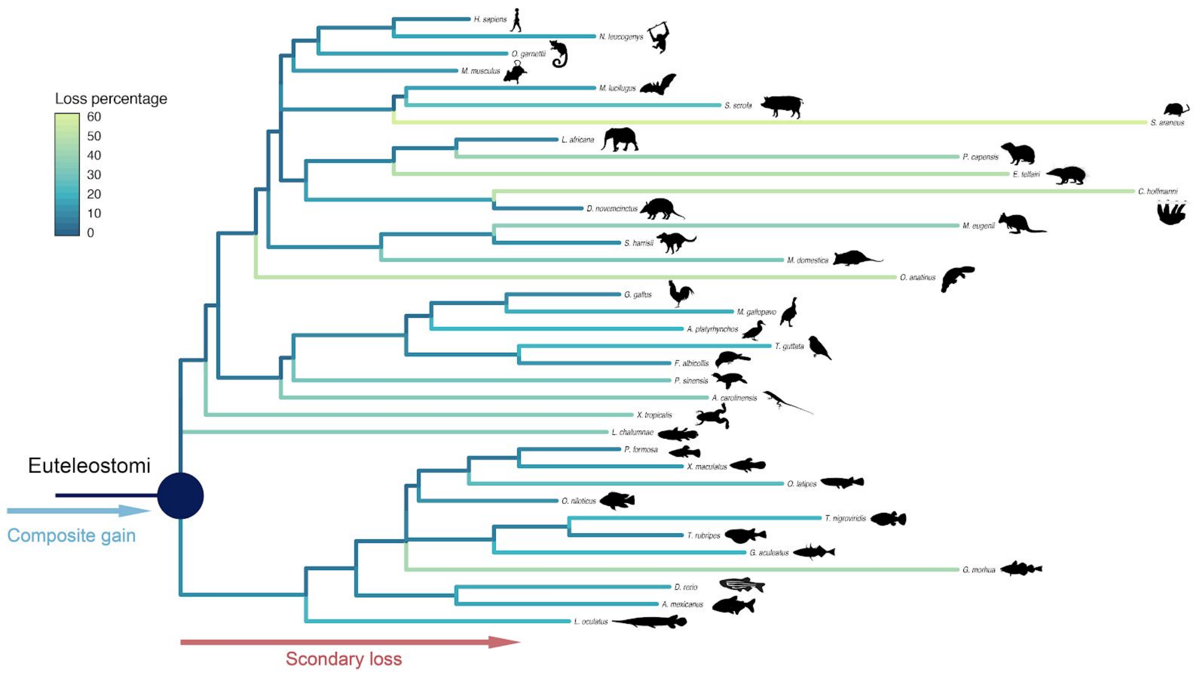


Figure 2.10. Rate of loss in specific lineages for CHGs gained on the ancestral Euteleostomi node. The tree displays the species phylogeny for the Euteleostomi species in our dataset. The large node at the base of the Euteleostomi tree shows the 563 composite CHGs gained at this node. Branch lengths are calibrated by the proportion of CHGs lost in subsequent lineages that emerged at the base of Euteleostomi, with longer branch lengths showing higher proportion of Euteleostomi CHGs lost. The colour gradient also represents the proportion of Euteleostomi CHG lost in each subsequent lineage, with low proportion of CHG loss in blue, and high loss in yellow.

Gene loss has other important implications for our understanding of the processes of gene remodeling within animals. Loss of parent genes in particular has important implications for the function of the composite gene, as loss of parents may be more likely to result in a functionally redundant composite gene, whereas composite genes with one or all of its parent genes still present in the genome may be free to rapidly adapt a new function (Rogers et al. 2009). We find that across the species studied 11,211/13,632 of our composite CHGs have lost their respective parents subsequent to composite formation, while just 2,421 composite CHGs retained both parent and composite genes, with at least one composite gene present in the same genome as the parents genes.

2.4 Discussion

Through analysis of the protein coding content of 63 animal genomes we uncovered a set of 157,206 composite genes which equate to 13,632 composite CHGs (or putative gene families). RNA seq data, specifically the mapping of RNAseq reads across unique breakpoints, provided support for the transcription of this set of genes. Mapping the evolutionary patterns and rates of gene remodeling across animals we show that gene remodeling occurs across all major groups in Metazoa at unbalanced rates throughout the tree. For example, gene remodeling is more prevalent in Deuterostomia than in either Protostomia or the non-bilaterian lineages. The largest number of remodeling events found at Euteleostomi may have had an important role in the increase of morphological complexity at this point in the tree (Flajnik 2014). The emergence of mineralised bone was a major jump in phenotypic innovation in animals, allowing for protection of internal organs, improved locomotion, and coinciding with the development of a more complex immune system. Composite genes could have provided the raw genetic material for the development of some of these phenotypic traits. One such explanation for the higher rates of gene remodeling at this point on the phylogeny could be due to increases in genome complexity within these clades (Simakov et al. 2020). Whole genome duplications are hypothesised to have occurred at least three times within the Chordata clade (3-R hypothesis) (Holland and Ocampo Daza 2018). These WGD events, twice in the branch preceding the emergence of Vertebrata, and once preceding Teleostei, align with nodes annotated with a large number of gene remodeling events. Higher than average gains at these nodes in the tree may suggest that the evolution of greater genome complexity and content loosens the selective constraints on the genome allowing a higher rate of remodeling (Sémon and Wolfe 2007). Protostomia or the non-bilaterian lineages. Additionally, these patterns of composite gene gain may also provide insight into the manner in which new gene genesis occurs in animals. A recent study found that gene gain and duplication of animal genome content was prevalent at nodes leading to the emergence of animals, and at subsequent deep nodes within the animal tree eg. the root of the animal tree, Planulozoa, and Bilateria (Fernández and Gabaldón 2020). Subsequently, we find that the highest rates of composite gene formation correlate with nodes that emerge after these deep nodes, suggesting that the emergence of genetic content through mechanisms other than remodeling may be followed by subsequent higher rates of gene remodeling.

Subsequent loss of novel composite genes appears to be common with loss of composite genes per MA (0.32 per MA) occurring at a higher rate than gains (0.27 per MA).

Additionally, most of the subsequent loss was observed at the tips of the tree in a species-specific manner. Gene loss can be an adaptive process (Albalat and Cañestro 2016) and differential gene loss has been shown to be a potential driver of genome and organismal diversity within animals (Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020). Alternatively, we know that gene fusion and gene fission are related processes, for example with fission often leading to subsequent fusion of gene fragments (Kaessmann 2010). Therefore, it may be common that a given fusion event may be reversed by the disassociation of the gene fragments by fission in a large number of subsequent lineages (Leonard and Richards 2012). A final explanation for this observed lineage specific loss of composite genes could be due to errors in gene annotation. Accurate annotation of composite genes is difficult (Nagy and Patthy 2011), so missing composite genes from a genome may not mean true absence. This has important implications for our dataset, in that the high-lineage specific loss/gain could be an artifact of poor gene annotation. Future work to assess the impact of incorrectly annotated composite genes would require more in depth analysis of gene annotation ensuring accurate ORF prediction along with evidence of expression of potential composite genes, particularly within species displaying particularly high rates of loss such as *Nomascus leucogenys* and *Choloepus hoffmanni*.

Interestingly, a study focusing on ortholog gain and loss across Metazoa by (Fernández and Gabaldón 2020) found that, using conventional gene family clustering methods, there was only a very small proportion (2%) of the overall gene content gained at Deuterostomia and Ecdysozoa. Conversely, we demonstrate that the levels of gene remodeling is comparatively high within these clades in comparison to other nodes (Figure 2.3). This may suggest that while there is a low level of gain of novel genes at these nodes by *de novo* gene genesis for example, the innovation that was required at these points of major transition may have been achieved by rearranging and shuffling of existing gene content through fusion or fission. Composite genes tend to be larger and contain more functional domains than non-remodeled genes, suggesting that this recombinogenic process may produce proteins with a wider scope of function. It is tempting to say that the higher rate of gene remodeling coincident with points of major transition implicates these processes and genes in the emergence of the associated increased morphological complexity. However, to formally test such a claim is beyond the scope of this thesis, rather we are identifying and reporting the observed patterns in the data. Functional annotation of the composite genes identified suggests these genes have a key role in cellular processing and signalling that are necessary for the emergence of complex morphological traits. However, functional annotation of composite genes from a computational perspective is fraught with challenges as functional annotation is essentially assigned based on sequence homology.

While gene remodeling results in the formation of a protein with a novel structure and function, the patterns of domain retention (Figure 2.7) and parent gene loss within composites is highly variable. For example, there were no protein domains that were retained consistently during the formation of the composite gene, that could point to a particular functional domain or combination of domains that are adaptive in gene remodeling. Additionally, we find that the loss of parent genes following the gene remodeling event was common (only 2,421/13,632 or 18% of composite CHGs had both composite genes and parent genes retained in the genome of at least one species). This level of parent gene loss implies a more prevalent role for gene remodeling in carrying out the same function of the parent genes. Nonetheless, the set of 2,421 composite CHGs with their respective parents still present in the same genome represent an interesting set of genes which may have a potential novel function.

Mapping the composite CHGs to the known species phylogeny suggested that over 80% of our CHGs emerged independently multiple times across the tree. This parallel emergence of the same composite genes multiple times across the animal tree was a surprising and interesting discovery. While there is evidence of recurring emergence of domain combinations and proteins through processes of remodeling (Zmasek and Godzik 2012), little work has been carried out to quantify the rate at which parallel evolution of composite genes occurs in Metazoa. In fact, this finding goes against the standard ideals of protein evolution and domain architecture evolution, i.e. that the parallel emergence of the same sequence and functional make-up is rare (Gough 2005). In Chapter 3 we will delve further into the patterns and rates of parallel emergence of composite genes by assessing the steps of composite gene evolution across the tree and the evolutionary traits shared between composite genes within a CHG. Importantly, these patterns, along with high rates of secondary loss suggest that these molecular characters are highly homoplastic.

Our analysis was carried out ensuring high genome and annotation quality, with a set of genomes that were available at the time of sampling. Increased taxon sampling especially in non-bilaterian species and closely related single cell outgroups species will undoubtedly provide a more robust view of evolution by gene remodeling across the deep history of animal evolution. Similarly, increased sampling of invertebrate bilaterian organisms such as those present in the following groups: Xenacoelomorpha, Lophotrochozoa, and Ecdysozoa, would provide a more comprehensive view of the processes of gene remodeling throughout the animal tree.

2.5 Conclusion

We conclude that gene remodeling is widespread across the animal phylogeny and plays an important role in the evolution of protein, and perhaps functional, complexity. Certain clades demonstrate higher rates of gene remodeling than others, correlating with points of major phenotypic transition in animal evolution, for example at the emergence of Euteleostomi. Across AToL we see a pattern of gene novelty and duplication at deep nodes followed by higher rates of remodeling in subsequent intermediate nodes, and high rates of lineage-specific secondary loss. Finally, parallel emergence of protein coding genes through remodeling, where the same composite gene is formed multiple times independently, may represent a previously underestimated and underappreciated mechanism of gene evolution in animals.

Chapter 3: Gene remodeling events as phylogenetic markers - addressing contentious nodes in the Animal Tree of Life

3.1 Introduction

Despite incredible advances in molecular phylogenetics both in terms of methodology and data availability, independent analyses of the animal tree of life (AToL) have failed to result in a single consistent species tree (Wanninger 2016; Giribet 2016). Today, a number of key nodes are still controversial within AToL and their resolution is essential for our interpretation and understanding of animal evolution. A prime example of such a problematic region is the branching patterns at the root of AToL where there are two main competing hypotheses for the sister group to the rest of Metazoa - sponges or ctenophores (Figure 3.1) (Dohrmann and Wörheide 2013; Dunn et al. 2014; Telford et al. 2015; Dunn 2017; King and Rokas 2017; Whelan et al. 2015; Pisani et al. 2015; Chang et al. 2015; Whelan et al. 2017; Feuda et al. 2017; Simion et al. 2017; Pett et al. 2019; Laumer et al. 2019). More recently, doubt has been cast on the previously well-established Deuterostomia clade, with Philippe et al. (2019) finding very low or no support for the monophyly of the clade. Indeed, this topology has been reported elsewhere (Marlétaz et al. 2019), and there is no definite synapomorphy that can unquestionably group these organisms. The conflict is likely a result of numerous issues related to data quality, phylogenetic signal, taxon sampling, and appropriate model selection (for more detail see Section 1.1). It is clear that these remaining issues will not be resolved with a single dataset or model. We propose that composite genes may have the potential to contribute a novel datatype to resolve contentious regions of AToL. In this chapter we explore the phylogenetic properties of the composite CHGs identified in Chapter 2.

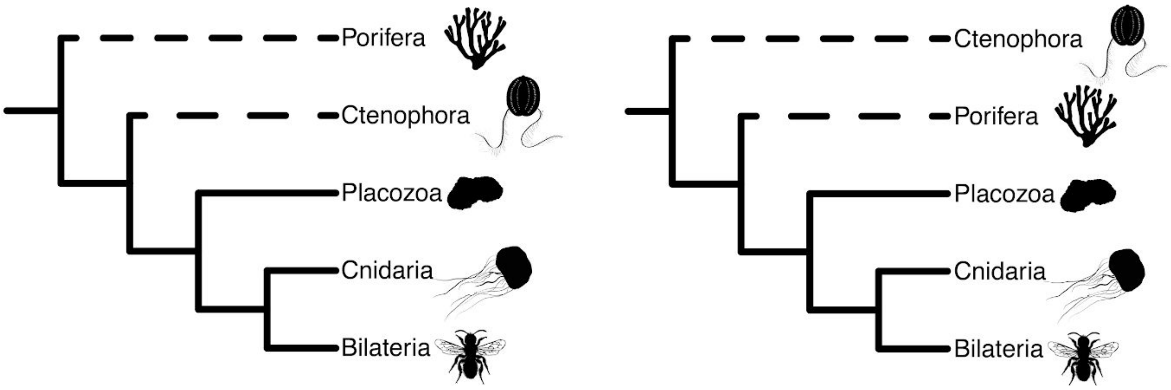


Figure 3.1. Competing hypotheses for the root of the animal tree. The "Porifera-sister hypothesis" (left) places sponges as the primary emerging animal lineage with ctenophores

as the next branching lineage - sister to the remaining animal phyla. The “Ctenophore-sister hypothesis” (right) places ctenophores at the root of the animal tree.

Composite genes may be useful markers to define individual clades and the relationships between them (Figure 3.2). Their application to phylogenetic analyses has been previously limited to single cases of fusion genes (Stechmann and Cavalier-Smith 2002), or the use of specific domain architectures (likely derived from some remodeling event) to describe a species phylogeny (Basu et al. 2008; Yang et al. 2005; Wang and Caetano-Anollés 2006),. Indeed, the split defining opisthokonts and amoebozoa/bikonts within eukaryotes has been resolved using a single derived fusion gene, where patterns of fusion presence in bikonts and the existence of only component in the other eukaryotic lineages (Stechmann and Cavalier-Smith 2002). The topology that was proposed from this study was later disproved using a consilient approach across different data types and methods (Hedges et al. 2004), thus illustrating that a single gene no matter how important is not a reliable marker for phylogeny reconstruction. Rather, a consilient approach that looks for agreement across a variety of data types and approaches is a more robust approach. This has worked for other contentious nodes such as the root of placental mammals (Tarver et al. 2016). Nevertheless, the question remains whether large-scale composite gene analyses have a contribution to make to animal phylogenetics.

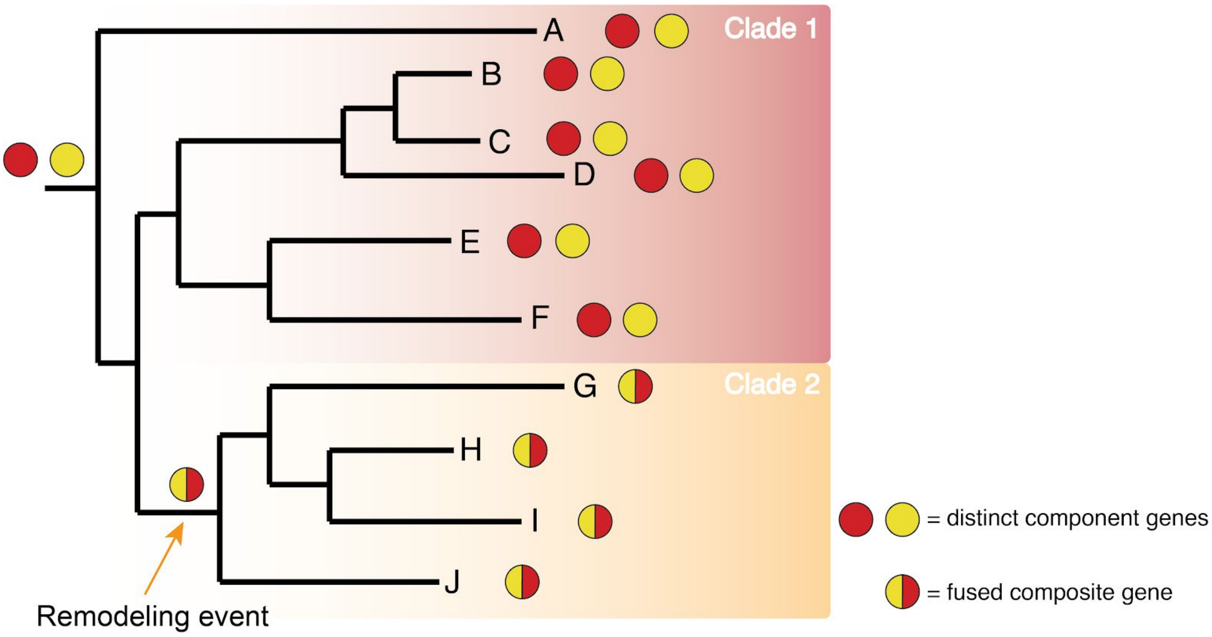


Figure 3.2. Composite genes as molecular synapomorphies. Shown is a cartoon topology containing species with two distinct parent genes (red and yellow circles) or a single composite gene (half red, half yellow circle). The clades with the parent genes only are defined by the red gradient block (Clade 1), while the composite genes cluster the species in the orange gradient block (Clade 2). The node of origin of the gene remodeling events is labeled.

Generally, rare genomic changes, such as gene remodeling events, should be infrequent enough that they show low levels of homoplasy such as secondary loss or convergent or parallel evolution. From our results in Chapter 2 we found that gene remodeling seems to be quite prevalent across the animal tree, suggesting that the events are not rare and therefore may not have this desirable property. Mapping the remodeling events also uncovered widespread patterns of secondary loss and convergent evolution of composite CHGs (83% of the composite CHGs were annotated as emerging multiple times across the animal tree). In this chapter we will examine these patterns of homoplasy in more detail so that we can extract the CHGs of single origin from those containing obvious homoplasy. Our assumption here is that composite CHGs of single origin are the most likely group to have useful phylogenetic signal and low levels of hidden homoplasy.

In our initial dataset of 63 animal genomes, we have appropriate taxon sampling to apply our composite genes to reconstruct a number of known monophyletic clades such as Bilateria, Protostomia, Ecdysozoa, Chordata, and Mammalia, and also test their usefulness at resolving contentious topologies such as the relationships at the root of the animal tree, and the monophyly of Deuterostomia. In this chapter we sought to (i) quantify the levels and distribution of measure the amount of homoplasy in composite genes within our data and (ii) assess the ability of composite genes to recapitulate known uncontroversial relationships within AToL species topologies in the animal tree. Ultimately, should these investigations show that composite genes have potential as phylogenetic markers we would apply them to resolving contentious nodes in AToL.

3.2 Materials and Methods

3.2.1 Assessing homoplasy within the data

3.2.1.1 Tree based test for homoplasy

A consistency index (CI) test in TNT (Goloboff et al. 2008) was run on the dataset of 13,632 composite CHGs. Using the *CharStats_csv.run* script [Appendix 3.1] to run the test, a presence/absence matrix, containing the either a 1 or 0 for each composite CHG determined by its presence in each species, was used to map each CHG to the species phylogeny. This test informs on the phylogenetic utility of a given character by measuring the amount of homoplasy it contains. To achieve this, the patterns of presence and absence of a set of characters are mapped to a tree and measured for their fit to the tree. CI is calculated as *m/s*, where *m* is the smallest number of changes required to map the character to the tree and *s* is the amount of change required parsimoniously for the observed tree (Kluge and Farris 1969; Farris 1989; Klingenberg and Gidaszewski 2010). This test outputs positive values less than or equal to 1.0, where higher values indicate low levels of homoplasy and low numbers indicating high levels of homoplasy.

This analysis was run on our full set of 13,632 composite CHGs, but also on a set of previously published datasets used for phylogenetic reconstruction to compare between different data types. These datasets included a set of 1,143 microRNAs from 36 species (Tarver et al. 2018), and a set of 35,244 homologous gene families from 36 species (Pett et al. 2019), both of which were originally assembled to address the root of the animal tree, and a third dataset consisting of the full set of hierarchical orthologous groups (HOGs) created using OMA, i.e. 51,717 gene families (Altenhoff et al. 2019). The dataset of homologous gene families from (Pett et al. 2019), were created using the OrthoMCL program (Li et al. 2003). While the OMA dataset was not originally constructed for phylogenetic reconstruction purposes, it is a useful comparator as it contains the protein coding sequence data from the same set of 63 species as we have used in our analyses to identify composite CHGs, however the analyses differ significantly in their approach to generating gene families.

3.2.1.2 Molecular assessment of homoplasy in composite CHGs

We sought support from two complementary approaches to identify and isolate those composite CHGs that are most likely to be of single origin from those with more complex patterns of evolution. The first approach was based on assessing the level of conservation of breakpoints and domains across composite CHGs, and the second approach was based on assessing the monophyly of composite genes within a gene tree consisting of homologous parent-composite sequences.

In the assessment of levels of conservation: firstly, we assessed the level of conservation at breakpoint positions between all composite genes in a given CHG (annotated in comp_ambig_breakpoints.csv Appendix 3.2), and secondly we measured the level of conservation of annotated Pfam domains either side of each breakpoint (Figure 3.3B). Here, the composite gene breakpoint is defined as the unique location within the composite sequence where the parent genes meet. For this analysis, composite CHGs where each composite gene within the CHG had the same number of component/parent genes (and thus the same number of breakpoints) were analysed (10,855/13.632). It is assumed that if composite genes in a given CHG have a different number of parent genes, then this is an indication that the composite CHG is a result of multiple events. Breakpoint coordinates within the composite genes were annotated from the BLASTp output files, and as such the location of the breakpoint may not be accurate across all genes in a CHG. To overcome this issue and in order to assess the level of conservation of a given breakpoint, the composite genes in a CHG were split into four quartiles based on the length of the gene and the breakpoint was considered conserved if it was present in the same quartile of each composite gene. To infer the conservation of flanking domains either side of the breakpoint, the Pfam domains for each composite gene were annotated with the Pfam database using domain-specific hidden markov models (using "*pfam_scan.pl*" Appendix 2.6, and parsing using PfamScanner with E-value threshold of 1e-3) and these were subsequently compared to check if they were conserved across the breakpoint for each composite gene (details on domain content of composite genes can be found in Table 2.1 in section 2.3.5). Both of these conservation checks were carried out using the script "*bp_DA_conserve.py*" (Appendix 3.2). From this, composite CHGs were grouped into (i) those that had the same breakpoint locations and domain architecture (DA), (ii) different breakpoint locations and DA, and (iii), a mix of (i) or (ii). We consider conserved breakpoints and conserved neighboring Pfam domains as support for a single origin of a composite CHG.

In a complementary phylogenetic approach, we split all composite CHGs into their constitutive composite-parent homologous regions. For example, if a composite gene was a result of the remodeling of two parent genes, the composite was split into its component parts and aligned (using Mafft (Katoh et al. 2005)) with the homologous region of the parent gene. This was carried out using the (Appendix 3.2 "*domain_extract.py*") script, which parses the original CompositeSearch output files to create fasta files of homologous parent-composite regions for each composite CHG. Following alignment of the homologous regions, gapped regions were trimmed using trimal (using the -gappyout function) (Capella-Gutiérrez et al. 2009), we generated the corresponding gene trees in IQTree (Nguyen et al. 2015) using automatic model selection and running 100 bootstrap replicates. Before tree inference, however, some alignments had to be discarded due to length of the sequence being too short for tree inference, the presence of too many gaps or ambiguous characters within the alignment, or two few genes present within the alignment. This resulted in a dataset of 10,849 composite CHGs from which we could create composite-parent gene trees. Support for a single point of origin for a composite CHG was taken when composite genes in a given CHG form a monophyletic clade within the parent-composite gene tree (Figure 3.3C).

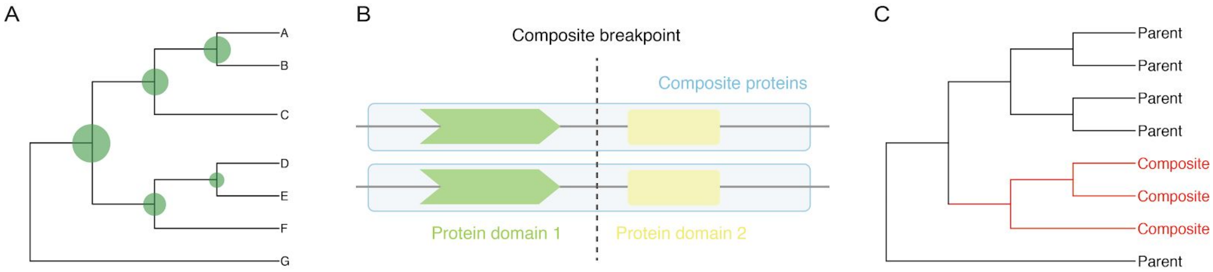


Figure 3.3. Methods of assessing single origin remodeling events. (A) Tree represents a species tree onto which our composite CHGs were placed using character mapping in RevBayes (Höhna et al. 2016). Green circles represent the number of composite CHGs mapped at each node on the tree. From this analysis we can determine the CHGs that are mapped at a single node or at multiple independent nodes in the tree. (B) Comparing the conservation of the breakpoint (dashed line) and the domain architecture flanking the breakpoint (green and yellow shapes) between two composite genes in a given CHG. (C) Tree shows a gene tree that was constructed from the homologous parent and composite gene sequence regions. This allows us to assess whether the composite genes (red) form a monophyletic clade to the exclusion of the parent genes (black).

3.2.2 Dataset construction for phylogenetic inference

Following assessment of homoplasy within the data, three different datasets were constructed that would be used for subsequent tree inference (Table 3.1). First, datasets of presence/absence of composite CHGs across taxa were created. This was carried out on the full set of 13,632 composite CHGs, and with a subset of these data where we removed CHGs that had evidence of homoplasy (as per analysis in Section 3.2.1). These datasets included the composite CHGs mapped as single gain events by RevBayes (2,285 CHGs) (Figure 3.3A) and the composite CHGs that were confirmed as single events by each of the three approaches described above (Section 3.2.1.2) (Figure 3.3). The three datasets are henceforth named “comp_pa_full”, “comp_pa_singleGain”, “comp_pa_singleEvent” (Table 3.1).

Datasets (# CHGs)	Dataset name	Models applied	Convergence score
13,632	comp_pa_full	Dollo, Mk	0.57, 0.38
2,285	comp_pa_singleGain	Dollo	0.04
4,563	comp_pa_singleEvent	Dollo, Dollo + 4 gamma rates + ascertainment bias	0.1, 0.1

Table 3.1. Presence/absence datasets used for tree inference. Details on the number of CHGs, name, models used during phylogenetic reconstruction, and convergence statistics for each of the three presence/absence dataset used for tree inference in biphy. The convergence score is given for resulting trees inferred under each model.

Second, the primary sequence data, *i.e.* the aligned composite CHG sequences, was used to study the molecular evolution of the composite genes. Due to the size of the dataset, this step was carried out using only the dataset of 2,285 composite CHGs that were annotated as single events in the character mapping analysis (Section 2.3.1). We filtered each of the 2,285 composite CHGs based on taxon sampling and missing data. First, we set a threshold of less than 50% missing data in the alignment and at least 20 species present in a CHG on the 2,285 composite CHGs of single origin, and this produced a dataset consisting of 312 composite CHGs, hereafter named comp_312 (Appendix 3.4). Next, the same approach was taken with the 2,285 composite CHGs but we set the threshold for the number of taxa to at least 30, this created a second filtered dataset of 108 composite CHGs, hereafter named comp_108 (Appendix 3.4). These filtering steps were introduced to the analyses to ensure high quality datasets with robust sampling across the animal tree. In both datasets comp_312 and comp_108 of the filtered composite CHGs there were no CHGs containing the sponge taxa *Amphimedon queenslandica* or the ctenophore *Mnemiopsis leidyi*, meaning that these datasets could not address topology at the root of the animal tree. The sequences for both filtered datasets were aligned and concatenated using scafos (default parameters, *scafos_concat.sh* Appendix 3.4 (Roure et al. 2007)) for subsequent supermatrix phylogenetic reconstruction and are available in (Appendix 3.4).

3.2.3 Tree inference using patterns of presence/absence

Phylogenetic trees were then inferred from the three presence/absence datasets ("comp_pa_full", "comp_pa_singleGain", and "comp_pa_singleEvent") with biphy (Pett et al. 2019), which employs RevBayes (Höhna et al. 2016). Biphy allows a broad range of Bayesian phylogenetic analyses using binary characters (github.com/willpett/biphy). It allows for phylogenetic reconstruction, based on the RevBayes package (Höhna et al. 2016), but also other useful functions such as calculating posterior predictive simulations, marginal likelihood estimations and cross validation scores to check for model fit. Biphy requires presence/absence data files in Phylip, Nexus, or Fasta formats, and is run in a single command with additional parameters specifying the model of evolution, rate distributions, branch length priors, corrections for unobserved sites, and number of MCMC chains (Appendix 3.3, "*biphy_run.sh*"). The software outputs trace files and tree files, which are suitable for estimation of convergence in the MCMC runs using tracecomp and bpcomp programmes, respectively, in PhyloBayes (Lartillot et al. 2013).

Phylogenetic reconstruction using biphy was carried out on each of the three datasets (Table 3.1) using either individual or a combination of models of binary substitution. These models included the reversible binary substitution model, whereby a gene may be gained or lost any number of times (Lewis 2001), and an irreversible Dollo model, in which each character may be gained only once, and can be subsequently lost (Alekseyenko et al. 2008). For the full dataset of 13,632 CHG, we applied both the reversible and irreversible models for tree inference. Next, we estimated the relative fit of both of these models to our data by calculating marginal likelihood estimates (Appendix 3.3 "*biphy_margLikelihood.sh"*). The marginal likelihood is provided as likelihood of the data given the joint prior distribution of the model parameters. In this way, unlike other model fit tests such as likelihood ratio test (LRT) or Bayesian Information Criterion (BIC) which base the fit of a model from its maximum likelihood point in parameter space, marginal likelihood estimation is able to measure the average fit of a model to the data provided (Xie et al. 2011). The method of estimating these likelihoods in biphy is carried out using the stepping-stone sampling approach (Xie et al. 2011). This produces a set of log-likelihood estimates for each model which are then used as input to compute Bayes Factors for a statistical comparison of model fit. Bayes factors are simply the ratios of the marginal likelihoods of the competing models, producing the probability of the data over the posterior distribution. Bayes Factor is calculated by taking twice the difference of the log likelihood scores for the models, and comparing to the Kass and Raferty table (Kass and Raftery 1995). If the resulting value is greater than 6, this suggests strong evidence for a better fit of one model over the other.

For the datasets where the homoplastic composite CHGs were removed (i.e. comp_pa_singleGain, and comp_pa_singleEvent), the irreversible Dollo model was applied and phylogenetic reconstruction was carried out using biphy. Additional parameters were used in combination with the Dollo model when running tree inference using the comp_pa_singleEvent dataset (4,563 CHGs). These included the use of site-specific rates sampled from a Gamma distribution, which allows the rate of change to vary between sites in the matrix, and parameters correcting for unobserved sites whereby characters present in a single species or absent from all species cannot be observed. As these types of characters were removed from our data matrix, accounting for their absence has been shown to be important to avoid biases in the calculated rates of evolution, and the resulting branch lengths and tree topology (Lewis 2001; Pett et al. 2019). To ensure that the tree inference reached consensus across chains, convergence statistics were calculated using the bpcomp program in PhyloBayes (Lartillot et al. 2013). As a proxy to test how well the inferred trees recapitulated known branching orders, each tree was compared to the species tree using the ETE3 package (Huerta-Cepas et al. 2016), to measure the number of shared edges (implemented in Appendix 3.3 “*compare_tree.py*”).

3.2.4 Tree inference using molecular sequence data of the homologous regions across parents and composites

The original set of 2,285 composite CHGs annotated as having single nodes of origin by RevBayes character mapping were filtered to produce two subsets of these data (1) composite CHGs that that were present in at least 20 taxa and no more than 50% missing data, i.e. Dataset “comp_312” and (2) those that had the same threshold for missing data but were present in at least 30 taxa, i.e. Dataset “comp_108”. Phylogenies for both datasets were inferred using the CAT + GTR model in PhyloBayes MPI (Lartillot et al. 2013) with 4 rate categories and accounting for constant sites (Appendix 3.4, "*phylobayes_mpi.sh*").

3.2.5 Expanding the dataset to include an outgroup species

In the initial dataset of 63 metazoan species, there were some clades within the animal tree that were under sampled. This was due to a combination of two factors, firstly the genome data that was available at the time of sampling, and secondly in order to reduce the risk of error due to assembly or annotation in our analysis we placed strict criteria on genome quality. However, this meant that our dataset was missing important outgroup species such as choanoflagellates and other single cell organisms. To introduce homologs from outgroup species to each of our composite CHGs, we carried out reciprocal BLASTp (Altschul et al. 1990) searches (E-value <= 1e-5, percent identity >= 30%, coverage >= 80%) of each of the 13,632 composite CHGs against the protein coding regions of the choanoflagellate species *Monosiga brevicollis* (which has BUSCO score of 78.6% (King et al. 2008)).

Due to the nature of composite genes, which share partial homology to their parent genes, it was important that the homology searches were filtered to ensure that only genes that share full homology to the subject composite gene were retained. The following three steps were taken to provide this filter for the BLASTp output; (i) the homology coverage threshold was set to 80%, (ii) the percent sequence identity was set to 30%, and (iii) the homology hit was required to have coverage across the breakpoint location (Appendix 3.5 "comp_ambig_breakpoints.csv") of the composite gene and query sequence (assuming that the homologous composite is a result of a single remodeling event and would thus have a similar breakpoint region in the gene). These strict homology filters increased the accuracy of detecting true homologous composite genes in the additional species of *Monosiga brevicollis*.

3.3 Results

3.3.1 Independent origin and secondary loss across the animal tree is a common property of remodeled genes

We find an average CI score of 0.4 across the full dataset of 13,632 composite CHGs suggesting that composite gene events occur and are lost frequently across the tree. However, the distribution of CI scores across this dataset of 13,632 composite CHGs, shows that for 4,536 of these CHGs their CI scores are between 0.5 (relatively low levels of homoplasy) and 1 (non-homoplastic) (Figure 3.4). The subset of composite CHGs with a CI score of 0.5 display two changes across the tree (e.g. have been gained and subsequently lost or gained twice). It is plausible that the subset of 4,536 CHGs with a score of 0.5 up to 1.0 may have sufficient signal for use as phylogenetic markers.

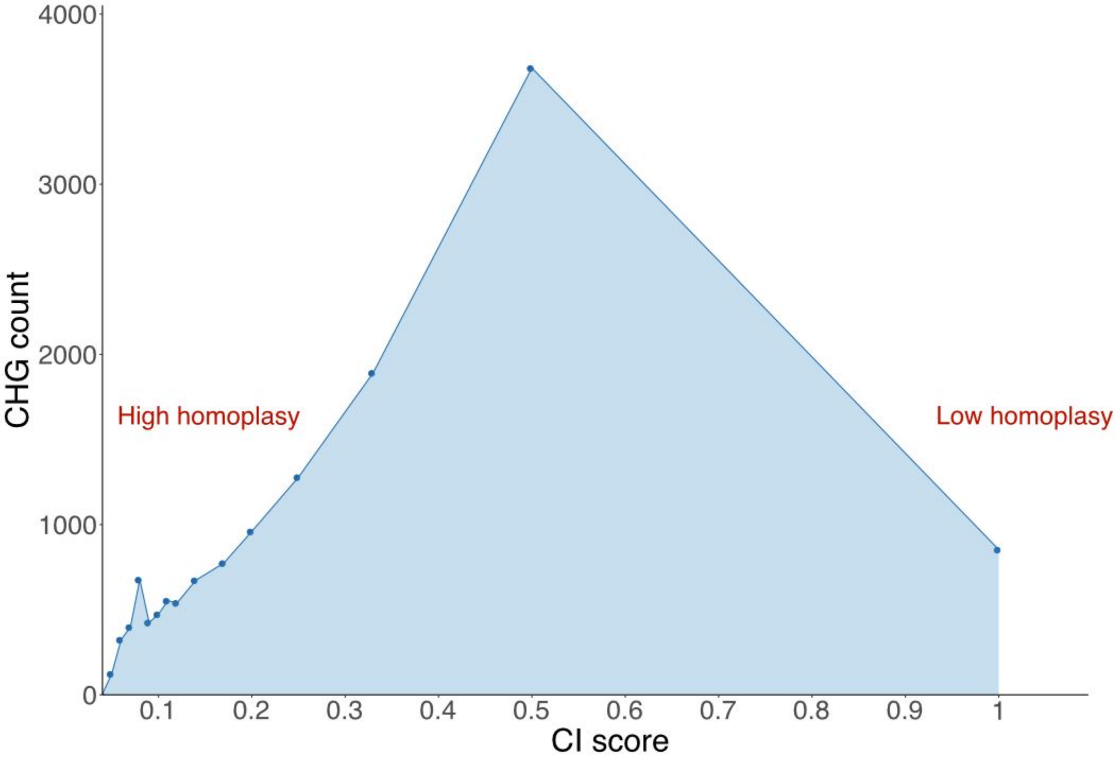


Figure 3.4. The distribution of CI scores for the complete dataset of 13,632 composite CHGs. Each point in the plot represents the number of CHGs (y-axis) with the given CI score on the x-axis.

To contextualize the amount of signal and noise within our dataset of 13,632 composite CHGs, we compared the distribution of CI scores to three other datasets containing sampling from across the animal tree (Figure 3.5). The Tarver et al. (2018) microRNA dataset has a mean CI score of 0.88, demonstrating a strong phylogenetic signal. The distribution of scores also shows that the majority of these 1,143 miRNA families have a score of 1 indicating no detectable homoplasy. The Pett et al. (2019) dataset of homologous gene families also showed a distribution indicative of low levels of homoplasy, with a mean CI score of 0.79 across the 35,244 gene families. The OMA HOG dataset produced a pattern of CI scores more similar to that of the composite dataset, however, there was a higher overall average CI score for the OMA HOG set 0.7 as compared to 0.4 for the composite CHGs, and there were a greater number of characters in the OMA HOG set that had a CI score of 1.

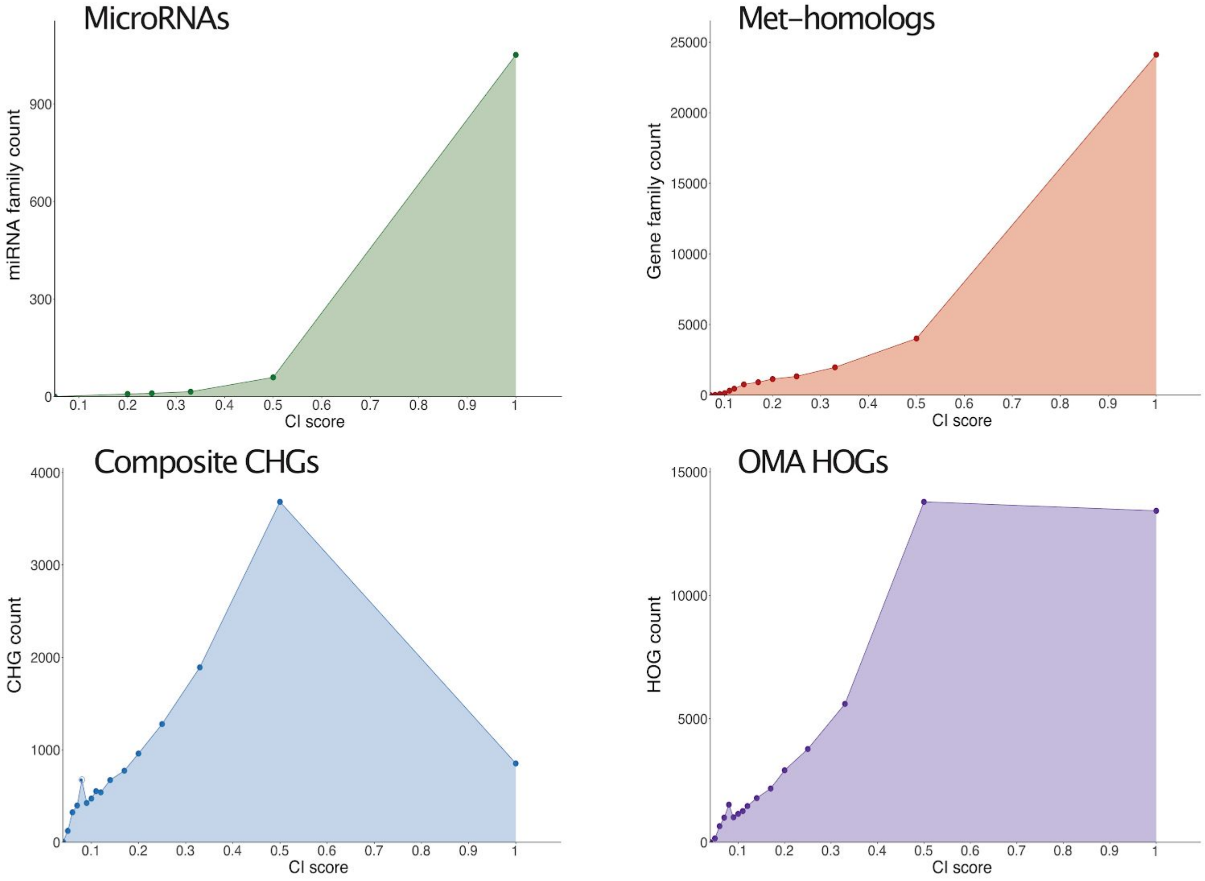


Figure 3.5. Comparison of CI score distributions between different datasets. Distribution of CI scores across the following datasets (A) microRNA gene families (green), (B) homologous gene families (orange), (C) composite CHGs (blue), and (D) HOGs (purple).

Overall, we find a high level of homoplasy (analysis above) and a high rate of composite gene gain and loss (analysis presented in Chapter 2). However, we have been able to identify a set of 2,285 composite CHGs from a character mapping approach that were gained only once across the animal phylogeny and we wished to focus our efforts on characterising these in more detail. This set of 2,285 composite CHGs may possess useful phylogenetic signal. In order to see if we could find further support for a set of non-homoplastic composite CHGs, we analysed the breakpoint and domain conservation, as well as the patterns of substitutions with the composite genes in the full dataset of 13,632 composite CHGs. In doing so, we could check whether taking alternative approaches to find composite CHGs of single origin would produce a set consistent with those found to have a node of single origin i.e. could we recover the set of 2,285 composite CHGs from the full dataset using these other definitions of single events.

In this analysis we assessed the full set of 13,632 composite CHGs for signatures of single origin events of composite genes. After filtering (see Section 3.2.1.2), we were able to test 10,855 CHGs for conservation of breakpoint and domain architecture across composite genes, and 10,849 composite CHGs for the existence of a monophyletic composite gene clade in the composite-parent gene tree. In our analysis of breakpoint and domain architecture conservation, we identified 1,461/10,855 composite CHGs supported as single origin events. In our phylogeny based approach, we identified 2,506/10,849 composite CHGs as single events as their constituent genes form a monophyletic clade in the composite-parent gene tree. These results reflect similar proportions of CHGs annotated as single events as the character mapping approach suggesting that high secondary loss or parallel evolution of composite genes is common. However, combining each of these approaches results in just 152 composite CHGs that were consistent across the different analyses (Figure 3.6A). If we take the assumption that any support for single origin is sufficient then we have 4,938 CHGs of single origin. Conversely, CHGs of multiple origin comprise the majority of the dataset, either 99% or 64% of the full set of 13,632 CHGs depending on how lenient or strict we wish to define composite CHGs of single origin i.e. whether we accept that support of single origin from any of the three approaches (combined), or whether single origin must be supported by all three approaches (all) (Figure 3.6B).

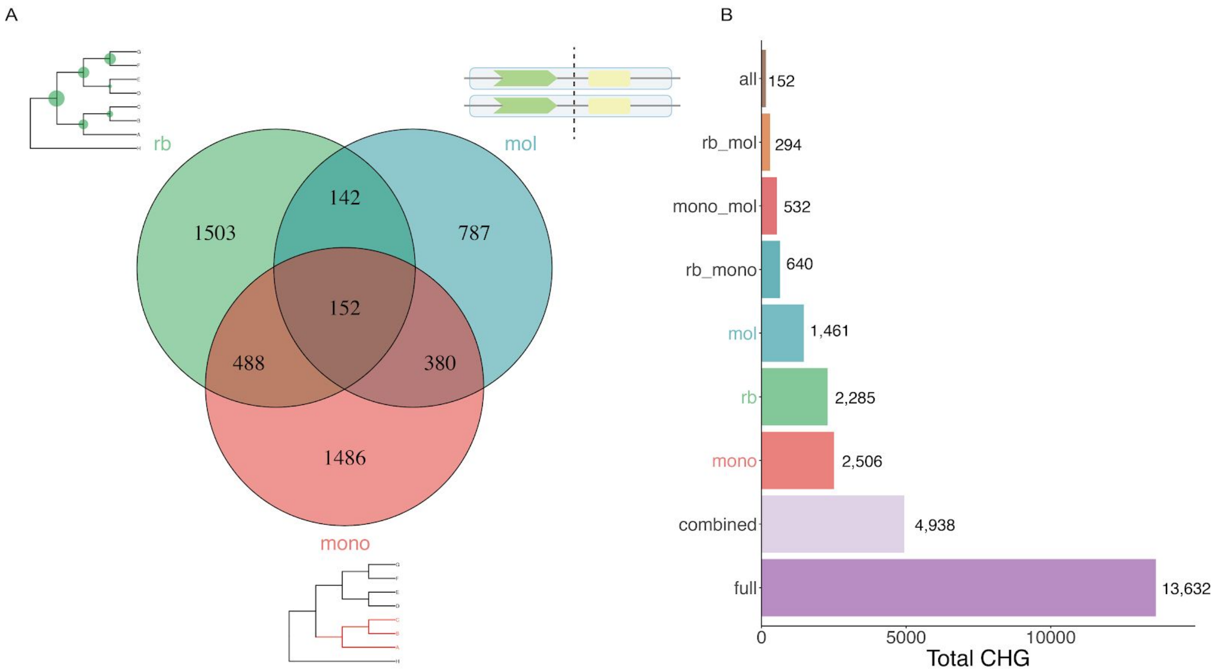


Figure 3.6. Numbers of composite CHGs confirmed as single events using three different methods. (A) Venn diagram of the number of composite CHGs confirmed as being of single origin using Revbayes (rb) character mapping in green, breakpoint and domain conservation (mol) in blue, and the monophyly of composite genes in the composite-parent gene tree (mono) in red. (B) Barplot showing the number of composite CHGs that are of single origin from the 152 composite CHGs in common across all approaches, to the 4,938 composite CHGs suggested by at least one approach to be of single origin. The number of the full set of composite CHGs is shown at the bottom (dark purple) as a comparison.

3.3.2 Composite genes in phylogeny reconstruction

Having established the level of homoplasy within our dataset of composite CHGs, tree inference was performed using a set of stochastic models implemented in RevBayes (Höhna et al. 2016). The two primary models used for tree inference of binary characters are the reversible Mk model (Lewis 2001) and the irreversible Dollo model (Alekseyenko et al. 2008). To assess which model was most appropriate to describe the evolution of composite CHGs in our full dataset, tree inference was carried out on the presence/absence matrix of the 13,632 CHGs ("comp_pa_full dataset") using both the Mk and the Dollo model. To infer the model of best fit to the data, marginal likelihood tests were carried out. The result was a lower log likelihood score (LnL = -271,955) for the reversible Mk model (Table 3.2), with a Bayes Factor score greater than 6, suggesting that composite genes are gained and lost in equal capacity. This also provides justification for the use of a reversible model when mapping the composite CHGs to the phylogeny to infer patterns of gain and loss.

Model	Marginal Likelihood
Mk; reversible	-271,955
Dollo; irreversible	-293,479

Table 3.2. Marginal likelihood estimates for the full composite CHG data. Assessment of the fit of the two models (Mk and Dollo) used for phylogenetic inference on the full dataset of 13,632 composite CHGs.

The trees generated using both the Mk model (Figure 3.7A) and the Dollo model (Figure 3.7B), demonstrate some ability to recreate known branching orders within the animal tree. However, the most of these correct branching orders are closer to the tips of the tree, between smaller groups of species. The deeper nodes in the tree represent more problematic regions to resolve, with a number of polytomies present, notable examples are the root of the Dollo tree (Figure 3.7B), and the branching order between the three bilaterian clades in the Mk tree (Figure 3.7A). Additionally, certain clades which are known to be monophyletic are shown as paraphyletic such as the Deuterostomia, Ecdysozoa, and Lophotrochozoa in the Mk tree, similarly the Ecdysozoa and Lophotrochozoa are shown as paraphyletic in the Dollo tree. The deeper internal nodes in the phylogenies produced under both models show lower posterior predictive support values (e.g. PP of 0.54, 0.58, and 0.72 at the three deepest nodes in the Mk tree, and PP of 0.72 and 0.54 at the two deepest nodes within the Dollo tree). When we compared the number of shared edges between the trees constructed under the Mk and Dollo models with the known species phylogeny (taking sponges as the primary emerging lineage), we found 72% and 75% of the edges are shared with the species tree, respectively (calculated using "*compare_trees.py*", Appendix 3.3). It is clear that the ratio of signal to noise is too high within the full set of 13,632 composite CHGs, owing to the high levels of homoplasy, to resolve deep branches within the animal tree. It is important to note, however, that due to the large size of the dataset, neither of the tree inference runs were able to reach convergence (Mk; 0.38 and Dollo; 0.57, where a score of 0.1 represents complete convergence between chains).

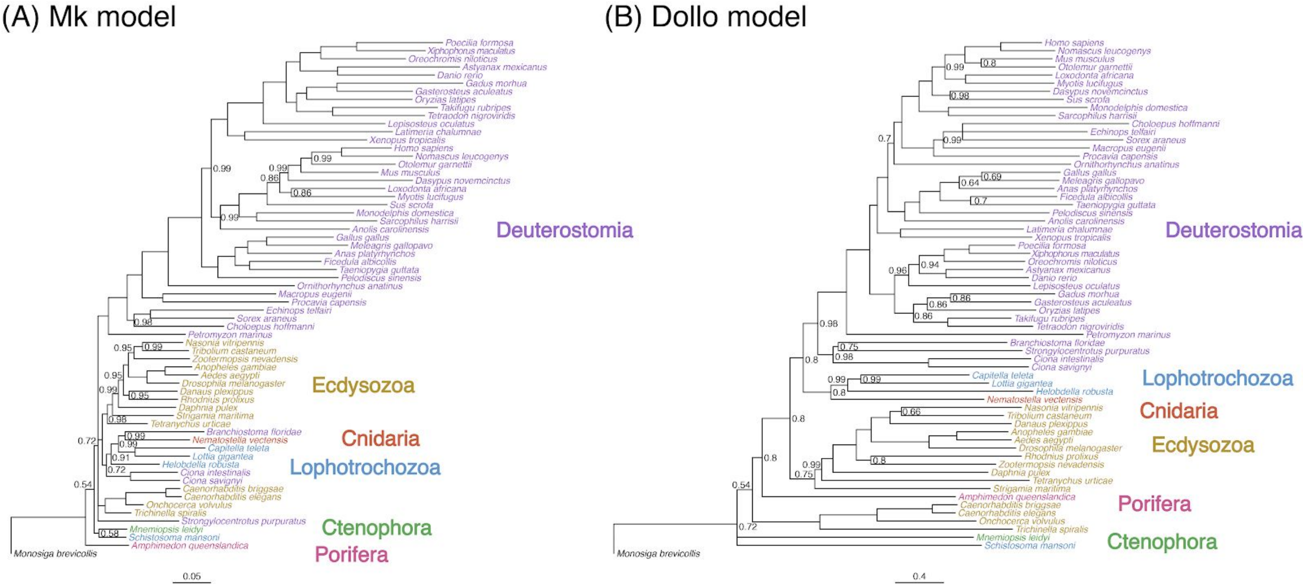


Figure 3.7. 13,632 composite CHGs (comp_pa_full) used to generate animal phylogenies under two different models of evolution. (A) The tree inferred under the reversible Mk model, with species from major phyla coloured along with the phylum name. (B) Tree inferred under the irreversible Dollo model, with major clades labelled by the same colour scheme. Bayesian posterior probabilities (PPs) are labelled for nodes with a score less than 1.

Next, we removed composite CHGs that had evidence of multiple gains from the dataset (as identified in Section 2.3.1) and inferred a phylogeny from the remaining 2,285 composite CHGs ("comp_pa_singleGain" dataset). Using the "comp_pa_singleGain" dataset we inferred a phylogeny using the Dollo model (as these characters are annotated as being gained only once), allowing for subsequent losses but never multiple gains. This resulted in a topology closer to that of the species tree and with better a higher range of node support (with PP values in the range 1.0 to 0.63), although some internal nodes did show lower support, such as the clade grouping both protostome lineages (PP 0.63), and the root of Ecdysozoa (PP 0.73) (Figure 3.8). When we compare the number of shared edges between our inferred tree and the constructed tree, we find 89% of the edges are recapitulated in our analysis (Appendix 3.3, "*compare_trees.py*"). Both Deuterostomia and Ecdysozoa were recapitulated as monophyletic clades. Lophotrochozoa was shown as paraphyletic, with *Schistosoma mansoni* placed as sister to Planulozoa (Cnidaria + Bilateria) and *Lottia gigantea* placed as the sister lineage to the rest of the bilaterian clades (although with lower support; PP=0.85). These unexpected placements could be due to poor genome annotation of these species or it could be due to actual peculiar biology of the genomes and gene content, for example *Schistosoma mansoni* has undergone significant genome reduction and gene loss (Guijarro-Clarke et al. 2020) possibly owing to its parasitic lifestyle. If these rogue species, which were deemed to be incorrectly placed by comparison to the known species phylogeny, are removed, all known major clades (Bilateria, Deuterostomia, Protostomia, Lophotrochozoa, and Ecdysozoa) are resolved as monophyletic using Dollo parsimony and the dataset of 2,285 composite CHGs (Figure 3.8 inset). This tree also recapitulated more known edges when compared to the species tree (94% of the edges shared with the species tree). The porifera species *Amphimedon queenslandica* was placed at the root of the animal tree, sister to the remaining animal clade (i.e. supporting the "Porifera-sister hypothesis"), with the ctenophore *Mnemiopsis leidyi* branching just after Porifera.

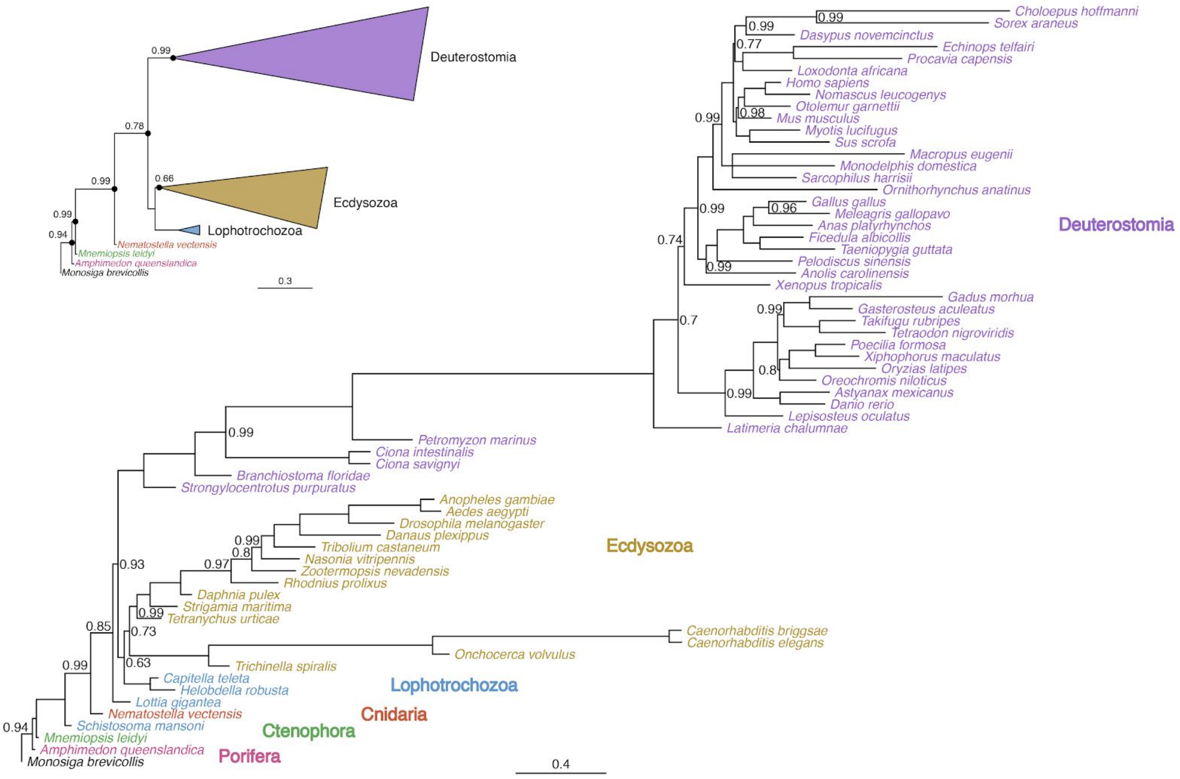


Figure 3.8. Phylogeny inferred using Dollo Parsimony from 2,285 Composite CHGs of single origin as detected using character mapping. The primary, larger tree shows the resulting tree inferred using presence/absence of the 2,285 CHGs annotated as single origin events by the character mapping analysis. Major clades are coloured by lineage and labelled with the phylum name. Posterior probabilities (PPs) are labelled on the nodes where a value of less than 1 was observed. The inset phylogeny shows the analysis using the same characters with the two misplaced species from the initial analysis (*Schistosoma mansoni* and *Lottia gigantea*) removed. Major clades are collapsed to represent only the branching pattern between Deuterostomia, Ecdysozoa, Lophotrochozoa, and the non-bilaterian lineages. Major clades are represented by coloured triangles, based on the same phylum colour scheme as the larger tree. Nodes with PPs of less than 1 are labelled.

Following this we ran the same tree inference steps on a dataset of 4,938 CHGs ("comp_pa_singleEvent" dataset) annotated as single remodeling events by one of the three methods described in Section 3.2.1.2. Firstly, a phylogeny was inferred under the Dollo model alone (Figure 3.9A). Then, using the Dollo model in combination with a gamma distribution of rates across sites using four discrete categories, and correcting for the absence of gene families and singletons lost in all lineages we inferred the phylogeny from these 4,938 composite CHGs (Figure 3.9B).

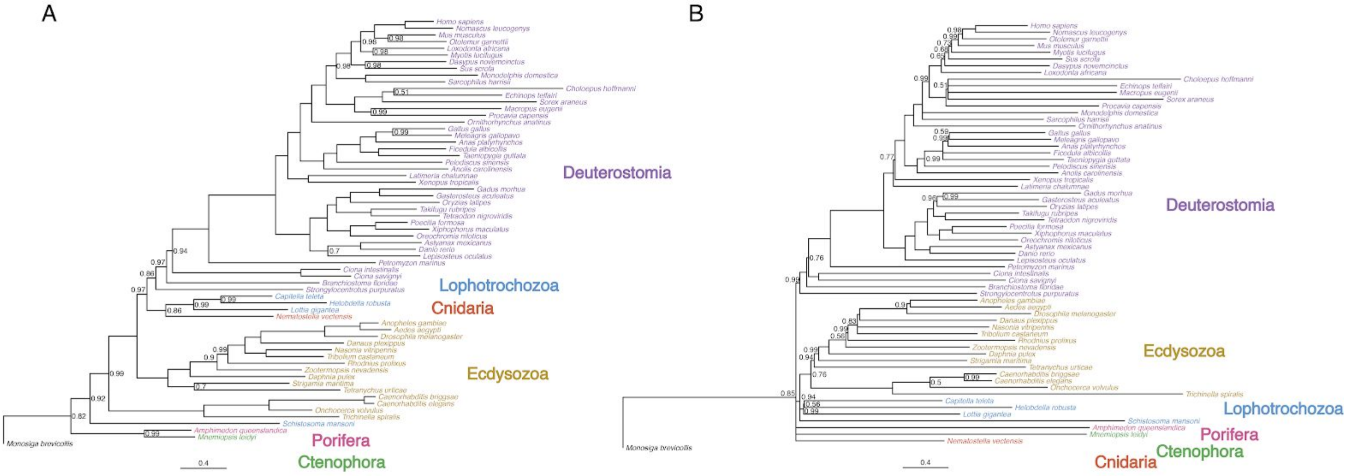


Figure 3.9. Phylogenies created from the 4,938 CHGs confirmed as single events by at least one of the methods from Section 3.2.1.2. (A) Tree inferred from the 4,938 CHGs under the Dollo model. (B) Tree inferred under the Dollo model with a gamma distribution of rates using 4 discrete categories and the application of an ascertainment bias which corrects

for the absence of singletons in the dataset. Major clades are coloured by lineage and labelled with the phylum name. Node values show PP support values that are below 1.

The analysis using the Dollo model alone reached convergence (0.1), and recovered a monophyletic Deuterostomia (Figure 3.9A). Overall, the proportion of shared edges with the species tree was 81%. Lophotrochozoa and Ecdysozoa were annotated as paraphyletic. Similar to previous analyses, the platyhelminth worm *Schistosoma mansoni* was placed near the root of the tree. The cnidarian *Nematostella vectensis* was grouped with the remaining lophotrochozoan species (PP 0.86), with this group placed as sister to the deuterostome clade (PP 0.97). This branching pattern, where *Nematostella vectensis* branches within Bilateria, is also found in the analysis of the full set of 13,632 composite CHGs (comp_pa_full dataset). The sponge *Amphimedon queenslandica* and the ctenophore *Mnemiopsis leidyi* were placed outside of Bilateria (PP 0.82), however they were grouped as sister lineages to one another, meaning that the root of the tree remained unresolved.

In the analysis of “comp_pa_singleEvent” dataset with Dollo model plus parameters adjusting for ascertainment bias, which includes the additional rate parameters and accounts for lineage specific loss, the chains reached convergence (0.1), and recovered a monophyletic Deuterostomia, Ecdysozoa, and Bilateria. Lophotrochozoa was represented as a polytomy, unable to place *Schistosoma mansoni* with the other lophotrochozoan species. However, this meant that a monophyletic Protostomia (Ecdysozoa + Lophotrochozoa) was recovered. The proportion of edges matching the species was greater than the proportion recovered in the tree inferred under the dollo model alone (83%, versus 81% for Dollo model). The root of the animal tree was also shown as a polytomy, with the relationships between three non-bilaterian species (representing Porifera, Ctenophora, and Cnidaria) remaining unresolved. In both analyses the internal relationships within the major clades (i.e. Deuterostomia, Lophotrochozoa, and Ecdysozoa) represented relatively accurate topologies with relation to the known species tree (83% of the edges matching the known species tree). Within Deuterostomia, the main clades of Mammalia, Sauria, and Actinopterygii (ray-finned fish) are recovered.

3.3.3 Phylogenetics of composite CHG amino acid sequences

Following the use of presence/absence of composite genes for phylogenetic reconstruction, we next explored the phylogenetic informativeness of the molecular sequence data (in amino acid format) of the composite genes. As the “comp_312” dataset did not reach enough generation to accurately measure convergence, we only present the findings from the “comp_108” dataset. Reconstruction using a set of 108 composite CHGs (“comp_108”), with a single node of origin (non-homoplastic), less than 50% missing data, and present in 52 species, resulted in a relatively well resolved phylogeny with 87% of the edges matching the known species tree. Even though the tree chains did not reach convergence (bpcomp score of 1), we find correct branching in certain clades, such as Mammalia, Sauria, and Ecdysozoa (Figure 3.10). The basal deuterostome lineages, *Branchiostoma floridae* and *Strongylocentrotus purpuratus*, were incorrectly placed alongside clades of Protostomia and Deuterostomia in a polytomy. However, allowing the chains to run for more iterations in order to reach convergence could result in a better placement of these lineages. The lophotrochozoan *Helobdella robusta*, was incorrectly placed within Deuterostomia, sister to *Xenopus tropicalis*, however with low support (PP=0.5). Additionally, deeper nodes such as the base of Deuterostomia clade and Vertebrata show low support (PP = 0.57 for both), however, support for these well established nodes could increase when the chains reach convergence. Overall, while this datasets is importantly missing the other non-bilaterian lineages Amphimedon queenslandica and Mnemiopsis leidyi, the use of molecular sequence data of composite CHGs provides another potential source of phylogenetic information to resolve issues remaining in AToL.

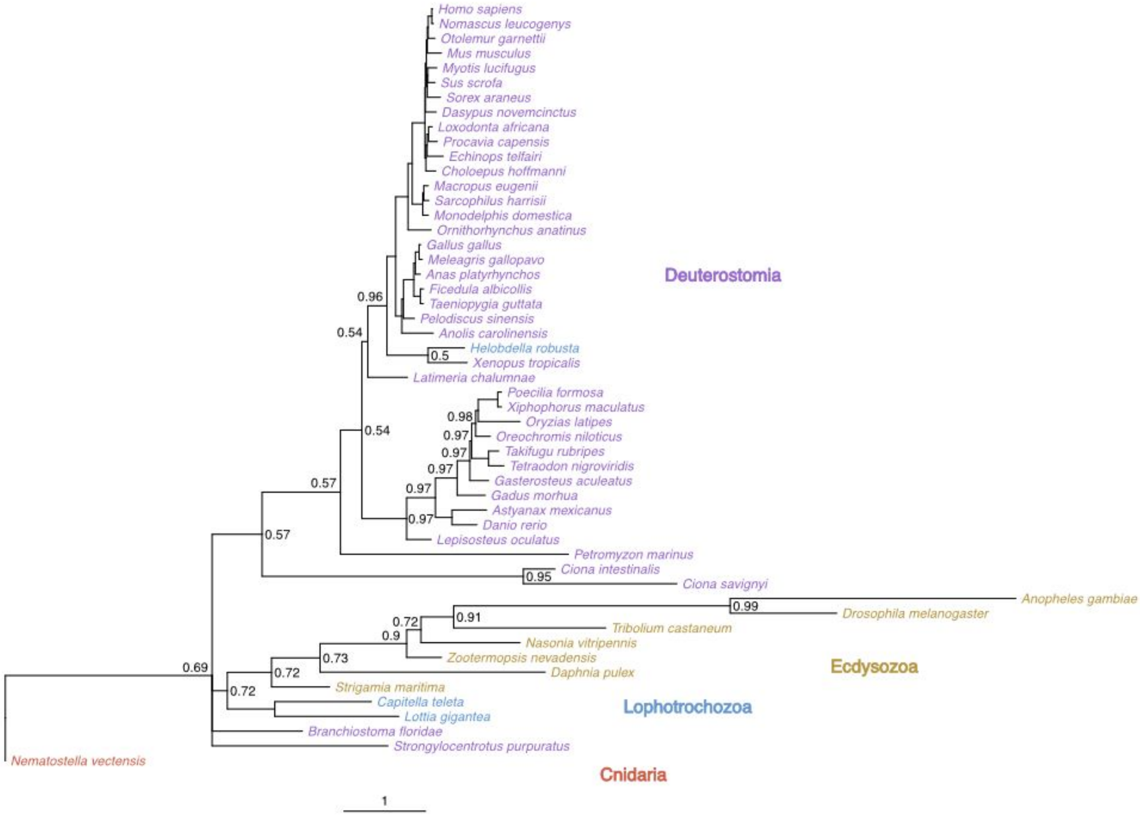


Figure 3.10. Phylogeny created from the amino acid data for 108 CHGs (comp_108) using PhyloBayes. Major clades are coloured by lineage and labelled with the phylum name. PP support values below 1 are shown, all other nodes have PP = 1.0.

3.4 Discussion

We have found that composite genes evolve in a non-clocklike manner with heterogeneous rates of gain and loss across the animal phylogeny. Furthermore, parallel emergence of the same composite genes in separate parts of the tree appears to be prevalent across our dataset. Using a range of methods we have assessed the level and extent of parallel evolution of composite CHGs, and we have found similar numbers of CHGs annotated as emerging multiple times independently in different species (e.g. 2,285, 1,461, or 2,506 CHGs). This suggests that parallel evolution of the same protein coding gene may be a common and understudied process of gene evolution within animal genomes. Indeed, repeated independent origin of specific gene fusions from the same parental protein coding genes has been observed in bacteria (Makiuchi et al. 2007; Stover et al. 2011) and in single cell eukaryotes (Stover et al. 2005). In animals, convergent fusion of the *TRIM5-CYPA* gene in New World monkeys (Sayah et al. 2004) and Old World monkeys (Brennan et al. 2008), and the repeated fusion of β-globin genes in Laurasiatheria (Gaudry et al. 2014) have been reported. More broadly speaking, 25% of all multi-domain proteins in eukaryotes have emerged independently several times, and 70% of domain combinations in the human genome have been found to be independently gained in at least one other eukaryotic genome (Zmasek and Godzik 2012). Taken together these studies, along with our findings, would suggest that selection for the same combinations of gene sequences in composite genes could be common in gene remodeling. And from our study we now have an empirical estimate for the level of parallel evolution in remodeled genes in animal genomes. The results of our three approaches to test for parallel evolution within our composite CHGs did not provide consistent agreement (Figure 3.6), and this could be due to a number of factors such as incomplete protein domain annotation and inaccurate breakpoint annotation of composite genes from BLAST homology search. Why we see these patterns of parallel emergence of composite genes in animal genomes is still uncertain. One hypothesis could stem from the nature of protein domains, whereby promiscuous domains could easily associate and disassociate with other domains multiple times throughout evolutionary history. The rate of composite gain and loss suggest a constant turnover of genes that undergo remodeling, which may provide insight as to why we observe the same composite genes forming independently. If two or more genes, or portions of the genes, are not deleterious when remodeled, given the high rate of composite formation, this provides some reason for the high rates of parallel evolution.

The high rate of gene turnover (as inferred from character mapping and CI scores) and high level of parallel evolution (64-99%) in composite CHGs implies that these characters are highly homoplastic. Indeed, the CI test for homoplasy revealed an overall high rate of homoplasy (average CI score of 0.4) within the full set of composite CHGs. This suggests that the rate of gene gain and loss is higher in composite genes than any other molecular data type analysed here (Figure 3.5). It is important to note that the CI score is negatively correlated with the number of species sampled, so datasets with a larger tree may result in lower CI scores. However, this is a useful test to get an empirical measurement of the usefulness of a character to be used as a phylogenetic marker. Importantly, we know from this analysis and subsequent confirmation based on molecular characteristics and phylogenetic distributions that there is a subset of composite CHGs that were gained only once and are therefore suspected to have low levels of homoplasy. We proposed that this set of 2,285 characters (comp_pa_singleGain dataset) may contain strong phylogenetic signal and be useful for species tree inference.

Comparing the resultant phylogenies of the full dataset of 13,632 composite CHGs (“comp_pa_full”) to the subsets of data cleared of homoplastic characters (“comp_pa_singleGain” and “comp_pa_singleEvent” datasets), the resulting topologies of known clades between the two trees suggests that this is indeed the case. While the “comp_pa_full” dataset recovered some correct branching orders close to the tips of the tree, it was overall unable to recapitulate known branching patterns within the animal tree, with polytomies and low branch support particularly at the very early nodes within the animal tree. It is important to note, however, that both analyses using the full dataset of composite CHGs failed to reach convergence. We therefore cannot rule out that some of these differences in branching orders may disappear with longer run times. The species tree inferred from the 2,285 CHGs (“comp_pa_singleGain”) annotated as single gain events by RevBayes character mapping in particular provided a more accurate reconstruction of the animal phylogeny. With the exception of the misplaced positioning of five species, some of which may possess fast evolving genomes such as *Schistosoma mansoni*, all deep internal nodes and shallower tip relationships recapitulate the known species tree. There remains a number of unresolved regions within AToL that persist today, such as whether Deuterostomia are monophyletic (Philippe et al. 2019), and which non-bilaterian lineage should be placed at the root of the tree as sister lineage to the remaining animal groups (King and Rokas 2017). The difficulty in resolving these clades has been placed on poor modeling of the data, difficulty in extracting signal from such deeply diverging branches, and a lack of strong markers to place them. Here, we present a dataset of molecular markers that can recapitulate known relationships within AToL and is thus a good proxy for addressing some of these contentious regions. The phylogeny produced from the 2,285 CHGs of single origin (as identified by RevBayes character mapping analysis) was also assessed to determine what the placements were for these two contentious regions of the animal phylogeny. We found that these set of characters produced a topology most similar to known branching patterns within the species tree (89% of edges matched with species phylogeny). We recovered a monophyletic Deuterostomia, and Porifera as the primary emerging animal lineage. Our approach is limited due to undersampling of ambulacrarian and xenacoelomorphan species (which would be required to robustly test the monophyly of Deuterostomia), non-bilaterian species and non animal outgroup species. However, this set of characters have been shown to contain phylogenetic signal and have recaptilutated uncontested parts of the animal tree.

Thus, gene remodeling with deeper sampling of non-bilaterian species and of early diverging bilaterian lineages may contribute to long standing contentious nodes in the AToL.

3.5 Conclusion

We have shown how the evolution of gene remodeling occurs across the animal tree, and have discovered a high rate of parallel evolution of the same composite gene. This has major implications for our understanding of how protein coding genes evolve, how we construct gene families to infer phylogenies, and how we annotate function between homologous genes. Accounting for and removing homoplastic characters we have created a unique dataset of molecular markers for phylogenetic reconstruction. Here, we show that these composite CHGs have the ability to reconstruct known and uncontested relationships within the AToL, and also show support for a monophyletic Deuterostomia and Porifera as sister to all other animal lineages.

Chapter 4: Ortholog selection and the application of composite genes – a case study of Xenacoelomorpha

4.1 Introduction

Xenacoelomorpha is a clade of morphologically simple marine bilaterian worms with a controversial history (Hejnol and Pang 2016; Marlétaz 2019). The superphylum consists of two phyla; Xenoturbella and Acoelomorpha, Acoelomorpha composed of two orders; Acoela and Nemertodermatida (Jondelius et al. 2019). Xenacoelomorpha has traditionally been placed as the primary emerging bilaterian lineage (Figure 4.1 A), implying an ancestrally simple body plan of the most recent common ancestor of Bilateria, with a simple brain, blind gut, and lacking excretory and vascular systems (Hejnol and Pang 2016). Alternative placement, within Deuterostomia (Figure 4.1 B and C) suggests more complex morphological evolution within Bilateria and a secondary simplification of the Xenacoelomorpha clade (Philippe et al. 2019). This suggests the loss of a large number of significant morphological traits which are considered to be present in the last common ancestor of deuterostomes such as complex organ systems, a digestive system with mouth and anus, coelomes, and body compartmentalisation. Whichever position is eventually settled upon for this group of worms, they are of major importance for understanding the evolution of Bilateria. More recently, questions surrounding the placement of Xenacoelomorpha within Bilateria has led to further profound queries addressing the branching patterns within Bilateria, namely whether Deuterostomia are monophyletic or whether their support is a result of systematic error (Philippe et al. 2019). This alternate topology, placing Chordata as sister to Protostomia to the exclusion of Xenambulacraria (Figure 4.1 C), suggests that the common ancestor of all bilaterians possessed many deuterostome traits such as radial cleavage and the development of the anus from the blastopore, and that these traits were significantly altered or lost at the emergence of Protostomia.

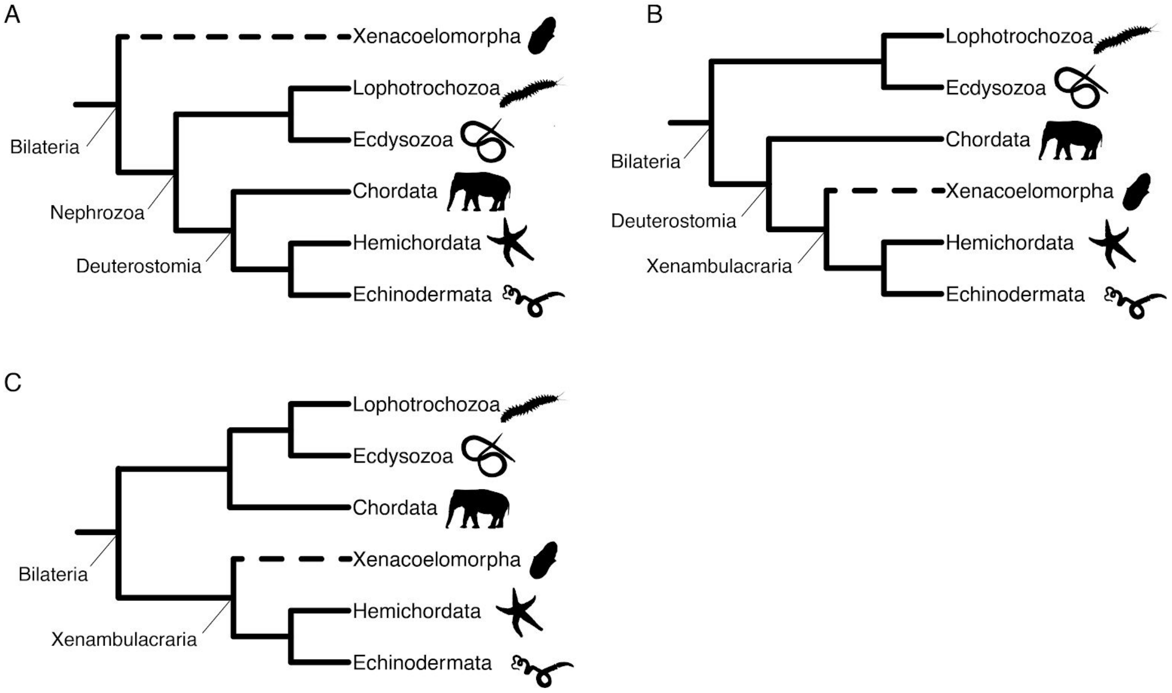
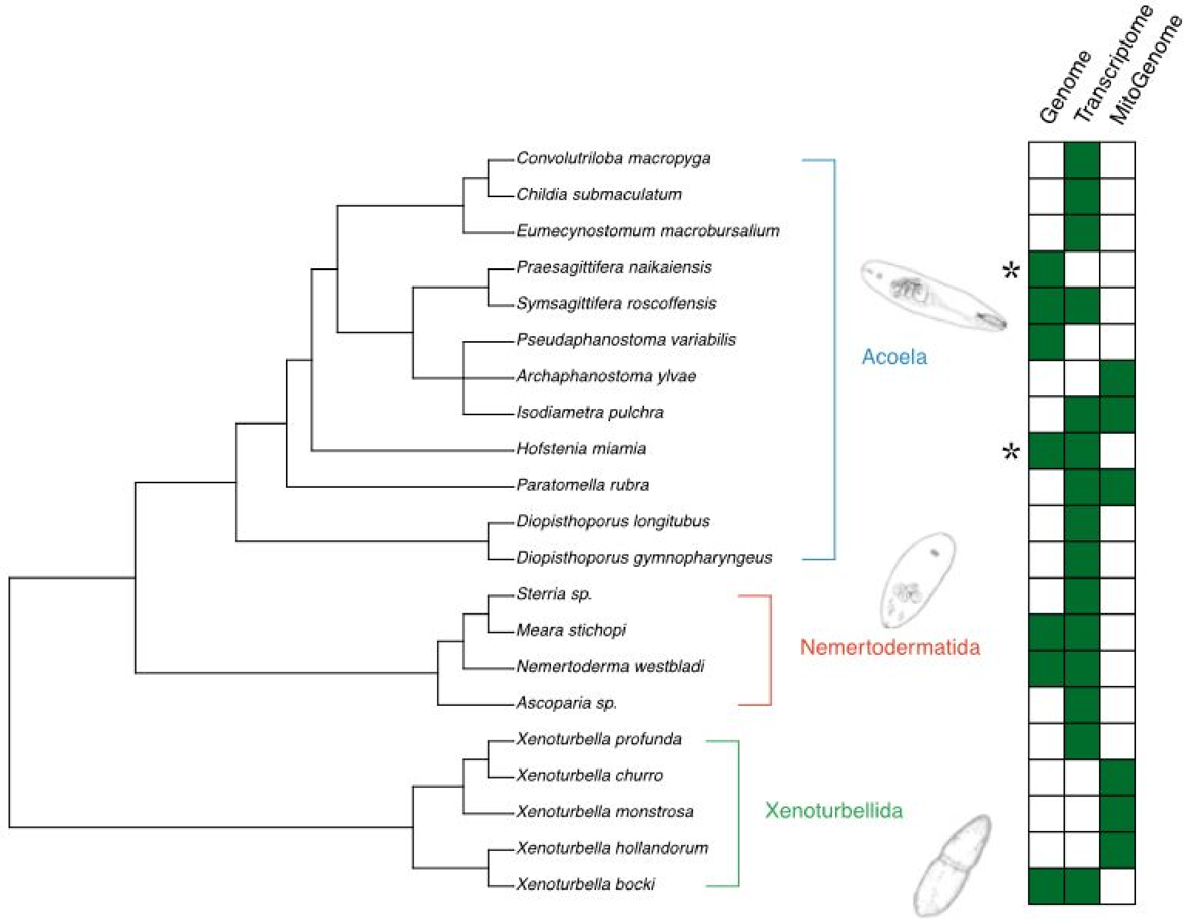


Figure 4.1. Alternative placements of Xenacoelomorpha within the animal tree. (A) "Nephrozoa hypothesis": Xenacoelomorpha the primary emerging bilaterian lineage sister to Nephrozoa (Hejnol et al. 2009; Rouse et al. 2016; Cannon et al. 2016). (B) "Xenambulacraria hypothesis": Xenacoelomorpha within Deuterostomia and sister to the Ambulacraria clade (Philippe et al. 2011). (C) Xenacoelomorpha sister to Ambulacraria, with a non-monophyletic Deuterostomia (Philippe et al. 2019).

Systematic error such as LBA within these phylogenetic studies caused by, e.g. the faster molecular evolutionary rate observed in Acoelomorpha genomes, could be erroneously pushing species of the Acoelomorpha clade to a basel position in Bilateria. In addition, the lack of high quality genomic sequence data for Acoelomorpha species contributes to the challenge of resolving their position (Figure 4.2). There are only two complete assembled genomes with high contiguity, the acoel species *Hofstenia miamia* (N50 = 294Mb, BUSCO score of 90%) (Gehrke et al. 2019), and *Praesagittifera naikaiensis* (N50 = 117Kb, BUSCO score of 76.5%) (Arimoto et al. 2019). Recently, Philippe et al. (2019) produced new and updated genomes for two nemertodermatid species *Nemertoderma westbladi* and *Meara stichopi*, two acoel species *Symsagittifera roscoffensis* and *Pseudaphanostoma variabilis* and the xenoturbellid *Xenoturbella bocki* (Figure 4.2). While the number of genomes for these undersampled species are increasing, the absence of high quality genomes, along with the use of a large number of transcriptomic sequence data, presents significant issues for phylogenomic analyses (Siu-Ting et al. 2019).



		Genome	Transcriptome	MitoGenome
Convolutriloba macropyga			🟩	
Childia submaculatum			🟩	
Eumecynostomum macrobursalium			🟩	
Praesagittifera naikaiensis	*	🟩		
Symsagittifera roscoffensis		🟩	🟩	
Pseudaphanostoma variabilis		🟩		
Archaphanostoma ylvae				🟩
Isodiametra pulchra			🟩	🟩
Hofstenia miamia	*	🟩	🟩	
Paratomella rubra			🟩	🟩
Diopisthoporus longitubus			🟩	
Diopisthoporus gymnopharyngeus			🟩	
Sterria sp.			🟩	
Meara stichopi		🟩	🟩	
Nemertoderma westbladi		🟩	🟩	
Ascoparia sp.			🟩	
Xenoturbella profunda			🟩	
Xenoturbella churro				🟩
Xenoturbella monstrosa				🟩
Xenoturbella hollandorum				🟩
Xenoturbella bocki		🟩	🟩	

Figure 4.2. Available sequence data for Xenacoelomorpha clade. The tree on the left shows the phylogeny for species within Xenacoelomorpha (based on molecular data) for which sequence data is available (Rouse et al. 2016; Cannon et al. 2016; Robertson et al. 2017; Philippe et al. 2019). Highlighted are the three xenacoelomorph clades, Xenoturbella, Nemertodermatida, and Acoela. The presence/absence profile on the right shows available sequence data for each species, with three columns representing genomic data, transcriptomic data, or mitochondrial (MitoGenome) genomic data. Green highlights the presence of a given sequence data type for each species, with white representing absence. * genomes of high quality with large scaffolds. Species images are taken from (Cannon et al. 2016).

Besides the obvious effects of poor taxon sampling, lack of genomic data can impact phylogenomic studies in a number of ways (see Section 1.1.3.1). For example, transcriptomes or low quality genomes may have poor annotations, or indeed assemblies, and as a result may be missing genes or gene families erroneously. Paralogs missing due to poor annotation or poor quality data could lead to the mis-annotation of a duplicate gene as a single copy ortholog, leading to potential biases in phylogenomic analyses caused by inadvertent paralog selection (see Section 1.1.3.2) (Brown and Thomson 2017; Siu-Ting et al. 2019; Walker et al. 2020). Genuine alternative loss of paralogs may also result in the misannotation of genes as single copy orthologs in a species. Given that the clade Xenacoelomorpha, particularly Xenoturbellida, has been found to have higher rates of gene loss when compared to other animal clades (Fernández and Gabaldón 2020), the effects of hidden paralogy could be a major issue in phylogenomic studies attempting to place this group of worms.

The “late loss” of the ancestral ortholog results in the observed extant, ancestrally paralogous, gene being annotated as an ortholog (Figure 4.3). This could result in the recovery of paraphyletic relationships between species i.e. violation of the monophyletic clade (eg. violation of the monophyletic relationship between species B and C in Figure 4.3). In order to assess the potential impact of these patterns we used a new software called Clan_Check (Siu-Ting et al. 2019). The central premise of the Clan_Check algorithm is to filter gene families from a dataset based on their ability to retrieve clans (Wilkinson et al. 2007) that are known to be incontestable monophyletic groups. Clans are equivalent to monophyletic groups or clades (present in rooted trees) that are found in an unrooted tree i.e. their monophyly is unquestioned regardless of where the root is placed outside of the group (Wilkinson et al. 2007). Creating gene trees for each gene family, a set of known clans were tested to check whether the species group together to the exclusion of the remaining species. This would suggest that a given clan is not violated, and thus is suitable for subsequent phylogenetic analysis. If a tested clan is violated by the gene family, this suggests that this gene family contains potentially unannotated paralogs genes and may be affecting the resulting species topology.

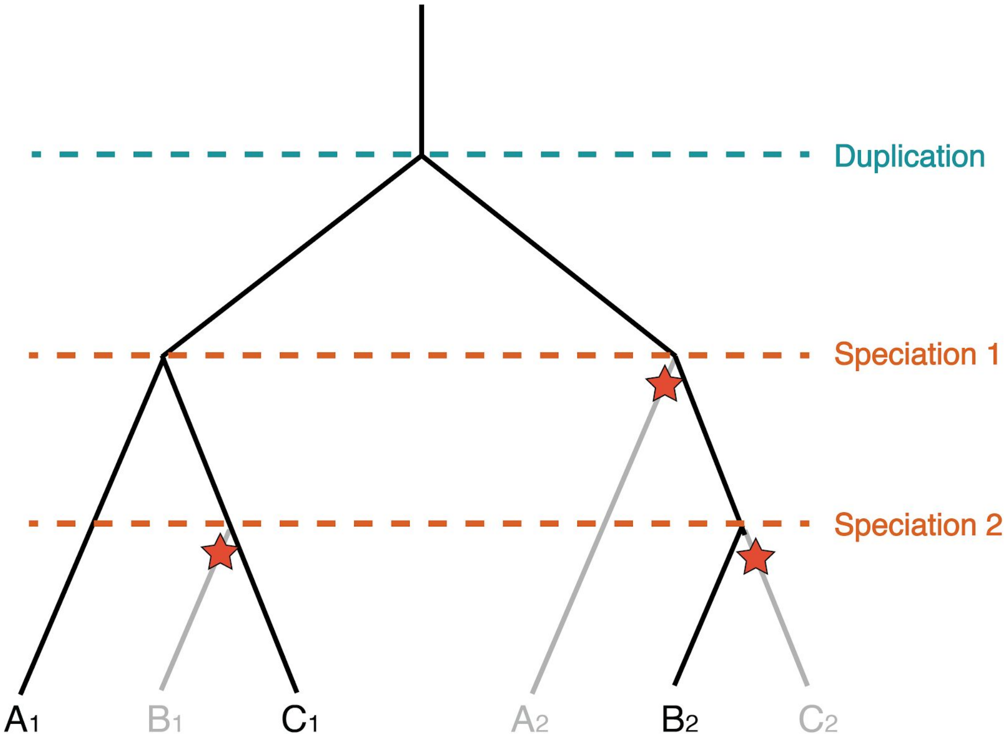


Figure 4.3. Hypothesis of ancient gene duplication and loss patterns resulting in hidden paralogy modified from Siu-Ting et al (2019). Gene tree which represents the evolution of a single example gene in three species (A, B, and C). Ancient duplication of the gene (Duplication) is followed by two rounds of speciation events (Speciation 1 and Speciation 2), resulting in two gene copies in each species eg. species A has gene A1 and A2. Subsequent loss of gene copies are labelled with red stars. This results in a mix of ortholog and paralog genes within the extant species. The species relationships should be given as species B sister to species C, with species A as the outgroup to this clade. However, in this example, the hidden paralog B2 which is a result of the late loss of ortholog and paralog genes causes species A to group with C, to the exclusion of B.

To address the placement of the Xenacoelomorpha we have taken two different approaches using two different data types (previously published phylogenomic datasets and our newly assembled set of composite genes) and a range of methods and compare the resultant topologies: (1) the reanalysis of existing datasets with the removal of paralogous data from gene families and (2) the application of presence absence composite CHG data for 2,285

CHGs from Chapter 2. Consilience across both approaches for the node would be taken as strong support for the phylogenetic placement.

Firstly, in an effort to address the effects of erroneous paralog inclusion on the placement of the Xenocoelomorpha, we assessed three recently published phylogenomic datasets from Rouse et al. (2016), Cannon et al. (2016), and Philippe et al. (2019), all of which aimed to address the position of Xenacoelomorpha within AToL and which produced conflicting results (Table 4.1). Rouse et al. (2016) and Cannon et al. (2016) found support for the Nephrozoa hypothesis (Xenacoelomorpha sister to the remaining bilaterian species), while Philippe et al. (2019) recovered Xenambulacraria (Xenacoelomorpha sister to Ambulacraria clade associated with Deuterostomia). Interestingly, Philippe et al. (2019) also failed to find support for monophyletic Deuterostomia (Xenambulacraria + Chordata). We measured the effects of hidden paralogy on the constituent gene families within each dataset using the software Clan_Check (Siu-Ting et al. 2019), and filtering the gene families to enrich for high quality orthologous genes. Phylogenetic reconstruction was then carried out using these updated gene families.

Secondly, we applied our 2,285 newly assigned composite CHGs of single origin identified in Chapter 2 to resolving the placement of the Xenocoelomorpha. In order to carry out this analysis we post-hoc increased our taxon sampling for this clade of organisms, including the addition of xenacoelomorph species and increased sampling for species from Ambulacraria, and the non-bilaterian clades as outgroups.

Dataset	# Species	# Gene families	Hypothesis
Philippe et al. 2019	59	1,173	Xenambulacraria
Cannon et al. 2016	78	212	Nephrozoa
Rouse et al. 2016	26	1,178	Nephrozoa

Table 4.1. Size of phylogenomic datasets and resulting hypothesis for each published dataset used. For each of the three datasets addressing the position of Xenacoelomorpha (Philippe et al. 2019; Cannon et al. 2016; Rouse et al. 2016), the table shows the number of species and gene families that were used to produce the resulting topology.

4.2 Materials and Methods

4.2.1 Enriching for orthologs by removing hidden paralogy in published datasets

To assess the impact of these patterns of gene evolution on previously published datasets, we employed Clan_Check (https://github.com/ChrisCreevey/clan_check) to infer the impact of hidden paralogy on the resulting topologies (Siu-Ting et al. 2019). Phylogenomic datasets for all three studies were downloaded from online data depositories (github.com/MaxTelford/Xenacoelomorpha2019, (Philippe et al. 2019); datadryad.org/stash/dataset/doi:10.5061/dryad.493b7, (Cannon et al. 2016); datadryad.org/stash/dataset/doi:10.5061/dryad.79dq1, (Rouse et al. 2016)). Each dataset consisted of a concatenated matrix of each of their gene families in amino acid format (see Table 4.1 for details on the number of gene families and species in each dataset). This concatenated matrix was split into constituent gene family fasta files and each one was aligned using Mafft (Katoh et al. 2005). Next, we used IQ-TREE (Nguyen et al. 2015), carrying out 1000 ultrafast bootstrap replicates and automatic model selection, to construct gene trees for all genes families in each dataset. For each dataset, we annotated a number of clans which were to be tested using Clan_Check. These were assigned based on the phylogenetic spread of species within each of the dataset. The clans tested included Porifera, Ctenophora, Cnidaria, Bilateria, Protostomia, Deuterostomia, Xenacoelomorpha, Ambulacraria, Lophotrochozoa, Ecdysozoa, and Chordata (Figure 4.4). A list of the tested clans for each of the individual datasets can be found in “philippe_2019”, “cannon_2016”, and “rouse_2016” directories in Appendix 4.1, under the filename “*clans.txt*”.

Clan_Check was run on the list of gene trees for each dataset using the script (Appendix 4.1, “*clan_check.sh*”). Following this, for each dataset we constructed two subsets of gene families based on (i) if they recovered at least one tested clan, or (ii) if they recovered at least half of the clans tested (carried out using “*find_nonviolate_trees.py*”, Appendix 4.1). This meant for each of the three phylogenomic datasets analysed we produced two subsets of gene families enriched for orthologous genes at different degrees of strictness. In this way, we could have sets of filtered gene families for each dataset, where we could compare the effect of this systematic error in gene family construction on topology and node support. In the analysis of the Philippe et al. (2019) dataset, this resulted in 812 gene families, where at least one clan was not violated (“Philippe_812”), and 65 gene families, where half of the clans tested were not violated (“Philippe_65”). For the Cannon et al. (2016) dataset, we obtained 178 gene families where at least one tested clan was not violated in each gene family (“Cannon_178”). None of the gene trees in the Cannon et al. (2016) dataset were able to recapitulate at least half of the clans, perhaps an indication of the quality of the gene family data, so we obtained a subset of 16 gene families where at least three clans were not violated (“Cannon_16”). Analysis of the Rouse et al. (2016) dataset produced 717 gene families where at least one clan was not violated (“Rouse_717”), and 70 gene families that did not violate half of the clans tested (“Rouse_70”). For each of these six subsetted gene family datasets from each of the three studies, concatenated matrices of amino acid sequences in Phylip format were constructed using scafos (Roure et al. 2007) (Appendix 4.1, “*scafos_concat.sh*”).

4.2.2 Composite dataset construction to expand taxon sampling

In constructing our initial dataset to search for composite CHGs in Chapter 2, we did not include any species from the Xenacoelomorpha, and undersampled from other clades such as Xenambulacraria. This was due to strict thresholds on genome quality to ensure high confidence in our approach for identifying composite CHGs should they exist. In order to address the placement of Xenacoelomorpha using composite genes as phylogenetic markers, we thus needed to expand our dataset to include some of these key species. Using BLASTp (Altschul et al. 1990), (E-value <= 1e-5, percent identity >= 30%, coverage >= 80%) we used our composite genes to search for homologous genes in representative genomes. This was carried out using six species from Xenacoelomorpha (*Xenoturbella bocki*, *Symsagittifera roscoffensis*, *Praesagittifera naikaiensis*, *Hofstenia miamia*, *Nemertoderma westbladi*, and *Meara stichopi*) , three additional species from Ambulacraria (*Saccoglossus kowalevskii*, *Acanthaster planci*, and *Apostichopus japonicus*), and three additional non-bilaterian species (two sponges; *Leucosolenia complicata* and *Sycon ciliatum*, and one cnidarian; *Hydra magnipapillata*). Homology searching for composite genes in the additional species was carried out as in Sections 3.2.4, whereby the coverage threshold is set at 80% of the query and subject sequences and coverage across the composite breakpoint regions is required, to ensure confidence in the set of composite genes we find.

4.2.3 Tree inference

For each of the six resulting dataset following assessment of hidden paralogy using Clan_Check (“Philippe_812”, “Philippe_65”, “Cannon_178”, “Cannon_16”, “Rouse_717”, and “Rouse_70”) phylogenetic reconstruction of all supermatrices was carried out using PhyloBayes-MPI (Lartillot et al. 2013). After constant sites were removed (-dc option) the

CAT+GTR model was applied, along a gamma distribution consisting of four rate categories. Two independent chains were run until convergence between the runs was reached. Convergence between chains was assessed using the bpcomp function in PhyloBayes, with a score of 0.1 indicating convergence.

4.2.4 Tree inference using taxonomically-expanded composite CHG data

For phylogenetic inference of the composite CHG dataset, presence/absence matrices were first created as in Section 3.2.2. This included the additional species from Xenacoelomorpha, Ambulacraria, and non-bilaterian clades which were added to our datasets using BLASTp as described in 4.2.2. Given the large number of species in this expanded dataset, which consisted of 75 species, we removed certain lineages that were considered to be fast evolving based on analyses carried out in Chapter 3. The species we removed were *Mnemiopsis leidyi* from Ctenophora, and *Schistosoma mansoni* from Lophotrochozoa. We also filtered the dataset to contain a more evenly distributed number of species in all major clades: 8 chordate species, 4 ambulacrarians, 6 xenacoelomorphs, 7 ecdysozoans, 3 lophotrochozoans, and 5 non-bilaterian species (2 cnidarian and 3 porifera species). This resulted in a dataset consisting of 33 species. Reducing the dataset in this way would provide a more evenly distributed sampling of species across the tree, but also reduce running time. Trees were then inferred using biphy (Pett et al. 2019), which employs RevBayes (Höhna et al. 2016). This was carried out on the set of composite CHGs annotated as having a single node of origin by the character mapping analysis in RevBayes (2,285 CHGs) (see Section 2.3.2). This dataset was found to reproduce the most accurate phylogeny in analyses carried out in Chapter 3. Therefore, this set of characters were analysed under the Dollo model, which allows secondary loss but not multiple gains, to infer the species tree. The script (Appendix 4.2, "*biphy_dollo.sh*") was used to run the phylogenetic analysis.

4.3 Results

4.3.1 High numbers of violated clans in all previously published datasets

For each of the three datasets (Table 4.1) we ran Clan_Check (Siu-Ting et al. 2019) to assess whether each of the constructed gene trees recapitulated known monophyletic clades. The monophyletic clades or “clans” tested were as follows: Porifera, Ctenophora, Cnidaria, Bilateria, Protostomia, Deuterostomia, Xenacoelomorpha, Ambulacraria, Lophotrochozoa, Ecdysozoa, and Chordata (Figure 4.4 D). We found that for each of the three datasets the majority of the gene trees violated most of the clans tested (Figure 4.4).

In the Philippe et al. (2019) dataset, each clan was violated by between 60-99% of the gene trees (Figure 4.4 A). The clan which was violated most often was Deuterostomia, with 1,155/1,164 (99%) of the gene trees failing to recapitulate the clan. This is of particular interest as from their original study, using this supermatrix alignment, Philippe et al. (2019) failed to recover a monophyletic Deuterostomia. The Cannon et al. (2016) dataset revealed more variable patterns in the violation of clans (Figure 4.4 B), e.g. Lophotrochozoa, Xenacoelomorpha, Protostomia, and Bilateria were annotated as monophyletic in all the gene trees in the dataset and Deuterostomia was violated by almost all (98%) of the gene trees tested. It is important to note that the number of gene families in Cannon et al. (2016) study was much lower (212 gene families) when compared to the other two datasets (1,173 and 1,178 gene families). Finally, analysis of the Rouse et al. (2016) dataset revealed more similar patterns to the results observed in the Philippe et al. (2019) analysis. Again, Deuterostomia was the most frequently violated clan with 97% of gene trees not recapitulating this monophyly , followed closely by Lophotrochozoa (91%) and Protostomia (91%) (Figure 4.4 C).

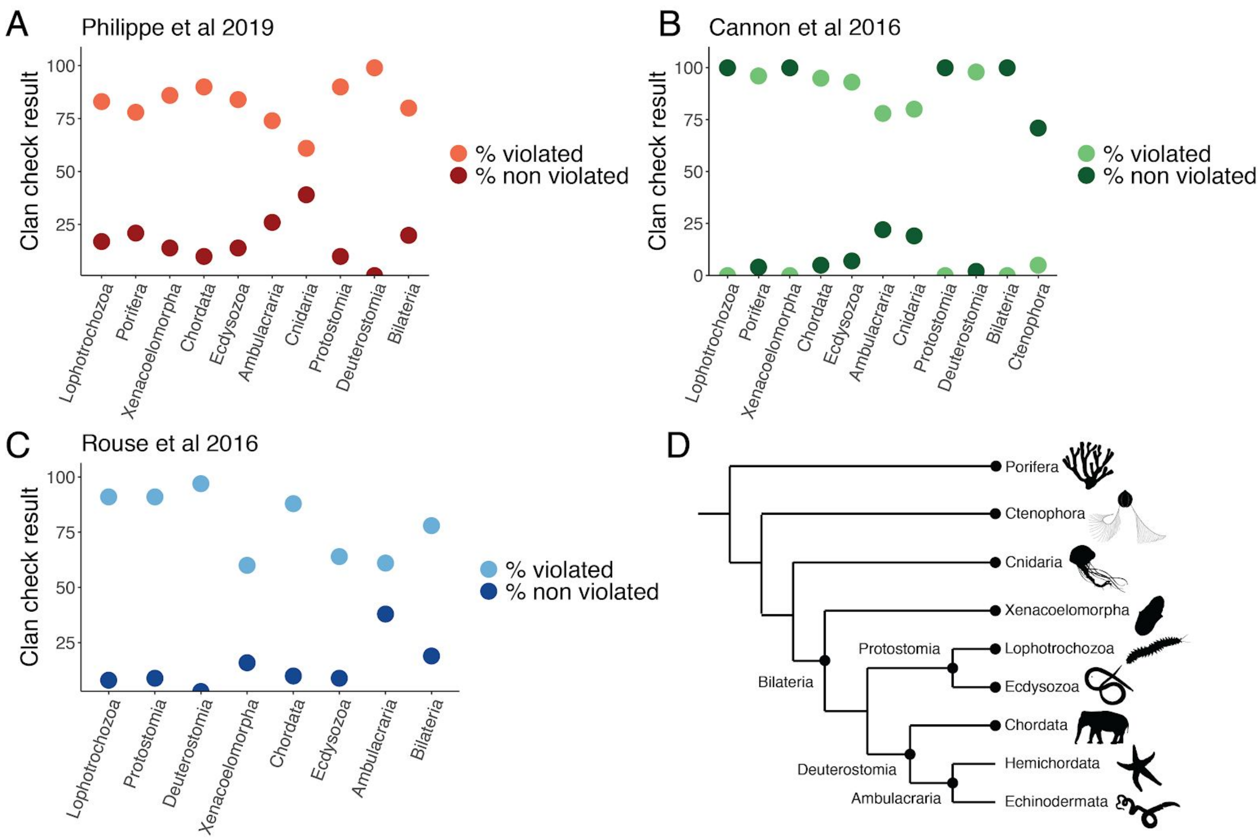


Figure 4.4. Results of Clan_Check analysis on each of the three datasets. The level of violation (or otherwise) (on the y axis) of each of the (8-11) uncontested monophyletic clans (given on the X axis) for each dataset (A-C) are shown. The percent of gene families that violated (lighter datapoint) or did not violate (darker datapoint) each tested clan is presented for each dataset. The complete set of 11 clans tested across the three datasets are shown in (D), with the tested clans labelled with a black dot and the name of the clan beside it.

4.3.2 Phylogenetic analysis of datasets enriched for orthologs

Using our filtered datasets enriched for gene families consisting of orthologous genes, we carried out phylogenomic analysis to assess the placement of Xenacoelomorpha. We wished to determine what impact applying Clan_Check had on the resultant topology and node support. First, we analysed the two filtered datasets from Philippe et al. (2019), i.e. the 812 gene families ("Philippe_812") and the 65 gene families ("Philippe_65") using the CATGTR model in PhyloBayes (Lartillot et al. 2013). Both datasets produced the same topologies for the major clades, however the "Philippe_812" dataset did not reach convergence and therefore we will focus on the results from the "Philippe_65" dataset of genes. A monophyletic Xenambulacraria was recovered with the "Philippe_65" dataset with PP = 1, and a grouping of Chordata and Protostomia again with a PP score of 1. (Figure 4.5). In the original work Philippe et al. (2019) proposed that Deuterostomia is non-monophyletic although their support values were low, our filtering of their data for orthologous genes has also recovered this relationship but with full support (PP=1). The proposal that Deuterostomia is paraphyletic and that Xenacoelomorpha are related to Ambulacraria is also supported by this analysis (PP=1) (Philippe et al. 2011; Philippe et al. 2019). Although the grouping of Chordata and Protostomia received low support from our analysis of the "Philippe_812" dataset (which did not reach convergence), it seems that reanalysing this dataset and enriching for orthologous gene families supports Xenambulacraria as a clade (PP=1).

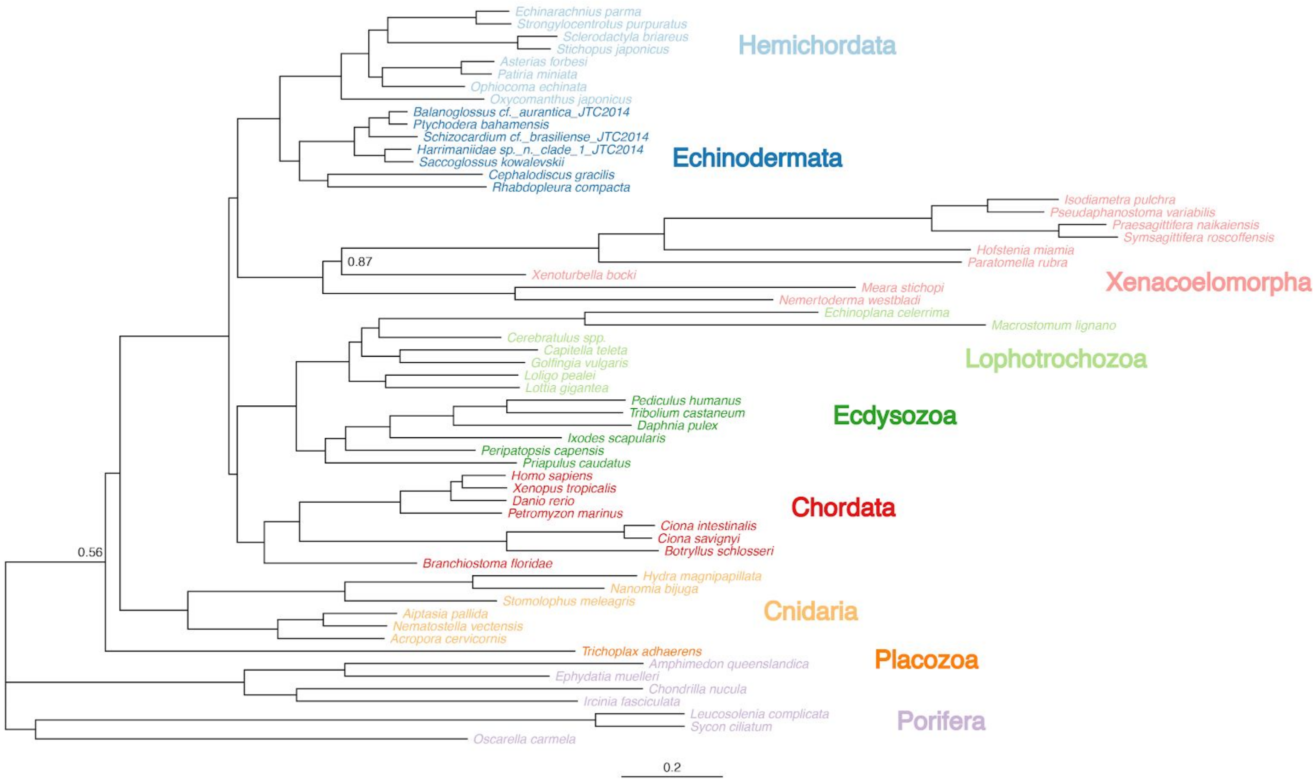


Figure 4.5. Rooted species tree obtained from the ortholog enriched "Philippe_65" dataset using the CATGTR model. Major clades are labelled by colour, with corresponding colour coordinated clade name beside the tips. Branch support showing Bayesian posterior probabilities (PP) is labelled on nodes with PP <1.

Next, we assessed the subset of genes enriched for orthologs from the Cannon et al. (2016) dataset. As there were a smaller number of gene families in this initial dataset compared to the other datasets (see Table 4.1), there were fewer number of families that passed our Clan_Check filters, and none passed the stricter filter of at least half of the clans tested recovered as non-violated. This allowed us to analyse datasets where at least one clan was not violated: 178 gene families ("Cannon_178"), and where at least three clans were not violated: 16 gene families ("Cannon_16"). Both datasets produced similar topologies in terms of the major branching orders, however the larger "Cannon_178" dataset did not reach convergence (0.49). Nonetheless, both analyses recovered the Nephrozoa hypothesis, i.e. Xenacoelomorpha were placed as the primary emerging bliaterian lineage (Figure 4.6). This result is concordant with the findings of the original analysis (Cannon et al. 2016). In our analysis of the Cannon_16 dataset, the Deuterostome monophyly is recovered with low support (PP=0.81), although it groups Chordata as sister lineage to a clade consisting of Ambulacraria + Protostomia (Figure 4.6). The Cannon_178 dataset, recovers monophyletic Deuterostomia, with a clade of Chordata + Ambulacraria sister to Protostomia, with full support (PP=1.0).

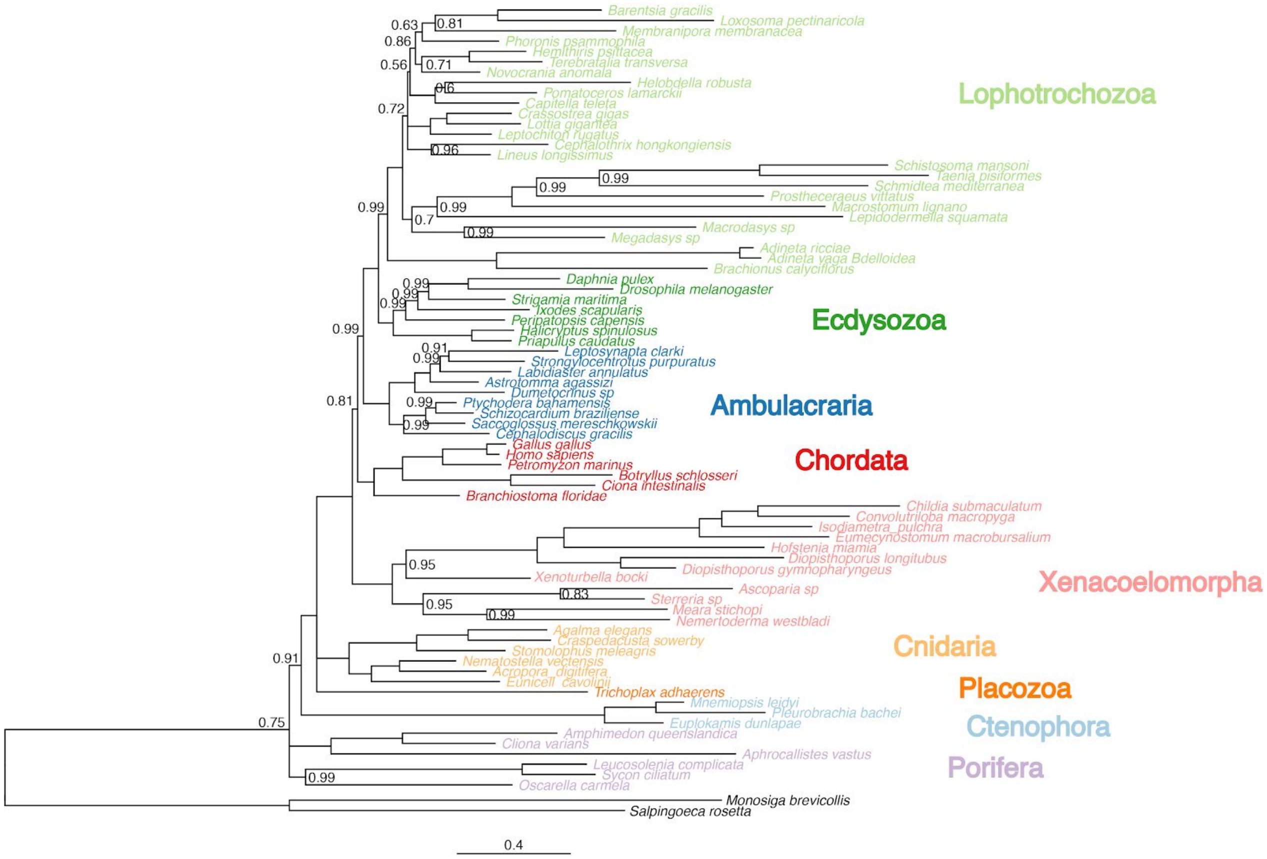


Figure 4.6. Phylogenetic tree from the "Cannon_16" dataset reconstructed using the CATGTR model. Major clades are labelled by colour. Branch support showing Bayesian posterior probabilities (PP) is labelled on nodes with PP <1.

Finally, the Rouse et al. (2016) dataset was assessed following filtering of gene families that may be a source of bias causing incorrect topologies. This resulted in two datasets of 717 and 70 gene families, i.e. “Rouse_717”, and “Rouse_70” respectively. Both analyses recovered Xenambulacraria (PP=0.94 in “Rouse_70” and PP=0.5 “Rouse_717”). “Rouse_70” recovered Xenambulacraria (PP=0.94) and a monophyletic Deuterostomia, i.e. Xenambulacraria + Chordata, albeit with lower support (PP = 0.7) (Figure 4.7). The “Rouse_717” dataset did not fully converge (convergence score of 1), but did group Xenambulacraria and Protostomia together, with Chordata sister to this clade, although with low support (PP=0.5). These findings conflict with the original study, where Xenacoelomorpha was sister to the remaining bilaterian lineages (Rouse et al. 2016), and instead supports grouping of Xenacoelomorpha within Deuterostomia, as found in our previous analysis and in other analyses (Philippe et al. 2011; Philippe et al. 2019).

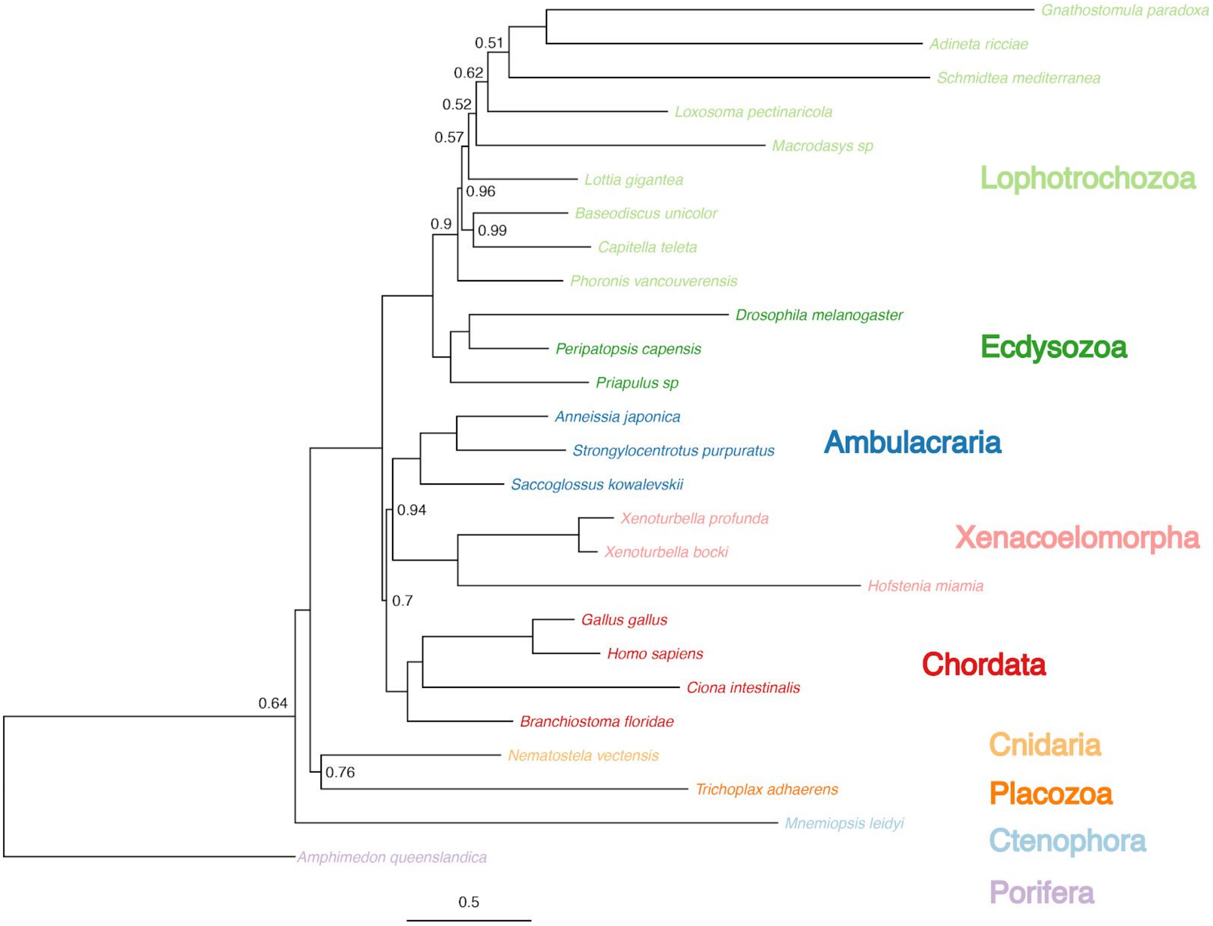


Figure 4.7. Phylogenetic tree from Rouse_70 dataset using the CATGTR model. Xenacoelomorpha are grouped with Ambulacraria, with this clade sister to Chordata. Major clades are coloured per lineage, and node support, where PP is below 1, is annotated on the tree.

To summarise our results from assessment of filtered subsets of previously published datasets: Comparing the resultant topologies from each of the six subsets of Rouse et al. (2016), Cannon et al. (2016) and Philippe et al. (2019) datasets with each other and with the original topologies conferred from the unfiltered original datasets, we find varying consensus (FIgure 4.8). Both the “Philippe_812” (which did not reach convergence) and the “Philippe_65” datasets produced the same topology (Xenambulacraria and paraphyletic Deuterostomia) as one another (Figure 4.8 A and B). This was consistent with the original findings of the study (Philippe et al. 2019). The “Cannon_717” and “Cannon_70” datasets also resulted in the same topology as one another (Xenacoelomorpha sister to Nephrozoa), which was also the same as the topology observed in the initial study (Figure 4.8 C and D) (Cannon et al. 2016). However, there were differences in the two subset datasets, with “Cannon_178” recovering monophyletic Deuterostomia (Chordata + Ambulacraria), and “Cannon_16” recovering paraphyletic Deuterostomia (Protostomia + Ambulacraria). Finally, “Rouse_717” and “Rouse_70” both recovered a clade of Xenambulacraria (Xenacoelomorpha + Ambulacraria), which directly conflicts with the findings of the original study which recovered the Nephrozoa hypothesis (Figure 4.8 E and F). The position of Deuterostomia differed across “Rouse_717” and “Rouse_70”, with “Rouse_717” recovering paraphyletic Deuterostomia and “Rouse_70” recovering monophyletic Deuterostomia.

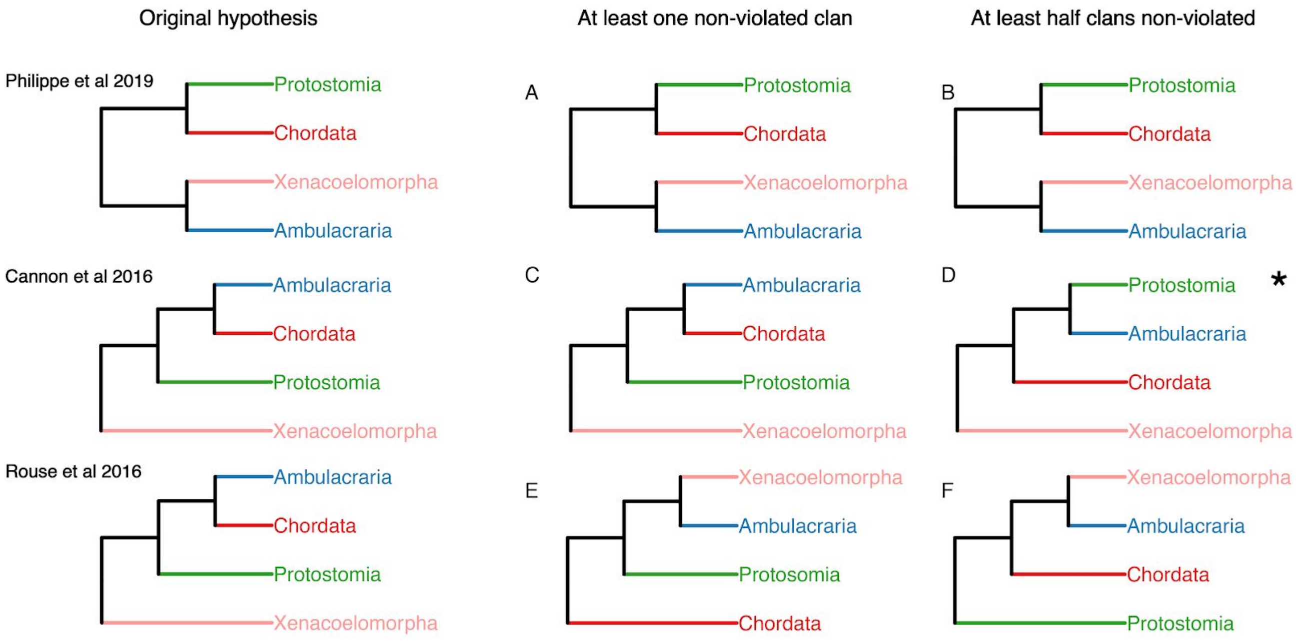


Figure 4.8. Comparison of three topologies from original studies with the two ortholog enriched topologies for each. Each row displays final tree topologies for each dataset analysed in this chapter. The topology resulting from the original study is shown on the left, with the filtered and analysed datasets shown to the right. These show the branching orders between the major bilaterian clades Xenacoelomorpha, Ambulacraria, Chordata, and Protostomia resulting from: (A) “Philippe_812” dataset, (B) “Philippe_65”, (C) “Cannon_178”, (D) “Cannon_16”, (E) “Rouse_717”, and (F) “Rouse_70”. * dataset in which we were unable to get any gene trees that recapitulated at least half of the tested clans (as denoted by the column name), so instead used those that recapitulated at least three clans.

4.3.3 Composite genes as phylogenetic markers

We analysed our set of composite CHGs, annotated as single gain events, with expanded taxon sampling (presenceAbsence_xenacoel_singleGain_remFast_data.fasta, Appendix 4.2). In Chapter 3, we found that the datasets that were filtered for homoplastic characters recovered more species topologies with higher support values and fewer polytomies, based on the known branching patterns within the animal tree. Therefore, here we assessed only the set of composite CHGs that were annotated as having a node of single origin by the character mapping analysis in Chapter 2 (i.e. 2,285 CHGs). The resulting species topology following tree inference using the Dollo model in RevBayes (Höhna et al. 2016) for the 2,285 composite CHGs across 33 species, was generally poorly resolved with PP values in the range 0.57 to 1 (Figure 4.9). A large number of polytomies and branching patterns with low support (i.e. PP< 70) suggesting this specific dataset may not be informative. Nonetheless, some resolution was found within the tree. The Ecdysozoa and Lophotrochozoa (together forming Protostomia) clades grouped together with high support, although contained a number of polytomies. Additionally, the cnidarian *Nematostella vectensis* was erroneously placed within the group of Protostomia. Xenacoelomorpha formed a clade with Ambulacraria and Chordata, however, with low support (PP=0.58) and represented mostly by polytomies. This clade also incorrectly includes the cnidarian *Hydra magnipapillata*.

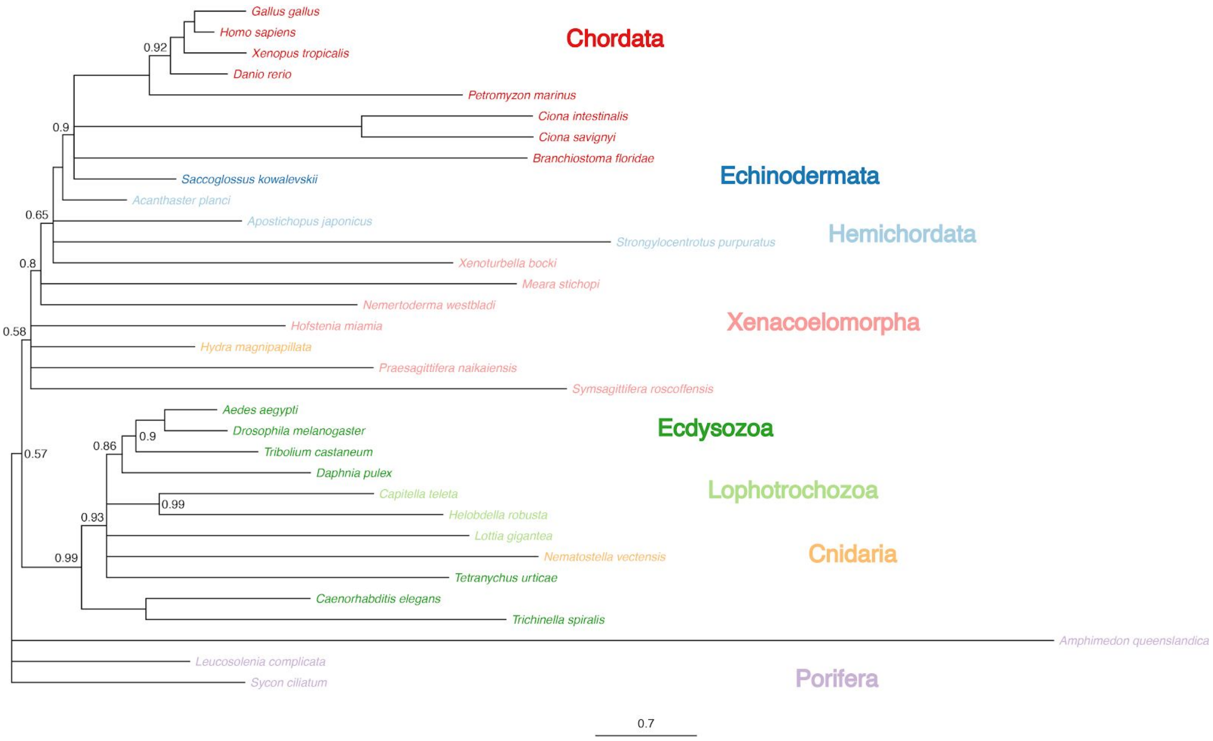


Figure 4.9. Tree inference using composite CHG expanded to address placement of Xenacoelomorpha. Species present in this tree were filtered so that each clade had similar numbers of taxa. Polytomies represent unresolved regions within the tree. PP scores below 1 are shown.

4.4 Discussion

In this chapter we have assessed the position of Xenacoelomorpha in the animal tree. We analysed the effects of gene family construction and ortholog assignment on the resultant topology, these are two steps which can vary greatly between datasets and rarely assessed for potential biases. We also applied a new data type, composite genes, to reconstructing the position of the Xenacoelomorpha within Bilateria.

The difficulty in placing this group of enigmatic worms stems from a number of evolutionary characteristics. For example, the morphological simplicity of Xenacoelomorpha leads to the parsimonious assumption that they are the primary emerging bilaterian phylum, indicating a morphologically simple intermediate state between Cnidaria and the more complex Nephrozoa lineages (Hejnol and Martindale 2008). This position would also go some way to explain microRNAs and Hox genes shared across bilaterian species that are missing in Xenacoelomorpha (Sempere et al. 2007; Cook et al. 2004). If, however, Xenacoelomorpha are sister phyla to Ambulacraria, this would suggest that similar morphological and genomic traits would group these two lineages. We do not find any shared traits such as a common composite gene that is present in these two groups only and not any of the other animal lineages. Similarly, there is little support found from other synapomorphic traits, and the discovery of any more such traits would provide further support for the secondary loss of complexity in Xenacoelomorpha and their positioning associated with Deuterostomia.

Short internal branches, indicating short divergence times, means that signal within the sequence data may be difficult to separate from the noise, thus making the resolution of the major clades within Bilateria a difficult and ongoing issue. Additionally, the fast rates of sequence evolution with the Acoelomorpha clade, and potential compositional heterogeneity mean that modelling the evolutionary events that followed the radiation of the bilaterian clade is difficult. Applying site heterogeneous models and accounting for compositional bias by data re-coding has found some congruence across datasets (Philippe et al. 2019). However, it is possible that better modeling, for example models accounting for compositional heterogeneity and rate heterogeneity simultaneously across the phylogeny as applied in P4 (Foster 2004), could provide a more accurate picture of the processes of sequence evolution at this part of the tree.

In this chapter, our first analysis applied Clan_Check to search for hidden paralogy within the gene families from three previously published phylogenomic studies (Rouse et al. 2016; Cannon et al. 2016; Philippe et al. 2019). We found that all datasets contained significant proportions of gene families which violated uncontroversial monophyletic clans. However, levels of violation within the gene families did vary between datasets. While the Philippe et al. (2019) and Rouse et al. (2016) datasets showed consistent patterns of violation across the gene families, the Cannon et al. (2016) dataset contained a number of clans that were recapitulated by almost all of the gene trees. Nonetheless, the pervasive violation of clans across all datasets suggest that the constructed orthologous gene families may contain a high number of hidden paralogous genes. In all three datasets Deuterostomia was the clan that was violated by the most gene families. These findings not only provide insight into issues surrounding methods to infer orthologous gene families, but also give interesting insights into the nature of gene evolution in the Deuterostomia clade. The fact that the Deuterostomia is violated in almost all gene trees in each dataset may provide support for non-monophyletic Deuterostomia. Alternatively, this consistent violation of Deuterostomia could be a reflection of the nature of gene evolution within this clade. With high rates of gene loss known to have occurred at this node (Fernández and Gabaldón 2020), loss of the ancestral ortholog and retention of the paralogous gene in the extant species could be common.

The methods of constructing orthologous gene families varied in each of the three datasets. Philippe et al. (2019) applied the OMA standalone software to construct orthologous groups (Altenhoff et al. 2019), while Cannon et al. (2016) used HaMStR (Ebersberger et al. 2009), and Rouse et al. (2016) applied the Agalma software (Dunn et al. 2013). While there is no perfect piece of software to infer true orthology (see Section 1.1.3.2) and arguments can be made to justify the use of any one, it has been shown that the construction of orthologous gene families using particular algorithms can have an effect on the resulting species topology. Altenhoff et al. (2019) addressed this impact of ortholog selection on the difficult to resolve, deep clade of Lophotrochozoa, by reconstructing the phylogeny using datasets created by OMA, HaMStR, OrthoMCL, OrthoFinder, and BUSCO. They found that each software resulted in differing numbers of orthologous groups, with varying numbers of species per group (OMA resulted in 2,162 orthologous groups, while HaMStR produced 1,241 orthologous groups). Significantly, phylogenomic reconstruction on a supermatrix using both maximum likelihood (ML) and Bayesian analyses resulted in different topologies with different support values. Comparing OMA (Altenhoff et al. 2019) and HaMStR (Ebersberger et al. 2009) methods found marked differences in topology, while both analyses showed high support across the tree. The OMA (Altenhoff et al. 2019) dataset produced a Lophotrochozoan phylogeny closer to that of previously published datasets, while the HaMStR (Ebersberger et al. 2009) phylogeny erroneously grouped species in both the ML and Bayesian analyses (however, it is important to note that this analysis did not reach convergence). This important piece of analysis, and others which addressed the yeast phylogeny (Shen et al. 2018) found that deep, difficult to resolve topologies are often sensitive to the method of orthologous gene family construction. This could certainly be the case with the different topologies resulting from the three different studies addressing the position of Xenacoelomorpha, and further analysis comparing the effects of different methods for ortholog inference on the resulting topology may provide significant insight into this difficult to resolve question.

We attempted to address some of these issues of ortholog assignment by filtering each dataset based on the ability of each gene family to recover known clans, retaining only gene families that (i) did not violate at least one tested clan or (ii) did not violate at least half of the clans tested (in the case of the Cannon et al. (2016) dataset, there were no gene families that did not violated half of the tested clans, and we took those that did not violate at least three clans). While this approach does not ensure all paralogous genes will be filtered or that all orthologous genes will be identified, it provides a novel method to enrich for orthologous gene families and provide a more accurate dataset from which to infer the species tree (Siu-Ting et al. 2019). We found increased, but not consistent, support for the grouping of Xenacoelomorpha with Ambulacraria (Xenambulacraria). Two of the three analyses recovered this clade, with the filtered Philippe et al. (2019) datasets recovering their initial findings but with higher node support, and both filtered Rouse et al. (2016) datasets differing from the topology of the initial analysis which recovered the Nephrozoa hypothesis (Figure 4.8). Reanalysis of the Cannon et al. (2016) dataset recovered the same hypothesis as the initial analysis with Xenacoelomorpha as the primary emerging bilaterian lineage (Nephrozoa hypothesis) (Figure 4.8). However, a criticism of this dataset has been the use of species representing Ctenophora. These lineages are known to be fast evolving, and along with the fast evolving Acoelomorpha phylum within Xenacoelomorpha, the sister relationship between Xenacoelomorpha and the remaining bilaterian clades has been hypothesised to be a result of LBA. To test this, we removed all Acoelomorpha species (11 in total) from the Cannon_16 dataset, and used the Xenoturbella species Xenoturbella bocki as representative of the phylum. This was possible as we know Xenacoelomorpha is a monophyletic clade, thus, removing the fastest evolving lineages within a group and using representative members may aid in limiting the effects of LBA (Aguinaldo et al. 1997). When we reanalyse this set of taxon filtered orthologous gene families, we indeed find that Xenacoelomorpha forms a monophyletic clade with Ambulacraria, with this clade placed sister to Protostomia, and Chordata placed as sister to all other bilaterian clades. This is congruent with similar analysis carried out by Philippe et al. (2019) where removing the fast evolving Acoelomorpha clade recovered Xenambulacraria.

We also fail to confirm the existence of a monophyletic Deuterostomia. Both filtered datasets from the Philippe et al. (2019) study recover non-monophyletic Deuterostomia with full support. The Cannon et al. (2016) and Rouse et al. (2016) filtered datasets differ in their reconstruction of Deuterostomia. The grouping of Xenambulacraria with chordata*, i.e.*

monophyletic Deuterostomia ("Rouse_70" dataset), and the grouping of Ambulacraria with Protostomia i.e. non-monophyletic Deuterostomia ("Cannon_16") are both recovered with relatively lower support (PP=0.7 and PP=0.81 respectively). However, with only 2/6 subset reanalyses supporting monophyletic Deuterostomia (Figure 4.8), and increased support for non-monophyletic Deuterostomia in the reanalysis of the Philippe et al. (2019), this goes some way to providing some support for the hypothesis that Deuterostomia are paraphyletic.

Our analysis of composite genes to resolve this region of the animal tree did not reach a conclusive result. All analyses produced trees that were poorly supported and contained a high number of polytomies. This could be due to a number of factors, such as the lack of enough shared composite genes between all of these diverse lineages, or the ex post facto introduction of more species using BLASTp, which may not capture the relevant complex patterns of composite genes in the added species. Alternatively, despite our efforts to use highest quality genomes in our analysis, annotation errors could be impacting on the analysis within these groups of particular species. Nonetheless, we do find a grouping of some Xenoacoelomorph species with Ambulacraria and Chordata, suggesting a closer affinity of this clade with other Deuterostome clades (Figure 4.9).

Our findings provide more support for the affinity of Xenacoelomorpha with Ambulacraria. This goes against parsimonious assumptions, and a large number of phylogenetic studies (Ruiz-Trillo et al. 2004; Paps et al. 2009; Ruiz-Trillo and Paps 2016; Rouse et al. 2016; Cannon et al. 2016), which suggest that Xenacoelomorphs are early diverging bilaterians. We have found that gene family construction, and the accurate prediction of orthologous genes, represents a major topic to be addressed to answer this question. Indeed, the majority of phylogenomic studies often use just one approach for gene family construction, and this could be affecting the resolution of other parts of the animal tree such as the root of animals. While our analyses were not conclusive, overall we provide more evidence for the secondary simplification of Xenacoelomorpha placing them as the sister clade to Ambulacraria, and question the existence of a monophyletic Deuterostomia.

4.5 Conclusion

Given the limited genomic sampling for species involved in resolving the position of Xenacoelomorpha, it is unquestionable that the underlying assumptions of orthology across this diverse group of species is being severely violated (Figure 4.4). Additionally, use of different methods for the construction of orthologous groups for phylogenetic reconstruction has caused some intrinsic bias between these studies, resulting in incongruent but highly supported topologies. This bias has generally been overlooked in favour of discussions surrounding model selection and parameters and data issues such as missing data, compositional heterogeneity, and incomplete taxon sampling. While addressing these issues surrounding hidden paralogy does not provide consensus across all three studies, we do find increased support for the existence of the Xenambulacraria clade (Figure 4.8 A, B, E, and F), and increased branch support for a non-monophyletic Deuterostomia. These preliminary findings suggest that Xenambulacraria may be supported by the molecular data, but additional work on testing the effects of different methods to create orthologous groups, increased taxon sampling and higher quality genomes, and the application of models that could incorporate between lineage heterogeneity may provide better insight into the evolution of Xenacoelomorpha. Additionally, creation of a set of composite CHGs addressing specifically the species of focus (such as sampling of available Xenacoelomorpha genomes, and increased sampling of Ambulacraria species) may provide more accurate markers for the resolution of the placement of the Xenacoelomorpha clade.

Chapter 5: Discussion

The underlying genetic mechanisms driving diversification and evolution within animals are numerous and complex. Since the sequencing of the first animal genome in 2000 (Adams et al. 2000), our understanding of animal evolution and AToL has evolved rapidly. We now know that the genomic content governing 'what is an animal?' is far more complex than just the protein coding content we observe in extant animals (Paps 2018). How animals became so diverse is a result of novelty, loss, regulation, and shuffling within the genome, in response to external factors such as environmental changes and population effects. As the field of animal comparative genomics and phylogenetics continues to explore the abundance of high quality sequenced genomes, it is essential that we take a holistic view of the mechanisms and processes of molecular evolution.

In Chapter 2, we addressed one mechanism driving new gene genesis in animals. A large number of studies have assessed the rates of gene gain and loss across animal lineages, including mammals (Dunwell et al. 2017), primates (Hahn et al. 2007), vertebrates (Blomme et al. 2006), Drosophila (Zhou et al. 2008), Lophotrochozoa (Luo et al. 2017), and Arthropoda (Thomas et al. 2020), in addition to gene gain on the animal stem lineage (Paps and Holland 2018; Richter et al. 2018), and more recently in two studies which address the evolution of gene content across all phyla in the animal tree (Fernández and Gabaldón 2020; Guijarro-Clarke et al. 2020). However, each of these studies assume tree-like processes of gene evolution, whereby new genes emerge through duplication followed by neofunctionalization, or by *de novo* emergence from non-coding sequence. Current software limitations mean that these studies fail to accommodate non-tree-like processes such as gene remodeling by gene fusion and fission in their definition of gene families. A wealth of studies into domain and protein evolution have shown that gene remodeling has indeed played a major role in the evolution of novel traits, and has contributed to the rapid expansion of protein complexity within animals (Tordai et al. 2005; Itoh et al. 2007; Kaessmann 2010; Zmasek and Godzik 2011). As described in Chapter 2 we measured the rates of gene remodeling at the sequence level across entire genomes and established a set of high confidence composite clusters of homologous groups (CHGs) corresponding to 13,632 CHGs, which following survey of RNAseq data had evidence of transcription.

With these newly identified composite families, we set out to understand their patterns of gain and loss and elucidate their rates of evolution at different points in the animal tree. What we discovered showed striking patterns of increased rates of remodeling events at particular points in the animal tree, and uneven rates of composite gene gain and loss. By far the largest number of gene remodeling events occur at the base of boney vertebrates (Euteleostomi). As this node represents a period of major transition in morphology and rapid species diversification, with emerging complex traits such as mineralised bone and the adaptive immune system. Composite gene formation may have played a role in the emergence of these morphological traits. This period of time also coincided with large scale genomic changes, namely genome duplication events, followed by large-scale chromosomal rearrangements and fusions (Sacerdot et al. 2018; Simakov et al. 2020). These changes in genomic structure and subsequent rearrangements in content suggest that selective pressures could have been relaxed allowing for greater shuffling and remodeling of the protein coding genes. Other regions in the animal tree which were found to have high numbers of gene remodeling events included the Diptera and Nematoda lineages. While these younger clades perhaps possess less diversity in morphological complexity, at least compared to Euteleostomi, they are incredibly speciose (Nematoda is predicted to have over a million species (Lambshead and Boucher 2003; Blaxter 2011)) and have been found to contain a large amount of novelty with relation to gene family content (Prabh et al. 2018). Additionally, relative to other animal lineages, these groups of organisms are known to have higher rates of recombination (Stapley et al. 2017), which could facilitate higher rates of gene remodeling relative to other animal clades. Further investigation into the role and rate of gene remodeling in these lineages, especially within the Caenorhabditis lineage (which displayed the second highest number of gene remodeling events at an internal node) is warranted. Overall, while new gene genesis by tree-like processes still constitute the majority of genetic novelty in animals, gene remodeling has undoubtedly contributed significantly to the animal protein repertoire, particularly at points of major phenotypic transition.

The impact of remodeling on the animal functional proteome is likely significant. We found that composite genes tend to be larger and contain more protein domains, suggesting that they potentially possess a larger range of function than non-composite genes. This also suggests that remodeling may play an important role in driving the evolution of complexity on the phenotypic level. While inferring function of composite genes is inherently complicated, as function is often annotated by sequence similarity, we do find that searching for broad functional categories shows that most composite genes carry out signalling and other cellular processing functions. This is in line with the evolution of phenotypic complexity such as more cell types and complex tissue systems we find in a wide array across the animal tree.

In placing these gene remodeling events onto the species tree, some striking patterns of gain and loss were immediately evident, with high rates of secondary loss and the potential independent emergence of the same composite genes independently in different lineages. These patterns have significant implications for this thesis, but also have a wider impact on our understanding of gene family evolution. In Chapter 3 we sought to further assess the rates of homoplasy within these composite genes, and tease out the rates of secondary loss and parallel evolution of these genes. Using multiple methods to characterise each composite CHG we could group most of the composite CHGs into either multi-gain or single-gain events. Across each analysis we found a high incidence of the same composite gene emerging multiple times independently. This has clear implications for our study applying composite genes as phylogenetic markers. However, when we take the set of composite CHGs that fall into the single-gain set, we recover uncontentious nodes within the animal phylogeny with high support. With the single-gain subset of composite CHGs (i.e. those composite CHGs that do not have evidence of homoplasy), our analysis supports the Porifera-sister hypothesis. While this result is intriguing and suggests that filtered CHGs may have an important contribution to make to phylogenomics and to early animal evolution - we would also see it necessary to increase taxon sampling, particularly surrounding the root of the animal tree and of outgroup species.

In order to accurately analyse patterns of gene evolution, a well resolved species tree is required. However, there are some key parts of AToL that remain unresolved. Untangling these major branching patterns within AToL, such as the primary emerging animal and bilaterian lineages, is essential for understanding the processes of major transition events in animals, gene family evolution, and other aspects such as speciation and extinction. In Chapters 3 and 4 of this thesis we use a data driven approach to contribute to the resolution of some of these difficult to resolve regions. Using patterns of presence and absence of composite CHGs of single origin supported the placement of Porifera as the primary emerging animal lineage. In Chapter 4 we took the application of "single gain events" to phylogeny reconstruction a little further, this time addressing the position of the Xenacoelomorpha within AToL. In this Chapter we look for consilience across single-gain composite CHGs and previously published phylogenomic datasets which had produced conflicting resulting topologies (Rouse et al. 2016; Cannon et al. 2016; Philippe et al. 2019). Ortholog assignment and the construction of orthologous gene families is an often overlooked part of the phylogenomic workflow. While the importance of accuracy in defining orthology within a set of genes has been discussed extensively in the literature (Gabaldón 2008; Salichos and Rokas 2011; Springer and Gatesy 2018), most phylogenomic studies often just use a single construction method, and do not test the validity of the gene families. The penalty of misidentified orthology is large, and results in misleading topologies. We tested the effects of hidden paralogy present in orthologous gene families from previously published phylogenomic datasets on the resulting species topology using the newly published algorithm "Clan_Check" (Siu-Ting et al. 2019). Using the original gene families from each of the three publications, we found pervasive violation of known monophyletic clans across all three datasets indicating problematic orthology assignment. In each case the original studies had used only one approach to identify orthologous gene families, with the methods differing in each study. Enriching for orthologous genes in each of these datasets provided a better picture of the patterns of evolution at the root of Bilateria. While our enriched ortholog datasets recovered the same topologies for two out of the three studies (i.e. we recovered Nephrozoa hypothesis originally found by Cannon (et al. 2016), and the Xemabulacraria hypothesis originally found by Philippe et al. (2019)), we recovered a different topology in the reanalysis of the (Rouse et al. 2016) dataset, finding support for the Xemabulacraria hypothesis. Filtering phylogenomic datasets for the most informative gene families we do find increased support for an association of Xenacoelomorpha with Ambulacraria. Moving forward, it would be interesting to combine this analysis with a newly constructed dataset which assesses how specific approaches for ortholog gene family construction affects the resulting topology. Comparing software and methods may provide a more in depth insight into the gene family evolution that is being used to reconstruct the patterns of species evolution. Additionally, searching for composite genes in this newly constructed set of species may provide a more accurate depiction of the composite gene content within these groups and could provide useful makers to help resolve the placement of Xenacoelomorpha.

Overall, we have provided an empirical assessment of the contribution of gene remodeling to new gene genesis in Metazoa. While our analysis is limited in taxon sampling, in particular for non-bilaterian and outgroup species, our study shows that this mechanism of novel gene genesis makes a significant contribution to the gene family repertoire in animals. Using novel approaches, applying network theory, we uncovered complex patterns of gene evolution by fusion and fission.

Future analyses into comparing both composite and non-composite gene evolution may provide interesting insight into the modes and rates of evolution of these genes. Additionally, comparing the sequence dynamics of parent and composite genes, for example patterns of presence and absence, spatial proximity within the genome, shifting selective pressures, and differential levels of expression, could provide interesting insight into how gene remodeling occurs and how it impacts the subsequent evolution of the parent genes. This would require accurate annotation of both parent genes and composite genes within the same genome in order to compare the structure and evolution of these gene types. In our dataset we found only a small portion of composite genes that had retained parent genes within the same extant species. This suggests that analysis of this type would have to be carried out on younger composite genes, those present in specific lineages that emerged more recently, as these are more likely to be present along with their parent genes (Rogers et al. 2009). Furthermore, analysing the rate of remodeling following large genomic changes within the genome, for example after a whole genome duplication event, would be an interesting avenue of research to test the hypothesis that these significant changes within genome structure are correlated with higher rates of gene remodeling as we find in our results in Chapter 2. Similar to the previously outlined analysis, to get an accurate depiction of rate of remodeling following large genomic events recently diverging groups of species would provide key insights. The salmonid group is one such example, where another whole genome duplication event has occurred within the recently diverged group of fish, where annotating the mode and tempo of gene remodeling following these changes could be very insightful.

While we find high rates of homoplasy within composite genes, there is a subset that may be useful as phylogenetic markers. Ensuring appropriate taxon sampling and filtering for non-homoplastic characters is essential. We propose that composite genes (non-homoplastic) may be most powerful to use in combination with different data types and methods. Within the more traditional phylogenomic analyses, we believe that greater attention to ortholog assignment, and the effects of misidentification of orthology on topologies, is required to resolve outstanding issues within AToL. Recently, a significant focus has been placed on model misspecification and taxon sampling; in the future of phylogenetics greater emphasis should be put on taking a consilient approach using high quality data and new data types to resolve contentious nodes.

Conclusion

To conclude, the rate and contribution of gene remodeling to novel protein coding gene genesis across the Metazoa is significant with a total of 13,632 gene families emerging in this way throughout animal evolution. We show that gene remodeling is intensified at nodes of major transition in animal evolution, suggesting it has played an important role in driving protein and perhaps phenotypic complexity. We characterised and compared the traits of composite genes versus non-composite genes. The highly homoplastic nature of composite genes has significant impacts on our understanding of protein coding gene genesis within animals with independent fusions occurring independently along the tree. Indeed the levels of homoplasy brought us to question the application of composite genes to phylogenetic problems. However, we have also shown that enriching for non-homoplastic composite CHGs allowed us to accurately reconstruct known topologies within AToL and address contentious parts of the phylogeny. Finally, using non-homoplastic CHGs and treating previously published datasets to strict ortholog filters we have been able to take a consilient approach to the issue of the placement of the Xenacoelomorpha within AToL. We conclude that previously published datasets constructed to address the placement of Xenacoelomorpha within AToL contain high levels of hidden paralogy. Removing these problematic gene families produces increased support for the controversial placement of Xenacoelomorpha within Deuterostomia.

Chapter 6: Bibliography

Achatz, Johannes G., Marta Chiodin, Willi Salvenmoser, Seth Tyler, and Pedro Martinez. 2013. “The Acoela: On Their Kind and Kinships, Especially with Nemertodermatids and Xenoturbellids (Bilateria Incertae Sedis).” *Organisms, Diversity & Evolution* 13 (2): 267–86.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, et al. 2000. “The Genome Sequence of Drosophila Melanogaster.” *Science* 287 (5461): 2185–95.

Agaram, Narasimhan P., Hsiao-Wei Chen, Lei Zhang, Yun-Shao Sung, David Panicek, John H. Healey, G. Petur Nielsen, Christopher D. M. Fletcher, and Cristina R. Antonescu. 2015. “EWSR1-PBX3: A Novel Gene Fusion in Myoepithelial Tumors.” *Genes, Chromosomes & Cancer* 54 (2): 63–71.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. “Evidence for a Clade of Nematodes, Arthropods and Other Moulting Animals.” *Nature* 387 (6632): 489–93.

Akiva, Pinchas, Amir Toporik, Sarit Edelheit, Yifat Peretz, Alex Diber, Ronen Shemesh, Amit Novik, and Rotem Sorek. 2006. “Transcription-Mediated Gene Fusion in the Human Genome.” *Genome Research* 16 (1): 30–36.

Albalat, Ricard, and Cristian Cañestro. 2016. “Evolution by Gene Loss.” *Nature Reviews. Genetics* 17 (7): 379–91.

Alberti, Chiara, and Luisa Cochella. 2017. “A Framework for Understanding the Roles of miRNAs in Animal Development.” *Development* 144 (14): 2548–59.

Alekseyenko, Alexander V., Christopher J. Lee, and Marc A. Suchard. 2008. “Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos.” *Systematic Biology* 57 (5): 772–84.

Altenhoff, Adrian M., Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A. Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, et al. 2016. “Standardized Benchmarking in the Quest for Orthologs.” *Nature Methods* 13 (5): 425–30.

Altenhoff, Adrian M., Jeremy Levy, Magdalena Zarowiecki, Bartłomiej Tomiczek, Alex Warwick Vesztrocy, Daniel A. Dalquen, Steven Müller, et al. 2019. “OMA Standalone: Orthology Inference among Public and Custom Genomes and Transcriptomes.” *Genome Research* 29 (7): 1152–63.

Altenhoff, Adrian M., Nives Škunca, Natasha Glover, Clément-Marie Train, Anna Sueki, Ivana Piližota, Kevin Gori, et al. 2015. “The OMA Orthology Database in 2015: Function Predictions, Better Plant Support, Synteny View and Other Improvements.” *Nucleic Acids Research* 43 (Database issue): D240–49.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Arendt, Jeff, and David Reznick. 2008. “Convergence and Parallelism Reconsidered: What Have We Learned about the Genetics of Adaptation?” *Trends in Ecology & Evolution* 23 (1): 26–32.

Arimoto, Asuka, Tomoe Hikosaka-Katayama, Akira Hikosaka, Kuni Tagawa, Toyoshige Inoue, Tatsuya Ueki, Masa-Aki Yoshida, et al. 2019. “A Draft Nuclear-Genome Assembly of the Acoel Flatworm Praesagittifera Naikaiensis.” *GigaScience* 8 (4). https://doi.org/10.1093/gigascience/giz023.

Aristotle, Schneider JG, Cresswell R. 1862. Aristotle’s history of animals. In: Bohn HG, editor. Ten books (Historia animalium). London: George Bell & Sons.

Avesson, Lotta, Johan Reimegård, E. Gerhart H. Wagner, and Fredrik Söderbom. 2012. “MicroRNAs in Amoebozoa: Deep Sequencing of the Small RNA Population in the Social Amoeba Dictyostelium Discoideum Reveals Developmentally Regulated microRNAs.” *RNA* 18 (10): 1771–82.

Babushok, Daria V., Kazuhiko Ohshima, Eric M. Ostertag, Xinsheng Chen, Yanfeng Wang, Prabhat K. Mandal, Norihiro Okada, Charles S. Abrams, and Haig H. Kazazian Jr. 2007. “A Novel Testis Ubiquitin-Binding Protein Gene Arose by Exon Shuffling in Hominoids.” *Genome Research* 17 (8): 1129–38.

Bailey, Jeffrey A., Zhiping Gu, Royden A. Clark, Knut Reinert, Rhea V. Samonte, Stuart Schwartz, Mark D. Adams, Eugene W. Myers, Peter W. Li, and Evan E. Eichler. 2002. “Recent Segmental Duplications in the Human Genome.” *Science* 297 (5583): 1003–7.

Bapteste, E., E. Susko, J. Leigh, D. MacLeod, R. L. Charlebois, and W. F. Doolittle. 2005. “Do Orthologous Gene Phylogenies Really Support Tree-Thinking?” *BMC Evolutionary Biology* 5 (May): 33.

Bapteste, Eric, Leo van Iersel, Axel Janke, Scot Kelchner, Steven Kelk, James O. McInerney, David A. Morrison, et al. 2013. “Networks: Expanding Evolutionary Thinking.” *Trends in Genetics: TIG* 29 (8): 439–41.

Basu, Malay Kumar, Liran Carmel, Igor B. Rogozin, and Eugene V. Koonin. 2008. “Evolution of Protein Domain Promiscuity in Eukaryotes.” *Genome Research* 18 (3): 449–61.

Baumgarten, Sebastian, Oleg Simakov, Lisl Y. Esherick, Yi Jin Liew, Erik M. Lehnert, Craig T. Michell, Yong Li, et al. 2015. “The Genome of Aiptasia, a Sea Anemone Model for Coral Symbiosis.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (38): 11893–98.

Berry, Vincent, and Olivier Gascuel. 1996. “On the Interpretation of Bootstrap Trees: Appropriate Threshold of Clade Selection and Induced Gain.” *Molecular Biology and Evolution* 13 (7): 999–999.

Betancur-R., Ricardo, Dahiana Arcila, Richard P. Vari, Lily C. Hughes, Claudio Oliveira, Mark H. Sabaj, and Guillermo Ortí. 2019. “Phylogenomic Incongruence, Hypothesis Testing, and Taxonomic Sampling: The Monophyly of Characiform Fishes*.” *Evolution; International Journal of Organic Evolution* 73 (2): 329–45.

Bininda-Emonds, Olaf R. P. 2004. “The Evolution of Supertrees.” *Trends in Ecology & Evolution* 19 (6): 315–22.

Biology, Of. n.d. “THE QUARTERLY REVIEW.”

Blaxter, Mark. 2011. “Nematodes: The Worm and Its Relatives.” *PLoS Biology* 9 (4): e1001050.

Bleidorn, Christoph. 2017. *Phylogenomics: An Introduction*. Springer.

Bleidorn, Christoph. 2019. “Recent Progress in Reconstructing Lophotrochozoan (spiralian) Phylogeny.” *Organisms, Diversity & Evolution* 19 (4): 557–66.

Blomme, Tine, Klaas Vandepoele, Stefanie De Bodt, Cedric Simillion, Steven Maere, and Yves Van de Peer. 2006. “The Gain and Loss of Genes during 600 Million Years of Vertebrate Evolution.” *Genome Biology* 7 (5): R43.

Böhne, Astrid, Frédéric Brunet, Delphine Galiana-Arnoux, Christina Schultheis, and Jean-Nicolas Volff. 2008. “Transposable Elements as Drivers of Genomic and Biological Diversity in Vertebrates.” *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 16 (1): 203–15.

Boore, Jeffrey L. 2006. “The Use of Genome-Level Characters for Phylogenetic Reconstruction.” *Trends in Ecology & Evolution* 21 (8): 439–46.

Bornberg-Bauer, Erich, and M. Mar Albà. 2013. “Dynamics and Adaptive Benefits of Modular Protein Evolution.” *Current Opinion in Structural Biology* 23 (3): 459–66.

Borner, Janus, Peter Rehm, Ralph O. Schill, Ingo Ebersberger, and Thorsten Burmester. 2014. “A Transcriptome Approach to Ecdysozoan Phylogeny.” *Molecular Phylogenetics and Evolution* 80 (November): 79–87.

Børve, Aina, and Andreas Hejnol. 2014. “Development and Juvenile Anatomy of the Nemertodermatid Meara Stichopi (Bock) Westblad 1949 (Acoelomorpha).” *Frontiers in Zoology* 11 (July): 50.

Bourlat, Sarah J., Thorhildur Juliusdottir, Christopher J. Lowe, Robert Freeman, Jochanan Aronowicz, Mark Kirschner, Eric S. Lander, et al. 2006. “Deuterostome Phylogeny Reveals Monophyletic Chordates and the New Phylum Xenoturbellida.” *Nature* 444 (7115): 85–88.

Bourlat, Sarah J., Claus Nielsen, Anne E. Lockyer, D. Timothy J. Littlewood, and Maximilian J. Telford. 2003. “Xenoturbella Is a Deuterostome That Eats Molluscs.” *Nature* 424 (6951): 925–28.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. “Ten Things You Should Know about Transposable Elements.” *Genome Biology* 19 (1): 199.

Brandt, Jürgen, Sabrina Schrauth, Anne-Marie Veith, Alexander Froschauer, Torsten Haneke, Christina Schultheis, Manfred Gessler, Cornelia Leimeister, and Jean-Nicolas Volff.

2005. “Transposable Elements as a Source of Genetic Innovation: Expression and Evolution of a Family of Retrotransposon-Derived Neogenes in Mammals.” *Gene* 345 (1): 101–11.

Bråte, Jon, Ralf S. Neumann, Bastian Fromm, Arthur A. B. Haraldsen, James E. Tarver, Hiroshi Suga, Philip C. J. Donoghue, et al. 2018. “Unicellular Origin of the Animal MicroRNA Machinery.” *Current Biology: CB* 28 (20): 3288–95.e5.

Brennan, Greg, Yury Kozyrev, and Shiu-Lok Hu. 2008. “TRIMCyp Expression in Old World Primates Macaca Nemestrina and Macaca Fascicularis.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (9): 3569–74.

Brinza, Lilia, José Viñuelas, Ludovic Cottret, Federica Calevro, Yvan Rahbé, Gérard Febvay, Gabrielle Duport, et al. 2009. “Systemic Analysis of the Symbiotic Function of Buchnera Aphidicola, the Primary Endosymbiont of the Pea Aphid Acyrthosiphon Pisum.” *Comptes Rendus Biologies* 332 (11): 1034–49.

Brown, Jeremy M., and Robert C. Thomson. 2017. “Bayes Factors Unmask Highly Variable Information Content, Bias, and Extreme Influence in Phylogenomic Analyses.” *Systematic Biology* 66 (4): 517–30.

Brunet, Thibaut, and Nicole King. 2017. “The Origin of Animal Multicellularity and Cell Differentiation.” *Developmental Cell* 43 (2): 124–40.

Bryant, David, and Matthew W. Hahn. 2020. “The Concatenation Question.” *Phylogenetics in the Genomic Era*, 3.4:1–3.4:23.

Cafarelli, T. M., A. Desbuleux, Y. Wang, S. G. Choi, D. De Ridder, and M. Vidal. 2017. “Mapping, Modeling, and Characterization of Protein-Protein Interactions on a Proteomic Scale.” *Current Opinion in Structural Biology* 44 (June): 201–10.

Campbell, Lahcen I., Omar Rota-Stabelli, Gregory D. Edgecombe, Trevor Marchioro, Stuart J. Longhorn, Maximilian J. Telford, Hervé Philippe, Lorena Rebecchi, Kevin J. Peterson, and Davide Pisani. 2011. “MicroRNAs and Phylogenomics Resolve the Relationships of Tardigrada and Suggest That Velvet Worms Are the Sister Group of Arthropoda.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (38): 15920–24.

Canapa, Adriana, Marco Barucca, Maria A. Biscotti, Mariko Forconi, and Ettore Olmo. 2015. “Transposons, Genome Size, and Evolutionary Insights in Animals.” *Cytogenetic and Genome Research* 147 (4): 217–39.

Cannon, Johanna Taylor, Bruno Cossermelli Vellutini, Julian Smith 3rd, Fredrik Ronquist, Ulf Jondelius, and Andreas Hejnol. 2016. “Xenacoelomorpha Is the Sister Group to Nephrozoa.” *Nature* 530 (7588): 89–93.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73.

Carvunis, Anne-Ruxandra, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, et al. 2012. “Proto-Genes and de Novo Gene Birth.” *Nature* 487 (7407): 370–74.

Castresana, J. 2000. “Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis.” *Molecular Biology and Evolution* 17 (4): 540–52.

Cavalier-Smith, T. 2007. “A Revised Six-Kingdom System of Life.” *Biological Reviews of the Cambridge Philosophical Society* 73 (3): 203–66.

Cavalier-Smith, Thomas. 2017. “Origin of Animal Multicellularity: Precursors, Causes, Consequences-the Choanoflagellate/sponge Transition, Neurogenesis and the Cambrian Explosion.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372 (1713). https://doi.org/10.1098/rstb.2015.0476.

Cavalier-Smith, Thomas, and Ema E-Y Chao. 2003. “Phylogeny of Choanozoa, Apusozoa, and Other Protozoa and Early Eukaryote Megaevolution.” *Journal of Molecular Evolution* 56 (5): 540–63.

Chalopin, Domitille, Magali Naville, Floriane Plard, Delphine Galiana, and Jean-Nicolas Volff. 2015. “Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates.” *Genome Biology and Evolution* 7 (2): 567–80.

Chang, E. Sally, Moran Neuhof, Nimrod D. Rubinstein, Arik Diamant, Hervé Philippe, Dorothée Huchon, and Paulyn Cartwright. 2015. “Genomic Insights into the Evolutionary

Origin of Myxozoa within Cnidaria." *Proceedings of the National Academy of Sciences of the United States of America* 112 (48): 14912–17.

Chapman, Jarrod A., Ewen F. Kirkness, Oleg Simakov, Steven E. Hampson, Therese Mitros, Thomas Weinmaier, Thomas Rattei, et al. 2010. "The Dynamic Genome of Hydra." *Nature* 464 (7288): 592–96.

Chen, Meng-Yun, Dan Liang, and Peng Zhang. 2017. "Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences." *Genome Biology and Evolution* 9 (8): 1998–2012.

Chen, Sidi, Yong E. Zhang, and Manyuan Long. 2010. "New Genes in Drosophila Quickly Become Essential." *Science* 330 (6011): 1682–85.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2016. "Regulatory Evolution of Innate Immunity through Co-Option of Endogenous Retroviruses." *Science* 351 (6277): 1083–87.

Ciccarelli, Francesca D., Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. 2006. "Toward Automatic Reconstruction of a Highly Resolved Tree of Life." *Science* 311 (5765): 1283–87.

Ciccarelli, Francesca D., Christian von Mering, Mikita Suyama, Eoghan D. Harrington, Elisa Izaurralde, and Peer Bork. 2005. "Complex Genomic Rearrangements Lead to Novel Primate Gene Function." *Genome Research* 15 (3): 343–51.

Conaco, Cecilia, Pantelis Tsoulfas, Onur Sakarya, Amanda Dolan, John Werren, and Kenneth S. Kosik. 2016. "Detection of Prokaryotic Genes in the Amphimedon Queenslandica Genome." *PloS One* 11 (3): e0151092.

Conant, Gavin C., and Kenneth H. Wolfe. 2008. "Turning a Hobby into a Job: How Duplicated Genes Find New Functions." *Nature Reviews. Genetics* 9 (12): 938–50.

Cook, Charles E., Eva Jiménez, Michael Akam, and Emili Saló. 2004. "The Hox Gene Complement of Acoel Flatworms, a Basal Bilaterian Clade." *Evolution & Development* 6 (3): 154–63.

Criscuolo, Alexis, and Simonetta Gribaldo. 2010. “BMGE (Block Mapping and Gathering with Entropy): A New Software for Selection of Phylogenetic Informative Regions from Multiple Sequence Alignments.” *BMC Evolutionary Biology* 10 (July): 210.

Cromar, Graham, Ka-Chun Wong, Noeleen Loughran, Tuan On, Hongyan Song, Xuejian Xiong, Zhaolei Zhang, and John Parkinson. 2014. “New Tricks for ‘Old’ Domains: How Novel Architectures and Promiscuous Hubs Contributed to the Organization and Evolution of the ECM.” *Genome Biology and Evolution* 6 (10): 2897–2917.

Crow, Karen D., Günter P. Wagner, and SMBE Tri-National Young Investigators. 2006. “Proceedings of the SMBE Tri-National Young Investigators’ Workshop 2005. What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity?” *Molecular Biology and Evolution* 23 (5): 887–92.

Delgado-Baquerizo, Manuel, Peter B. Reich, Chanda Trivedi, David J. Eldridge, Sebastián Abades, Fernando D. Alfaro, Felipe Bastida, et al. 2020. “Multiple Elements of Soil Biodiversity Drive Ecosystem Functions across Biomes.” *Nature Ecology & Evolution* 4 (2): 210–20.

Delsuc, Frédéric, Henner Brinkmann, Daniel Chourrout, and Hervé Philippe. 2006. “Tunicates and Not Cephalochordates Are the Closest Living Relatives of Vertebrates.” *Nature* 439 (7079): 965–68.

Delsuc, Frédéric, Henner Brinkmann, and Hervé Philippe. 2005. “Phylogenomics and the Reconstruction of the Tree of Life.” *Nature Reviews. Genetics* 6 (5): 361–75.

Delsuc, Frédéric, Hervé Philippe, Georgia Tsagkogeorga, Paul Simion, Marie-Ka Tilak, Xavier Turon, Susanna López-Legentil, Jacques Piette, Patrick Lemaire, and Emmanuel J. P. Douzery. 2018. “A Phylogenomic Framework and Timescale for Comparative Studies of Tunicates.” *BMC Biology* 16 (1): 39.

Demichelis, F., K. Fall, S. Perner, O. Andrén, F. Schmidt, S. R. Setlur, Y. Hoshida, et al. 2007. “TMPRSS2:ERG Gene Fusion Associated with Lethal Prostate Cancer in a Watchful Waiting Cohort.” *Oncogene* 26 (31): 4596–99.

Denoeud, France, Philipp Kapranov, Catherine Ucla, Adam Frankish, Robert Castelo, Jorg Drenkow, Julien Lagarde, et al. 2007. “Prominent Use of Distal 5’ Transcription Start Sites and Discovery of a Large Number of Additional Exons in ENCODE Regions." *Genome Research* 17 (6): 746–59.

Des Marais, David L., and Mark D. Rausher. 2008. "Escape from Adaptive Conflict after Duplication in an Anthocyanin Pathway Gene." *Nature* 454 (7205): 762–65.

Dessimoz, Christophe, and Manuel Gil. 2010. "Phylogenetic Assessment of Alignments Reveals Neglected Tree Signal in Gaps." *Genome Biology* 11 (4): R37.

Di Franco, Arnaud, Raphaël Poujol, Denis Baurain, and Hervé Philippe. 2019. "Evaluating the Usefulness of Alignment Filtering Methods to Reduce the Impact of Errors on Evolutionary Inferences." *BMC Evolutionary Biology* 19 (1): 21.

Dohmen, Elias, Steffen Klasberg, Erich Bornberg-Bauer, Sören Perrey, and Carsten Kemena. 2020. "The Modular Nature of Protein Evolution: Domain Rearrangement Rates across Eukaryotic Life." *BMC Evolutionary Biology* 20 (1): 30.

Dohrmann, Martin, and Gert Wörheide. 2013. "Novel Scenarios of Early Animal Evolution--Is It Time to Rewrite Textbooks?" *Integrative and Comparative Biology* 53 (3): 503–11.

Doolittle, W. F. 1999. "Phylogenetic Classification and the Universal Tree." *Science* 284 (5423): 2124–29.

Dudas, Gytis, Luiz Max Carvalho, Trevor Bedford, Andrew J. Tatem, Guy Baele, Nuno R. Faria, Daniel J. Park, et al. 2017. "Virus Genomes Reveal Factors That Spread and Sustained the Ebola Epidemic." *Nature* 544 (7650): 309–15.

Dunn, C. W., and C. Munro. 2016. "Comparative Genomics and the Diversity of Life." *Zoologica Scripta*. https://onlinelibrary.wiley.com/doi/abs/10.1111/zsc.12211.

Dunn, Casey W. 2017. "Ctenophore Trees." *Nature Ecology & Evolution*.

Dunn, Casey W., Gonzalo Giribet, Gregory D. Edgecombe, and Andreas Hejnol. 2014. "Animal Phylogeny and Its Evolutionary Implications." *Annual Review of Ecology, Evolution, and Systematics* 45 (1): 371–95.

Dunn, Casey W., Andreas Hejnol, David Q. Matus, Kevin Pang, William E. Browne, Stephen A. Smith, Elaine Seaver, et al. 2008. "Broad Phylogenomic Sampling Improves Resolution of the Animal Tree of Life." *Nature* 452 (7188): 745–49.

Dunn, Casey W., Mark Howison, and Felipe Zapata. 2013. “Agalma: An Automated Phylogenomics Workflow.” *BMC Bioinformatics* 14 (November): 330.

Dunn, Casey W., Sally P. Leys, and Steven H. D. Haddock. 2015. “The Hidden Biology of Sponges and Ctenophores.” *Trends in Ecology & Evolution* 30 (5): 282–91.

Dunn, Casey W., and Joseph F. Ryan. 2015. “The Evolution of Animal Genomes.” *Current Opinion in Genetics & Development* 35 (December): 25–32.

Dunwell, Thomas L., Jordi Paps, and Peter W. H. Holland. 2017. “Novel and Divergent Genes in the Evolution of Placental Mammals.” *Proceedings. Biological Sciences / The Royal Society* 284 (1864). https://doi.org/10.1098/rspb.2017.1357.

Ebersberger, Ingo, Sascha Strauss, and Arndt von Haeseler. 2009. “HaMStR: Profile Hidden Markov Model Based Search for Orthologs in ESTs.” *BMC Evolutionary Biology* 9 (July): 157.

Eddy, Sean R. 2008. “A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation.” *PLoS Computational Biology* 4 (5): e1000069.

Edgar, Robert C. 2004. “MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity.” *BMC Bioinformatics* 5 (August): 113.

Edwards, Scott V. 2009. “Is a New and General Theory of Molecular Systematics Emerging?” *Evolution; International Journal of Organic Evolution* 63 (1): 1–19.

Eitel, Michael, Warren R. Francis, Frédérique Varoqueaux, Jean Daraspe, Hans-Jürgen Osigus, Stefan Krebs, Sergio Vargas, et al. 2018. “Comparative Genomics and the Nature of Placozoan Species.” *PLoS Biology* 16 (7): e2005359.

Ekman, Diana, Asa K. Björklund, and Arne Elofsson. 2007. “Quantification of the Elevated Rate of Domain Rearrangements in Metazoa.” *Journal of Molecular Biology* 372 (5): 1337–48.

El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. “The Pfam Protein Families Database in 2019.” *Nucleic Acids Research* 47 (D1): D427–32.

Elliott, Tyler A., and T. Ryan Gregory. 2015. “Do Larger Genomes Contain More Diverse Transposable Elements?” *BMC Evolutionary Biology* 15 (April): 69.

Emmert-Streib, Frank, Matthias Dehmer, and Benjamin Haibe-Kains. 2014. “Gene Regulatory Networks and Their Applications: Understanding Biological and Medical Problems in Terms of Networks.” *Frontiers in Cell and Developmental Biology* 2 (August): 38.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.

Enright, A. J., I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. 1999. “Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events.” *Nature* 402 (6757): 86–90.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. “An Efficient Algorithm for Large-Scale Detection of Protein Families.” *Nucleic Acids Research* 30 (7): 1575–84.

Fairclough, Stephen R., Zehua Chen, Eric Kramer, Qiandong Zeng, Sarah Young, Hugh M. Robertson, Emina Begovic, et al. 2013. “Premetazoan Genome Evolution and the Regulation of Cell Differentiation in the Choanoflagellate Salpingoeca Rosetta.” *Genome Biology* 14 (2): R15.

Farris, James S. 1989. “THE RETENTION INDEX AND THE RESCALED CONSISTENCY INDEX.” *Cladistics: The International Journal of the Willi Hennig Society* 5 (4): 417–19.

Fedotova, A. A., A. N. Bonchuk, V. A. Mogila, and P. G. Georgiev. 2017. “C2H2 Zinc Finger Proteins: The Largest but Poorly Explored Family of Higher Eukaryotic Transcription Factors.” *Acta Naturae* 9 (2): 47–58.

Felsenstein, J. 1981. “Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach.” *Journal of Molecular Evolution* 17 (6): 368–76.

Felsenstein, J., and G. A. Churchill. 1996. “A Hidden Markov Model Approach to Variation among Sites in Rate of Evolution.” *Molecular Biology and Evolution* 13 (1): 93–104.

Felsenstein, Joseph. 1978. “Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading.” *Systematic Zoology* 27 (4): 401–10.

Felsenstein, Joseph. 1985. “Phylogenies and the Comparative Method.” *The American Naturalist* 125 (1): 1–15.

Felsenstein, Joseph, and Joseph Felenstein. 2004. *Inferring Phylogenies*. Vol. 2. Sinauer associates Sunderland, MA.

Fernández, Rosa, and Toni Gabaldón. 2020. “Gene Gain and Loss across the Metazoan Tree of Life.” *Nature Ecology & Evolution*, January. https://doi.org/10.1038/s41559-019-1069-x.

Fernández, Rosa, Robert J. Kallal, Dimitar Dimitrov, Jesús A. Ballesteros, Miquel A. Arnedo, Gonzalo Giribet, and Gustavo Hormiga. 2018. “Phylogenomics, Diversification Dynamics, and Comparative Transcriptomics across the Spider Tree of Life.” *Current Biology: CB* 28 (9): 1489–97.e5.

Feuda, Roberto, Martin Dohrmann, Walker Pett, Hervé Philippe, Omar Rota-Stabelli, Nicolas Lartillot, Gert Wörheide, and Davide Pisani. 2017. “Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals.” *Current Biology: CB* 27 (24): 3864–70.e4.

Fitch, W. M. 2000. “Homology a Personal View on Some of the Problems.” *Trends in Genetics: TIG* 16 (5): 227–31.

Fitz-Gibbon, S. T., and C. H. House. 1999. “Whole Genome-Based Phylogenetic Analysis of Free-Living Microorganisms.” *Nucleic Acids Research* 27 (21): 4218–22.

Flajnik, Martin F. 2014. “Re-Evaluation of the Immunological Big Bang.” *Current Biology: CB* 24 (21): R1060–65.

Flammang, Brooke E., Apinun Suvarnaraksha, Julie Markiewicz, and Daphne Soares. 2016. “Tetrapod-like Pelvic Girdle in a Walking Cavefish.” *Scientific Reports* 6 (March): 23711.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. “Preservation of Duplicate Genes by Complementary, Degenerative Mutations.” *Genetics* 151 (4): 1531–45.

Foster, Peter G. 2004. “Modeling Compositional Heterogeneity.” *Systematic Biology* 53 (3): 485–95.

Gabaldón, Toni. 2008. “Large-Scale Assignment of Orthology: Back to Phylogenetics?” *Genome Biology* 9 (10): 235.

Gaiti, Federico, Andrew D. Calcino, Miloš Tanurdžić, and Bernard M. Degnan. 2017. “Origin and Evolution of the Metazoan Non-Coding Regulatory Genome.” *Developmental Biology* 427 (2): 193–202.

Galperin, Michael Y., Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. 2015. “Expanded Microbial Genome Coverage and Improved Protein Family Annotation in the COG Database.” *Nucleic Acids Research* 43 (Database issue): D261–69.

Garamszegi, László Zsolt, ed. 2014. *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*. Springer, Berlin, Heidelberg.

Gatesy, John, and Mark S. Springer. 2014. “Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/concatalescence Conundrum.” *Molecular Phylogenetics and Evolution* 80 (November): 231–66.

Gaudry, Michael J., Jay F. Storz, Gary Tyler Butts, Kevin L. Campbell, and Federico G. Hoffmann. 2014. “Repeated Evolution of Chimeric Fusion Genes in the β-Globin Gene Family of Laurasiatherian Mammals.” *Genome Biology and Evolution* 6 (5): 1219–34.

Gehrke, Andrew R., Emily Neverett, Yi-Jyun Luo, Alexander Brandt, Lorenzo Ricci, Ryan E. Hulett, Annika Gompers, et al. 2019. “Acoel Genome Reveals the Regulatory Landscape of Whole-Body Regeneration.” *Science* 363 (6432). https://doi.org/10.1126/science.aau6173.

“Genomic Control Process | ScienceDirect.” n.d. Accessed March 20, 2020. https://www.sciencedirect.com/book/9780124047297/genomic-control-process.

Giribet, Gonzalo. 2016. “New Animal Phylogeny: Future Challenges for Animal Phylogeny in the Age of Phylogenomics.” *Organisms, Diversity & Evolution* 16 (2): 419–26.

Giribet, Gonzalo, and Gregory D. Edgecombe. 2017. “Current Understanding of Ecdysozoa and Its Internal Phylogenetic Relationships.” *Integrative and Comparative Biology* 57 (3): 455–66.

Gladyshev, Eugene A., Matthew Meselson, and Irina R. Arkhipova. 2008. “Massive Horizontal Gene Transfer in Bdelloid Rotifers.” *Science* 320 (5880): 1210–13.

Gogarten, J. Peter, and Jeffrey P. Townsend. 2005. “Horizontal Gene Transfer, Genome Innovation and Evolution.” *Nature Reviews. Microbiology* 3 (9): 679–87.

Goloboff, Pablo A., James S. Farris, and Kevin C. Nixon. 2008. “TNT, a Free Program for Phylogenetic Analysis.” *Cladistics: The International Journal of the Willi Hennig Society* 24 (5): 774–86.

González, Josefa, Talia L. Karasov, Philipp W. Messer, and Dmitri A. Petrov. 2010. “Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila.” *PLoS Genetics* 6 (4): e1000905.

González, Josefa, Kapa Lenkov, Mikhail Lipatov, J. Michael Macpherson, and Dmitri A. Petrov. 2008. “High Rate of Recent Transposable Element-Induced Adaptation in Drosophila Melanogaster.” *PLoS Biology* 6 (10): e251.

González, Josefa, J. Michael Macpherson, and Dmitri A. Petrov. 2009. “A Recent Adaptive Transposable Element Insertion near Highly Conserved Developmental Loci in Drosophila Melanogaster.” *Molecular Biology and Evolution* 26 (9): 1949–61.

Gough, Julian. 2005. “Convergent Evolution of Domain Architectures (is Rare).” *Bioinformatics* 21 (8): 1464–71.

Grau-Bové, Xavier, Guifré Torruella, Stuart Donachie, Hiroshi Suga, Guy Leonard, Thomas A. Richards, and Iñaki Ruiz-Trillo. 2017. “Dynamics of Genomic Innovation in the Unicellular Ancestry of Animals.” *eLife* 6 (July). https://doi.org/10.7554/eLife.26036.

Gregory, T. Ryan. 2002. “Genome Size and Developmental Complexity.” *Genetica* 115 (1): 131–46.

Gregory, T. Ryan, James A. Nicol, Heidi Tamm, Bellis Kullman, Kaur Kullman, Ilia J. Leitch, Brian G. Murray, Donald F. Kapraun, Johann Greilhuber, and Michael D. Bennett. 2007. “Eukaryotic Genome Size Databases.” *Nucleic Acids Research* 35 (Database issue): D332–38.

Grimson, Andrew, Mansi Srivastava, Bryony Fahey, Ben J. Woodcroft, H. Rosaria Chiang, Nicole King, Bernard M. Degnan, Daniel S. Rokhsar, and David P. Bartel. 2008. “Early

Origins and Evolution of microRNAs and Piwi-Interacting RNAs in Animals.” *Nature* 455 (7217): 1193–97.

Guang, August, Felipe Zapata, Mark Howison, Charles E. Lawrence, and Casey W. Dunn. 2016. “An Integrated Perspective on Phylogenetic Workflows.” *Trends in Ecology & Evolution* 31 (2): 116–26.

Guijarro-Clarke, Cristina, Peter W. H. Holland, and Jordi Paps. 2020. “Widespread Patterns of Gene Loss in the Evolution of the Animal Kingdom.” *Nature Ecology & Evolution*, February. https://doi.org/10.1038/s41559-020-1129-2.

Hadfield, James, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. 2018. “Nextstrain: Real-Time Tracking of Pathogen Evolution.” *Bioinformatics* 34 (23): 4121–23.

Haeckel E. 1874a. Anthropogenie oder Entwicklungsgeschichte des Menschen. Gemeinverständliche wissenschaftliche Vorträge über die Grundzüge der menschlichen Keimes-und Stammesgeschichte. Leipzig, Germany.

Haggerty, Leanne S., Pierre-Alain Jachiet, William P. Hanage, David A. Fitzpatrick, Philippe Lopez, Mary J. O’Connell, Davide Pisani, Mark Wilkinson, Eric Bapteste, and James O. McInerney. 2014. “A Pluralistic Account of Homology: Adapting the Models to the Data.” *Molecular Biology and Evolution* 31 (3): 501–16.

Hahn, Matthew W., Jeffery P. Demuth, and Sang-Gook Han. 2007. “Accelerated Rate of Gene Gain and Loss in Primates.” *Genetics* 177 (3): 1941–49.

Harzsch, Steffen, and Carsten H. G. Müller. 2007. “A New Look at the Ventral Nerve Centre of Sagitta: Implications for the Phylogenetic Position of Chaetognatha (arrow Worms) and the Evolution of the Bilaterian Nervous System.” *Frontiers in Zoology* 4 (May): 14.

Hastings, W. K. 1970. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications.” *Biometrika* 57 (1): 97–109.

Haszprunar, Gerhard. 2016. “Review of Data for a Morphological Look on Xenacoelomorpha (Bilateria Incertae Sedis).” *Organisms, Diversity & Evolution* 16 (2): 363–89.

Hedges, S. Blair, Jaime E. Blair, Maria L. Venturi, and Jason L. Shoe. 2004. “A Molecular Timescale of Eukaryote Evolution and the Rise of Complex Multicellular Life.” *BMC Evolutionary Biology* 4 (January): 2.

Heimberg, Alysha M., Lorenzo F. Sempere, Vanessa N. Moy, Philip C. J. Donoghue, and Kevin J. Peterson. 2008. “MicroRNAs and the Advent of Vertebrate Morphological Complexity.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (8): 2946–50.

Heinen, Tobias J. A. J., Fabian Staubach, Daniela Häming, and Diethard Tautz. 2009. “Emergence of a New Gene from an Intergenic Region.” *Current Biology: CB* 19 (18): 1527–31.

Hejnol, Andreas. 2010. “A Twist in Time--the Evolution of Spiral Cleavage in the Light of Animal Phylogeny.” *Integrative and Comparative Biology* 50 (5): 695–706.

Hejnol, Andreas. 2015. “Acoelomorpha and Xenoturbellida.” In *Evolutionary Developmental Biology of Invertebrates 1: Introduction, Non-Bilateria, Acoelomorpha, Xenoturbellida, Chaetognatha*, edited by Andreas Wanninger, 203–14. Vienna: Springer Vienna.

Hejnol, Andreas, and Mark Q. Martindale. 2008. “Acoel Development Indicates the Independent Evolution of the Bilaterian Mouth and Anus.” *Nature* 456 (7220): 382–86.

Hejnol, Andreas, Matthias Obst, Alexandros Stamatakis, Michael Ott, Greg W. Rouse, Gregory D. Edgecombe, Pedro Martinez, et al. 2009. “Assessing the Root of Bilaterian Animals with Scalable Phylogenomic Methods.” *Proceedings. Biological Sciences / The Royal Society* 276 (1677): 4261–70.

Hejnol, Andreas, and Kevin Pang. 2016. “Xenacoelomorpha’s Significance for Understanding Bilaterian Evolution.” *Current Opinion in Genetics & Development* 39 (August): 48–54.

Helm, Conrad, Stephan H. Bernhart, Christian Höner zu Siederdissen, Birgit Nickel, and Christoph Bleidorn. 2012. “Deep Sequencing of Small RNAs Confirms an Annelid Affinity of Myzostomida.” *Molecular Phylogenetics and Evolution* 64 (1): 198–203.

Hernandez, Alexandra M., and Joseph F. Ryan. 2018. “Horizontally Transferred Genes in the Ctenophore Mnemiopsis Leidyi.” *PeerJ* 6 (June): e5067.

Hernandez, Alexandra M., and Joseph F. Ryan. 2019. “Six-State Amino Acid Recoding Is Not an Effective Strategy to Offset the Effects of Compositional Heterogeneity and Saturation in Phylogenetic Analyses.” *bioRxiv*. https://doi.org/10.1101/729103.

Hittinger, Chris Todd, and Sean B. Carroll. 2007. “Gene Duplication and the Adaptive Evolution of a Classic Genetic Switch.” *Nature* 449 (7163): 677–81.

Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, and Fredrik Ronquist. 2016. “RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language.” *Systematic Biology* 65 (4): 726–36.

Holland, Linda Z., and Daniel Ocampo Daza. 2018. “A New Look at an Old Question: When Did the Second Whole Genome Duplication Occur in Vertebrate Evolution?” *Genome Biology*.

Horváth, Vivien, Miriam Merenciano, and Josefa González. 2017. “Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response.” *Trends in Genetics: TIG* 33 (11): 832–41.

Huelsenbeck, John, and Bruce Rannala. 2004. “Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees under Simple and Complex Substitution Models.” *Systematic Biology* 53 (6): 904–13.

Huelsmann, Matthias, Nikolai Hecker, Mark S. Springer, John Gatesy, Virag Sharma, and Michael Hiller. 2019. “Genes Lost during the Transition from Land to Water in Cetaceans Highlight Genomic Changes Associated with Aquatic Adaptations.” *Science Advances* 5 (9): eaaw6671.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data.” *Molecular Biology and Evolution* 33 (6): 1635–38.

Husnik, Filip, Naruo Nikoh, Ryuichi Koga, Laura Ross, Rebecca P. Duncan, Manabu Fujie, Makiko Tanaka, et al. 2013. “Horizontal Gene Transfer from Diverse Bacteria to an Insect Genome Enables a Tripartite Nested Mealybug Symbiosis.” *Cell* 153 (7): 1567–78.

Itoh, Masumi, Jose C. Nacher, Kei-Ichi Kuma, Susumu Goto, and Minoru Kanehisa. 2007. “Evolutionary History and Functional Implications of Protein Domains and Their Combinations in Eukaryotes.” *Genome Biology* 8 (6): R121.

Jacob, F. 1977. “Evolution and Tinkering.” *Science* 196 (4295): 1161–66.

Jager, Muriel, Roxane Chiori, Alexandre Alié, Cyrielle Dayraud, Eric Quéinnec, and Michaël Manuel. 2011. “New Insights on Ctenophore Neural Anatomy: Immunofluorescence Study in Pleurobrachia Pileus (Müller, 1776).” *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 316B (3): 171–87.

Jékely, Gáspár, Jordi Paps, and Claus Nielsen. 2015. “The Phylogenetic Position of Ctenophores and the Origin(s) of Nervous Systems.” *EvoDevo* 6 (January): 1.

Jenner, Ronald A. 2004. “Accepting Partnership by Submission? Morphological Phylogenetics in a Molecular Millennium.” *Systematic Biology*.

Jondelius, Ulf, Olga I. Raikova, and Pedro Martinez. 2019. “Xenacoelomorpha, a Key Group to Understand Bilaterian Evolution: Morphological and Molecular Perspectives.” In *Evolution, Origin of Life, Concepts and Methods*, 287–315. Springer.

Jukes TH, Cantor CR (1969). Evolution of Protein Molecules. New York: Academic Press. pp. 21–132.

Kaessmann, Henrik. 2010. “Origins, Evolution, and Phenotypic Impact of New Genes.” *Genome Research* 20 (10): 1313–26.

Kass, Robert E., and Adrian E. Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association* 90 (430): 773–95.

Katju, Vaishali, James C. Farslow, and Ulfar Bergthorsson. 2009. “Variation in Gene Duplicates with Low Synonymous Divergence in Saccharomyces Cerevisiae Relative to Caenorhabditis Elegans.” *Genome Biology* 10 (7): R75.

Katoh, Kazutaka, Kei-Ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. “MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment.” *Nucleic Acids Research* 33 (2): 511–18.

Kawashima, Takeshi, Shuichi Kawashima, Chisaki Tanaka, Miho Murai, Masahiko Yoneda, Nicholas H. Putnam, Daniel S. Rokhsar, Minoru Kanehisa, Nori Satoh, and Hiroshi Wada.

2009. “Domain Shuffling and the Evolution of Vertebrates.” *Genome Research* 19 (8): 1393–1403.

Kellis, Manolis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S. Lander. 2003. “Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements.” *Nature* 423 (6937): 241–54.

Kemena, Carsten, and Cedric Notredame. 2009. “Upcoming Challenges for Multiple Sequence Alignment Methods in the High-Throughput Era.” *Bioinformatics* 25 (19): 2455–65.

King, Nicole, and Antonis Rokas. 2017. “Embracing Uncertainty in Reconstructing Early Animal Evolution.” *Current Biology: CB* 27 (19): R1081–88.

King, Nicole, M. Jody Westbrook, Susan L. Young, Alan Kuo, Monika Abedin, Jarrod Chapman, Stephen Fairclough, et al. 2008. “The Genome of the Choanoflagellate Monosiga Brevicollis and the Origin of Metazoans.” *Nature* 451 (7180): 783–88.

Klingenberg, Christian Peter, and Nelly A. Gidaszewski. 2010. “Testing and Quantifying Phylogenetic Signals and Homoplasy in Morphometric Data.” *Systematic Biology* 59 (3): 245–61.

Kluge, Arnold G., and James S. Farris. 1969. “Quantitative Phyletics and the Evolution of Anurans.” *Systematic Biology* 18 (1): 1–32.

Kocot, Kevin M., Torsten H. Struck, Julia Merkel, Damien S. Waits, Christiane Todt, Pamela M. Brannock, David A. Weese, et al. 2017. “Phylogenomics of Lophotrochozoa with Consideration of Systematic Error.” *Systematic Biology* 66 (2): 256–82.

Kondo, Natsuko, Naruo Nikoh, Nobuyuki Ijichi, Masakazu Shimada, and Takema Fukatsu. 2002. “Genome Fragment of Wolbachia Endosymbiont Transferred to X Chromosome of Host Insect.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 14280–85.

Krause, Ann E., Kenneth A. Frank, Doran M. Mason, Robert E. Ulanowicz, and William W. Taylor. 2003. “Compartments Revealed in Food-Web Structure.” *Nature* 426 (6964): 282–85.

Kubatko, Laura Salter, and James H. Degnan. 2007. “Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence.” *Systematic Biology* 56 (1): 17–24.

Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. "TimeTree: A Resource for Timelines, Timetrees, and Divergence Times." *Molecular Biology and Evolution* 34 (7): 1812–19.

Lake, James A., and Maria C. Rivera. 2004. "Deriving the Genomic Tree of Life in the Presence of Horizontal Gene Transfer: Conditioned Reconstruction." *Molecular Biology and Evolution* 21 (4): 681–90.

Lambert, J. David. 2010. "Developmental Patterns in Spiralian Embryos." *Current Biology: CB* 20 (2): R72–77.

Lambshead, P. John D., and Guy Boucher. 2003. "Marine Nematode Deep-Sea Biodiversity - Hyperdiverse or Hype?: Guest Editorial." *Journal of Biogeography* 30 (4): 475–85.

Lanciano, Sophie, and Marie Mirouze. 2018. "Transposable Elements: All Mobile, All Different, Some Stress Responsive, Some Adaptive?" *Current Opinion in Genetics & Development* 49 (April): 106–14.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. "The Closest Unicellular Relatives of Animals." *Current Biology: CB* 12 (20): 1773–78.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Lartillot, N. 2020. "The Bayesian Approach to Molecular Phylogeny." https://hal.archives-ouvertes.fr/hal-02535330/file/chapter_1.4_lartillot_chap.pdf.

Lartillot, Nicolas, Henner Brinkmann, and Hervé Philippe. 2007. "Suppression of Long-Branch Attraction Artefacts in the Animal Phylogeny Using a Site-Heterogeneous Model." *BMC Evolutionary Biology* 7 Suppl 1 (February): S4.

Lartillot, Nicolas, and Hervé Philippe. 2004. "A Bayesian Mixture Model for across-Site Heterogeneities in the Amino-Acid Replacement Process." *Molecular Biology and Evolution* 21 (6): 1095–1109.

Lartillot, Nicolas, Nicolas Rodrigue, Daniel Stubbs, and Jacques Richer. 2013. "PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment." *Systematic Biology* 62 (4): 611–15.

Laumer, Christopher E. 2018. “Inferring Ancient Relationships with Genomic Data: A Commentary on Current Practices.” *Integrative and Comparative Biology* 58 (4): 623–39.

Laumer, Christopher E., Nicolas Bekkouche, Alexandra Kerbl, Freya Goetz, Ricardo C. Neves, Martin V. Sørensen, Reinhardt M. Kristensen, et al. 2015. “Spiralian Phylogeny Informs the Evolution of Microscopic Lineages.” *Current Biology: CB* 25 (15): 2000–2006.

Laumer, Christopher E., Rosa Fernández, Sarah Lemer, David Combosch, Kevin M. Kocot, Ana Riesgo, Sónia C. S. Andrade, Wolfgang Sterrer, Martin V. Sørensen, and Gonzalo Giribet. 2019. “Revisiting Metazoan Phylogeny with Genomic Sampling of All Phyla.” *Proceedings. Biological Sciences / The Royal Society* 286 (1906): 20190831.

Laumer, Christopher E., Harald Gruber-Vodicka, Michael G. Hadfield, Vicki B. Pearse, Ana Riesgo, John C. Marioni, and Gonzalo Giribet. 2018. “Support for a Clade of Placozoa and Cnidaria in Genes with Minimal Compositional Bias.” *eLife* 7 (October). https://doi.org/10.7554/eLife.36278.

Lawn, R. M., K. Schwartz, and L. Patthy. 1997. “Convergent Evolution of Apolipoprotein(a) in Primates and Hedgehog.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (22): 11992–97.

Lee, M. S. 1999. “Molecular Phylogenies Become Functional.” *Trends in Ecology & Evolution* 14 (5): 177–78.

Lee, Michael S. Y., and Alessandro Palci. 2015. “Morphological Phylogenetics in the Genomic Age.” *Current Biology: CB* 25 (19): R922–29.

Leonard, Guy, and Thomas A. Richards. 2012. “Genome-Scale Comparative Analysis of Gene Fusions, Gene Fissions, and the Fungal Tree of Life.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (52): 21402–7.

Levine, Mia T., Corbin D. Jones, Andrew D. Kern, Heather A. Lindfors, and David J. Begun. 2006. “Novel Genes Derived from Noncoding DNA in Drosophila Melanogaster Are Frequently X-Linked and Exhibit Testis-Biased Expression.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (26): 9935–39.

Lewis, P. O. 2001. “A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data.” *Systematic Biology* 50 (6): 913–25.

Li, Bin, Tao Qing, Jinhang Zhu, Zhuo Wen, Ying Yu, Ryutaro Fukumura, Yuanting Zheng, Yoichi Gondo, and Leming Shi. 2017. "A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-Seq." *Scientific Reports* 7 (1): 4200.

Li, Li, Christian J. Stoeckert Jr, and David S. Roos. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes." *Genome Research* 13 (9): 2178–89.

Lim, Lee P., Nelson C. Lau, Philip Garrett-Engele, Andrew Grimson, Janell M. Schelter, John Castle, David P. Bartel, Peter S. Linsley, and Jason M. Johnson. 2005. "Microarray Analysis Shows That Some microRNAs Downregulate Large Numbers of Target mRNAs." *Nature* 433 (7027): 769–73.

Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, et al. 2011. "A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals." *Nature* 478 (7370): 476–82.

Linnaeus C. 1758. *Systema naturae per regna tria naturae: secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. 10th ed. Stockholm: Laurentius Salvius.*

Long, M., and C. H. Langley. 1993. "Natural Selection and the Origin of Jingwei, a Chimeric Processed Functional Gene in Drosophila." *Science* 260 (5104): 91–95.

Long, Manyuan, Esther Betrán, Kevin Thornton, and Wen Wang. 2003. "The Origin of New Genes: Glimpses from the Young and Old." *Nature Reviews. Genetics* 4 (11): 865–75.

Luis Villanueva-Cañas, José, Jorge Ruiz-Orera, M. Isabel Agea, Maria Gallo, David Andreu, and M. Mar Albà. 2017. "New Genes and Functional Innovation in Mammals." *Genome Biology and Evolution* 9 (7): 1886–1900.

Lunter, Gerton, Andrea Rocco, Naila Mimouni, Andreas Heger, Alexandre Caldeira, and Jotun Hein. 2008. "Uncertainty in Homology Inferences: Assessing and Improving Genomic Sequence Alignment." *Genome Research* 18 (2): 298–309.

Luo, Yi-Jyun, Miyuki Kanda, Ryo Koyanagi, Kanako Hisata, Tadashi Akiyama, Hirotaka Sakamoto, Tatsuya Sakamoto, and Noriyuki Satoh. 2017. "Nemertean and Phoronid Genomes Reveal Lophotrochozoan Evolution and the Origin of Bilaterian Heads." *Nature Ecology & Evolution* 2 (1): 141–51.

Lynch, M., and J. S. Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” *Science*.

Lynch, Michael, and John S. Conery. 2003. “The Origins of Genome Complexity.” *Science* 302 (5649): 1401–4.

Lynch, Michael, and Vaishali Katju. 2004. “The Altered Evolutionary Trajectories of Gene Duplicates.” *Trends in Genetics: TIG* 20 (11): 544–49.

Lynch, Vincent J., Robert D. Leclerc, Gemma May, and Günter P. Wagner. 2011. “Transposon-Mediated Rewiring of Gene Regulatory Networks Contributed to the Evolution of Pregnancy in Mammals.” *Nature Genetics* 43 (11): 1154–59.

Lyons-Weiler, James, Guy A. Hoelzer, and Robin J. Tausch. 1998. “Optimal Outgroup Analysis.” *Biological Journal of the Linnean Society. Linnean Society of London* 64 (4): 493–511.

Lyu, Yang, Yang Shen, Heng Li, Yuxin Chen, Li Guo, Yixin Zhao, Eric Hungate, Suhua Shi, Chung-I Wu, and Tian Tang. 2014. “New microRNAs in Drosophila--Birth, Death and Cycles of Adaptive Evolution.” *PLoS Genetics* 10 (1): e1004096.

M. O. Dayhoff, R. M. Schwartz. 1978. “Chapter 22: A Model of Evolutionary Change in Proteins.” *In Atlas of Protein Sequence and Structure*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.4315.

Makino, Takashi, and Aoife McLysaght. 2010. “Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9270–74.

Makiuchi, Takashi, Takeshi Nara, Takeshi Annoura, Tetsuo Hashimoto, and Takashi Aoki. 2007. “Occurrence of Multiple, Independent Gene Fusion Events for the Fifth and Sixth Enzymes of Pyrimidine Biosynthesis in Different Eukaryotic Groups.” *Gene* 394 (1-2): 78–86.

Marlétaz, Ferdinand. 2019. “Zoology: Worming into the Origin of Bilaterians.” *Current Biology: CB*.

Marlétaz, Ferdinand, Elise Martin, Yvan Perez, Daniel Papillon, Xavier Caubit, Christopher J. Lowe, Bob Freeman, et al. 2006. “Chaetognath Phylogenomics: A Protostome with Deuterostome-like Development.” *Current Biology: CB* 16 (15): R577–78.

Marlétaz, Ferdinand, Katja T. C. A. Peijnenburg, Taichiro Goto, Noriyuki Satoh, and Daniel S. Rokhsar. 2019. “A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans.” *Current Biology: CB* 29 (2): 312–18.e3.

Marques-Bonet, Tomas, Santhosh Girirajan, and Evan E. Eichler. 2009. “The Origins and Impact of Primate Segmental Duplications.” *Trends in Genetics: TIG* 25 (10): 443–54.

Marra, Marco A., Steven J. M. Jones, Caroline R. Astell, Robert A. Holt, Angela Brooks-Wilson, Yaron S. N. Butterfield, Jaswinder Khattra, et al. 2003. “The Genome Sequence of the SARS-Associated Coronavirus.” *Science* 300 (5624): 1399–1404.

Martín-Durán, José M., Yale J. Passamaneck, Mark Q. Martindale, and Andreas Hejnol. 2016. “The Developmental Basis for the Recurrent Evolution of Deuterostomy and Protostomy.” *Nature Ecology & Evolution* 1 (1): 5.

Martín-Durán, José M., Joseph F. Ryan, Bruno C. Vellutini, Kevin Pang, and Andreas Hejnol. 2017. “Increased Taxon Sampling Reveals Thousands of Hidden Orthologs in Flatworms.” *Genome Research* 27 (7): 1263–72.

Matus, David Q., Richard R. Copley, Casey W. Dunn, Andreas Hejnol, Heather Eccleston, Kenneth M. Halanych, Mark Q. Martindale, and Maximilian J. Telford. 2006. “Broad Taxon and Gene Sampling Indicate That Chaetognaths Are Protostomes.” *Current Biology: CB* 16 (15): R575–76.

McCartney, Ann M., Edel M. Hyland, Paul Cormican, Raymond J. Moran, Andrew E. Webb, Kate D. Lee, Jessica Hernandez-Rodriguez, et al. 2019. “Gene Fusions Derived by Transcriptional Readthrough Are Driven by Segmental Duplication in Human.” *Genome Biology and Evolution* 11 (9): 2678–90.

McClintock, B. 1950. “The Origin and Behavior of Mutable Loci in Maize.” *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344–55.

McCormack, John E., Brant C. Faircloth, Nicholas G. Crawford, Patricia Adair Gowaty, Robb T. Brumfield, and Travis C. Glenn. 2012. “Ultraconserved Elements Are Novel Phylogenomic Markers That Resolve Placental Mammal Phylogeny When Combined with Species-Tree Analysis.” *Genome Research* 22 (4): 746–54.

McInerney, James O., Alan McNally, and Mary J. O'Connell. 2017. "Why Prokaryotes Have Pangenomes." *Nature Microbiology* 2 (March): 17040.

Meredith, Robert W., Jan E. Janečka, John Gatesy, Oliver A. Ryder, Colleen A. Fisher, Emma C. Teeling, Alisha Goodbla, et al. 2011. "Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification." *Science* 334 (6055): 521–24.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92.

Meunier, Julien, Frédéric Lemoine, Magali Soumillon, Angélica Liechti, Manuela Weier, Katerina Guschanski, Haiyang Hu, Philipp Khaitovich, and Henrik Kaessmann. 2013. "Birth and Expression Evolution of Mammalian microRNA Genes." *Genome Research* 23 (1): 34–45.

Mi, S., X. Lee, X. Li, G. M. Veldman, H. Finnerty, L. Racie, E. LaVallie, et al. 2000. "Syncytin Is a Captive Retroviral Envelope Protein Involved in Human Placental Morphogenesis." *Nature* 403 (6771): 785–89.

Molnár, Attila, Frank Schwach, David J. Studholme, Eva C. Thuenemann, and David C. Baulcombe. 2007. "miRNAs Control Gene Expression in the Single-Cell Alga Chlamydomonas Reinhardtii." *Nature* 447 (7148): 1126–29.

Mooi, Randall D., and Anthony C. Gill. 2010. "Phylogenies without Synapomorphies—A Crisis in Fish Systematics: Time to Show Some Character." *Zootaxa* 2450 (1): 26–40.

Moore, Andrew D., Asa K. Björklund, Diana Ekman, Erich Bornberg-Bauer, and Arne Elofsson. 2008. "Arrangements in the Modular Evolution of Proteins." *Trends in Biochemical Sciences* 33 (9): 444–51.

Moreau, R., and K. Dabrowski. 1998. "Body Pool and Synthesis of Ascorbic Acid in Adult Sea Lamprey (Petromyzon Marinus): An Agnathan Fish with Gulonolactone Oxidase Activity." *Proceedings of the National Academy of Sciences of the United States of America* 95 (17): 10279–82.

Morgan, Claire C., Peter G. Foster, Andrew E. Webb, Davide Pisani, James O. McInerney, and Mary J. O'Connell. 2013. "Heterogeneous Models Place the Root of the Placental Mammal Phylogeny." *Molecular Biology and Evolution* 30 (9): 2145–56.

Morris, Simon Conway. 2007. "The Origins and Relationships of Lower Invertebrates." *Lethaia* 16 (1): 92–92.

Nagy, Alinda, and Laszlo Patthy. 2011. "Reassessing Domain Architecture Evolution of Metazoan Proteins: The Contribution of Different Evolutionary Mechanisms." *Genes* 2 (3): 578–98.

Neme, Rafik, and Diethard Tautz. 2013. "Phylogenetic Patterns of Emergence of New Genes Support a Model of Frequent de Novo Evolution." *BMC Genomics* 14 (February): 117.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32 (1): 268–74.

Nichio, Bruno T. L., Jeroniza Nunes Marchaukoski, and Roberto Tadeu Raittz. 2017. "New Tools in Orthology Analysis: A Brief Review of Promising Perspectives." *Frontiers in Genetics* 8 (October): 165.

Nielsen, Claus. 2004. "Trochophora Larvae: Cell-Lineages, Ciliary Bands, and Body Regions. 1. Annelida and Mollusca." *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 302 (1): 35–68.

Nielsen, Claus. 2008. "Six Major Steps in Animal Evolution: Are We Derived Sponge Larvae?" *Evolution & Development* 10 (2): 241–57.

Nikaido, M., A. P. Rooney, and N. Okada. 1999. "Phylogenetic Relationships among Cetartiodactyls Based on Insertions of Short and Long Interpersed Elements: Hippopotamuses Are the Closest Extant Relatives of Whales." *Proceedings of the National Academy of Sciences of the United States of America* 96 (18): 10261–66.

Nikoh, Naruo, John P. McCutcheon, Toshiaki Kudo, Shin-Ya Miyagishima, Nancy A. Moran, and Atsushi Nakabachi. 2010. "Bacterial Genes in the Aphid Genome: Absence of Functional Gene Transfer from Buchnera to Its Host." *PLoS Genetics* 6 (2): e1000827.

O’Leary, Maureen A., Jonathan I. Bloch, John J. Flynn, Timothy J. Gaudin, Andres Giallombardo, Norberto P. Giannini, Suzann L. Goldberg, et al. 2013. “The Placental Mammal Ancestor and the Post-K-Pg Radiation of Placentals.” *Science* 339 (6120): 662–67.

Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Springer, Berlin, Heidelberg.

Oliver, Keith R., and Wayne K. Greene. 2009. “Transposable Elements: Powerful Facilitators of Evolution.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 31 (7): 703–14.

Oliver, Keith R., Jen A. McComb, and Wayne K. Greene. 2013. “Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity.” *Genome Biology and Evolution* 5 (10): 1886–1901.

Paps, Jordi. 2018. “What Makes an Animal? The Molecular Quest for the Origin of the Animal Kingdom.” *Integrative and Comparative Biology* 58 (4): 654–65.

Paps, Jordi, Jaume Baguñà, and Marta Riutort. 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 (10): 2397–2406.

Paps, Jordi, and Peter W. H. Holland. 2018. “Reconstruction of the Ancestral Metazoan Genome Reveals an Increase in Genomic Novelty.” *Nature Communications* 9 (1): 1730.

Park, J., S. A. Teichmann, T. Hubbard, and C. Chothia. 1997. “Intermediate Sequences Increase the Detection of Homology between Sequences.” *Journal of Molecular Biology* 273 (1): 349–54.

Parra, Genís, Alexandre Reymond, Noura Dabbouseh, Emmanouil T. Dermitzakis, Robert Castelo, Timothy M. Thomson, Stylianos E. Antonarakis, and Roderic Guigó. 2006. “Tandem Chimerism as a Means to Increase Protein Complexity in the Human Genome.” *Genome Research* 16 (1): 37–44.

Paten, Benedict, Javier Herrero, Stephen Fitzgerald, Kathryn Beal, Paul Flicek, Ian Holmes, and Ewan Birney. 2008. “Genome-Wide Nucleotide-Level Mammalian Ancestor Reconstruction.” *Genome Research* 18 (11): 1829–43.

Pathmanathan, Jananan Sylvestre, Philippe Lopez, François-Joseph Lapointe, and Eric Bapteste. 2018. “CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection.” *Molecular Biology and Evolution* 35 (1): 252–55.

Patthy, L. 1985. “Evolution of the Proteases of Blood Coagulation and Fibrinolysis by Assembly from Modules.” *Cell* 41 (3): 657–63.

Patthy, L. 1996. “Exon Shuffling and Other Ways of Module Exchange.” *Matrix Biology: Journal of the International Society for Matrix Biology* 15 (5): 301–10; discussion 311–12.

Pawson, T. 1995. “Protein Modules and Signalling Networks.” *Nature* 373 (6515): 573–80.

Pawson, Tony, and Piers Nash. 2003. “Assembly of Cell Regulatory Systems through Protein Interaction Domains.” *Science* 300 (5618): 445–52.

Petersen, Malte, Karen Meusemann, Alexander Donath, Daniel Dowling, Shanlin Liu, Ralph S. Peters, Lars Podsiadlowski, et al. 2017. “Orthograph: A Versatile Tool for Mapping Coding Nucleotide Sequences to Clusters of Orthologous Genes.” *BMC Bioinformatics* 18 (1): 111.

Pett, Walker, Marcin Adamski, Maja Adamska, Warren R. Francis, Michael Eitel, Davide Pisani, and Gert Wörheide. 2019. “The Role of Homology and Orthology in the Phylogenomic Analysis of Metazoan Gene Content.” *Molecular Biology and Evolution* 36 (4): 643–49.

Philippe, H., and P. Lopez. 2001. “On the Conservation of Protein Sequences in Evolution.” *Trends in Biochemical Sciences*.

Philippe, Hervé, Henner Brinkmann, Richard R. Copley, Leonid L. Moroz, Hiroaki Nakano, Albert J. Poustka, Andreas Wallberg, Kevin J. Peterson, and Maximilian J. Telford. 2011. “Acoelomorph Flatworms Are Deuterostomes Related to Xenoturbella.” *Nature* 470 (7333): 255–58.

Philippe, Hervé, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J. Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. 2011. “Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough.” *PLoS Biology* 9 (3): e1000602.

Philippe, Hervé, Anne Chenuil, and André Adoutte. 1994. “Can the Cambrian Explosion Be Inferred through Molecular Phylogeny?” *Development* 1994 (Supplement): 15–25.

Philippe, Hervé, Romain Derelle, Philippe Lopez, Kerstin Pick, Carole Borchiellini, Nicole Boury-Esnault, Jean Vacelet, et al. 2009. “Phylogenomics Revives Traditional Views on Deep Animal Relationships.” *Current Biology: CB* 19 (8): 706–12.

Philippe, Hervé, Nicolas Lartillot, and Henner Brinkmann. 2005. “Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia.” *Molecular Biology and Evolution* 22 (5): 1246–53.

Philippe, Hervé, Albert J. Poustka, Marta Chiodin, Katharina J. Hoff, Christophe Dessimoz, Bartlomiej Tomiczek, Philipp H. Schiffer, et al. 2019. “Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria.” *Current Biology: CB* 29 (11): 1818–26.e6.

Philippe, Hervé, and Béatrice Roure. 2011. “Difficult Phylogenetic Questions: More Data, Maybe; Better Methods, Certainly.” *BMC Biology*.

Piacentini, Lucia, Laura Fanti, Valeria Specchia, Maria Pia Bozzetti, Maria Berloco, Gino Palumbo, and Sergio Pimpinelli. 2014. “Transposons, Environmental Changes, and Heritable Induced Phenotypic Variability.” *Chromosoma* 123 (4): 345–54.

Picciani, Natasha, Jamie R. Kerlin, Noemie Sierra, Andrew J. M. Swafford, M. Desmond Ramirez, Nickellaus G. Roberts, Johanna T. Cannon, Marymegan Daly, and Todd H. Oakley. 2018. “Prolific Origination of Eyes in Cnidaria with Co-Option of Non-Visual Opsins.” *Current Biology: CB* 28 (15): 2413–19.e4.

Piriyapongsa, Jittima, Leonardo Mariño-Ramírez, and I. King Jordan. 2007. “Origin and Evolution of Human microRNAs from Transposable Elements.” *Genetics* 176 (2): 1323–37.

Pisani, Davide, Robert Carton, Lahcen I. Campbell, Wasiu A. Akanni, Eoin Mulville, and Omar Rota-Stabelli. 2013. “An Overview of Arthropod Genomics, Mitogenomics, and the Evolutionary Origins of the Arthropod Proteome.” In *Arthropod Biology and Evolution: Molecules, Development, Morphology*, edited by Alessandro Minelli, Geoffrey Boxshall, and Giuseppe Fusco, 41–61. Berlin, Heidelberg: Springer Berlin Heidelberg.

Pisani, Davide, Walker Pett, Martin Dohrmann, Roberto Feuda, Omar Rota-Stabelli, Hervé Philippe, Nicolas Lartillot, and Gert Wörheide. 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 (50): 15402–7.

Platt, Roy N., 2nd, Michael W. Vandewege, Colin Kern, Carl J. Schmidt, Federico G. Hoffmann, and David A. Ray. 2014. “Large Numbers of Novel miRNAs Originate from DNA Transposons and Are Coincident with a Large Species Radiation in Bats.” *Molecular Biology and Evolution* 31 (6): 1536–45.

Pollock, David D., Derrick J. Zwickl, Jimmy A. McGuire, and David M. Hillis. 2002. “Increased Taxon Sampling Is Advantageous for Phylogenetic Inference.” *Systematic Biology* 51 (4): 664–71.

Ponting, Chris P., Peter L. Oliver, and Wolf Reik. 2009. “Evolution and Functions of Long Noncoding RNAs.” *Cell* 136 (4): 629–41.

Powell, Sean, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, et al. 2012. “eggNOG v3.0: Orthologous Groups Covering 1133 Organisms at 41 Different Taxonomic Ranges.” *Nucleic Acids Research* 40 (Database issue): D284–89.

Pozdnyakov, Igor, Agniya Sokolova, Alexander Ereskovsky, and Sergey Karpov. 2017. “Kinetid Structure of Choanoflagellates and Choanocytes of Sponges Does Not Support Their Close Relationship.” *Protistology* 11 (4): 248–64.

Prabh, Neel, Waltraud Roeseler, Hanh Witte, Gabi Eberhardt, Ralf J. Sommer, and Christian Rödelsperger. 2018. “Deep Taxon Sampling Reveals the Evolutionary Dynamics of Novel Gene Families in Pristionchus Nematodes.” *Genome Research* 28 (11): 1664–74.

Quah, Shan, Jerome H. L. Hui, and Peter W. H. Holland. 2015. “A Burst of miRNA Innovation in the Early Evolution of Butterflies and Moths.” *Molecular Biology and Evolution* 32 (5): 1161–74.

Queiroz, Alan de, and John Gatesy. 2007. “The Supermatrix Approach to Systematics.” *Trends in Ecology & Evolution* 22 (1): 34–41.

Ranwez, Vincent, and Nathalie Chantret. 2020. “Strengths and Limits of Multiple Sequence Alignment and Filtering Methods.” *Phylogenetics in the Genomic Era*, 2.2:1–2.2:36.

Reinhardt, Josephine A., Betty M. Wanjiru, Alicia T. Brant, Perot Saelao, David J. Begun, and Corbin D. Jones. 2013. “De Novo ORFs in Drosophila Are Important to Organismal

Fitness and Evolved Rapidly from Previously Non-Coding Sequences." *PLoS Genetics* 9 (10): e1003860.

Reis, Mario dos, Yuttapong Thawornwattana, Konstantinos Angelis, Maximilian J. Telford, Philip C. J. Donoghue, and Ziheng Yang. 2015. "Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales." *Current Biology: CB* 25 (22): 2939–50.

Richter, Daniel J., Parinaz Fozouni, Michael B. Eisen, and Nicole King. 2018. "Gene Family Innovation, Conservation and Loss on the Animal Stem Lineage." *eLife* 7 (May). https://doi.org/10.7554/eLife.34226.

Robertson, Helen E., François Lapraz, Bernhard Egger, Maximilian J. Telford, and Philipp H. Schiffer. 2017. "The Mitochondrial Genomes of the Acoelomorph Worms Paratomella Rubra, Isodiametra Pulchra and Archaphanostoma Ylvae." *Scientific Reports* 7 (1): 1847.

Rödelsperger, Christian, Marina Athanasouli, Maša Lenuzzi, Tobias Theska, Shuai Sun, Mohannad Dardiry, Sara Wighard, Wen Hu, Devansh Raj Sharma, and Ziduan Han. 2019. "Crowdsourcing and the Feasibility of Manual Gene Annotation: A Pilot Study in the Nematode Pristionchus Pacificus." *Scientific Reports* 9 (1): 18789.

Rogers, Rebekah L., Trevor Bedford, and Daniel L. Hartl. 2009. "Formation and Longevity of Chimeric and Duplicate Genes in Drosophila Melanogaster." *Genetics* 181 (1): 313–22.

Rogers, Rebekah L., and Daniel L. Hartl. 2012. "Chimeric Genes as a Source of Rapid Evolution in Drosophila Melanogaster." *Molecular Biology and Evolution* 29 (2): 517–29.

Rokas, A., and P. W. Holland. 2000. "Rare Genomic Changes as a Tool for Phylogenetics." *Trends in Ecology & Evolution* 15 (11): 454–59.

Rokas, Antonis, and Sean B. Carroll. 2008. "Frequent and Widespread Parallel Evolution of Protein Sequences." *Molecular Biology and Evolution* 25 (9): 1943–53.

Rosa, R. de, J. K. Grenier, T. Andreeva, C. E. Cook, A. Adoutte, M. Akam, S. B. Carroll, and G. Balavoine. 1999. "Hox Genes in Brachiopods and Priapulids and Protostome Evolution." *Nature* 399 (6738): 772–76.

Roure, Béatrice, Denis Baurain, and Hervé Philippe. 2013. “Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets.” *Molecular Biology and Evolution* 30 (1): 197–214.

Roure, Béatrice, Naiara Rodriguez-Ezpeleta, and Hervé Philippe. 2007. “SCaFoS: A Tool for Selection, Concatenation and Fusion of Sequences for Phylogenomics.” *BMC Evolutionary Biology* 7 Suppl 1 (February): S2.

Rouse, Greg W., Nerida G. Wilson, Jose I. Carvajal, and Robert C. Vrijenhoek. 2016. “New Deep-Sea Species of Xenoturbella and the Position of Xenacoelomorpha.” *Nature* 530 (7588): 94–97.

Ruiz-Orera, Jorge, Xavier Messeguer, Juan Antonio Subirana, and M. Mar Alba. 2014. “Long Non-Coding RNAs as a Source of New Peptides.” *eLife* 3 (September): e03523.

Ruiz-Trillo, I., and J. Paps. 2016. “Acoelomorpha: Earliest Branching Bilaterians or Deuterostomes?” *Organisms, Diversity & Evolution*. https://link.springer.com/article/10.1007/s13127-015-0239-1.

Ruiz-Trillo, Iñaki, Marta Riutort, H. Matthew Fourcade, Jaume Baguñà, and Jeffrey L. Boore. 2004. “Mitochondrial Genome Data Support the Basal Position of Acoelomorpha and the Polyphyly of the Platyhelminthes.” *Molecular Phylogenetics and Evolution* 33 (2): 321–32.

Ruiz-Trillo, Iñaki, Andrew J. Roger, Gertraud Burger, Michael W. Gray, and B. Franz Lang. 2008. “A Phylogenomic Investigation into the Origin of Metazoa.” *Molecular Biology and Evolution* 25 (4): 664–72.

Ryan, Joseph F., Kevin Pang, Christine E. Schnitzler, Anh-Dao Nguyen, R. Travis Moreland, David K. Simmons, Bernard J. Koch, et al. 2013. “The Genome of the Ctenophore Mnemiopsis Leidyi and Its Implications for Cell Type Evolution.” *Science* 342 (6164): 1242592.

Sacerdot, Christine, Alexandra Louis, Céline Bon, Camille Berthelot, and Hugues Roest Crollius. 2018. “Chromosome Evolution at the Origin of the Ancestral Vertebrate Genome.” *Genome Biology* 19 (1): 166.

Sagane, Yoshimasa, Karin Zech, Jean-Marie Bouquet, Martina Schmid, Ugur Bal, and Eric M. Thompson. 2010. “Functional Specialization of Cellulose Synthase Genes of Prokaryotic Origin in Chordate Larvaceans.” *Development* 137 (9): 1483–92.

Saitou, N., and M. Nei. 1987. “The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees.” *Molecular Biology and Evolution* 4 (4): 406–25.

Salichos, Leonidas, and Antonis Rokas. 2011. “Evaluating Ortholog Prediction Algorithms in a Yeast Model Clade.” *PloS One* 6 (4): e18755.

Salipante, Stephen J., and Marshall S. Horwitz. 2006. “Phylogenetic Fate Mapping.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (14): 5448–53.

Sayah, David M., Elena Sokolskaja, Lionel Berthoux, and Jeremy Luban. 2004. “Cyclophilin A Retrotransposition into TRIM5 Explains Owl Monkey Resistance to HIV-1.” *Nature* 430 (6999): 569–73.

Schierwater, Bernd, Michael Eitel, Wolfgang Jakob, Hans-Jürgen Osigus, Heike Hadrys, Stephen L. Dellaporta, Sergios-Orestis Kolokotronis, and Rob Desalle. 2009. “Concatenated Analysis Sheds Light on Early Metazoan Evolution and Fuels a Modern ‘Urmetazoon’ Hypothesis.” *PLoS Biology* 7 (1): e20.

Schlötterer, Christian. 2015. “Genes from Scratch--the Evolutionary Fate of de Novo Genes.” *Trends in Genetics: TIG* 31 (4): 215–19.

Schmitz, Jonathan F., and Erich Bornberg-Bauer. 2017. “Fact or Fiction: Updates on How Protein-Coding Genes Might Emerge de Novo from Previously Non-Coding DNA.” *F1000Research* 6 (January): 57.

Scotland, Robert W. 2011. “What Is Parallelism?” *Evolution & Development* 13 (2): 214–27.

Scotland, Robert W., Richard G. Olmstead, and Jonathan R. Bennett. 2003. “Phylogeny Reconstruction: The Role of Morphology.” *Systematic Biology* 52 (4): 539–48.

Sebé-Pedrós, Arnau, Bernard M. Degnan, and Iñaki Ruiz-Trillo. 2017. “The Origin of Metazoa: A Unicellular Perspective.” *Nature Reviews. Genetics* 18 (8): 498–512.

Sémon, Marie, and Kenneth H. Wolfe. 2007. “Consequences of Genome Duplication.” *Current Opinion in Genetics & Development* 17 (6): 505–12.

Sempere, Lorenzo F., Pedro Martinez, Charles Cole, Jaume Baguñà, and Kevin J. Peterson. 2007. “Phylogenetic Distribution of microRNAs Supports the Basal Position of Acoel Flatworms and the Polyphyly of Platyhelminthes.” *Evolution & Development* 9 (5): 409–15.

She, Xinwei, Julie E. Horvath, Zhaoshi Jiang, Ge Liu, Terrence S. Furey, Laurie Christ, Royden Clark, et al. 2004. “The Structure and Evolution of Centromeric Transition Regions within the Human Genome.” *Nature* 430 (7002): 857–64.

Shedlock, A. M., and N. Okada. 2000. “SINE Insertions: Powerful Tools for Molecular Systematics.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 22 (2): 148–60.

Shen, Xing-Xing, Chris Todd Hittinger, and Antonis Rokas. 2017. “Contentious Relationships in Phylogenomic Studies Can Be Driven by a Handful of Genes.” *Nature Ecology & Evolution* 1 (5): 126.

Shen, Xing-Xing, Dana A. Opulente, Jacek Kominek, Xiaofan Zhou, Jacob L. Steenwyk, Kelly V. Buh, Max A. B. Haase, et al. 2018. “Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum.” *Cell* 175 (6): 1533–45.e20.

Shubin, N., C. Tabin, and S. Carroll. 1997. “Fossils, Genes and the Evolution of Animal Limbs.” *Nature* 388 (6643): 639–48.

Shubin, Neil, Cliff Tabin, and Sean Carroll. 2009. “Deep Homology and the Origins of Evolutionary Novelty.” *Nature* 457 (7231): 818–23.

Simakov, Oleg, Ferdinand Marletaz, Sung-Jin Cho, Eric Edsinger-Gonzales, Paul Havlak, Uffe Hellsten, Dian-Han Kuo, et al. 2013. “Insights into Bilaterian Evolution from Three Spiralian Genomes.” *Nature* 493 (7433): 526–31.

Simakov, Oleg, Ferdinand Marlétaz, Jia-Xing Yue, Brendan O’Connell, Jerry Jenkins, Alexander Brandt, Robert Calef, et al. 2020. “Deeply Conserved Synteny Resolves Early Events in Vertebrate Evolution.” *Nature Ecology & Evolution*, April. https://doi.org/10.1038/s41559-020-1156-z.

Simion, Paul, Nicolas Bekkouche, Muriel Jager, Eric Quéinnec, and Michaël Manuel. 2015. “Exploring the Potential of Small RNA Subunit and ITS Sequences for Resolving Phylogenetic Relationships within the Phylum Ctenophora.” *Zoology* 118 (2): 102–14.

Simion, Paul, Hervé Philippe, Denis Baurain, Muriel Jager, Daniel J. Richter, Arnaud Di Franco, Béatrice Roure, et al. 2017. “A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals.” *Current Biology: CB* 27 (7): 958–67.

Siu-Ting, Karen, María Torres-Sánchez, Diego San Mauro, David Wilcockson, Mark Wilkinson, Davide Pisani, Mary J. O’Connell, and Christopher J. Creevey. 2019. “Inadvertent Paralog Inclusion Drives Artifactual Topologies and Timetree Estimates in Phylogenomics.” *Molecular Biology and Evolution* 36 (6): 1344–56.

Sogabe, Shunsuke, William L. Hatleberg, Kevin M. Kocot, Tahsha E. Say, Daniel Stoupin, Kathrein E. Roper, Selene L. Fernandez-Valverde, Sandie M. Degnan, and Bernard M. Degnan. 2019. “Pluripotency and the Origin of Animal Multicellularity.” *Nature* 570 (7762): 519–22.

Somarelli, Jason A., Kathryn E. Ware, Rumen Kostadinov, Jeffrey M. Robinson, Hakima Amri, Mones Abu-Asab, Nicolaas Fourie, Rui Diogo, David Swofford, and Jeffrey P. Townsend. 2017. “PhyloOncology: Understanding Cancer through Phylogenetic Analysis.” *Biochimica et Biophysica Acta, Reviews on Cancer* 1867 (2): 101–8.

Song, Nan, Jacob M. Joseph, George B. Davis, and Dannie Durand. 2008. “Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins.” *PLoS Computational Biology* 4 (4): e1000063.

Song, Sen, Liang Liu, Scott V. Edwards, and Shaoyuan Wu. 2012. “Resolving Conflict in Eutherian Mammal Phylogeny Using Phylogenomics and the Multispecies Coalescent Model.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (37): 14942–47.

Sonnhammer, E. L., and D. Kahn. 1994. “Modular Arrangement of Proteins as Inferred from Analysis of Homology.” *Protein Science: A Publication of the Protein Society* 3 (3): 482–92.

Sonnhammer, Erik L. L., and Eugene V. Koonin. 2002. “Orthology, Paralogy and Proposed Classification for Paralog Subtypes.” *Trends in Genetics: TIG* 18 (12): 619–20.

Sonnhammer, Erik L. L., and Gabriel Östlund. 2015. “InParanoid 8: Orthology Analysis between 273 Proteomes, Mostly Eukaryotic.” *Nucleic Acids Research* 43 (Database issue): D234–39.

Sperling, Erik A., Davide Pisani, and Kevin J. Peterson. 2011. “Molecular Paleobiological Insights into the Origin of the Brachiopoda.” *Evolution & Development* 13 (3): 290–303.

Springer, Mark S., and John Gatesy. 2018. “On the Importance of Homology in the Age of Phylogenomics.” https://pubag.nal.usda.gov/catalog/5957814.

Stapley, Jessica, Philine G. D. Feulner, Susan E. Johnston, Anna W. Santure, and Carole M. Smadja. 2017. “Variation in Recombination Frequency and Distribution across Eukaryotes: Patterns and Processes.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372 (1736). https://doi.org/10.1098/rstb.2016.0455.

Stechmann, Alexandra, and Thomas Cavalier-Smith. 2002. “Rooting the Eukaryote Tree by Using a Derived Gene Fusion.” *Science* 297 (5578): 89–91.

Stedman, Hansell H., Benjamin W. Kozyak, Anthony Nelson, Danielle M. Thesier, Leonard T. Su, David W. Low, Charles R. Bridges, Joseph B. Shrager, Nancy Minugh-Purvis, and Marilyn A. Mitchell. 2004. “Myosin Gene Mutation Correlates with Anatomical Changes in the Human Lineage.” *Nature* 428 (6981): 415–18.

Stirnimann, Christian U., Evangelia Petsalaki, Robert B. Russell, and Christoph W. Müller. 2010. “WD40 Proteins Propel Cellular Networks.” *Trends in Biochemical Sciences* 35 (10): 565–74.

Stover, Nicholas A., André R. O. Cavalcanti, Anya J. Li, Brian C. Richardson, and Laura F. Landweber. 2005. “Reciprocal Fusions of Two Genes in the Formaldehyde Detoxification Pathway in Ciliates and Diatoms.” *Molecular Biology and Evolution* 22 (7): 1539–42.

Stover, Nicholas A., Thomas A. Dixon, and Andre R. O. Cavalcanti. 2011. “Multiple Independent Fusions of Glucose-6-Phosphate Dehydrogenase with Enzymes in the Pentose Phosphate Pathway.” *PloS One* 6 (8): e22269.

Streicher, Jeffrey W., James A. Schulte 2nd, and John J. Wiens. 2016. “How Should Genes and Taxa Be Sampled for Phylogenomic Analyses with Missing Data? An Empirical Study in Iguanian Lizards.” *Systematic Biology* 65 (1): 128–45.

Suga, Hiroshi, Zehua Chen, Alex de Mendoza, Arnau Sebé-Pedrós, Matthew W. Brown, Eric Kramer, Martin Carr, et al. 2013. “The Capsaspora Genome Reveals a Complex Unicellular Prehistory of Animals.” *Nature Communications* 4: 2325.

Suh, Alexander, Martin Paus, Martin Kiefmann, Gennady Churakov, Franziska Anni Franke, Jürgen Brosius, Jan Ole Kriegs, and Jürgen Schmitz. 2011. “Mesozoic Retroposons Reveal Parrots as the Closest Living Relatives of Passerine Birds.” *Nature Communications* 2 (August): 443.

Susko, Edward. 2010. “First-Order Correct Bootstrap Support Adjustments for Splits That Allow Hypothesis Testing When Using Maximum Likelihood Estimation.” *Molecular Biology and Evolution* 27 (7): 1621–29.

Susko, Edward, and Andrew J. Roger. 2007. “On Reduced Amino Acid Alphabets for Phylogenetic Inference.” *Molecular Biology and Evolution* 24 (9): 2139–50.

Szöllősi, Gergely J., Eric Tannier, Vincent Daubin, and Bastien Boussau. 2015. “The Inference of Gene Trees with Species Trees.” *Systematic Biology* 64 (1): e42–62.

Tarver, James E., Alexandre Cormier, Natalia Pinzón, Richard S. Taylor, Wilfrid Carré, Martina Strittmatter, Hervé Seitz, Susana M. Coelho, and J. Mark Cock. 2015. “microRNAs and the Evolution of Complex Multicellularity: Identification of a Large, Diverse Complement of microRNAs in the Brown Alga Ectocarpus.” *Nucleic Acids Research* 43 (13): 6384–98.

Tarver, James E., Mario Dos Reis, Siavash Mirarab, Raymond J. Moran, Sean Parker, Joseph E. O’Reilly, Benjamin L. King, et al. 2016. “The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference.” *Genome Biology and Evolution* 8 (2): 330–44.

Tarver, James E., Erik A. Sperling, Audrey Nailor, Alysha M. Heimberg, Jeffrey M. Robinson, Benjamin L. King, Davide Pisani, Philip C. J. Donoghue, and Kevin J. Peterson. 2013.

"miRNAs: Small Genes with Big Potential in Metazoan Phylogenetics." *Molecular Biology and Evolution* 30 (11): 2369–82.

Tarver, James E., Richard S. Taylor, Mark N. Puttick, Graeme T. Lloyd, Walker Pett, Bastian Fromm, Bettina E. Schirrmeister, Davide Pisani, Kevin J. Peterson, and Philip C. J. Donoghue. 2018. "Well-Annotated microRNAomes Do Not Evidence Pervasive miRNA Loss." *Genome Biology and Evolution* 10 (6): 1457–70.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. "A Genomic Perspective on Protein Families." *Science* 278 (5338): 631–37.

Tautz, Diethard, and Tomislav Domazet-Lošo. 2011. "The Evolutionary Origin of Orphan Genes." *Nature Reviews. Genetics* 12 (10): 692–702.

Tavare, S. 1986. "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences." *Some Mathematical Questions in Biology / DNA Sequence Analysis Edited by Robert M. Miura*. http://agris.fao.org/agris-search/search.do?recordID=US201301755037.

Telford, Maximilian J., Graham E. Budd, and Hervé Philippe. 2015. "Phylogenomic Insights into Animal Evolution." *Current Biology: CB* 25 (19): R876–87.

"The Deep Evolution of Metazoan microRNAs." n.d. Accessed March 17, 2020. https://scholar.harvard.edu/sperling/publications/deep-evolution-metazoan-micrornas.

Thomas, Gregg W. C., Elias Dohmen, Daniel S. T. Hughes, Shwetha C. Murali, Monica Poelchau, Karl Glastad, Clare A. Anstead, et al. 2020. "Gene Content Evolution in the Arthropods." *Genome Biology* 21 (1): 15.

Thomson, Robert C., Ian J. Wang, and Jarrett R. Johnson. 2010. "Genome-Enabled Development of DNA Markers for Ecology, Evolution and Conservation." *Molecular Ecology* 19 (11): 2184–95.

Thomson, T. M., J. J. Lozano, N. Loukili, R. Carrió, F. Serras, B. Cormand, M. Valeri, et al. 2000. "Fusion of the Human Gene for the Polyubiquitination Coeffector UEV1 with Kua, a Newly Identified Gene." *Genome Research* 10 (11): 1743–56.

Toll-Riera, Macarena, Nina Bosch, Nicolás Bellora, Robert Castelo, Lluis Armengol, Xavier Estivill, and M. Mar Albà. 2009. “Origin of Primate Orphan Genes: A Comparative Genomics Approach.” *Molecular Biology and Evolution* 26 (3): 603–12.

Tordai, Hedvig, Alinda Nagy, Krisztina Farkas, László Bányai, and László Patthy. 2005. “Modules, Multidomain Proteins and Organismic Complexity: Mobile Domains.” *The FEBS Journal* 272 (19): 5064–78.

Townsend, Jeffrey P., Francesc López-Giráldez, and Robert Friedman. 2008. “The Phylogenetic Informativeness of Nucleotide and Amino Acid Sequences for Reconstructing the Vertebrate Tree.” *Journal of Molecular Evolution* 67 (5): 437–47.

Van de Peer, Yves, Steven Maere, and Axel Meyer. 2009. “The Evolutionary Significance of Ancient Genome Duplications.” *Nature Reviews. Genetics* 10 (10): 725–32.

Van Oss, Stephen Branden, and Anne-Ruxandra Carvunis. 2019. “De Novo Gene Birth.” *PLoS Genetics* 15 (5): e1008160.

Vogel, Christine, and Cyrus Chothia. 2006. “Protein Family Expansions and Biological Complexity.” *PLoS Computational Biology* 2 (5): e48.

Voinnet, Olivier. 2009. “Origin, Biogenesis, and Activity of Plant microRNAs.” *Cell* 136 (4): 669–87.

Volff, Jean-Nicolas. 2006. “Turning Junk into Gold: Domestication of Transposable Elements and the Creation of New Genes in Eukaryotes.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 28 (9): 913–22.

Wagner, Günter P. 1996. “Homologues, Natural Kinds and the Evolution of Modularity.” *Integrative and Comparative Biology* 36 (1): 36–43.

Wake, David B., Marvalee H. Wake, and Chelsea D. Specht. 2011. “Homoplasy: From Detecting Pattern to Determining Process and Mechanism of Evolution.” *Science* 331 (6020): 1032–35.

Walker, Joseph F., Xing-Xing Shen, Antonis Rokas, Stephen A. Smith, and Edwige Moyroud. 2020. “Disentangling Biological and Analytical Factors That Give Rise to Outlier Genes in Phylogenomic Matrices.” https://doi.org/10.1101/2020.04.20.049999.

Wang, Jingwen, Liselotte Vesterlund, Juha Kere, and Hong Jiao. 2016. "Identification of Novel Transcribed Regions in Zebrafish (Danio Rerio) Using RNA-Sequencing." *PloS One* 11 (7): e0160197.

Wang, Minglei, and Gustavo Caetano-Anollés. 2006. "Global Phylogeny Determined by the Combination of Protein Domains in Proteomes." *Molecular Biology and Evolution* 23 (12): 2444–54.

Wanninger, Andreas. 2016. "Twenty Years into the 'new Animal Phylogeny': Changes and Challenges." *Organisms, Diversity & Evolution* 16 (2): 315–18.

Whelan, Nathan V., Kevin M. Kocot, Leonid L. Moroz, and Kenneth M. Halanych. 2015. "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 (18): 5773–78.

Whelan, Nathan V., Kevin M. Kocot, Tatiana P. Moroz, Krishanu Mukherjee, Peter Williams, Gustav Paulay, Leonid L. Moroz, and Kenneth M. Halanych. 2017. "Ctenophore Relationships and Their Placement as the Sister Group to All Other Animals." *Nature Ecology & Evolution* 1 (11): 1737–46.

White-Cooper, Helen, and Irwin Davidson. 2011. "Unique Aspects of Transcription Regulation in Male Germ Cells." *Cold Spring Harbor Perspectives in Biology* 3 (7). https://doi.org/10.1101/cshperspect.a002626.

Wiens, John. 2004. "The Role of Morphological Data in Phylogeny Reconstruction." *Systematic Biology* 53 (4): 653–61.

Wilberg, Eric W. 2015. "What's in an Outgroup? The Impact of Outgroup Choice on the Phylogenetic Position of Thalattosuchia (Crocodylomorpha) and the Origin of Crocodyliformes." *Systematic Biology* 64 (4): 621–37.

Wilkinson, Mark, James O. McInerney, Robert P. Hirt, Peter G. Foster, and T. Martin Embley. 2007. "Of Clades and Clans: Terms for Phylogenetic Relationships in Unrooted Trees." *Trends in Ecology & Evolution* 22 (3): 114–15.

Wissler, Lothar, Jürgen Gadau, Daniel F. Simola, Martin Helmkampf, and Erich Bornberg-Bauer. 2013. "Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes." *Genome Biology and Evolution* 5 (2): 439–55.

Wolf, Yuri I., and Eugene V. Koonin. 2013. “Genome Reduction as the Dominant Mode of Evolution.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 35 (9): 829–37.

Wu, Dong-Dong, David M. Irwin, and Ya-Ping Zhang. 2011. “De Novo Origin of Human Protein-Coding Genes.” *PLoS Genetics* 7 (11): e1002379.

Xie, Chen, Cemalettin Bekpen, Sven Künzel, Maryam Keshavarz, Rebecca Krebs-Wheaton, Neva Skrabar, Kristian Karsten Ullrich, and Diethard Tautz. 2019. “A de Novo Evolved Gene in the House Mouse Regulates Female Pregnancy Cycles.” *eLife* 8 (August). https://doi.org/10.7554/eLife.44392.

Xie, Wangang, Paul O. Lewis, Yu Fan, Lynn Kuo, and Ming-Hui Chen. 2011. “Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection.” *Systematic Biology* 60 (2): 150–60.

Yang, Song, Russell F. Doolittle, and Philip E. Bourne. 2005. “Phylogeny Determined by Protein Domain Content.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (2): 373–78.

Yang, Z. 1994. “Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods.” *Journal of Molecular Evolution* 39 (3): 306–14.

Yang, Ziheng. 1996. “Among-Site Rate Variation and Its Impact on Phylogenetic Analyses.” *Trends in Ecology & Evolution* 11 (9): 367–72.

Yang, Ziheng, and Bruce Rannala. 2012. “Molecular Phylogenetics: Principles and Practice.” *Nature Reviews. Genetics* 13 (5): 303–14.

Zhang, Jianzhi, Ya-Ping Zhang, and Helene F. Rosenberg. 2002. “Adaptive Evolution of a Duplicated Pancreatic Ribonuclease Gene in a Leaf-Eating Monkey.” *Nature Genetics* 30 (4): 411–15.

Zhang, Zhi-Qiang. 2013. “Animal Biodiversity: An Update of Classification and Diversity in 2013. In : Zhang, Z.-Q. (Ed.) Animal Biodiversity: An Outline of Higher-Level Classification and Survey of Taxonomic Richness (Addenda 2013).” *Zootaxa* 3703 (1): 5–11.

Zhou, Qi, Guojie Zhang, Yue Zhang, Shiyu Xu, Ruoping Zhao, Zubing Zhan, Xin Li, Yun Ding, Shuang Yang, and Wen Wang. 2008. “On the Origin of New Genes in Drosophila.” *Genome Research* 18 (9): 1446–55.

Zhuang, Xuan, Chun Yang, Katherine R. Murphy, and C-H Christina Cheng. 2019. “Molecular Mechanism and History of Non-Sense to Sense Evolution of Antifreeze Glycoprotein Gene in Northern Gadids.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (10): 4400–4405.

Zmasek, Christian M., and Adam Godzik. 2011. “Strong Functional Patterns in the Evolution of Eukaryotic Genomes Revealed by the Reconstruction of Ancestral Protein Domain Repertoires.” *Genome Biology* 12 (1): R4.

Zmasek, Christian M., and Adam Godzik. 2012. “This Déjà vu Feeling--Analysis of Multidomain Protein Evolution in Eukaryotic Genomes.” *PLoS Computational Biology* 8 (11): e1002701.

Zwickl, Derrick J., and David M. Hillis. 2002. “Increased Taxon Sampling Greatly Reduces Phylogenetic Error.” *Systematic Biology* 51 (4): 588–98.