Investigating the linear and non-linear mechanisms of protein coding innovation across the animal kingdom

Raymond Moran B.Sc. (Hons) Genetics and Cell Biology

A thesis presented to Dublin City University for the Degree
of
Doctor of Philosophy

Internal Supervisor: Dr. Tim Downing
School of Biotechnology, Dublin City University
External Supervisor: Dr. Mary J. O'Connell
School of Biology, Faculty of Biological Sciences, University of Leeds

August 2017

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of PhD is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:
(Candidate) ID No.:
Date:

Acknowledgments

Completing a PhD is not something that is often done alone. Although my name is on the front of this thesis, there are so many people involved in making this happen (too many to mention personally). My family, friends and lab-mates have all got me through the mentally difficult times of this work. There are a few that need to be mentioned.

My friends, family and lab-mates, you all got me through the difficult times. There were times that I was mentally struggling with life as a student for so long. You all dragged me through them times. You all believe in me so much, I could never give up on myself. For that, I'll be forever grateful.

My supervisor Mary is the person that inspired me to do a PhD. Her lectures as an undergraduate got me excited. Her passion and positive attitude made me excited. I would not be here now if it wasn't for Mary. From undergraduate she has guided me and advised me on how to achieve my goals. In my time doing a PhD with Mary I have gained a lifelong friend and mentor. I am incredibly fortunate to have such a warm, positive and intelligent friend to guide me through my PhD. I could not have found a better and kinder person to learn from. I have no doubt that Mary has made me a better scientist, but much more importantly – a better person.

Lastly, one person who has gone through this entire process with me is my girlfriend Leanne. You are the single most important influence in my life and have been for the past 10 years. I push myself to be better for you, for us. You made all the problems that seemed so big to me, seem small. I couldn't have asked for a more supportive partner to get me through this. You are and will always be my rock and I know you will guide me through anything.

Table of Contents

Figures and Tables

Abbreviations

AA	Amino acid
ASRV	Among-site rate variation
BF	Bayes Factor
CC	Clustering coefficient
CDS	Coding sequence
CI	Consistency Index
DV	Dorsoventral
ERV	Endogenous retrovirus
GO	Gene Ontology
GTR	General Time reversible
HGT	Horizontal gene transfer
HMM	Hidden Markov models
ILS	Incomplete lineage sorting
JC	Jukes and Cantor
JTT	Jones-Taylor-Thornton
K	Degree
<Knn>	Average degree of nearest neighbours
LBA	Long branch attraction
LM	Likelihood mapping
MCMC	Markov-chain Monte Carlo
MCMCMC	Metropolis-coupled MCMC
ML	Maximum likelihood
MPI	Message passing interface
MSA	Multiple sequence alignment
MY	Million Years
MYA	Million years ago
OUT	operational taxonomic unit
PIN	Protein interaction network
PPS	Posterior predictive simulations
SGO	Single gene ortholog
SSN	Sequence similarity network
TE	Transposable element

WAG	Whelan and Goldman

Abstract: Raymond Moran: investigating the linear and non-linear mechanisms of protein coding innovation across the animal kingdom

In this thesis, evolutionary models and frameworks for studying protein coding sequence evolution in the *Metazoa* are explored. The *Metazoa* refers to all multicellular animals. The models and frameworks applied are phylogeny and network based – we propose that both approaches are necessary to understand the various mechanisms of protein evolution that can be involved in protein family evolution. Firstly, we apply a phylogenetic approach to assess whether the additional parameterisation and computational requirements associated with heterogeneous modelling can be justified in terms of model fit and adequacy. The question addressed in Chapter 2 is whether we can accurately and adequately model the position of the placental mammal root. We show that model misspecification has resulted in the conflicting positions for the root of the placental mammal phylogeny in the recent literature. Therefore, inadequate models can result in wrong inferences about the evolutionary history of the sequences. Indeed, phylogenetic models may not be appropriate for the evolutionary process at work, for example in the case of recombinatorial processes. Therefore, Chapters 3 and 4 focus on the application of network theory, and specifically sequence similarity networks, to model the evolution of protein coding sequences by gene fission/fusion and domain shuffling – referred to throughout as gene remodelling events. Little is known about the contribution of these remodelling events in protein evolution and in the evolution of the Metazoa in particular. These are innovations in the protein coding sequences that are brought about by shuffling of regions of DNA from different proteins (this is not direct descent with modification and therefore not strictly phylogenetic in nature as it requires multiple roots on the phylogenetic tree). In this thesis, we describe the application of phylogenetic and network based models of protein coding evolution to sequence data from all major groups across the Metazoa and we define the basic characteristics and rules for gene remodelling in the Metazoa.

Thesis Aims

(1) To assess the impact of model misspecification and heterogeneous modelling in phylogeny reconstruction – Chapter 2

We wished to establish a data driven method for phylogeny reconstruction using currently available homogenous and heterogeneous model based software. Using the case of the root of the placental mammal tree, we sought to determine if model misspecification has been a primary cause of the conflict in resolving the root of the placental phylogeny. Secondly, we wished to determine if heterogeneous modelling can resolve the problem. These aims have been completed and the results have been published in Moran et al. (2015) and (Tarver et al. 2016).

(2) To investigate protein-coding innovation by gene remodelling (gene fusion and gene fission) in the animal kingdom – Chapter 3

We wished to investigate the role of gene remodelling in creating novel protein coding genes in the animal kingdom. We wished to quantitatively survey the family expansions and contractions at each node across the *Metazoa*. We also wanted to establish the rate at which these expansions and contractions occur and we wished to time these events in relation to the fossil record for animals. Finally, we wanted to assess how specific expansions and contractions correlate with function and major phenotypic innovations. The results of Chapter 3 are in preparation for publication (Moran Raymond J. et al. 2017).

(3) To investigate the role of domain shuffling in protein coding genes across the animal kingdom – Chapter 4

Using network theory, we wanted to investigate how functional domains form protein coding modular proteins. Using these networks, we wished to compare the network of remodelled genes to the network of non-remodelled genes from Chapter 3. We wanted to understand if some domains are more promiscuous than others and what are the rules around combinations of domains permitted/retained in animal proteins.

Chapter 1 – Introduction

1.1 Molecular innovation

It is intuitive that gene gain can drive molecular innovation. For example, the emergence of new genes by gene duplication (Zhang 2003), point mutation (Imwong et al. 2003), and horizontal gene transfer (Blomme et al. 2006), amongst other mechanisms (Long 2001)have resulted in novel phenotypes. However, loss is also a powerful effector of phenotype and protein coding and regulatory element loss have been shown to play a substantial role in the evolution of novel phenotypes (Smith et al. 2015), *e.g.* a point mutation causing the loss of an enhancer from the human Androgen Receptor has resulted in the emergence of a novel phenotype - a lack of penile spines in human (McLean et al. 2011).

In this thesis, we focus entirely on the study of protein coding element evolution and modelling the origin and evolutionary history of protein coding gene families. Specifically, we examine the emergence of novel protein coding gene families, patterns of gene duplication and loss, and gene remodelling (by domain shuffling and fusion/fission). Novel protein coding gene families can arise from existing genes by gene duplication, and also by recombinatorial processes, such as remodelling existing protein coding genes (McClintock 1950, Kaessmann et al. 2002, Bashton and Chothia 2007, Hoen and Bureau 2015). Novel gene families can also arise from various types of non-coding DNA by *de novo* gene genesis (Tautz and Domazet-Lošo 2011, Schlötterer 2015) but these are not examined here. Therefore, modelling of the emergence of new genes in sequence space can follow tree-like processes (e.g. duplication and loss of genes) or non-tree-like processes (e.g. introgression and gene remodelling). It is widely accepted that non-tree-like mechanisms of evolution like Horizontal Gene Transfer (HGT) are dominant in prokaryotes (Frost et al. 2005) and to a lesser extent in eukaryotes (Keeling and Palmer 2008).

1.2 Tree-like mechanisms of protein coding gene evolution

Tree-like mechanisms of protein coding gene evolution include insertion, deletion, point mutation, and gene duplication and loss. They are tree-like because their evolutionary history is the result of descent with modification, parent genes either duplicating or mutating to form new genes. In Chapter 3 of this thesis the tree-like mechanism on which we focus on is gene duplication and loss. Gene duplication is a major mechanism of molecular innovation (Tautz and Domazet-Lošo 2011), and is abundant in eukaryotic genomes (Kaessmann 2010). Both DNA-mediated and RNA-mediated duplication occur frequently and have been reported in the Metazoa and both are considered in this thesis, although it is known that the probability of a gene becoming functional is higher in the case of DNA-mediated duplication (Jun et al. 2009). Duplication can occur on a small scale (whole genes or gene fragments) (Hughes 1994) or a large scale (whole genome duplication) (Kaessmann 2010) and can create functional redundancy (Ohno 1970) and/or can be detrimental as described in gene dosage models of evolution (Klein 1981). The fate of duplicated genes has been described mathematically and empirically. Ohno's neofunctionalization model (Ohno 2013), the duplication–degeneration–complementation model (Force et al. 1999) and the specialization or escape from adaptive conflict (Hughes 1994) model all proposing potential evolutionary trajectories of duplication genes. In brief, a duplicate gene can undergo subfunctionalization, neofunctionalization or pseudogenization (Zhang 2003). In Sections (1.2.1 to 1.2.7) below the concepts, theories and models associated with tree-like mechanisms of protein evolution relevant to this thesis are described. Central to these approaches is phylogenetics. Chapter 2 is focused on phylogeny, and the application of heterogeneous models to place of the root of placental mammals.

1.2.1 Phylogenetic Reconstruction using Molecular data

Molecular phylogenetics is the study of evolutionary relationships using molecular data such as DNA, RNA and molecular markers. Traditionally phylogenetics has been used to describe the patterns of divergence between species. In the last decade phylogenetics/phylogenomics has been applied to epidemiological dynamics of pathogens (Yang and Rannala 2012, Lyubetsky et al. 2015) and infectious diseases (Grenfell et al. 2004, Sheen et al. 2013, Leventhal et al. 2014), to classify metagenomic sequences and identifying genes, to identify regulatory elements and non-coding RNA in sequenced genomes (Kellis et al. 2003, Pedersen et al. 2006), and to reconstruct ancestral genomes (Yu et al. 2015).

A phylogenetic tree is a graph composed of nodes and branches (Figure 1.1). Nodes can represent species, populations, individuals or genes. A branch connecting two adjacent nodes defines the evolutionary relationship between those nodes. For example, A and B are closely related and share a most recent common ancestor at node F (Figure 1.1). Nodes A-D and O are termed operational taxonomic units (OTUs). On the rooted tree O represents the outgroup, the most distantly related OTU. R represents the root, the most recent common ancestor to all the nodes. In a rooted tree, a unique path of ancestry can be drawn from the root to any of OTUs (Morris et al. 2013).

In general, phylogenetic reconstruction has been performed on nucleotide or amino acid sequence data. Amino acid alignments can be recoded into Dayhoff categories (Dayhoff and Schwartz 1978), based on their physiochemical properties, thus reducing the 20 character states of AA (amino acid) data down to six states (as in (Morgan et al. 2013)). This effectively removes a layer of composition heterogeneity in the amino acid data, which can be useful when models are not adequate for the data. However, there is a subsequent a loss of information here. Using Dayhoff categories reduces the number of character states allowing for parameters such as the rate matrix to be estimated directly from the AA data (Liò and Goldman 1998, Morgan et al. 2013). For nucleotide substitution models, it is possible to estimate all parameters from the data, as there are only four character states (A, G, T, C) and RY-coding can further reduce the character states to just two (Purines (R) and Pyrimidines (Y)) (Phillips and Penny 2003). RY-coding can remove a layer of composition heterogeneity within the nucleotide data, but it has also been shown to remove informative transition substitutions (Ishikawa et al. 2012). For nucleotides, the exchange rate matrix and composition vectors are calculated from the alignment. However, protein empirical models are not based directly on the data, rather they use fixed composition and rate parameters – the q-matrix. It is then possible to estimate the compositions of the amino acids and combine it with an empirical rate matrix, giving a p-matrix (Swofford et al. 1996). In summary, there are models available for both nucleotides and amino acids.

Figure 1.1: Depiction of a rooted and an unrooted phylogeny.

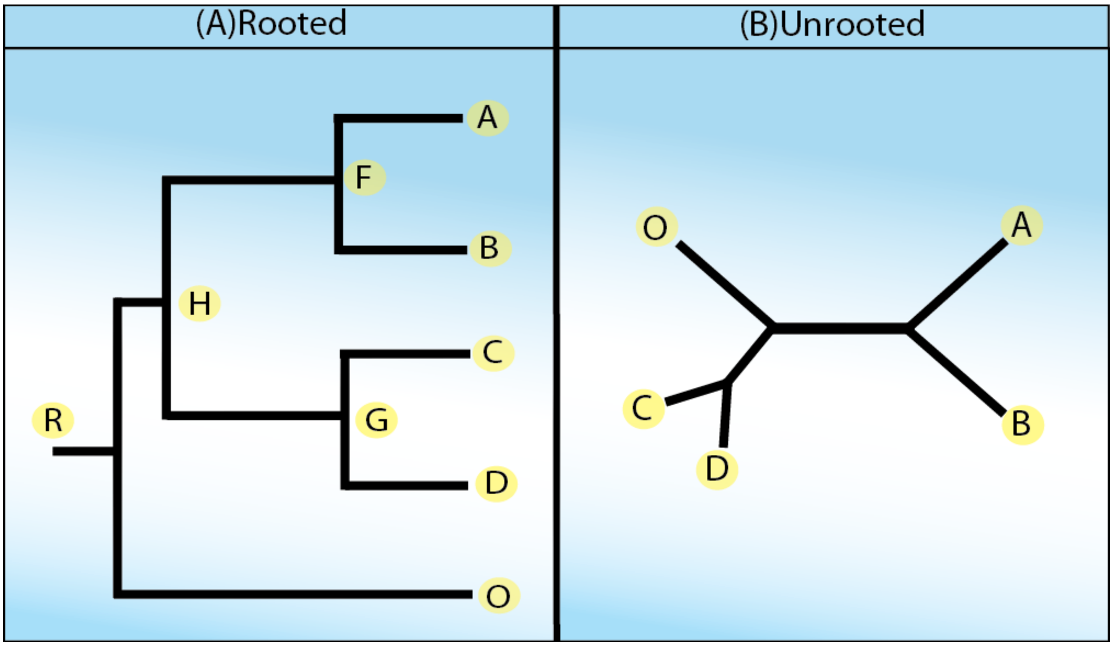


Figure 1.1: (A) A rooted phylogenetic tree where A-D and O are external nodes, F-H are internal nodes, R represents the root of the tree and O represents an outgroup. (B) An unrooted phylogenetic tree where A-D and O are all equivalent nodes with no directionality of the evolutionary process.

As mitochondrial sequence data have low recombination rates but more rapid point mutational rates than nuclear DNA sequences, they have been used extensively in phylogenetic reconstruction (Reyes et al. 2004, Gibson et al. 2005, Kjer and Honeycutt 2007). The application of whole mitochondrial genomes to mammal phylogeny failed to retrieve many of the well-known and uncontroversial groupings within the mammal tree such as the 4 superorders within the placental mammals (Morgan et al. 2014) and it was proposed that mitochondrial data are better suited to shallower depths (Arnason and Janke 2002, Springer et al. 2004). More recently, models for mtDNA sequence evolution have been developed such as a three-state general time reversible DNA substitution model that accommodates homoplasy (Tavaré 1986a), and an approach that uses the secondary structure to guide the alignment of RNA sequences, both of which have improved the resolution of mtDNA based phylogeny reconstruction (Gibson et al. 2005, Grajales et al. 2007).In addition, accounting for rate heterogeneity across the data with the application of appropriate models and data partitioning has succeeded in producing mammal phylogenies from mitochondrial data that are similar to those derived from nuclear genes (Reyes et al. 2004, Kjer and Honeycutt 2007). However, the power and suitability of mitochondrial genes (individually or concatenated into a super-matrix) to resolve this phylogeny continue to be debated (Morgan et al. 2014). In summary, these studies of mtDNA illustrate the importance of data choice, the impact of model specification on phylogeny reconstruction and the need to adapt and develop models that are suitable for specific data types.

Using protein-coding nucleotides can be desirable for phylogeny reconstruction because the parameters for the model can be estimated directly from the data (unlike empirical amino acid models). However, in practice it is often too computationally intensive to gain parameter estimates directly from the data and empirical models of substitution are often used for amino acid data (Henikoff and Henikoff 1992, Adachi et al. 2000, Müller and Vingron 2000, Dimmic et al. 2002, Abascal et al. 2005, Abascal et al. 2007, Nickle et al. 2007, Le and Gascuel 2008). While deep divergences always have increased risk of erased signal due to multiple substitutions at the same site (amongst other issues such as the identification of homologs), nucleotide sequences have been successfully applied to resolving topologies for divergence times of over 500 million years (Misof et al. 2014).

The issue of saturation is often cited as justification for applying amino acids rather than nucleotides, however amino acids can also saturate rapidly and can mutate rapidly between functionally similar groups causing functional convergence that is not seen at the nucleotide level. Therefore, while there is a common view that amino acids are useful for deeper evolutionary distances and nucleotides for shallower timescales (Brown 2002) – simply selecting amino acids over nucleotides ignores these subtleties/idiosyncrasies of real data. Software packages such as P4 (Foster 2004) and PhyloBayes (Lartillot and Philippe 2004, Lartillot et al. 2009) have now become available that allow for more complex modelling of amino acid substitution processes thereby greatly improving the accuracy of amino acid sequence modelling.

1.2.3 Methods of Phylogeny Reconstruction

Phylogeny reconstruction methods are based on a set of parameters that describe the process of evolution in molecular sequence data. These parameters can be explicitly defined as for Maximum Likelihood (ML) and Bayesian approaches or implicit as is the case for Maximum Parsimony (Posada 2003). An evolutionary model consists of two parts: a scheme (sometimes referred to as the model) and a tree. The scheme is the way we describe how the sequences have evolved and it is composed of two parts: (i) the composition vector, and (ii) the exchange rate matrix. The composition vector describes the base frequencies in the data – this is a particularly important factor in mammal phylogeny for example given the observation of biased gene conversion (Galtier et al. 2001, Galtier and Duret 2007). The exchange rate matrix describes the probabilities of a character state changing.

The first statistical model described was the Jukes and Cantor (JC) model (Jukes and Cantor 1969). The JC substitution model assumes all composition frequencies are equal and the exchange rate between the character states is equal. Models that expanded on JC include the F81, that allowed base frequencies to vary (Felsenstein 1981). This ultimately led to the development of the General time reversible (GTR) model (Tavaré 1986b, Felsenstein and Churchill 1996). The GTR model allows for each type of substitution to have a probability in a reversible manner. For example, A≥T is the same as T≥A, and it can estimate composition from the data. Evidently, many of the less complex models are nested within the more complex models (Figure

1.2) (Swofford et al. 1996). A set of empirical models for protein sequences have been developed based on specific data types. For example, the Jones-Taylor-Thornton (JTT) model is based on transmembrane proteins (Jones et al. 1994) and the Whelan And Goldman (WAG) model is based on globular proteins (Whelan and Goldman 2001) and many other protein models exists (Henikoff and Henikoff 1992, Cao et al. 1994, Adachi et al. 2000, Müller and Vingron 2000, Dimmic et al. 2002, Abascal et al. 2007, Nickle et al. 2007, Le and Gascuel 2008).

Models of evolution have become increasingly more sophisticated and complex. However, the increased number of free parameters required for the additional complexity equates to higher sampling variance. This diminishes the power to differentiate between competing hypotheses. Over-fitting a model to the data could lead to a decrease in support values, but under-fitting a model would lead to strong support for an erroneous topology. It is the case that the application of larger/genome scale data greatly reduces the risk of over-fitting a model but the risk of under-fitting is still significant, therefore a model should be chosen based on the fit to the data.

The goal is to identify and use the least complex model that most adequately describes the data (Posada 2003). To generate a heterogeneous model the first step is to select a substitution model (alternatively known as the exchange rate matrix) – this provides the homogeneous exchange rate matrix that best fits the data. A data driven process is recommended to determine the model of best fit (Foster 2004) – the initial homogeneous exchange rate matrix (substitute model) will vary from dataset to dataset. By testing and comparing different models on the data we can find the best model (that has the least number of free parameters) from the available pool of models (Swofford et al. 1996, Foster 2004). There are well known methods designed for this purpose for example ProtTest (Abascal et al. 2005) (protein sequences), ModelTest (Posada and Crandall 1998) (nucleotide sequences), MrModelTest2 (Nylander 2004) (nucleotide sequences) and ModelGenerator (Keane et al. 2004) (protein and nucleotide sequences).

Figure 1.2: Example of nested models of molecular evolution.

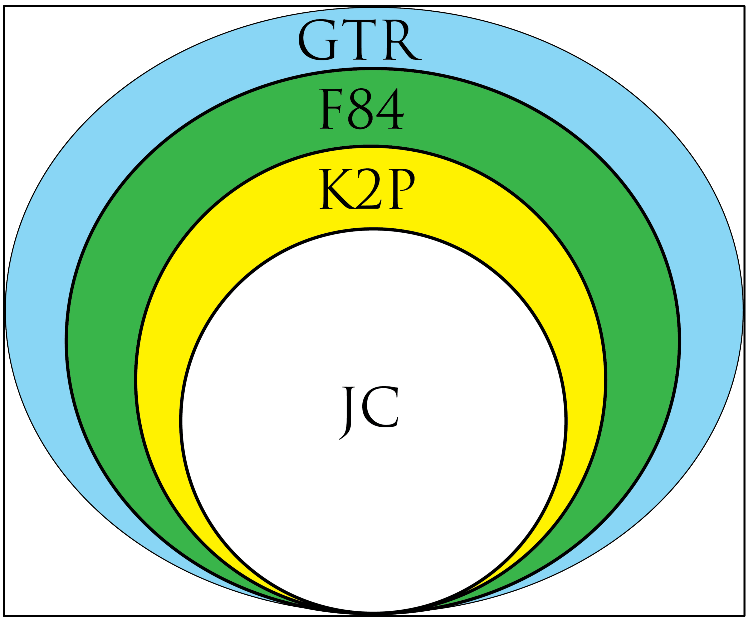


Figure 1.2: A typical example of nested models of evolution. The Kimura Two Parameter(K2P) model was created based on the Jukes Cantor (JC) model. The Felsenstein(F84) model was created with a strong basis on K2P and JC models. The General Time Reversible (GTR) model was established with a strong basis on F84, K2P and JC models. From the inner part of the figure (JC) to the outmost part (*GTR*) the complexity of the model and number of parameters increases.

The two main approaches to phylogenetic modelling the evolutionary history of sequences are homogeneous modelling and heterogeneous modelling. These differ in their construction/assumptions and in their interpretation of biology. We define homogeneous models as those that do not account for compositional variation and exchange rate variation across the phylogeny (between species in the alignment) (Morgan et al. 2013). The most advanced homogeneous models can model rate heterogeneity but are unable to model compositional heterogeneity. Homogeneous models use a single rate matrix and composition vector and can incorporate a gamma distribution for among site rate variation (ASRV) (Yang 1996). An advantage of using homogeneous models over heterogeneous models is their speed (Stamatakis 2006a, Stamatakis 2006b). For example, RAxML (a homogenous modelling algorithm) models rate heterogeneity across sites using a CAT model but is not capable of modelling compositional heterogeneity(Stamatakis 2006a). [Note: RAxML CAT allows for site-specific rates of evolution and should not be confused with the CAT model implemented in PhyloBayes (Lartillot and Philippe 2004, Lartillot et al. 2009)]. RAxML CAT allows for rapid modelling of large datasets as it acts as an approximation of the gamma parameter (Γ) and speeds up the tree search. In addition to this, RAxML also implements CAT + rΓ. This rΓ algorithm uses Γ to refine the estimations of the CAT model (Stamatakis 2006b, Lartillot et al. 2009). The PhyloBayes CAT model (Lartillot and Philippe 2004, Lartillot et al. 2009) takes into account that characters at different sites in an alignment can have different probabilities of evolving into another character. For example, an Arginine at one site in an alignment might have a high probability of evolving into a Histidine while an Arginine at another site in an alignment can have a higher probability of evolving into a Lysine. Therefore, the CAT model in RAxML (Stamatakis 2006b) and PhyloBayes (Lartillot et al. 2009) are different because the PhyloBayes CAT model allows for among site compositional heterogeneity and the CAT model in RAxML does not.

The major issue is that homogeneous models are often too simplistic to account for the complexity of biological data. Therefore, phylogenetic reconstruction based on such models can lead to erroneous topologies, particularly if for example there is compositional heterogeneity evident in the data as is the case with mammal sequence data (Foster 2004, Morgan et al. 2013). Heterogeneous models can simultaneously model exchange rate variation and compositional variation between species or across the sites in the dataset (Foster 2004, Lartillot and Philippe 2004, Morgan et al. 2013) and are therefore a marked improvement in modelling complex sequence data.

The two major frameworks used in modern molecular phylogenetic analyses are maximum likelihood (ML) estimation and Bayesian inference. Some approaches rely exclusively on one of the frameworks (Stamatakis 2006a, Lartillot et al. 2009) while others can use a mixture of both (Foster 2004). It has been shown that both methods will not necessarily conclude the same answer (Douady et al. 2003), thus emphasizing the importance of judiciously choosing a method for the task at hand.

ML uses the likelihood function which is the probability of generating the observed data given a particular hypothesis. In our case, the likelihood function is the probability of the data given the topology and other parameters such as branch lengths and nucleotide frequencies (Felsenstein 1981, Huelsenbeck and Crandall 1997), but an absolute probability is not returned. To calculate the probability of an event (t) we need to know all possible outcomes (O) and the number of times t is observed in the test, therefore probability = t/O.

Likelihood becomes particularly useful when we do not know all the possible outcomes. A single tree with the highest likelihood (often the lowest negative log-transformed likelihood) is the most likely tree. A statistical technique called bootstrapping (Felsenstein 1978) is often coupled with ML to assign node support values to the tree, and this allows us to assess how strongly the data supports each split on the ML tree (Felsenstein 1981, Brown 1994). Bayesian inference is closely related to ML - it estimates the posterior probability of a tree by combining the likelihood of the tree and prior probability of the tree given the model. Bayes theorem (equation below) gives the posterior probability of an event occurring (an hypothesis) conditioned upon the prior probability of the event, which is the probability of the event before new evidence is considered (Shafer 1976). The equation below gives the posterior probability of event A occurring given that event B occurs(P(A|B)). P(B) and P(A) represent the probability of events B and A occurring without regard for each other respectively. P(B|A) is the probability of observing event B given that event A occurs (Shafer 1976).

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

The Bayesian approach is dependent on priors, which is both an advantage and disadvantage of this approach (Huelsenbeck et al. 2001). Advantages of Bayesian methods are that they are efficient means of complementing complex models and strong data can overcome poor priors. Unlike the aforementioned ML, Bayesian phylogenetic inference will produce a set of credible trees that are sampled according to their likelihood. A major difference between the two frameworks is that Bayesian inference generates a posterior distribution for the parameters rather than a fixed value. Bayesian phylogenetic inference only became viable when combined with Markov-chain Monte Carlo (MCMC) approaches (Metropolis et al. 1953, Hastings 1970) which allowed for sampling over parameter space. Advances on the MCMC, such as Metropolis-coupled MCMC (MCMCMC), have proven effective in tackling the problem of local maxima (Altekar et al. 2004).

In summary, there are a number of approaches for modelling phylogenetic reconstruction. The features of the 3 main approaches are summarised in Table 1.1.

1.2.4 Generating a multiple sequence alignment for phylogenetic reconstruction

Alignment is the initial stage in most comparative genomic analyses. In this thesis alignments are used to generate species trees, to assess gene gain and loss and to determine the extent of gene remodelling. There are a number of powerful MSA (multiple sequence alignment) tools available. For example, Clustal Omega is a progressive alignment algorithm that can create large MSAs in a relatively fast manner (Sievers and Higgins 2014). Clustal Omega also allows the user to inform the alignment process using hidden Markov model (HMM) profiles based on similar data from a database such as Pfam (Bateman et al. 2004, Sievers and Higgins 2014). MUSCLE is an MSA algorithm that involves multiple sequence comparison by log expectation (Edgar 2004). This is an iterative alignment tool, where an initial progressive MSA.

Table 1.1 A summary of some of the advantages and disadvantages of different approaches to phylogeny reconstruction.

Method	Advantages	Disadvantages
RAxML (Stamatakis 2006a) (Maximum likelihood)	Very Fast Easy to use Recommended for homogeneous modelling	No heterogeneous modelling Often too simplistic
P4 (Foster 2004) (Bayesian)	Models heterogeneity across lineages Extensible Strong statistical framework	Python knowledge needed Time/computationally consuming Convergence diagnosis difficult
PhyloBayes (Lartillot et al. 2009) (Bayesian)	Employs CAT model of heterogeneity Very easy to use MPI (Message Passing Interface) version available Convergence is easily assessed and can be done automatically	Computationally intensive and time consuming, even with MPI. (However, newer computer clusters are more readily prepared for this.)

is created and then informs subsequent distance calculations for another MSA. A new MSA is created by realigning the first and second MSAs. If the new MSA is an improvement it is kept, and the previous MSAs are deleted and if there is no improvement the newly created MSA is deleted and the original alignment is used (Edgar 2004). A different approach to MSA is implemented in HMMER which is based on probabilistic models called hidden Markov models (HMM) (Eddy and Rost 2008). These HMMs can assign likelihoods to all possible combinations of alignment in order to find the most likely MSA. They can produce global and local alignments and are particularly useful when searches for sequence profiles and short sequences such as the domain style searches we will be carrying out in Aim 3 (Chapter 4). In summary, there are numerous ways to create an MSA. The examples mentioned here are just small samples of the variety of algorithms available. Regardless of which approach is taken to generate MSAs it is crucial that an optimal MSA is used for subsequent analyses. A number of statistical algorithms have been designed to assess the quality of the alignment, e.g. AQUA (Muller et al. 2010) and MetAL (Blackburne and Whelan 2012). The way in which they work is they perform alignment using a variety of approaches including those outlined above and then for each individual gene family they assess which alignment algorithm produces the alignment of best fit for the data using purpose built statistical scores such as the Z score and the NorMD score (Muller et al. 2010, Blackburne and Whelan 2012).

1.2.5 Super-tree and Super-matrix approaches for species trees

Given a set of genes from a group of organisms of interest, there are two main approaches to phylogenetic reconstruction. The first is to build gene trees individually and to apply a coalescent model or a super-tree approach to generate a phylogeny that captures the variation across gene trees (Gatesy and Springer 2014). An advantage of this is that the final tree is a step removed from the actual sequence/morphological data. Thus it is possible to summarize results obtained from different characters (Delsuc et al. 2005). The second approach is to concatenate the alignments into a super-matrix and from this single alignment generate a phylogeny. The increased MSA size of the super-matrix approach can give more appropriate model estimations for sophisticated heterogeneous models like the PhyloBayes CAT-GTR model and P4 models (Foster 2004, Lartillot and Philippe 2004). However, super-matrix methods can be more susceptible to incomplete lineage sorting (ILS). ILS occurs when alleles are not perfectly segregated into all lineages and subsequently fail to coalesce or converge on a most recent common ancestor (Hudson), see Figure 1.3 (detailed in Section 1.2.7.1). This is problematic super-matrix methods in particular because a single alignment of multiple gene histories creates a single phylogeny (Philippe et al. 2004). This means, if ILS was sufficiently present, super-matrix methods could potentially give positively misleading or statistically inconsistent results. However, the risk of this can be assessed by analysing the presence of ILS in a dataset (Tarver et al. 2016). As is often the case for complex phylogenetic questions, we have large datasets that we are interested in analysing and this is not tractable using co-estimation type coalescence. An alternative would be to use shortcut coalescence, but it has already been shown that the super-matrix approach is superior at deep timescales (Gatesy and Springer 2014). A hybrid of the concatenation and coalescence approaches has previously been applied to the mammal phylogeny conundrum (Song et al. 2012).

1.2.6 Problems with phylogeny

1.2.6.1 Discordance between Gene Tree and Species Tree

It is well understood that the branching pattern of a gene tree can differ significantly from a complimentary species tree topology (Degnan and Rosenberg 2006, Degnan and Rosenberg 2009, Leaché 2009). Incomplete lineage sorting (ILS) is thought to substantially contribute to this discordance (Figure 1.3). Recombination can give rise to neighbouring segments of a gene having different histories and this can contribute to gene tree discordance (Posada and Crandall 2002). Another major contributor to gene tree discordance is horizontal gene transfer or HGT (horizontal gene transfer), where a gene is acquired from a different species. HGT is prominent in prokaryotes and can also occur in eukaryotes, however this is less studied (Philippe and Douady 2003). Examples of HGT in animals are believed to be extremely rare, with a few examples in *Drosophila ananassae* (Hotopp et al. 2007) and the *Bdelloid rotifers* (Boschetti et al. 2011, Flot et al. 2013). However, there is recent evidence suggesting that HGT is not as rare as previously thought (Crisp et al. 2015). At shallow time depths, ILS is particularly problematic because a lineage can have insufficient time to bring loci to fixation before its divergence (Pamilo and Nei 1988).

1.2.6.2 Base or Amino acid Composition Biases

Composition biases at the level of nucleotide or protein, can lead to species clustered together on a tree due to similar composition patterns rather than shared ancestry. Many early studies in the field of molecular phylogenetics used rRNA genes for phylogenetic analysis and reported erroneous topologies (Loomis and Smith 1990, Lockhart et al. 1992, Hasegawa et al. 1993) because rRNA genes had extreme AT or GC biases in different species (Foster and Hickey 1999). There is evidence of compositional heterogeneity across the animal kingdom (Nesnidal et al. 2010) and it is particularly apparent in mammals (Romiguier et al. 2010). It is suggested that this is a major source of error in mammal phylogeny (Romiguier et al. 2010, Romiguier et al. 2013). There is evidence to suggest that amino acids (as opposed to nucleotides) will not be directly affected by compositional bias (Hasegawa et al. 1993). However, Foster and Hickey (1999) have shown that this is not necessarily true and that compositional bias at a nucleotide level can be detected at an amino acid level (Foster and Hickey 1999). In summary, composition biases can be a major source of error in phylogenetic reconstruction and therefore when reconstructing phylogenetic histories we should use models that can accommodate these variations in composition. This was achieved in Chapter 2 where composition vectors are incorporated into the models (Foster and Hickey 1999, Lartillot and Philippe 2004).

1.2.6.3 Long Branch attraction (LBA) – the problem of shared rapid rates of evolution

LBA occurs when taxa/genes are incorrectly clustered together because they share high mutation rates (Felsenstein 1985). Similarly, the same logic applies to short branch attraction (Philippe et al. 2005). LBA has impacted on estimates of the metazoan phylogeny, leading to conflict and ongoing debate (Philippe et al. 2011b, Nosenko et al. 2013). Iconic examples of this include the branching order between non-bilaterian metazoans (*Porifera*, *Cnidaria*, *Ctenophore* and *Placazoa*) (Pick et al. 2010) and the *Coelomata*/*Ecdysozoa* groupings (Rogozin et al. 2007). Careful sampling of taxa has proven a valuable method of reducing LBA (Magallón 2010).

Figure 1.3: Graphical representation of incomplete lineage sorting (ILS)

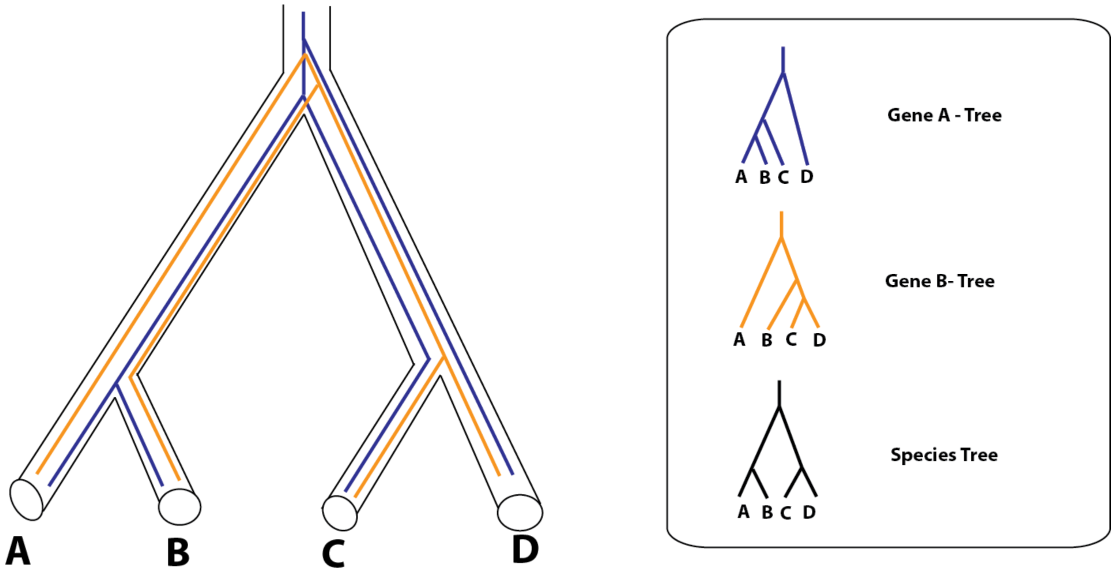


Figure 1.3: The outline of a phylogenetic tree depicting the species level relationship between species A, B, C, and D is shown on the left. The red and yellow lines correspond to the evolutionary history of gene A and B respectively. The inset on the right shows the gene trees for gene A and gene B.

For example, Pick et al. (2010) used broad sampling of non-bilaterian taxa to reduce LBA and reveal Porifera as the sister group to all other Metazoans (Pick et al. 2010). In this thesis we have applied similar careful sampling strategies in Chapter 2 and also in the preparation of datasets for Chapters 3 and 4.

1.2.6.4 Homoplasy

Homoplasy is when sequences/species have similarity due to independent events and not due to common ancestry. There are three types of homoplastic events: 1) Convergence, 2) Parallelism, and 3) Reversal. Convergent evolution is a process by which unrelated or distantly related taxa independently evolve similar traits (Morris et al. 2013). An example of this is flight in bats and birds, where it is believed that powered flight arose independently in the two different lineages (Speakman 2001). Parallel evolution or parallelism is when closely related lineages evolve similar traits but independently from ancestors that shared that similarity (Bell 2015). Classic examples of this are demonstrated within mammals. For example, placental mammals and marsupials evolved similar species on separate continents. Placental shrews are smaller to marsupial tree shrews and both pursue insects and other small prey, and placental gliders are very similar to marsupial gliders (Bell 2015). Reversals are when an inherited trait reverses to an ancestral form. An example of this is the loss of gyrencephaly (folded neocortex) to become secondarily lissencephalic (smooth neocortex) in some mammal lineages (Kelava et al. 2013). Homoplasy represents a significant problem to molecular phylogenetics as it can result in erroneous topologies if it not minimized or mitigated in the phylogenetic modelling. It has been shown that explicit statistical models of evolution (Bayesian and Maximum Likelihood) are more robust that other methods such as maximum parsimony with respect to homoplasy (Brandley et al. 2009).

1.2.6.5 Heterotachy

Heterotachy refers to shifts in site-specific substitution rates within a sequence over time. It is understood that substitution rates for a site in a sequence do not always follow a Γ law, rates are not always uniform. This phenomenon occurs because at any given time only a fraction of positions within a sequence are free to vary over time and as functional constraints on the protein change over time, the sites that are free to vary also change. Therefore, the rate at each position can vary over time

(Philippe and Lopez 2001). Heterotachy is commonly found in sequence alignments and has been the source of error for many phylogenetic studies (Lockhart et al. 1996, Philippe and Germot 2000, Kolaczkowski and Thornton 2004). Heterotachy can have a negative affect on phylogenetic reconstruction, but these effects can be minimized by using appropriate models that account for the heterotachy (Mayrose et al. 2005, Philippe et al. 2005).

1.2.6.7 Contentious nodes on the animal tree.

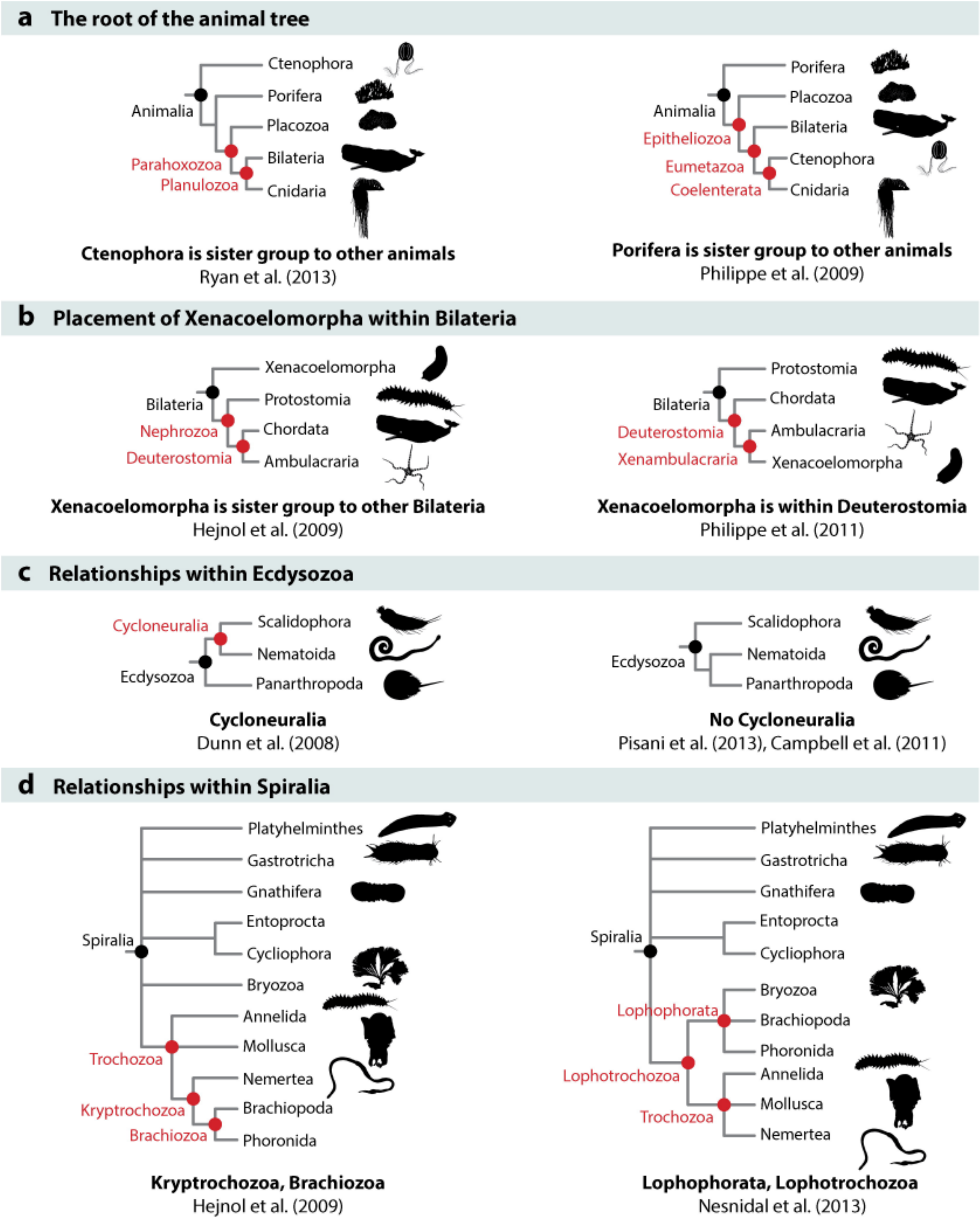
Due to a combination of the above issues around phylogeny reconstruction and compounded by the dependence on overly simplistic evolutionary models, there are a number of nodes on the metazoan phylogeny that are debated. The phylogenetic position of turtles (*Testudines*) with respect to the rest of the major tetrapod clades is contentious (Crawford et al. 2015). Other contentious nodes are depicted in Figure 1.4. Note: this list is not extensive, rather it serves as an illustration of the wide ranging and large number of unresolved nodes on the animal tree of life. It is possible to minimize or mitigate the aforementioned problems that can lead to erroneous topologies using appropriate taxon sampling and parameterized models appropriate for the input data/sequences.

There are limitations to the tree-like approaches we have described in Section 1.2 and indeed there are problems associated with phylogenetics these can sometimes be minimized or mitigated by using appropriate models of evolution and high-quality data. However, some problems such as non-tree-like processes of protein coding gene evolution cannot be solved using traditional tree-like analyses. The following Section details the types of non-tree-like processes of evolution that we are particularly interested in, and how these processes may be modeled using Network theory.

1.2.7 Molecules and Morphology

Predominantly, the transitions in animal evolution have been studied using traditional morphological approaches (including fossils where available). Both molecular and morphological studies suffer from problems of limited data, methods and technology (Abouheif et al. 1998, Irestedt et al. 2004, Lavrov 2007, Dunn et al. 2008, Hejnol et al. 2009, Dávalos et al. 2012, DeBiasse and Hellberg 2015, Dornburg et al. 2015).

Figure 1.4: Contentious nodes on the metazoan tree of life


a The root of the animal tree

Ctenophora is sister group to other animals
Ryan et al. (2013)

Porifera is sister group to other animals
Philippe et al. (2009)

b Placement of Xenacoelomorpha within Bilateria

Xenacoelomorpha is sister group to other Bilateria
Hejnol et al. (2009)

Xenacoelomorpha is within Deuterostomia
Philippe et al. (2011)

c Relationships within Ecdysozoa

Cycloneuralia
Dunn et al. (2008)

No Cycloneuralia
Pisani et al. (2013), Campbell et al. (2011)

d Relationships within Spiralia

Kryptrochozoa, Brachiozoa
Hejnol et al. (2009)

Lophophorata, Lophotrochozoa
Nesnidal et al. (2013)

Dunn CW, et al. 2014.
Annu. Rev. Ecol. Evol. Syst. 45:371–95

Figure 1.4.: The panels (a-d) represent contentious nodes at different phylogenetic depths on the metazoan phylogeny. (a) The placement of *Ctenophora* as the sister to all other animals on the left and *Porifera* as the sister to all other animals on the right. (b) The placement of *Xenacoelomorpha* as the sister group to *Bilateria* on the left and *Xenacoelmorpha* is placed within *Deuterostomia* on the right. (c) A Cycloneuralia relationship (*Nematoida* and *Scalidophora*) within *Ecdysoza* on the left and a

Nematoida-Panarthropoda grouping as a sister group to *Scalidophora*. (d) The various different branching orders within *Spiralia*: support for *Kryptrochozoa* (*Nemertea* and *Brachiozoa* (includes *Brachiopoda* and *Phoronida*) topology on the left and support for a *Lophophorata* (*Lophotrocozoa, Bryozoa, Brachiopoda and Phoronida*) and Trochozoa (*Annelida, Mollusca and Nemertea*) topology. Taken with permission from (Dunn et al 2014).

However, the increase in availability of molecular data (and the ease in generating large volumes of it – even from extinct species) and the associated rise of molecular phylogenetics, have meant that the study of evolution and of major transitions has moved from being a morphological to a molecular science.

The rise of molecular phylogenetics has been accelerated due to significant advances in sequencing technologies and statistical biological programs and computational capabilities (Dávalos et al. 2012, Pyron 2015). However, having lots of data is not good enough – good model of evolution and computational resources are also essential (outlined in Section 1.2.7). It is important to note that regardless of whether the data is morphological or molecular there are problems and limitations associated with systematics. There is conflict throughout the metazoan tree of life (as outlined in Section 1.2.7.7), with molecules and morphology supporting alternative relationships with strong statistical support and this is not unusual (Dávalos et al. 2012, DeBiasse and Hellberg 2015, Dornburg et al. 2015). For example, morphological data supports cephalochordates as the closest living relatives to vertebrates. Conversely to this, molecular data supports tunicates as the closest relatives to vertebrates. The resolution of this node in the metazoan tree of life is critical to understanding the origin of vertebrates (Delsuc et al. 2006). Another example related to Chapter 2 of this thesis is the use of morphological data to resolve the Xenarthran root for the placental mammal phylogeny (O'Leary et al. 2013). However, the latest molecular studies have found support for Atlantogenata root for the placental mammals ((Moran et al. 2015), Chapter 2). Together with our collaborators we have taken a total evidence approach to resolving the position of the placental mammal root where consilience across different data types is required for inferences to be made and this work has been published (Tarver et al. 2016). A consilience approach is a likely future solution for the contentious nodes in the metazoan phylogeny - provided data is available (Dávalos et al. 2012, Ronquist et al. 2012, Pyron 2015).

1.3 Gene remodelling in protein coding evolution

Traditionally the term introgression has been used to refer to the movement and integration of genetic material from one species to another species that is replicated within the host genomes (Rhymer and Simberloff 1996). Recently the definition of introgression has been broadened out to also encompass the reorganization, of genetic material within the same genome (Bapteste et al. 2013b). To avoid confusion we will use the term gene remodelling to refer to the reorganization, of genetic material within the same genome.

1.3.1 Interspecies introgressive descent – the movement of genetic material between species

A well known example of interspecies introgression is horizontal gene transfer (HGT), a significant driver of evolution in prokaryotes and major contributor to the evolution of the first eukaryote (Keeling and Palmer 2008). There are some individual cases of introgression in animals: for example the endogenous retrovirus (ERV) gene family in mammals is essential in mammal placenta development and is retroviral in origin (Nakagawa et al. 2013). Another example of introgression is hybridization, resulting in fertile offspring, making it possible for hybrids to introduce genetic innovation into populations and species. A now iconic example of adaptive introgression by hybridization is from the Heliconius butterfly - a diverse and brightly colored genus that are well known for mimicry (Heliconius Genome 2012). These butterflies frequently hybridize across species boundaries, with different species within the genus evolving nearly identical wing patterns to mimic poisonous relatives (Heliconius Genome 2012). Evidently, interspecies introgressive descent plays a prominent role in evolution.

1.3.2 Gene Remodelling (Intraspecies introgression) – the rearrangement of genetic material within a single genome

There are two major mechanisms of intra-species introgression. Firstly, the transport and integration of genetic material from place to place in the genome by mobile elements or transposable elements (TE) (McClintock 1950). TEs are believed to be a major source of genetic innovation in animals (Hoen and Bureau 2015) and they are present in all eukaryotic and prokaryotic organisms (McClintock 1950, Muñoz-López and García-Pérez 2010). TEs replicate via transposition and this means they can reside within a genome without conferring a selective advantage. However, TEs have been known to evolve in a similar way to regular coding DNA sequence, providing there is a selective advantage – a process called molecular domestication or TE exaptation (Hoen and Bureau 2015). We are not focused on TEs in this thesis, rather we are focused on the second major mechanism of introgression within a species genome - referred to throughout this thesis as gene remodelling.

Gene remodelling is an umbrella term for gene fusions, gene fissions, exon shuffling and domain shuffling (Jachiet et al. 2014b) and it is caused by recombination and transcription-mediated read-thorough (Kaessmann 2010, Wu et al. 2013). Gene fusion can be DNA-mediated or transcriptionally mediated and it occurs when two genes/gene fragments become fused into a single segment/gene (a single transcriptional unit) (Zhou and Wang 2008, Kaessmann et al. 2009, Agaram et al. 2015). Gene fusions that have occurred and have been fixed more recently retain molecular signatures of their origin and thus it is possible to find the parent or component genes that were involved in the fusion (Long et al. 2003). Whilst fused genes are often associated with cancers (Demichelis et al. 2007, Agaram et al. 2015), they can drive adaptive evolution and produce phenotypes (Long and Langley 1993). An iconic example of this is the *jingwei* gene in *Drosophila melanogaster* (Long and Langley 1993). *Jingwei* is derived from a transposon copy of the Adh locus and a 5' end derived from another gene called *yande*. The novel phenotype is a new specificity towards long-chain primary alcohols (Long and Langley 1993) . Gene fission is the opposite of fusion genes in that a single gene breaks into several smaller coding genes. Gene fission and fusion are mechanistically linked in that gene fission followed by juxtaposition of gene fission fragments often precedes a gene fusion event (Kaessmann 2010). Exon shuffling is when two or more exons from different genes are ectopically recombined or when an exon is duplicated, creating a new exon-intron structure(Long et al. 2003). Similar to this is domain shuffling except the biological element in this case is a functional protein domain. Structural domains are protein modules of discrete structural folding that often represent the functional building blocks of protein (Kaessmann et al. 2002). It is clear that gene remodelling plays a significant role in the evolution of sequences (Jachiet et al. 2014b), and it is also abundantly clear that traditional tree-like thinking cannot capture gene remodelling (Bapteste et al. 2012, Jachiet et al. 2014b). Therefore, we have exploited graph theory (Section 1.3.3) to study gene remodelling in the Metazoa.

1.3.3 Network based approaches to studying gene remodelling

1.3.3.1 Introduction to networks

Graph theory is an area of mathematics that tries to evaluate and represent relationships/connections (called edges) between objects (referred to as nodes)

(Figure 1.5). In biology, networks have been used to study: the transmission of infectious agents (Craft 2015); ecological food webs (Krause et al. 2003) ; protein interaction networks (Rual et al. 2005); gene regulatory networks (Huang et al. 2005), and more recently sequence similarity networks (SSNs) – networks where connections are based on sequence homology/similarity (Song et al. 2008). Through the application of networks to biological data the surprising discovery was made that biological networks share the same fundamental characteristics of technological graphs (e.g. Facebook and twitter). In other words in a biological network (much like on Facebook) the connections are not random. Usually biological networks have many nodes with few connections and few nodes with many connections, this property is referred to as scale-free (Albert 2005). In random networks, each node has a similar number of connections (the standard deviation is very low) (Albert 2005). Scale-free networks exhibit a power-law distribution with regards to the number of connections held by each node (this quantitative characteristic is called the degree of the node). This means that a scale-free network has many nodes with a low degree and few nodes with a very high degree - these are called hub nodes. Hub nodes have a very strong influence on the rest of the network due to their high connectivity (Albert 2005). There are two justifications for the presence of these hub nodes: 1) growth, and 2) preferential attachment. Growth simply refers to the growth or evolution of the graph over time. Preferential attachment refers to the fact that the higher the degree of a node the greater the probability that it will make a new connection (as the network grows/evolves) (Albert 2005). Sequence similarity networks (SSNs) are central to this thesis as they provide a quantitative framework for the identification of remodelled genes through time. By identifying cliques of gene families that are only connected by through a single gene family on a network we can identify remodelled gene families (Figure 1.8)

Figure 1.5: Sequence Similarity Network terminology and topology

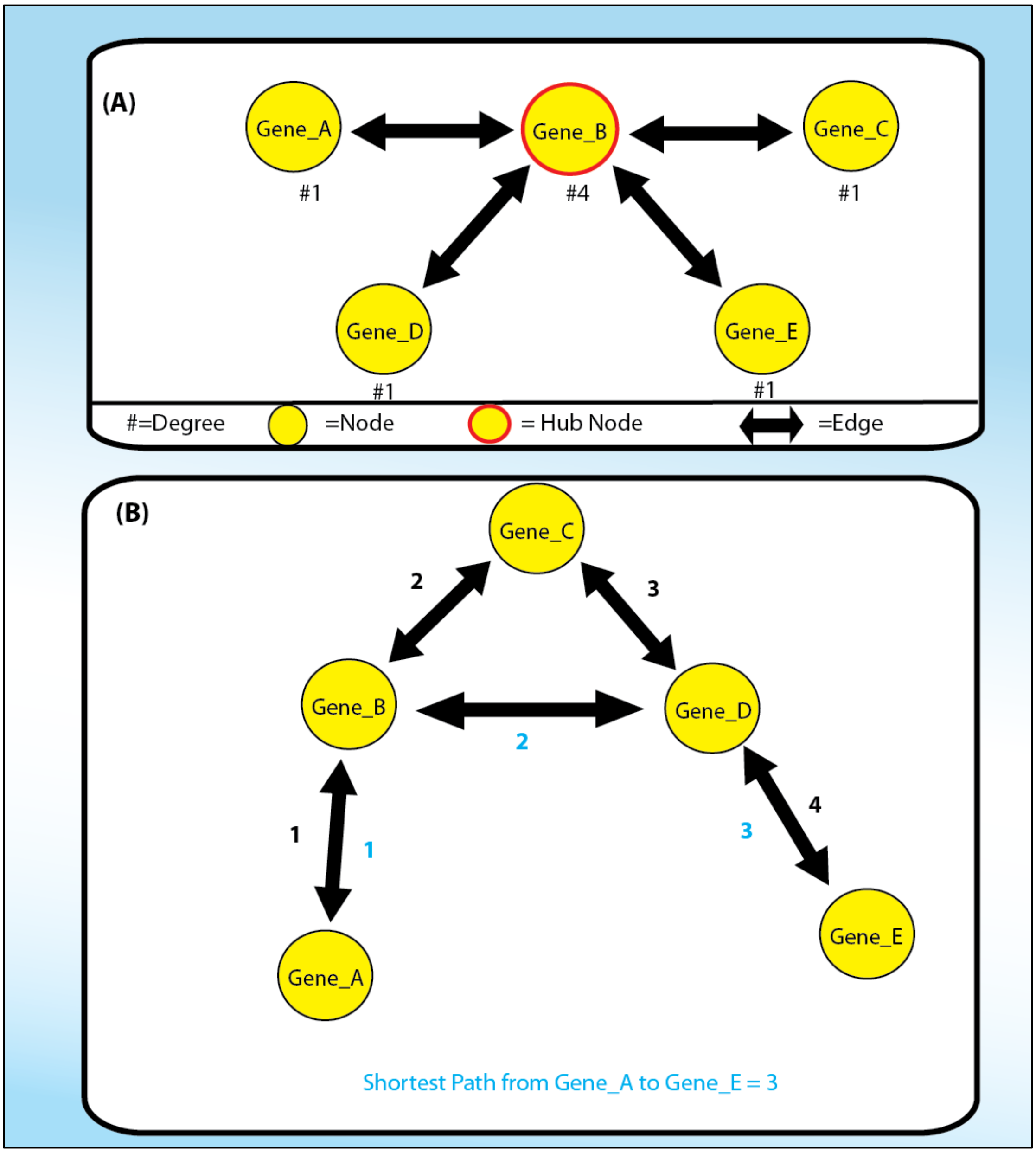


Figure 1.5: (A) Sequence similarity networks are composed of nodes (representing gene sequences in this example) and edges/connections (representing sequence homology in this example). The #X numbers indicate the degree of each node. This is the total number of edges it has to other nodes. Gene B is a hub node as it has a high degree with respect to other nodes in the network. (B) The shortest path between two nodes is the minimum number of edges required to connect two nodes. In this example there are two possible paths from Gene_A to Gene_E (outlined with black and blue numbers). The shortest path in this example is 3 as is achieved by bypassing Gene_C.

1.3.3.2 Characterization of the sequence similarity networks

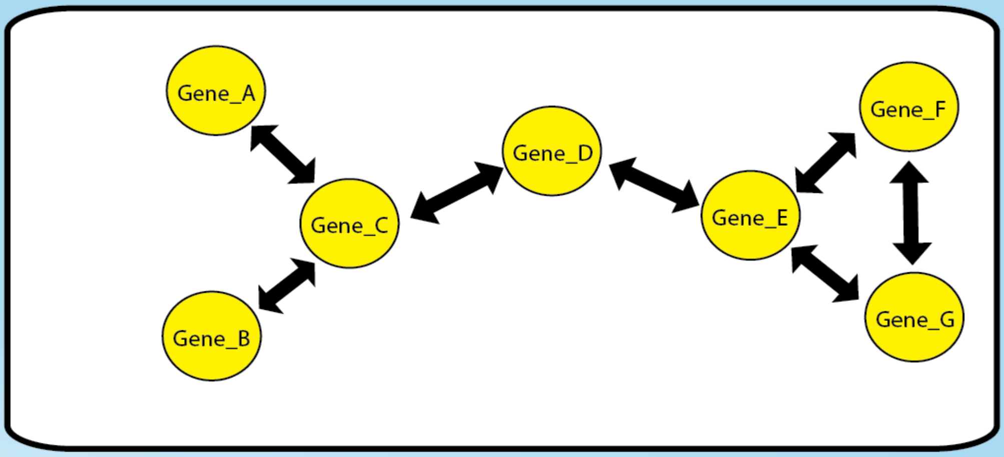
A. Centrality

A centrality measurement quantifies the importance of a node over others in a network. This is an important measure in our analysis of SSNs because it allows us to identify domains that have a significantly strong influence in gene remodelling from a several aspects of strong influence. There are three main centrality measurements used - degree, closeness and betweenness, and they are briefly described in Figure 1.6.

Degree centrality of a node is the number of edges directly attached to that node. Nodes with a higher degree are considered the most important or influential. For example, hub nodes in an infection network pose the greatest initial risk for significant spread of infection - having so many edges it is connected to many nodes and potentially to many parts of the network. From a protein interaction network (PIN) perspective high degree nodes tend to be lethal upon removal (Jeong et al. 2001). However, this measure does not consider the global structure of the network. For example, a high degree node may not be closely connected to all other nodes in the network. This is why we also use the measure of closeness centrality.

Closeness centrality measures have been used in infection networks where nodes with high closeness centrality are closer to the infection and thus more likely to get the infection early (Borgatti 2005). Closeness centrality is a measure of how quick a node can interact with all other nodes in a network (Opsahl et al. 2010). In a connected network, the shortest path between two nodes is the smallest number of edges that will connect both nodes. The closeness of a given node is the cumulative length of the shortest path to each other nodes in the network. Closeness centrality takes into account both direct and indirect interactions in a network. The limitation of this measure is that it cannot be applied to networks with disconnected components. This is why we also use betweenness centrality. Betweenness centrality has been used to find biologically important genes in gene co-expression networks (Azuaje 2014). Betweenness centrality indicates how often a node is found on the shortest path of all other nodes. It indicates how important a node is, with regards to the flow of information between nodes. Therefore, betweenness centrality favours nodes that join

Figure 1.6: Graphical depiction of the three measures of centrality in a network: Degree, Closeness and Betweenness.



(A)

Degree centrality		
Node	Degree	Normalized score
Gene_A	1	1/6
Gene_B	1	1/6
Gene_C	3	1/2
Gene_D	2	1/3
Gene_E	3	1/2
Gene_F	2	1/3
Gene_G	2	1/3

Normalization= Degree/N-1

(B)

Closeness centrality		
Node	1/Average Shortest Path	Normalized
Gene_A	1/16	6/16
Gene_B	1/16	6/16
Gene_C	1/11	6/11
Gene_D	1/10	6/10
Gene_E	1/11	6/11
Gene_F	1/15	6/15
Gene_G	1/15	6/15

Normalization=
(1/Average Shortest Path)
/
(N-1)

(C)

Betweenness centrality		
Node	Lies on Shortest Path	Normalized
Gene_A	0	0
Gene_B	0	0
Gene_C	16/3	16/45
Gene_D	13/3	13/45
Gene_E	13/3	13/45
Gene_F	0	0
Gene_G	0	0

Lies on Shortest Path=
Number of node pair paths
/
Number of times node lies on the shortest path between node pairs

=====================

Normalization=
(1/Average Shortest Path)
/
((N-1)(N-2)/2)

Figure 1.6.: A network of 7 nodes (genes) is used as an example to illustrate the different measure of network centrality. (A) Degree centrality is calculated by dividing the Degree (number of nodes it shares edges with) by N-1 (the number of nodes - 1). In this example Gene_C and Gene_E are most central nodes with regards to degree as they have the highest degree. (B) Closeness centrality is calculated by taking the inverse of the reciprocal of the average shortest path to each node in the network divided by N-1. In this example, Gene_D is the most central node with regards to closeness. (C) Betweenness centrality is essentially calculated by counting how many times a node is found on the shortest path between all node pairs in a network. To normalize this to account for the size of the network, this score is divided by ((N-1)(N-2)/2). In this example, Gene_C is most central to the network with regards to betweenness.

communities or cliques rather than nodes within a community or clique (Jeong et al. 2001, Borgatti 2005, Opsahl et al. 2010).

B. Assortativity

Assortativity is a quantitative correlation between degree and node connectivity. A network can be described as assortative, neutral or disassortative. An assortative network is when nodes generally connect with nodes of a similar degree (number of connections/edges). Assortative mixing is often seen in scale free networks, where high degree nodes connect with other high degree nodes. A disassortative network is when nodes generally connect with dissimilar nodes, where similar often refers to degree. Biological networks are predominately disassortative. Furthermore, it is rare that networks are neutral, neither assortative or disassortative (Newman 2002a, Newman 2003). Assessing the mixing pattern of a network is best achieved by combining a number of tests. Firstly, we can apply a linear regression on a plot of the average degree of nearest neighbours for a node against the degree of that node. A negative slope for the regression indicates disassortative mixing, a positive slope indicates assortative mixing and a slope of 0 indicates neutral (random) mixing (Barrat et al. 2004). In addition to this, calculation of the assortativity coefficient for a network can indicate of the type of mixing in the network. This can be done using Pearson correlation to determine the correlation between degree connectivity. The assortativity lies between -1 and 1, a score of 0 indicates neutral mixing, a negative score indicates disassortative mixing and a positive score indicates assortative mixing (Newman 2002b). Foster et al. (2010) demonstrated that the significance of the assortativity coefficient could be tested by comparing the actual assortativity coefficient with the assortativity of randomized graphs with the same degree distribution (Foster et al. 2010). The mixing pattern of a network is an important measure for this thesis because it enables us to characterize if gene remodelling events are assortative, disassortative or neutral. For example a very interesting question we wish to answer is: are fusion genes created from similar parent nodes?

C. Community structures

A community forms in a network when a group of nodes are relatively densely connected to each other, but sparsely connected to other nodes. The presence of communities in social networks is intuitively obvious, and they are also dominant in biological networks. In network theory motifs are building blocks of complex networks that represent functional biological modules (Porter et al. 2009). Using network theory small community structures can be identified that cluster together significantly more than expected – these are what is termed a community (Milo et al. 2002, Porter et al. 2009). A clique is a particular region of the network (a subnetwork) where each node is connected to every other node in the clique. Maximal cliques are cliques that cannot expand any further within the network (a clique that is not a nested clique within a bigger clique). Maximum cliques are the largest cliques in the network. Community structures are often used to define modules within protein-protein interaction networks. Biological network modules aim to identify sets of molecular interactions that work together towards a common function. This theory has been used to explore modules within large intermolecular interaction networks that govern cell structure and function. In evolutionary biology, conserved module structures can be compared across species. Identifying bridge nodes between cliques is the way we will identify possible gene remodelling events (Figure 1.7).

D. Clustering

A clustering coefficient (CC) is used to measure the degree to which nodes cluster together in a network. It is possible to evaluate clustering at a local level or a global level. The average clustering coefficient of a node as defined by Watts and Strogatz (1998) reflects the average of local CC for every node (Watts and Strogatz 1998). The local CC for a node is the ratio of its neighbours (nodes that share an edge with given node) that share a connection with each other. The CC ranges from 0 to 1, where 0 indicates no shared connections between node neighbours and 1 indicates the node neighbours all share an edge. A global CC is based on triplets of nodes, where three nodes are connected by two edges (called an open triplet) or three edges (called a closed triplet). The ratio of closed triplets to the total of all triplets (both closed and open) represents the global CC (also called transitivity). Transitivity gives an indication to the clustering within the whole network. Transitivity ranges from 0 to 1, where 0 indicates all triplets are open and 1 indicates all triplets are closed. This was first defined by Luce and Perry (1949) (Luce and Perry 1949) and refined by Opsahl and Panzarasa (2009) to use on weighted networks (Opsahl and Panzarasa 2009).

Figure 1.7: A clique community structure within a network

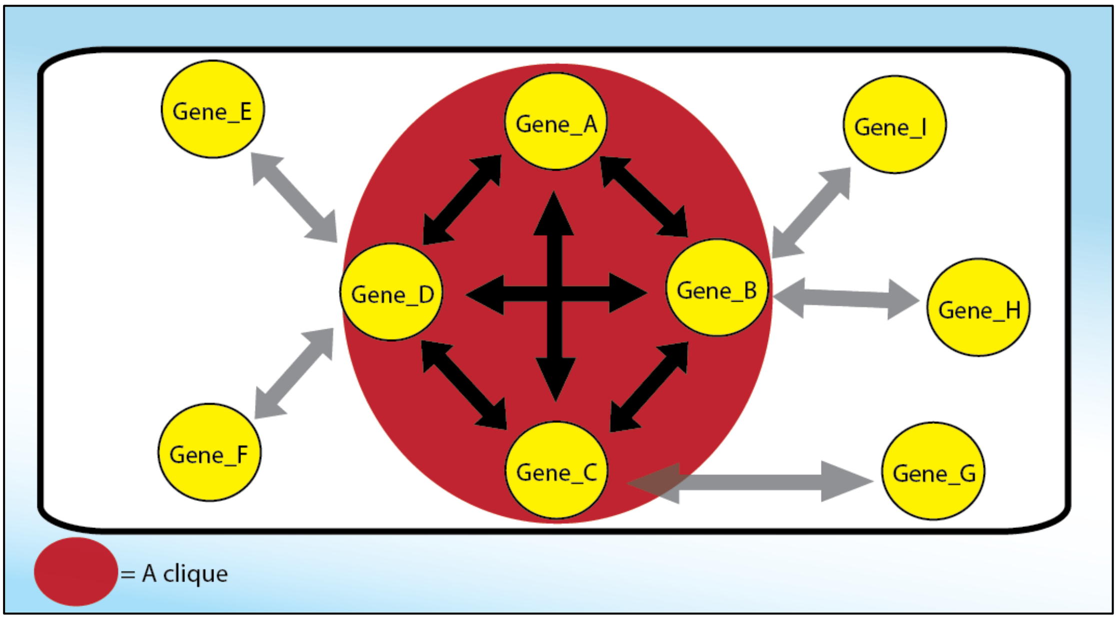


Figure 1.7: An example network where nodes represent genes and edges represent sequence homology. The area of the network coloured in red represent a clique structure (each node pair within this area is connected by an edge).

1.3.3.3 The application of network theory to gene remodeling

Recently SSNs have been proven particularly useful for studying non-tree-like processes of evolution (Song et al. 2008, Alvarez-Ponce et al. 2013b). In these networks, nodes represent sequences and the connections represent statements of homology. Homology is usually identified across sequences using programs like BLAST (Altschul et al. 1990b) or HMMER (Finn et al. 2011). In these networks a sequence can have multiple connections/edges, and thus can model non-tree like evolution. For example, a sequence of vertical descent will usually appear as a clique. A chimeric sequence formed by gene remodelling can be seen as a sparsely connected node/s that connect otherwise unrelated cliques/nodes (almost like a bridge connecting two or more different families of sequence) (Alvarez-Ponce et al. 2013b) . There are relatively few computer programs available to analyse non-tree like evolution. However, this research is becoming more prominent in molecular evolution. Recently, an algorithm CompositeSearch (Pathmanathan JS et al, 2017) was developed by our collaborators to identify gene remodelling events using BLAST and sequence similarity network mathematics (Pathmanathan JS et al, 2017). CompositeSearch is a novel network based algorithm that creates a sequence similarity network and identifies network structures that are indicative of gene remodelling (see Figure 1.8).We will apply these theories, metrics and algorithms from network theory to sequence data to address Aim 3 of this thesis where we wish to uncover the role of gene remodelling processes in the evolution of the *Metazoa* (Figure 1.8).

1.4 Major phenotypic transitions in the Metazoa

1.4.1 introduction

The experience of phylogenetic models of evolution gained in Chapter 2 are combined with network based models for Aims 2 and 3 of this thesis where we assess the process of gene remodelling in the *Metazoa*.

There have been a number of major phenotypic transitions in the animal kingdom, and while the molecular underpinnings of these transitions are largely unknown it is agreed that coding and non-coding changes are likely to be implicated (Carroll 2005).

Figure 1.8: Basic characterization of gene remodelling using a sequence similarity network (example depicts a Composite or Fused gene).

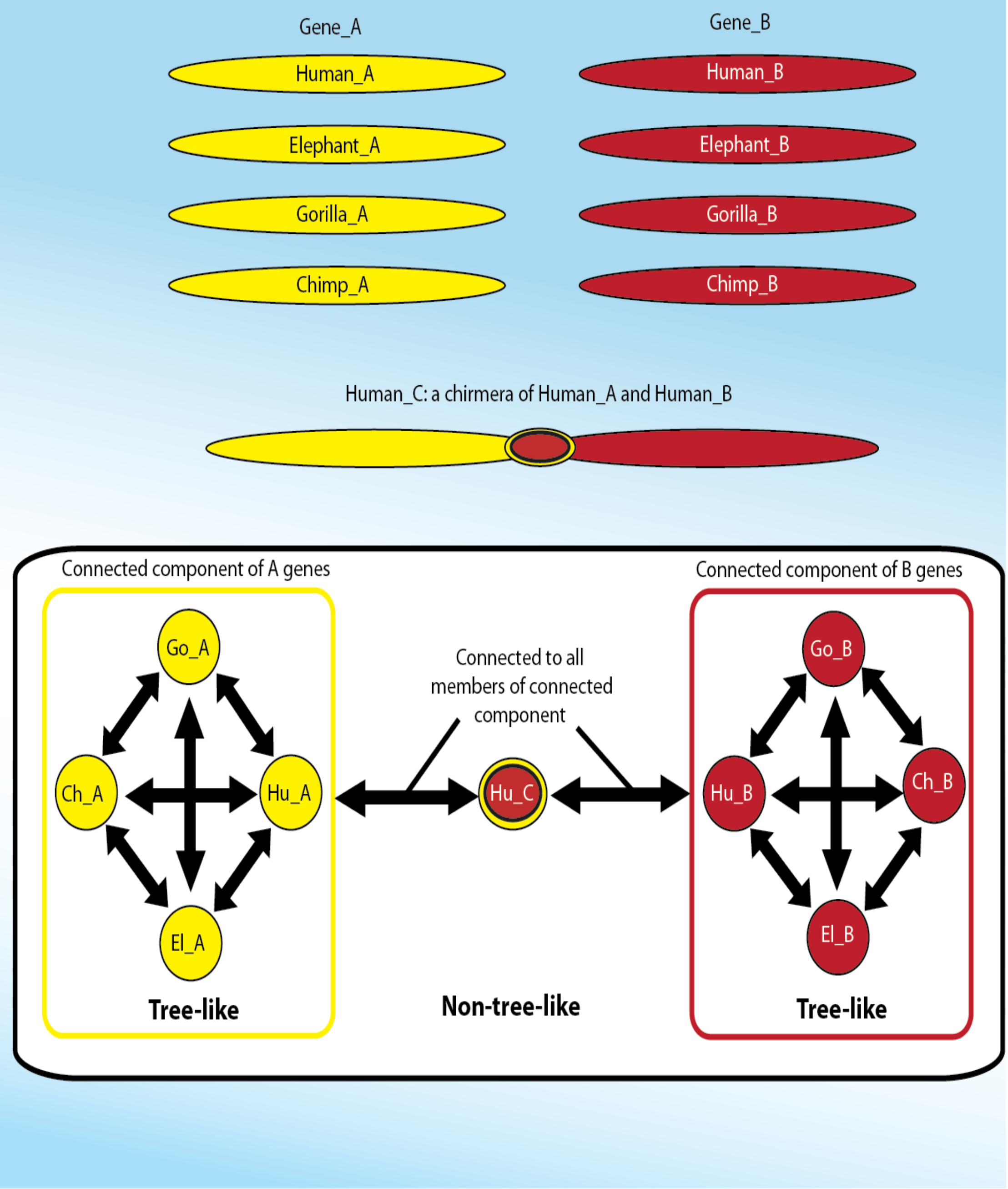

Gene_A
Gene_B
Human_A
Human_B
Elephant_A
Elephant_B
Gorilla_A
Gorilla_B
Chimp_A
Chimp_B
Human_C: a chirmera of Human_A and Human_B
Connected component of A genes
Connected component of B genes
Go_A
Ch_A
Hu_A
El_A
Connected to all members of connected component
Hu_C
Go_B
Hu_B
Ch_B
El_B
Tree-like
Non-tree-like
Tree-like

Figure 1.8: Networks have the capacity to illustrate both tree-like processes (e.g. monophyletic orthologs) and non-tree-like processes (e.g. fusion gene genesis). In this example, there are two monophyletic gene ortholog families (Gene_A in yellow and Gene_B in red) between human, elephant, gorilla and chimp. Each family forms a clique structure within the network, as is typical or orthologs. The network also shows the Hu_C gene, a fusion gene originating from the Human_A gene and the Human_B. Finding a single node that acts as a bridge, connecting two unrelated clique structures identified this.

We define major transitions as events that have allowed a lineage to radically change their environment, a biological function and/or phenotype. From studies such as McLean et al. (2011) (on the lack of penile spines in Humans) and D'Apice et al. (2004) (on the cause of Progeria) we know that small changes at the genetic level can cause radical changes at the phenotypic level (D'Apice et al. 2004, McLean et al. 2011). A major question in theoretical evolutionary biology that remains unanswered is: Are these seemingly major transitions in phenotype caused by saltational change at the genetic level or are they a result of incremental and gradual change over time? The theory of punctuated equilibrium proposes that most species will remain in an extended state of stasis (with little evolutionary change) for most of their geological history and that changes or transitions only occur due to rare and rapid speciation events (Gould 1972). On the other hand, since the time of Darwin it has been proposed that evolution is gradual, constant and slow, as proposed by the theory of phyletic gradualism (Gould 1972). There has been much heated debate over these opposing theories with a well known gradualist referring to punctuated equilibrium as evolution by jerks (Turner 1984). However, with the increase in data availability and quality and with improvements in models of evolution there a possibility to test both theories. There are some proponents of a half-way house between both theories, where evolution occurs mostly by gradual change but with periods of slower and much more rapid evolution (Simpson 1944). For Aim 2 and 3 of this thesis we use sequence similarity networks (SSNs) and -omics data from variety of species (63 in total) that span the animal kingdom.

Using the SSN approach in Chapter 3 we test the hypothesis that the major transitions in animal phenotypes were facilitated by significant bursts of novel gene family birth by gene remodelling and duplication. A summary of the major transitions in the Metazoa (as we define them) is outlined in the following Sections.

1.4.2 Development of germ layers

Germ layers are cell layers that arise in early animal development that give rise to all tissues and organs in adults. All animals, with the exception of sponges, have either two germ layers - diploblastic (e.g. Cnidaria and Ctenophora) or three germ layers - triploblastic (e.g. Human). The three possible germ layers are the endoderm, mesoderm and ectoderm. In general terms, an increase in germ layers is associated with increased biological complexity (Technau and Scholz 2003).

1.4.3 Origin of symmetry

There are 3 main body plans described for the Metazoa: 1) asymmetrical, 2) radially symmetrical, and 3) bilaterally symmetrical. Many sponges, and the placazoan *Trichoplax adhaerens*, are asymmetrical - they have no point, axis or plane of symmetry. Radial symmetry is the repetition of a set of anatomical features around a single axis, they have no real left and right side and no defined head or tail. Porifera, Ctenophora, most Cnidaria and adult Echinoderma are radially symmetrical. Bilaterally symmetrical animals have up to three axes of symmetry – a dorsoventral axis, anterior-posterior axis and left-right axis. Most bilateria have a bilateral body plan throughout their lives with the exception of echinoderms which are bilaterally symmetrical as larvae but almost penta-radially symmetrical as adults. The emergence of bilateral symmetry approximately 550 MYA (million years ago) created the opportunity for regional differentiation of the body allowing for increased complexity (Martindale 2005, Manuel 2009, Bell 2015).

1.4.4 Origin of Chordates

The chordate clade emerged approximately 530 MYA and is composed of urochrodates (Tunicates), vertebrates (e.g. reptiles, birds, mammals) and cephalochordates (e.g. lancelets) (Lamb et al. 2007). The combination of Chordata, Hemichordata (e.g. acorn worms) and Echinodermata (e.g. sea urchins and starfish) compose the superphylum Deuterostomia (Lowe et al. 2015) (Figure 1.9). Molecular evidence has suggested that Cephalochordates diverged before the split between Tunicates and Vertebrates and that the Hemichordates and Echinoderms clade (called Ambulacraria) form a sister grouping to the Chordates (Lowe et al. 2015). Some believe that Xenoturbella should be classified as a Deuterostome (Philippe et al. 2011a), while others do not support this hypothesis (Hejnol et al. 2009). It is clear that the origin of the chordates is an area of active research and debate, with several alternative hypotheses under consideration (Lowe et al. 2015). In terms of marking a major transition, the evolution of chordates marks the evolution of efficient movement. A chordate possesses all of the following distinguishable features at some stage in their lifecycle: a notochord (a stiff rod of cartilage that can develop into the spine and helps aquatic animals flex for

Figure 1.9: The phylogeny of the major phyla within Deuterostomia

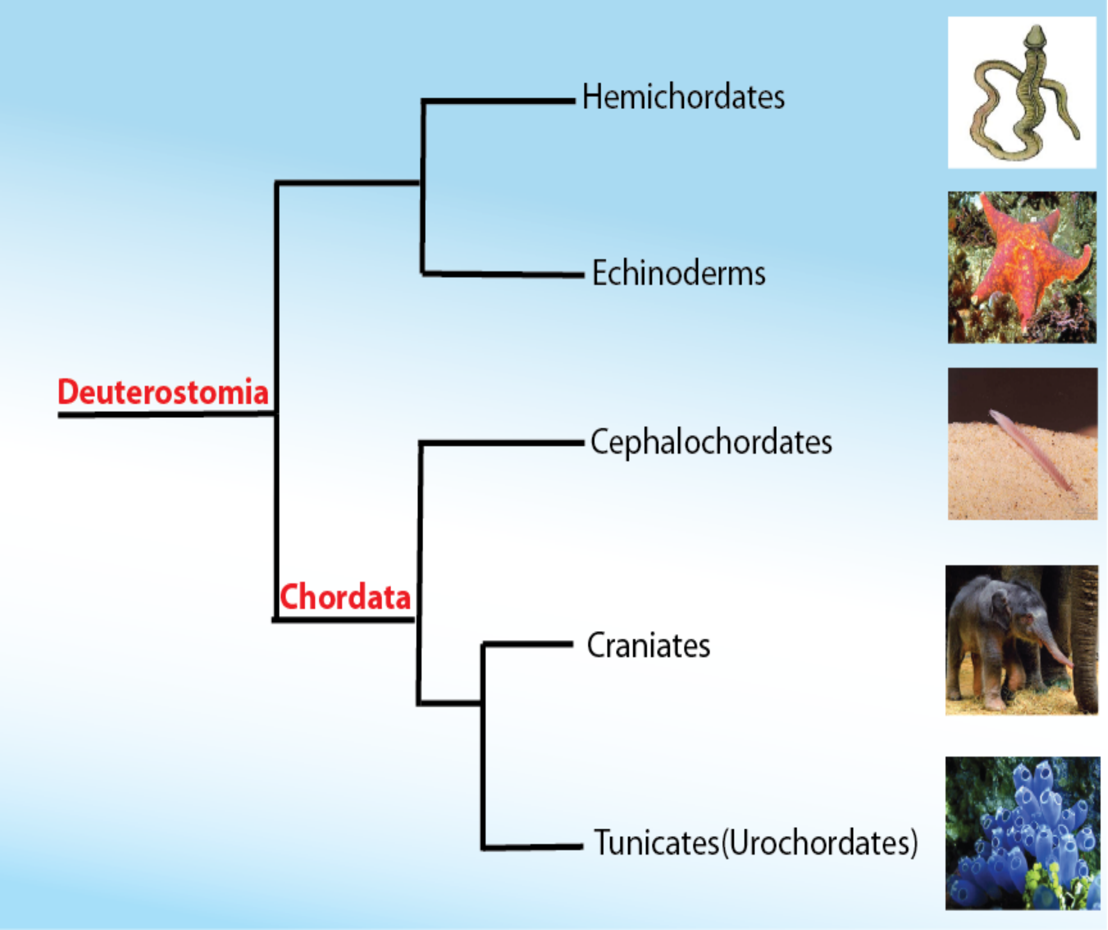


Figure 1.9: A well-supported topology for the Superphylum Deuterostomia. The phylogeny depicts the relationships amongst each of the major groups of Deuterostomes and chordates. A single representative of each major group is depicted in a photo beside that group to highlight the diversity of phenotypes: Hemichordates represented here by the acorn worm, Echinoderms by the starfish, Cephalochordates by the lancelet, Craniates by the elephant and Tunicates by the bluebell tunicate.

swimming), a dorsal neural tube (this develops into the spinal chord in fish and other vertebrates), pharyngeal slits (these are a modified form of gills that can also act as a filter feeding system in some aquatic chordates), a post-anal tail (a muscular tail that extends backwards behind the anus) and an endostyle (a gland that produces mucus to assist with the transportation of food from filter feeding).

1.4.5 Inversion of dorsoventral axis

A major difference between chordates and Protostomes is that their heart and nerve cord are inversely orientated, the dorsoventral (DV) axis is inverted (De Robertis and Sasai 2000). An alternative to the axis being inverted is that the common ancestor of Protostomes and Deuterostomes had only diffused nerve clustering with no distinct DV localization and the apparent inversion is actually the result of a concentration of the nervous system in opposite poles in separate lineages that developed independently in the Deuterostomes and Protostomes (Gerhart 2000).

1.4.6 Origin of a distinct head – the Craniata

Craniates possess a bony cartilaginous or fibrous structure surrounding the brain call the cranium, a jaw and facial bones (Bell 2015). *Petromyzontida* (lampreys), *Myxini* (hagfishes) and *Gnathostomata* (jawed vertebrates) form the Craniata. It is believed that the earliest branching vertebrates (diverging ~440 MYA) are the jawless Agnathans hagfish and lamprey. While the Agnathan group lack a jaw, they do possess a cartilaginous cranium. Lamprey and hagfish differ significantly in that Lamprey possess cartilaginous vertebrae while hagfish don't have any vertebrae (Morris et al. 2013, Bell 2015). It is postulated that this trait was lost in the hagfish ancestral lineage (Ota et al. 2011). The position of the major groups within the Craniata is also debated with some placing lamprey more closely to Gnathostomata and hagfish as sister to all other Craniates. However, more recent molecular studies have placed Agnathans as sister to all other vertebrates (Gnathostomata) (Kuraku et al. 1999, Delarbre et al. 2002, Furlong and Holland 2002, Heimberg et al. 2010, Janvier 2015).

1.4.7 Evolution of hinged jaw – the Gnathostomata

An articulated jaw is a unique characteristic of Gnathostomes and is seen as a major innovation that provided the jawed vertebrate with a distinct advantage over many jawless fish predators (Gans and Northcutt 1983, Cerny et al. 2010, Brazeau and

Friedman 2015). Ancestral vertebrates had a simple skeleton that was composed hinged jaw evolved from the unhinged gill bar in the jawless ancestor (Cerny et al. 2010) and studies continue to address this key question (Brazeau and Friedman 2015).

1.4.8 From Water to Land – breathing air

Evidence suggests that tetrapods (four-limbed vertebrates) evolved the ability to breathe air as ancestral fish in water. Polypterids (Actinopterygian lineage that are a part of the Craniata phylum) have the ability to breath in water and on land/aerially. They breathe using lungs similar to lungfish and tetrapods. However, a key component of their breathing apparatus are spiracles. Spiracles are large canals located on the head and they appear in the fossil record in the earliest known tetrapods. Spiracles aided early stem tetrapods to respire near the shoreline and it is believed that spiracles were crucial in the transition from water to land (Graham et al. 2014).

1.4.9 From Water to Land - Development of walking legs

The ability to walk on dry land, the development of walking limbs (~365 MYA), is seen as one of the most important innovations in vertebrate evolution. *Ichthyostega* and *Acanthostega* are early tetrapods having diverged ~365 MYA (Clack 2005, Coyne 2009). For the most part, *Acanthostega* and *Ichthyostega* are not well adapted for terrestrial life. Their limbs were short, stubby and had 8 digits but these limbs were unable to support their weight on land. It is proposed that the tetrapod limb initially evolved in aquatic species where it's function was in movement. Given these features extracted from fossil remains it is widely held that *Acanthostega* and *Ichthyostega* were entirely aquatic and that the origin of land living tetrapods therefore involved preadaptation. Preadaptation is when a biological structure adapted for one particular function in a given environment can be used for a different function in a different environment. For example, lungs and limbs existed prior to land-living tetrapods. The function of lungs and limbs in an aquatic species has been proposed as follows: ancestors of tetrapods were adapted to life in water, but also in shallow water adjacent to the banks and lungs and limbs allowed these animals to temporarily cross mud or shallow water barriers when required (Kemp 2005).

1.4.10 Development of a mineralized skeleton

Another major innovation within the vertebrate clade is the development of a mineralized skeleton. The earliest diverging extant vertebrates with mineralized skeleton tissues are sharks and bony fish (~440 MYA), whereas Lamprey and Hagfish (~510MYA) do not have any mineralized skeletal tissue (Sansom and Downs , Sidow 1992). The distinct advantage with a mineralized skeleton is protection of internal organs and improved locomotion. Vertebrates have two types of skeletal tissue: an exoskeleton (e.g. teeth, scales and fins) and endoskeleton (e.g. neurocranium and vertebrae). It is believed that a mineralized endoskeleton evolved from an exoskeleton progenitor. An endoskeleton has better weight bearing properties than an exoskeleton (it can support greater weight without becoming too heavy for an organism). In addition to this, an exoskeleton limits the size as it can become impractically heavy for an organism at a certain point. Tetrapod evolution favoured a mineralized endoskeleton, thus explaining a minimal exoskeleton and full endoskeleton (Shimada et al. 2013).

1.4.11 Development of amniotic egg

Being able to reproduce on land and independent of water for fertilization and development was a major innovation that greatly facilitated the colonization of land in both plants and animals. In the metazoan tree, it was the reptiles that were the first to fully adapt to life out of water, with the evolution of the amniotic egg, which freed them from a dependency on water for reproduction. Amniotes include reptiles, birds and mammals. The amniotic egg contains a food source (egg yolk), an enclosing membrane (the chorion), an internal membrane that retains water but exchanges gases and removes waster (the allantois), and a fluid filled sac (the amnion) that protects the embryo. The advantage of the amniotic egg is that the relatively water-impermeable shell protects the embryo from drying out, while also providing sufficient nutrition and gas exchange for complete development (Ashley-Ross et al. 2013). Species of bony fish, amphibia and reptiles have been reported to use both viviparity (live birth strategy i.e. no egg) and oviparity (egg-laying strategy). All birds and some mammals (monotremes) are oviparous and all therian mammals (marsupial and placental) are viviparous. The next evolutionary advancement in reproduction that we will briefly discuss is placentation and live birth in mammals (Tian et al. 2010).

1.4.12 Placentation

The placental mammals are the most phenotypically diverse group of vertebrates spanning volant, aquatic and terrestrial niches (Wilson and Reeder 2005). The first placental mammal is thought to have lived ~125MYA (Kemp 2005). The origin of placental mammals has sparked much debate, with some regarding it an almost impossible task to solve (Teeling and Hedges 2013). In Chapter 2 we address this question and pose a solution to the radiation of mammals. While there are variations on placental types based on shape and invasiveness, in general the embryo is surrounded by amniotic fluid (amnion), the umbilical cord functions similarly to the allantois and yolk of the amniotic egg, providing food and removing waste and the accumulation of these placental membranes allow for food, air and waste to be transferred (Mess and Carter 2007). The distinct advantage of the placenta is it allows the offspring to be larger and more quickly independent when born (Morris et al. 2013).

These are only a sample of some well-characterised examples of major phenotypic change observed in the evolution of the Metazoa. In Chapter 3 we apply a network based approach to determine if gene remodelling correlates with these major phenotypic transitions and if so what functions have evolved in protein families at these time points.

Chapter 2: On Model misspecification and phylogeny reconstruction - resolving the root of the placental mammal tree

2.1 Introduction

There are numerous phylogenetic reconstruction methods and models available—but which should you use and why? Important considerations in phylogenetic analyses include data quality, structure, signal, alignment length and sampling. If poorly modelled, variation in rates of change across proteins and across lineages can lead to incorrect phylogeny reconstruction that can then lead to downstream misinterpretation of the underlying data. The risk of choosing and applying an inappropriate model can be reduced with some critical yet straightforward steps outlined in this chapter. We use the question of the position of the root of placental mammals as our working example to illustrate the topological impact of model misspecification. Using this case study, we focus on using models in a Bayesian framework and we outline the steps involved in identifying and assessing better fitting models for specific datasets.

Placental mammals are made up of three main clades: Afrotheria (Afroinsectiphilia and Paenungulata), Xenarthra (Cingulata and Pilosa), and Boreoeutheria (Laurasiatheria and Euarchontoglires). Placing the root of the placental mammal tree has been a controversial topic (Teeling and Hedges 2013) and some have argued that this is an impossible task (Teeling and Hedges 2013). Many studies have concluded conflicting positions, sometimes even from the same pool of data (Springer et al. 2004, Murphy et al. 2007, Prasad et al. 2008, Morgan et al. 2013, Romiguier et al. 2013). In 2013, two molecular studies narrowed down the root of the placental mammal phylogeny to one of two possibilities: (i) an Afrotherian root which places mammals such as elephants as the earliest diverging placental mammal group (Romiguier et al. 2013); and (ii) an Atlantogenata root, placing the common ancestor of the Xenarthra and Afrotheria as the earliest diverging placental mammal group (Morgan et al. 2013) (Figure 2.1).

In addition, it has been shown that previous studies of the position of the root of placental mammals lacked definitive resolution because of suboptimal models and datasets of low power (Morgan et al. 2013). Romiguier et al. (2013) proposed that GC-rich genes lead to erroneous topologies (as GC is a proxy for recombination), and therefore that data partitioning should play a major role in the resolution of the

Figure 2.1: Conflicting positions for the root of the placental mammal tree

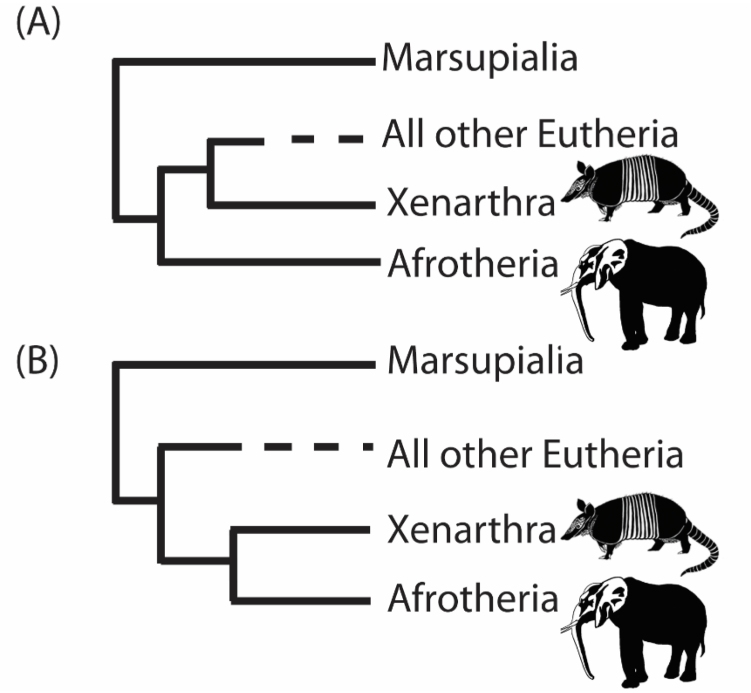


Figure 2.1: Conflicting positions for the root of the placental mammal tree. (A) Afrotheria position retrieved from alignment and model (GTR) used in the Romiguier et. al study (2013) (result = Afrotheria position for the root) and (B) Atlantogenata position retrieved using the 39 taxa mammal dataset from Morgan et al. (Morgan et al. 2013) with the CAT-GTR model in PhyloBayes (Figure taken with permission from Morgan et al. 2013).

mammal phylogeny. Evidently, the data and subsequent model used can greatly impact the result of the phylogenetic reconstruction. We propose that the conflicting positions for the root are due to model misspecification. We set out to assess whether the models applied in the Romiguier et al. study were adequate for the data used (Romiguier et al. 2013). In this study a set of mammal single gene orthologues were assembled from 39 mammal species and each of the 13,111 single gene families were categorized as GC-rich (>40% GC3 content) or AT-rich. The phylogenetic analysis of these datasets showed that GC-rich genes are most likely to support erroneous topologies and that AT-rich genes have a rate of error five times lower than that of GC-rich genes. The AT-rich genes supported an Afrotheria rooting, while the GC-rich genes supported an Atlantogenata rooting. The authors argued that the use of GC-rich genes contributes to conflict in the placement of the root of the placental mammal tree.

2.2 Materials and methods

2.2.1 Dataset assembly

All datasets were created from the single gene orthologs (SGO) used in (Romiguier et al. 2013). We downloaded these SGOs from OrthoMaM v7 (Ranwez et al. 2007). The two datasets we use are summarized in Table 2.1 and briefly are composed of the following:

(i) **RomiguierTopAT:** This is a concatenated alignment using 100 of the most AT-rich genes. We wished to assess the impact of heterogeneous modelling on these data and on the support for the proposed Afrotheria hypothesis.

(ii) **Subsets 1–8:** Each of the eight subsets are a concatenated alignment generated from 25 GC-rich genes (i.e. GC content >40%) chosen at random. We wished to assess whether the P4 heterogeneous models can adequately model the GC variation.

2.2.2 Assessment of phylogenetic signal using Likelihood mapping

Using TreePuzzle v5.2 (Schmidt et al. 2002) we wanted to ensure that the RomiguierTopAT contained adequate phylogenetic signal. Likelihood mapping is a method to visualize the phylogenetic content of an alignment (Figure 2.2). In brief, the input alignment is split up into quartets (4 sequences). Each quartet is assed in a maximum

Table 2.1: The nucleotide composition of datasets RomiguierTopAT and Subsets 1-8

	GC3 %	Length	A(%)	C(%)	G(%)	T(%)
(A) Subsets 1–8 properties						
Subsets 1	44.55653224	30587	22.5	32	27.6	17.9
Subsets 2	43.54061239	28182	25	28.5	25.8	20.7
Subsets 3	42.11977501	27203	25.7	27.4	25.4	21.5
Subsets 4	45.42805129	28904	22.9	30.6	26	20.5
Subsets 5	41.09068538	28384	24.6	28.7	25.5	21.2
Subsets 6	40.77550101	34474	25.9	26.3	23.5	24.3
Subsets 7	41.32848377	30201	26.4	27.1	23.9	22.6
Subsets 8	40.77129042	23531	26.6	26.4	25.4	21.7
(B) RomiguierTopAT						
RomiguierTopAT	14.66387233	81924	34.8	18.5	20.7	26

Table 2.1: Section (A) represents the nucleotide composition for Subsets 1-8, and Section (B) represents the nucleotide composition for the RomiguierTopAT dataset. GC3% - is the percentage of the specified dataset that is composed of guanine and cytosine and is calculated using the third codon position of the sequences. The length of each dataset is given in base pairs. The percentage of adenosine, cytosine, guanine and tyrosine as a percentage over the total number of bases are denoted as A(%); C(%); G(%), and T(%) respectively.

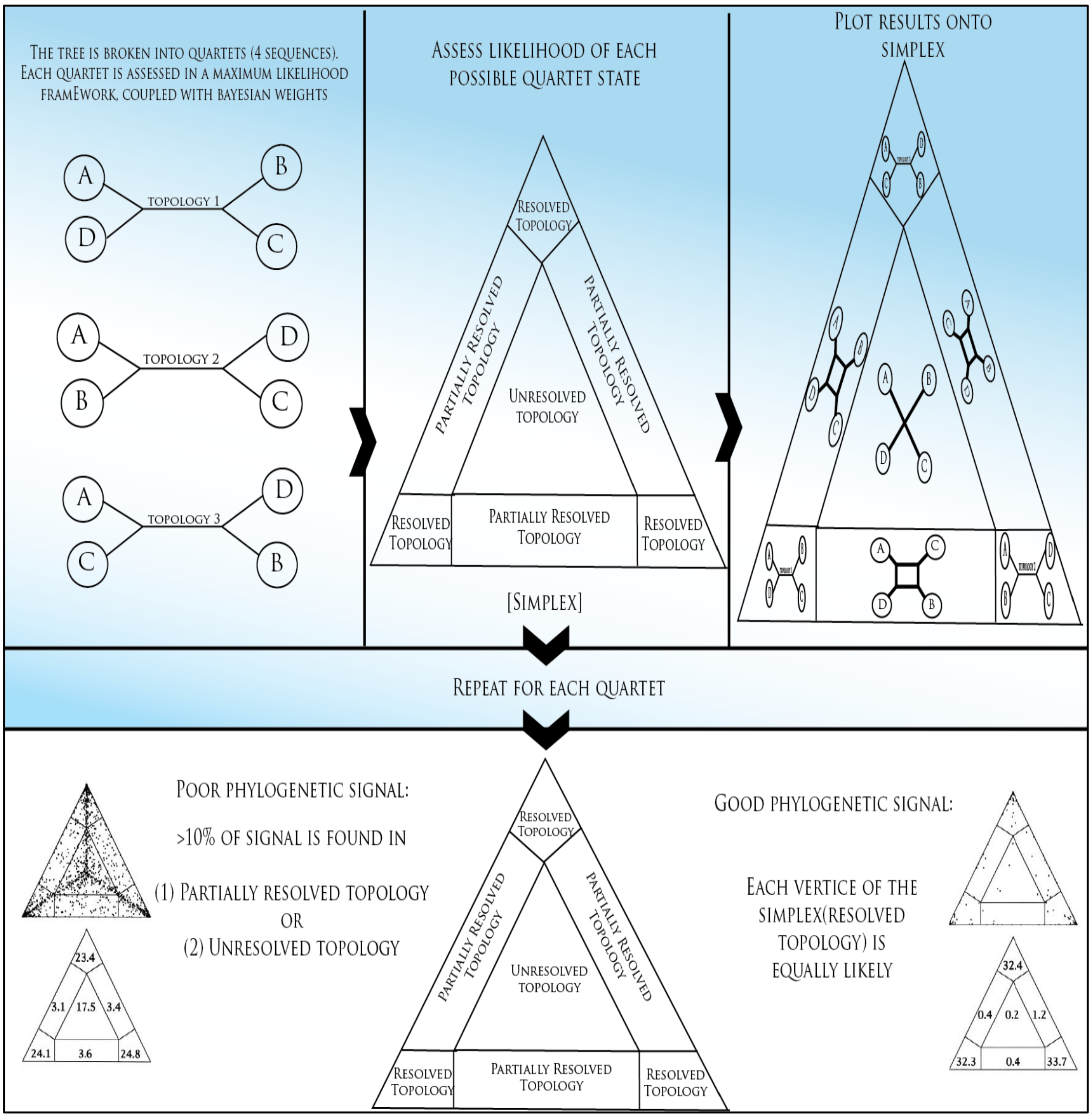
likelihood framework combined with Bayesian weights. Each quartet can have 3 possible topologies. However, there are 7 possible states: 3 can be fully resolved, 3 can be partially resolved and 1 can be an unresolved. Accordingly, there are seven areas of the likelihood mapping triangle (Figure 2.2) that a quartet state can occupy. This can theory can be used to test the strength of the phylogenetic signal within data. When the majority of the quartets can be fully resolved, the phylogenetic signal is strong. A weak phylogenetic signal is observed when >10% of the quartets cannot be fully resolved (partially resolved or unresolved) (Figure 2.2) (Strimmer and Von Haeseler 1997).

2.2.3 Phylogenetic reconstruction

2.2.3.1 Phylogenetic analysis using PhyloBayes

We ran the RomiguierTopAT alignment on PhyloBayes-MPI (Lartillot et al. 2013a). We ran this with nucleotide and amino acid data. In a PhyloBayes (Lartillot et al. 2009) analysis the number of biochemical categories is optimized during the tree search, therefore if the GTR model (which is nested within the CAT-GTR model) is a better fit to the data than the CAT-GTR model, then the CAT-GTR model will reduce to the simpler GTR model, thus relinquishing the need to independently test the GTR model on the data (Lartillot 2014). Hence, by running analyses under CAT-GTR+G we can avoid running independent model testing analyses to determine whether CAT-GTR or GTR fits the data best. This is a key point because Bayesian model selection analyses as implemented in PhyloBayes (Lartillot et al. 2009) to test alternative models (including the CAT-based ones) would be prohibitively computationally intensive given the size of dataset. For all PhyloBayes analyses two chains were run. Burn-in varied and all chains were run until convergence was achieved. Convergence was assessed using the BPCOMP software – part of the PhyloBayes package (Lartillot et al. 2009, Lartillot et al. 2013a). Following the recommendations in the PhyloBayes manual, all the chains of the MCMC run were considered to have converged on the same solution when the Maxdiff (maximal difference between observed bipartitions) dropped below 0.2 (Lartillot et al. 2013a).

Figure 2.2: Summary of the_Likelihood mapping approach.



The tree is broken into quartets (4 sequences). Each quartet is assessed in a maximum likelihood framEwork, coupled with Bayesian weights
A
B
D
C
topology 1
A
D
B
C
topology 2
A
D
C
B
topology 3
Assess likelihood of each possible quartet state
Resolved Topology
Partially Resolved Topology
Partially Resolved Topology
Unresolved topology
Resolved Topology
Partially Resolved Topology
Resolved Topology
[Simplex]
Plot results onto simplex
A
B
C
D
Repeat for each quartet
Poor phylogenetic signal:
>10% of signal is found in
(1) Partially resolved topology
or
(2) Unresolved topology
23.4
3.1
17.5
3.4
24.1
3.6
24.8
Resolved Topology
Partially Resolved Topology
Partially Resolved Topology
Unresolved topology
Resolved Topology
Partially Resolved Topology
Resolved Topology
Good phylogenetic signal:
Each vertice of the simplex(resolved topology) is equally likely
32.4
0.4
0.2
1.2
32.3
0.4
33.7

Figure 2.2: Starting in the top left panel and following the arrows through the various stages of the likelihood mapping process are outlined. The top left panel depicts where the input tree is broken into quartets (4 sequences). Each quartet is then assessed in a maximum likelihood framework, coupled with Bayesian weights. The top center panel depicts the assessment of the likelihood of each possible quartet state (resolved, unresolved or partially resolved). The top right panel depicts plotting results onto a simplex (triangle) within the seven quartet states. These stages are repeated for each quartet and the results are summarized graphically as per the bottom panel.

2.2.3.2 Phylogenetic analysis using p4

We analysed the RomiguierTopAT and the Subsets1-8 in p4 (Foster 2004). Each alignment was treated as outlined in Figure 2.3. Each model was run with 5 independent runs until convergence was reached and topological agreement between independent runs was sought. We examined how the MCMC run progressed via the checkpoints of the MCMC run by comparing the ASDOSS values (average standard deviation of split supports) at the checkpoints where ASDOSS <0.01 is indicative of convergence. We plotted the likelihood values from the samples taken during the MCMC run to ensure that the likelihood values had reached plateau before completion. We used the parameter acceptance rates to get a rough indication of mixing and convergence, e.g. ideally all parameters would be accepted between 10%-70% of the time. However, in p4 this is difficult to achieve because the composition and rate parameters can only be adjusted globally. For example, one composition vector may be accepted 40% of the time, while 2 out of 3 of the composition vectors are only accepted 5% of the time. There is no way around this issue it is due to the nature of the algorithm (Foster 2004).

2.2.4 Test of Compositional homogeneity

We employ tree- and model-based composition fit tests as described by Foster (2004). These tests are done in a maximum likelihood framework. Foster (2004) illustrates that the traditional χ2 test used for assessing compositional homogeneity presents a problem. More specifically, this test can have a large probability of type II error (failure to reject a false null hypothesis i.e. a false positive). This is because the traditional test uses a χ2 curve as a null distribution; this is not appropriate to assess the significance of the χ2 value. Essentially, the test described by Foster (2004) is superior to the traditional χ2 test because a null distribution is gained through calculating χ2 values from simulated data (simulations based on the appropriate model and tree). This test essentially asks the following question: does my possibly heterogeneous data fit the possibly heterogeneous model? (Foster 2004).

Figure 2.3: Flowchart describing the steps involved in phylogenetic reconstruction using heterogeneous models.

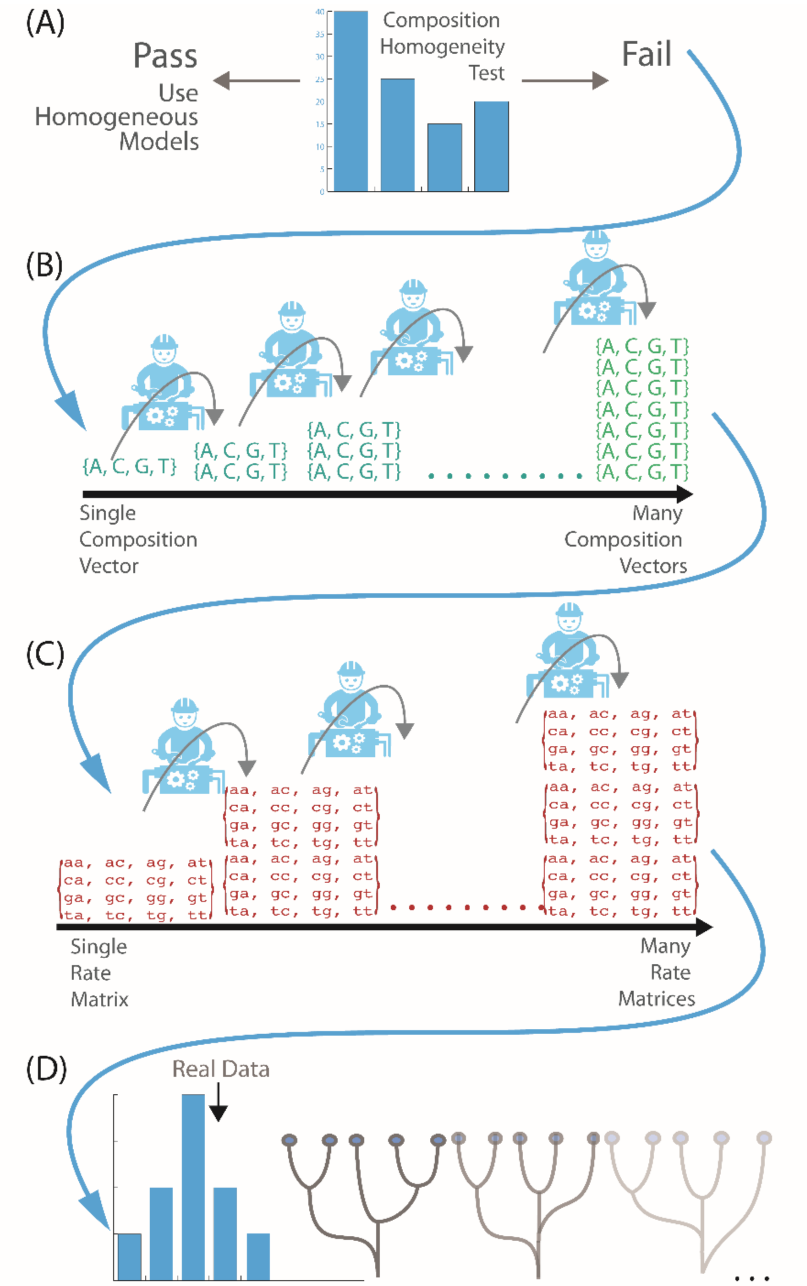


Figure 2.3: This image is taken from (Moran et al. 2015). All the major steps we have described in performing a phylogenetic reconstruction using heterogeneous models are summarized here. (A) Starting with a multiple sequence alignment the test for compositional homogeneity is carried out, failing this test means the data must be modelled using a heterogeneous model. There are two phases to generating the heterogeneous model of best fit in P4 these are depicted in (B) and (C); (B) The first

phase to estimate the number of composition vectors needed for the dataset; (C) The second phase is to estimate the number of exchange rate matrices; (D) Depicts the assessment of the model adequacy using posterior predictive simulations and the generation of the trees from which a consensus tree is generated.

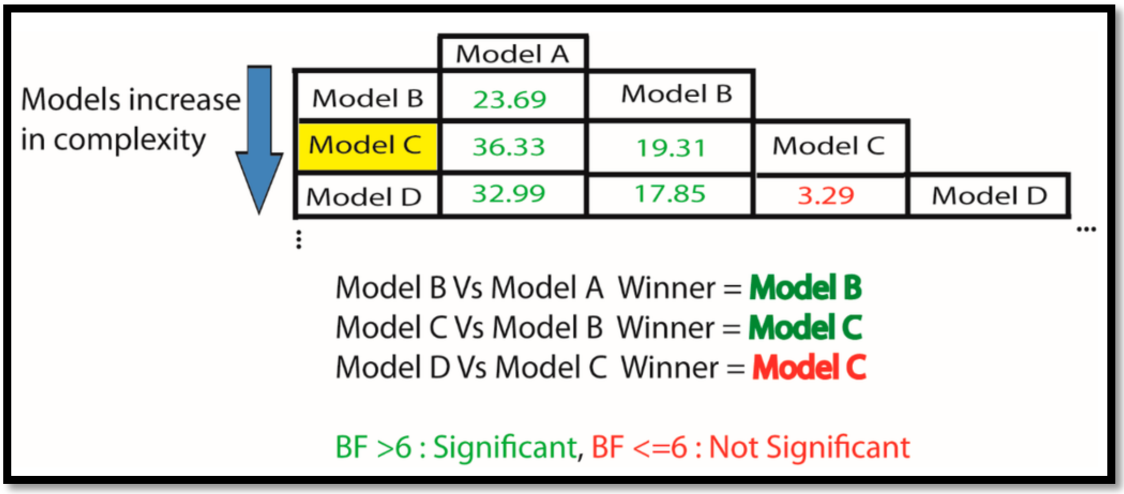
2.2.5 How Do I Select the Optimum Model for My Data? – Bayes Factor Analysis

Bayes Factor (BF) analysis is a Bayesian approach to select between nested models (Figure 2.4) and this is particularly useful in P4 (Foster 2004). In P4 the user can vary the number of parameters of the core model, for example, the user can test a GTR model with two composition vectors and one rate matrix and compare this to a GTR model with three composition vectors and one rate matrix. There are many permutations of different numbers of composition vectors and rate matrices that the user can generate. BF analysis allows the user to find which one of the models fits the data best, always selecting for the model with the lowest number of parameters. Using Newton Raftery Equation (16) (Newton and Raftery 1994) within the P4 environment, the user can calculate the logarithm of the marginal likelihood and then compare the P4 models using the following test: 2[lnL(model B) − lnL(model A)] for each possible pairwise comparison of models. We compare the resulting value to the Kass and Raftery table (Kass and Raftery 1995). If the value is greater than 6, then there is a strong indication that model B is a better fit to the data than model A. We then compare model C to model B in the same way and so on until we get a BF below 6. In which case, the model being tested is not a better fit to the data. There is no need to carry on at this point with the comparisons to the remaining models because we observe an increase parameterization that does not improve the fit (Figure 2.4) (Kass and Raftery 1995, Foster 2004, Lopes and West 2004). It is important that Bayes factor analysis and posterior predictive simulations described in step (vi) below are both used when one considers does my model fit my data significantly?

2.2.6 Does My Model Fit My Data? – Posterior predictive simulations

Posterior predictive simulation (PPS) is a Bayesian model fit test that can be applied to both heterogeneous models and homogeneous models to give statistical support for how well a model describes a specific dataset (Gelman et al. 2014) Data is simulated under this model, and if the model describes the real data well, then our simulated data should be similar to the real data. Data is simulated based on the model used for the real data analysis and all the parameter estimations retrieved during the MCMC run. The simulated data can be compared to the real data through a single parameter—commonly the composition parameter is used for comparison. It is suggested that

Figure 2.4 Overview of a sample Bayes factor analysis.



	Model A	Model B	Model C	Model D
Model B	23.69			
Model C	36.33	19.31		
Model D	32.99	17.85	3.29	

Model B Vs Model A Winner = **Model B**

Model C Vs Model B Winner = **Model C**

Model D Vs Model C Winner = **Model C**

BF >6 : Significant, BF <=6 : Not Significant

Figure 2.4: Taken from (Moran et al. 2015). Models are denoted as A, B, C and D and parameters increase from A to D (i.e., Model A is the least parameterized model). They are compared in a pairwise manner as denoted by the lower triangular matrix. Each pairwise comparison results in a score (or Bayes Factor) by calculating 2[lnL(Model B) − lnL(Model A)]. Significance is based on the Kass and Raftery equation (Kass and Raftery 1995). Cases where the Bayes Factor comparison yields a significant result are shown in green and models that do not significantly improve the fit to the data are given in red. Model C is selected as the best-fit model as an increase in parameters in Model D does not significantly improve the model fit. Therefore, we do not continue to search more parameter rich models and we choose Model C (highlighted in yellow).

multinomial likelihood is used to test overall model fit and that the composition parameter is used to test how well this parameter is modelling the composition of the data (Huelsenbeck et al. 2001).Considering the resultant graphs from a PPS analysis, a model that fits well will plot the real test statistic close to the central mass of the histogram that is composed of the simulated test statistics. From a statistical perspective, a tail area probability of 0.5 is optimal and indicates the model describes the data well (Foster 2004). In PhyloBayes the output of the PPS is given as a Z-score, where a Z-score >2 is an indication that the model does not fit the data (Lartillot et al. 2009).

2.3 Results

2.3.1 Likelihood mapping reveals phylogenetic signal is present in RomiguierTopAT

The likelihood mapping (LM) results are displayed in Figure 2.5. Topologies at the vertices of the simplex form 94.9% of the quartets tested. 4.4% of the quartets were partially resolved and 0.7% of the quartets were unresolved. Therefore, the RomiguierTopAT alignment contains a significant phylogenetic signal.

2.3.2 RomiguierTopAT is compositionally heterogeneous

Following model selection analysis using ModelGenerator (Keane et al. 2004), the GTR model was selected. The results of the compositional homogeneity test showed that for the RomiguierTopAT of the 39 taxa, there were 12 that did not fit the homogeneous model. Therefore, whilst the RomiguierTopAT dataset is compositionally heterogeneous but it was analysed using a compositionally homogeneous model in the original study.

2.3.3 Optimum model for RomiguierTopAT is a heterogeneous model

BF analysis revealed that our optimum model for this dataset is annotated as 2GTR + 4C + 4G (2 GTR rate matrices, four composition vectors and four discrete gamma categories). The biggest improvement in BF score was 1281 and was for the comparison of the homogeneous model (1GTR + 1C + 4G) to the first heterogeneous model (1GTR + 2C + 4G). In addition to this, the number of composition vectors appears more important to model the data than the number of rate vectors (Figure 2.6).

Figure 2.5: Results of Likelihood mapping for the RomiguierTopAT alignment

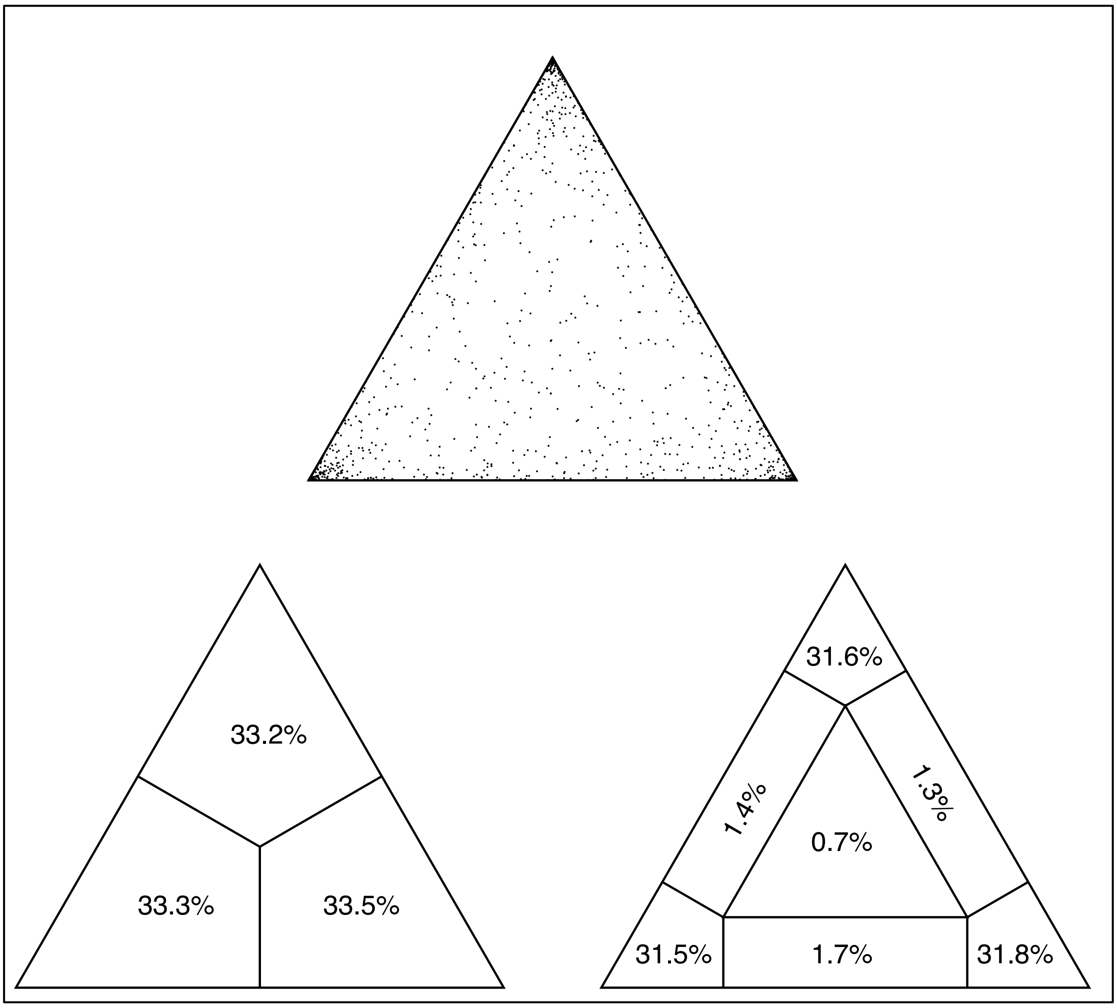


Figure 2.5: The 3 triangles show the distribution of quartets (each dot in the top triangle represents a quartet) within Likelihood mapping space (a complete triangle). The vertices denote the proportion of quartets for a particular topology (each alternative topology, partially resolved and unresolved). The bottom left triangle represents support for each of the three quartets. The bottom right triangle has seven state spaces: each vertex is a fully resolved quartet, the sides between the vertices represent quartets that cannot be fully resolved and the central area of the triangle represents quartets that cannot be resolved. The numbers represent the proportion of all quartets that fall within that space.

2.3.4 P4 cannot model the RomiguierTopAT Alignment

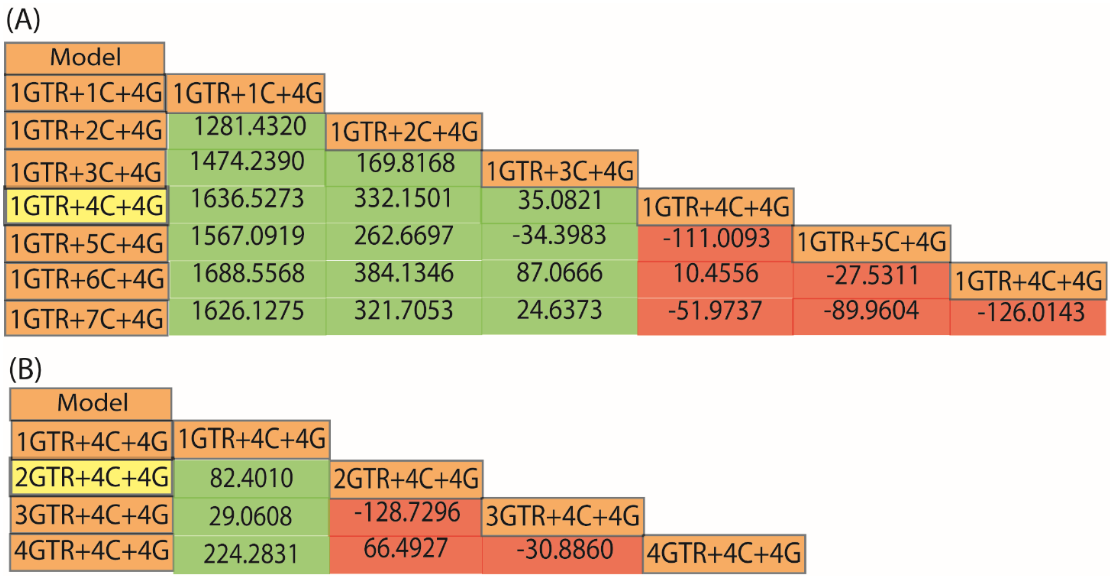
A consensus tree was made from the MCMC run for the RomiguierTopAT dataset under the 2GTR + 4C + 4G model. The topology supported the Afrotheria position of the root. However, posterior predictive simulations (PPS) reveal a poor tail area probability of 0.00, showing that this model does not adequately describe the data and that the resultant topology is not to be trusted. We went on to assess whether any of the 24 P4 models tested had a significant tail area probability for RomiguierTopAT. Surprisingly, none of the models had a significant tail area probability; in fact, all models we tested had a tail area probability of 0.00, suggesting that P4 cannot model this data adequately with the models available. However, the simulated dataset produced using the heterogeneous P4 models are a better fit to the real data than the homogeneous model (Figure 2.7) although as would be expected in this scenario the topology is not well supported. This is unusual as P4 is generally capable of showing some model fit. The RomiguierTopAT dataset seems to be very difficult to model adequately, perhaps this is due to a high level of heterogeneity across sites and between lineages in combination. But it is clear from these analyses that homogeneous modelling is too simplistic and not realistic for this dataset and that additional improvements need to be made to heterogeneous models to capture the complexity of the data.

The analysis of Subsets 1–8 in P4 showed some model fit for all these subsets and in all cases of Subsets1–8 the heterogeneous model was a better fitting model than a homogeneous. In most cases (6/8 of the subsets) the heterogeneous model significantly fits the data (Figure 2.8). These results for Subsets 1–8 illustrate that heterogeneous models are capable of adequately modelling complex data and are more appropriate than homogeneous models for heterogeneous data.

2.3.5 PhyloBayes can model the RomiguierTopAT alignment and finds strong support for Atlantogenata rooting

For the RomiguierTopAT, dataset the CAT–GTR model was found to fit the data better than the site-homogeneous models used in the original studies (LG) (Table 2.2). For RomiguierTopAT the

Figure 2.6: Bayes factor analysis for RomiguierTopAT with heterogeneous models.



(A)

Model						
1GTR+1C+4G	1GTR+1C+4G					
1GTR+2C+4G	1281.4320	1GTR+2C+4G				
1GTR+3C+4G	1474.2390	169.8168	1GTR+3C+4G			
1GTR+4C+4G	1636.5273	332.1501	35.0821	1GTR+4C+4G		
1GTR+5C+4G	1567.0919	262.6697	-34.3983	-111.0093	1GTR+5C+4G	
1GTR+6C+4G	1688.5568	384.1346	87.0666	10.4556	-27.5311	1GTR+4C+4G
1GTR+7C+4G	1626.1275	321.7053	24.6373	-51.9737	-89.9604	-126.0143

(B)

Model				
1GTR+4C+4G	1GTR+4C+4G			
2GTR+4C+4G	82.4010	2GTR+4C+4G		
3GTR+4C+4G	29.0608	-128.7296	3GTR+4C+4G	
4GTR+4C+4G	224.2831	66.4927	-30.8860	4GTR+4C+4G

Figure 2.6: Taken from (Moran et al. 2015). The results depicted in (A) and (B) are for RomiguierTopAT, in both cases the BF result indicating the model of best fit is depicted in yellow. The overall model of best fit (for composition vectors and rate matrices) is 2GTR + 4C + 4G. (A) The lower triangular matrix of Bayes Factors used to find the optimum number of composition vectors for the model; (B) The lower triangular matrix of Bayes factors used to find the optimum number of rate matrices. Combined they produce the best-fit model for the specific dataset. In both (A) and (B) significant improvements in model fit as judged by Kass and Raftery criteria (Kass and Raftery 1995) are shown in green and non-significant in red.

Figure 2.7: Assessment of fit of the homogeneous and heterogeneous models for the RomiguierTopAT dataset.

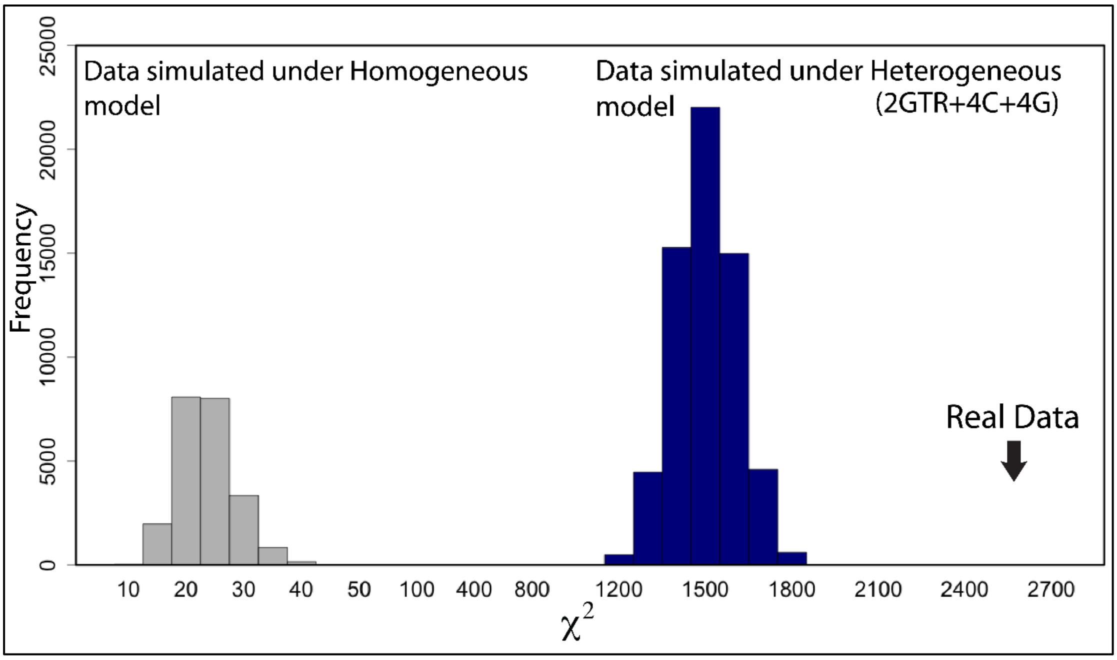
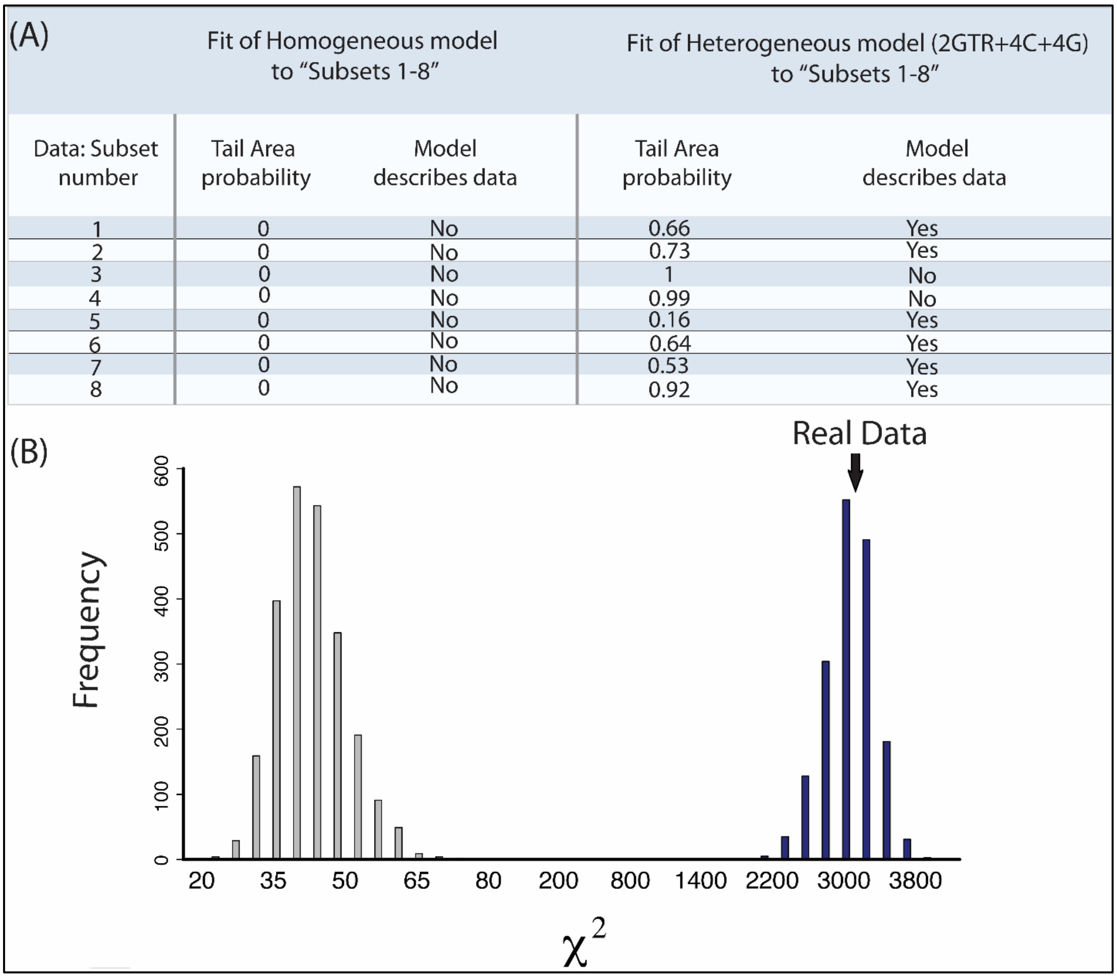


Figure 2.7: Posterior predictive simulations calculated for the RomiguierTopAT dataset are shown. The leftmost histogram in grey shows the results of the posterior predictive simulation for RomiguierTopAT using the homogeneous model. The rightmost histogram in blue shows the improved posterior predictive simulation on modelling RomiguierTopAT with the heterogeneous model of best fit (i.e., 2GTR + 4C + 4G for RomiguierTopAT). In both simulations, the X-axis represents the parameter of comparison taken from the simulations, and the Y-axis represents the frequency. The large black arrow indicates the value of test statistic (χ 2) retrieved from the real data.

Figure 2.8: Assessment of fit of Homogeneous and Heterogeneous models for Subsets 1–8.



(A)	Fit of Homogeneous model to "Subsets 1-8"		Fit of Heterogeneous model (2GTR+4C+4G) to "Subsets 1-8"	
Data: Subset number	Tail Area probability	Model describes data	Tail Area probability	Model describes data
1	0	No	0.66	Yes
2	0	No	0.73	Yes
3	0	No	1	No
4	0	No	0.99	No
5	0	No	0.16	Yes
6	0	No	0.64	Yes
7	0	No	0.53	Yes
8	0	No	0.92	Yes

Figure 2.8: (A) Posterior predictive simulations for the Subsets 1–8 are shown. The leftmost Section of the table shows the results for the homogeneous model. The rightmost Section of the table shows the improved tail area probability on modeling the data with heterogeneous models. Values closer to zero represent a poorer fit of the model to the data; (B) The results of the posterior predictive simulations for one subset (Subset 1) are shown in detail. Again, the simulation for the homogeneous GTR model is shown in grey and the heterogeneous P4 model of best fit is shown in blue (i.e., 2GTR + 4C + 4G). In both simulations, the X-axis represents the parameter of comparison taken from the simulations, and the Y-axis represents the frequency. The large black arrow indicates the value of test statistic (χ 2) retrieved from the real data.

Table 2.2: Comparing posterior predictive simulations (PPS) for the LG and CAT-GTR models for the RomiguierTopAT dataset

Model	LG	CAT-GTR
Composition parameter of original data	2.19977	2.19977
Composition parameter of simulated data	2.33305	2.20858
Tail area probability	0	0.436695

Table 2.2: The composition is estimated from the original RomiguierTopAT dataset during the MCMC process in PhyloBayes. During the MCMC process data is simulated for each model and the composition is estimated from the simulated data. The original and the simulated compositions are then compared. Tail area probability of 0.5 is reflective of compositions that are not significantly different.

model did not boil down to a GTR+G model during the tree search, confirming that the original analyses were not performed under the best fitting model in the original study. The PhyloBayes CAT–GTR model supports an Atlantogenata root and not an Afrotheria root as found in the original study that used homogenous models.

2.4 Discussion

Combining the sophisticated modelling approach in P4 (Foster 2004) as described in Morgan et al. (2013) (Morgan et al. 2013) and the data partitioning method of Romiguier, Ranwez et al. (2013) (Romiguier et al. 2013), we show that homogeneous models do not describe the RomiguierTopAT data adequately and therefore any topology resulting from that model is not significant and certainly not reliable. While software packages such as RAxML (Stamatakis 2006b) can be incredibly powerful and fast, they are not always appropriate and this depends on the dataset. The results for the datasets Subsets 1–8 indicate that modelling heterogeneity across the phylogeny will not always work, again this is dataset dependent. In this case, we searched for the optimum heterogeneous GTR model and still failed to adequately model the data. This serves to highlight the importance of checking the adequacy of the model for each specific dataset. One cannot assume that any heterogeneous model is adequate. One can only trust the phylogeny produced with a model that adequately and statistically significantly models that data. However, the question the user needs to ask is does this model adequately describe my data and if the answer is no then the resultant topology is not reliable. The key is to resist making assumptions of the data and instead let the data dictate the approach and the model. As genome-scale phylogenies become commonplace and as consilience across datatypes rapidly takes hold as best practice, the importance of developing methods that can efficiently and adequately model data is paramount.

The heterogeneous modelling approach applied using P4 models of heterogeneity were not capable of modelling the data adequately. However, the PhyloBayes model of heterogeneity (CAT-GTR) was capable of modelling the data. Using the CAT-GTR model on the RomiguierTopAT data we were able to show that this model could adequately describe the data using posterior predictive simulations. Furthermore, posterior predictive simulations show that the original model used in the Romiguier et. al. study (2013) was inadequate for the data (Romiguier et al. 2013). We show that the model that adequately describes the data provides support for the Atlantogenata hypothesis. It has been stated that solving the root of the placental mammals was impossible to solve largely due to factors such as incomplete lineage sorting (ILS) (Teeling and Hedges 2013), reflected in large scale gene tree heterogeneity and a result of the apparent rapidity of successive vicariance-driven divergence events associated with the fragmentation of the Pangaean and Gondwanan supercontinents (Nishihara et al. 2009). However, it has been found that the levels of ILS present during placental diversification do not obscure the true relationships within the superorders of placental mammals (Tarver et al. 2016).

In exploring models of sequence change in this chapter it is clear that phylogenetic models can be designed that can accurately capture the evolutionary history of some sequences. However, there are mechanisms of protein coding gene evolution that do not follow a tree-like pattern of evolution, i.e. gene remodelling. Little is currently known about the general rules around non-tree like processes and modelling of these processes is in its infancy. These non-tree like processes include gene fusion/fission (Chapter 3) domain and exon shuffling (Chapter 4).

2.5 Conclusion

In summary, we have got consilience for the root of the placental mammal tree across different data types using a data driven method that incorporates both homogenous and heterogeneous models. Reanalysis of previous studies revealed that inadequate modelling was the primary cause of support for alternative topologies for the root of the placental mammals. When a model that fits the data is used, only one hypothesis was supported: an Atlantogenata rooting.

Chapter 3 - Major protein-coding innovation by gene remodeling in the animal kingdom.

3.1 Introduction

Understanding the origin of genetic and functional novelty across the *Metazoa* (i.e. all multicullar animals) is a fundamental problem in modern biology. Although there has been a lot of research carried out investigating the role of tree-like mechanisms (such as gene duplication and loss) in creating novel genes, little is understood about the role of non-tree-like mechanisms (such as gene remodelling) in creating novel genes. By performing analyses of sequence similarity networks (SSNs) and phylogenomics across 63 Metazoan genomes, we assess the contribution of two interrelated mechanisms of protein coding evolution to the diversity of animal protein coding gene families. Firstly, we use a novel phylogenetic approach to plot gene duplication, loss and point mutation in the evolution of gene families on the animal tree of life. Secondly, we use a network approach to study novel gene family evolution by gene remodeling. In this chapter, we focus on gene fusion/fission as the mechanism of gene remodeling. We show that gene remodeling is present right across animal life, and is a major source of novel protein-coding gene family genesis. In general, we see that the rate and specific mechanism involved in the generation of novel protein-coding gene varies significantly depending on the lineage. For example, the *Deuterostomia* as compared to the *Protostomia* have a much higher incidence of gene remodeling (specifically fusion). Bilaterian animals are either deuterostomes (in development first opening becomes the anus) or protostomes (first opening becomes the mouth). On a broader scale, this work provides insight into how novel protein-coding gene families have evolved through time and contributed to the diversity we see across the *Metazoa*.

Genome sequencing is revealing the existence of an enormous repertoire of protein coding genes in animal genomes (Consortium 1998, Adams et al. 2000, Holt et al. 2002, Consortium 2004). Recombinogenic processes and transcription-mediated read-thorough create remodeled genes that likely contribute novel protein coding genes to genomes (Zhou and Wang 2008, Kaessmann et al. 2009, Wu et al. 2013, Agaram et al. 2015). Indeed, given the diversity of protein domain combinations, it is reasonable to assume that protein remodeling has made a contribution to the whole-organism diversity observed in *Metazoa*. Well-understood and well-characterized examples of gene remodeling include *Jingwei*, a remodeled *Drosophila* gene derived 2 MYA from a fusion of a retrotransposed copy of an *Adh* locus and the 5' end of the *yande* gene.

The novel phenotype conferred by the resultant remodeled protein is a new specificity towards long-chain primary alcohols (Wang et al. 2000, Long et al. 2003). The *Kua-UEV* fusion gene in human is remodeled from two adjacent genes (*Kua* and *UBE2V1*)(Thomson et al. 2000) . The functional impact of this remodeling event is that the ubiquitin conjugating enzyme *UBE2V1,* which normally has activity localized solely to the nucleus, now has novel activity localized to the cytoplasm (Thomson et al. 2000). While we understand that mutational molecular clocks tend to tick with complex but increasingly well-understood rates (Lynch 2010), we have not yet been able to understand how, when, at what rate, and to what extent remodeling has impacted on animal proteomes. Until recently, we have not had available a comprehensive dataset of proteins to determine the details and genome wide impact of gene remodeling processes.

Animals exhibit significant diversity in development, morphology and indeed body plan (Section 1.4). We define major transitions as events that have allowed a lineage to radically change their environment, a biological function, and/or phenotype. From studies such as McLean et al. (2011) (on the lack of penile spines in Humans) and D'Apice et al. (2004) (on the cause of Progeria) we know that even small changes at the genetic level can cause major phenotypic effect (D'Apice et al. 2004, McLean et al. 2011). While phenotypic transitions in the Metazoa such as the emergence of the mesoderm, mineralized skeleton and chordate have been well documented (Bell 2015), the underlying genetic changes contributing to these major phenotypic transitions are generally quite poorly understood. Major questions in theoretical evolutionary biology that are addressed in this Chapter include: are these major phenotypic transitions fuelled by the emergence of novel protein coding gene families, and, has gene remodeling contributed to these novel families at a steady rate over time or in a punctuated manner across the fossil record.

3.2 Methods

3.2.1 Data acquisition

We retrieved our data from the OMA database (Altenhoff et al. 2014). The full details and references for the data used in this study are available in Appendix 3.1. We only used coding DNA sequences (CDS). All data was passed through our initial quality filter (Section 3.2.2). From the data that passed this filter, we took representatives across the *Metazoa* for each major phylum of the tree (Figure 3.1). Finally, form all the genomes that passed the quality filter, we selected 63 of these Metazoan species as representatives of all major groupings within the *Metazoa* (Figure 3.2). Some pre-computed Smith-Waterman alignments were available for download for ~50% of the species comparisons (Altenhoff et al. 2014) and we used these pre-computed alignments in our analysis.

3.2.2 Quality check filter

Data quality is of paramount importance in any analysis. We carefully researched all aspects of data quality. For example, commonly used statistics such as contig/scaffold N50 and fold coverage are not good measures of data quality as they are not always easily accessible and do not directly correlate with data quality (Bradnam et al. 2013). Therefore, we used the raw sequence data and from it extrapolated our own statistics/metrics and subsequently assigned data quality measures. This was challenging due to the sheer amount of storage needed for the data (data for a single species was in the region of many Gigabytes) but we had some working solutions in place. However, even more challenging was the acquisition of raw genome data. Although there are hundreds of animal genomes sequenced, the raw data is not always available. In addition to these challenges we discovered that our assumption that high sequencing quality would correlate to high protein coding data quality was simply not supported by the data – i.e. high-quality sequences can be poorly annotated. Therefore, we explored an alternative procedure for assessing data quality.

We developed the following procedure to provide us with the necessary metrics on data quality. The procedure involves two data filters - both of which are based on sets of protein coding genes that are present across: (1) all of life, and (2) all of Metazoa. The number of conserved protein coding genes present in a genome acts a proxy for the quality of that genome and its annotation. A genome with all (or at least 70% of -) the orthologs is deemed high quality and genomes with large portions of missing orthologs are lower quality (*i.e.* <70% of orthologs). The first filter uses a gene set of 412 orthologs that are found across all Metazoa (Powell et al. 2012). Using a reciprocal BLASTp (Altschul et al. 1990b), we identify the distribution of the 412 ortholog families in each genome.

Figure 3.1: Relationships and divergence times of the 63 Metazoan species sampled.

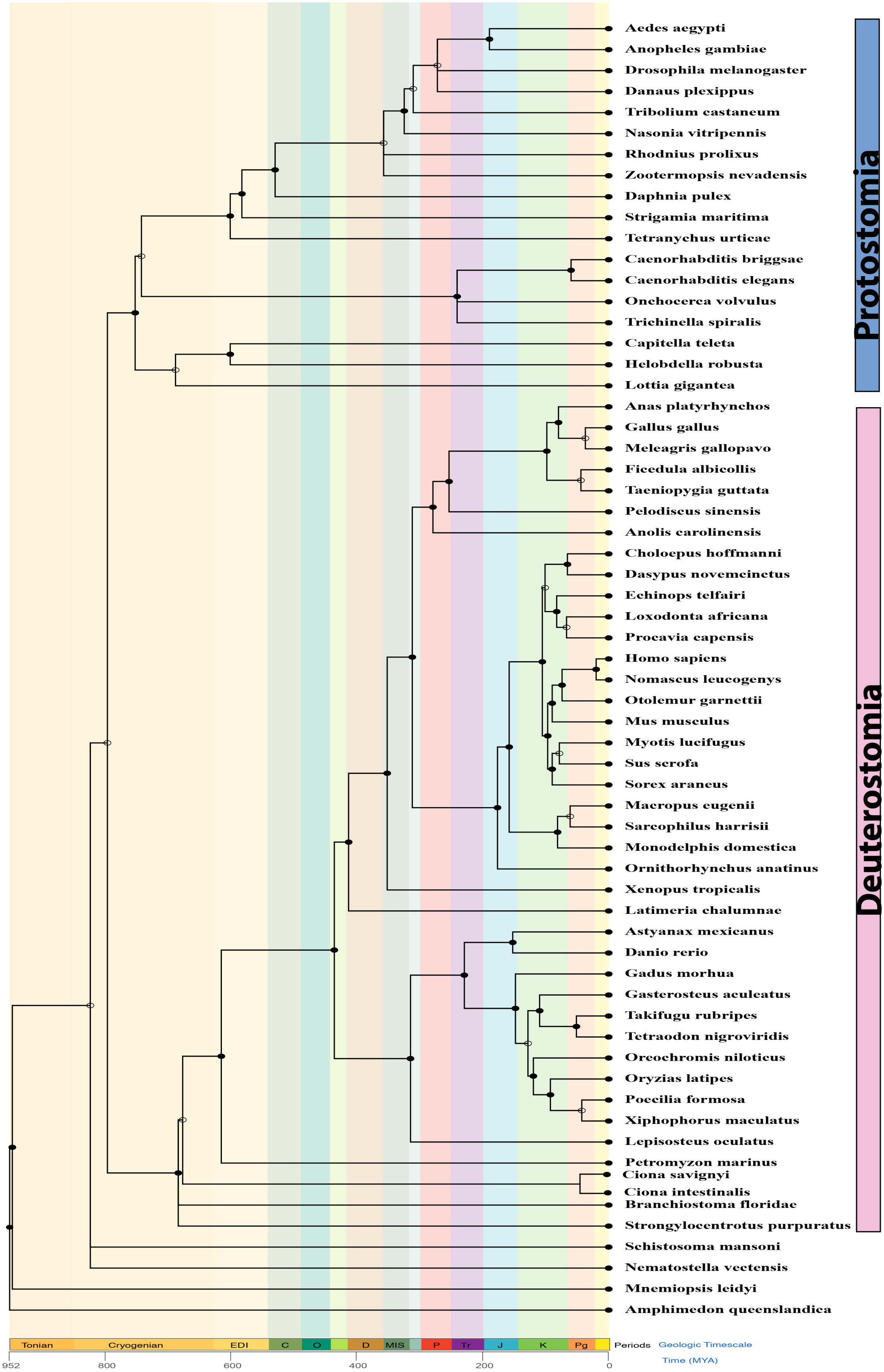


Figure 3.1: The phylogenetic relationship and dates of all 63 Metazoa represented in our dataset are shown as generated using TimeTree *(Kumar et al. 2017)*. Branches in the phylogeny are scaled according to divergence times. The geological periods from the Tonian period (~952 MYA) to present day are color-coded and are scaled in Millions of Years.

Figure 3.2 The distribution of our 63 genomes into their major taxonomic groupings across the Metazoa

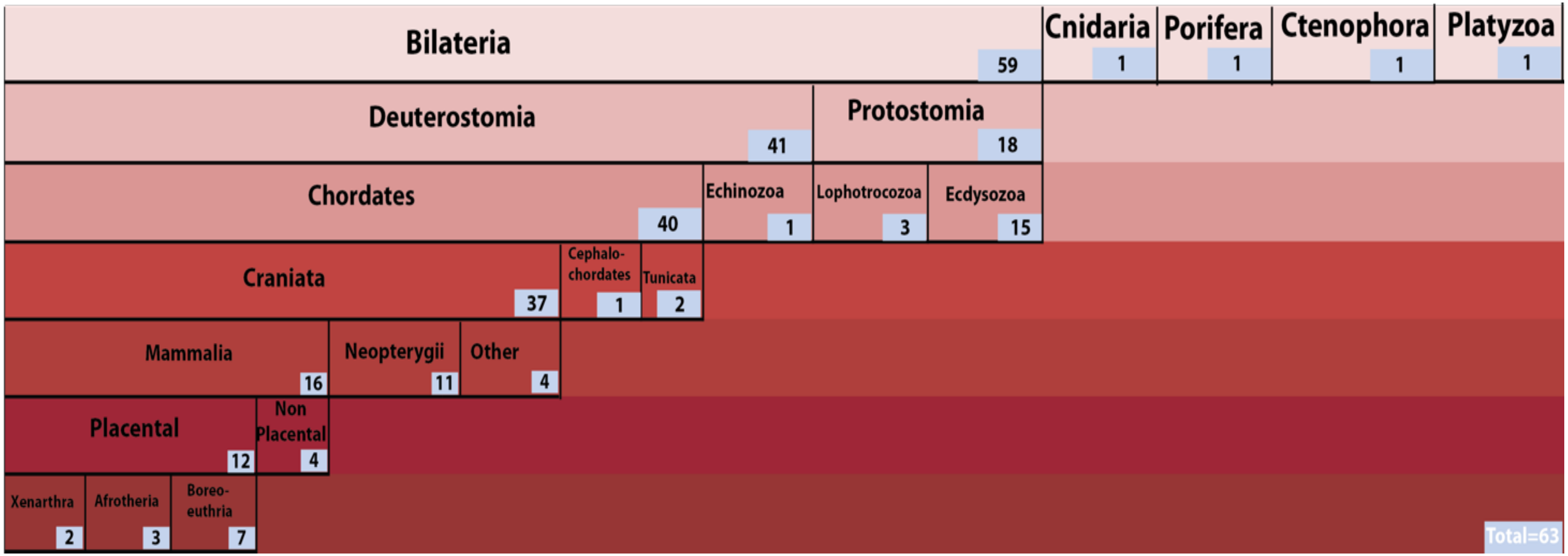


Figure 3.2: The distribution of the 63 Metazoa used in this study into their major taxonomic groupings. The numbers provided are post-filtering and are ordered by major taxonomic category. Each row represents a level of taxonomic grouping on the tree. Each subsequent row (from top to bottom) describes the more minor taxonmic groupings within the major taxonmoic groupings. For example, Bliateria is a major taxonomic grouping that can be further divded into either Deuterostomia or Protostomia.

The second filter is the stricter of the two filters. The second filter uses the 4 conserved orthologous families that are found in all of life as the quer homology search (Ciccarelli et al. 2006). As there are 412 genes in the first step, it is expected that there will be a reasonable level of variation in distrib terms of the number matches in each genome), this is useful in that we can t the genomes by quantity of genes present. However, this is not a strict filte random chance some genes be missed even in higher quality genomes. The 4 conserved genes are present in all of life and are highly conserved in their s as well as in their distribution. Therefore, we can query one of the 40 genes particular species to determine if it is present, and we can assess the conservation of the orthologous sequence. This combination of filters, carri this order, allowing the highest quality data possible to be retained wl accounting for variation in sequence and annotation quality.

3.2.3 Metazoan topology and dating

The second aspect of our data assembly that is crucial to the questions Chapters 3 and 4 of this thesis is sampling across the Metazoa. Sampling is g phylogeny - this is particularly challenging as there is no agreed species pl for the Metazoa (Section 1.2.7.7). The large number of alternative topolog therefore the large number of contentious yet critical nodes, increases the nu permutations necessary to represent the evolution of the animal kingdom. In some lineages of metazoan life are understudied and poorly sampled (e.g. *l* whilst others (e.g. the mammals as detailed in Chapter 2) are well studied and sampled. To this end, a dataset capable of addressing the major transitions i evolution was constructed with multiple representatives from before and a transition.

We used the topology and node dates from TimeTree (Kumar et al. 201 access on 23/09/2016). In total 51/63 of the species in our sampling were rep in TimeTree. For the remaining 12 species that were not present in the T database, we searched for their closest neighbours (sister taxa) in TimeTree. cases the time estimate was available for a member of the same genus. The c the nearest neighbour estimations can be found in Appendix 3.7.

3.2.4 Gene gain and loss analysis – OMA

The major frameworks for inferring orthology are either graph-based or tree-based. Graph-based methods use graph theory to create a network where genes/sequences are represented as nodes on the network connected by edges representing evolutionary relationships between nodes. Usually there are two parts to graph-based orthology inference. Initially an orthology network is created where nodes (genes) are connected by edges based on a statement of orthology. The OMA algorithm (Roth et al. 2008) we employ here to define gene birth and death, bases the edges/connections on Smith-Waterman alignments. Following on from this the orthologs are clustered into groups or gene families. In OMA the orthologous groups are Hierarchical Orthologous groups or HOGs (Altenhoff et al. 2014). HOGs are defined as groups of genes that have descended from a common ancestor in a given taxonomic range. Traditionally, gene/tree reconciliation is used to identify HOGs. However, OMA employs a graph-based method to identify HOGs based directly on the orthology graph it generates. The removal of traditional gene/species tree reconciliation from the inference process significantly reduces the computational cost. As well as this, OMA also reports several other advantages over standard bidirectional best hit approaches: it uses evolutionary distances instead of scores, considers distance inference uncertainty, includes many-to-many orthologous relations, and accounts for differential gene losses. (Roth et al. 2008). Using OMA v 1.1.2 (Altenhoff et al. 2014) and the default parameters (as discussed with the authors of OMA) we carried out our gene family evolution analyses. All available pre-computed data for our database was downloaded from OMA (http://omabrowser.org (Altenhoff et al. 2014)). We used the *familyanalyzer* python module (Altenhoff et al. 2014), made available from the authors, to analyse gene events across our topology according to the HOGs produced by OMA [famAnalyser_retrieveEventIds.py – Appendix 2.2].

3.2.5 Remodeled gene detection using CompositeSearch

By definition a remodeled gene is formed through a recombinogenic process such as gene fusion, where segments of the remodeled gene are derived from different gene families. Sequence similarity networks (SSN) where each node represents a unique sequence and each edge represents the similarity between sequences, appear to be well suited to identify and study this genetic mosaicism (Alvarez-Ponce et al. 2013a, Bapteste et al. 2013a).

I constructed a SSN using the results of an all-against-all BLASTp (Altschul et al. 1990a) of 1.2 million protein coding sequences (i.e. all genes within our database of 63 genomes). In this undirected network, two proteins are connected based on their similarity scores (E-value <= 1e-5, %Identity >= 30%). The SSN was made symmetrical by keeping only the best match of each pairwise comparison. We detected the remodeled genes and their families in this SSN using the software package CompositeSearch (*under review* Pathmanathan JS et al, 2017). The structure of the SSN captures much of the history of the evolution of the gene, such as divergence by point mutations and also recombinogenic events like fusions/fission events (Adai et al. 2004, Jachiet et al. 2014a). Typically, gene families form sub-graphs with high connectivity, in which connected sequences display significant BLAST E-values ≤ 1E-5, mutual covers ≥ 80%, %Identity ≥ 30%.

The results of CompositeSearch were parsed to retain only remodeled gene families with more than one remodeled gene and having no overlapping contributing sequences or components. We removed any singleton remodeled gene families, i.e. those with only a single member, as these were more potentially false positives. This removed 44,453 of the 53,456 remodeled gene families (~83% of total remodeled gene families were removed in this step). We also removed any remodeled gene families where only 1 member was remodeled. This removed 1,065 (~10%) of the leaving a total of 7938 remodeled gene families.

3.2.6 Remodeled Gene Family Classification using CompositeClassifier

Remodeled gene families were classified based on their origin and that of their components. This classification allowed us to identify if a remodeled gene family was formed by the fusion of pre-existing or entirely new protein coding gene families. All gene families were then placed on the reference tree and their last common ancestor was inferred using parsimony (Farris 1977). The type of remodeled gene family can be inferred by comparing its position on the phylogeny to that of its components [Appendix 3.5]. For example, a remodeled gene family is classified as old if its component gene families evolved before its emergence on the same path. The categories are as follows: Old refers to instances when components of a remodeled family are found at ancestral nodes only. Mixed refers to instances when components of a composite family are found ancestrally in the same lineage and also at the current node of comparison (a mix of old and new components). Complex refers to instances when components for a composite family are found in another path on the tree (not in a common ancestor). Contemporary refers to instances when a components of a composite gene family are found at the present node (all components arose at the same node on the tree as the composite). Subsequently remodeled refers to instances when components of a composite gene family are found at younger time points in the tree that the composite family (this is gene fission). Undefined are instances of families that cannot be categorized by these rules (Section 3.2.6). This approach aims to show the evolutionary combinatorial processes under which genes evolve (Figure 3.5(B))

3.2.7 Functional Enrichment Analysis

Using the stats.hypergeom function from the python SciPy package (Jones et al. 2014a) as incorporated into enrichmentDector3.py [Appendix 3.2], the genes at each node on the tree were assessed for enrichment of Gene ontology (GO) functions using a Bonferroni multiple testing correction (Weisstein 2004). Domains from Pfam (Finn et al. 2016) and their associated GO terms were retrieved from the gene ontology website (Consortium 2015). We represented each family defined in the CompositeSearch (Pathmanathan JS et al, 2017) analysis by its common Pfam domains and GO terms. Our criteria were that the GO term must be in > 50% of all genes in the family and a Pfam domain had to be ubiquitous within the family. For example, Family_A has 100 member genes. Qualitatively, all members have Domain_W and Domain_X, 4 members also have Domain_Y and 62 have Domain_Z. The first filter states that each ontology must be present in the majority (>50%) of the members to be included as a representative. This filter would exclude Domain_Y as it is not a majority. It has 4 associated Pfam domains, 12 have term_A, 100 have terms B, and C, and 70 have term D. As the criteria requires 100% of the genes to have a term only terms B and C pass the filter.

Figure 3.3: Phylogenetic tree of our metazoan sampling with internal nodes labelled

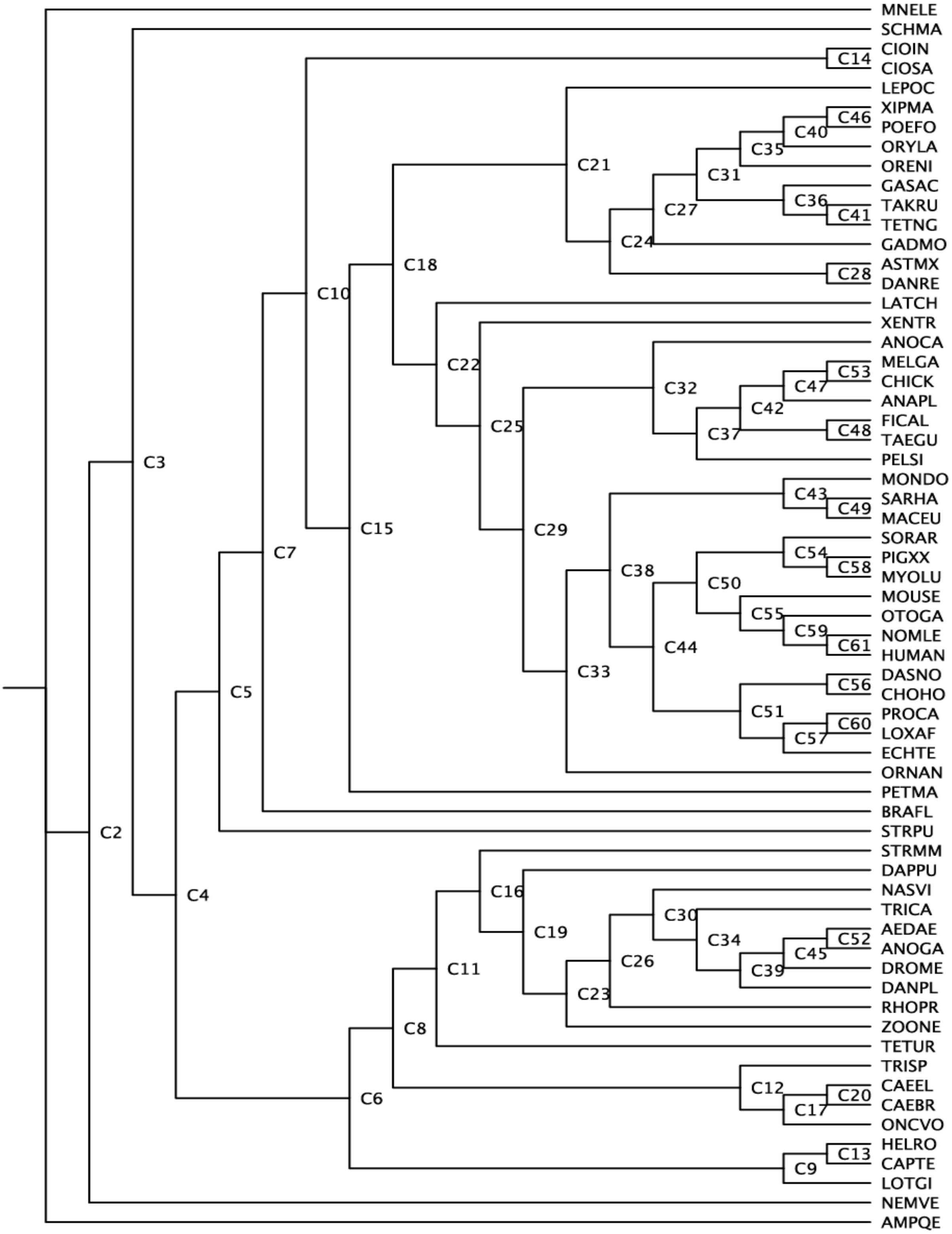


Figure 3.3: A phylogenetic tree (cladogram) of our sampling of the *Metazoa* with the internal nodes labeled for future reference when describing results.

The next filtering step works on the average of each Pfam domain per gene in the family. So, if all genes have 1 copy of Domain_W; 48 members have 2 copies of Domain_X, 40 members have 3 copies of Domain_X, and 12 have 1 copy of Domain_X; of the 62 members containing the passing Domain_Z 40 members contain 2 copies of Domain_Z and 22 only have 1 copy. The most common filtered domain sums per gene would be:

Domain_W: 1 copy
Domain_X: Average = 2.28 copies or 2 copies
Domain_Z: Average=1.65 copies or 2 copies

In essence, this removes gene families that show a high probability of being homoplastic and give the representative domains present in a gene family.

3.3 Results

3.3.1 Novel protein coding gene families emerge throughout the Metazoa and primarily by gene remodeling.

The orthology network created by OMA was based on Smith-Waterman alignments and subsequently identified orthologous families using a hierarchical clustering method (Altenhoff et al. 2014) (Section 3.2.4). The Hierarchical Orthologous groups (HOGs) produced by this method were defined as groups of genes that descended from a common ancestor in a given taxonomic range. These groups allowed us to identify where gene gain, loss or duplication arose in time. Novel gene families are those that had not been found prior to this point on the tree and what type of new gene family they are (*e.g.* remodeled or non-remodeled). The analysis of gene gain and loss identified 45,612 instances of novel genes at internal nodes on the animal tree. Of this cohort of novel genes 36,948 (81%) are remodeled (Figure 3.4). The majority of internal nodes (57/61) have more novel remodeled gene families than novel non-remodeled gene families. The average number of novel genes per node in the phylogeny is 760 and the median is 390 (Standard deviation = 1003). Most nodes that have above average number of novel gene families are major transitional nodes, including the following (*Clade* (total number of novel genes, % of novel genes that are remodeled at each node)): *Eumetazoa* (2267, 68%); *Bilateria* (3005, 65%); *Protostomia* (2179, 92%); *Euteleostomi* (3674, 73%); *Sarcopterygii* (1584, 85%), and

Neopterygii (3026, 89%). In the *Protostomia* there are a total of 2,179 novel genes and in the *Deuterostomia* there are 957. However, on average both *Protostomia* and *Deuterostomia* have the same number of novel non-remodeled genes per node (118 in both cases). The *Protostomia* have more novel remodeled genes (797) per internal node than *Deuterostomia* (493) (Table 3.1).

3.3.2 Gene remodeling is prevalent across the *Metazoa*, particularly at nodes of major phenotypic transition

Using a sequence similarity network (SSN) approach employed in CompositeSearch (Pathmanathan JS et al, 2017) we identified a total of 71,460 gene families in animal evolution. The analysis spans 63 Metazoan species representing all major groups of animals and 20,801 million cumulative years of animal evolution (Figure 3.1). On the SSN, remodeled gene families are represented as nodes that hold otherwise unconnected gene families together on the graph and we identify a total of 48,985 nodes with this feature (Figure 3.5). Using the canonical species phylogeny (Section 3.2.3) each of the 71,460 gene families were mapped to their node of origin. Each internal node (61 in total) in the phylogeny contained remodeled gene families and 49/61 of the internal nodes had more remodeled than non-remodeled gene families indicating that for the majority of internal nodes more novel gene families emerge by gene remodeling than other mechanisms.

Next, we wished to determine if the genesis of novel gene families by remodeling is distributed equally across the phylogeny or are there particular nodes that have a higher instance of novel gene family genesis by gene remodeling when compared to the other nodes in the tree. In particular we identified the internal nodes that contained the largest number of gene families (Figure 3.4): *Eumetazoa* (3913 families– 84% remodeled); *Bilateria* (8075 families– 87% remodeled); *Deuterostomia* (1019 families– 86% remodeled); *Vertebrata* (1500 families – 85% remodeled); *Euteleostomi (*7723 families- 84% remodeled); *Sarcopterygii* (1057 – 78% remodeled); and *Amniota* (2267– 75% remodeled). Each of these nodes represents a major transition in metazoan life history. In contrast, a large number of new gene families also emerge on two more recent nodes on the tree: (1) the ancestral node of *Caenorhabditis briggsae* and *C. elegans* has 4621 new gene families, 30% of which

Table 3.1: Gene counts for each node in the Metazoan tree from the OMA and CompositeSearch analyses.

Name	#genes	#duplicated	#lost	#novel/singleton(leaf)	#Comp Fams	#Genes in Comp Fams
SCHMA	11404	715	7428	7972	44	136
STRPU	26882	2896	9697	17563	520	2213
BRAFL	28464	3582	10079	18015	465	1688
CIOSA	13936	387	2724	6557	52	115
CIOIN	16500	459	1706	8077	55	130
C14	9867	1578	11156	1789	351	897
PETMA	10766	1188	14587	4340	53	130
XENTR	19291	2365	12098	4461	36	122
ORNAN	19730	1452	12789	5786	58	120
MACEU	15262	374	6289	1174	7	14
SARHA	19337	650	3741	2518	13	51
C46	20143	1548	2093	10	8	16
MONDO	16844	1284	7898	2458	26	231
C40	21394	3345	7240	144	30	78
CHOHO	12329	477	10551	1378	3	6
DASNO	23533	1905	2960	3991	43	130
C53	21218	2788	4656	12	14	28
ECHTE	16499	650	7780	2089	12	25
LOXAF	21050	1427	2677	2149	5	12
PROCA	16002	275	5806	1034	6	12
C58	20603	1282	1875	3	8	16
C54	21761	1403	3408	28	87	180
C47	24333	2936	4761	80	239	531
OTOGA	19514	926	2852	1030	2	4
HUMAN	30808	1314	1200	11464	72	192
NOMLE	18717	341	2631	1453	6	21
C61	19699	629	1613	160	86	232
C59	20773	677	1389	27	31	82
MOUSE	25679	1876	3332	5945	16	40
C55	21743	1193	4869	46	89	202
SORAR	13096	641	12154	1502	5	11
PIGXX	21452	1849	4653	3222	14	28
MYOLU	19862	1704	4771	1679	18	46
C60	21813	1212	2137	15	13	27
C56	23270	1251	3304	10	11	23
C48	25853	3399	3507	140	175	515
C41	27309	2633	2008	1181	1226	5246
C35	26558	1721	1601	1516	1205	5809
C31	25639	2128	3135	1019	495	2978
PELSI	18318	1001	5757	2710	27	62
CHICK	15504	246	1927	1394	11	59

MELGA	14627	105	2735	1408	13	26
C57	15887	441	1604	46	47	110
ANAPL	15753	177	3469	1899	13	39
C49	17212	650	2255	84	47	109
TAEGU	17104	849	2503	2440	33	72
FICAL	15383	179	2463	1124	5	10
C50	16608	739	3027	186	54	116
C42	19036	679	2328	347	142	414
C36	20650	937	2004	317	82	340
ANOCA	18029	1172	7119	2524	29	174
C32	21789	2717	7261	974	113	301
C27	26527	1166	684	1306	1696	10214
C24	25211	1091	1123	480	435	3482
LATCH	20358	2376	10990	4608	58	176
C21	25165	3008	3881	1584	825	7158
DANRE	27499	2379	3795	6085	69	353
ASTMX	23079	1162	4677	3488	23	48
C28	23595	5085	6072	259	143	316
ORENI	22257	1644	5579	2129	9	19
ORYLA	20499	683	6986	3245	30	110
XIPMA	20370	180	3469	928	2	4
POEFO	25163	1541	1817	3284	29	65
C51	22814	1068	1854	215	125	262
C43	23807	827	1308	79	43	108
C37	24559	1613	3091	145	101	337
GASAC	21773	1167	4645	2710	6	12
TETNG	20020	691	3155	2447	19	48
TAKRU	22942	585	3219	5448	29	64
C44	20345	801	3040	55	79	186
C38	22888	1329	4484	14	19	44
C33	26592	1711	861	392	299	1053
GADMO	20479	1424	9187	2741	48	113
C29	26050	2695	2655	819	404	1219
C25	26402	2130	823	1293	668	3398
LEPOC	18893	1326	9101	2383	14	29
C22	24764	5720	7132	3026	496	3206
C18	25622	3470	558	3674	6486	103008
C15	20330	3381	1151	768	1280	33573
C10	18274	1678	1457	534	802	13561
C7	18008	2138	1306	728	728	14607
C5	17107	5332	2340	959	874	11631
TRISP	15661	575	4319	11333	287	919
CAEBR	21610	1268	942	8020	132	376
CAEEL	20800	1838	674	6508	109	455
C19	13750	1311	991	5381	1350	5597
ONCVO	12948	447	2662	6806	73	237

C16	8535	1073	1382	953	404	1461
C11	8307	1645	8255	368	77	304
STRMM	14888	1166	8600	7353	182	549
DAPPU	30088	2143	7108	20308	843	2727
ZOONE	14336	866	6895	5987	46	216
RHOPR	15045	835	7535	7546	109	490
NASVI	16986	1080	6647	8899	370	1021
TRICA	14798	1037	5887	6555	93	242
DANPL	16232	727	4847	7971	103	333
DROME	14506	1349	3120	5430	59	157
ANOGA	12499	692	2202	3016	27	63
AEDAE	15129	1843	1600	4420	103	280
C52	11248	1056	1528	772	404	949
C45	11332	854	2196	388	315	976
C39	12660	1043	1459	131	84	284
C34	13408	872	1304	231	175	493
C30	14000	919	1289	260	194	628
C26	14513	694	839	276	184	570
C23	14689	989	2087	738	476	2087
C20	15457	1180	1132	451	162	759
C17	15432	1344	913	631	218	1258
TETUR	18019	1847	9326	11282	230	1255
C12	14846	1427	1655	454	124	1450
C8	15156	1816	4064	520	116	790
HELRO	23263	1638	6737	13900	389	1618
CAPTE	31325	2472	3070	17704	635	2473
C13	15099	2133	2633	414	78	233
LOTGI	23514	2332	5381	11331	290	1785
C9	16028	3844	5370	1421	134	402
C6	17525	5619	2769	2179	407	2455
C4	14426	1612	67	3005	7001	123510
C3	10402	638	205	308	612	15143
NEMVE	26036	3299	3195	17192	197	625
C2	9832	1920	34	2267	3301	92486
AMPQE	28464	3053	987	21286	979	4148
MNELE	16020	992	2650	11929	263	1020

Table 3.1: For each node in the tree (Col1) we have shown the counts for each node describing the following: 1) the number of genes present in the genome, 2) the number of gene duplication events, 3) the number of gene loss events, 4) the number of novel/singleton(leaf nodes), 5) the number of composite gene families emerging and 6) the number of composite genes emerging.

are the result of gene remodeling, and (2) the common ancestor of *Ciona savignyi* and *C. intestinalis* has 916 new gene families with 37% the result of gene remodeling.

The protein-coding elements that contribute to a remodeling event are known as components and can be of different ages or can themselves be the result of gene remodeling (Figure 3.5). To extract more detail on each case of gene remodeling detected we used CompositeClassifier from CompositeSearch (Pathmanathan JS et al, 2017) (Section 3.2.6). We categorised the components of every remodeled gene family based on their phylogenetic placement as: old, mixed, complex, undefined, contemporary and subsequently remodeled (Figure 3.5). In general, we see that most remodeling events on the tree are categorized as old. This means that most gene remodeling occur using only genetic material that is ancestral.

In general, the emergence of remodeled gene families is more prevalent within Deuterostomes than Protostomes (501 as compared to 288 remodeled gene families per internal node on average) (Figure 3.5). However, in Section 3.3.1 above we show that Deuterostomes have less novel remodeled genes than Protostomes indicating that Protostomes rely on gene remodeling as a mechanism to create novel genes more than Deutrostomes. The most prevalent category of remodeling in Metazoa is to reuse ancestral genetic protein coding elements (old category) with 50% and 51% of remodeling events in Protostomes and Deuterostomes respectively the result of old remodeling events (Figure 3.5). Therefore protein-coding gene families that are already established, or segments thereof, are used most often to create new gene families.

The large number of remodeled gene families predicted may be due to rapid turnover throughout the tree. We calculated the consistency index (CI) for remodeled and non-remodeled gene families (Kluge and Farris 1969) (where the maximum CI of 1 indicates that a family is gained/lost only once). Remodeled gene families have an average CI of 0.4 as compared to 0.7 for non-remodeled gene families suggesting that remodeled gene families are gained/lost more readily than non-remodeled gene families.

Figure 3.4: Proportion of remodeled and non-remodeled events in novel gene family genesis

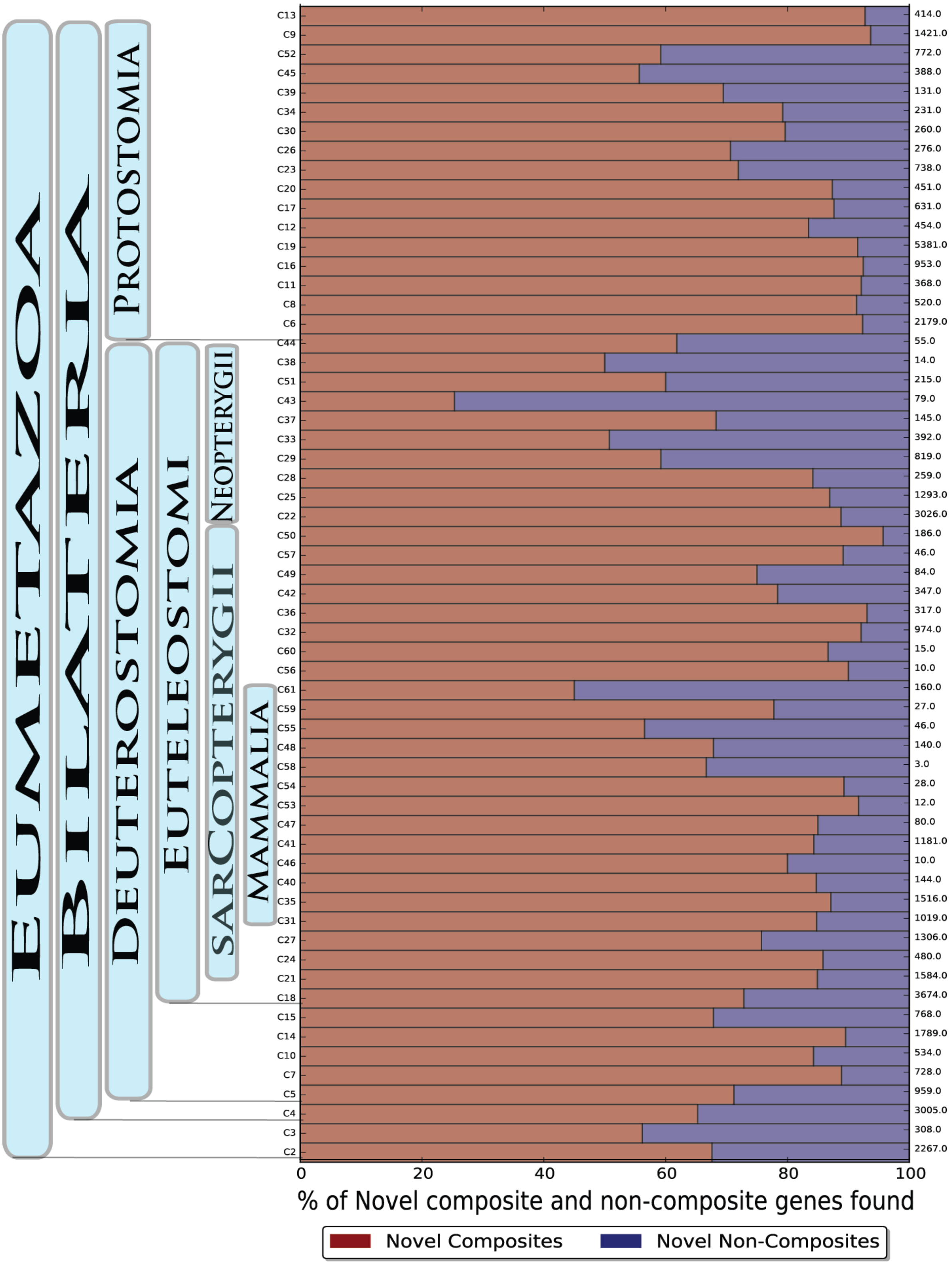
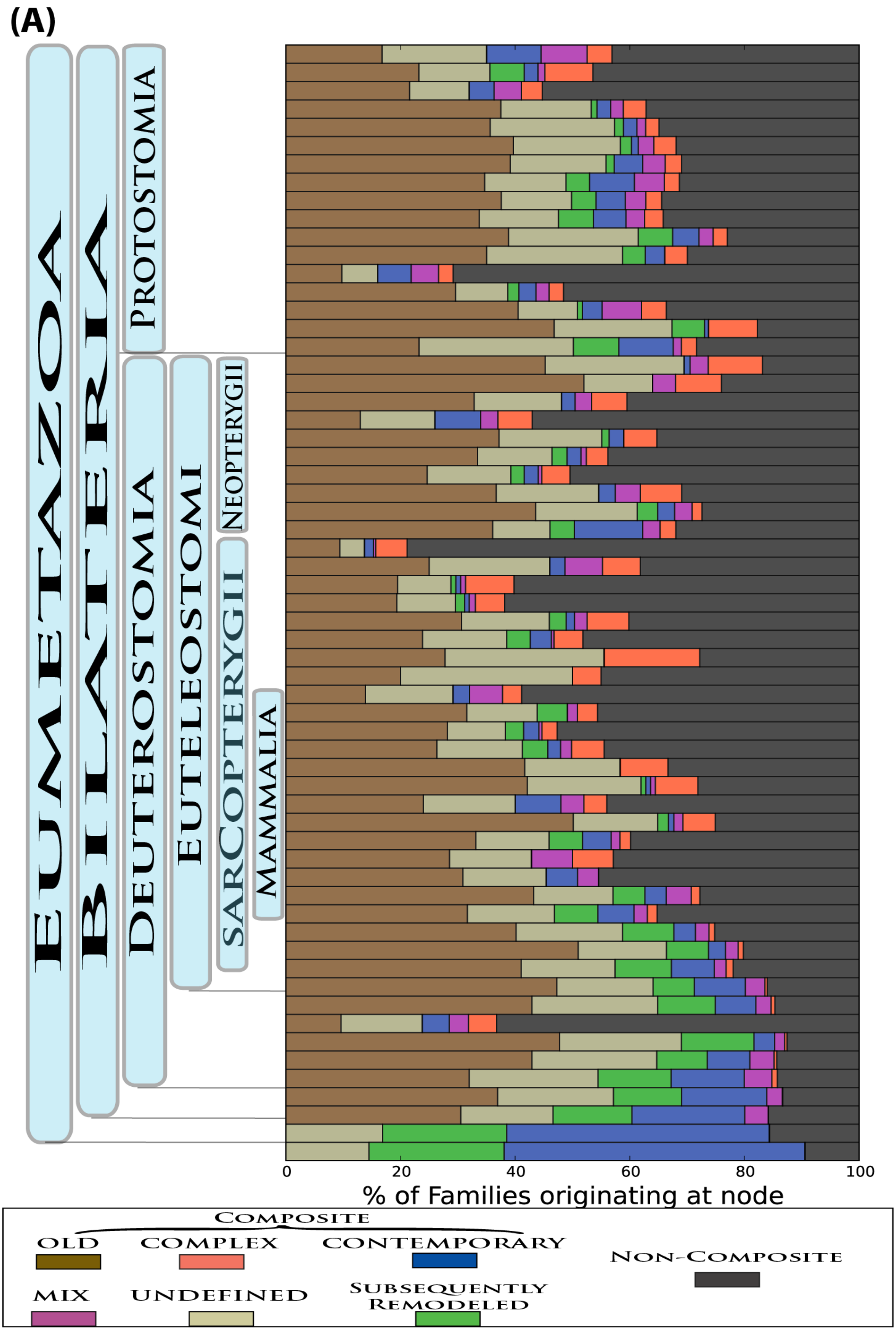


Figure 3.4: Each bar represents the proportion of novel genes that arose at each internal node on our tree (found in our OMA analysis) in each category: composite or

non-composite (determined from our CompositeSearch analysis) (each bar represents 100%). The number in black on the right Y-axis represents the number of novel genes that originate at this node in the tree. The red bar represents the proportion of novel genes that are composite and the blue bar represents the proportion of novel genes that are non-composite. The left Y-axis represent the label we have given to internal nodes of the tree (Figure 3.3).We have outlined the major taxonomic groupings.

Figure 3.5: Gene remodeling across the *Metazoa*



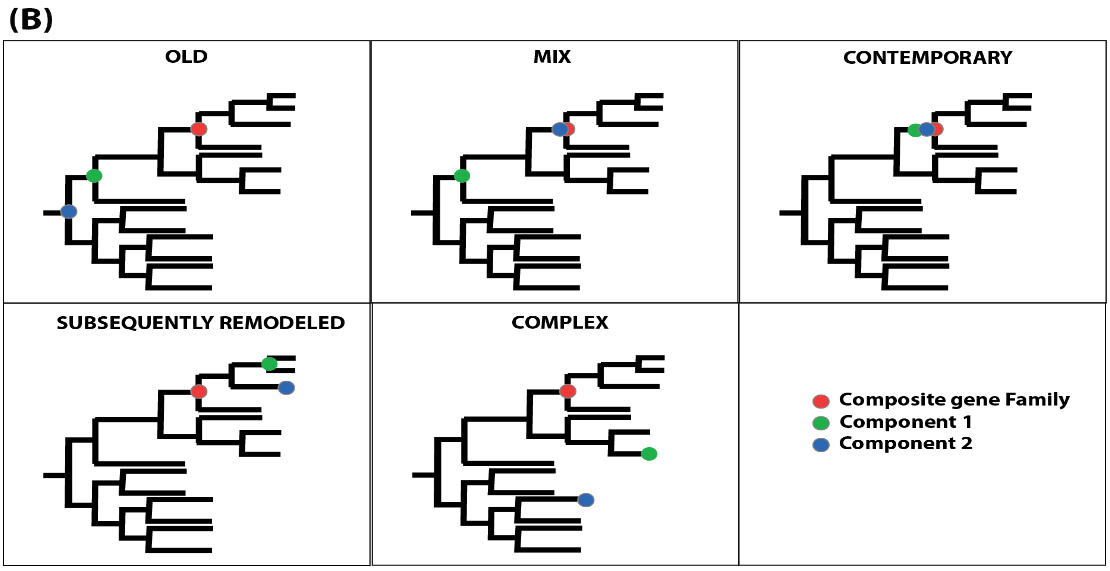


Figure 3.5: **(A)** Each bar represents the proportion of each category of family from the CompositeSearch analysis (each bar represents 100%) for each internal node of our tree. All colored bars are subcategory of composite gene families, black represents the proportion of gene families that are not composite. We have outlined major taxonomic groups. The node labelling system is illustrated in Figure 3.3. (**B**) We categories the components of every remodeled gene family based on their phylogenetic placement as: old, mixed, complex, undefined, contemporary and subsequently remodeled.

3.3.3 The rate of novel gene genesis across the *Metazoa* is not strictly clocklike

To determine the rate at which novel genes are emerging across the *Metazoa* we compared the rate of novel gene genesis for remodeled and non-remodeled genes. In general, we find that the rate at which novel gene families arise from gene remodeling is higher than the emergence of novel genes from other mechanisms (Figure 3.6). The average number of novel remodeled genes per node per million years (MY) is 13.0, and for novel non-remodeled genes it is 3.0. While there are some minor fluctuations (e.g. *Bilateria)* in the rate of generation of novel non-remodeled genes, the rates remain relatively similar across nodes (standard deviation = 5.7 from the mean). This is not the case for novel remodeled genes that have a comparatively high average standard deviation of 17.9 from the mean. Some major nodes in the animal phylogeny show a relatively high rate of emergence of novel gene genesis by gene remodeling, *Bilateria* (71.5 novel remodeled genes per MY); *Sarcopterygii* (60.2/MY); *Theria* (72.0/MY); *Protostomia* (46.2/MY), and *Ecdysozoa* (47.5/MY) are all examples of this.

Overall, novel remodeled genes have emerged at a faster rate than novel non-remodeled genes. But certain time points in metazoan evolution show higher than expected rates of emergence of novel gene families by both remodeling and non-remodeling mechanisms. One such node is the *Bilateria* node, at ~797 MYA (Kumar et al. 2017), arguably one of the most significant transitions in the *Metazoa* representing the origin of the third germ layer (the mesoderm) and increased morphological complexity (Martindale et al. 2002). The *Bilateria* node has on average 109 novel gene families emerge per MY. Another example of a high rate of novel gene family genesis is the origin of placental mammals (Crompton and Jenkins Jr 1979) (82 novel genes per MY).

3.3.4 Gene remodeling impacts the functional landscape at major phenotypic transitions in the *Metazoa*

The potential functional roles of the remodeled genes (at the level of domains) was assessed using Pfam domain data(Finn et al. 2016). For each internal node on the tree we established a list of significant functions gained at that time point (Section 3.2.7). Functional analysis of remodeled gene families at the *Euteleostomi*

Figure 3.6: The rate of novel gene genesis is not strictly gradual

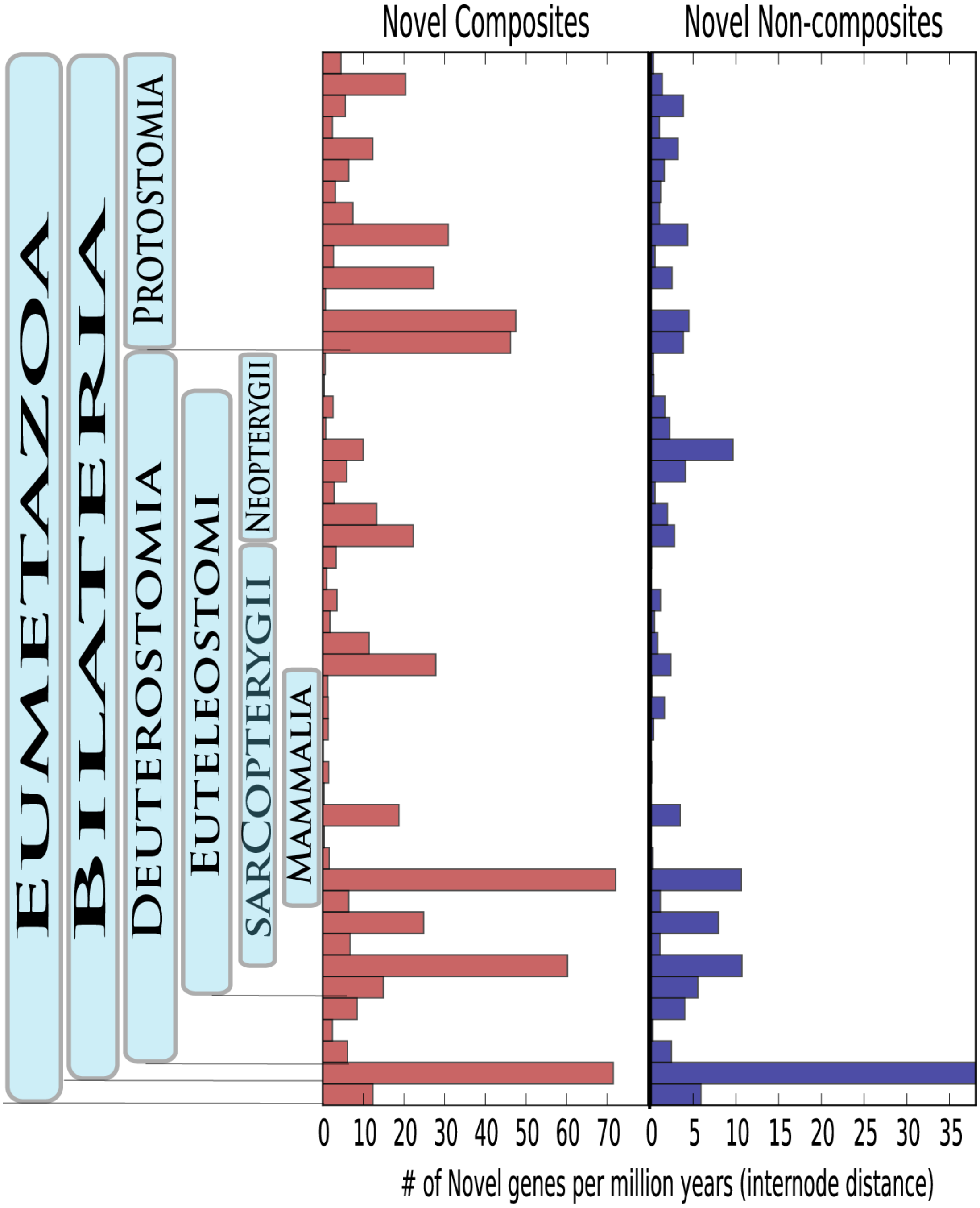


Figure 3.6: The bar charts represent the number of novel genes that originate at internal nodes divided by the internode distance(time) between the node and its closest ancestor. This gives the number of genes per unit of time for each node. Nodes that have a short internode distance (<10 million years) were not included on this as the short period of time skews the data. The Red bars represents the rate of novel composite genes and the blue bars represent the rate of novel non-composite genes. We have outlined the major taxonomic groupings

Table 3.2: Sample of Functional enrichment for novel remodelled genes found at some *Metazoa* transition nodes.

Enriched Gene	Corrected P-value	Tree Node
MHCII(Todd et al. 1988)	3.8e-05	*Euteleostomi*
RAG-2 involved in the initiation of V(D)J recombination during B and T cell development (Shinkai et al. 1992)	5e-06	*Euteleostomi*
Fibrinogen	(3.9e-07)	*Euteleostomi*
Ribosomal_protein_L44	4.2e-07	*Eumetazoa*
Ribosomal_protein_L21e	2.9e-09	*Eumetazoa*
Ribosomal_L27e_protein_family	2.0e-08	*Eumetazoa*
Ribosomal_protein_S17	3.5e-06	*Eumetazoa*
DHODH)(Fang et al. 2013),	3.2e-05	*Eumetazoa*
DHFR(Schnell et al. 2004),	7.3e-08	*Eumetazoa*
GPK(Wu et al. 2004)	3.2e-05	*Eumetazoa*
NDPK(Almgren et al. 2004)	5.1e-21	*Eumetazoa*
WNT	5.8e-05	*Deuterostomia*
Lipoxygenase	2.0e-05	*Deuterostomia*
Hydroxymethylglutaryl-coenzyme A reductase	3.8e-07	*Deuterostomia*
GDP dissociation inhibitor	8.3e-07	*Deuterostomia*
GrpE	1.5e-08	*Deuterostomia*
Peptidase M41	4.2e-05	*Deuterostomia*
MOSC	1.0e-05	*Deuterostomia*
GPI transamidase subunit PIG-U	3.8e-06	*Deuterostomia*
Cytochrome b	5.7e-06	*Chordata*
Cytochrome C and Quinol oxidase polypeptide I	5.7e-10	*Chordata*
V-ATPase subunit	5.7e-06	*Chordata*
ATP synthase protein 8	2.9e-07	*Chordata*
Glycosyltransferase_family_6	6.1e-05	*Chordata*
Tight Junction protein	4.4e-05	*Chordata*
Nuclear receptor coactivator	3.3e-06	*Chordata*

Table 3.2: The table shows examples of novel remodeled genes (enriched gene) that were found to be significantly enriched (Corrected p-value) for a particular function at particular nodes in the tree (Tree node column). All nodes shown in this example represent nodes on the animal tree where major phenotypic changes have occurred. For a full list of enrichment see Appendix 3.6.

ancestral node, identifies that many immune system related functions are introduced at this point(Table 3.2) and this node of course represents a major transition in the emergence of the adaptive immunity (Flajnik 2014). At the origin of the *Eumetazoa* novel gene families gained by gene remodeling have significant enrichment for ribosomal protein related functions and for enzyme functions related to cell proliferation(Table 3.2). The origin of the *Deuterostomia* has significant enrichment in functions related to cell signaling, development and metabolism (Jones et al. 2014b). The origin of *Chordata* shows significant gains in a number of key processes (Jones et al. 2014b) such as the remodeling of proteins involved metabolism and generating cellular energy and protein packaging and transport (Table 3.2). In summary, there are a plethora of significantly enriched functions at most internal nodes, with some nodes containing functions that correlate with a major phenotypic transition at that node (Appendix 3.6).

3.4 Discussion

This chapter gives an insight into the role of composite gene remodeling (gene fusions/gene fissions) in the evolution of novel protein coding genes across the *Metazoa*.

It has been established that modular proteins have an important role in the evolution of the *Metazoa.* For example, Patthy (2003) shows that a large proportion of proteins involved in the extracellular matrix of multicellular animals are a result of chimeric or gene fusions (Patthy 2003). However, it is generally believed that events to create a gene fusion/fission are rare (Jachiet et al. 2013). Fusion genes have been well documented in animals (Buljan et al. 2010, Marsh and Teichmann 2010). In humans, fusion genes are often linked with cancer (Soller et al. 2006, Soda et al. 2007, Lawson et al. 2011). However, it has not been fully established as to how this composite gene (fusion/fission gene) mechanism drives the evolution of novel proteins and phenotypes right across the *Metazoa*. We have shown that composite genes are indeed present in all major groups across the *Metazoa (*Figure 3.5*).* We have shown that they quantitatively form a major part of metazoan protein coding families. Furthermore, we have found that the majority of composite gene events occur using ancestral protein coding elements within the *Metazoa*. Until now, there has been no research into this aspect of composite gene formation.

In addition to this, we wanted to understand not only the prevalence of composite genes, but also how they impact the creation of novel proteins across the *Metazoa*. It has been established that fusion genes can indeed create novel proteins (Long 2000, Thomson et al. 2000). However, the extent to which this process creates novel proteins has not been documented. Our findings suggest that composite gene formation is a major mechanism for creating novel genes in the *Metazoa* (Figure 3.4*)*. We find that in the vast majority of our sample animal species, more than >50% of novel genes are created through gene remodeling events. This result gives an insight into the important role composite gene formation has in genetic innovation. However, there are examples of fusion genes making their parent genes redundant. If this occurred often, the number of non-composite novel genes that we find would be diminished as they would not be found in our search if they became functionally redundant.

After establishing that composite genes are prevalent across all major groups in the *Metazoa* and do have a major role in creating novel proteins, we wanted to gain an insight to the rate of composite gene formation through time in the evolutionary history of animals. There has been much debate on the rate of evolution. Two strongly supported hypothesis of evolutionary rate are phyletic gradualism and punctuated equilibrium (Gould 1972). Phyletic gradualism refers to slow, gradual changes that accumulate over time to create new species (within intermediate species present). Punctuated equilibrium argues that evolution occurs in bursts of evolution (bursts of high rate) that are tied to speciation events to create new species (Gould 1972). Our results indicate that the rate of composite gene evolution is not strictly clocklike (Figure 3.6). We show that novel composite genes have emerged at a faster rate than novel non-remodeled genes. Interestingly, we see that certain time points in metazoan evolution show higher than expected rates of emergence of novel gene families by both remodeling and non-remodeling mechanisms. For example, we found a high rate of novel gene genesis at the *Bilateria* node which represents a major transition in the *Metazoa* where the third germ layer (the mesoderm) was introduced, allowing for increased morphological complexity (Martindale et al. 2002).

In order to gain an insight into the functional importance of the novel composite genes we found across the *Metazoa*, we carried out a functional enrichment analysis. The wide distribution and abundance of composite genes in the *Metazoa* suggests that these genes are not restricted to a single functional pathway. Literature shows examples of very different pathways and functions being carried out by composite genes (Long 2000, Soller et al. 2006, Demichelis et al. 2007, Soda et al. 2007, Lawson et al. 2011, Agaram et al. 2015). Our functional analysis supports this. We show many composite genes that are enriched for functions and pathways at each node of the *Metazoa* such as immune system genes at the *Euteleostomi* node – a point in animal history where adaptive immunity originates (Appendix 3.6).

One possible reason for the higher level of apparent homoplasy that we found in the remodeled gene families (as compared to non-remodeled gene families) is the presence of epaktologs causing interpretation errors. Epaktalogs are multidomain gene families that share sequence similarity through the independent acquisition of the same domains rather than being homologous due to a common ancestry (Figure 3.7). The classical types of homologs that algorithms detect are orthologs (homologous genes derived from the same gene in a common ancestor), paralogs (homologous genes derived from a duplicate copy of the same gene) and pseudoparalogs (homologous genes in a genome where at one of the genes was transferred from another species). It is difficult to distinguish between epaktologs and paralogs. This can lead to interpretation errors, where epaktologs are treated as paralogs. In other words, trying to cluster a group of epaktologs as a family with a single point of origin on the tree is incorrect because the epaktologous genes are not directly related through descent. They are only related due to homology shared by containing the same domain (Figure 3.7) (Nagy et al. 2011).

Lastly, our approach relies on high quality data as annotation and sequencing errors can cause incorrect inferences. To diminish the impact of this we have used strict filtering parameters and high-quality genomes. This work can be built on as more high-quality genomes become available, particularly for non-vertebrates.

3.5 Conclusion

In summary, we have utilized novel data driven methods to assess the contribution of tree-like and non-tree-like mechanisms in the creation of novel protein coding elements across the *Metazoa* using 63 high quality genomes. We have illustrated that gene remodeling is prevalent across the entire *Metazoa* and has a significant contribution to novel gene genesis from protein coding elements. We have shown that the rate of novel gene genesis for remodeled genes is not clocklike and is higher than novel gene genesis of non-remodeled genes. Finally, we have given an insight into how gene remodeling may have had a significant impact in driving adaptive evolution at nodes of major phenotypic transition.

Figure 3.7: Epaktologs: genes with high sequence similarity(due to the independent acquisition of homologous domains) that are not related through common descent.

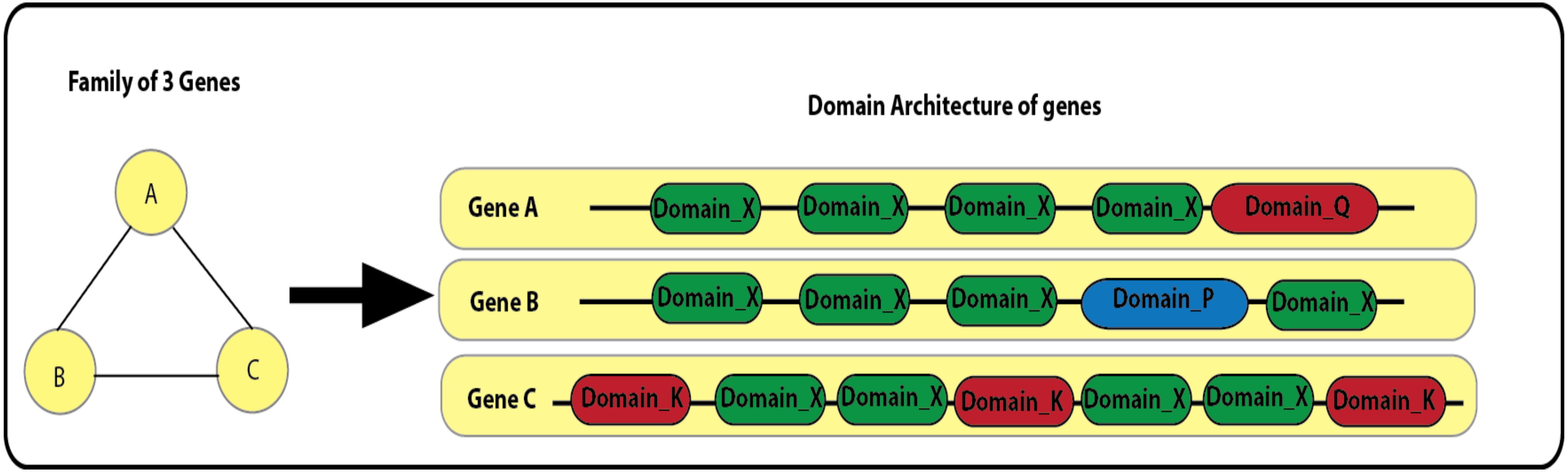


Figure 3.7: The figure illustrates epaktologous genes. Three genes (A, B and C) are identified as a family within the same species. Traditionally, these are classified as paralogs. However, in this scenario Gene A , Gene B and Gene C gained the Domain_X independently and are not related to each other by common descent – they are only related by the presence of the tandem Domain_X domains.

Chapter 4 – Establishing the domain co-occu characteristics of domain rearrangements metazoan protein coding genes.

4.1 Introduction

We will determine the extent of domain shuffling (a non-tree like process) on the evolution of proteins in Metazoa. The focus in this chapter is on the non-tree like evolutionary patterns and processes that can occur in the evolution of new protein coding genes and the focus is on functional domains in comparison to protein coding genes in Chapter 2 and 3. As we are interested in remodelling at the level of domains, BLAST (Altschul et al. 1990b) is not suitable to identify sequence similarity across such small units. In this case hidden markov models are necessary to find similarity across short sequences in a statistically robust way. We have selected the package called HMMER (Finn et al. 2011) to search the Pfam-A database of functional domains(Finn et al. 2016). Next the results of this analysis will be examined to identify the properties of domain shuffling within composite genes and non-composite genes (as defined in Chapter 3). This approach involves the use of sequence similarity networks (Jachiet et al. 2013) to trace domain movements across all animal genomes. Specifically, we create a bipartite network, where the nodes will represent domains/genes and edges will represent sequence homology between the domains and the respective genes. From this Bipartite network, we create Unipartite projections or co-occurrence networks. Essentially a network of domains that are found to co-occur on a modular protein/gene. From this we can identify the characteristics of domain sharing and shuffling across the animal kingdom.

Specifically, we address the following:

a) Do the network properties such as Degree distribution and Centrality of Pfam-A domains and protein coding metazoan genes follow traditional biological networks characteristics?

b) How do the domain-gene and domain co-occurrence networks differ between datasets composed of composite genes and non-composite genes respectively?

c) Are domains of composite genes more functionally permissive with respect to non-composite gene proteins?

d) Are there certain types of functional domains that have an affinity for each other and are often found together on a modular protein?

e) Are there certain types of functional domains that we never see together on a modular protein?

A working definition of a domain is a protein module of discrete structural folding that often represents the functional building blocks of a modular protein (Kaessmann et al. 2002). Many proteins are modular in nature, comprising of 2 or more domains (Moore et al. 2008). Chapter 3 established composite gene formation as a prevalent mechanism of protein coding gene genesis in the *Metazoa* with a particular focus on gene fusion. In this chapter, we apply network theory again but in this case to study the mechanism of domain shuffling. We use the sequence dataset of 63 Metazoa from Chapter 3 and their associated composite/non-composite gene status, along with Pfam-A domains (Finn et al. 2016) to address key features of domain shuffling in Metazoa.

4.2 Materials and Methods

4.2.1 Creation of Bipartite and Unipartite networks from Pfam-A domains

The profile hidden markov models (HMMs) of 16,712 PFam-A domain families were retrieved from Pfam v31.0 (Finn et al. 2016). Using the HMMscan function of HMMER v3.1b1(Finn et al. 2011) and an e-value cut-off of e^{-20} we identified sequence homology between the Pfam-A domains and the genes in the 63 Metazoa set (assembly of the Metazoan dataset is described in 3.2.1). The following filtering steps were carried out using PfamFilter.py [Appendix 4.1]. As certain domains can form part of a larger domain in the Pfam database we removed these nested or smaller domain from the sequence homology results. If multiple nested domains exist within a larger domain, only the largest domain with the lowest e-value was retained. In addition, if 80% or more of a domain overlapped with another domain on the same protein only the domain with the lowest e-value was kept. If multiple domains overlapped to the same region of a protein with 80% or more overlap, the domain with the lowest e-value from all overlapping domains was kept. This final filter minimises false positives due to multiple Pfam domain families (those with similar sequence motifs) aligning to the same position on the gene. As certain genes that contribute to composite genes are themselves remodelled (see results Section 3.3), we retained composite genes with a noOverLap score (a statistic calculated for each composite that gives an indication of how confident we are that it is a composite (see Pathmanathan JS et al, 2017 for more details) >0.30 for analysis here - a total of 96,053 composite genes. This created a dataset of composite genes in which we had high confidence (Appendix 4.4 - Top-Composite _0.3_Overlap.compositesinfo.ids).

After filtering, a global bipartite network was generated connecting the Pfam domains on one side to the genes on the other. Each individual connected component was then generated from this global network as a separate bipartite network (Appendix 4.2). Unipartite or co-occurrence networks were created from these bipartite networks by connecting Pfam domains that co-occur on the same gene (Appendix 4.2).

The relationships of the protein coding genes and the domains they contained were analysed according to the following 4 partitions of the data:

1) Entire Dataset: The network of Pfam-A domains and the entire set of protein coding genes from all 63 Metazoa
2) Composite Dataset: A network of domains and the composite genes from all 63 Metazoa genomes.
3) Non-composite Dataset: A network of domains and the non-composite genes from all 63 Metazoa genomes.
4) Highest Confidence Composite Dataset: (subset of Composite Dataset in (2) above) network of domains that hit a set of composite genes in which we have high confidence(described more above).

4.2.2 Domain co-occurrence network centrality

Three measures of centrality (as described in detail in Section 1.3.3.2). were calculated for each node in each domain co-occurrence network:

1) degree centrality
2) closeness centrality
3) betweenness centrality

The networks we generated were undirected so degree centrality was calculated as the number of edges a node had. Both closeness and betweenness centrality were normalised as follows using networkStats.py [Appendix 4.1]: closeness centrality was normalised by the remaining number of nodes in the network (n-1), and betweenness centrality was normalised by the maximum number of pairs of nodes (excluding the node of interest) ((2/(n-1)(n-2))).

4.2.3 Removal of highly-central nodes from domain co-occurrence network

This analysis was only carried out for co-occurrence networks with a major connected component (a subgraph with > 100 nodes). The 50 most central nodes were defined independently using degree, closeness and betweenness centrality measures. To assess the role these highly central nodes have in our networks, transitivity and average clustering were calculated before and after the removal of the 50 most highly central nodes. For comparison with the statistics from the removal of highly central nodes, a random sample and removal procedure was carried out 100 times on non-central nodes of the same network (NodeDeletion.py, Appendix 4.1).

4.2.4 Test for Assortativity

We used two measures of assortativity (Section 1.3.3.2) to determine if the co-occurrence networks generated are disassortative, neutral or assortative. Firstly, a linear regression was calculated for the average degree of nearest neighbours (<Knn>) for each node, and this was plotted against the degree (K) for each node. Secondly, the assortativity coefficient was calculated. Simulations were performed (using Assortativity.py [Appendix 4.1]) to generate random networks with the same degree distribution as the real data (simulated data networks are provided in Appendix 4.9). For any real data network with an assortativity coefficient was < 0.1, the assortativity coefficient of the simulated network and the real data are compared. If the simulated networks have a similar assortativity coefficient to the real data then the real data is deemed to have neutral or random assortativity. Note: This analysis was only performed on those co-occurrence networks with a major connected component i.e. > 100 nodes Therefore, the Highest Confidence Composite Dataset was not analysed in this way].

4.2.5 Community detection in domain co-occurrence network

Large connected components detected in the co-occurrence networks for Entire Dataset, Composite Dataset, Non-Composite Dataset and Highest Confidence Composite Dataset were analysed using the community detection algorithm NeMo (Rivera et al. 2010). NeMo is a plugin for Cytoscape v2.7 (Shannon et al. 2003) that employs a hierarchical protocol that detects communities of nodes within a network. Smaller co-occurrence networks were classified as being a community as they lacked connections to anywhere else on the global co-occurrence network.

4.2.6 GO enrichment in domain co-occurrence communities

Communities of domains that co-occur in the datasets were assessed for functional enrichment using Fisher's exact test to determine if certain communities are statistically enriched for a particular domain function/process. Gene ontology (GO) terms for domains were retrieved from the gene ontology website (Consortium 2015). Using recall and precision protocol (DomainEnrichment.py [Appendix 4.1]) we established which communities were linked to GO terms more often than expected. Recall evaluates the occurrence of a GO term with respect to the remainder of the network. Precision refers to the ratio between nodes in the network with a GO term and the number of nodes in the network not linked with the GO term.

4.2.7 Comparison of network degree distribution to a typical Scale-Free network

We compared and visualized the degree distribution of each real co-occurrence network, with a randomly generated network, and with a network generated with a typical scale-free property both of which had the same number of nodes as the real co-occurrence networks. The networkX functions gnp_random_network and scale_free_network were used to generate the simulated datasets (Degree_Graph_generator.py [Appendix 4.1]).

4.3 Results

4.3.1 Entire Dataset (all genes)

4.3.1.1 Bipartite Networks

The bipartite network of the Entire Dataset and Pfam domains had 3743 connected components. In this network, there was a single large connected component that accounted for 30% of all the Pfam domains and represented 65.4% of all homology connections. The next largest connected component in this network accounted for 0.004% of all the Pfam domains and represented 0.011% of all homology connections. In total 85.9% (3216) of the connected components only contained a single Pfam domain equating to exactly half (50.0%) of the Pfam domains in this network. These represent domains that are functionally exclusive *i.e.* they don't contribute to the creation of modular protein coding genes.

4.3.1.2 Unipartite Networks

To establish the relationship between the function of a domain and its presence in modular proteins we constructed unipartite projections. In total, we generated 527 unipartite projections of Pfam domains and genes from the Entire Dataset. We identified a single giant connected component comprising of 58.9% of all co-occurrence connections and 77.4% of the co-occurring Pfam domains in this network. The co-occurrence of over three quarters of all Pfam domains suggests that certain functional domains are promiscuous and co-occur in many different combinations. Excluding the giant connected component, the remaining 526 unipartite connected components have an average of 2.6 co-occurring Pfam domains indicating that many domains can co-occur but they are limited in terms of their combination with other domains.

4.3.1.3 The impact of central domains

Our domain co-occurrence network displays a scale-free topology (Figure 4.1) that is expected for biological networks (Barabási 2009). This means that there are many nodes with a low degree centrality and few nodes with a high degree centrality. In other words, the majority of domains are functionally restricted in terms of the observed combinations formed in protein coding sequences. It also means that there are central or hub domains in our network that contribute to a number of modular proteins, often referred to as promiscuous domains(Basu et al. 2009).

Measures such as degree, closeness and betweenness centrality can indicate domains that are central to the network and thus have the most influence in establishing modular proteins. This identified a number of domains that appear to be central to many modular proteins (Table 4.1). For example, *Pkinase* and *Pkinase_Tyr* were found to be most central by all measures of centrality with a degree of 144 and 88 respectively in the co-occurrence network. To place this in context, the next most central domain was *RVT_1* with a degree of 45. It has been found that kinase domains are important for many modular proteins in the human genome, 461 proteins in human genome are currently known to contain kinase domains (Manning et al. 2002). Other highly central domains that we have found include *RVT_1* mentioned above (a reverse transcriptase domain) (Gladyshev and Arkhipova 2011), *Trypsin* (a serine protease domain) (Rawlings and Barrett 1994) and *MAM* (an extracellular domain found in many receptor proteins) (Beckmann and Bork 1993). These have all been shown to have a biological role in many proteins.

The impact of removing highly central domains from the co-occurrence network were demonstrated using average clustering and transitivity (Figure 4.2). It is clear that removing central domains causes a significant decrease in degree, closeness and betweenness centrality, particularly for degree and betweenness centrality. This result indicates that there are far fewer co-occurrences of domains on the same modular protein. Conversely, removal of highly central domains from the co-occurrence network results in a significant increase in transitivity. Transitivity ranges from 0 to 1.0 where 0 indicates all triplets (3 domains connected by 2 or 3 edges) are open (3 domains with 2 edges) and 1 indicates all triplets are closed (3 domains with 3 edges - a clique). This indicates that as central domains are removed from the network there are more domains forming connected triplets.

Based on average clustering and transitivity, it appears that removing central domains in the co-occurrence networks is creating more closed triplets/cliques by removing domains that would usually form open triplets with two other domains. These results suggest that highly central domains have an important role in many modular proteins and connect other domains that would never usually be present in the same protein.

4.3.1.4 Assortativity

The degree of assortativity for the Entire Dataset network was calculated to determine the specific nature of domain combinations in the domain co-occurrence network. The degree assortativity indicates if domains in the network have a preference for connecting to other domains with a similar degree or not. The Entire Dataset network exhibits a positive degree assortativity coefficient of 0.009. However, it is apparent from the neighbour connectivity plot generated for the co-occurrence network that the linear regression trend line does not fit, indicating a pattern of dissasortative mixing in the co-occurrence network (Figure 4.3). These metrics indicate that the domains in our co-occurrence network are disassortative in nature, meaning that nodes (domains) of high degree have a preference to connect with other domains with a low degree~~.~~

Figure 4.1: Comparison of Degree distribution for domain co-occurrence network of (1) Entire Dataset (blue), (2) a random network (green), and (3) a typical scale-free network (red).

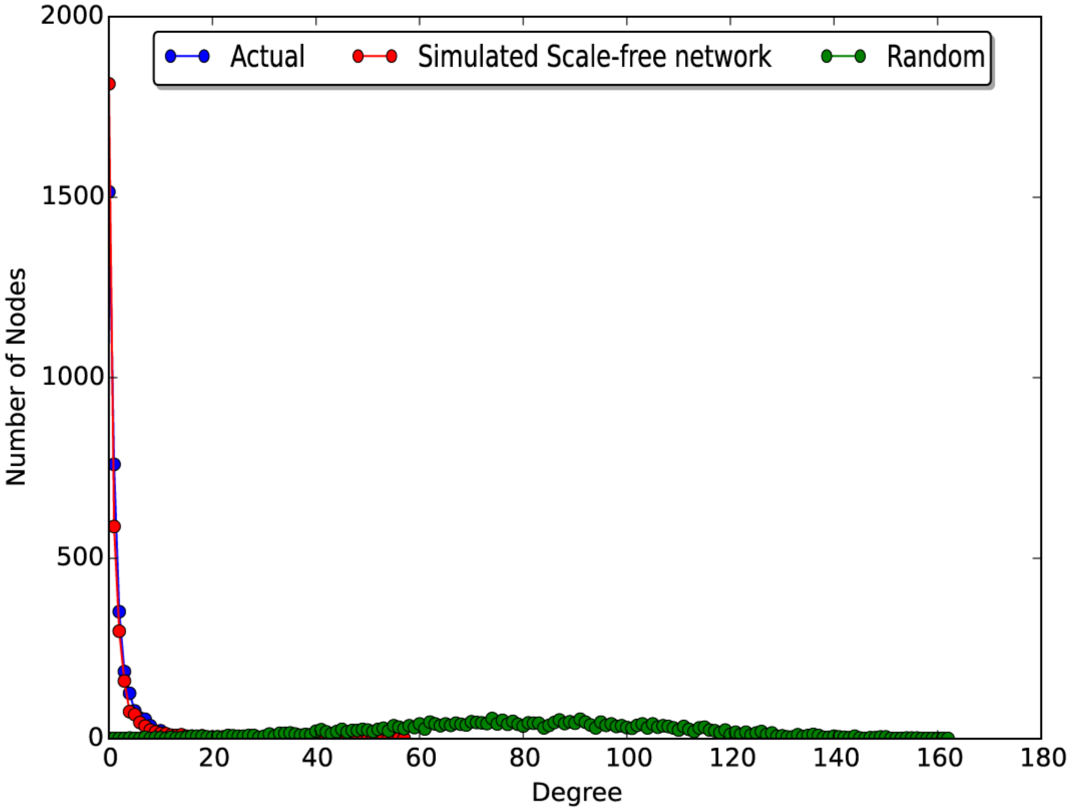


Figure 4.1: The degree distribution of Entire Dataset , a random network and a typical scale free network. On the X-axis is the degree, and the y-axis is the number of nodes. Blue represents the domain-occurrence network, Red is the simulated network with a scale-free property and in Green a randomly generated.

Figure 4.2: The impact of node deletion for highly central nodes measured by transitivity and average clustering

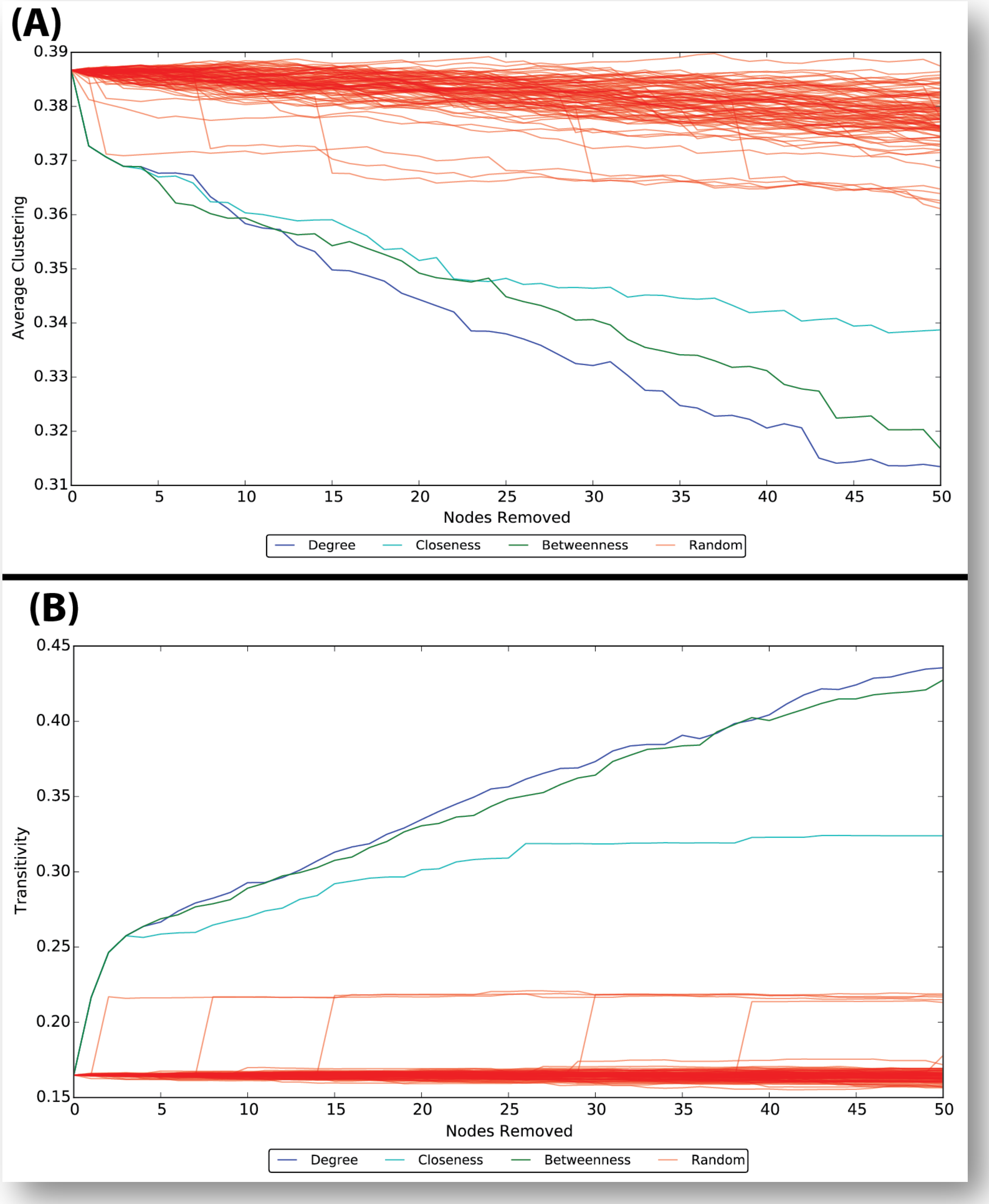


Figure 4.2: The impact of removing random nodes and the 50 most central nodes in the Entire Dataset network. The y-axis of the top panel (A) represents the average clustering after removing the corresponding number of nodes on the x-axis. The y-axis of the bottom panel (B) represents the transitivity after removing the corresponding number of nodes on the x-axis. Both graphs are indications of how average clustering is affected when certain domains are removed from the network. The top panel (A) is showing how removing central domains from the network is affecting average clustering. The bottom panel (B) is showing how removing central domains from the network is affecting the transitivity of the network (a measure of clique structure among connected triplets of domains). Domains were removed from the network by random (red) or according to a centrality value (degree (violet), closeness (blue), or betweenness (green)). For random removal of domains, we repeated this 100 times to give appropriate sampling.

Table 4.1: Domains removed from the co-occurrence network for the Entire Dataset to measure average clustering and transitivity.

	Degree Centrality		Closeness Centrality		Betweenness Centrality	
Index	Domain ID	Degree	Domain ID	Closeness	Domain ID	Betweenness
1	Pkinase	144	Pkinase	0.196	Pkinase	0.158
2	Pkinase_Tyr	83	Pkinase_Tyr	0.181	Pkinase_Tyr	0.057
3	RVT_1	45	RVT_1	0.174	RVT_1	0.042
4	Trypsin	35	MAM	0.166	Trypsin	0.026
5	VWA	32	C2	0.164	RhoGEF	0.021
6	7tm_1	32	Kringle	0.164	ABC_tran	0.02
7	RabGAP-TBC	31	F5_F8_type_C	0.163	RabGAP-TBC	0.019
8	ABC_tran	30	RhoGEF	0.162	Helicase_C	0.018
9	UCH	30	Kinesin	0.162	Myosin_head	0.018
10	RhoGEF	30	Myosin_head	0.162	7tm_1	0.017
11	F5_F8_type_C	29	SRCR	0.161	Kinesin	0.016
12	SRCR	28	SH2	0.161	Y_phosphatase	0.016
13	DEAD	28	RabGAP-TBC	0.159	p450	0.015
14	Ras	27	Peptidase_C1	0.159	Peptidase_C1	0.014
15	Ion_trans	26	7tm_1	0.157	UCH	0.014
16	Helicase_C	26	Helicase_C	0.157	adh_short_C2	0.014
17	MAM	25	Guanylate_cyc	0.157	Ras	0.014
18	BTB	25	ANF_receptor	0.157	GTP_EFTU	0.013
19	C2	25	Ank_2	0.156	VWA	0.013
20	RhoGAP	25	UCH	0.156	C2	0.012
21	Y_phosphatase	25	I-set	0.156	F5_F8_type_C	0.012
22	GTP_EFTU	24	ABC_tran	0.156	SRCR	0.012
23	SNF2_N	23	Histone	0.156	PK	0.012
24	Kinesin	23	Roc	0.156	MFS_1	0.012
25	VWD	22	ASC	0.155	Ion_trans	0.011
26	CUB	22	Trypsin	0.155	MAM	0.011
27	Pro_isomerase	21	SH3-RhoG_link	0.155	Homeobox	0.011
28	Myosin_head	21	AAA_12	0.155	RhoGAP	0.01

29	FERM_M	20	Sema	0.155	Bromodomain	0.009
30	Fibrinogen_C	20	RBD	0.155	SH2	0.008
31	MFS_1	20	Inhibitor_Mig-6	0.154	DEAD	0.008
32	CH	19	AAA_11	0.154	AAA	0.008
33	PI3_PI4_kinase	19	SAM_1	0.154	Guanylate_cyc	0.008
34	7tm_2	19	RGS	0.154	Peptidase_S9	0.008
35	AAA	19	Guanylate_kin	0.154	ThiF	0.008
36	CPSase_L_D2	18	COR	0.154	Cnd1	0.007
37	Guanylate_cyc	18	F_actin_bind	0.153	BTB	0.007
38	Peptidase_C1	18	Dynamin_N	0.153	Pro_isomerase	0.007
39	Kringle	18	Y_phosphatase	0.153	adh_short	0.007
40	Bromodomain	17	PTEN_C2	0.153	Ald_Xan_dh_C2	0.007
41	adh_short	17	cNMP_binding	0.153	CH	0.007
42	p450	17	CNH	0.153	Kringle	0.007
43	Adaptin_N	16	Macro	0.153	Fibrinogen_C	0.006
44	AMP-binding	16	Death	0.153	Adaptin_N	0.006
45	Methyltransf_FA	16	EphA2_TM	0.153	MIT	0.006
46	adh_short_C2	16	Ephrin_lbd	0.153	Sugar_tr	0.006
47	Zona_pellucida	15	Recep_L_domain	0.153	ANF_receptor	0.006
48	SH2	15	Furin-like	0.153	Spc97_Spc98	0.006
49	Aldedh	15	SAM_2	0.153	COesterase	0.006
50	Gal_Lectin	15	OSR1_C	0.153	SNF2_N	0.006

Table 4.1: The Pfam ID and respective measure of centrality for the top 50 nodes in the domain co-occurrence network that were removed according to Degree centrality (Merged Column 1), Closeness centrality (Merged Column 2) and Betweenness centrality (Merged Column 3).

4.3.1.5 Functional domain combinations and community detection

We detected 199 communities of co-occurring domains within our largest co-occurrence network. To the exclusion of the giant connected component the remaining components were treated as a community as they were small in size and not connected elsewhere on the network. We wanted to determine if domains that function similarly are more likely to co-occur. The enrichment analysis (see Section 4.2.6) of each community in the co-occurrence network revealed a number of significantly enriched communities of domains for a particular process/function. However, we need to account for instances where GO terms are only associated with a single domain and not anywhere else in the network. We only assigned significant enrichment if the P-value was <0.05 and precision measurement >=0.5 (the GO term must be found in a least half the network). This revealed 148 significantly enriched domains in the co-occurrence network [Appendix 4.5], the top 20 of which are shown in Table 4.2. Communities that have significant enrichment and a high recall (scale from 0 to 1.0) are enriched for a function that is rarely found elsewhere in the network. In turn, this implies that the co-occurrence of the domains in that community are rarely found in an alternative combination. Communities that have significant enrichment and a high precision (scale from 0 to 1.0) have more domains enriched for a specific function, indicating that most of the specific functional domains within the co-occurrence community are needed for the enriched function. In general, communities that have significant enrichment, a high precision, and a high recall are significantly enriched for unique functions that require the majority of the domain co-occurrence connections.

4.3.2 Comparison of network properties of composite and non-composite gene remodelling by domain shuffling

To determine if the properties of domain shuffling differ between composite and non-composite genes, we have compared our findings for the Entire Dataset to three more datasets: Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset (Section 4.2.1).

Figure 4.3: Assortativity test for Entire Dataset domain co-occurrence network

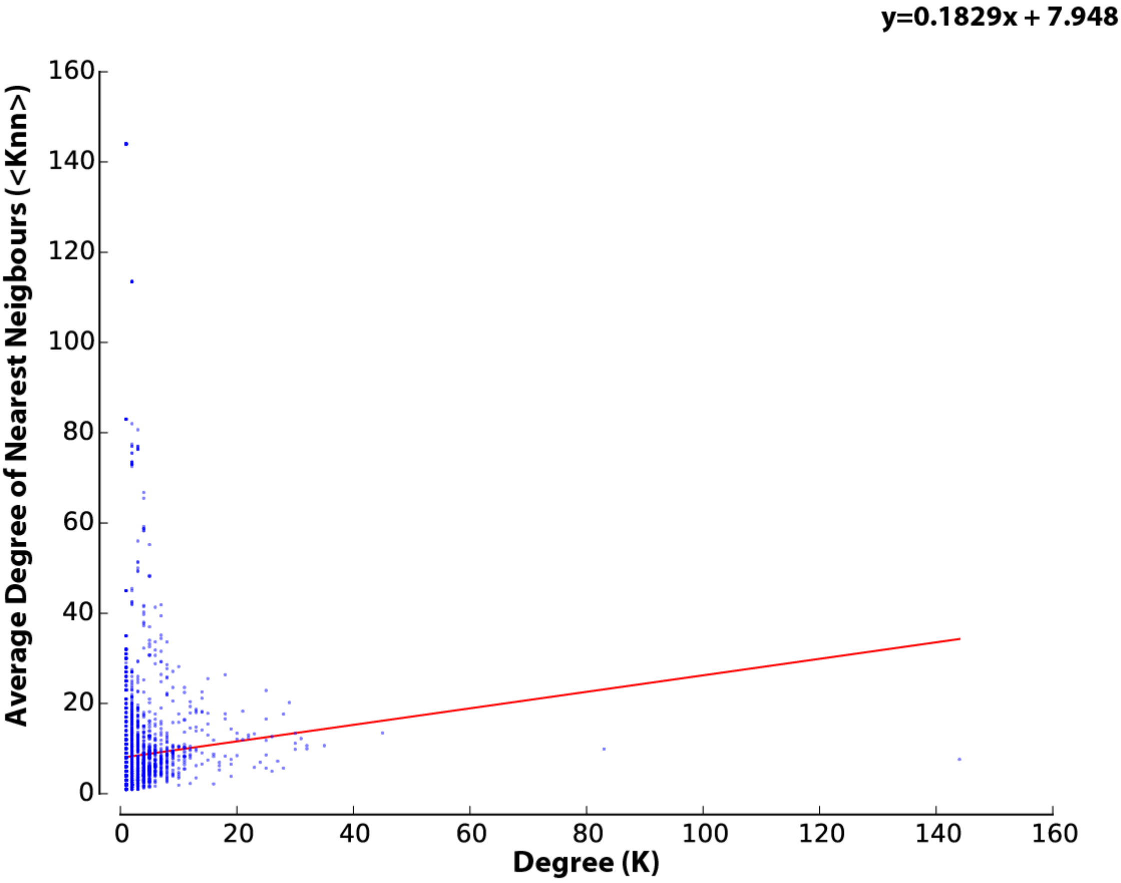


Figure 4.3: The neighbour connectivity plot showing no strong mixing for the Entire Dataset co-occurrence network. K (X-axis) refers to the degree of the given node. <Knn> (Y-axis) refers to the average degree of the nearest neighbours for that given node.

Table 4.2: Domain community enrichment of GO terms – top 20 significant scores

GroupID	TotalMembers	GO_Terms	Recall	Precision	Fisher's Exact p-value
742	3	mannose metabolism	1.00	1.00	1.70E-10
742	3	alpha-mannosidase activity	1.00	1.00	1.70E-10
964	3	positive regulation of transcription, DNA-templated	1.00	1.00	1.70E-10
1403	3	protein-arginine deiminase activity	1.00	1.00	1.70E-10
1788	3	arginyl-tRNA aminoacylation	1.00	1.00	1.70E-10
1788	3	arginine-tRNA ligase activity	1.00	1.00	1.70E-10
2028	3	S-adenosylmethionine biosynthetic process	1.00	1.00	1.70E-10
2028	3	methionine adenosyltransferase activity	1.00	1.00	1.70E-10
40	2	inflammatory response	1.00	1.00	1.86E-07
105	2	translational termination	1.00	1.00	1.86E-07
155	2	glycerone kinase activity	1.00	1.00	1.86E-07
155	2	glycerol metabolism	1.00	1.00	1.86E-07
267	2	cytoskeleton organization	1.00	1.00	1.86E-07
659	2	peptide cross-linking	1.00	1.00	1.86E-07

804	2	phosphoenolpyruvate carboxykinase activity	1.00	1.00	1.86E-07
1178	2	superoxide dismutase activity	1.00	1.00	1.86E-07
1231	2	MHC class II protein binding	1.00	1.00	1.86E-07
1314	2	RNA polyadenylation	1.00	1.00	1.86E-07
1490	2	chorion	1.00	1.00	1.86E-07
1490	2	chorion-containing eggshell formation	1.00	1.00	1.86E-07

Table 4.2: The 20 most highly significant hits from the analysis of GO term functional enrichments within communities of co-occurring domains. Ranked according to the highest precision (column 5) and recall (column 4), followed by lowest p-value (column 6). Group ID (column 1) is an id given to a connected component in the network. Total Members (column 2) refers to the total number of domains in the connected component.

4.3.2.1 Bipartite Networks

The Bipartite networks for the Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset are described in Table 4.3. Both the Composite Dataset and Non-composite Dataset had a giant connected component; this was not a feature of the Highest Confidence Composite Dataset. For context, the next largest connected component in the Composite Dataset and Non-composite Dataset accounted for 0.1% and 0.3% of all our Pfam domains respectively and 0.03% and 1.8% of all homology connections respectively. In all three networks the majority of connected components had only a single domain. The domains are functionally exclusive or specialized domains. In other words, they do not contribute to the creation of modular proteins with other Pfam domains.

4.3.2.2 Unipartite Networks

To establish the relationship between co-occurring domains on the same modular protein for Composite Dataset, Non-composite Dataset, and Highest Confidence Composite Dataset, we made Unipartite projections of the Bipartite networks for Pfam domains (see Section 4.2.1). In these subgraphs, we look for Pfam domains that are connected and see how they may co-occur on the same gene. We determine what domains seem to work together in modular proteins/genes - for summary of results see Table 4.4.

The co-occurrence of over a 40% of all Pfam domains in a single component of the Composite Dataset and Non-composite Dataset suggests that certain functional domains in these networks are able to co-occur in many different combinations. They act like a glue that holds the connected components together. The low average number of Pfam domains co-occurring in the majority of connected components for all 3 datasets indicates that these domains can co-occur but are limited in their combinations.

4.3.2.3 The impact of central domains

Our Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset domain co-occurrence networks all exhibit a scale free topology that is expected for biological networks (Barabási 2009) (Figure 4.4). This means that there are many nodes with a low degree centrality and relatively few nodes with a high degree centrality. Biologically, this means that majority of domains are functionally restricted in the combinations they make to other functional domains. It also means that there are central or hub domains in our network that contribute to a number of modular proteins (Barabasi and Oltvai 2004). There

Table 4.3: Summary of Bipartite network characteristics for Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset

Dataset	# Connected components	% of all Pfam domains found within the largest connected component	% of all homology connections found within the largest connected component	% of connected components with 1 Pfam domain	% of Pfam domains that don't co-occur with other domains
Composite Dataset	2438	15.4	44.2	80.5	49.3
Non-composite Dataset	3969	24.6	45.9	86.2	53.5
Highest Confidence Composite Dataset	2313	0.22	0.012	88.1	75.6

Table 4.3: The following characteristics are detailed for the Bipartite network for Datasets: the dataset is given in the leftmost column. Column 2 shows the number of connected components in their respective dataset's Bipartite network. Column 3 shows the percentage of all Pfam (Finn et al. 2016) domains that are found within the largest connect component in the network. Column 4 shows the percentage of all network connections that are found within the largest connected component. Column 5 shows the percentage of connected components that have only a single domain. Column 6 shows the % of all Pfam domains that do not co-occur with other domains in the Pfam database.

Table 4.4: Comparison of Unipartite (co-occurring domains) network characteristics for Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset

Dataset	# Connected components	% of all co-occurring Pfam domains found within the largest connected component	% of all co-occurrence connections found within the largest connected component	Average # of domains in all other co-occurrence connected components
Composite Dataset	473	45.4	30.4	3
Non-composite Dataset	546	71.4	53	3
Highest Confidence Composite Dataset	275	0.009	0.03	2

Table 4.4: The table describes the characteristics of the Unipartite network for the Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset. The dataset is given in the leftmost column. Column 2 shows the number of connected components in their respective dataset's Unipartite network. Column 3 shows the percentage of all Pfam (Finn et al. 2016) domains that are found within the largest connect component in that network. Column 4 shows the percentage of all network connections that are found within the largest connected component. Column 5 shows the average number of domains found in all other components from the Unipartite network (all those connected components except the largest connected component).

Figure 4.4: Degree distribution comparison for domain co-occurrence network for Composite Dataset, Non-composite Dataset, Highest Confidence Composite Dataset, a random network (green), and (3) a typical scale-free network (red).

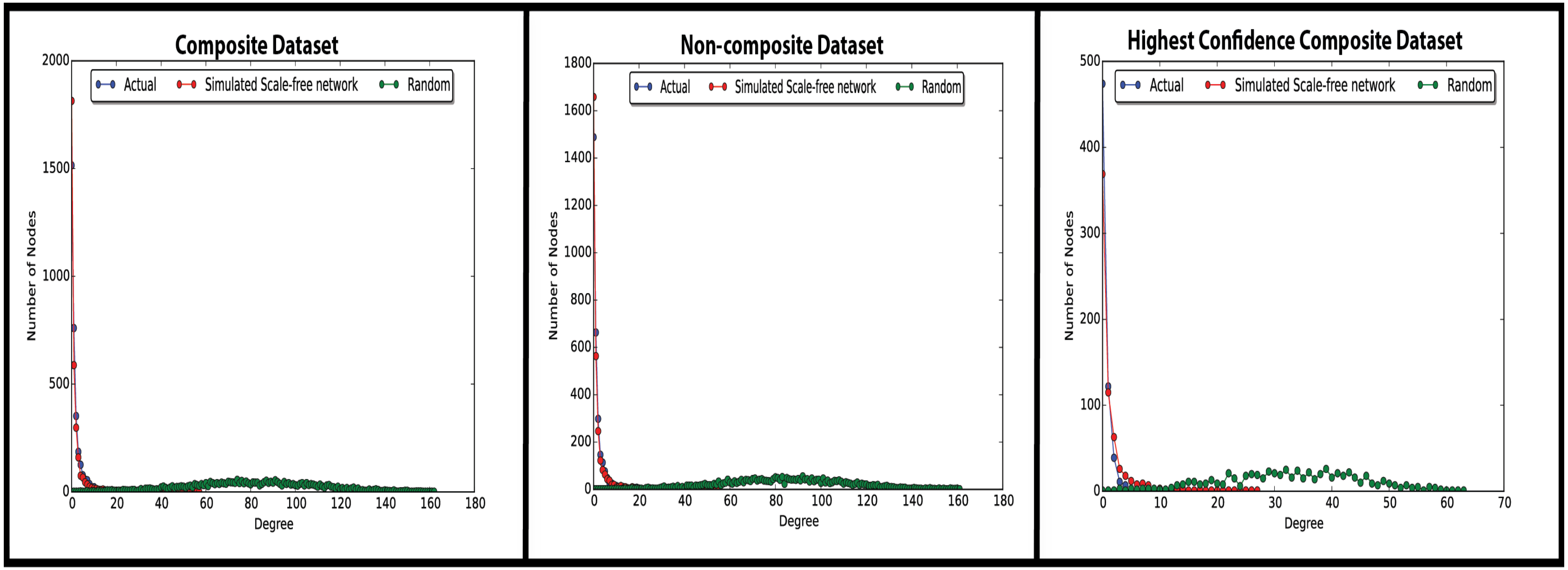


Figure 4.4: The degree distribution of Composite Dataset, Non-composite Dataset, Highest Confidence Composite Dataset, a random network and a typical scale free network. On the X-axis is the degree, and the y-axis is the number of nodes. Blue represents the domain-occurrence network, Red is the simulated network with a scale-free property and in Green a randomly generated.

are a number of measures that can indicate domains that are central to the netw have the most influence in establishing modular protein. Measures such as degre and betweenness centrality we calculated for each node in each domain co network. This identified a number of domains that appear to be central to ma proteins (Appendix 4.7). For example, in the Composite Dataset and Non-compo domains *Pkinase* and *Pkinase_Tyr* were found to be most central in both co networks by degree and closeness measures of centrality, with *Pkinase_Tyr* bein highest in betweenness centrality. The degrees of these domains are 67 and 54 re the Composite Dataset and 113 and 53 respectively in the Non-composite l context, the next most central domain by degree centrality in the Composite *RhoGEF* (degree of 22) and *RVT_1 (*degree of 44*)*. It has been found that kinase important for many modular proteins (461) in the human genome (Manning et al.

Other highly central domains identified have been shown to have an important bi in many proteins such as:

From the Composite Dataset*:*

1. *RhoGEF* (a guanine nucleotide exchange factor) (Soisson et al. 1998),
2. *C2* (involved in targeting proteins to cell membranes) (Haynie and Xue 201
3. *MAM* (an extracellular domain found in many receptor proteins) (Beckma 1993).

From the Non-composite Dataset*:*

1. *RVT_1* (a reverse transcriptase domain) (Gladyshev and Arkhipova 2011),
2. *Trypsin* (a serine protease domain) (Rawlings and Barrett 1994), and
3. *VWA* (von Willebrand factor- a glycoprotein) (Colombatti et al. 1993).

We have indicated the impact of the removing highly central domains from the co network using average clustering and transitivity for the Composite Datase composite Dataset (Figure 4.5). The Highest Confidence Composite Dataset h giant connected component. Therefore, we have not tested the impact of node de network as components are too small for this. It is clear that removing central d the Composite Dataset and Non-composite Dataset causes a significant decreas closeness and betweenness centrality, particularly for degree and betweenness ce result indicates that much fewer domains in the network are connected to othe

Table 4.5: Domains removed from the co-occurrence network for the Composite Dataset to measure average clustering and transitivity

	Degree Centrality		Closeness Centrality		Betweenness Centrality	
Index	Domain ID	Degree	Domain ID	Closeness	Domain ID	Betweenness
1	Pkinase	67	Pkinase	0.089	Pkinase	0.047
2	Pkinase_Tyr	54	Pkinase_Tyr	0.083	Bromodomain	0.024
3	RhoGEF	22	Myosin_head	0.077	Pkinase_Tyr	0.021
4	C2	20	MIT	0.075	AAA	0.02
5	Ion_trans	18	RhoGEF	0.075	MIT	0.019
6	SNF2_N	17	MAM	0.075	BAH	0.018
7	Bromodomain	17	F5_F8_type_C	0.074	Myosin_head	0.015
8	RhoGAP	17	C2	0.074	F5_F8_type_C	0.012
9	Helicase_C	16	Roc	0.073	DEAD	0.011
10	FERM_M	15	Kringle	0.072	C2	0.011
11	PI3_PI4_kinase	15	SH2	0.072	RhoGEF	0.01
12	Myosin_head	15	I-set	0.071	DMAP_binding	0.01
13	CH	14	PX	0.071	AMP-binding	0.01
14	Y_phosphatase	14	SH3-RhoG_link	0.071	RabGAP-TBC	0.008
15	7tm_1	14	RabGAP-TBC	0.071	SPRY	0.007
16	CUB	14	Guanylate_cyc	0.071	MAM	0.007
17	Adaptin_N	13	F_actin_bind	0.07	SET	0.007
18	AMP-binding	13	Guanylate_kin	0.07	Ras	0.007
19	F5_F8_type_C	13	ANF_receptor	0.07	Helicase_C	0.007
20	VWA	13	COR	0.07	SH2	0.006
21	Laminin_G_2	13	Macro	0.07	Roc	0.006
22	MAM	12	MIB_HERC2	0.07	CUB	0.005
23	Cadherin	12	Inhibitor_Mig-6	0.07	HECT	0.005
24	CPSase_L_D2	12	CNH	0.07	adh_short_C2	0.004
25	HECT	12	RBD	0.07	SNF2_N	0.004
26	VWD	12	RGS	0.069	VWD	0.004

27	Trypsin	12	Death	0.069	RhoGAP	0.004
28	7tm_2	12	Dynamin_N	0.069	7tm_1	0.004
29	Laminin_G_1	12	cNMP_binding	0.069	CH	0.004
30	AAA	12	Ank_2	0.069	PRY	0.004
31	ADH_zinc_N	11	EphA2_TM	0.069	Y_phosphatase	0.004
32	PX	11	Ephrin_lbd	0.069	Macro	0.004
33	ANF_receptor	11	Recep_L_domain	0.069	Kringle	0.004
34	UCH	11	Furin-like	0.069	ANF_receptor	0.003
35	MutS_V	11	SAM_1	0.069	DUF3694	0.003
36	SPRY	11	SAM_2	0.069	PWWP	0.003
37	zf-C4	11	OSR1_C	0.069	PX	0.003
38	FERM_C	10	GF_recep_IV	0.069	PARP	0.003
39	Ketoacyl-synt_C	10	PTEN_C2	0.069	FERM_M	0.003
40	ketoacyl-synt	10	Sema	0.069	Trypsin	0.003
41	Thioesterase	10	GTPase_binding	0.069	I-set	0.003
42	HMG_box	10	ecTbetaR2	0.069	NACHT	0.003
43	Ras	10	DCX	0.069	Laminin_G_2	0.003
44	SH2	10	DUF4071	0.069	MIB_HERC2	0.003
45	RabGAP-TBC	10	Death_2	0.069	PI3_PI4_kinase	0.003
46	KR	10	PBD	0.069	eIF2_C	0.003
47	Hormone_recep	10	DUF1908	0.069	PABP	0.003
48	Acyl_transf_1	10	Focal_AT	0.069	VWA	0.003
49	I-set	10	CaMKII_AD	0.069	Guanylate_cyc	0.002
50	ABC_tran	9	Ig_Tie2_1	0.069	Laminin_G_1	0.002

Table 4.5: The Pfam ID and respective measure of centrality for the top 50 nodes in the domain co-occurrence network that were removed according to Degree centrality (Merged Column 1), Closeness centrality (Merged Column 2) and Betweenness centrality (Merged Column 3).

Table 4.6: Domains removed from the co-occurrence network for the Non-composite Dataset to measure average clustering and transitivity

	Degree Centrality		Closeness Centrality		Betweenness Centrality	
Index	Domain ID	Degree	Domain ID	Closeness	Domain ID	Betweenness
1	Pkinase	113	Pkinase	0.167	Pkinase	0.134
2	Pkinase_Tyr	53	Pkinase_Tyr	0.155	RVT_1	0.044
3	RVT_1	44	RVT_1	0.153	Pkinase_Tyr	0.04
4	Trypsin	34	Kringle	0.146	Trypsin	0.027
5	VWA	29	Kinesin	0.144	ABC_tran	0.024
6	ABC_tran	27	Peptidase_C1	0.14	Kinesin	0.017
7	RabGAP-TBC	27	SH2	0.139	Y_phosphatase	0.017
8	SRCR	26	SRCR	0.139	RabGAP-TBC	0.016
9	UCH	23	ANF_receptor	0.138	p450	0.016
10	Y_phosphatase	22	Guanylate_cyc	0.138	Peptidase_C1	0.015
11	RhoGEF	22	ABC_tran	0.137	RhoGEF	0.014
12	DEAD	22	I-set	0.137	adh_short_C2	0.014
13	MAM	21	Sema	0.137	MFS_1	0.014
14	GTP_EFTU	21	Trypsin	0.137	SRCR	0.013
15	F5_F8_type_C	21	Myosin_head	0.137	VWA	0.013
16	Kinesin	21	ASC	0.137	Ion_trans	0.013
17	VWD	21	AAA_12	0.137	7tm_1	0.013
18	RhoGAP	21	RhoGEF	0.137	Kringle	0.013
19	7tm_1	21	C2	0.137	GTP_EFTU	0.012
20	BTB	19	AAA_11	0.136	Helicase_C	0.012
21	MFS_1	19	Y_phosphatase	0.135	UCH	0.012
22	Ion_trans	19	UCH	0.135	RhoGAP	0.012
23	Pro_isomerase	19	Roc	0.135	Myosin_head	0.01
24	Helicase_C	19	COR	0.135	Guanylate_cyc	0.01
25	SNF2_N	18	F_actin_bind	0.135	CPSase_L_D2	0.009
26	Ras	18	MFS_1	0.134	Peptidase_S9	0.008

27	CUB	18	Recep_L_domain	0.134	Ras	0.008
28	Fibrinogen_C	17	EphA2_TM	0.134	ThiF	0.008
29	CPSase_L_D2	17	OSR1_C	0.134	SH2	0.007
30	Guanylate_cyc	17	WIF	0.134	adh_short	0.007
31	7tm_2	17	DUF3543	0.134	Ald_Xan_dh_C2	0.007
32	Methyltransf_FA	16	RabGAP-TBC	0.134	AMP-binding	0.007
33	Kringle	16	7tm_1	0.134	BTB	0.006
34	FERM_M	15	Helicase_C	0.134	Bromodomain	0.006
35	Zona_pellucida	15	ABC_membrane	0.133	ABC_membrane	0.006
36	Gal_Lectin	15	MAM	0.132	Arf	0.006
37	PI3_PI4_kinase	15	Aldedh	0.131	Pro_isomerase	0.006
38	Myosin_head	15	FH2	0.131	COesterase	0.006
39	p450	15	DUF4200	0.131	DEAD	0.006
40	C2	14	RhoGAP	0.131	MAM	0.006
41	CH	14	Pro_isomerase	0.131	Aldolase_II	0.006
42	FH2	14	Galactosyl_T	0.131	AAA	0.005
43	Peptidase_C1	14	E1-E2_ATPase	0.131	Sugar_tr	0.005
44	adh_short	14	Histone	0.13	ANF_receptor	0.005
45	adh_short_C2	14	Cation_ATPase_C	0.13	SNF2_N	0.005
46	Aldedh	14	SIR2	0.13	MBT	0.005
47	AAA	14	CUB	0.13	Hist_deacetyl	0.005
48	IgGFc_binding	14	Peptidase_S9	0.13	C2	0.005
49	AAA_12	14	Sec23_trunk	0.13	CH	0.005
50	Adaptin_N	13	Hist_deacetyl	0.13	Fibrinogen_C	0.005

Table 4.6: The Pfam ID and respective measure of centrality for the top 50 nodes in the domain co-occurrence network that were removed according to Degree centrality (Merged Column 1), Closeness centrality (Merged Column 2) and Betweenness centrality (Merged Column 3).

Figure 4.5: The impact of node deletion for highly central nodes measured by transitivity and average clustering

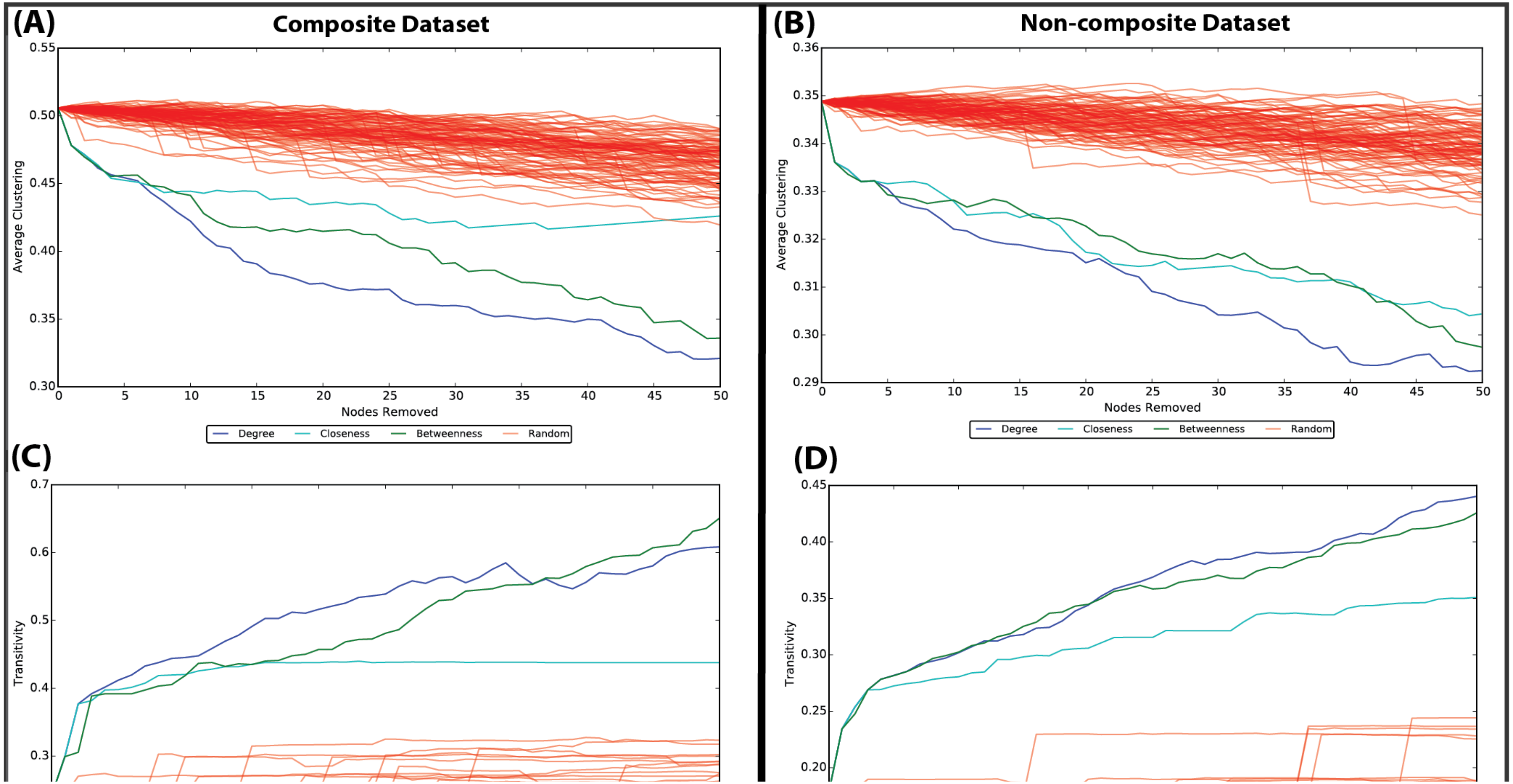


Fig 4.5: The impact of removing random nodes and the 50 most central nodes in the Composite Dataset ((A) and (C)) and Non-composite Dataset ((B) and (D)) networks. The y-axis of the top panel in both datasets ((A) and (B)) represents the average clustering after removing the corresponding number of nodes on the x-axis. The y-axis of the bottom panel in both datasets ((C) and (D)) represents the transitivity after removing the corresponding number of nodes on the x-axis**.** Both graphs are indications of how average clustering is affected when certain domains are removed from the network. The top panel in both datasets ((A) and (B)) is showing how removing central domains from the network is affecting average clustering. The bottom panel in both datasets ((C) and (D)) is showing how removing central domains from the network is affecting the transitivity of the network (a measure of clique structure among connected triplets of domains). Domains were removed from the network by random (red) or according to a centrality value (degree (violet), closeness (blue), or betweenness (green)). For random removal of domains, we repeated this 100 times to give appropriate sampling.

the network and thus forming fewer connected components. Conversely to this, highly central domains (again according to all measures of centrality) from the co network results in a significant increase of transitivity. Transitivity ranges fro where 0 indicates all triplets (3 domains connected by 2 or 3 edges) are open (3 d 2 edges) and 1 indicates all triplets are closed (3 domains with 3 edge a clique). I this indicates that there are more domains forming connected triplets as central removed. Based on average clustering and transitivity, it appears that remo domains in the co-occurrence networks fro the Composite Dataset and No Dataset is creating more closed triplets/cliques by removing domains that would open triplets with two other domains. These results suggest that highly central d an important role in many modular proteins and connect other domains that usually connect in the absence of the central domains.

4.3.2.4 Assortativity

In order to investigate the nature of specific domain combinations in our network co-occurrence we calculated the degree assortativity of each co-occurrence netv 4.6). The degree assortativity indicates if domains in our network have a p connecting to other domains with a similar degree or not. All datasets exhib degree assortativity coefficient for the network. However, the neighbour conn generated for the Composite Dataset and Non-composite Dataset co-occurren show that the linear regression trend line does not fit the data, there is no strong assortative mixing. The Highest Confidence Composite Dataset shows a b regression and indicates an assortative mixing pattern (Figure 4.6). Therefore, tl that the mixing pattern found for the Composite Dataset and Non-composite disassortative and the Highest Confidence Composites Dataset is assortative. Tl indicate that the domains in the Highest Confidence Composite Dataset co networks are assortative in nature, meaning that high degree domains have a p connect with other domains with a high degree for remodelled genes. T (disassortative) is true for non-remodelled genes.

Figure 4.6: Assortativity test for Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset domain co-occurrence networks

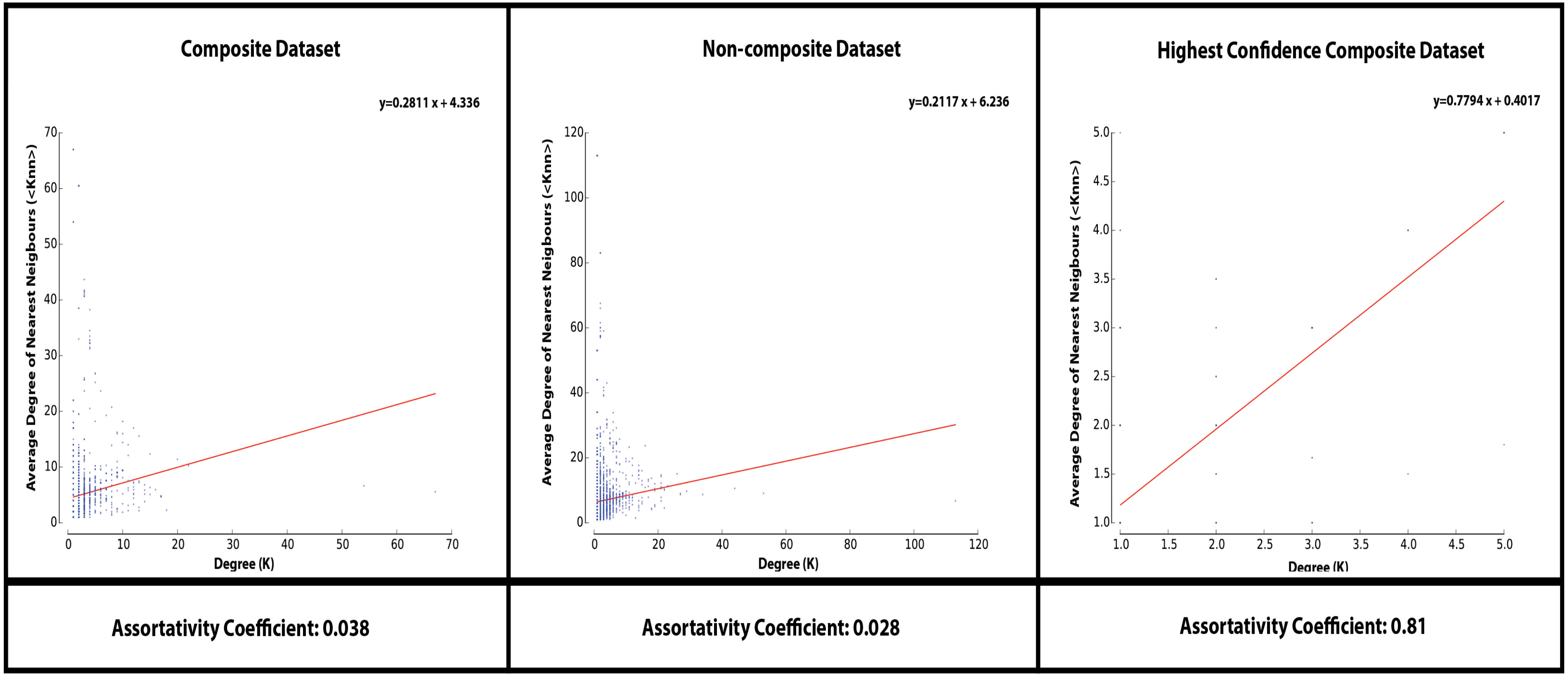


Fig 4.6: The neighbour connectivity plot shows no strong mixing pattern for the Composite and Non-composite Dataset co-occurrence networks, while showing a positive assortative mixing pattern for the Highest Confidence Composite Dataset. K (x-axis) refers to the degree of the given

4.3.2.5 Functional domain combinations and community detection

Using NeMo (Rivera et al. 2010) to detect communities within the largest co-occurrence network for the Composite Dataset and Non-Composite Dataset we established 72 and 149 communities of co-occurring domains respectively. As the Highest Confidence Composite Dataset network only contains small connected components and no giant connected component there was no need to make communities from the large component. Instead each connected component within the domain co-occurrence network was treated as a community as they were small in size and not connected to anywhere else on the network. Similarly, each of the remaining co-occurrence networks (excluding the giant connected component) for the Composite Dataset and Non-Composite Dataset were treated as a community as they were small in size and not connected anywhere else on the network. The enrichment analysis of each community revealed a number of significantly enriched communities of domains for a particular process/function. Summary of the results for Composite, Non-Composite and Highest Confidence Composite Datasets are given in Table 4.7, 4.8 and 4.9 respectively (Appendix 4.5 and 4.6 contains complete set of results).

4.4 Discussion

This chapter establishes some of the features of domain usage observed across metazoan protein coding sequences and provides us with a more detailed understanding of the difference between composite and non-composite genes in terms of their domain remodelling properties.

In summary, all of the domain co-occurrence networks displayed a scale-free property with many low degree nodes and few high degree nodes, meaning that domains co-occur in genes according to a power law degree - a typical characteristic of biological networks (Barabási 2009). The Highest Confidence Composite Dataset network had a degree distribution that ranged from 2 domains (73.5%) to 6 domains (0.7%). Most networks have the minimum of two domains that co-occur, however the range between the highest and lowest degree node is quite small. This indicates that there a small number of central domains that co-occur with many other domains to form modular proteins, but in general most domains are limited in terms of the domains with which they can co-occur in a modular protein.

Table 4.7: Domain community enrichment of GO terms – top 20 significant scores for Composite Dataset

GroupID	TotalMembers	GO_Terms	Recall	Precision	FisherExact p-value
196	3	positive regulation of transcription, DNA-templated	1.00	1.00	7.31E-10
472	3	mannose metabolic process	1.00	1.00	7.31E-10
472	3	alpha-mannosidase activity	1.00	1.00	7.31E-10
571	3	thiamine pyrophosphate binding	1.00	1.00	7.31E-10
1204	3	arginyl-tRNA aminoacylation	1.00	1.00	7.31E-10
1204	3	arginine-tRNA ligase activity	1.00	1.00	7.31E-10
1538	3	protein-arginine deiminase activity	1.00	1.00	7.31E-10
2036	3	S-adenosylmethionine biosynthetic process	1.00	1.00	7.31E-10
2036	3	methionine adenosyltransferase activity	1.00	1.00	7.31E-10
319	2	RNA polyadenylation	1.00	1.00	4.91E-07
533	2	cytoskeleton organization	1.00	1.00	4.91E-07
770	2	nuclease activity	1.00	1.00	4.91E-07
918	2	MHC class II protein binding	1.00	1.00	4.91E-07
1569	2	translational termination	1.00	1.00	4.91E-07

1793	2	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor	1.00	1.00	4.91E-07
1810	2	anion transport	1.00	1.00	4.91E-07
2004	2	oxidoreductase activity, acting on CH-OH group of donors	1.00	1.00	4.91E-07
2325	2	regulation of DNA replication	1.00	1.00	4.91E-07
2408	2	RNA metabolic process	1.00	1.00	4.91E-07
1364	4	amine metabolic process	1.00	0.75	2.92E-09

Table 4.7: The 20 most highly significant hits from the analysis of GO term functional enrichments within communities of co-occurring domains for the Composite Dataset. Ranked according to the highest precision (column 5) and recall (column 4), followed by lowest p-value (column 6). Group ID (column 1) is an id given to a connected component in the network. Total Members (column 2) refers to the total number of domains in the connected component.

Table 4.8: Domain community enrichment of GO terms – top 20 significant scores for Non-composite Dataset

GroupID	TotalMembers	GO_Terms	Recall	Precision	FisherExact p-value
707	3	S-adenosylmethionine biosynthetic process	1	1	2.29E-10
707	3	methionine adenosyltransferase activity	1	1	2.29E-10
804	3	mannose metabolic process	1	1	2.29E-10
804	3	alpha-mannosidase activity	1	1	2.29E-10
1407	3	positive regulation of transcription, DNA-templated	1	1	2.29E-10
1458	3	amine metabolic process	1	1	2.29E-10
1458	3	primary amine oxidase activity	1	1	2.29E-10
1584	3	protein-arginine deiminase activity	1	1	2.29E-10
2748	3	DNA ligase (ATP) activity	1	1	2.29E-10
152	2	translational termination	1	1	2.26E-07
207	2	glycerone kinase activity	1	1	2.26E-07
207	2	glycerol metabolic process	1	1	2.26E-07
272	2	RNA polyadenylation	1	1	2.26E-07
371	2	cytoskeleton	1	1	2.26E-07

		organization			
444	2	proteasome activator complex	1	1	2.26E-07
445	2	phosphoenolpyruvate carboxykinase activity	1	1	2.26E-07
458	2	glycylpeptide N-tetradecanoyltransferase activity	1	1	2.26E-07
829	2	MHC class II protein binding	1	1	2.26E-07
835	2	peptide cross-linking	1	1	2.26E-07
884	2	glutamate-ammonia ligase activity	1	1	2.26E-07

Table 4.8: The 20 most highly significant hits from the analysis of GO term functional enrichments within communities of co-occurring domains for the Non-composite Dataset. Ranked according to the highest precision (column 5) and recall (column 4), followed by lowest p-value (column 6). Group ID (column 1) is an id given to a connected component in the network. Total Members (column 2) refers to the total number of domains in the connected component.

Table 4.9: Domain community enrichment of GO terms – top 20 significant scores for the Highest Confidence Composite Dataset

GroupID	TotalMembers	GO_Terms	Recall	Precision	FisherExact p-value
457	4	mismatch repair	1	1	1.33E-10
457	4	mismatched DNA binding	1	1	1.33E-10
595	4	DNA-directed 5'-3' RNA polymerase activity	1	1	1.33E-10
595	4	transcription, DNA-templated	1	1	1.33E-10
943	3	calcium ion binding	1	1	2.16E-08
943	3	protein-arginine deiminase activity	1	1	2.16E-08
981	3	mannose metabolic process	1	1	2.16E-08
981	3	alpha-mannosidase activity	1	1	2.16E-08
1519	3	lipid transporter activity	1	1	2.16E-08
1519	3	lipid transport	1	1	2.16E-08
1795	3	arginyl-tRNA aminoacylation	1	1	2.16E-08
1795	3	arginine-tRNA ligase activity	1	1	2.16E-08
45	2	transmembrane transport	1	1	4.70E-06
278	2	tRNA aminoacylation for protein translation	1	1	4.70E-06
713	2	intramolecular transferase activity, phosphotransferases	1	1	4.70E-06
1009	2	nuclease activity	1	1	4.70E-06
1152	2	fatty acid metabolic	1	1	4.70E-06

		process			
1152	2	3-hydroxyacyl-CoA dehydrogenase activity	1	1	4.70E-06
1346	2	phosphorylation	1	1	4.70E-06
1363	2	MHC class II protein binding	1	1	4.70E-06

Table 4.9: The 20 most highly significant hits from the analysis of GO term functional enrichments within communities of co-occurring domains for the Highest Confidence Composite Dataset. Ranked according to the highest precision (column 5) and recall (column 4), followed by lowest p-value (column 6). Group ID (column 1) is an id given to a connected component in the network. Total Members (column 2) refers to the total number of domains in the connected component.

In terms of assortativity, the Entire Dataset~~,~~ Composite Dataset and Non-composite Datasets are weakly disassortative (assortativity coefficients of 0.009, 0.038 and 0.028 respectively, but poor linear regressions). The Highest Confidence Composite Dataset was found to have a significant assortative mixing pattern (assortativity coefficient of 0.81). Together, these results indicate that the domains in our Highest Confidence Composite Dataset typically co-occur with domains with a similar degree, suggesting that functionally restricted domains do not favour co-occurring with domains that are functionally permissive. This is in direct contrast to Newman et. al (2002), where they showed that most biological networks high degree nodes have a preference to connect with nodes of a low degree and therefore are disassortative in nature (Newman 2002b) – a pattern displayed by our non-composite genes.

Functional enrichment of communities demonstrated many communities of co-occurring domains are favoured to co-occur with each other. For example, a community in the Entire Dataset, Composite Dataset and Non-composite Dataset were enriched for positive regulation of DNA-templated transcription (Table 4.2, Table 4.7 and Table 4.8) and in all cases, 3 domains ((*HNF-)B_C*, *HNF-1_N,* and *HNF-1A_C*) formed the functionally enriched community. These domains are part of the Hepatocyte nuclear factor-1 family and mutations to these domains can cause type 3 form of maturity-onset diabetes of the young (Urhammer et al. 1997). Evidently, this is an essential biological function that requires specialized domains rather than promiscuous domains of high degree. Similarly, for the Highest Confidence Composite Dataset the top four significant functional enrichments have only four domains (See Table 4.7). Two of the communities contained *MutS* family of domains and were functionally enriched for mismatch repair, mismatched DNA binding. Another two communities were composed of *RNA_pol_Rpb* family of domains and were functionally enriched for DNA-directed 5'-3' RNA polymerase activity and DNA templated transcription. These functions are essential biological functions that require specialised domains (the domains responsible for these functions are highly specialized and not functionally permissive). In all examples, the communities are small and restrictive in terms of the functional domains with which they co-occur.

We have shown that all these networks display similar characteristics, all are scale-free. However, we do find some key differences between the Highest Confidence Composite Dataset and the other two datasets. For example, the Highest Confidence Composite Dataset has assortative mixing compared to a weakly disassortative mixing in the Entire Dataset, Composite Dataset and Non-Composite Dataset. The Highest Confidence Composite Dataset differs in the functions for which we see functional enrichment and it does not contain a giant connected component. The Highest Confidence Composite Dataset is notably smaller than the other datasets. However, on a random network of similar size for the subsetted from the Entire Graph, the trend remains similar – close to neutral/0, with a network assortativity coefficient of -0.096. In essence, in the case of the Highest Confidence Composite Dataset - the domains that co-occur on remodelled genes are functionally restricted in almost all cases and lack functionally promiscuous domains that are present in many other remodelled genes. This is not a feature of the complete Composite Dataset most likely because this dataset contains all putative composite genes, genes that have multiple sources of domains. However, some of these putative domains could be a result of domain rearrangements rather than gene fusion/fission events (Bornberg-Bauer et al. 2010). Therefore, not all genes within the Composite Dataset are strictly gene fusions/fissions in nature, unlike the Highest Confidence Composite Dataset.

In summary, we have applied SSNs to investigate how domains co-occur within composite genes and non-composite genes within the *Metazoa.* With the exception of assortativity all networks analysed display typical properties of a biological network such as a scale-free degree distribution. The largest difference between the networks composed of Non-composite genes and Composite genes was found in community structure of the co-occurrence domains. The composite network had no giant connected component as is the case for non-composite genes. This indicates that domains within the composite network are generally not likely to be functionally compatible with many other domains. In the Non-composite gene SSN, most domains form communities with other specialised domains to contribute to a specific function without the use of promiscuous domains- domains that are involved in a number of different proteins. However, there are a small number of these promiscuous domains that connect with a large number of other domains (Figure 4.1 and Figure 4.4*)*. We have shown here that composite genes mostly rely on the co-occurrence of a low number of specialised domains, while non-composite genes use a mixture of promiscuous domains that function in many proteins and specialised domains to create novel proteins.

Chapter 5: Discussion

Since the completion of the Drosophila melanogaster(Adams et al. 2000) and Human (Venter et al. 2001) genomes in 2000 and 2001 respectively a large number of animal genomes have been published and ambitious global collaborative efforts such as the Genome 10k Project (Koepfli et al. 2015) and the 1KITE (http://www.1kite.org) are generating large volumes of sequence data. In tandem with this, databases that make data available to the community have been developed and improved, including the OMA database (Roth et al. 2008), Ensembl (Hubbard et al. 2002) and NCBI's RefSeq (Hubbard et al. 2002) to name just a few. Large amounts of molecular data have also resulted in novel algorithms in fields such as phylogenetics/phylogenomics. Statistical algorithms such as RAxML (Stamatakis 2006b), and MrBayes (Ronquist and Huelsenbeck 2003) provide homogeneous modeling solutions and PhyloBayes (Lartillot et al. 2009, Lartillot et al. 2013b), P4 (Foster 2004) and RevBayes (Höhna et al. 2016) have heterogeneous modeling capabilities. However, the reality is that these phylogenetic algorithms are often misused.

In Chapter 2, we establish a data driven approach that incorporates both homogeneous and heterogeneous models of phylogeny reconstruction. To test the robustness of this approach we examined the position of the placental mammal root as a good example of a case where the same datasets and taxa sampling have produced conflicting positions (with strong support) for the root depending on what method is used (Tarver et al. 2016). For example, the Romiguier *et al.* study (Romiguier et al. 2013) used homogeneous models to create a phylogeny for placental mammals that placed the root at the Afrotherian position, whereas in the same issue of MBE and using the same data, Morgan e*t al.* (2013) reported the conflicting position of Atlantogenata using a heterogeneous modeling approach (Morgan et al. 2013, Teeling and Hedges 2013, Tarver et al. 2016). We demonstrated that homogeneous models do not reflect the biological complexity of the dataset used by Romiguier *et al* (Romiguier et al. 2013) and while software packages such as RAxML (Stamatakis 2006b) can be incredibly powerful and fast, they are not always appropriate. First, we tested the data for sufficient signal using likelihood mapping (Strimmer and Von Haeseler 1997). This revealed that the data used in the study contained significant phylogenetic signal. Next, we performed the compositional homogeneity test employed in P4 (Foster 2004) and determined that the data was not homogeneous. Using heterogeneous modelling approaches implemented in P4 (Foster 2004) and PhyloBayes (Lartillot et al. 2009) we modeled the Romiguier *et al* datasets (Romiguier et al. 2013) and using posterior predictive simulations we determined if the model adequately describes real data.

The results from Chapter 2 nicely demonstrated that even after searching for the best model in the pool of models, the optimum available heterogeneous model that we can construct may not appropriate for the data, it is simply the best model out of an inappropriate bunch of models. This serves to highlight the importance of checking the adequacy of the model for each specific dataset. One cannot assume that any heterogeneous model is adequate nor can one assume that a heterogeneous model is appropriate at all – perhaps a homogeneous model is sufficient. The point here is that the process of model selection should be data driven. One can only trust the phylogeny produced with a model that adequately and statistically significantly models that data. The key is to resist making assumptions of the data and instead let the data dictate the approach and the model.

Finally, reanalysis of the Romiguier et al dataset (Romiguier et al. 2013) using the heterogeneous CAT+GTR model implemented in PhyloBayes (Lartillot et al. 2009, Lartillot et al. 2013a) shows that this model is capable of modelling the data (posterior predictive simulations) and lends support to the Atlantogenata position for the root rather than the reported Afrotherian position – model misspecification had resulted in the incorrect topology. This work was published in (Moran et al. 2015). In collaboration with Prof Davide Pisani's group in Bristol University we reanalyzed data from studies that had supported alternative hypothesis for the position of the root of placental mammals, and incorporating novel data from microRNAs we have gained consilience across different datasets and data-types for the Atlantogenata root of the placental mammals. We published this work in (Tarver et al. 2016).

Although advances in phylogenetics have greatly improved our understanding of molecular processes of evolution such as gene gain and loss (Chapter 3), phylogenetics is not capable of modeling non-tree-like mechanisms of evolution such as gene remodeling (Chapter 3, Chapter 4) (Bapteste et al. 2013b). There are many studies on gene gain/loss in vertebrates (Blomme et al. 2006, Demuth et al. 2006, Hahn et al. 2007) and indeed inclusion of gene repertoire analyses are commonplace in publications describing new completed genomes (Decatur et al. 2013, Moroz et al. 2014). However, a lack of high quality non-model organism animal genomes (Ellegren 2014) has meant that until recently we have not have sufficient sampling and data to investigate gene gain/loss across the entire *Metazoa*. In the past five years a significant number of non-model organism genomes have become available. (Decatur et al. 2013, Simakov et al. 2013, Moroz et al. 2014), providing us with sufficient sampling to examine gene repertoires across the entire *Metazoa*. Using OMA (Altenhoff et al. 2014) we established all gene gain and loss events across the animal tree for each gene in the database of 1.2 Million sequences we had assembled. We then established the rate and abundance of novel gene families that evolved at each internal node.

Previous studies have shown that gene remodeling is a prevalent and important mechanism in animal evolution (Buljan et al. 2010) (Marsh and Teichmann 2010). However, until recently very few algorithms were available to assess and quantify gene remodeling events (Jachiet et al. 2013). Therefore, with the available data and sampling across the Metazoa and with the development of novel algorithms (*under review* Pathmanathan JS et al, 2017) we had a unique opportunity to establish how gene remodeling has played a role in the evolution of protein coding gene families.

In Chapter 3 we analyzed 63 high quality datasets that span the entire *Metazoa* and identified all gene fusion/fission events. Using a novel network theory based algorithm designed by our collaborators we identified composite genes (gene fusions/fissions) and clustered them into remodeled gene families and mapped them to their node of origin on the phylogenetic tree. The CompositeSearch algorithm was designed by our collaborators with our contribution, it is under review - Pathmanathan JS et al, 2017.

It is well established that gene remodeling is a significant mechanism in creating novel genes(Long 2001, Agaram et al. 2015). It has also known that gene remodeling (fusion/fission genes) has a significant role in functional evolution(Kaessmann 2010) and in driving phenotypic innovation (Ciccarelli et al. 2005, Kaessmann 2010, Chen et al. 2013). However, there is little research into the prevalence of gene fusions/fission across animal life. We identified 71,460 gene families in the database, 68.5% of these were classified as remodeled. We show that most internal nodes in the tree (80.3%) have more remodeled gene families than non-remodeled gene families. These statistics clearly show that gene remodeling is a prevalent mechanism in the evolution of animals. We also show that remodeled gene families originate more often at specific time points on the tree, many of which are nodes of major phenotypic transition. The most notable is the *Bilateria* where 8,075 novel gene families originate – 87% of which are remodeled. This node represents one of the biggest transitions in animal life where the third germ layer is introduced, giving rise to triploblastic animals and allowing for more complex body plans (Section 1.4.2) (Martindale et al. 2002).. Clearly, gene remodeling was a major evolutionary force in the evolution of early Bilaterian species.

Categorizing the remodeled gene families based on their point of origin on the phylogeny and that of their components revealed that the most prominent mechanism of remodeling involves the use of protein coding genes that themselves originated much deeper in the phylogeny or prior to the divergence of the major animal groups. The longer a gene is present in a genome the higher the probability that it will become involved in gene remodeling events.

We have shown that there is distinct difference in how the *Deuterostomia* and *Protostomia* use gene remodeling. *Protostomia* have a higher number of novel gene families created through gene remodeling than the *Deuterostomia.* Overall, we find that the majority of novel gene families that emerged at internal nodes in the tree were formed by gene remodeling. This high prevalence of novel composite genes makes it likely that gene remodeling has a significant impact on phenotypic innovation. There are a number of well documented examples from literature that show examples of gene remodeling creating novel phenotypes (Kaessmann 2010, Chen et al. 2013). For example, the *jingwei* gene in *Drosophila melanogaster* introduced a novel specificity to metabolising long chain primary alcohols (Long and Langley 1993). This suggests that gene remodeling is a strong force in creating novel gene families and thus potentially novel phenotypes. The enrichment analysis revealed that certain functions are specifically enriched in remodeled gene families. For example, *Euteleostomi* represents a major transition for the adaptive immunity in animals and is also enriched for remodeled genes emerging at that point on the tree that have immune system functions.

Studies on fossil data suggests that the rate of evolution (for the most part) is gradual (Estes and Arnold 2007, Hunt et al. 2015), while studies on ecological data suggests bursts of punctuated evolution are potentially common (Herrel et al. 2008, Brown and Brown 2013). In animals, the rate of novel gene genesis by gene remodeling (gene fusion/fission) was unknown. We show that novel remodeled genes have emerged at a faster rate than novel non-remodeled genes and these rates can fluctuate through time.

A significant amount of research has been carried out on the impact of domains in creating novel functions (Buljan et al. 2010, Marsh and Teichmann 2010). It has been shown that there are domains that form part of many of modular proteins, these are referred to as promiscuous domains (Basu et al. 2008). For example, *kinase* and *fibronectin III* domains are found in many modular proteins (Little et al. 1994, Manning et al. 2002).Furthermore, it is suggested that domain shuffling played a significant role in the transition from animal unicellularity to multicellularity(Patthy 2003, Ekman et al. 2007). However, it is not fully understood how the rules of domain shuffling may differ between remodeled and non-remodeled genes. We compared the domain shuffling characteristics such as assortativity, network structure and transitivity from a composite gene perspective and non-composite gene perspective. All of the domain co-occurrence networks displayed a scale-free property with many low degree nodes and few high degree nodes – a typical characteristic of biological networks (Barabási 2009). For the Highest Confidence Composite Dataset there was no single giant connected component. 73.5% of all co-occurrence networks in this dataset had the minimum number of two domains. This indicates that there a small number of central domains that co-occur with many other domains to form a modular protein, but in general most domains are limited in the domains they can co-occur with on a gene/modular protein, when composite genes are concerned. In the case of the Entire Dataset, Composite Dataset and Non-composite Dataset we show that there are a number of highly promiscousous domains. The presence of a relatively low number of promiscousous domains is in agreement with current literature (Basu et al. 2008, Barabási 2009) .

The co-occurrence networks*:* Full Dataset, Composite Dataset, Non-composite Dataset exhibited a weak disassortative mixing pattern. and Highest Confidence Composite Dataset exhibited an assortative mixing pattern. Overall this suggests that Pfam-A domains in general co-occur with other domains of a similar degree for composite genes and for non-composite genes domains co-occur with other domains that often have a different degree. It has been found that biological networks often have dissasortative mixing pattern (Newman 2002a, Newman 2003, Proulx et al. 2005).

It is understood that domains have an importnant role in creating novel funcitons in eukaryotes(Jin et al. 2009). Functional enrichment of communities of co-occruing domains from the Entire Dataset, Composite Dataset, Non-composite Dataset and Highest Confidence Composite Dataset demonstrated many communities of co-occurring domains are favoured to co-occur with each other in order to carry out a specific function.

Although the different Dataset networks display similar characteristics such as a scale-free property and assortative mixing, we do find some key differences between the Highest Confidence Composite Dataset and the other datasets (Entire Dataset, Composite Dataset and Non-composite Dataset). For example, the Highest Confidence Composite Dataset network has a very strong assortative mixing compared to a weakly assortative mixing in the other networks. The Highest Confidence Composite Dataset network differs in the functions it is enriched for and contains no giant connected component. In essence, the Highest Confidence Composite Dataset network indicates that domains that co-occur on composite genes are functionally restricted for almost all cases with a lack of functionally permissive domains that are present in many other composite genes.

We have offered a global view of the impact of gene remodeling in the *Metazoa.* However, to gain a further understanding of how this mechanism differs between specific groups or clades of animals - denser sampling and a robust phylogenetic tree is required. Therefore, as technology advances and more genomes become available there is an opportunity to improve upon analyses. Whilst we have resolved one contentious node in the animal tree (the root of the placental mammal tree) (Moran et al. 2015, Tarver et al. 2016), a number of other contentious nodes remain (Dunn et al. 2008). The data driven heterogeneous modelling approach we applied in Chapter 2 offer a good potential approach to take with larger and better quality genomes and perhaps with improved modeling and data these nodes will resolve. The use of a more fully resolved phylogenetic tree for the *Metazoa* would allow the algorithms we used such as OMA (Roth et al. 2008) and CompositeSearch (Pathmanathan JS et al, 2017) to have improved accuracy. We have offered an insight into the role of composite genes in animal species. However there is still much work to be done to build on the work of this thesis such as selective pressure analyses and more in-depth functional analyses that will help us understand in greater detail the functional impact and importance of composite genes in the *Metazoa*.

Conclusion:

In this thesis we explored protein-coding gene evolution using tree-like and non-tree-like frameworks. For the first time, we successfully got consilience across several data types on a large-scale study for a heavily debated node on the animal tree – the root of the placental mammals. We applied graph theory to quantitatively survey the extent of gene remodeling across the *Metazoa.* We determined that gene remodeling is a prevalent process across all clades in the *Metazoa* and that the emergence of novel gene families is significantly facilitated by gene remodeling processes. We show that the rate of novel gene genesis is not strictly clocklike for remodeled gene families. We have gained some initial insights into how gene remodeling may contribute to novel phenotypes at nodes of major phenotypic transition. Finally, we have quantified the extent of domain shuffling and have elaborated on the patterns of domain shuffling observed in the Metazoa.

Chapter 6: Bibliography

Abascal, Federico, David Posada and Rafael Zardoya (2007). "MtArt: a new model of amino acid replacement for Arthropoda." Molecular biology and evolution **24**(1): 1-5.

Abascal, Federico, Rafael Zardoya and David Posada (2005). "ProtTest: selection of best-fit models of protein evolution." Bioinformatics **21**(9): 2104-2105.

Abouheif, Ehab, Rafael Zardoya and Axel Meyer (1998). "Limitations of metazoan 18S rRNA sequence data: implications for reconstructing a phylogeny of the animal kingdom and inferring the reality of the Cambrian explosion." Journal of Molecular Evolution **47**(4): 394-405.

Adachi, Jun, Peter J. Waddell, William Martin and Masami Hasegawa (2000). "Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA." Journal of Molecular Evolution **50**(4): 348-358.

Adai, A. T., S. V. Date, S. Wieland and E. M. Marcotte (2004). "LGL: creating a map of protein function with an algorithm for visualizing very large biological networks." J Mol Biol **340**(1): 179-190.

Adams, Mark D, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins and Richard F Galle (2000). "The genome sequence of Drosophila melanogaster." science **287**(5461): 2185-2195.

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 and Cancer **54**(2): 63-71.

Albert, Reka (2005). "Scale-free networks in cell biology." Journal of cell science **118**(21): 4947-4957.

Almgren, Malin AE, K Cecilia E Henriksson, Jennifer Fujimoto and Christina L Chang (2004). "Nucleoside Diphosphate Kinase A/nm23-H1 Promotes Metastasis of NB69-Derived Human Neuroblastoma11NIH RO1 CA78241 and RO1 CA78241S grants (CL Chang)." Molecular cancer research **2**(7): 387-394.

Altekar, Gautam, Sandhya Dwarkadas, John P. Huelsenbeck and Fredrik Ronquist (2004). "Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference." Bioinformatics **20**(3): 407-415.

Altenhoff, Adrian M, Nives Škunca, Natasha Glover, Clément-Marie Train, Anna Sueki, Ivana Piližota, Kevin Gori, Bartlomiej Tomiczek, Steven Müller and Henning Redestig (2014). "The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements." Nucleic acids research: gku1158.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990a). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers and David J. Lipman (1990b). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

Alvarez-Ponce, D., P. Lopez, E. Bapteste and J. O. McInerney (2013a). "Gene similarity networks provide tools for understanding eukaryote origins and evolution." Proc Natl Acad Sci U S A **110**(17): E1594-1603.

Alvarez-Ponce, David, Philippe Lopez, Eric Bapteste and James O. McInerney (2013b). "Gene similarity networks provide tools for understanding eukaryote origins and evolution." Proceedings of the National Academy of Sciences **110**(17): E1594-E1603.

Arnason, Ulfur and Axel Janke (2002). "Mitogenomic analyses of eutherian relationships." Cytogenetic and genome research **96**(1-4): 20-32.

Ashley-Ross, Miriam A., S. Tonia Hsieh, Alice C. Gibb and Richard W. Blob (2013). "Vertebrate land invasions—past, present, and future: an introduction to the symposium." Integrative and Comparative Biology: ict048.

Azuaje, Francisco J. (2014). "Selecting biologically informative genes in co-expression networks with a centrality score."

Bapteste, E., L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, L. Stougie and J. Whitfield (2013a). "Networks: expanding evolutionary thinking." Trends Genet **29**(8): 439-441.

Bapteste, Eric, Philippe Lopez, Fr√©d√©ric Bouchard, Fernando Baquero, James O McInerney and Richard M Burian (2012). "Evolutionary analyses of non-genealogical bonds produced by introgressive descent." Proceedings of the National Academy of Sciences **109**(45): 18266-18272.

Bapteste, Eric, Leo van Iersel, Axel Janke, Scot Kelchner, Steven Kelk, James O. McInerney, David A. Morrison, Luay Nakhleh, Mike Steel and Leen Stougie (2013b). "Networks: expanding evolutionary thinking." Trends in Genetics **29**(8): 439-441.

Barabási, Albert-László (2009). "Scale-free networks: a decade and beyond." science **325**(5939): 412-413.

Barabasi, Albert-Laszlo and Zoltan N Oltvai (2004). "Network biology: understanding the cell's functional organization." Nature Reviews Genetics **5**(2): 101-113.

Barrat, Alain, Marc Barthelemy, Romualdo Pastor-Satorras and Alessandro Vespignani (2004). "The architecture of complex weighted networks." Proceedings of the National Academy of Sciences of the United States of America **101**(11): 3747-3752.

Bashton, Matthew and Cyrus Chothia (2007). "The generation of new protein functions by the combination of domains." Structure **15**(1): 85-99.

Basu, Malay Kumar, Liran Carmel, Igor B Rogozin and Eugene V Koonin (2008). "Evolution of protein domain promiscuity in eukaryotes." Genome research **18**(3): 449-461.

Basu, Malay Kumar, Eugenia Poliakov and Igor B Rogozin (2009). "Domain mobility in proteins: functional and evolutionary implications." Briefings in bioinformatics **10**(3): 205-216.

Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon and Erik L. L. Sonnhammer (2004). "The Pfam protein families database." Nucleic acids research **32**(suppl 1): D138-D141.

Beckmann, Georg and Peer Bork (1993). "An adhesive domain detected in functionally diverse receptors." Trends in Biochemical Sciences **18**(2): 40-41.

Bell, Graham (2015). The Evolution of Life, Oxford University Press.

Blackburne, Benjamin P. and Simon Whelan (2012). "Measuring the distance between multiple sequence alignments." Bioinformatics **28**(4): 495-502.

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.

Borgatti, Stephen P. (2005). "Centrality and network flow." Social networks **27**(1): 55-71.

Bornberg-Bauer, Erich, Ann-Kathrin Huylmans and Tobias Sikosek (2010). "How do new proteins arise?" Current opinion in structural biology **20**(3): 390-396.

Boschetti, Chiara, Natalia Pouchkina-Stantcheva, Pia Hoffmann and Alan Tunnacliffe (2011). "Foreign genes and novel hydrophilic protein genes participate in the desiccation response of the bdelloid rotifer Adineta ricciae." The Journal of experimental biology **214**(1): 59-68.

Bradnam, Keith R., Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A. Chapman, Guillaume Chapuis and Rayan Chikhi (2013). "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." GigaScience **2**(1): 1-31.

Brandley, Matthew C., Dan L. Warren, Adam D. Leaché and Jimmy A. McGuire (2009). "Homoplasy and clade support." Systematic Biology **58**(2): 184-198.

Brazeau, Martin D. and Matt Friedman (2015). "The origin and early phylogenetic history of jawed vertebrates." Nature **520**(7548): 490-497.

Brown, Charles R and Mary Bomberger Brown (2013). "Where has all the road kill gone?" Current Biology **23**(6): R233-R234.

Brown, J. K. (1994). "Bootstrap hypothesis tests for evolutionary trees and other dendrograms." Proceedings of the National Academy of Sciences **91**(25): 12293-12297.

Brown, Terence A. (2002). "Molecular phylogenetics."

Buljan, Marija, Adam Frankish and Alex Bateman (2010). "Quantifying the mechanisms of domain gain in animal proteins." Genome biology **11**(7): R74.

Cao, Ying, Jun Adachi, Axel Janke, Svante Pääbo and Masami Hasegawa (1994). "Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene." Journal of Molecular Evolution **39**(5): 519-527.

Carroll, Sean B. (2005). "Evolution at two levels: on genes and form." PLoS biology **3**(7): 1159.

Cerny, Robert, Maria Cattell, Tatjana Sauka-Spengler, Marianne Bronner-Fraser, Feiqiao Yu and Daniel Meulemans Medeiros (2010). "Evidence for the prepattern/cooption model of vertebrate jaw evolution." Proceedings of the National Academy of Sciences **107**(40): 17262-17267.

Chen, Sidi, Benjamin H Krinsky and Manyuan Long (2013). "New genes as drivers of phenotypic evolution." Nature Reviews Genetics **14**(9): 645-660.

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): 343-351.

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): 1283-1287.

Clack, Jennifer A (2005). "Getting a leg up on land." Scientific American **293**(6): 100-107.

Colombatti, Alfonso, Paolo Bonaldo and Roberto Doliana (1993). "Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins." Matrix **13**(4): 297-306.

Consortium, Gene Ontology (2015). "Gene ontology consortium: going forward." Nucleic acids research **43**(D1): D1049-D1056.

Consortium, International Human Genome Sequencing (2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-945.

Consortium, Sequencing (1998). "Genome sequence of the nematode C. elegans: A platform for investigating biology." science **282**: 2012-2018.

Coyne, Jerry A (2009). Why evolution is true, Penguin.

Craft, Meggan E (2015). "Infectious disease transmission and contact networks in wildlife and livestock." Phil. Trans. R. Soc. B **370**(1669): 20140107.

Crawford, Nicholas G., James F. Parham, Anna B. Sellas, Brant C. Faircloth, Travis C. Glenn, Theodore J. Papenfuss, James B. Henderson, Madison H. Hansen and W. Brian Simison (2015). "A phylogenomic analysis of turtles." Molecular phylogenetics and evolution **83**: 250-257.

Crisp, Alastair, Chiara Boschetti, Malcolm Perry, Alan Tunnacliffe and Gos Micklem (2015). "Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes." Genome biology **16**(1): 50.

Crompton, AW and Farish A Jenkins Jr (1979). "Origin of mammals." JA Lillegraven, Z. Kielan− Jaworowska, and WA Clemens, Jr.(eds.), Me− sozoic Mammals: The First Two− thirds of Mammalian History: 59-73.

D'Apice, MR, R Tenconi, I Mammi, J van den Ende and G Novelli (2004). "Paternal origin of LMNA mutations in Hutchinson–Gilford progeria." Clinical genetics **65**(1): 52-54.

Dávalos, Liliana M., Andrea L. Cirranello, Jonathan H. Geisler and Nancy B. Simmons (2012). "Understanding phylogenetic incongruence: lessons from phyllostomid bats." Biological Reviews **87**(4): 991-1024.

Dayhoff, Margaret O. and Robert M. Schwartz (1978). A model of evolutionary change in proteins. In Atlas of protein sequence and structure, Citeseer.

De Robertis, Eddy M. and Yoshiki Sasai (2000). "A common plan for dorsoventral patterning in Bilateria." Shaking the Tree: Readings from Nature in the History of Life: 89.

DeBiasse, Melissa B. and Michael E. Hellberg (2015). "Discordance between morphological and molecular species boundaries among Caribbean species of the reef sponge Callyspongia." Ecology and evolution **5**(3): 663-675.

Decatur, Wayne A, Jeffrey A Hall, Jeramiah J Smith, Weiming Li and Stacia A Sower (2013). "Insight from the lamprey genome: glimpsing early vertebrate development via neuroendocrine-associated genes and shared synteny of gonadotropin-releasing hormone (GnRH)." General and comparative endocrinology **192**: 237-245.

Degnan, James H. and Noah A. Rosenberg (2006). "Discordance of species trees with their most likely gene trees." PLoS Genet **2**(5): e68.

Degnan, James H. and Noah A. Rosenberg (2009). "Gene tree discordance, phylogenetic inference and the multispecies coalescent." Trends in ecology & evolution **24**(6): 332-340.

Delarbre, Christiane, Cyril Gallut, Veronique Barriel, Philippe Janvier and Gabriel Gachelin (2002). "Complete mitochondrial DNA of the hagfish, Eptatretus burgeri: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly." Molecular phylogenetics and evolution **22**(2): 184-192.

Delsuc, F., H. Brinkmann and H. Philippe (2005). "Phylogenomics and the reconstruction of the tree of life." Nat Rev Genet **6**(5): 361-375.

Delsuc, Frédéric, Henner Brinkmann, Daniel Chourrout and Hervé Philippe (2006). "Tunicates and not cephalochordates are the closest living relatives of vertebrates." Nature **439**(7079): 965-968.

Demichelis, Francesca, K. Fall, S. Perner, Ove Andrén, F. Schmidt, S. R. Setlur, Y. Hoshida, J. M. Mosquera, Y. Pawitan and C. Lee (2007). "TMPRSS2: ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort." Oncogene **26**(31): 4596-4599.

Demuth, Jeffery P, Tijl De Bie, Jason E Stajich, Nello Cristianini and Matthew W Hahn (2006). "The evolution of mammalian gene families." PloS one **1**(1): e85.

Dimmic, Matthew W., Joshua S. Rest, David P. Mindell and Richard A. Goldstein (2002). "rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny." Journal of Molecular Evolution **55**(1): 65-73.

Dornburg, Alex, Matt Friedman and Thomas J. Near (2015). "Phylogenetic analysis of molecular and morphological data highlights uncertainty in the relationships of fossil and living species of Elopomorpha (Actinopterygii: Teleostei)." Molecular phylogenetics and evolution **89**: 205-218.

Douady, Christophe J., Frédéric Delsuc, Yan Boucher, W. Ford Doolittle and Emmanuel J. P. Douzery (2003). "Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability." Molecular biology and evolution **20**(2): 248-254.

Dunn, Casey W., Andreas Hejnol, David Q. Matus, Kevin Pang, William E. Browne, Stephen A. Smith, Elaine Seaver, Greg W. Rouse, Matthias Obst and Gregory D. Edgecombe (2008). "Broad phylogenomic sampling improves resolution of the animal tree of life." Nature **452**(7188): 745-749.

Eddy, Sean R. and Burkhard Rost (2008). "A probabilistic model of local sequence alignment that simplifies statistical significance estimation." PLoS Comput Biol **4**(5): e1000069.

Edgar, Robert C (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic acids research **32**(5): 1792-1797.

Ekman, Diana, Åsa K Björklund and Arne Elofsson (2007). "Quantification of the elevated rate of domain rearrangements in metazoa." Journal of Molecular Biology **372**(5): 1337-1348.

Ellegren, Hans (2014). "Genome sequencing and population genomics in non-model organisms." Trends in ecology & evolution **29**(1): 51-63.

Estes, Suzanne and Stevan J Arnold (2007). "Resolving the paradox of stasis: models with stabilizing selection explain evolutionary divergence on all timescales." The American Naturalist **169**(2): 227-244.

Fang, JingXian, Takeshi Uchiumi, Mikako Yagi, Shinya Matsumoto, Rie Amamoto, Shinya Takazaki, Haruyoshi Yamaza, Kazuaki Nonaka and Dongchon Kang (2013). "Dihydro-orotate dehydrogenase is physically associated with the respiratory complex and its loss leads to mitochondrial dysfunction." Bioscience reports **33**(2): e00021.

Farris, James S (1977). "Phylogenetic analysis under Dollo's Law." Systematic Biology **26**(1): 77-88.

Felsenstein, Joseph (1978). "Cases in which parsimony or compatibility methods will be positively misleading." Systematic Biology **27**(4): 401-410.

Felsenstein, Joseph (1981). "Evolutionary trees from DNA sequences: a maximum likelihood approach." Journal of Molecular Evolution **17**(6): 368-376.

Felsenstein, Joseph (1985). "Phylogenies and the comparative method." The American Naturalist **125**(1): 1-15.

Felsenstein, Joseph and Gary A. Churchill (1996). "A Hidden Markov Model approach to variation among sites in rate of evolution." Molecular biology and evolution **13**(1): 93-104.

Finn, Robert D, Jody Clements and Sean R Eddy (2011). "HMMER web server: interactive sequence similarity searching." Nucleic acids research: gkr367.

Finn, Robert D, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi and Amaia Sangrador-Vegas (2016). "The Pfam protein families database: towards a more sustainable future." Nucleic acids research **44**(D1): D279-D285.

Flajnik, Martin F (2014). "Re-evaluation of the immunological Big Bang." Current Biology **24**(21): R1060-R1065.

Flot, Jean-François, Boris Hespeels, Xiang Li, Benjamin Noel, Irina Arkhipova, Etienne G. J. Danchin, Andreas Hejnol, Bernard Henrissat, Romain Koszul and Jean-Marc Aury (2013). "Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga." Nature **500**(7463): 453-457.

Force, Allan, Michael Lynch, F. Bryan Pickett, Angel Amores, Yi-lin Yan and John Postlethwait (1999). "Preservation of duplicate genes by complementary, degenerative mutations." Genetics **151**(4): 1531-1545.

Foster, Jacob G, David V Foster, Peter Grassberger and Maya Paczuski (2010). "Edge direction and the structure of networks." Proceedings of the National Academy of Sciences **107**(24): 10815-10820.

Foster, Peter G and Donal A Hickey (1999). "Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions." Journal of Molecular Evolution **48**(3): 284-290.

Foster, Peter G. (2004). "Modeling compositional heterogeneity." Systematic Biology **53**(3): 485-495.

Frost, Laura S., Raphael Leplae, Anne O. Summers and Ariane Toussaint (2005). "Mobile genetic elements: the agents of open source evolution." Nature Reviews Microbiology **3**(9): 722-732.

Furlong, Rebecca F. and Peter W. H. Holland (2002). "Bayesian phylogenetic analysis supports monophyly of ambulacraria and of cyclostomes." Zoological science **19**(5): 593-599.

Galtier, N., G. Piganeau, D. Mouchiroud and L. Duret (2001). "GC-content evolution in mammalian genomes: the biased gene conversion hypothesis." Genetics **159**(2): 907-911.

Galtier, Nicolas and Laurent Duret (2007). "Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution." Trends in Genetics **23**(6): 273-277.

Gans, Carl and R. Glenn Northcutt (1983). "Neural crest and the origin of vertebrates: a new head." science **220**(4594): 268-273.

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**(0): 231-266.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari and Donald B Rubin (2014). Bayesian data analysis, CRC press Boca Raton, FL.

Gerhart, J. (2000). "Inversion of the chordate body axis: Are there alternatives?" Proceedings of the National Academy of Sciences **97**(9): 4445-4448.

Gibson, Andrew, Vivek Gowri-Shankar, Paul G. Higgs and Magnus Rattray (2005). "A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods." Molecular biology and evolution **22**(2): 251-264.

Gladyshev, Eugene A and Irina R Arkhipova (2011). "A widespread class of reverse transcriptase-related cellular genes." Proceedings of the National Academy of Sciences **108**(51): 20311-20316.

Gould, Niles Eldredge-Stephen Jay (1972). "Punctuated equilibria: an alternative to phyletic gradualism."

Graham, Jeffrey B., Nicholas C. Wegner, Lauren A. Miller, Corey J. Jew, N. Chin Lai, Rachel M. Berquist, Lawrence R. Frank and John A. Long (2014). "Spiracular air breathing in polypterid fishes and its implications for aerial respiration in stem tetrapods." Nature communications **5**.

Grajales, Alejandro, Catalina Aguilar and Juan A Sánchez (2007). "Phylogenetic reconstruction using secondary structures of Internal Transcribed Spacer 2 (ITS2, rDNA): finding the molecular and morphological gap in Caribbean gorgonian corals." BMC Evolutionary Biology **7**(1): 90.

Grenfell, Bryan T., Oliver G. Pybus, Julia R. Gog, James L. N. Wood, Janet M. Daly, Jenny A. Mumford and Edward C. Holmes (2004). "Unifying the epidemiological and evolutionary dynamics of pathogens." science **303**(5656): 327-332.

Hahn, Matthew W, Jeffery P Demuth and Sang-Gook Han (2007). "Accelerated rate of gene gain and loss in primates." Genetics **177**(3): 1941-1949.

Hasegawa, Masami, Tetsuo Hashimoto, Jun Adachi, Naoyuki Iwabe and Takashi Miyata (1993). "Early branchings in the evolution of eukaryotes: ancient divergence of Entamoeba that lacks mitochondria revealed by protein sequence data." Journal of Molecular Evolution **36**(4): 380-388.

Hastings, W. Keith (1970). "Monte Carlo sampling methods using Markov chains and their applications." Biometrika **57**(1): 97-109.

Haynie, Donald T and Bin Xue (2015). "Superdomains in the protein structure hierarchy: The case of PTP‐C2." Protein Science **24**(5): 874-882.

Heimberg, Alysha M., Richard Cowper-Sal, Marie Sémon, Philip C. J. Donoghue and Kevin J. Peterson (2010). "microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate." Proceedings of the National Academy of Sciences **107**(45): 19379-19383.

Hejnol, Andreas, Matthias Obst, Alexandros Stamatakis, Michael Ott, Greg W. Rouse, Gregory D. Edgecombe, Pedro Martinez, Jaume Baguñà, Xavier Bailly and Ulf Jondelius (2009). "Assessing the root of bilaterian animals with scalable phylogenomic methods." Proceedings of the Royal Society of London B: Biological Sciences **276**(1677): 4261-4270.

Heliconius Genome, Consortium (2012). "Butterfly genome reveals promiscuous exchange of mimicry adaptations among species." Nature **487**(7405): 94-98.

Henikoff, Steven and Jorja G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proceedings of the National Academy of Sciences **89**(22): 10915-10919.

Herrel, Anthony, Katleen Huyghe, Bieke Vanhooydonck, Thierry Backeljau, Karin Breugelmans, Irena Grbac, Raoul Van Damme and Duncan J Irschick (2008). "Rapid large-scale evolutionary divergence in morphology and performance associated with exploitation of a different dietary resource." Proceedings of the National Academy of Sciences **105**(12): 4792-4795.

Hoen, Douglas R. and Thomas E. Bureau (2015). "Discovery of novel genes derived from transposable elements using integrative genomic analysis." Molecular biology and evolution **32**(6): 1487-1506.

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): 726-736.

Holt, Robert A, G Mani Subramanian, Aaron Halpern, Granger G Sutton, Rosane Charlab, Deborah R Nusskern, Patrick Wincker, Andrew G Clark, JoséM C Ribeiro and Ron Wides (2002). "The genome sequence of the malaria mosquito Anopheles gambiae." science **298**(5591): 129-149.

Hotopp, Julie C. Dunning, Michael E. Clark, Deodoro C. S. G. Oliveira, Jeremy M. Foster, Peter Fischer, Mónica C. Muñoz Torres, Jonathan D. Giebel, Nikhil Kumar, Nadeeza Ishmael and Shiliang Wang (2007). "Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes." science **317**(5845): 1753-1756.

Huang, Sui, Gabriel Eichler, Yaneer Bar-Yam and Donald E Ingber (2005). "Cell fates as high-dimensional attractor states of a complex gene regulatory network." Physical review letters **94**(12): 128701.

Hubbard, Tim, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen and Thomas Down (2002). "The Ensembl genome database project." Nucleic acids research **30**(1): 38-41.

Hudson, Richard R. "Gene genealogies and the coalescent process."

Huelsenbeck, John P. and Keith A. Crandall (1997). "Phylogeny estimation and hypothesis testing using maximum likelihood." Annual Review of Ecology and Systematics: 437-466.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen and Jonathan P. Bollback (2001). "Bayesian inference of phylogeny and its impact on evolutionary biology." science **294**(5550): 2310-2314.

Hughes, Austin L. (1994). "The evolution of functionally novel proteins after gene duplication." Proceedings of the Royal Society of London B: Biological Sciences **256**(1346): 119-124.

Hunt, Gene, Melanie J Hopkins and Scott Lidgard (2015). "Simple versus complex models of trait evolution and stasis as a response to environmental change." Proceedings of the National Academy of Sciences **112**(16): 4885-4890.

Imwong, Mallika, Sasithon Pukrittayakamee, Laurent Rénia, Franck Letourneur, Jean-Paul Charlieu, Ubolsree Leartsakulpanich, Sornchai Looareesuwan, Nicholas J White and Georges Snounou (2003). "Novel point mutations in the dihydrofolate reductase gene of Plasmodium vivax: evidence for sequential selection by drug pressure." Antimicrobial agents and chemotherapy **47**(5): 1514-1521.

Irestedt, Martin, Jon Fjeldså and Per G. P. Ericson (2004). "Phylogenetic relationships of woodcreepers (Aves: Dendrocolaptinae)–incongruence between molecular and morphological data." Journal of Avian Biology **35**(3): 280-288.

Ishikawa, Sohta A., Yuji Inagaki and Tetsuo Hashimoto (2012). "RY-coding and non-homogeneous models can ameliorate the maximum-likelihood inferences from nucleotide sequence data with parallel compositional heterogeneity." Evolutionary bioinformatics online **8**: 357.

Jachiet, P. A., P. Colson, P. Lopez and E. Bapteste (2014a). "Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network." Genome Biol Evol **6**(9): 2195-2205.

Jachiet, Pierre-Alain, Philippe Colson, Philippe Lopez and Eric Bapteste (2014b). "Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network." Genome biology and evolution **6**(9): 2195-2205.

Jachiet, Pierre-Alain, Romain Pogorelcnik, Anne Berry, Philippe Lopez and Eric Bapteste (2013). "MosaicFinder: identification of fused gene families in sequence similarity networks." Bioinformatics **29**(7): 837-844.

Janvier, Philippe (2015). "Facts and fancies about early fossil chordates and vertebrates." Nature **520**(7548): 483-489.

Jeong, Hawoong, Sean P. Mason, A. L. Barabási and Zoltan N. Oltvai (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-42.

Jin, Jing, Xueying Xie, Chen Chen, Jin Gyoon Park, Chris Stark, D Andrew James, Marina Olhovsky, Rune Linding, Yongyi Mao and Tony Pawson (2009). "Eukaryotic protein domains as functional units of cellular evolution." Sci Signal **2**(98): 1-18.

Jones, D. T., W. R. Taylor and J. M. Thornton (1994). "A mutation data matrix for transmembrane proteins." FEBS letters **339**(3): 269-275.

Jones, Eric, Travis Oliphant and Pearu Peterson (2014a). "{SciPy}: open source scientific tools for {Python}."

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell and Gift Nuka (2014b). "InterProScan 5: genome-scale protein function classification." Bioinformatics **30**(9): 1236-1240.

Jukes, Thomas H. and Charles R. Cantor (1969). "Evolution of protein molecules." Mammalian protein metabolism **3**: 21-132.

Jun, Jin, Paul Ryvkin, Edward Hemphill, Ion Mandoiu and Craig Nelson (2009). "The birth of new genes by RNA-and DNA-mediated duplication during mammalian evolution." Journal of Computational Biology **16**(10): 1429-1444.

Kaessmann, Henrik (2010). "Origins, evolution, and phenotypic impact of new genes." Genome research **20**(10): 1313-1326.

Kaessmann, Henrik, Nicolas Vinckenbosch and Manyuan Long (2009). "RNA-based gene duplication: mechanistic and evolutionary insights." Nature Reviews Genetics **10**(1): 19-31.

Kaessmann, Henrik, Sebastian Zöllner, Anton Nekrutenko and Wen-Hsiung Li (2002). "Signatures of domain shuffling in the human genome." Genome research **12**(11): 1642-1650.

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes factors." Journal of the american statistical association **90**(430): 773-795.

Keane, T. M., T. J. Naughton and J. O. McInerney (2004). "ModelGenerator: amino acid and nucleotide substitution model selection." National University of Ireland, Maynooth, Ireland.

Keeling, Patrick J. and Jeffrey D. Palmer (2008). "Horizontal gene transfer in eukaryotic evolution." Nature Reviews Genetics **9**(8): 605-618.

Kelava, Iva, Eric Lewitus and Wieland B. Huttner (2013). "The secondary loss of gyrencephaly as an example of evolutionary phenotypical reversal." Frontiers in neuroanatomy **7**.

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): 241-254.

Kemp, Thomas Stainforth (2005). The origin and evolution of mammals.

Kjer, Karl M. and Rodney L. Honeycutt (2007). "Site specific rates of mitochondrial genomes and the phylogeny of eutheria." BMC Evolutionary Biology **7**(1): 8.

Klein, George (1981). "The role of gene dosage and genetic transpositions in carcinogenesis." Nature **294**: 313-318.

Kluge, Arnold G and James S Farris (1969). "Quantitative phyletics and the evolution of anurans." Systematic Biology **18**(1): 1-32.

Koepfli, Klaus-Peter, Benedict Paten, Genome 10K Community of Scientists and Stephen J O'Brien (2015). "The Genome 10K Project: a way forward." Annu. Rev. Anim. Biosci. **3**(1): 57-111.

Kolaczkowski, Bryan and Joseph W. Thornton (2004). "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous." Nature **431**(7011): 980-984.

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): 282-285.

Kumar, Sudhir, Glen Stecher, Michael Suleski and Hedges S Blair (2017). "TimeTree: A resource for timelines, timetrees, and divergence times." Molecular biology and evolution.

Kuraku, Shigehiro, Daisuke Hoshiyama, Kazutaka Katoh, Hiroshi Suga and Takashi Miyata (1999). "Monophyly of Lampreys and Hagfishes Supported by Nuclear DNA–Coded Genes." Journal of Molecular Evolution **49**(6): 729-735.

Lamb, Trevor D, Shaun P Collin and Edward N Pugh (2007). "Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup." Nature Reviews Neuroscience **8**(12): 960-976.

Lartillot, Nicolas (2014). Bayesian automatic control of model complexity. **2015**.

Lartillot, Nicolas, Thomas Lepage and Samuel Blanquart (2009). "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating." Bioinformatics **25**(17): 2286-2288.

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 (2013a). "PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment." Systematic Biology: syt022.

Lartillot, Nicolas, Nicolas Rodriguez, Daniel Stubbs and Jacques Richer (2013b). "PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment." Systematic Biology: syt022.

Lavrov, Dennis V. (2007). "Key transitions in animal evolution: a mitochondrial DNA perspective." Integrative and Comparative Biology **47**(5): 734-743.

Lawson, Andrew RJ, Guy FL Hindley, Tim Forshew, Ruth G Tatevossian, Gabriel A Jamie, Gavin P Kelly, Geoffrey A Neale, Jing Ma, Tania A Jones and David W Ellison (2011). "RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology." Genome research **21**(4): 505-514.

Le, Si Quang and Olivier Gascuel (2008). "An improved general amino acid replacement matrix." Molecular biology and evolution **25**(7): 1307-1320.

Leaché, Adam D. (2009). "Species tree discordance traces to phylogeographic clade boundaries in North American fence lizards (Sceloporus)." Systematic Biology: syp057.

Leventhal, Gabriel E., Huldrych F. Günthard, Sebastian Bonhoeffer and Tanja Stadler (2014). "Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission." Molecular biology and evolution **31**(1): 6-17.

Liò, Pietro and Nick Goldman (1998). "Models of molecular evolution and phylogeny." Genome research **8**(12): 1233-1244.

Little, Elizabeth, Peer Bork and Russell F Doolittle (1994). "Tracing the spread of fibronectin type III domains in bacterial glycohydrolases." Journal of Molecular Evolution **39**(6): 631-643.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland and A. W. D. Larkum (1992). "Substitutional bias confounds inference of cyanelle origins from sequence data." Journal of Molecular Evolution **34**(2): 153-162.

Lockhart, Peter J., A. W. Larkum, M. Steel, Peter J. Waddell and David Penny (1996). "Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis." Proceedings of the National Academy of Sciences **93**(5): 1930-1934.

Long, Manyuan (2000). "A new function evolved from gene fusion." Genome research **10**(11): 1655-1657.

Long, Manyuan (2001). "Evolution of novel genes." Current opinion in genetics & development **11**(6): 673-680.

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): 865-875.

Long, Manyuan and Charles H. Langley (1993). "Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila." science **260**(5104): 91-95.

Loomis, William F. and Douglas W. Smith (1990). "Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison." Proceedings of the National Academy of Sciences **87**(23): 9093-9097.

Lopes, Hedibert Freitas and Mike West (2004). "Bayesian model assessment in factor analysis." Statistica Sinica **14**(1): 41-68.

Lowe, Christopher J., D. Nathaniel Clarke, Daniel M. Medeiros, Daniel S. Rokhsar and John Gerhart (2015). "The deuterostome context of chordate origins." Nature **520**(7548): 456-465.

Luce, R. Duncan and Albert D. Perry (1949). "A method of matrix analysis of group structure." Psychometrika **14**(2): 95-116.

Lynch, Michael (2010). "Evolution of the mutation rate." Trends in Genetics **26**(8): 345-352.

Lyubetsky, Vassily, William H. Piel and Peter F. Stadler (2015). "Molecular Phylogenetics 2014." BioMed Research International **2015**: 2.

Magallón, Susana (2010). "Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms." Systematic Biology **59**(4): 384-399.

Manning, Gerard, David B Whyte, Ricardo Martinez, Tony Hunter and Sucha Sudarsanam (2002). "The protein kinase complement of the human genome." Science **298**(5600): 1912-1934.

Manuel, Michaël (2009). "Early evolution of symmetry and polarity in metazoan body plans." Comptes rendus biologies **332**(2): 184-209.

Marsh, Joseph A and Sarah A Teichmann (2010). "How do proteins gain new domains?" Genome biology **11**(7): 126.

Martindale, Mark Q, John R Finnerty and Jonathan Q Henry (2002). "The Radiata and the evolutionary origins of the bilaterian body plan." Molecular phylogenetics and evolution **24**(3): 358-365.

Martindale, Mark Q. (2005). "The evolution of metazoan axial properties." Nature Reviews Genetics **6**(12): 917-927.

Mayrose, Itay, Nir Friedman and Tal Pupko (2005). "A Gamma mixture model better accounts for among site rate heterogeneity." Bioinformatics **21**(suppl 2): ii151-ii158.

McClintock, Barbara (1950). "The origin and behavior of mutable loci in maize." Proceedings of the National Academy of Sciences **36**(6): 344-355.

McLean, Cory Y., Philip L. Reno, Alex A. Pollen, Abraham I. Bassan, Terence D. Capellini, Catherine Guenther, Vahan B. Indjeian, Xinhong Lim, Douglas B. Menke and Bruce T. Schaar (2011). "Human-specific loss of regulatory DNA and the evolution of human-specific traits." Nature **471**(7337): 216-219.

Mess, Andrea and Anthony M. Carter (2007). "Evolution of the placenta during the early radiation of placental mammals." Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology **148**(4): 769-779.

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): 1087-1092.

Milo, Ron, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii and Uri Alon (2002). "Network motifs: simple building blocks of complex networks." science **298**(5594): 824-827.

Misof, Bernhard, Shanlin Liu, Karen Meusemann, Ralph S. Peters, Alexander Donath, Christoph Mayer, Paul B. Frandsen, Jessica Ware, Tomáš Flouri and Rolf G. Beutel (2014). "Phylogenomics resolves the timing and pattern of insect evolution." science **346**(6210): 763-767.

Moore, Andrew D, Åsa K Björklund, Diana Ekman, Erich Bornberg-Bauer and Arne Elofsson (2008). "Arrangements in the modular evolution of proteins." Trends in Biochemical Sciences **33**(9): 444-451.

Moran, Raymond J, Claire C Morgan and Mary J O'Connell (2015). "A Guide to Phylogenetic Reconstruction Using Heterogeneous Models—A Case Study from the Root of the Placental Mammal Tree." Computation **3**(2): 177-196.

Moran Raymond J., Pathmanathan Jananan, Leigh Robert, Creevey Christopher J., Siu-Ting Karen, Bapteste Eric, McInerney James O. and O'Connell Mary J. (2017). "Major protein-coding innovation by gene remodeling in the animal kingdom. (In preparation)."

Morgan, Claire C, Christopher J Creevey and Mary J O'Connell (2014). "Mitochondrial data are not suitable for resolving placental mammal phylogeny." Mammalian genome **25**(11-12): 636-647.

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): 2145-2156.

Moroz, Leonid L, Kevin M Kocot, Mathew R Citarella, Sohn Dosung, Tigran P Norekian, Inna S Povolotskaya, Anastasia P Grigorenko, Christopher Dailey, Eugene Berezikov and Katherine M Buckley (2014). "The ctenophore genome and the evolutionary origins of neural systems." Nature **510**(7503): 109-114.

Morris, James R., D. Hartl, A. Knoll and Robert Lue (2013). Biology: How life works, WH Freeman/Macmillan.

Muller, Jean, Christopher J. Creevey, Julie D. Thompson, Detlev Arendt and Peer Bork (2010). "AQUA: automated quality improvement for multiple sequence alignments." Bioinformatics **26**(2): 263-265.

Müller, Tobias and Martin Vingron (2000). "Modeling amino acid replacement." Journal of Computational Biology **7**(6): 761-776.

Muñoz-López, Martín and José L. García-Pérez (2010). "DNA transposons: nature and applications in genomics." Current genomics **11**(2): 115.

Murphy, William J., Thomas H. Pringle, Tess A. Crider, Mark S. Springer and Webb Miller (2007). "Using genomic data to unravel the root of the placental mammal phylogeny." Genome research **17**(4): 413-421.

Nagy, Alinda, László Bányai and László Patthy (2011). "Reassessing domain architecture evolution of metazoan proteins: major impact of errors caused by confusing paralogs and epaktologs." Genes **2**(3): 516-561.

Nakagawa, So, Hanako Bai, Toshihiro Sakurai, Yuki Nakaya, Toshihiro Konno, Takayuki Miyazawa, Takashi Gojobori and Kazuhiko Imakawa (2013). "Dynamic evolution of endogenous retrovirus-derived genes expressed in bovine conceptuses during the period of placentation." Genome biology and evolution **5**(2): 296-306.

Nesnidal, Maximilian P., Martin Helmkampf, Iris Bruchhaus and Bernhard Hausdorf (2010). "Compositional heterogeneity and phylogenomic inference of metazoan relationships." Molecular biology and evolution **27**(9): 2095-2104.

Newman, Mark E. J. (2002a). "Assortative mixing in networks." Physical review letters **89**(20): 208701.

Newman, Mark E. J. (2003). "Mixing patterns in networks." Physical Review E **67**(2): 026126.

Newman, Mark EJ (2002b). "Assortative mixing in networks." Physical review letters **89**(20): 208701.

Newton, Michael A. and Adrian E. Raftery (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." Journal of the Royal Statistical Society. Series B (Methodological): 3-48.

Nickle, David C., Laura Heath, Mark A. Jensen, Peter B. Gilbert, James I. Mullins and Sergei L. Kosakovsky Pond (2007). "HIV-specific probabilistic models of protein evolution." PloS one **2**(6): e503.

Nishihara, Hidenori, Shigenori Maruyama and Norihiro Okada (2009). "Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals." Proceedings of the National Academy of Sciences **106**(13): 5235-5240.

Nosenko, Tetyana, Fabian Schreiber, Maja Adamska, Marcin Adamski, Michael Eitel, Jörg Hammel, Manuel Maldonado, Werner E. G. Müller, Michael Nickel and Bernd Schierwater (2013). "Deep metazoan phylogeny: when different genes tell different stories." Molecular phylogenetics and evolution **67**(1): 223-233.

Nylander, J. A. A. (2004). "MrModeltest v2. Program distributed by the author." Evolutionary Biology Centre, Uppsala University **2**.

O'Leary, Maureen A., Jonathan I. Bloch, John J. Flynn, Timothy J. Gaudin, Andres Giallombardo, Norberto P. Giannini, Suzann L. Goldberg, Brian P. Kraatz, Zhe-Xi

Luo and Jin Meng (2013). "The placental mammal ancestor and the post–K-Pg radiation of placentals." science **339**(6120): 662-667.

Ohno, Susumu (1970). The creation of a new gene from a redundant duplicate of an old gene. Evolution by gene duplication, Springer**:** 71-82.

Ohno, Susumu (2013). Evolution by gene duplication, Springer Science & Business Media.

Opsahl, Tore, Filip Agneessens and John Skvoretz (2010). "Node centrality in weighted networks: Generalizing degree and shortest paths." Social networks **32**(3): 245-251.

Opsahl, Tore and Pietro Panzarasa (2009). "Clustering in weighted networks." Social networks **31**(2): 155-163.

Ota, Kinya G., Satoko Fujimoto, Yasuhiro Oisi and Shigeru Kuratani (2011). "Identification of vertebra-like elements and their possible differentiation from sclerotomes in the hagfish." Nature communications **2**: 373.

Pamilo, Pekka and Masatoshi Nei (1988). "Relationships between gene trees and species trees." Molecular biology and evolution **5**(5): 568-583.

Patthy, László (2003). "Modular assembly of genes and the evolution of new functions." Genetica **118**(2): 217-231.

Pedersen, Jakob Skou, Gill Bejerano, Adam Siepel, Kate Rosenbloom, Kerstin Lindblad-Toh, Eric S. Lander, Jim Kent, Webb Miller and David Haussler (2006). "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput Biol **2**(4): e33.

Philippe, Hervé, Henner Brinkmann, Richard R. Copley, Leonid L. Moroz, Hiroaki Nakano, Albert J. Poustka, Andreas Wallberg, Kevin J. Peterson and Maximilian J. Telford (2011a). "Acoelomorph flatworms are deuterostomes related to Xenoturbella." Nature **470**(7333): 255-258.

Philippe, Herve, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J Littlewood, Michael Manuel, Gert Wörheide and Denis Baurain (2011b). "Resolving difficult phylogenetic questions: why more sequences are not enough." PLoS-Biology **9**(3): 402.

Philippe, Hervé and Christophe J. Douady (2003). "Horizontal gene transfer and phylogenetics." Current opinion in microbiology **6**(5): 498-505.

Philippe, Hervé and Agnès Germot (2000). "Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution." Molecular biology and evolution **17**(5): 830-834.

Philippe, Hervé and Philippe Lopez (2001). "On the conservation of protein sequences in evolution." Trends in Biochemical Sciences **7**(26): 414-416.

Philippe, Hervé, Elizabeth A. Snell, Eric Bapteste, Philippe Lopez, Peter W. H. Holland and Didier Casane (2004). "Phylogenomics of eukaryotes: impact of missing data on large alignments." Molecular biology and evolution **21**(9): 1740-1752.

Philippe, Hervé, Yan Zhou, Henner Brinkmann, Nicolas Rodrigue and Frédéric Delsuc (2005). "Heterotachy and long-branch attraction in phylogenetics." BMC Evolutionary Biology **5**(1): 50.

Phillips, Matthew J. and David Penny (2003). "The root of the mammalian tree inferred from whole mitochondrial genomes." Molecular phylogenetics and evolution **28**(2): 171-185.

Pick, K. S., Hervé Philippe, F. Schreiber, D. Erpenbeck, D. J. Jackson, P. Wrede, M. Wiens, Alexandre Alié, B. Morgenstern and Michaël Manuel (2010). "Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships." Molecular biology and evolution **27**(9): 1983-1987.

Porter, Mason A., Jukka-Pekka Onnela and Peter J. Mucha (2009). "Communities in networks." Notices of the AMS **56**(9): 1082-1097.

Posada, David (2003). "Selecting models of evolution." The phylogenetic handbook. A practical approach to DNA and protein phylogeny. Cambridge University Press, Cambridge.

Posada, David and Keith A. Crandall (1998). "Modeltest: testing the model of DNA substitution." Bioinformatics **14**(9): 817-818.

Posada, David and Keith A. Crandall (2002). "The effect of recombination on the accuracy of phylogeny estimation." Journal of Molecular Evolution **54**(3): 396-402.

Powell, Sean, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, Thomas Rattei, Ivica Letunic and Tobias Doerks (2012). "eggNOG v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges." Nucleic acids research **40**(D1): D284-D289.

Prasad, Arjun B., Marc W. Allard and Eric D. Green (2008). "Confirming the phylogeny of mammals by use of large comparative sequence data sets." Molecular biology and evolution **25**(9): 1795-1808.

Proulx, Stephen R, Daniel EL Promislow and Patrick C Phillips (2005). "Network thinking in ecology and evolution." Trends in ecology & evolution **20**(6): 345-353.

Pyron, R. Alexander (2015). "Post-molecular systematics and the future of phylogenetics." Trends in ecology & evolution.

Ranwez, Vincent, Fr√©d√©ric Delsuc, Sylvie Ranwez, Khalid Belkhir, Marie-Ka Tilak and Emmanuel JP Douzery (2007). "OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics." BMC Evolutionary Biology **7**(1): 241.

Rawlings, Neil D and Alan J Barrett (1994). "[2] Families of serine peptidases." Methods in enzymology **244**: 19-61.

Reyes, Aurelio, Carmela Gissi, Francois Catzeflis, Eviatar Nevo, Graziano Pesole and Cecilia Saccone (2004). "Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods." Molecular biology and evolution **21**(2): 397-403.

Rhymer, Judith M. and Daniel Simberloff (1996). "Extinction by hybridization and introgression." Annual Review of Ecology and Systematics: 83-109.

Rivera, Corban G, Rachit Vakil and Joel S Bader (2010). "NeMo: network module identification in Cytoscape." BMC bioinformatics **11**(1): S61.

Rogozin, Igor B., Yuri I. Wolf, Liran Carmel and Eugene V. Koonin (2007). "Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements." Molecular biology and evolution **24**(4): 1080-1090.

Romiguier, Jonathan, Vincent Ranwez, Fr√©d√©ric Delsuc, Nicolas Galtier and Emmanuel J. P. Douzery (2013). "Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals." Molecular biology and evolution **30**(9): 2134-2144.

Romiguier, Jonathan, Vincent Ranwez, Emmanuel JP Douzery and Nicolas Galtier (2010). "Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes." Genome research **20**(8): 1001-1009.

Ronquist, Fredrik and John P Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-1574.

Ronquist, Fredrik, Seraina Klopfstein, Lars Vilhelmsen, Susanne Schulmeister, Debra L. Murray and Alexandr P. Rasnitsyn (2012). "A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera." Systematic Biology **61**(6): 973-999.

Roth, Alexander CJ, Gaston H Gonnet and Christophe Dessimoz (2008). "Algorithm of OMA for large-scale orthology inference." BMC bioinformatics **9**(1): 518.

Rual, Jean-François, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze and Nono Ayivi-Guedehoussou (2005). "Towards a proteome-scale map of the human protein–protein interaction network." Nature **437**(7062): 1173-1178.

Sansom, Ivan James and Jason Philip Downs "Early evolution of vertebrate skeletal tissues and cellular interactions, and the canalization of skeletal development."

Schlötterer, Christian (2015). "Genes from scratch–the evolutionary fate of de novo genes." Trends in Genetics **31**(4): 215-219.

Schmidt, Heiko A., Korbinian Strimmer, Martin Vingron and Arndt von Haeseler (2002). "TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing." Bioinformatics **18**(3): 502-504.

Schnell, Jason R, H Jane Dyson and Peter E Wright (2004). "Structure, dynamics, and catalytic function of dihydrofolate reductase." Annu. Rev. Biophys. Biomol. Struct. **33**: 119-140.

Shafer, Glenn (1976). A mathematical theory of evidence, Princeton university press Princeton.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski and Trey Ideker (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research **13**(11): 2498-2504.

Sheen, Patricia, David Couvin, Louis Grandjean, Mirko Zimic, Maria Dominguez, Giannina Luna, Robert H. Gilman, Nalin Rastogi and D. A. Moore (2013). "Genetic diversity of Mycobacterium tuberculosis in Peru and exploration of phylogenetic associations with drug resistance." PloS one **8**(6): e65873.

Shimada, Atsuko, Toru Kawanishi, Takuya Kaneko, Hiroki Yoshihara, Tohru Yano, Keiji Inohaya, Masato Kinoshita, Yasuhiro Kamei, Koji Tamura and Hiroyuki Takeda (2013). "Trunk exoskeleton in teleosts is mesodermal in origin." Nature communications **4**: 1639.

Shinkai, Yoichi, Kong-Peng Lam, Eugene M Oltz, Valerie Stewart, Monica Mendelsohn, Jean Charron, Milton Datta, Faith Young, Alan M Stall and Frederick W Alt (1992). "RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V (D) J rearrangement." Cell **68**(5): 855-867.

Sidow, Arend (1992). "Diversification of the Wnt gene family on the ancestral lineage of vertebrates." Proceedings of the National Academy of Sciences **89**(11): 5098-5102.

Sievers, Fabian and Desmond G. Higgins (2014). "Clustal Omega." Current Protocols in Bioinformatics: 3.13. 11-13.13. 16.

Simakov, Oleg, Ferdinand Marletaz, Sung-Jin Cho, Eric Edsinger-Gonzales, Paul Havlak, Uffe Hellsten, Dian-Han Kuo, Tomas Larsson, Jie Lv and Detlev Arendt (2013). "Insights into bilaterian evolution from three spiralian genomes." Nature **493**(7433): 526-531.

Simpson, George Gaylord (1944). Tempo and mode in evolution, Columbia University Press.

Smith, Chris R, Sara Helms Cahan, Carsten Kemena, Seán G Brady, Wei Yang, Erich Bornberg-Bauer, Ti Eriksson, Juergen Gadau, Martin Helmkampf and Dietrich Gotzek (2015). "How do genomes create novel phenotypes? Insights from the loss of the worker caste in ant social parasites." Molecular biology and evolution **32**(11): 2919-2931.

Soda, Manabu, Young Lim Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, Hideki Watanabe, Kentaro Kurashina and Hisashi Hatanaka (2007). "Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer." Nature **448**(7153): 561-566.

Soisson, Stephen M, Anjaruwee S Nimnual, Marc Uy, Dafna Bar-Sagi and John Kuriyan (1998). "Crystal structure of the Dbl and pleckstrin homology domains from the human Son of sevenless protein." Cell **95**(2): 259-268.

Soller, Maria Johansson, Margareth Isaksson, Peter Elfving, Wolfgang Soller, Rolf Lundgren and Ioannis Panagopoulos (2006). "Confirmation of the high frequency of the TMPRSS2/ERG fusion gene in prostate cancer." Genes, Chromosomes and Cancer **45**(7): 717-719.

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**(5): 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: 201211733.

Speakman, John R. (2001). "The evolution of flight and echolocation in bats: another leap in the dark." Mammal Review **31**(2): 111-130.

Springer, Mark S., Michael J. Stanhope, Ole Madsen and Wilfried W. de Jong (2004). "Molecules consolidate the placental mammal tree." Trends in ecology & evolution **19**(8): 430-438.

Stamatakis, Alexandros (2006a). Phylogenetic models of rate heterogeneity: a high performance computing perspective. Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International, IEEE.

Stamatakis, Alexandros (2006b). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688-2690.

Strimmer, Korbinian and Arndt Von Haeseler (1997). "Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment." Proceedings of the National Academy of Sciences **94**(13): 6815-6819.

Swofford, David L, Gary J Olsen, Peter J Waddell and David M Hillis (1996). Molecular systematics, Sinauer Associates Sunderland, MA.

Tarver, James E, Mario Dos Reis, Siavash Mirarab, Raymond J Moran, Sean Parker, Joseph E O'Reilly, Benjamin L King, Mary J O'Connell, Robert J Asher and Tandy Warnow (2016). "The interrelationships of placental mammals and the limits of phylogenetic inference." Genome biology and evolution **8**(2): 330-344.

Tautz, Diethard and Tomislav Domazet-Lošo (2011). "The evolutionary origin of orphan genes." Nature Reviews Genetics **12**(10): 692-702.

Tavaré, Simon (1986a). "Some probabilistic and statistical problems in the analysis of DNA sequences." Lectures on mathematics in the life sciences **17**(2): 57-86.

Tavaré, Simon (1986b). "Some probabilistic and statistical problems in the analysis of DNA sequences." Lectures on mathematics in the life sciences **17**: 57-86.

Technau, Ulrich and Corinna B. Scholz (2003). "Origin and evolution of endoderm and mesoderm." International Journal of Developmental Biology **47**(7/8): 531-540.

Teeling, Emma C. and S. Blair Hedges (2013). "Making the impossible possible: rooting the tree of placental mammals." Molecular biology and evolution **30**(9): 1999-2000.

Thomson, Timothy M, Juan José Lozano, Noureddine Loukili, Roberto Carrió, Florenci Serras, Bru Cormand, Marta Valeri, Víctor M Díaz, Josep Abril and Moisés Burset (2000). "Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene." Genome research **10**(11): 1743-1756.

Tian, Xin, Joel Gautron, Philippe Monget and Géraldine Pascal (2010). "What makes an egg unique? Clues from evolutionary scenarios of egg-specific genes." Biology of reproduction **83**(6): 893-900.

Todd, John A, Hans Acha-Orbea, John I Bell, Nelson Chao, Zdenka Fronek, Chaim O Jacob, Michael McDermott, Animesh A Sinha, Luika Timmerman and Lawrence Steinman (1988). "A molecular basis for MHC class II-associated autoimmunity." science **240**(4855): 1003-1010.

Turner, John (1984). "Why we need evolution by jerks."

Urhammer, Søen A, Marianne Fridberg, Torben Hansen, Søren K Rasmussen, Ann Merete Møller, Jesper O Clausen and Oluf Pedersen (1997). "A prevalent amino acid polymorphism at codon 98 in the hepatocyte nuclear factor-1α gene is associated with reduced serum C-peptide and insulin responses to an oral glucose challenge." Diabetes **46**(5): 912-916.

Venter, J Craig, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans and Robert A Holt (2001). "The sequence of the human genome." science **291**(5507): 1304-1351.

Wang, Wen, Jianming Zhang, Carlos Alvarez, Ana Llopart and Manyuan Long (2000). "The origin of the jingwei gene and the complex modular structure of its parental gene, yellow emperor, in drosophila melanogaster." Molecular biology and evolution **17**(9): 1294-1301.

Watts, Duncan J. and Steven H. Strogatz (1998). "Collective dynamics of ‘small-world’networks." Nature **393**(6684): 440-442.

Weisstein, Eric W (2004). "Bonferroni correction."

Whelan, Simon and Nick Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." Molecular biology and evolution **18**(5): 691-699.

Wilson, Don E. and DeeAnn M. Reeder (2005). Mammal species of the world: a taxonomic and geographic reference, JHU Press.

Wu, Chia-Chin, Kalpana Kannan, Steven Lin, Laising Yen and Aleksandar Milosavljevic (2013). "Identification of cancer fusion drivers using network fusion centrality." Bioinformatics: btt131.

Wu, Guoyao, Yun-Zhong Fang, Sheng Yang, Joanne R Lupton and Nancy D Turner (2004). "Glutathione metabolism and its implications for health." The Journal of nutrition **134**(3): 489-492.

Yang, Ziheng (1996). "Among-site rate variation and its impact on phylogenetic analyses." Trends in ecology & evolution **11**(9): 367-372.

Yang, Ziheng and Bruce Rannala (2012). "Molecular phylogenetics: principles and practice." Nat Rev Genet **13**(5): 303-314.

Yu, Yan, Alan J. Harris, Christopher Blair and Xingjin He (2015). "RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography." Molecular phylogenetics and evolution **87**: 46-49.

Zhang, Jianzhi (2003). "Evolution by gene duplication: an update." Trends in ecology & evolution **18**(6): 292-298.

Zhou, Qi and Wen Wang (2008). "On the origin and evolution of new genes—a genomic and experimental perspective." Journal of Genetics and Genomics **35**(11): 639-648.

Chapter 7: Publications

Moran, R. J., C. C. Morgan and M. J. O'Connell (2015). A Guide to Phylogenetic Reconstruction Using Heterogeneous Models—A Case Study from the Root of the Placental Mammal Tree. Computation **3**(2): 177-196

Tarver, J. E., M. Dos Reis, S. Mirarab, R. J. Moran, S. Parker, J. E. O’Reilly, B. L. King, M. J. O’Connell, R. J. Asher and T. Warnow (2016). The interrelationships of placental mammals and the limits of phylogenetic inference. Genome biology and evolution 8(2): 330-344.

Appendix:

Attached here is a USB flash drive containing a tar.gz file. Within this tar.gz file is the Appendix. For each results chapter of this thesis there is a folder called Chapter_num (where num is the chapter number). Within each Chapter folder is all the scripts, data, results and supplementary information needed for this thesis.