Phylogenomic supertrees: integration of maximal data and assessment of input tree shape.

A thesis submitted to the National University of Ireland for the Degree of
Doctor of Philosophy

Presented by:
Thérèse A. Holton
Department of Biology,
NUI Maynooth,
Maynooth,
Co. Kildare, Ireland.

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

February 2011

Supervisor: Dr. Davide Pisani B.Sc., Ph.D. (Bristol)
Head of Department: Professor Kay Ohlendieck, Dip.Biol., M.Sc. (Konstanz), Ph.D.

Table of Contents

For my parents and sisters

Acknowledgements

There are many people who have helped me in enumerable ways over the past three years; I truly am so privileged to have had the support of so many wonderful people during my time in Maynooth.

First and foremost, I wish to express my sincere gratitude to my supervisor Dr. Davide Pisani for all the advice, guidance and opportunities that he has provided me with. I am so lucky to have learned from such a knowledgeable scientist, who throughout maintained such an amiable demeanour.

To everyone in the Bioinformatics Unit, a big thank you to each and every one of you for the various ways you have helped me out. Both from a professional and personal perspective, you are all incredible people and it has been a privilege to work alongside you. I would particularly like to thank past lab members Dr. Angela McCann, Dr. Victoria Svinti and Dr. Fergal Martin for their warm welcome and much needed assistance when I started my PhD. A special thank you to Mr. Brian Daly for all his help (and patience!) with the computational resources available to the lab and the countless issues he resolved to make my research run more smoothly. I would also like to express my sincere appreciation to Dr. James McInerney for his insightful advice and comments, which have always been so helpful and beneficial to me.

My research would not have been possible without funding The Irish Research Council for Science, Engineering & Technology and Science Foundation Ireland, and the computational resources of the Irish Centre for High End Computing (ICHEC) and The High Performance Computing (HPC) Facility at NUI Maynooth. I would particularly like to thank Dr. Simon Wong of ICHEC and Dr. Vanush "Misha" Paturyan of HPC NUI Maynooth for their invaluable technical support and assistance.

To my friends, both in Maynooth and further afield, thank you for your support. To Carla, Ruth and Theresa thank you for always being there, be it for a chat or for diners club, I am extremely grateful for all you have done for me. To Susan, Ciara, Clare, Emma and Claire, thank you all for your advice, support and understanding over the last few years, I couldn't have done it without all of you. Finally to Fiona, Clodagh and Nicola, thank you for putting up with my student ways for so long!

To Ronan, I can't even begin to thank you for how wonderful you have been over the last few months. I am so grateful for all the love and encouragement you have shown me. Thank you so much for always being there, I don't know how you've put up with me at times!

To my parents, Mary and Liam, words can't express how much I appreciate all that you have done for me over the years. Thank you for your constant encouragement and support throughout, but especially in these last few months. I am so lucky to have such loving and supportive parents. To my fantastic sisters Claire, Mairead, Eilish and Maeve thank you for all the tea, laughter and chats. You each have helped me so much and for that I am eternally grateful. Katie, you are never far from my thoughts, thank you for being a constant inspiration.

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed:__________________________________

Thérèse Anne Holton

Abbreviations

AIC	Akaike Information Criterion
ANOVA	Analysis Of Variance
APE	Analysis of Phylogenetics and Evolution
BF	Bayes Factor
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
BLASTP	protein BLAST
BLOSUM	Blocks of amino acid Substitution Matrix
BS	Bootstrap Support
DNA	Deoxyribonucleic Acid
E value	Expect value
EPT	Equiprobable Types
ERM	Equal Rate Markov
EST	Expressed Sequence Tag
FSA	Fast Statistical Alignment
GG	Greenhouse-Geisser correction
GTP	Gene Tree Parsimony

GTP-PTP	Gene Tree Parsimony Permutation Tail Probability Test
GTR	General Time Reversible model
HF	Huynh-Feldt correction
JC69	Jukes and Cantor model (1969)
K2P	Kimura's Two- Parameter model
LBA	Long Branch Attraction
LGT	Lateral Gene Transfer
LRT	Likelihood Ratio Test
MCL	Markov Clustering Algorithm
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MP	Maximum Parsimony
MPT	Most Parsimonious Tree
MRP	Matrix Representation with Parsimony
NJ	Neighbor Joining
OTU	Operational Taxonomic Unit
PAUP*	Phylogenetic Analysis Using Parsimony (* and other methods)
PDA	Proportion To Distinguishable Arrangements

PhyML	PHYlogenies by Maximum Likelihood
PTP	Permutation Tail Probability
RAS	Rhizaria Alveolates and Stramenopiles
RAxML	Randomised Axelerated Maximum Likelihood
RM-ANOVA	Repeated Measures Analysis Of Variance
RNA	Ribonucleic Acid
rRNA	ribosomal Ribonucleic Acid
SAR	Stramenopiles Alveolates and Rhizaria
T-PTP	topology dependent PTP
TBM	Tree Balance Metric
TRM	Tree Reconstruction Method
Tukey HSD	Tukey’s honestly significant difference
YAPTP	Yet Another Permutation Tail Probability

Index of Figures

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

Index of Tables

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

Abstract

In this thesis, three distinct studies were carried out to investigate various aspects pertaining to the properties and applicability of phylogenomic data used in a supertree context. While the availability of genomic scale data is rapidly diminishing as a problem in the field of phylogenomics, there is now a greater need for an appropriate means of analysing such data.

Supertrees have emerged as a useful approach in handling large data sets and have been shown to work extremely well in a phylogenomic context (e.g. Creevey et al., 2004, Fitzpatrick et al., 2006, Pisani et al., 2007). While supertree studies do generally sample significantly more genomic data than their supermatrix counterpart, much of the genome, which has evolved in the light of gene duplication, is not considered in this method. Further to this, typically, in the supertree approach complete genomic data is exclusively used, which can result in a very limited taxon sampling compared to alternative approaches that use expressed sequence tag (EST) data. Here, in attempt to address these shortcomings, the viability of integrating genes with a history of duplication in the supertree approach, as well as the extent to which a combined data set of complete and partial genomes (ESTs) can be used to increase taxon sampling in this context, is investigated.

Additionally, in this thesis, the effect of input tree shape biases is assessed. It has been shown previously that some commonly used supertree methods are biased with respect to the tree shape they produce (Wilkinson et al., 2005). However, since some supertree methods (e.g. matrix representation with parsimony; MRP) have an inherent phylogenetic component, the observed shape predispositions of these supertree methods may be attributed to such methodological elements. As such, here the effective shape bias of various phylogenetic methods is assessed using a phylogenomic data set.

Chapter 1: Introduction

Lamarck is credited with the introduction of the first true evolutionary tree (see Figure 1.1; see also, for example, McInerney et al., 2008), however, it was Charles Darwin's (1859) celebrated phylogenetic tree that captivated public interest. This drawing (Figure 1.2), which is the only figure in "*On the Origin of Species*", perfectly captured the meaning of his theory of natural selection, and spurred others, like Haeckel, to investigate the evolutionary relationships among living organisms. This ultimately spawned an active area of research that has extended across the last two centuries: phylogenetics. The impact of this theory has not been confined to the classrooms of science and has, and continues, to fascinate and incite the readers of systematic journals and popular science alike. Since Darwin's time, phylogenetics in itself has experienced a rather interesting evolutionary history. It has seen a transition from morphology to molecules and from genes to genomes (Eisen, 1998). In this chapter I will discuss the concepts that have delimited these transitions and outline the common methods employed by each era of molecular phylogenetics.

1.1 "Molecules as documents of evolutionary history"

The above quotation is the title of the seminal 1965 publication of Zuckerkandl and Pauling, which was instrumental in highlighting an important avenue being explored by biologists at that time: the use of molecular data as a means of understanding historical biology. In their communication, they explore the use of alternative molecules (such as episemantic molecules, e.g. ATP, and asemantic molecules, e.g. vitamins), identifying semantides, a class of molecules that encompasses DNA, RNA and

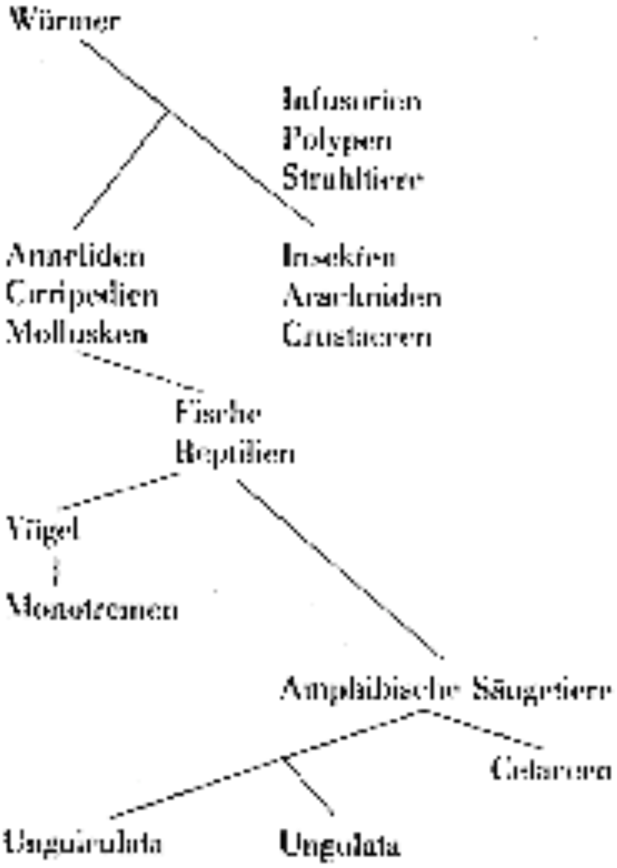


Figure 1.1 Lamarck's evolutionary tree.

This tree, which appears in book "Philosophie Zoologique" (Lamarck, 1809), depicts Lamarck's understanding of how the animals evolved. This is considered the first example of an evolutionary tree.

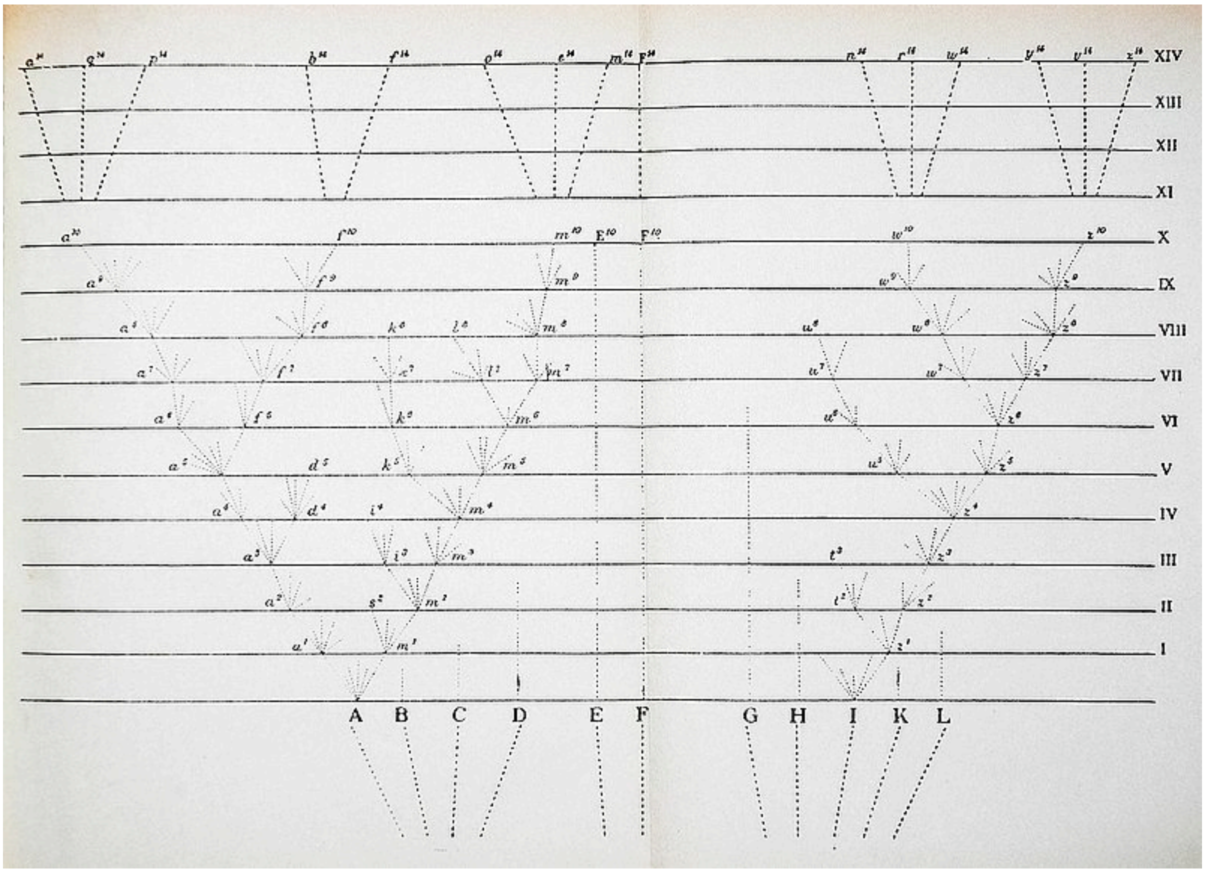

a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} w^{10} z^{10}
X
a^{9} f^{9} m^{9} w^{9} z^{9}
IX
a^{8} f^{8} k^{8} l^{8} m^{8} u^{8} w^{8} z^{8}
VIII
a^{7} f^{7} k^{7} l^{7} m^{7} u^{7} w^{7} z^{7}
VII
a^{6} f^{6} k^{6} m^{6} u^{6} z^{6}
VI
a^{5} d^{5} k^{5} m^{5} u^{5} z^{5}
V
a^{4} d^{4} i^{4} m^{4} z^{4}
IV
a^{3} i^{3} m^{3} t^{3} z^{3}
III
a^{2} s^{2} m^{2} t^{2} z^{2}
II
a^{1} m^{1} z^{1}
I
A B C D E F G H I K L

Figure 1.2 Darwin’s Tree of Life.

This tree appears in the book “On the Origin of Species” (Darwin, 1859) and is the only illustration to feature in the seminal publication.

polypeptides, to be the unequivocal molecular informant of evolutionary histories (Zuckerkandl and Pauling, 1965). This work was symbolic of a new era where homologous genetic sequences were used to trace evolutionary events. Many other important publications, including that of Fitch (1967) featuring one of the first molecular phylogenies, and that of Kimura (1969) focusing on variable mutation rate, as well as the methodologically important publication of Edwards and Cavalli-Sforza (1964), helped establish a framework for molecular evolution. Such studies provided a formative foundation, upon which modern molecular phylogenetics has been built.

1.1.1 Homology to alignment

Homology can be defined as the "special" similarity between characters that have descended, typically with divergence, from a common ancestor (Fitch, 2000). A classic example from traditional morphological studies would be that of the humerus bone, which is present in the human arm, in the bird wing and in the anterior leg of, for example, the cat. Homology was introduced in pre-Darwinian times by Sir Richard Owen (1843), where he defined it as "[the] *part or organ in one animal which has the same function as another part or organ in a different animal*". Upon this concept the basis of modern phylogenetics has been built. In order to infer a molecular phylogeny, a set of homologous proteins or nucleotides is required. As such, each molecular phylogenetic study commences with the identification of homologous genetic sequences, often referred to as gene (or protein) families.

Currently, in the field of bioinformatics, the most popular method of determining homology between genetic sequences is through the use of the Basic Local Alignment

Search Tool (Altschul et al., 1997). Using this method, a database of genetic sequences is constructed or utilised, with the potential homology of constituent sequences being evaluated by a measure of whether that sequence is expected to be returned in a search by chance, in a database of that particular size. This measure, known as an expect value or E-value, can be used as a proxy for homology, where by only sequences within the bounds of the given E-value are considered potentially homologous. The use of BLAST and E-values commonly features as a constituent stage in other homology assignment methodologies such as MCL (Enright et al., 2002), BLASTClust (Dondoshansky and Wolf, 2000) and reciprocal BLAST strategies.

It is important to note that there are three distinct subtypes of homology defined; these are paralogy, orthology and xenology. Paralogy occurs when homology arises due to gene duplication; orthology, when homology arises due to a speciation event and xenology, when homology arises due to interspecies transfer of genetic material (Fitch, 2000). Currently, BLAST does not distinguish between these classes of homology, leading to the requirement of downstream methods of detecting these events (see for example Cotton, 2005).

In order to construct a sequence-based phylogeny, homologous (i.e. corresponding) sites in homologous sequences need to be compared. To achieve this, alignment of the homologous sequences is carried out, in a procedure that is commonly referred to as multiple sequence alignment. Available alignment software implementations include, for example, Clustal, Muscle, PRANK and FSA (Thompson et al., 1994, Edgar, 2004, Löytynoja and Goldman, 2008, Bradley et al., 2009 respectively, for a recent review see Kemena and Notredame, 2009), which are based on different algorithms and thus possess different strengths and weaknesses.

ClustalW (Thompson et al., 1994) persists as the most widely used multiple sequence alignment method due to its long established reputation and low computational cost. However, the accuracy of this method (Edgar and Batzoglou, 2006), particularly for longer sequences, and its handling of indels (Golubchik et al., 2007, Löytynoja and Goldman, 2008), have been criticised. Accordingly, many modern methods addressing these issues have been developed. A contemporary method of note, FSA, provides an improvement on speed and accuracy for larger sequences (Bradley et al., 2009). Another recent method, PRANK (Löytynoja and Goldman, 2008) represents a departure from traditional alignment approaches, in that it attempts to produce alignments that more accurately reflect the evolutionary history of the considered sequences. To do this, the algorithm treats insertions and deletions as discrete events and phylogenetic information is used to determine which of these events is responsible for observed gaps in the alignment.

Alignment software accounts for positional homology, providing a configuration that best explains the biological likeness of the nucleotides or amino acids of each sequence, at each site. As part of this process, the software may determine it necessary to insert what is commonly known as a gap character (represented in the sequence by a '-') at a given site, to uphold the parallel confirmation of sites downstream. Additionally, mutations and insertions in the sequence are accounted for by means of an inbuilt weighting scheme in the alignment algorithm, which can be defined by the user to tailor specifically to the demands of each study. Once the alignment is complete, curation of the resulting sequence configurations is often necessitated to resolve misaligned regions. Manual curation of alignments is commonplace, however, for larger genomic scales studies (such as those discussed in this thesis), an automated approach, for example, the

Gblocks software (Castresana, 2000) is generally used. When a suitably high standard of data quality is achieved, one of the numerous tree reconstruction methods is applied to the aligned sequences. In the following section I will introduce the main procedures and standards used for the inference of phylogenetic trees. The focus here is on methods that I have used during my PhD, therefore, outdated methods, such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean; see Felsenstein 2004), will not be discussed.

1.1.2 Phylogenetic tree reconstruction

1.1.2.1 Maximum parsimony

Philosophically, the concept of parsimony is derived from a principle introduced by the 14th century monk William of Ockham. This principle, known as Ockham's razor, dictates: "Plurality should not be posited without necessity". More plainly, this theory confers precedence to simplicity, where, of two competing theories, the simplest explanation should be the favoured one. Edwards and Cavalli-Sforza (1963) are attributed with the inception of parsimony in an evolutionary context (this is often mistakenly accredited to Willi Hennig and Walter Zimmerman, however, parsimony in this sense is more similar to compatibility or clique analyses; see Estabrook et al., 1977, Meacham and Estabrook, 1985), when they asserted that the tree that is arrived upon by the minimum amount of evolution is the most acceptable, however, application of their minimum evolution algorithm was limited to genetic frequency data.

Use of parsimony in the context of character-based phylogenetics is ascribed to Camin and Sokal (1965), who first defined the term, in addition to algorithms to estimate the number of evolutionary changes and to perform tree search and construction. In a phylogenetic context, the parsimony criterion is applied through the search for the tree that minimises the number of character transformations across all sites, i.e. the most parsimonious tree. In this way, the number of substitutions between character states implied by each tree is calculated over the sum of all characters. There are numerous variations of the parsimony algorithm defined, including the unidirectional Wagner parsimony (Kluge and Farris, 1969) and Fitch algorithm (Fitch, 1971), and the bidirectional Sankoff algorithm (Sankoff, 1975, Sankoff and Rousseau, 1975, see Felsenstein, 2004 for a more exhaustive list and discussion).

Although maximum parsimony (MP) was the method of choice in the 1970s and 1980s, it has been shown to have several flaws. Felsenstein (1978), in a landmark paper showed that parsimony is statistically inconsistent, that is, under certain conditions, i.e. when rates of substitution are highly heterogeneous in neighbouring branches, with the accumulation of more data, the estimation arrives upon the wrong tree with increasingly high support. This concept is more commonly referred to as long branch attraction (LBA). LBA occurs because parsimony systematically minimises branch lengths and, as such, does not accurately reflect what is observed in real data (Zhang and Nei, 1997, Steel and Penny, 2000).

The consequence of these shortcomings is that MP does not perform as well as other methods (Sourdis and Nei, 1988, Tateno et al., 1994). Despite this, parsimony maintains the support of ardent proponents (Farris et al., 1970, Sober, 1988, Kolaczkowski and Thornton, 2004) and even today, persists as a popular method of tree inference. Moreover, in the case of morphological data analysis, MP remains the methodological standard, despite the availability of a specific likelihood model (Lewis, 2001) and accompanying software implementations (in MrBayes; Huelsenbeck and

Ronquist, 2001, and in RAxML; Stamatakis et al., 2005) to account for data of this nature.

1.1.2.2 Distance matrix methods

Distance matrix methods can trace their origin to the work of Sokal and Sneath (1963) who first introduced classic phenetic methods such as UPGMA. However, modern distance methods, for example, least squares, were introduced by Cavalli-Sforza and Edwards (1967), and Fitch and Margoliash (1967). All distance methods calculate the distance between each pair of sequences in an alignment, generate a matrix of pairwise distances and determine the tree that most accurately reflects the computed distances (Felsenstein, 2004). To allow for a more accurate representation of reality, models that incorporate various parameters are generally used to measure the distance between sequences. Indeed, these models of sequence evolution (see below; Section 1.1.2.3) are also employed in probabilistic tree inference methods, like maximum likelihood and Bayesian inference (see below).

1.1.2.3 Models of sequence evolution

The most basic model of DNA sequence evolution is the Jukes and Cantor model (1969). Under this model, the probability of change to any given character state (i.e. A, C, G, T), at any given site in the sequence, is equal. The formula for estimating distance is as follows:

$$D = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D_S\right) \qquad [1.1]$$

where D is the distance between two sequences, ln (the natural logarithm) corrects for superimposed substitutions and D_S is the amount of nucleotides that differ between the two sequences.

The simplicity of the Jukes and Cantor (1969) model (JC69) means that it is a poor estimator of how DNA sequences actually evolve. Accordingly, many subsequent models have attempted to incorporate more realistic parameters to estimate change from one state to another, in a process that is known as transition probability in mathematics (Felsenstein, 2004). Kimura's two-parameter (K2P) model (Kimura, 1980) represented the first improvement over JC69. This model takes the transition-transversion bias of a set of sequences into consideration when estimating distances by assigning a higher probability to the occurrence of transitions in comparison to transversions. Another important improvement was the introduction of general time reversible (GTR) model (Lanave et al., 1984, Rodriguez et al., 1990, see also Yang, 1994a), under which base frequencies are free to vary and all substitutions are assigned a different rate. The substitution matrix in this model is symmetric, allowing it to be time reversible. Consequently, there is no direction of evolution (substitutions) assumed between sequences.

Equally, there are various models of protein evolution described. The majority of these models are based upon empirical data, and include the simple Dayhoff (Dayhoff et al., 1978) and BLOSUM (Henikoff and Henikoff, 1992) models. Other empirical models of protein evolution such as JTT (Jones et al., 1992), WAG (Whelan and Goldman, 2001) and the recently described LG model (Le and Gascuel, 2008), are time reversible and so are generally referred to as empirical GTR models. Recently, the use of mechanistic amino acid GTR models (i.e. GTR models in which the transition probabilities are inferred directly from the data) has become possible in a Bayesian framework (e.g. using MrBayes 3.0; Ronquist and Huelsenbeck, 2003). The emergence of Bayesian phylogenetics (see below, Section 1.1.2.6) has also allowed complex heterogeneous models, such as the CAT model (Lartillot and Philippe, 2004), to be widely used in analyses performed at the amino acid level. Under the CAT model, the dimensionality of the model in itself is a free parameter, therefore, allowing each site to be assigned to the category (of a number of distinct categories) that best describes its substitution rate.

1.1.2.4 Neighbor joining

The neighbor joining (NJ) method, introduced by Saitou and Nei (1987), can be considered an approximation of Nei's minimum evolution method (Kidd and Sgaramella-Zonta, 1971, Rzhetsky and Nei, 1992, 1993), which is not to be confused with the minimum evolution method of Edwards and Cavalli-Sforza (1963). Under Nei's minimum evolution method, alternative tree topologies are fit to the data and branch lengths are subsequently calculated using the ordinary least square method, with the selected tree being the one with the shortest total sum of lengths (Rzhetsky and Nei, 1993, Felsenstein, 2004).

The mode of operation of NJ begins with a tree with a star like topology, which assumes that no taxa cluster together. Amongst all the possible pairs of taxa, the first pair of "neighbors", i.e. the two taxa with the smallest summed branch lengths, are selected and joined. This pair of taxa is then considered by the algorithm to be one (i.e. composite) operational taxonomic unit (OTU). The branch length of the composite OTU is calculated as the average of the branch lengths of the initial two taxa. The new composite OTU is then considered by the algorithm to be another taxa and the process of neighbor generation is then repeated. This procedure continues until all interior branches are identified (Saitou and Nei, 1987).

Although NJ performs relatively well (Sourdis and Nei, 1988, Saitou and Imanishi, 1989), some methodological issues regarding pairs of OTUs with equal length (Backeljau et al., 1996) have been identified, including the problematic use of bootstrap with NJ where the random reconciliation of equivalent OTUs can lead to inflated support (Farris et al., 1996). However, NJ is widely used as a preliminary rapid tree building method that serves as starting point for other methods (for example PhyML, Guindon and Gascuel, 2003, uses a variant of NJ). Further to this, NJ is often selected over other methods when there are a large number of sequences in a considered data set as its performance is fundamentally sufficient, with moderate computation costs (Tamura et al., 2004).

1.1.2.5 Maximum likelihood

In the early part of the 20th century the statistician, and population geneticist, R.A. Fisher devised the maximum likelihood (ML) method (Fisher, 1912, 1921, 1922). However, it was Edwards and Cavalli-Sforza (1964), who in addition to parsimony, first applied likelihood methods in phylogeny, albeit for gene frequency data. Progression of the likelihood approach in phylogenetics included the work of Neyman (1971), who first applied likelihood to genetic sequences, with Kashyap and Subas (1974), in turn, providing improvement upon Neyman’s work. It was, however, Felsenstein (1981) that was responsible for developing the groundbreaking "pruning algorithm" that allowed ML to be implemented on a realistic number of sequences.

ML can be considered a parametric approach to tree building, which estimates the probability of observing the data given a model of sequence evolution. Under the ML criterion the data is fixed, while the model (or part of it) is free to vary (Edwards, 1972). In phylogenetics, the data are aligned genetic sequences, while the model is comprised of both the model of DNA or protein evolution (which is generally fixed), and the tree (which is free to vary). For each site in the alignment (given a substitution model and a tree) an associated likelihood can be calculated, with the product of these site likelihoods yielding the total likelihood for the considered model (substitution model and tree). While a tree-search is performed, the fit of different trees to the data is investigated (fixing the substitution model), and the tree that gives the highest likelihood (under the fixed model) is the ML tree (Felsenstein, 1981).

The appeal of this method can be attributed to its demonstrated robustness to systematic errors and model violation in comparison to other methods (see Hasegawa et al., 1991, Huelsenbeck, 1995, Whelan et al., 2001), particularly parsimony. Another clear advantage of this method is the ability to differentially select between models of evolution (Keane et al., 2006), so as to better account for the evolutionary process that generated the data. This is crucial to avoiding phylogenetic artifacts (Pisani, 2004, Sperling et al., 2009, Rota-Stabelli et al., 2010). Although it is important to select a model that acts as an appropriate reflection of the evolutionary process, it is equally imperative to avoid selecting a model that is too parameter rich. This is particularly relevant in the case of small (or relatively small) alignments (e.g. single gene analyses), where the use of parameter rich models like CAT (Lartillot and Philippe, 2004), or mechanistic GTR, can on occasion lead to the problem known as the "infinitely many parameters trap" (Felsenstein, 2004) where the number of parameters increases as more sites are considered. However, for the modern superalignment approach this problem is probably less important (Philippe et al., 2005a)

There are a number of statistical measures defined that facilitate the selection of the model that best fits the data, e.g. the Likelihood Ratio Test (LRT) and Bayesian cross-validation. In the LRT, multiple, pairwise tests of goodness of fit are conducted to determine the model that best fits the data. This test can more explicitly be considered a hierarchical likelihood ratio test (Posada and Crandall, 1998) because, with each iteration, an increasing number of parameters are added to the alternative model, until the present null model is not rejected (Pol, 2004). However, this approach is limited as it can only be used for models that are subtypes of each other (i.e. nested models) and consequently is generally unsuitable for models of protein evolution (Keane et al., 2006).

The most widely used test for amino acid model selection is the Akaike information criterion (AIC; Akaike, 1973). The AIC is measured using the following formula:

$$\text{AIC} = -2L_m + 2m \qquad [1.2]$$

where m is the number of parameters of the model and L_m is the maximised log likelihood of the estimated model. An alternative measure is the Bayesian information criterion (BIC; Schwarz, 1978), which penalises more severely than the AIC for extra parameters. Both measures are featured in various model selection software, including Modeltest (Posada and Crandall, 1998) and Modelgenerator (Keane et al., 2006). Differently, and more conclusively, the Goldman-Cox test (Goldman, 1993, Whelan et al., 2001) can be used to examine if the chosen model in fact fits the data. In this way, it provides a more absolute test than the AIC (Foster, 2004) or LRT, as such measures will always return a best fitting model, even in the case where every considered model is a poor fit to the data.

Concluding, ML is now a well-established, hugely popular, method of phylogenetic inference, with many software implementations, including the relatively recent PhyML (Guindon and Gascuel, 2003) and RAxML (Stamatakis et al., 2005), with the latter being generally considered the better performing of all currently available ML software.

1.1.2.6 Bayesian inference

Statistically, Bayesian methods are closely related to likelihood methods. The important difference between these probabilistic methods is that the Bayesian approach uses an informative prior distribution of the entity being estimated (Felsenstein, 2004). The implementation of Markov chain Monte Carlo (MCMC) algorithms has greatly helped popularise these methods (Li, 1996, Yang and Rannala, 1997, Mau and Newton, 1997). The appeal of the Bayesian statistic is that it tries to mimic the human decision making process, in that decisions are altered by data (Huelsenbeck and Bollback, 2001, Huelsenbeck et al., 2001).

Bayesian phylogenetics is based upon what is referred to as the posterior probability of a tree, which can be considered more simply as the probability that a tree is “true” (Huelsenbeck et al., 2001). The posterior probability is arrived upon using Bayes’ theorem:

$$\Pr(H \mid D) = \frac{\Pr(H) \times \Pr(D \mid H)}{\Pr(D)} \qquad [1.3]$$

which calculates the posterior probability of a hypothesis *H* (i.e. a tree), given some data *D* (i.e. an alignment of *n* sequences), a substitution model, and a prior probability distribution for the set of all available alternative hypotheses (in our case all trees of the *n* taxa). Here, the denominator of the theorem is represented in its most simple form, however, in reality, this is extremely difficult to calculate, as it requires summing the likelihood of all possible hypotheses (i.e. trees; Yang and Rannala, 1997). This problem is overcome by the use of MCMC methods, which only consider a random sample from the posterior distribution, thus providing an approximation for the true posterior probabilities (Huelsenbeck et al., 2001, Felsenstein, 2004). Key to the success of the MCMC methods is the exploitation of a mathematical trick, by means of which calculation of the denominator in the Bayes formula is no longer necessary (Huelsenbeck et al., 2001).

The most widely used MCMC method is the Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970), which operates as follows: (1) a random starting tree is selected and its posterior probability is calculated. A new tree from the distribution is then selected. (2) The probability of the new tree is calculated and the Metropolis-Hastings algorithm is used to decide whether to accept or reject this new tree. If accepted, the new tree becomes the current tree, while if rejected another tree is selected. An important aspect of MCMC is that, on occasion, a new tree can be accepted even if its likelihood is lower than that of the current tree. This is important for obtaining a fair representation of the posterior distribution, and to this end, trees of poor likelihood are accepted with a probability which is proportional to their likelihood: the worse the likelihood of a tree, the less likely it is to be accepted. The MCMC procedure will continue infinitely as the algorithm has no termination clause. In general, more than one independent "chain" for each data set is run, with cessation of the algorithm being determined by the convergence of these chains (i.e. when the chains find trees with a similar distribution). While most defined Bayesian MCMC methods employ the Metropolis algorithms, they do typically differ in the type of prior assumed, in the manner in which they move through tree space and in the way they summarise the posterior (Felsenstein, 2004).

In Bayesian analyses, support for each node is represented by a posterior probability. Unlike other methods of estimating support (see Section 1.1.4.1), this has the advantage of being a measure of the probability of a particular node being true, given the data and the model (Erixon et al., 2003). However, some authors have contended that posterior probabilities overestimate the true support of a node (Rannala and Yang, 1996, Douady et al., 2003). The Bayesian approach confers the additional advantage over other methods in that it allows for the use of models with higher dimensionality (Lartillot and Philippe, 2004). In this sense, more realistic aspects about the evolutionary process can be incorporated into Bayesian inference (see above). Bayesian inference continues to see a steady uptake in phylogenetic studies and, currently, boasts several software implementations, including MrBayes (Huelsenbeck and Ronquist, 2001) and PhyloBayes (Lartillot and Philippe, 2004).

1.1.3 Detecting signal in phylogenetic data sets

While the use of robust methods of tree inference, coupled with the selection of an appropriate model of evolution, is imperative to the accurate reconstruction of a phylogeny, such measures are largely futile in the light of poor quality data. Accordingly, the quality of data is often assessed as part of the tree building process. A common method used to test for the presence of hierarchical structure in a data set is the permutation tail probability (PTP) test (Archie, 1989, Faith, 1991).

Under the PTP test, each character at each site in the alignment is permuted, so that characters are allocated to a species at random. This has the effect of generating an alternative alignment that is not intended to represent a phylogeny, but rather a distribution of character states characteristic of the data (Archie, 1989, Felsenstein, 2004). This procedure is repeated numerous times and the minimum tree length for the original data is compared to the distribution of tree lengths from the data perturbations. If a significant proportion of the permuted data tree lengths (i.e. more than 1% or 5% - depending on the significance threshold considered) are shorter or of the same length of the most parsimonious tree obtained using the unpermuted data, then the tested data set is considered to convey no phylogenetic signal.

Two general criticisms of the PTP test have been raised. The first issue brought up in relation to the PTP test is that it is too lenient a measure of phylogenetic signal. Felsenstein (2004) suggested that this could reasonably be the case when two species are practically identical sibling species. In this situation, the PTP test will detect the strong relationship between these species and thus may provide misleading signal, as more distant relationships may not necessarily be detected. A similar contention has been shown experimentally, where alignments devoid of signal successfully attain significant scores (Slowinski and Crother, 1998).

A second criticism, broached by Swofford et al. (1996), is that a tree consisting entirely of a multifurcation can be shown to pass the PTP test if the branch lengths of the tree are sufficiently disproportionate. However, Felsenstein (2004) suggests that it could be argued that such a case does in fact have signal. Like Felsenstein, I contend that the case of Swofford et al. (1996) does not detract from the capability of the PTP test. The role of the PTP test is expressly to detect signal when present. However, this role does not extend to determining the nature of the detected signal, this can only be determined by phylogenetic analysis. For this reason, the PTP test should (as in the analyses described in this thesis) only be used to identify data sets with sufficient clustering signal suitable for further analysis. In the case suggested by Swofford et al. (1996), the signal in the data is not phylogenetic in nature, but it still is signal (i.e. it represents a bias).

While the merit of the PTP test has been the topic of much debate (see Trueman, 1996, Carpenter et al., 1998), active research into the development of the PTP test has continued. Variations include the topology dependent PTP (T-PTP) test (Faith, 1991), the 'yet another permutation tail probability' (YAPTP) test (Creevey et al., 2004) and the gene tree parsimony PTP (GTP-PTP) test (Holton and Pisani, 2010), which is a variant of the YAPTP test. A better approach to test the quality of a data set is likelihood mapping (Strimmer and Von Haeseler, 1997). Unfortunately, likelihood mapping is time consuming and was not suitable for the analyses discussed in this thesis.

1.1.4 Assessment of support

It is extremely important to be able to have a means of evaluating a phylogenetic tree produced for a given data set of interest. This confers a level of confidence in the resulting tree and allows for comparison between different topologies and methodologies. There are various methods to determine the robustness of a tree, some of which are described below.

1.1.4.1 Bootstrap and jackknife

Two related approaches that can be used to estimate the level of support for a phylogeny are the bootstrap and the jackknife. Bootstrap is a statistical technique that was first applied in phylogenetics by Felsenstein (1985). When the bootstrap is applied to a phylogenetic data set, the original alignment is used to generate multiple (pseudoreplicate) alignments of the same dimensions. Each new alignment is created by sampling sites from the original alignment, with replacement (i.e. the same site can appear more than once in the considered pseudoreplicated data set, while some other sites might not appear in the same pseudoreplicate data set). This process is replicated a defined number of times, and each resultant alignment is individually used to build a phylogeny using a reconstruction method of choice. A majority rule consensus method (Margush and McMorris, 1981) is then used to coalesce the resulting sample of trees to give a single tree with support values at every node. Values at the nodes represent the proportion of times a given clade is found from the analysis of the pseudoreplicated data sets.

The jackknife, which is an older statistical method, was also first used in a phylogenetic context by Felsenstein (1985). The general concept of the jackknife is to reduce the sample of characters by one, or more, iteratively and subsequently calculate the estimate for the remaining data. One variant of the jackknife used in phylogenetics is the delete-half jackknife, which exhibits many of the same properties as bootstrap, however, samples without replacement at a rate of $\frac{n}{2}$ for a data set of size *n*. Farris et al. (1996) proposed an alternative adaptation of the jackknife, called delete-1/e jackknife, however, Felsenstein (2004) demonstrates that this method can result in inflated support compared to the delete-half jackknife, and under some conditions becomes equivalent to bootstrap, therefore, negating its need.

1.1.4.2 Bayes factors

The Bayes factor (BF) is a Bayesian approach to hypothesis testing. The BF, in its simplest form, is a likelihood ratio and thus represents the part of Bayes formula (see Equation [1.3] above) through which the effect of the data, on the definition of the posterior probability, is expressed. The BF can be considered the probability of the data given the null hypothesis, over the probability of the data given the alternative hypothesis (Goodman, 1999). Essentially, the BF is a measure of evidence for one hypothesis as opposed to another (Kass and Raftery, 1995). The difference between the BF and the likelihood ratio test is that BF values are calculated using likelihood values marginalised across all hypotheses, rather than on a fixed optimal topology. In this way, the BF can take into consideration statistical uncertainty when comparing two hypotheses. The BF returned when two hypotheses are compared is generally interpreted according to the table of Kass and Raftery (1995; see Appendix A1). In phylogenetics, BFs are proving to be a useful tool for evaluating alternative topologies (Sperling et al., 2009, Holton and Pisani, 2010) and for model selection (Sperling et al., 2010).

1.2 Sources of phylogenetic error

There are two classes of error that can occur in phylogenetics: systematic error and stochastic error. Stochastic error affects all tree reconstruction methods equally, however, this problem has largely been eliminated by the use of large genomic scale data sets. Differently, systematic error persists as the key problem faced by modern phylogenetics, although some methods are more adept at handling this type of error than others. Below, the two types of error are discussed, in addition to some measures employed to reduce and preclude systematic errors.

1.2.1 Systematic errors

Systematic errors, or inconsistencies, occur when a reconstruction method arrives upon an incorrect solution, with stronger support, as the amount of data considered increases. This situation occurs when certain characteristics of the data cause the method to be misled (Delsuc et al., 2005). More specifically to ML and Bayesian inference, systematic errors can occur when the model of sequence evolution does not fit the data (Rodriguez-Ezpeleta et al., 2007). Three of the most common causes of systematic error are compositional bias, long branch attraction and heterotachy.

1.2.1.1 Compositional bias

Compositional bias causes sequences to be erroneously grouped together based upon their analogous nucleotide or amino acid composition. For some time, it was believed that this problem was confined to nucleotide-based phylogenies, with protein sequences being considered to be relatively robust to such compositional effects (Loomis and Smith, 1990, Hasegawa and Hashimoto, 1993). However, it was later shown that there was indeed an implicit compositional bias observed in phylogenies derived from amino acid sequences as well (Foster et al., 1997, Foster and Hickey, 1999). Nonetheless, compositional biases are significantly less likely to occur in the analysis of amino acid data sets (Rota-Stabelli et al., 2010).

Popular methods of overcoming compositional biases include the use of LogDet transformation (Lockhart et al., 1994, also known as paralinear distances, Lake, 1994). Another notable approach is to use RY coding (Woese et al., 1991), in which sequences are recoded as either purines or pyrimidines. At the amino acid level, Dayhoff recoding (i.e. recoding amino acids in their functional classes) has also been shown to significantly reduce compositional biases (see Hrdy et al., 2004). The use of heterogeneous models, accounting for varying composition throughout the tree (e.g. Foster, 2004), have also been proposed to limit compositional effects, however, these are computationally expensive and, as such, can be of limited utility (Rodríguez-Ezpeleta et al., 2007).

1.2.1.2 Long branch attraction

Long branch attraction (LBA) is the most infamous and well documented systematic error. It occurs in the situation where species in a given data set, that are rapidly evolving, are artifactually drawn together. Felsenstein (1978), first observed this phenomenon, where he identified conditions pertaining to unequal evolutionary rates, under which parsimony is inconsistent. Significant contributions followed from Hendy and Penny (1989), where they showed that disproportionate branch lengths also caused LBA, and from Kim (1996) who demonstrated that even when branch lengths are equivalent parsimony can become inconsistent, however, this is conditional on whether the tree is imbalanced (see Chapter 4).

Since the identification of parsimony as a method sensitive to LBA, all other methods of tree reconstruction have been thoroughly scrutinised. Various distance methods have been shown to be statistically consistent, including NJ (Saitou and Nei, 1987), minimum evolution least squares (Rzhetsky and Nei, 1993). However, more recently it has been repeatedly shown that in cases of model misspecification, distance methods do invariably become inconsistent (Gascuel et al., 2001, Susko et al., 2004, Pisani, 2004)

Tree reconstruction using ML was vehemently claimed to be impervious to becoming inconsistent by Felsenstein (1973), and more recently Yang (1994b). However, similar to what has been observed with distance-based methods, ML can become inconsistent when the model used is not sufficiently parameter rich (Gaut and Lewis, 1995, Lockhart et al., 1996, Sullivan and Swofford, 1997). Despite this, ML is more robust to model misspecification than distance methods, and the conditions of inconsistency for this method are known (see Sperling et al. 2009 for an example). Bayesian inference, also being a probabilistic method, is thought to emulate ML, becoming inconsistent only when an ill-fitting model is used (but see Kolaczkowski and Thornton, 2009).

As such, ML and Bayesian inference subsist as the most robust tree reconstruction methods to LBA, and use of these adept methods is encouraged to avoid introducing such bias (Bergsten, 2005). Further to this, however, Bergsten (2005) does stress that these methods are more resistant, rather than immune, to LBA and simply implementing these methods will not suffice for the preclusion of LBA (for example Sperling et al., 2009). However, Bayesian inference may perhaps have a more promising future prospect with regard to LBA, as it better lends itself to the implementation of complex models (like CAT) that are key to avoiding LBA.

The introduction of LBA into a phylogenetic reconstruction can be avoided in several ways. One of the most widely used of these approaches is to increase the taxonomic sampling. This was first shown to be a means of alleviating LBA by Hendy and Penny (1989) and has, subsequently, been repeatedly confirmed to be effective (Hillis, 1996, Rannala et al., 1998, Pollock et al., 2002, Poe, 2003). Increasing the taxon sampling essentially serves to break up the problematic long branches. The application of an improved taxonomic sampling in many data sets has resulted in more accurate phylogenies, calling for a revision of many preceding results (for example Halanych, 1998, Philippe et al., 2005b, Holton and Pisani, 2010). Rosenburg and Kumar (2001) contend that insufficient taxon sampling plays a far less critical role in phylogenetic accuracy than the inclusion of more lengthy sequences. However, this assertion has met with much criticism (Zwickl and Hillis, 2002, Pollock et al., 2002) and has been shown to be incorrect with genomic scale data by Holton and Pisani (2010).

Despite being a highly commendable approach, increasing taxon sampling is far from a panacea, as it has been known to aggravate the LBA problem in some cases (Kim, 1996, Poe and Swofford, 1999, Poe, 2003). As such, the addition of new taxa may bring the addition of other unanticipated problems. Further to this, supplementing the taxon sampling offers little utility in instances where LBA occurs when all defined members of a group are already sampled (Bergsten, 2005, Bergsten and Miller, 2006).

Another means of minimising LBA is the use of an optimal outgroup (Wheeler, 1990). When an outgroup that is too divergent is selected, a fast evolving ingroup taxon may be artifactually attracted to the long branch of the outgroup (Philippe and Laurent, 1998). The use of an extremely inappropriate outgroup becomes equivalent to using a random, highly saturated, sequence with regard to the ingroup taxa (Sanderson and Shaffer, 2002). Various strategies can be employed to ensure the selection of an appropriate outgroup (Sanderson and Shaffer, 2002), some of which benefit from success (Rota-Stabelli and Telford, 2008), however, often the simple availability of a suitable outgroup can be a limiting factor.

LBA can additionally be circumvented by the adoption of a selective sampling strategy. In this approach, the evolutionary rate of large clade members is assessed, with taxa exhibiting a particularly rapid rate being overlooked in preference for taxa with a more subdued evolutionary tempo. In this way, certain optimal species are used as representatives of these larger clades in phylogenetic studies. The most notable use of this method was by Aguinaldo et al. (1997), which saw the definition of the Ecdysozoa clade and as such the new animal phylogeny.

Lastly, LBA can be avoided by the removal of fast evolving sites (e.g. Hirt et al., 1999, Ruiz-Trillo et al., 1999, Brinkmann and Philippe, 1999). One method proposed for the identification of such sites is the use of a parsimony-based approach called slow fast (Brinkmann and Philippe, 1999), however, this method does have a limitation in that it requires the *a priori* definition of monophyletic groups. An alternative method is the compatibility based approach of Pisani (2004), which allows for the identification of fast evolving sites in cases where no *a priori* information is available.

One outstanding issue with the elimination of fast evolving sites to preclude LBA is the determination of when to stop removing sites. Pisani (2004) offers a series of guidelines that can be adopted in deciding a cut off (see also Sperling et al., 2009), however, it is additionally warned that even deleted characters can convey a considerable phylogenetic signal. Recent work of Cummins and McInerney (2011), however, provides a method of categorising and scoring sites according to their degree of similarity. This approach represents a significant improvement in that it leads to a spectrum over which sites to be removed can be selected, rather than the binary approach offered by preceding methods.

1.2.1.3 Heterotachy

Heterotachy is defined as variation in the evolutionary rate of a given site throughout time (Delsuc et al., 2005). This situation occurs due to functional constraints that are imposed on a given gene or protein, with phylogenetic inference being misled where the proportion of invariable sites in distantly related species has converged (Delsuc et al., 2005). Surprisingly, the true importance of the incidence of heterotachy was only recently realised (Lopez et al., 2002), although the covarion model, of Fitch and Markowitz (1970), does somewhat attempt to address a certain degree of variability. In this model, substitutions in the "c" or "concomitably variable codons" are accounted for, but since the proportion of these sites remains constant, this model is limited (Steel et al., 2000).

Heterotachy presents a particularly difficult problem to detect, as it does not leave any observable signatures in sequences (Kolaczkowski and Thornton, 2004). However, recently, Whelan et al., (2011) have developed a likelihood ratio test based approach for detecting heterotachy, which allows substitution and switching rates to vary independently across branches of the tree. An older approach suggested for addressing heterotachy is based on the use of a mixture, or non-homogenous gamma, model to account for variability of site rates over time. Two general types of gamma models are used in this context: covarion-like models, similar to the model originally introduced by Fitch and Markowitz (1970), and the mixture of branch lengths model (Kolaczkowski and Thornton, 2004). Covarion-like models allow sites to interchange from being variable to invariable, while the mixture of branch lengths model assumes each site arises from one of a number of specified branch lengths on the same topology. A recent comparison of these two models shows that the covarian-like models perform better than the mixture of branch lengths models (Zhou et al., 2007), but there is reason to believe that the models introduced by Whelan et al. (2011) should outperform both (see Whelan et al., 2011 for more details).

1.2.2 Stochastic error

In the situation where too small a number of positions are considered, stochastic or sampling error can be introduced into a phylogenetic analysis. Since there is a scant amount of data, random noise can be incorporated into certain aspects of the tree, resulting in poor resolution (Philippe et al., 2005a). Traditionally in phylogenetics, this posed a great problem as studies were largely based on a single gene. As sequences accumulate mutations at random, each mutation is subject to a certain amount of stochastic error, therefore, to reduce the effect of such error, a large amount of sites must be included (Nei, 1986). It can be imagined that stochastic errors might result in incorrect trees with high support. This could potentially happen in a situation where sequence sampling is very poor (i.e. the sequences are short) and a subset of the sites in the sequences, by chance, agrees on a specific set of relationships.

With the incorporation of increasing numbers of sites, the influence of stochastic error should eventually be negated, allowing support for nodes to reach maximum bootstrap values consistently across a tree (Philippe et al., 2005a). The recent progression in phylogenetics from single gene studies to multiple genes alignments, and even genomic scale studies (e.g. Pisani et al., 2007, Holton and Pisani, 2010, Rota-Stabelli et al., 2011), has seen a diminishing influence for stochastic error. Indeed, the availability, and subsequent use, of genomic data, has been suggested to mark the end of stochastic error driven incongruence (Gee, 2003, Philippe et al., 2005a, Jeffroy et al., 2006).

1.3 Phylogenomics

The advent of genome sequencing, and the accompanying implications, marked an exciting paradigm shift in molecular phylogenetics. The sequencing of the first complete genome of a free-living organism was concluded in the mid nineties (Fleischmann et al., 1995). For the first time, the entire genetic landscape of an organism was available to researchers, and as more genomes came on stream meaningful comparisons could be made between organisms in a more comprehensive manner. Almost 16 years on, the number of complete genomes sequenced is in the order of thousands, with sequencing innovations and affordability progressing at an inexorable rate, even still (Liolios et al., 2010). This, coupled with the perpetual improvement in computer hardware and software, means that molecular phylogeneticists have never been better equipped. With such compelling assets in their arsenal, many have turned their attention to constructing phylogenies that incorporate extensive proportions of an organisms genetic information (as indeed I have in the studies described in this thesis). In 1998, on the brink of the genomic explosion, Eisen conferred the name "phylogenomics" on this new approach (Eisen, 1998). Two approaches generally employed for the construction of genomic scale phylogenies are data concatenation and super tree reconstruction (Delsuc et al., 2005). An alternative class of methods used are gene content methods (e.g. Rivera and Lake, 2004), however, to date these methods receive little utilisation and have been shown to perform poorly (McCann et al., 2008). The more widely used data concatenation and super tree methods are introduced and discussed in the ensuing sections as they represent the key tools used in the analyses presented in this thesis.

1.3.1 Supermatrix

The supermatrix approach is defined as "the direct, simultaneous use of all the character evidence from all included taxa" (de Queiroz and Gatesy, 2007). More specifically, it can be considered a total evidence (*sensu* Kluge, 1989) approach, where the alignments (matrices) of each individual data component of interest are combined to form one large composite matrix. In the event that a taxon is not present in a given source matrix, this taxon is represented in the supermatrix by a series of '?', that extends the length of the other sequences present in the source matrix. Upon the integration of all source matrices, the supermatrix is then analysed by a tree reconstruction method (see

Figure 1.3). It is important to note that this method is limited to single gene families only and that paralogy can cause supermatrix analyses to return incorrect trees. In the supermatrix approach, by directly using character data in the estimation of a global phylogeny, stochastic error is diminished (or even eliminated). However, it has been recognised for some time, that in light of increasing data accumulation, this approach, at least in its current form, becomes unsustainable (Sanderson et al., 1998, but see Philippe et al., 2005a).

The major advantage of the supermatrix approach lies in its ability to detect underlying signals that may not be apparent from the analysis of the individual data sets, a feature of the method that can even extend to relationships that are not supported by the individual analysis of the source matrices (Kluge, 1989, Pisani and Wilkinson, 2002, de Queiroz and Gatesy, 2007). This enhancement of weak and underlying signals confers the “total evidence” property to this approach (see Pisani and Wilkinson, 2002). A further benefit of the supermatrix approach lies in the ability to use probabilistic tree reconstruction methods that incorporate parameter rich mixture models (Delsuc et al., 2005). This has been shown to have the effect of improving the support of genome scale analyses (e.g. Brinkmann et al., 2005), although increased support does not necessarily reflect phylogenetic accuracy (phylogenetic artifacts are generally well supported).

One issue of concern in relation to the supermatrix approach is the effect of missing data, which has been shown to lead to a lack of resolution (Wiens, 2006). Although it has been claimed by Philippe et al. (2004) that a large amount of missing data can be tolerated without compromising accuracy (see also Driskell et al., 2004), Sanderson et al. (2010), in an investigation of the distribution pattern of missing data, show that specific distributions of missing data can lead to situations where the true tree

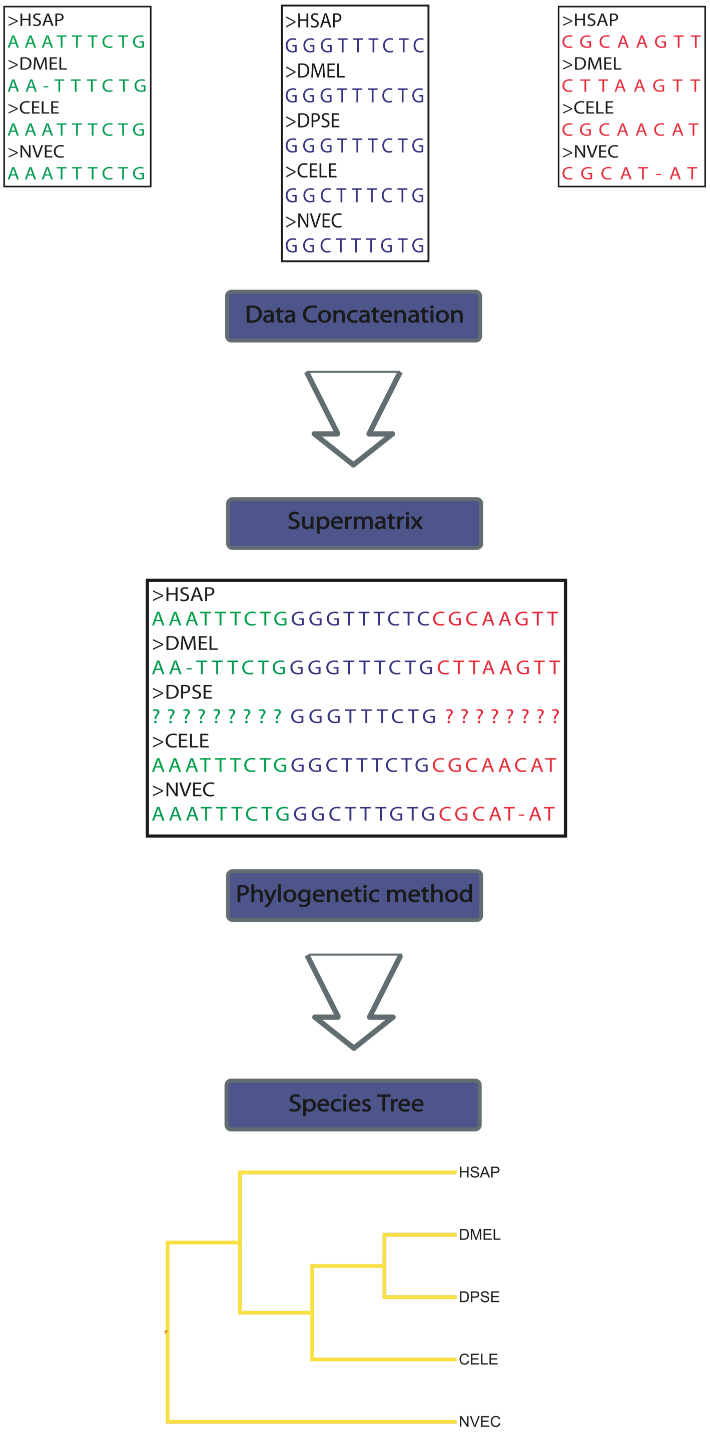


Figure 1.3 Supermatrix analysis.

The individual gene matrices are combined to produce a supermatrix. A tree reconstruction method is then used to derive a single global species phylogeny.

cannot be found. Sanderson et al. (2010) define this property of a data set “decisiveness”, and point out that no matter how many genes are used, a specific combination of missing genes in taxa, can always result in a data set lacking decisiveness. Importantly, it should be pointed out that many non-trivial combinations of missing genes (i.e. combinations that cannot be readily identified by eyeballing a data set) could lead to a data set lacking decisiveness.

1.3.2 Supertrees

A supertree approach is one that combines several input tree topologies, rather than the character data upon which they are derived, to obtain a single tree that represents the information contained in each input tree (Bininda-Emonds, 2004a). This approach can be considered a generalisation of the consensus approach (e.g. strict consensus and majority rule consensus) but differs in that input trees may have a partially, rather than fully, overlapping leaf set (Cotton and Wilkinson, 2009).

From a theoretical point of view, the concept of a supertree, or rather a “composite” tree, has existed since systematics originated. While strictly speaking Aho et al. (1981) defined the first supertree method (but in the context of merging databases in computer science), it was Gordon (1986) who provided a formal definition of a supertree method and introduced the term “supertree” in the study of classification. Gordon (1986) described a strict consensus supertree, which displays only the groups that are common to all source trees (Bininda-Emonds, 2004b). Although this marked a significant development, progression beyond the consensus approach proves methodologically difficult. This is because the familiar split substructure used in the consensus approach becomes ineffective, as each input tree presents different splits (Cotton and Wilkinson, 2009).

As such, all currently implemented supertree approaches represent somewhat *ad hoc* methodologies, with the exclusion of Cotton and Wilkinson's majority rule consensus supertree (2007), and the recently developed maximum likelihood supertree method (Steel and Rodrigo, 2008; of which the majority rule consensus supertree method is a special case). Unfortunately, software implementation of both these methods is still under development (Akanni and Pisani, personal communication), therefore, they could not be implemented in the analyses outlined in this thesis.

The introduction of the matrix representation with parsimony (MRP) supertree approach (Baum, 1992, Ragan, 1992) marked an important milestone for supertree methodology. It offered a more applicable method than that of Gordon and, as such, the supertree approach saw a steady uptake by systematists. The appeal of supertrees is that they offer a divide and conquer approach, which confers two advantages. Firstly, the impact of missing data in standard phylogenetic reconstruction is reduced (although missing data can cause a problem in the subsequent supertree stage of analysis; see Chapter 3). This is because topologies are reconciled at a local level, using only the taxa available for that particular subdivision, and are then incorporated into a global solution (Wilkinson and Cotton, 2007; see Figure 1.4). Secondly, supertrees offer the ability to consider both a very large taxon and gene sampling (see for example Pisani et al., 2007, Holton and Pisani, 2010), which is something that cannot be achieved using the supermatrix approach. For example, the supermatrix based study of Dunn et al., 2008 samples only 150 genes for 77 taxa, whereas the supertree study of Holton and Pisani, 2010 samples 2,216 genes for 42 taxa.

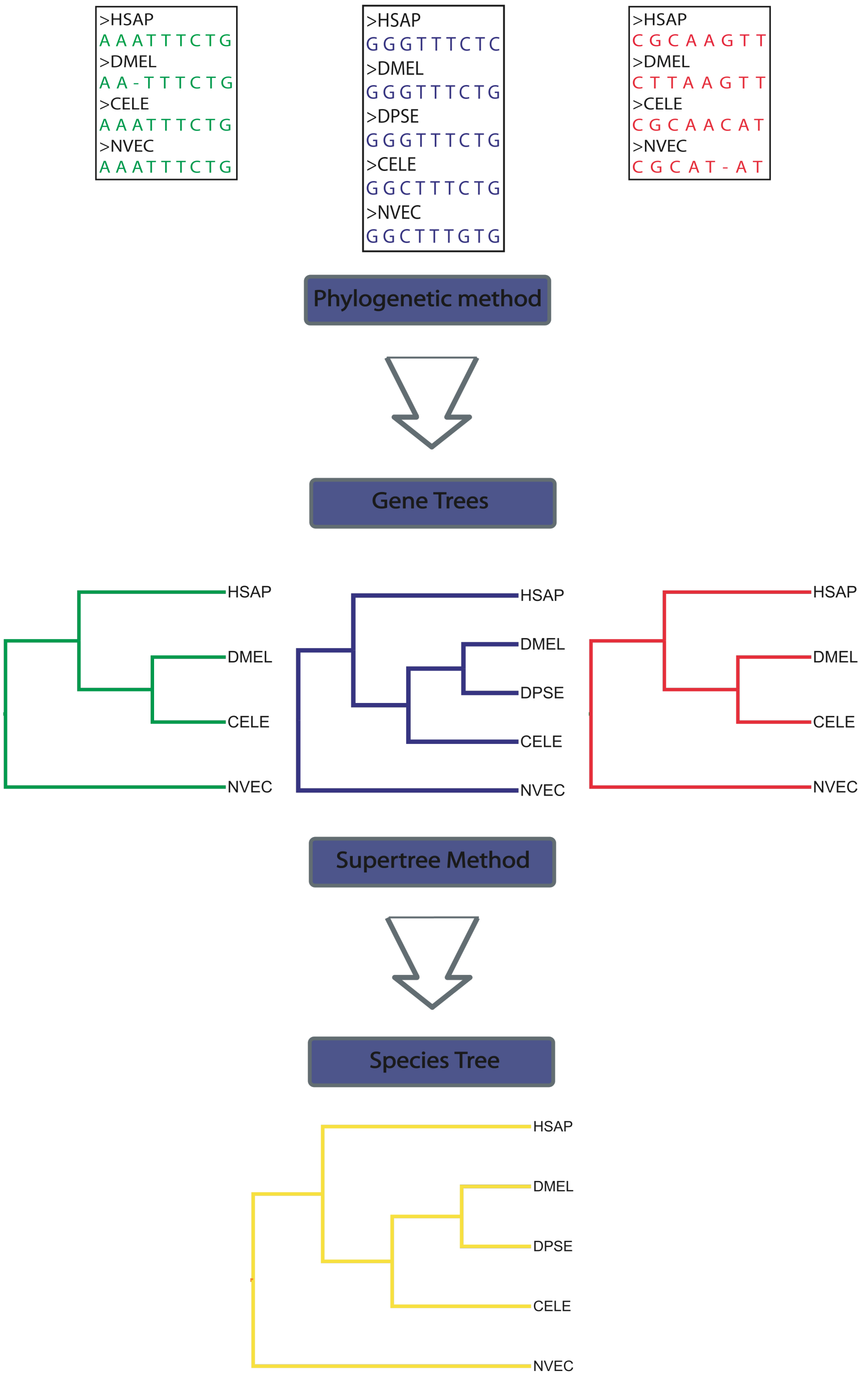


Figure 1.4 Supertree construction.

For each individual gene, a topology is derived using a tree reconstruction method. The resulting trees are then combined on the basis of their overlapping taxa by a supertree method, resulting in a single species based tree.

Although the supertree approach has met with some criticism, because of the indirect use of character data (Rodrigo, 1993, Slowinski and Page, 1999), development of the supertree approach has flourished in recent years, with approximately 11 alternative supertree methods currently available (see Table 1.1). In the following sections I will limit the discussion to the supertree methods I have used throughout the work presented in this thesis.

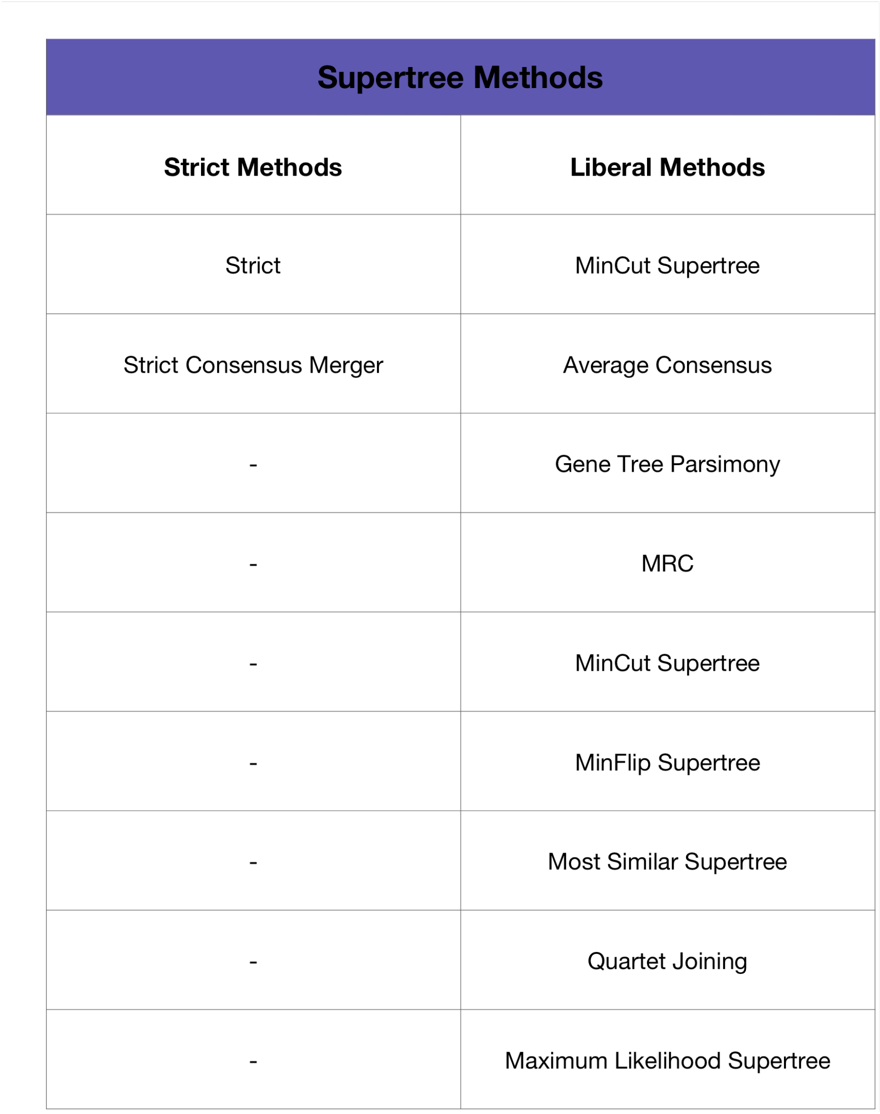
1.3.2.1 Matrix representation with parsimony

Traditionally, matrix representation with parsimony (MRP) was the most widely used supertree method, and indeed persists as such even today (Cotton and Wilkinson, 2009). Under MRP, each input tree topology is recoded into a data matrix, with each node of the input trees being represented by either of the following characters: ‘0’, ‘1’ or ‘?’. A variety of coding schemes have been proposed (e.g. Baum, 1992, Ragan, 1992, Purvis, 1995, Wilkinson et al., 2001), all of which use variations of the aforementioned characters. One scheme, defined by Baum (1992), and independently by Ragan (1992), is to assign a ‘1’ to each taxon present in a clade, a ‘0’ to taxa present but not in a clade, and ‘?’ for a taxon that is not present in the current input tree. An alternative method, defined by Purvis (1995) to account for redundancy, scores ‘1’ for taxa in a node, ‘0’ for taxa in a sister clade, and ‘?’ for missing taxa. Wilkinson et al. (2001) introduced further coding strategies based on quartets and triplets (i.e. rooted quartets of taxa). The (coded) individual matrices are combined and the resulting data matrix is then analysed by maximum parsimony to produce one or more most parsimonious trees.

The properties of MRP have attracted some discussion and it is generally known that MRP is a limited method, however, this assertion can be extended to supertrees in general. One criticism of MRP is that it does not act directly on the data itself, but rather

Table 1.1 Supertree methods.

A list of some of the currently defined supertree methods, separated by category.



Supertree Methods	
Strict Methods	**Liberal Methods**
Strict	MinCut Supertree
Strict Consensus Merger	Average Consensus
-	Gene Tree Parsimony
-	MRC
-	MinCut Supertree
-	MinFlip Supertree
-	Most Similar Supertree
-	Quartet Joining
-	Maximum Likelihood Supertree

on topologies inferred from the data (Rodrigo, 1993). However, as Wilkinson et al. (2001) point out, this is the cost involved in achieving such a tractable and malleable method. In a practical sense, the direct inference from data becomes unfeasible, particularly in the case of whole genome datasets using a supermatrix approach, due to computational limitations (Baum and Ragan, 2004). Accordingly, this criticism is effectively nullified in practice.

A second shortcoming of MRP is that it is not based on a central model (Rodrigo, 1993, Rodrigo, 1996). MRP is an *ad hoc* method, and like many supertree methods, has not been designed to include specific desirable characteristics (Cotton and Wilkinson, 2009). Additional deficiencies with this method include the potential to return unsupported clades (Pisani and Wilkinson, 2002), along with input tree shape biases that seem to depend on the coding scheme used, producing trees that are generally more imbalanced (e.g. standard coding), or balanced (e.g. Purvis coding) than expected (Wilkinson et al., 2001, Wilkinson et al., 2005). While MRP represents a very feasible and flexible supertree method, Pisani and Wilkinson's (2002) warning, that "applicability in practice should not be confused with acceptability in principle", should be least considered, if not observed. Supertree analyses should thus be performed using a few different methods to provide for some sensitivity analysis.

While MRP is used in the studies presented in this thesis for reasons of practicability, I am conscious of this method's limitations. As such, follow on work from this thesis will include reanalysis using the ML supertree method, which is currently being implemented in new software by Akanni and Pisani. In addition, in the study described in Chapter 2 (the animal phylogeny) results have been derived using two supertree methods (see Section 2.3.1.3).

1.3.2.2 Quartet joining

The quartet joining supertree method, introduced by Wilkinson and Cotton (2007) is somewhat similar to the quartet puzzling method of Vinh and von Haeseler (2004). In this divide and conquer approach, the input trees are reduced to their composing quartets, and each quartet of relevance is then used to decide the position of a new leaf in the supertree. The core concept of this approach is to build a tree according to successive refinements to the decision of where to add leaves (Wilkinson and Cotton, 2006). This method confers the advantage of being a fast supertree method, which is expected to work well where there is no conflict. In Chapter 2 of this thesis, I have used this method as an alternative to MRP as it has been shown (Wilkinson and Cotton, 2006) that it does not have a shape related bias (unlike MRP; see Wilkinson et al., 2005). A potential shape-related bias was the greatest concern in relation to the MRP supertree analyses I performed in Chapter 2, as it was clear from the inspection of the supertrees I generated that no unsupported clade was present in these trees (see Chapter 2).

1.3.2.3 Gene Tree Parsimony

Gene tree parsimony (GTP; Slowinski and Page, 1999) is a supertree method that can be used to derive a species phylogeny when gene duplications are present in the input trees. More formally, GTP “takes a collection of rooted, binary gene trees and seeks a rooted, binary species tree with the minimum reconciliation cost for the corresponding taxa” (Wehe et al., 2008; see Figure 1.5). The aforementioned reconciliation cost can be obtained by counting the number of duplication and loss events, or can be restricted to consider duplications only, or additionally by other events, including horizontal transfer.

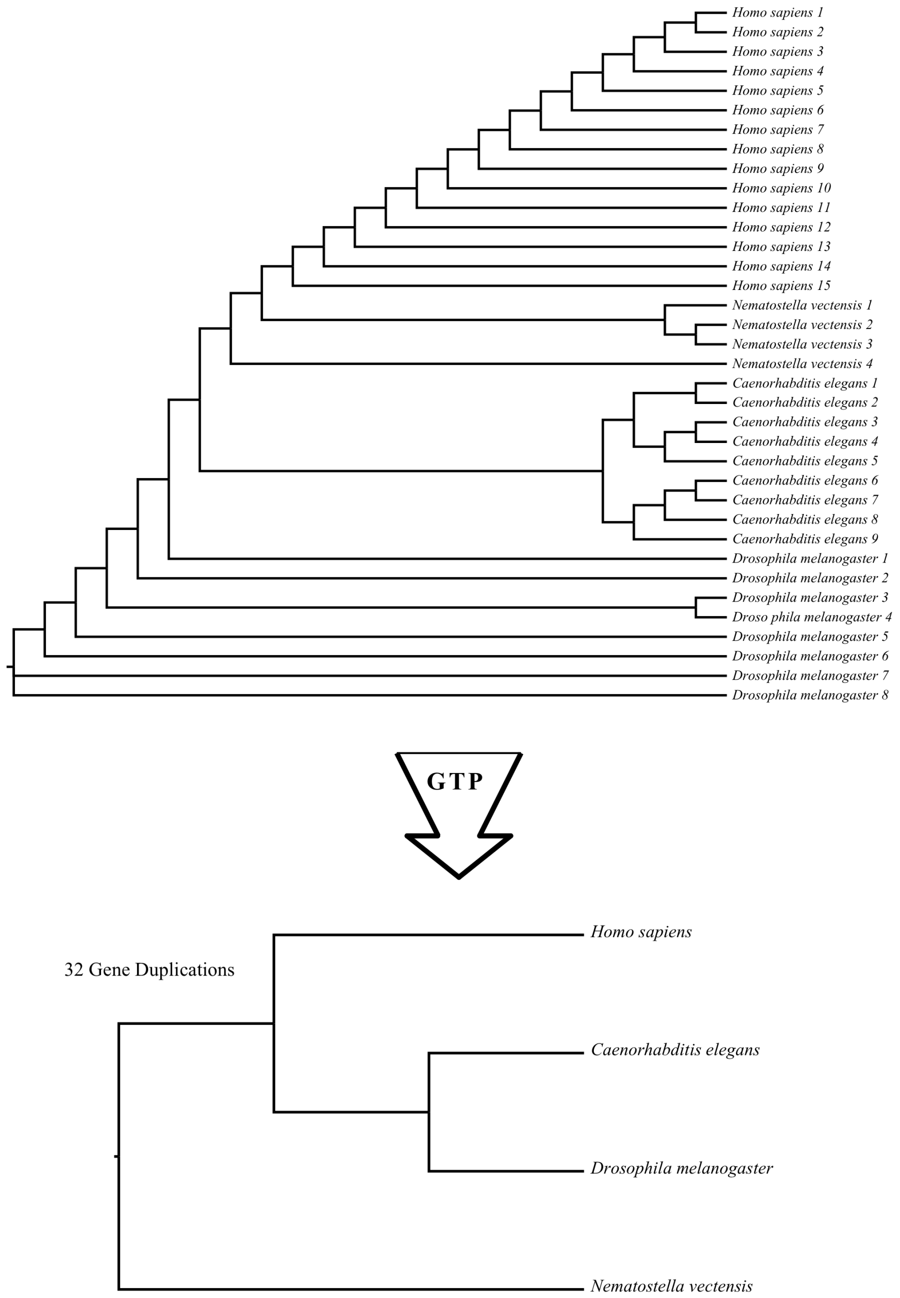


Figure 1.5 Gene Tree Parsimony.

An example of multi gene family reconciliation using GTP. Here, the 36-taxon gene tree is reconciled to a 4-taxon species tree, most parsimoniously explained by 32 gene duplications.

However, since the absence of a species in the taxon sample can, in some cases, be interpreted as gene loss, in general, calculation of the reconciliation is generally limited only to duplications (Cotton and Page, 2004).

Although bounded by the limitations of parsimony, the performance of GTP is expected to be in line with other supertree methods, and indeed, has been shown to outperform other methods at incorporating paralogous genes into phylogenies (Cotton and Page, 2003). In terms of speed, GTP can be considered sluggish compared to polynomial time methods (for example Hallett and Lagergren, 2000; Cotton and Page, 2004), however, recent algorithmic improvements (see Wehe et al., 2008) have alleviated this problem, making the use of GTP on genomic-scale data sets more feasible (see, for example, Holton and Pisani, 2010). The ability to incorporate genes with a history of gene duplication into phylogenetic analyses marks an extremely important progression, as it facilitates the execution of truly genomic scale analyses.

1.4 Aims of this thesis

With increased availability of genomic-scale data, coupled with major advances in computational power, it is now feasible to analyse truly genomic-scale data sets. In this thesis, it is my aim to conduct a comprehensive investigation of various types, and aspects, of the data used as input for phylogenomic supertrees. By exploiting the flexibility of the supertree approach, I have endeavoured to reconstruct phylogenies that survey the most expansive sample, in terms of depth (genes) and breadth (taxa), of the genome possible. To do this, the use of two atypical data types was investigated (i.e. firstly complete genomes, including paralogous gene families and secondly complete genomes, including paralogous gene families, plus partial genomes, in the form of expressed sequence tags; ESTs).

A further aim of this thesis is to obtain a better understanding of the gene trees that are combined to derive supertrees. As such, I have undertaken an assessment of the existence of shape-related biases associated with standard methods of phylogenetic reconstruction used to infer input trees (MP, ML, Bayesian inference and NJ). The manner in which I addressed these aims is outlined below.

In Chapter 2, phylogenomic supertrees are used to address the Bilaterian phylogeny, but more explicitly the contention between the Coelomata and Ecdysozoa hypotheses. Through the use of data sets that contain the minimum (taxonomic) sampling of complete genomes, the effect of outgroup choice in recovering each hypothesis is tested. In an experimental approach, the gene sampling of these data sets is extended to include families that have undergone a history of duplication. Finally, three data sets with a taxonomic sampling designed to contain the broadest range of Bilaterian genomes available are used to investigate on a more realistic scale the effect of outgroup selection in supertree-based phylogenomics.

In Chapter 3, a phylogenomic analysis of the eukaryote phylogeny is performed. In this study, both taxonomic and gene sampling is maximised. Using an experimental approach, I have created a data set that samples all available eukaryotic genomes, to which I have augmented taxon sampling by incorporating a large EST database. As such, genes from approximately 550 species are sampled, spanning all of eukaryotic diversity. By scaling up the approach used in Chapter 2, duplicated genes are additionally included, essentially leading to a gene sampling of over 20,000 protein families. This is arguably the largest phylogenomic data set ever analysed, with a gene sampling that is (approximately) 4 times larger than that used by Pisani et al. (2007); 10 times larger than that of Holton and Pisani (2010; see Chapter 2), and 13 times larger than that of Hejnol et al. (2009).

In Chapter 4, the problem of input tree quality for supertree reconstruction is addressed. This is done by studying shape-related biases in the input trees used in the supertree analyses of Chapter 2. Aside from being fundamental to understanding the problem of data quality in supertree reconstruction, the study of potential biases in input tree topologies is important in a more general sense, as the balance of a phylogenetic tree can disclose certain aspects of macroevolutionary processes. However, such aspects can be masked if the phylogenetic method used to derive the tree, upon which macroevolutionary analyses are based, is influenced by a shape bias, as the shape of a tree is the general criterion used to identify adaptive radiations. I thus used alignments derived in Chapter 2 to examine the four most commonly utilised tree reconstruction methods in modern phylogenetics (and subsequently the most common type of trees used in supertree-based phylogenomics). I then evaluated their relative balance using tree balance specific metrics. In this way, I was able to identify if there is a tree shape bias associated with a given tree reconstruction method.

The findings of these studies and their overall implications are discussed in Chapter 5.

Chapter 2: The animal phylogeny: a comprehensive phylogenomic investigation of alternative hypotheses for the origin of the bilatera

2.1 Introduction

In this chapter, the phylogeny of the Bilateria is investigated in the context of phylogenomics. Two main, alternative, hypotheses for the origin of animals with bilateral symmetry have been proposed: the Coelomata (Hyman, 1940) and Ecdysozoa hypotheses (Aguinaldo et al., 1997). The emergence of molecular phylogenetics has marked a movement away from the long-standing, morphologically supported, Coelomata hypothesis. The new animal phylogeny, or the Ecdysozoa hypothesis has repeatedly, and definitively, been supported by various single and multiple gene studies. Conversely, genomic-scale studies, including supertree studies, have recurrently endorsed the traditional Coelomata topology. Here, by using a suitable outgroup, coupled with a broad taxonomic sampling, I evaluate the last missing piece of evidence in favour of the new animal phylogeny, i.e. the extent to which complete genomic analyses support it.

Results presented in this chapter have been published in Genome Biology and Evolution (see Holton and Pisani, 2010; see also the Publication section of this thesis).

2.1.1 The Bilateria: a morphological context

The Bilateria are metazoans that are typically characterised by bilateral symmetry, a pronounced anteroposterior axis, and a head with a nervous concentration: i.e. a brain (Nielsen, 2001). This group consists of all extant animals, with the exclusion of the sponges, the Placozoa, the Cnidaria and the Ctenophora (see, for example, Nielsen, 2001, Dunn et al., 2008, Philippe et al., 2009, Sperling et al., 2009). Uncertainty still persists pertaining to the origin and early evolution of this important metazoan group. Central to this disparity are the phylogenetic relationships of the "pseudocoelomates" (sensu Hyman, 1940, Hyman, 1951), and particularly that of the Nematoda (i.e. the round worms), which were still subject to debate when this study began (Telford et al., 2008).

Hyman (1940) proposed that bilaterian evolution could be explained through a process in which morphological complexity was achieved *via* a series of incremental steps (see also, for example, Adoutte et al., 2000, Halanych, 2004, Philippe et al., 2005b, Telford et al., 2008). The presence or absence of a hydrostatic skeleton (i.e. the coelom), and the nature of this skeleton when observed, was the main feature over which Hyman's hypothesis was derived. Consequently, Hyman (1940) proposed that the Bilateria should be partitioned into three groups: the Acoelomata (Platyhelminthes, Nemertea and Acoela), the Pseudocoelomata (Nematoda, Nematomorpha, Rotifera, Priapulida, Kinorhyncha, and Gastrotricha; see also Hyman, 1951) and the Coelomata (all other bilaterian phyla, e.g. Arthropoda, Mollusca, Annelida and Vertebrata).

Under this scheme, the less complex Acoelomata, lacking a hydrostatic skeleton, upheld the ancestral organisational condition of the Bilateria (see Hyman, 1940), and were considered the "ancestral stock"[1] from which all other Bilaterians originated. The Pseudocoelomata, which possess a hydrostatic skeleton of blastocoelic origin, but not a "true" body cavity (i.e. a coelom of mesodermal origin), were considered of intermediate complexity and the sister group of the more complex Coelomata. Finally, the Coelomata, possessing a mesodermally derived coelom, were considered the most advanced group of the Bilateria (which are the ancestral stock of the Pseudocoelomata).

[1] At the time of Hyman, phylogenetics was concerned with ancestors. These types of schemes are not used anymore and the search for ancestors has been replaced with the search for sister groups.

Hyman's (1940) Coelomata hypothesis is generally regarded as the "classic textbook phylogeny" of the Bilateria. However, a variety of morphology-based bilaterian phylogenies have been proposed since Darwin's time, with little consensus ever reached among morphologists as to which should be adopted (Jenner and Schram, 1999). Amongst the alternative schemes proposed, that of Grobben (1908) has long been considered an obvious and valid alternative to that of Hyman (1940). According to Grobben (1908), Bilateria were to be split (depending on the fate of their blastopore during development) into two, rather than three, groups: the Protostomia and the Deuterostomia, positioning both Hyman's Pseudocoelomata and Acoelomata within Protostomia.

It is interesting to note that Hyman (1940) explicitly refers to her classification scheme as defining three *organisational grades*, not clades, and in discussing Grobben's (1908) phylogenetic scheme she overtly states that it "… *proved more acceptable and may attain wide adoption*". Indeed, in line with this, Figure 5 of Hyman (1940; see Figure 2.1) adheres to the scheme of Grobben (1908). Despite her theoretical acceptance of Grobben's scheme, Hyman viewed invertebrate phylogenetics as a volatile science: "*Anything said on these questions lies in the realm of fantasy…*" (see Hyman, 1959, see also Valentine, 2004). However, she did consider her *organisational grades* to stand "*firmly on realistic anatomical basis*" (1940).

Accordingly, in outlining the organisation of her encyclopaedic discussion of the Metazoa (Hyman, 1940, Hyman, 1967), she pointed out that in her work she would "*attempt to arrange the phyla in general according to their grade of construction while at the same time avoiding the separation of allied phyla*", displaying her ambivalent position with reference to the grades and clades she describes in Figure 5

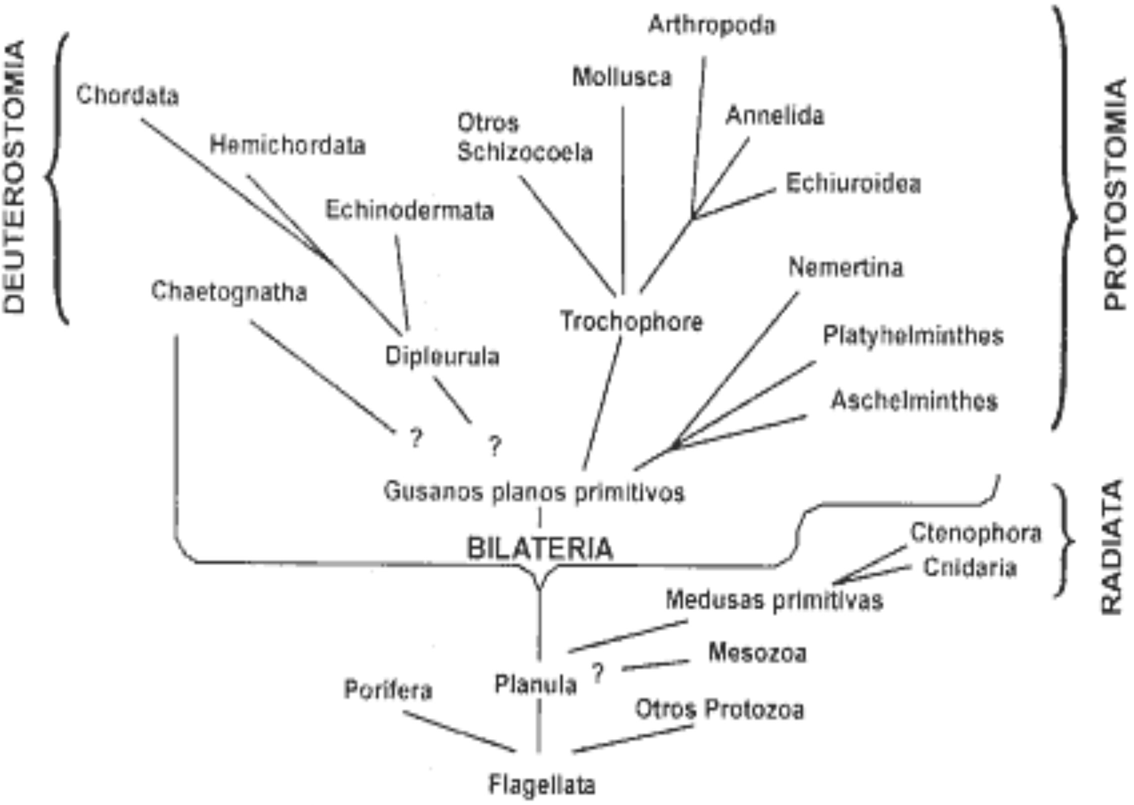


Figure 2.1 **Figure 5 of The Invertebrates (Hyman, 1940).**

The above scheme depicts Hyman's opinion as to the organisation of the Bilaterian phyla. A clear Deuterostome-Protostome split is observed, which is in line with the scheme proposed by Grobben (1908).

(Hyman, 1940).

When Hyman (1940) proposed her Coelomata hypothesis, it thus seems that she did not necessarily intend it to be “a phylogeny”, but rather a robust and convenient classification scheme, i.e. essentially a phenetic classification. Taking this into consideration, it is somewhat surprising that Coelomata has transmuted into the standard “textbook” view of bilaterian evolution. This transformation might well be attributed to the intrinsic (but misleading, see Gould, 1989) appeal of schemes based on incremental evolution (i.e. evolutionary ladders). This is clearly exemplified by Zheng et al. (2007) who stated:

“The Coelomata topology appears ‘natural’ from the viewpoint of the straightforward and intuitive concept of the hierarchy of morphological and physiological complexity among animals, which is the main reason why this phylogeny had been accepted since the work of Haeckel (1866)”.

This statement, which corresponds to the understanding of animal evolution held by most biology graduates, is incorrect in many ways. Coelomata, or rather Coelomera (in Haeckel’s words), was not introduced in Haeckel (1866), but rather in Haeckel (1872). Further to this, Haeckel’s Coelomera did not equate to Coelomata as we now understand it. Haeckel (1872) merely suggested Coelomera to be all animals with a true body cavity, which in any event, was not defined as being mesodermally derived. Indeed, Haeckel (1872) does not feature the Pseudocoelomata, an integral facet of Hyman’s (1940, 1967) Coelomata hypothesis. In addition, as pointed out by Nielsen (2001), the coelom of the protostomes and the deuterostomes cannot be considered homologous structures, as the mesoderm of these groups have disparate origins. The coelom evolved independently, and cannot be regarded as an apomorphy for Coelomata, regardless of whether Coelomata represents the correct hypothesis of Bilaterian evolution or not. It is thus clear that the "natural progressiveness" of the Coelomata hypothesis lies only in the eye of the beholder, and so it is not surprising that several morphological phylogenies of the Bilateria do not support it, e.g. Eernisse et al. (1992) and Schmidt-Rhaesa (2003).

2.1.2 Molecular contention

Morphologists have yet to reach a definitive consensus on the high-level relationships of the Metazoa (see Jenner and Schram, 1999). However, it was not until the completion of the first analyses of taxon-rich 18S rRNA data sets (Halanych et al., 1995, Aguinaldo et al., 1997) that the need for a reassessment of Hyman's "textbook phylogeny" became apparent. The new molecular phylogeny of animals, which soon became known as the "new animal phylogeny" or the "Ecdysozoa" hypothesis, supported a division of the Bilateria in two groups that were fundamentally consistent with the Protostomia and Deuterostomia scheme of Grobben (1908) (excluding the problematic phylum Acoela: Ruiz-Trillo et al., 1999, Littlewood et al., 2001, but see Philippe et al., 2007, Philippe et al., 2011). The 18S rRNA data also proposed a major rearrangement of the protostomes, partitioning them into two new monophyletic groups: the Lophotrocozoa (Halanych et al., 1995) and the Ecdysozoa (Aguinaldo et al., 1997).

Surprisingly, some of the principal emendations supported by the new animal phylogeny had previously been suggested on morphological grounds (Eernisse et al., 1992). These include the dissolution of the Articulata (i.e. the Panarthropoda-Annelida grouping) in favour of a Mollusca-Annelida clade (i.e. Eutrochozoa), and the discovery of a potential relationship between the Arthropoda and several pseudocoelomates (i.e.

Nematoda and Kinorhyncha). The results of Aguinaldo et al. (1997) and Halanych et al. (1995) were thus not totally unforeseen (see Eernisse et al., 1992), yet, many scholars (e.g. Nielsen, 2001) remained reluctant to embrace the possibility that the Arthropoda may be closely related to several of Hyman's (1951) pseudocoelomates (i.e. the Nematoda, the Nematomorpha, the Priapulida and the Kinorhyncha; see, for example, Dunn et al., 2008). Even today, it is widely agreed that Ecdysozoa still lacks robust, unequivocal morphological support. Presently, the only obvious non-molecular character considered to potentially represent an apomorphy for this clade is that they moult, i.e. they undergo the process of ecdysis (Eernisse and Peterson, 2004, Telford et al., 2008).

2.1.3 Phylogenomics and the Bilateria

Great expectations of resolving difficult phylogenetic problems arose from the availability of complete genomes (Gee, 2003) and the consequent emergence of phylogenomics (Eisen, 1998, Delsuc et al., 2005, Philippe et al., 2005a). The genomes of several model organisms, including the arthropod *Drosophila melanogaster* (a coelomate protostome), the vertebrate *Homo sapiens* (a coelomate deuterostome), the nematode *Caenorhabditis elegans* (a pseudoceolomate protostome), and the fungus *Saccharomyces cerevisiae* (a non-metazoan outgroup) have been available for almost a decade. In theory, these genomes should be sufficient to test, at a minimal level, the Ecdysozoa hypothesis within a phylogenomic framework.

To this end, many authors have attempted to assess the new animal phylogeny using genomic-scale data sets, or in any case data sets deemed to be of genomic-scale at the time of their assembly (Blair et al., 2002, Dopazo et al., 2004, Copley et al., 2004, Wolf et al., 2004, Dopazo and Dopazo, 2005, Philip et al., 2005, Zheng et al., 2007, Rogozin et al., 2008, Rogozin et al., 2007). The majority of published deep genomic-scale analyses have failed repeatedly to endorse Ecdysozoa (e.g. Blair et al., 2002, Wolf et al., 2004, Dopazo et al., 2004, Philip et al., 2005, Zheng et al., 2007), supporting Coelomata instead. The only exceptions to this are the studies of Copley et al. (2004) and Dopazo & Dopazo (2005), in which, however, the authors only find moderate, and somewhat unconvincing, support for Ecdysozoa.

Phylogenomic analyses supporting Coelomata present a compelling argument on the basis of the volume of genomic data they consider (Telford et al., 2008). However, studies supporting Coelomata characteristically suffer from a sparse taxonomic sampling (see also Halanych, 2004), which can exacerbate phylogenetic artifacts, particularly long branch attraction (LBA) (Philippe and Laurent, 1998, Pisani, 2004, Delsuc et al., 2005, Philippe et al., 2005b, Jeffroy et al., 2006, Sperling et al., 2009). Interestingly, improved taxon sampling has recurrently been suggested to have a marked effect on accuracy (Hendy and Penny, 1989, Graybeal, 1998, Zwickl and Hillis, 2002, Pollock et al., 2002, Sperling et al., 2009) and indeed has been shown to be successful at resolving controversial groupings (Murphy et al., 2001, Baurain et al., 2007).

Recent studies of bilaterian evolution, conducted using the expressed sequence tag (EST) method (Philippe et al., 2005b, Dunn et al., 2008, Lartillot and Philippe, 2008, Philippe et al., 2009, Hejnol et al., 2009), are characterised by a denser taxon sampling and the use of more appropriate (animal) outgroups and, therefore, should be less prone to LBA. Interestingly, such studies have consistently supported Ecdysozoa, giving further substantiation to the possibility that Coelomata, as recovered by genomic-scale analyses, may be the result of a LBA artifact. However, with the exception of Hejnol et al. (2009), who considered 1487 genes (but only for a very small subset of the taxa they sampled), EST studies represent a shallow genomic sampling (Zilversmit et al., 2002), with Philippe et al. (2005b) considering only 146 genes, Dunn et al. (2008) 150 genes and Philippe et al. (2009) 128 genes. Additionally, EST libraries generated for phylogenetic purposes are generally not normalised (e.g. Dunn et al., 2008, Hejnol et al., 2009), and the protein coding genes sampled in these studies do not represent a random sample of the genes in the considered genomes. Rather, they correspond to a sample of the most highly expressed genes. This non-random sampling is not a problem *per se*, nevertheless, it does pose the question: what will the outstanding proportion of the animal proteome disclose? To date, the answer has often been that standard sequence analyses of deeply sampled genomic data sets favour Coelomata.

EST studies, whilst undoubtedly are of considerable merit, are far from ideal and do not represent an exhaustive coverage of the proteins in the average animal proteome (each study representing ~ 0.1%). Given that such "trees of 0.1%" (*sensu* Dagan and Martin, 2006) of animal genomes seem to support Ecdysozoa, once genomic coverage is extended across a large number of taxa, one wonders what will the remaining 99.9% of the genes disclose.

The strongest test of a phylogenetic hypothesis is one considering all the relevant information (e.g. Kluge, 1989). In phylogenomics, EST studies can maximise taxonomic sampling, whilst studies using complete genomes can maximise gene sampling. Accordingly, I deduce that a reasonable solution to the Coelomata versus Ecdysozoa controversy can only be achieved through the congruence of taxonomically well-sampled EST studies and deep genomic-scale analyses.

2.1.4 Phylogenomics: methodological approaches

From a methodological point of view, two principal approaches are generally employed in phylogenomics: the supertree and the supermatrix approach (Delsuc et al., 2005; see Section 1.3), with both approaches having different strengths and weaknesses.

In the supertree approach, gene trees are recovered for each individual protein family using the most appropriate phylogenetic method. Gene trees are then combined using one of a number of existing supertree methods (for a brief introduction see McInerney et al., 2008). Advantages of the supertree approach include: (1) the ability to analyse each gene individually under the best fitting substitution model. (2) The capacity to amalgamate trees derived from the analysis of both single and multi protein families. (3) A significant decrease in the computational time necessary to build large phylogenies (facilitating the handling of data sets scoring thousands of genes) for hundreds of taxa (e.g. Pisani et al., 2007).

As protein families are first analysed in isolation, the major limitation of the supertree approach is that the combined trees can be based on relatively small alignments. This can result in significant stochastic errors (see Section 1.2.2), which may translate into poorly supported phylogenomic supertrees. Filtering strategies, i.e. eliminating genes that do not pass the Permutation Tail Probability (PTP) test (Archie, 1989; see Section 1.1.3) or that do not support the monophyly of universally accepted clades (Pisani et al., 2007), which also serve to alleviate the negative impact of hidden paralogy when analysing sets of single protein families, can be used to improve resolution significantly.

In the supermatrix approach, single gene alignments are merged into a multiple gene alignment, which is then analysed using the most appropriate phylogenetic method.

The principal merit of this approach is that gene concatenation allows for the minimisation of statistical errors, often resulting in well-supported trees (Delsuc et al., 2005). The main shortcomings of this approach are: (1) whilst it minimises stochastic errors, it tends to exacerbate systematic ones (e.g. Delsuc et al., 2005, Jeffroy et al., 2006). While the use of well-performing, parameter-rich models, like CAT (Lartillot and Philippe, 2004, Philippe et al., 2007), alleviates this problem, it does not fully eliminate it (e.g. Jeffroy et al., 2006, Sperling et al., 2009). (2) The supermatrix approach does not lend itself to the integration of multi protein families and, as such, limits the information that can be analysed to that of single protein families, or in some rare cases (i.e. when the gene phylogeny is well understood) to single paralogy groups within a multi gene family (e.g. Dunn et al., 2008, Philippe et al., 2009, Hejnol et al., 2009). (3) If the number of considered genes, or species, or both is considerably large, supermatrix analyses become very difficult to perform due to computer memory and time constraints (see, for example, Hejnol et al., 2009). Technological advances should ameliorate this problem, but this limit of the supermatrix approach can be expected to persist for the foreseeable future.

2.1.5 Circumventing long-branch attraction

LBA (Felsenstein, 1978) is a common phylogenetic artifact (Brinkmann and Philippe, 1999, Pisani, 2004, Delsuc et al., 2005, Jeffroy et al., 2006), which can affect every phylogenetic method (Pisani, 2004, Delsuc et al., 2005, Jeffroy et al., 2006). Since time and rate are confounded in branch length estimation (e.g. Yang, 2006), LBA results in trees in which fast-evolving species are artifactually grouped together, or with distantly related taxa (e.g. with the outgroups). Two straightforward approaches employed to reduce LBA are optimal outgroup selection (to minimise root to tip distances in a phylogeny), and increased taxon sampling (to break long branches; see Pisani, 2004).

Early, deep genomic-scale analyses used fungal outgroups, or on occasion even more distantly related outgroups. For example, Blair et al. (2002) use the plant *Arabidopsis thaliana*, a selection that is counterintuitive, especially given that elsewhere these authors find strong support for the idea that animals and fungi are sister taxa (Hedges et al., 2004). Fungi and plants clearly represent poor choices to investigate the phylogenetic relationships of the Bilateria as they may serve to exacerbate LBA.

Dopazo and Dopazo (2005) performed standard sequence analyses of a deeply sampled genomic data set using a distant (fungal) outgroup. Realising that a fungal outgroup might not have been adequate for their analyses, and in the absence of a closer outgroup, these authors used a relative rate test (for an overview see Robinson et al., 1998) based approach to identify clock-like genes. Analysis of these genes found support for Ecdysozoa. Although their results are interesting, their approach is not without problems. Firstly, the relative rate test is not particularly sensitive; a more discriminating approach (i.e. the likelihood ratio test) should have been used instead. In addition, their relative rate tests were implemented under the simplistic Kimura's distance in PROTDIST (Felsenstein, 2005), which is unlikely to be a good fit to their data. Finally, these authors considered only homologues of protein coding genes found in 18 human chromosomes, unnecessarily discarding potentially informative genes not found in this subset of human chromosomes.

The number of complete animal genomes has now increased significantly making the improvement of taxonomic sampling in genomic-scale phylogenetic analyses possible. Recent genome sequencing projects have included that of the cnidarian

Nematostella vectensis and the placozoan *Trichoplax adherans.* While there is ongoing debate over the relative phylogenetic relationships of these two organisms, there is general agreement that both are non-bilaterian Metazoans (see Nielsen, 2001, Dunn et al., 2008, Philippe et al., 2009, Sperling et al., 2009, Hejnol et al., 2009, Pick et al., 2010). Accordingly, *N. vectensis* and *T. adherans* represent more appropriate outgroups for testing hypotheses of bilaterian evolution than fungal outgroups (see also Philippe et al., 2005b). Therefore, here I have avoided such gene selection strategies (e.g. Dopazo and Dopazo, 2005, Copley et al., 2004), focusing instead on taxonomic sampling and outgroup selection to test hypotheses of bilaterian evolution, using the largest possible number of protein families.

2.1.6 Experimental phylogenomics and data set assembly

Rather than simply collecting all available animal genomes and reconstructing yet another metazoan phylogeny, I took an experimental approach. I made the following *ad hoc* (working) assumption: Coelomata is the true tree and not the result of LBA (my null hypothesis). I then predicted what the consequences of this null hypothesis would be, selected a suitable set of complete genomes, and tested whether the predictions derived from my assumption could be met. If my predictions were to be upheld by the data, the null hypothesis was not to be rejected, whilst if overturned, the data would reject the null hypothesis. I finally used the data to test whether my results provided support for the most obvious alternative hypothesis (i.e. Ecdysozoa).

Based upon my working assumption, I predicted that, under the null hypothesis, in a sparsely sampled (four taxon) data set, Coelomata should invariably be recovered, irrespective of whether a distant (fungal) or closer (animal) outgroup was used. Therefore, if Coelomata was, indeed, the result of a LBA artifact, it would be recovered only when using a divergent outgroup. I further predicted (again based on the postulation that Coelomata is the *bona fide* tree), that the Coelomata topology should also be recovered in the presence of an extensive taxonomic sampling, irrespective of the outgroup used. Thus, if Coelomata is a LBA artifact, it should not be recovered if a targeted sampling strategy is adopted in order to break the long-branch connecting the fungal outgroup with the Bilateria (this can be done by including *N. vectensis* and *T. adherans* in the analyses), or when animal outgroups (i.e. *N. vectensis* and *T. adherans*) are used instead of the fungal outgroup.

I assembled 2 sparsely sampled data sets (scoring four taxa) to investigate, at the most fundamental level, the effect of outgroup choice and taxon sampling on genome scale phylogenies. In the four-taxon case, I show, using genomic data, that Coelomata is only recovered when the fungal outgroup *Saccharomyces cerevisiae* is used. Use of a less divergent animal outgroup (i.e. *N. vectensis*), in an analogous data set, results in the recovery of Ecdysozoa. Furthermore, I assembled three extensively sampled and more convincing data sets, and show that Coelomata is only recovered if the branch connecting the fungal outgroup with the Bilateria is not broken (i.e. when no non-bilaterian Metazoans are included in the analyses and a fungus is used as an outgroup) [2]. Upon inclusion of a non-bilaterian Metazoan in the analyses, support is lost for Coelomata and is garnered for Ecdysozoa. These results hold true for data sets that incorporate up to 2,216 genes, in addition to data sets that consider both single and multi protein families

[2] For each of the two types of data set, sparsely sampled and densely sampled, materials and methods, and results are split accordingly. These are followed by a general discussion encompassing the results of both data set types.

(i.e. these results are upheld when 100% of the phylogenetically informative proteins in the considered proteomes are used).

2.2 Sparsely sampled data sets

2.2.1 Materials and Methods

2.2.1.1 Data collection and data set assembly

Complete genomic data for *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens* were downloaded from COGENT (http://maine.ebi.ac.uk:8000/services/cogent/). Additionally, the complete genomes of the fungus *Saccharomyces cerevisiae* (sourced from COGENT) and the cnidarian *Nematostella vectensis* (sourced from DOE Joint Genome Institute) were collected for use as alternative outgroups. Two sparsely sampled data sets, containing the following core species: *H. sapiens*, *D. melanogaster* and *C. elegans,* were assembled. In addition to the three core species, the first data set included *S. cerevisiae*, as an outgroup. As fungi are evolutionarily distant from the Bilateria, *S. cerevisiae* was selected as an outgroup to investigate the combined effect of sparse taxon sampling and poor outgroup choice. To properly test the effect of outgroup choice on the results of our analyses, *N. vectensis* was used as outgroup for the second data set. Cnidarians are the most likely sister group of Bilateria (see, for example, Philippe et al., 2009, Sperling et al., 2009, but see also Pick et al., 2010), and thus represent an optimal outgroup to be used to study the phylogeny of Bilateria.

2.2.1.2 Protein family identification

Due to their small dimensions, the two sparsely sampled data sets were amenable to the comparison of two alternative protein family identification strategies. Firstly, the BLASTP based, all-versus-all approach of Creevey et al. (2004), Fitzpatrick et al. (2006) and Pisani et al. (2007) was implemented to cluster homologous protein families. Under the strategy of Creevey et al. (2004), protein families are isolated by sequentially selecting a random seed sequence from a database, scoring all considered genomes, and identifying all the homologs of that sequence. Once the homologs of a seed sequence are identified, they are removed (together with the seed) from the database. This process is repeated until all sequences are assigned to a putative protein family, at which stage the database will be empty. This protein family identification strategy is heuristic in nature, but has the advantage of having fast implementation, as the size of the searchable database decreases with each BLASTP-iteration. Its heuristic nature notwithstanding, it has been shown previously (see Pisani et al., 2007) that this approach performs better than, for example, the single-linkage clustering algorithm implemented in BLASTClust (Dondoshansky and Wolf, 2000).

To validate my results, and further assess the performance of the Creevey et al. (2004) strategy, I additionally tested this strategy against the approach of Enright et al. (2002) based upon the Markov clustering algorithm (MCL). The MCL procedure identifies protein families using a method that was originally developed for graph clustering using flow simulation. Under the MCL approach, an all-versus-all BLAST is first carried out to establish sequence similarity relationships, which are then represented in a graph. Clusters (of sequence similarity) are then identified by the occurrence of a large number of shared connections between sequences. This is computed by a series of random walks (connections) through the graph. A large number of random walks through an area is indicative of the presence of a protein family. This is because a random walk through a member of a protein family is more likely to continue within the same family, rather than move to an entirely different protein family. BLASTP searches required by both the Creevey et al. (2004) and MCL approaches were performed using an E-value cut-off of 10^{-8}.

Following each type of homology search, each data set was partitioned into two groups. Families scoring only one member for any given genome (i.e. putative single protein families) were separated from those containing multiple members per genome (i.e. the multi protein families). Since phylogenetic analyses can only be performed on protein families that score four or more sequences, only single and multi protein families consisting of a minimum of four sequences were retained for further analysis. Typically, only single protein families are used for phylogenetic reconstruction (e.g. Pisani et al., 2007, Hejnol et al., 2009). This is to minimise the complexity associated with the analysis of multi gene data sets and the inclusion of signals representing the relationships of paralogous genes.

However, this approach has the disadvantage of considering only a minority of the genes in the genomes, whilst the strongest test of a phylogenetic hypothesis is one considering all relevant information (e.g. Kluge, 1989). Only upon the integration of multi protein families can such a test be performed. Here, by exploiting the flexibility of the supertree approach, I have combined both single and multi protein families to generate trees based on the deepest possible sample of genomic data. Owing to the dimensions of the sparsely sampled datasets, these were once again selected as exemplar cases to examine the feasibility of this approach. However, following on from this analysis I have extended the integration of multi protein families into data sets with a larger taxonomic sampling (see Chapter 3).

2.2.1.3 Alignment, curation and phylogenetic analysis

All considered single and multi protein families were aligned using ClustalW (Thompson et al., 1994). As the accuracy of this traditional multiple sequence alignment algorithm has been questioned (e.g. Löytynoja and Goldman, 2008), single and multi protein families in the 4-taxon data sets were also aligned using PRANK (Löytynoja and Goldman, 2008). This was done to investigate whether alignment dependent biases (Löytynoja and Goldman, 2008) influenced the results. As aligning sequences using PRANK is computationally expensive, the tractability of the 4-taxon data was once again utilised.

Due to the number of protein families obtained from the data sets, manual curation of alignments was unfeasible. Accordingly, Gblocks (Castresana, 2000) was used to eliminate highly variable, and potentially misaligned regions. Gblocks parameters were set as follows: gapped positions were not eliminated, the minimum block length was set to 8 amino acid positions, while the maximum number of permitted consecutive non-conserved positions was set to 15 (see also Pisani et al., 2007). Curated alignments were then subjected to the PTP test (Archie, 1989). This allowed the identification of families conveying significant hierarchical signal (see also Pisani et al., 2007). Such families were considered to contain sufficient hierarchical structure to be deemed potentially phylogenetically informative (obviously this signal could represent a bias, but the PTP test cannot be used to determine this; see Section 1.1.3). The PTP test was implemented in PAUP4.0b10 (Swofford, 1998). Settings were as follows: 2,000 permutations with heuristic search, with one random addition sequence and the MulTrees option set to off. For the PTP test, a probability value $P \leq 0.05$ was considered significant. Alignments not passing the PTP test ($P \geq 0.05$) were disregarded, as they would not contribute anything except noise to the analyses.

PHYML (Guindon and Gascuel, 2003) was used to perform Maximum Likelihood (ML) phylogenetic analyses of each alignment passing the PTP test. ML analyses were performed under the best fitting substitution model, as inferred using the Akaike Information Criterion in Modelgenerator (Keane et al., 2006). For each single and multi gene family tree, support was evaluated using bootstrap (100) replicates.

2.2.1.4 Deriving phylogenomic supertrees for the 4-taxon data sets

Consensus tree methods allow the combination of fully overlapping input trees (i.e. trees on the same leaf set). Examples include the majority rule consensus tree method of Margush and McMorris (1981) which, given a set of input trees, includes all the splits that are present in a minimal (*a priori* defined) number of input trees (e.g. 50%). For each of the final, 4-taxon data sets (eight in total arising from alternative homology assessment and alignment procedures), phylogenomic consensus trees were derived. These were built using (1) the set of all single protein families, (2) the set of all multi protein families and (3) the combined set of all single and multi protein families. Accordingly, a total of twenty-four, 4-taxon, phylogenomic trees were derived. Table 2.1 and Figure 2.2 report the number of genes used to build each of these trees.

Table 2.1 Progression of protein family numbers at each stage of the analysis.

All data sets and protein families were subjected to the same protocol. GTP-PTP = Gene Tree Parsimony Permutation Tail Probability Test

Fungal Outgroup									
Single Gene Families				Multi Gene Families					
Homology Search	No. single gene families	No. of families with > 4 taxa	No. families passing PTP	No. multi gene families	No. of families with > 4 taxa	No. families passing PTP	Species level trees > 4 taxa	No. families passing GTP-PTP	No. families Consensus Tree
Creevey et al. (2004)	16780	201	30	7947	4197	3301	917	258	258
MCL	6588	254	28	8529	5312	4143	1366	392	392
Animal Outgroup									
Single Gene Families				Multi Gene Families					
Homology Search	No. single gene families	No. of families with > 4 taxa	No. families passing PTP	No. multi gene families	No. of families with > 4 taxa	No. families passing PTP	Species level trees > 4 taxa	No. families passing GTP-PTP	No. families Consensus Tree
Creevey et al. (2004)	18094	314	48	10146	5269	4328	1923	516	516
MCL	6254	332	40	9808	6561	4666	2319	682	682

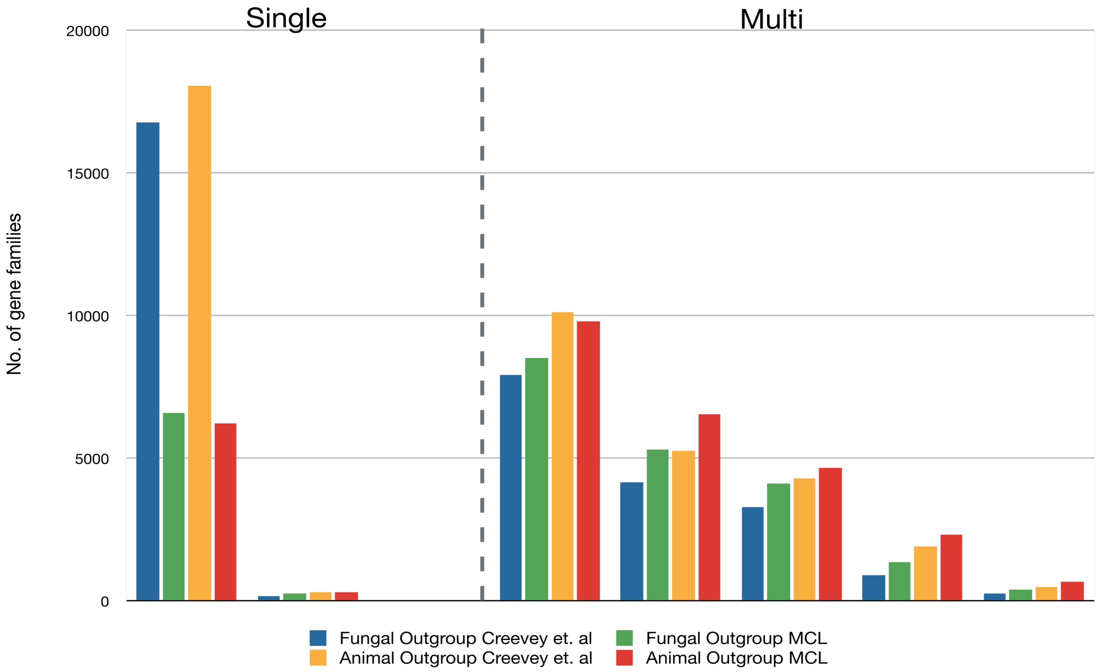


Figure 2.2 Comparison of protein family numbers at each stage.

Graphical representation of Table 2.1. The graph on the left represents single protein families, whilst the one on the right represents multi protein families.

Each of the eight, single protein family based, 4-taxon phylogenomic trees (see below) were built as follows: (1) the 100 bootstrap ML trees generated for each single protein family in that data set were pooled to generate a single bootstrap tree file. (2) The trees in the pooled, bootstrap tree file were summarised using the majority rule consensus tree method (Margush and McMorris, 1981), as implemented in the software Consense (Felsenstein, 2005). This was possible as all considered bootstrap trees were on the same taxon set (i.e. they were fully overlapping). As these phylogenomic trees were derived from pooling trees obtained from the individual bootstrap replicates, assessment of the support for the clades in these trees was straightforward because the 4-taxon phylogenomic trees were also bootstrap consensus trees.

Each of the 8 multi protein family based phylogenomic trees were derived as follows: (1) for each considered multi protein family, 100 bootstrap ML trees were used to generate reconciled species trees. This was done using the duplication only, Gene Tree Parsimony (GTP) method (e.g. Cotton and Page, 2004) as implemented in the software DupTree (Wehe et al., 2008), with the nogenetree option turned on, using a partial queue based heuristic search (see Figure 1.3 for an exemplar multi protein family and the corresponding GTP derived species tree). (2) The resulting species trees (one per bootstrap ML tree) were pooled into a single file. (3) The pooled, bootstrap (species)-trees were summarised using the majority rule consensus method (as implemented in the software Consense; Felsenstein, 2005), thus generating a bootstrap consensus phylogenomic tree. Also in this case, the use of the majority rule consensus method could be implemented, as all the bootstrap species trees were on the same taxa set.

Each of the 8 combined multi and single protein family phylogenomic trees were derived as follows: (1) the corresponding sets of individual bootstrap trees (obtained from the ML analyses of the single protein families), and the species trees derived from the DupTree analysis of the bootstrap trees from the multi protein families (see above) were pooled into a single file. Trees in the pooled file were summarised using the majority rule consensus method, to derive a bootstrap consensus phylogenomic tree.

2.2.1.5 Gene tree parsimony permutation tail probability test

Not all of the multi protein families were used for phylogenetic reconstruction (i.e. some families, despite passing the PTP test, were not deemed viable). An additional Permutation Tail Probability (PTP) test was developed, to evaluate whether the duplication history of each considered multi protein family was phylogenetically informative. To implement the Gene Tree Parsimony PTP test (GTP-PTP), for each optimal multi protein family tree derived using PHYML, 100 permuted trees were generated. This was done by randomly swapping the labels associated with the terminal nodes of the optimal multi protein family tree, whilst maintaining the unlabelled phylogenetic history as fixed. This is effectively a variant of the YAPTP test of Creevey et al. (2004).

Each permuted tree was used to infer a species phylogeny using the GTP method (as implemented in DupTree). The score of each GTP reconstruction was recorded, and these values were compared against the GTP score of the species history derived from the original (unpermuted) multi protein family tree. Families were retained for phylogenetic analysis when the species history derived from the unpermuted tree was significantly more parsimonious than those obtained from the GTP analysis of the permuted trees. For these analyses, the significance level was set to $P \leq 0.01$. To facilitate the implementation of the GTP-PTP a number of PERL scripts were written (see Electronic Appendix). It is clear that the species phylogeny embedded in multi protein families failing to pass the PTP-GTP test has essentially been erased due to a complex gene deletion/duplication history. These multi protein families can only contribute noise to the analyses and were thus not used for phylogenetic reconstruction.

2.2.2 Results

2.2.2.1 Methodological examinations

The four species data sets afforded the opportunity to test various methodological aspects of phylogenomic studies. Firstly, they provided for a direct comparison between two alternative homology assignment procedures. From Figure 2.1, it can be seen that there is little difference between the Creevey et al. (2004) and MCL based approaches with respect to the ultimate number of protein families deemed viable for phylogenetic reconstruction. This trend is upheld throughout each preceding stage of the analysis (see Figure 2.2 and Table 2.1). There is, however, a distinct difference between the total number of protein families identified by each approach, with the Creevey et al. (2004) approach consistently isolating far more protein families (almost 3 times as many as MCL in the case of the data set containing *N. vectensis*; see Figure 2.2 and Table 2.1). This can be attributed to the "seed" sampling strategy of the Creevey et al. (2004) approach, which results in a more modular approach to finding clusters of homologous protein families, unlike that of MCL which results in an exhaustive search of clusters.

The performance of two alternative multiple sequence alignment strategies, namely, ClustalW and PRANK, were also evaluated. The proficiency of each method was measured based upon the support values their alignments attained from supertree construction. From Table 2.1, it can be seen that PRANK seems to perform slightly better, which is similar to what is observed in the comparison of homology assessment protocols. While these differences are irrelevant (support values for each combination of approaches are consistent) for the data sets considered here, for larger datasets this difference may prove to have a greater bearing.

2.2.2.2 Phylogenetic analysis

The four species data sets were analysed to assess, at a very basic level, the effect of outgroup selection in phylogenomics. The first interesting result obtained from these analyses was that only a somewhat diminutive number of single protein families, conveying a significant amount of phylogenetic information, could be identified (see Table 2.1). This was not fully unforeseen, as the stringency of the PTP test increases as the number of considered species decreases. More families were found when *N. vectensis* was used as an outgroup instead of *S. cerevisiae*, however, the difference was negligible (from 31 to 48). The number of single protein families passing the PTP test in the 4-taxon data sets did not change significantly when either an alternative homology assignment strategy or alignment software were used (see Table 2.1), suggesting that the small number of single protein families arising from these analyses does not stem from methodological biases. It merely implies, that when only 4 taxa are considered, there are very few, universally distributed single protein families conveying significant phylogenetic information pertinent to testing hypotheses of bilaterian relationships.

The number of multi protein families (see Table 2.1) passing all the quality checks is also quite low, but significantly higher than the equivalent number of single protein families. This was to be expected, as there are far more multi protein families than single protein families in the average animal genome. However, interestingly, it is noted that while the number of phylogenetically informative multi protein families identified if *S. cerevisiae* is used as outgroup is 258 (using the homology assessment strategy of Creevey et al., 2004) or 392 (using MCL), the number of phylogenetically informative multi protein families identified when *N. vectensis* is the outgroup is 516 (using the Creevey et al., 2004 homology assessment strategy) or 682 (using MCL), i.e. approximately twice as many. This strongly implies that using a closer outgroups is key to maximising the amount of phylogenetic information and increasing the signal to noise ratio in phylogenomic data sets.

Phylogenomic trees derived from single protein families, passing the PTP test, showed that when *S. cerevisiae* was used as an outgroup, support was found for Coelomata (see Figure 2.3). This result holds true irrespective of the protein family identification method used, and of the alignment software used (see Figure 2.3 and Table 2.2). When only multi protein families are used, similar results are found, although there is a significant decrease in the level of support observed (Figure 2.3 and Table 2.2). Finally, in the phylogenomic trees obtained when both the single and the multi protein families were considered concurrently, the support for Coelomata ranges between 55% and 61% depending on the clustering method and alignment software used (Figure 2.3 and Table 2.2). This represents a marked decrease in the support for Coelomata. Similar results were obtained in the study of Philippe, Lartillot and Brinkmann (2005b), although based solely on single protein families.

When the cnidarian *N. vectensis* is used as an outgroup, Coelomata is no longer recovered. Instead, a nematode-arthropod clade emerges, supported most strongly in the

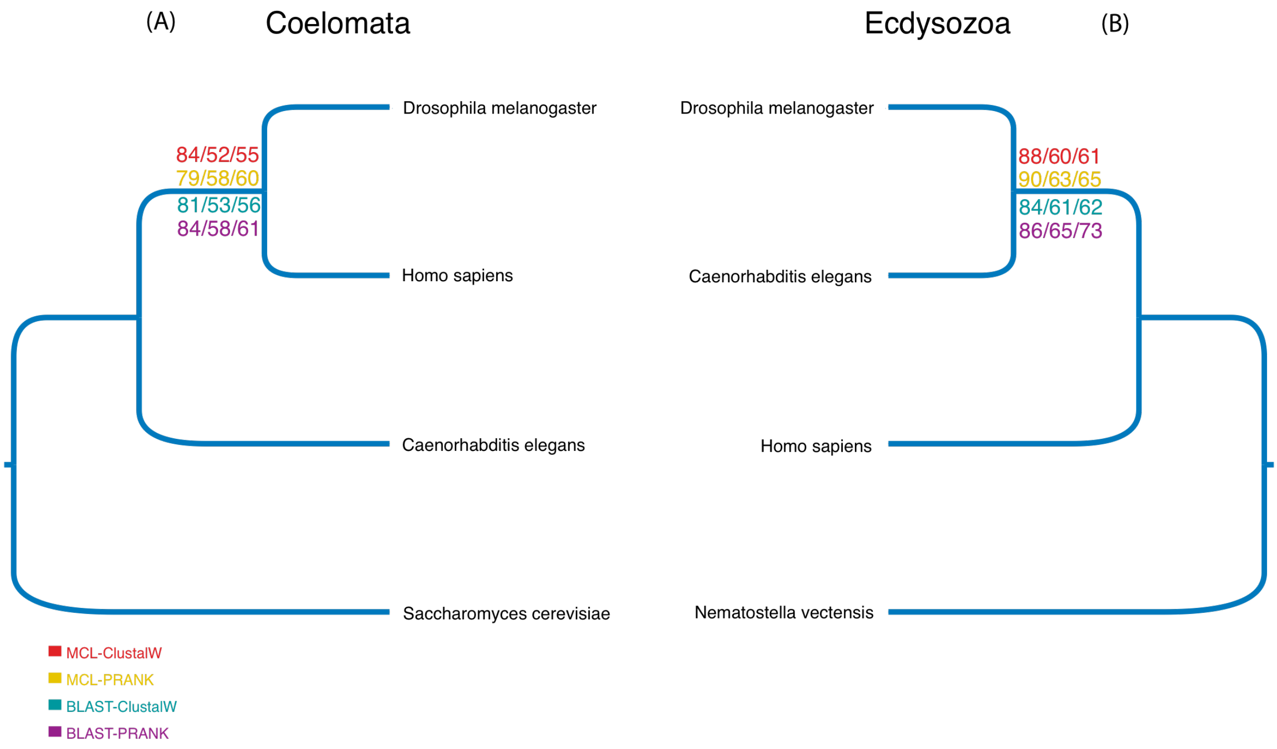


Figure 2.3 Testing outgroup choice in minimally sampled data sets.

Majority rule consensus trees derived from ML input trees. Bootstrap support from both multi and single protein families is shown for each node. The following core ingroup species are common to all: *Homo sapiens*, *Drosophila melanogaster* and *Caenorhabditis elegans*. Outgroups used are (A) the yeast *Saccharomyces cerevisiae* (B) the cnidarian *Nematostella vectensis*. Bootstrap support values are shown for each combination of protein family identification and alignment method. Bootstrap support is displayed for single protein families, multi protein families and combined single and multi protein families respectively.

Table 2.2 Support for alternative hypotheses

The percentage bootstrap support for each hypothesis (Coelomata, Ecdysozoa or the alternative topology) arising from the analysis of the sparsely sampled data sets.

Fungal Outgroup							
Alignment Protocol:		Clustal W			PRANK		
Homology Search	Gene Families	Coelomata	Ecdysozoa	Vertebrata-Nematoda	Coelomata	Ecdysozoa	Vertebrata-Nematoda
Creevey et al. (2004)	Single	81	9	10	84	6	10
	Multi	53	26	21	58	23	19
	Single + Multi	56	24	20	61	20	19
MCL	Single	84	6	10	79	7	14
	Multi	52	26	21	58	22	20
	Single + Multi	55	25	20	60	20	20
Animal Outgroup							
Alignment Protocol:		Clustal W			PRANK		
Homology Search	Gene Families	Coelomata	Ecdysozoa	Vertebrata-Nematoda	Coelomata	Ecdysozoa	Vertebrata-Nematoda
Creevey et al. (2004)	Single	14	84	2	6	86	10
	Multi	21	61	18	18	65	19
	Single + Multi	20	62	17	13	73	19
MCL	Single	9	88	3	7	90	14
	Multi	21	60	19	19	63	20
	Single + Multi	20	18	18	18	65	20

analysis of the single protein families (BS=90%; Figure 2.3 and Table 2.2). Support for Ecdysozoa arising from the analysis of single and multi protein families, both in isolation and when combined, ranges from 60% to 90% (Figure 2.3 and Table 2.2). In the analysis of the single protein families, the support for this clade increases, as more efficient and accurate clustering techniques and alignment software are used (see Figure 2.3). It is interesting to note that, this trend is antithetic to what is observed when *S. cerevisiae* is used as outgroup (see Figure 2.3). Support for Coelomata decreases when less heuristic protein family recognition and alignment methods are used, whilst under the same conditions support increases for Ecdysozoa. It is thus probable that misalignment and incorrect protein family identification may have also contributed to the support found for Coelomata in previous studies.

It is important to note that, when multi protein families are used, a general decrease in support is observed for the nodes in the recovered trees, irrespective of whether a fungal or animal outgroup is used. This suggests that multi protein families contain more noise than single protein families. Or more likely, that the approach used to infer species trees from the multi protein family trees (i.e. duplication only GTP) is not ideal and cannot completely eliminate the paralogy signal. It is to be expected that the development of more refined methods (for example probabilistic methods such as Arvestad et al., 2003) for inferring species trees from multi protein family trees will alleviate this problem in the future.

Analyses of the 4-taxon data sets illustrate that when a closer outgroup is used, sequence analyses with a deep genomic sampling support Ecdysozoa. Conversely, Coelomata is found only when a distant outgroup is used, thus failing to uphold the predictions made at the beginning of this analysis. The recovery of Coelomata can be better viewed as inconsistent (i.e. “strongly supported but erroneous”; Philippe et al., 2005b), arising from the selection of a distant outgroup. In the presence of a distantly related outgroup like *S. cerevisiae* (which probably shared a last common ancestor with the Bilateria one billion years ago; see Peterson et al., 2008, Sperling et al., 2010), the rapidly evolving nematode *C. elegans* is placed at the base of the tree, close to the outgroup. When in its stead, a closer outgroup (*N. vectensis*), which probably shared a last common ancestor with the Bilateria only ≈ 670 million years ago (Peterson et al., 2008, Sperling et al., 2010) is used, *C. elegans* emerges as the sister group of the arthropod *D. melanogaster,* and thus as an ecdysozoan. This strongly implies that the recovery of Coelomata is the result of a tree reconstruction artifact.

2.3 Densely sampled data sets

2.3.1 Materials and Methods

2.3.1.1 Data Collection and data set assembly

Genomic data for 43 eukaryotic species were downloaded from COGENT (http://maine.ebi.ac.uk:8000/services/cogent/), DOE Joint Genome Institute (http://genome.jgi-psf.org/), EMBL-EBI IPI (http://www.ebi.ac.uk/IPI/IPIhelp.html), Ensembl (http://www.ensembl.org/info/data/ftp/index.html), and NCBI (ftp://ftp.ncbi.nih.gov/genomes/). Using this data, 3 intersecting data sets were compiled. These data sets scored between 41 and 43 species, a full list of which can be seen in Table 2.3.

Table 2.3 Taxonomic Sampling

A list of the 43 genomes used in this study and where they were sourced.

Species	Taxonomy	No. of protein sequences	Source
Homo sapiens	Mammalia	33869	Cogent
Loxodonta africana	Mammalia	15717	Ensembl
Loxodonta africana	Mammalia	15717	Ensembl
Macaca mulatta	Mammalia	36423	Ensembl
Macaca mulatta	Mammalia	36423	Ensembl
Monodelphis domestica	Mammalia	32612	Ensembl
Monodelphis domestica	Mammalia	32612	Ensembl
Mus musculus	Mammalia	25371	Cogent
Mus musculus	Mammalia	25371	Cogent
Myotis lucifugus	Mammalia	16233	Ensembl
Myotis lucifugus	Mammalia	16233	Ensembl
Ornithorhynchus anatinus	Mammalia	27473	Ensembl
Ornithorhynchus anatinus	Mammalia	27473	Ensembl
Oryctolagus cuniculus	Mammalia	15439	Ensembl
Oryctolagus cuniculus	Mammalia	15439	Ensembl
Otolemur garnetti	Mammalia	15449	Ensembl
Otolemur garnetti	Mammalia	15449	Ensembl
Pan troglodytes	Mammalia	33167	Ensembl
Pan troglodytes	Mammalia	33167	Ensembl
Rattus norvegicus	Mammalia	33438	Ensembl
Rattus norvegicus	Mammalia	33438	Ensembl
Sorex araneus	Mammalia	13195	Ensembl
Sorex araneus	Mammalia	13195	Ensembl
Spermophilus tridecemlineatus	Mammalia	14831	Ensembl
Spermophilus tridecemlineatus	Mammalia	14831	Ensembl
Tupaia belangeri	Mammalia	15462	Ensembl
Tupaia belangeri	Mammalia	15462	Ensembl

Gallus gallus	Aves	25680	EMBL-EBI
Xenopus tropicalis	Amphibia	27916	DOE Joint
Danio rerio	Actinopterygii	31743	Ensembl
Gasterosteus aculeatus	Actinopterygii	27581	Ensembl
Oryzias latipes	Actinopterygii	25107	Ensembl
Takifugu rubripes	Actinopterygii	26721	DOE Joint
Tetraodon nigroviridis	Actinopterygii	28005	Ensembl
Ciona intestinalis	Urochordata	15852	DOE Joint
Ciona savignyi	Urochordata	20143	Ensembl
Capitella sp. I	Annelida	32415	DOE Joint
Helobdella robusta	Annelida	23432	DOE Joint
Lottia gigantea	Mollusca	23851	DOE Joint
Aedes aegypti	Arthropoda	16789	Ensembl
Apis mellifera	Arthropoda	9900	NCBI
Daphnia pulex	Arthropoda	30940	DOE Joint
Drosophila melanogaster	Arthropoda	18484	Cogent
Drosophila pseudoobscura	Arthropoda	9878	EMBL-EBI
Caenorhabditis briggsae	Nematoda	19507	Cogent
Caenorhabditis elegans	Nematoda	19957	Cogent
Trichoplax adhaerens	Placozoa	11520	DOE Joint
Nematostella vectensis	Cnidaria	27273	DOE Joint
Saccharomyces cerevisiae	Fungi	6357	Cogent

Each data set shared a common set of 40 species; corresponding to all the complete bilaterian genomes available at the time this study was undertaken. To each data set alternate outgroups were added. Outgroups used were as follows: for data set (1) *Saccharomyces cerevisiae,* data set (2) *Trichoplax adherans* and *Nematostella vectensis* and data set (3) *S. cerevisiae, T. adherans and N. vectensis*.

2.3.1.2 Protein family selection and phylogenetic analysis

Homologous protein families for each of the three data sets were identified using the Creevey et al. (2004) approach discussed above (see 2.2.1.2), under the same parameters. From the resultant homologous protein families, like in many standard phylogenomic studies (e.g. Pisani et al., 2007, Hejnol et al., 2009), only single protein families were selected for further analysis (see Table 2.4 for the number of families in each of the considered data sets at each stage of the analysis). As with the sparsely sampled data sets, only families that contained at least four species were suitable for phylogenetic construction. All considered single protein families were subject to the same protocol as used for the single protein families of the 4-taxon data sets (see Section 2.2.1). Multiple sequence alignment was carried out using ClustalW (Thompson et al., 1994) and the subsequent alignments were curated using Gblocks (Castresana, 2000). The PTP test was then carried out on the alignments, followed by model selection for those that passed, both as above. Finally, PHYML (Guindon and Gascuel, 2003) was used to perform Maximum Likelihood (ML) phylogenetic analyses on the remaining alignments, with support evaluated using bootstrap (100) replicates. Single protein trees were manually inspected to evaluate possible instances of hidden paralogy; trees that failed to recover the monophyly of uncontroversial, universally accepted groups (e.g.

Table 2.4 Progression of protein family numbers at each stage of analysis.

All data sets and protein families were subjected to the same protocol.

Large Data Sets			
Outgroup	No. single gene families	No. of families with more than 4 taxa	No. families passing PTP
S. cerevisiae	82043	3241	2164
S. cerevisiae, N.vectensis and *T.adherans*	88858	3304	1949
N.vectensis and *T.adherans*	86855	3615	2216

Vertebrata or Arthropoda) were excluded from further analyses (see also Pisani et al., 2007).

2.3.1.3 Supertree reconstruction

Supertrees represent a generalisation of the consensus tree problem, in the case of partially, rather than fully overlapping trees (Semple and Steel, 2003). Since genes do not have a universal distribution, in the case of the 41, 42 and 43 species data sets, single-protein families could score in the range of 4 to 41, 4 to 42, or 4 to 43 sequences respectively. That is, unlike the 4-taxon data sets, single protein family trees in these data sets are partially, rather than fully, overlapping. Accordingly, gene trees derived from protein families identified in these larger data sets could not be summarised using a standard consensus method. Instead, a supertree approach was used to derive phylogenomic trees.

For each of the three densely sampled data sets, consensus supertrees were generated as follows: (1) the bootstrap trees obtained from the ML analysis of each considered single protein family were pooled into one single data set. (2) Input tree bootstrapping (Creevey et al., 2004, Burleigh et al., 2006, Moore et al., 2006, Pisani et al., 2007) of the pooled trees was used to generate 100 pseudoreplicate data sets. (3) For each pseudoreplicate data set, supertrees were derived using the matrix representation with parsimony (MRP) method (Baum, 1992, Ragan, 1992). To do so, for each pseudoreplicate data set, a standard MRP matrix was generated using CLANN (Creevey and McInerney, 2005). This matrix was then analysed using maximum parsimony in PAUP (Swofford, 1998) to generate the MRP supertrees. For the parsimony analysis 100 heuristic searches were performed with random sequence addition and TBR branch swapping. (4) The supertrees derived from the analysis of each pseudoreplicate data set were summarised using the majority rule consensus method, generating a majority rule consensus genomic supertree, in which support for the clades recovered was expressed as their percentage bootstrap support.

Due to the known limitations of MRP (see Section 1.3.2.1), an additional supertree method was implemented. The quartet joining method of Wilkinson and Cotton (2006) was selected (see Section 1.3.2.2). This method is expected to perform well where there is no conflict. However, results from the analysis of two (of the three) densely sampled data sets indicate that this method (at the least in its current implementation) is unable to cope with the varying signals in the Bilateria (see Figure 2.4), therefore, use of this method was discontinued.

2.3.1.4 Supermatrix analysis

In addition to the supertree analysis for each of the 41, 42, and 43 taxa data sets, a superalignment of the single protein families that passed the PTP test was generated, using a PERL script (see Electronic Appendix). However, only families that contained at least one nematode sequence were concatenated. This was done to reduce the dimensions of the superalignment (thus making it more manageable), whilst retaining all the information that could possibly bear on the phylogenetic position of the Nematoda. As pointed out in Section 1.3.1, the supermatrix approach becomes impracticable for complete genome scale studies, as such, it was necessary to adopt this reduction strategy. The three concatenated data sets, generated in this way, were thus subsamples of the

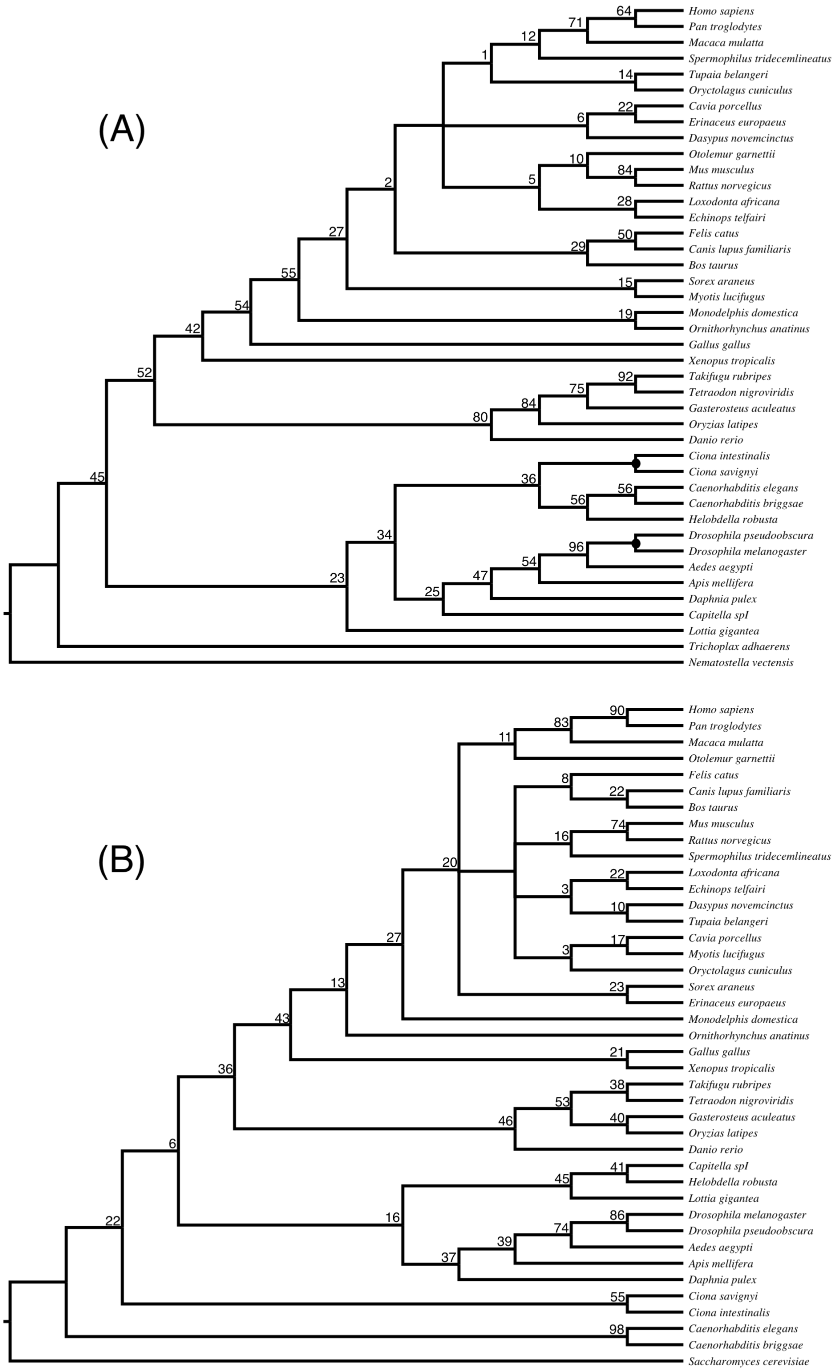


Figure 2.4 Quartet Joining supertrees.

(A) A tree based on 2,216 genes from 42 species, where only non-bilaterian animal outgroups were used. (B) A tree derived using only the fungal outgroup. This tree is based on 2,164 from 41 species. It is evident from both trees that there is low resolution, even within the mammals.

complete data sets and scored: 43392 amino acid positions (41-taxon data set), 38701 amino acid positions (42-taxon data set), and 25857 amino acid positions (43-taxon data set). As the considered genes were not universally distributed, there was a significant amount of missing data in each alignment.

Phylogenetic analyses of the three data sets were performed in PhyloBayes, (Lartillot and Philippe, 2004) under the CAT + G model. CAT was selected as it has been shown (e.g. Philippe et al., 2007, Sperling et al., 2009) that this model provides a better fit to data in comparison to ordinary general time reversible models (e.g. WAG or mechanistic GTR). The use of CAT-GTR was also tested, but under this model convergence could not be reached (the data set was too large for effective computation under this very parameter rich model). For each data set, two independent runs were performed. Convergence was tested using the bpcomp program (which is part of the PhyloBayes software). Two runs were considered to have converged when the max difference in observed bipartitions dropped below 0.2 (see PhyloBayes manual).

2.3.1.5 Bayes factors: testing Coelomata and Ecdysozoa in a Bayesian framework

Bayes factors (BF) are general statistical tools that can be used, within a Bayesian framework, to compare alternative models, e.g. the trees representing the relationships for a group of taxa, and evaluate the weight of evidence in favour of one of the compared models (and hence against the alternative one; Sperling et al., 2009). To calculate BFs for each considered data set, two constrained Bayesian analyses were run using MrBayes (Ronquist and Huelsenbeck, 2003). Each of these analyses could only visit trees compatible with one of the two compared hypotheses (i.e. Ecdysozoa or Coelomata). For each of the two constrained analyses, two runs of one chain were run for 1,000,000 generations (sampling every 100 generations). A burn in of 500,000 generations was considered for all analyses. Due to the dimensions of the data set, it was not feasible to implement the approximately unbiased test (see Shimodaira, 2002). All analyses were performed under WAG + G. This is not ideal, but BF analyses could not be run under CAT, as the current PhyloBayes output is not suitable for estimating BFs (see also Sperling et al., 2009), whilst running the analyses under GTR in MrBayes was not feasible because of time limitations.

BFs were calculated in Tracer 1.4.1 (Rambaut and Drummond, 2007) using, for each constrained analysis, the trace files from the run of highest harmonic mean. Standard errors, around the estimated BF, were calculated using bootstrap (1000 replicates). BFs were interpreted according to the table of Kass and Raftery (1995; see Appendix A1).

2.3.2 Results

While the small data sets demonstrate at the most fundamental level the effects of outgroup selection, they still consider a very scant taxonomic sampling. These analyses allow rejection of the null hypothesis (i.e. Coelomata is the true tree), but only relative to small data sets. To test the validity of these results in a more practicable context, attention was turned to data sets with a broader taxonomic sampling.

Three experiments were performed. In the first, a data set in which taxon sampling was incremented from four to forty one species was used. *S. cerevisiae* was selected as the outgroup, whilst all supplementary taxa included were bilaterian. That is, no attempt at breaking the putative long branch between the fungi and the Bilateria was made. In the second experiment, a data set sampling 43 taxa was used. This data set was designed to contain the full complement of taxa from the first data set, but additionally included *T. adherans* and *N. vectensis*. Here, *S. cerevisiae*, *T. adherans* and *N. vectensis* were simultaneously used as outgroups for the Bilateria. The branch joining the fungi and the Bilateria was still present, but now it was split into three shorter branches, allowing the effect of targeted taxon sampling to be investigated. Finally, the third data set sampled 42 genomes. All metazoan genomes used to generate the first two data sets were retained, whilst *S. cerevisiae* was removed. Excluding *S. cerevisiae* eliminated the long-branch joining the fungi and the Bilateria, thus enabling investigation of the effect of using only non-bilaterian metazoans (*T. adherans and N. vectensis*) as outgroups.

The analysis of the data set generated for experiment one resulted in 2,164 single protein families passing the PTP test. Results of an input tree bootstrapping supertree analysis, of the ML bootstrap trees generated for these families, is reported in Figure 2.5(A), and shows the placement of the Nematoda as the sister group of all the other Bilateria, i.e. 100% support for Coelomata. This tree also displays monophyletic Deuterostomia, Arthropoda and, interestingly, Eutrochozoa. (BS= 98%, 100%, and 100% respectively). The BF analysis shows that the data fit the Coelomata tree better than the Ecdysozoa tree, decisively discriminating against Ecdysozoa: Log$_{10}$-BF=10.792 (± 0.29).

When *S. cerevisiae*, *T. adhaerens* and the Cnidarian *N. vectensis* were concurrently used as outgroups, a total of 1,949 single protein families conveying significant phylogenetic signal (see Table 2.4) were found. When these protein families were used for supertree reconstruction, Ecdysozoa was recovered, but with very low support (BS= 43%; See Figure 2.5 B). Bilateria finds significant support in this analysis (BS= 99%), and is partitioned into Protostomia and Deuterostomia.

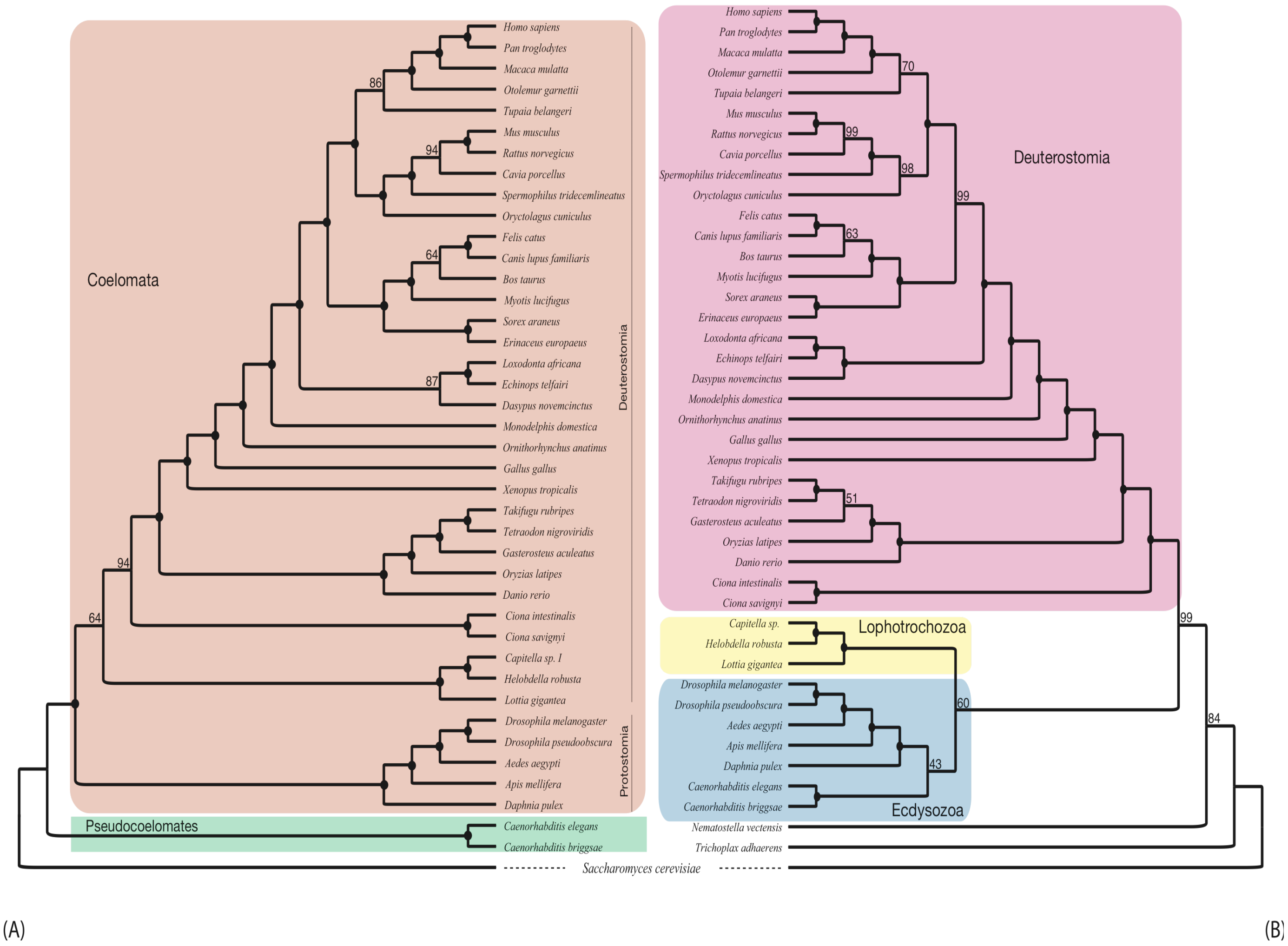


Figure 2.5 Phylogenomic supertrees of the Bilateria.

(A) A tree derived using only the fungal outgroup. This tree is based on 2,164 from 41 species. (B) A tree derived using fungal and animal (non-bilaterian) outgroups. This tree is based on 1,949 genes from 43 species. The monophyly of Ecdysozoa, Lophotrochozoa and Protostomia is recovered in (B), while (A) supports Coelomata. Numbers at the nodes represent bootstrap support. Full circles indicate 100% bootstrap support for a node.

Monophyly of the Eumetazoa is also supported (BS= 84%; in agreement with Sperling et al., 2009, but see Pick et al., 2010), whilst support for Protostomia is not very high (BS=60%). Inspection of the partition table for this bootstrap analysis shows that Coelomata is still recovered, albeit with minimal support (BS=13%). This is suggestive of an enduring LBA effect.

LBA is obviously reduced when the additional animal outgroups are included in the analyses, to the point where the Ecdysozoa tree is the most commonly recovered in the individual bootstrap replicates. However, the reduction of the LBA effect is not significant enough to completely exclude Coelomata from the set of possible solutions. Interestingly, BFs still favour Coelomata with respect to Ecdysozoa (at the least under WAG + G): Log_{10}-BF = 6.67 (± 0.59). However, in agreement with the results of the bootstrap analysis, which suggest that the LBA effect was indeed reduced when non-bilaterian animals were in the sample, the weight of the evidence in favour of Coelomata is now greatly decreased (by 4.122 points in a $\log_{10}$ scale). That is, when the fungi-Bilateria branch is broken, Coelomata is still favoured but the data fits the tree ~ 13,243 times less well than they did when the branch was not interrupted.

In the third experiment, *S. cerevisiae* was interchanged with two animal outgroups (*T. adhaerens* and *N. vectensis*). With this specific taxonomic sampling, 2,216 single protein families conveying significant phylogenetic signal are recovered. Their analysis resulted in a phylogenomic supertree supporting all major, recognised groups (Protostomia, Deuterostomia, Eutrochozoa, and Arthropoda). Additionally, this analysis found significant support for Ecdysozoa (BS= 90%) within Protostomia (See Figure 2.6), with the BF now decisively discriminating against Coelomata: Log_{10}-BF=90.811 (± 0.977). If one compares the fit of the Ecdysozoa tree to the data set where *S. cerevisiae* is

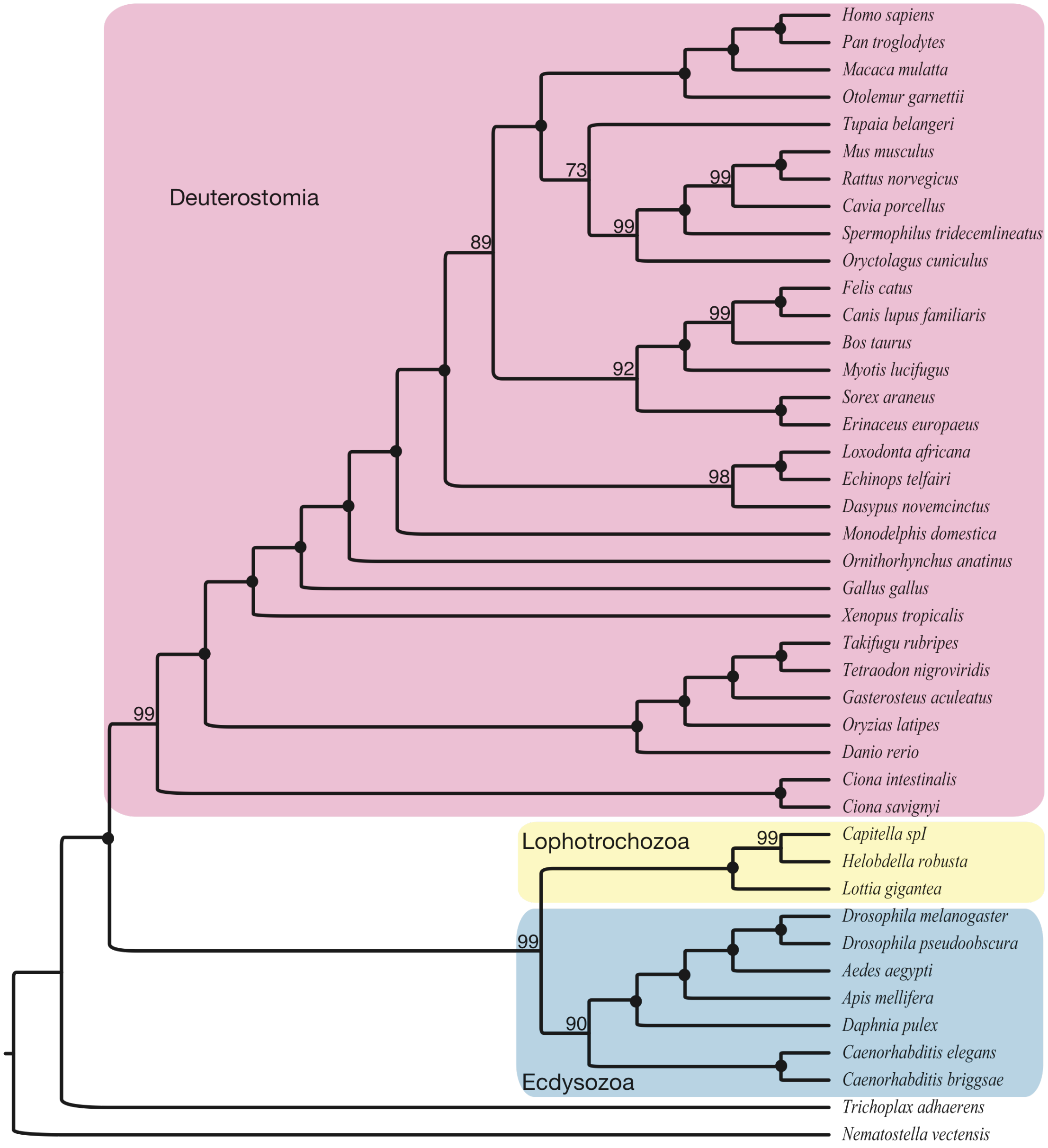


Figure 2.6 Phylogenomic supertree of the Bilateria recovered using only animal (non-bilaterian) outgroups.

This tree is based on 2,216 genes from 42 species. High support for the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia can be observed. Numbers at the nodes represent bootstrap support. Full circles indicate 100% bootstrap support for a node.

the only outgroup, with the fit of the same tree to the data set where only the animal outgroups were used, a dramatic change (~10^{100}) in the BF in favour of Ecdysozoa is observed. This clearly highlights the major role played by outgroup selection in phylogenomics.

These results are finally confirmed by the supermatrix analyses. In these analyses, when *S. cerevisiae* was used as the only outgroup, convergence could not be reached and the resulting phylogeny (not shown) was nonsensical. When all outgroups were included (Figure 2.7 A), Ecdysozoa was recovered, but the effect of LBA was still evident. If one were to root the tree using *N.vectensis*, to better pinpoint the LBA effect, a tree essentially consistent with the new animal phylogeny is recovered. However, in this rooted tree, *S.cerevisiae* is incorrectly clustered within Protostomia. If the tree is correctly rooted using *S.cerevisiae* (not shown), the Lophotrochozoa are incorrectly attracted toward the root. This result, which was somewhat unexpected (as lophotrocozoans generally do not seem to show serious attraction problems in other supermatrix analyses), is probably a partial consequence of the gene sub sampling strategy, in which I maximised information bearing on the relationships of the Nematoda, whilst ignoring the Lophotrochozoa and the Deuterostomia (see 2.3.1.4). However, it is also clearly telling of an enduring LBA effect.

Finally, when only the animal outgroups are used (Figure 2.7 B) the Ecdysozoa tree is recovered. In Figure 2.7 (B) support for the Urochordata as members of the Deuterostomia is not significant, and this group is thus collapsed into a polytomy, which again can most likely be attributed to the gene sub sampling strategy (see above). This is confirmed by the supertree analysis of our full data sets in which support for monophyletic Deuterostomia varies between 94% to 100% depending on the outgroup

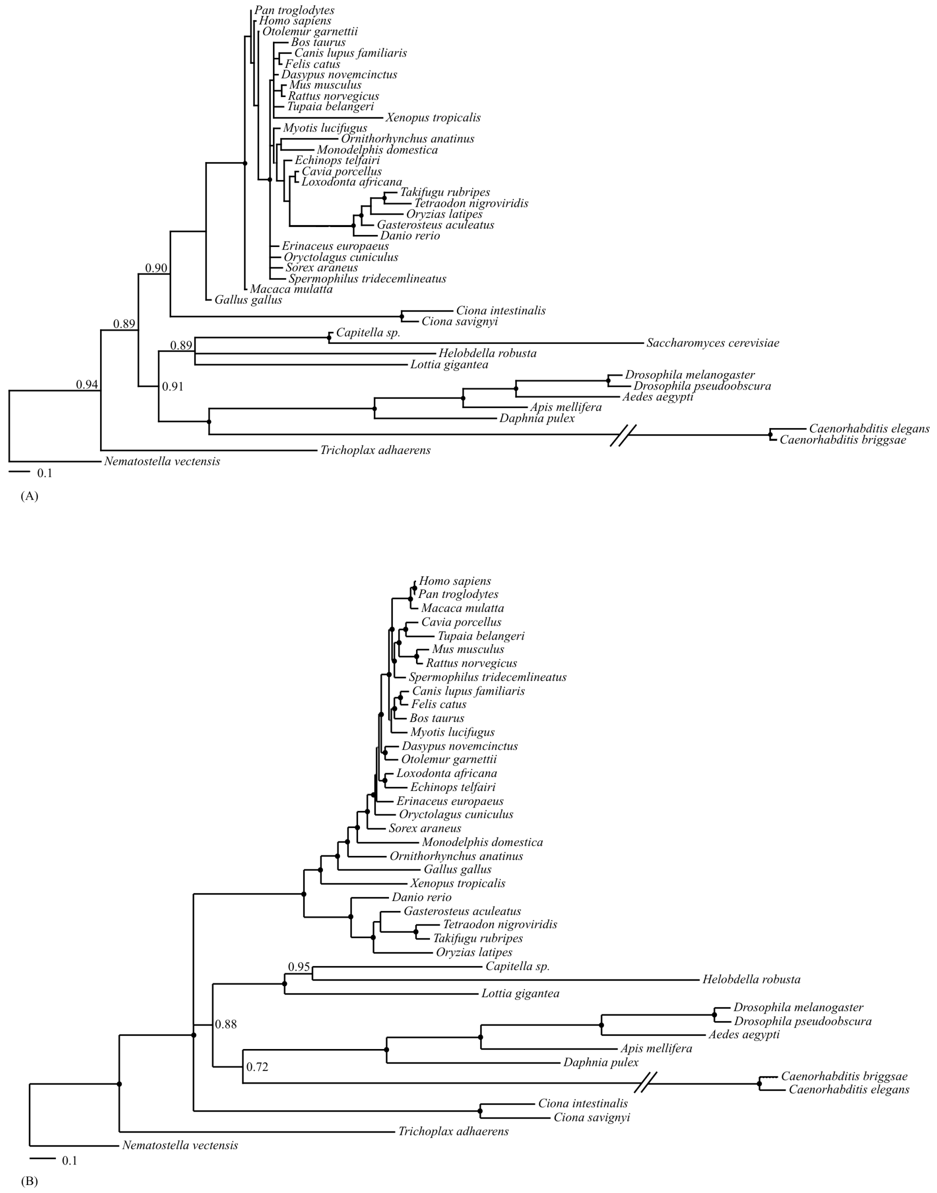


Figure 2.7 Results of the supermatrix analyses.

(A) The effect of long branch attraction is obvious if one roots the tree using *N.vectensis*, as a tree essentially consistent with the new animal phylogeny is recovered, but *S.cerevisiae* is incorrectly nested within the Protostomia. (B) A tree illustrating that Ecdysozoa is easily recovered when analyses are performed using only non-bilaterian animals as outgroups. Numbers at the nodes represent posterior probabilities. Full circles indicate a posterior probability of 1. Posterior probabilities lower than 1 have only been reported for nodes that are relevant to the Ecdysozoa Vs. Coelomata problem. Urochordata is collapsed in a basal polytomy because the posterior probability of Deuterisomia is less than 0.5.

used (see Figures 2.5 and 2.6). Notably, a similar effect was observed in the EST study of Hejnol et al. (2009), in which Urochordata became unstable when gene sampling was reduced (see Figure S1 Hejnol et al., 2009).

2.4 Discussion

2.4.1 Phylogenomics in a pluralist context

ESTs provide an excellent means of increasing taxon sampling, and have been shown to produce highly resolved, well-supported phylogenies (e.g. Philippe et al., 2005b, Dunn et al., 2008). However, as pointed out by Sperling et al. (2009), the incongruence of different EST studies implies that EST data does not assure accuracy. For example, the studies of Dunn et al. (2008) and Philippe et al. (2009) conflict on the relationships amongst the non-bilaterian Metazoa, offering two alternative, and well supported, positions for the ctenophores (but see Pick et al., 2010). Similarly, Dunn et al. (2008) and Rota-Stabelli et al. (2011) differ on the placement of the Myriapoda.

Additionally, EST studies consider only a shallow sampling of genomic content, and include a large amount of missing data, the effect of which, until recently, had not been thoroughly investigated. However, Sanderson et al. (2010; see Section 1.3.1) provide some tentative, but important, results. For Coelomata to be robustly rejected, EST data, although obviously important, cannot be considered sufficient: accord between taxonomically rich EST studies, and gene rich deep-scale analyses must be reached. With the wealth of genomic data that is currently available, coupled with advances in sequencing technologies, taxon sampling is becoming less of a limitation for deep genomic-scale phylogenetic analyses. In short, we now have at our disposal the data to conduct extensive, experimental phylogenomic studies of metazoan evolution.

Supertree methods offer an ideal solution for the reconstruction of large-scale phylogenies based upon complete genomes, as they provide a means of overcoming the limits of gene concatenation based approaches. Gene concatenation methods, at present, do not allow for the easy amalgamation of thousands of genes. Supertrees (and in the four taxon case, consensus methods), implementing a divide and conquer strategy, facilitate the analysis of entire genomes, for many taxa, by coalescing the results of multiple sub-analyses to attain a global solution (Wilkinson and Cotton, 2006).

However, supermatrix approaches also have important advantages, particularly as they overcome the most important limitation of supertrees; that is, supertrees do not allow hidden sub-signals to interact and thus lack total-evidence like properties (Pisani and Wilkinson, 2002). In addition, supermatrix approaches allow for the use of statistical tools (like BFs) to test alternative phylogenetic hypotheses. However, the implementation of maximum likelihood supertrees should allow the development of statistical testing within a supertree framework as well. Bearing in mind that both approaches have highly desirable, and significantly different properties, I therefore opted for a pluralist, supertree/consensus tree and supermatrix approach in this study.

The four-taxon analyses show that multi protein families can be appropriately treated to derive species phylogenies, and suitably included in a consensus tree (if all considered protein families are universally distributed) or supertree (if the protein families are not universally distributed) analyses. In particular, I show that all consensus supertrees (including those that sample multi protein families) continue to support Ecdysozoa, a result that is further confirmed by the supermatrix analyses.

Supertrees have previously been employed to address the phylogenetic position of the nematodes (Philip et al., 2005). While carefully conducted, using the best methods and data available at that time, this analysis did contain (by the authors' own admission) a very limited sampling of just 10 genomes. In particular, a noticeable problem that Philip et al. (2005) faced was the absence of an adequate outgroup (i.e. non-bilaterian metazoan genomes). As postulated by these authors, in time, an increased sampling could well serve to alter their results. In line with that prediction, supertree analyses performed here, using appropriate outgroups and a significantly increased taxon (and gene in the case of the four taxon data sets) sampling, have revealed an alternative topology (see Figure 2.3, 2.6 and 2.7). My results suggest that the study of Philip et al. (2005), and indeed other genomic scale analyses (e.g. Blair et al., 2002, Wolf et al., 2004) may have been influenced by systematic errors arising from poor outgroup availability, sparse taxon sampling, and hidden paralogy.

2.4.2 Circumventing systematic errors

The study described here illustrates the importance of outgroup choice in phylogenomic scale studies. It shows that the use of a distant outgroup has a marked effect, irrespective of whether ingroup sampling is sparse or dense. I found, like in other studies (Philippe et al., 2005b, Rota-Stabelli and Telford, 2008), that outgroup choice completely alters the resulting topology, consequently lending analogous support to competing hypotheses. The recovery of the Coelomata topology can be considered a LBA artifact, brought about by the use of a divergent outgroup. Comparison of BF values gives an indication of the strength of the bias and of how difficult it is to limit its effects. These results also reject the contention of Rosenberg and Kumar (2001) and Rokas and Carrol (2005), that poor taxon sampling is irrelevant as long as enough genes are considered.

The densely sampled data sets illustrate that optimal outgroup selection is more important than targeted taxon sampling in avoiding LBA artifacts. If a distant outgroup (*S. cerevisiae)* is included in the analysis, targeted taxon sampling (i.e. breaking the long Bilateria-Fungi branch) does not completely eradicate LBA (as shown most powerfully by the BF analyses). Only upon the exclusion of *S. cerevisiae* do the BFs show a radical decrease in the fit of the Coelomata tree. Optimal outgroup selection is a rarely addressed topic in phylogenetics and phylogenomics, and one has to bear in mind that the optimal outgroup for a given data set, is not necessarily the closest one (for an interesting example see Rota-Stabelli and Telford, 2008). Aside from LBA, compositional bias is another source of phylogenetic artifact, thus an outgroup (which may not be the closest one available) that simultaneously minimises the likelihood of both artifacts occurring should be selected.

2.4.3 Stringency and the selection of families for phylogenetic reconstruction

When analysing a small selection of genomes, I could not identify a number of single protein families comparable to those identified by, for example, Blair et al. (2002). Disparity between this study and that of Blair et al. (2002) is particularly striking when comparing their 4-taxon data set to the sparsely sampled data set including *S. cerevisiae* used in this study. Although the ultimate results of both data sets are congruent, i.e. both data sets support Coelomata; my analysis considers 70% less single protein families than Blair et al. (2002). Failure of these data sets to have correlating numbers of single protein families merits discussion.

I suggest that the observed difference can partially be explained by the use of different outgroups. Blair et al. (2002), somewhat illogically (see also above), primarily used a plant outgroup and only in cases where plant genes were not available was a fungal outgroup used. However, this difference can also be accounted for by the implementation of measures to assess data quality in this study. Under my protocol, a protein family was only considered for phylogenetic analysis if it demonstrated significant clustering signal. The approach I implemented here, thus ensured that noisy families, or families devoid of clustering signal, were eliminated from the analysis. It is interesting to note that prior to this filtering stage the number of single (4-taxon) protein families identified in my study was twice the number identified by Blair et al. (2002).

2.5 Conclusions

The Ecdysozoa hypothesis has accumulated significant support in recent years (Philippe et al., 2005b, Irimia et al., 2007, Lartillot and Philippe, 2008), particularly from the analyses of EST data sets. To supplement this amassment of evidence, here I present support for Ecdysozoa from genomic-scale data sets. From these, overall, Ecdysozoa represents the most cogent hypothesis. It is supported from the analyses of both single and multi protein families, and once suitable outgroups are considered. Coelomata, on the other hand, is only supported upon the inclusion of a distantly related outgroup, which suggests that this topology is systematically generated by a long branch attraction artifact.

My results, based on arguably the deepest gene sampling of the Bilateria to date, present overwhelming support for Ecdysozoa, and clearly illustrate that it is the use of a distant outgroup that mislead previous analyses. Taken in combination with results from the aforementioned EST studies, it now appears that all aspects of molecular based phylogenetics support the rejection of Coelomata. While lack of unambiguous morphological support for Ecdysozoa persists as a moot point (but see Eernisse et al., 1992), in the light of overwhelming molecular evidence and lack of morphological evidence conclusively discrediting Ecdysozoa, I think that it is now finally time to shed the notion of Coelomata.

Chapter 3: Towards the reconstruction of eukaryotic tree of life using complete and partial genomes

3.1 Introduction

The application of supertree reconstruction to molecular data has prompted some of the most expansive and comprehensive phylogenies of the past decade. While robust, this approach is almost exclusively applied to data derived from complete genomes, or to trees obtained from published sources (e.g. Pisani et al., 2002, Lloyd et al., 2008; however this type of analysis is not the subject of discussion here). Typically, genomic scale analyses use only single gene (or protein) families, to limit the confounding effect of paralogy. A data pool of this kind is extremely limiting, as many groups are severely under represented by genome sequencing projects. Furthermore, much pertinent information is omitted simply because genes have undergone duplication.

The study discussed in this chapter attempts to overcome limitations in sampling, both in terms of genomic breadth and depth, to recover a phylogenomic supertree of the eukaryotes. In an effort to increase taxon sampling across this domain, ESTs (referred to here as partial genomes) are used to supplement the relatively limited amount of complete eukaryote genomes. Additionally, to increase gene sampling, gene families with a history of duplication are included, in what can be considered an expansion of the approach used for the sparsely sampled data sets of Chapter 2. As such, a data set spanning over 20,000 genes, for approximately 550 species is analysed.

3.1.1 The eukaryotes and their origin

Members of the eukaryote domain are set apart from prokaryotes by distinct features that are indicative of a more complex form and structure. Eukaryotes are characterised by membrane-delimited compartmentalisation, that is supported by a cytoskeleton (Parfrey et al., 2006).[3] Cellular subunits, or organelles, that exclusively feature in all eukaryotes are the nucleus, which is the repository of genetic material, and the mitochondrion (or its associates; see Embley and Martin, 2006, Hjort et al., 2010), which, amongst other roles, is responsible for the production of energy. The ubiquity of the mitochondrion across the domain has, for the most part, conserved metabolism in eukaryotes, unlike bacteria who have a broad metabolic variation (Baldauf, 2008). Despite these unifying features, the eukaryotes are hugely diverse, with their span extending from unicellular organisms right up to complex plant, fungal and animal forms (Parfrey et al., 2006).

Although there is a general agreement that living eukaryotes are symbiotic organisms (as first proposed by Margulis, 1970), and that the mitochondrion and the chloroplast were once free-living eubacteria (Pisani et al., 2007), the specific means as to how the eukaryotes emerged still remains largely unknown. There are two general schools of thought as to their origin, which differ about the timing of the mitochondrion acquisition from a bacterial endosymbiont. The most traditional view is based on a literal interpretation of the Carl Woese tree of life (Fox et al., 1977), which postulated that the eukaryotes are the sister group of the Archaebacteria (Poole and Penny, 2007, De Duve, 2007, Cavalier-Smith, 2010, Gribaldo et al., 2010). According to this perspective, the

[3] It must be noted that prokaryotic counterparts have been identified for numerous such features that were once thought to be distinctly eukaryotic.

eukaryotic lineage is very ancient and evolved its specific features (e.g. phagocytosis) before the acquirement of the mitochondria. Proponents of this theory maintain that the eukaryotes (together with the Eubacteria and the Archaebacteria) are a distinct, “primary” domain of life. This three-domain based hypothesis is often referred to as the 3D hypothesis (see Gribaldo et al., 2010).

The second hypothesis proposes that the eukaryotes originated through a process of symbiosis, that lead to a genomic fusion (Rivera and Lake, 2004) between an archaebacterium (the host cell) and an alpha-proteobacterium (which subsequently became the mitochondrion). According to this hypothesis, the eukaryotes are not a “primary” domain of life, and are paraphyletic with respect to both Eubacteria and Archaebacteria (Embley and Martin, 2006, Pisani et al., 2007, Cox et al., 2008, Foster et al., 2009, Cotton and McInerney, 2010). Consequently, the relationships among Woese’s three domains of life are seen to be more ring like than tree like.

Irrespective of which hypothesis is correct, eukaryotes are undoubtedly chimeric, and their genomes feature a mosaic of archaebacterial and eubacterial genes (e.g. Pisani et al., 2007, Esser et al., 2004, Cotton and McInerney, 2010, Lane and Martin, 2010). Additionally, consistent with both hypotheses, Eukaryota are indubitably monophyletic. As the molecular era progresses, the weight of evidence suggests that the ring like origin of the eukaryotes might be the most plausible (see for example Pisani et al., 2007, Cox et al., 2008, Foster et al., 2009, Cotton and McInerney, 2010, but see Logsdon, 2010), however, elucidating the “leap in complexity at the origin of eukaryotes is one of the principal challenges of evolutionary biology” (Koonin, 2010). Some possible steps towards achieving this goal have recently been explicated by Lane and Martin (2010), who suggest that possessing mitochondria allowed for a dramatic increase in the eukaryotic energy availability.

3.1.2 The eukaryote phylogeny

An equally challenging problem facing evolutionary biologists is resolving the relationships within the eukaryotes. Currently, the consensus view of the eukaryotes, arrived upon by a combination of molecular and morphological evidence, is that they are distributed across five or six (if the Amoebozoa and Opisthokonts are not unified as the Unikonts) supergroups (Keeling et al., 2005, Keeling, 2007, Lane and Archibald, 2008, Rogozin et al., 2009, Roger and Simpson, 2009, but see Baldauf, 2008). These supergroups are, namely, Plantae, Excavata, Rhizaria, Chromalveolata and the Unikonts, which encompasses the Amoebozoa and the Opisthokonts (See Figure 3.1).

Although this represents the common view, these assemblages are still subject to dispute, along with their relationships and branching order (Lane and Archibald, 2008). In considering the six-supergroup scheme, the monophyly of two such supergroups has traditionally attained poor support. The first of these, the Excavata, consists of eight groups of protists (i.e. Kinetoplastids, Euglenids, Heterolobosea, Jakobids, Oxymonads, Parabasalia, including Trichomonas, Retortamonads and Diplomonads, including Giardia; Keeling, 2007) that are weakly grouped together using combined aspects of molecular and morphological evidence (Keeling et al., 2005; Simpson, 2003). Recently, convergence of two independent phylogenomic analyses, recovered monophyletic Excavata with high support (Burki et al., 2008, Hampl et al., 2009; based on 135 and 143 genes respectively), bolstering confidence in this supergroup. Yet, whether Exacavata

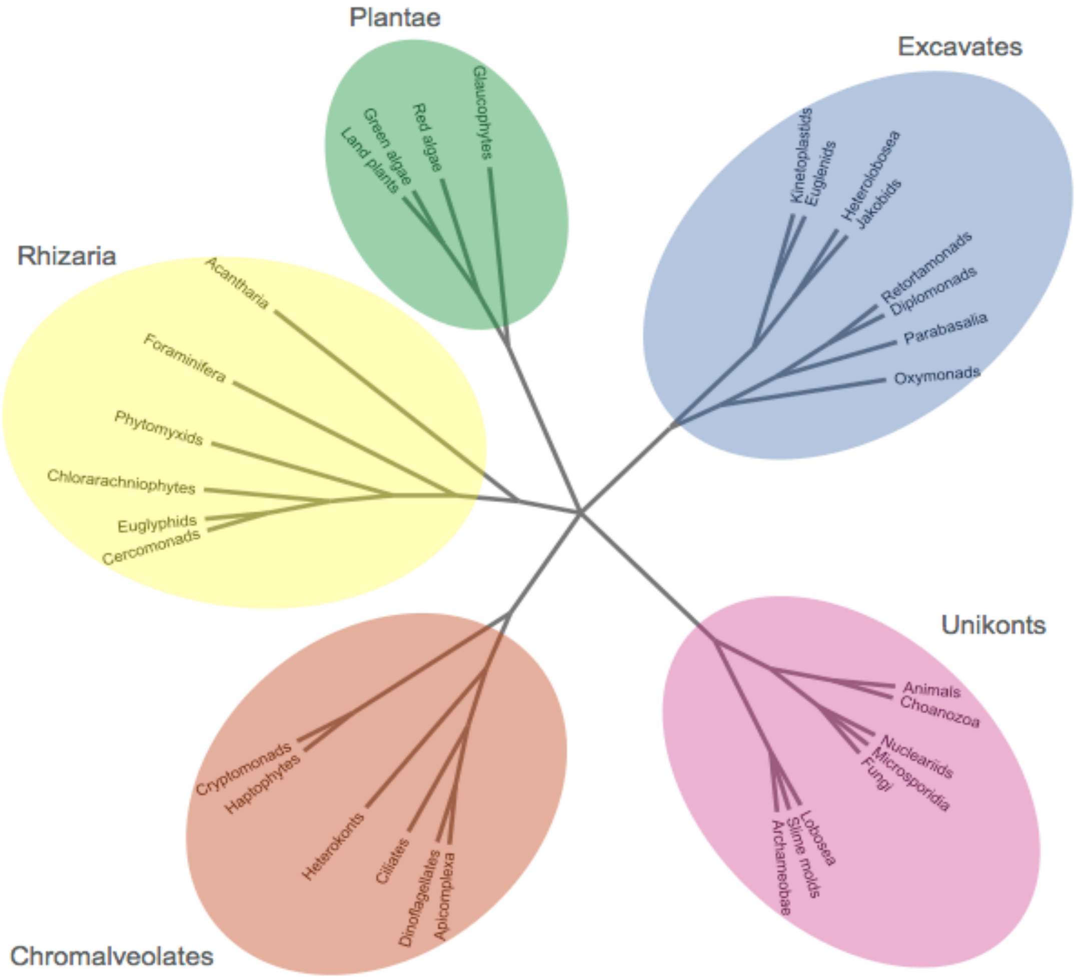


Figure 3.1 The consensus view of the eukaryote phylogeny.

(redrawn from Keeling, 2007).

The five eukaryote supergroups that attain support from a mixture of morphological and molecular data. There is no consensus on the branching order of these groups or the rooting position for the tree. In some schemes, the members of the Unikonts are considered independent groups, namely the Opisthokonts and the Amoebozoa.

represents a real group, still requires further confirmation.

The second supergroup with uncertain monophyly is Chromalveolata, a unicellular group that contains several algae and protists (i.e. Apicomplexa, Dinoflagellates, Ciliates, Heterokonts, Haptophytes and Cryptomonads; Keeling, 2007), connected by plastid based features (Lane and Archibald, 2008). To date, no phylogenetic reconstruction or single character has been found to unify all the members of this assemblage (Lane and Archibald, 2008), and, further to this, Chromalveolata fails to be recovered in several phylogenomic studies (Rodriguez-Ezpelata et al., 2007; Burki et al., 2007; Hackett et al., 2007; Hampl et al., 2009). Indeed, current phylogenomic data (EST and multi gene data sets) are instead converging on a group uniting Alveolata, Stramenopiles and Rhizaria (Rodríguez-Ezpeleta et al., 2007, Burki et al., 2007, Hackett et al., 2007), known as the SAR or RAS group. It is interesting to note that although the monophyly of Rhizaria is less contentious than the groups just discussed, evidence for this grouping is based solely upon molecular data (Keeling et al., 2005).

The particular difficultly in resolving relationships amongst these supergroups lies in the fact that many eukaryote groups may have rapidly radiated in a “Big Bang” manner, resulting in very short internal branches (Philippe, 2000, Koonin, 2007, Rogozin et al., 2009). Obviously, the chimeric nature of their genomes, due to the impression of both endosymbiotic and later gene transfer (LGT) events, especially between unicellular eukaryotes (see Keeling and Palmer, 2008 for a review), further confounds the elucidation of such internal relationships (see for example Pisani et al., 2007, Cotton and McInerney, 2010).

Additionally, one can postulate that given the chimeric nature of eukaryote genomes, substantial phylogenetic artifacts are to be expected. As seen in Chapter 2, outgroup choice is key in phylogenetics. In the case of the eukaryotes, different genes will have different, optimal (and possibly very distantly related), outgroups to reflect their varied origin. For example, the optimal outgroup for genes of archaebacterial origin would be an archaebacterium, while for genes of mitochondrial origin, the ideal outgroup would be an alpha-proteobacterium. If, in a supermatrix approach (e.g. Burki et al., 2007, Hampl et al., 2009), genes of archaebacterial, chloroplastic and alpha-proteobacterial origin are merged, and subsequently analysed using an archaebacterial outgroup (as is to be expected if adhering to the 3D hypothesis), a situation arises where a significant proportion of the data is analysed using a highly suboptimal outgroup, which may introduce LBA. Nevertheless, two recent phylogenomic analyses of over 130 proteins claim to have achieved success in establishing some order within the eukaryotes, both arriving upon a phylogeny that features three lineages: Excavata, Unikonts and a "megagroup" consisting of Rhizaria, Chromalveolata and Plantae (Burki et al., 2008, Hampl et al., 2009). The validity of these groups, however, remains a matter of opinion, particularly since use of a prokaryotic outgroup is not a feature of either study. It is thus clear that these groups may in truth be paraphyletic.

A final problem posed by the eukaryotes is the rooting position for the phylogeny. Postulated placements include a rooting point that would create a unikont-bikont split (Stechmann and Cavalier-Smith, 2003), or a position within the excavates, either at the branch leading to the diplomonads and parabasalids (Arisue et al., 2005), or basal to the jakobids (Rodríguez-Ezpeleta et al., 2007). Although the aforementioned phylogenomic studies (i.e. Burki et al., 2008, Hampl et al., 2009) claim to have come some way in resolving internal eukaryote relationships, both notably report unrooted trees. Clearly, these studies can only suggest potential sets of relationships, while, ultimately, only positioning the root of the eukaryotic tree will identify which groups are genuine and which are not. As such, it appears that in attempting to determine the rooting position of the eukaryotes, there is a general sense that the data currently available are insufficient to deal with such a demanding task (Baldauf, 2008). Indeed, the abovementioned outgroup selection problems are additionally likely to impact on the recovery of the root of the tree (Jeffroy et al., 2006, Sperling et al., 2009), and it is somewhat surprising that Baldauf (2008), for example, seems to overlook this problem. Given the chimeric nature of eukaryotic genomes it not surprising that rooting the eukaryotic tree, for the time being, has not been possible.

3.1.3 Increasing the breadth and depth of sampling

The previous section has painted a rather uncertain perspective for the eukaryote tree; however, this uncertainty must be put into context. Many aspects of the eukaryote supergroups have only recently been described and, as such, reflect the rapid pace at which our understanding of this domain is changing (Keeling et al., 2005). Indeed, for a very stark example of this one only has to compare the trees in the figures of Baldauf (2003) and Baldauf (2008), where eight major groups are detailed in the former and only six in the latter. Further to this, molecular data, and in particular genomic data, has until recently been quite concentrated on a limited number of eukaryotic groups, generally those including model organisms (or important parasites). With reference to the eukaryotic tree of life, Sanderson (2008) asserts that “a stronger sampling effort aimed at genomic depth, in addition to taxonomic breadth, will be required to build high-resolution phylogenetic trees at this scale”.

In conducting this analysis, I aimed to address both of these issues, in an attempt to build a phylogeny for the eukaryotes based upon the broadest and deepest genomic sampling to date. At the commencement of this study there were approximately 120 complete eukaryote genomes publicly available, however, as most of the eukaryote supergroups were severely underrepresented by the sampling included, this was insufficient. Therefore, in order to augment the taxonomic sampling of this study, an experimental approach was adopted, where an EST database (kindly provided by Dr John Parkinson, Hospital for Sick Children / University of Toronto) was additionally used. In this way, it was possible to increase taxon sampling by almost fourfold, significantly improving the coverage of the majority of the eukaryote supergroups. It is important to note, however, that the taxonomic sampling of EST projects reflects what is observed in genome sequencing projects, where there is a bias towards major groups, therefore, an overrepresentation of some groups (i.e. plants, fungi, nematodes and more broadly animals) in the taxon sampling is observed.

Typically, for data sets with a taxonomic sampling of this magnitude, phylogenomic reconstruction is limited to single gene (or protein) families (e.g. Fitzpatrick et al., 2006, Pisani et al., 2007, Holton and Pisani, 2010). However, here, in an effort to consider the maximal gene sampling possible, multi gene families were additionally considered. Increasing gene sampling, as pointed out by Sanderson (2008), is important for recovering the eukaryotic tree of life. In line with this, Dagan and Martin (2006) put forward an equally compelling appeal for increased genomic depth when they introduced the term “tree of one percent”.

This thought-provoking concept was broached in the context of a criticism of Ciccarelli et al. (2006), who presented a tree of life based on only 31 proteins. These 31 proteins, representing ~ 1% of the average prokaryotic proteome, were selected by Ciccarelli et al. (2006) because they corresponded to the few genes in the considered genomes that did not appear to have undergone lateral gene transfer. Dagan and Martin (2006) aptly defined the phylogeny of Ciccarelli et al. (2006) "a tree of one percent", and suggested that if the other 99% of the protein families disagreed with their proposed tree of life, then their tree was unlikely to be an accurate descriptor of the evolutionary history of life. Here, by incorporating multi gene families not only is the depth of sampling improved, but also important phylogenetic signals specific to eukaryotic genes, that have evolved under conditions of duplication and loss, are considered. The approach implemented in this study can be considered a scaling-up of the protocol used in Chapter 2, where multi protein families were integrated into a phylogenomic approach, but only in the case of the 4-taxon data sets.

Typically, a supermatrix approach is applied to EST data, however, given the particular dimensions of this data set, this simply was not feasible (see Section 2.3.1.4 where a reduction strategy had to be employed for datasets sampling only approx. 40 species). Additionally, as discussed in Chapter 2 (see Section 2.1.4), the supermatrix approach currently does not allow for the inclusion of multi gene (protein) families. Therefore, a supertree approach is adopted here. In using a supertree approach, the extent to which ESTs, as well as multi protein families, could be integrated into supertree analyses was investigated.

3.2 Materials and methods

3.2.1 Data assembly

For this analysis, a data set scoring both complete and partial genomes was used. Firstly, in a similar fashion to the data sets discussed in Chapter 2, genomic depositories were searched to attain all available genomes for the eukaryotic domain, amassing to 121 species. To augment taxon sampling, so as to include a broader range of eukaryotic diversity, a database of EST sequences, for a further 448 species, was additionally included. The sequences from both these genomic sources were combined to create a data set featuring 569 eukaryotic species. See Electronic Appendix for a full list of species, their associated data type (i.e. EST or complete genome) and their source.

Further to the above, the genomes of 8 non-eukaryote species (i.e. bacterial and archaeal) were chosen for use as outgroups. Outgroups selected for this analysis can be broadly classified into two groups, namely, the proteobacterial alpha-proteobacteria and cyanobacteria, and the archaeal halobacteria, crenarchaeota and thermoplasmatales. These classes of outgroup were specifically selected to discern varying ancient signals that have punctuated the origin of the eukaryotes (see Pisani et al., 2007). It is expected that the use of a selection of outgroups should better account for the diverse origins of different genes, rather than the use of a single outgroup (e.g. an archaebacterium), which is unlikely to collectively account for all genes and might result in the generation of tree reconstruction artifacts. Note that, as single protein families are analysed in the supertree approach, it is expected that for each protein family, the most adequate outgroup will be used. This is because the closest prokaryotic homolog of each eukaryotic gene is the one with the highest likelihood of being included in its gene family by the homology assignment strategy (i.e. MCL; see below).

3.2.2 Homology assignment to tree reconstruction

The fundamental protocol of this study follows the experiments conducted on the two sparsely sampled data sets in Chapter 2. In that study, these small data sets were used to test alternative homologous protein family identification strategies and multiple sequence alignment algorithms (see Sections 2.2.1.2 and 2.2.1.3). As such, here, in implementing certain steps of the same protocol (as Chapter 2), some technical aspects were modified in line with the findings of Section 2.2.2.2. These are outlined below.

Firstly, in this current experiment homologous protein families were identified using the MCL-based approach (Enright et al., 2002). Although in the comparison of homology assignment strategies little difference was found between the resulting phylogenies of MCL and the approach of Creevey et al. (2004), MCL was selected for use here as it is becoming the current standard for homology assignment (see for example Wu et al., 2009, Brown et al., 2010, Dagan et al., 2010; however some technical aspects do appear worrying; Cummins and McInerney, personal communication). The precursory BLASTP search, required before MCL implementation, was carried out using an E-value cut off 10^{-8}. MCL was then implemented with an inflation parameter value of 5.0, which returns more finely grained clusters (i.e. a broader range of smaller protein families; that is, a proportion of multi protein families will be split into single protein families, each of which will only include the members of one of the paralogous groups in that family).

From the resulting homologous (both single and multi protein) families, those that contained at least four species were retained for further analysis. Once again, using the 4-taxon data sets as a reference point, the integration of multi protein families into phylogenetic analyses was extended to a more practicable scale. As such, unlike the densely sampled data sets in Chapter 2, it was not necessary to partition the protein families into single and multi protein families at this point. All protein families deemed viable were then put forward for multiple sequence alignment.

Alignment was carried out using the PRANK software implementation (Löytynoja and Goldman, 2008). As determined by the analysis of the sparsely sampled data sets in Chapter 2, alignments produced by PRANK attained higher support values in supertree reconstructions (see Section 2.2.2.2 and Table 2.1). Accordingly, for this experiment the more computationally intensive PRANK software was selected over ClustalW (Thompson et al., 1994), in an effort to maximise phylogenetic accuracy. Gblocks (Castresana, 2000) was then used to curate the ensuing alignments, phylogenetic signal was assessed by means of the PTP test (Archie, 1989) and the amino acid substitution model for each protein family was determined using Modelgenerator (Keane et al., 2006), all as per Section 2.2.1.3.

Alignments of each family deemed to convey significant hierarchical signal were then subjected to phylogenetic reconstruction. Differently from the protocol used in Chapter 2, here, maximum likelihood (ML) trees, for each protein family, were derived using the RAxML software (Stamatakis et al., 2005). RAxML represents a technological improvement over the PhyML (Guindon and Gascuel, 2003) software employed in Chapter 2 and, as such, it was selected for implementation in this experiment. Due to the dimensions of this data set, it was not feasible to use bootstrapped ML trees for each protein family in downstream supertrees analyses (deviating from the approach used for all data sets in Chapter 2), therefore, in this experiment supertree analysis was limited to the optimal RAxML tree for each family.

3.2.3 Supertree reconstruction

3.2.3.1 Single protein family supertree reconstruction

The aim of this study is to build an extensive phylogeny for the eukaryotes that incorporates both single and multi protein families, in a practicable context. However, due to the computational and time costs incurred by the analysis of expansive multi protein families, in such abundance, the single protein families were considered on their own in the interim waiting period. This provided for a useful means of comparing the results obtained from the composite multi and single protein family approach to the standard supertree approach generally adopted (see for example Fitzpatrick et al., 2006, Pisani et al., 2007, Holton and Pisani, 2010).

The optimal ML trees derived for each single protein family were coalesced into a single data set, upon which the input tree bootstrapping approach, discussed in Section 2.3.1.3, was implemented. One hundred pseudoreplicates were generated, and for each a MRP matrix was derived using the software package CLANN (Creevey and McInerney, 2005). Parsimony analysis of these matrices was carried out in PAUP (Swofford, 1998) as follows: 10 heuristic searches with random sequence addition and TBR branch swapping. The resultant supertrees were then summarised, resulting in a majority rule consensus genomic supertree (see Section 2.3.1.3). Table 3.1 reports the number of genes used to obtain this supertree.

Table 3.1 Progression of protein family numbers at each stage of analysis.

Unlike the data sets in Chapter 2, to reduce computational time, protein families with more than 4 species were identified first. It must be noted that a small number of families were unable to undergo the sequence alignment or model selection stages due to software specific issues with the families in question.

Eukaryote Data Set							
All families		Single protein families		Multi protein families			
No. homologous families	No. families with more than 4 taxa	No. single protein families	No. families passing PTP	No. multi protein families	Species level trees > 4 taxa	No. families passing PTP	No. families passing GTP-PTP
553263	96711	11475	7398	85236	43650	16521	16353

3.2.3.2 Integration of multi protein families into supertree reconstruction

Gene duplications present in the optimal ML tree of each multi protein family were reconciled according to the GTP method (see Section 1.3.2.3), as implemented in the DupTree software (Wehe et al., 2008) using the same settings outlined in Section 2.2.1.4. The resulting species trees were subject to the GTP-PTP test (see Section 2.2.1.5), and those that passed were deemed viable for supertree reconstruction. For the purpose of subsequent analyses, these species trees were considered analogous to a typical single protein family and were combined with the optimal single protein family trees to create a single comprehensive data set.

Again, due to restrictions imposed by the sheer size of this data set, input tree bootstrapping, as outlined for the supertree analyses in Section 2.3.1.3, could not be performed. Therefore, this analysis was limited to a single MRP matrix derived using the CLANN software. Subsequent parsimony analysis was carried out in PAUP under the aforementioned parameters (10 heuristic searches with random sequence addition and TBR branch swapping), with ensuing supertrees being summarised as before using the majority rule consensus method. See Table 3.1 for the number of genes (both single and multi) used to derive this supertree. Support for resultant supertrees was measured using the stsupport software of James Cotton, which implements the support measures defined by Wilkinson et al. (2005).

3.3 Results and Discussion

3.3.1 Eukaryote phylogeny based upon single protein families

From the analysis of the single protein families, a total of 7,398 viable single protein families were identified for phylogenetic reconstruction (this spanned only 553 species, as some were eliminated through filtering strategies). This number is strikingly small, however, it reflects the complex history of gene duplication that has featured in the evolution of eukaryotes. Use of other strategies, such as that implemented in the Inparanoid software (O'Brien et al., 2005), may have proved better at accounting for such

duplication events, than the approach used here. However, this is beyond the scope of this current study. The resulting phylogeny can be seen in Figure 3.2. From this tree, it is apparent that single protein families alone are incapable of resolving the relationships among the eukaryotes.

There is significant misplacement of the plants throughout the tree. Similarly, the chromalveolates, excavates and Amoebozoa are considerably dispersed; however, there does appear to be a localised concentration of taxa from each of these groups at various points (e.g. the bulk of the chromalveolates appear at the top left of the tree, the excavates towards the bottom right, as does the Amoebozoa). The opisthokonts form two distinct clusters (one extending across the majority of the top of the tree and one towards the centre right), which is indicative of poor resolution, with the extensive interspersal of several other groups contributing further to this problem.

While this tree is clearly far from ideal, it does serve as an important stepping-stone. Despite extensive misplacement of taxa throughout, the tree is not entirely devoid

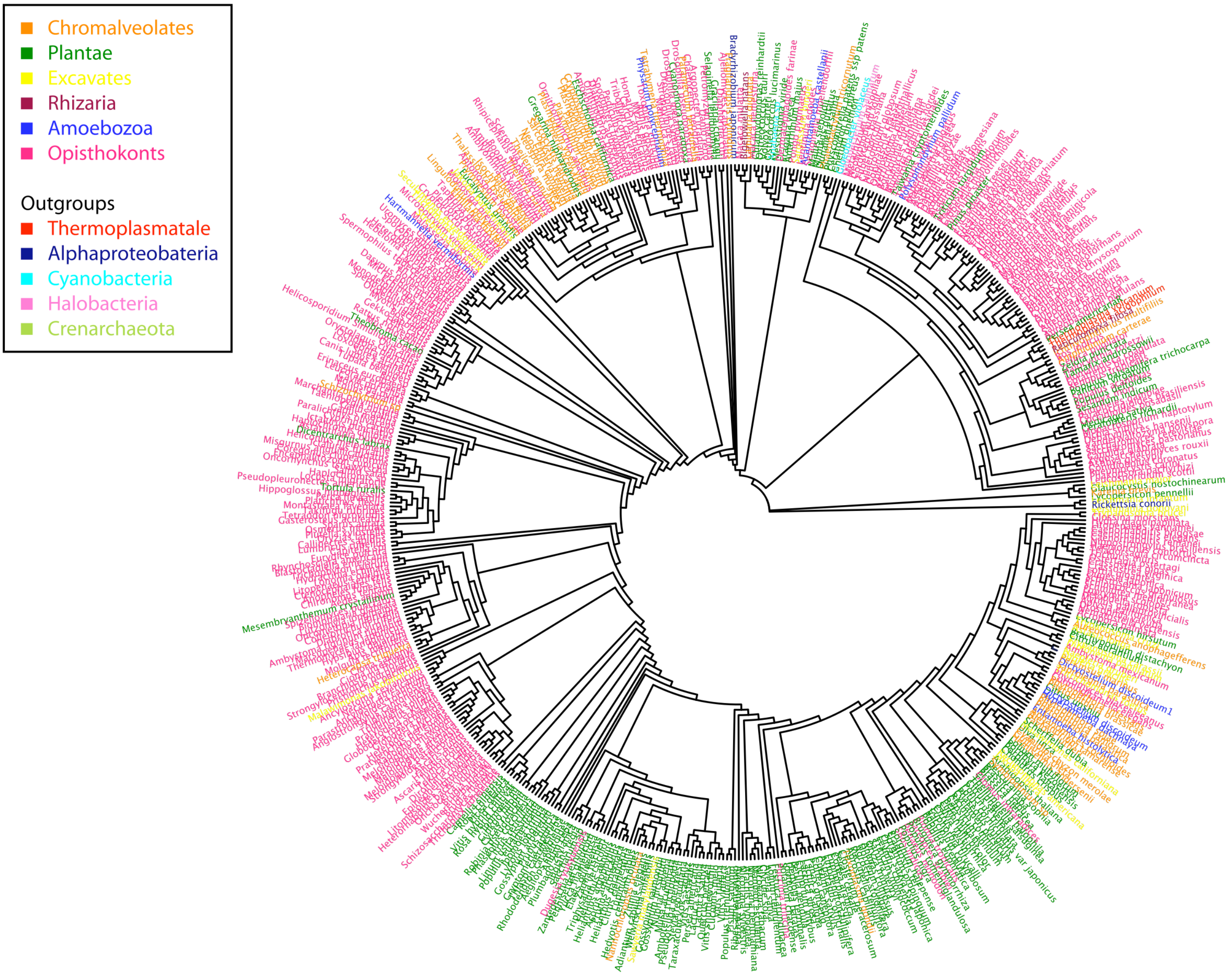


Figure 3.2 Phylogenomic supertree of the eukaryotes using single protein families.

This tree is based on 7,398 genes from 553 species. Although this tree does not recover the monophyly of any of the eukaryote supergroups, it is evident that some phylogenetic signal exists. This is an unrooted tree; with outgroup taxa being recovered at various positions within the tree. See Electronic Appendix for a nexus file of this tree.

of phylogenetic signal. Indeed, there seems to be a level of resolution, albeit minimal, suggesting that the addition of the multi protein families may help improve the resolution.

3.3.2 A comprehensive eukaryote phylogeny

To the 7,398 single protein families, a further 16,353 multi protein families were added, resulting in a gene sampling almost three times the size of that featured in the tree of Figure 3.2. The comprehensive phylogeny of all protein families can be seen in Figure 3.3. In this tree, it can be seen that, overall, the resolution of the phylogeny has significantly improved, with all major groups manifesting more distinctly. Further to this, there is a notable improvement in the internal relationships within each group, as well as a marked reduction in spurious taxon placement.

Importantly, the outgroup taxa appear together, providing directionality and a possible rooting position for the eukaryotes. An interesting aspect of the outgroup analysis is that, if the tree is rooted in its traditional position (i.e. between archaeabacteria and eubacteria) the eukaryotes do not emerge as the sister group of the archaebacteria, as one would expect under the 3D hypothesis. Instead, the eukaryotes appear as the sister group of the alpha-proteobacteria. This is in accordance with what observed by Pisani et al. (2007) and Cotton and McInerney (2010), and reflects the fact that the majority of eukaryotic genes are of alpha-proteobacterial origin, as predicted by the ring of life hypothesis (e.g. Rivera and Lake, 2004, Pisani et al., 2007, McInerney et al., 2008, Cotton and McInerney, 2010).

The tree (see Figure 3.3) still shows a certain number of taxa with an unexpected phylogenetic assignment. For example, the microsporidian *Antonospora locustae* is

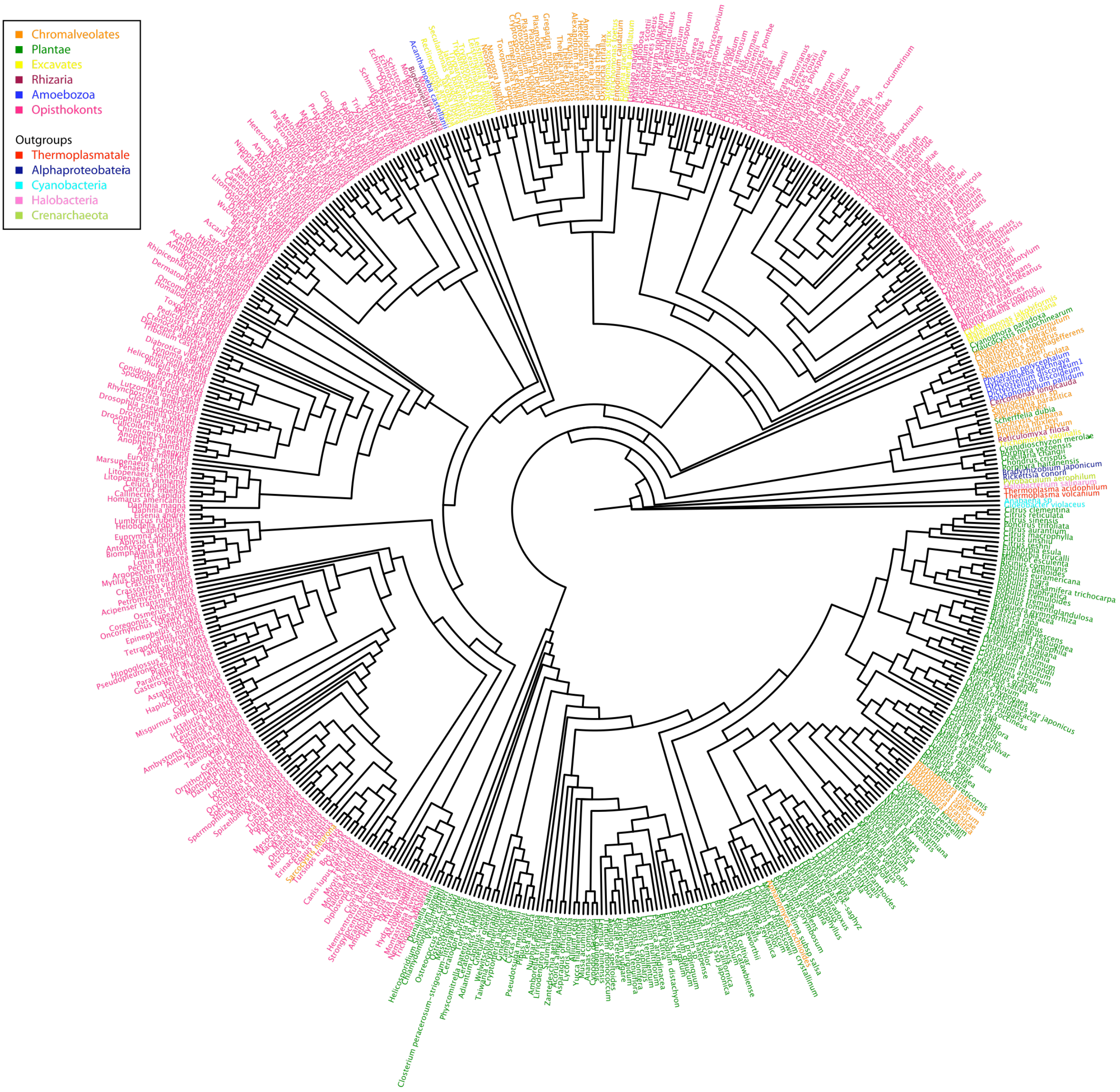


Figure 3.3 Phylogenomic supertree of the eukaryotes using single and multi protein families.

A tree based on 23,758 genes from 550 species. Due to poor resolution, support values are not shown, but are provided in Electronic Appendix.

nested among the lophotrocozoans, rather than among the fungi. Inspection of the visibly misplaced taxa shows that, in most cases (such as that of *Antonospora*), the misplaced taxon is represented in the data set by an EST collection, for which very few genes are available (see Table 3.2). In some cases, even species represented by complete genomes, such as *Giardia lamblia,* are seen to be misplaced. However, similar to what was observed with misplaced EST-based species, very few genes from these genomes were deemed viable for phylogenetic analyses.

Other groups of species, such as the oomycetes (represented here by members of the genera *Phytophthora* and *Aphanomyces*), which are found in Figure 3.3 to nest among the plants, are also noticeably misplaced. However, it is interesting to note that these species are represented by relatively large EST collections, and therefore, their unusual placement cannot be imputed to lack of information. Instead, as oomycetes (water moulds) are parasites of plants, it seems likely that such a placement might have biological significance, possibly reflecting LGT events between the plants and the oomycete species. Notably, it seems unlikely that this placement of the oomycetes could be the result of contamination, as multiple oomycete species are present in the data set, falling in two different parts of the plant tree (see Figure 3.3), depending on the genus to which they belong.

All major groups (for example fungi, animals, plants, red algae, jakobids, Apicomplexa plus Dinophyceae) are seen to be monophyletic, suggesting that there is a significant amount of information in this data set. This result confirms the viability of the experimental procedure of amalgamating two, somewhat disparate, data types (complete and partial genomic data). Although the study of Hejnol et al. (2009) goes some way to employing a similar approach, their study is concerned with the improvement of gene

Table 3.2 Species with low coverage.

Species removed to improve resolution and the number of trees (and gene/protein families) in which they were represented.

Species	NCBI Classification	No. of genes
Entodinium caudatum	Alveolata	14
Oxytricha trifallax	Alveolata	30
Polysphondylium pallidum	Amoebozoa	18
Sarcocystis neurona	Apicomplexa	77
Blomia tropicalis	Arthropoda	22
Sarcoptes scabiei	Arthropoda	21
Guillardia theta	Cryptophyta	43
Scherffelia dubia	Cryptophyta	20
Giardia lamblia	Diplomonadida	6
Diplonema papillatum	Euglenozoa	25
Streblomastix strix	Excavata	32
Trichomonas vaginalis	Excavata	42
Tritrichomonas foetus	Excavata	26
Conidiobolus coronatus	Fungi	9
Spizellomyces punctatus	Fungi	47
Amoebidium parasiticum	Fungi/Metazoa group	31
Anolis sagrei	Gnathostomata	19
Bos sp.	Mammalia	200
Antonospora locustae	Microsporidia	11
Mesocestoides corti	Platyhelminthe	1
Reticulomyxa filosa	Rhizaria	20
Bigelowiella natans	Rhizaria, Cercozoa	86
Cercomonas longicauda	Rhizaria, Cercozoa	18

sampling. Here, it is shown that this protocol is a practical means of concurrently increasing taxon and gene sampling.

However, overall the resolution of the final tree remains unsatisfactory for a number of reasons. For example, the relationships among the unicellular eukaryotes are not resolved according to the current understanding of high-level eukaryotic relationships. Although conformity to previous phylogenies should not be considered a measure of accuracy of the present result, it seems that the level of disagreement with previous studies is too high across the unicellular eukaryotic groups to confidently conclude that the results of this analysis are reliable.

Therefore, to evaluate the extent to which the relationships in Figure 3.3 might be considered to be true, an additional analysis was performed in which all potentially misplaced taxa with limited genomic information were excluded. Unfortunately, this resulted in the loss of the only (three) Rhizarian representatives included in the taxon sampling. Further to the poorly represented taxa, species like the oomycetes, despite having sufficient genomic coverage, were excluded, as the biological implication of their misplacement, albeit of possible importance (to investigate a potential role of HGT in the evolution of their parasitism), is not the focus of this study. Finally, to reduce the dimensions of the data set, the Nematoda and the Platyhelminthes were also excluded. The decision to exclude representatives of these two animal phyla was made because: (1) they are fast evolving and could confound analyses if they are incorrectly resolved in the input trees, (2) the internal relationships within these groups are not of interest here, and (3) other, better behaving, ecdysozoans and lophotrochozoans are present in the data set.

Results of this analysis are reported in Figure 3.4, where a further, significant, improvement in the resolution of the tree is observed. As before, the monophyly of major

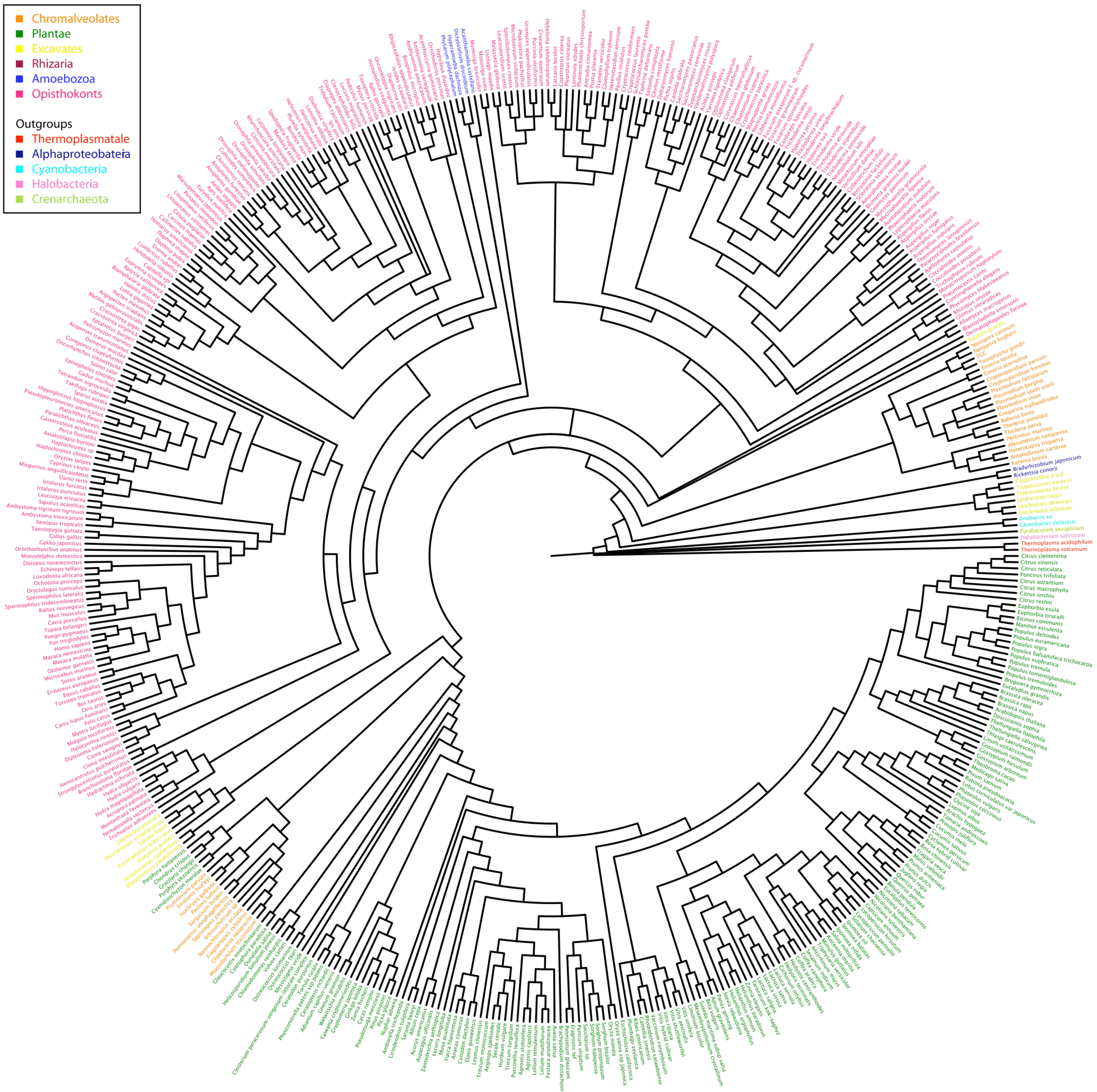


Figure 3.4 Phylogeny of the eukaryotes with problem taxa removed.

A phylogenomic supertree for 474 eukaryote species, based on 20,737 gene families. Some poorly represented species that were visibly misplaced in Figure 3.3 have been removed, in addition to the Nematoda, the Platyhelminthes and the oomycetes. Here, a better overall resolution for the eukaryotes is observed, however, the five supergroup scheme is not upheld. See the Electronic Appendix for the support values for each node.

groups, such as the animals and fungi, continue to be upheld here, therefore, for ease of discussion the phylogeny is redrawn in Figure 3.5 with such groups collapsed. In this reduced data set, some interesting differences in the global phylogeny are observed (in comparison with what is observed in Figure 3.3). Firstly, the Amoebozoa are now seen to bemonophyletic (see Figure 3.5), and are found basal to the animals, however, the even more basal placement of the *Monosiga* species suggests that this region of the tree persists as somewhat problematic. Although the traditional position of the Amoebozoa, basal to fungi + animals (i.e. the opisthokonts), is not recovered, the placement of the Amoeboza observed here (Figure 3.5) does support the union of Amoebozoa and Opisthokonta (i.e. the monophyly of the unikonts).

Plantae and Chromoalveolata are not recovered as monophyletic groups (Figure 3.5), with Rhodophyta emerging as the sister group to Haptophyta plus the stramenopiles. It is interesting to note that the non-monophyly of both these groups is also supported in all the analyses presented in the study of Hampl et al. (2009). Here (Figure 3.5), differently to Hampl et al. (2009), the arrangement of the chromoalveolates is not consistent with the SAR supergroup, as Alveolata (dinoflagellates and apicomplexans) emerges more basal than the stramenopiles. Although SAR is supported in a number of studies (Burki et al., 2007, Burki et al., 2008, Hampl et al., 2009, Burki et al., 2010), it must be noted that all report unrooted trees, therefore, SAR which can be defined more correctly as a clan, rather than a clade, might well turn out to be nothing more than a paraphyletic assemblage. As the validity of the SAR group is still to be tested, it can be suggested that the topology recovered here (Figure 3.5) may well be indicative of the true relationships.

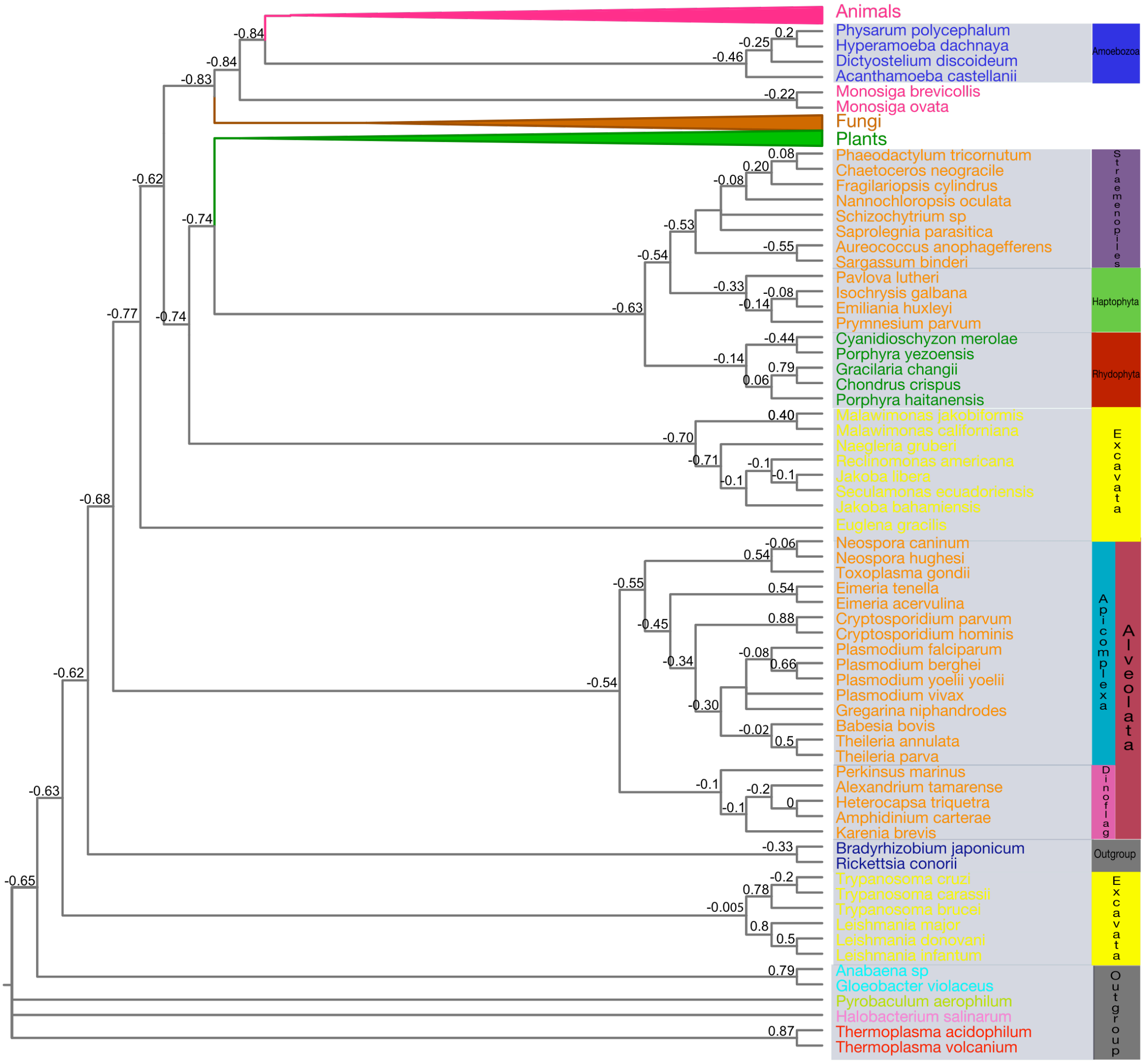


Figure 3.5 General view of the eukaryote phylogeny.

A simplified view of the tree in Figure 3.4. Here, the monophyly of the animals, Amoebozoa, fungi and plants can be seen. Excavata is not monophyletic, with the Kinetoplastida nesting within the outgroups. Chromoalveolata is also non-monophyletic, and is split into the Alveloata, towards the base of the tree, and the stramenopiles plus the Haptophyta, grouping with the Rhodophyta (red algae). Support at the nodes is determined by the V measure of Wilkinson et al. (2005), where V can range from 1 to -1. V=1 indicates that all input trees support the supertree clade, while V=-1 indicates that all input trees conflict with the supertree clade.

Irrespective of the monophyly of SAR, it is clear from this tree (Figure 3.5) that the problematic group here is not Rhodophyta, but the Haptophyta plus the stramenopiles, which nest inside Plantae. Indeed, I suggest that the observed monophyly of Haptophyta plus the stramenopiles, is likely to be correct, however, the accuracy of their emergence as sister group to Rhodophyta is questionable. It is generally accepted that the chloroplast of Haptophyta + stramenopiles is secondary (Yoon et al., 2004), and is most likely a symbiont that used to be a free-living rhodophyte. It can thus be suggested that the grouping of Rhodophyta, Haptophyta and the stramenopiles is evidence of this symbiotic event, which, most likely, was followed by the transfer of genes from the symbiont to the nucleus host (as in the case of the origin of the Eukaryota).

Monophyletic Excavata is not recovered in this tree (Figure 3.5), instead the excavates are found to be split into three groups, with the euglenids emerging as an independent lineage. In Figure 3.3, *Euglena* is found in a group with the Apicomplexa, with its position in Figure 3.5 suggesting a persistent attraction between these two groups. As Apicomplexa have a chloroplast (or a chloroplast derived organ, the apicoplast), it can be suggested that this attraction may reflect a complex history of secondary and tertiary endosymbioses. Finally, the Kinetoplastida (*Leishmania* and *Trypanosoma*) are found nested amongst the outgroup taxa. This may be indicative of massive HGT from the prokaryotes, however, as this was not the case in Figure 3.3, this placement could also imply that the tree search is stuck in a suboptimal island of trees, thus necessitating the execution of a more extensive search.

The emergence of the excavates basal to a Chromoalveolata-Plantae grouping (see Figure 3.5) is in keeping with the results of Hampl et al. (2009). Indeed, in general, the topology in Figure 3.5 is consistent with what is currently known of the chromalveolate assemblage. Accordingly, it may be concluded the monophyly of Chromalveolata can be rejected. If Chromoalveolata is monophyletic, attraction with the rhodophytes notwithstanding, one would expect this group to be recovered. The question thus is, if Rhodophyta plus the stramenopiles + haptophytes represents one of the branches of a further endosymbiotic ring, where do the stramenopiles plus haptophytes nest in the eukaryotic tree? This problem, as well as problems concerning the relationships of the excavates, may well be resolved using the phylogenetic signal stripping approach of Pisani et al. (2007).

Localised relationships aside, this phylogeny does provide a robust framework for the eukaryotes. It is interesting that, in general, Figure 3.5 is in keeping with the phylogeny presented by Hampl et al. (2009), and it is possible that disparity between these two schemes (i.e. the specific relationships of the chromalveolates and the recovery of the SAR group) can be attributed to the use of outgroup species in the analyses discussed here. Although far from complete, the phylogeny of the eukaryotes presented here can be considered a more than defensible starting point.

3.4 Conclusion

While the integration of multi protein families into a data set of this size proved quite demanding, both on time and computational resources, it is shown here (with admittedly, an extreme case) to be feasible for large-scale phylogenomic studies. I feel that, where possible, this protocol should be adopted as standard. An obvious merit of such an approach is that genomic sampling will be significantly increased (in this study it triples), thus avoiding ‘trees of one percent’. This is important, as inferences on phylogenetic relationships are more robust if they are made based upon all the evidence available (*sensu* Kluge, 1989).

A second advantage to the use of multi protein families, as clearly exemplified by this analysis, is that they can significantly improve phylogenetic resolution. This will prove particularly pertinent to studies, such as the one discussed here, which are concerned with species that have experienced a complex history of gene transfer. Although the method of resolving gene duplications used in this analysis represents the current state of the art, as mentioned in Chapter 2, it is not ideal. As more and more studies opt for a comprehensive genomic approach, such as the one discussed here, it is expected that more sophisticated methods to account for gene duplication will be devised.

The eukaryote phylogeny is one of the greatest outstanding problems in evolutionary biology. Here, I have examined both the most extensive taxonomic and gene sampling of this group to date. My findings suggest that a more significant amount of phylogenetic information lies within genes that have experienced duplication, suggesting that phylogenomic studies to date have only begun to scratch the surface in relation to this domain. The phylogeny presented here, although rich in sampling, is a long way off addressing a significant portion of eukaryote diversity, particularly the unicellular eukaryotes. With increased genomic sequencing, and the consideration of genes with a history of duplication, it is certain that our view of the eukaryotes will continue to alter at the same pace as it has in the last number of years.

The phylogeny of the eukaryotes presented here is far from complete as there are still many technical aspects that need to be addressed (outlined in Section 3.5), and it is hoped that the implementation of these methodological improvements will contribute to the further improvement of the resolution of the tree. However, it must be acknowledged that even in the light of such remedies, the incongruence observed may be a reflection of reality, due to the high level of lateral gene transfer and endosymbiont transfer in the genomes of the eukaryotes. I thus suggest that current phylogenies proposed for the eukaryotes should be considered with caution.

3.5 Future work

The major draw back in conducting a study of this scale is the huge time cost involved. Each step of the protocol was carried out for almost 24,000 protein families, however, this number was even greater for stages of the protocol that preceded the filtering strategies. Additionally, a study of this magnitude has tested the limits of computational resources currently available, both in terms of hardware and software. As such, a number of procedures relating to this study are ongoing.

In particular, the phylogenetic signal stripping analysis (Pisani et al., 2007) has not yet been completed, and it is expected that this approach will clarify the relationships of the stramenopiles, the Haptophyta, and potentially of the water moulds. Similarly, it is anticipated that adoption of this approach will clarify the relationships within the Excavata (particularly *Euglena*), and amongst the Excavata and the Apicomplexa. Further studies in which I shall attempt to reintroduce the Rhizaria are still ongoing, as are studies in which all animal species will be reintroduced. These will ultimately provide a much better picture of the evolutionary relationships within the Eukaryota.

Another issue that needs to be addressed here is the use of alternative strategies to eliminate LBA. While two of the approaches described in Section 2.1.5 have been employed thus far (namely, a large taxon sampling and appropriate outgroup selection), it is clear that LBA persists as a major problem in this data set (e.g. with regard to the nematodes and flatworms). A similar approach to that used by Aguinaldo et al. (1997), where slow evolving, representative taxa, are used instead of an entire group, may prove useful in reducing LBA (as well as concurrently reducing, the taxon sampling and thus computational time).

Chapter 4: Shape related biases in phylogenetic reconstruction and their impact on our understanding of evolution

4.1 Introduction

The explicit function of a phylogenetic method is to infer a phylogenetic tree; therefore, given appropriate data, any such algorithm will invariably recover a tree. However, despite the inevitable recovery of a tree by any phylogenetic method, it is widely known that the accuracy of the recovered phylogeny is not guaranteed (see Chapters 1 and 2). There are many potential biases that can hinder the recovery of an accurate phylogeny (see Section 1.2), some of which are encountered in the study described in Chapter 2 (e.g. long branch attraction). These biases are well known and many strategies exist to alleviate their impact, some of which are implemented, with success, in Chapter 2 (e.g. targeted taxon sampling).

Such data-driven biases are problematic; however, it can be expected that with more robust methods of analysis, they will, to a large extent, become solvable. A more fundamental (and often underestimated) problem is that phylogenetic methods may be biased toward returning a topology with a certain “shape” (see below), irrespective of the signals in the data. This presents a problem, as methodological biases of this nature are effectively undetectable because they are not a property of the data, but rather of the method used to analyse it. Accordingly, such biases will equally apply to every data set and can result in the accumulation of substantial volumes of misleading results. This is a particularly serious problem for supertree reconstruction, as biased input trees will result in the inference of a biased supertree.

Aside from being used to understand the genealogical relationships of a set of taxa, phylogenetic trees are often used as the fundamental basis for understanding large-scale evolutionary trends (i.e. macroevolutionary patterns; Mooers and Heard, 1997, 2002). A classic use of phylogenies in this context is to identify adaptive radiations (see Harvey and Purvis, 1991), which, it is alleged, should leave strong signatures in phylogenetic trees. Although numerous alternative approaches have been devised to identify adaptive radiations (e.g. species through time plots), many of these methods have a disadvantage in that they require historical knowledge (in the form of fossils), which is not always available for a given group (see Mooers and Heard, 1997). To address this potential problem, methods have been developed that use tree shape (i.e. the level of balance of phylogenetic trees) to identify nodes in a tree that are representative of underlying adaptive radiations (see Guyer and Slowinski, 1993). The rationale behind tree shape based approaches is that lineages that undergo an adaptive radiation will contain more taxa because they have had a greater potential to speciate (Kirkpatrick and Slatkin, 1993). Consequently, this will be reflected in the phylogenetic tree as a more asymmetrical topology.

The shape, and in particular the observed balance or symmetry (i.e. the degree to which nodes divide into subgroups of equal size; see Figure 4.1) of a phylogenetic tree could feasibly be ascribed to the tree reconstruction method (TRM) used to derive the phylogeny (Heard, 1992). Accordingly, approaches that use the shape of a tree to make inferences about evolutionary processes will return fallacious results, if such a bias exists. A number of studies have attempted to ascertain if a particular TRM is responsible for causing a bias with respect to tree balance.

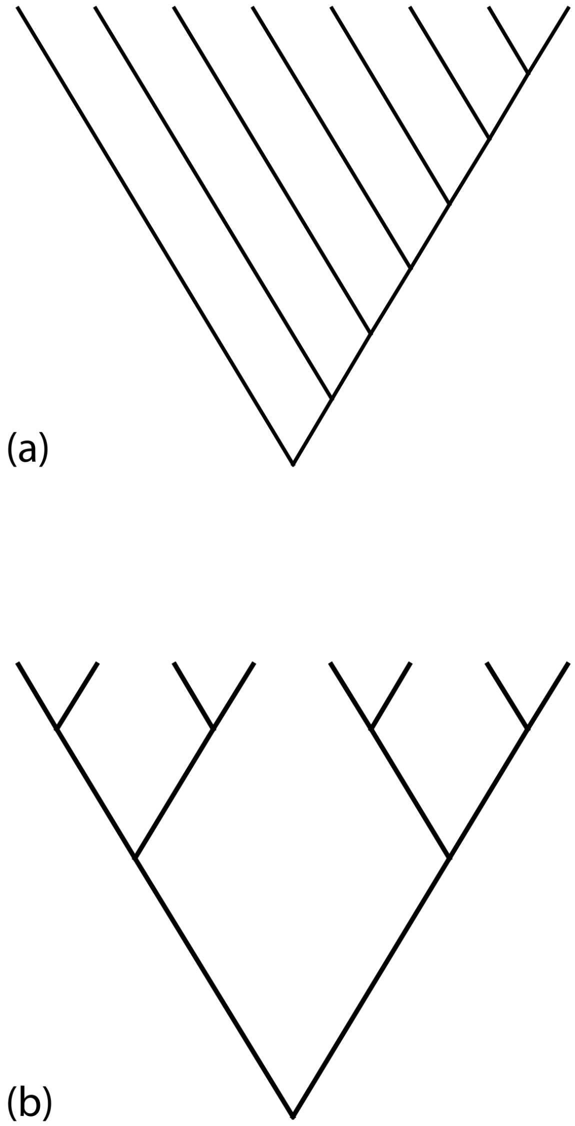


Figure 4.1 The two extremes of tree balance.

Tree (a) is a completely unbalanced, or pectinate, 8 taxon phylogenetic tree, while conversely tree (b) is the fully balanced topology for the same number of taxa.

Colless (1982) was the first to theoretically propose that trees of a cladistic origin were comparably more imbalanced than their phenetic counterparts, a contention that later gained experimental verification (Colless, 1995). However, because trees derived using classic phenetic approaches (e.g. UPGMA) were of poor quality and often misrepresented evolutionary relationships, the findings of these authors were quite correctly deemed to be of little biological significance. Additionally, in the case of UPGMA, midpoint-rooted trees are returned by default leading to the further implication of bias. Later, a study by Huelsenbeck and Kirkpatrick (1996) rather surprisingly found that maximum likelihood produced trees that are more asymmetrical than those derived using any other method (including parsimony), while an earlier study by Heard (1992) found there to be no disagreement between alternative TRMs in terms of balance.

Wilkinson et al. (2005), in the context of studying shape related biases in supertree reconstruction, pointed out that the MRP methods have the potential to generate biased supertrees. More precisely, they show that under the classic Baum and Ragan (1992) coding scheme (used in the supertree analyses of Chapter 2) there is the potential for resulting supertrees to be more asymmetrical than expected given the set of input trees (see Figure 4.2). The authors proposed that this bias could be explained by the fact that parsimony distances are asymmetrical and, by definition, the recoding of an asymmetrical tree of *n* taxa will tend to have a smaller parsimony score than the recoding of a symmetrical tree with the same number of taxa (see Wilkinson et al., 2005). Although Wilkinson et al. (2005) confined the scope of this suggestion to supertree reconstruction, it could logically be extended to the analysis of standard character data under parsimony due to the equivalence between trees and characters (Estabrook et al., 1976). Therefore, it is possible that the observed shape bias of MRP may not be an inherent feature of this

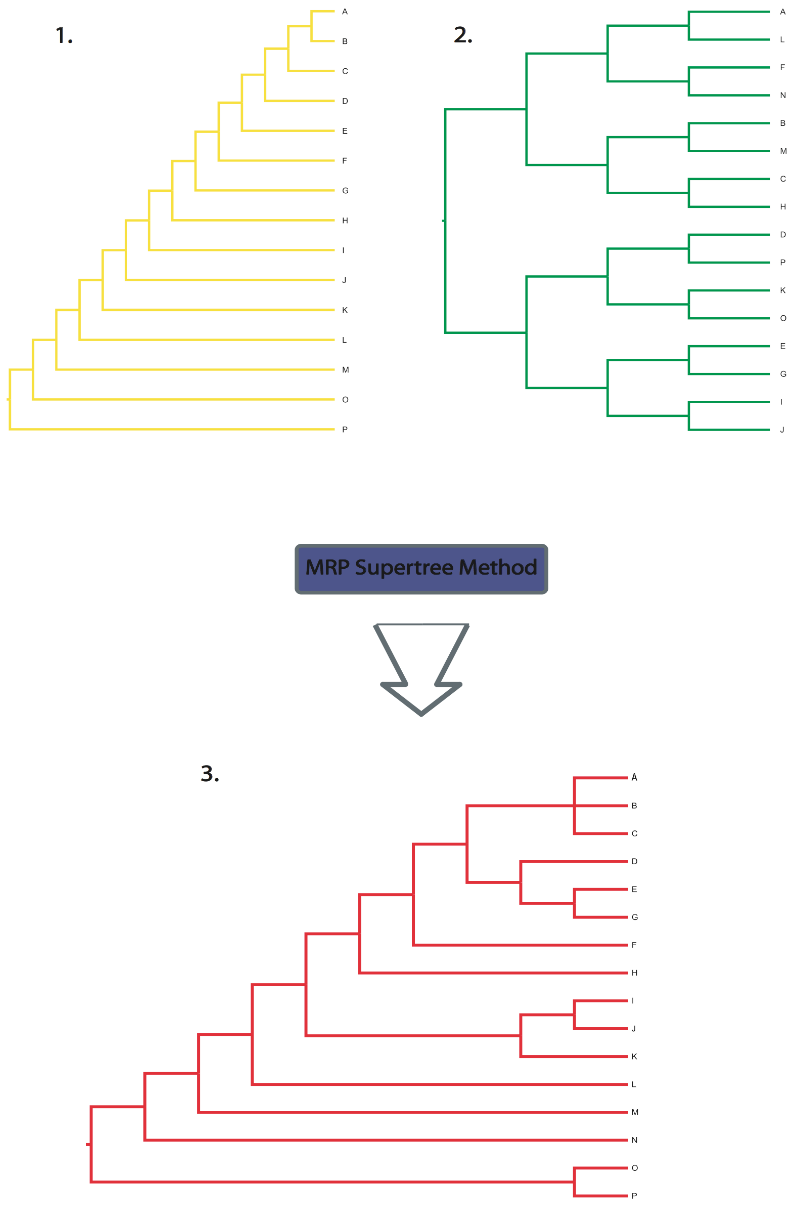


Figure 4.2 Shape bias in the MRP supertree method.

In this example (re-drawn from Wilkinson et al., 2005), when the MRP method is applied to a perfectly imbalanced input tree (1) and a perfectly balanced input tree (2) the resulting tree topology (3) is seen to be more imbalanced.

supertree method, but rather of parsimony more generally.

If it is shown conclusively that parsimony has a shape bias, the impact on phylogenetics will be significant. For example, parsimony is still the method of choice in the analysis of morphological and paleontological data. This persists as the case despite the availability of a model of evolution for morphological characters (Lewis, 2001) and the concomitant ability to implement ML and Bayesian methods for such data. Phylogenies are also used to study the evolution of single morphological or life history characters. This is carried out by tests such as Felsenstein's Independent Constrasts (Felsenstein, 1985), which, like the study of adaptive radiations, would be strongly influenced by the shape of the phylogeny used to study their evolution.

Finally, as pointed out before, phylogenies are used to study large-scale adaptive radiations. If parsimony is biased towards the recovery of asymmetrical trees, using parsimony derived trees, say for a set of fossil species, would be ill advised, as the observed tree shape can no longer be considered a correct reflection of the underlying macroevolutionary process (e.g. speciation). A bias of this kind, which increases the probability of type I error (i.e. a false positive)[4], will at best result in an overestimation of radiation in a lineage. However, and more worryingly, it could result in the erroneous identification of adaptive radiations.

It is thus clear that knowing whether parsimony (or indeed any phylogenetic method) is fundamentally biased with respect to tree shape is of the utmost importance. Surprisingly, since the study of Huelsenbeck and Kirkpatrick (1996), and despite the

[4] Note that a shape bias favouring symmetrical topologies, although problematic, in the context of studying large adaptive radiation, is less worrying as it simply implies that the null hypothesis will be more difficult to reject (i.e. the test will be more stringent) – as type II errors will be artificially increased.

discovery of similar biases in supertree methods (Wilkinson et al. 2005), investigation into the inherent tree shape biases of alternative phylogenetic methods has subsided. Instead, recent studies of tree shape, through the use of null models of tree balance, have focused on whether some biological forces are causing trees to be more imbalanced than expected in the absence of adaptive radiation (e.g. Harcourt-Brown et al., 2001, Blum and Francois, 2006).

Here, in a return to earlier studies such as Huelsenbeck and Kirkpatrick (1996), I wish to determine if the TRMs most frequently used in modern phylogenetics produce topologies that are biased with respect to tree symmetry. Using the same genomic-scale data set of single protein families (from Chapter 2), trees were inferred under a variety of phylogenetic methods. Subsequently, the observed balance of the trees was measured, and compared under various statistical tests to determine which (if any) methods have a bias towards producing more asymmetrical (or symmetrical) trees.

4.1.1 Tree balance metrics

Various statistical measures of tree balance have been defined in the literature. However, many of these can be considered variants, measuring marginally differing characteristics of tree shape. Accordingly, only a cursory introduction for many tree balance metrics (TBMs) is afforded here. The study of Agapow and Purvis (2002) comparing alternative TBMs features an extensive list of tree balance metrics, which are (following the nomenclature of Agapow and Purvis, 2002): $\overline{N}$ (the Sackin index; Sackin, 1972), σ_N^2 (unnamed variant of the Sackin index; Shao and Sokal, 1990, Sackin, 1972), I_c(the Colless index; Colless, 1982, but see Heard, 1992), B_1 (Shao and Sokal, 1990), B_2

(Shao and Sokal, 1990), I' (unnamed method of Fusco and Cronk, 1995; amended by Purvis et al., 2002), $\sum I'$ (unnamed method of Agapow and Purvis, 2002) and Mean I' (a further unnamed method of Agapow and Purvis, 2002). An earlier study of Kirkpatrick and Slatkin (1993), comparing six TBMs, considers a subset of the above, and additionally includes the R statistic (Furnas, 1984); a metric not strictly defined as a TBM but deemed a naturally appropriate measure of tree balance (see Kirkpatrick and Slatkin, 1993). Finally, McKenzie and Steel (2000) relatively recently defined a simple measure known as the cherry count, C_n. Of the above, the most widely used and useful TBMs are $\overline{N}$, I_c and C_n, which are the specific focus of the remainder of this section.

The Sackin index, $\overline{N}$, can more precisely be defined as a measure of tree imbalance, and for a given rooted tree, is calculated by summing the number of internal nodes between each terminal node and the root. More formally, if *t* is the number of terminal nodes, *i* is a given terminal node and N_i the number of nodes between *i* and the root, then the Sackin index is defined as:

$$\overline{N} = \sum_i N_i \qquad i = 1,\ldots,t. \qquad [4.1]$$

This definition is attributed to Shao and Sokal (1990) who introduced summing over all N_i, while Sackin (1972) is credited with defining the *b* (branching) vector for phenograms (Shao and Sokal, 1990, Rogers, 1996), where a given *b* vector element, b_i, can be considered analogous to N_i above. Simulation studies have found that the Sackin index is amongst the best performing measures of tree balance (Agapow and Purvis, 2002, Kirkpatrick and Slatkin, 1993, Shao and Sokal, 1990), particularly for trees with a moderately large taxon sampling (Kirkpatrick and Slatkin, 1993).

Like the Sackin index, I_c or the Colless index (Colless, 1982), can be more strictly classified as a measure of tree imbalance, where the higher the value, the greater the degree of imbalance. This metric is computed by summing up, over all interior nodes of a tree, the absolute difference (T_j), between the number of terminal nodes that descend from each branch of the interior node *j*. As such the Colless index can be defined as:

$$I_c = \sum_j T_j \qquad j=1,\ldots,k(3), \qquad [4.2]$$

where $k(3)$ is the number of internal bifurcations (which have a degree[5] equal to 3 by definition) in a rooted, fully resolved tree (Shao and Sokal, 1990).

Along with proposing this metric, Colless (1982) additionally provided a normalising denominator to adjust for varying tree sizes. Heard (1992), however, provides an amended normalising denominator to account for errors in the original denominator of Colless (1982). Heard's denominator, which was later confirmed by Colless (1995) to be correct, is as follows:

$$\frac{(n-1)(n-2)}{2} \qquad [4.3]$$

where *n* is equal to the number of taxa in a tree. The Colless index can range from 0 to 1, with progression towards 1 representing increased imbalance.

Similar to the Sackin index, it has been shown that the Colless index ranks amongst the most powerful TBMs (Agapow and Purvis, 2002, Kirkpatrick and Slatkin, 1993). Indeed, it is well known that there is a strong association between the Colless and Sackin indices, with the study of Shao and Sokal (1990) reporting them as "highly

[5] The degree of a node is the number of branches connected to that node.

correlated”. However, unlike the Sackin index, the Colless index frequently features as the TBM of choice for many studies. The Colless index has the advantage of being simple and transparent (Heard, 1992, Harcourt-Brown et al., 2001), and has the added benefit of being mathematically tractable (Rogers, 1996).

The cherry count, C_n, of McKenzie and Steel (2000), is calculated by counting the number of paired terminal nodes or leaves that are subtended from the same node. A single terminal node pairing of this kind is known as a cherry (See Figure 4.3); hence the cherry count of a tree is simply the number of cherries present in that tree. Unlike the Colless and Sackin indices, the cherry count behaves as a measure of tree balance and, as such, increases in value with respect to increased balance. Obviously, every TBM, irrespective of whether it was originally defined as a measure of balance or imbalance, can be designed as a measure of either. For example, in the case of the normalised Colless index, by taking $1\text{-}I_c$ of a given tree, this index can be transformed to a measure of tree balance.

With the exception of I_c, TBMs are usually not defined with a prescribed normalising denominator. However, for the TBMs included in their study, Shao and Sokal (1990) proposed independent, general normalising corrections for both measures of tree balance and tree imbalance (including the Colless and Sackin indices). These normalising factors, however, failed to gain widespread adoption. With particular reference to the Colless index (the most commonly used TBM), this failure could be attributed to the subsequent amendment by Heard (1992) to the original normalising denominator of Colless (1982), and his finding that the normalising factors of Shao and Sokal (1990) are in fact size dependent (Heard, 1992). As such, TBMs are typically normalised with respect to a selected model of tree balance (see below).

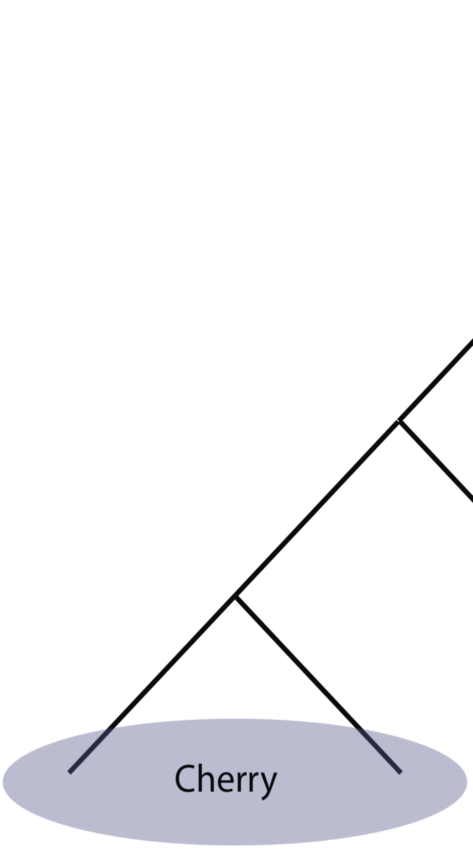


Figure 4.3 Cherries on a phylogenetic tree.

An unrooted phylogenetic tree with five leaves and $C_n = 2$. Cherries are seen at the leaves in shaded ovals (Redrawn from McKenzie and Steel, 2000).

4.1.2 Models of tree balance

Null models of tree balance are useful tools in the study of tree balance. By using such models as reference points, extrapolations regarding the distribution of tree balance in a data set, or evaluations of inferred tree topologies, are made possible (Mooers and Heard, 1997, Matsen, 2006). Use of these models as a benchmark for real data has become the most active line of research into phylogenetic tree shape (Matsen, 2006). There are three distinct, conventional models of tree balance defined; namely the Yule model, the proportion to distinguishable arrangements (PDA) model and the equiprobable types (EPT) model. However, the latter is generally not used in practice.

The Yule model (Yule, 1924), as well as being the first model of tree shape to be described, persists as the most widely used model of speciation (Blum and Francois, 2006). This stochastic model, often referred to as the equal rate Markov (ERM) model, is one of the simplest representations of speciation. This model dictates that each of the branches, or extant species, has the same probability of bifurcating, or giving rise to a novel species (Steel and McKenzie, 2001). As such under this model, from an initial seed species, trees are constructed by selecting equivalently from species the next to speciate. The equal rates element of this model pertains to the fact that although the speciation rate may vary with time, this variation occurs throughout all lineages. Similar to the equiprobability of the speciation of any species occurring, in the occurrence of an extinction event, any of the extant species is equally likely to be eliminated. Note that the model does not strictly incorporate extinction, however, this can be resolved by considering the computed speciation rate to be equal to subtracting the extinction rate from the true rate of speciation, otherwise known as the net rate of diversification (Mooers and Heard, 1997).

The PDA model (Rosen, 1978; see Figure 4.4), sometimes referred to as the uniform model, operates according to the underlying assumption that all tree topologies of *n* taxa are equally likely. That is, a species can be added to any point of a tree with equal probability. As such, this procedure is not strictly a model of evolution (McKenzie and Steel, 2000) as there is no process of growing trees. It has been suggested that this model is biologically driven, as Steel and McKenzie (2001) have shown that the PDA model can be achieved under conditions of explosive radiation (see Blum and Francois, 2006).

The final null model of evolution, EPT (Simberloff et al., 1981) attracts little utilisation. This model dictates that, for a given tree, each different topology is equally likely (Mooers and Heard, 1997). As such under the EPT model, all unlabeled topologies are equiprobable, as opposed to the PDA model where all labelled topologies are equiprobable (Harcourt-Brown et al., 2001). While this model is known to produce the most balanced trees of all the models, it does not realistically reflect the evolutionary process, and is, therefore, safely overlooked by many researchers (Mooers and Heard, 1997).

While the models discussed represent the most frequently used in investigations of tree shape and balance, they are considered far from ideal. Trees produced under the ERM model have been shown to be less balanced than those observed from real data sets (Purvis and Agapow, 2002, Pinelis, 2003; however, this may well be attributed to the fact that the method of inference cannot produce trees that fit this model due to intrinsic method-borne biases i.e. the subject of this chapter). Further to this, trees produced under the PDA model are even more imbalanced than those derived under the ERM model (Mooers and Heard, 1997, Matsen, 2006). Many more contemporary and promising

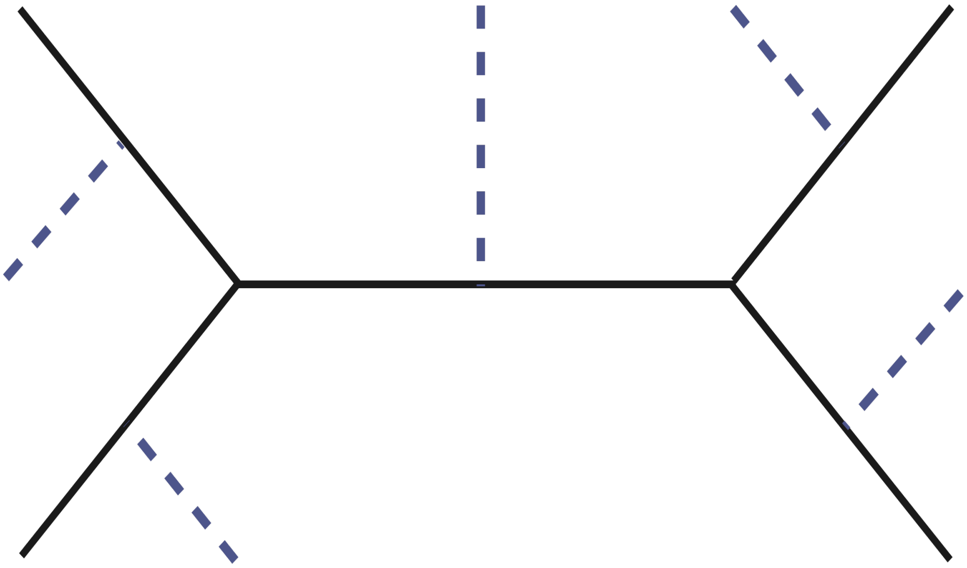


Figure 4.4 Schematic for the Proportion to Distinguishable Arrangements (PDA) model.

Under this model, for an unrooted tree with 4 leaves, there are 5 possible edges where the next edge (blue line) can attach. (Redrawn from McKenzie and Steel, 2000).

models have been proposed (e.g. Losos and Adler, 1995, Heard, 1996, Harcourt-Brown et al., 2001, Steel and McKenzie, 2001), with several focusing on a variation (e.g. Harcourt-Brown et al., 2001) or combination of the two models (Steel and McKenzie, 2001, Pinelis, 2003). These models, however, will not be discussed further, as the focus of this chapter is the methods themselves, not the extent to which trees derived under a particular method fit the expectations of a given model.

4.2 Materials and Methods

4.2.1 Tree reconstruction

For this study, the large-scale genomic data set termed (2) as per section 2.3.1.1 was selected for analysis. This data set was chosen specifically as it has a proven and definite phylogenetic signal (i.e. all protein families have in a previous analysis passed the PTP test; see section 2.2.1.3), which resulted in a robust supertree topology (see Figure 2.6). Single protein families emerging from the protocol described in section 2.3.1.2 were used to test four tree reconstruction methods, namely maximum likelihood (ML), maximum parsimony (MP), neighbour joining (NJ) and Bayesian inference. Since supertree analysis of the aforementioned 2.3.1.1 data set (see Figure 2.6) required the phylogenetic analysis of each protein family under ML, trees constructed using PhyML (Guindon and Gascuel, 2003), as per section 2.3.1.2, could be used directly in this analysis.

PAUP (Swofford, 1998) was used to generate trees for these protein families (i.e. from section 2.3.1.2) under both the MP and NJ criteria. For the MP analysis, each individual protein family data set was allowed to run for 6 hours, with each resulting tree being retained. Settings were as follows: 100 repetitions with heuristic search with the MulTrees option set to on. MP analyses resulted in multiple trees for several of the protein families. For the NJ method, similar to the ML reconstruction, analysis of each protein family resulted in a single tree. Distances were calculated using observed distances (P-distances) and no gamma correction. These conditions are obviously artificial and model selection might influence the topology of the recovered tree (see below for details). However, the use of P-distances should be sufficient here, as the scope of the study is not to recover true trees, but rather a global pattern in the data in terms of tree shape.

Finally, to carry out a Bayesian-based phylogenetic reconstruction of the protein families, the software PhyloBayes (Lartillot and Philippe, 2004) was used. For each data set two independent runs were performed, all of which were carried out under the LG model (Le and Gascuel, 2008). Convergence of the two runs was determined using the automatic stopping rule of PhyloBayes. From each independent run a burn-in of 100 trees was disregarded. A sample of inferred trees for each protein family was then collected (using a PERL script) by extracting one tree in every 100 from each of the independent runs. In addition to this approach, a consensus tree was derived for each protein family using the bpcomp program, as per a typical PhyloBayes analysis. These two sets of Bayesian trees (those directly sampled from the chains, and the majority rule consensus trees derived from the trees in these chains) were used interchangeably as representatives of the Bayesian tree reconstruction method in subsequent analyses.

To ensure as much homogeneity as possible in the study design, for each protein family a corresponding rooting position was imposed on the resultant trees of each TRM. That is, for example, the ML tree derived for a particular protein family was rooted according to the same taxon as the NJ derived tree for the same protein family. The most divergent species in the data set is *Nematostella vectensis*, and where present in a protein family, this taxon was used as the outgroup. If this taxon did not feature in a protein family the next most divergent species, *Trichoplax adherans* was selected instead. Finally if neither of the aforementioned species were present in a protein family, the first species found when reading the sequence alignment of that family was set as the default outgroup. This selection strategy is arbitrary; however, as the aim of this study is to identify global trends in the data, and not the correct relationships among the ingroup species, this rooting scheme should not incur bias (as long as the same outgroup is used for all the trees derived for a given gene family).

Although all gene families were subjected to a robust assessment of phylogenetic signal, some additional selection criteria to preclude additional sources of error were implemented. Firstly, only trees with more than four taxa were selected for further analysis. Four taxon trees are uninformative for this experiment as a rooting strategy based on the use of one outgroup species will, by default, make any such tree fully imbalanced. Further to this, it has been suggested that only trees with 7 or more taxa should be considered when using the Colless TBM. This is because with lower numbers of taxa, the standard deviation of the index becomes so high, that any value for a given tree, under various kinds of models of tree balance, is to be expected (Rogers, 1994, Rogers, 1996, Harcourt-Brown et al., 2001). Indeed, it appears that this effect is not just limited to the Colless index, with Kirkpatrick and Slatkin (1993) suggesting it to be the case in a more general context, citing an even stricter cut off limit of 8 taxa. This problem, however, is somewhat extraneous here, as speciation or extinction events are not the concern of this study, but rather the propensity of tree reconstruction methods to produce trees with an observed degree of balance or imbalance. Accordingly, the use of models of tree balance to normalise the Colless index is avoided. However, the large sample size of protein families did afford the opportunity to err on the side of caution, therefore only trees with 7 or more taxa were considered for analysis.

A further filtering condition was implemented: families were eliminated from the study if their MP analysis resulted in more than 100 trees. This was done in accordance with the study of Harcourt-Brown et al. (2001), where it was suggested that in instances where data sets produced high numbers of most parsimonious trees (MPTs), it might be expected that each topology put forward becomes essentially arbitrary. While these authors do stress that there is inherently no inverse relationship between the amount of MPTs and phylogenetic accuracy, to observe prudence, their approach was adopted here. As such, families that under parsimony attained more than 100 MPTs were not considered. After these precautionary measures were taken into account, a sample of 1,008 protein families was deemed suitable for further analysis.

4.2.2 Calculation of tree balance metrics

Trees passing the above stringency measures were then assessed for their relative degree of balance. To do this two balance metrics were selected: the Colless index and the cherry count. As Shao and Sokal (1990) find a strong correlation between the Sackin and Colless indices, it was deemed sufficient to implement only one measure of imbalance here, selecting the Colless index due to its intuitiveness and model independent normalising denominator.

The cherry count was additionally selected in order to conduct a more comprehensive study. The use of the cherry count allows for a comparison of results obtained using two independent methods, thus controlling for biases that may have been introduced by the use of a specific TBM. However, it must be noted that as a TBM the cherry count is less sensitive than the Colless index (Mark Wilkinson, personal communication). This is important, as it could serve to explain differences between results obtained using these measures. Currently, no method of normalising the cherry count that precludes the use a model of tree balance has been defined. However, for this study the use of non-normalised cherry counts should be irrelevant as comparisons are always made between the same protein families under different TRMs, thus the number of taxa remains invariable.

For the ML and NJ trees the cherry count was calculated using the APE package (Paradis et al., 2004) as implemented in the R program suite. While theoretically the cherry count statistic can be calculated for both binary and non-binary trees, currently the APE package only supports its computation for binary trees. Accordingly, if necessary, trees were firstly resolved using the APE function "multi2di". It has been suggested that polytomies are more likely to be representative of insufficient knowledge rather than a genuine multifurcation and, as such, polytomous trees should not be considered in studies of tree balance (Heard, 1992). However, for the purposes of our comparison, which only considers tree reconstruction methods when applied to the same molecular data set, the resolved trees provided by the multi2di function were deemed acceptable. This is because it is expected that random resolution of polytomies on a data set of 1,008 trees should not introduce any directional bias in the result. By default the APE function "cherry" reports the number of cherries on a given phylogenetic tree, as well as the probability of that cherry count being observed under two null models of tree balance. As the use of such models is avoided in this study, only the raw cherry count was extracted for each considered tree.

To calculate the Colless index for the ML and NJ trees, apTreeshape (Bortolussi et al., 2006), a companion package of APE, specifically dedicated to the analysis of tree shape, was used. As the Colless index can only be applied to binary trees, the APE function, multi2di, was again used to resolve any polytomies present in the input trees. The Colless function in APE generally employs the Yule or PDA models of tree balance for standardisation of the measure. To remove any association with these models the "norm parameter" was set to "null". Normalisation of the resultant Colless index was then carried out using an R script (see Electronic Appendix), according to the following normalising denominator: $\frac{(n-1)(n-2)}{2}$ (Heard, 1992; confirmed by Colless, 1995), where n is equal to the number of taxa in the tree in question.

For the MP trees the cherry count and Colless index were calculated in the same way as above, however, in this instance it was done for every MPT arising from the analysis of a given protein family. Often, in parsimony analyses, a strict (Sokal and Rohlf, 1981) or majority rule consensus (Margush and McMorris, 1981) tree is used to summarise all MPTs derived for a given data set. As I wanted to ensure that the average shape of the recovered MPTs was not misrepresented by the consensus method, the shape of each MPT was individually measured using both TBMs. The resultant TBM values were merged (independently of each other; i.e. all the cherry counts were combined separately to all the combined Colless index values), and from these the mean, mode and median values, for each protein family were calculated. These average statistics, in addition to the TBM values of a random MPT per protein family were used for further analysis. All average statistics were computed using R scripts (see Electronic Appendix).

For the Bayesian trees, the cherry count and Colless TBMs were first calculated for each tree in the sample selected from all those produced by the PhyloBayes analysis (this was repeated for all protein families considered). In a similar fashion to the MP trees, the summary statistics (mean, median and mode) of the total cherry count and normalised Colless index values were computed for each protein family. Additionally, a single tree per protein family was selected at random from those produced by PhyloBayes (after convergence), for which the cherry count and Colless index were calculated and used for further analysis. Note that, in selecting the random trees, both PhyloBayes "chains" were considered. As above, all average statistics and TBMs were calculated using R scripts (see Electronic Appendix). Further to the above, an additional class of Bayesian tree was used in this study. In general, the tree reported in the Bayesian analysis of a given data set is the majority rule consensus of the trees sampled in both chains (which should be an unbiased estimator of the trees in the sample – Holder et al., 2008), therefore, for each protein family TBMs of the Bayesian consensus tree were also calculated and used for further analysis.

4.2.3 Comparison of methods and statistical testing

The computed Colless and cherry counts, for each tree reconstruction method, were then used to perform a series of statistical tests. The study design is such that the TBM values derived from every protein family, analysed under each of the considered phylogenetic methods, are directly compared. Huelsenbeck and Kirkpatrick (1996) suggested that balance metrics for the same data set, analysed under different reconstruction methods, are not independent and therefore cannot be used to test for variation between methods using standard statistical approaches. While this assertion is true, there are various statistical tests that can be used to detect within-subject (in this case the subjects being the protein families) variation, thus avoiding this limitation.

A total of 16 comparisons were carried out, four per TBM, using the two categories of Bayesian trees described above. All sixteen data comparison contained the full complement of calculated TBMs for both the ML and NJ trees, and varied in the type of MP and Bayesian value included. In one set of data comparisons the mean of the MP and Bayesian sample values of a given TBM were used, in another the mode, and so forth, up to a total of eight data comparisons (4 per TBM). Finally for the last eight data comparisons, MP-values of each TBM were separated as before, while the full complement of the Bayesian consensus values for each TBM were used. In this way the ML and NJ TBM values remained the same in all comparisons. See Table 4.1 for a list of data types included in each comparison performed.

To measure variance between the data comparisons a parametric test called analysis of variance (ANOVA), was implemented. For my study design, where the same subject is exposed to various treatments, a specific type of ANOVA, known as repeated measures ANOVA (RM-ANOVA) was required. In order to carry out a RM-ANOVA, the data comparisons were first tested to see if they upheld the assumptions of this parametric test. The power of ANOVA relies on the data adhering to these conditions, which are as follows: each data group must be independent of the others and the data groups must follow a normal distribution. RM-ANOVA additionally assumes that the variance observed between groups is equal across all groups, an assumption known as

Table 4.1 The 16 classes of data comparisons carried out.

In order to conduct a truly comprehensive investigation of the possible inherent shape biases of alternative phylogenetic methods, it was necessary to consider each composite aspect associated with certain methods. For the parsimony trees, the use of consensus approaches was avoided to determine the exact behaviour of this inference method. Indeed, this led to greater homogeneity between the parsimony trees for different protein families, as not all families result in MPTs and, as such, do not require the use of consensus methods. For the Bayesian method of tree inference, two classes of trees were considered, both a random sample and the consensus of all trees sampled per protein family. Here, unlike the parsimony method, the consensus approach is considered, as this is a reflection of how Bayesian trees are inferred in reality, i.e. a consensus approach is consistently used for every protein family and has been shown as the appropriate way to summarise the sample of trees (Holder et al., 2008). However, to rule out any underlying features of the consensus approach for each protein family a sub-sample from each chain was additionally selected.

Data Comparisons			
Bayesian Consensus Trees		Bayesian Sample Trees	
TBM Value		TBM Value	
Colless Mean	Cherry Mean	Colless Mean	Cherry Mean
Colless Median	Cherry Median	Colless Median	Cherry Median
Colless Mode	Cherry Mode	Colless Mode	Cherry Mode
Colless Random	Cherry Random	Colless Random	Cherry Random

sphericity. It is generally accepted that ANOVA is remarkably robust to deviations from normality (Box and Andersen, 1955, Lindman, 1974), however, breach of the sphericity assumption still poses a problem as the RM-ANOVA p-values become untrustworthy. In truth, it has been well documented that the majority of empirical data sets are unlikely to adhere to the stringent sphericity assumption (Vassey and Thayer, 1987, Overall and Doyle, 1994, Keselman, 1998). However, for such cases, it should be noted that two corrections for violations of sphericity exist, namely the Greenhouse-Geisser (Geisser and Greenhouse, 1958) and Huynh-Feldt (Huynh and Feldt, 1976) corrections. These methods, which operate in a similar manner, amend the degrees of freedom relating to the *F* value of the RM-ANOVA to produce a p-value that can be considered a robust alternative to that produced by an uncorrected RM-ANOVA (Keselman, 1998).

As previous analyses of this nature have reported a violation of the assumptions of RM-ANOVA (Heard, 1992), it is to be expected the same will be true for the data discussed here. However, for the sake of completeness, adherence to the assumptions was tested. Each data comparison was assessed for a normal distribution using the Shapiro-Wilk test (Shapiro and Wilk, 1965). This test produces a *W* test statistic, which if, for a given sample, is particularly small, is representative of a departure from normality. This test, implemented in R, returns a p-value, from which the null hypothesis, that the sample follows a normal distribution, can be tested. See Table 4.2a & b for the results of the Shapiro-Wilk test for each data partition.

The sphericity of each of data comparison was then tested using Mauchly's sphericity test (Mauchly, 1940). This test essentially determines whether a covariance matrix of the within-subject variables is proportional to the identity matrix. A matrix is said to be spherical if it has equal variances and covariance of 0. This test, implemented

Table 4.2a Shapiro-Wilk Test (Bayesian Consensus).

Results of the Shapiro-Wilk normality test for data comparisons containing the Bayesian consensus trees. W is the test statistic. The null hypothesis is rejected for p-values < 0.01.

Shapiro-Wilk Normality Test							
TBM Value	**TRMs**	**W statistic**	**P value**	**TBM Value**	**TRMs**	**W statistic**	**P value**
Colless Mean	ML	0.9691886252	8.68E-14	**Cherry Mean**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.973077203996	1.01E-12		MP	0.900699055479	4.97E-25
	Bayesian	0.971901321855	4.69E-13		Bayesian	0.886840917221	1.65E-26
Colless Median	ML	0.9691886252	8.68E-14	**Cherry Median**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.963846241086	4.12E-15		MP	0.883890466745	8.34E-27
	Bayesian	0.971901321855	4.69E-13		Bayesian	0.886840917221	1.65E-26
Colless Mode	ML	0.9691886252	8.68E-14	**Cherry Mode**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.953687690674	2.71E-17		MP	0.877334658256	1.91E-27
	Bayesian	0.971901321855	4.69E-13		Bayesian	0.886840917221	1.65E-26
Colless Random	ML	0.9691886252	8.68E-14	**Cherry Random**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.955363383551	5.86E-17		MP	0.877961187496	2.20E-27
	Bayesian	0.971901321855	4.69E-13		Bayesian	0.886840917221	1.65E-26

Table 4.2b Shapiro-Wilk Test (Bayesian Sample).

Results of the Shapiro-Wilk normality test for data comparisons containing the Bayesian sample trees. W is the test statistic. The null hypothesis is rejected for p-values < 0.01.

Shapiro-Wilk Normality Test							
TBM Value	**TRMs**	**W statistic**	**P value**	**TBM Value**	**TRMs**	**W statistic**	**P value**
Colless Mean	ML	0.9691886252	8.68E-14	**Cherry Mean**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.973077203996	1.01E-12		MP	0.900699055479	4.97E-25
	Bayesian	0.986284054279	4.10E-08		Bayesian	0.914733355708	2.32E-23
Colless Median	ML	0.9691886252	8.68E-14	**Cherry Median**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.963846241086	4.12E-15		MP	0.883890466745	8.34E-27
	Bayesian	0.966166938492	1.49E-14		Bayesian	0.882607128277	6.22E-27
Colless Mode	ML	0.9691886252	8.68E-14	**Cherry Mode**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.953687690674	2.71E-17		MP	0.877334658256	1.91E-27
	Bayesian	0.943466653889	3.61E-19		Bayesian	0.874379860577	1.01E-27
Colless Random	ML	0.9691886252	8.68E-14	**Cherry Random**	ML	0.882827880805	6.54E-27
	NJ	0.956497935362	9.99E-17		NJ	0.881243853796	4.57E-27
	MP	0.955363383551	5.86E-17		MP	0.877961187496	2.20E-27
	Bayesian	0.960871337142	8.62E-16		Bayesian	0.885070314435	1.09E-26

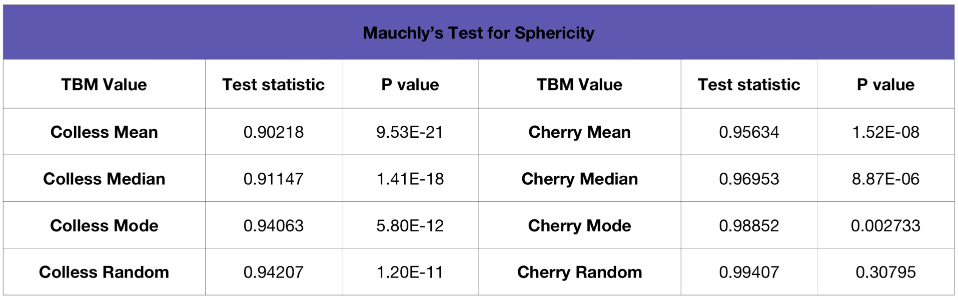
in R, using the Car 2.0-2 package (Fox and Weisberg, 2010), produces a p-value, which if less than 0.05 indicates a violation of sphericity. See Table 4.3 for the results of Mauchly's sphericity test.

Despite violation of both the normality and sphericity assumptions, sixteen one-away RM-ANOVAs in total were carried out as follows: four per TBM for each data partition, according to the average statistics (i.e. mean, median and mode) and one for the random sample from a particular data set. Similar to the tests of the assumptions of ANOVA, these were carried out in R, and required the Car 2.0-2 package (Fox and Weisberg, 2010). All p-values were then adjusted according to the Greenhouse-Geisser and Huynh-Feldt corrections. One final correction to the all ANOVA p-values was carried out, to account for the fact that multiple comparisons are being conducted. This correction, namely the Bonferroni correction, imposes a significance level of alpha divided by the number of considered comparisons. This correction strategy is considered more stringent than other corrections of this nature (Sokal and Rohlf, 1995), and is thus the most widely used multiple comparison correction strategy. Here, as ultimately 32 multiple comparisons (including the Friedman test described below) will be carried out, significance at the 0.01 level, will correspond to a Bonferroni corrected p-value of 0.00031.

As significant differences between treatments were found in all our RM-ANOVAs (see Section 4.3.1), it was necessary to determine in each case which of the TRMs was contributing to the detected variance. In order to do this a comparison between means was conducted. Comparisons of variance (or means) can be carried out in two ways, either *a priori* or *a posteriori* depending on whether there is a question about the data defined before the ANOVA is carried out. There is an important difference

Table 4.3 Mauchlys's sphericity Test.

Results of Mauchly's sphericity test for data comparisons containing the Bayesian consensus trees. The null hypothesis is rejected for p-values < 0.01.



Mauchly's Test for Sphericity					
TBM Value	**Test statistic**	**P value**	**TBM Value**	**Test statistic**	**P value**
Colless Mean	0.90218	9.53E-21	**Cherry Mean**	0.95634	1.52E-08
Colless Median	0.91147	1.41E-18	**Cherry Median**	0.96953	8.87E-06
Colless Mode	0.94063	5.80E-12	**Cherry Mode**	0.98852	0.002733
Colless Random	0.94207	1.20E-11	**Cherry Random**	0.99407	0.30795

between planned (*a priori*) and unplanned (*a posteriori*) comparisons of means as the significance levels are different for the two. Here, as this experiment is addressed without penchant towards any opinion on the data, an *a posteriori* test, known as Tukey's honestly significant difference method (Tukey HSD test), was conducted. The Tukey HSD test is based upon a statistic called the studentized range, which tests differences between the means of paired samples, provided the samples are of equal size (Sokal and Rohlf, 1995). In the Tukey HSD the probability of this difference is also reported. The result of this test, again implemented in R, can be seen in Section 4.3.1.

To bolster confidence in RM-ANOVA results, in light of violations to its assumptions, I additionally performed the Friedman test (Friedman, 1937) on each of the data comparisons (16 in total, carried out as per the ANOVA above), implemented in R. The Friedman test can be considered the non-parametric equivalent of RM-ANOVA, and, as such, differently from standard RM-ANOVA, it does not assume normality or sphericity. The null hypothesis of the Friedman test assumes that all subjects have come from a population with the same median, essentially considering all treatments to have the same effect (Siegel and Castellan, 1988). Like the parametric tests above the Friedman test was implemented in R, results of which can be seen in Section 4.3.1.

As the Friedman test identified significant variance between our subjects it was necessary to conduct a *post-hoc* test to determine between which TRM methods this variance was occurring. To do these pairwise comparisons, a statistical measure called the Wilcoxon-Nemenyi-McDonald-Thompson test (Hollander and Wolfe, 1999) was used. Again this test was implemented in R and required the Coin package (Hothorn et al., 2008). Results can be seen in Section 4.3.1.

4.2.4 Testing of implementation bias

Considering the importance of the ML tree reconstruction method in molecular phylogenetics it was deemed necessary to rule out any bias incurred due to the mode of implementation, i.e. the software used to execute this method. Accordingly, another ML software, RAxML (Stamatakis et al., 2005), was used to derive phylogenetic trees, for which the degree of balance was calculated by means of the TBMs used above.

Tree reconstruction via RAxML was implemented for all 1,008 families passing our stringency tests. To facilitate this, the amino acid model specifications returned by Modelgenerator (Keane et al., 2006) as per section 2.2.1.3, were used. Upon execution of RAxML, the number of substitution rate categories for each protein family was extracted from the output of Modelgenerator. For all other RAxML parameters default settings were maintained. The resultant trees were rooted as described above (see Section 4.2.1), and the TBMs for each tree were computed as outlined for the NJ and ML reconstruction methods. These values were then merged for use in further analyses. The TBM values calculated for the PhyML derived trees (labelled ML above) were additionally used here. Statistical testing of both sets of combined TBMs was then carried out to determine if there was a significant difference between the degree of balance between trees obtained using RAxML and those derived using PhyML.

Due to the non-normal distribution of the balance values in the considered samples, the Wilcoxon-Mann-Whitney (*u* test), which is the non-parametric equivalent of Student's *t* test (effectively a special case of ANOVA), was selected to determine significant variance between the two samples. The *u* test, one of the most powerful of all non-parametric tests, is frequently used to test if two independent samples have the same distribution, in cases where it is wished to avoid the assumptions of the *t* test (Siegel and

Castellan, 1988). The *u* test, like all previous statistical tests, was applied to the RAxML and PhyML samples using R, the results of which can be seen in Section 4.3.2.

Finally, to test the extent to which amino acid model selection could have impinged upon the observed results (particularly for the Bayesian analyses which were all performed under the LG model), two further PhyloBayes analyses were performed. Both supplementary analyses were carried out according to the same protocol as the first PhyloBayes run, but differed in the amino acid model specification. The additional models selected were CAT (Lartillot and Philippe, 2004) and JJT (Jones et al., 1992), as these two models are frequently used in phylogenetic studies. Additionally, selection of these models represented a sample from the two ends of a spectrum of model development, with JTT being an older model, derived under a parsimony-based approach, and CAT being a modern, heterogeneous model (with LG lying in between). The general expectation is that JTT should be, on average, the worst fit to the data, while conversely CAT should provide the best, with the fit of LG (which is an improvement upon the WAG model, which in turn was an improvement upon the JTT model) resting approximately in the middle.

The resultant trees of these PhyloBayes analyses were dealt with in the same manner as the first PhyloBayes analysis. That is, for each amino acid model, TBMs were calculated for a sub-sample of trees per protein family and, additionally, for the consensus tree of that protein family. From the TBMs obtained from the trees in each sample, the mean, median, and mode were calculated and respectively combined, while the TBMs for a random tree per gene were also computed. Finally, the TBMs for the consensus trees were combined, for both TBMs respectively. Ten RM ANOVAs were then carried out as follows: one per average statistic and random partition per TBM (i.e.

eight in total for the sample trees, and one per TBM for the consensus trees; See Section 4.3.3).

4.3 Results and discussion

4.3.1 Analyses of variance: TRMs

Box plots of all Colless index values can be seen in Figure 4.5a-h, while box plots of all cherry count values can be seen in 4.6a-h (data transformations, i.e. logarithm and square root of these values can be seen in Appendix A2). Each constituent ANOVA in this study established that there is a significant difference in terms of tree shape between the various tree reconstruction methods considered. This is true for both the Colless index and cherry count TBMs, and for each variation of the data used (i.e. average values, random value, Bayesian sample or consensus tree). In all cases, the raw p-value reported is < 2.2e-16 (this being the lower bound for ANOVA p-values returned by R), which is upheld by both the Greenhouse-Geisser (GG) and Huynh-Feldt (HF) corrections that account for violation of the sphericity assumption (see Table 4.4a & b). Here, in the interest of being conservative, the upper bound of the p-value range is assumed, therefore, the p-values for all ANOVAs are considered to be 2.2e-16. Each observed p-value is considerably lower than the Bonferroni corrected significance level of 0.00031 and as such the null hypothesis, that the TRMs have equal variance, can be rejected.

Results of the Friedman test are in accordance with those reported for the ANOVA, however, more precise p-values are returned by this test (see Table 4.5a & b). For the Colless index TBM, the smallest p-value (1.54e-22) is observed in the data

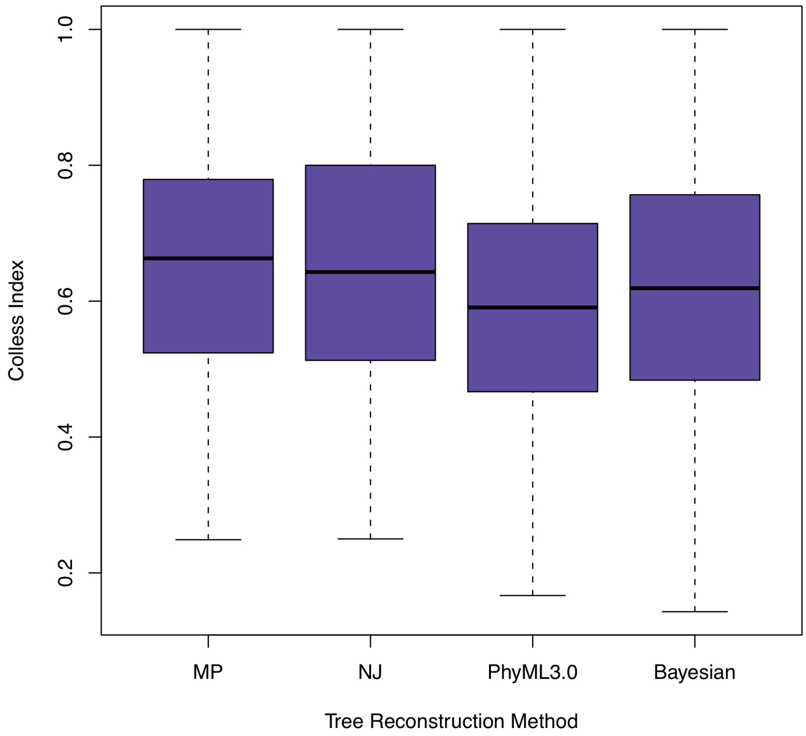


Figure 4.5a Box plot of mean Colless values featuring Bayesian con. trees.

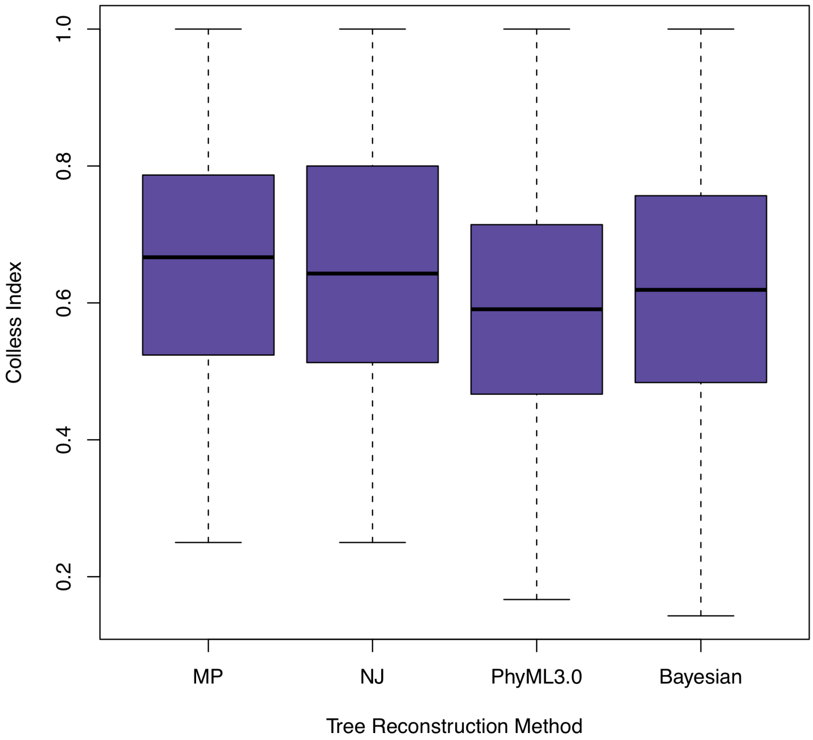


Figure 4.5b Box plot of median Colless values featuring Bayesian con. trees.

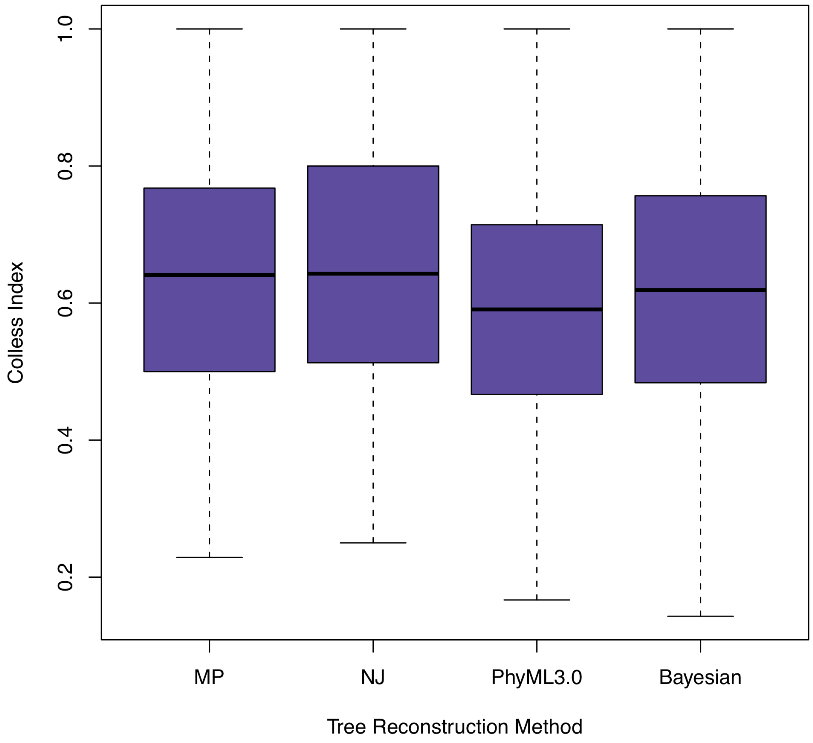


Figure 4.5c Box plot of mode Colless values featuring Bayesian con. trees.

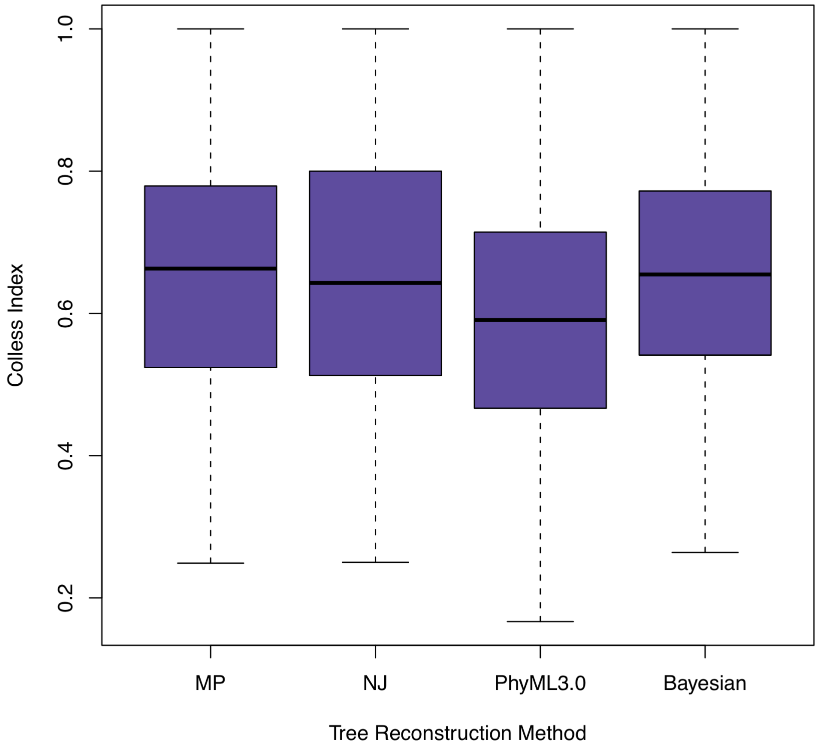


Figure 4.5d Box plot of random Colless values featuring Bayesian con. trees.

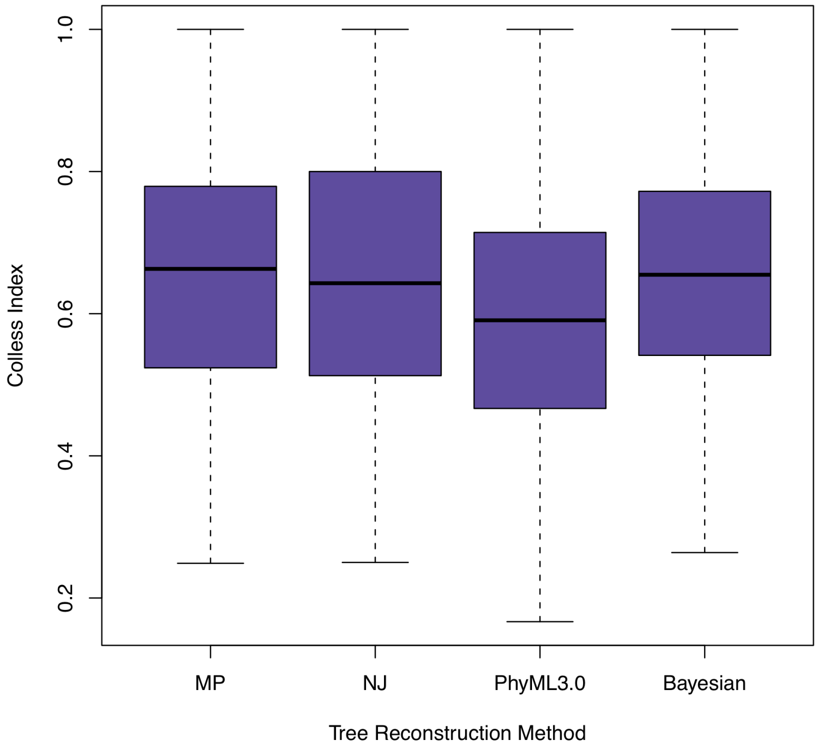


Figure 4.5e Box plot of mean Colless values featuring Bayesian samp. trees.

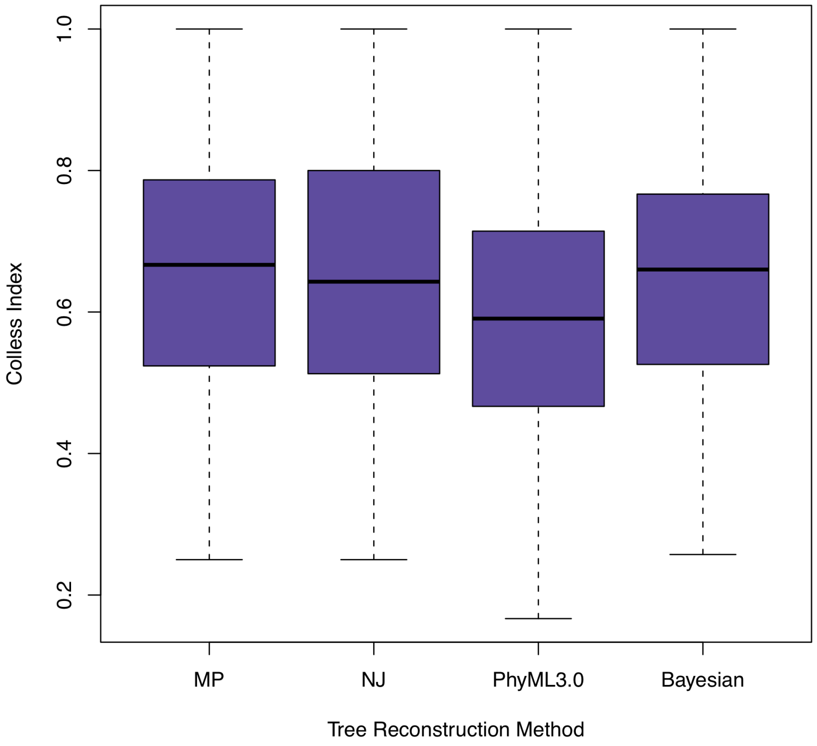


Figure 4.5f Box plot of median Colless values featuring Bayesian samp. trees.

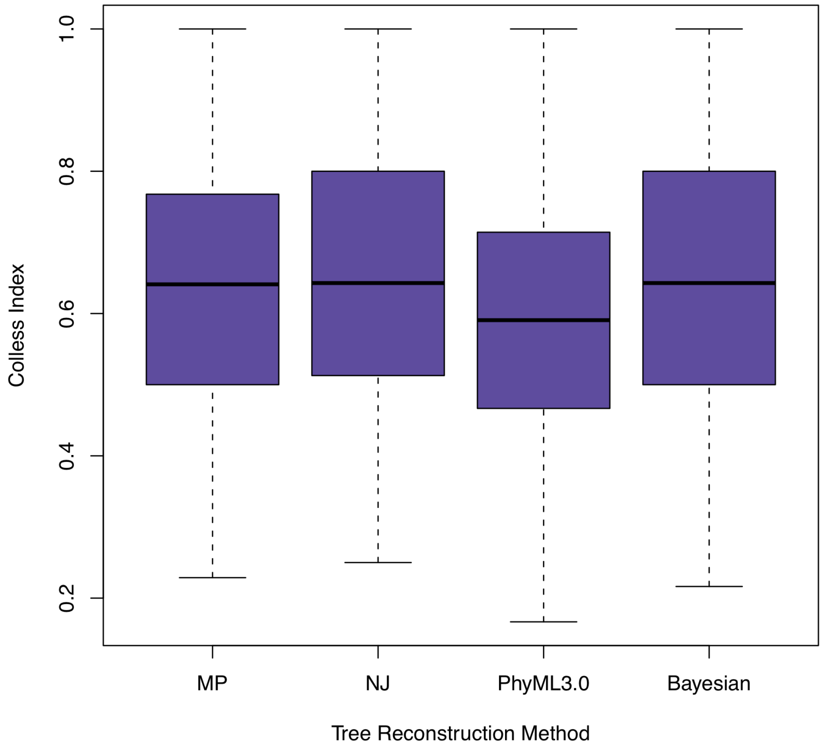


Figure 4.5g Box plot of mode Colless values featuring Bayesian samp. trees.

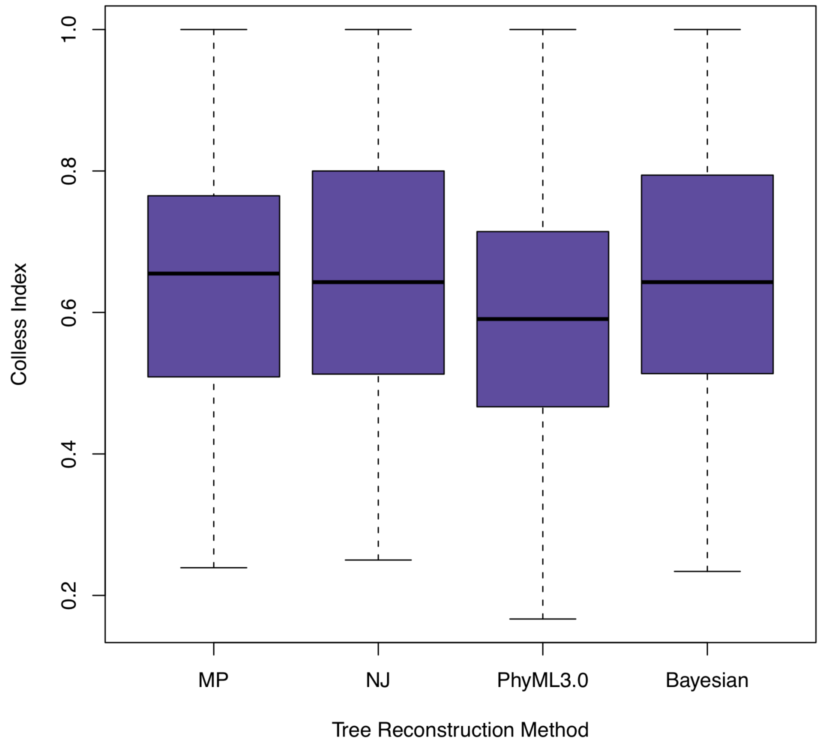


Figure 4.5h Box plot of random Colless values featuring Bayesian samp. trees.

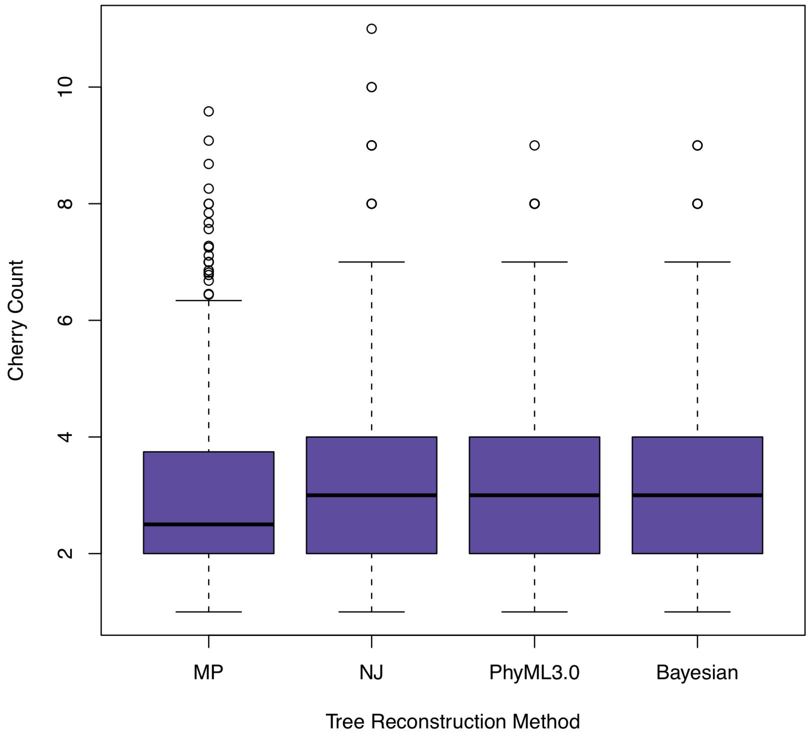


Figure 4.6a Box plot of mean cherry count values featuring Bayesian con. trees.

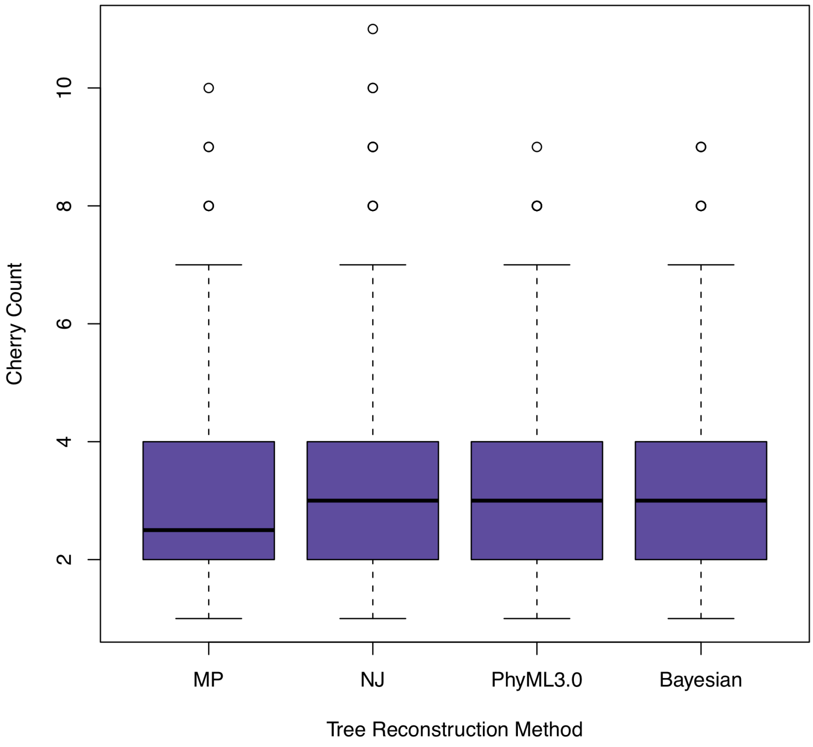


Figure 4.6b Box plot of median cherry count values featuring Bayesian con. trees.

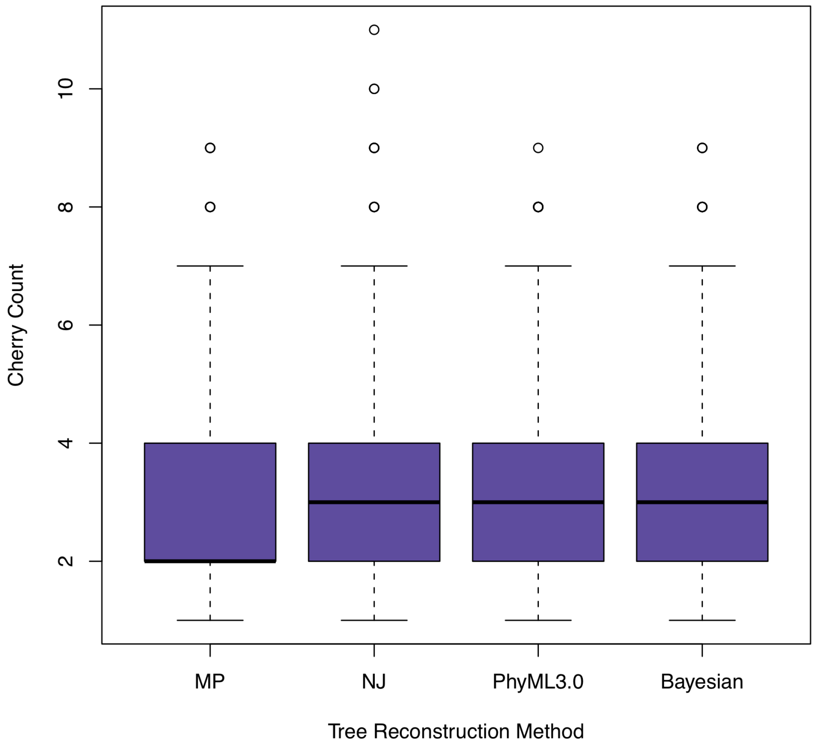


Figure 4.6c Box plot of mode cherry count values featuring Bayesian con. trees.

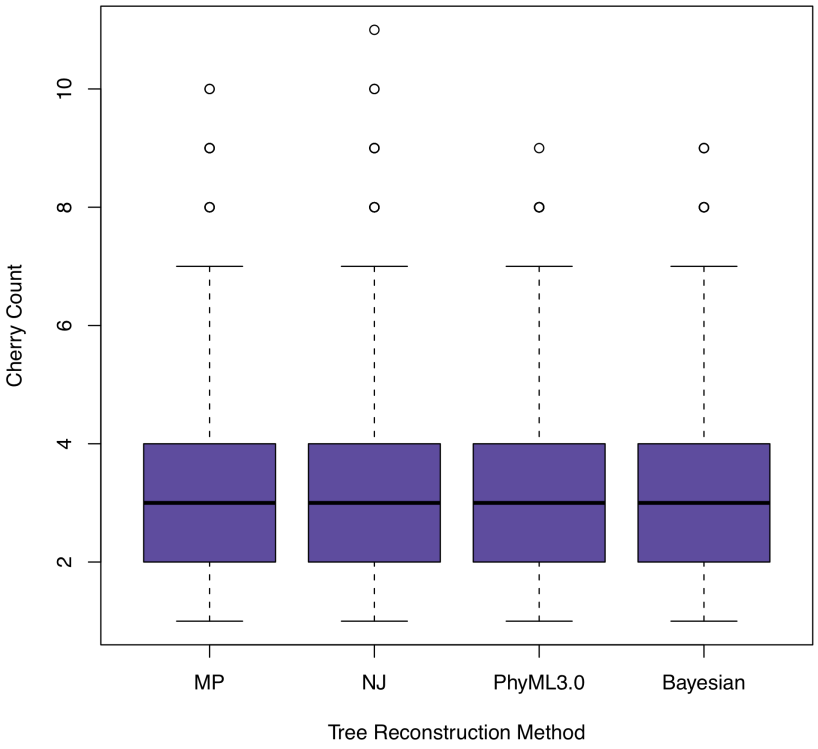


Figure 4.6d Box plot of random cherry count values featuring Bayesian con. trees.

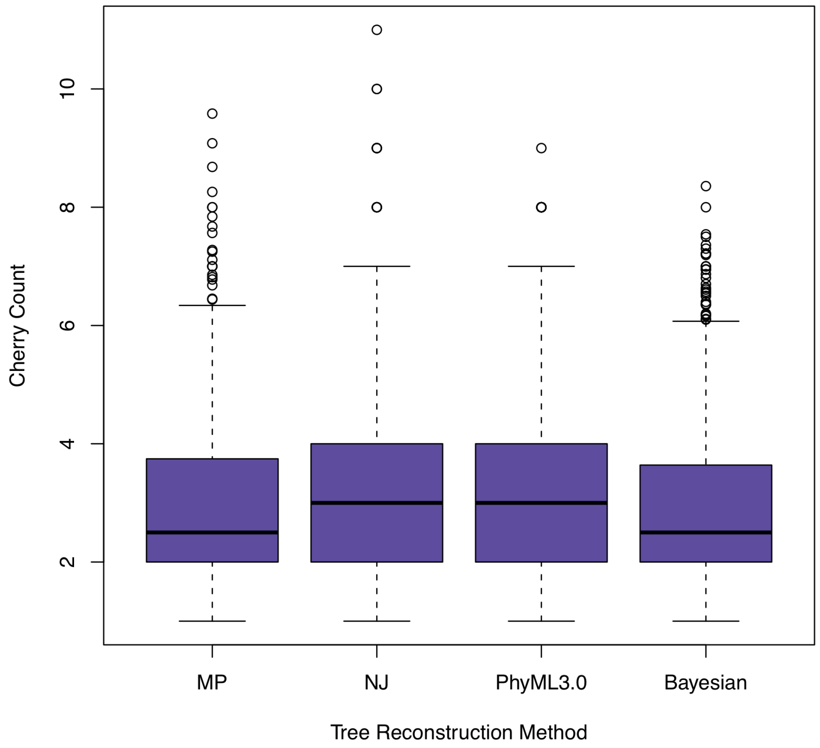


Figure 4.6e Box plot of mean cherry count values featuring Bayesian samp. trees.

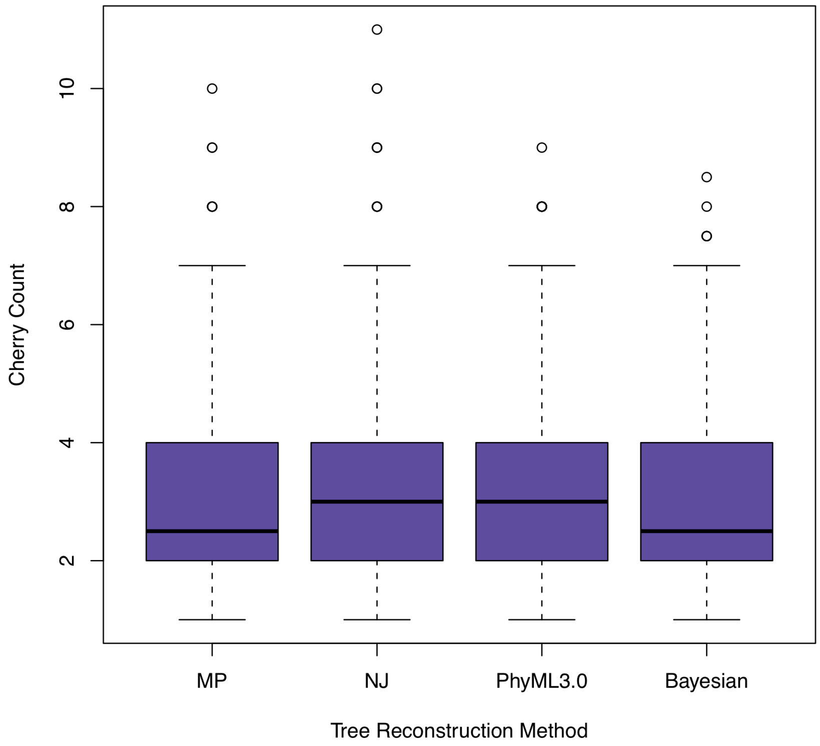


Figure 4.6f Box plot of median cherry count values featuring Bayesian samp. trees.

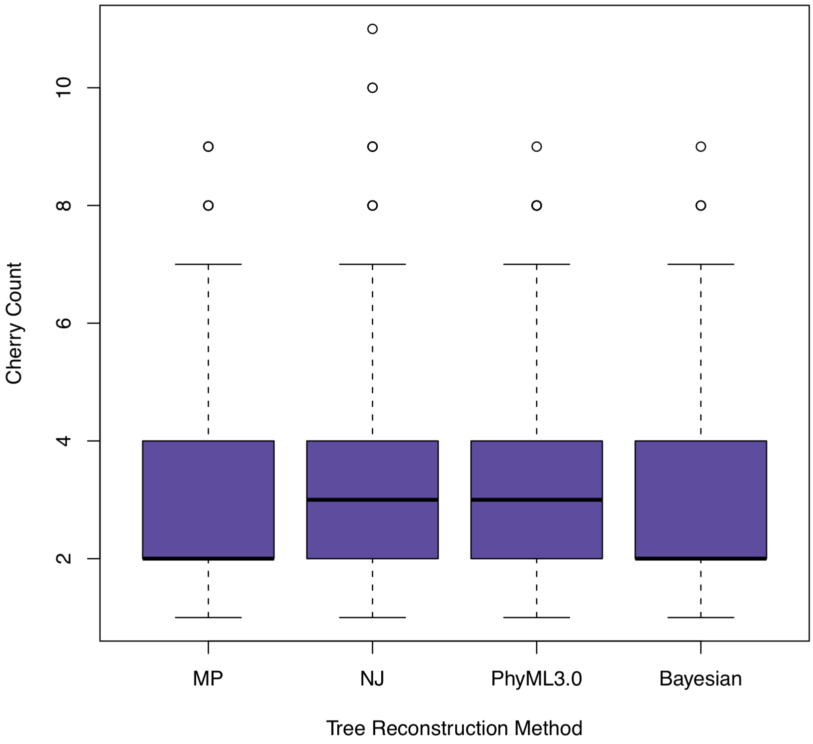


Figure 4.6g Box plot of mode cherry count values featuring Bayesian samp. trees.

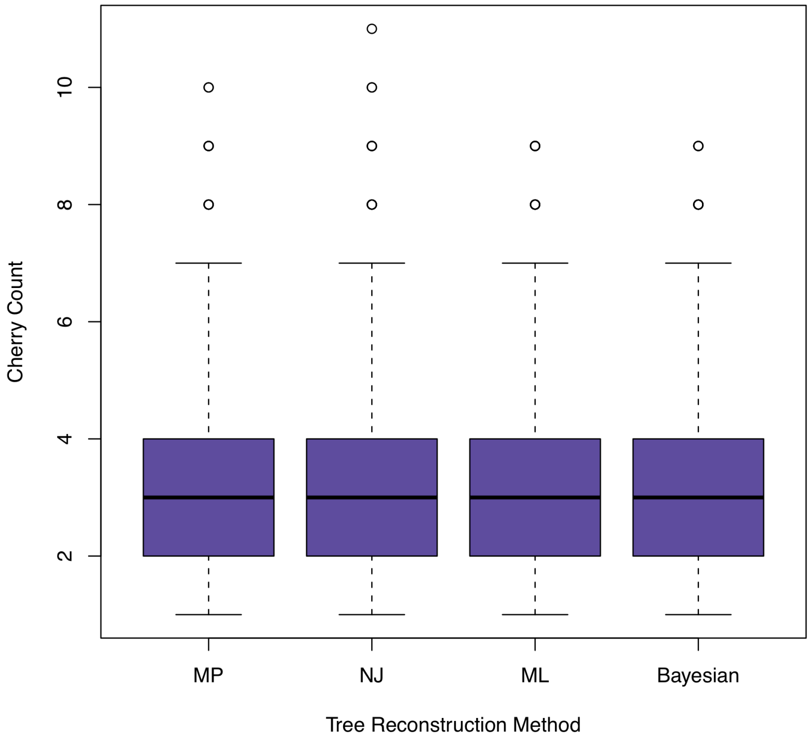
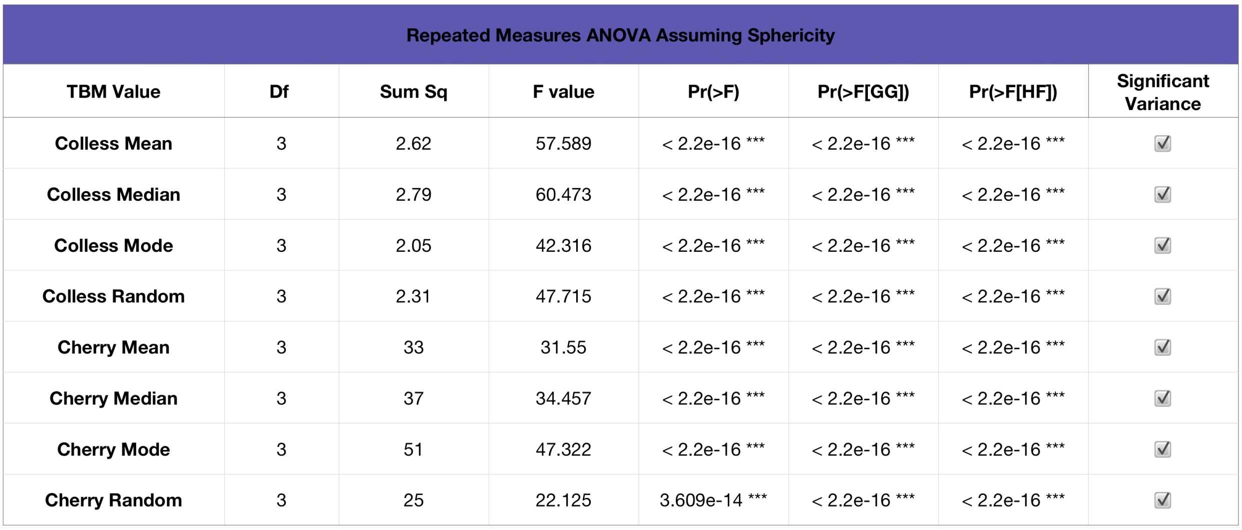


Figure 4.6h Box plot of random cherry count values featuring Bayesian samp. trees.

Table 4.4a ANOVA Bayesian Sample.

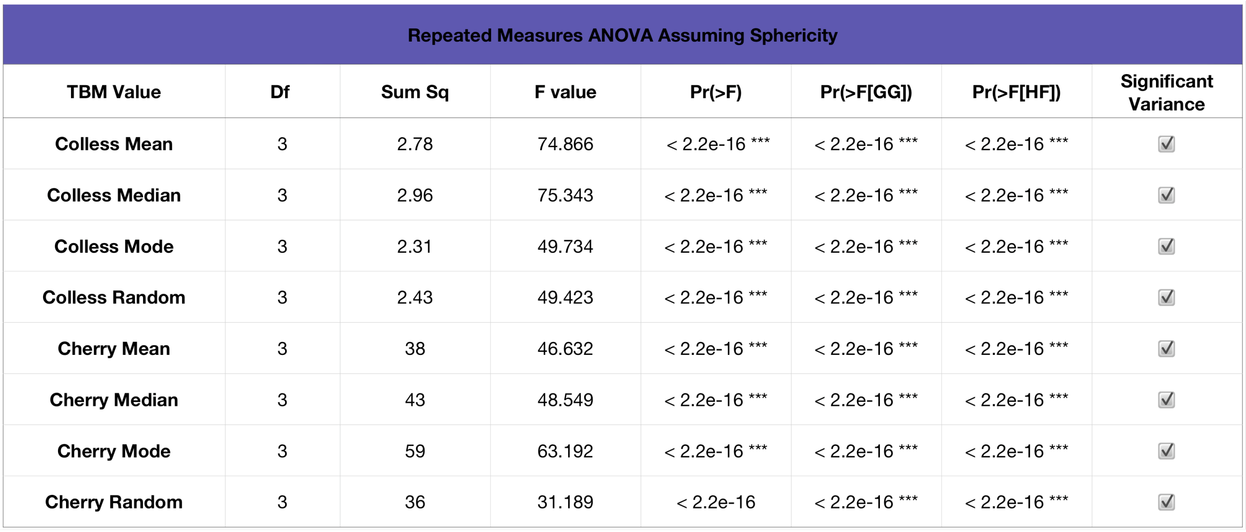
Results from the 8 ANOVAs conducted using the average statistic, random and Bayesian Sample values. The F value is the statistic measure in the ANOVA. Pr(>F) is the p-value, while Pr(>F[GG]) and Pr(>F[HF]) are the Greenhouse-Geisser and Huynh-Feldt corrected p-values respectively(*** p < 0.001).



Repeated Measures ANOVA Assuming Sphericity							
TBM Value	Df	Sum Sq	F value	Pr(>F)	Pr(>F[GG])	Pr(>F[HF])	Significant Variance
Colless Mean	3	2.62	57.589	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Colless Median	3	2.79	60.473	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Colless Mode	3	2.05	42.316	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Colless Random	3	2.31	47.715	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Mean	3	33	31.55	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Median	3	37	34.457	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Mode	3	51	47.322	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Random	3	25	22.125	3.609e-14 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑

Table 4.4b ANOVA Bayesian Consensus.

Results from the 8 ANOVAs conducted using the average statistic, random and Bayesian consensus values. Pr(>F), Pr(>F[GG]) and Pr(>F[HF]) are as per Table 4.4a.

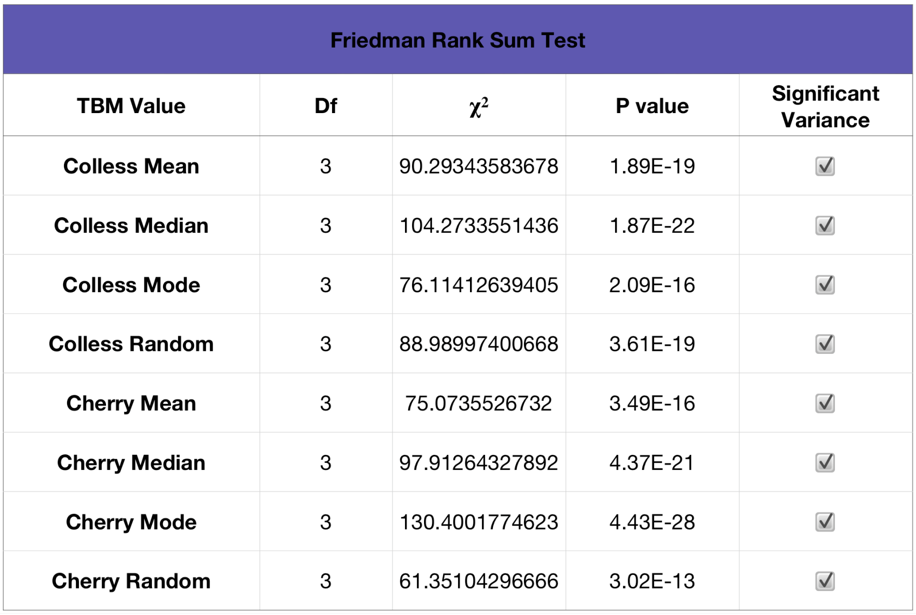


Repeated Measures ANOVA Assuming Sphericity

TBM Value	Df	Sum Sq	F value	Pr(>F)	Pr(>F[GG])	Pr(>F[HF])	Significant Variance
Colless Mean	3	2.78	74.866	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Colless Median	3	2.96	75.343	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Colless Mode	3	2.31	49.734	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Colless Random	3	2.43	49.423	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Mean	3	38	46.632	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Median	3	43	48.549	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Mode	3	59	63.192	< 2.2e-16 ***	< 2.2e-16 ***	< 2.2e-16 ***	☑
Cherry Random	3	36	31.189	< 2.2e-16	< 2.2e-16 ***	< 2.2e-16 ***	☑

Table 4.5a Friedman Test Bayesian sample.

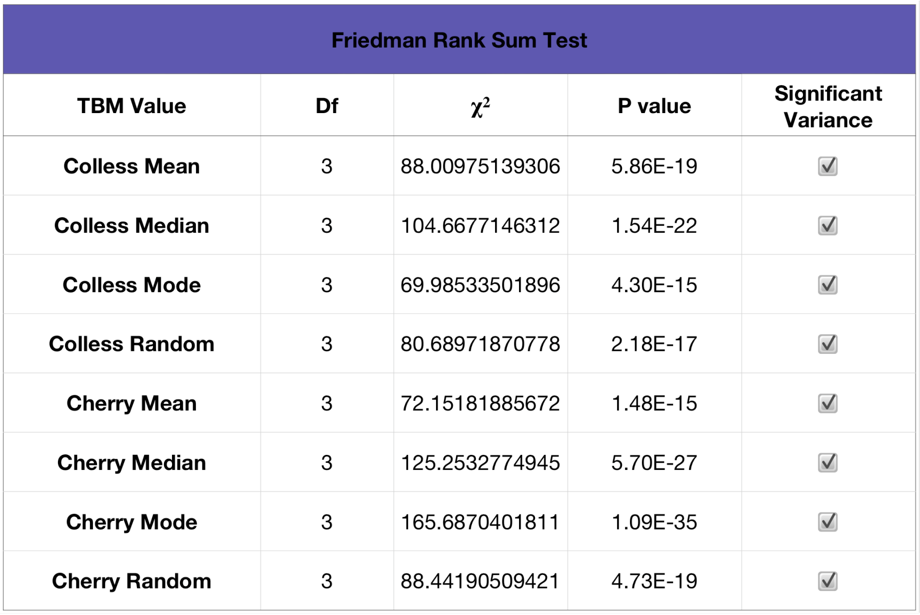
Results from the 8 Friedman Tests conducted using the average statistic, random and Bayesian sample values.



Friedman Rank Sum Test				
TBM Value	**Df**	χ^2	**P value**	**Significant Variance**
Colless Mean	3	90.29343583678	1.89E-19	☑
Colless Median	3	104.2733551436	1.87E-22	☑
Colless Mode	3	76.11412639405	2.09E-16	☑
Colless Random	3	88.98997400668	3.61E-19	☑
Cherry Mean	3	75.0735526732	3.49E-16	☑
Cherry Median	3	97.91264327892	4.37E-21	☑
Cherry Mode	3	130.4001774623	4.43E-28	☑
Cherry Random	3	61.35104296666	3.02E-13	☑

Table 4.5b Friedman Test Bayesian consensus.

Results from the 8 Friedman Tests conducted using the average statistic, random and Bayesian consensus values.



Friedman Rank Sum Test				
TBM Value	**Df**	χ^2	**P value**	**Significant Variance**
Colless Mean	3	88.00975139306	5.86E-19	☑
Colless Median	3	104.6677146312	1.54E-22	☑
Colless Mode	3	69.98533501896	4.30E-15	☑
Colless Random	3	80.68971870778	2.18E-17	☑
Cherry Mean	3	72.15181885672	1.48E-15	☑
Cherry Median	3	125.2532774945	5.70E-27	☑
Cherry Mode	3	165.6870401811	1.09E-35	☑
Cherry Random	3	88.44190509421	4.73E-19	☑

comparison of the median values of the Bayesian sample, while the largest p-value (4.30e-15) is seen in the data comparison of the mode values of the Bayesian sample also. Conversely, for the cherry count the most significant variance (p-value of 1.09e-35) was observed between the mode values of the data comparison featuring the Bayesian sample, while the largest p-value (3.02e-13) was observed in the data comparison of the random values containing the Bayesian consensus. At the Bonferroni corrected 0.00031 significance level, all Friedman test p-values reject the null hypothesis that the TRMs have the same mean.

As it was clear that there was a significant variance in the balance of trees produced by varying TRMs, it was necessary to determine which TRMs were accountable. This was achieved in two ways: for the ANOVA data sets, Tukey's HSD test was carried out (see Table 4.6a-d) and, for the non-parametric data sets, the Wilcoxon-Nemenyi-McDonald-Thompson test was carried out (see Table 4.7a-d). From the results of both *post hoc* tests it can be seen that there is a significant variance between PhyML and all other TRMs. This is true for the cherry count and Colless index, when both the Bayesian consensus and sample values are used, and is reflected in all 32 data comparisons (see Table 4.8). From examining the box plots it can be seen that PhyML is consistently and considerably the most balanced tree reconstruction method as measured by the Colless index (see Figure 4.5a-h). This trend, however, is significantly less pronounced from the cherry count TBM, with the median band of the PhyML and NJ methods continually appearing to be equal (see Figure 4.6a-h), an observation that can additionally be extended to the Bayesian TRM when the consensus tree values are used (see Figure 4.6a-d). This is not unexpected considering the cherry count is less sensitive than Colless index values.

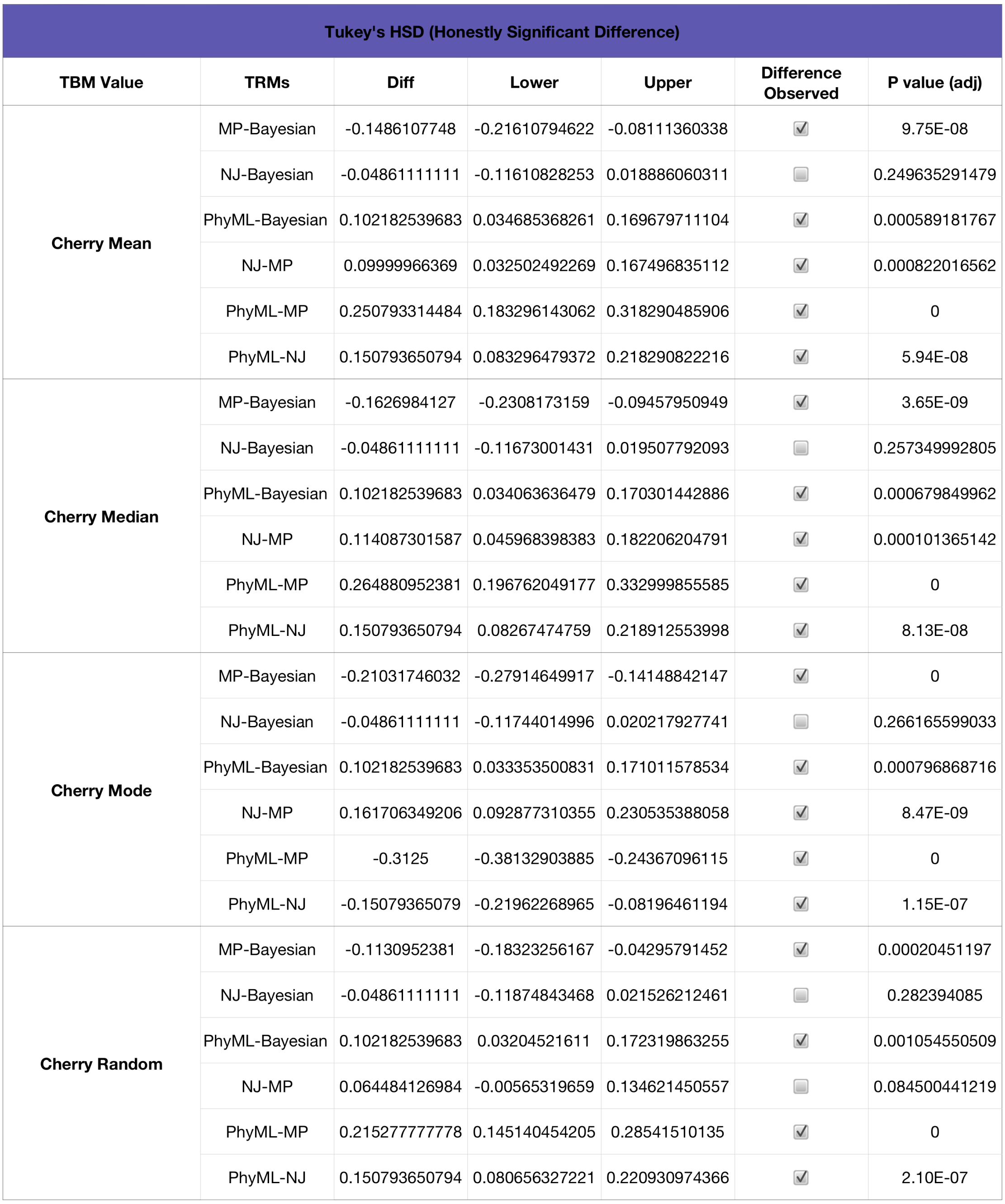
Table 4.6a Tukey HSD Test Bayesian consensus trees (Colless index).

Results from Tukey's Honestly Significant Difference test performed on the mean, median, mode and random Colless index values, featuring the Bayesian consensus trees. Significant variance between the groups is measured at the 0.01 level.

Tukey's HSD (Honestly Significant Difference)						
TBM Value	**TRMs**	**Diff**	**Lower**	**Upper**	**Difference Observed**	**P value (adj)**
Colless Mean	MP-Bayesian	0.033441850397	0.019344406664	0.047539294129	☑	5.31E-09
	NJ-Bayesian	0.030042264187	0.015944820454	0.044139707919	☑	2.77E-07
	PhyML-Bayesian	-0.02944591696	-0.0435433607	-0.01534847323	☑	5.07E-07
	NJ-MP	-0.00339958621	-0.01749702994	0.010697857522	□	0.925759779166
	PhyML-MP	-0.06288776736	-0.07698521109	-0.04879032363	☑	0
	PhyML-NJ	-0.05948818115	-0.07358562488	-0.04539073742	☑	0
Colless Median	MP-Bayesian	0.036653182937	0.022458531732	0.050847834141	☑	0
	NJ-Bayesian	0.030042264187	0.015847612982	0.044236915391	☑	3.42E-07
	PhyML-Bayesian	-0.02944591696	-0.04364056817	-0.01525126576	☑	6.21E-07
	NJ-MP	-0.00661091875	-0.02080556995	0.007583732455	□	0.628688696593
	PhyML-MP	-0.0660990999	-0.08029375111	-0.0519044487	☑	0
	PhyML-NJ	-0.05948818115	-0.07368283236	-0.04529352995	☑	0
Colless Mode	MP-Bayesian	0.01907728998	0.00451641911	0.033638160851	☑	0.004268111448
	NJ-Bayesian	0.030042264187	0.015481393316	0.044603135057	☑	7.27E-07
	PhyML-Bayesian	-0.02944591696	-0.04400678783	-0.01488504609	☑	1.28E-06
	NJ-MP	0.010964974206	-0.00359589666	0.025525845077	□	0.213311541247
	PhyML-MP	-0.04852320694	-0.06308407781	-0.03396233607	☑	0
	PhyML-NJ	-0.05948818115	-0.07404905202	-0.04492731028	☑	0
Colless Random	MP-Bayesian	0.02659324752	0.012047169789	0.041139325251	☑	1.62E-05
	NJ-Bayesian	0.030042264187	0.015496186455	0.044588341918	☑	7.06E-07
	PhyML-Bayesian	-0.02944591696	-0.0439919947	-0.01489983923	☑	1.25E-06
	NJ-MP	0.003449016667	-0.01109706106	0.017995094398	□	0.929099776253
	PhyML-MP	-0.05603916448	-0.07058524222	-0.04149308675	☑	0
	PhyML-NJ	-0.05948818115	-0.07403425888	-0.04494210342	☑	0

Table 4.6b Tukey HSD Test Bayesian consensus trees (cherry count).

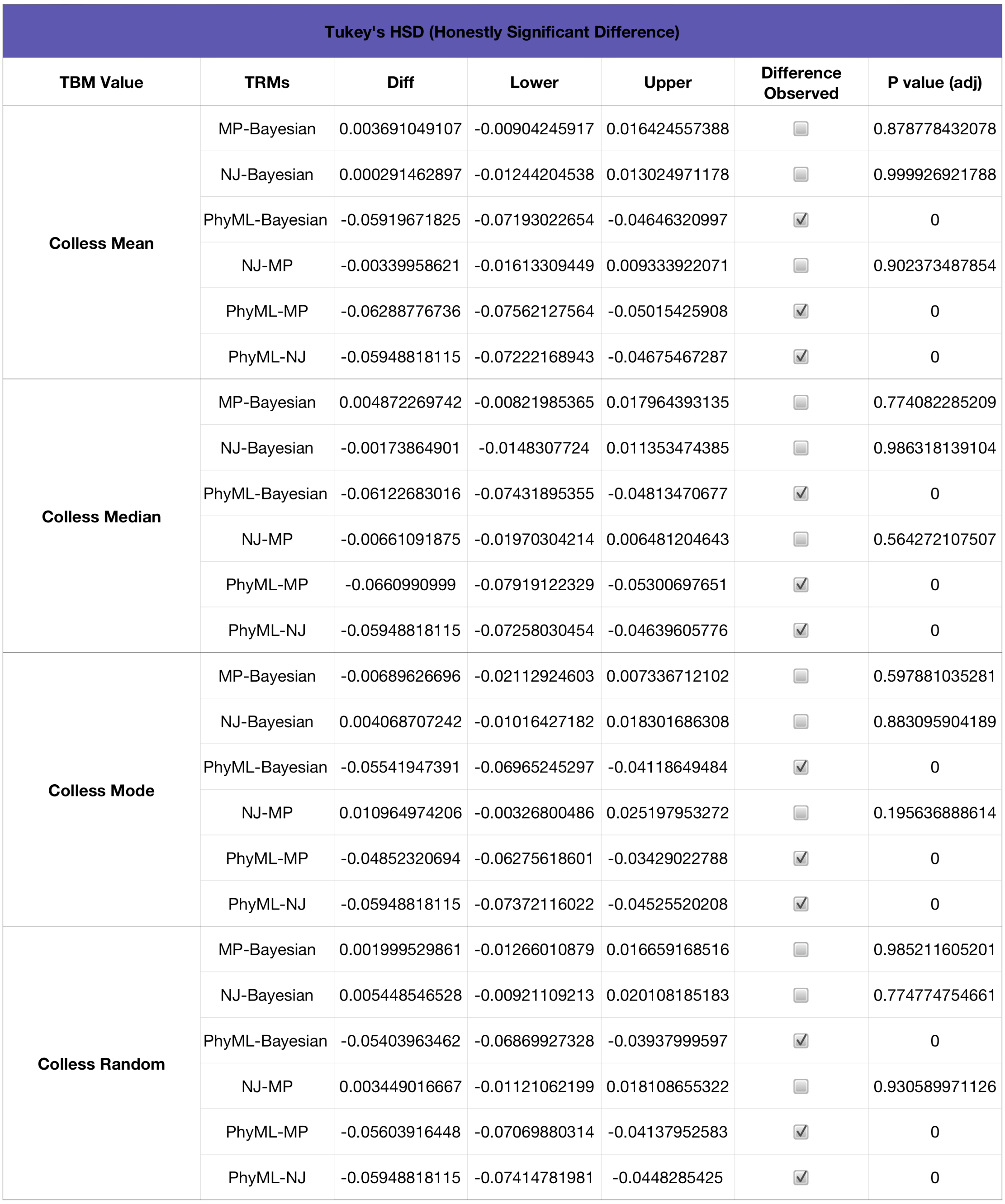
Results from Tukey's Honestly Significant Difference test performed on the mean, median, mode and random cherry count values, featuring the Bayesian consensus trees. Significant variance between the groups is measured at the 0.01 level.



Tukey's HSD (Honestly Significant Difference)						
TBM Value	**TRMs**	**Diff**	**Lower**	**Upper**	**Difference Observed**	**P value (adj)**
Cherry Mean	MP-Bayesian	-0.1486107748	-0.21610794622	-0.08111360338	☑	9.75E-08
	NJ-Bayesian	-0.04861111111	-0.11610828253	0.018886060311	☐	0.249635291479
	PhyML-Bayesian	0.102182539683	0.034685368261	0.169679711104	☑	0.000589181767
	NJ-MP	0.09999966369	0.032502492269	0.167496835112	☑	0.000822016562
	PhyML-MP	0.250793314484	0.183296143062	0.318290485906	☑	0
	PhyML-NJ	0.150793650794	0.083296479372	0.218290822216	☑	5.94E-08
Cherry Median	MP-Bayesian	-0.1626984127	-0.2308173159	-0.09457950949	☑	3.65E-09
	NJ-Bayesian	-0.04861111111	-0.11673001431	0.019507792093	☐	0.257349992805
	PhyML-Bayesian	0.102182539683	0.034063636479	0.170301442886	☑	0.000679849962
	NJ-MP	0.114087301587	0.045968398383	0.182206204791	☑	0.000101365142
	PhyML-MP	0.264880952381	0.196762049177	0.332999855585	☑	0
	PhyML-NJ	0.150793650794	0.08267474759	0.218912553998	☑	8.13E-08
Cherry Mode	MP-Bayesian	-0.21031746032	-0.27914649917	-0.14148842147	☑	0
	NJ-Bayesian	-0.04861111111	-0.11744014996	0.020217927741	☐	0.266165599033
	PhyML-Bayesian	0.102182539683	0.033353500831	0.171011578534	☑	0.000796868716
	NJ-MP	0.161706349206	0.092877310355	0.230535388058	☑	8.47E-09
	PhyML-MP	-0.3125	-0.38132903885	-0.24367096115	☑	0
	PhyML-NJ	-0.15079365079	-0.21962268965	-0.08196461194	☑	1.15E-07
Cherry Random	MP-Bayesian	-0.1130952381	-0.18323256167	-0.04295791452	☑	0.00020451197
	NJ-Bayesian	-0.04861111111	-0.11874843468	0.021526212461	☐	0.282394085
	PhyML-Bayesian	0.102182539683	0.03204521611	0.172319863255	☑	0.001054550509
	NJ-MP	0.064484126984	-0.00565319659	0.134621450557	☐	0.084500441219
	PhyML-MP	0.215277777778	0.145140454205	0.28541510135	☑	0
	PhyML-NJ	0.150793650794	0.080656327221	0.220930974366	☑	2.10E-07

Table 4.6c Tukey HSD Test Bayesian Sample trees (Colless index).

Results from Tukey’s Honestly Significant Difference test performed on the mean, median, mode and random Colless index values, featuring the Bayesian sample trees. Significant variance between the groups is measured at the 0.01 level.

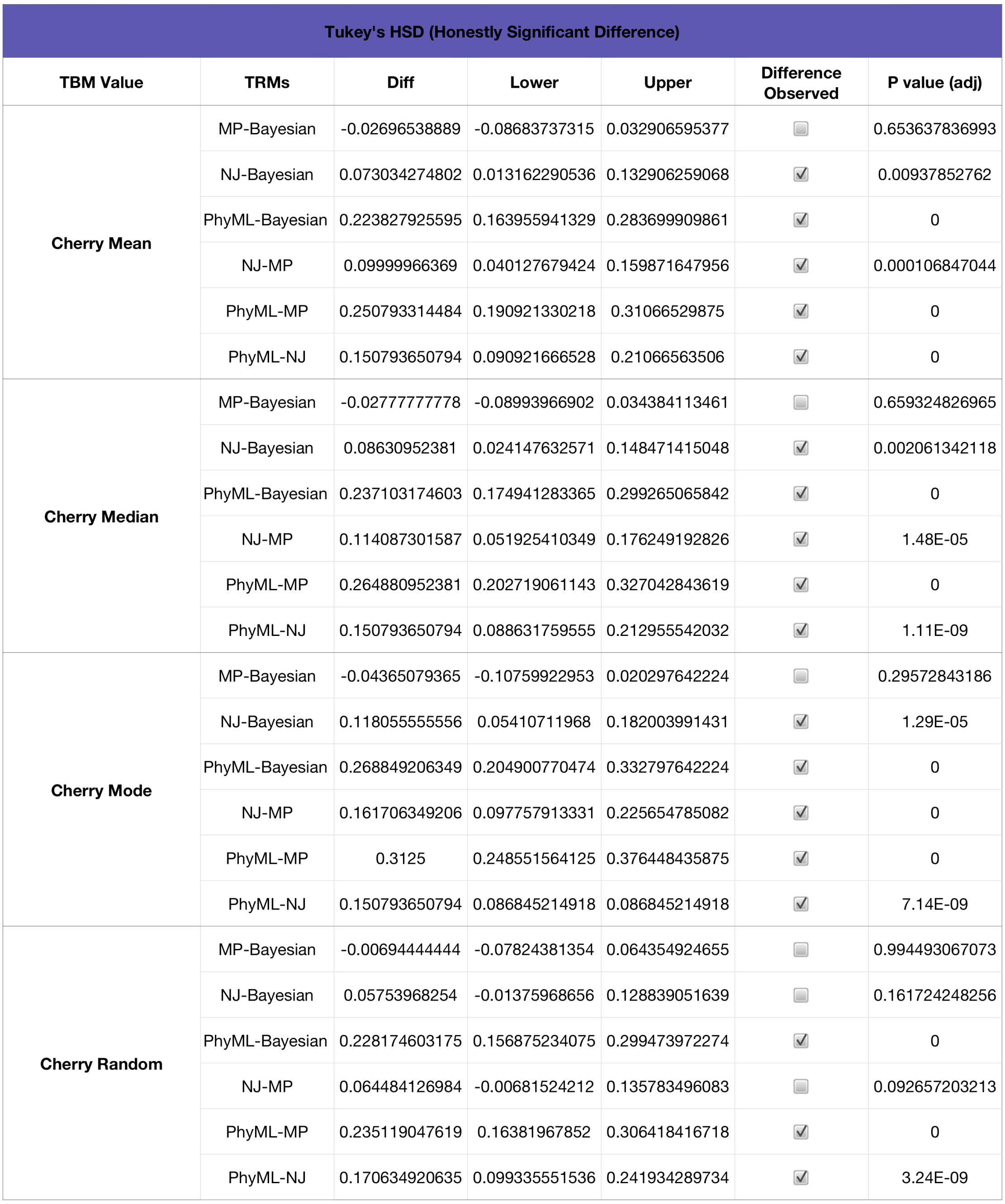


Tukey's HSD (Honestly Significant Difference)

TBM Value	TRMs	Diff	Lower	Upper	Difference Observed	P value (adj)
Colless Mean	MP-Bayesian	0.003691049107	-0.00904245917	0.016424557388	☐	0.878778432078
	NJ-Bayesian	0.000291462897	-0.01244204538	0.013024971178	☐	0.999926921788
	PhyML-Bayesian	-0.05919671825	-0.07193022654	-0.04646320997	☑	0
	NJ-MP	-0.00339958621	-0.01613309449	0.009333922071	☐	0.902373487854
	PhyML-MP	-0.06288776736	-0.07562127564	-0.05015425908	☑	0
	PhyML-NJ	-0.05948818115	-0.07222168943	-0.04675467287	☑	0
Colless Median	MP-Bayesian	0.004872269742	-0.00821985365	0.017964393135	☐	0.774082285209
	NJ-Bayesian	-0.00173864901	-0.0148307724	0.011353474385	☐	0.986318139104
	PhyML-Bayesian	-0.06122683016	-0.07431895355	-0.04813470677	☑	0
	NJ-MP	-0.00661091875	-0.01970304214	0.006481204643	☐	0.564272107507
	PhyML-MP	-0.0660990999	-0.07919122329	-0.05300697651	☑	0
	PhyML-NJ	-0.05948818115	-0.07258030454	-0.04639605776	☑	0
Colless Mode	MP-Bayesian	-0.00689626696	-0.02112924603	0.007336712102	☐	0.597881035281
	NJ-Bayesian	0.004068707242	-0.01016427182	0.018301686308	☐	0.883095904189
	PhyML-Bayesian	-0.05541947391	-0.06965245297	-0.04118649484	☑	0
	NJ-MP	0.010964974206	-0.00326800486	0.025197953272	☐	0.195636888614
	PhyML-MP	-0.04852320694	-0.06275618601	-0.03429022788	☑	0
	PhyML-NJ	-0.05948818115	-0.07372116022	-0.04525520208	☑	0
Colless Random	MP-Bayesian	0.001999529861	-0.01266010879	0.016659168516	☐	0.985211605201
	NJ-Bayesian	0.005448546528	-0.00921109213	0.020108185183	☐	0.774774754661
	PhyML-Bayesian	-0.05403963462	-0.06869927328	-0.03937999597	☑	0
	NJ-MP	0.003449016667	-0.01121062199	0.018108655322	☐	0.930589971126
	PhyML-MP	-0.05603916448	-0.07069880314	-0.04137952583	☑	0
	PhyML-NJ	-0.05948818115	-0.07414781981	-0.0448285425	☑	0

Table 4.6d Tukey HSD Test Bayesian sample trees (cherry count).

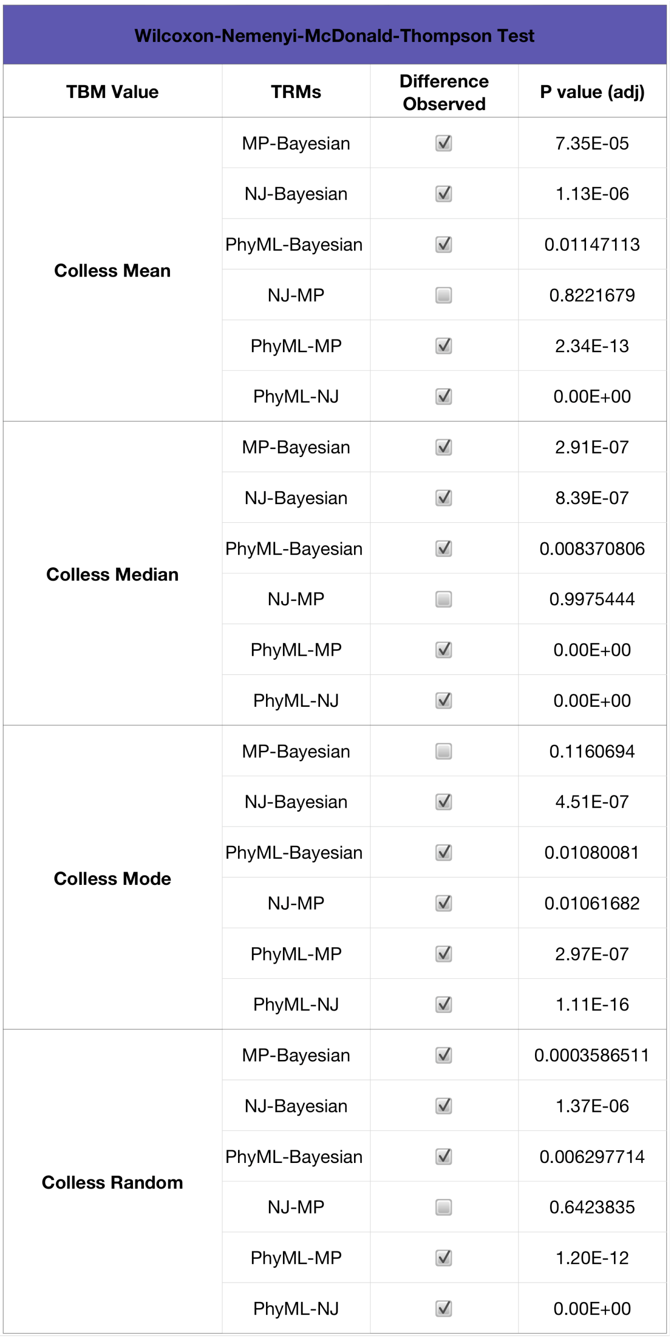
Results from Tukey's Honestly Significant Difference test performed on the mean, median, mode and random cherry count values, featuring the Bayesian sample trees. Significant variance between the groups is measured at the 0.01 level.



Tukey's HSD (Honestly Significant Difference)						
TBM Value	**TRMs**	**Diff**	**Lower**	**Upper**	**Difference Observed**	**P value (adj)**
Cherry Mean	MP-Bayesian	-0.02696538889	-0.08683737315	0.032906595377	☐	0.653637836993
	NJ-Bayesian	0.073034274802	0.013162290536	0.132906259068	☑	0.00937852762
	PhyML-Bayesian	0.223827925595	0.163955941329	0.283699909861	☑	0
	NJ-MP	0.09999966369	0.040127679424	0.159871647956	☑	0.000106847044
	PhyML-MP	0.250793314484	0.190921330218	0.31066529875	☑	0
	PhyML-NJ	0.150793650794	0.090921666528	0.21066563506	☑	0
Cherry Median	MP-Bayesian	-0.02777777778	-0.08993966902	0.034384113461	☐	0.659324826965
	NJ-Bayesian	0.08630952381	0.024147632571	0.148471415048	☑	0.002061342118
	PhyML-Bayesian	0.237103174603	0.174941283365	0.299265065842	☑	0
	NJ-MP	0.114087301587	0.051925410349	0.176249192826	☑	1.48E-05
	PhyML-MP	0.264880952381	0.202719061143	0.327042843619	☑	0
	PhyML-NJ	0.150793650794	0.088631759555	0.212955542032	☑	1.11E-09
Cherry Mode	MP-Bayesian	-0.04365079365	-0.10759922953	0.020297642224	☐	0.29572843186
	NJ-Bayesian	0.118055555556	0.05410711968	0.182003991431	☑	1.29E-05
	PhyML-Bayesian	0.268849206349	0.204900770474	0.332797642224	☑	0
	NJ-MP	0.161706349206	0.097757913331	0.225654785082	☑	0
	PhyML-MP	0.3125	0.248551564125	0.376448435875	☑	0
	PhyML-NJ	0.150793650794	0.086845214918	0.086845214918	☑	7.14E-09
Cherry Random	MP-Bayesian	-0.00694444444	-0.07824381354	0.064354924655	☐	0.994493067073
	NJ-Bayesian	0.05753968254	-0.01375968656	0.128839051639	☐	0.161724248256
	PhyML-Bayesian	0.228174603175	0.156875234075	0.299473972274	☑	0
	NJ-MP	0.064484126984	-0.00681524212	0.135783496083	☐	0.092657203213
	PhyML-MP	0.235119047619	0.16381967852	0.306418416718	☑	0
	PhyML-NJ	0.170634920635	0.099335551536	0.241934289734	☑	3.24E-09

Table 4.7a Wilcoxon-Nemenyi-McDonald-Thompson Test Bayesian consensus trees (Colless index).

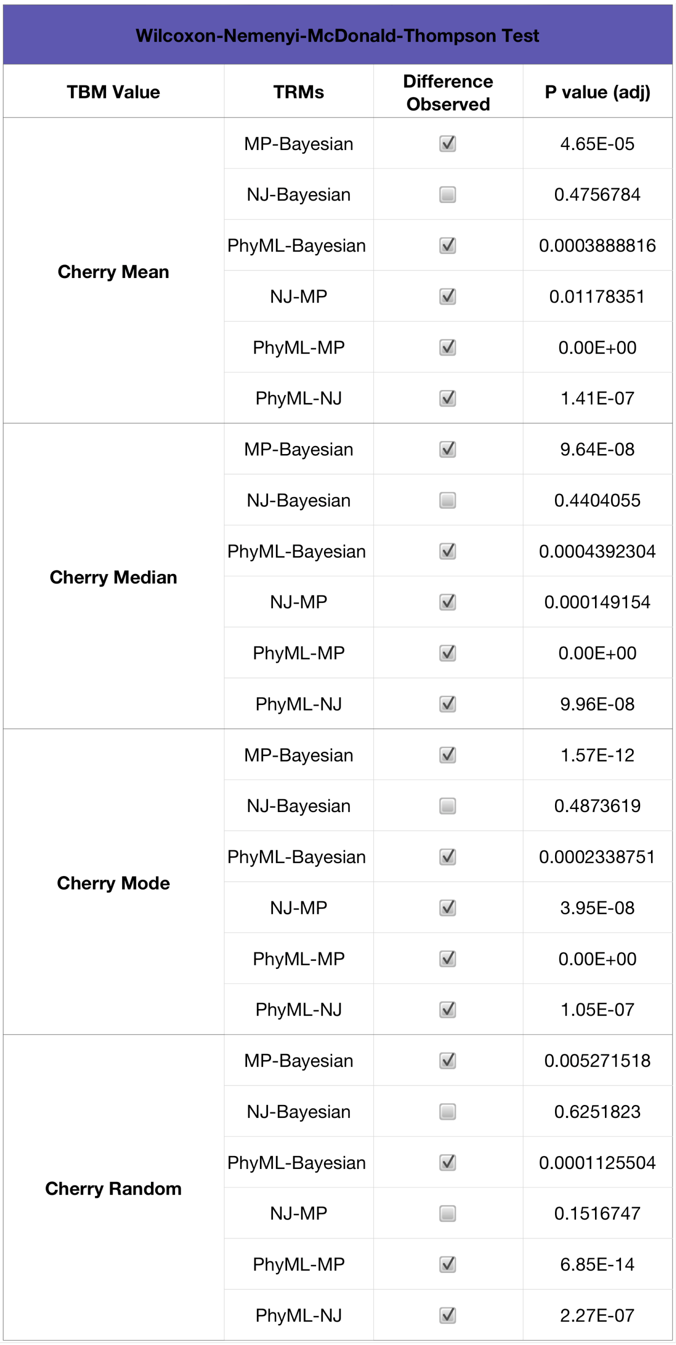
Results from the Wilcoxon-Nemenyi-McDonald-Thompson test performed on the mean, median, mode and random Colless index values, featuring the Bayesian consensus trees. Significant variance between the groups is measured at the 0.01 level.



Wilcoxon-Nemenyi-McDonald-Thompson Test			
TBM Value	**TRMs**	**Difference Observed**	**P value (adj)**
Colless Mean	MP-Bayesian	☑	7.35E-05
	NJ-Bayesian	☑	1.13E-06
	PhyML-Bayesian	☑	0.01147113
	NJ-MP	☐	0.8221679
	PhyML-MP	☑	2.34E-13
	PhyML-NJ	☑	0.00E+00
Colless Median	MP-Bayesian	☑	2.91E-07
	NJ-Bayesian	☑	8.39E-07
	PhyML-Bayesian	☑	0.008370806
	NJ-MP	☐	0.9975444
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	0.00E+00
Colless Mode	MP-Bayesian	☐	0.1160694
	NJ-Bayesian	☑	4.51E-07
	PhyML-Bayesian	☑	0.01080081
	NJ-MP	☑	0.01061682
	PhyML-MP	☑	2.97E-07
	PhyML-NJ	☑	1.11E-16
Colless Random	MP-Bayesian	☑	0.0003586511
	NJ-Bayesian	☑	1.37E-06
	PhyML-Bayesian	☑	0.006297714
	NJ-MP	☐	0.6423835
	PhyML-MP	☑	1.20E-12
	PhyML-NJ	☑	0.00E+00

Table 4.7b Wilcoxon-Nemenyi-McDonald-Thompson Test Bayesian consensus trees (cherry count).

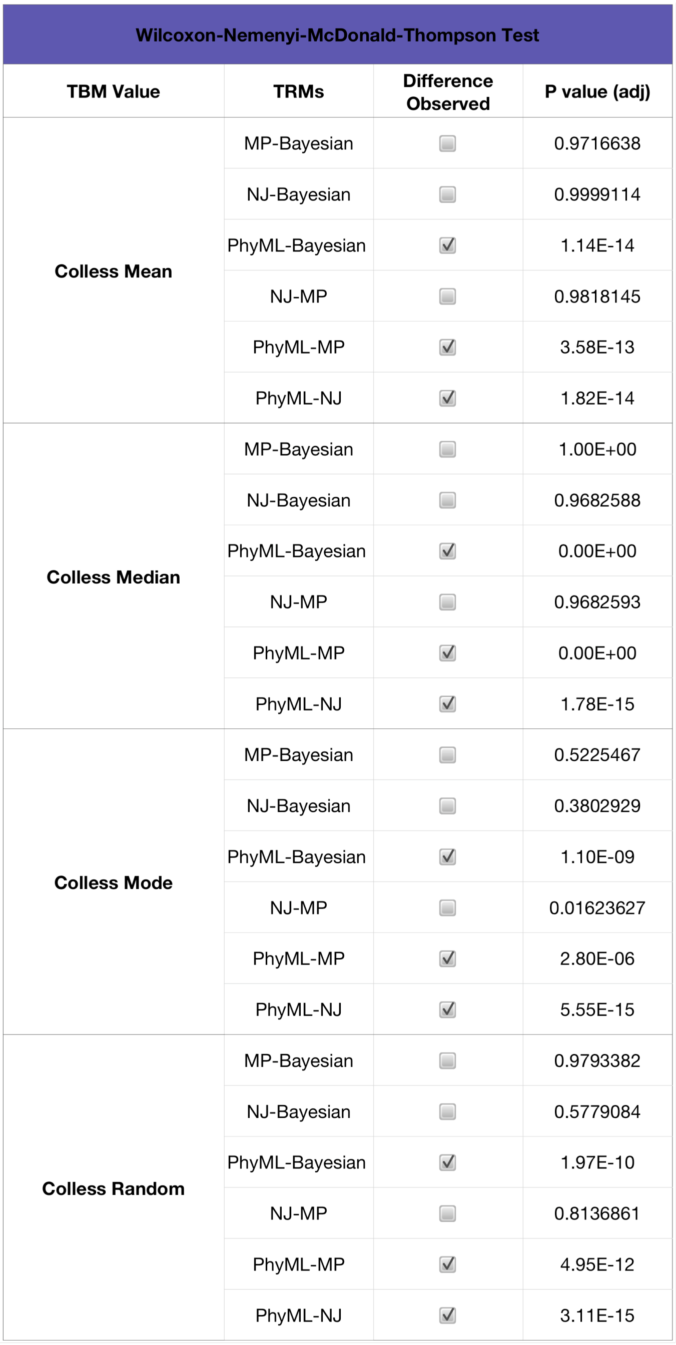
Results from the Wilcoxon-Nemenyi-McDonald-Thompson test performed on the mean, median, mode and random cherry count values, featuring the Bayesian consensus trees. Significant variance between the groups is measured at the 0.01 level.



Wilcoxon-Nemenyi-McDonald-Thompson Test			
TBM Value	**TRMs**	**Difference Observed**	**P value (adj)**
Cherry Mean	MP-Bayesian	☑	4.65E-05
	NJ-Bayesian	□	0.4756784
	PhyML-Bayesian	☑	0.0003888816
	NJ-MP	☑	0.01178351
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	1.41E-07
Cherry Median	MP-Bayesian	☑	9.64E-08
	NJ-Bayesian	□	0.4404055
	PhyML-Bayesian	☑	0.0004392304
	NJ-MP	☑	0.000149154
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	9.96E-08
Cherry Mode	MP-Bayesian	☑	1.57E-12
	NJ-Bayesian	□	0.4873619
	PhyML-Bayesian	☑	0.0002338751
	NJ-MP	☑	3.95E-08
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	1.05E-07
Cherry Random	MP-Bayesian	☑	0.005271518
	NJ-Bayesian	□	0.6251823
	PhyML-Bayesian	☑	0.0001125504
	NJ-MP	□	0.1516747
	PhyML-MP	☑	6.85E-14
	PhyML-NJ	☑	2.27E-07

Table 4.7c Wilcoxon-Nemenyi-McDonald-Thompson Test Bayesian sample trees (Colless Index).

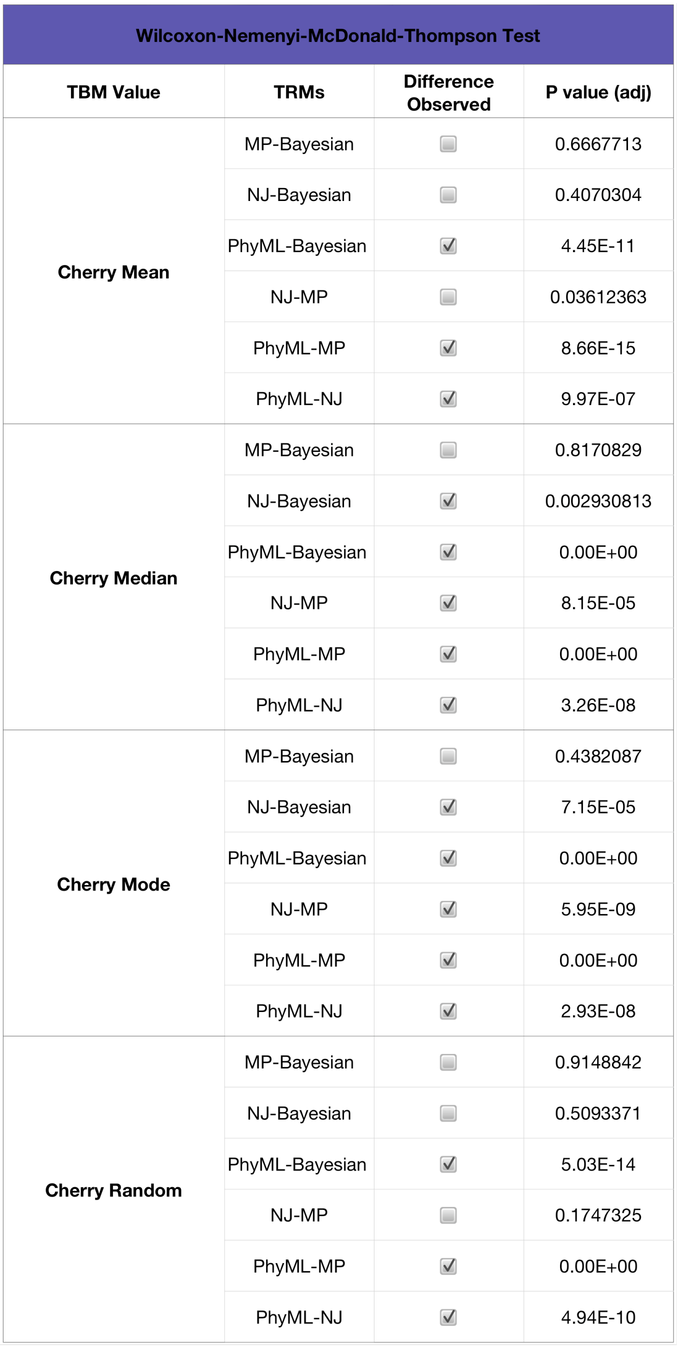
Results from the Wilcoxon-Nemenyi-McDonald-Thompson test performed on the mean, median, mode and random Colless index, featuring the Bayesian sample trees. Significant variance between the groups is measured at the 0.01 level.



Wilcoxon-Nemenyi-McDonald-Thompson Test			
TBM Value	**TRMs**	**Difference Observed**	**P value (adj)**
Colless Mean	MP-Bayesian	□	0.9716638
	NJ-Bayesian	□	0.9999114
	PhyML-Bayesian	☑	1.14E-14
	NJ-MP	□	0.9818145
	PhyML-MP	☑	3.58E-13
	PhyML-NJ	☑	1.82E-14
Colless Median	MP-Bayesian	□	1.00E+00
	NJ-Bayesian	□	0.9682588
	PhyML-Bayesian	☑	0.00E+00
	NJ-MP	□	0.9682593
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	1.78E-15
Colless Mode	MP-Bayesian	□	0.5225467
	NJ-Bayesian	□	0.3802929
	PhyML-Bayesian	☑	1.10E-09
	NJ-MP	□	0.01623627
	PhyML-MP	☑	2.80E-06
	PhyML-NJ	☑	5.55E-15
Colless Random	MP-Bayesian	□	0.9793382
	NJ-Bayesian	□	0.5779084
	PhyML-Bayesian	☑	1.97E-10
	NJ-MP	□	0.8136861
	PhyML-MP	☑	4.95E-12
	PhyML-NJ	☑	3.11E-15

Table 4.7d Wilcoxon-Nemenyi-McDonald-Thompson Test Bayesian sample trees (cherry count).

Results from the Wilcoxon-Nemenyi-McDonald-Thompson test performed on the mean, median, mode and random cherry count, featuring the Bayesian sample trees. Significant variance between the groups is measured at the 0.01 level.

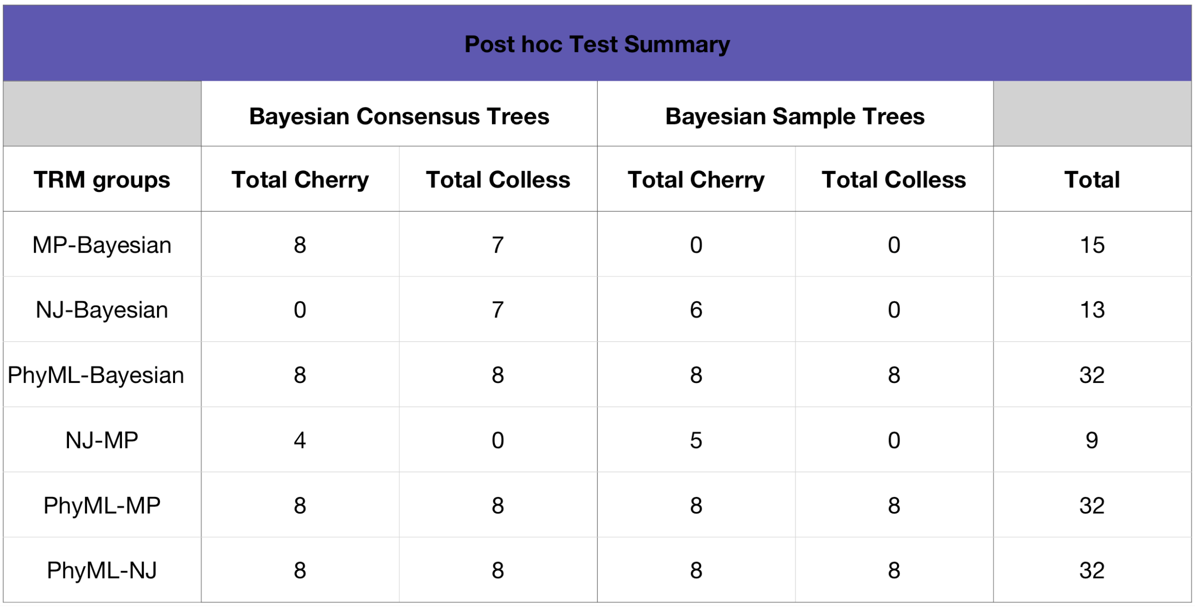


Wilcoxon-Nemenyi-McDonald-Thompson Test

TBM Value	TRMs	Difference Observed	P value (adj)
Cherry Mean	MP-Bayesian	☐	0.6667713
	NJ-Bayesian	☐	0.4070304
	PhyML-Bayesian	☑	4.45E-11
	NJ-MP	☐	0.03612363
	PhyML-MP	☑	8.66E-15
	PhyML-NJ	☑	9.97E-07
Cherry Median	MP-Bayesian	☐	0.8170829
	NJ-Bayesian	☑	0.002930813
	PhyML-Bayesian	☑	0.00E+00
	NJ-MP	☑	8.15E-05
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	3.26E-08
Cherry Mode	MP-Bayesian	☐	0.4382087
	NJ-Bayesian	☑	7.15E-05
	PhyML-Bayesian	☑	0.00E+00
	NJ-MP	☑	5.95E-09
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	2.93E-08
Cherry Random	MP-Bayesian	☐	0.9148842
	NJ-Bayesian	☐	0.5093371
	PhyML-Bayesian	☑	5.03E-14
	NJ-MP	☐	0.1747325
	PhyML-MP	☑	0.00E+00
	PhyML-NJ	☑	4.94E-10

Table 4.8 Post hoc test summary.

Combined results of the Tukey HSD and Wilcoxon-Nemenyi-McDonald-Thompson test tables above. The number of times each TRM group is reported as having significant variance in the aforementioned tables is counted here. This is categorised according to the type of Bayesian tree included in the data comparison, and sub-categorised according to the TBM used. The maximum count possible for each TRM group per sub-category is eight, while the total count possible for each TRM group is 32.



Post hoc Test Summary					
	Bayesian Consensus Trees		Bayesian Sample Trees		
TRM groups	**Total Cherry**	**Total Colless**	**Total Cherry**	**Total Colless**	**Total**
MP-Bayesian	8	7	0	0	15
NJ-Bayesian	0	7	6	0	13
PhyML-Bayesian	8	8	8	8	32
NJ-MP	4	0	5	0	9
PhyML-MP	8	8	8	8	32
PhyML-NJ	8	8	8	8	32

Both *post hoc* tests report a variance between the MP and Bayesian TRMs in just under half of the data comparisons, however, this is limited to those where the Bayesian consensus trees are used (see Table 4.6a & b). As such, it appears that this effect can be attributed to the use of a consensus tree approach in some of the Bayesian analyses. Inspection of the box plots shows that when the consensus tree values are used, the degree of imbalance measured by the TBMs is reduced, an effect that is particularly evident with respect to the cherry count (see Figure 4.6c and Figure 4.6g for a particularly stark example of this effect). This important observation suggests that there is a balancing influence that can be attributed to the use of consensus trees in Bayesian analyses, and may also be indicative of broader implications in, for example, parsimony analyses. When consensus trees are not used to summarise Bayesian trees, the MP and Bayesian methods are found to produce trees that are not significantly differently imbalanced.

Also of note from both *post hoc* tests were the data comparisons that detected variance between the NJ and MP methods. As can be seen in Table 4.8, the difference between these two TRMs is specific only to the cherry count TBM. While from the box plots there are no obvious common features between these two TRMs, the observed variance could be ascribed to the fact that the cherry count is a less sensitive measure of balance and, as such, may exacerbate differences between the tree shape of different TRMs in a similar manner to what is observed when using standard tree distance metrics (e.g. symmetric difference versus quartet distances).

Finally, there is a significant difference reported between the NJ and Bayesian methods in a total of 13 data comparisons. These 13 data comparisons span both classes of Bayesian trees (i.e. consensus or sample) used, and also both TBMs (See Table 4.8). It appears that using alternative types of Bayesian trees has a polarising effect on the two TBMs, as no variance in the Bayesian consensus data sets is detected by the cherry count, however, as measured by the Colless index, 7 out of 8 data sets report variance. Once the Bayesian sample trees are used the opposite is true for these TBMs. It can be suggested that this is again due to the use of the consensus approach in one of the Bayesian tree types, which causes the degree of balance in these trees to shift in the alternative data sets. The inverse relationship between the TBMs is accounted for by the fact that two metrics measure alternate aspects of tree shape and therefore should be deemed irrelevant.

From the box plots it can be seen that the maximum parsimony method is consistently the most imbalanced of the considered tree reconstruction methods, a trend that is observed from both TBMs (see Figure 4.6c for the most extreme instance of this effect). Here, as I wished to avoid the implication of the majority rule consensus approach for trees producing many MPTs, it is reasonable to say that this effect is purely driven by the tree reconstruction method. As this is observed from both a measure of tree balance and tree imbalance, it can be concluded that this does not stem from methodological bias.

However, it must be noted that even though MP is the most imbalanced method, the divergence of this method from other methods is not as apparent as one might expect, particularly in relation to the NJ and Bayesian TRMs. For example, MP and NJ are significantly different only for 9 out of 32 tests. However, when only the most sensitive TBM, the Colless index, is considered, no difference between these methods is detected. Similarly in the case of the Bayesian approach, if the data comparisons that use Bayesian consensus trees are disregarded, no significant difference is detected between this method of inference and MP. This suggests that, in truth, there is little difference between the imbalance of trees derived using MP, NJ and Bayesian analysis.

It is clear that ML produces trees that are more balanced then any other TRM, with a significant difference found between MP and ML in all data comparisons (32 out of 32). While this was, to some extent, expected, it was not foreseen that the same trend would be observed between ML and NJ, and particularly not between ML and the other probabilistic method, Bayesian inference. This raises the question of whether, in fact, ML has a bias in the opposite direction, and has an inclination to produce more balanced trees. This poses an interesting prospect, as there is no obvious reason why the use of the likelihood function should result in the selection of more balanced trees. Further to this, the Bayesian method of inference, which also uses the likelihood function, fails to produce trees as imbalanced as ML.

An express difference between the ML and Bayesian methods is that branch lengths are optimised in the ML approach. Accordingly, it is possible that this optimisation process might ultimately lead to the selection of more balanced trees. However, conversely the Bayesian method of inference may in itself be problematic, finding trees with a similar degree of balance to the MP and NJ methods. Such issues, relating to potential biases in either ML or Bayesian analyses, remain as open problems and need to be addressed further.

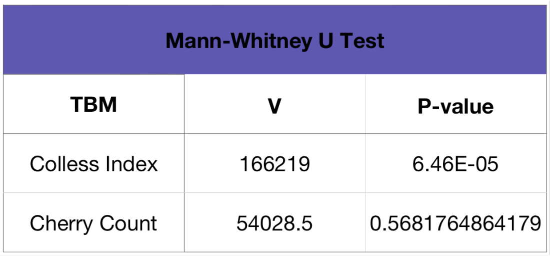
4.3.2 Variance in maximum likelihood

For the two TBMs, the *u* test produced slightly varying results. Comparison of the ML software implementations under the Colless index determined that there was a significant difference in terms of tree shape between these two methods (see Table 4.9). From the box plot of these values (Figure 4.7a) it can be seen that, although the median and spread of these two methods are quite disparate, the RAxML Colless values are notably more balanced than the PhyML trees derived from the same data. As the search strategy of RAxML is believed to carry out a more thorough search, this result supports the conjecture that it is the lack of optimisation of branch lengths in the Bayesian method (see above) that causes its resultant trees to be more imbalanced than those obtained using ML. If this is found to be true, then the observed similarity between the Bayesian and MP methods (see above) might indicate that ML methods should be preferred to Bayesian ones. Notably, if PhyML was replaced by RAxML, as the ML representative in the data comparisons the same overall trend would continue to be observed; however, ML would emerge as even more balanced.

For the cherry count values no significant difference was observed between the two software implementations (see Table 4.9), which is confirmed by the box plots of these values where the medians are seen to be equivalent (see Figure 4.7b). This suggests that when there is considerable similarity between the compared values, the less sensitive cherry count is unable to identify refined differences. Results of this analysis illustrate that the influence of software implementation bias can be ruled out as an explanation for the trends observed in the above comparisons of TRMs, and, as such, the observed balance of the ML method can be considered genuine.

Table 4.9 Mann-Whitney *u* Test

Results of the Mann-Whitney *u* test comparing both the Colless index and cherry count measures of two alternative ML software implementations, PhyML and RAxML. V is equal to the sum of ranks. P-values < 0.01 are deemed significant.



Mann-Whitney U Test

TBM	V	P-value
Colless Index	166219	6.46E-05
Cherry Count	54028.5	0.5681764864179

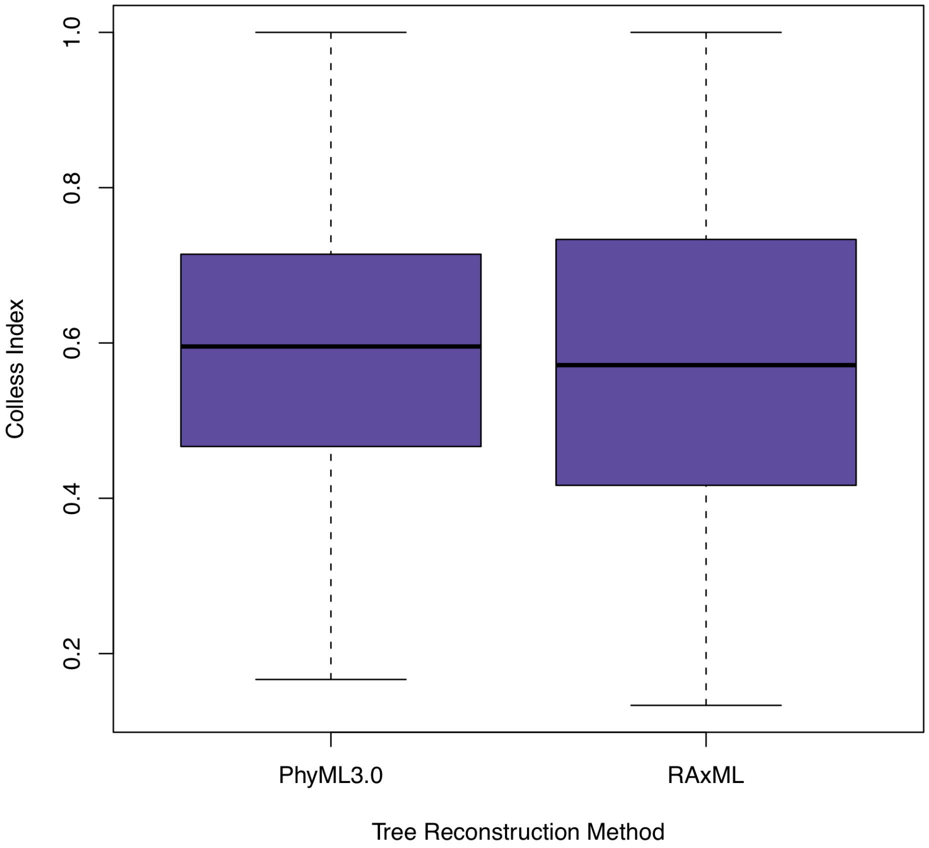


Figure 4.7a Box plot of Colless index values of trees produced by PhyML and RAxML.

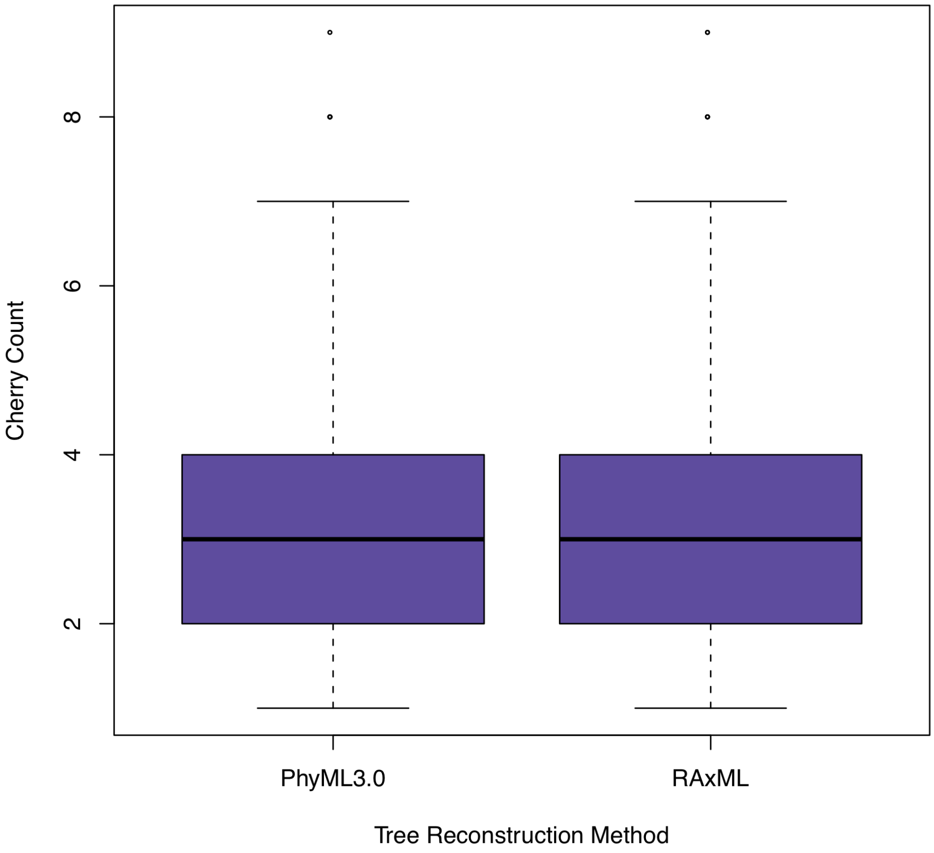


Figure 4.7b Box plot of cherry count values of trees produced by PhyML and RAxML.

4.3.3 Variance in Bayesian inference under varying amino acid models

From the ANOVAs conducted on the Bayesian trees derived under different models, it was observed that, on average, each model produced trees with a consistent degree of balance (as measured by the cherry count) and imbalance (as measured by the Colless index). This is confirmed by examining the box plots of these data comparisons (see Figure 4.8a-e and Figure 4.9a-e). Only one of the ANOVAs using the cherry count TBM, namely the mean values comparison, reported a significant difference between trees derived using the considered models with a p-value of 0.01 (corrected for violations of sphericity; see Table 4.10). However, according to the Boferroni correction of significance for these comparisons, 0.001, the null hypothesis can no longer be rejected.

Of the ANOVAs conducted using the Colless index values, a significant difference between the alternative amino acid models is observed only when the Bayesian consensus trees were used (see Table 4.10). Therefore, Tukey's HSD test was then implemented to determine between which models the variation was occurring. Results of this test reveal that trees built under the JTT and CAT models are divergent from each other, while trees built under the CAT and LG models were similarly deviating (see Table 4.11). Inspection of the box plot for this data set shows that the CAT tree values are significantly more balanced than the LG or JTT model trees (see Figure 4.8e). This result is surprising given that all other data comparisons of this kind report no substantial variation between models (see Table 4.10).

The CAT model is frequently the model of choice for Bayesian analyses, as it boasts the ability to incorporate site-specific amino acid motifs in its modelling strategy. This model has been shown to be particularly adept at overcoming systematic errors such as long branch attraction (Lartillot and Philippe, 2004), therefore, it is fitting that, of the

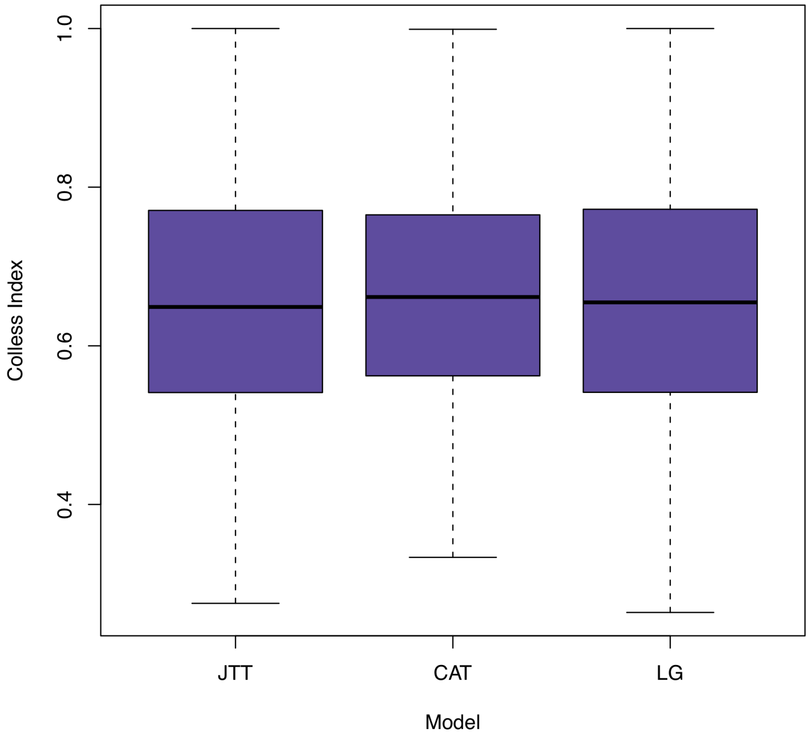


Figure 4.8a Box plot of mean Colless index values from Bayesian sample trees.

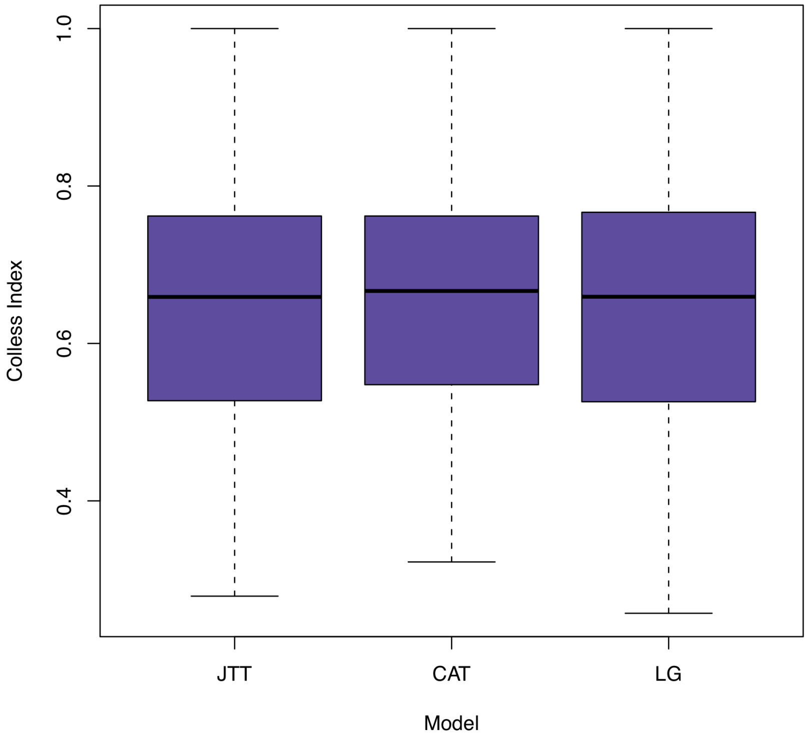


Figure 4.8b Box plot of median Colless index values from Bayesian samples trees.

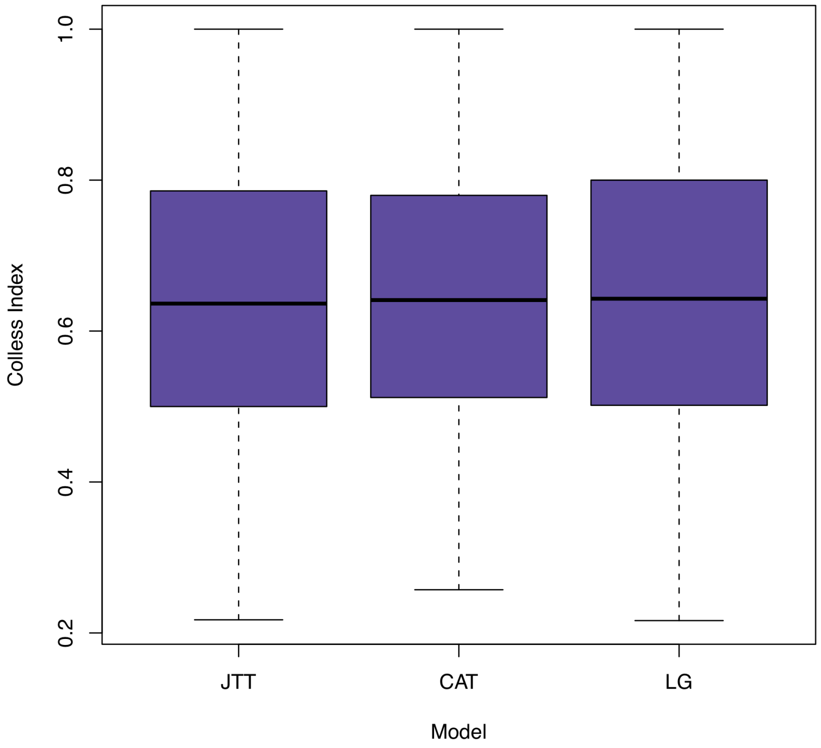


Figure 4.8c Box plot of mode Colless index values from Bayesian samples trees.

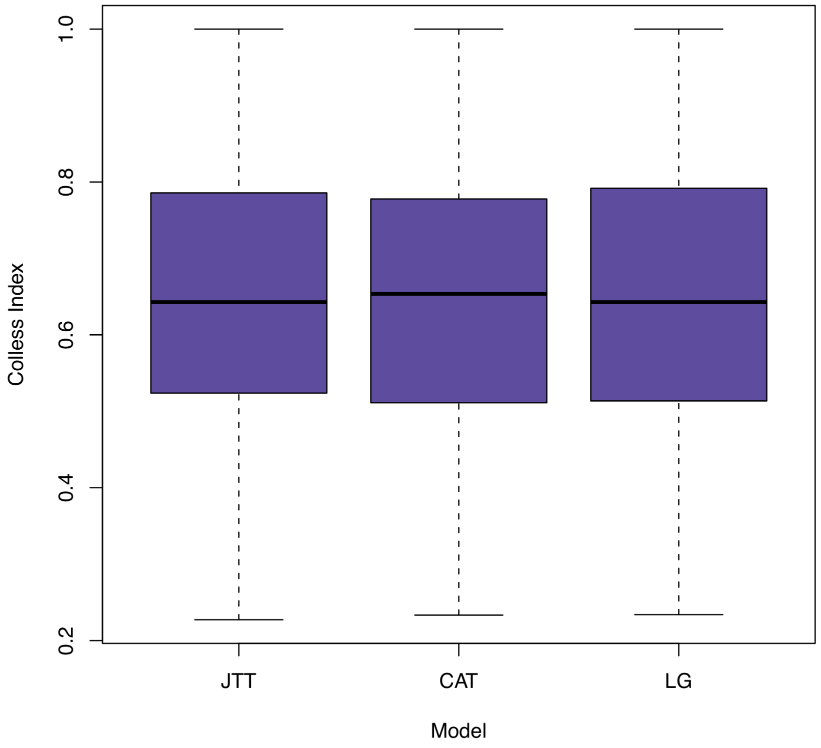


Figure 4.8d Box plot of random Colless index values from Bayesian samples trees.

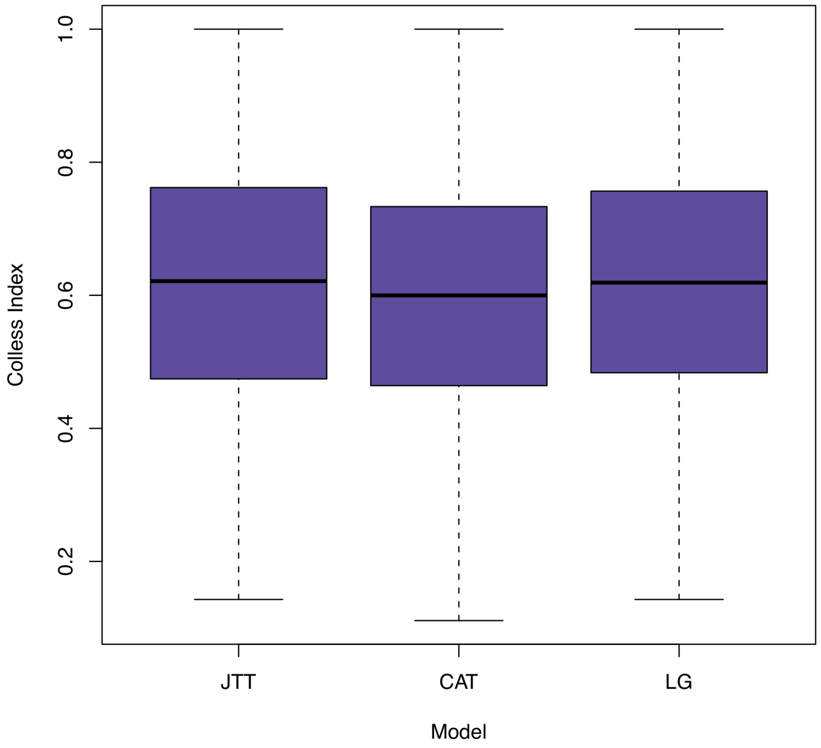


Figure 4.8e Box plot of Colless index values from Bayesian consensus trees.

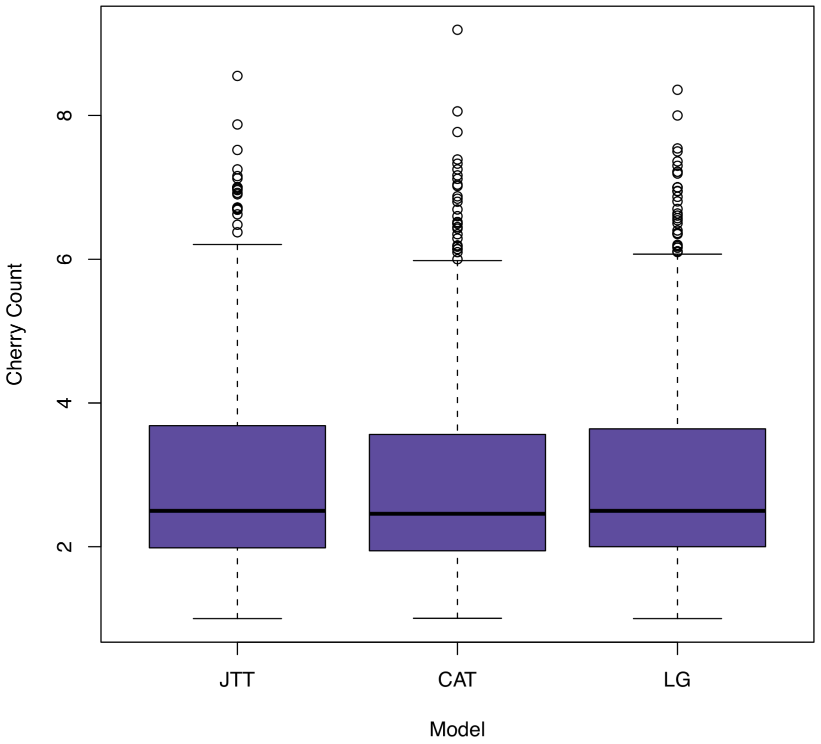


Figure 4.9a Box plot of mean cherry count values from Bayesian sample trees.

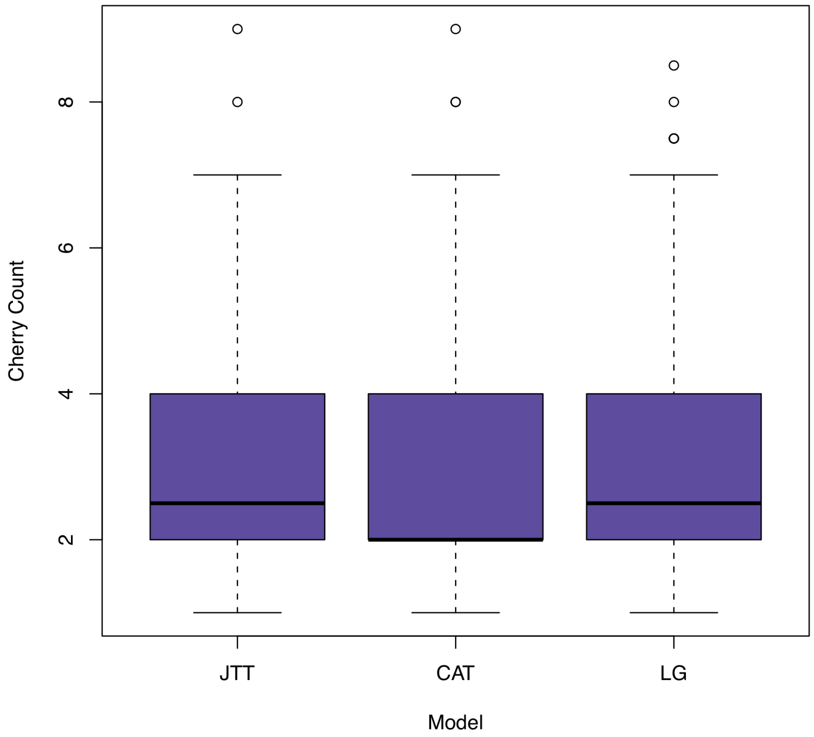


Figure 4.9b Box plot of median cherry count values from Bayesian sample trees.

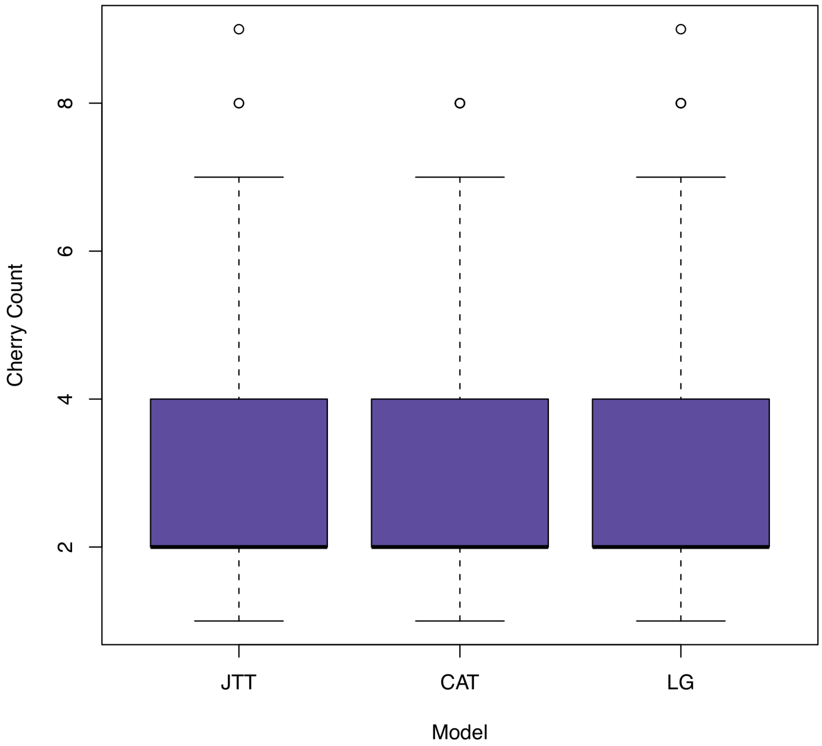


Figure 4.9c Box plot of mode cherry count values from Bayesian sample trees.

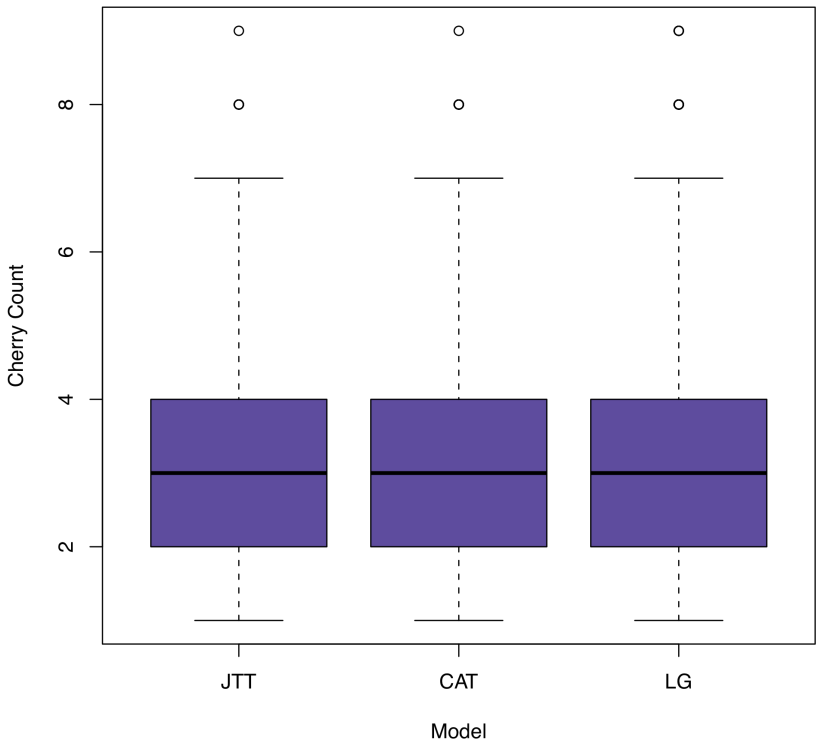


Figure 4.9d Box plot of random cherry count values from Bayesian sample trees.

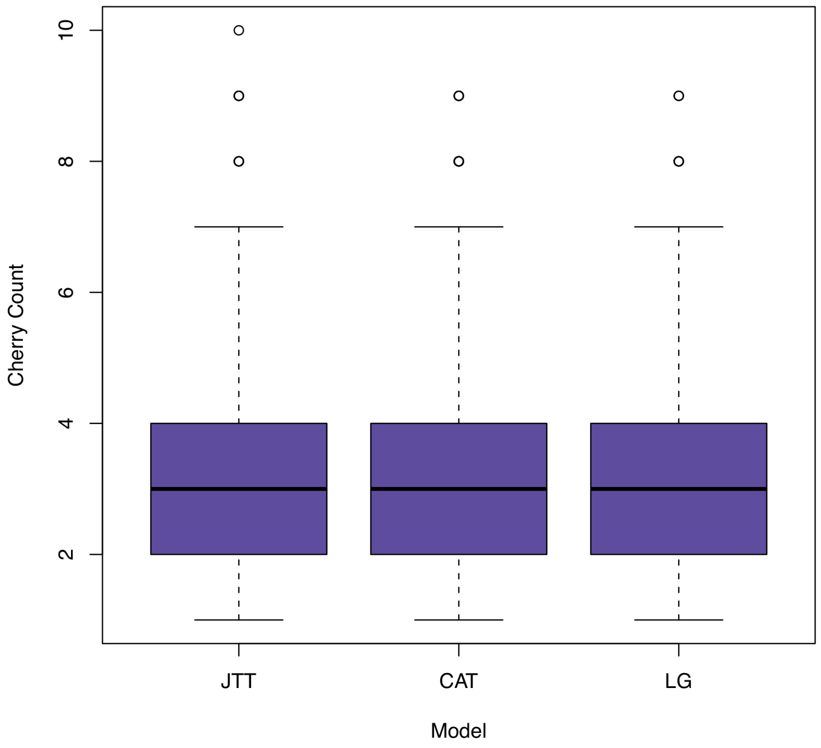
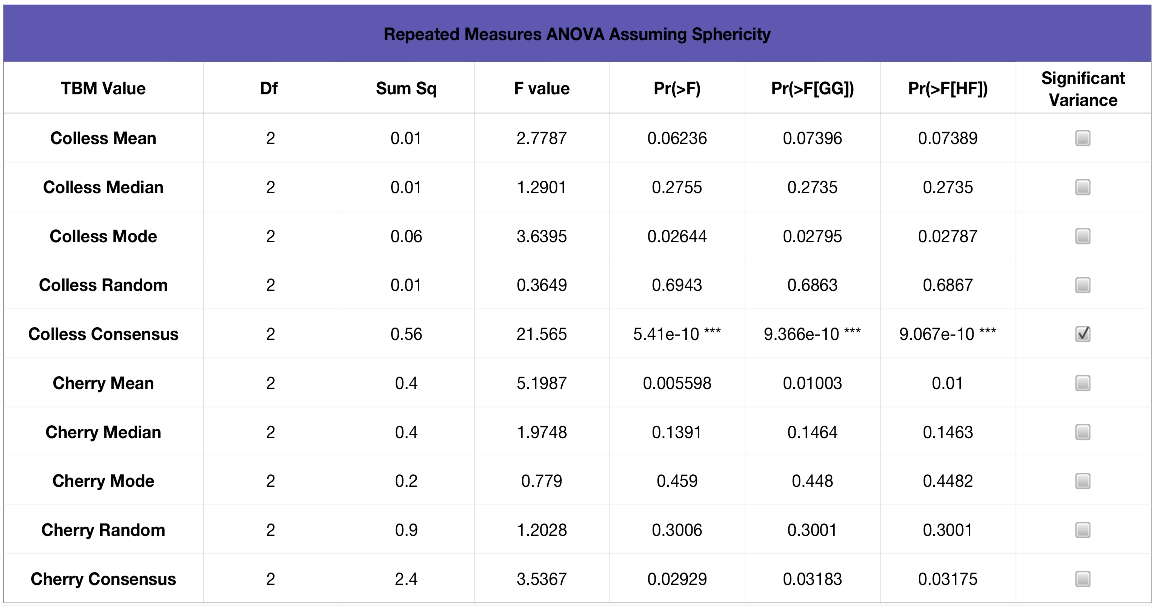


Figure 4.9e Box plot of random cherry count values from Bayesian consensus trees.

Table 4.10 ANOVA Bayesian trees under alternative models

Results from the 10 ANOVAs conducted using the average statistic, random Bayesian sample, and Bayesian consensus values. The F value is the statistic measure in the ANOVA. Pr(>F) is the p-value, while Pr(>F[GG]) and Pr(>F[HF]) are the Greenhouse-Geisser and Huynh-Feldt corrected p-values respectively(*** $p < 0.001$).

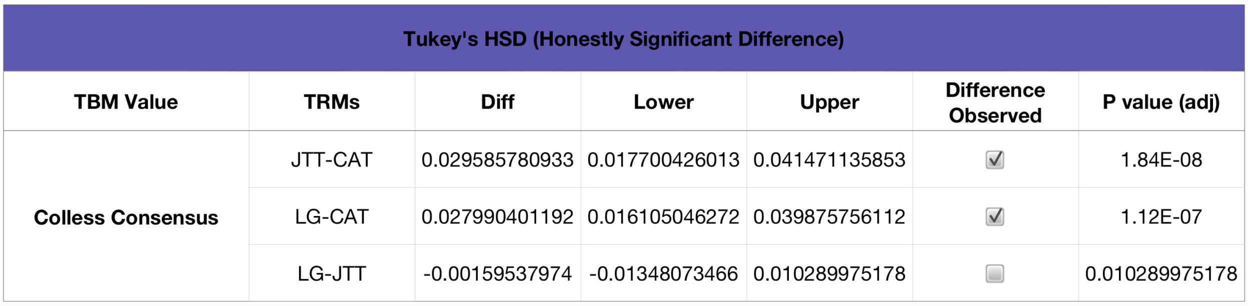


Repeated Measures ANOVA Assuming Sphericity

TBM Value	Df	Sum Sq	F value	Pr(>F)	Pr(>F[GG])	Pr(>F[HF])	Significant Variance
Colless Mean	2	0.01	2.7787	0.06236	0.07396	0.07389	□
Colless Median	2	0.01	1.2901	0.2755	0.2735	0.2735	□
Colless Mode	2	0.06	3.6395	0.02644	0.02795	0.02787	□
Colless Random	2	0.01	0.3649	0.6943	0.6863	0.6867	□
Colless Consensus	2	0.56	21.565	5.41e-10 ***	9.366e-10 ***	9.067e-10 ***	☑
Cherry Mean	2	0.4	5.1987	0.005598	0.01003	0.01	□
Cherry Median	2	0.4	1.9748	0.1391	0.1464	0.1463	□
Cherry Mode	2	0.2	0.779	0.459	0.448	0.4482	□
Cherry Random	2	0.9	1.2028	0.3006	0.3001	0.3001	□
Cherry Consensus	2	2.4	3.5367	0.02929	0.03183	0.03175	□

Table 4.11 Tukey HSD Test.

Results from Tukey's Honestly Significant Difference test performed on the Bayesian consensus Colless index values. Significant variance between the groups is measured at the 0.01 level. As the Bayesian consensus trees were the only data comparison to display variance in the ANOVA, the Tukey HSD test was restricted to this group.



Tukey's HSD (Honestly Significant Difference)						
TBM Value	**TRMs**	**Diff**	**Lower**	**Upper**	**Difference Observed**	**P value (adj)**
Colless Consensus	JTT-CAT	0.029585780933	0.017700426013	0.041471135853	☑	1.84E-08
	LG-CAT	0.027990401192	0.016105046272	0.039875756112	☑	1.12E-07
	LG-JTT	-0.00159537974	-0.01348073466	0.010289975178	□	0.010289975178

models selected for this analysis, CAT produces the most balanced consensus trees (as measured by the Colless index). However, as this result is not replicated in the cherry count analysis of the same data, or in analyses where consensus trees are not used, it can be assumed that this result is unlikely to reflect an important difference. Results of the non-parametric Friedman tests converged upon the results of all ANOVAs discussed above. Therefore, the incorporation of statistical errors associated with the violation of the assumptions of ANOVA can be ruled out. Results of the Friedman test can be seen in Table 4.12, while the Wilcoxon-Nemenyi-McDonald-Thompson test conducted on the Colless index values for the Bayesian consensus trees can be seen in Table 4.13.

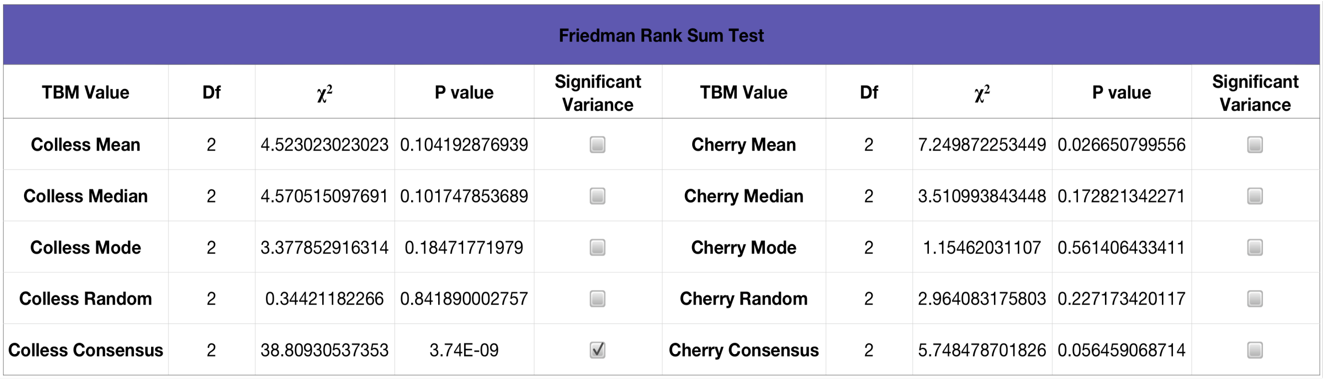
4.4 Conclusions

This analysis shows that the most frequently used tree reconstruction methods in modern phylogenetics produce trees with varying degrees of balance. ML emerges clearly as the most balanced tree reconstruction method, while MP manifests to the contrary. This result directly contradicts the study of Huelsenbeck and Kirkpatrick (1996) where ML is found to be the most imbalanced TRM. However, the trees of Huelsenbeck and Kirkpatrick (1996) were produced using nucleotide data and under the simplistic Jukes and Cantor (1969) model, two aspects of their study that vary substantially from the one discussed here.

It is possible that different data types may generate trees of different shape, but this hypothesis was not tested here. However, it seems that improving the search strategy of the ML method (from PhyML to RaxML) resulted in increased symmetry of the recovered trees. Therefore, considering that algorithms for effective searches of the tree

Table 4.12 Friedman Test Bayesian trees under alternative models

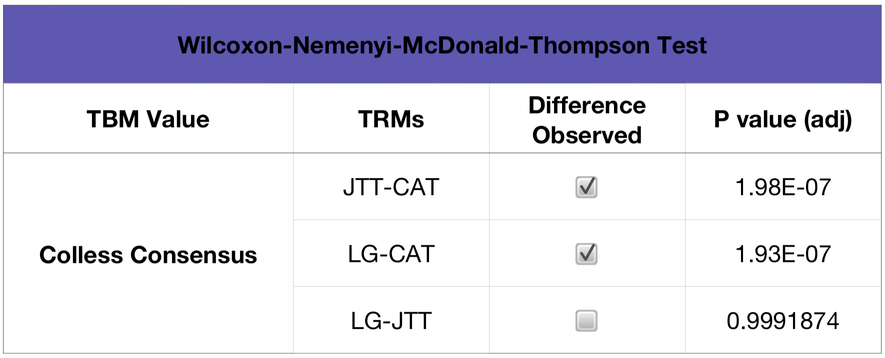
Results from the 10 Friedman tests conducted using the average statistic, random and Bayesian consensus values.



Friedman Rank Sum Test									
TBM Value	Df	χ^2	P value	Significant Variance	TBM Value	Df	χ^2	P value	Significant Variance
Colless Mean	2	4.523023023023	0.104192876939	□	Cherry Mean	2	7.249872253449	0.026650799556	□
Colless Median	2	4.570515097691	0.101747853689	□	Cherry Median	2	3.510993843448	0.172821342271	□
Colless Mode	2	3.377852916314	0.18471771979	□	Cherry Mode	2	1.15462031107	0.561406433411	□
Colless Random	2	0.34421182266	0.841890002757	□	Cherry Random	2	2.964083175803	0.227173420117	□
Colless Consensus	2	38.80930537353	3.74E-09	☑	Cherry Consensus	2	5.748478701826	0.056459068714	□

Table 4.13 Wilcoxon-Nemenyi-McDonald-Thompson Test.

Results from the Wilcoxon-Nemenyi-McDonald-Thompson test performed on the Colless index of the Bayesian consensus trees. Significant variance between the groups is measured at the 0.01 level. As the Bayesian consensus trees were the only data comparison to display variance in the Friedman test, the Wilcoxon-Nemenyi-McDonald-Thompson test was restricted to this group.



Wilcoxon-Nemenyi-McDonald-Thompson Test			
TBM Value	TRMs	Difference Observed	P value (adj)
Colless Consensus	JTT-CAT	☑	1.98E-07
	LG-CAT	☑	1.93E-07
	LG-JTT	□	0.9991874

space under ML have improved since those used by Huelsenbeck and Kirkpatrick (1996), it can be suggested that their results may have been influenced by the use of poorly performing tree search strategies. This remains as an open question, but can potentially be addressed by performing a comparison between trees obtained using RaxML or PhyML against those obtained by an older ML algorithm (e.g. those implemented in Phylip).

The results of this study show that the choice of model does not seem to affect tree shape. While this is found to be the case specifically for Bayesian inference, this is likely to be the case for other phylogenetic methods as well. As such, it is reasonable to rule out model selection as a potential explanation for difference between the results observed here and those of Huelsenbeck and Kirkpatrick (1996). Further to this, as two modern ML software programs display the same over all trend, that ML has propensity to produce balanced trees, it can be assumed that what is observed in this study is an accurate reflection of a general tendency of this tree reconstruction method.

The findings of this study are also inconsistent with those of Heard (1992), where no variation between tree reconstruction methods was detected. In his study, Heard sampled trees from previous publications and examined the degree of balance of these trees in relation to one another. Little attention was paid to the type of tree reconstruction method used to produce these trees as they were merely classified as being phenetic or cladistic in origin. Grouping tree reconstruction methods in this way is likely to have had a confounding effect. Neighbor joining can be considered a phenetic method, and is seen in this current study to have levels of imbalance similar to those of MP (the cladistic method). It is true that other phenetic methods have the tendency to return more balanced trees (e.g. UPGMA; see Huelsenbeck and Kirkpatrick, 1996), but merging results from, say, NJ and UPGMA must have an obscuring effect, potentially incurring noise in the analysis.

Further to this, Heard (1992) did not apply a strict lower limit on the number of taxa per tree to be included in his study and, therefore, may have introduced random effects associated with the Colless index for trees with less than 7 taxa (see Rogers, 1994, Rogers, 1996, Harcourt-Brown et al., 2001). Additionally, in Heard's (1992) study, only 208 trees in total were sampled. This is a relative small number; particularly considering that this was further split into cladograms and phenograms, and may be responsible for the introduction of further random errors. A final point of note with regard to the study of Heard (1992) is that, because of the timing of his study, it seems unlikely that his sample included likelihood trees.

Here, the proposal of Colless (1982), that parsimony produces more imbalanced phylogenetic trees, is confirmed. This is in line with the classic study of Colless (1995), who originally provided experimental evidence for this contention. While variation between TRMs impacts significantly on phylogenetics, the gravity of this is somewhat limited in the case of molecular studies as it is widely accepted that ML and Bayesian inference are the most robust methods. However, as pointed out above, Bayesian trees seem quite similar to the parsimony ones, and quite different to the ML ones, which implies that these results might well be relevant also to molecular phylogenetics. Although these methods return trees that are quite different in terms of balance, it is likely, that irrespective of tree shape biases, the overall congruence of ML and Bayesian is a good proxy for phylogenetic accuracy.

The results of this study have, however, a much more serious implication for morphology-based analyses, as the use of parsimony may be contributing bias to the observed imbalance of trees (see Harcourt-Brown et al., 2001), particularly as they are rarely compared to ML trees (that should be more balanced) derived from the same data. It has often been said that molecular trees are more balanced than morphological trees and one would wonder to what extent this is the product of phylogenetic methodology, rather than the biological truth or problems with coding of morphological characters? The answer to this question, again, remains open.

Finally, the results presented here may also impact upon the field of phylogenomics. The study of Wilkinson et al. (2005) suggests that the frequently used matrix representation with parsimony (MRP) supertree method is influenced by a tree shape bias. Trees produced by this method are shown to be, in general, more imbalanced than other supertree methods, an effect which the authors attribute to the way in which this method resolves conflict between input trees. Here, I expand on the results of Wilkinson et al. (2005) and conclude that it is not MRP *per se* (or the Baum coding scheme) that is problematic, rather it might be maximum parsimony. Wilkinson (personal communication) suggests that if parsimony is an asymmetric distance measure, then one would expect that its effect would be visible not only in supertree reconstruction but also in the analyses of standard molecular data.

Here, this is shown to be the case and, thus it can be concluded that if parsimony produces more imbalanced trees in a general sense, it follows that the parsimony component of the MRP supertree method is responsible for the observed imbalance of the supertrees produced by this method. This suggests that where possible, parsimony should be avoided as a method of phylogenetic reconstruction, particularly when the trees (or supertrees) are then used to test macroevolutionary hypotheses (e.g. Purvis et al., 1995, Bininda-Emonds et al., 1999, Hone et al., 2005, Ruta et al., 2007, Lloyd et al., 2008, where the balance of a tree is used to identify adaptive radiations). Using MP in this context may lead to the detection of false positives, with radiation being observed in groups that did not radiate.

Chapter 5: General Discussion

Today, the potential wealth of genomic data obtainable has far exceeded even the enthusiastic expectations that were typical of the scientific community early last decade (see for example Gee, 2003). This can be attributed to the momentous advancements in sequencing technology and, in particular, the development of next generation sequencing techniques. Indeed, all previously held expectations are set to be readjusted further with the development of affordable technologies such as the Ion Torrent system (Drmanac et al., 2010). While this all spells good news for the field of phylogenomics, there are major issues associated with this rapid data accumulation, some of which are encountered through out this thesis.

In Chapter 2, a study of the animal phylogeny was conducted featuring 43 species. While the taxonomic sampling was certainly adequate to recover a robust phylogeny, supported by both a supertree and supermatrix approach, it highlights a worrying trend. The animals are a group that attain particular interest in complete genome sequencing projects (see Sanderson et al., 2010), however a sampling of 43 unique species is certainly a poor representation of this group. Further to this, a domain as broad and diverse as the eukaryotes currently achieves a sample of only 156 genomes. If this is the situation in a well-represented group, the gravity of under representation of less focal groups can be deduced.

In this thesis, as in many studies (e.g. Philippe et al., 2005, Burki et al., 2007, Dunn et al., 2008, Hejnol et al., 2009), the use of ESTs offers a means of increasing taxon sampling. The approach used in Chapter 3, where the taxonomic sampling of complete genomes is supplemented with a broader EST sample, is promising but it can only be considered a stopgap solution. ESTs represent a shallow genomic sampling (Zilversmit et al., 2002, see also Chapter 2), often featuring the most highly expressed genes of the species of interest. Additionally, there is a substantial amount of missing data associated with the use of ESTs, which has recently been shown to have a greater bearing on analyses than was previously thought (see Sanderson et al., 2010). Therefore, while providing a useful tool, ESTs cannot be considered anything more than an “improvised” phylogenomic resource. Despite these limitations, the use of ESTs to build phylogenies is generally referred to as phylogenomics, regardless of its original inception as shallow genomic studies (see, for example, Hughes et al., 2006). While of late there appears to be a more conscientious effort to use a combined EST and complete genome approach (e.g. Hejnol et al., 2009), it is expected that strictly EST-based “phylogenomics” will persist for the foreseeable future.

Although many have demonstrated the utility of complete genomic data (e.g. Creevey et al., 2004, Philip et al., 2005, Fitzpatrick et al., 2006, Pisani et al., 2007), until recently it was not feasible to incorporate genes with a history of duplication. In this thesis, in a movement towards a truly phylogenomic approach, genes of this nature were integrated into a phylogenomic analysis. In the context of the animal phylogeny, which was the subject of Chapter 2, including genes that have undergone duplication is seen to have little impact on the topology. However, this represents a rather contrived situation, where only four taxa were included, as here my intention was merely to investigate the feasibility of the GTP approach.

Chapter 3 sees a more practicable implementation of GTP, where approximately 20,000 duplicated genes are supplemented to those typically used in the supertree approach. In considering single protein families alone, the gene coverage for over 550 species amounts to a meagre 7,214 genes, a figure that clearly points to the significant role of duplication in the evolution of the eukaryotes. The resolution of the eukaryote phylogeny notably improved upon inclusion of multi protein families, which again is indicative of the extent of duplication in this domain, but in a more broad sense demonstrates the importance of a maximal gene sampling. While the specific means of resolving duplications, GTP, is not beyond reproach, it does present a valuable starting point, which I feel should be adopted more frequently in phylogenomic analyses.

A considerable problem in light of recent data accumulation, which will undoubtedly be exasperated with the future amassing of genomic data, is the limited approaches available for analysing genomic scale data. As mentioned recurrently through this thesis, the supermatrix approach does offer an appropriate solution, in theory, due to its total evidence properties and its associated Bayesian-based inference. However, in practice this approach struggles with moderately large data sets, as seen in Chapter 2, and subsequently cannot even be considered a viable option for very large data sets such as that of Chapter 3. Indeed the study of Hejnol et al. (2009), which is the largest supermatrix data set assembled to date, features only 94 taxa and 1487 genes and necessitated the use of extremely high performing supercomputing resources, exemplifying clearly the current limitations of this approach.

Accordingly, supertrees currently provide the only practical alternative for the phylogenetic analysis of large genomic data sets. In this thesis the limits of the supertree approach were tested in several ways. This was done, firstly, with regard to the volume of data this approach could contend with. The three supertree analyses discussed in Chapter 2 represent the largest sampling of genes considered in a published phylogeny of the Metazoa to date, while the study in Chapter 3 sees a ten-fold increase in this sampling, representing, as far as I am aware, the largest phylogenetic analysis that has ever been attempted. When compared with the current capabilities of the supermatrix approach, supertrees offer the potential to include a far greater genomic depth and breath, something which Sanderson (2008) asserts will be required for the recovery of the tree of life.

Two experimental approaches to supertree reconstruction were considered in this thesis. Firstly, in Chapter 2 and 3, the aforementioned duplicated genes were incorporated into a supertree approach, which currently is the only means of integrating genes of this nature into a phylogenomic approach. Secondly, the use of a combined EST and complete genome data set was tested in the supertree context. The relative success, coupled with the ease of integration, of both approaches attests to the flexibility of supertree methods in light of atypical data.

While overall in this thesis the application of supertree methods is met with success, it must be noted that the limited MRP supertree method is widely used. In Chapter 4 it is demonstrated that the shape biases associated with this method are likely to stem from the behaviour of the parsimony component of this approach, rather than some feature of the supertree process (for example the Baum and Ragan coding scheme). While this is an important distinction, it does not detract from the fundamental bias of this method. Although the use of other supertree methods is tested in Chapter 2, the resolution of the resulting phylogenies is poor. As such, in the absence of a more applicable method, MRP persists as an expedient method. However, I anticipate that the advent of a software implementation of ML supertrees is likely to account for the inadequacies of current methods such as MRP.

It is shown in this thesis that the effective balance of a phylogenetic tree is dependent on the tree reconstruction method used. This finding has a two-fold relevance for supertree methods in general. Firstly, trees derived by these methods are used in supertree analyses as input trees, therefore, it is likely that an inherent shape bias in the input trees may manifest in the global supertree output. Further to this, as seen in the case of MRP, some supertree methods have a phylogenetic component, therefore, biases in this component will be imposed upon the supertree approach as a whole. The use of methods that produce more balanced trees, for example ML, in the absence of evidence of an intrinsic bias associated with these methods, should be adopted as balanced trees better adhere to the Yule (1924) model.

Although there appears to be somewhat of a gap between the accumulation of genomic data and the methods used to analyse it, this thesis is testimony to an obvious progression in the field of phylogenomics. In 2006, Ciccarelli et al. published a tree of life based on only 31 proteins, motivating Dagan and Martin (2006) to refer to it is a tree of one percent. Here, due to advancements in computational tools, I am able to present a data set that features almost 1000 times more proteins. A progression of this extent, in less than five years, is undoubtedly promising for the future of phylogenomics.

Chapter 6: Bibliography

ADOUTTE, A., BALAVOINE, G., LARTILLOT, N., LESPINET, O., PRUD'HOMME, B. & DE ROSA, R. (2000) The new animal phylogeny: reliability and implications. *Proceedings of the National Academy of Sciences USA,* 97**,** 4453-4456.

AGAPOW, P. & PURVIS, A. (2002) Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology,* 51**,** 866-872.

AGUINALDO, A., TURBEVILLE, J., LINFORD, L., RIVERA, M., GAREY, J., RAFF, R. & LAKE, J. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature,* 387**,** 489-493.

AHO, A., SAGIV, Y., SZYMANSKI, T. G. & ULLMAN, J. D. (1981) Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing,* 10**,** 405-421.

AKAIKE, H. (1973) Information theory and an extension of the maximum likelihood principle. IN PETROV, B. N. & CSAKI, F. (Eds.) *Proceedings 2nd International Symposium on Information Theory.* Budapest, Akademia Kiado.

ALDOUS, D. (1991) The continuum random tree II: An overview. *Stochastic analysis***,** 23-70.

ALTSCHUL, S., MADDEN, T., SCHAFFER, A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25**,** 3389-3402.

ARCHIE, J. (1989) Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Systematic Biology,* 38**,** 253-269.

ARISUE, N., HASEGAWA, M. & HASHIMOTO, T. (2005) Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. *Molecular Biology and Evolution,* 22**,** 409-420.

ARVESTAD, L., BERGLUND, A., LAGERGREN, J. & SENNBLAD, B. (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics,* 19**,** i7-15.

BACKELJAU, T., DE BRUYN, L., DE WOLF, H., JORDAENS, K., VAN DONGEN, S. & WINNEPENNINCKX, B. (1996) Multiple UPGMA and neighbor-joining trees and the performance of some computer packages. *Molecular Biology and Evolution,* 13**,** 309-313.

BALDAUF, S. (2003) The deep roots of eukaryotes. *Science,* 300**,** 1703-1706.

BALDAUF, S. (2008) An overview of the phylogeny and diversity of eukaryotes. *Journal of Systematics and Evolution,* 46**,** 263-273.

BAUM, B. (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon,* 41**,** 3-10.

BAUM, B. & RAGAN, M. (2004) The MRP Method IN BININDA-EMONDS, O. R. P. (Ed.) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* Dordrecht, Kluwer Academic.

BAURAIN, D., BRINKMANN, H. & PHILIPPE, H. (2007) Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors? *Molecular Biology and Evolution,* 24**,** 6-9.

BERGSTEN, J. (2005) A review of long-branch attraction. *Cladistics,* 21**,** 163-193.

BERGSTEN, J. & MILLER, K. (2006) Taxonomic revision of the Holarctic diving beetle genus Acilius Leach (Coleoptera: Dytiscidae). *Systematic Entomology,* 31**,** 145-197.

BININDA-EMONDS, O., GITTLEMAN, J. & PURVIS, A. (1999) Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews,* 74**,** 143-175.

BININDA-EMONDS, O. R. P. (2004a) The evolution of supertrees. *Trends in Ecology & Evolution,* 19**,** 315-322.

BININDA-EMONDS, O. R. P. (2004b) *Phylogenetic Supertrees: Combining Information To Reveal The Tree Of Life,* Dorderecht, Kluwer Academic.

BLAIR, J., IKEO, K., GOJOBORI, T. & HEDGES, S. (2002) The evolutionary position of nematodes. *BMC Evolutionary Biology,* 2**,** 7.

BLUM, M. & FRANCOIS, O. (2006) Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology,* 55**,** 685-691.

BLUM, M. & FRANÇOIS, O. (2005) On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. *Mathematical Biosciences,* 195**,** 141-153.

BLUM, M., FRANÇOIS, O. & JANSON, S. (2006) The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability,* 16**,** 2195-2214.

BORTOLUSSI, N., DURAND, E., BLUM, M. & FRANCOIS, O. (2006) apTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics,* 22**,** 363-364.

BOX, G. & ANDERSEN, S. (1955) Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society. Series B (Methodological),* 17**,** 1-34.

BRADLEY, R., ROBERTS, A., SMOOT, M., JUVEKAR, S., DO, J., DEWEY, C., HOLMES, I. & PACHTER, L. (2009) Fast statistical alignment. *PLoS Computational Biology,* 5, e1000392.

BRINKMANN, H. & PHILIPPE, H. (1999) Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Molecular Biology and Evolution,* 16**,** 817-825.

BRINKMANN, H., VAN DER GIEZEN, M., ZHOU, Y., DE RAUCOURT, G. & PHILIPPE, H. (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic Biology,* 54**,** 743-757.

BROWN, C., MURRAY, A. & VERSTREPEN, K. (2010) Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology,* 20, 895-903.

BURKI, F., INAGAKI, Y., BRATE, J., ARCHIBALD, J., KEELING, P., CAVALIER-SMITH, T., SAKAGUCHI, M., HASHIMOTO, T., HORAK, A. & KUMAR, S. (2010) Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. *Genome Biology and Evolution,* 1**,** 231-238.

BURKI, F., SHALCHIAN-TABRIZI, K., MINGE, M., SKJAEVELAND, A., NIKOLAEV, S. I., JAKOBSEN, K. S. & PAWLOWSKI, J. (2007) Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE,* 2**,** e790.

BURKI, F., SHALCHIAN-TABRIZI, K. & PAWLOWSKI, J. (2008) Phylogenomics reveals a new 'megagroup'including most photosynthetic eukaryotes. *Biology letters,* 4**,** 366-369.

BURLEIGH, J., DRISKELL, A. & SANDERSON, M. (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Systematic Biology,* 55**,** 426-440.

CAMIN, J. & SOKAL, R. (1965) A method for deducing branching sequences in phylogeny. *Evolution,* 19**,** 311-326.

CARPENTER, J., GOLOBOFF, P. & FARRIS, J. (1998) PTP is Meaningless, T-PTP is Contradictory: A Reply to Trueman. *Cladistics,* 14**,** 105-116.

CASTRESANA, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution,* 17**,** 540-552.

CAVALIER-SMITH, T. (2010) Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. *Biology Direct,* 5**,** 7.

CAVALLI-SFORZA, L. & EDWARDS, A. (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics,* 19**,** 233-257.

CICCARELLI, F., DOERKS, T., VON MERING, C., CREEVEY, C., SNEL, B. & BORK, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science,* 311**,** 1283-1287.

COLLESS, D. (1982) Review of Phylogenetics: The Theory and Practice of Phylogenetic Systematics, by E. 0. Wiley. *Systematic Zoology,* 31**,** 100-104.

COLLESS, D. (1995) Relative symmetry of cladograms and phenograms: an experimental study. *Systematic Biology,* 44**,** 102-108.

COPLEY, R., ALOY, P., RUSSELL, R. & TELFORD, M. (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. *Evolution & development,* 6**,** 164-169.

COTTON, J. (2005) Analytical methods for detecting paralogy in molecular datasets. *Methods in Enzymology,* 395**,** 700-724.

COTTON, J. & MCINERNEY, J. (2010) Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proceedings of the National Academy of Sciences USA,* 107**,** 17252-17255.

COTTON, J. & PAGE, R. (2003) Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. *Molecular Phylogenetics and Evolution,* 29**,** 298-308.

COTTON, J. & PAGE, R. (2004) Tangled tales from multiple markers. IN BININDA-EDMONSDS, O.R.P. (Ed.) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life***,** Dordrecht, Kluwer Academic.

COTTON, J. & WILKINSON, M. (2007) Majority-rule supertrees. *Systematic Biology,* 56**,** 445-452.

COTTON, J. & WILKINSON, M. (2009) Supertrees join the mainstream of phylogenetics. *Trends in Ecology & Evolution,* 24**,** 1-3.

COX, C., FOSTER, P., HIRT, R., HARRIS, S. & EMBLEY, T. (2008) The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences USA,* 105**,** 20356- 20361.

CREEVEY, C., FITZPATRICK, D., PHILIP, G., KINSELLA, R., O'CONNELL, M., PENTONY, M., TRAVERS, S., WILKINSON, M. & MCINERNEY, J. (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proceedings of the Royal Society B: Biological Sciences,* 271**,** 2551-2558.

CREEVEY, C. & MCINERNEY, J. (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics,* 21**,** 390-392.

CUMMINS, C. & MCINERNEY, J. O. (2011) A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology,* In Press.

DAGAN, T. & MARTIN, W. (2006) The tree of one percent. *Genome Biology,* 7**,** 118.

DAGAN, T., ROETTGER, M., BRYANT, D. & MARTIN, W. (2010) Genome networks root the tree of life between prokaryotic domains. *Genome Biology and Evolution,* 2**,** 379-392.

DARWIN, C. (1859) *On the Origin of Species by means of natural selection,* London, John Murray.

DAYHOFF, M., SCHWARTZ, R. & ORCUTT, B. (1978) A model of evolutionary change in proteins. IN DAYHOFF, M. O. (Ed.) *Atlas of Protein Sequence and Structure Volume 5.* Washington DC, National Biomedical Research Foundation.

DE DUVE, C. (2007) The origin of eukaryotes: a reappraisal. *Nature Reviews Genetics,* 8**,** 395-403.

DE QUEIROZ, A. & GATESY, J. (2007) The supermatrix approach to systematics. *Trends in Ecology & Evolution,* 22**,** 34-41.

DELSUC, F., BRINKMANN, H. & PHILIPPE, H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics,* 6**,** 361-375.

DONDOSHANSKY, I. & WOLF, Y. I. (2000) BLASTCLUST. 2.2 ed. Bethesda (MD), National Institutes of Health.

DOPAZO, H. & DOPAZO, J. (2005) Genome-scale evidence of the nematode-arthropod clade. *Genome Biology,* 6**,** R41.

DOPAZO, H., SANTOYO, J. & DOPAZO, J. (2004) Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. *Bioinformatics ,* 20**,** 116-121.

DOUADY, C., DELSUC, F., BOUCHER, Y., DOOLITTLE, W. & DOUZERY, E. (2003) Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution,* 20**,** 248-254.

DRISKELL, A., ANE, C., BURLEIGH, J., MCMAHON, M., O'MEARA, B. & SANDERSON, M. (2004) Prospects for building the tree of life from large sequence databases. *Science,* 306**,** 1172-1174.

DRMANAC, R., SPARKS, A. B., CALLOW, M. J., HALPERN, A. L., BURNS, N. L., KERMANI, B. G., CARNEVALI, P., NAZARENKO, I., NILSEN, G. B., YEUNG, G., DAHL, F., FERNANDEZ, A., STAKER, B., PANT, K. P., BACCASH, J., BORCHERDING, A. P., BROWNLEY, A., CEDENO, R., CHEN, L., CHERNIKOFF, D., CHEUNG, A., CHIRITA, R., CURSON, B., EBERT, J. C., HACKER, C. R., HARTLAGE, R., HAUSER, B., HUANG, S., JIANG, Y., KARPINCHYK, V., KOENIG, M., KONG, C., LANDERS, T., LE, C., LIU, J., MCBRIDE, C. E., MORENZONI, M., MOREY, R. E., MUTCH, K., PERAZICH, H., PERRY, K., PETERS, B. A., PETERSON, J., PETHIYAGODA, C. L., POTHURAJU, K., RICHTER, C., ROSENBAUM, A. M., ROY, S., SHAFTO, J., SHARANHOVICH, U., SHANNON, K. W., SHEPPY, C. G., SUN, M., THAKURIA, J. V., TRAN, A., VU, D., ZARANEK, A. W., WU, X., DRMANAC, S., OLIPHANT, A. R., BANYAI, W. C., MARTIN, B., BALLINGER, D. G., CHURCH, G. M. & REID, C. A. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science,* 327**,** 78-81.

DUNN, C. W., HEJNOL, A., MATUS, D. Q., PANG, K., BROWNE, W. E., SMITH, S. A., SEAVER, E., ROUSE, G. W., OBST, M., EDGECOMBE, G. D., SØRENSEN, M. V., HADDOCK, S. H. D., SCHMIDT-RHAESA, A., OKUSU, A., KRISTENSEN, R. M., WHEELER, W. C., MARTINDALE, M. Q. & GIRIBET, G. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature,* 452**,** 745-749.

EDGAR, R. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5**,** 113.

EDGAR, R. & BATZOGLOU, S. (2006) Multiple sequence alignment. *Current Opinion in Structural Biology,* 16**,** 368-373.

EDWARDS, A. & CAVALLI-SFORZA, L. (1964) Reconstruction of evolutionary trees. IN HEYWOOD, V. H. & MCNEILL, J. (Eds.) *Phenetic and Phylogenetic Classification,* London, Systematics Association Publ. No. 6.

EDWARDS, A. W. F. (1972) *Likelihood,* Cambridge, Cambridge University Press.

EDWARDS, A. W. F. & CAVALLI-SFORZA, L. (1963) The reconstruction of evolution. *Annals of Human Genetics (Lond.),* 21**,** 105-106.

EERNISSE, D., ALBERT, J. & ANDERSON, F. (1992) Annelida and Arthropoda are not sister taxa: a phylogenetic analysis of spiralian metazoan morphology. *Systematic Biology,* 41, 305-330.

EERNISSE, D. & PETERSON, K. J. (2004) The history of animals. IN CRACRAFT, J. & DONOGHUE, M. J. (Eds.) *Assembling the Tree of Life.* New York, Oxford University Press.

EISEN, J. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research,* 8, 163-167.

EMBLEY, T. & MARTIN, W. (2006) Eukaryotic evolution, changes and challenges. *Nature,* 440, 623-630.

ENRIGHT, A., VAN DONGEN, S. & OUZOUNIS, C. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research,* 30, 1575-1584.

ERIXON, P., SVENNBLAD, B., BRITTON, T. & OXELMAN, B. (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology,* 52, 665-673.

ESSER, C., AHMADINEJAD, N., WIEGAND, C., ROTTE, C., SEBASTIANI, F., GELIUS-DIETRICH, G., HENZE, K., KRETSCHMANN, E., RICHLY, E. & LEISTER, D. (2004) A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution,* 21, 1643-1660.

ESTABROOK, G., JOHNSON, C. & MCMORRIS, F. (1976) An algebraic analysis of cladistic characters. *Discrete Mathematics,* 16, 141-147.

ESTABROOK, G., STRAUCH JR, J. & FIALA, K. (1977) An application of compatibility analysis to the Blackiths' data on orthopteroid insects. *Systematic Zoology,* 26, 269-276.

FAITH, D. (1991) Cladistic permutation tests for monophyly and nonmonophyly. *Systematic Zoology,* 40, 366-375.

FARRIS, J., ALBERT, V., KÄLLERSJÖ, M., LIPSCOMB, D. & KLUGE, A. (1996) Parsimony jackknifing outperforms neighbor-joining. *Cladistics,* 12, 99-124.

FARRIS, J., KLUGE, A. & ECKARDT, M. (1970) A numerical approach to phylogenetic systematics. *Systematic Zoology,* 19, 172-191.

FELSENSTEIN, J. (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology,* 22, 240-249.

FELSENSTEIN, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology,* 27, 401-410.

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

FELSENSTEIN, J. (1985) Phylogenies and the comparative method. *American Naturalist,* 125, 1-15.

FELSENSTEIN, J. (2004) *Inferring phylogenies,* Sunderland, Sinauer Associates.

FELSENSTEIN, J. (2005) PHYLIP (Phylogeny Inference Package). 3.67 ed. Seattle (WA), Department of Genome Sciences, University of Washington.

FISHER, R. (1912) On an absolute criterion for fitting frequency curves. *Messenger of Mathematics,* 41, 155-160.

FISHER, R. (1921) On the ‘‘ probable error’’ of a coefficient of correlation deduced from a small sample. *Metron,* 1, 3-32.

FISHER, R. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society London, Series A,* 222**,** 309-368.

FITCH, W. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology,* 20**,** 406-416.

FITCH, W. (2000) Homology: a personal view on some of the problems. *TRENDS in Genetics,* 16**,** 227-231.

FITCH, W. & MARGOLIASH, E. (1967) Construction of phylogenetic trees. *Science,* 155**,** 279-284.

FITCH, W. & MARKOWITZ, E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics,* 4**,** 579-593.

FITZPATRICK, D., LOGUE, M., STAJICH, J. & BUTLER, G. (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology,* 6**,** 99.

FLEISCHMANN, R., ADAMS, M., WHITE, O., CLAYTON, R., KIRKNESS, E., KERLAVAGE, A., BULT, C., TOMB, J., DOUGHERTY, B. & MERRICK, J. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* 269**,** 496-512.

FOSTER, P. (2004) Modeling compositional heterogeneity. *Systematic Biology,* 53**,** 485-495.

FOSTER, P., COX, C. & EMBLEY, T. (2009) The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364**,** 2197-2207.

FOSTER, P. & HICKEY, D. (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution,* 48**,** 284-290.

FOSTER, P., JERMIIN, L. & HICKEY, D. (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *Journal of Molecular Evolution,* 44**,** 282-288.

FOX, G. E., MAGRUM, L. J., BALCH, W. E., WOLFE, R. S. & WOESE, C. R. (1977) Classification of methanogenic bacteria by 16s ribosomal RNA characterization. *Proceedings of the National Academy of Sciences USA,* 74**,** 4537-4541.

FOX, J. & WEISBERG, S. (2010) An {R} Companion to Applied Regression. Second ed. Thousand Oaks (CA), Sage.

FRIEDMAN, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association,* 32**,** 675-701.

FURNAS, G. W. (1984) The generation of random, binary, unrooted trees. *Journal of Classification,* 1**,** 187-233.

FUSCO, G. & CRONK, Q. (1995) A new method for evaluating the shape of large phylogenies. *Journal of theoretical biology,* 175**,** 235-243.

GASCUEL, O., BRYANT, D. & DENIS, F. (2001) Strengths and limitations of the minimum evolution principle. *Systematic Biology,* 50**,** 621-627.

GAUT, B. & LEWIS, P. (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution,* 12**,** 152-162.

GEE, H. (2003) Ending incongruence. *Nature,* 425**,** 782.

GEISSER, S. & GREENHOUSE, S. (1958) An extension of Box's results on the use of the F distribution in multivariate analysis. *The Annals of Mathematical Statistics*, 885-891.

GOLDMAN, N. (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution,* 36**,** 182-198.

GOLUBCHIK, T., WISE, M., EASTEAL, S. & JERMIIN, L. (2007) Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Molecular Biology and Evolution,* 24**,** 2433-2442.

GOODMAN, S. (1999) Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine,* 130**,** 1005-1013.

GORDON, A. (1986) Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification,* 3**,** 335-348.

GOULD, S. (1989) *Wonderful life: the Burgess Shale and the nature of history* New York, W.W. Norton.

GRAYBEAL, A. (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology,* 47**,** 9-17.

GRIBALDO, S., POOLE, A., DAUBIN, V., FORTERRE, P. & BROCHIER-ARMANET, C. (2010) The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nature Reviews Microbiology,* 8**,** 743-752.

GROBBEN, K. (1908) Die systematische Einteilung des Tierreiches. *Verh. Zool. Bot. Ges. Wien,* 58**,** l-5.

GUINDON, S. & GASCUEL, O. (2003) A simple, fast, and accurate method to estimate large phylogenies by Maximum Likelihood. *Systematic Biology,* 52**,** 696-704.

GUYER, C. & SLOWINSKI, J. (1993) Adaptive radiation and the topology of large phylogenies. *Evolution,* 47**,** 253-263.

HACKETT, J. D., YOON, H. S., LI, S., REYES-PRIETO, A., RÜMMELE, S. E. & BHATTACHARYA, D. (2007) Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. *Molecular Biology and Evolution,* 24**,** 1702-1713.

HAECKEL, E. H. P. A. (1866) *Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin reformirte Descendenz-Theorie,* Berlin, G. Reimer.

HAECKEL, E. H. P. A. (1872) *Die Kalkschwämme: Eine Monographie,* Berlin, G. Reimer.

HALANYCH, K. (1998) Considerations for reconstructing metazoan history: signal, resolution, and hypothesis testing. *Integrative and Comparative Biology,* 38**,** 929-941.

HALANYCH, K. (2004) The new view of animal phylogeny. *Annual Reviews of Ecology and Systematics*, 35, 229-256.

HALANYCH, K., BACHELLER, J., AGUINALDO, A., LIVA, S., HILLIS, D. & LAKE, J. (1995) Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science,* 267**,** 1641-1643.

HALLETT, M. & LAGERGREN, J. (2000) New algorithms for the duplication-loss model. *Proceedings of the fourth annual international conference on Computational Molecular Biology***,** 138-146.

HAMPL, V., HUG, L., LEIGH, J., DACKS, J., LANG, B., SIMPSON, A. & ROGER, A. (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences USA,* 106**,** 3859.

HARCOURT-BROWN, K., PEARSON, P. & WILKINSON, M. (2001) The imbalance of paleontological trees. *Paleobiology,* 27**,** 188-204.

HARRIS, T. E. (1963) *The theory of branching processes,* Berlin, Springer.

HARVEY, P. H. & PURVIS, A. (1991) Comparative methods for explaining adaptations. *Nature,* 351**,** 619-624.

HASEGAWA, M. & HASHIMOTO, T. (1993) Ribosomal RNA trees misleading? *Nature,* 361**,** 23.

HASEGAWA, M., KISHINO, H. & SAITOU, N. (1991) On the maximum likelihood method in molecular phylogenetics. *Journal of Molecular Evolution,* 32**,** 443-445.

HASTINGS, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* 57**,** 97-109.

HEARD, S. (1992) Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution,* 46**,** 1818-1826.

HEARD, S. (1996) Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution,* 50**,** 2141-2148.

HEDGES, S., BLAIR, J., VENTURI, M. & SHOE, J. (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evolutionary Biology,* 4**,** 2.

HEJNOL, A., OBST, M., STAMATAKIS, A., OTT, M., ROUSE, G. W., EDGECOMBE, G. D., MARTINEZ, P., BAGUÑÀ, J., BAILLY, X., JONDELIUS, U., WIENS, M., MÜLLER, W. E. G., SEAVER, E., WHEELER, W. C., MARTINDALE, M. Q., GIRIBET, G. & DUNN, C. W. (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences,* 276**,** 4261-4270.

HENDY, M. & PENNY, D. (1989) A framework for the quantitative study of evolutionary trees. *Systematic Biology,* 38**,** 297-309.

HENIKOFF, S. & HENIKOFF, J. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA,* 89**,** 10915-10919.

HILLIS, D. (1996) Inferring complex phylogenies. *Nature,* 383**,** 130-131.

HIRT, R., LOGSDON, J., HEALY, B., DOREY, M., DOOLITTLE, W. & EMBLEY, T. (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proceedings of the National Academy of Sciences USA,* 96**,** 580-585.

HJORT, K., GOLDBERG, A., TSAOUSIS, A., HIRT, R. & EMBLEY, T. (2010) Diversity and reductive evolution of mitochondria among microbial eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 365**,** 713-727.

HOLDER, M., SUKUMARAN, J. & LEWIS, P. (2008) A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. *Systematic Biology,* 57**,** 814-821.

HOLLANDER, M. & WOLFE, D. A. (1999) *Nonparametric Statistical Methods,* New York, Wiley.

HOLTON, T. & PISANI, D. (2010) Deep Genomic-Scale Analyses of the Metazoa Reject Coelomata: Evidence from Single-and Multigene Families Analyzed Under a Supertree and Supermatrix Paradigm. *Genome Biology and Evolution,* 2**,** 310-324.

HONE, D., KEESEY, T., PISANI, D. & PURVIS, A. (2005) Macroevolutionary trends in the Dinosauria: Cope's rule. *Journal of Evolutionary Biology,* 18**,** 587-595.

HOTHORN, T., HORNIK, K., VAN DE WIEL, M. & ZEILEIS, A. (2008) Implementing a class of permutation tests: The coin package. *Journal of Statistical Software,* 28**,** 1–23.

HRDY, I., HIRT, R., DOLEZAL, P., BARDONOVÁ, L., FOSTER, P., TACHEZY, J. & EMBLEY, T. (2004) *Trichomonas* hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature,* 432**,** 618-622.

HUELSENBECK, J. (1995) The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Molecular Biology and Evolution,* 12**,** 843-849.

HUELSENBECK, J. & BOLLBACK, J. (2001) Empirical and hierarchical Bayesian estimation of ancestral states. *Systematic Biology,* 50**,** 351-366.

HUELSENBECK, J. & KIRKPATRICK, M. (1996) Do phylogenetic methods produce trees with biased shapes? *Evolution,* 50**,** 1418-1424.

HUELSENBECK, J. & RONQUIST, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17**,** 754-755.

HUELSENBECK, J., RONQUIST, F., NIELSEN, R. & BOLLBACK, J. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294**,** 2310-2314.

HUGHES, J., LONGHORN, S. J., PAPADOPOULOU, A., THEODORIDES, K., DE RIVA, A., MEJIA-CHANG, M., FOSTER, P. G. & VOGLER, A. P. (2006) Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). *Molecular Biology and Evolution,* 23**,** 268-278.

HUYNH, H. & FELDT, L. (1976) Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational and Behavioral Statistics,* 1**,** 69-82.

HYMAN, L. H. (1940) *The Invertebrates: Protozoa through Ctenophora.,* New York, McGraw-Hill.

HYMAN, L. H. (1951) *The Invertebrates: Platyhelminthes and Rhynchocoela, the acoelomate Bilateria.,* New York, McGraw-Hill.

HYMAN, L. H. (1959) *The Invertebrates: Smaller Coelomate Groups, Chaetognatha, Hermichordata, Pogonophora, Phoronida, Ectoprocta, Brachiopoda, Sipunculida,the Coelomate Bilateria,* New York, McGraw-Hill.

HYMAN, L. H. (1967) *The Invertebrates: Mollusca I.,* New York, McGraw-Hill.

IRIMIA, M., MAESO, I., PENNY, D., GARCIA-FERNANDEZ, J. & ROY, S. (2007) Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. *Molecular Biology and Evolution,* 24**,** 1604-1607.

JEFFROY, O., BRINKMANN, H., DELSUC, F. & PHILIPPE, H. (2006) Phylogenomics: the beginning of incongruence? *TRENDS in Genetics,* 22**,** 225-231.

JENNER, R. & SCHRAM, F. (1999) The grand game of metazoan phylogeny: rules and strategies. *Biological Reviews,* 74**,** 121-142.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences,* 8**,** 275-282.

JUKES, T. & CANTOR, C. (1969) Evolution of protein molecules. IN MUNRO, H. N. (Ed.) *Mammalian Protein Metabolism.* New York, Academic Press.

KASHYAP, R. & SUBAS, S. (1974) Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *Journal of Theoretical Biology,* 47**,** 75-101.

KASS, R. & RAFTERY, A. (1995) Bayes factors. *Journal of the American Statistical Association,* 90**,** 773-795.

KEANE, T., CREEVEY, C., PENTONY, M., NAUGHTON, T. & MCLNERNEY, J. (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology,* 6**,** 29.

KEELING, P. (2007) GENOMICS: Deep Questions in the Tree of Life. *Science,* 317**,** 1875-1876.

KEELING, P., BURGER, G., DURNFORD, D., LANG, B., LEE, R., PEARLMAN, R., ROGER, A. & GRAY, M. (2005) The tree of eukaryotes. *Trends in ecology & evolution,* 20**,** 670-676.

KEELING, P. & PALMER, J. (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics,* 9**,** 605-618.

KEMENA, C. & NOTREDAME, C. (2009) Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics,* 25**,** 2455-2465.

KESELMAN, H. (1998) Testing treatment effects in repeated measures designs: An update for psychophysiological researchers. *Psychophysiology,* 35**,** 470-478.

KIDD, K. & SGARAMELLA-ZONTA, L. (1971) Phylogenetic analysis: concepts and methods. *American Journal of Human Genetics,* 23**,** 235-252.

KIM, J. (1996) General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. *Systematic Biology,* 45**,** 363-374.

KIMURA, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics,* 61**,** 893-903.

KIMURA, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution,* 16**,** 111-120.

KIRKPATRICK, M. & SLATKIN, M. (1993) Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution,* 47**,** 1171-1181.

KLUGE, A. (1989) A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Zoology,* 38**,** 7-25.

KLUGE, A. & FARRIS, J. (1969) Quantitative phyletics and the evolution of anurans. *Systematic Zoology,* 18**,** 1-32.

KOLACZKOWSKI, B. & THORNTON, J. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature,* 431**,** 980-984.

KOLACZKOWSKI, B. & THORNTON, J. (2009) Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS ONE,* 4**,** e7891.

KOONIN, E. (2007) The Biological Big Bang model for the major transitions in evolution. *Biology Direct,* 2**,** 21.

KOONIN, E. (2010) The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biology,* 11**,** 209.

LAKE, J. (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences USA,* 91**,** 1455.

LAMARCK, J. B. (1809) *Philosophie zoologique ou exposition des considérations relatives à l'histoire naturelle des animaux,* Paris, F. Savy.

LANAVE, C., PREPARATA, G., SACCONE, C. & SERIO, G. (1984) A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution,* 20**,** 86-93.

LANE, C. & ARCHIBALD, J. (2008) The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends in Ecology & Evolution,* 23, 268-275.

LANE, N. & MARTIN, W. (2010) The energetics of genome complexity. *Nature,* 467, 929-934.

LARTILLOT, N. & PHILIPPE, H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution,* 21, 1095.

LARTILLOT, N. & PHILIPPE, H. (2008) Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363, 1463-1472.

LE, S. & GASCUEL, O. (2008) An improved general amino acid replacement matrix. *Molecular Biology and Evolution,* 25, 1307-1320.

LEWIS, P. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology,* 50, 913-925.

LI, S. (1996) Phylogenetic tree construction using Markov chain Monte carlo. Ohio State University, Columbus.

LINDMAN, H. R. (1974) *Analysis of variance in complex experimental designs,* San Francisco, W.H Freeman.

LIOLIOS, K., CHEN, I., MIN, A., MAVROMATIS, K., TAVERNARAKIS, N., HUGENHOLTZ, P., MARKOWITZ, V. & KYRPIDES, N. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research,* 38, D346-254.

LITTLEWOOD, D.T.J., OLSON, P., TELFORD, M., HERNIOU, E. & RIUTORT, M. (2001) Elongation factor 1-alpha sequences alone do not assist in resolving the position of the Acoela within the Metazoa. *Molecular Biology and Evolution,* 18, 437-442.

LLOYD, G., DAVIS, K., PISANI, D., TARVER, J., RUTA, M., SAKAMOTO, M., HONE, D., JENNINGS, R. & BENTON, M. (2008) Dinosaurs and the Cretaceous terrestrial revolution. *Proceedings of the Royal Society B: Biological Sciences,* 275, 2483-2490.

LOCKHART, P., STEEL, M., HENDY, M. & PENNY, D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution,* 11, 605-612.

LOCKHART, P. J., LARKUM, A. W., STEEL, M., WADDELL, P. J. & PENNY, D. (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proceedings of the National Academy of Sciences USA,* 93, 1930-1934.

LOGSDON, J. (2010) Eukaryotic Evolution: The importance of being Archaebacterial. *Current Biology*, R1078-R1079.

LOOMIS, W. & SMITH, D. (1990) Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proceedings of the National Academy of Sciences USA,* 87, 9093-9097.

LOPEZ, P., CASANE, D. & PHILIPPE, H. (2002) Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution,* 19, 1-7.

LOSOS, J. & ADLER, F. (1995) Stumped by trees? A generalized null model for patterns of organismal diversity. *American Naturalist,* 145, 329-342.

LÖYTYNOJA, A. & GOLDMAN, N. (2008) Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science,* 320, 1632-1635.

MARGULIS, L. (1970) *Origin of Eukaryotic Cells,* New Haven and London, Yale University Press.

MARGUSH, T. & MCMORRIS, F. (1981) Consensus n-trees. *Bulletin of Mathematical Biology,* 43, 239-244.

MATSEN, F. (2006) A geometric approach to tree shape statistics. *Systematic Biology,* 55, 652-661.

MAU, B. & NEWTON, M. (1997) Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics,* 6, 122-131.

MAUCHLY, J. (1940) Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics,* 11, 204-209.

MCCANN, A., COTTON, J. & MCINERNEY, J. (2008) The tree of genomes: An empirical comparison of genome-phylogeny reconstruction methods. *BMC Evolutionary Biology,* 8, 312.

MCINERNEY, J., COTTON, J. & PISANI, D. (2008) The prokaryotic tree of life: past, present… and future? *Trends in Ecology & Evolution,* 23, 276-281.

MCKENZIE, A. & STEEL, M. (2000) Distributions of cherries for two models of trees. *Mathematical Biosciences,* 164, 81-92.

MEACHAM, C. & ESTABROOK, G. (1985) Compatibility methods in systematics. *Annual Review of Ecology and Systematics,* 16, 431-446.

METROPOLIS, N., ROSENBLUTH, A., ROSENBLUTH, M., TELLER, A. & TELLER, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,* 21, 1087.

MOOERS, A. & HEARD, S. (1997) Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology,* 72, 31-54.

MOOERS, A. & HEARD, S. (2002) Using tree shape. *Systematic Biology,* 51, 833-834.

MOORE, B., SMITH, S. & DONOGHUE, M. (2006) Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Systematic Biology,* 55, 662-676.

MURPHY, W., EIZIRIK, E., O'BRIEN, S., MADSEN, O., SCALLY, M., DOUADY, C., TEELING, E., RYDER, O., STANHOPE, M. & DE JONG, W. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science,* 294, 2348-2351.

NEI, M. (1986) Stochastic errors in DNA evolution and molecular phylogeny. *Progress in Clinical Biological Research,* 218, 133-147.

NEYMAN, J. (1971) Molecular studies of evolution: a source of novel statistical problems. IN GUPTA, S. S. & YACKEL, J. (Eds.) *Statistical Decision Theory and Related Topics.* New York, Academic.

NIELSEN, C. (2001) *Animal Evolution, Interrelationships of the Living Phyla.,* Oxford, Oxford University Press.

O'BRIEN, K., REMM, M. & SONNHAMMER, E. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research,* 33**,** D476-480.

OVERALL, J. & DOYLE, S. (1994) Estimating sample sizes for repeated measurement designs* 1. *Controlled Clinical Trials,* 15**,** 100-123.

OWEN, R. (1843) Lectures on the comparative anatomy and physiology of the invertebrate animals. *Longman, Brown, Green and Longman, London*.

PARADIS, E., CLAUDE, J. & STRIMMER, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics,* 20**,** 289-290.

PARFREY, L., BARBERO, E., LASSER, E., DUNTHORN, M., BHATTACHARYA, D., PATTERSON, D. & KATZ, L. (2006) Evaluating support for the current classification of eukaryotic diversity. *PLoS Genetics,* 2**,** e220.

PETERSON, K., COTTON, J., GEHLING, J. & PISANI, D. (2008) The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363**,** 1435-1443.

PHILIP, G., CREEVEY, C. & MCINERNEY, J. (2005) The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Molecular Biology and Evolution,* 22**,** 1175-1184.

PHILIPPE, H. (2000) Opinion: long branch attraction and protist phylogeny. *Protist,* 151**,** 307-316.

PHILIPPE, H., BRINKMANN, H., COPLEY, R. R., MOROZ, L. L., NAKANO, H., POUSTKA, A. J., WALLBERG, A., PETERSON, K. J. & TELFORD, M. J. (2011) Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature,* 470, 255-258.

PHILIPPE, H., BRINKMANN, H., MARTINEZ, P., RIUTORT, M. & BAGUÑÀ, J. (2007) Acoel flatworms are not Platyhelminthes: evidence from phylogenomics. *PLoS ONE,* 2**,** e717.

PHILIPPE, H., DELSUC, F., BRINKMANN, H. & LARTILLOT, N. (2005a) Phylogenomics. *Annual Review of Ecology, Evolution and Systematics*, 36, 541-562.

PHILIPPE, H., DERELLE, R., LOPEZ, P., PICK, K., BORCHIELLINI, C., BOURY-ESNAULT, N., VACELET, J., RENARD, E., HOULISTON, E. & QUÉINNEC, E. (2009) Phylogenomics revives traditional views on deep animal relationships. *Current Biology,* 19**,** 706-712.

PHILIPPE, H., LARTILLOT, N. & BRINKMANN, H. (2005b) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution,* 22**,** 1246-1253.

PHILIPPE, H. & LAURENT, J. (1998) How good are deep phylogenetic trees? *Current Opinion in Genetics & Development,* 8**,** 616-623.

PHILIPPE, H., SNELL, E., BAPTESTE, E., LOPEZ, P., HOLLAND, P. & CASANE, D. (2004) Phylogenomics of eukaryotes: impact of missing data on large alignments. *Molecular Biology and Evolution,* 21**,** 1740-1752.

PICK, K., PHILIPPE, H., SCHREIBER, F., ERPENBECK, D., JACKSON, D., WREDE, P., WIENS, M., ALIÉ, A., MORGENSTERN, B. & MANUEL, M. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Molecular Biology and Evolution,* 27**,** 1983-1987.

PINELIS, I. (2003) Evolutionary models of phylogenetic trees. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 270**,** 1425-1431.

PISANI, D. (2004) Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. *Systematic Biology,* 53**,** 978-989.

PISANI, D., COTTON, J. & MCINERNEY, J. (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular Biology and Evolution,* 24**,** 1752-1760.

PISANI, D. & WILKINSON, M. (2002) Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology,* 51**,** 151-155.

PISANI, D., YATES, A., LANGER, M. & BENTON, M. (2002) A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 269**,** 915-921.

POE, S. (2003) Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Systematic Biology,* 52**,** 423-428.

POE, S. & SWOFFORD, D. (1999) Taxon sampling revisited. *Nature,* 398**,** 299-300.

POL, D. (2004) Empirical problems of the hierarchical likelihood ratio test for model selection. *Systematic Biology,* 53**,** 949-962.

POLLOCK, D., ZWICKL, D., MCGUIRE, J. & HILLIS, D. (2002) Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology,* 51**,** 664-671.

POOLE, A. & PENNY, D. (2007) Eukaryote evolution: engulfed by speculation. *Nature,* 447**,** 913.

POSADA, D. & CRANDALL, K. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics,* 14**,** 817-818.

PURVIS, A. (1995) A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology,* 44**,** 251-255.

PURVIS, A. & AGAPOW, P. (2002) Phylogeny imbalance: taxonomic level matters. *Systematic Biology,* 51**,** 844-854.

PURVIS, A., NEE, S. & HARVEY, P. (1995) Macroevolutionary inferences from primate phylogeny. *Proceedings: Biological Sciences,* 260**,** 329-333.

RAGAN, M. (1992) Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *BioSystems,* 28**,** 47-55.

RAMBAUT, A. & DRUMMOND, A. J. (2007) Tracer. 1.4 ed., http://beast.bio.ed.ac.uk/Tracer.

RANNALA, B., HUELSENBECK, J., YANG, Z. & NIELSEN, R. (1998) Taxon sampling and the accuracy of large phylogenies. *Systematic Biology,* 47**,** 702-719.

RANNALA, B. & YANG, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution,* 43**,** 304-311.

RIVERA, M. & LAKE, J. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature,* 431**,** 152-155.

ROBINSON, M., GOUY, M., GAUTIER, C. & MOUCHIROUD, D. (1998) Sensitivity of the relative-rate test to taxonomic sampling. *Molecular Biology and Evolution,* 15**,** 1091-1098.

RODRIGO, A. (1993) A comment on Baum's method for combining phylogenetic trees. *Taxon,* 42**,** 631-636.

RODRIGO, A. (1996) On combining cladograms. *Taxon,* 45**,** 267-274.

RODRIGUEZ, F., OLIVER, J., MARIN, A. & MEDINA, J. (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology,* 142**,** 485-501.

RODRÍGUEZ-EZPELETA, N., BRINKMANN, H., BURGER, G., ROGER, A., GRAY, M., PHILIPPE, H. & LANG, B. (2007) Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Current Biology,* 17**,** 1420-1425.

RODRÍGUEZ-EZPELETA, N., BRINKMANN, H., ROURE, B., LARTILLOT, N., LANG, B. & PHILIPPE, H. (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology,* 56**,** 389-399.

ROGER, A. & SIMPSON, A. (2009) Evolution: revisiting the root of the eukaryote tree. *Current Biology,* 19**,** R165-R167.

ROGERS, J. (1994) Central moments and probability distribution of Colless's coefficient of tree imbalance. *Evolution,* 48**,** 2026-2036.

ROGERS, J. (1996) Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic Biology,* 45**,** 99-110.

ROGOZIN, I., BASU, M., CSUROS, M. & KOONIN, E. (2009) Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome Biology and Evolution,* 2009**,** 99-113.

ROGOZIN, I., THOMSON, K., CSÜRÖS, M., CARMEL, L. & KOONIN, E. (2008) Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. *Biology Direct,* 3**,** 7.

ROGOZIN, I., WOLF, Y., CARMEL, L. & KOONIN, E. (2007) Analysis of rare amino acid replacements supports the Coelomata clade. *Molecular Biology and Evolution,* 24**,** 2594-2597.

ROKAS, A. & CARROLL, S. (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular Biology and Evolution,* 22**,** 1337-1344.

RONQUIST, F. & HUELSENBECK, J. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19**,** 1572-1574.

ROSEN, D. (1978) Vicariant patterns and historical explanation in biogeography. *Systematic Biology,* 27**,** 159-188.

ROSENBERG, M. & KUMAR, S. (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences USA,* 98**,** 10751-10756.

ROTA-STABELLI, O., CAMPBELL, L., BRINKMANN, H., EDGECOMBE, G. D., LONGHORN, S. J., PETERSON, K. J., PISANI, D., PHILIPPE, H. & TELFORD, M. J. (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society B: Biological Sciences,* 278**,** 298-306.

ROTA-STABELLI, O., KAYAL, E., GLEESON, D., DAUB, J., BOORE, J., TELFORD, M., PISANI, D., BLAXTER, M. & LAVROV, D. (2010) Ecdysozoan mitogenomics: Evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biology and Evolution,* 2**,** 425-440.

ROTA-STABELLI, O. & TELFORD, M. (2008) A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Molecular Phylogenetics and Evolution,* 48**,** 103-111.

RUIZ-TRILLO, I., RIUTORT, M., LITTLEWOOD, D.T.J., HERNIOU, E.A. & BAGUÑÀ, J. (1999) Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. *Science,* 283**,** 1919-1923.

RUTA, M., PISANI, D., LLOYD, G. & BENTON, M. (2007) A supertree of Temnospondyli: cladogenetic patterns in the most species-rich group of early tetrapods. *Proceedings of the Royal Society B: Biological Sciences,* 274**,** 3087-3095.

RZHETSKY, A. & NEI, M. (1992) A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution,* 9**,** 945-967.

RZHETSKY, A. & NEI, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution,* 10**,** 1073-1095.

SACKIN, M. (1972) " Good" and" Bad" Phenograms. *Systematic Zoology,* 21**,** 225-226.

SAITOU, N. & IMANISHI, T. (1989) Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Molecular Biology and Evolution,* 6**,** 514-525.

SAITOU, N. & NEI, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4**,** 406-425.

SANDERSON, M. (2008) Phylogenetic signal in the eukaryotic tree of life. *Science,* 321, 121-123.

SANDERSON, M., MCMAHON, M. & STEEL, M. (2010) Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evolutionary Biology,* 10, 155.

SANDERSON, M., PURVIS, A. & HENZE, C. (1998) Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology & Evolution,* 13, 105-109.

SANDERSON, M. & SHAFFER, H. (2002) Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics,* 33, 49-72.

SANKOFF, D. (1975) Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics,* 28, 35-42.

SANKOFF, D. & ROUSSEAU, P. (1975) Locating the vertices of a Steiner tree in an arbitrary metric space. *Mathematical Programming,* 9, 240-246.

SCHMIDT-RHAESA, A. (2003) Old trees, new trees–is there any progress? *Zoology,* 106, 291-301.

SCHWARZ, G. (1978) Estimating the dimension of a model. *The annals of statistics,* 6, 461-464.

SEMPLE, C. & STEEL, M. (2003) *Phylogenetics.,* New York, Oxford Univerity Press.

SHAO, K. & SOKAL, R. (1990) Tree balance. *Systematic Biology,* 39, 266-276.

SHAPIRO, S. & WILK, M. (1965) An analysis of variance test for normality (complete samples). *Biometrika,* 52, 591-611.

SHIMODAIRA, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology,* 51, 492-508.

SIEGEL, S. & CASTELLAN, N. J. (1988) *Nonparametric statistics for the behavioral sciences,* New York, McGraw-Hill.

SIMBERLOFF, D., HECK, K. L., MCCOY, E. D. & CONNOR, E. F. (1981) There have been no statistical tests of cladistic biogeographic hypotheses. IN NELSON, G. & ROSEN, D. E. (Eds.) *Vicariance Biogeography: A Critique.* New York, Columbia University Press.

SLOWINSKI, J. & CROTHER, B. (1998) Is the PTP Test Useful? *Cladistics,* 14, 297-302.

SLOWINSKI, J. & PAGE, R. (1999) How should species phylogenies be inferred from sequence data? *Systematic Biology,* 48, 814-825.

SOBER, E. (1988) *Reconstructing the past: parsimony, evolution, and inference,* Cambridge, MIT Press.

SOKAL, R. & ROHLF, F. (1981) Taxonomic congruence in the Leptopodomorpha re-examined. *Systematic Zoology,* 30, 309-325.

SOKAL, R. R. & ROHLF, F. J. (1995) *Biometry: the principles and practice of statistics in biological research,* New York, W. H. Freeman and Co.

SOKAL, R. R. & SNEATH, P. H. A. (1963) *Principles of Numerical Taxonomy,* San Francisco, W.H. Freeman.

SOURDIS, J. & NEI, M. (1988) Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Molecular Biology and Evolution,* 5**,** 298-311.

SPERLING, E., PETERSON, K. & PISANI, D. (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Molecular Biology and Evolution,* 26**,** 2261-2274.

SPERLING, E. A., ROBINSON, J. M., PISANI, D. & PETERSON, K. J. (2010) Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200-Myr missing Precambrian fossil record of siliceous sponge spicules. *Geobiology,* 8**,** 24-36.

STAMATAKIS, A., LUDWIG, T. & MEIER, H. (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics,* 21**,** 456-463.

STECHMANN, A. & CAVALIER-SMITH, T. (2003) The root of the eukaryote tree pinpointed. *Current Biology,* 13**,** R665-R666.

STEEL, M., HUSON, D. & LOCKHART, P. J. (2000) Invariable sites models and their use in phylogeny reconstruction. *Systematic Biology,* 49**,** 225-32.

STEEL, M. & MCKENZIE, A. (2001) Properties of phylogenetic trees generated by Yule-type speciation models*. *Mathematical biosciences,* 170**,** 91-112.

STEEL, M. & PENNY, D. (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution,* 17**,** 839-850.

STEEL, M. & RODRIGO, A. (2008) Maximum likelihood supertrees. *Systematic Biology,* 57**,** 243-250.

STRIMMER, K. & VON HAESELER, A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences USA,* 94**,** 6815-6819.

SULLIVAN, J. & SWOFFORD, D. (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution,* 4**,** 77-86.

SUSKO, E., INAGAKI, Y. & ROGER, A. J. (2004) On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. *Molecular Biology and Evolution,* 21**,** 1629-1642.

SWOFFORD, D., THORNE, J., FELSENSTEIN, J. & WIEGMANN, B. (1996) The topology-dependent permutation test for monophyly does not test for monophyly. *Systematic Biology,* 45**,** 575-579.

SWOFFORD, D. L. (1998) PAUP*. Phylogenetic analysis using parsimony (* and other methods). 4 Ed. Sunderland (MA), Sinauer Associates.

TAMURA, K., NEI, M. & KUMAR, S. (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences USA,* 101**,** 11030-11035.

TATENO, Y., TAKEZAKI, N. & NEI, M. (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution,* 11**,** 261-277.

TAYLOR, D. & PIEL, W. (2004) An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. *Molecular Biology and Evolution,* 21**,** 1534-1537.

TELFORD, M., BOURLAT, S., ECONOMOU, A., PAPILLON, D. & ROTA-STABELLI, O. (2008) The evolution of the Ecdysozoa. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363**,** 1529-1537.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* 22**,** 4673-4680.

TRUEMAN, J. (1996) Permutation tests and outgroups. *Cladistics,* 12**,** 253-261.

VALENTINE, J. W. (2004) *On the origin of phyla.,* Chicago, The University of Chicago Press.

VASSEY, M. W. & THAYER, J. F. (1987) The continuing problem of false positives in repeated measures ANOVA in Psychophysiology: A multivariate solution. *Psychophysiology,* 24**,** 479–486.

VINH, L. & VON HAESELER, A. (2004) IQPNNI: Moving fast through tree space and stopping in time. *Molecular Biology and Evolution,* 21**,** 1565-1571.

WEHE, A., BANSAL, M., BURLEIGH, J. & EULENSTEIN, O. (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics,* 24**,** 1540-1541.

WHEELER, W. (1990) Nucleic acid sequence phylogeny and random outgroups. *Cladistics,* 6**,** 363–367.

WHELAN, S., BLACKBURNE, B. & SPENCER, M. (2011) Phylogenetic substitution models for detecting heterotachy during plastid evolution. *Molecular Biology and Evolution*, 28, 449-458.

WHELAN, S. & GOLDMAN, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution,* 18**,** 691-699.

WHELAN, S., LIÒ, P. & GOLDMAN, N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *TRENDS in Genetics,* 17**,** 262-272.

WIENS, J. (2006) Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics,* 39**,** 34-42.

WILKINSON, M., COTTON, J., CREEVEY, C., EULENSTEIN, O., HARRIS, S., LAPOINTE, F., LEVASSEUR, C., MCINERNEY, J., PISANI, D. & THORLEY, J. (2005) The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Systematic Biology,* 54**,** 419-431.

WILKINSON, M. & COTTON, J. A. (2006) Supertree methods for building the tree of life: divide-and-conquer approaches to large phylogenetic problems. IN

HODKINSON, T., PARNELL, J. & WALDREN, S. (Eds.) *Towards the Tree of Life: Taxonomy and Systematics of Large and Species Rich Taxa.* CRC Press.

WILKINSON, M., PISANI, D., COTTON, J. & CORFE, I. (2005) Measuring support and finding unsupported relationships in supertrees. *Systematic Biology,* 54, 823-831.

WILKINSON, M., THORLEY, J., LITTLEWOOD, D.T.J & BRAY, R.A. (2001) Towards a Phylogenetic Supertree for the Platyhelminthes? IN LITTLEWOOD, D.T.J. & BRAY, R.A. (Eds.) *Interrelationships of the Platyhelminthes.* London, Chapman-Hall.

WOESE, C., ACHENBACH, L., ROUVIERE, P. & MANDELCO, L. (1991) Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition-induced artifacts. *Systematic and Applied Microbiology,* 14**,** 364-371.

WOLF, Y., ROGOZIN, I. & KOONIN, E. (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Research,* 14**,** 29-36.

WU, D., HUGENHOLTZ, P., MAVROMATIS, K., PUKALL, R., DALIN, E., IVANOVA, N. N., KUNIN, V., GOODWIN, L., WU, M., TINDALL, B. J., HOOPER, S. D., PATI, A., LYKIDIS, A., SPRING, S., ANDERSON, I. J., D'HAESELEER, P., ZEMLA, A., SINGER, M., LAPIDUS, A., NOLAN, M., COPELAND, A., HAN, C., CHEN, F., CHENG, J.-F., LUCAS, S., KERFELD, C., LANG, E., GRONOW, S., CHAIN, P., BRUCE, D., RUBIN, E. M., KYRPIDES, N. C., KLENK, H.-P. & EISEN, J. A. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature,* 462**,** 1056-1060.

YANG, Z. (1994a) Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution,* 39**,** 105-111.

YANG, Z. (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution,* 39**,** 306-314.

YANG, Z. (2006) On the varied pattern of evolution of 2 fungal genomes: a critique of Hughes and Friedman. *Molecular Biology and Evolution,* 23**,** 2279-2282.

YANG, Z. & RANNALA, B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Molecular Biology and Evolution,* 14**,** 717-724.

YOON, H., HACKETT, J., CINIGLIA, C., PINTO, G. & BHATTACHARYA, D. (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution,* 21, 809-818.

YULE, G. (1924) A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character,* 213**,** 21-87.

ZHANG, J. & NEI, M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of Molecular Evolution,* 44**,** 139-146.

ZHENG, J., ROGOZIN, I., KOONIN, E. & PRZYTYCKA, T. (2007) Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation. *Molecular Biology and Evolution,* 24**,** 2583-2592.

ZHOU, Y., RODRIGUE, N., LARTILLOT, N. & PHILIPPE, H. (2007) Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evolutionary Biology,* 7**,** 206.

ZILVERSMIT, M., O'GRADY, P. & DESALLE, R. (2002) Shallow Genomics, Phylogenetics, and Evolution in the Family Drosophilidae. *Pacific Symposium on Biocomputing,* 7**,** 512-523.

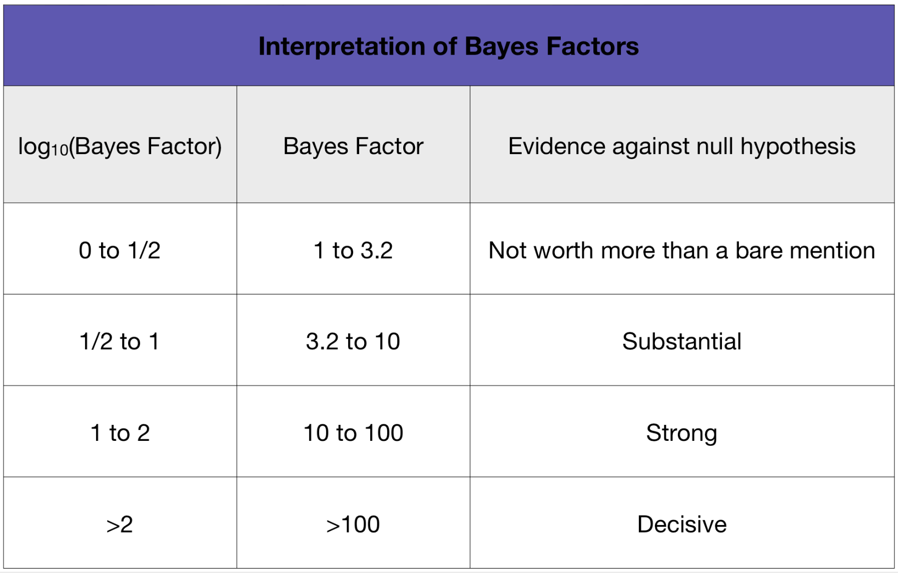
ZUCKERKANDL, E. & PAULING, L. (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology,* 8**,** 357-66.

ZWICKL, D. & HILLIS, D. (2002) Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology,* 51**,** 588-598.

Appendices

Appendix A1

Table A1. Interpretation of Bayes Factors (redrawn from Kass and Raftery, 1995)



Interpretation of Bayes Factors

$\log_{10}$(Bayes Factor)	Bayes Factor	Evidence against null hypothesis
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

Appendix A2

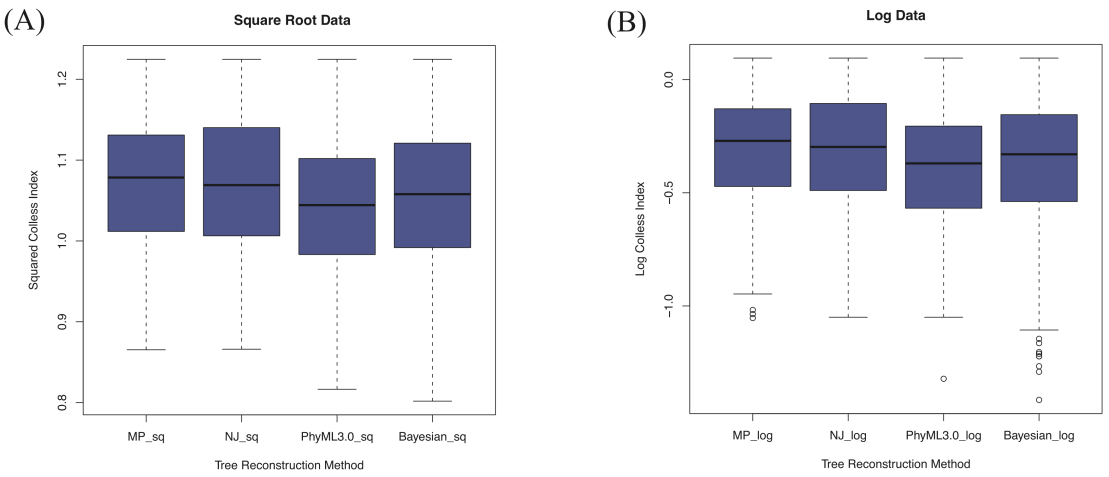


Figure A1. The square root (A) and log transformations (B) of the mean Colless values featuring Bayesian consensus trees

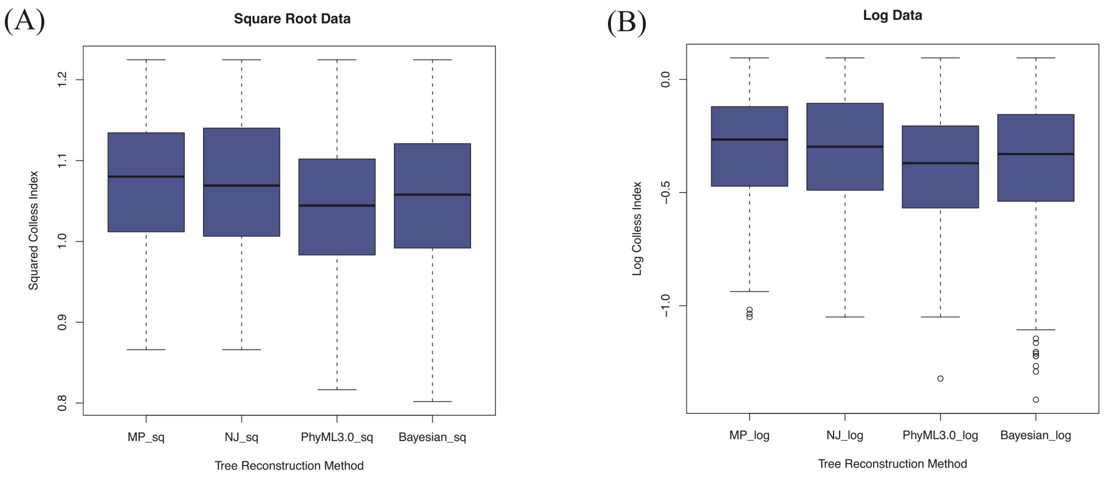


Figure A2. The square root (A) and log transformations (B) of the median Colless values featuring Bayesian consensus trees

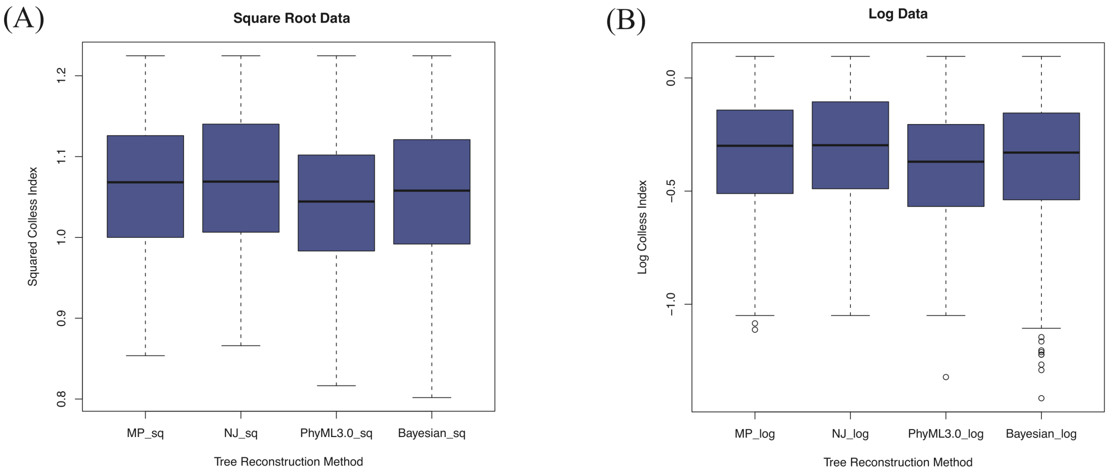


Figure A3. The square root (A) and log transformations (B) of the mode Colless values featuring Bayesian consensus trees

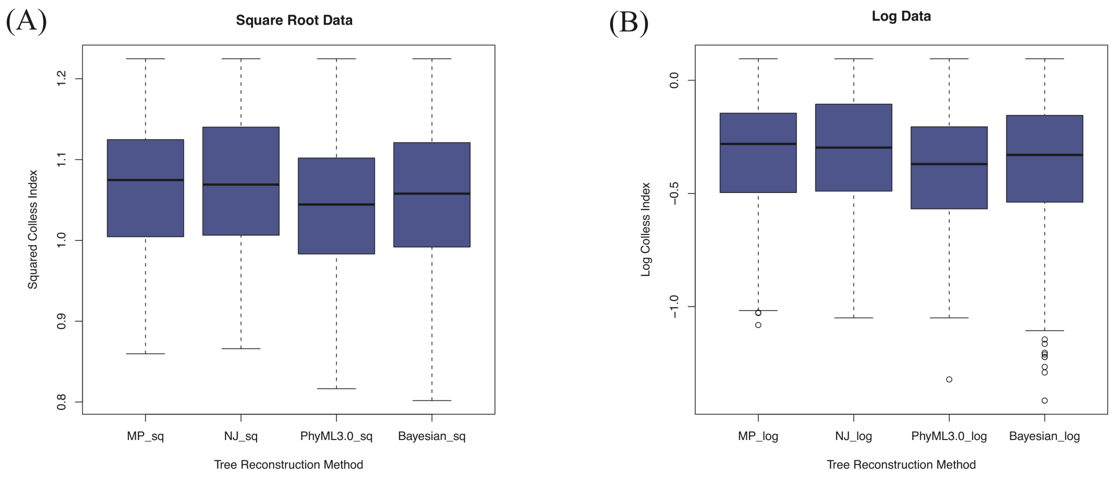


Figure A4. The square root (A) and log transformations (B) of the random Colless values featuring Bayesian consensus trees

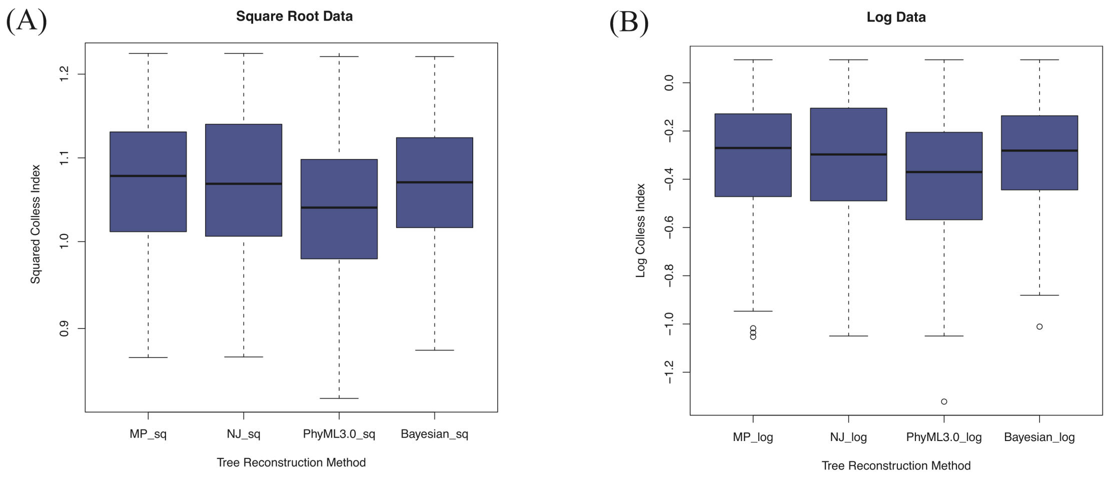


Figure A5. The square root (A) and log transformations (B) of the mean Colless values featuring Bayesian sample trees

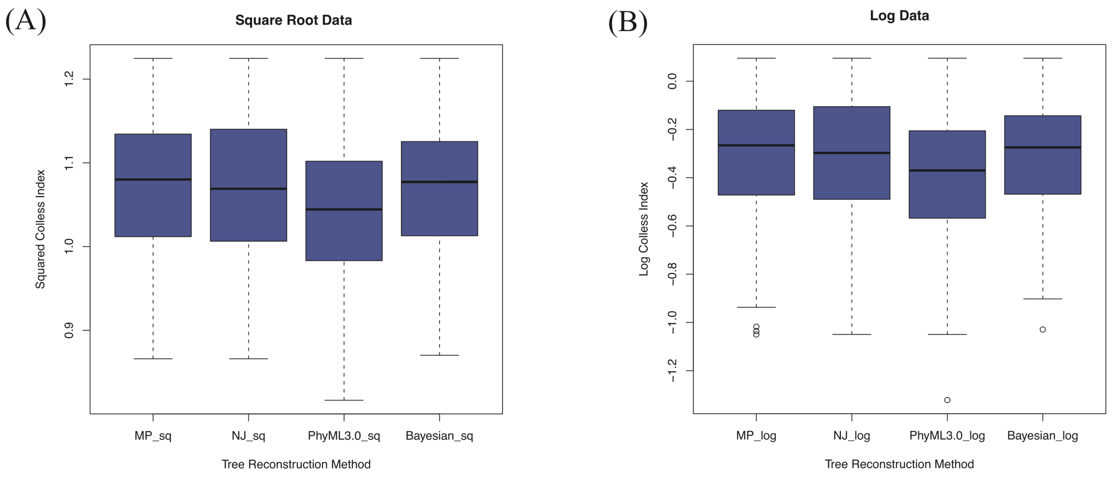


Figure A6. The square root (A) and log transformations (B) of the median Colless values featuring Bayesian sample trees

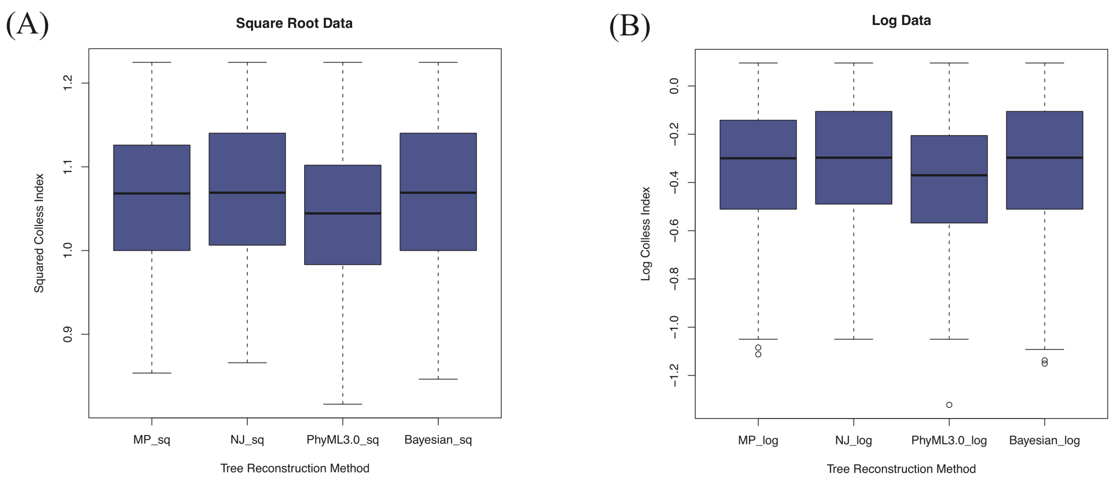


Figure A7. The square root (A) and log transformations (B) of the mode Colless values featuring Bayesian sample trees

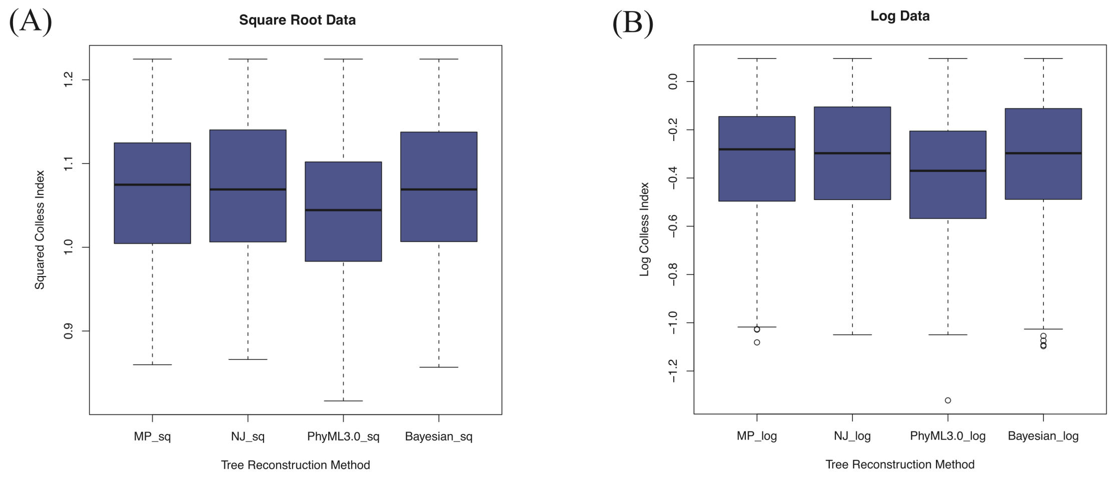


Figure A8. The square root (A) and log transformations (B) of the random Colless values featuring Bayesian sample trees

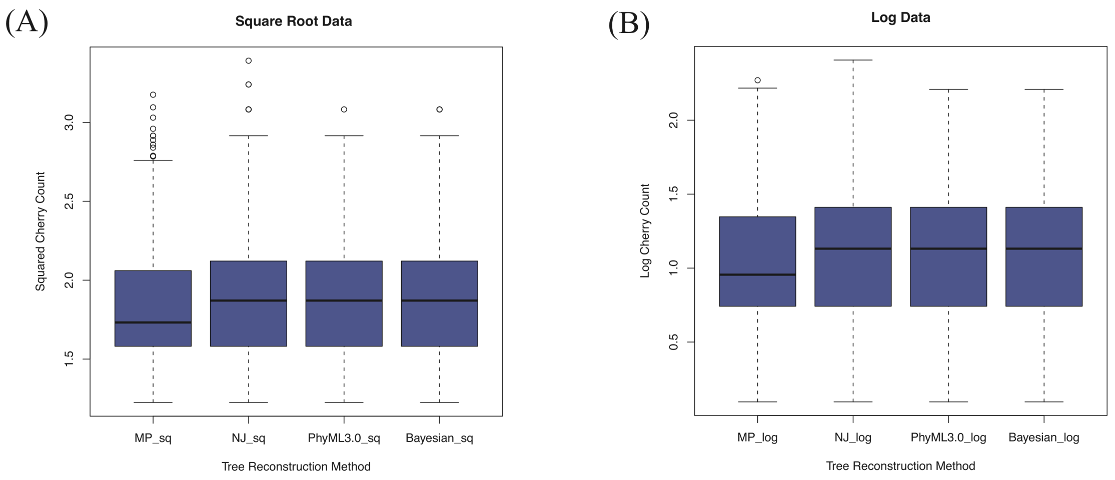


Figure A9. The square root (A) and log transformations (B) of the mean cherry count values featuring Bayesian consensus trees

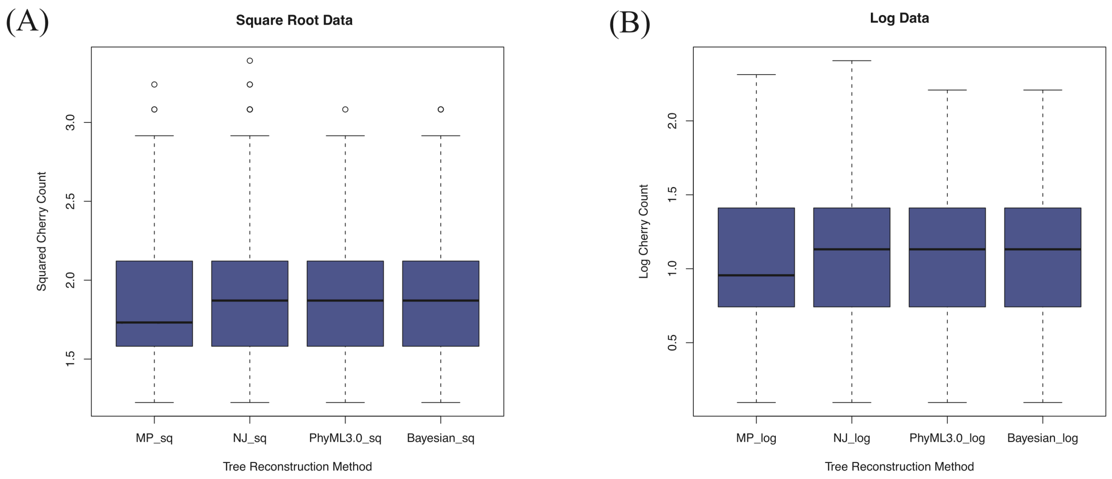


Figure A10. The square root (A) and log transformations (B) of the median cherry count values featuring Bayesian consensus trees

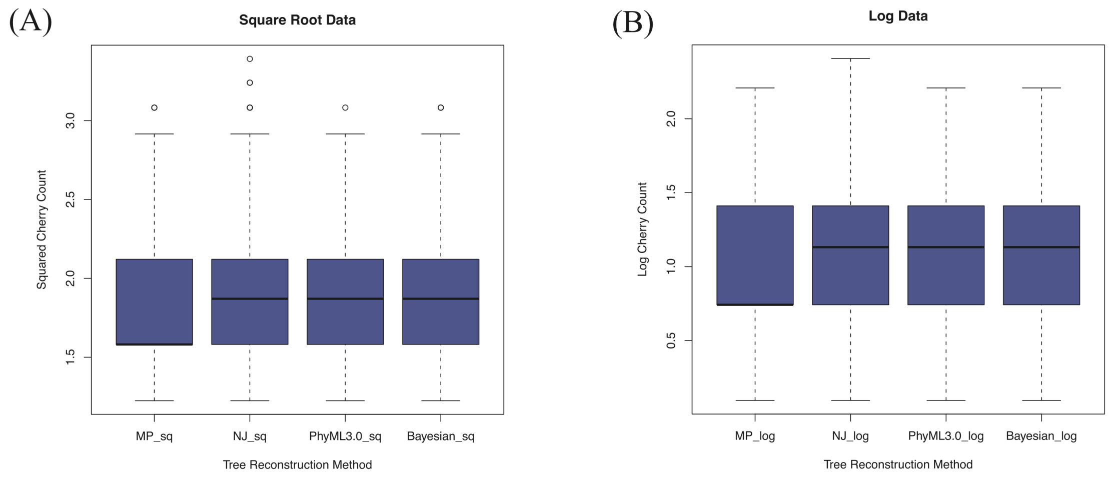


Figure A11. The square root (A) and log transformations (B) of the mode cherry count values featuring Bayesian consensus trees

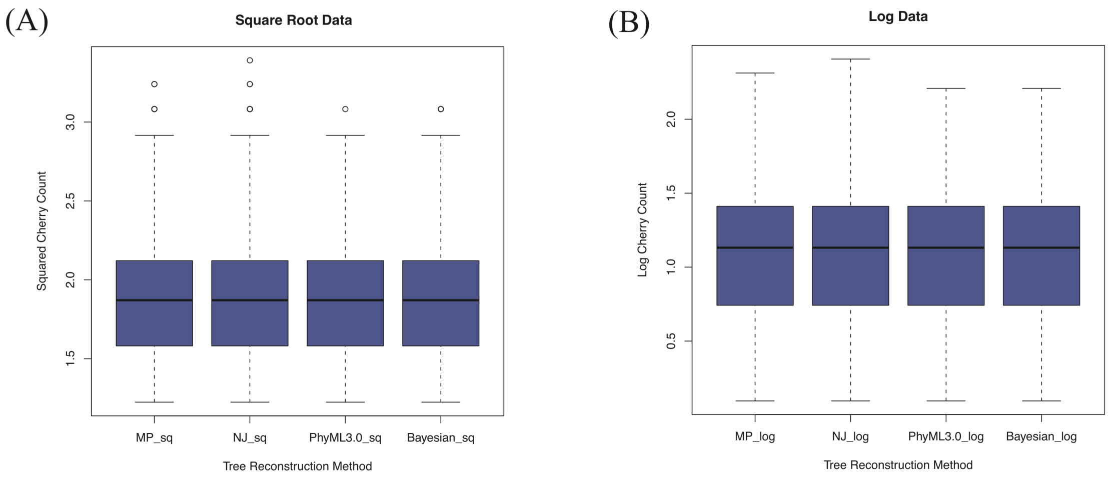


Figure A12. The square root (A) and log transformations (B) of the random cherry count values featuring Bayesian consensus trees

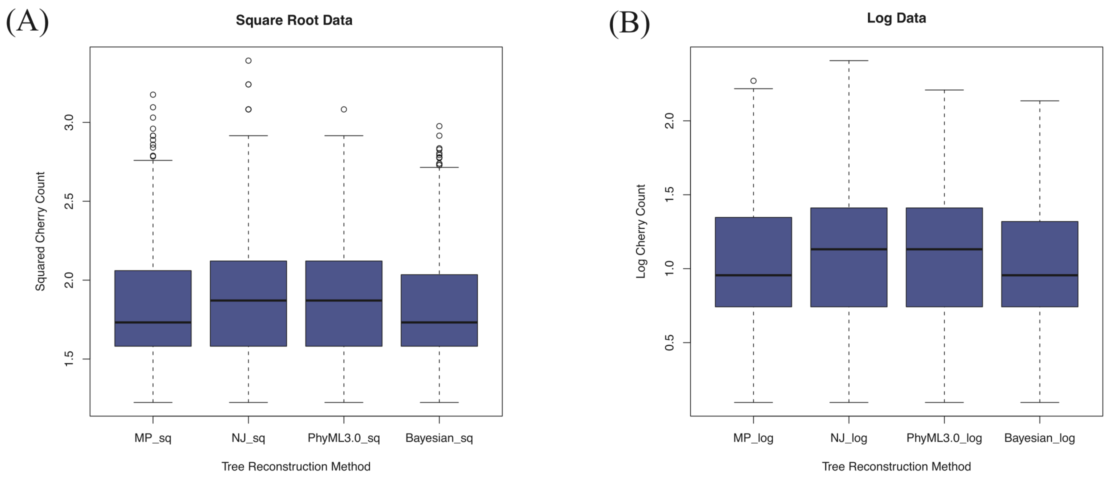


Figure A13. The square root (A) and log transformations (B) of the mean cherry count values featuring Bayesian sample trees

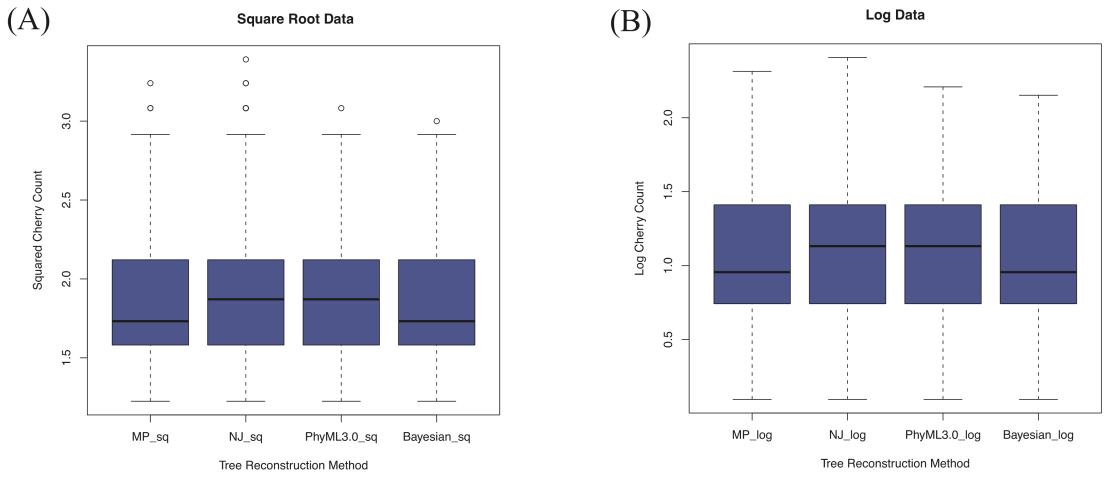


Figure A14. The square root (A) and log transformations (B) of the median cherry count values featuring Bayesian sample trees

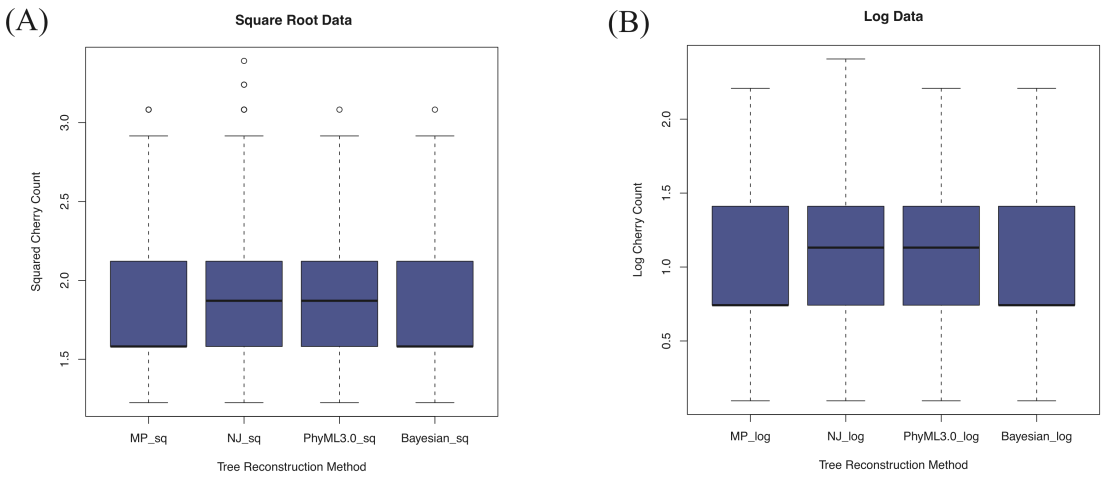


Figure A15. The square root (A) and log transformations (B) of the mode cherry count values featuring Bayesian sample trees

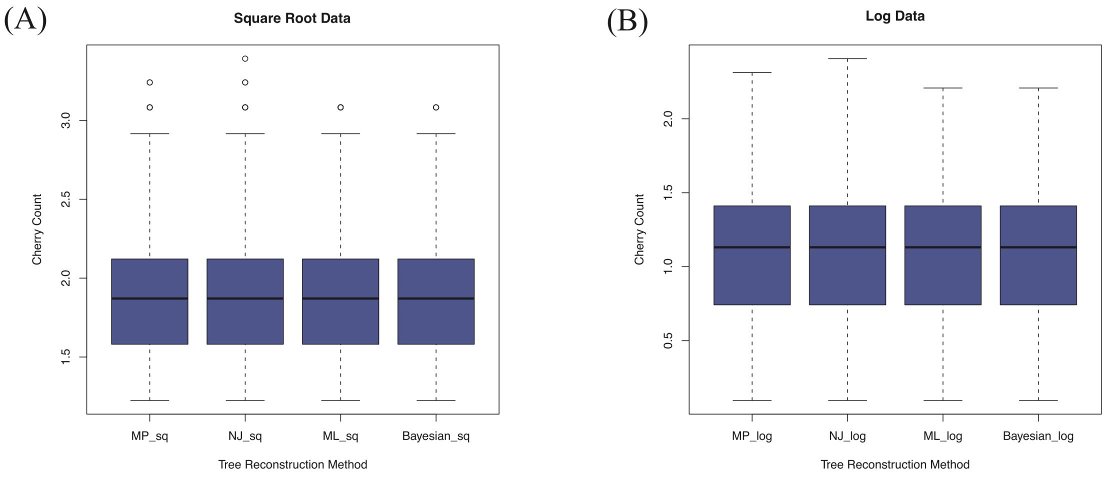


Figure A16. The square root (A) and log transformations (B) of the random cherry count values featuring Bayesian sample trees

Publication

Genome Biology and Evolution

SMBE

Deep Genomic-Scale Analyses of the Metazoa Reject Coelomata: Evidence from Single- and Multigene Families Analyzed Under a Supertree and Supermatrix Paradigm

Thérèse A. Holton, and Davide Pisani*

Department of Biology, National University of Ireland, Maynooth, Maynooth Co. Kildare, Ireland

*Corresponding author: E-mail: davide.pisani@nuim.ie.

Accepted: 26 April 2010

Abstract

Solving the phylogeny of the animals with bilateral symmetry has proven difficult. Morphological studies have suggested a variety of alternative hypotheses, of which, Hyman's Coelomata hypothesis has become the most established. Studies based on 18S rRNA have failed to endorse Coelomata, supporting instead the rearrangement of the protostomes into two new clades: the Lophotrochozoa (including, e.g., the molluscs and the annelids) and the Ecdysozoa (including the Panarthropoda and most pseudocoelomates, such as the nematodes and priapulids). Support for this new animal phylogeny has been attained from expressed sequence tag studies, although these generally have a limited gene sampling. In contrast, deep genomic-scale analyses have often supported Coelomata. However, these studies are problematic due to their limited taxonomic sampling, which could exacerbate tree reconstruction artifacts.

Here, we address both of these sampling limitations; we study the effect of long-branch attraction (LBA) in deep genomic-scale analyses and provide convincing evidence, using both single- and multigene families, that Coelomata is an artifact. We show that optimal outgroup selection is key in avoiding LBA and identify the use of inadequate outgroups as the reason previous deep genomic-scale analyses found strong support for Coelomata.

Key words: Coelomata, Ecdysozoa, phylogenomics, supertrees, outgroup selection, Bayes factors, supermatrix.

Introduction

Bilaterian Phylogenetics Uncertainty still persists pertaining to the early evolution of the Bilateria; an important group which includes all extant animals with the exclusion of the sponges, the Placozoa, the Cnidaria, and the Ctenophora (see, e.g., Nielsen 2001; Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009; Sperling et al. 2009). Central to this incertitude are the phylogenetic relationships of the "pseudocoelomates" (sensu Hyman 1940), particularly that of the Nematoda (i.e., the round worms), which remain an issue of debate (Telford et al. 2008).

From a morphological point of view, some of the most prominent features shared by the majority of bilaterians include bilateral symmetry, a pronounced anteroposterior axis and a head with a nervous concentration, that is, a brain (Nielsen 2001). A variety of morphological phylogenies of Bilateria have been proposed since Darwin's time (Jenner and Schram 1999); however, the dominant view has long been that of Hyman (1940) and her Coelomata hypothesis (see also Halanych 2004; Philippe et al. 2005; Telford et al. 2008). According to Coelomata, Bilateria were classified in three groups: the Acoelomata (Platyhelminthes and Nemertinea), the Pseudocoelomata (Nematoda, Nematomorpha, Rotifera, Priapulida, Kinorhyncha, and Gastrotricha), and the Coelomata (all the other bilaterian phyla, e.g., the Arthropoda, the Mollusca, the Annelida, and the Vertebrata).

The first major challenge to Coelomata came from the analyses of taxon-rich 18S rRNA data sets (Halanych et al. 1995; Aguinaldo et al. 1997), which proposed an alternative division of the Bilateria (with the possible exclusion of the Acoela—see Ruiz-Trillo et al. 1999; Littlewood et al. 2001; Hejnol et al. 2009; but see also Philippe et al. 2007) into the Protostomia and Deuterostomia. The 18S rRNA data further suggested a partitioning of the protostomes into the Lophotrochozoa (Halanych et al. 1995), including, for example, the molluscs and the annelids (i.e., the Eutrochozoa), and the Ecdysozoa (Aguinaldo et al. 1997), including the Panarthropoda and several of Hyman's Pseudocoelomata. This new animal phylogeny is now generally known and will hereafter be referred to as the Ecdysozoa hypothesis.

Ever since the genomes of the arthropod *Drosophila melanogaster* (a coelomate protostome), the vertebrate *Homo sapiens* (a coelomate deuterostome), the nematode *Caenorhabditis elegans* (a pseudocoelomate protostome), and the fungus *Saccharomyces cerevisiae* (a nonmetazoan outgroup) became available, many have attempted to test hypotheses of bilaterian relationships using genomic-scale data sets, or in any event, data sets deemed to be of genomic scale at the time they were assembled (Blair et al. 2002; Copley et al. 2004; Dopazo et al. 2004; Wolf et al. 2004; Dopazo H and Dopazo J 2005; Philip et al. 2005; Rogozin et al. 2007, 2008; Zheng et al. 2007). A number of these studies (Copley et al. 2004; Dopazo H and Dopazo J 2005; Irimia et al. 2007; Roy and Irimia 2008; and Belinky et al. 2010) have endorsed Ecdysozoa, however, only that of Dopazo H and Dopazo J (2005) used standard phylogenetic analyses of aligned sequence data.

The majority of published deep genomic-scale analyses have supported Coelomata, leading Lynch (2007), for example, to conclude a literature survey on this argument by claiming: "... [Ecdysozoa] continues to be presented as a fact in many major textbooks, even though phylogenies based on large numbers of protein-coding genes generally either place nematodes on their traditional position or are equivocal on the matter. . .." Studies supporting Coelomata, however, characteristically suffer from a sparse taxonomic sampling (see also Halanych 2004), which can exacerbate phylogenetic artifacts, particularly long-branch attraction (LBA), in the presence of a fast-evolving species such as *C. elegans* (e.g., Pisani 2004; Delsuc et al. 2005; Philippe et al. 2005; Jeffroy et al. 2006; Sperling et al. 2009).

Studies conducted using the expressed sequence tags (ESTs) methodology (Philippe et al. 2005, 2009; Dunn et al. 2008; Lartillot and Philippe 2008; and Hejnol et al. 2009), on the other hand, are characterized by a denser taxonomic sampling and generally include more appropriate (animal) outgroups and as such should be less prone to LBA. Accordingly, EST-based studies have recurrently supported Ecdysozoa (see Philippe et al. 2005; Lartillot and Philippe 2008 in particular). However, with the exception of Hejnol et al. (2009), who considered 1,487 genes (but only for a very small subset of the taxa they sampled), EST studies represent shallow genomic sampling (Zilversmit et al. 2002), with Philippe et al. (2005) considering only 146 genes, Dunn et al. (2008) 150 genes, and Philippe et al. (2009) 128 genes. Additionally, EST libraries generated for phylogenetic purposes are generally not normalized (e.g., Dunn et al. 2008; Hejnol et al. 2009), and the protein-coding genes sampled in these studies do not represent a random sample of the genes in the considered genomes. Rather, they correspond to a sample of the most highly expressed genes. This nonrandom sampling is not a problem per se, nevertheless, it does pose the question: what will the outstanding proportion of the animal proteome disclose? To date, the answer has often been that standard sequence analyses of deeply sampled genomic data sets favor Coelomata.

Phylogenomics: Methodological Approaches From a methodological point of view, two principal approaches are generally employed in phylogenomics: the supertree and the supermatrix approach (Delsuc et al. 2005), with both approaches having different strengths and weaknesses.

In the supertree approach, gene trees are recovered for each individual protein family using the most appropriate phylogenetic method. Gene trees are then combined using one of a number of existing supertree methods (for a brief introduction, see McInerney et al. 2008). Advantages of the supertree approach include: 1) the ability to analyze each gene individually under the best-fitting substitution model, 2) the capacity to amalgamate trees derived from the analysis of both single- and multigene families, and 3) a significant decrease in the computational time necessary to build large phylogenies (facilitating the handling of data sets scoring thousands of genes) for hundreds of taxa (e.g., Pisani et al. 2007). As gene families are first analyzed in isolation, the major limitation of the supertree approach is that the combined trees can be based on relatively small alignments. This can result in significant statistical errors, which may translate into poorly supported phylogenomic supertrees. Filtering strategies, that is, eliminating genes that do not pass the permutation tail probability (PTP) test (Archie 1989) or that do not support the monophyly of universally accepted clades (Pisani et al. 2007), which also serves to alleviate the negative impact of hidden paralogy when analyzing sets of single-gene families, can be used to improve resolution significantly.

In the supermatrix approach, single-gene alignments are merged into a multiple gene alignment, which is then analyzed using the most appropriate phylogenetic method. The principal merit of this approach is that gene concatenation allows for the minimization of statistical errors, often resulting in well-supported trees (Delsuc et al. 2005). The main shortcomings of this approach are: 1) while it minimizes stochastic errors, it tends to exacerbate systematic ones (e.g., Delsuc et al. 2005; Jeffroy et al. 2006). Although the use of well-performing, parameter-rich models, like categories model (Lartillot and Philippe 2004; Philippe et al. 2007), alleviates this problem, it does not fully eliminate it (e.g., Jeffroy et al. 2006). 2) The supermatrix approach does not lend itself to the integration of multigene families and as such limits the information that can be analyzed to that of single-gene families or in some rare cases (i.e., when the gene phylogeny is well understood) to single paralogy groups within a multigene family (e.g., Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009). 3) If the number of considered genes, species, or both is considerably large, supermatrix analyses become very difficult to perform due to memory and time constraints (see, e.g., Hejnol et al. 2009). Technological advances should ameliorate this problem, but this limit of the supermatrix approach can be expected to persist for the foreseeable future.

Circumventing LBA LBA (Felsenstein 1978) is a common phylogenetic artifact (see Brinkmann and Philippe 1999; Pisani 2004; Delsuc et al. 2005; Jeffroy et al. 2006), which can affect every phylogenetic method (Pisani 2004; Delsuc et al. 2005; Jeffroy et al. 2006). Because time and rate are confounded in branch length estimation (e.g., Yang 2006), LBA results in trees in which fast-evolving species are artifactually grouped together or with distantly related taxa (e.g., with the outgroups). Two straightforward approaches to reduce LBA are optimal outgroup selection (to minimize root to tip distances in a phylogeny) and increased taxon sampling (to break long branches), see also Pisani (2004).

Early, deep genomic-scale analyses used fungal outgroups or on occasion even more distantly related outgroups (e.g., Blair et al. 2002). These clearly represent poor choices to investigate the phylogenetic relationships of the Bilateria as they may serve to exacerbate LBA. Dopazo H and Dopazo J (2005) performed standard sequence analyses of a deeply sampled genomic data set using a distant (fungal) outgroup. Realizing that a fungal outgroup might not have been adequate for their analyses, and in the absence of a closer outgroup, these authors used a relative-rate test (for an overview, see Robinson et al. 1998) based approach to identify clock-like genes. Analyses of these genes found support for Ecdysozoa. Although their results are interesting, their approach is not without problems. First, the relative-rate test is not particularly sensitive; a more discriminating approach (i.e., the likelihood ratio test) should have been used instead. In addition, their relative-rate tests were implemented under the simplistic Kimura's distance in PROTDIST (Felsenstein 2005), which is unlikely to have fit their data well. Finally, these authors considered only homologues of protein-coding genes found in 18 human chromosomes, unnecessarily discarding potentially informative genes not found in this subset of human chromosomes.

The number of complete animal genomes has now increased significantly making the improvement of taxonomic sampling in genomic-scale phylogenetic analyses possible. Recent genome sequencing projects have included that of the cnidarian *Nematostella vectensis* and the placozoan *Trichoplax adherans.* Although there is ongoing debate over the phylogenetic relationships of these organisms, there is general agreement that both are nonbilaterian Metazoans (see Dunn et al. 2008; Hejnol et al. 2009; Philippe et al. 2009; Sperling et al. 2009). Accordingly, *N. vectensis* and *T. adhaerens* represent more appropriate outgroups for testing hypotheses of bilaterian evolution than fungi (see also Philippe et al. 2005). We thus avoided gene selection strategies (e.g., Dopazo H and Dopazo J 2005), focusing instead on taxonomic sampling and outgroup selection to test hypotheses of bilaterian evolution.

Maximizing Gene Sampling within a Phylogenomic Approach The strongest test of a phylogenetic hypothesis is one considering all the relevant information (e.g., Kluge 1989). In phylogenomics, EST studies can maximize taxonomic sampling, whereas studies using complete genomes can maximize gene sampling. Accordingly, a pragmatic solution to the Coelomata versus Ecdysozoa controversy can only be achieved through the congruence of taxonomically well-sampled EST studies and deep genomic-scale analyses.

Here, we performed analyses to maximize gene sampling. We implemented a pluralist approach where phylogenomic trees of Bilateria were generated using supertrees and consensus trees, summarizing both single- and multigene family trees. Because supertrees do not allow for the integration of the subsignals in the data (Pisani and Wilkinson 2002), we augmented our study to include a supermatrix approach, where single-gene families were concatenated and concomitantly analyzed. This was done to confirm the results from the supertree analyses and to provide a statistical test, within a Bayesian framework, of the fit of the considered hypotheses (i.e., Coelomata and Ecdysozoa) to the data.

An experimental approach was used to investigate the support for the considered alternative hypotheses in the light of LBA and to reject the one most likely to be artifactual. In particular, the effect of using fungi, nonbilaterian animals, or both, in order to break long branches, was examined. By comparing our results with those of previous EST studies, we evaluate the congruence between different phylogenomic approaches.

Materials and Methods

Data Collection Genomic data for 43 eukaryotic species were downloaded from COGENT (http://maine.ebi.ac.uk:8000/services/cogent/), DOE Joint Genome Institute (http://genome.jgi-psf.org/), EMBL-EBI IPI (http://www.ebi.ac.uk/IPI/IPIhelp.html), Ensembl (http://www.ensembl.org/info/data/ftp/index.html), and National Center for Biotechnology Information (ftp://ftp.ncbi.nih.gov/genomes/).

Experimental Phylogenomics and Data Set Assembly Rather than simply collecting all available animal genomes and reconstructing yet another metazoan phylogeny, we took an experimental approach. We made the following ad hoc (working) assumption: Coelomata is the true tree and not the result of LBA (our null hypothesis). We predicted what the consequences of this null hypothesis would be, selected a suitable set of complete genomes, and tested whether the predictions derived from our assumption could be met. If our predictions were to be upheld by the data, the null hypothesis was not to be rejected, whereas if overturned, the data would reject the null hypothesis.

Based on our assumption, we first predicted that in sparsely sampled (four taxon) data sets, Coelomata should invariably be recovered irrespective of whether a distant (fungal) or closer (animal) outgroup was used. Conversely, we anticipated that if Coelomata was due to a LBA artifact, then it would only be recovered when using a distant outgroup. We further hypothesized (based again on the postulation that Coelomata is the "bona fide" tree) that Coelomata should continue to be recovered in the presence of an extensive taxonomic sampling, irrespective of the outgroup used. Alternatively, if Coelomata was the result of LBA, we would expect that it should not be recovered if a targeted sampling strategy was adopted to break the long branch connecting the distant (in our case fungal) outgroup and the Bilateria. This could be done by including *N. vectensis* and/or *T. adhaerens* in the analyses or by replacing the fungal outgroups with animal outgroups (i.e., *N. vectensis* and/or *T. adhaerens*).

We assembled (from our starting set of 43 genomes) five intersecting data sets to test our predictions. Two of these data sets contained a minimal sampling, scoring only four taxa. The remaining three data sets included 41, 42, and 43 species, respectively. The four-taxon data sets were designed to mimic the taxonomic sampling of the earliest phylogenomic studies, whereas the 41, 42, and 43 taxon data sets were constructed to contain the broadest possible sampling of complete animal genomes (for a list of the species in each of the five data sets, see supplementary table S1, Supplementary Material online).

The sparsely sampled data sets were used to investigate, at the most fundamental level, the effect of outgroup choice in phylogenomics. Accordingly, these data sets only differed in the outgroup they included, which was either *S. cerevisiae* or *N. vectensis*. In both data sets, the remaining three taxa were *H. sapiens*, *D. melanogaster*, and *C. elegans*. For these sparsely sampled data sets, *N. vectensis* was preferred over *T. adhaerens* as outgroup to the Bilateria, as there is little doubt that cnidarians are closer to the Bilateria (Hejnol et al. 2009; Philippe et al. 2009; Sperling et al. 2009).

Similarly, the three densely sampled data sets scored a common set of 40 bilaterian species (see supplementary table S1, Supplementary Material online), to which one to three outgroups were added. The 41-taxon data set only included *S. cerevisiae* as the outgroup. The 42 species data set contained two animal outgroups (*N. vectensis* and *T. adhaerens*) but did not include *S. cerevisiae*. Finally, the 43 species data set included both the fungal and the animal outgroups (*S. cerevisiae*, *N. vectensis*, and *T. adhaerens*). These densely sampled data sets were used to investigate the effect of using alternative taxon sampling strategies and optimal outgroup selection.

If Coelomata is the correct topology, it should always be recovered in the densely sampled data sets. If Coelomata is a LBA artifact, we expect it to appear only when the fungal outgroup is used in isolation. That is, when the long branch joining the fungi and the Bilateria is present and unbroken. Accordingly, our expectation is that if the data is affected by LBA, Coelomata should be recovered from the 41-taxon data set but not from the 42 and the 43-taxon data sets.

Protein Family Identification For each sparsely and densely sampled data set, homologous sequences were identified and clustered using the BlastP based, all-versus-all approach of Creevey et al. (2004), Fitzpatrick et al. (2006), and Pisani et al. (2007). For the sparsely sampled data sets, protein families were also identified using the markov cluster (MCL)-based algorithm of Enright et al. (2002). Details of how both protein identification strategies were implemented are reported in the Supplementary online information (SI). As a result, a total of seven initial data sets (four sparsely sampled and three densely sampled ones) were used in this study.

For each of these seven data sets, gene families were partitioned into two groups. Families scoring only one member for any given genome (i.e., the putative single-gene families) were separated from those containing multiple members per genome (i.e., the multigene families). Because phylogenetic analyses can only be performed on gene families that score four or more sequences, only single- and multigene families consisting of a minimum of four sequences were retained for further analysis (for a comparison of the number of single- and multigene families in each of the 7 considered data sets, see table 1).

Only single-gene families are typically used for phylogenetic reconstruction (e.g., Pisani et al. 2007; Hejnol et al. 2009). This is to minimize the complexity associated with the analysis of multigene data sets and the inclusion of signals representing the relationships of paralogous genes. However, this approach has the disadvantage of considering only a minority of the genes in the genomes, whereas the strongest test of a phylogenetic hypothesis is one considering all relevant information (e.g., Kluge 1989). Only upon the integration of multigene families can such a test be performed. Here, by exploiting the flexibility of the supertree approach, we have combined both single- and multigene families to generate trees based on the deepest possible sample of genomic data. However, due to the volume of multigene families generated, it was not currently practicable to analyze the multigene families in all seven data sets. Owing to their smaller size, the four 4-taxon data sets were selected as exemplar cases for analysis using both single- and multigene families.

Table 1
Progression of Protein Family Numbers At Each Stage of Analysis

Data Set	Homology Search	Single-Gene Families Number of Families	Families with Four or More Taxa	Families Passing the PTP Test	Families Used in Phylogenetic Analysis	Multigene Families Number of Families	Families with Four or More Taxa	Families Passing the PTP Test	Species-Level Trees with Four or More Taxa	Species-Level Trees Passing the GTP-PTP Test	Species-Level Trees Used in Supertree Analyses
40 genomes and fungal outgroup	Creevey et al. (2004)	82,043	3,241	2,164	2,164	26,165	N/A	N/A	N/A	N/A	N/A
40 genomes and animal outgroups	Creevey et al. (2004)	86,855	3,615	2,216	2,216	25,722	N/A	N/A	N/A	N/A	N/A
40 genomes, fungal, and animal outgroups	Creevey et al. (2004)	88,858	3,304	1,949	1,949	27,895	N/A	N/A	N/A	N/A	N/A
3 genomes and fungal outgroup	Creevey et al. (2004)	16,780	201	30	30	7,947	4,197	3,301	917	258	258
	MCL	6,588	254	28	28	8,529	5,312	4,143	1,366	392	392
3 genomes and animal outgroups	Creevey et al. (2004)	18,094	314	48	48	10,146	5,269	4,328	1,923	516	516
	MCL	6,254	332	40	40	9,808	6,561	4,666	2,319	682	682

Note.—na, not applicable; All data sets and protein families were subjected to the same protocol. GTP-PTP test (see text).

Alignment, Curation, and Identification of Gene Families Conveying Significant Hierarchical Signal All considered single- and multigene families were aligned using ClustalW (Thompson et al. 1994). As the accuracy of traditional multiple sequence alignment software has been questioned (e.g., Löytynoja and Goldman 2008), the single- and multigene families in our four-taxon data sets were also aligned using PRANK (Löytynoja and Goldman 2008). This was done to investigate whether alignment-dependent biases (Löytynoja and Goldman 2008) influenced our results. This experiment was limited to our four-taxon data sets as aligning sequences using PRANK is computationally expensive.

Due to the number of protein families obtained from our data sets, manual curation of alignments was unfeasible. Gblocks (Castresana 2000) was thus used to eliminate highly variable and potentially misaligned regions. Gblocks parameters were set as follows: gapped positions were not eliminated, the minimum block length was set to eight amino acid positions, whereas the maximum number of permitted consecutive nonconserved positions was set to 15 (see also Pisani et al. 2007). Curated alignments were subjected to the PTP test (Archie 1989). This allowed the identification of families conveying significant hierarchical signal (see Pisani et al. 2007). Such families were considered to contain sufficient hierarchical structure to be deemed phylogenetically informative. The PTP test was implemented in PAUP4.0b10 (Swofford 1998). Settings were as follows: 2,000 permutations with heuristic search with one random addition sequence and the MulTrees option set to off. For the PTP test, a probability value $P \leq 0.05$ was considered significant. Alignments not passing the PTP test ($P \geq 0.05$) were disregarded, as they would not contribute anything except noise to the analyses.

Model Selection and Phylogenetic Analysis PHYML (Guindon and Gascuel 2003) was used to perform maximum likelihood (ML) phylogenetic analyses of each alignment passing the PTP test. ML analyses were performed under the best-fitting substitution model, as inferred using the Akaike information criterion in Modelgenerator (Keane et al. 2006). For each single- and multigene family tree, support was evaluated using bootstrap (100) replicates.

Single-gene trees were manually inspected to evaluate possible instances of hidden paralogy; trees that failed to recover the monophyly of uncontroversial, universally accepted groups (e.g., Vertebrata or Arthropoda) were excluded from further analyses (see also Pisani et al. 2007).

Supertree and Consensus Tree Analyses Supertrees represent a generalization of the consensus tree problem in the case of partially overlapping trees (Semple and Steel 2003). Both consensus and supertree methods were used to derive phylogenomic supertrees representing the relationships

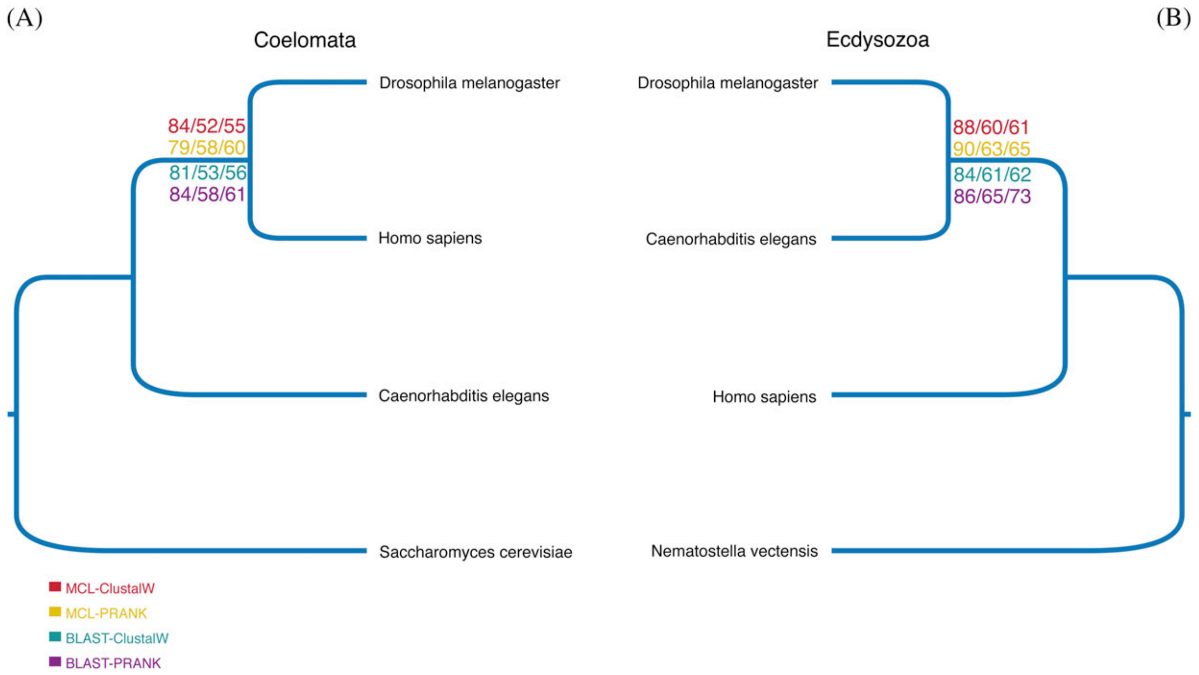


Fig. 1.—Testing outgroup choice in minimally sampled data sets. Majority rule consensus trees derived from ML gene trees. Bootstrap support from both multigene families and single-gene families is shown for each node. The following core ingroup species are common to all: *Homo sapiens*, *Drosophila melanogaster*, and *Caenorhabditis elegans*. Outgroups used are (*A*) the yeast *Saccharomyces cerevisiae* (*B*) the cnidarian *Nematostella vectensis*. Bootstrap support values are shown for each combination of protein family identification and alignment method. Bootstrap support is displayed for single-gene families, multigene families, and combined single-gene families and multigene families, respectively.

among the species in our seven starting data sets. The number of gene families considered at each stage of the protocol used to generate the supertrees is reported in table 1.

Deriving Phylogenomic Supertrees for the Four-Taxon Data Sets For each of the final four-taxon data sets (see fig. 1; eight in total arising from alternative homology assessment and alignment procedures), we derived phylogenomic consensus trees. These were built using 1) the set of all the single-gene families, 2) the set of all the multigene families, and 3) the combined set of all single- and multigene families. Accordingly, a total of 24, four-taxon, phylogenomic trees were derived. Table 1 reports the number of genes used to build each of these trees.

Each of the eight single-gene family based, four-taxon phylogenomic trees (see fig. 1) were built as follows: 1) the 100 bootstrap ML trees generated for each single-gene family in that data set were pooled to generate a single bootstrap tree file. 2) The trees in the pooled, bootstrap tree file were summarized using the majority rule consensus tree method (Margush and McMorris 1981), as implemented in the software Consense (Felsenstein 2005). This was possible as all considered bootstrap trees were on the same taxon set (i.e., they were fully overlapping). As these phylogenomic trees were derived from pooling trees obtained from the individual bootstrap replicates, assessment of the support for the clades in these trees was straightforward because the four-taxon phylogenomic trees were also bootstrap consensus trees.

Each of the eight multigene family-based phylogenomic trees (see fig. 1) were derived as follows: 1) for each considered multigene family, the 100 bootstrap ML trees were used to generate reconciled species trees. This was done using the duplication only, gene tree parsimony (GTP) method (e.g., Cotton and Page 2004) as implemented in the software DupTree (Wehe et al. 2008), with the nogenetree option turned on, using a partial queue based heuristic search (see supplementary fig. SI1, Supplementary Material online for an exemplar multigene family and the corresponding GTP-derived species tree). 2) The resulting species trees (one per bootstrap ML tree) were pooled into a single file. 3) The pooled, bootstrap (species) trees were summarized using the majority rule consensus method (as implemented in the software Consense), thus generating a bootstrap consensus phylogenomic tree. Also, in this case, the use of the majority rule consensus method could be implemented, as all the bootstrap species trees were on the same taxa set.

Each of the eight combined multigene family and single-gene family phylogenomic trees (see fig. 1) were derived as follows: the corresponding sets of individual bootstrap trees (obtained from the ML analyses of the single-gene families) and the species trees derived from the DupTree analysis of the bootstrap trees from the multigene families (see above) were pooled into a single file. Trees in the pooled file were summarized using the majority rule consensus method to derive a bootstrap consensus phylogenomic tree.

The GTP-PTP Test Not all of our multigene families were used for phylogenetic reconstruction (i.e., some families, despite passing the PTP test, were deemed not viable). An additional PTP test was developed to evaluate whether the duplication history of each considered multigene family was phylogenetically informative. To implement the GTP-PTP test, for each optimal multigene family tree derived using PHYML, 100 permuted trees were generated. This was done by randomly swapping the labels associated with the terminal nodes of the optimal multigene family tree, whereas maintaining the unlabeled phylogenetic history as fixed. This is similar to the YAPTP test of Creevey et al. (2004). Each permuted tree was used to infer a species phylogeny using the GTP method (as implemented in DupTree). The score of each GTP reconstruction was recorded, and these values were compared against the GTP score of the species history derived from the original (unpermuted) multigene family tree. Families were retained for phylogenetic analysis when the species history derived from the unpermuted tree was significantly better than those obtained from the GTP analysis of the permuted trees. For these analyses, the significance level was set to $P \leq 0.01$. PERL scripts to implement the GTP-PTP are available upon request.

The species phylogeny embedded in multigene families failing to pass the GTP-PTP test has essentially been erased due to a complex gene deletion/duplication history. These multigene families can only contribute noise to the analyses and were thus not used for phylogenetic reconstruction.

Deriving Phylogenomic Trees for the 41, 42, and 43-Taxa Data Sets Because genes do not have a universal distribution, in the case of the 41, 42, and 43 species data sets, single-gene families could score in the range of 4–41, 4–42, or 4–43 sequences, respectively. That is, unlike the four-taxon data sets, single-gene family trees in these data sets are partially, rather than fully, overlapping. Accordingly, gene trees derived from protein families identified in these larger data sets could not be summarized using a standard consensus method. Instead a supertree approach was used to derive phylogenomic supertrees for these data sets.

For each of the three densely sampled data sets, consensus supertrees were generated as follows: 1) the bootstrap trees obtained from the ML analysis of each considered single-gene family were pooled into one single data set. 2) Input tree bootstrapping (Creevey et al. 2004; Burleigh et al. 2006; Moore et al. 2006; Pisani et al. 2007) of the pooled trees was used to generate 100 pseudoreplicate data sets. 3) For each pseudoreplicate data set, supertrees were derived using the matrix representation with parsimony (MRP) method (Baum 1992; Ragan 1992). To do so, for each pseudoreplicate data set, a standard MRP matrix was generated using CLANN (Creevy and McInerney 2005). This matrix was then analyzed using maximum parsimony in PAUP (Swofford 1998) to generate the MRP supertrees. For the parsimony analysis, 100 heuristic searches were performed with random sequence addition and tree bisection and reconnection branch swapping. 4) The supertrees derived from the analysis of each pseudoreplicate data set were summarized using the majority rule consensus method, generating a majority rule consensus genomic supertree in which support for the clades recovered was expressed as their bootstrap support.

Supermatrix Analysis For each of the 41, 42, and 43 taxon data sets, a superalignment of the single-gene families that passed the PTP test was generated. However, only families that contained at least one nematode sequence were concatenated. This was done to reduce the dimensions of the superalignment (thus making it more manageable) whereas retaining all the information that could possibly bear on the phylogenetic position of the Nematoda. The three concatenated data sets generated in this way were thus subsamples of our complete data sets and scored: 43392 amino acid positions (41-taxon data set), 38701 amino acid positions (42-taxon data set), and 25857 amino acid positions (43-taxon data set). Because the considered genes are not universally distributed, there was a significant amount of missing data in each alignment.

Phylogenetic analyses of the three data sets were performed in Phylobayes (Lartillot and Philippe 2004) under the CAT + G model. We selected CAT as it has been shown (e.g., Philippe et al. 2007; Sperling et al. 2009) that this model provides a better fit to data in comparison with ordinary general time reversible models (e.g., Whelan and Goldman model [WAG] or mechanistic general time revisible [GTR]). We also tested the use of CAT-GTR but under this model we could not reach convergence.

For each data set, two independent runs were performed. Convergence was tested using the bpcomp program (which is part of phylobayes). Two runs were considered to have converged when the maximum difference in observed bipartitions dropped below 0.2.

BFs: Testing Coelomata and Ecdysozoa in a Bayesian Framework Bayes factors (BFs) are general statistical tools that can be used, within a Bayesian framework, to compare alternative models—for example, the trees representing the relationships for a group of taxa (see Sperling et al. 2009) and evaluate the weight of evidence in favor of one of the compared models (and hence against the alternative one). To calculate BFs for each considered data set, we ran two constrained Bayesian analyses using MrBayes (Ronquist and Huelsenbeck 2003). Each of these analyses could only visit trees compatible with one of the two compared hypotheses (i.e., Ecdysozoa or Coelomata). For each

Table 2
Percentage Bootstrap Support for Each Hypothesis (Coelomata, Ecdysozoa, or the Alternative Topology) Arising from the Analysis of the Sparsely Sampled Data Sets

			Percent Support for Each Hypothesis Under Each Alignment Protocol					
			ClustalW			PRANK		
Data Set	**Homology Search**	**Gene Families**	**Coelomata**	**Ecdysozoa**	**Vertebrata–Nematoda**	**Coelomata**	**Ecdysozoa**	**Vertebrata–Nematoda**
Fungal outgroup	Creevey et al. (2004)	Single	81	9	10	84	6	10
		Multi	53	26	21	58	23	19
		Single + multi	56	24	20	61	20	19
	MCL	Single	84	6	10	79	7	14
		Multi	52	26	21	58	22	20
		Single + multi	55	25	20	60	20	20
Animal outgroup	Creevey et al. (2004)	Single	14	84	2	6	86	8
		Multi	21	61	18	18	65	17
		Single + multi	20	62	17	13	73	14
	MCL	Single	9	88	3	7	90	3
		Multi	21	60	19	19	63	18
		Single + multi	20	61	18	18	65	17

of the two constrained analyses, two runs of one chain were run for 1,000,000 generations (sampling every 100 generations). A burn in of 500,000 generations was considered for all analyses. All analyses were performed under WAG + G. This is not ideal, but we could not perform BF analyses under CAT, as the current Phylobayes output is not suitable for estimating BFs (see also Sperling et al. 2009), while running our analyses under GTR in MrBayes was not feasible because of time limitations.

BFs were calculated in Tracer 1.4.1 (Rambaut and Drummond 2007) using, for each constrained analysis, the trace files from the run of highest harmonic mean. Standard errors around the estimated BF were calculated using the bootstrap (1,000 replicates). BFs were interpreted according to the table of Kass and Raftery (1995).

Results

Four-Taxon Data Sets The four species data sets were analyzed to assess at a very basic level the effect of outgroup selection in phylogenomics. The first interesting result we obtained from these analyses was that only a somewhat diminutive number of single-gene families conveying a significant amount of phylogenetic information could be identified (see table 1). This was not fully unforeseen as the stringency of the PTP test increases as the number of considered species decreases. More families were found when *N. vectensis* was used as an outgroup instead of *S. cerevisiae*; however, the difference was small (from 31 to 48). The number of single-gene families passing the PTP test in the four-taxon data sets did not change significantly when either an alternative homology assignment strategy or alignment software was used (see table 1), suggesting that the small number of single-gene families arising from these analyses does not stem from methodological biases. It merely implies that when only 4 taxa are considered, there are very few, universally distributed single-gene families conveying significant phylogenetic information pertinent to testing hypotheses of bilaterian relationships. The number of multigene families (see table 1) passing all of our quality checks is also quite low but significantly higher than the equivalent number of single-gene families. This was to be expected as there are far more multigene families than single-gene families in the average animal genome. However, interestingly, we noted that although the number of phylogenetically informative multigene families identified if *S. cerevisiae* is used as outgroup is 258 (using the Creevey et al. 2004 homology assessment strategy) or 392 (using MCL), the number of phylogenetically informative multigene families identified when *N. vectensis* is the outgroup is 516 (using the Creevey et al. 2004 homology assessment strategy) or 682 (using MCL), that is approximately twice as many. This strongly implies that using closer outgroups is key to maximizing the amount of phylogenetic information and increasing the signal to noise ratio in phylogenomic data sets.

Phylogenomic trees derived from single-gene families passing the PTP test showed that when *S. cerevisiae* was used as an outgroup, support was found for Coelomata (see fig. 1). This result holds true irrespective of the protein family identification method used and of the alignment software used (see fig. 1 and table 2). When only multigene families are used similar results are found, although there is a significant decrease in the level of support observed (fig. 1 and table 2). Finally, in the phylogenomic, trees obtained when both the single-gene families and the multigene families were considered concurrently the support for Coelomata ranges between 55% and 61% depending on the clustering method and alignment software used (fig. 1 and table 2). This represents a marked decrease in the support for Coelomata. Similar results were obtained in the study of Philippe et al. (2005), although based solely on single-gene families.

When the cnidarian *N. vectensis* is used as an outgroup, Coelomata is no longer recovered. Instead, a nematode–arthropod clade emerges, supported most strongly in the analysis of the single-gene families (bootstrap proportion [BP] = 90%; fig. 1 and table 2). Support for Ecdysoza arising from the analysis of single-gene families and multigene families, both in isolation and when combined, ranges from 60% to 90% (fig. 1 and table 2).

It is important to note that when multigene families are used, we observe a general decrease in support for the nodes in the recovered tree, irrespective of whether a fungal or animal outgroup is used. This suggests that multigene families contain more noise than single-gene families. Or more likely that the approach used to infer species trees from the multigene family trees (i.e., duplication only GTP) is not ideal and cannot completely eliminate the paralogy signal. It is to be expected that the development of more refined methods for inferring species trees from multigene family trees will alleviate this problem in the future.

Analyses of the four-taxon data sets illustrate that when a closer outgroup is used sequence analyses with a deep genomic sampling support Ecdysozoa. Conversely, Coelomata is found only when a distant outgroup is used, thus failing to uphold our predictions. The recovery of Coelomata can be better viewed as inconsistent (i.e., "strongly supported but erroneous" Philippe et al. 2005), arising from the selection of a distant outgroup. In the presence of a distantly related outgroup like *S. cerevisiae* (which probably shared a last common ancestor with the Bilateria one billion years ago; see Peterson et al. 2008; Sperling et al. 2010), the rapidly evolving nematode *C. elegans* is placed at the base of the tree, close to the outgroup. When in its stead, a closer outgroup (*N. vectensis*), which probably shared a last common ancestor with the Bilateria only ≈ 670 MYA (Peterson et al. 2008; Sperling et al. 2010) is used, *C. elegans* emerges as the sister group of the arthropod *D. melanogaster* and thus as an Ecdysozoan. This strongly implies the recovery of Coelomata to be the result of a tree reconstruction artifact.

Densely Sampled Data Sets Although the small data sets demonstrate at the most fundamental level the effects of outgroup selection, they still consider only a scant taxonomic sampling. These analyses allow us to reject our null hypothesis (i.e., Coelomata is the true tree) but only relative to small data sets. To test the validity of these results in a more practicable context, we turned our attention to data sets with a broader taxonomic sampling.

Three experiments were performed. In the first, a data set in which taxon sampling was incremented from 4 to 41 species was used. *Saccharomyces cerevisiae* was maintained as the outgroup, whereas all supplementary taxa included were Bilaterian. That is, no attempt at breaking the putative long branch between the fungi and the Bilateria was made. In the second experiment, a data set sampling 43 taxa was used. This data set was designed to contain the full complement of taxa from the first data set but additionally included *T. adhaerens* and *N. vectensis*. Here *S. cerevisiae*, *T. adhaerens*, and *N. vectensis* were simultaneously used as outgroups for the Bilateria. The branch joining the fungi and the Bilateria was still present, but now it was split into three shorter branches, allowing us to investigate the effect of targeted taxon sampling. Finally, the third data set sampled 42 genomes. All metazoan genomes used to generate the first two data sets were retained, whereas *S. cerevisiae* was removed. Excluding *S. cerevisiae* eliminates the long branch joining the fungi and the Bilateria, thus allowing the investigation of using only nonbilaterian metazoans (*T. adhaerens and N. vectensis*) as outgroups.

The analysis of the data set generated for experiment one resulted in 2,164 single-gene families passing the PTP test. Results of an input tree bootstrapping supertree analysis of the ML bootstrap trees generated for these families is reported in figure 2*A* and shows the placement of the Nematoda as the sister group of all the other Bilateria, that is, 100% support for Coelomata. This tree also displays monophyletic Deuterostomia, Arthropoda and, interestingly, Eutrochozoa. (BP = 98%, 100%, and 100%, respectively). The BF analysis shows that the data fit the Coelomata tree better than the Ecdysozoa tree, thus decisively discriminating against Ecdysozoa: $\log_{10}$-BF = 10.792 (±0.29).

When *S. cerevisiae*, *T. adhaerens*, and the Cnidarian *N. vectensis* were concurrently used as outgroups, we found a total of 1,949 single-gene families that conveyed significant phylogenetic signal (see table 1). When these gene families were used for supertree reconstruction, Ecdysozoa was recovered but with very low support (BS = 43%; see fig. 2*B*). Bilateria finds significant support in this analysis (BP = 99%) and is partitioned into Protostomia and Deuterostomia. Monophyly of the Eumetazoa is also supported (BP = 85%), whereas support for Protostomia is not very high (BP = 60%). Inspection of the partition table for this bootstrap analysis shows that Coelomata is still recovered, albeit with minimal support (BP = 13%). This is suggestive of an enduring LBA effect. LBA is obviously reduced when the additional animal outgroups are included in the analyses to the point where the Ecdysozoa tree is the most commonly recovered in the individual bootstrap replicates. However, the reduction of the LBA effect is not significant enough to completely exclude Coelomata from the set of possible solutions. Interestingly, BFs still favor Coelomata with respect to Ecdysozoa (at the least under WAG + G): $\log_{10}$-BF = 6.67 (±0.59). However, in agreement with the results of the bootstrap analysis, which suggest that the LBA effect was indeed reduced when nonbilaterian animals were

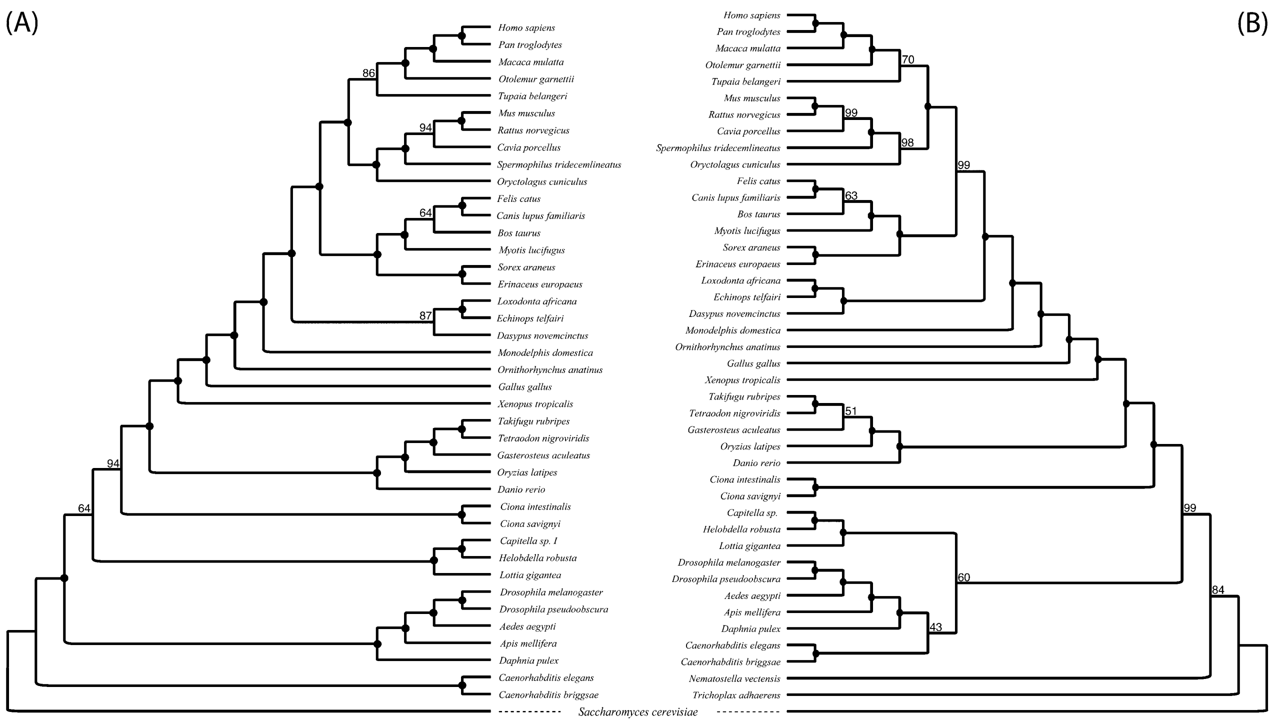


Fig. 2—Phylogenomic supertrees of the Bilateria. (*A*) A tree derived using only the fungal outgroup. This tree is based on 2,164 from 41 species. (*B*) A tree derived using fungal and animal (nonbilaterian) outgroups. This tree is based on 1,949 genes from 43 species. The monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia is recovered in (*B*), whereas (*A*) supports Coelomata. Numbers at the nodes represent bootstrap support. Full circles indicate 100% bootstrap support for a node.

included in the sample, the weight of the evidence in favor of Coelomata is now greatly decreased (by 4.122 points in a $\log_{10}$ scale). That is, when the fungi–Bilateria branch is broken Coelomata is still favored but the data fits the tree ~13,243 times less well than it did when the branch was not interrupted.

In the third experiment, *S. cerevisiae* was interchanged with two animal outgroups (*T. adhaerens* and *N. vectensis*). With this specific taxonomic sampling, we recovered a total of 2,216 single-gene families conveying significant phylogenetic signal. Their analysis recovered a phylogenomic supertree supporting all major, recognized groups (Protostomia, Deuterostomia, Euthrocozoa, and Arthropoda). Additionally this analysis found significant support for Ecdysozoa (BS = 90%) within Protostomia (see fig. 3), with the BF now decisively discriminating against Coelomata: $\log_{10}$-BF = 90.811 (±0.977). If one compares the fit of the Ecdysozoa tree to the data set where *S. cerevisiae* is the only outgroup, with the fit of the same tree to the data set where only the animal outgroups were used, a dramatic change ($\sim 10^{100}$) in the BF in favor of Ecdysozoa is observed. This clearly highlights the major role played by outgroup selection in phylogenomics.

These results are finally confirmed by our supermatrix analyses. In these analyses, when *S. cerevisiae* was used as the only outgroup, convergence could not be reached and the resulting phylogeny (not shown) was nonsensical. When all outgroups were included (fig. 4*A*), Ecdysozoa was recovered, but the effect of LBA was still evident. If one was to root the tree using *N. vectensis* to better pinpoint the LBA effect, a tree essentially consistent with the new animal phylogeny was recovered. However, in this rooted tree, *S. cerevisiae* is incorrectly clustered within Protostomia. If the tree is correctly rooted using *S. cerevisiae* (not shown), the Lophotrochozoa would be incorrectly attracted toward the root. This result, which was somewhat unexpected, is probably a partial consequence of our gene subsampling strategy, in which we maximized information bearing on the relationships of the Nematoda, while ignoring the Lophotrochozoa and the Deuterostomia (see Materials and Methods); however, it is also clearly telling of an enduring LBA effect. Finally when only the animal outgroups are used (fig. 4*B*), the Ecdysozoa tree is recovered. In figure 4*B*, support for the Urochordata as members of the Deuterostomia is not significant, and this group is thus collapsed into a polytomy. We conjecture that this result is also most likely an effect of our gene subsampling strategy (see above). This is confirmed by the supertree analysis of our full data sets in which support for monophyletic Deuterostomia varies between 94% and 100% depending on the outgroup used (see figs. 2 and 3). Notably, a similar effect was observed in the EST study of Hejnol et al. (2009) in which Urochordata became unstable when gene sampling was reduced; see

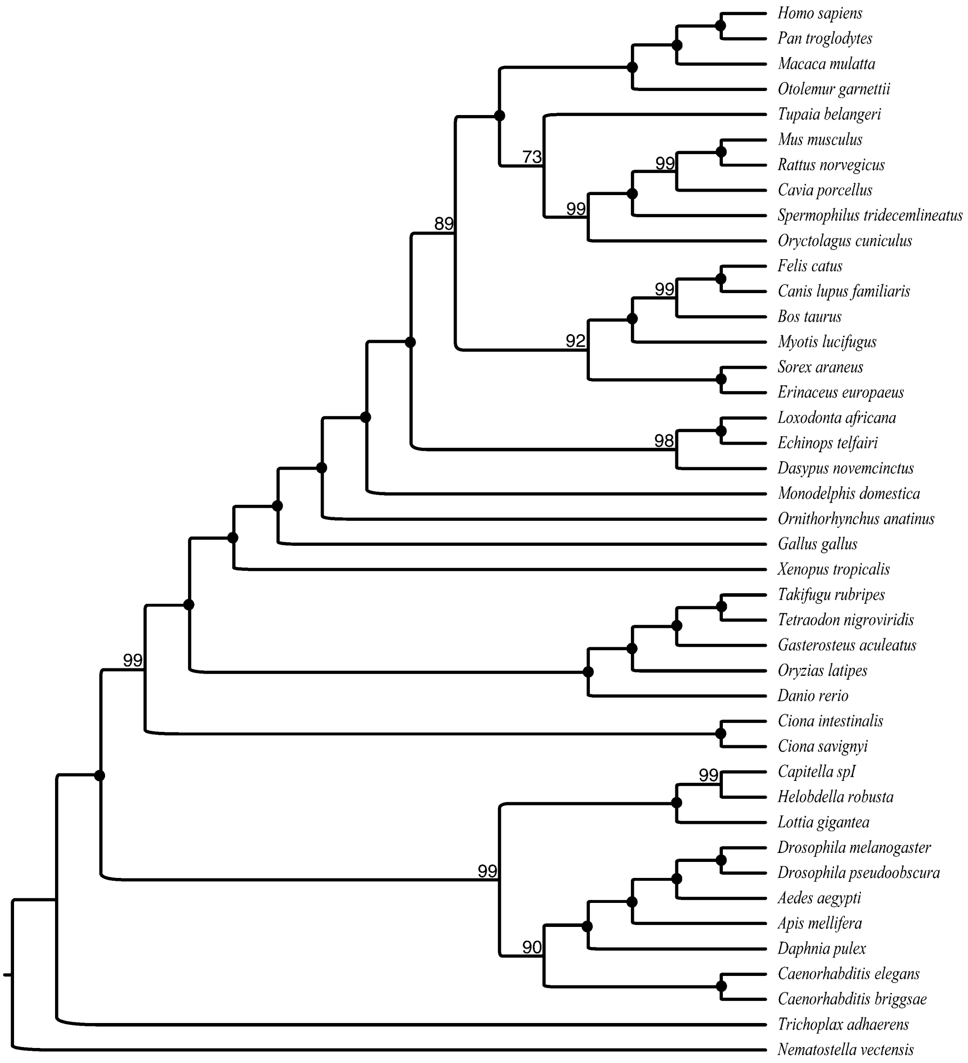


Fig. 3.—Phylogenomic supertree of the Bilateria recovered using only animal (nonbilaterian) outgroups. This tree is based on 2,216 genes from 42 species. High support for the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia can be observed. Numbers at the nodes represent bootstrap support. Full circles indicate 100% bootstrap support for a node.

supplementary figure S1 (Supplementary Material online) Hejnol et al. (2009).

Discussion

Phylogenomics in a Pluralist Context ESTs provide an excellent means of increasing taxon sampling and have been shown to produce highly resolved, well-supported phylogenies (e.g., Philippe et al. 2005, 2009; Dunn et al. 2008; Hejnol et al. 2009). However, EST studies consider only a shallow sampling of genomic content and include a large amount of missing data, the effect of which has never been thoroughly investigated. For Coelomata to be robustly rejected, EST data, although obviously important, cannot be considered sufficient: accord between taxonomically rich EST studies, and gene-rich deep-scale analyses must be reached. With the wealth of genomic data that is currently available, arising from an ever-increasing number of sequencing projects, coupled with advances in sequencing technologies, taxon sampling is becoming less of a limitation for deep genomic-scale phylogenetic analyses. In short, we now have at our disposal the data to conduct extensive, experimental phylogenomic studies of metazoan evolution.

Supertree methods offer an ideal solution for the reconstruction of large-scale phylogenies based upon complete genomes, as they provide a means of overcoming the limits of gene concatenation-based approaches. Gene concatenation methods, at present, do not allow for the easy amalgamation of thousands of genes. Supertrees (and in the four taxon case consensus methods), implementing a divide and conquer strategy, facilitate the analysis of entire genomes for many taxa by coalescing the results of multiple

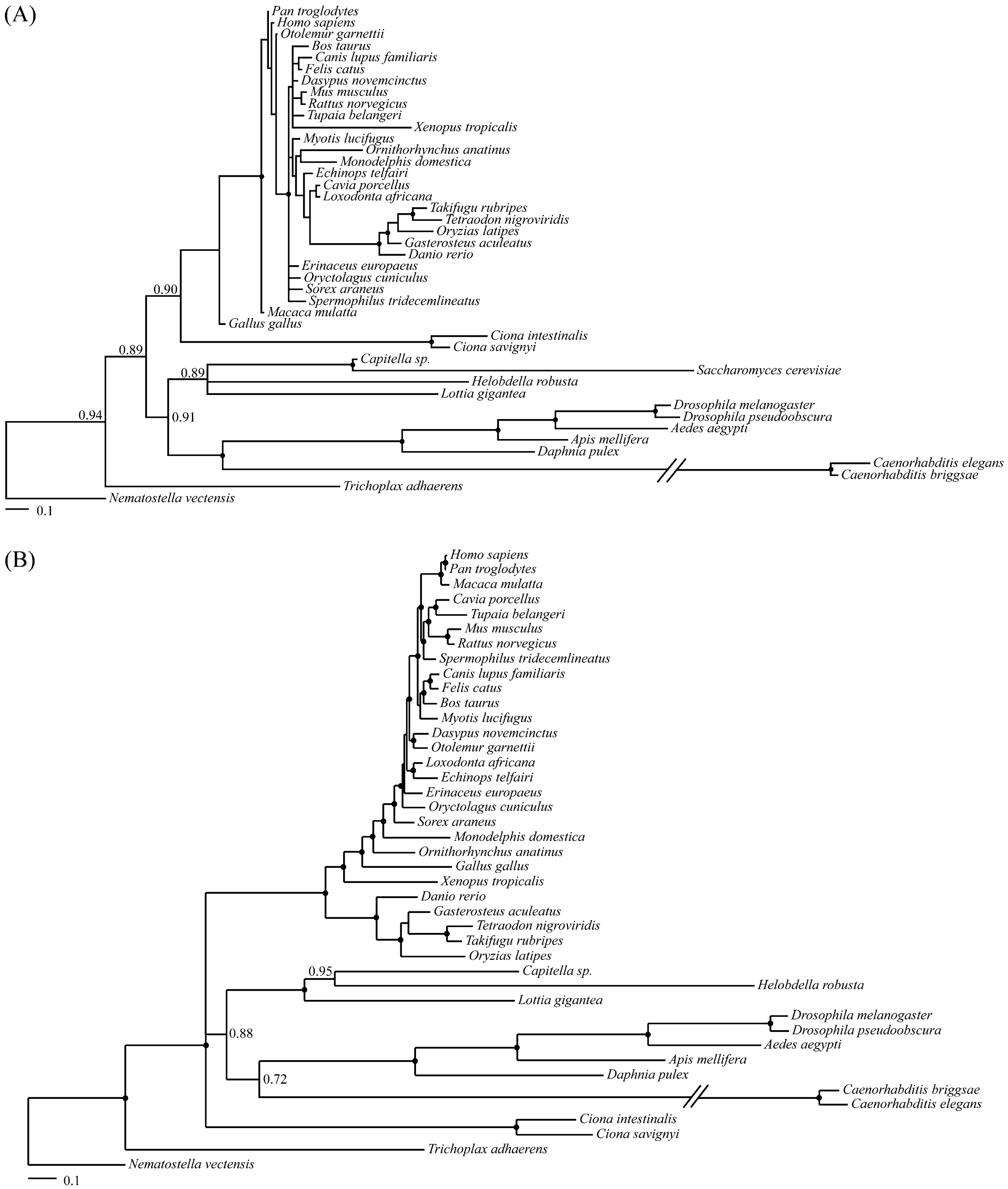


Fig. 4.—Results of the supermatrix analyses. (*A*) The effect of LBA is obvious if one roots the tree using *Nematostella vectensis*, as a tree essentially consistent with the new animal phylogeny is recovered, but *Saccharomyces cerevisiae* is incorrectly nested within the Protostomia. (*B*) A tree illustrating that Ecdysozoa is easily recovered when analyses are performed using only nonbilaterian animals as outgroups. Numbers at the nodes represent posterior probabilities. Full circles indicate a posterior probability of 1. Posterior probabilities lower than 1 have only been reported for nodes that are relevant to the Ecdysozoa versus Coelomata problem. Urochordata is collapsed in a basal polytomy because the posterior probability of Deuterostomia is less than 0.5.

subanalyses to attain a global solution (Wilkinson and Cotton 2006). However, supermatrix approaches also have important advantages, particularly as they overcome the most important limitation of supertrees, that is that the latter do not allow hidden subsignals to interact and thus lack total evidence like properties (Pisani and Wilkinson 2002). In addition, supermatrix approaches allow for the use of statistical tools (like BFs) to test alternative phylogenetic hypotheses. Bearing in mind that both approaches have highly desirable and significantly different properties, we therefore opted for a pluralist, supertree/consensus tree and supermatrix approach in our study.

Our four-taxon analyses show that multigene families can be appropriately treated to derive species phylogenies and suitably included in a consensus tree (if all considered gene families are universally distributed) or supertree (if the gene families are not universally distributed) analyses. In particular, we show that all our consensus supertrees (including those that sample multigene families) continue to support Ecdysozoa, a result that is further confirmed by our supermatrix analyses.

Supertrees have previously been employed to address the phylogenetic position of the nematodes (Philip et al. 2005). Although carefully conducted, using the best methods and data available at that time, this analysis did contain (by the authors' own admission) a very limited sampling of just 10 genomes. In particular, a noticeable problem that Philip et al. (2005) faced was the absence of an adequate outgroup (i.e., nonbilaterian metazoan genomes). As postulated by Philip et al. (2005), in time, an increased sampling could well serve to alter their results. In line with that prediction, our supertree analyses performed using appropriate outgroups and a significantly increased taxon (and gene in the case of the four-taxon data sets) sampling has revealed an alternative topology (see figs. 2*B*, 3, and 4*B*). Our results suggest that the study of Philip et al. (2005) and indeed other genomic-scale analyses (e.g., Blair et al. 2002; Wolf et al. 2004) may have been influenced by systematic errors arising from poor outgroup choice, sparse taxon sampling, and hidden paralogy.

Circumventing Systematic Errors Our study illustrates the importance of outgroup choice in phylogenomic-scale studies. We see that the use of a distant outgroup has a marked effect, irrespective of whether ingroup sampling is spare or dense. We found, like in other studies (Philippe et al. 2005; Rota-Stabelli and Telford 2008), that outgroup choice completely alters the resulting topology, consequently lending analogous support to competing hypothesis. The recovery of the Coelomata topology can be considered a LBA artifact brought about by the use of a divergent outgroup. Comparison of BF values gives an indication of the strength of the bias and of how difficult it is to limit its effects. Our results also reject the contention of Rosenberg and Kumar (2001) and Rokas and Carroll (2005) that poor taxon sampling is irrelevant as long as enough genes are considered.

Our densely sampled data sets illustrate that optimal outgroup selection is more important than targeted taxon sampling in avoiding LBA artifacts. If a distant outgroup (*S. cerevisiae*) is included in the analysis, targeted taxon sampling (i.e., breaking the long Bilateria–fungi branch), does not completely eradicate (as shown most powerfully by the BF analyses) LBA. Only upon the exclusion of *S. cerevisiae* do the BFs show a radical decrease in fit of the Coelomata tree. Optimal outgroup selection is a rarely addressed topic in phylogenetics and phylogenomics, and one has to bear in mind that the optimal outgroup for a given data set is not necessarily the closest one (for an interesting example, see Rota-Stabelli and Telford 2008). Aside from LBA, another important source of phylogenetic artifacts is gene (or amino acid) composition bias, and one should thus try to select outgroups that simultaneously minimize the likelihood of both artifacts occurring.

Stringency and the Selection of Families for Phylogenetic Reconstruction When analyzing a small selection of genomes we could not identify a number of single-gene families comparable with those identified by, for example, Blair et al. (2002). Disparity between our study and that of Blair et al. (2002) is particularly striking when comparing their four-taxon data set to our data set including *S. cerevisiae*. Although the ultimate results of both data sets are congruent, that is, both data sets support Coelomata; our analysis considers 70% less single-gene families than Blair et al. (2002). Failure of these data sets to have correlating numbers of single-gene families merits discussion. We suggest that the observed difference can partially be explained by the use of different outgroups. Blair et al. (2002) primarily used a plant outgroup and only in cases where plant genes were not available was a fungal outgroup used. However, this difference can also be accounted for by the implementation of measures to assess data quality in our study. Under our protocol, a gene family was only considered for phylogenetic analysis if it demonstrated significant clustering signal. Our approach thus ensured that noisy families or families devoid of clustering signal were eliminated from our analysis. It is interesting to note that prior to this filtering stage, the number of single (four-taxon)- gene families identified in our study was twice the number identified by Blair et al. (2002).

Conclusions

The Ecdysozoa hypothesis has accumulated significant support in recent years (Philippe et al. 2005; Irimia et al. 2007; Bourlat et al. 2008; Dunn et al. 2008; Lartillot and Philippe 2008; Telford et al. 2008), particularly from the analyses of

EST data sets. To supplement this amassment of evidence, here, we present support for Ecdysozoa from genomic-scale data sets. From these, overall, Ecdysozoa represents the most cogent hypothesis. It is supported from the analyses of both single-gene families and multigene families, once suitable outgroups are considered. Coelomata, on the other hand, is only supported upon the inclusion of a distantly related outgroup, which suggests that this topology is systematically generated by a LBA artifact.

Our results, based on arguably the deepest gene sampling of the Bilateria to date, present overwhelming support for Ecdysozoa and clearly illustrate that it is the use of a distant outgroup that mislead previous analyses. Taken in combination with results from the aforementioned EST studies, it now appears that all aspects of molecular-based phylogenetics support the rejection of Coelomata. Although lack of unambiguous morphological support for Ecdysozoa persists as a moot point, in the light of overwhelming molecular evidence and lack of morphological evidence conclusively discrediting Ecdysozoa, is it now finally time to shed the notion of Coelomata?

Supplementary Material

Supplementary figure SI1 and table S1 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

This study was supported by an "IRCSET EMBARK Initiative" PhD scholarship awarded to T.A.H., partially supported by a Science Foundation Ireland Research Frontiers Programme grant awarded to D.P. (08-RFP-EOB1595), by the High Performance Computing (HPC) Facility at NUI Maynooth, and by the Irish Centre for High-End Computing (ICHEC). We would like to thank Omar Rota-Stabelli, James O. McInerney, and David Fitzpatrick. We also thank the three anonymous reviewers for their suggestions and criticisms, which significantly helped improving the quality of this paper.

Literature Cited

Aguinaldo AMA, et al. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 387:489–493.

Archie JW. 1989. A randomization test for phylogenetic information in systematic data. Syst Zool. 38:219–225.

Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon. 41:3–10.

Belinky F, Cohen O, Huchon D. 2010. Large-scale parsimony analysis of metazoan indels in protein-coding genes. Mol Biol Evol. 27:441–451.

Blair JE, Ikeo K, Gojobori T, Hedges SB. 2002. The evolutionary position of nematodes. BMC Evol Biol. 2:7.

Bourlat SJ, Nielsen C, Economou AE, Telford MJ. 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. Mol Phylogenet Evol. 49:23–31.

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artefacts in ancient phylogenies. Mol Biol Evol. 16:817–825.

Burleigh JG, Driskell AC, Sanderson MJ. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst Biol. 55:426–440.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Copley RR, Aloy P, Russell RB, Telford MJ. 2004. Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. Evol Dev. 6:164–169.

Cotton JA, Page RDM. 2004. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. In: Bininda-Emonds ORP, editor. Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht (The Netherlands): Kluwer Academic. p. 107–125.

Creevey CJ, et al. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc Biol Sci. 271:2551–2558.

Creevy CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics. 21:390–392.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6:R41.

Dopazo H, Santoyo J, Dopazo J. 2004. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. Bioinformatics. 20:i116–i121.

Dunn CW, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452:745–749.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27:401–410.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.67. Seattle (WA): Department of Genome Sciences, University of Washington.

Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol. 6:99.

Guindon S, Gascuel O. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Halanych KM. 2004. The new view of animal phylogeny. Annu Rev Ecol Evol Sys. 35:229–256.

Halanych KM, et al. 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science. 267:1641–1643.

Hejnol A, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc R Soc B Biol Sci. 276: 4261–4270.

Hyman LH. 1940. The invertebrates. Vol. 1: Protozoa through Ctenophora. New York: McGraw-Hill.

Irimia M, Maeso I, Penny D, Garcia-Fernàndez J, Roy SW. 2007. Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. Mol Biol Evol. 24:1604–1607.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Jenner RA, Schram FR. 1999. The grand game of metazoan phylogeny: rules and strategies. Biol Rev Camb Philos Soc. 74:121–142.

Kass RE, Raftery AE. 1995. Bayes factors. J Am Stat Assoc. 90:773–795.

Keane TM, Creevy CJ, Pentony MM, Naughton TJ, McInerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol. 6:29.

Kluge AG. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst Zool. 38:7–25.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095–1109.

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of the Bilateria. Philos Trans R Soc Lond B. 363:1463–1472.

Littlewood DT, Olsen PD, Telford MJ, Herniou EA, Riutort M. 2001. Elongation factor 1-alpha sequences alone do not assist in resolving the position of the acoela within the metazoa. Mol Biol Evol. 18:437–442.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632–1635.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Margush T, McMorris FR. 1981. Consensus *n*-trees. Bull Math Biol. 43:239–244.

McInerney JO, Cotton JA, Pisani D. 2008. The prokaryotic tree of life: past, present. . . and future? Trends Ecol Evol. 23:276–281.

Moore BR, Smith SA, Donoghue MJ. 2006. Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. Syst Biol. 55:662–676.

Nielsen C. 2001. Animal evolution, interrelationships of the living phyla. Oxford: Oxford University Press.

Peterson KJ, Cotton JA, Gehling JG, Pisani D. 2008. The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos Trans R Soc Lond B Biol Sci. 363:1435–1443.

Philip GK, Creevy CJ, McInerney JO. 2005. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support got the Coelomata than Ecdysozoa. Mol Biol Evol. 22:1175–1184.

Philippe H, Brinkmann H, Martinez P, Riutort M, Baguñà J. 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS One. 2:e717.

Philippe H, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol. 19:706–712.

Philippe H, Lartillot N, Brinkmann H. 2005. Multi gene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol. 22:1246–1253.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst Biol. 53:978–989.

Pisani D, Cotton JA, McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24:1752–1760.

Pisani D, Wilkinson M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. Syst Biol. 51:151–155.

Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol. 1:53–58.

Rambaut A, Drummond AJ. 2007. Tracer v1.4. [Internet]. Available from: http://beast.bio.ed.ac.uk/Tracer.

Robinson M, Gouy M, Gautier C, Mouchiroud D. 1998. Sensitivity of the relative-rate test to taxonomic sampling. Mol Biol Evol. 15:1091–1098.

Rogozin IB, Thomson K, Csürös M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. Biol Direct. 3:7.

Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007. Analysis of rare amino acid replacements supports the Coelomata clade. Mol Biol Evol. 24:2594–2597.

Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol. 22:1337–1344.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc Natl Acad Sci U S A. 98:10751–10756.

Rota-Stabelli O, Telford MJ. 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for the Mandibulata over Myriochelata using mitogenomics. Mol Phylogenet Evol. 48:103–111.

Roy SW, Irimia M. 2008. Rare genomic characters do not support Coelomata: intron loss/gain. Mol Biol Evol. 25:620–623.

Ruiz-Trillo I, Riutort M, Littlewood DTJ, Herniou EA, Baguna J. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science. 283:1919–1923.

Semple C, Steel M. 2003. Phylogenetics. New York: Oxford University Press.

Sperling EA, Peterson KJ, Pisani D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol. 26:2261–2274.

Sperling EA, Robinson JM, Pisani D, Peterson KJ. 2010. Where's the glass? Biomarkers, molecular clocks and microRNAs suggest a 200 million year missing precambrian fossil record of siliceous sponge spicules. Geobiology. 8:24–36.

Swofford DL. 1998. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland (MA): Sinauer Associates Inc.

Telford MJ, Bourlat SJ, Economou A, Papillon D, Rota-Stabelli O. 2008. The evolution of the Ecdysozoa. Philos Trans R Soc Lond B Biol Sci. 363:1529–1537.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Wehe A, Bansal MS, Burleigh JG, Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 24:1540–1541.

Wilkinson M, Cotton JA. 2006. Supertree methods for building the tree of life: divide-and-conquer approaches to large phylogenetic problems. In: Hodkinson T, Parnell J, Waldren S, editors. Towards the tree of life: taxonomy and systematics of large and species rich taxa. Systematic Association special volume. CRC Press. p. 61–75.

Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14:29–36.

Yang Z. 2006. Computational molecular evolution. Oxford series in ecology and evolution. New York: Oxford University Press.

Zheng J, Rogozin IB, Koonin EV, Przytycka TM. 2007. Support for the Coelomata clade of animals from a rigorous analysis of the pattern of intron conservation. Mol Biol Evol. 24:2583–2592.

Zilversmit M, O'Grady P, Desalle R. 2002. Shallow genomics, phylogenetics, and evolution in the family Drosophilidae. Pac Symp Biocomput. 7:512–523.

Associate editor: Takashi Gojobori