João Miguel Rodrigues da Silva Brazão

Applying data-specific substitution models and mitigating the effects of among-lineage heterogeneity to infer better protein-based phylogenies

UAlg FCT
UNIVERSIDADE DO ALGARVE
FACULDADE DE CIÊNCIAS E TECNOLOGIA

2024

João Miguel Rodrigues da Silva Brazão

Applying data-specific substitution models and mitigating the effects of among-lineage heterogeneity to infer better protein-based phylogenies

Doutoramento em Ciências Biológicas
Especialidade em Genética, Genómica e Evolução

Trabalho efetuado sob a orientação de:
Dr. Cymon J. Cox

UAlg FCT
UNIVERSIDADE DO ALGARVE
FACULDADE DE CIÊNCIAS E TECNOLOGIA

2024

DECLARAÇÃO DE AUTORIA DE TRABALHO

Applying data-specific substitution models and mitigating the effects of among-lineage heterogeneity to infer better protein-based phylogenies

Doutoramento em Ciências biológicas
Especialidade em Genética, Genómica e Evolução

Declaro ser o autor deste trabalho que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

This work has not previously been submitted for a degree in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

(João Miguel Rodrigues da Silva Brazão)

Copyright-João Brazão

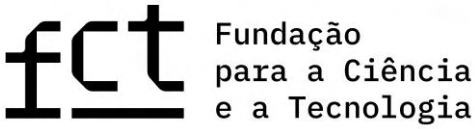
A Universidade do Algarve reserva para si o direito, em conformidade com o disposto no Código do Direito de Autor e dos Direitos Conexos, de arquivar, reproduzir e publicar a obra, independentemente do meio utilizado, bem como de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição para fins meramente educacionais ou de investigação e não comerciais, conquanto seja dado o devido crédito ao autor e editor respetivos.

The University of Algarve reserves the right, in accordance with the provisions of the Code of Copyright and Related Rights, to archive, reproduce and publish the work, regardless of the medium used, as well as disseminating it through scientific repositories and admitting its copying and distribution for purely educational or research purposes and not for commercial purposes, as long as due credit is given to the respective author and publisher.

Apoio

João Miguel Rodrigues da Silva Brazão foi financiado pela Fundação para a Ciência e Tecnologia (FCT) através de uma bolsa de doutoramento (SFRH/BD/134422/2017).

João Miguel Rodrigues da Silva Brazão was funded by the Portuguese Foundation for Science and Technology (FCT) through a PhD grant (SFRH/BD/134422/2017).



Suporte Institucional

Este trabalho não teria sido possível sem o apoio institucional do Centro de Ciências do Mar do Algarve (CCMAR).

This work would not be possible without the support from the Centro de Ciências do Mar do Algarve (CCMAR).

Ao meu pai

Acknowledgements

Terminando esta jornada, não poderia deixar de mostrar o meu reconhecimento àqueles que me acompanharam e estiveram presentes.

Em primeiro lugar um agradecimento ao meu orientador, Dr. Cymon J. Cox, pelo seu apoio para ingressar nesta jornada e pela oportunidade de juntar-me ao 'mundo' da filogenética e evolução. Foram vários os desafios que contribuíram para fortalecer as minhas competências, mas mais relevante foi crescer enquanto investigador. Por isso, agradeço-lhe pelas partilhas de conhecimento, discussões e apoio, as quais foram importantes para concluir esta jornada e serão sempre uma referência para mim.

Quero reconhecer os bons momentos e apoio dos meus colegas do CCMAR, o Bruno, o Gianluca, a Monika, a Christina e todos os meus colegas doutorandos.

Aos meus amigos, um obrigado e a promessa que ainda há muito para vir.

Ao meu irmão e aos meus tios um profundo obrigado por todo o apoio e reconhecimento.

À minha mãe que, enquanto professora de ciências da natureza, terá sido a minha primeira inspiração para escolher ser biólogo. Sem ela, não teria chegado aqui e por isso agradeço do meu coração.

À Ana, a minha companheira de vida, foi um apoio incondicional, visível e invisível. Sempre me acompanhou no entusiamo, mas aparou-me nas descidas. Obrigado por fazeres parte deste percurso e obrigado pelo apoio por todos os momentos que eu não estava.

E por fim, quero também dedicar esta dissertação à minha filha Maria Luísa.

“If you cannot - in the long run - tell everyone what you have been doing,
your doing has been worthless.”
Erwin Schrodinger

Abstract

A major challenge in phylogenetic reconstruction of evolutionary relationships lies in understanding the impact of model-fit on the accuracy of phylogenetic trees. The work conducted in this thesis aims to infer better trees by using amino-acid substitution models that are specific to the study data, and to evaluate strategies to mitigate the effects of systematic bias. Several software programmes for calculating data-specific models were evaluated, with IQ-TREE exhibiting the best features. These models consistently showed a better fit to the data than the pre-computed empirical models, indicating their greater robustness against biases caused by poorer-fitting models. Data-specific substitution models combined with more complex heterogeneous models or data partitioning strategies helped to reduce systematic bias. Among methods evaluated to identify heterogeneous data, the matched-pairs test of marginal symmetry combined with the Benjamini-Hochberg method exhibited the highest statistical power, identifying composition-heterogeneous sequences that biased the relationships among Archaeplastida and Bryophyta. By contrast, the process of evolution underlying the emergence of land plants from charophyte algae was shown to be composition-homogeneous among lineages in the analyses of nuclear and chloroplast data. Tree-homogeneous and heterogeneous analyses using these data robustly recovered the green algae Zygnematophyceae as the most-closely related to land plants. However, analyses of mitochondrial data placed Charophyceae as the sister-group to land plants; a result that was shown not to be caused by compositional heterogeneity among lineages. Nevertheless, further analyses identified a weak signal favouring Zygnematophyceae as the sister-group of land plants in buried-sites and slower-evolving sites data partitions. The cause of the incongruence between nuclear plus chloroplast data and the mitochondrial data remain unknown but maybe biological in nature rather than due to systematic bias, and perhaps a result of evolutionary processes such as horizontal gene transfer.

key-words: substitution model, amino-acid sequences, systematic bias, among-lineage heterogeneity, Streptophyta

Resumo

Atualmente a disponibilidade de dados não é mais uma restrição para a maioria das análises filogenéticas e um dos principais desafios dos filogeneticistas reside sobretudo em aspetos metodológicos relacionados com o uso dos modelos de evolução e o impacto de erros sistemáticos. O objetivo do trabalho desenvolvido nesta tese é inferir melhores árvores filogenéticas através do uso de modelos de substituição de aminoácidos específicos para os dados em análise e implementação de estratégias para mitigar os efeitos de erros sistemáticos. Foram avaliados cinco métodos para calcular modelos de substituição específicos para dados de sequências de aminoácidos, implementados nos programas FastMG, IQ-TREE, PAML e P4. Os quatro programas utilizam máxima verossimilhança, enquanto o P4 também utiliza inferência Bayesiana, sendo este método também avaliado. Os modelos específicos foram calculados utilizando alinhamentos de aminoácidos simulados com diferentes comprimentos. O processo de simulação dos dados de alinhamento incluiu um modelo de substituição e uma árvore filogenética conhecidos, isto é, o modelo e árvore de simulação. Cada modelo específico foi depois utilizado para calcular a pontuação de máxima verossimilhança da árvore e alinhamento usados para gerar o modelo específico. Este valor foi comparado com a pontuação de máxima verossimilhança resultante das análises da mesma árvore e alinhamento, mas usando o modelo de simulação. Os valores comparados foram estatisticamente similares, indicando que os métodos utilizados
para cálculo de modelos específicos de acordo com esta métrica são precisos. Quando as árvores filogenéticas estimadas livremente por estes modelos foram comparadas entre si, as árvores resultantes das análises que usaram os modelos calculados através de métodos de máxima verossimilhança implementados no IQ-TREE e P4 foram as mais precisas. Os modelos específicos foram também comparados com os modelos empíricos cpREV e WAG. Independentemente do método usando para cálculo dos modelos específicos, estes demonstraram um ajuste superior aos dados e inferiram árvores mais precisas quando comparados com os modelos empíricos. Estes resultados indicam que os modelos específicos deverão ter uma maior robustez contra desvios sistemáticos que resultem da falta de ajuste aos dados. O programa IQ-TREE demonstrou o melhor balanço entre rapidez no cálculo de modelos específicos e precisão das árvores filogenéticas inferidas usando estes modelos. A análise de conjuntos de dados empíricos utilizando modelos de substituição específicos corroborou os resultados anteriores. Tendo em conta a existência de programas eficientes e

rápidos para o cálculo de modelos específicos eficientes torna-se assim pouco razoável o uso de modelos de substituição empíricos em análises filogenéticas.

As análises anteriores e a maioria das análises filogenéticas assumem que o processo evolutivo é homogéneo ao longo do tempo, isto é, ao longo da árvore. No entanto, este pressuposto não corresponde à realidade do processo evolutivo, e quando fortemente rejeitado, pode resultar em erros sistemático que podem afetar a correta inferência da árvore filogenética. Os testes de pares emparelhados de simetria permitem investigar a heterogeneidade na composição e substituições ao longo da árvore. O teste de pares emparelhados de simetria marginal permite auferir a presença de processos composicionalmente heterogéneos, enquanto o teste de simetria interna permite avaliar processos de substituiç heterogéneos. O teste de pares emparelhados de simetria permite avaliar ambos. Uma vez estes incluem comparações múltiplas, o valor-p (valor da probabilidade) da comparação entre cada duas sequências no alinhamento deverá ser ajustado para uma correta identificação das sequências heterogéneas. Assim, em conjunto com os três testes de simetria, quatro métodos para correção do valor-p, nomeadamente, Bonferroni, Bonferroni-Holm, Benjamini-Yekutieli, e Benjamini-Hochberg foram avaliados. Para isso, foram simulados conjuntos de dados de acordo com quatro critérios: heterogéneos na composição, heterogéneos nas substituições, heterogéneos em ambos e totalmente homogéneos. O teste de pares emparelhados de simetria marginal combinado com o método de correção Benjamini-Hochberg exibiu a potência estatística mais elevada comparado com outros métodos. O teste de simetria e simetria interna revelaram uma baixa capacidade de deteção das sequências com substituições heterogéneas. Os métodos Bonferroni apresentaram a menor potência estatística. Análises de conjuntos de dados nucleares e mitocondriais utilizando o teste de simetria e o teste de simetria marginal combinados com os métodos de correção Benjamini, mas principalmente o teste de simetria marginal combinado com o método Bonferroni-Holm, identificaram sequências composicionalmente heterogéneas que distorceram a inferência das relações evolutivas nos clados Archaeplastida, Bryophyta e Setaphyta. Desta forma, o processo evolutivo é heterogéneo ao longo da árvore não podendo ser acomodado pelos modelos habituais que assumem a homogeneidade. A remoção destas sequências anulou ou reduziu o efeito negativo na inferência, resultando na monofilia destes grupos, o que está de acordo com outros estudos de análise filogenética utilizando modelos mais sofisticados.

O estudo do surgimento das plantas terrestres tem sido marcadamente debatido, incluindo três grupos de algas como possível grupo mais próximo das plantas, nomeadamente, Zygnematophyceae, Charophyceae e Coleochaetophyceae. Utilizando os métodos descritos acima e outros, foi investigado a história evolucionária entre as algas verdes, charophytes, e as plantas terrestres. As análises foram realizadas com recurso a dados sequencias de aminoácidos nucleares, mitocondriais e de cloroplasto. Contrariamente aos processos evolutivos descritos acima, o surgimento das plantas terrestres foi demonstrado ser homogéneo ao longo da árvore na análise de dados nucleares e de cloroplasto. Estas análises recuperaram as algas verdes Zygnematophyceae como grupo irmão das plantas terrestres, indicando este ser o grupo de algas mais próximo das plantas. Adicionalmente, análises que utilizaram modelos que acomodam heterogeneidade ao longo do alinhamento e estratégias para partição dos dados obtiveram resultados congruentes. No entanto, as análises de dados mitocondriais reconstruiram invés o grupo Charophyceae como o grupo de algas mais próximo das plantas. Posteriormente, análises de partições de dados associados a rácios de substituição mais lentos ou associados a aminoácidos com localizações mais conservadas na estrutura da proteína, recuperaram Zygnematophyceae como o grupo mais próximo das plantas, embora com pouco suporte. Estes resultados sugerem que os modelos de evolução atuais não são capazes de modelar corretamente o surgimento das plantas usando dados mitocondriais. Por outro lado, os sinais inerentes poderão estar corretos e o genoma mitocondrial possuir uma forma quimérica resultante de um processo biológico, nomeadamente a transferência horizontal de genes. No entanto, sendo este conflito o resultado de desvios sistemáticos ou de um processo biológico, as análises aqui realizadas não permitem apurar o mais provável.

Palavras-chave: Modelos de substituição, sequências de aminoácidos, erro sistemático, heterogeneidade ao longo da árvore, Streptophyta

Table of contents

Index of figures

Index of tables

LIST OF ABBREVIATIONS

AIC	Akaike information criterion
ASRV	Among-site rate variation
ASTRAL	Accurate Species TRee ALgorithm
BEAST	Bayesian Evolutionary Analysis Sampling Trees
BIC	Bayesian information criterion
BPP	Bayesian Phylogenetics and Phylogeography
BS	Bootstrap
FDR	False discovery rate
FNR	False negative rate
FWER	Family-wise error rate
GTR	General time-reversible
HGT	Horizontal gene transfer
JTT	Jones-Taylor-Thornton
LG	Le-Gascuel
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
MLE	Maximum likelihood estimate
MP	Maximum parsimony
MPTIS	Matched-pairs test of internal symmetry
MPTMS	Matched-pairs test of marginal symmetry
MPTS	Matched-pairs test of symmetry
NCBI	National Center for Biotechnology Information
NDCH	Node-discrete composition heterogeneity
NDRH	Node-discrete rate heterogeneity model
NNI	Nearest neighbour interchange
nRCFV	Normalised relative composition frequency variability
OV	Observed variability
PAM	Point Accepted Mutation
PAML	Phylogenetic Analysis by Maximum Likelihood
PCA	Principal Component Analysis

PP	Posterior probability
RaxML	Randomized Axelerated Maximum Likelihood
RF	Robinson-Foulds
SAR	Stramenopiles, alveolates, and rhizarians
SPR	Subtree pruning and regrafting
SSU rRNA	Small subunit ribosomal RNA
TBR	Tree bisection and reconnection
UFBOOT	Ultrafast bootstrap
WAG	Whelan and Goldman
WRF	Weighted Robinson-Foulds

Chapter I

General Introduction

1. General Introduction

1.1 Methods of phylogenetic inference

1.1.1 Calculating a phylogenetic tree

Phylogenetic inference of evolutionary relationships among present-day organisms is the basis for understanding how species have evolved and diversified. Typically, this involves reconstructing a bifurcating tree that relates extant species based on the analysis of traits or characters represented in a data matrix (Swofford *et al.,* 1996; Felsenstein, 2004). A phylogenetic tree, or phylogeny, comprises a topology, typically with accompanying branch lengths. The branching order indicates the relationships between taxa (i.e. the leaves, tips, or external nodes of the tree) and internal nodes represent common ancestors. The length of tip branches indicates how much taxa have diverged from the common ancestor they have with their closest relative sampled in the tree. As a bifurcating tree the degree of all internal nodes is three (each node is connected to three branches, one antecedent and two descendants). A degree greater than three indicates a polytomy where the branching order is unknown. Polytomies result from either a lack of information (a 'soft' polytomy) or from truly simultaneous species divergence (a 'hard' polytomy). However, the latter is considered very unlikely to occur in nature (Yang, 2014).

The ancestor of all taxa in a tree is located at the root node of the tree. Although trees are usually computed without implying a root, such that the tree length is the same regardless of which node is designated the root (i.e. the tree is time-reversible), it is common practice to root them thereby polarising relationships among taxa in the tree. Molecular clock, midpoint, and outgroup rooting are well-known methods to root a tree, with the latter being the most widely used. The outgroup rooting method includes taxa that are not part of the study group, named outgroups, in the analysis, and the root is then placed between the outgroups and the study species, the ingroup. It is well established that including outgroups that are closely related to the ingroup are preferable over including distantly related outgroups (Smith, 1994; Swofford *et al.,* 1996).

The terms monophyly, paraphyly, and polyphyly describe the distribution of a group of taxa on a phylogenetic tree. Monophyly describes a group that includes a common ancestor and all, and only, it's descendants. In contrast, a group is paraphyletic if it comprises the common ancestor of all members of the group and some but not all descendants of that

ancestor. When the group does not include the common ancestor of all members, it is polyphyletic (Hennig, 1966; Farris, 1974).

Phylogenetic trees are reconstructed from the analysis of a matrix of characters that can vary among taxa but should share a common ancestry. Similarity in characters due to inheritance from a common ancestor is known as homology (Hall, 2007). However, character similarity does not always imply homology. For example, the endothermy observed in mammals and birds arose independently in each of the two lineages. This phenomenon of independently evolved similarity is termed homoplasy and can bias the reconstructed tree, if homoplasisous characters are mistaken for being homologous (Wake *et al.,* 2011). Characters are derived from morphological or molecular features, representing the variation within and among species. With the advance of high throughput sequencing technologies, today most data used for phylogenetics are molecular, being either nucleotides from DNA sequences or amino acids from proteins. In this context, a data matrix used for phylogenetics is the result of a sequence alignment, where each row represents a different taxon, and each column corresponds to a specific position, or site, along aligned sequences. Sequences from individual taxa are aligned so that the homologous sites are aligned in the same column thereby enabling the comparison of site identities among sequences. Given there are four nucleotides and 20 amino acids, the number of possible combinations is 4^m and 20^m, respectively, where m represents the number of taxa (Swofford *et al.,* 1996).

Inference methods can operate directly on phylogenetic data matrices, in which case the methods are character-based, or they can act on a matrix of distance metrics calculated from the former. Character-based methods include maximum parsimony (MP; Fitch, 1971; Hartigan, 1973; Felsenstein, 1978), maximum likelihood (ML; Felsenstein, 1981), and Bayesian inference (Rannala and Yang, 1996; Mau and Newton, 1997; Huelsenbeck *et al.,* 2001). Apart from MP where the model is implied but not explicit (Felsenstein, 1978; Yang, 1996a), these methods rely on an explicit model of nucleotide or amino-acid substitution.

The MP, ML, and Bayesian inference methods are based on an optimality criterion, that is, a score is assigned to each tree considered and used to rank trees (Swofford *et al.,* 1996). These methods encompass two main steps: calculating the tree score and searching for the tree with the best score in tree space. The tree space is the conceptual landscape of all possible phylogenetic trees that can be constructed for a given set of taxa. The number of possible unrooted trees of 10 taxa exceeds two million, yet most current analyses include many more, even thousands, of taxa. This highlights the computational challenge and illustrates why an exhaustive tree search is computationally unfeasible for larger data sets.

Consequently, most methods employ heuristic search algorithms (Swofford *et al.*, 1996). These methods traverse the tree space in order to identify the tree that have the highest probability of generating the observed data. However, while they reduce computation time, they cannot guarantee to find the optimal tree. Heuristic tree searches usually operate by "hill climbing" methods, such as tree-rearrangement or branch-swapping algorithms, namely the nearest neighbour interchange (NNI), the subtree pruning and regrafting (SPR), and the tree bisection and reconnection (TBR) (Swofford *et al.*, 1996; Guindon and Gascuel, 2003; Stamatakis, 2006). New trees, called neighbours, are proposed after local perturbations of the current tree. If the candidate tree has a better score, the algorithm selects the new tree.

1.1.2 Maximum likelihood

The ML method was introduced in the early 1920s (Fisher, 1922) and became commonly applied in many scientific fields, including phylogenetics. As it is applied in phylogenetic analysis, ML is statistically consistent in most cases and will converge to the true tree with increasing certainty as the number of sites approaches infinity (Bryant *et al.*, 2004). Furthermore, this method benefits from using an explicit model of change, which facilitates the evaluation and subsequent improvement of the fit between the model and the data as necessary. ML uses the likelihood function, which describes the probability of observing the data given a particular hypothesis. The hypothesis is the parameters one aims to estimate, such as topology, branch lengths, or substitution rates (models of molecular substitution are described in section 1.2). The likelihood function summarises all information from the data about the parameters, and the values of the parameters that maximise the likelihood is named the maximum likelihood estimate (MLE). The likelihood is, therefore, a function of the parameters, and a higher likelihood score is preferred (Yang, 2014).

Under the assumption that sequence sites evolve independently, the likelihood score is calculated separately for each site. The joint probability of the whole data matrix is then the product of the probabilities of data at individual sites. Because the raw likelihood values are extremely small, the logarithm of the likelihood is used instead, and the probabilities are accumulated as the sum of the likelihood logarithm of each site (Swofford *et al.*, 1996).

In the reconstruction of the phylogeny using ML, one of the two levels of optimisation is those of the branch lengths (and other model parameters if they are also to be estimated), as the evolutionary process is modelled as a Markov process running along a tree's branches. Branch and model parameters are optimised by maximizing the log-likelihood using numerical iterative algorithms such as Newton, Davidon-Fletcher-Powell, Broyden-Fletcher-

Goldfard-Shanno, among others (Gill et al., 1981; Fletcher, 2000). The likelihood of a tree is typically computed under the time-reversible assumption which implies that the evolutionary process is probabilistically identical whether observed forwards or backwards. Consequently, the placement of the root in the tree does not affect the likelihood, thereby simplifying its calculation (*pulley principle;* Felsenstein, 1981). The second level of optimisation consists in searching for the best tree in the tree space, that is, the tree with the highest log-likelihood score. The tree search is conducted using heuristic methods, such as NNI, SPR, or TBR. Because neighbouring trees share subtrees, the algorithm can save time by avoiding repeating likelihood computations of the same subtrees. In addition, because the resulting trees are unrooted, the number of candidates in the tree space also decrease considerably. Significant efforts have been dedicated to the development of more efficient algorithms for the search of likelihood-based phylogenetic trees, such as the FastDNAML (Olsen *et al.,* 1994), PhyML (Guindon and Gascuel, 2003), RAxML (Stamatakis, 2006), or IQ-TREE (Nguyen *et al.,* 2015).

A phylogenetic tree reconstructed using ML lacks an inherent measure of accuracy, requiring additional methods to assign support values to the tree, such as the bootstrap method (Felsenstein, 1985). The latter method involves generating many bootstrap pseudo-samples, which consist of resampling, with replacement, sites from the original alignment. Sites are chosen at random so that the bootstrap data sets are the same length as the original. Phylogenetic trees are reconstructed from each bootstrap data using the same method as used for the original data set. The bootstrap support value of a node is the proportion of bootstrap trees that include the split. These support values are typically used to visually decorate the optimal tree estimated from the original data set, or the consensus tree derived from the bootstrap trees itself is presented.

1.1.4 Bayesian phylogenetic estimation

Bayesian inference offers a more intuitive approach to phylogenetic reconstruction than ML since it considers the probability of the parameters given the data. A second significant difference between the two, is that the likelihood approach focuses on calculating an optimal point estimation, while Bayesian inference estimates a posterior probability distribution of parameters and trees (Yang, 2014; Nascimento *et al.,* 2017).

Bayes' theorem connects the posterior distribution and the likelihood function, forming the basis of Bayesian probabilistic inference (Bayes, 1763). The theorem states that the posterior probability of observing A given that B has occurred, $\Pr(A|B)$, is proportional to

the product of the likelihood of observing *B* given *A* Pr(*B*|*A*) and the prior probability, Pr (*A*), normalised by the unconditional probability of B. In Bayesian inference, one assigns a distribution based on prior knowledge about the unknown parameter before observing the data (probability of observing A): this is called the prior distribution. The unknown parameter(s) is estimated using the likelihood function which summarises the information in the data (Pr(B|A)). The product of these two is then normalised by the marginal probability by integrating the posterior distribution into one. In sum, the posterior information results from the sum of the prior information and sample information (likelihood).

The prior probability (or just prior) in Bayesian inference reflects the knowledge of the process before the data are analysed. In the context of phylogenetic inference, priors can be assigned on branch lengths, substitution rates, and the tree topology. For instance, one can construct a prior on the tree topology based on biological processes, such as the birth-death process (Rannala and Yang, 1996). If an uninformative (also called a flat or diffuse) prior is assigned, then the posterior distribution is primarily driven by the likelihood of the data. However, in the case of an informative prior, it is advisable to assess the sensitivity of the posterior to the prior. It is common practice to ignore the correlation between parameters in the construction of priors, assigning independent priors for each parameter. This can sometimes cause problems when multiple parameters of an analogous type are specified, such as independent and identically distributed priors on branch lengths (Rannala *et al.*, 2012). Therefore, understanding the data and the insights they provide regarding the parameters helps prevent the overloading of analyses with an excessive number of parameters (Yang, 2014).

Having chosen prior distributions for parameters, computation of the marginal probability (normalising constant) of a data set involves multidimensional integrals. It is typically impossible to calculate the exact value of the marginal probability of a data set due to computational complexity. Use of the Markov chain Monte Carlo (MCMC) algorithm avoids the direct calculation of the marginal probability and instead generates a sample from the posterior distribution. MCMC is a simulation algorithm used for sampling from probability distributions which generates dependent samples from the target density (the posterior) and forms a stationary Markov chain (Metropolis *et al.*, 1953). Values of states sampled during the chain represent the possible values of the parameters which eventually converge to the target posterior distribution at stationarity.

Phylogenies are reconstructed under the general framework of hierarchical Bayesian inference. The applied MCMC algorithm starts with a topology, branch lengths, and

substitution parameters, all randomly drawn from the prior distributions. Then follows a number of iterations (samples) where the MCMC proposes changes (moves) to the tree (e.g., using NNI, SPR, or TBR), branch lengths, and substitution rate parameters. The chain is sampled at specific samples, and the current parameters are saved (Nascimento *et al.*, 2017). The MCMC algorithm is memoryless in the sense that parameter values visited in the following iteration do not depend on the past values but only on the current ones. Another important feature of the algorithm is the use of the ratio of posterior densities rather than the posterior density itself, thereby enabling the MCMC algorithm to generate samples from the posterior. If the algorithm is run for sufficient time, the parameter values with high posterior are sampled more often than those values with low posterior probability. Once the run is complete, one can select the tree with the highest posterior probability (the maximum *a posteriori* tree) or construct a majority-rule consensus tree from the sampled trees (Rannala and Yang, 1996a). A natural measure of confidence for the tree splits is obtained by calculating the proportion of sampled trees that include each split (Larget and Simon, 1999).

1.2 Evolutionary models of molecular substitution

1.2.1 Nucleotide and amino-acid substitution models

Evolutionary substitution models are probabilistic models used to describe the process of how changes accumulate in the data which and in the context of phylogenetic analyses can provide detailed insights into evolutionary relationships. Sites in the sequences are assumed to evolve independently of each other and are described by a Markov chain with a probability of change at any particular site. The evolutionary process is assumed to exhibit a "Markovian property", which characterises the states of the chain (the four nucleotides or the 20 amino acids) as memoryless, with the probability of change from the current state not depending on how the current state is reached. Changes among molecular states are described as instantaneous events over evolutionary time, even though molecular substitutions require multiple generations to become fixed in nature (Swofford *et al.*, 1996). In ML and Bayesian inference methods, substitution models are used to calculate the probabilities of change along the tree branches, thereby enabling the computation of the likelihood of the observed data (Swofford *et al.*, 1996; Yang, 2014). Notably, constraints placed on the evolutionary process, while providing mathematical convenience and simplifying calculations, do not accurately represent the natural evolutionary processes. Nonetheless, we expect that these models approximate real processes closely enough for a valid and meaningful interpretation: "All models are wrong, but some are useful" (Box, 1979).

Modelling of the substitution process is subject to distinct constraints, hence a variety of different models have been proposed. The JC69 model (Jukes and Cantor, 1969) is the simplest and most constrained model of nucleotide substitution. It assumes that there is no difference in substitution rate among all four nucleotides and that base compositions are equal. The JC69 substitution rates can be expressed as a Q matrix of 4x4 instantaneous substitution rates from nucleotides *i* to *j*, with *i, j* = T, C, A, or G. The diagonal elements are adjusted to ensure that each row sums to zero. The amino-acid equivalent is the Poisson-distributed model (Bishop and Friday, 1985; 1987). The K80 model (Kimura, 1980) accommodates distinct substitution rates between transitions and transversions, having two free parameters. However, as with the JC69 model, K80 also restricts the base compositions, assuming equal proportions among states. By contrast, the F84 (Kishino and Hasegawa, 1989) and the HKY (Hasegawa *et al.,* 1985) models extend the K80 model to having asymmetric composition frequencies, with both models having five parameters. These two models are special cases of the TN93 model (Tamura and Nei, 1993), with six free parameters. The general time-reversible model (GTR; Tavaré, 1986; Yang, 1994a; Zharkikh, 1994) imposes fewer constraints on substitution rates than the previous models, allowing each type of substitution to have its unique rate. The GTR is also commonly known as the REV (reversible) model, mainly in the context of protein sequences.

To model the aligned sequence data with Markov chains, one needs to compute the matrix of transition probabilities P(t). The elements of the transition probability matrix describe the probability of character *X* that is in state *i* will turn to *j* after a period of time, *t*, and can be denoted as:

$$p_{ij}(t) = \Pr\{X(t) = j \mid X(0) = i\}, \; t > 0 \qquad (1.1)$$

The Markov model is therefore applied to the data by calculating the likelihood of each possible change between states at each site. The Q matrix specifies the instantaneous transition rates between different states without depending on time (i.e., substitution rates remain constant over time). Therefore, and assuming time-homogeneity, that is, that the transition probabilities are the same for all values of *t,* the Q matrix can be used to derive the transition probability matrix over any period of time as:

$$P(t) = e^{Qt} \qquad (1.2)$$

Markov models assume the limiting distribution of the chain, where π_j represents the probability that the chain is in state *j* as *t* approaches infinity. This translates to the equilibrium

frequencies of the four nucleotides or the 20 amino acids. Once the sequence evolution reaches the limiting distribution, the frequencies remain in that distribution, reflecting a state of equilibrium (stationary distribution π). Most models assume molecular sequences to have the same base composition - the assumption of stationary. The stationary distribution is defined as:

$$\pi P(t) = \pi \qquad (1.3)$$

Substitution models are usually time-reversible, with the Q matrix satisfying the condition:

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \text{ for all } i \neq j \text{ and for any } t \qquad (1.4)$$

With π_i denoting the proportion of time that the Markov chain spends in state *i* and $\pi_i p_{ij}$ representing the amount of change from state *i* to *j*, while change in the opposite direction is detonated as $\pi_j p_{ji}$. This equation implies that the probability of changing from *i* to *j* is the same as from *j* to *i*.

In summary, the Q matrix of the time-reversible and time-homogeneous GTR model includes six substitution rate parameters, one for each possible pair of nucleotides, and four equilibrium frequency parameters (Tavaré, 1986; Yang, 1994a). This extends to 190 rate parameters for amino-acid substitutions to account for all possible pairs of the 20 amino acids and 20 composition frequencies. Because the composition frequencies must sum to one, the total number of free parameters is reduced by one. Similarly, it is common to express the rate parameters as relative rates to one of rate values (mainly in likelihood analyses), again reducing the number of free parameters by one. The off-diagonal elements of the GTR matrix for nucleotide sequences can be described as a product of a symmetrical matrix of rates *r,* multiplied by a diagonal matrix comprising the composition frequencies, π:

$$Q = \begin{bmatrix} * & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & * & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & * & r_{GT}\pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & * \end{bmatrix} = \begin{bmatrix} * & r_{AC} & r_{AG} & r_{AT} \\ r_{AC} & * & r_{CG} & r_{CT} \\ r_{AG} & r_{CG} & * & r_{GT} \\ r_{AT} & r_{CT} & r_{GT} & * \end{bmatrix} \times \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix} \qquad (1.5)$$

Unlike most current phylogenetic analyses of nucleotide data where the substitution rate matrices are calculated directly from the study data under the GTR model, analyses of protein sequences typically use pre-computed empirical substitution models. For the empirical models the amino-acid exchange rate values are estimated from custom-assembled protein data sets using the GTR model. The reason for this practice lies in the lack of availability of

data and the larger computational resources required to estimate models (discussed in Chapter II). The first attempts to model protein sequence evolution were undertaken in the sixties (Eck and Dayhoff, 1966; 1968), resulting in a "mutation probability matrix" calculated using parsimony-based methods (Dayhoff *et al.*, 1978). The "Dayhoff matrix" (PAM 001 - Point Accepted Mutation), at the time the most important model, represents the transition probability matrix for a 1% expected change per site. Later, the calculation of the JTT model (Jones *et al.*, 1992) employed the same procedure but used a much larger data set of more diverse proteins. Early instances of REV models estimated using ML include the mitochondrial models mtREV (Adachi and Hasegawa, 1996) and mtMAM (Yang *et al.*, 1998), and the chloroplast model cpREV (Adachi *et al.*, 2000). Subsequently, the WAG (Whelan and Goldman, 2001) and the LG (Le and Gascuel, 2008) models were also estimated under the ML framework. Other empirical matrices have been developed, with varying degrees of specificity for certain taxonomic groups. Though the majority were estimated using the ML, the gcpREV model (Cox and Foster, 2013) is distinguished as an amino-acid substitution model estimated using Bayesian MCMC with rate parameters were drawn from a posterior distribution. An often beneficial variation on the use of empirical amino-acid models which improves the model's fit is to replace the equilibrium amino-acid frequencies of the model with the frequencies observed in the data (Cao *et al.*, 1994), or have them optimised during the analysis.

1.2.2 Models of among-site variation

Instantaneous substitution rates among sites and their frequencies can vary across the data, that is, site-heterogeneous processes. Incorrectly assuming a constant rate of change among sites can negatively impact the accuracy of the phylogenetic reconstruction for instance (e.g., Gaut and Lewis, 1995; Yang, 1996b). In response, substitution models have been extended to accommodate variable rates across sites, such as mixture models, with each site having a probability of evolving at any rate drawn from either a discrete or continuous probability distribution (Swofford *et al.*, 1996). A mixture model assumes the absence of information regarding the site's nature and assumes that the evolutionary processes vary widely, accounting for the variation at the site level.

The gamma mixture model is commonly used for modelling the variation in the absolute rate of substitution at sites in the data. In this model, site-specific rates are drawn from a gamma distribution shaped according to a parameter α which is inversely related to the extent of rate variation at sites (Uzzell and Corbin, 1971; Jin and Nei, 1990). Instead of using

a continuous gamma distribution, the gamma distribution is often divided into several discrete categories (typically four) which has been shown to provide a good approximation to the full distribution (Yang, 1994b).

Molecular sequences may have varied types of substitution processes among sites as they are exposed to different selective pressures, such as their specific roles in protein structure and function. For instance, hydrophobic amino acids are typically preferred at buried sites (amino-acids in the interior of the protein structure), whereas active-site residues are more commonly electrostatically charged and on the protein surface. Mixture models that use site-specific equilibrium frequencies can accommodate this variation (Bruno, 1996; Koshi and Goldstein, 1998; Crooks and Brenner, 2005), with the CAT model being the most widely used (Lartillot and Philippe, 2004). The latter model uses multiple site profiles each with its own equilibrium frequencies. Sites are then modelled by the probability of belong to a profile, As implemented under Bayesian framework, a Dirichlet process specifies a prior distribution on the assignment of sites to different classes, which also induces a prior distribution on the number of classes. A Dirichlet process enables the estimation of parameters from the data without need of *a priori* knowledge. In contrast to site-homogeneous models, the CAT model has been shown to better accommodate multiple substitutions and improve resilience against artefacts such as long-branch attraction (Lartillot and Philippe, 2006; Lartillot *et al*., 2007; Struck *et al.,* 2011), hence the model has been used widely in phylogenetics studies (e.g., Brinkmann and Philippe 2008; Delsuc *et al.,* 2008; Philippe *et al.,* 2009; Finet *et al.,* 2010; Timme *et al*., 2012; Nesnidal *et al.,* 2013; Nosenko *et al.,* 2013; Chang *et al*., 2015; Luo, 2015; Whelan *et al*., 2015; Cannon *et al.,* 2016). Despite this, analysis of large data sets using the CAT model demonstrated challenges in achieving convergence of the MCMC chains (Kocot *et al*., 2011; Pisani *et al*., 2015; Whelan *et al*., 2015).

Empirical profile mixture models, sharing a similar principle to the CAT model, can be implemented in a ML approach (Le, Gascuel *et al.,* 2008; Wang *et al.,* 2008; Wang *et al.,* 2018; Schrempf *et al.,* 2020). Analyses using these models are much less computationally demanding because they use fixed stationary distributions thereby avoiding the need to calculate it from the study data. The empirical distribution models are grounded on the expectation that site-specific amino-acid constraints may be due to universal biochemical constraints, and that, therefore, the same empirical distributions can be applied to various data sets (Schrempf *et al.,* 2020). Nevertheless, the amino-acid composition distributions can be calculated from the study data and applied in a ML tree search (Susko *et al.,* 2018; Schrempf *et al.,* 2020).

Mixture models can also be used to include several rate matrices in the same analysis, such as matrices that specify rates for substitution in protein secondary structures, or the rate of substitution among protein surface-accessible residues (e.g., Koshi and Goldstein, 1996; Thorne *et al.*, 1996; Goldman *et al*., 1998; Le, Lartillot *et al.*, 2008; Le and Gascuel, 2010). For example, the EX2 model incorporates two substitution rate matrices estimated from sites that are either buried or exposed in the protein tertiary structure, while the EX3 model includes three rate matrices corresponding to highly exposed, intermediate, and buried sites (Le, Lartillot *et al.*, 2008). Mixture models can also incorporate different substitution rate matrices where each matrix corresponds to a class of sites differentiated by their evolutionary rates, such as the LG4M and LG4X (Le *et al*., 2012).

In contrast to the mixture models, which assume the absence of information regarding the site's nature, the partition models assume *a priori* knowledge about the evolutionary process across data. A partition model implies the partitioning of the data, where partitions can be delineated according to genes or codons, which is readily justifiable in biological terms (Yang, 1996c; Nylander *et al*., 2004; Brandley *et al*., 2005; Lanfear *et al.*, 2012), or other criteria such as distinguishing between fast- and slowly-evolving sites, or buried and exposed sites (e.g., Shapiro *et al.*, 2006; Ho and Lanfear, 2010; Pandey and Braun, 2019). However, modelling short data partitions (e.g., partitioning into genes) risks over-fitting the data, and suboptimal partitioning schemes can result in incorrect trees (Kainer and Lanfear, 2015). To counter this, reducing the number of partitions by merging them can decrease the number of parameters to estimate, thereby enhance tree inference. To optimise partition schemes, tools like PartitionFinder have been developed (Lanfear *et al.*, 2012; 2016) which aim to select the best-fitting scheme without over-parameterising the model. In addition, partition models can be integrated with mixture models, with the first addressing large-scale differences among partitions, and the latter accommodating the remaining variation within each partition (e.g., Yang, 1995; Redmond and McLysaght, 2021).

1.2.3 Models for among-lineage variation

The process of evolution can also differ over time, that is, among-lineages in the tree, and thereby conflict with the assumptions of rate homogeneity and stationarity of the model that are typical in phylogenetic analyses. When there is significant among-lineage heterogeneity the accuracy of phylogenetic methods can be impacted if not accounted for in the model. To counteract this several tree-heterogeneous models have been proposed (e.g., Yang and Roberts, 1995; Galtier and Gouy, 1998; Galtier *et al*., 1999; Foster 2004; Blanquart

and Lartillot, 2006; Blanquart and Lartillot, 2008; Foster *et al.*, 2009; Jayaswal *et al.*, 2011; Zou *et al.*, 2012; Groussin *et al.*, 2013; Williams *et al.*, 2020). Early approaches implemented using ML allowed for variation in base compositions over the tree, with the composition frequencies varying independently on each branch with a distinct composition vector for each branch (Yang and Roberts, 1995). However, for larger trees this approach implies many parameters, and consequently, the number of vectors of frequency parameters and their assignment to the branches should be specified prior to analysis. Other proposed models included fewer parameters, such as those assuming equal frequencies between nucleotide bases T and C and between A and G (Galtier and Gouy, 1998; Boussau and Gouy, 2006), or using only a two state GC content vector of parameters for each branch (Galtier *et al*., 1999). Bayesian MCMC analyses of non-stationary models mitigate the problem of increasing number of parameters by constructing a prior (Foster, 2004; Blanquart and Lartillot, 2006). For instance, the node-discrete composition heterogeneity (NDCH) model uses a predefined number of composition vectors which are shared among some branches (Foster, 2004). This approach assumes that compositional heterogeneity, if present, will be localised in parts of the tree. As a result, the model does not require a unique composition vector for every branch. In a later version of this model, NDCH2, a composition vector is estimated for each branch of the tree, which is constrained via a sampled concentration parameter of a Dirichlet prior (Williams *et al.*, 2020). Blanquart and Lartillot (2006) proposed a tree-heterogeneous composition model where the substitution process shifts at specific points (breakpoints) on a branch, assuming new composition frequencies generated from a uniform Dirichlet prior. These breakpoints are determined according to a compound Poisson process along the tree branches. This model was combined with the CAT mixture model into a new model, CAT-BP, which can accommodate compositional heterogeneity across lineages and among sites (Blanquart and Lartillot, 2008). By contrast, the node-discrete rate heterogeneity model (NDRH) assigns rate matrices to different branches of the tree (Foster *et al.*, 2009). The use of tree-heterogeneous models can result in trees that were more likely to be correct (e.g., Cox *et al.*, 2008; Foster *et al.*, 2009; Morgan *et al.*, 2013; Sousa *et al.*, 2019; 2020; Williams *et al.*, 2020). These studies not only demonstrated a better fit to the data using tree-heterogeneous models but also indicated the presence of systematic bias in the original analyses due to non-stationary (mostly) and non-homogeneous processes.

Substitutions can also vary in their site-specific rates, i.e., the substitution rate for each site is not necessarily constant across the tree. For example, a site may evolve faster in some lineages and more slowly in others. These lineage-specific shifts in the substitution rates

(rates varying across sites and lineages) are known by heterotachy (Philippe and Lopez, 2001; Lopez *et al.*, 2002). Heterotachy has been suggested to arise from selective pressure on a focal site being influenced by the nature of nearby sites, or from functional divergence where the rate variation results from the protein differentiation across lineages (Gu, 2001; Gaston *et al.*, 2011). The covarion (concomitantly variable codon) model accommodates this variation, enabling a site to switch from one state to another (invariable or variable; Fitch and Markowitz, 1970; Fitch, 1971). Current heterotachy/covarion models enable sites to switch between a number of different rates and an invariable state as they evolve across the tree (Galtier, 2001; Huelsenbeck, 2002; Wang *et al*., 2007; Zhou *et al*., 2007; Wu and Susko, 2009; Crotty *et al.*, 2020).

1.2.4 Model fit

In ML analyses the fit of different models to the data can be compared using a likelihood ratio test or information criteria. The likelihood ratio test assesses the likelihood of two models where one is a simplification of the other. This test statistic follows approximately a χ^2 distribution where the degrees of freedom are equal to the difference in the number of parameters between the two models. The resulting p-value enables the assessment of whether the difference in likelihood is statistically significant. For instance, if one assumes the likelihood value of the less parameter-rich model as the null hypothesis, then a significant result from the likelihood ratio test indicates a marked improvement in data fit from including more parameters in the alternative model. This justifies the preference for the parameter-rich alternative (Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997). When the models being compared are not nested, or when the sequences are too short such that the distribution may not be reliable, the null distribution for the test statistic can be generated through simulations; a method known as the parametric bootstrap (Goldman, 1993a; Yang *et al.*, 1994). Alternatively, information criteria offer a distinct method for comparing non-nested models, allowing for the simultaneous evaluation of all candidate models. The Akaike information criterion (AIC) is a score calculated using the optimal likelihood under the model and the number of parameters, where the model candidates are penalised by the increasing number of parameters (Akaike, 1973). A lower AIC value indicates a better fit of the model to the data. The Bayesian information criterion (BIC; Schwarz, 1978) is based on an approximation to the natural log of the Bayes factor (marginal likelihood ratios from two competing models; Kass and Raftery, 1995), where models with more parameters are penalised greater than AIC methods. Tools for automatic model selection, such as the

ModelTest, allow for the hierarchical comparison of known substitution models based on these criteria (Posada and Crandall, 1998; Posada, 2008). The debate over the use of information criteria is ongoing, focusing on the severity of parameter penalisation (e.g., Seo and Thorne, 2018; Susko and Roger, 2020).

Although the above methods can select the model with the relative highest statistical fit, the best-fitting model may still describe the data poorly. Methods of evaluating model adequacy, such as the Goldman-Cox test (Goldman, 1993b) and posterior predictive simulation (Gelman *et al*., 1996; Huelsenbeck *et al*., 2001; Bollback, 2002; 2005), can assess if a model fits the data in an absolute sense. The Goldman-Cox test uses a null distribution generated through a parametric bootstrap to evaluate the hypothesis of a perfect fit between the model and the data. By contrast, posterior predictive simulation uses parameter estimates drawn from their respective posterior distributions under the Bayesian framework. The two tests typically use the multinomial log-likelihood test statistic. These methods assess the overall model fit to the data, but the model's fit to the data composition can also be evaluated separately. The chi-squared (χ^2) test for compositional homogeneity is commonly used to test the fit of a composition-homogeneous model to the data. Nonetheless, this test ignores tree-based correlation of compositions among taxa and therefore because the chi-squared curve does not provide an appropriate null distribution, there is a high probability of suffering a Type II error (Foster, 2004). Based on simulations, Foster (2004) proposed a better null distribution for this test. Chi-squared statistics, extracted from data simulated under a specified model and tree, are used to generate a valid null distribution to assess the significance of the original Chi-squared value. A p-value is calculated as the percentage of simulations resulting in more extreme statistic values than the original statistic (tail-area probability). Other tests, such as the matched-pairs tests of symmetry (Bowker, 1948; Stuart, 1955; Ababneh *et al.*, 2006), can also be used to test the fit of the model to the data (described in 1.3.3 below).

1.3 Evaluating tree estimation error

1.3.1 Biases in tree inference caused by analytical errors

Analyses conducted using different methods, such as ML and Bayesian inference, and different genes and genomes would ideally converge to the same species tree for the same set of taxa. When that is the case, one can have high confidence in having reconstructed the correct species tree. However, reconstructing certain evolutionary divergences remains challenging and competing phylogenetic hypotheses can emerge. The incongruence between

analyses often arises due stochastic and systematic errors (Cox, 2018). The first is produced by insufficient amount of data, such as analyses relying on a small number of sequences, resulting in few defining substitutions and leading to low signal-to-noise ratios. This occurs most often in deep and short branches of a tree which depict ancient and rapid evolutionary divergences that had little time to accumulate changes (Delsuc *et al.*, 2005). In such cases, using larger amounts of data usually helps to resolve these relationships (Rokas *et al.,* 2003; Dunn *et al*., 2008). Indeed, current analyses often use genome-scale data sets, which are less susceptible to stochastic error. However, larger data sets may actually amplify non-historical signals (biases), leading to systematic error during the phylogenetic inference mainly due to inaccurate model assumptions about the data (Phillips *et al.,* 2004; Jeffroy *et al.,* 2006). This error is "systematic" because it consistently and repeatably produces the same incorrect solution, rather than varying randomly. For instance, in deeper divergences where multiple substitutions may have occurred at the sequence sites, taxa can share character states because of convergent evolution rather than inheriting it from a common ancestor (homoplasy). In these cases, the actual number of substitutions is underestimated by the model due to its lack of ability to handle multiple substitutions correctly. Consequently, when unaccounted for, taxa may be inferred to be closely related due the accumulation of homoplasy when they are actually more-distantly related. A pronounced effect of homoplasy is its accumulation in rapidly evolving lineages, leading to the well-known phenomenon of "long-branch attraction" (Felsenstein, 1978). This is characterised by the apparent loss of the historical signal due to substitutional "saturation" (i.e. accumulation of multiple hidden substitutions). Adding more taxa can benefit the phylogenetic inference by segmenting the long branches in such cases, at the risk of adding or increasing the effect of other biases (Rokas and Carroll, 2005; Heath *et al*., 2008). Heterogeneity among-sites, or among-lineages, is another source of systematic error when the substitution process is incorrectly assumed to be uniform. Consequently, artefacts in phylogenetic inference emerge because the model fails to accommodate heterogeneity adequately.

1.3.2 Assessing among-lineage heterogeneity

Among-lineage heterogeneity, or tree-heterogeneity, in the data implies the rejection of the usual assumptions of stationarity and homogeneity of phylogenetic analysis and is one of the causes of systematic error when not properly addressed. The χ^2 test for compositional among-lineage homogeneity (introduced in section 1.2.4) can assess whether the data are homogeneous in composition across the tree (Foster, 2004). A rejection of the model indicates

the data are compositionally heterogeneous and non-stationary. If the model describes a stationary (and reversible) process, the probability of finding a site with nucleotides *i* and *j* in two sequences is the same as the probability for a site with *j* and *i* (Yang, 2014), that is, the data follow an assumption of symmetry. The symmetry of two (Tavaré, 1986) or more sequences (Rzhetsky and Nei, 1995) can be tested using a contingency table to count site patterns. A X^2 statistic (reflecting the number of sites with nucleotide *i* in sequence one and *j* in sequence two) is compared against the asymptotic χ^2 distribution with six degrees of freedom. These tests assume that each element in the array corresponds to the frequency of a matching pair of nucleotides or amino acids in an alignment and are based on the matched-pairs tests.

The matched-pairs tests of symmetry (MPTS; Bowker, 1948), marginal symmetry (MPTMS; Stuart, 1955), and internal symmetry (MPTIS; Ababneh *et al.*, 2006) can be used to assess whether homologous sequences have evolved under the same conditions (Jermiin *et al.*, 2004; 2009). Homologous sequences are summarised in a divergence matrix, D, as a contingency table with elements d*ij* representing the number of alignment sites having *i* in the first sequence and *j* in the second sequence. Then, a null hypothesis can be formulated concerning the overall symmetry, marginal symmetry, and internal symmetry of D. Symmetry implies that in a contingency table, the count of a nucleotide (or amino acid) pair is identical to its transposed pair. Therefore, the null hypothesis for symmetry assumes f_{ij} is equal to f_{ji}, with *i* different from *j*. The MPTS is then given by:

$$Ss^2 = \sum \frac{(d_{ij} - d_{ji})^2}{d_{ij} + d_{ji}} \qquad (1.5)$$

The statistic (S_S^2) follows an asymptotic distribution that can be described as a χ^2 variate with $v = l(l-1)/2$ degrees of freedom, where *l* represents the number of states. The marginal symmetry indicates that the frequency of a nucleotide or amino acid in one sequence equals its frequency in another. Therefore, the MPTMS assess the equality of composition between sequences and is given by:

$$S_M^2 = u^T V^{-1} u \qquad (1.6)$$

The resulting statistic (S_M^2) also follows an asymptotic distribution, described as a χ^2 variate with $v = l-1$ degrees of freedom. MPTMS requires a vector of marginal differences (u) and its variance-covariance matrix (V). The MPTMS is not applicable when the variance-covariance matrix is not invertible. The MPTIS is the difference between the two previous statistics, $S_I^2 = S_S^2 - S_M^2$. It is asymptotically distributed with $v = (l-1)(l-2)/2$ degrees of

freedom. A small p-value (e.g., <0.05) indicates a rejection of the assumption of stationarity or homogeneity when computed using the MPTMS or the MPTIS, respectively. When using the MPTS, a low p-value suggests rejecting either or both of these assumptions (Jermiin *et al.*, 2017). For more than two sequences under analysis, the tests can be conducted pairwise on all possible pairs of sequences. A sequential Bonferroni correction can be applied to counteract the effect of multiple comparisons (Ababneh *et al.*, 2006; Jermiin *et al.*, 2017).

Various other methods have been proposed to assess compositional heterogeneity across lineages. The normalised relative composition frequency variability measures the average variability in composition frequency across lineages (Fleming & Struck, 2023). This method evaluates the relative frequency of a given nucleotide or amino acid for a given taxon compared to its average frequency across the entire data set. The disparity index measures the frequencies of nucleotides or amino acids shared and differing between two sequences, using Monte Carlo simulations (Kumar & Gadagkar, 2001). Another option is comparing tables and graphs of sequence-specific distributions generated using Monte Carlo simulations or the multinomial distribution (Lanave *et al.*, 1984, 1986). Furthermore, compositional heterogeneity can be visualised by displaying sequence compositional content in one, two, or three dimensions (Ho *et al.*, 2009). In contrast to the χ^2 test for compositional homogeneity and matched-pairs tests, these additional methods are not directly related to the model fit or to the degree to which the assumption of stationarity is violated, if at all.

1.3.2 Strategies to reduce systematic bias

In a probabilistic framework the expectation is that the model fits the data sufficiently so that there is no systematic bias. Consequently, due to the complexity of the substitution process and it's variation over time there has been a focus on developing more realistic models, such as the CAT and NDCH2. Nevertheless, as parameter-rich models their implementation can sometimes be difficult and traceability limitations emerge with larger data sets. Furthermore, while more complex models may be able to accommodate different types of heterogeneity separately, accounting for joint effects can be exceedingly difficult (Williams *et al.*, 2021). Therefore, better models should not necessarily have to be overly complex (Steel, 2005), and the balance between model complexity, more realistic models, taxa sampling, and data size should be considered carefully.

Being able to identify heterogeneous data (whether among sites or lineages) aids in investigating the causes of systematic bias and subsequently selecting the data most appropriate for the model (Lemmon and Lemmon, 2013). Once identified, heterogeneous data

can be excluded or analysed separately (Jermiin *et al.*, 2017; Naser-Khdour *et al.*, 2019). Sites can also be sorted according to different criteria, such as composition (Viklund *et al.*, 2012; Muñoz-Gómez *et al.*, 2019) or site-specific rates (Brinkmann and Philippe, 1999; Goremykin *et al*., 2010; Cummins and McInerney, 2011) and a proportion removed to improve the fit of the data to the model. An improvement in the model's robustness to systematic bias is expected when sites that fit the model poorly are removed. Moreover, one can also explore the effects on the tree reconstruction of the different heterogeneous signals (Fleming *et al.*, 2023). Artefacts induced by model misspecification can also be reduced by data recoding, despite the loss of information the recoding involves. These approaches compress the data into fewer states and thereby simplify the data. For example, in the RY recoding method nucleotides can be recoded as purines (R) and pyrimidines (Y) so that only transversion events are considered for phylogenetic reconstruction (Woese *et al.*, 1991; Phillips and Penny, 2003; Phillips *et al*., 2004). Since transversions evolve more slowly than transitions, they exhibit fewer numbers of multiple substitutions at the same sites and have more balanced compositions than transitions, which promotes a better model fit. Data recoding strategies have also been applied to protein data to reduce compositional heterogeneity and substitution saturation. Amino acids are usually grouped based on empirically observed substitution rates (Hrdy *et al*., 2004; Kosiol *et al*., 2004; Susko and Roger 2007). For instance, the well-known six-state Dayhoff-6 recoding was suggested to reduce potential artefacts caused by amino-acid composition differences among taxa (Feuda *et al.*, 2017). Nevertheless, recoding strategies and their merits are still under active debate (Hernandez and Ryan, 2021; Giacomelli *et al.*, 2022; Foster *et al.*, 2023).

Additional approaches have been designed to identify problematic sequences, often referred to as ‘rogue taxa’. These taxa are described by their unstable placement in the phylogenetic tree because of ambiguous or insufficient information and can be identified using bootstrap methods (Wilkinson, 1996; Aberer *et al*., 2013).

1.3.4 Tree reconstruction errors caused by biological processes

Evolutionary processes acting at the population level, such as variations in the number of alleles, their fixation, and sorting across loci, can lead to discrepancies between the evolutionary history of specific loci and the species tree (Maddison & Knowles, 2006). This is due to the retention and differential sorting of alleles (polymorphisms) into different descendant lineages, a phenomenon known as incomplete lineage sorting (ILS). The multispecies coalescent model enables estimating the species tree despite the conflicting locus trees, by modelling the probability of genes coalescing (Pamilo and Nei, 1988; Rannala and

Yang, 2003). The coalescent describes the process of lineages merging as the species genealogy is traced backwards in time (converging in their common ancestor). The probability of gene coalescences is proportional to the effective population size and branch lengths. The conflict between the species tree and locus trees is expected to increase with larger ancestral population sizes and shallower divergences (Degnan and Rosenberg, 2006, 2009). Summary coalescent or two-step methods, such as those implement in ASTRAL (Accurate Species TRee Algorithm; Mirarab *et al*., 2014; Zhang *et al*., 2018; 2020) or MP-EST (Liu *et al*., 2010) are often used to infer the species tree while accommodating ILS. These methods rely on using pre-computed gene trees to estimate the species tree as a computational convenience and are therefore referred to as a "short-cut" methods. However, gene trees, especially those derived from short alignments, are susceptible to topological errors caused by stochastic and systematic biases, which is problematic for these methods (Gatesy and Springer, 2014; Richards *et al*., 2018). On the other hand, fully hierarchical multi-species coalescent models, as implemented in BPP (Yang 2015; Flouri *et al*., 2018) or BEAST (Heled and Drummond, 2010) fall within single-step methods and jointly estimate the gene and species trees, minimizing the probability of gene tree estimation errors. However, these implementations require a greater computational complexity than the short-cut methods (Shi and Yang, 2018).

The locus trees can also differ from species trees due to biological processes other than ILS. Such processes included horizontal gene transfer, introgression, and gene loss after gene duplication. Horizontal gene transfer is the transfer of genetic material directly from one species to another, which in the context of a phylogeny is seen as lateral inheritance rather than vertical inheritance. Introgression, the transfer of genes between species, and hybridisation, cross-breeding between different species, also involve the lateral transfer of genetic material and can also result in the same conflict between gene and species tree phylogenies (Fitch, 1970; Doyle, 1997; Galtier and Daubin, 2008). Likewise, gene duplication followed by gene loss can also be problematic for the construction of species trees if paralogous, and not orthologous, copies of the gene are used for tree inference. The effects of horizontal gene transfer, gene duplication, and gene losses after duplication on phylogenetic reconstruction can be assessed using reconciliation methods which estimate gene and species trees simultaneously by considering these underlying processes (Boussau *et al*., 2013; Szöllősi *et al*., 2013).

1.3.5 Phylogenetic tree reconstruction

The phylogenetic analysis of multi-locus data typically follows one of two main strategies: simultaneous analysis of concatenated genes or analysis using a multispecies coalescent framework. Incongruence in resulting species trees is often found between the two methods (e.g. Wickett *et al*., 2014) which makes establishing a robust species tree difficult (Bravo *et al*., 2019; Shen *et al*., 2021). In the reconstruction of deep ancestral relationships, factors such as homoplasy, heterogeneity, and weak historical signals typically have a greater impact on inference analyses than processes like ILS (Bryant and Hahn, 2020). On the other hand, coalescence methods are more effective at reconstructing accurate trees for recent speciation events.

Encountering competing phylogenetic hypotheses requires caution against a summary consensus result that disregards the conflicting phylogenies. Often it is easier to reconstruct an incorrect phylogenetic tree than to accurately identify the true species phylogeny, especially when specific relationships are difficult to reconstruct. In such cases, a "better" analysis should take precedence over multiple analyses that yield congruent but potentially a biased phylogeny. In practice, to effectively assess the phylogenetic hypothesis under study, one needs to understand the extent to which the methods suit the study data. This involves considering the data's known properties and those that have yet to be assessed (Cox, 2018). Therefore, following a comprehensive strategy that integrates multiple approaches and investigating the discrepancies across results and considers their distinct methodological assumptions is often a better strategy (Lozano-Fernandez, 2022).

1.4 Objectives and thesis structure

The objective of the work conducted in this thesis is to improve the accuracy of phylogenetic trees by using substitution rate models that are specific to the data under analysis and consequently more robust to systematic biases than trees constructed using conventional empirical models. Additionally, I address the impact of among-lineage heterogeneity by applying methods for identification of tree-heterogeneous data and analysing the impact of their exclusion from subsequent analyses. I also aim to mitigate the effects of systematic bias by implementing better-fitting, more complex models, and evaluating strategies for exclusion of problematic data sorted by appropriate criteria.

The analyses performed in the technical chapters are conducted on proteins as amino acids evolve substantially slower than the underlying nucleotide data and are deemed more appropriate data to reconstruct ancient relationships. Subsequently, the objective is to

reconstruct relationships among Streptophyta using new and better-fitting amino-acid substitution models.

Chapter II evaluates current methods to estimate amino-acid substitution models from empirical data. With the availability of increased amounts of data and improved computational performance, phylogeneticists now have the possibility to define and deploy data-specific protein substitution models instead of choosing from pre-computed models. However, to date, the accuracy of existing methods for calculating amino-acid substitution models has not been fully evaluated. In this study the ability of four widely software to calculate amino-acid substitution models similar to an original simulation model are evaluated: namely, 1) IQ-TREE (Nguyen *et al.*, 2015), 2) FastMG (Dang *et al.*, 2014), 3) PAML (vers. 4.9i; Yang, 1997), and 4) P4 (vers. 1.3, Foster, 2004). IQ-TREE, PAML, and FastMG use an ML estimation method, while P4 can use ML and Bayesian MCMC procedures to estimate a substitution model. To test the accuracy of methods, data is simulated using a known amino-acid model and a pre-defined phylogenetic tree. Thereafter, the software is tested with respect to their ability to calculate an amino-acid substitution model similar to the original simulation model using the simulation data and a given tree. Data-specific amino-acid models are also estimated from published data sets and compared with the best-fitting model employed in the original study. Data-specific models are expected to have a better fit and be more robust to systematic biases than commonly used and poorer-fitting models.

In Chapter Three the objective is to assess methods that identify tree-heterogeneous data. The matched-pairs tests of symmetry, marginal symmetry, and internal symmetry can identify tree-heterogeneous sequences that reject the assumptions of homogeneity and stationarity. Nevertheless, as multi-comparison and statistical significance tests, the resulting p-values are required to be adjusted using procedures such as the Bonferroni, Holm, Benjamini-Yekutieli, and Benjamini-Hochberg (Holm, 1979; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). This study uses simulated and empirical data sets to test the matched-pairs tests of symmetry combined with the latter p-value adjustment methods regarding their ability to correctly identify tree-heterogeneous sequences. Additionally, the effect of tree heterogeneity in well-debated phylogenies, namely Bryophyta, Archaeplastida, and Metazoa, is analysed. This chapter enriches the general understanding of the impact of systematic bias caused by among-lineage heterogeneity in phylogenetic inference. It explores and recommends methodologies capable of mitigating tree-heterogeneous biases.

The objective of Chapter Four is to reconstruct the phylogenetic relationships among Streptophyta, with special emphasis on the difficult to resolve divergence between land plants

and green algae. Green algae gave rise to land plants approximately 470 Mya ago (Lewis and McCourt, 2004). However, the precise sister-group lineage to land plants has been controversial with different phylogenetic analyses placing Zygnematophyceae, Charophyceae, or Coleochaetophyceae as the sister-group lineage to land plants. This Chapter aims to integrate the findings from previous chapters by evaluating the effectiveness of the methodological approaches described therein on these data. Data-specific amino-acid substitution models have a better fit than the empirical models and would be expected to enhance the robustness of phylogenetic hypotheses against biases in the inference of the origin of land plants. Additionally, by using site- and tree-heterogeneous models, site-partitioning schemes, or removal of fast-evolving sites, or tree-heterogeneous sequences can increase the accuracy of tree-inference by reducing the effects of model misspecification-induced artefacts. All these strategies are explored in Chapter Four. It is expected that these comprehensive analyses of the nuclear, chloroplast, and mitochondria data, and the use of strategies to mitigate the effects of systematic bias, would enable the reconstruction of a robust hypothesis for the sister-group relationship to land plants.

Chapter Five provides a discussion of the main findings of this thesis. I discuss the major conclusions and establish connections and interrelations between the arguments of the preceding chapters. I analyse the use of data-specific substitution rate models in the different chapters and strategies used to explore the data signals and mitigate systematic bias, particularly those biases caused by tree-heterogeneous data.

Chapter II

Data-specific substitution models improve protein-based phylogenetics

Brazão, J. M., Foster, P. G., & Cox, C. J. (2023). Data-specific substitution models improve protein-based phylogenetics. *PeerJ*, *11*, e15716. https://doi.org/10.7717/peerj.15716

Data-specific substitution models improve protein-based phylogenetics

Abstract

Calculating amino-acid substitution models that are specific for individual protein data sets is often difficult due to the computational burden of estimating large numbers of rate parameters. In this study we tested the computational efficiency and accuracy of five methods used to estimate substitution models, namely Codeml, FastMG, IQ-TREE, P4 (maximum likelihood), and P4 (Bayesian inference). Data-specific substitution models were estimated from simulated alignments (with different lengths) that were generated from a known simulation model and simulation tree. Each of the resulting data-specific substitution models was used to calculate the maximum likelihood score of the simulation tree and simulated data that was used to calculate the model and compared with the maximum likelihood scores of the known simulation model and simulation tree on the same simulated data. Additionally, the commonly-used empirical models, cpREV and WAG, were assessed similarly. Data-specific models performed better than the empirical models, which under-fitted the simulated alignments, had the highest difference to the simulation model maximum-likelihood score, clustered further from the simulation model in Principal Component Analysis ordination, and inferred less accurate trees. Data-specific models and the simulation model shared statistically indistinguishable maximum-likelihood scores, indicating that the five methods were reasonably accurate at estimating substitution models by this measure. Nevertheless, tree statistics showed differences between optimal maximum likelihood trees. Unlike other model estimating methods, trees inferred using data-specific models generated with IQ-TREE and P4 (maximum likelihood) were not significantly different from the trees derived from the simulation model in each analysis, indicating that these two methods alone were the most accurate at estimating data-specific models. To show the benefits of using data-specific protein models several published data sets were reanalysed using IQ-TREE-estimated models. These newly estimated models were a better fit to the data than the empirical models that were used by the original authors, often inferred longer trees, and resulted in different tree topologies in more than half of the re-analysed data sets. The results of this study show that software availability and high computation burden are not limitations to generating better-fitting data-specific amino-acid substitution models for phylogenetic analyses.

2.1 Introduction

Maximum likelihood (ML) and Bayesian inference (BI) phylogenetic methods include a set of assumptions about the evolutionary process of change in molecular sequences (amino acids, nucleotides, or codons) that are specified by a substitution rate model. The resulting phylogenetic trees (topology and branch lengths) are dependent on the model, and a poor fit of the model to the data will affect the accuracy of tree reconstruction (Keane *et al*., 2006; Cox and Foster, 2013). Substitution rates at sites are assumed to follow a continuous-time Markov chain model: sites evolve independently of each other through time and are described by a probability of change at any particular site, where the states of the chain, and the probability of change from the current state, do not depend on the past states (Felsenstein, 2004; Yang, 2014). For analyses of proteins, an amino-acid substitution model is usually expressed as a 20x20 instantaneous rate matrix, where each off-diagonal element is the product of the relative rate of exchange between amino acids and the equilibrium frequency of the resulting amino-acid (Swofford *et al*., 1996). Typically, only 189 instantaneous rate parameters are considered, because the evolutionary process is assumed to be reversible at the same rate (a time-reversible process). Change in the substitution rate among sites is typically accommodated by modelling the distribution of rates with a discrete gamma-distribution (Yang, 1994).

During phylogenetic tree reconstruction it is expected that the substitution rates specified in the model are a good fit to the evolutionary process underlying change in the sequence data being analysed. Traditionally, large sets of proteins were used to calculate general models which were then used to analyse new data. These general-fitting, empirical, models were generated for analyses of particular genomes, or taxon groups. The use of empirical models is especially important when the new data to be analysed are of limited size and therefore unlikely to be sufficient to estimate all the substitution model rate parameters. Moreover, in the past, analyses of large data sets appropriate for calculating data-specific substitution rates, imposed a considerable computational burden. Nevertheless, despite today there being a general increase in the size of protein data sets used in phylogenetics, and the availability of faster computers with more efficient algorithms for calculating substitution models, the use of pre-computed, empirical models for the analysis of amino-acid sequences is still almost ubiquitous in phylogenetic practice.

The first widely-used amino-acid substitution models, namely the Point Accepted Mutation (PAM) matrices (Dayhoff *et al*., 1978) and the JTT model (Jones *et al.*, 1992), were derived from nuclear protein data using parsimony counting methods. By contrast, the

mitochondrial mtREV model (Adachi and Hasegawa, 1996) and the chloroplast cpREV model (Adachi *et al*., 2000), were estimated using ML but on smaller data sets (< 100000 amino-acids) due the computational burden of optimising both substitution rates and branch lengths simultaneously. The nuclear genomic models, WAG (Whelan and Goldman, 2001) and LG (Le and Gascuel, 2008), were also estimated using ML but with far larger data sets (~900,000 and ~ 6.5 million amino-acid residues respectively), although optimisation procedures were simplified again to reduce computational burden. The only known amino-acid substitution model to have estimated using Bayesian Markov Chain Monte Carlo (MCMC) routines is the gcpREV calculated for chloroplast data of Streptophyta plants (Cox and Foster, 2013). The gcpREV model was shown to have a better fit to Streptophyta plant data than the more general plant cpREV model which was calculated with the inclusion of red algae. Likewise, protein-specific substitution models were found to be a better fit to protein virus data than the available empirical substitution models (Del Amparo and Arenas, 2022). These analyses demonstrate a common expectation that a model calculated specifically for the data-at-hand is going to be a better fit to the data than a more general-fitting model which might have wider application but less specificity, and therefore result in a theoretically better justified phylogenetic hypothesis.

With the ever-increasing availability of sequence data due to improvements in sequencing technologies, allied with the increased computational performance of modern computers, more than ever before phylogeneticists have the possibility of calculating and using data-specific protein substitution models instead of choosing from pre-computed empirical models. However, to date, the accuracy of existing methods for calculating amino-acid substitution models has not been evaluated collectively. In this study, we assess five methods for their ability to estimate accurate amino-acid substitution models: the methods are implemented in the programmes Codeml (PAML; Yang, 1997, 2007), FastMG (Dang *et al*., 2014), IQ-TREE (Nguyen *et al*., 2015), and P4 (Foster, 2004). Each method estimates the 189 free rate parameters of a general time-reversible model using ML optimisation procedures. P4 can also estimate a substitution model using Bayesian methodology by sampling parameters from the posterior distribution of a MCMC.

To test the efficiency of the five methods, amino acid sequence data were simulated using a known amino-acid model (gcpREV) and a specified phylogenetic tree. Thereafter, the methods were tested with respect to their ability to calculate an amino-acid substitution model similar to the original simulation model. The new data-specific models were compared to the simulation model with respect to the likelihood scores of the simulation tree, clustering

distance (Principal Component Analysis (PCA)), and the similarity of the topology and branch lengths of reconstructed optimal trees with respect to the simulation tree. The analyses using data-specific models were also compared to those using the commonly-used empirical models (cpREV and WAG). Additionally, data-specific amino-acid models were estimated for a set of published phylogenetic analyses and the results using the new models compared with the results from the chosen models used in the original studies.

2.2 Methods

Analyses of simulated sequence data using data-specific models and assessing five model estimation methods

Amino-acid sequence alignments were generated in P4 (vers. 1.3) by simulating sequence evolution using the gcpREV substitution model and a 26 taxon tree (taken from Sousa *et al*., 2019) with fixed branch lengths (total tree length 7.74 expected numbers of substitutions per site). The simulation process consisted of generating a random root sequence and evolving it over the simulation tree under the process specified by a simulation model (Foster, 2004). The gcpREV substitution model was used as simulation model. The root sequence had the model-specified gcpREV composition (F_{mod}) and sites were evolved under a discrete gamma-distribution of among-site rate variation, discretised with 4 categories (Γ_4). 100 simulated alignments were generated with each of 400, 1500, and 8000 site lengths, which corresponds to a mean expected number of substitutions per branch of 63.2, 236.9, and 1263.6 respectively. It should be noted that we do not consider the effect of incomplete sequences on model selection or reconstruction in this study.

The methodology used to assess the accuracy of the methods for calculating data-specific models is shown in Figure A1. Data-specific amino-acid substitution models were estimated using ML optimisation in Codeml (PAML, vers. 4.9i), FastMG (vers. beta), IQ-TREE (vers. 1.6.8), and P4 (vers. 1.3), and by BI in P4. Model parameters were calculated for a general time-reversible model (GTR; Tavaré 1986), with among-site rate variation (Γ_4), and optimised composition frequencies (F_{est}), except when using Codeml where the composition frequencies are set to the empirical values of the data (F_{emp}). Model parameters were estimated using the simulation tree as a constraint (both topology and branch lengths) to reduce the variability in the methodology and enable a more direct comparison of calculating methods. Similarly, the method used in FastMG did not include the "alignment split algorithm" (use to reduce computation burden on calculating ML trees of large data sets) so that the estimation could be constrained to the simulation tree. For the Bayesian estimation of

substitution model parameters MCMC analyses were run using P4 for 600,000 generations sampling parameter values every 100 generations, for a total of 6,000 samples. Of the posterior distribution samples, 1,000 were dis-carded as "burn-in", and the substitution rates calculated as the mean values of the remaining 5,000 samples.

The data-specific models estimated using the five software methods are logical equivalents to the empirical models (WAG, gcpREV, and cpREV) to which they were compared in this study. In other words, all model comparisons had the same numbers of parameters (190 amino-acid fixed exchange rate parameters and 20 fixed amino-acid frequency parameters, and one free parameter for the alpha variable of Γ_4). The AIC (Akaike, 1973) or BIC (Schwarz, 1978) metrics are often used to compare substitution models when the numbers of parameters vary between models. However, because all the models compared in this study had the same number of parameters, the model-fit to the data was assessed using the log likelihood scores.

The estimated data-specific models and the commonly-used empirical models (cpREV and WAG) were visually compared to the simulation model (gcpREV) via ordination using PCA, after first normalizing the rate parameters of each data-specific model were normalized by dividing each value by the sum of the 189 parameters. Mean rate parameters were then calculated from each set of data-specific models calculated using each of the five methods and each of the three alignment lengths (400, 1500, and 8000 sites).

Data-specific models were compared to the simulation model with respect to their fit to the data. ML scores were calculated using IQ-TREE for each data-specific model (with Γ_4 and F_{mod}) on the same alignments the models were derived from, with the simulation tree as a constraint (topology and branch lengths). Similarly, ML scores using the equivalent models were calculated for gcpREV, cpREV, and WAG empirical models. The cpREV was chosen because it was determined by ModelFinder (implemented in IQ-TREE; Kalyaanamoorthy *et al.*, 2017) as the best-fitting empirical model to the simulated alignments. By contrast, the WAG model, being derived from nuclear data, would be expected to have a lower fit to the data than cpREV. The statistically significant differences between ML scores were assessed using a two-tailed independent t-test (Student, 1908). Because the latter test assumes a normal distribution of the data, ML scores were tested using the Shapiro-Wilk test, which assessed the normality of the data (Shapiro and Wilk, 1965). The t-test null hypothesis assumed equal score means between the simulation model analyses and each data-specific model, cpREV, and WAG analyses. The null hypothesis was rejected with a p-value (P) significant at < 0.05. However, an not-rejected null hypotehsis using log-transformed data (where each variable x is

replaced by log(x)), does not necessarily imply the same for the untransformed values, mainly when the variances of the log-transformed data are unequal (Zou *et al*., 1997). Nevertheless, the analyses of log likelihood scores using the F-test of equality of variances (assess whether the variances between samples are equal) did not reject the assumption of equality of the variances between the gcpREV scores and each model scores, indicating that the analyses of the log likelihood values are likely in agreement with the untransformed likelihood values.

Tree reconstruction accuracy was tested for each data-specific model (with Γ_4 and F_{mod}) by comparing optimal ML tree topologies and branch lengths reconstructed with IQ-TREE with the simulation tree. The topological accuracy was assessed using the unweighted Robinson-Foulds (RF; Robinson and Foulds, 1981) and weighted Robinson-Foulds (WRF; Robinson and Foulds, 1979) tree metrics, as implemented in P4. The RF measures the number of branches that differ between the trees, while the WRF distance is the sum of the differences between all branch lengths. Additionally, the tree length (the sum of all branch lengths) of each optimised ML tree was compared to the simulation tree length. The statistically significant differences between tree distances were assessed using a two-tailed independent t-test, where the null hypothesis assumed a WRF mean distance and mean tree length difference equal to the simulation model tree results. The null hypothesis was rejected with a P significant at < 0.05. The assumption of normality of the data was assessed using the Shapiro-Wilk test. When the latter was rejected, the Wilcoxon test (Wilcoxon, 1945) was used instead of the t-test. The Wilcoxon test is a non-parametric test used to assess the null hypothesis of equality of the score means under the non-normality assumption.

Re-analysis of published studies using data-specific models

To determine whether the use of data-specific amino-acid models would likely impact the results of phylogenetic analyses of empirical data, data-specific substitution models were estimated from published data sets using a GTR (with Γ_4 and F_{est}) model in IQ-TREE (Appendix Table A1). Optimal ML trees were inferred from the published data sets using the data-specific models with the same among rate site variation parameters as used in the original studies. The resulting ML trees were compared to the original published trees with respect to their ML score, topological distance, and total tree length. The topological distances were calculated as a normalized RF (nRF) distance metrics (RF/RFmax, where RFmax is obtained by 2*(number of taxa-3); Kupczok *et al*., 2008). Where optimal ML trees and their scores were not available from the original publication, they were computed using the published trees and the original model.

The data products (protein alignments, calculated substitution models, and trees) are available from Zenodo: João Brazão. (2023). Data-specific substitution models improve protein-based phylogenetics—data [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7628408. The novel scripts used to make calculations and a machine actionable RO-Crate metadata specification are available from GitHub: https://github.com/joaobrazao/Data-specific-substitution-models-improve-protein-based-phylogenetics.

2.3 Results

Simulated sequence data analyses and assess five methods for estimating substitution models

Simulated sequence alignments were used to test the accuracy of five methods to generate data-specific amino-acid substitution models. The data-specific models were assessed by comparison to the simulation model with respect to ML scores calculated on the alignments the data-specific models were derived from, when using the simulation tree as a constraint (Table 2.1; Appendix Fig. A2). The ML scores from each set of analyses had a normal distribution according to the Shapiro-Wilk test and equal variance to the simulation model values. All five methods resulted in data-specific models that were a close fit to simulation model (gcpREV), that is, the mean ML scores of each set of alignment lengths were similar to the ML score of the simulation model. Besides, apart from those data-specific models of 8000-site alignments estimated using FastMG, ML mean scores of data-specific models from all methods were higher than the simulation model mean scores. Data-specific models estimated using the P4-BI and FastMG method had the lowest difference to the simulation model score. The analyses using the P4-BI-estimated models had a mean score difference from the simulation model by 35, 55, and 82 log likelihood units, when inferred from 400, 1500, and 8000-site alignments respectively, while FastMG-estimated model analyses had a mean score difference by 73, 45, and -21 units. The P4-ML-estimated models had the highest mean scores and varying between 91 and 102 likelihood units when compared to gcpREV mean likelihood scores. The differences of the Codeml- and IQ-TREE-estimated models to the simulation model were between 83 and 93 and between 81 and 88 likelihood units, respectively. Nevertheless, the likelihood scores resulting from data-specific model analyses were not significantly different from simulation model scores. By contrast, the mean ML scores derived from commonly-used empirical models were significantly lower than the simulation model scores ($P<0.05$ as assessed by a two-tailed t-test; Table 2.1), apart of the

cpREV analyses using the 400-site alignments. The mean likelihood scores calculated using the cpREV model deviated from the gcpREV analyses by -88, -332, and -1770 units when inferred from 400, 1500, and 8000-site alignments respectively, while WAG mean scores deviated by -243, -903, and -4837 likelihood units.

Table 2. 1 - ML scores were calculated using data-specific models (+Γ_4+F_{mod}) with the topology and branch lengths constrained to those of the simulation tree used to derive the simulated data. ML scores using the equivalent models were calculated for gcpREV, cpREV, and WAG empirical models for comparison. The null hypothesis of no difference between the mean ML scores of the simulation model (gcpREV) and the mean scores resulting from the estimated models, cpREV, and WAG was rejected with a p-value (P) significant at < 0.05 (*), under a two-tailed t-test.

Amino-acid substitution models (+Γ_4+F_{mod})	**Mean ML scores of the simulation tree for 100 simulated data sets (log likelihood units)**		
	400 sites (Δ simulation model)	**1500 sites (Δ simulation model)**	**8000 sites (Δ simulation model)**
gcpREV (simulation model)	-10,781	-40,417	-215,294
cpREV	-10,869 (-88; 0.10)	-40,749* (-332; 0.0)	-217,064* (-1,770; 0.0)
WAG	-11,024* (-243; 0.0)	-41,318* (-902; 0.0)	-220,131* (-4837; 0.0)
Codeml-estimated models	-10,698 (83; 0.12)	-40,326 (91; 0.37)	-215,201 (93; 0.63)
FastMG-estimated models	-10,708 (73; 0.17)	-40,372 (45; 0.66)	-215,315 (-21; 0.91)
IQ-TREE-estimated models	-10,700 (81; 0.13)	-40,329 (88; 0.38)	-215,209 (85; 0.66)
P4-ML-estimated models	-10,690 (91; 0.09)	-40,317 (100; 0.32)	-215,191 (102; 0.59)
P4-BI-estimated models	-10,745 (35; 0.52)	-40,361 (55; 0.58)	-215,212 (83; 0.67)

Accuracy of the estimated amino-acid models was also assessed by visually inspecting a PCA ordination between the gcpREV simulation model, cpREV and WAG empirical models, and the mean rates of the estimated data-specific models (Fig. 2.1). Mean-rate models were calculated from the mean of the normalised rates of data-specific models according to each alignment length. The cpREV and WAG models were two of the three most distant

models from the simulation model (the other being the P4-BI-estimated 400 sites model), corroborating the likelihood score comparison analyses. By contrast, the mean models calculated from the Codeml- and P4-ML-estimated models clustered closest to the simulation model. The P4-BI-estimated mean rate models computed using the 1500 and 8000-site alignments were also relatively close to the simulation model, while the IQ-TREE derived models were more distant. FastMG mean models were clustered together further than the other mean rate models. Generally, the longer the simulated alignment, the closer the estimated model was to the simulation model in the ordination. However, this was not the case for Codeml- and FastMG-estimated models where the length of the simulated data had less effect on the overall accuracy of the estimated models.

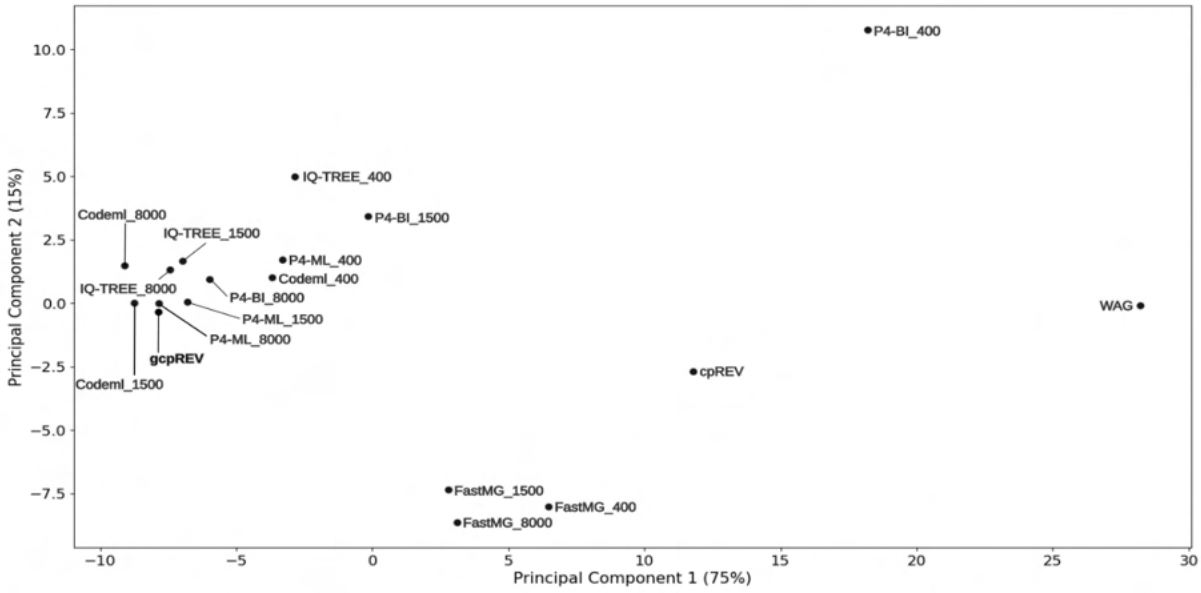


Figure 2. 1 - Principal Component Analysis of model exchange values. Ordination of the mean values of the normalised exchange rate model parameters of the data-specific models (named according to the calculation method and simulated data length), the gcpREV simulation model, and the cpREV and WAG empirical models.

The estimated data-specific models were also assessed with respect to their ability to reconstruct the topology and correctly estimate the branch of the simulation tree during and unconstrained tree search (Appendix Table A2). The optimal ML trees inferred from 1500 and 8000 site simulated alignments always recovered the correct topology, i.e., the same as the simulation tree. However, the trees inferred from the 400-site alignments using data-specific models failed in 19-22% of replicates to recover the correct topology, having a RF mean varying between 0.38 and 0.44. The cpREV and WAG derived trees topologically different to the simulation tree 24 (mean RF 0.48) and 25 times (mean RF 0.54) respectively,

whereas the simulation model itself failed to recover the simulation tree 22 times (mean RF 0.44).

The WRF mean distances between the optimal ML trees and the simulation tree decreased as the length of the data set increased. The Shapiro-Wilk test indicated the normality of the data for each set of WRF distances, except for the 1500-site alignment analyses using the FastMG-estimated models. The cpREV and WAG derived trees had the highest WRF distances and were significantly different from the simulation (gcpREV) model derived statistics ($P<0.05$; Fig. 2.2; Table 2.2). The WRF distances computed using the FastMG-estimated models had a higher difference to WRF means derived from the simulation model than the remaining data-specific models at all data lengths, and statistically higher when computed using the 1500- and 8000-site alignments. The P4-ML-estimated models had the lowest WRF mean distance within the analyses using the 1500-site alignments. The analyses of the 8000-site alignments using the IQ-TREE- and P4-estimated (ML and BI) models shared the lowest WRF mean distances overall and the closest to the simulation model. The WRF distances resulting from Codeml-estimated model analyses were the closest to the gcpREV values, when using the 400- and 1500-site alignments. Nonetheless, the WRF mean distances computed using the optimal trees derived from the Codeml-, IQ-TREE-, and P4-estimated models were not significantly different to the simulation model results.

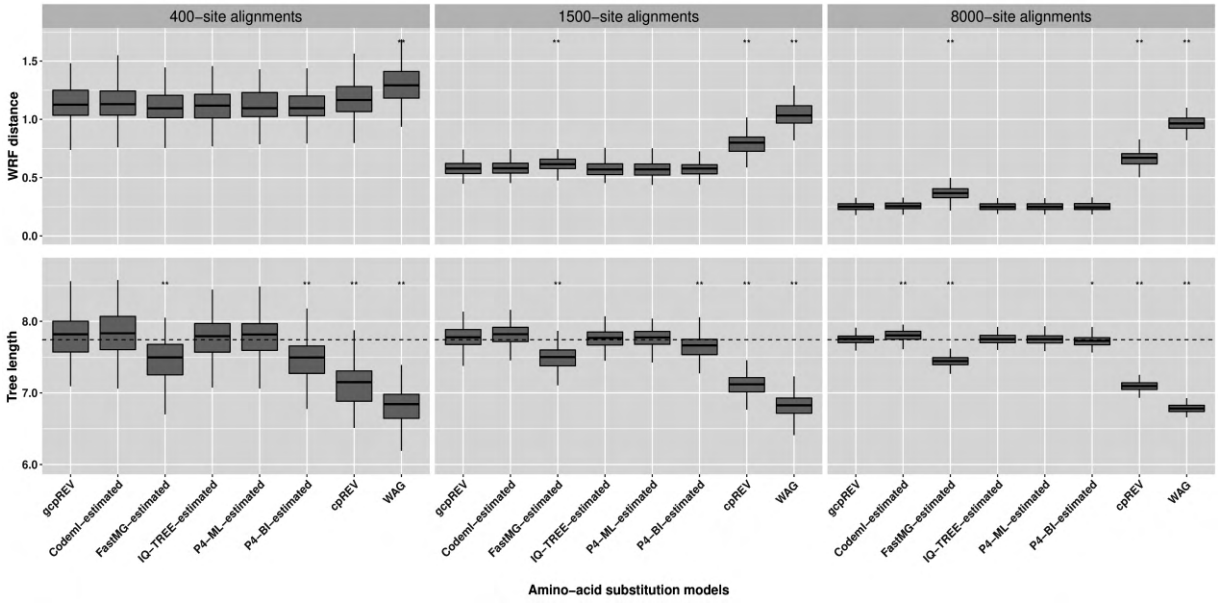


Figure 2. 2 - Box-plots of weighted Robinson-Foulds distances between the simulation tree and the optimal ML trees and the optimal tree lengths. The ML trees were inferred using data-specific models, the gcpREV (simulation model), cpREV, and WAG ($+\Gamma_4+F_{mod}$). The simulation tree length is 7.74 substitutions per site (dashed line). The statistical test evaluates the null hypothesis of no difference between the means of the tree metrics of the optimal trees derived using the simulation model and the means of tree metrics derived from the estimated models, cpREV, and WAG. The null hypothesis was rejected with a p-value (P) significant at < 0.05 (*) and < 0.01 (**), under a two-tailed t-test. The weighted Robinson-Foulds distances and tree length differences computed from the 1500-site alignments using the FastMG-estimated and P4-BI-estimated models, respectively, were assessed using the Wilcox test because the assumption of normality was rejected according to the Shapiro-Wilk test.

Table 2.2 - Weighted Robinson-Foulds (WRF) distances between the simulation tree and the optimal ML trees and the optimal tree lengths. The ML trees were inferred using data-specific models, the gcpREV (simulation model), cpREV, and WAG ($+\Gamma_4+F_{mod}$). The null hypotheses of no difference between the means of the tree metrics of the gcpREV derived trees and the means of tree metrics derived from the estimated models, cpREV, and WAG were rejected with a p-value (P) significant at < 0.05 (*), under a two-tailed t-test or under the Wilcox test (a) where the assumption of normality was rejected according to the Shapiro-Wilk test (substitutions=subs).

Data sets	Amino-acid substitution models	Mean WRF	Δ gcpREV mean WRF		Mean optimal tree length	Δ simulation tree length	Δ gcpREV optimal mean length	
		Subs/site	Subs/site	P	Subs/site	Subs/site	Subs/ site	P
400 sites	gcpREV	1.135	-	-	7.797	0.054	-	-
	cpREV	1.179	0.044	$5x10^{-02}$	7.122	-0.621	-0.675	$3x10^{-36}$ *
	WAG	1.312	0.177	$4x10^{-12}$ *	6.819	-0.924	-0.978	$5x10^{-57}$ *
	Codeml-estimated	1.140	0.005	$8x10^{-01}$	7.828	0.085	0.031	$5x10^{-01}$ *
	FastMG-estimated	1.112	-0.023	$3x10^{-01}$	7.452	-0.291	-0.344	$3x10^{-12}$ *
	IQ-TREE-estimated	1.117	-0.018	$4x10^{-01}$	7.756	0.013	-0.040	$4x10^{-01}$
	P4-ML-estimated	1.116	-0.019	$4x10^{-01}$	7.790	0.047	-0.007	$7x10^{-01}$
	P4-BI-estimated	1.114	-0.021	$3x10^{-01}$	7.455	-0.288	-0.342	$2x10^{-13}$
1500 sites	gcpREV	0.582	-	-	7.780	0.037	-	-
	cpREV	0.799	0.217	$8x10^{-42}$ *	7.116	-0.627	-0.664	$5x10^{-75}$ *
	WAG	1.043	0.461	$1x10^{-72}$ *	6.797	-0.946	-0.983	$9x10^{-99}$ *
	Codeml-estimated	0.583	0.001	$9x10^{-01}$	7.811	0.068	0.032	$2x10^{-01}$
	FastMG-estimated	0.621	0.039	$1x10^{-04}$ *a	7.470	-0.273	-0.310	$3x10^{-25}$ *
	IQ-TREE-estimated	0.574	-0.008	$4x10^{-01}$	7.754	0.011	-0.026	$2x10^{-01}$
	P4-ML-estimated	0.573	-0.009	$4x10^{-01}$	7.769	0.026	-0.011	$4x10^{-01}$
	P4-BI-estimated	0.577	-0.005	$6x10^{-01}$	7.667	-0.076	-0.113	$3x10^{-08}$ *a
8000 sites	gcpREV	0.251	-	-	7.750	0.007	-	-
	cpREV	0.666	0.416	$1x10^{-102}$ *	7.094	-0.649	-0.656	$9x10^{-135}$ *
	WAG	0.968	0.717	$3x10^{-144}$ *	6.783	-0.960	-0.968	$2x10^{-167}$ *
	Codeml-estimated	0.258	0.007	$1x10^{-01}$	7.803	0.060	0.052	$2x10^{-06}$ *
	FastMG-estimated	0.368	0.118	$2x10^{-40}$ *	7.445	-0.298	-0.306	$2x10^{-73}$ *
	IQ-TREE-estimated	0.249	-0.002	$8x10^{-01}$	7.753	0.010	0.002	$8x10^{-01}$
	P4-ML-estimated	0.249	-0.001	$8x10^{-01}$	7.749	0.006	-0.001	$9x10^{-01}$
	P4-BI-estimated	0.249	-0.001	$8x10^{-01}$	7.725	-0.018	-0.025	$1x10^{-02}$ *

The optimal ML trees inferred using the cpREV and WAG models had the highest tree length differences to the simulation tree, did not converge with the simulation tree with longer alignments and were significantly shorter ($P<0.05$ as assessed by a two-tailed t-test) than gcpREV derived trees (Table 2.2; Fig. 2.2). The mean length of the optimal trees inferred using the Codeml-, IQ-TREE-, and P4-estimated (ML and BI) models converged with the simulation tree length, as the alignment length increased. The optimal trees resulting from IQ-TREE- and P4-ML-estimated models were the closest to the simulation tree and gcpREV derived trees ($P>0.05$) about the mean tree length. The Codeml-estimated models also inferred tree lengths close to the simulation length and not significantly different from the gcpREV derived trees, except when computed using the 8000-site alignments. The optimal trees inferred using the P4-BI estimated models were shorter than the simulation tree and significantly different from the gcpREV derived trees, regardless the alignments length. Nevertheless, their lengths converged strongly with the simulation tree and simulation model trees, as the alignment length increased. The optimal tree lengths inferred from the 1500-site alignments using the later models were the only values to not have a normal distribution. The optimal ML trees inferred using the FastMG-estimated models were the shortest trees (within data-specific model analyses), significantly different from the gcpREV derived trees, and did not improve with longer alignments.

The estimation of amino-acid substitution models using P4-BI took by far the longest time, taking approximately 34, 68, and 322 hours, when inferred from 400, 1500, and 8000-site alignments respectively. By contrast, P4-ML (32-756 minutes), Codeml (14-265 minutes), IQ-TREE (17-155 minutes), and FastMG (1-3 minutes), to considerably less time to compute data-specific models (Appendix Table A3).

Re-analysis of published studies using data-specific models

The effectiveness of using data-specific models, estimated using IQ-TREE, was also assessed by reanalysing published empirical data sets and comparing them to commonly-used substitution models (Table 2.3). The re-analyses of the Timme (2012) and the Zeng (2014) data sets recovered the same topologies as the original published trees, though 2% and 13% longer, respectively, and with likelihood scores of -720,951 and -758,752 log likelihood units, improving on the published scores by 1,486 and 5,972 units, respectively. The optimal ML tree inferred from the Lelieart (2016) data set using a data-specific model had a score of -483,209 log likelihood units, 3,590 units higher than the original model derived tree score,

and was 3% longer with a topological distance of 0.05 (nRF) to the tree inferred using the original model. The re-analysis of the Chang (2015) data set using a branch-linked partition model and a data-specific model recovered a tree of score -2,305,041 log likelihood units. Feuda (2017) also reanalysed the same data set with a branch-linked partition model; however the published tree score was 16,788 units lower. The topology inferred using the data-specific model was the same as the published tree topology, but 16% longer. The ML tree inferred from the Munro (2018) data set using a data-specific model was longer 10% than the original published tree and had a score of -8,070,115 log likelihood units, an improvement of 43,580 units over the published tree score. The resulting tree topology was congruent with the original tree, except for the placement of *Erenna richardi* (nRF = 0.03). The optimal tree resulting from the Schulz (2018) data set using a data-specific model had a score of -470,898 log likelihood units, 1,407 units higher than the original published analysis. The tree was 10% shorter than the published tree and had a nRF distance of 0.02. The re-analysis of the Schwentner (2017) recovered an optimal ML tree 1% longer and with a topological distance of 0.05 (nRF) to the original published tree, and with a score of -443,135 log likelihood units, 1,375 units higher than the original tree score. The re-analyses of the Toussaint (2018) data set included a data-specific model (estimated from the concatenated data set) and different partition schemes, namely, a single partition, a by-locus partition scheme (366 partitions), and a PartitionFinder scheme (27 partitions; from the original study). The reanalysis of the concatenated data set (single partition) had the lowest score (-620,234 log likelihood units) of the data-specific model analyses. Nevertheless, it was higher than any ML tree score of the original analyses. The Bayesian information criteria of the partition analyses using data-specific models were also higher than the original analyses using the same scheme. Additionally, an optimal tree was inferred using the Partition-Finder scheme with separate data-specific models for each partition which had the highest fit to the data, -610,590 log likelihood units. The data-specific model derived trees had longer lengths, between 4% and 9%, and a nRF varying between 0.007 and 0.022 to the originally published trees. The re-analyses of Irisarri (2020) data set recovered a ML tree with a score of -114,615 log likelihood units, fitting the data better than the original model by 541 units. The resulting topology was to the same as the published tree, although 6.6% longer. The ML tree inferred from the Koenen (2020) data set was 4.2% longer than the original published tree and had a topological distance of 0.07 (nRF). The tree score was -571,563 log likelihood units, an improvement of 38,056 units over the published tree score.

Table 2.3 - ML score and length of the optimal trees inferred using data-specific models. Data-specific models were estimated using a GTR ($+\Gamma_4+F_{est}$) model in IQ-TREE. Normalised Robinson-Foulds (nRF), tree score and length differences were calculated using the optimal trees derived from the data-specific models and the trees derived from the original models.

Study	**Model**	**Likelihood tree score** (log likelihood units)	**Tree score improvement** (log likelihood units)	**Tree length** (Substitutions per site)	**Tree length variance** (Percentage of substitutions per site)	**nRF**
Timme *et al.*, 2012	Data-specific model $+\Gamma_4$	-720,951	1,486	4.7	2.0%	0
Zeng *et al.*, 2014	Data-specific model $+\Gamma_4+I$	-758,752	5,972	9.4	12.8%	0
Leliaert *et al.*, 2016	Data-specific model $+\Gamma_4$	-483,209	3,590	13.9	3.2%	0.05
Feuda *et al.*, 2017 (Chang data set *et al.*, 2015)	Branch-linked model + data-specific model	-2,305,041	16,438	19.2	15.9%	0
Munro *et al.*, 2018	Data-specific model $+R_7+F$	-8,070,115	43,580	9.2	10.1%	0.03
Schulz *et al.*, 2018	Data-specific model $+R_6$	-470,898	1,407	76.3	-10.2%	0.02
Schwentner *et al.*, 2017 (Matrix 1)	Data-specific model $+R_4$	-443,135	1,375	6.3	1.1%	0.05
Toussaint *et al.*, 2018 (DT369)	Data-specific model $+R_4$	-620,324	10,385	1.9	6.9%	0.01
	Branch-unlinked 366-partition + data-specific model	-615,428	9,259	2.5	3.9%	0.02
	Branch-unlinked 27-partition + data-specific model	-616,082	11,171	2.0	8.5%	0.01

	Branch-unlinked 27-partition + data-specific models/partition	-610,590	16,664	2.0	8.9%	0.01
Irisarri *et al*., 2020	Data-specific model +Γ_4+I	-114,615	540	20.8	6.6%	0
Koenen *et al*., 2020	Data-specific model +R_4	-571,563	38,056	7.2	4.2%	0.07

2.4 Discussion

In this study data-specific substitution models were calculated from simulated sequence data (simulated using a specified substitution model and tree) using five different estimation methods. An accurate model estimation method would be expected to estimate models that are similar to the model used to simulate the data, and result in phylogenetic trees that were similar to the simulation tree with respect to likelihood and topology. Our results showed that data-specific substitution models consistently out-performed empirical, commonly-used, pre-computed substitution matrices. We simulated data using the gcpREV model (Cox and Foster, 2013) which was computed for green plant chloroplast data and was intended to more accurately reflect the amino-acid substitution patterns found in green plant chloroplasts when compared to the commonly used cpREV substitution model (Adachi *et al*., 2000) that was estimated from green and non-green plant chloroplasts. The gcpREV model is not commonly-used, and is not used in any of the popular model selection software used to select best-fitting amino-acid substitution models (e.g. ProtTest; Abascal 2005). The gcpREV simulation model can therefore been seen as a single point in amino-acid substitution space; but it is not a random "distance" from any of the empirical models, just an arbitrary point. Had a simulation model more similar to one of the commonly-used empirical models been chosen, however, the results of the estimated models may not have been distinguishable from those of the empirical model. Nevertheless, given that all of the analyses of published studies (discussed below) showed that data-specific substitution models result in better-fitting models, we suspect that unless the data are "very close" to being modeled accurately by one of the commonly-used empirical models a data-specific model will be a significantly better fit. Putting aside a caveat regarding data size, given the efficiency at which data-specific exchange rate models can be generated, and their accuracy (see below), it seems unlikely that using an empirical exchange rate model can be justified no matter how closely the data fit the

model. The cpREV model was found to be better-fitting than the WAG model to all simulated data, as would be expected. However, the two empirical models, cpREV and WAG, were of much poorer fit to the data than the estimated data-specific substitution models. Indeed, compared to data-specific models, they inferred shorter trees with greater mean WRF and tree length differences to the simulation tree and simulation model (gcpREV) derived trees and failed more often to recover the correct (simulation) topology.

The ML scores of the simulated data on the simulation tree using the data-specific models estimated by each of the five methods were overall similar to each other (Table 2.1; Appendix Fig. A2) and were not statistically different from the ML tree scores of the simulated data using the simulation model. Nevertheless, the tree statistics of the optimal trees sometimes differed between the analyses using the data-specific substitution matrices estimated by the different methods and those estimated by the simulation model (Table 2.2; Fig. 2.2). Specifically, mean WRF distances of optimal trees under the FastMG-estimated model analyses using 1500- and 8000-site alignments were significantly different than the WRF values of the optimal trees derived when using the simulation model. Moreover, FastMG-estimated trees were significantly shorter than the optimal tree lengths derived from the simulation model at all data lengths. P4-BI-estimated models also inferred optimal trees that were significantly shorter than the optimal trees derived from the simulation model. By contrast, the remaining estimation methods (Codeml, IQ-TREE, and P4-ML) all estimated optimal trees with mean WRF and lengths statistically indistinguishable from the simulation model derived trees, except for optimal trees estimated by Codeml for 8000 length data sets which were significantly longer than those of estimated from the simulation model. Codeml uses empirical composition frequencies instead of optimised values, which may explain the lower accuracy compared to the IQ-TREE and P4-ML methods using the longest alignments where more data enable a more precise estimation of the composition.

Data-specific models inferred optimal trees with higher ML scores compared to trees from the simulation model (with the exception of trees inferred from the 8000-site alignments using FastMG-estimated models, Table 2.1). This is expected because the data-specific models had free and optimized parameters while the simulation model analysis did not. However, the increase in ML values was not significant for any of the estimation methods, suggesting that even the shortest data set that we used was sufficient, by this measure, to approximate the simulation model. The difference in ML values between the data-specific models and the simulation model did not show a noticeable trend over data set sizes (Table 2.1). For example, the average differences between the P4-ML-estimated trees and the

simulation model trees were more or less constant at 91, 100, and 102 log units over increasing data set sizes, and as mentioned above none of these differences were significant (with corresponding t-test p-values 0.09, 0.32, and 0.59). Increasing p-values of this difference over data set sizes is seen in all estimation methods and is expected because more data will make the optimized parameters in the data-specific models closer to the simulation model parameters. On the other hand, the alignment length had an impact on the model estimation accuracy of all methods: a putative sample size effect, which is most pronounced in the 400-site alignment analyses, is diminished by more data. In the PCA ordination, the data-specific models estimated from the 400-site alignments clustered further from the simulation model than the models estimated from longer alignments. Moreover, around 20% of optimal trees resulting from the 400-site alignments had a different topology to the simulation tree, while analyses of longer alignment lengths always recovered the simulation topology and had more similar tree and branch lengths of the simulation tree. Despite the observation that 400 sites are still sufficient to estimate accurate data-specific models when considering the lack of significance difference between the ML means compared to those of the simulation model (Table 2.1), data this length will likely suffer from a higher sample size effect and produce a poorer approximation of the simulation model than those estimated with more data, as suggested by the inaccurate topologies estimated with data this length.

Although the Bayesian model estimation method was relatively accurate compared to ML estimation methods, it took considerable time to implement and would seem impracticable for typical experimental conditions. By contrast, analyses using all four ML estimation methods were completed in reasonable times even for 8000-site alignments; where we consider reasonable as being assessed in the context of the time needed to generate the data sets and perform typical phylogenetic analyses. However, FastMG, and Codeml to a lesser extent, may be less effective estimation methods where tree and branch lengths are especially important (e.g. dating analyses or molecular rate estimation) for the reasons outlined above. Overall, the IQ-TREE and P4-ML estimation methods appear accurate and time efficient (especially IQ-TREE) and can be recommended for practical use. Indeed, data-specific model estimation appears to be effective even at small sequence data lengths (e.g. 400 sites), as while the estimated models may be inaccurate to some extent perhaps even leading to topological errors, they are also certainly going to be more accurate than pre-computed empirical models unless the data are perchance accurately modeled by one of the very few published models. This interpretation of the simulation results seems to be borne-out by the re-analysis of data from published studies.

The data sets from the published studies we re-analysed varied in numbers of taxa from 14 to 138, had between 5,095 to 365,699 amino acid sites, and included data from nuclear, chloroplast, mitochondria, and viral genomes. In all cases the data-specific models generated in this study were a better fit to the data than the models used in the original studies. Nine of 13 optimal trees inferred using data-specific models were topologically different to the trees published in the original studies (nRF = 0.01-0.07). In addition, optimal trees resulting from data-specific models were longer than the original published trees by between 1.1 and 15.9%, suggesting the that the increased lengths observed in these re-analyses are not artefactual but due to better-fitting models. By contrast, the single viral data set (Schulz *et al*., 2018) had an exceptionally high estimated number of substitutions per site (90.3) where the data-specific model estimated 10.2% fewer substitutions per site. It has previously been suggested that in most cases commonly used models were inadequate and had a lower ability to describe Flavivirus data sets (Duchêne *et al*., 2015), and indeed the results presented here may suggest that the use of data-specific models may improve accuracy for virus data.

Studies using compound, multi-partition models, such as partition models (Feuda *et al*., 2017; Toussaint *et al.*, 2018) and LG4X models (Schwentner *et al*., 2017; Koenen *et al*., 2020), published trees that have lower ML scores than the optimal trees recovered here with data-specific models and a single data partition. The analysis of the chloroplast data set of Koenen (2020) using a data-specific model resulted in an optimal tree with a substantial ML score improvement (given the data set length) over the published tree using the LG4X model. This result is likely because the LG4X model (Le *et al*., 2012) was generated from nuclear data, and therefore a very poor fit to the chloroplast data despite the seeming sophistication of the model structure. These results indicate that assigning best-fitting empirical amino-acid substitution models according to a partition scheme, or according to sites assigned under a distribution-free scheme (LG4X), may not be sufficient to compensate for the use of inadequately fitting (though best-fitting) empirical substitution models.

Phylogenetic relationships of the optimal trees inferred using data-specific models were mostly congruent with the original studies. However, re-analysis of the Munro (2018) data set resulted in an optimal tree where the taxon *Erenna richardi* was placed differently from where it was resolved in the published ML tree. In the original study, its relationship was already noted to be incongruent between ML analyses using a JJT model and BI analyses using the CAT-Poisson model (Munro *et al*., 2018). When using a data-specific model the optimal tree is congruent with the tree from the original BI CAT-Poisson model analyses, which suggests that the amino-acid substitution process of the data cannot not be correctly

accommodated by the JTT model in the original ML analysis. In this study we show that data-specific models which have better fit to the data (in terms of likelihood) infer more accurate trees (with respect to topology and branch lengths) than the empirical models. Superficially, this result seems to contrast with recent analyses by Spielman (2020) who was unable to distinguish topological accuracy among analyses conducted of simulated amino-acid data with empirical models selected through relative model fit methods. However, as Spielman cautions, the results may be due to the possibility of all models have a poor absolute fit to the data, and indeed, our results suggest a possible reason why that is the case. In our study, we show that the 400 amino-acid sites simulations with a 63.2 mean number of expected substitutions per branch only recovered the correct topology (simulation tree) 78% of the time using the actual simulation model (our Appendix Table A2). The simulation trees used by Spielman (2020) were (with one exception) much larger (60, 305, 274, 200, 179, 103, 70, and 23 taxa) than used on our study (26), while the amino-acid data set lengths were relatively small (262, 497, 564, and 661 sites). More importantly, in terms of diversity, at the longest simulation sequence length of the control simulations (661 sites) only 3 (Spiralia 81.4; Opisthokonta 100.9; Yeast 145.2-see our Supplementary Information) of the 8 sets had a mean number of expected substitutions per branch greater than our 400 sites data set (63.2). None of Spielman's simulated control data sets had a mean expected number of substitutions per branch greater than our 1500 site data set (all were considerably lower). Moreover, Spielman notes that the only simulations to reconstruct the correct topology (simulation tree) were from analyses of simulated data from the Spiralia, Opisthokonta, or Yeast phylogenies, suggesting data size and site diversity may be crucial. While these are broad summary statistics, it does point to the probability that the simulated data sets of Spielman were insufficiently variable and/or too short to distinguish among the models being compared (see also Del Amparo and Arenas (2023) for a similar critique).

2.5 Conclusions

Of the five software methods for computing data-specific protein models we compared, the IQ-TREE and P4 (ML) methods were shown to be the most accurate, with IQ-TREE being the most time efficient of the two. Given the availability of time efficient software to calculate sufficiently accurate data-specific amino-acid substitution models there seems longer any justification for using pre-computed empirical models in phylogenetic analyses, even when the data are limited (~400-sites). Indeed, the time needed to choose a model from among an assortment of empirical models may be longer than the time needed to

compute a data-specific model. Moreover, one could imagine a tool that calculates whether the best-fitting empirical model is a sufficiently good fit to the data when compared to a data-specific model by simulating data and performing a statistical test, but if our results generalise to most data, as we suspect they do, then such analyses are superfluous: just use the already calculated a data-specific substitution model.

2.6 References

Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatic*s, 21(9), 2104–2105. https://doi.org/10.1093/bioinformatics/bti263

Adachi, J., & Hasegawa, M. (1996). Model of Amino Acid Substitution in Proteins Encoded by Mitochondrial DNA. *Journal of Molecular Evolution*, 42, 459–468. https://doi.org/10.1007/BF02498640

Adachi, J., Waddell, P. J., Martin, W., & Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution*, 50(4), 348–358. https://doi.org/10.1007/s002399910038

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. Selected papers of hirotugu akaike 199-213. *Springer Series in Statistics*. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-1694-0_15

Chang, E. S., Neuhof, M., Rubinstein, N. D., Diamant, A., Philippe, H., Huchon, D., & Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences of the United States of America*, 112(48), 14912–14917. https://doi.org/10.1073/pnas.1511468112

Cox, C. J., & Foster, P. G. (2013). A 20-state empirical amino-acid substitution model for green plant chloroplasts. *Molecular Phylogenetics and Evolution*, 68(2), 218–220. https://doi.org/10.1016/j.ympev.2013.03.030

Dang, C. C., Le, V. S., Gascuel, O., Hazes, B., & Le, Q. S. (2014). FastMG: a simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets. *BMC Bioinformatics*, 15, 341. https://doi.org/10.1186/1471-2105-15-341

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). A Model of Evolutionary Change in Proteins. *Stat.Wisc.Edu*, 234–236. https://doi.org/10.1.1.145.4315

Del Amparo, R., Arenas, M. 2022. HIV protease and integrase empirical substitution models of evolution: protein-specific models outperform generalist models. *Genes*, 13(1), 61. https://doi.org/10.3390/genes13010061

Del Amparo, R., Arenas, M. 2023. Influence of substitution model selection on protein phylogenetic tree reconstruction. *Gene*, 865, 147336. https://doi.org/10.1016/j.gene.2023.147336

Duchêne, S., Di Giallonardo, F., & Holmes, E. C. (2015). Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Molecular Biology and Evolution*, 33(1), 255–267. https://doi.org/10.1093/molbev/msv207

Felsenstein J. (2004). Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Feuda, R., Pisani, D., Rota-Stabelli, O., Lartillot, N., Pett, W., Dohrmann, M., Wörheide, G., & Philippe, H. (2017). Improved modelling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology*, 27(24), 3864-3870.e4. https://doi.org/10.1016/j.cub.2017.11.008

Foster, P. G. (2004). modelling compositional heterogeneity. *Systematic Biology*, 53(3), 485–495. https://doi.org/10.1080/10635150490445779

Irisarri, I., Uribe, J. E., Eernisse, D. J., & Zardoya, R. (2020). A mitogenomic phylogeny of chitons (Mollusca: Polyplacophora). *BMC Evolutionary Biology*, 20(1), 22. https://doi.org/10.1186/s12862-019-1573-2

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* : CABIOS, 8(3), 275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14, 587. http://dx.doi.org/10.1038/nmeth.4285

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., & McInerney, J. O. (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(1), 1–17. https://doi.org/10.1186/1471-2148-6-29

Koenen, E. J. M., Ojeda, D. I., Steeves, R., Migliore, J., Bakker, F. T., Wieringa, J. J., Kidner, C., Hardy, O. J., Pennington, R. T., Bruneau, A., & Hughes, C. E. (2020). Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytologist*, 225(3), 1355–1369. https://doi.org/10.1111/nph.16290

Kupczok, A., Haeseler, A. Von, & Klaere, S. (2008). An exact algorithm for the geodesic distance between phylogenetic trees. *Journal of Computational Biology*, 15(6), 577–591. https://doi.org/10.1089/cmb.2008.0068

Le, Q. S., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Le, S. Q., Dang, C. C., & Gascuel, O. (2012). modelling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution*, 29(10), 2921–2936. https://doi.org/10.1093/molbev/mss112

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., Depriest, M. S., Bhattacharya, D., Karol, K. G., Fredericq, S., Zechman, F. W., & Lopez-Bautista, J. M. (2016). Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Scientific Reports*, 6(May), 1–13. https://doi.org/10.1038/srep25367

Munro, C., Goetz, F. E., Howison, M., Damian-Serrano, A., Haddock, S. H. D., Pugh, P. R., Siebert, S., Dunn, C. W., Church, S. H., & Zapata, F. (2018). Improved phylogenetic resolution within Siphonophora (Cnidaria) with implications for trait evolution. *Molecular Phylogenetics and Evolution*, 127(June), 823–833. https://doi.org/10.1016/j.ympev.2018.06.030

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1), 268–274. https://doi.org/10.1093/molbev/msu300

Robinson, D. F., & Foulds, L. R. (1979). Comparison of weighted labelled trees In: Horadam, A.F., Wallis, W.D. (eds) *Combinatorial Mathematics VI. Lecture Notes in Mathematics*, vol 748. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0102690

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1–2), 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Schulz, F., Alteio, L., Goudeau, D., Ryan, E. M., Yu, F. B., Malmstrom, R. R., Blanchard, J., & Woyke, T. (2018). Hidden diversity of soil giant viruses. *Nature Communications*, 9(1), 4881. https://doi.org/10.1038/s41467-018-07335-2

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics,* 6(2), 461-464. https://www.jstor.org/stable/2958889

Schwentner, M., Combosch, D. J., Pakes Nelson, J., & Giribet, G. (2017). A Phylogenomic solution to the origin of insects by resolving crustacean-hexapod relationships. *Current Biology*, 27(12), 1818-1824.e5. https://doi.org/10.1016/j.cub.2017.05.040

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sousa, F. de, Foster, P. G., Donoghue, P. C. J., Schneider, H., & Cox, C. J. (2019). Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist*, 222(1), 565–575. https://doi.org/10.1111/NPH.15587

Student. (1908). The Probable Error of a Mean. *Biometrika,* 6(1), 1. https://doi.org/10.2307/2331554

Spielman, S. J. 2020. Relative model fit does not predict topological accuracy in single-gene protein phylogenetics. *Molecular biology and evolution,* 37(7), 2110-2123. https://doi.org/10.1093/molbev/msaa075

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996) Phylogenetic Inference. In: Hillis, D.M., Moritz, C. and Mable, B.K., Eds., *Molecular Systematics,* 2nd Edition, Sinauer Associates, Sunderland (MA), 407-514.

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

Timme, R. E., Bachvaroff, T. R., & Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE*, 7(1). https://doi.org/10.1371/journal.pone.0029696

Toussaint, E. F. A., Breinholt, J. W., Earl, C., Warren, A. D., Brower, A. V. Z., Yago, M., Dexter, K. M., Espeland, M., Pierce, N. E., Lohman, D. J., & Kawahara, A. Y. (2018). Anchored phylogenomics illuminates the skipper butterfly tree of life. *BMC Evolutionary Biology*, 18(1), 1–11. https://doi.org/10.1186/s12862-018-1216-z

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecluar Biology and Evolution*, 18(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. http://www. jstor.org/stable/3001968.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* 39:3, 39(3), 306–314. https://doi.org/10.1007/BF00160154

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5), 555–556. https://doi.org/10.1093/bioinformatics/13.5.555

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang Z. (2014). Molecular evolution: a statistical approach. Oxford: Oxford University Press.

Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N., & Ma, H. (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications*, 5, 1–12. https://doi.org/10.1038/ncomms5956

Zhou, X.-H., Gao, S., & Hui, S. L. (1997). Methods for Comparing the Means of Two Independent Log-Normal Samples. *Biometric*s, 53(3), 1129. https://doi.org/10.2307/2533570

2.7 Appendix

Table A1 - Published phylogenetic trees and data sets used in this study. Where optimal ML trees and/or their scores were not available from the original publication, they were computed using the original model and the constraint tree if available (*). The amino-acid data set used is indicated when the published study included several data sets.

Study	Sequence data type	Main taxonomic groups	Taxa	Sites	Amino-acid substitution Model	Likelihood tree scores (log likelihood units)	ML tree length (substitutions per site)
Timme *et al.*, 2012	Nuclear	Streptophyta	14	56,274	$LG+\Gamma_4+F_{emp}$	-722,437*	4.6
Zeng *et al.*, 2014	Nuclear	Angiosperms	31	25,589	$JTT+\Gamma_4+I+F_{emp}$	-764,724*	8.3
Leliaert *et al.*, 2016	Chloroplast	Viridiplantae	59	13,730	$cpREV+\Gamma_4+F_{emp}$	-486,799*	13.5*
Feuda *et al.*, 2017 (Chang data set *et al.*, 2015)	Nuclear	Metazoa	77	51,940	Branch-linked partition model (47 partitions) + best-fitting models	-2,321,478	16.6
Munro *et al.*, 2018	Nuclear	Siphono-phores	41	365,699	$JTT+R_7+F_{emp}$	-8,113,695	8.4
Schulz *et al.*, 2018	Virus	Giant soil viruses	128	5,095	$LG+R_6+F_{emp}$	-472,305	93.0
Schwentner *et al.*, 2017 (Matrix 1)	Nuclear	Arthropods	40	20,227	LG4X	-444,511	6.3
Toussaint *et al.*, 2018 (DT369)	Nuclear	Papilionoidea butterflies	138	52,287	Best-fit model $+R_4+F_{emp}$	-630,709	1.8*
					Branch-unlinked partition model (366 partitions)	-624,687	2.4*
					Branch-unlinked partition model (27 partitions)	-627,253	1.8*
Irisarri *et al.*, 2020	Mitochondrial	Mollusca	34	3,653	$MtZOA+\Gamma_4+I+F_{emp}$	-115,156	19.5
Koenen *et al.*, 2020 (Chloroplast)	Chloroplast	Leguminosae	157	25,094	LG4X	-571,563*	6.9

Table A2 - Topological distance between the simulation tree and the optimal ML trees inferred from the 400-site alignments (100 replicates). The ML trees were inferred using the estimated data-specific models, the gcpREV (simulation model), cpREV, and WAG models ($+\Gamma_4+F_{mod}$).

Amino-acid substitution models ($+\Gamma_4$+Fmod)	**Non-simulation trees recovered (N of 100 replicates)**	**RF topological distances (mean of 100 replicates)**
gcpREV	22	0.44
cpREV	24	0.48
WAG	25	0.54
Codeml-estimated models	22	0.44
FastMG-estimated models	19	0.38
IQ-TREE-estimated models	20	0.40
P4-ML-estimated models	20	0.40
P4-BI-estimated models	20	0.42

Table A3 - Times used to calculate data-specific models. Mean times (rounded to a minute) are given for the calculation of data-specific substitution models using five methods and three simulated data alignment lengths.

Software	**Simulated alignments**		
	400 sites	**1500 sites**	**8000 sites**
Codeml	14 minutes	65 minutes	265 minutes
FastMG	<1 minute	<1 minute	3 minutes
IQ-TREE	18 minutes	45 minutes	155 minutes
P4-ML	32 minutes	121 minutes	576 minutes (9.6 hours)
P4-BI	34 hours	68 hours	322 hours

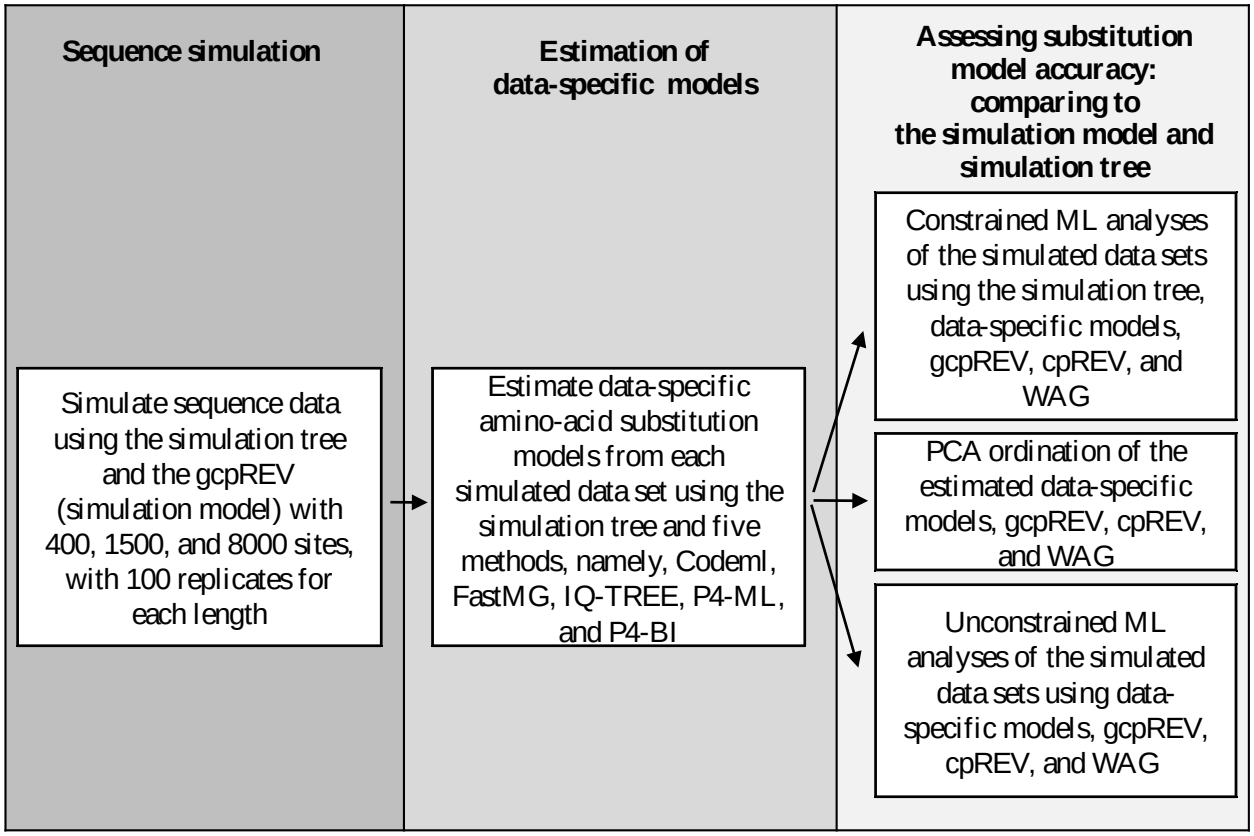


Figure A1 - Summary of the methodology. Data sets were simulated using a simulation tree and substitution model. Data-specific models were then estimated from the data sets using five methods (Codeml, IQ-TREE, FastMG, P4-ML, and P4-BI) and compared to the simulation model using ordination. Phylogenetic analyses of the data-specific models using constrained and unconstrained tree topologies were also compared to the simulation trees and model.

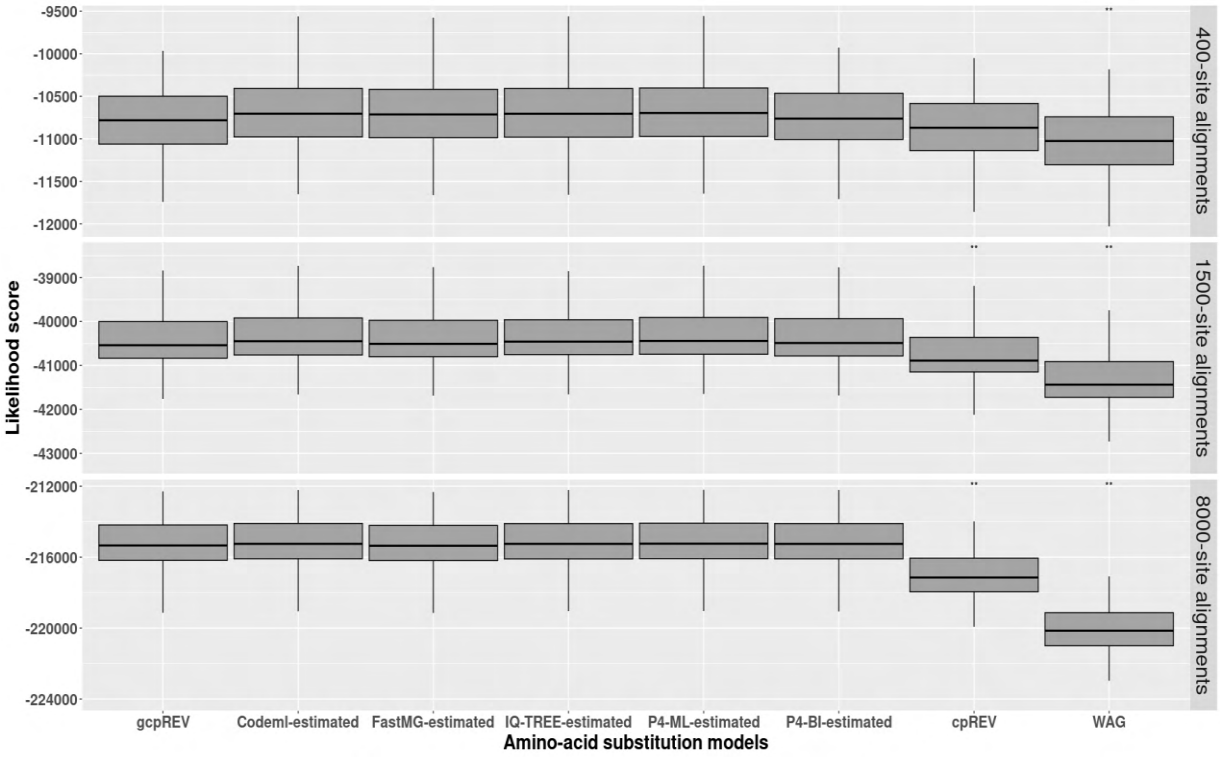


Figure A2 - Box-plots of ML scores of each simulated sequence alignment and its data-specific substitution rate model derived using five model estimation methods. ML scores were calculated using data-specific models ($+\Gamma_4+F_{mod}$) with the topology and branch lengths constrained to those of the simulation tree used to derive the simulated data. ML scores using the equivalent models were calculated for gcpREV, cpREV, and WAG empirical models for comparison. The null hypothesis of no difference between the mean ML scores of the simulation model (gcpREV) and the mean scores resulting from the estimated models, cpREV, and WAG was rejected with a p-value (P) significant at < 0.05 (*), under a two-tailed t-test.

Chapter III

Measuring the efficacy of matched-pairs tests of symmetry and p-value adjustments in identifying tree-heterogeneous sequences in protein data sets

Measuring the efficacy of matched-pairs tests of symmetry and p-value adjustments in identifying tree-heterogeneous sequences in protein data sets

Abstract

Heterogeneity in phylogenetic trees is characterised by lineage-specific evolutionary processes that violate the assumptions of stationarity and the homogeneity of rates of change. The matched-pairs tests of marginal and internal symmetry assess the assumptions of stationarity and homogeneity, respectively, while the matched-pairs test of symmetry evaluates both. Comparing sequences among taxa falls within the realm of multi-comparison statistical tests, requiring the resulting p-values to be adjusted. In this study, the relative ability of each matched-pairs test of symmetry, combined with p-value adjustment methods, was assessed with respect to their accuracy in identifying lineage-specific heterogeneity using simulated amino-acid data sets. Furthermore, tree-heterogeneous sequences were investigated in empirical subsets and complete concatenated data sets. The analyses using the matched-pairs test of marginal symmetry and the Benjamini-Hochberg method had the highest statistical power in relation to the identification of composition-heterogeneous sequences. Although the analyses using the Benjamini-Hochberg method had more false positives than other methods, this rate decreased in the analyses using this method combined with the marginal symmetry test and longer alignments. Identifying rate-heterogeneous sequences was overall difficult. Nevertheless, matched-pairs test of symmetry exhibited slightly lower statistical power compared to the test of internal symmetry. However, in the analyses of composition- and rate-heterogeneous sequences, the former test had a higher power, due to its ability to also detect compositional tree-heterogeneity, albeit lower than that of the marginal symmetry test. The effect on tree inference of empirical data sets was evaluated after filtering out the tree-heterogeneous sequences. Some of the subsequent inference analyses produced topological rearrangements when compared to the original trees, mainly using the matched-pairs test of marginal symmetry combined with the Benjamini-Hochberg method. Such new rearrangements included the resolving the Setaphyta, bryophytes, and Archaeplastida as monophyletic. It could be argued that these clades represent the current best phylogenetic hypotheses for these taxa and therefore evidence that removing heterogeneous sequences can improve the accuracy of phylogenetic analyses.

3.1 Introduction

Simply including as much data as possible in phylogenetic analyses is often insufficient to address many highly-debated evolutionary questions, precisely because these are the relationships that have proven difficult to resolve consistently regardless of the amount of data applied to the question. Often a lack of robustness in tree reconstruction or the source of incongruence between analyses resides in methodological issues such as model misspecification, misassigned data, or stochastic error. Increasingly, the emphasis in phylogenetics is to better understand the evolutionary processes affecting change across sites and among lineages so as to employ better-fitting and more realistic substitution models that decrease the degree of model misspecification. To this end, site-heterogeneous models (Lartillot and Philippe, 2004) and tree-heterogeneous models (Boussau and Gouy, 2006; Foster *et al.,* 2009; Williams *et al.,* 2020) enable the accommodation of variable site and tree evolutionary processes, respectively. However, while usually better-fitting, these models are highly parametrised and often result in significant computational costs when applied to large data sets.

As an alternative to modelling increasingly heterogeneous data, data complexity can be mitigated by excluding the 'disruptive' data (be they taxa or sites) that do not fit the model intended to be used. Such an approach is especially true given the advent of next-generation sequencing where the amount of data available for analysis is often not a limitation. This abundance of molecular phylogenetic data allows for the application of filtering criteria to discard certain data while ensuring that sufficient information remains to infer well-resolved trees (methodologies reviewed in Fleming *et al.,* 2023). In principle, this strategy should reduce the non-historical signal, resulting in tree inferences that are less susceptible to bias from artefacts. Taxon sequences are often referred to as 'rogue' taxa when they are characterised by instability of their position in the inferred topology due to ambiguous or insufficient information (Wilkinson 1996). Bootstrap methods have been used to identify these rogue taxa (Aberer *et al*., 2011, 2013). By contrast, 'problematic' sequences (Fleming *et al.,* 2020, 2023) are robustly placed in the tree, but have a distorting effect, moving other sequences towards or away from them. The so-called "canary sequence" methodology searches for these sequences through a multi-staged process with *a priori* target sequences (Fleming *et al.,* 2020). Another well-known source of incongruence is referred to as "long branch attraction" (LBA) where long-branches are artefactually joined in the tree. Typically, the effects of LBA can be reduced by improving taxon sampling (by 'breaking' long branches

through the addition of closely related taxa) or by excluding fast-evolving sites from the data set. Several methodologies have also been developed to identify taxa affected by LBA (e.g. Soria-Carrasco *et al.,* 2007; Struck, 2014; Mai and Mirarab, 2018).

Most tree inference methods assume that the molecular sequences have evolved under stationary, reversible, and homogeneous conditions, that is, that they evolved under the same Markovian process (Bryant e*t al.,* 2005; Jermiin *et al*., 2017). Stationarity implies that the marginal site frequencies are the same over time, while reversibility means that the rate of change from one state to another is the same in either direction, and homogeneity indicates that instantaneous substitution rates are constant over time. Although these assumptions are usually unrealistic, they are convenient because of their mathematical clarity and computational tractability. A single Markov model is often used for the entire data set, thereby avoiding the need to estimate larger numbers of parameters and ignoring the direction of the evolutionary process. However, when these assumptions are rejected by the data, phylogenetic analyses are more prone to inferring an incorrect tree (Foster and Hickey, 1999; Foster, 2004; Ho and Jermiin 2004; Jermiin et al., 2004; Cox *et al.,* 2008; Foster *et al.,* 2009; Williams *et al.*, 2021). For example, the analyses of one-quarter of the partitions of 35 published data sets indicated the rejection of one or more model assumptions (Naser-Khdour *et al.,* 2019). Trees inferred from these partitions showed different topologies compared to those inferred from partitions where the assumptions were valid. These studies highlight the importance of not neglecting violations of the model's assumptions.

Several methods for testing for the presence of among-lineage, or tree, heterogeneity in nucleotide and amino-acid data sets have been developed. The normalised relative composition frequency variability (nRCFV; Fleming & Struck, 2023), derived from the RCV (Phillips and Penny, 2003) and RCFV (Kück and Struck, 2014), is the average variability in compositions across sequences, where a higher nRCFV is more indicative of compositional heterogeneity than a small nRCFV. As a model-free method, it can be calculated very quickly. However, it does not indicate how strongly the model assumptions are violated, and if one wishes to define a threshold (between heterogeneous and homogeneous sequences), additional criteria are required. In contrast, the chi-squared (χ^2) based significance tests enable the identification of tree-heterogeneous data and to determine its magnitude. The chi-squared test for compositional homogeneity is commonly used to assess the (compositional) model fit to the data (if the model has a poor fit to the data, the implication is that the data evolved under compositional heterogeneous conditions). However, this test ignores the tree-based correlation of compositions among taxa, and thus the chi-square result tends to suffer of Type II error

when computed from the χ^2 curve (Foster, 2004). To ameliorate this effect, Foster (2004) developed a chi-square test using simulations (under a tree and model) to generate a valid null distribution for the statistic.

Statistical analyses to test the assumption of composition- and/or rate-homogeneity have also been developed using pair-wise comparisons of nucleotide or amino-acid sequences. Stuart's matched-pairs test of marginal symmetry (MPTMS; Stuart, 1955) measures the magnitude of compositional heterogeneity in the data, whereas Ababneh's matched-pairs test of internal symmetry (MPTIS; Ababneh et al., 2006) assess the validity of the rate-homogeneity assumption. Bowker's matched-pairs test of symmetry (MPTS; Bowker, 1948) assesses both of these sources of heterogeneity simultaneously. In each test, a divergence matrix of the aligned amino-acid sequences (containing the number of times each combination of amino-acid sites occur) is computed and p-distances are then calculated using a χ^2 test with different degrees of freedom. MPTS computes the distance between the divergence matrix and its transpose, testing the stationarity and homogeneity of the aligned sequences, i.e., the composition-homogeneity and rate-homogeneity of the data. The MPTMS assesses site equality by computing the difference between sequence frequencies and its variance-covariance matrix. When the variance-covariance matrix is not invertible, MPTMS is not applicable. In the MPTIS, the test statistic results from the difference between the MPTS and MPTMS statistic. Jermiin and collaborators have developed tools, such as Homo, that uses the MPTS combined with the Bonferroni-Holm method for p-value correction to assess the heterogeneity among lineages of a specific data partition (Jermiin *et al.*, 2017, 2019). A user can then choose to remove the most tree-heterogeneous sequences of that data partition. A different implementation of the matched-pairs tests of symmetry consist of assigning a data partition as tree-homogeneous or -heterogeneous according to the most divergent taxon pair in the data based on each one of the three matched-pairs tests (Naser-Khdour *et al.*, 2019). The analysed data partition if assigned as tree-heterogeneous can then be excluded from subsequent inference analyses.

Multiple simultaneous comparison tests, such as the matched-pairs tests of symmetry, can suffer from family-wise error rate (FWER). This error is described as the probability of incorrectly rejecting the symmetry assumption at least once (i.e. Type I error or false positive; Benjamini and Hochberg, 1995). The Bonferroni method is often used to counteract the FWER by implementing a more stringent p-value threshold or adjusting the derived p-values. However, this method is very conservative, loses power with increased numbers of tests, and does not account for Type II error (false negatives). The Holm-Bonferroni method (hereafter

called 'Holm') is a modification of the Bonferroni method that uses a stepwise algorithm for simultaneous inference and is more powerful than the Bonferroni (Holm, 1979). The procedures to account for the false discovery rate (FDR) control the rate of false discoveries among the null hypotheses that are rejected, allowing for a more flexible and less conservative approach compared to the former methods. The FDR controlling methods have a higher statistic power compared to methods that control the FWER, accounting for Type II error, but at the cost of more false positives. Such methods include, for instance, the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) and Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001). The former assumes the tests to be independent or positively dependent, while the Benjamini-Yekutieli method is more conservative and better suited to handle certain types of dependencies.

Although assessing tree-heterogeneity is not a standard procedure in phylogenetics, several authors have emphasised the need to assess model assumptions and adjust the phylogenetic methodology accordingly (e.g., Foster, 2004; Morgan *et al.,* 2013; Jermiin *et al*., 2020). However, the extent of tree-heterogeneity across molecular sequence data is still unclear. In a recent simulation study, maximum likelihood inference was shown to be sufficiently robust to handle a wide range of model assumption violations, being only inaccurate in an extreme scenario of convergent evolution (Naser-Khdour *et al.,* 2021). Nevertheless, even if it is assumed that the impact of tree-heterogeneity in tree inference is highly specific, ignoring it merely due to a matter of convenience makes little sense when there are several quick and easy-to-use methods available to counteract it.

The efficacy of matched-pairs tests used in phylogenetics, especially with regard to identifying lineage-specific heterogeneity, has received only limited discussion. In this study, we use simulated amino-acid sequence data to compare the three matched-pairs tests of symmetry and p-value adjustment methods (Bonferroni, Bonferroni-Holm, Benjamini-Hochberg, and Benjamini-Yekutieli) and assess their relative ability to correctly identify tree-heterogeneous sequences. The methods were also assessed on published data sets: 1) a mitochondrial data set (Liu *et al.,* 2014) addressing the relationships among bryophytes; 2) a nuclear data set (Strassert *et al.,* 2021) assembled for the study of eukaryotic ‘supergroups’; and 3) a nuclear data set (Whelan *et al.,* 2015) addressing the placement of the root node of the animal lineage. The monophyly of bryophytes (liverworts, mosses, and hornworts) and their relationship with vascular plants (lycopods, ferns, and seed plants) has been widely debated. The prevailing hypothesis supports the monophyly of this group, placing liverworts and mosses together (Setaphyta), and sister to the hornworts (Puttick *et al.*, 2018; Sousa *et al.,*

2019, 2020). However, analyses of mitochondrial data do not support the same relationships: a result that has been suggested to be due to the influence of compositional heterogeneity (Liu *et al.*, 2014; Sousa *et al.*, 2019). The second nuclear-derived data set addresses deep evolutionary issues among the eukaryotic supergroups (Archaeplastida, Cryptista, Haptista, SAR, Telonemia, Excavata, Amoebozoa, and Obazoa). With the availability of more data, the monophyly of Archaeplastida has come under reassessment, with different models and data sets showing incongruent trees. (e.g., Burki *et al.*, 2016; Brown *et al.*, 2018; Strassert *et al.*, 2019). In analyses where Archaeplastida are monophyletic, Viridiplantae and Glaucophyta are placed together, and form the sister clade to red algae (e.g., Katz & Grant, 2015; Lax *et at.*, 2018; Strassert *et al.*, 2021). This data set was therefore chosen because deep evolutionary divergences are prone to exhibiting lineage-specific heterogeneity. Similarly, the third data set addresses a contentious evolutionary relationship at the base of the Metazoa tree, where either Porifera or Ctenophora are resolved as the sister-group of all other extant animals. The placement of root node of the Metazoa has been intensively debated (e.g., Pisani *et al.*, 2015; Feuda *et al.*, 2017; Simion *et al.*, 2017; Redmond and McLysaght, 2021), and different data filtering and model strategies have been used. Several causes have been indicated for the incongruence between analyses, including the suggestion of their being due to lineage-specific evolutionary processes.

3.2 Methods

Data simulation and the identification of lineage-specific heterogeneity

Amino-acid sequence alignments were simulated in P4 (vers. 1.3; Foster, 2004) allowing for combinations of homogeneous and heterogeneous instantaneous rates and composition frequencies over the tree. The substitution rates were assumed to be uniform across sites. The simulation tree comprised 30 taxa, where each leaf branch was 0.4 substitutions per site long, and internal branches 0.3 substitutions per site (comprising five clades, each with six leaf branches and four internal branches). The simulation process consisted of generating a random root sequence and evolving it over the simulation tree under the process specified by the ‘default’ simulation model, while the lineage-specific heterogeneity parameters (i.e., composition-heterogeneous and/or rate-heterogeneous sequences) were constrained to models different from the 'default' model. The use of empirical substitution exchange rate models for the simulation of tree-heterogeneous sequences, rather than random values, ensures more realistic substitution parameters and greater applicability of the results. The nuclear-derived WAG model (Whelan and Goldman, 2001) was used as the ‘default’ simulation model to seed the homogeneous substitution rates. Tree-heterogeneous sequences were derived from the nuclear-derived JTT (Jones *et al*., 1992) and the mitochondria-derived stmtREV (Liu *et al.,* 2014) models. These models were selected according to their position in the protein space and plotted via ordination using Principal Component Analysis (PCA; Fig. 3.1). The sequences derived from the JTT were expected to simulate an evolutionary process more similar to the default model, as suggested by their position in the PCA, and thereby harder to distinguish from WAG. By contrast, the stmtREV derived sequences were expected to be more easily detected as they were moved divergent from the default model.

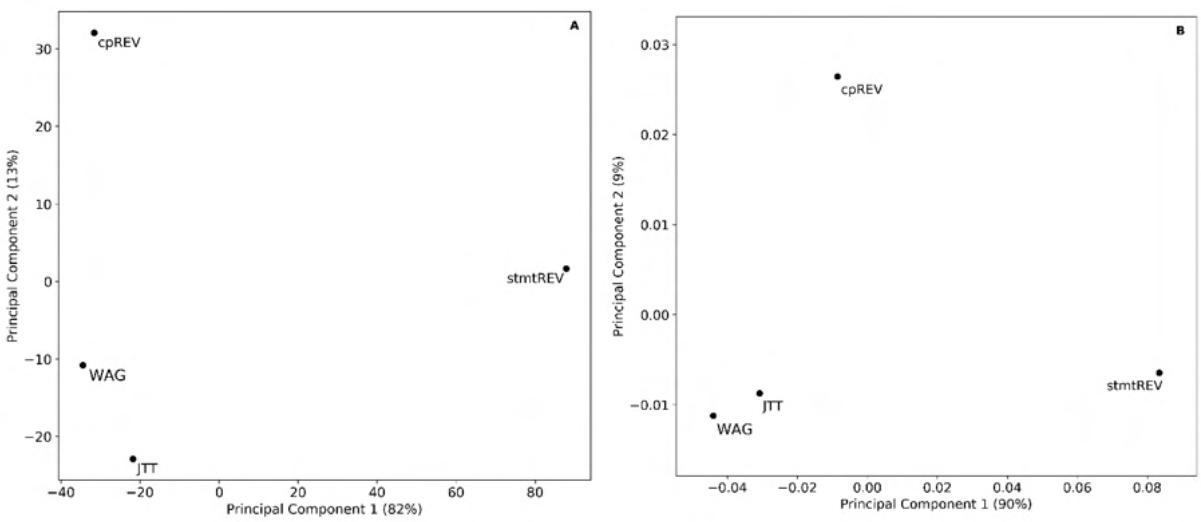


Figure 3.1 - Substitution exchange rate models used in this study. Principal component analyses of the exchange rates (A) and composition frequencies (B) of the WAG, JTT, cpREV, and stmtREV models**.**

To simulate the data, two of the five 6-taxon clades were randomly chosen, and the JTT model (rates and/or compositions) was assigned to one of them, while the stmREV model was assigned to the other. The rest of the tree maintained its default WAG rates and composition frequencies. Alignments were generated with 2000, 5000, and 8000 sites, with 100 replicates for each length. The procedure followed four strategies (Table 3.1), wherein the sequence alignments were: composition-homogeneous and rate-homogeneous, composition-heterogeneous and rate-homogeneous, composition-homogeneous and rate-heterogeneous, and composition-heterogeneous and rate-heterogeneous.

Table 3.1 - Substitution model parameters used for simulating tree-heterogeneous and tree-homogeneous alignments. The WAG model was used to seed tree-homogeneous substitution rates, while the JTT, stmtREV, and cpREV models were used to seed tree-heterogeneous substitution rates (instantaneous rates, composition frequencies, or both).

Sequences simulation	Substitution model parameters	
	Composition frequencies	**Instantaneous rates**
Composition-heterogeneous and rate-heterogeneous	WAG, JTT, stmtREV	WAG, JTT, stmtREV
Composition-homogeneous and rate-heterogeneous	WAG	WAG, JTT, stmtREV
Composition-heterogeneous and rate-homogeneous	WAG, JTT, cpREV	WAG
Composition-homogeneous and rate-homogeneous	WAG	WAG

The objective of the simulation analyses was to compare the relative accuracy of the matched-pairs tests of symmetry combined with different p-value correction methods to detect lineage specific heterogeneous sequences, and not their absolute ability to identify tree-heterogeneous sequences and the relation with the tree inference result. Therefore, the effect of among-site rate variation, incomplete sequences (missing data), branch lengths, and other evolutionary aspects were not considered. Moreover, because the composition-heterogeneous sequences derived from the stmtREV are easily identified due to their notable divergence from the parameters of the WAG model, the stmtREV was replaced by the cpREV model (Adachi *et al*., 2000) in simulations involving composition-heterogeneous and rate-homogeneous data.

The MPTS, MPTMS, and MPTIS tests were used (as implemented in P4) to detect sequences unlikely to have evolved under the same Markovian process (i.e., their having evolved under tree-heterogeneous processes). The simulated alignments were assessed using the three matched-pairs tests combined with four p-value adjustment procedures, namely Bonferroni, Holm, Benjamini-Hochberg, and Benjamini-Yekutieli. These methods were used as implemented in the Python library Statsmodels (vers. 0.12.2; with an alpha of 5% and default parameters). For each analysis, the number of times that the null hypothesis was correctly rejected, wrongly rejected (Type I error), wrongly not-rejected (Type II error), and

correctly not-rejected (Table 3.2) were recorded. These frequencies were then used to calculate the FDR, false negative rate (FNR), and statistical power (1-FNR).

Table 3.2 - Hypothesis testing for assessment of the stationarity and homogeneity assumptions.

Significance	**Null Hypothesis (stationarity and/or homogeneity)**	
	Stationary and/or homogeneous data	**Non-stationary and/or non-homogeneous data**
Fail to reject	Correctly not-rejected	Type II error
Reject	Type I error	Correctly rejected

Identification of tree-heterogeneous sequences in three published data sets

Three empirical published data sets were selected and assessed using the matched-pairs tests of symmetry with adjusted p-values. The Liu *et al.*, (2014) data set was chosen as it was previously shown to be compositionally heterogeneous. The Strassert *et al.*, (2022) and the Whelan *et al.*, (2015) analyses addressed controversial deep evolutionary nodes, namely, with regard to the monophyly of the Archaeplastida, and the placement of the root node of animals, respectively. For convenience, each data set is referred to by the name of the first author of its study, that is, Liu, Strassert, and Whelan. The Liu data set comprises 41 proteins and 60 taxa, the Strassert data set comprises 320 proteins and 136 taxa, and the Whelan data set includes 209 proteins and 76 taxa (data matrix 12 in the original study). Tree-heterogeneity was assessed using the three matched-pairs tests combined with the Bonferroni, Holm, Benjamini-Hochberg, and Benjamini-Yekutieli methods. These analyses were performed for each single protein alignment, the concatenated data set of single alignments, and on data partitions derived from optimal partition schemes. The latter partitioning schemes were determined using two methods: 1) merging protein alignments according to substitution model-fit optimisation using the ModelFinder tool (Kalyaanamoorthy *et al*., 2017), as implemented in IQ-TREE 2 (vers. 2.2.2.7; Minh *et al.*, 2020); and 2) clustering of individual genes based on their composition frequencies using the K-means algorithm from the Python Sklearn library (version 0.23.2), with the number of clusters being determined according to the Elbow method (Ketchen *et al.*, 1996). The ModelFinder analyses were conducted using a data-specific substitution model estimated from the combined data set (described below) and

using an among-site rate variation parameter constrained to a gamma-distribution discretised in four categories (Γ_4). In the analysis of each data partition (individual genes, constructed data partitions, and entire concatenated data sets), taxon sequences were identified as tree-heterogeneous sequences when lineage-specific evolutionary processes were detected, according to each symmetry test and p-value adjustment procedure. Subsequently, the rejected null hypothesis pairwise comparisons were sorted from the lowest p-value, and the taxon within the top pair with the highest number of rejected comparisons was discarded (as well as all p-values associated it, and therefore the remaining pairwise comparisons including it were no longer considered). This process continued until no more rejected pairwise comparisons remained. In situations where two taxa in the same pairwise comparison had identical counts of rejected comparisons, the taxon with a lower overall sum of p-values was removed. If indistinguishable, both taxa were discarded.

Optimal ML trees were inferred from the filtered data sets and from the original data sets without filtering of heterogeneous sequences. In detail, each single-protein alignment was filtered and combined into a single concatenated data set. From this data set, the optimal ML tree was inferred using a data-specific model. This procedure was repeated for each combination of matched-pairs test and p-value correction method. The optimal partition scheme analyses follow a similar procedure, but with each calculated partition being filtering instead of the single-protein partitions. Each data set was then analysed with a data-specific model estimated from the concatenated data but combined with a branch-linked model (Chernomor et al., 2016). Combined data sets were also analysed without filtering sub data partitions; rather, sequences were filtered from the entire combined dataset and optimal ML trees inferred. All analyses were conducted in IQ-TREE (vers. 1.6.12; Nguyen *et al*., 2015) and included data-specific amino-acid substitution models, Γ_4 parameter, and optimised composition frequencies (F_{est}). The data-specific models were estimated using a general time-reversible GTR+Γ_4 model in IQ-TREE. Branch support values were derived from 10000 ultrafast bootstrap (UFBOOT) replicates (Hoang et al., 2017). The optimal ML trees were visually inspected and compared to the trees resulting from the unfiltered data sets using a normalised Robinson-Foulds (nRF) distance metrics (RF/RFmax, where RFmax is obtained by 2*(number of taxa-3); Kupczok et al., 2008). The data products (protein alignments, calculated substitution models, and trees), novel scripts used to make calculations and a machine actionable RO-Crate metadata specification are available from GitHub: https://github.com/joaobrazao/Applying-data-specific-substitution-models-and-mitigating-the-effects-of-among-lineage-heterogeneity.git

3.3 Results

The matched-pairs tests of symmetry coupled with p-value adjustment procedures were compared among themselves regarding their ability to identify lineage-specific rate- or composition-heterogeneous sequences, i.e., correctly reject or accept the null hypothesis of stationarity and/or rate-homogeneity of individual sequences. The analysis of the simulated composition-heterogeneous and rate-homogeneous data demonstrated that the null hypothesis was rejected significantly more often in the MPTMS analyses than in those using the MPTS, regardless of the alignment length and p-value adjustment procedures. The statistical power of the MPTMS (0.3-0.9) was significantly higher compared to that of MPTS (<0.1-0.7), when using the same alignment length and p-value adjustment procedure (Fig. 3.2; Appendix Tables A1 and A2). As expected, the statistical power increased with longer alignments, whether the MPTS or the MPTMS were used. The FDR calculated from MPTMS analyses (<0.1-0.2) decreased as the alignment length increased in most cases (except in the Bonferroni based methods analyses using the 8000-site alignments). Similarly, the FDR metric computed from the MPTS results (<0.1-0.2) had the same behaviour, except in those derived from the MPTS analyses combined with the Benjamini-Hochberg method, and with the Benjamini-Yekutieli method using the 8000-site alignments. Although the FDR of the MPTS analyses of the shortest alignments had a range of values similar to those of the MPTMS results (<0.1-0.2), the former analyses failed to reject most null hypotheses that should have been rejected (Type II error; yielding a statistical power close to zero). As the sequence alignment lengths increased, the FDR calculated from the MPTS results either decreased or remained statistically similar (with the exception of the Benjamini-Hochberg results, which showed a significant increase). Apart from the analyses using the shortest alignments, where the statistical power associated with MPTS results was nearly zero, the FDR values were either significantly higher or statistically indistinguishable compared to those calculated from the MPTMS results in most cases.

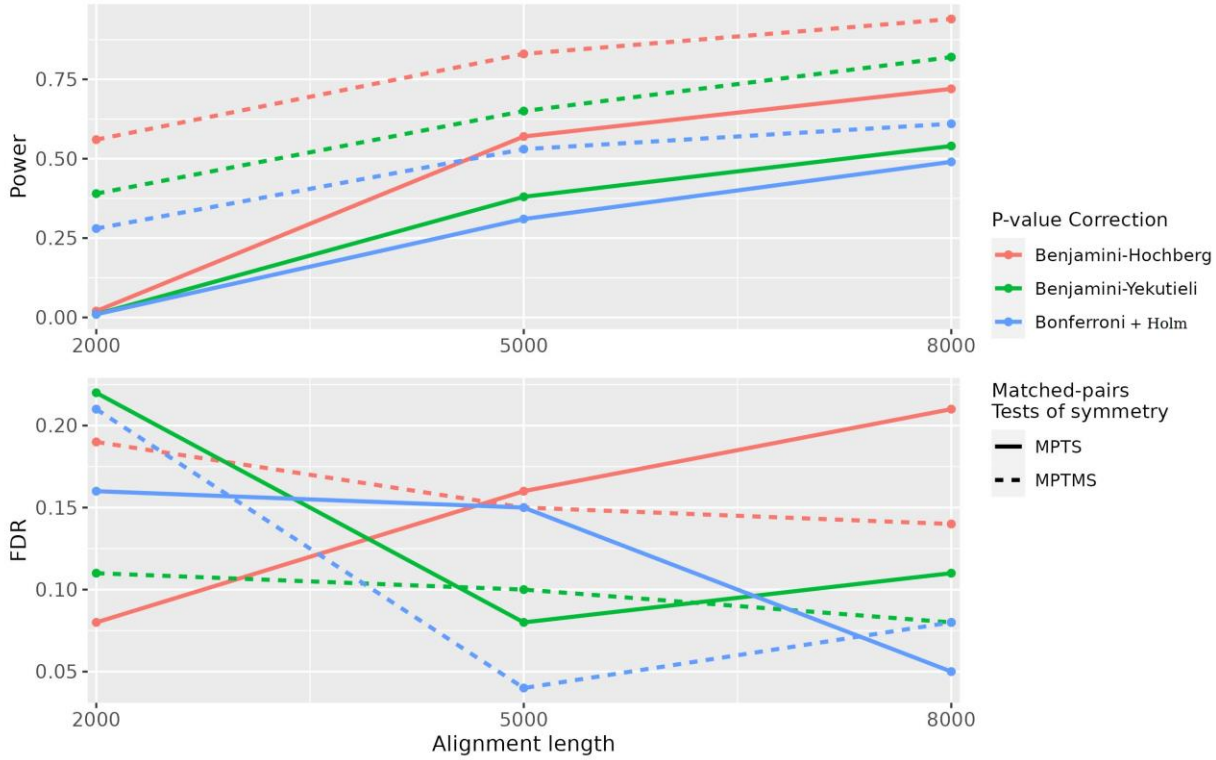


Figure 3.2 - Statistical power and false discovery rate (FDR) of the matched-pairs tests of symmetry (MPTS) and marginal symmetry (MPTMS) in the identification of composition-heterogeneous sequences. The tests were computed using 2000, 5000, and 8000-site alignments (100 replicates each) and combined with the p-value adjustment methods, namely, Benjamini-Hochberg, Benjamini-Yekutieli, and Bonferroni. Because the results from the Holm method were almost identical to those of the Bonferroni method, it was not included to enhance plot visualisation.

The statistical power of p-value adjustment procedures increased with longer alignments. For the Benjamini-Hochberg method, the power was significantly higher compared to that obtained from analyses using other methods when evaluated under the same matched-pairs test and sequence alignment length. The highest power observed (0.9) was achieved when the Benjamini-Hochberg method was combined with MPTMS using 8000-site alignments. While the FDR calculated from the analyses using the Benjamini-Hochberg method generally exceeded those derived from the analyses using other methods, it decreased as the alignment length increased when coupled to the MPTMS. However, when combined with the MPTS, it increased as the alignment length increased. The statistical power of the analyses using Bonferroni and Holm methods was the lowest and statistically similar in each analysis (compared under the same matched-pairs test and alignment length) and had fewer Type I errors. The FDR calculated from the analyses using the 8000-site alignments were lower than those calculated from the 2000-site alignment analyses. The analyses conducted using the Benjamini-Yekutieli method generally yielded intermediate values in terms of

statistical power. These values fell between those obtained with the Benjamini-Hochberg and Bonferroni methods. The FDR decreased with longer alignments or remained statistically indistinguishable between the analyses conducted using the alignments of 5000 and 8000 sites. The analyses of the composition-heterogeneous and rate-homogeneous data also demonstrated that the further in value the composition frequencies (derived from the JTT and cpREV models) used in simulating lineage-specific composition-heterogeneous sequences were from the simulation model (WAG), the greater the number of composition-heterogeneous sequences were detected (Appendix Tables A1 and A2). Indeed, the statistical power regarding only the cpREV-derived sequences was close, or equal, to one in most analyses (except in 2000-site alignment analyses), while in the analyses of the JTT-derived sequences, the statistical power was less than 0.1 in more than half of the analyses. The FDR values calculated regarding the JTT-derived sequences (0.2-1) were also significantly higher than those calculated from the cpREV derived results (<0.1-0.3). Overall, these results indicated that the composition-heterogeneous (and rate-homogeneous) sequences derived from the cpREV model were detected with significantly greater frequency compared to those derived from the JTT model.

MPTS and MPTIS analyses of the composition-homogeneous and rate-heterogeneous alignments failed to reject most rate-heterogeneous sequences when computed from the shortest alignments, regardless of which test was used (Fig. 3.3; Appendix Tables A3 and A4). The rejection of the null hypothesis increased significantly with longer alignments where the sequences derived from the stmtREV model were more often rejected than the JTT-derived sequences. These observations are in agreement with the above results where the analyses of the JTT-derived sequences showed a lower number of identified sequences, a lower statistical power (close to zero), and a higher FDR (0-0.7). The MPTIS analyses had a higher statistical power (<0.1-0.5) compared to the MPTS analyses (<0.1-0.4). However, these differences were statistically significant only in the 8000-site alignment analyses (and in the 5000-site alignment analyses using the Benjamini-Yekutieli method). The FDR values were statistically similar when compared between the MPTS and MPTIS results (<0.1-0.2). Excluding the shortest alignment results (where the calculated values were zero), the FDR values derived from the longest alignment analyses (<0.1-0.2) were lower than those derived from the 5000 sites long analyses (0.1-0.2). However, the results using the Benjamini-Hochberg method contrasted with the latter where the FDR value increased (from <0.1 to 0.1) irrespective of the test used. The analyses using the Benjamini-Hochberg method had always the highest statistical power (<0.1-0.5) and the lowest FDR (0.1), except for the 8000-site alignment

analyses. In these cases, regardless of the matched-pairs test, the Benjamini-Yekutieli method displayed the lowest FDR (<0.1), but it also had the lowest statistical power (0.1). The Bonferroni and Holm methods were statistically similar in each analysis regarding the FDR (0.1-0.2) and the statistical power (<0.1-0.2).

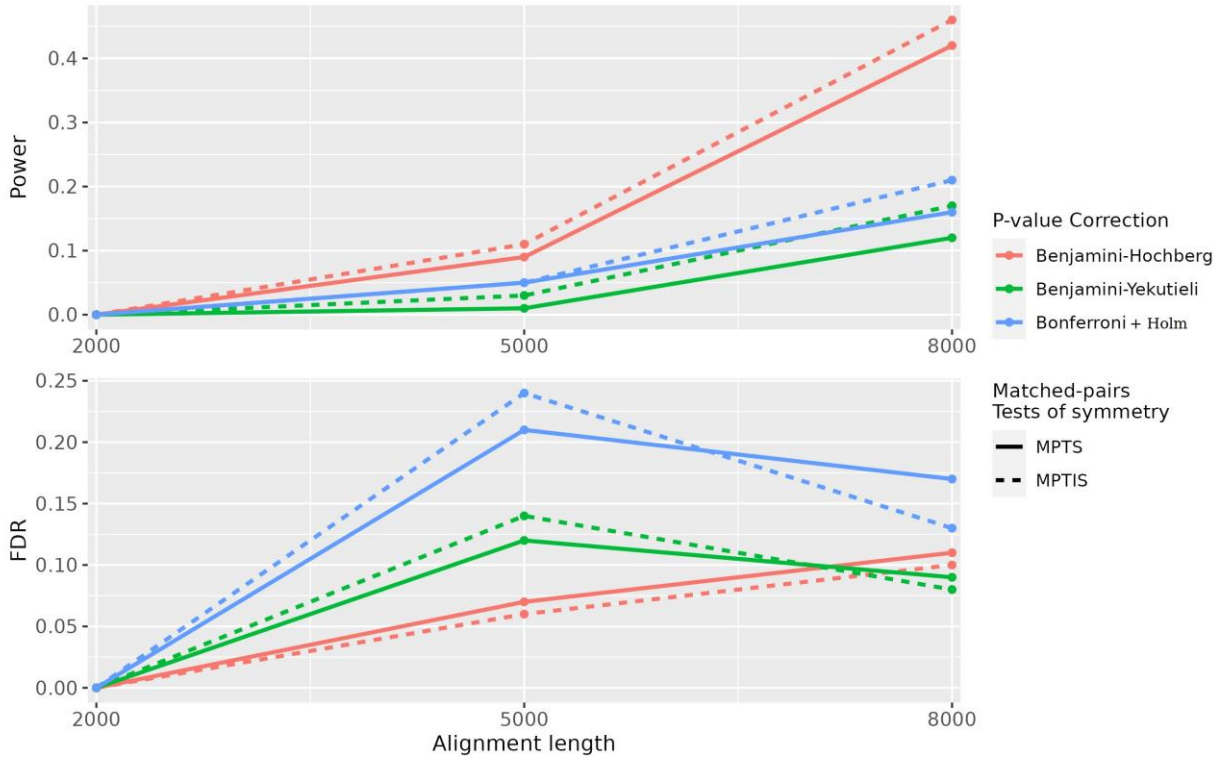


Figure 3.3 - Statistical power and false discovery rate (FDR) of the matched-pairs tests of symmetry (MPTS) and internal symmetry (MPTIS) in the identification of rate-heterogeneous sequences. The tests were computed using 2000, 5000, and 8000-site sequence alignments (100 replicates each) and combined with the p-value adjustment methods, namely, Benjamini-Hochberg, Benjamini-Yekutieli, and Bonferroni. Because the results from the Holm method were almost identical to those of the Bonferroni method, it was not included to enhance plot visualisation.

In matched-pairs analyses of composition-heterogeneous and rate-heterogeneous data, the highest statistical power was observed using the MPTMS, which also increased with longer alignments (0.5-1; Fig. 3.4). MPTS analyses (0.5-0.8) had the second highest statistical power, followed by the MPTIS analyses (0.1-0.5). The MPTS and MPTMS analyses detected all heterogeneous sequences derived from the stmtREV model (Appendix Tables A5 and A6), thereby having a maximum statistical power regarding these sequences (1) and accompanied by a low FDR (0-0.3). The JTT-derived sequences were detected significantly less often than the stmtREV-derived sequences regardless of the method or alignment length used (statistical power of 0-0.9 and FDR of 0.2-1). The MPTIS did not reject any sequence from the 2000-site alignments and failed to reject most tree-heterogeneous sequences in the 5000-site

alignments. With respect to the FDR, the MPTS results (0-0.2) were either statistically significantly lower or similar to those from MPTMS analyses (<0.1-0.2), except when combined with the Benjamini-Hochberg method using the 5000- and 8000-site alignments where they were significantly higher (0.2). The FDR values derived from MPTIS analyses using the 5000-site alignments were the highest (0.2-0.3; except those derived from the Benjamini-Hochberg analyses). However, these results were only supported by a few observations (statistical power <0.1). By contrast, the analyses using the 8000-site alignments combined with Benjamini methods exhibited the lowest FDR (<0.1-0.1) when compared with the MPTS (<0.1-0.2) and MPTMS (<0.1-0.1) using the 8000-site alignments. Nevertheless, the statistical power derived from the latter MPTIS analyses were always lower than those from other tests. Overall, the Benjamini-Hochberg method analyses had the highest statistical power, followed by the Benjamini-Yekutieli. The FDR values derived from Benjamini-Hochberg analyses were also the highest, but decreased with longer alignments when combined with MPTMS, while those calculated from the Bonferroni and Holm analyses were statistically similar and the lowest in most cases.

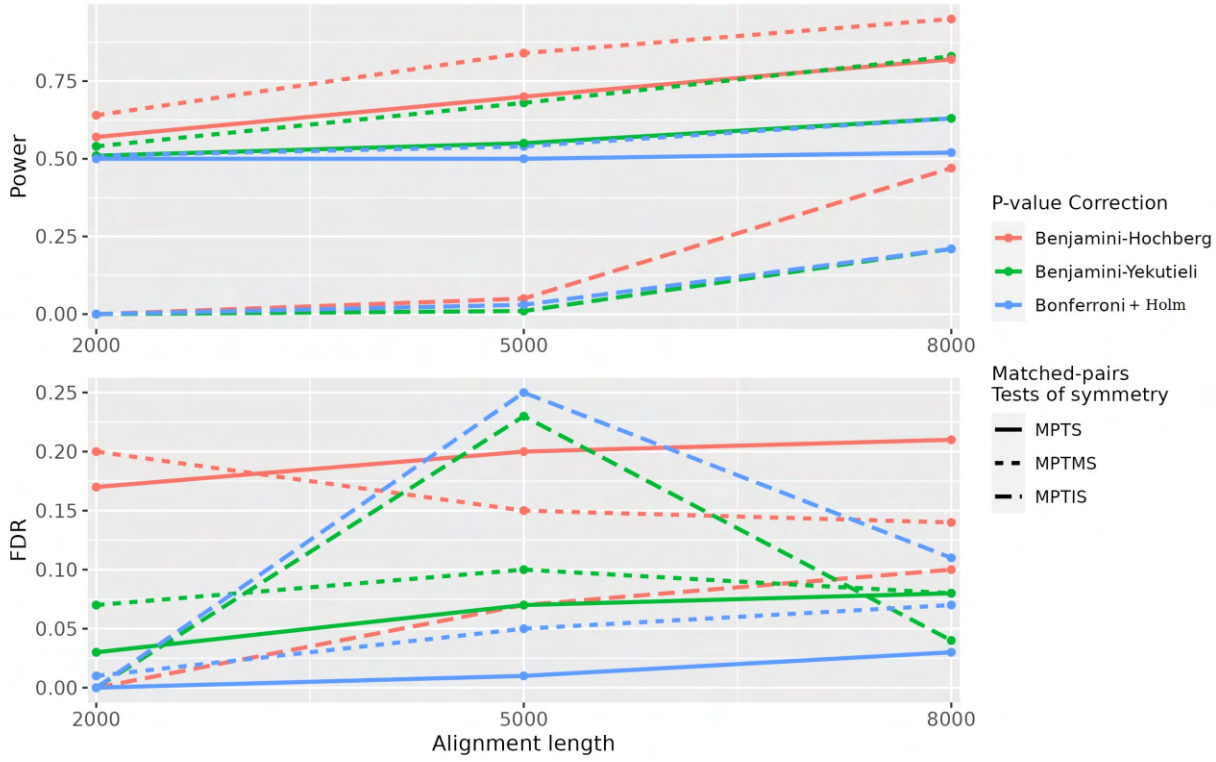


Figure 3.4 - Statistical power and false discovery rate (FDR) of the matched-pairs tests of symmetry (MPTS), marginal symmetry (MPTMS), and internal symmetry (MPTIS) in the identification of composition- and rate-heterogeneous sequences. The tests were computed using 2000, 5000, and 8000-site alignments (100 replicates each) and combined with the p-value adjustment methods, namely, Benjamini-Hochberg, Benjamini-Yekutieli, and Bonferroni. Because the results from the Holm method were almost identical to those of the Bonferroni method, it was not included to enhance plot visualisation.

The analyses of the composition- and rate-homogeneous data demonstrated that most of the sequences were correctly classified as tree-homogeneous. Across all methodologies either none or an insignificant fraction (< 0.3%) of sequences were wrongly identified as tree-heterogeneous sequences.

Case study 1-Bryophytes (Liu *et al*., 2014)

MPTMS analyses of the sequences of individual proteins of the Liu data set identified more tree-heterogeneous sequences (4-10%) than the MPTS (1-5%), regardless of the p-value adjustment procedure (Table 3.3). After excluding lineage-specific composition-heterogeneous sequences identified using the MPTS test, the reconstructed optimal ML trees exhibited lower nRF distances (4-11%; computed against the tree inferred from the original data set without filtering of heterogeneous sequences; Appendix Figs. A1-A5) compared to the trees inferred from data set derived after filtering sequences using the MPTMS (9-19%; Appendix Figs. A6-A9). These comparisons were made using the same p-value adjustment procedures. With respect to the p-value adjustment procedures only, the Benjamini-Hochberg method analyses detected more tree-heterogeneous sequences (5-10%, when combined with MPTS and MPTMS respectively) than the Benjamini-Yekutieli (1-8%) and Bonferroni based methods (1-4%) analyses. Trees inferred from the data sets filtered using the Bonferroni and Holm methods (when combined with the same matched-pairs test of symmetry) had the same topology and had the lowest nRF (7-9%). Trees inferred from the data sets derived after filtering according to the Benjamini-Yekutieli and Benjamini-Hochberg methods had a nRF of 7-11% and 11-19%, respectively. The MPTIS analyses only detected one lineage-specific rate-heterogeneous sequence irrespective of the p-value correction method used. The tree derived after filtering this sequence had a congruent topology as the tree inferred from the original data set. In most trees bryophytes were recovered as a grade. Nevertheless, the optimal ML tree (Appendix Fig. A9) inferred from the data set filtered using the MPTMS combined with the Benjamini-Hochberg method recovered Setaphyta (UFBOOT = 58%) and bryophytes as a paraphyletic group with lycopods sister to hornworts (UFBOOT = 65%). In the same analyses, monocots were nested in eudicots, rather than as sister-groups. The monocot taxa, the *Cycas taitungensi* (sister-taxon of angiosperms), and the lycopod *Isoetes engelmannii* had the highest amount of tree-heterogeneous sequences removed. Therefore, the nested position of lycopods and monocots may be due to a lack of phylogenetic signal in the remaining taxa.

Table 3.3 - Tree-heterogeneous sequences identified in the Liu et al., (2014) data set and the inferred topological rearrangements after their exclusion. The heterogeneous sequences were identified according to the matched-pairs tests of symmetry (MPTS), marginal symmetry (MPTMS), and internal symmetry (MPTIS) combined with four p-value adjustment procedures. These sequences were excluded from each data partition analysed, be a single-protein, merged proteins, or the entire concatenated data set. The inferred topologies were then compared with the original data set derived tree using the normalized Robinson-Foulds (nRF). When the number of taxa was different, the latter metrics was not calculated.

Data filtered	Matched-pairs test	P-value adjustment procedure	Tree-heterogeneous sequences (%)	nRF (%)
Individual proteins	MPTS	Bonferroni	0.8	7.0
		Holm	0.8	7.0
		Benjamini-Yekutieli	1.1	7.0
		Benjamini-Hochberg	5.1	10.53
	MPTMS	Bonferroni	3.6	8.8
		Holm	3.7	8.8
		Benjamini-Yekutieli	7.6	10.5
		Benjamini-Hochberg	10.2	19.3
	MPTIS	Bonferroni	<0.1	0
		Holm	<0.1	0
		Benjamini-Yekutieli	<0.1	0
		Benjamini-Hochberg	<0.1	0
K-means derived partitions	MPTS	Bonferroni	4.2	8.8
		Holm	4.3	8.8
		Benjamini-Yekutieli	6.5	12.3
		Benjamini-Hochberg	13.2	12.3
	MPTMS	Bonferroni	11.4	-
		Holm	11.6	-
		Benjamini-Yekutieli	15.8	-
		Benjamini-Hochberg	18.9	-
	MPTIS	Bonferroni	0.2	0
		Holm	0.2	0
		Benjamini-Yekutieli	0.2	7.0
		Benjamini-Hochberg	0.1	0
ModelFinder derived partitions	MPTS	Bonferroni	16.5	14.0
		Holm	16.9	14.0
		Benjamini-Yekutieli	22.3	22.8
		Benjamini-Hochberg	34.5	19.3
	MPTMS	Bonferroni	26.9	-
		Holm	27.1	-
		Benjamini-Yekutieli	33.1	-
		Benjamini-Hochberg	39.8	-
	MPTIS	Bonferroni	0.2	0
		Holm	0.2	0
		Benjamini-Yekutieli	0.2	0

		Benjamini-Hochberg	0.2	0
Concatenated data set	MPTS	Bonferroni	66.7	-
		Holm	68.3	-
		Benjamini-Yekutieli	76.7	-
		Benjamini-Hochberg	81.7	-
	MPTMS	Bonferroni	75.0	-
		Holm	76.7	-
		Benjamini-Yekutieli	78.3	-
		Benjamini-Hochberg	85.0	-
	MPTIS	Bonferroni	5.0	-
		Holm	5.0	-
		Benjamini-Yekutieli	23.3	-
		Benjamini-Hochberg	58.3	-

The optimal data partitioning schemes comprised 19 partitions according to the K-means algorithm and nine partitions derived from the search using the ModelFinder tool. The MPTMS analyses of the K-means-partitioned data identified 11-19% tree-heterogeneous sequences, while the MPTS analyses identified 4-13% (Table 3.3). The same procedures applied to the ModelFinder derived partitions identified 7-35% and 27-40% heterogeneous sequences according to the MPTMS and MPTS respectively. These results showed that despite the lower number of partitions in the ModelFinder scheme analyses, the percentage of tree-heterogeneous sequences was higher than in the K-means scheme analyses. This difference is likely due to the longer partitions length derived from ModelFinder, which enhance the statistical power, as demonstrated with the simulated data analyses. The optimal ML trees inferred from the data sets filtered according to the MPTS had a nRF of 9-12% (using the K-means derived partitions; Appendix Figs. A10, A11, A12, and A13) and 14-23% (using the ModelFinder derived partitions; Appendix Figs. A14, A15, A16, and A17). The nRF was not computed for the trees inferred from the data sets derived after filtering composition-heterogeneous sequences using the MPTMS because they had fewer taxa than the original tree (Appendix Figs. A18-A25). With respect to the p-value adjustment procedures, the Benjamini-Hochberg method identified more tree-heterogeneous sequences than other methods, namely 13-19% (K-means derived partitions) and 35-40% (ModelFinder derived partitions). By contrast, the Bonferroni based methods identified the least (4-17%). The trees inferred from the latter data sets had a nRF of 9-14%, while the trees derived after filtering sequences using the Benjamini methods had a nRF of 12-23%. Overall, these results agreed with the above analyses, where the data sets and optimal trees derived after filtering heterogeneous sequences according to the MPTMS and Benjamini procedures identified more

tree-heterogeneous sequences and exhibited higher nRF distances than those derived from the analyses using the MPTS combined with Bonferroni methods. Data sets derived after filtering sequences using the MPTIS and the trees reconstructed using all four correction methods were identical to the original data set and optimal tree, regardless of the partition scheme (except the tree inferred from the K-means derived data set filtered using the Benjamini-Yekutieli; nRF of 7%; Fig. A26). Analyses using the ModelFinder derived partitioning scheme recovered the bryophytes as a monophyletic group. This result was observed in trees reconstructed from the data sets derived after filtering heterogeneous sequences using the MPTMS coupled with the Bonferroni, Holm, and Benjamini-Hochberg methods (Appendix Figs. A22, A23, and A24). However, mosses were recovered sister to hornworts rather than to liverworts (i.e. contradicting the Setaphyta clade). Additionally, the tree inferred from the data set derived after filtering heterogeneous sequences using the MPTS coupled with the Benjamini-Yekutieli method recovered the Setaphyta and bryophytes as a paraphyletic group with lycopods sister to hornworts (Fig. A20). In none of these optimal trees were these relationships well-supported (UFBOOT support >95%). The trees inferred using the K-means-derived partitioning scheme did not recover the bryophytes as monophyletic or the Setaphyta clade.

When the Liu data set was treated as a single partition and filtered of heterogeneous sequences using MPTS and MPTMS methods (regardless of correction method), as well as the MPTIS coupled with Benjamini-Hochberg method, the taxa required to address the monophyly of Bryophyta were removed. The trees reconstructed from the data sets derived after filtering sequences using the MPTIS combined with Bonferroni and Holm methods were overall congruent with the original tree without sequence filtering (only three taxa were assigned as non-homogeneous; Appendix Figs. A27 and A28). The tree inferred from the data set derived after filtering sequences using the MPTIS combined with the Benjamini-Yekutieli method recovered the bryophytes as a paraphyletic group (the lycopod *Huperzia squarrosa* was nested within) and the Setaphyta clade (Appendix Fig. A29).

Case study 2-Archaeplastida (Strassert *et al*., 2021)

Analyses of individual proteins from the Strassert data set indicated less than 1% of 36,857 sequences as tree-heterogeneous sequences. MPTMS coupled with the Benjamini-Hochberg method identified 0.7% of sequences as composition-heterogeneous (Table 3.4). The MPTMS coupled with the remaining p-value correction methods indicated 0.2-0.3% heterogeneous sequences, while MPTS and MPTIS identified less than 0.1%, irrespective of

the p-value adjustment method. When the resulting topologies differed from the tree inferred from the original data set (Appendix Figs. A30-A35), the nRF varied between 2% and 5%, mainly due to the placement of Telonemia and Haptisa. In the original tree, Telonemia was recovered sister to the SAR supergroup (TSAR; Strassert *et al.*, 2019), with Haptista resolved as their sister-group. However, in some inferred trees, Telonemia was nested in Haptista, with this clade sister to the Archaeplastida, though weakly support. These results suggest that Telonemia and Haptista are difficult to place robustly because of very poor phylogenetic signal and/or other causes, but unlikely due to lineage-specific evolutionary processes. The Cryptista was nested in Archaeplastida (the same as the original tree), regardless of the methods used. Trees reconstructed from the data sets derived after filtering sequences using the MPTIS had no topological rearrangements due to only a few sequences being filtered (<0.1%).

Table 3.4 - Tree-heterogeneous sequences identified in the Strassert et al., (2021) data set set and the inferred topological rearrangements after their exclusion. The heterogeneous sequences were identified according to the matched-pairs tests of symmetry (MPTS), marginal symmetry (MPTMS), and internal symmetry (MPTIS) combined with four p-value adjustment procedures. These sequences were excluded from each data partition analysed, be a single-protein, merged proteins, or the entire concatenated data set. The inferred topologies were then compared with the original data set derived tree using the normalized Robinson-Foulds (nRF). When the number of taxa was different, the latter metrics was not calculated.

Data filtered	Matched-pairs test	P-value adjustment procedure	Tree-heterogeneous sequences (%)	nRF (%)
Individual proteins	MPTS	Bonferroni	<0.1	3.0
		Holm	<0.1	0.0
		Benjamini-Yekutieli	<0.1	0.0
		Benjamini-Hochberg	<0.1	3.0
	MPTMS	Bonferroni	0.2	4.5
		Holm	0.2	1.5
		Benjamini-Yekutieli	0.3	1.5
		Benjamini-Hochberg	0.7	1.5
	MPTIS	Bonferroni	<0.1	0.0
		Holm	<0.1	0.0
		Benjamini-Yekutieli	<0.1	0.0
		Benjamini-Hochberg	<0.1	0.0
K-means derived partitions	MPTS	Bonferroni	0.5	2.3
		Holm	0.5	2.3
		Benjamini-Yekutieli	0.9	1.5
		Benjamini-Hochberg	1.8	7.3
	MPTMS	Bonferroni	3.1	2.3
		Holm	3.2	1.5
		Benjamini-Yekutieli	5.6	6.0

		Benjamini-Hochberg	10.4	7.5
	MPTIS	Bonferroni	<0.1	0.0
		Holm	<0.1	0.0
		Benjamini-Yekutieli	<0.1	0.0
		Benjamini-Hochberg	<0.1	0.0
ModelFinder derived partitions	MPTS	Bonferroni	1.3	1.5
		Holm	1.3	1.5
		Benjamini-Yekutieli	1.9	1.5
		Benjamini-Hochberg	4.5	2.3
	MPTMS	Bonferroni	7.1	3.0
		Holm	7.2	3.0
		Benjamini-Yekutieli	13.5	6.0
		Benjamini-Hochberg	24.8	8.3
	MPTIS	Bonferroni	0.0	-
		Holm	0.0	-
		Benjamini-Yekutieli	0.0	-
		Benjamini-Hochberg	0.0	-
Concatenated data set	MPTS	Bonferroni	88.2	-
		Holm	89.7	-
		Benjamini-Yekutieli	94.1	-
		Benjamini-Hochberg	95.6	-
	MPTMS	Bonferroni	95.6	-
		Holm	95.6	-
		Benjamini-Yekutieli	97.8	-
		Benjamini-Hochberg	97.8	-
	MPTIS	Bonferroni	2.2	-
		Holm	2.2	-
		Benjamini-Yekutieli	0.0	-
		Benjamini-Hochberg	5.1	-

The optimal data partitioning schemes had 84 and 41 partitions when computed using the K-means algorithm and the ModelFinder tool, respectively. Analyses of the partitions derived from the K-means partitioning scheme identified 7-25% tree-heterogeneous sequences using the MPTMS, and 1-4% using the MPTS (Table 3.4). Analyses using the ModelFinder partitioning scheme identified 3-10% (MPTMS) and 1-2% (MPTS). The MPTIS analyses only indicated one rate-heterogeneous sequence, regardless of the p-value adjustment procedure, and this was observed solely in the K-means based analyses. Similar to the above analyses, the Benjamini method analyses identified more tree-heterogeneous sequences than analyses using the Bonferroni method. The resulting optimal ML trees, except those inferred from the data sets filtered using the MPTIS, had topological rearrangements (Appendix Figs. A36-A51). The optimal ML trees inferred from the data sets derived after filtering sequences

according to the MPTMS coupled with the Benjamini-Hochberg method had the highest nRF (8%), whether using the K-means or the ModelFinder partitioning scheme. The same procedure, but using the Benjamini-Yekutieli method, produced trees with a nRF of 6%. The tree resulting from the data set derived after filtering sequences according to the MPTS combined with Benjamini-Hochberg method had a nRF of 7%. The nRF of the remaining trees varied between 2% and 3%. Trees inferred after sequence filtering based on MPTMS analyses combined with the Benjamini methods and using the K-means partitioning scheme resolved Archaeplastida as monophyletic. This arrangement placed red algae (UFBOOT = 63-71%) sister to a well-supported clade comprising Viridiplantae and Glaucophyta (Appendix Figs. A42 and A43). The Cryptista was recovered sister to the Archaeplastida (UFBOOT >= 95%). This relationship differed from the original tree in which Cryptista was recovered nested in Archaeplastida as the closest relative of the clade Glaucophyta and Viridiplantae. The ModelFinder scheme analyses recovered Cryptista nested in Archaeplastida in all reconstructed trees irrespective of the filtering method. Additionally, the Telonemia clade consistently appeared as the sister-group of the SAR supergroup. By contrast, the Haptista clade moved around the tree, sometimes being closely related to the Cryptista and Archaeplastida clade, and at other times to the TSAR supergroup, though always poorly supported.

The MTPS and MPTMS analyses conducted on the concatenated data set, as a single partition, identified most taxa as tree-heterogeneous (88-98%). Consequently, the resulting data sets, after filtering heterogeneous sequences, were very incomplete and inadequate to address the relationships among the eukaryotic supergroups and the monophyly of Archaeplastida. By contrast, the MPTIS analyses identified only 5% of sequences as rate-heterogeneous when using the Benjamini-Hochberg method, 2% with Bonferroni-based methods, and none when applying the Benjamini-Yekutieli method. The optimal trees inferred from the data sets derived after filtering tree-heterogeneous sequences using the MPTIS combined with Benjamini-Hochberg and Bonferroni methods recovered the Archaeplastida as monophyletic (Appendix Figs. A52, A53, and A54), with the red algae sister to the remaining Archaeplastida taxa (though poorly supported, UFBOOT = 39-46%). In the two trees, the Haptista were the sister-group to the clade comprising Archaeplastida and Cryptista (with no support).

Case study 3-Metazoa (Whelan *et al.*, 2015)

MPTS and MPTIS analyses of individual proteins in the Whelan data set identified less than 0.1% of the total of 11,532 sequences as either rate- or composition-heterogeneous (Table 3.5). The MPTMS analyses identified 3% lineage-specific composition-heterogeneous sequences when combined with the Benjamini-Hochberg method, and 2% when combined with the remaining p-value adjustment procedures. The optimal ML trees inferred from the data sets derived after filtering the composition-heterogeneous sequences using the MPTMS combined with the Benjamini-Hochberg and Benjamini-Yekutieli methods had a nRF of 3% and 1%, respectively. The latter trees were overall congruent with the original tree without filtering of heterogeneous sequences, recovering the same relationships among metazoan groups, and having Ctenophora as the sister-group to all other extant animals (Appendix Figs. 55 and 56). The remaining inferred trees had the same topology as the original tree.

Table 3.5 - Tree-heterogeneous sequences identified in the Whelan et al., (2015) data set set and the inferred topological rearrangements after their exclusion. The heterogeneous sequences were identified according to the matched-pairs tests of symmetry (MPTS), marginal symmetry (MPTMS), and internal symmetry (MPTIS) combined with four p-value adjustment procedures. These sequences were excluded from each data partition analysed, be a single-protein, merged proteins, or the entire concatenated data set. The inferred topologies were then compared with the original data set derived tree using the normalized Robinson-Foulds (nRF). When the number of taxa was different, the latter metrics was not calculated.

Data filtered	Matched-pairs test	P-value adjustment procedure	Tree-heterogeneous sequences (%)	nRF (%)
Individual proteins	MPTS	Bonferroni	0.1	0.0
		Holm	0.1	0.0
		Benjamini-Yekutieli	0.1	0.0
		Benjamini-Hochberg	0.1	0.0
	MPTMS	Bonferroni	1.5	0.0
		Holm	1.5	0.0
		Benjamini-Yekutieli	1.5	1.4
		Benjamini-Hochberg	3.3	2.7
	MPTIS	Bonferroni	0.0	0.0
		Holm	0.0	0.0
		Benjamini-Yekutieli	0.0	0.0
		Benjamini-Hochberg	0.0	0.0
K-means derived partitions	MPTS	Bonferroni	3.9	2.7
		Holm	3.9	2.7
		Benjamini-Yekutieli	5.4	4.1
		Benjamini-Hochberg	8.3	4.1
	MPTMS	Bonferroni	13.2	16.4
		Holm	13.2	11.0

		Benjamini-Yekutieli	17.7	16.4
		Benjamini-Hochberg	25.5	6.8
	MPTIS	Bonferroni	0.1	0.0
		Holm	0.1	0.0
		Benjamini-Yekutieli	0.1	0.0
		Benjamini-Hochberg	0.1	0.0
ModelFinder derived partitions	MPTS	Bonferroni	2.1	2.7
		Holm	2.1	2.7
		Benjamini-Yekutieli	2.6	2.7
		Benjamini-Hochberg	4.8	2.7
	MPTMS	Bonferroni	10.8	2.7
		Holm	11.1	2.7
		Benjamini-Yekutieli	16.3	2.7
		Benjamini-Hochberg	25.7	5.5
	MPTIS	Bonferroni	0.0	0.0
		Holm	0.0	0.0
		Benjamini-Yekutieli	0.0	0.0
		Benjamini-Hochberg	0.0	0.0
Concatenated data set	MPTS	Bonferroni	64.5	-
		Holm	67.1	-
		Benjamini-Yekutieli	75.0	-
		Benjamini-Hochberg	81.6	-
	MPTMS	Bonferroni	76.3	-
		Holm	77.6	-
		Benjamini-Yekutieli	86.8	-
		Benjamini-Hochberg	93.4	-
	MPTIS	Bonferroni	0.0	-
		Holm	0.0	-
		Benjamini-Yekutieli	0.0	-
		Benjamini-Hochberg	0.0	-

The optimal partitioning scheme calculated using the K-means algorithm had 28 partitions, while the ModelFinder scheme had 35 partitions. The MPTS analyses identified 4-8% and 2-5% tree-heterogeneous sequences using the K-means and ModelFinder partitioning schemes respectively (Table 3.5). The MPTMS identified 11-26% and 13-26% composition-heterogeneous sequences using the K-means and ModelFinder partitioning schemes, respectively. The MPTIS analysis indicated none or few sequences (0-0.1%) to have evolved under rate-heterogeneous conditions. The optimal ML trees inferred from the K-means derived data sets and filtered using the MPTS and MPTMS had a nRF of 3-4% and 3-16%, respectively. The relationships across the main metazoan groups were overall congruent with the analyses of the original data set in most cases (Appendix Figs. A57-A60). However, the

analyses of the data sets derived after filtering heterogeneous sequences according to the MPTMS combined with the Benjamini-Yekutieli and -Hochberg methods recovered Ambulacraria (*Strongylocentrotus purpuratus* + *Saccoglosuss kowalevskii*) sister to Chordata (composed of *Petromyzon marinus, Danio rerio,* and *Homo sapiens*), rather than sister to all extant bilaterian animals (Appendix Figs. A61 and A62). The analyses conducted using the ModelFinder scheme inferred trees with a nRF of 3-5% and agreed with the original tree regarding the metazoan group relationships (Appendix Figs. A63-A68).

The analysis of the Whelan data set treated as a single partition and filtered of heterogeneous sequences using the MPTS and MPTMS methods identified most sequences as possessing lineage-specific heterogeneity (64-93%). The data set derived after filtering heterogeneous sequences consequently had a low number of taxa, and therefore inadequate to address the relationships among the metazoan groups. By contrast, no sequence was indicated to have evolved under rate-heterogeneous conditions according to the MPTIS.

3.4 Discussion

Heterogeneity among lineages can introduce biases in phylogenetic inferences, if not accounted for by the analytical model. Thus, understanding these processes in the context of modelling the substitution process is crucial to inferring the most likely tree or at least identifying potential sources of phylogenetic incongruence. In this study, the relative accuracy of the matched-pairs tests of symmetry was assessed when combined with different p-value adjustment procedures to correctly reject or not-reject the null hypothesis of stationarity and homogeneity among lineages of individual taxon sequences.

The simulation analyses presented here included two evolutionary patterns of among-lineage heterogeneity and three sequence alignment lengths. These variables enable the simulation of different strength levels of heterogeneity and, to a certain extent, data ambiguity with respect to the phylogeny-the more data, the stronger is the anticipated effected of among-lineage heterogeneity. The results demonstrated that as alignments became longer the statistical power of the tests increased. Moreover, the higher the divergence of the tree-heterogeneous sequences from the ‘default’ simulation model, the greater the number of heterogeneous sequences were correctly identified.

The MPTMS demonstrated the greatest statistical power in the analyses of the composition-heterogeneous data. Analyses of the empirical data sets corroborated these results, where the MPTMS identified more compositional-heterogeneous sequences than MPTS. The increase in correctly identified tree-heterogeneous sequences by the MPTMS

compensated for the rise in false positives (Type I error) thereby leading to a decrease in the FDR. Indeed, the false positives remained nearly constant between using the 5000- and 8000-site alignments, in particular in the analyses combined with Benjamini methods. These results indicate that there is no increase of false positives associated with the use of longer alignments. In addition, the FDR values derived from the MPTMS analyses were generally either lower or statistically indistinguishable from the FDR values of MPTS. Furthermore, the analyses of the empirical data sets demonstrated that the trees inferred from the data sets filtered using the MPTMS showed more often topological rearrangements likely correct than the tree inferred from the data sets filtered using the MPTS. Overall, these results indicated that the trade-off between error rates and statistical power promotes the use of MPTMS rather than MPTS for the identification of composition-heterogeneous sequences.

The analyses of rate-heterogeneous and composition-heterogeneous simulated data demonstrated that the MPTS had greater statistical power than the MPTIS, although both were inferior to the MPTMS. The advantage of the MPTS over the MPTIS is likely due to its sensitivity to compositional heterogeneity, i.e., the greater power of the MPTS results from its ability to identify compositionally heterogeneity among lineages, rather than assess the rates heterogeneity. However, in the analyses of rate-heterogeneous and composition-homogeneous simulated data, the MPTIS exhibited a slightly higher statistical power and a lower FDR than the MPTS. Nevertheless, the number of rate-heterogeneous sequences identified were overall low, including in the analyses of the empirical data. These results are consistent with previous findings, which demonstrated a lower degree of tree-heterogeneity indicated by the $\text{MaxSymTest}_{\text{int}}$ (this test uses the MPTIS to evaluate only the most divergent taxon pair) compared to other matched-pairs tests (Naser-Khdour *et al.,* 2019). The simulation analyses conducted here demonstrated that although rate-heterogeneous sequences are in the alignment, they were difficult to identify irrespective if using the MPTS or the MPTIS, indicating a poor sensitivity of these tests. This agrees with previous analyses of simulated data, which showed that the statistical power of the MPTIS for detecting the rejection of the assumption of rates homogeneity among lineages (implemented as $\text{MaxSymTest}_{\text{int}}$) was lower than that of other tests (Naser-Khdour *et al.,* 2021).

Using either the Bonferroni and Holm corrections methods produced nearly identical results, whether tested using simulated or empirical data. Therefore, although the Holm method is usually recommended over the Bonferroni for being more powerful (Aickin and Gensler, 1996), no significant differences were observed between the two methods in the analyses of these data. Overall, analyses using the Bonferroni and Holm p-value correction

methods had the lowest statistical power and, in some analyses, also the highest FDR. By contrast, the Benjamini-Yekutieli method had a high statistical power while maintaining low numbers of false positives, whereas the Benjamini-Hochberg method analyses had the highest statistical power, but also the highest number of false positives. The analyses of empirical datasets after filtering the tree-heterogeneous sequences using the Benjamini methods, particularly the Benjamini-Hochberg method, resulted in topological improvements likely correct, which could be seen as indicative of the efficacy of the method. Hence, the substantial higher statistical power coupled with topological improvements in subsequent inference analyses of the empirical data sets with tree-heterogeneous sequences removed indicate that the occurrences of false positives should have little impact. As a consequence, the Benjamini-based methods, should be preferred over the Bonferroni and Holm methods, as the latter exhibited unsurprisingly low power (e.g. Nakagawa, 2004) in addressing the presence of tree-heterogeneous sequences.

The analyses of the Liu data set found a greater prevalence of proteins evolving under among-lineage heterogeneity as well as a higher percentage of tree-heterogeneous sequences compared to the Strassert and Whelan nuclear data. Analyses of mitochondrial data have confirmed the presence of compositional heterogeneity among lineages, which was identified as a cause of systematic bias (e.g., Song *et al.,* 2010; Liu *et al.,* 2014; Sousa *et al.,* 2020). Furthermore, the tree inferred from the Liu data set derived after filtering heterogeneous sequences using the MPTMS coupled with Benjamini-Hochberg recovered Setaphyta monophyletic and bryophytes as a paraphyletic group with lycopods nested within. By contrast, the analyses of the Strassert and Whelan derived data sets using the same procedure inferred topologies very similar to the original trees without sequence filtering. This suggest the absence of among-lineage heterogeneity in single-protein partitions affecting the relationships among Archaeplastida and Metazoa; or if present, the heterogeneity is not sufficiently strong to be identified by the methods used here. Indeed, the small size of data partitions for analysis can constrain the power of the matched-pairs tests, highlighting the importance of carefully balancing the size and number of partitions (Jermiin *et al.,* 2008).

The matched-pairs test analyses conducted on partitions derived from the optimal partitioning schemes and on the single-partitioned data resulted in the identification of more tree-heterogeneous sequences compared with the analyses of single-protein partitions. This enhanced statistical power is due to the use of longer partitions, as demonstrated in the analyses of the simulated data, where the power increased using increased partitions. In addition, the multi-partition filtered data sets or the entire filtered combined data sets were

analysed using a partition model or a tree-homogeneous model, respectively. These data partitions are expected to not reject the model's assumptions of stationarity and/or homogeneity, as the tree-heterogeneous sequences are removed. As a result, the overall model fit and the accuracy of the inferred tree were likely improved.

The K-means partitioning analyses of the Strassert and Whelan data sets recovered Archaeplastida as monophyletic and Ambulacraria the sister-group to Chordata when composition-heterogeneous sequences were removed, topological rearrangements identified as likely to be correct. These results appear to support the effectiveness of the K-means clustering approach, as the identification and removal of composition-heterogeneous sequences led to improved topologies in subsequent inference analyses. By contrast, the analyses of the Liu data set using the K-means partitioning approach recovered bryophytes as a paraphyletic grade contrasting to the recovering of the Setaphyta and bryophytes as a paraphyletic group in analyses using the single-protein alignments. The 41 mitochondrial single-protein alignments used in the K-means clustering were substantial fewer and exhibited greater among-lineage heterogeneity, when compared with the nuclear single-protein alignments which were more namely 209 (Whelan data set) and 320 (Strassert data set) and had a lower rate of tree-heterogeneous sequences. When the number of data points, that is, the number of composition frequency vectors representing each single-protein alignment, is low, single variations might over-bias the clustering results when compared with data sets with higher number of data points, a putative effect of random error. Indeed, these variations may manifest as outliers, which is usually problematic for the K-means algorithm (Jain, 2010). Perhaps, in data sets with limited number of data points, other algorithms would be more efficient, such as the K-medoids which is more robust to outliers (Kaufman and Rousseeuw, 1990; Arora et al., 2016).

Partition model analyses from the Liu data set using the ModelFinder derived scheme recovered bryophytes as monophyletic, though with mosses sister to hornworts rather than liverworts, or as a paraphyletic group with Setaphyta monophyletic and lycopods sister to hornworts, after excluding heterogeneous sequences according to the MPTS and MPTMS respectively. However, the same procedure did not recover a monophyletic Archaeplastida or Ambulacraria sister to Chordata in the analyses of the Strassert and Whelan data sets. The search for the optimal partitioning scheme using nuclear proteins might not be computationally efficient enough due to the high number of initial data partitions, 209 and 320, making it challenging to find the optimal partitioning schemes (see Lanfear *et al.*, 2014).

By contrast, the Liu data set is considerably shorter, which might be easier to find the optimal partitioning scheme.

In analyses of a data set using the matched-pairs tests, consideration should be given to the size and number of partitioned data sets when selecting the most appropriate data. While partitioning the data into numerous small partitions can render among-lineage heterogeneity unidentifiable, larger partitions can compromise a reliable characterisation of the heterogeneity if for instance loci that evolved at different conditions are analyses together. Indeed, treating multiple proteins as a single partition combines regions that could evolve under different constraints, identifying more sequences as heterogeneous in the analysis. The analyses of the single-partitioned data using the MPTS and the MPTMS appears to corroborate that, as most sequences (>75%) were assigned as tree-heterogeneous. By contrast, the proportion of data assigned as rate-heterogeneous by the MPTIS was substantial lower compared to the latter analyses in most cases (0-5%, except in the analyses of the mitochondrial Liu data set using the Benjamini methods, 23-58%). Moreover, trees inferred from the data sets with the identified tree-heterogeneous sequences removed recovered Setaphyta monophyletic within bryophytes as a paraphyletic group, and Archaeplastida monophyletic, suggesting that the sequences removed were correctly identified as source of bias. Due to the consistently low statistical power of the MPTIS, perhaps only the sequences extremely rate-heterogeneous affecting the phylogenetic inference were identified in the analyses of the single-partitioned data. Nevertheless, excluding falsely identified tree-heterogeneous sequences from the analysis may not impact the results, especially in the context of current large data sets, provided that there are still sufficient data to obtain a robust phylogeny.

The monophyly of the bryophytes and clade Setaphyta uniting the mosses and liverworts to the exclusion of hornworts has been argued to represent the most likely evolutionary history of these taxa (Sousa *et al.,* 2019; 2020). These topologies were inferred in this study after removing the composition-heterogeneous sequences from the mitochondrial Liu data set, confirming the effect of systematic bias caused by compositional heterogeneity among lineages in the analyses of the original data set without heterogeneous sequences removed, which corroborates previous studies (Liu *et al.,* 2014; Sousa *et al.,* 2020). In addition, the relationships among bryophytes were shown to be also affected by rate heterogeneity among lineages, as demonstrated in the analyses using the entire combined data set and the MPTIS. Similarly, obtaining a monophyletic Archaeplastida after the filtering analyses conduct here indicates that failing to recovering Archaeplastida as a monophyletic

group in the analyses of the original Strassert data set is a biased result caused by among-lineage heterogeneity. However, the filtering of single-protein alignments, in contrast to subset partitions and combined data set, did not lead to a monophyletic Archaeplastida in subsequent inference analyses therefore suggesting that the among-lineage heterogeneity is weak at single protein level in this data set. In the analyses of the Whelan data set, none of the topological rearrangements affected the position of Porifera and Ctenophora. This suggests that despite their placement, which comprises a contentious evolutionary relationship at the base of metazoan, among-lineage heterogeneity appears to not affect these relationships. Nevertheless, the phylogenetic inference of the relationship between Ambulacraria and Chordata was suggested to be biased by composition-heterogeneous data.

Analyses of the empirical data sets demonstrated important topological rearrangements that were congruent with current phylogenetic theory for many taxa; however, these relationships were not well-supported in most cases. This is unsurprising, as it has been shown that these rearrangements are difficult to infer correctly. Problems such as a low signal-to-noise ratio, among-site heterogeneity, and models that are too simple for the complexity of the data likely contribute for the lack of support. Furthermore, the joint effects of different unmodelled heterogeneities, which are usually exceedingly difficult to accommodate, may be affecting these relationships. For instance, the impact of rate-heterogeneity among lineages in the analyses presented here was difficult to gauge due to the low sensitivity of the MPTIS and MPTS for identifying these sequences. It is important to be cautious in assuming that the lack of identification of rate-heterogeneous sequences represents an absence of a systematic bias in the tree inference due to such a process. Indeed, tree-heterogeneous models have shown improved fit to the data compared to tree-homogeneous models in the analyses of among-lineage heterogeneous data even when MPTIS failed to detect heterogeneity (Foster 2020, unpublished).

In this study among-lineage heterogeneity was evaluated at level of a single taxon following their exclusion from analyses, and is conceptually similar to tree evaluation after the pruning of rogue taxa (e.g., Aberer *et al.,* 2013). Nevertheless, rogue sequence detection methods are topology-based and detect taxa which tend to move around the tree (i.e., their placement is ambiguous with several positions in the tree having similar tree scores) due to ambiguous or insufficient phylogenetic information. While these methods aim to improve tree inference, they do not primarily focus on investigating the causes of bias, such as among-lineage heterogeneity. Methods such the matched-pairs tests enable assessing the compositional heterogeneity among lineages. The matched-pairs tests implemented in IQ-

TREE only consider the two most divergent sequences (Naser-Khdour *et al.*, 2019). Therefore, if these two sequences are homogeneous with each other, the full partition is also assigned as tree-homogeneous, resulting in a false negative when other sequences within are tree-heterogeneous. On other the hand, assigned the most divergent taxon pair as tree-heterogeneous does not enable understating the prevalence of among-lineage heterogeneity in the partition under analysis, which contrasts with the multi-pairwise comparisons conducted here (see also Jermiin *et al.*, 2020).

3.5 Conclusions

The analyses presented here demonstrated that the MPTMS exhibits the highest statistical power in identifying composition-heterogeneous data. The influence of rate-homogeneous sequences was only perceptible by MPTIS in longer alignments, hence analyses using this test on smaller data sets should be viewed with caution. Because the MPTS has a lower statistical power than the MPTMS in identifying composition-heterogeneous sequences, and similar or lower than the MPTIS identifying rate-heterogeneous sequences, it does not seem logical to choose the MPTS over the MPTMS or MPTIS. The evaluation of p-value adjustment procedures indicated that the Benjamini-Hochberg offers overall the best trade-off between statistical power and error rates. The implementation of these methods in filtering composition- and rate-heterogeneous sequences from empirical data sets, in particular the MPTMS combined with the Benjamini-Hochberg, lead to notable topological shifts in subsequent inference analyses, indicated that incorrect topologies were driven by among-lineage heterogeneity.

3.6 References

Ababneh, F., Jermiin, L. S., Ma, C., & Robinson, J. (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, *22*(10), 1225–1231. https://doi.org/10.1093/bioinformatics/btl064

Aberer, A. J., Krompass, D., & Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, *62*(1), 162–166. https://doi.org/10.1093/sysbio/sys078

Aberer, A. J., Krompaß, D., & Stamatakis, A. (2011). A Simple and Accurate Method for Rogue Taxon Identification. *IEEE International Conference on Bioinformatics and Biomedicine*, Atlanta, GA, USA, 2011, pp. 118-122, doi: 10.1109/BIBM.2011.70.

Adachi, J., Waddell, P. J., Martin, W., & Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution*, *50*(4), 348–358. https://doi.org/10.1007/s002399910038

Aickin, M. & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American Journal of Public Health* 86, 726_728, https://doi.org/10.2105/AJPH.86.5.726

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Physics Procedia*, 78(December 2015), 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Benjamini, Y, & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. http://dx.doi.org/10.1214/aos/1013699998

Benjamini, Yoav, & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boussau, B., & Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology*, *55*(5), 756–768. https://doi.org/10.1080/10635150600975218

Bowker, A. H. (1948). A Test for Symmetry in Contingency Tables. *Journal of the American Statistical Association*, *43*(244), 572–574. https://doi.org/10.1080/01621459.1948.10483284

Brown, J. M., & Thomson, R. C. (2018). Evaluating Model Performance in Evolutionary Biology Random variable: a variable that takes on different values based on the outcome of a random process. *Annu. Rev. Ecol. Evol. Syst*, *49*, 379–408. https://doi.org/10.1146/annurev-ecolsys-110617

Bryant, D., Galtier, N., & Poursat, M.-A. (2005). Likelihood calculation in molecular phylogenetics. *Mathemathics of Evolution and Phylogeny*, 33–62. http://dx.doi.org/10.1093/oso/9780198566106.003.0002

Burki, F., Kaplan, M., Tikhonenkov, D. V., Zlatogursky, V., Minh, B. Q., Radaykina, L. V., Smirnov, A., Mylnikov, A. P., & Keeling, P. J. (2016). Untangling the early diversification of eukaryotes: A phylogenomic study of the evolutionary origins of centrohelida, haptophyta and cryptista. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1823). https://doi.org/10.1098/rspb.2015.2802

Chernomor, O., Von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, *65*(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., & Embley, T. M. (2008). The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, *105*(51), 20356–20361. https://doi.org/10.1073/pnas.0810647105

Feuda, R., Pisani, D., Rota-Stabelli, O., Lartillot, N., Pett, W., Dohrmann, M., Wörheide, G., & Philippe, H. (2017). Improved modelling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology*, *27*(24), 3864-3870.e4. https://doi.org/10.1016/j.cub.2017.11.008

Fleming, J. F., Feuda, R., Roberts, N. W., & Pisani, D. (2020). A Novel Approach to Investigate the Effect of Tree Reconstruction artefacts in Single-Gene Analysis Clarifies Opsin Evolution in Nonbilaterian Metazoans. *Genome Biology and Evolution*, *12*(2), 3906–3916. https://doi.org/10.1093/gbe/evaa015

Fleming, J. F., & Struck, T. H. (2022). nRCFV: a new, data set-size-independent metric to quantify compositional heterogeneity in nucleotide and amino acid data sets. *BMC Bioinformatics*, *24*(1), 145. https://doi.org/10.1186/s12859-023-05270-8

Fleming, J. F., Valero-Gracia, A., & Struck, T. H. (2023). Identifying and addressing methodological incongruence in phylogenomics: A review. In *Evolutionary Applications*. John Wiley and Sons Inc. https://doi.org/10.1111/eva.13565

Foster, P. G., Cox, C. J., & Martin 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*(1527), 2197–2207. https://doi.org/10.1098/rstb.2009.0034

Foster, P. G. (2004). modelling compositional heterogeneity. *Systematic Biology*, *53*(3), 485–495. https://doi.org/10.1080/10635150490445779

Foster, P. G., & Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, *48*(3), 284–290. https://doi.org/10.1007/PL00006471

Foster, P. G. (2020). Unpublished. http://pgfsites.s3-website-eu-west-1.amazonaws.com/H20/TreeHeteroRMatrixA/treeHeteroRMatrix.html

Ho, S. Y. W., & Jermiin, L. S. (2004). Tracing the decay of the historical signal in biological sequence data. *Systematic Biology*, *53*(4), 623–637. https://doi.org/10.1080/10635150490503035

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Le, S. V. (2017). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. *Molecular Biology and Evolution*, *35*(2), msx281. https://doi.org/10.5281/zenodo.854445

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. http://www.jstor.org/stable/4615733

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jermiin, L. S., Catullo, R. A., & Holland, B. R. (2020). A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. *NAR Genomics and Bioinformatics*, *2*(2), 1–14. https://doi.org/10.1093/nargab/lqaa041

Jermiin, L. S., Lovell, D. R., Misof, B., & Foster, P. G. (2020). Detecting and Visualising the Impact of Heterogeneous Evolutionary Processes on Phylogenetic Estimates. BioRxiv. https://doi.org/10.1101/828996

Jermiin, L.S., Jayaswal, V., Ababneh, F., Robinson, J. (2008). Phylogenetic Model Evaluation. In: Keith, J.M. (eds) *Bioinformatics. Methods in Molecular Biology*, vol 452. Humana Press. https://doi.org/10.1007/978-1-60327-159-2_16

Jermiin, L. S. (2017). Identifying Optimal Models of Evolution. In *Methods in molecular biology* (Vols. 347–369, Issue November). https://doi.org/10.1007/978-1-4939-6622-6

Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J., & Larkum, A. W. D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated The Biasing Effect of Compositional Heterogeneity on Phylogenetic Estimates May be Underestimated, *Systematic Biology*, Volume 53, Issue 4, Pages 638–643, https://doi.org/10.1080/10635150490468648

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* : CABIOS, 8(3), 275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*, 587. http://dx.doi.org/10.1038/nmeth.4285

Katz, L. A. & Grant, J. R. (2015). Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Systematic Biology*, 64 (3), 406–415. https://doi.org/10.1093/sysbio/syu126

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: an introduction to cluster analysis. *John Wiley & Sons*. http://dx.doi.org/10.1002/9780470316801

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, 17(6), 441–458. http://www.jstor.org/stable/2486927

Kück, P., & Struck, T. H. (2014). BaCoCa-A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Molecular Phylogenetics and Evolution*, *70*, 94–98. https://doi.org/https://doi.org/10.1016/j.ympev.2013.09.011

Kupczok, A., Haeseler, A. Von, & Klaere, S. (2008). An exact algorithm for the geodesic distance between phylogenetic trees. *Journal of Computational Biology*, *15*(6), 577–591. https://doi.org/10.1089/cmb.2008.0068

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., & Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. BMC *Evolutionary Biology*, 14(1), 1–14. https://doi.org/10.1186/1471-2148-14-82

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. https://doi.org/10.1093/molbev/msh112

Lax, G., Eglit, Y., Eme, L., Bertrand, E. M., Roger, A. J., & Simpson, A. G. B. (2018). Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature*, *564*(7736), 410–414. https://doi.org/10.1038/s41586-018-0708-8

Liu, Y., Cox, C. J., Wang, W., & Goffinet, B. (2014). Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Systematic Biology*, *63*(6), 862–878. https://doi.org/10.1093/sysbio/syu049

Mai, U., & Mirarab, S. (2018). TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, *19*(Suppl 5). https://doi.org/10.1186/s12864-018-4620-2

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Morgan, C., Foster, P. G., Webb, A. E., Pisani, D., McInerney, J. O., & O'Connell, M. J. (2013). Heterogeneous models place the root of the placental mammal phylogeny. *Molecular Biology and Evolution*, *30*(9), 2145–2156. https://doi.org/10.1093/molbev/mst117

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6), 1044–1045. https://doi.org/10.1093/beheco/arh107

Naser-Khdour, S., Minh, B. Q., & Robert, L. (2021). The Influence of Model Violation on Phylogenetic Inference: A Simulation Study. *BioRxiv*. https://doi.org/10.1101/2021.09.22.461455

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., Lanfear, R., & Bryant, D. (2019). The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biology and Evolution*, *11*(12), 3341–3352. https://doi.org/10.1093/gbe/evz193

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Phillips, M. J., & Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. *Molecular Phylogenetics and Evolution*, *28*(2), 171–185. https://doi.org/10.1016/S1055-7903(03)00057-5

Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., Wörheide, G., Philippe, H., Pisani, D., Dohrmann, M., Wörheide, G., Pett, W., & Rota-Stabelli, O. (2015). Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the National Academy of Sciences*, *112*(50), 15402–15407. https://doi.org/10.1073/pnas.1518127112

Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Schneider, H., Pisani, D., & Donoghue, P. C. J. (2018). The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology*, *28*(5), 733-745.e2. https://doi.org/10.1016/j.cub.2018.01.063

Redmond, A. K., & McLysaght, A. (2021). Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-22074-7

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., & Manuel, M. (2017). A Large and Consistent Phylogenomic data set Supports Sponges as the sister group to All Other Animals. *Current Biology*, *27*(7), 958–967. https://doi.org/10.1016/j.cub.2017.02.031

Song, H., Sheffield, N. C., Cameron, S. L., Miller, K. B., & Whiting, M. F. (2010). When phylogenetic assumptions are violated: Base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Systematic Entomology*, 35(3), 429–448. https://doi.org/10.1111/j.1365-3113.2009.00517.x

Soria-Carrasco, V., Talavera, G., Igea, J., & Castresana, J. (2007). The K tree score: Quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinformatics*, *23*(21), 2954–2956. https://doi.org/10.1093/bioinformatics/btm466

Sousa, F., Civáň, P., Foster, P. G., & Cox, C. J. (2020). The Chloroplast Land Plant Phylogeny: Analyses Employing Better-Fitting Tree- and Site-Heterogeneous Composition Models. *Frontiers in Plant Science*, *11*(July), 1–10. https://doi.org/10.3389/fpls.2020.01062

Sousa, F., Civáň, P., Brazão, J., Foster, P. G., & Cox, C. J. (2020). The mitochondrial phylogeny of land plants shows support for Setaphyta under composition-heterogeneous substitution models. *PeerJ*, *2020*(4), e8995. https://doi.org/10.7717/peerj.8995

Sousa, F., Foster, P. G., Donoghue, P. C. J., Schneider, H., & Cox, C. J. (2019). Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist*, *222*(1), 565–575. https://doi.org/10.1111/nph.15587

Strassert, J. F. H., Irisarri, I., Williams, T. A., & Burki, F. (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nature Communications*, 12(1), 1–13. https://doi.org/10.1038/s41467-021-22044-z

Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V., & Burki, F. (2019). New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Molecular Biology and Evolution*, *36*(4), 757–765. https://doi.org/10.1093/molbev/msz012

Struck, T. H. (2014). Trespex-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolutionary Bioinformatics*, *10*, 51–67. https://doi.org/10.4137/EBo.s14239

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, *42*(3–4), 412–416. https://doi.org/10.1093/biomet/42.3-4.412

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Whelan, N. V., Kocot, K. M., Moroz, L. L., & Halanych, K. M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(18), 5773–5778. https://doi.org/10.1073/pnas.1503453112

Wilkinson, M. (1996). Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution*, *13* (3), 437–444. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A025604

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., & Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology and Evolution*, *4*(1), 138–147. https://doi.org/10.1038/s41559-019-1040-x

Williams, T. A., Schrempf, D., Szöllősi, G. J., Cox, C. J., Foster, P. G., & Embley, T. M. (2021). Inferring the deep past from molecular data. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evab067

3.7 Appendix

Table A1 - Hypothesis testing for compositional heterogeneity. The matched-pairs test of symmetry (MPTS) and the matched-pairs test of marginal symmetry (MPTMS) combined with four p-value adjustment procedures were evaluated regarding their ability to correctly reject or accept the assumption of stationarity using simulated composition-heterogeneous and rate-homogeneous alignments. The known composition-heterogeneous sequences were simulated using the JTT composition frequencies. The false discovery rate (FDR) and the statistical power were calculated from the sum of 100 statistic replicates. Type I errors refer to sequences incorrectly rejected, while Type II errors refer to sequences falsely not-rejected.

Data set length (N=100)	Matched-pairs test	P-value adjustment procedure	JTT-derived sequences correctly rejected	Type I errors	Sequences correctly not-rejected	Type II errors	FDR	Power
2000	MPTS	Bonferroni	1	3	1797	599	0.75	0.00
		Holm	1	3	1797	599	0.75	0.00
		Benjamini-Yekutieli	0	2	1798	600	1.00	0.00
		Benjamini-Hochberg	0	2	1798	600	1.00	0.00
	MPTMS	Bonferroni	8	87	1713	592	0.92	0.01
		Holm	8	87	1713	592	0.92	0.01
		Benjamini-Yekutieli	26	59	1741	574	0.69	0.04
		Benjamini-Hochberg	104	162	1638	496	0.61	0.17
5000	MPTS	Bonferroni	4	65	1735	596	0.94	0.01
		Holm	4	65	1735	596	0.94	0.01
		Benjamini-Yekutieli	5	41	1759	595	0.89	0.01
		Benjamini-Hochberg	99	132	1668	501	0.57	0.17
	MPTMS	Bonferroni	35	29	1771	565	0.45	0.06
		Holm	43	33	1767	557	0.43	0.07
		Benjamini-Yekutieli	177	86	1714	423	0.33	0.30
		Benjamini-Hochberg	390	174	1626	210	0.31	0.65
8000	MPTS	Bonferroni	3	30	1770	597	0.91	0.01
		Holm	4	31	1769	596	0.89	0.01
		Benjamini-Yekutieli	54	79	1721	546	0.59	0.09

		Benjamini-Hochberg	269	231	1569	331	0.46	0.45
	MPTMS	Bonferroni	130	64	1736	470	0.33	0.22
		Holm	155	63	1737	445	0.29	0.26
		Benjamini-Yekutieli	386	87	1713	214	0.18	0.64
		Benjamini-Hochberg	531	181	1619	69	0.25	0.89

Table A2 - Hypothesis testing for compositional heterogeneity. The matched-pairs test of symmetry (MPTS) and the matched-pairs test of marginal symmetry (MPTMS) combined with four p-value adjustment procedures were evaluated regarding their ability to correctly reject or accept the assumption of stationarity using simulated composition-heterogeneous and rate-homogeneous alignments. The known composition-heterogeneous sequences were simulated using the cpREV composition frequencies. The false discovery rate (FDR) and the statistical power were calculated from the sum of 100 statistic replicates. Type I errors refer to sequences incorrectly rejected, while Type II errors refer to sequences falsely not-rejected.

Data set length (N=100)	**Matched-pairs test**	**P-value adjustment procedure**	**cpREV-derived sequences correctly rejected**	**Type I errors**	**Sequences correctly not-rejected**	**Type II errors**	**FDR**	**Power**
2000	MPTS	Bonferroni	15	3	1797	585	0.17	0.03
		Holm	15	3	1797	585	0.17	0.03
		Benjamini-Yekutieli	7	2	1798	593	0.22	0.01
		Benjamini-Hochberg	22	2	1798	578	0.08	0.04
	MPTMS	Bonferroni	327	87	1713	273	0.21	0.55
		Holm	331	87	1713	269	0.21	0.55
		Benjamini-Yekutieli	440	59	1741	160	0.12	0.73
		Benjamini-Hochberg	569	162	1638	31	0.22	0.95
5000	MPTS	Bonferroni	370	65	1735	230	0.15	0.62
		Holm	373	65	1735	227	0.15	0.62
		Benjamini-Yekutieli	449	41	1759	151	0.08	0.75
		Benjamini-Hochberg	588	132	1668	12	0.18	0.98
	MPTMS	Bonferroni	600	29	1771	0	0.05	1.00

		Holm	600	33	1767	0	0.05	1.00
		Benjamini-Yekutieli	600	86	1714	0	0.13	1.00
		Benjamini-Hochberg	600	174	1626	0	0.22	1.00
8000	MPTS	Bonferroni	581	30	1770	19	0.05	0.97
		Holm	583	31	1769	17	0.05	0.97
		Benjamini-Yekutieli	598	79	1721	2	0.12	1.00
		Benjamini-Hochberg	600	231	1569	0	0.28	1.00
	MPTMS	Bonferroni	600	64	1736	0	0.10	1.00
		Holm	600	63	1737	0	0.10	1.00
		Benjamini-Yekutieli	600	87	1713	0	0.13	1.00
		Benjamini-Hochberg	600	181	1619	0	0.23	1.00

Table A3 - Hypothesis testing for rates heterogeneity. The matched-pairs test of symmetry (MPTS) and the matched-pairs test of internal symmetry (MPTIS) combined with four p-value adjustment procedures were evaluated regarding their ability to correctly reject or accept the assumption of homogeneity using simulated composition-homogeneous and rate-heterogeneous alignments. The known rate-heterogeneous sequences were simulated using the JTT instantaneous exchange rates. The false discovery rate (FDR) and the statistical power were calculated from the sum of 100 statistic replicates. Type I errors refer to sequences incorrectly rejected, while Type II errors refer to sequences falsely not-rejected.

Data set length (N=100)	**Matched-pairs test**	**P-value adjustment procedure**	**JTT-derived sequences correctly rejected**	**Type I errors**	**Sequences correctly not-rejected**	**Type II errors**	**FDR**	**Power**
2000	MPTS	Bonferroni	0	0	1800	1199	0.00	0.00
		Holm	0	0	1800	1199	0.00	0.00
		Benjamini-Yekutieli	0	0	1800	1200	0.00	0.00
		Benjamini-Hochberg	0	0	1800	1199	0.00	0.00
	MPTIS	Bonferroni	0	0	1800	1200	0.00	0.00
		Holm	0	0	1800	1200	0.00	0.00
		Benjamini-Yekutieli	0	0	1800	1200	0.00	0.00

		Benjamini-Hochberg	0	0	1800	1200	0.00	0.00
5000	MPTS	Bonferroni	13	15	1785	587	0.54	0.02
		Holm	13	15	1785	587	0.54	0.02
		Benjamini-Yekutieli	2	2	1798	598	0.50	0.00
		Benjamini-Hochberg	12	8	1792	588	0.40	0.02
	MPTIS	Bonferroni	15	20	1780	585	0.57	0.03
		Holm	15	20	1780	585	0.57	0.03
		Benjamini-Yekutieli	4	5	1795	596	0.56	0.01
		Benjamini-Hochberg	9	8	1792	591	0.47	0.02
8000	MPTS	Bonferroni	18	39	1761	582	0.68	0.03
		Holm	18	40	1760	582	0.69	0.03
		Benjamini-Yekutieli	10	13	1787	590	0.57	0.02
		Benjamini-Hochberg	29	61	1739	571	0.68	0.05
	MPTIS	Bonferroni	18	38	1762	582	0.68	0.03
		Holm	18	38	1762	582	0.68	0.03
		Benjamini-Yekutieli	8	19	1781	592	0.70	0.01
		Benjamini-Hochberg	36	59	1741	564	0.62	0.06

Table A4 - Hypothesis testing for rates heterogeneity. The matched-pairs test of symmetry (MPTS) and the matched-pairs test of internal symmetry (MPTIS) combined with four p-value adjustment procedures were evaluated regarding their ability to correctly reject or accept the assumption of homogeneity using simulated composition-homogeneous and rate-heterogeneous alignments. The known rate-heterogeneous sequences were simulated using the stmtREV instantaneous exchange rates. The false discovery rate (FDR) and the statistical power were calculated from the sum of 100 statistic replicates. Type I errors refer to sequences incorrectly rejected, while Type II errors refer to sequences falsely not-rejected.

Data set length (N=100)	**Matched-pairs test**	**P-value adjustment procedure**	**stmtREV-derived sequences correctly rejected**	**Type I errors**	**Sequences correctly not-rejected**	**Type II errors**	**FDR**	**Power**
2000	MPTS	Bonferroni	1	0	1800	1199	0.00	0.00
		Holm	1	0	1800	1199	0.00	0.00
		Benjamini-Yekutieli	0	0	1800	1200	0.00	0.00
		Benjamini-Hochberg	1	0	1800	1199	0.00	0.00
	MPTIS	Bonferroni	0	0	1800	1200	0.00	0.00
		Holm	0	0	1800	1200	0.00	0.00
		Benjamini-Yekutieli	0	0	1800	1200	0.00	0.00
		Benjamini-Hochberg	0	0	1800	1200	0.00	0.00
5000	MPTS	Bonferroni	45	15	1785	555	0.25	0.07
		Holm	45	15	1785	555	0.25	0.07
		Benjamini-Yekutieli	13	2	1798	587	0.13	0.02
		Benjamini-Hochberg	91	8	1792	509	0.08	0.15
	MPTIS	Bonferroni	47	20	1780	553	0.30	0.08
		Holm	47	20	1780	553	0.30	0.08
		Benjamini-Yekutieli	26	5	1795	574	0.16	0.04
		Benjamini-Hochberg	119	8	1792	481	0.06	0.20
8000	MPTS	Bonferroni	178	39	1761	422	0.18	0.30
		Holm	179	40	1760	421	0.18	0.30
		Benjamini-	128	13	1787	472	0.09	0.21

		Yekutieli						
		Benjamini-Hochberg	472	61	1739	128	0.11	0.79
	MPTIS	Bonferroni	228	38	1762	372	0.14	0.38
		Holm	228	38	1762	372	0.14	0.38
		Benjamini-Yekutieli	201	19	1781	399	0.09	0.34
		Benjamini-Hochberg	514	59	1741	86	0.10	0.86

Table A5 - Hypothesis testing for compositional heterogeneity and rates heterogeneity. The matched-pairs tests of symmetry (MPTS), marginal symmetry (MPTMS) and internal symmetry (MPTIS) combined with four p-value adjustment procedures were evaluated regarding their ability to correctly reject or accept the evolutionary model assumptions using simulated compositional- and rate-heterogeneous alignments. The known tree-heterogeneous sequences were simulated using the JTT model. The false discovery rate (FDR) and the statistical power were calculated from the sum of 100 statistic replicates. Type I errors refer to sequences incorrectly rejected, while Type II errors refer to sequences falsely not-rejected.

Data set length (N=100)	**Matched-pairs test**	**P-value adjustment procedure**	**JTT-derived sequences correctly rejected**	**Type I errors**	**Sequences correctly not-rejected**	**Type II errors**	**FDR**	**Power**
2000	MPTS	Bonferroni	0	1	1799	600	1.00	0.00
		Holm	0	1	1799	600	1.00	0.00
		Benjamini-Yekutieli	12	20	1780	588	0.63	0.02
		Benjamini-Hochberg	83	138	1662	517	0.62	0.14
	MPTMS	Bonferroni	6	5	1795	594	0.45	0.01
		Holm	6	7	1793	594	0.54	0.01
		Benjamini-Yekutieli	49	51	1749	551	0.51	0.08
		Benjamini-Hochberg	166	193	1607	434	0.54	0.28
	MPTIS	Bonferroni	0	0	1800	600	0.00	0.00
		Holm	0	0	1800	600	0.00	0.00
		Benjamini-Yekutieli	0	0	1800	600	0.00	0.00
		Benjamini-Hochberg	0	0	1800	600	0.00	0.00
5000	MPTS	Bonferroni	4	4	1796	596	0.50	0.01
		Holm	5	6	1794	595	0.55	0.01
		Benjamini-	61	53	1747	539	0.46	0.10

		Yekutieli						
		Benjamini-Hochberg	240	210	1590	360	0.47	0.40
	MPTMS	Bonferroni	45	31	1769	555	0.41	0.07
		Holm	55	33	1767	545	0.38	0.09
		Benjamini-Yekutieli	220	92	1708	380	0.29	0.37
		Benjamini-Hochberg	408	178	1622	192	0.30	0.68
	MPTIS	Bonferroni	7	12	1788	593	0.63	0.01
		Holm	6	12	1788	594	0.67	0.01
		Benjamini-Yekutieli	2	3	1797	598	0.60	0.00
		Benjamini-Hochberg	2	5	1795	598	0.71	0.00
8000	MPTS	Bonferroni	22	18	1782	578	0.45	0.04
		Holm	26	17	1783	574	0.40	0.04
		Benjamini-Yekutieli	150	68	1732	450	0.31	0.25
		Benjamini-Hochberg	387	262	1538	213	0.40	0.65
	MPTMS	Bonferroni	157	57	1743	443	0.27	0.26
		Holm	181	67	1733	419	0.27	0.30
		Benjamini-Yekutieli	396	83	1717	204	0.17	0.66
		Benjamini-Hochberg	540	181	1619	60	0.25	0.90
	MPTIS	Bonferroni	36	31	1769	564	0.46	0.06
		Holm	35	31	1769	565	0.47	0.06
		Benjamini-Yekutieli	18	11	1789	582	0.38	0.03
		Benjamini-Hochberg	38	64	1736	562	0.63	0.06

Table A6 - Hypothesis testing for compositional heterogeneity and rates heterogeneity. The matched-pairs tests of symmetry (MPTS), marginal symmetry (MPTMS) and internal symmetry (MPTIS) combined with four p-value adjustment procedures were evaluated regarding their ability to correctly reject or accept the evolutionary model assumptions using simulated compositional- and rate-heterogeneous alignments. The known tree-heterogeneous sequences were simulated using the stmtREV model. The false discovery rate (FDR) and the statistical power were calculated from the sum of 100 statistic replicates. Type I errors refer to sequences incorrectly rejected, while Type II errors refer to sequences falsely not-rejected.

Data set length (N=100)	**Matched-pairs test**	**P-value adjustment procedure**	**stmtREV-derived sequences correctly rejected**	**Type I errors**	**Sequences correctly not-rejected**	**Type II errors**	**FDR**	**Power**
2000	MPTS	Bonferroni	599	1	1799	1	0.00	1.00
		Holm	599	1	1799	1	0.00	1.00
		Benjamini-Yekutieli	600	20	1780	0	0.03	1.00
		Benjamini-Hochberg	600	138	1662	0	0.19	1.00
	MPTMS	Bonferroni	600	5	1795	0	0.01	1.00
		Holm	600	7	1793	0	0.01	1.00
		Benjamini-Yekutieli	600	51	1749	0	0.08	1.00
		Benjamini-Hochberg	600	193	1607	0	0.24	1.00
	MPTIS	Bonferroni	0	0	1800	600	0.00	0.00
		Holm	0	0	1800	600	0.00	0.00
		Benjamini-Yekutieli	0	0	1800	600	0.00	0.00
		Benjamini-Hochberg	0	0	1800	600	0.00	0.00
5000	MPTS	Bonferroni	600	4	1796	0	0.01	1.00
		Holm	600	6	1794	0	0.01	1.00
		Benjamini-Yekutieli	600	53	1747	0	0.08	1.00
		Benjamini-Hochberg	600	210	1590	0	0.26	1.00
	MPTMS	Bonferroni	600	31	1769	0	0.05	1.00
		Holm	600	33	1767	0	0.05	1.00
		Benjamini-Yekutieli	600	92	1708	0	0.13	1.00
		Benjamini-Hochberg	600	178	1622	0	0.23	1.00
	MPTIS	Bonferroni	29	12	1788	571	0.29	0.05
		Holm	30	12	1788	570	0.29	0.05

		Benjamini-Yekutieli	8	3	1797	592	0.27	0.01
		Benjamini-Hochberg	63	5	1795	537	0.07	0.11
8000	MPTS	Bonferroni	600	18	1782	0	0.03	1.00
		Holm	600	17	1783	0	0.03	1.00
		Benjamini-Yekutieli	600	68	1732	0	0.10	1.00
		Benjamini-Hochberg	600	262	1538	0	0.30	1.00
	MPTMS	Bonferroni	600	57	1743	0	0.09	1.00
		Holm	600	67	1733	0	0.10	1.00
		Benjamini-Yekutieli	600	83	1717	0	0.12	1.00
		Benjamini-Hochberg	600	181	1619	0	0.23	1.00
	MPTIS	Bonferroni	221	31	1769	379	0.12	0.37
		Holm	222	31	1769	378	0.12	0.37
		Benjamini-Yekutieli	233	11	1789	367	0.05	0.39
		Benjamini-Hochberg	523	64	1736	77	0.11	0.87

Figure A1 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (Liu *et al*., 2014 data set). Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -165805. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

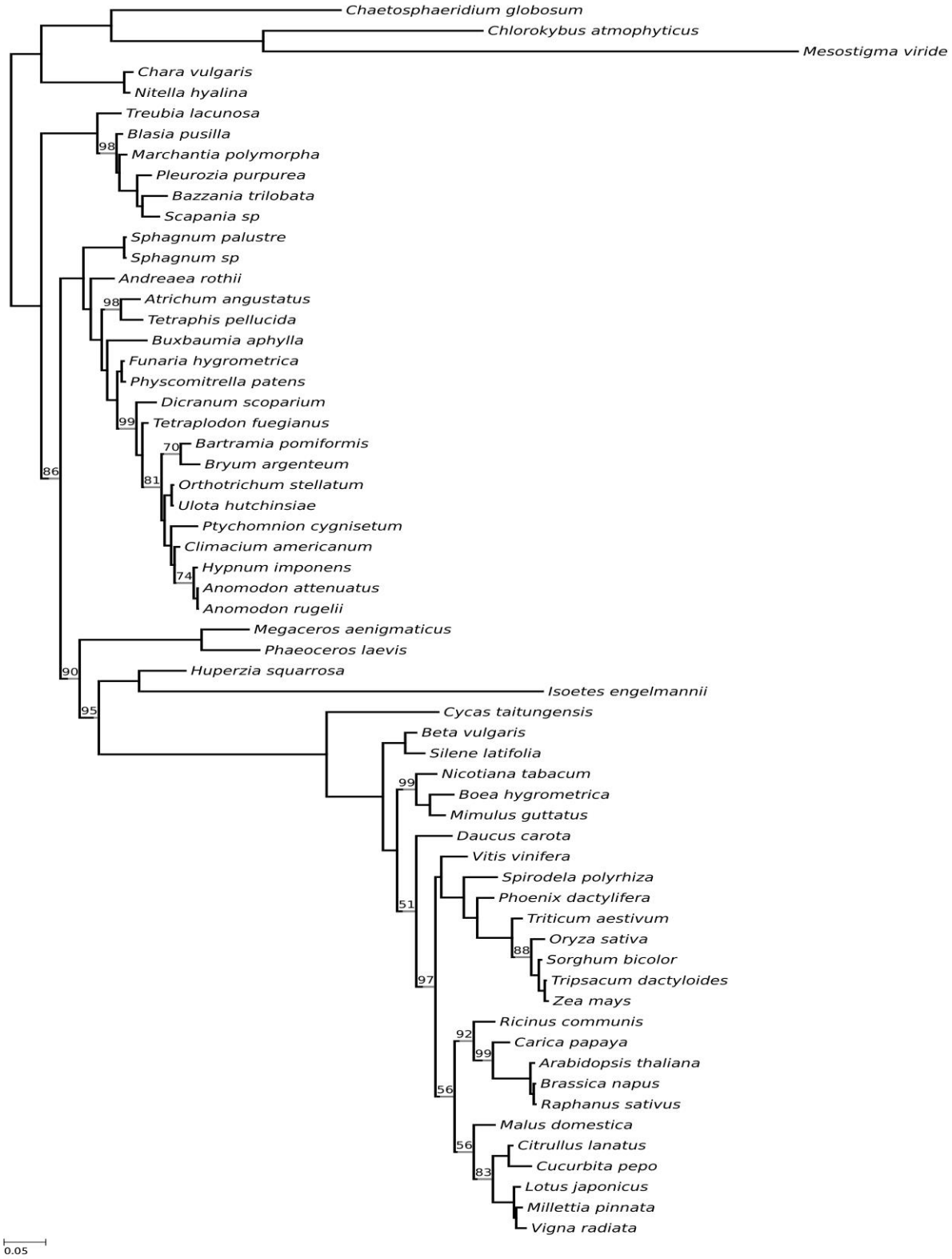


Figure A2 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -160503. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

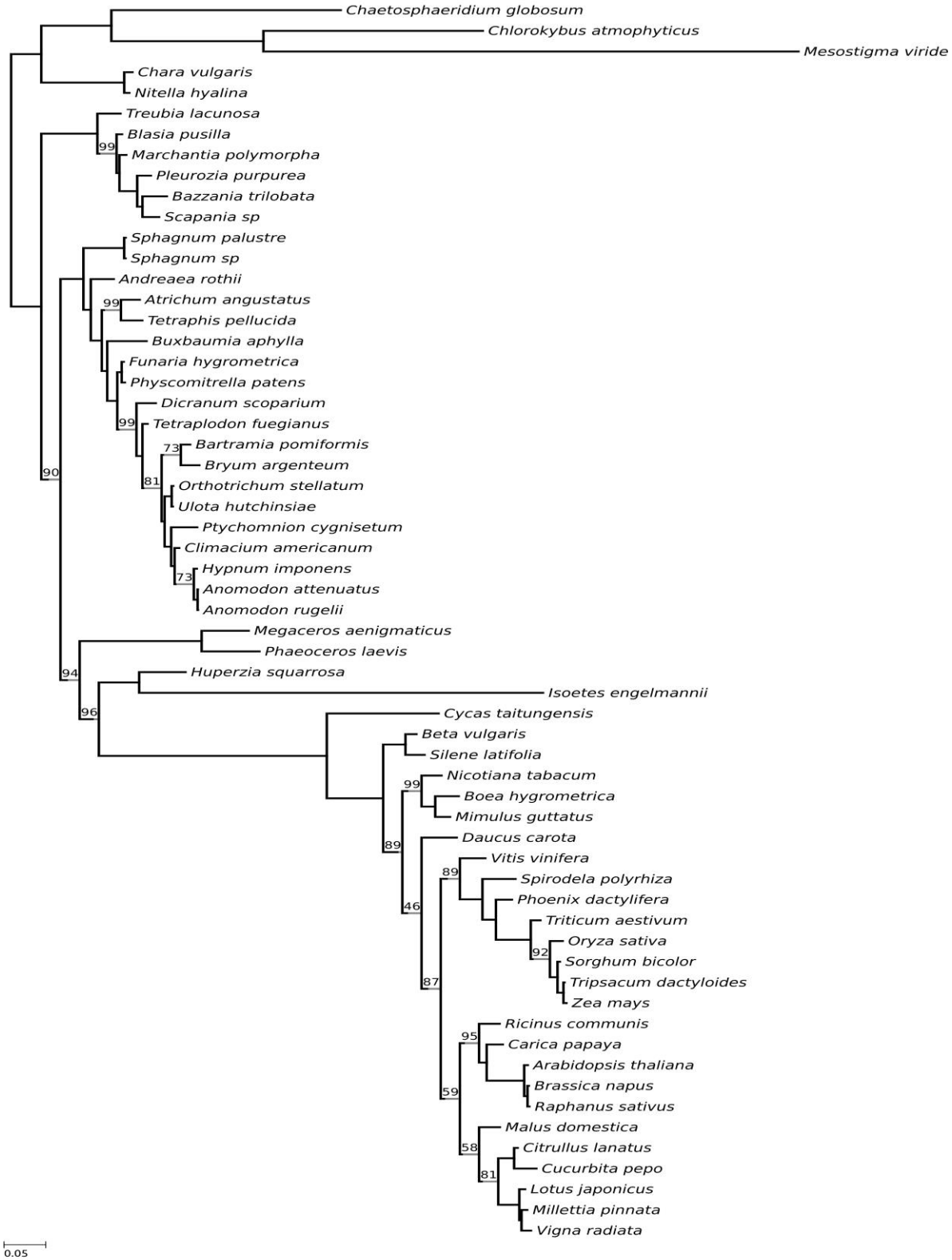


Figure A3 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -160501. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

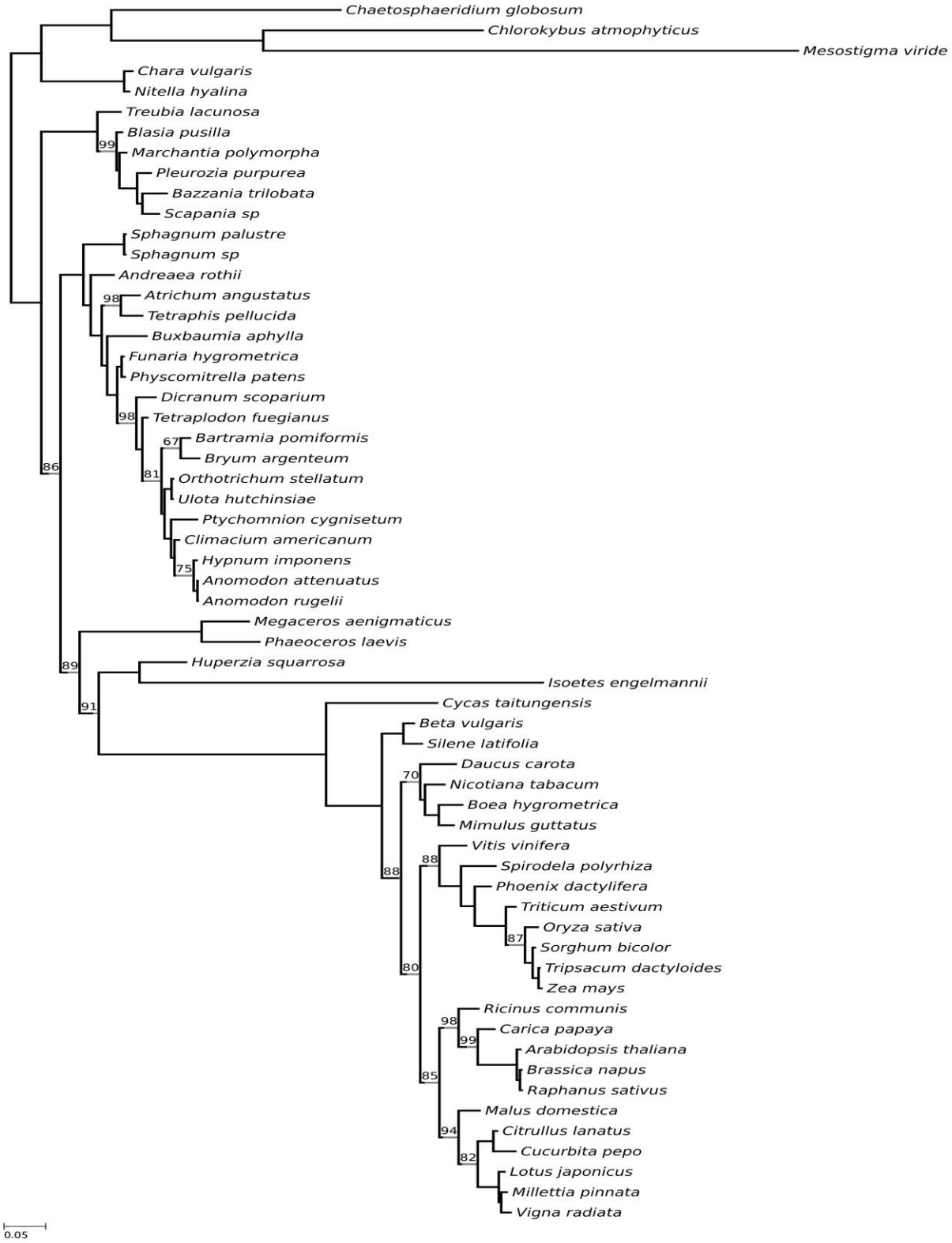


Figure A4 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -159945. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

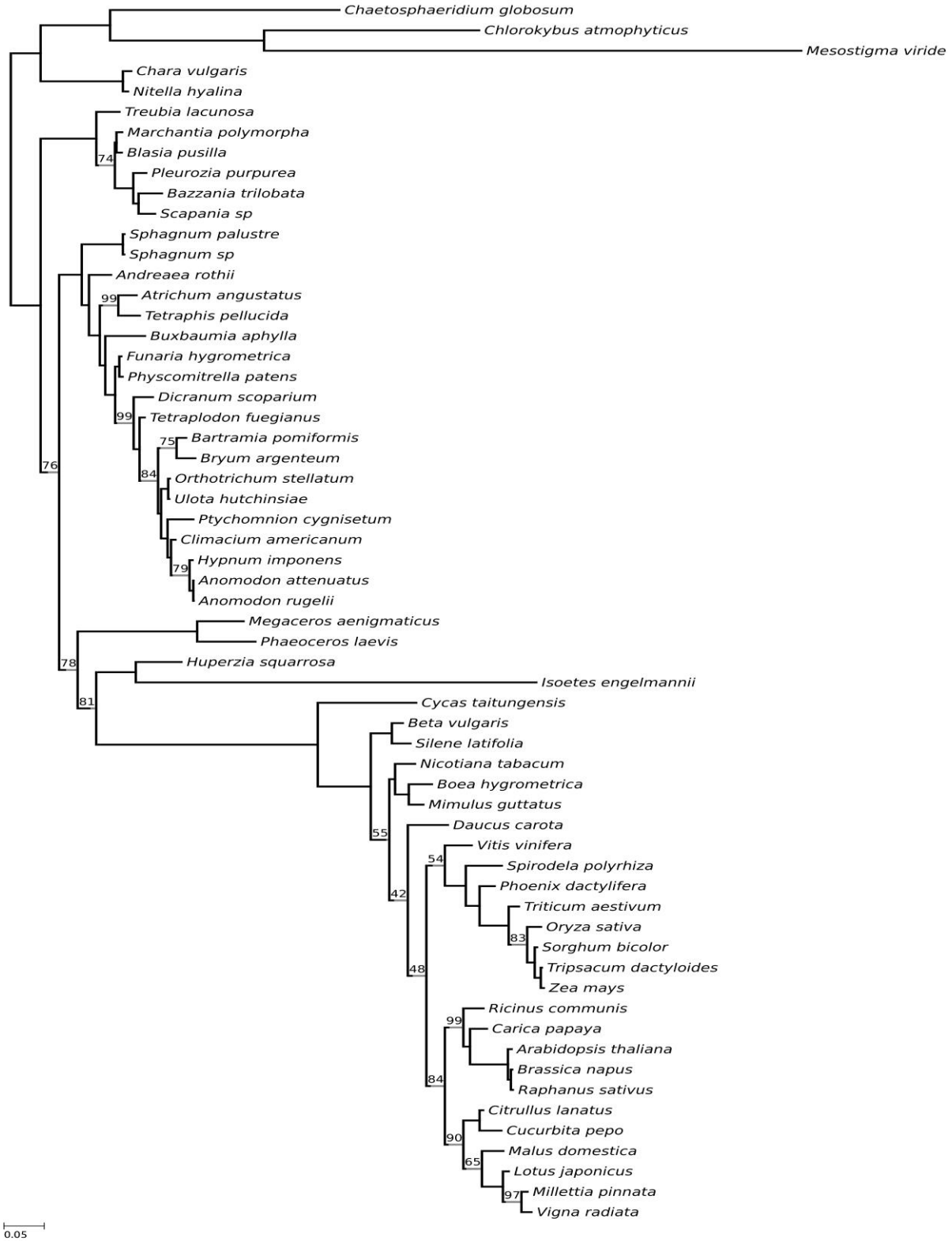


Figure A5 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -153585. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

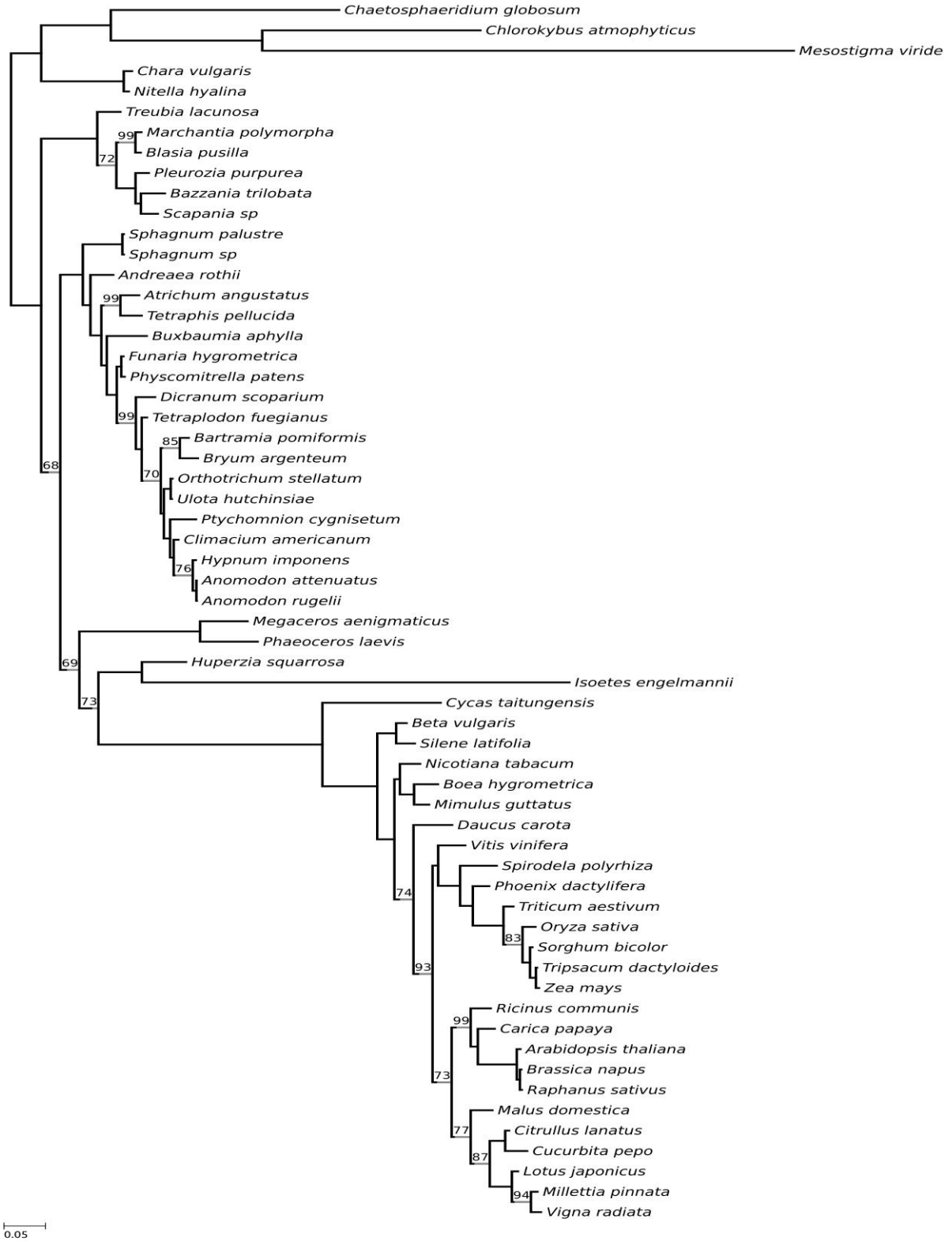


Figure A6 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -156476. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

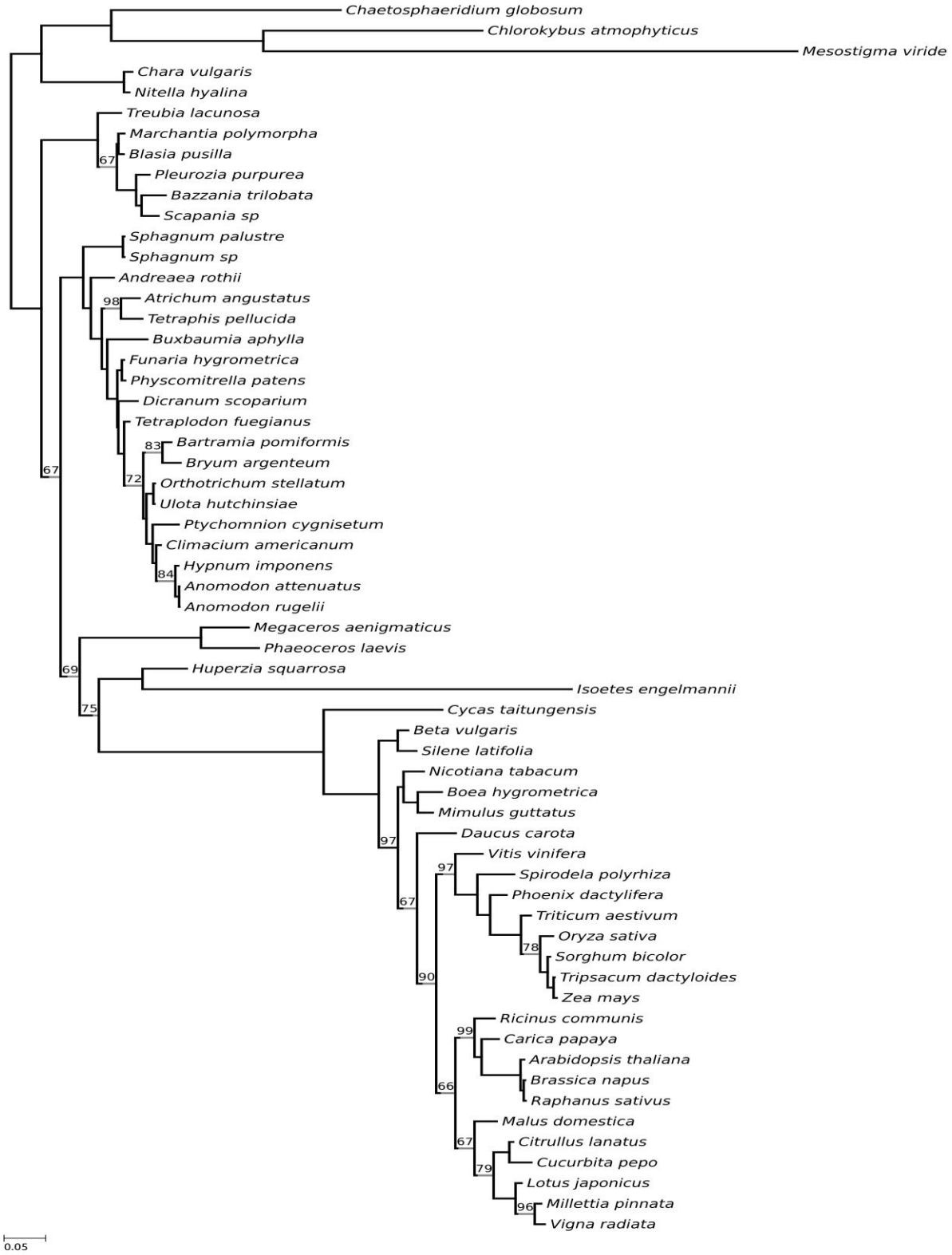


Figure A7 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -156276. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

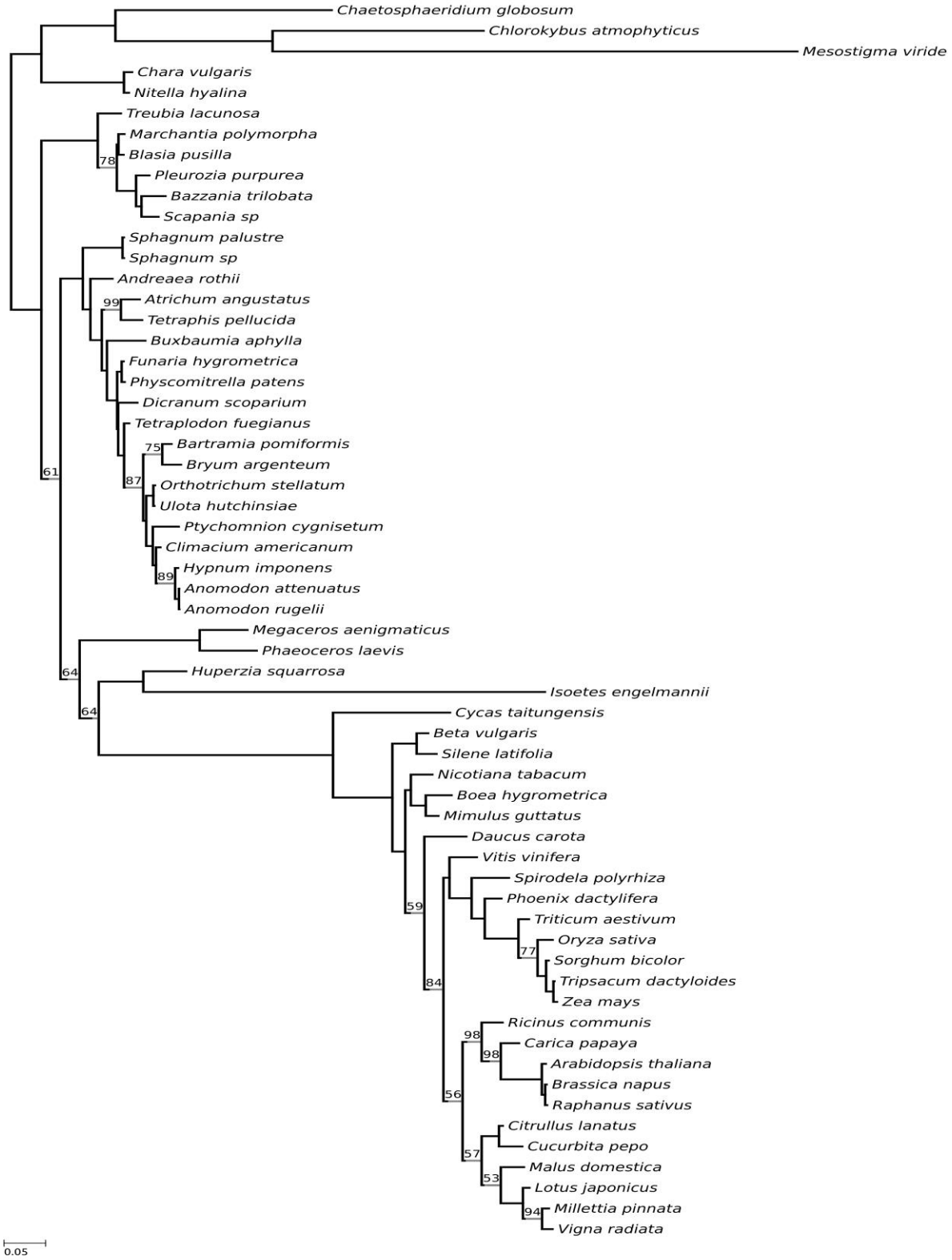


Figure A8 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -149562. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

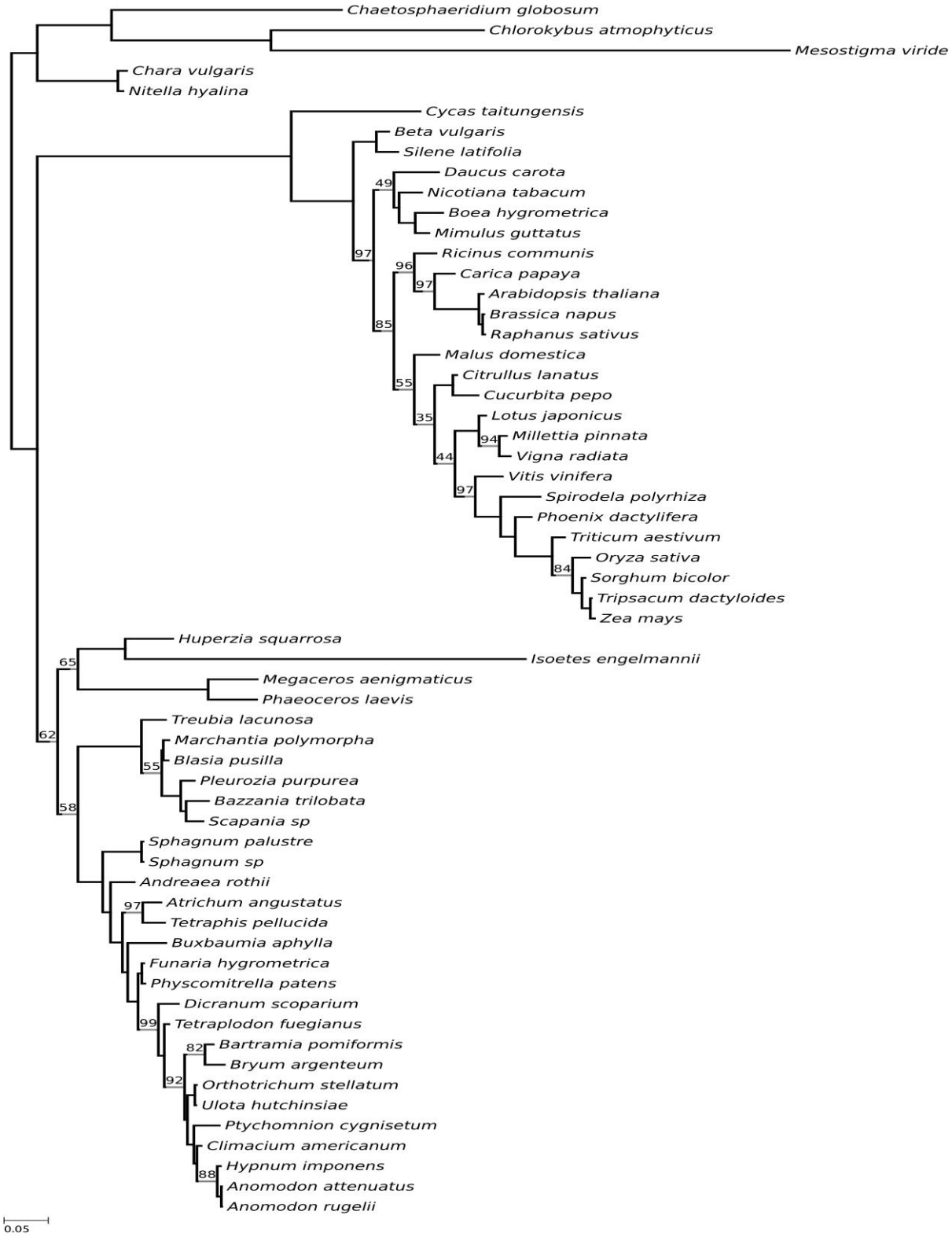


Figure A9 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -141360. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

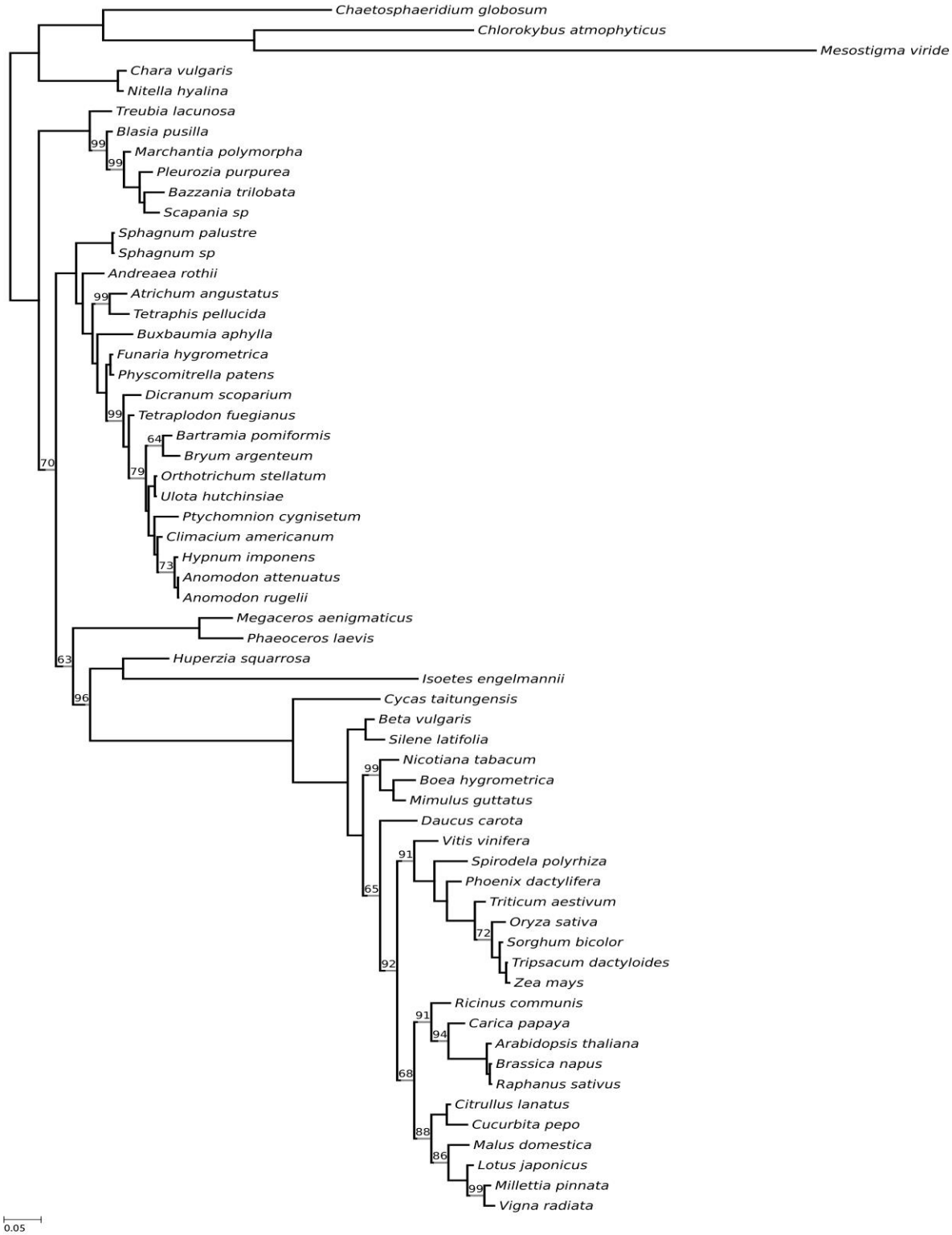


Figure A10 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al.*, 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -142691. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

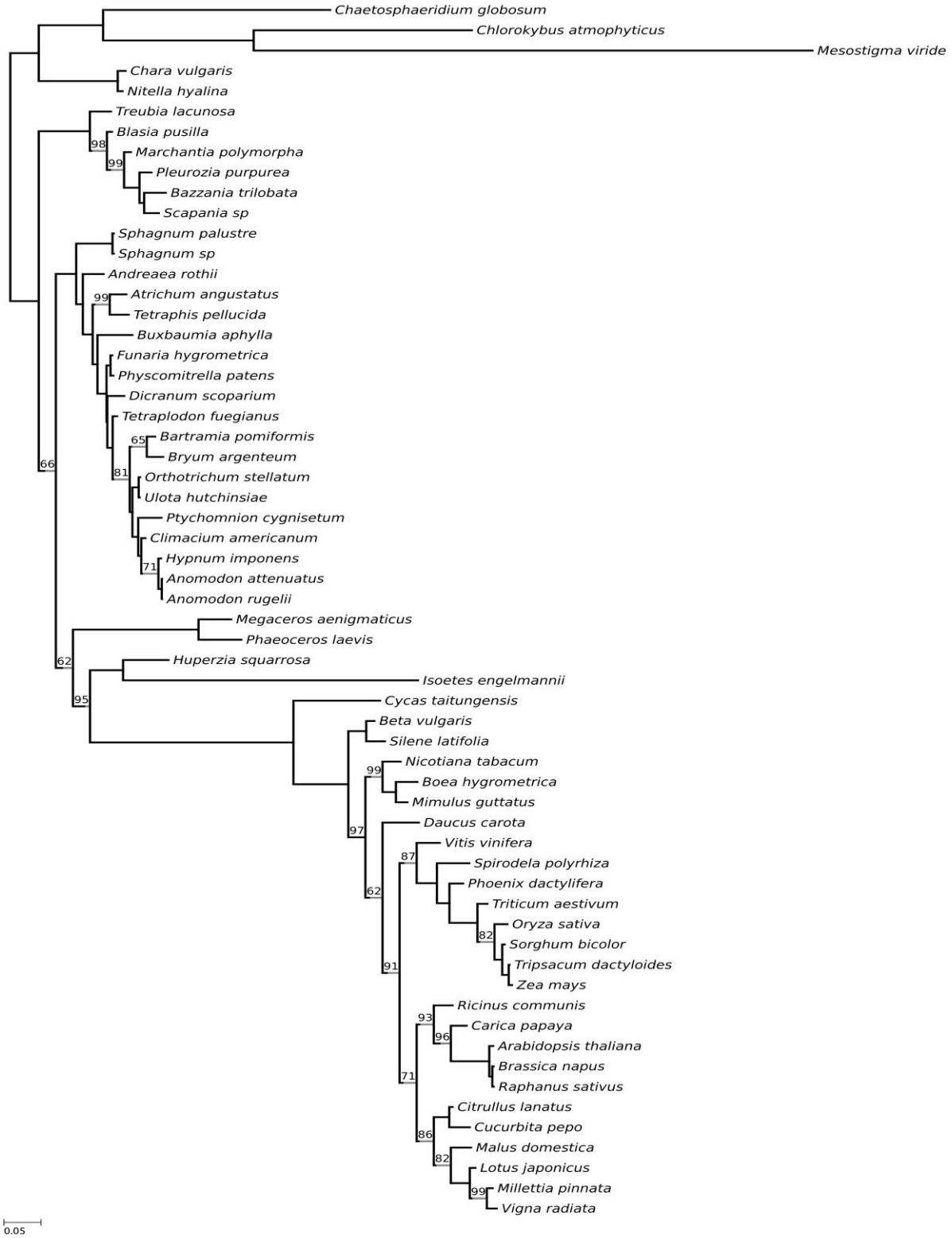


Figure A11 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -142938. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

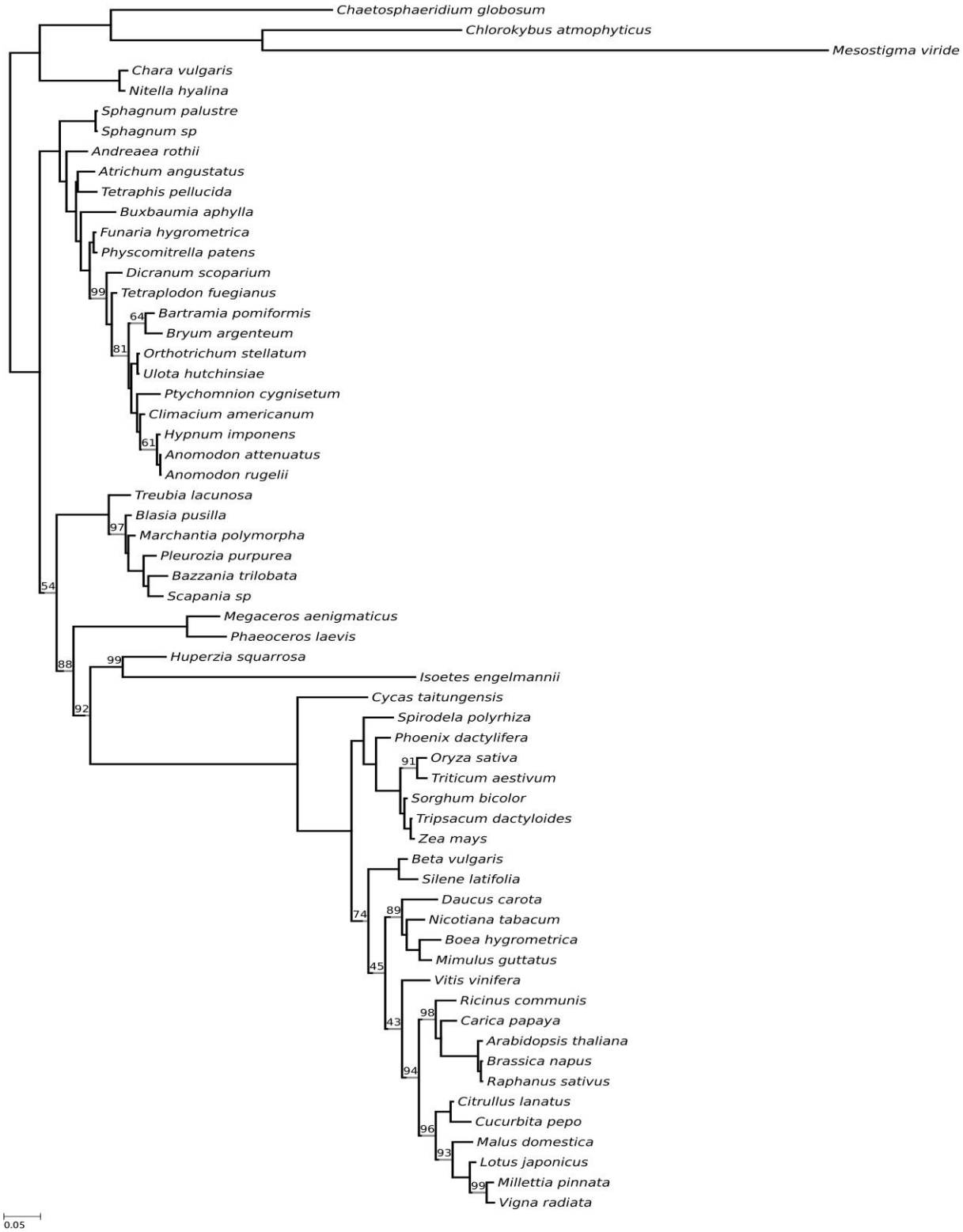


Figure A12 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -134968. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

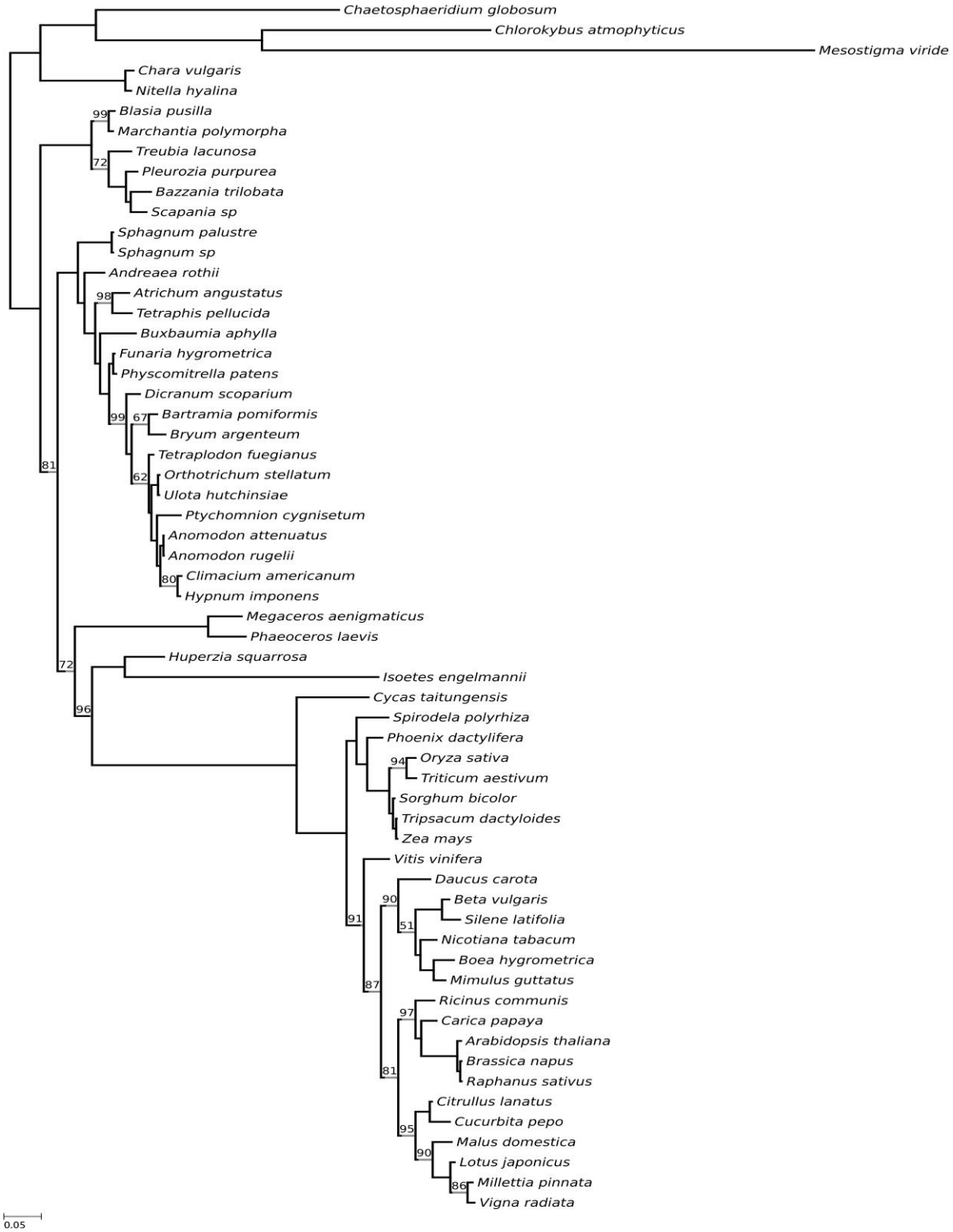


Figure A13 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -121719. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

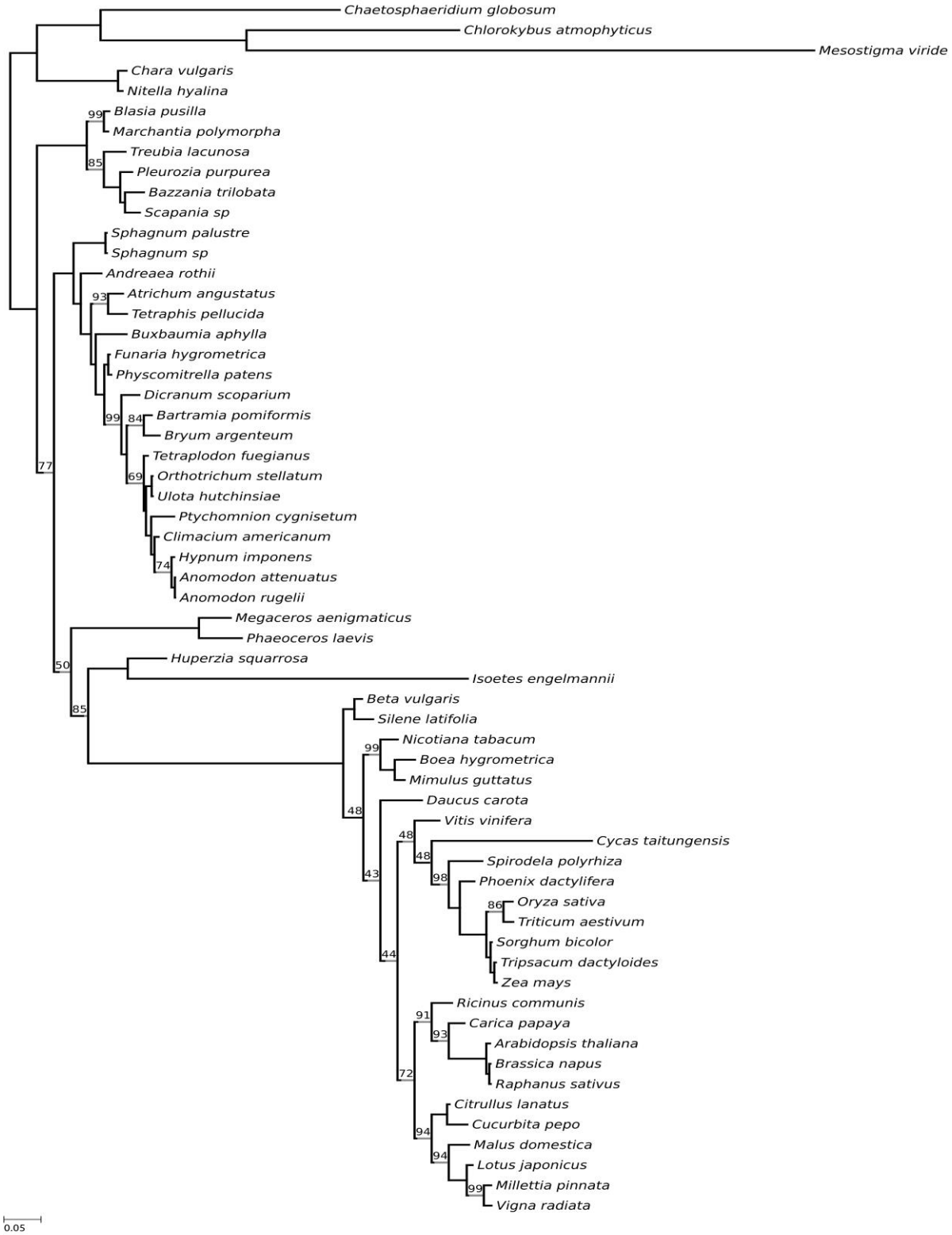


Figure A14 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -125720. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

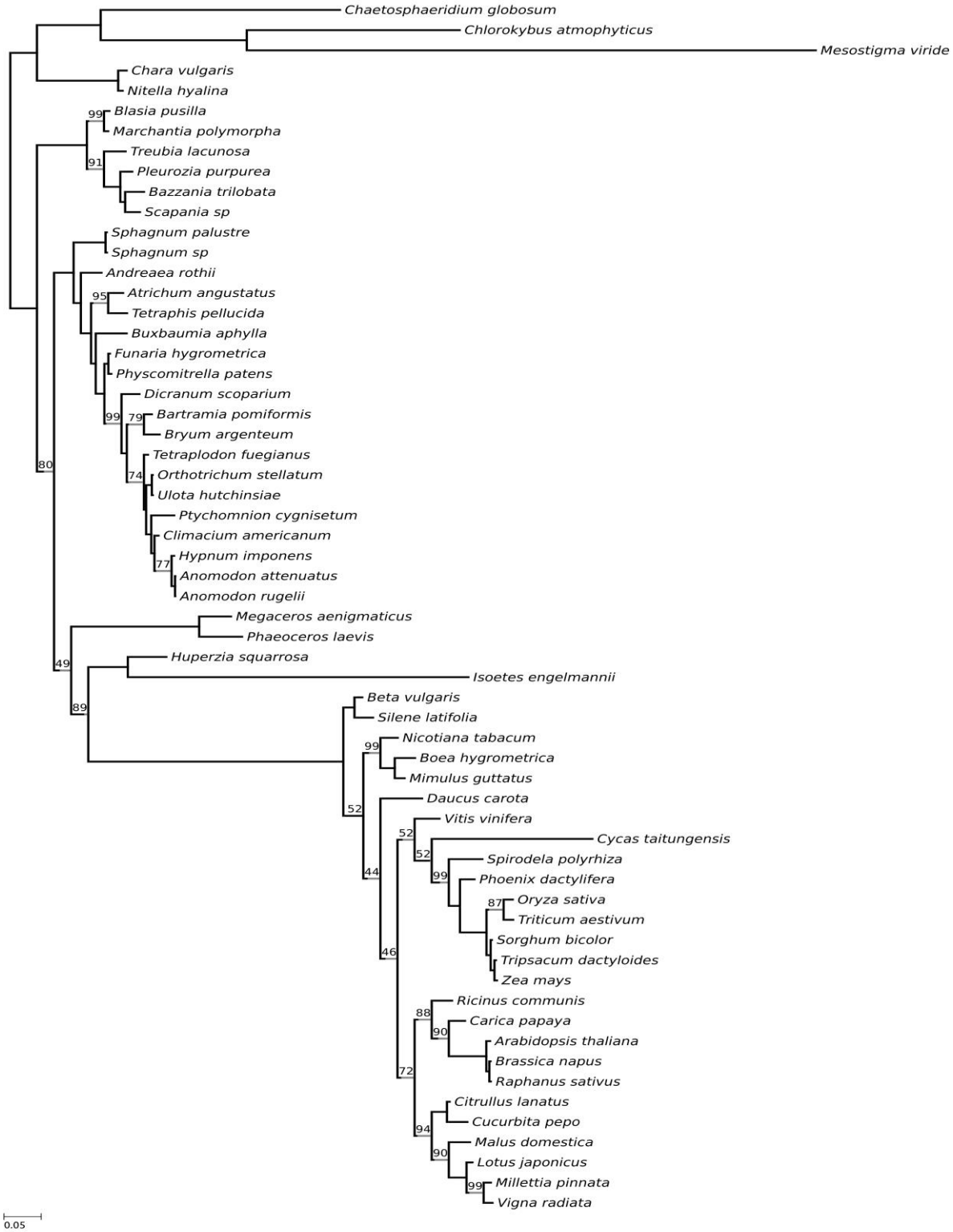


Figure A15 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al.*, 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -125501. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

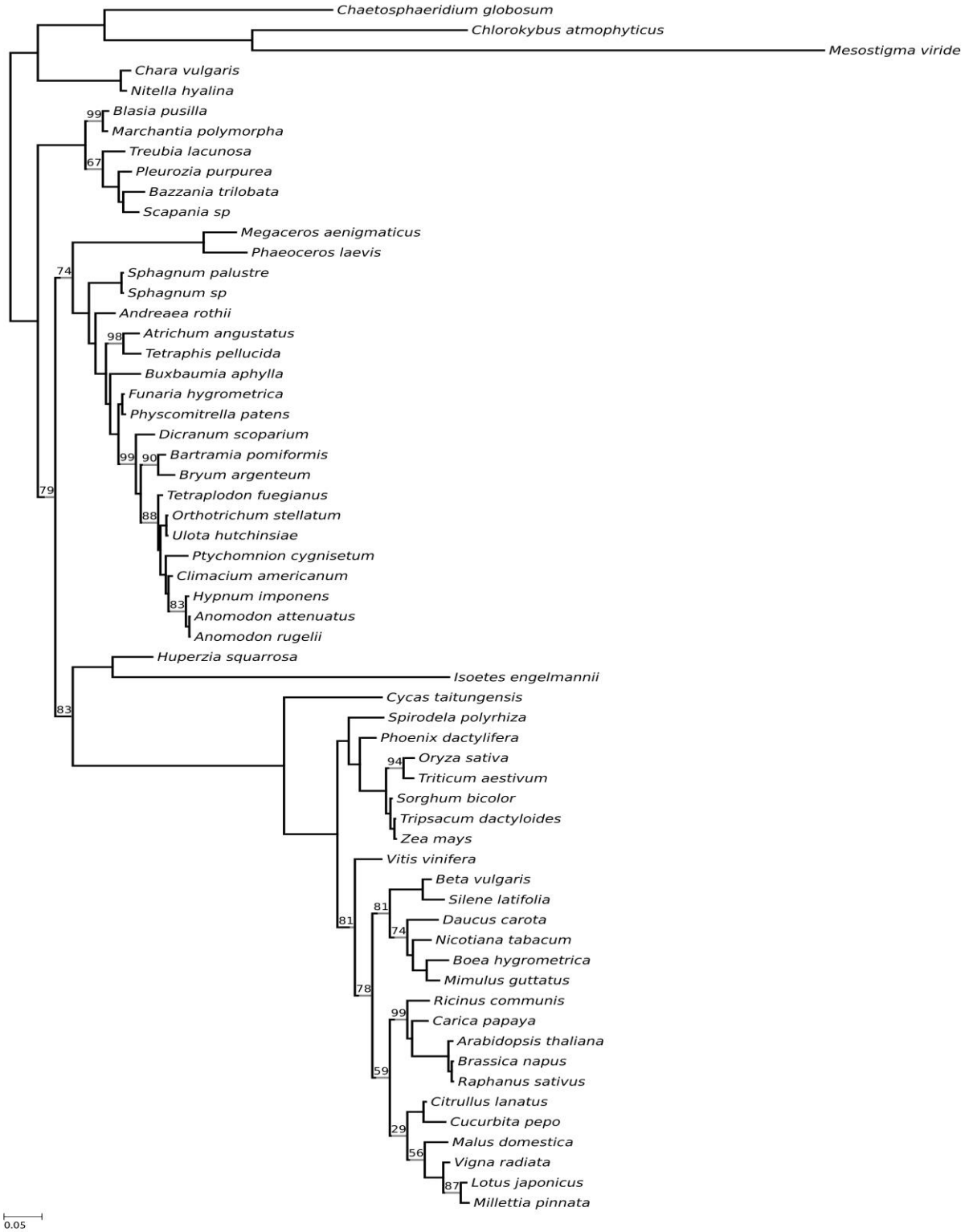


Figure A16 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -117431. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

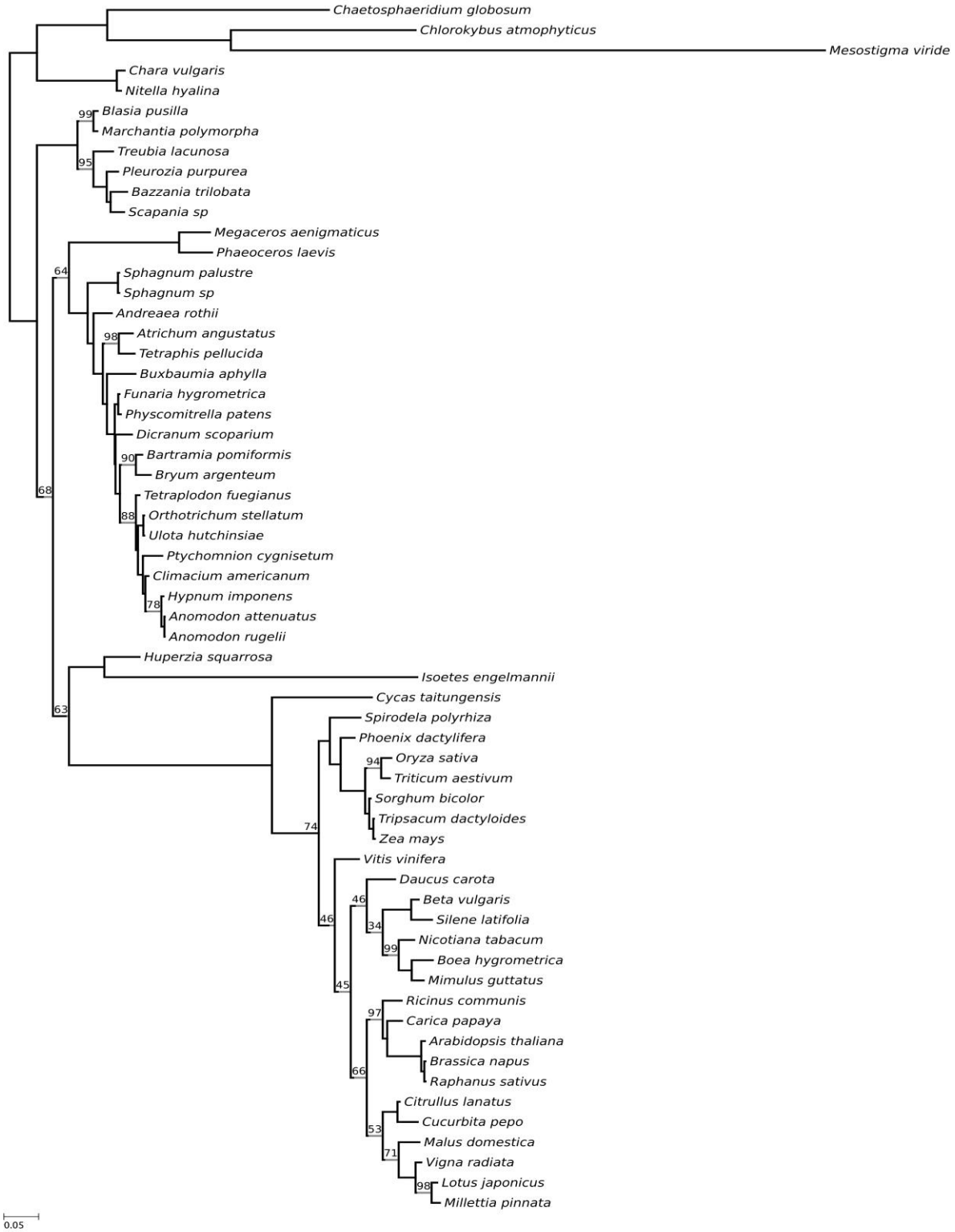


Figure A17 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al.*, 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -106612. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

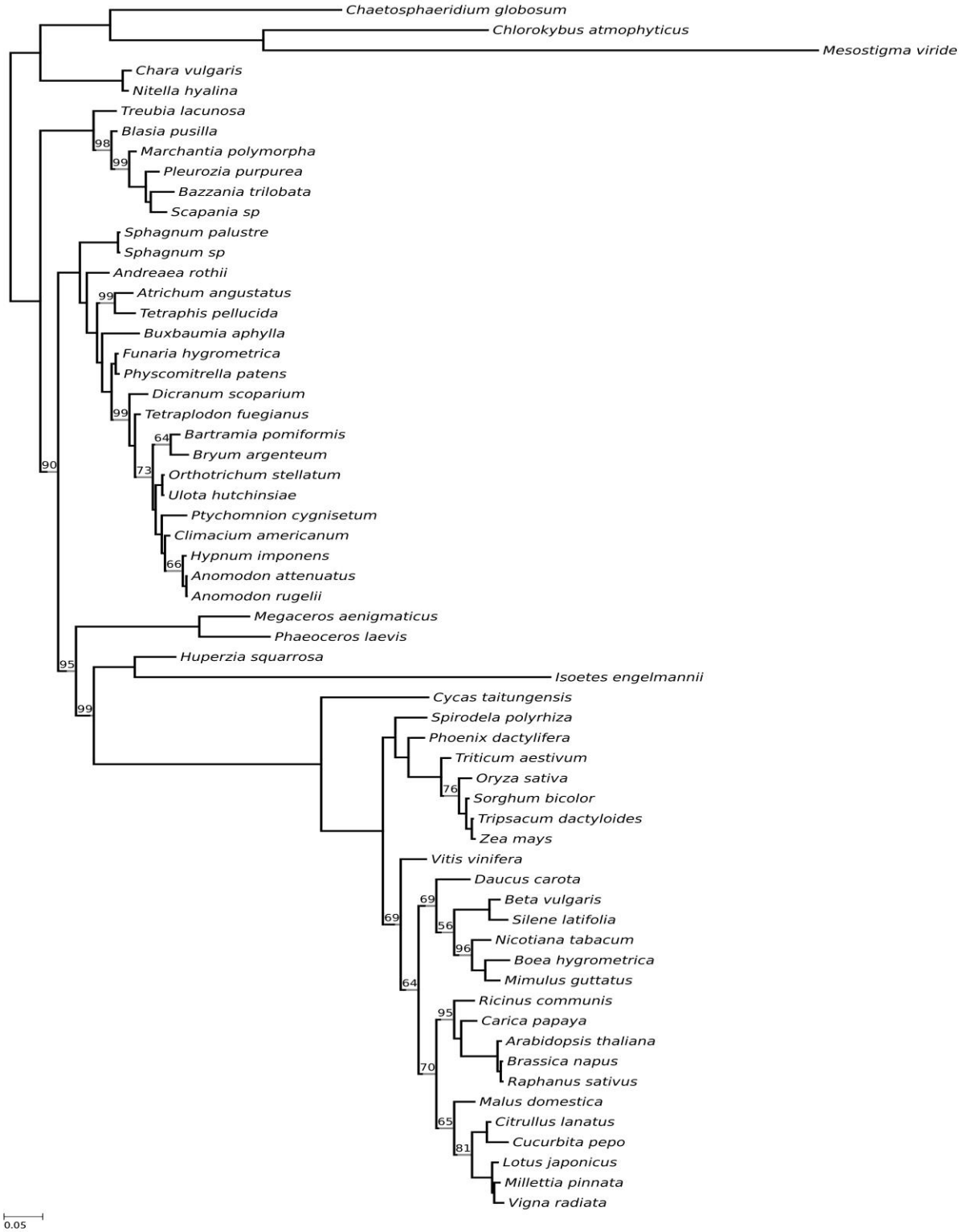


Figure A18 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 19 data partitions (derived from the Liu *et al.*, 2014 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of internal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -128630. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A19 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -128001. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

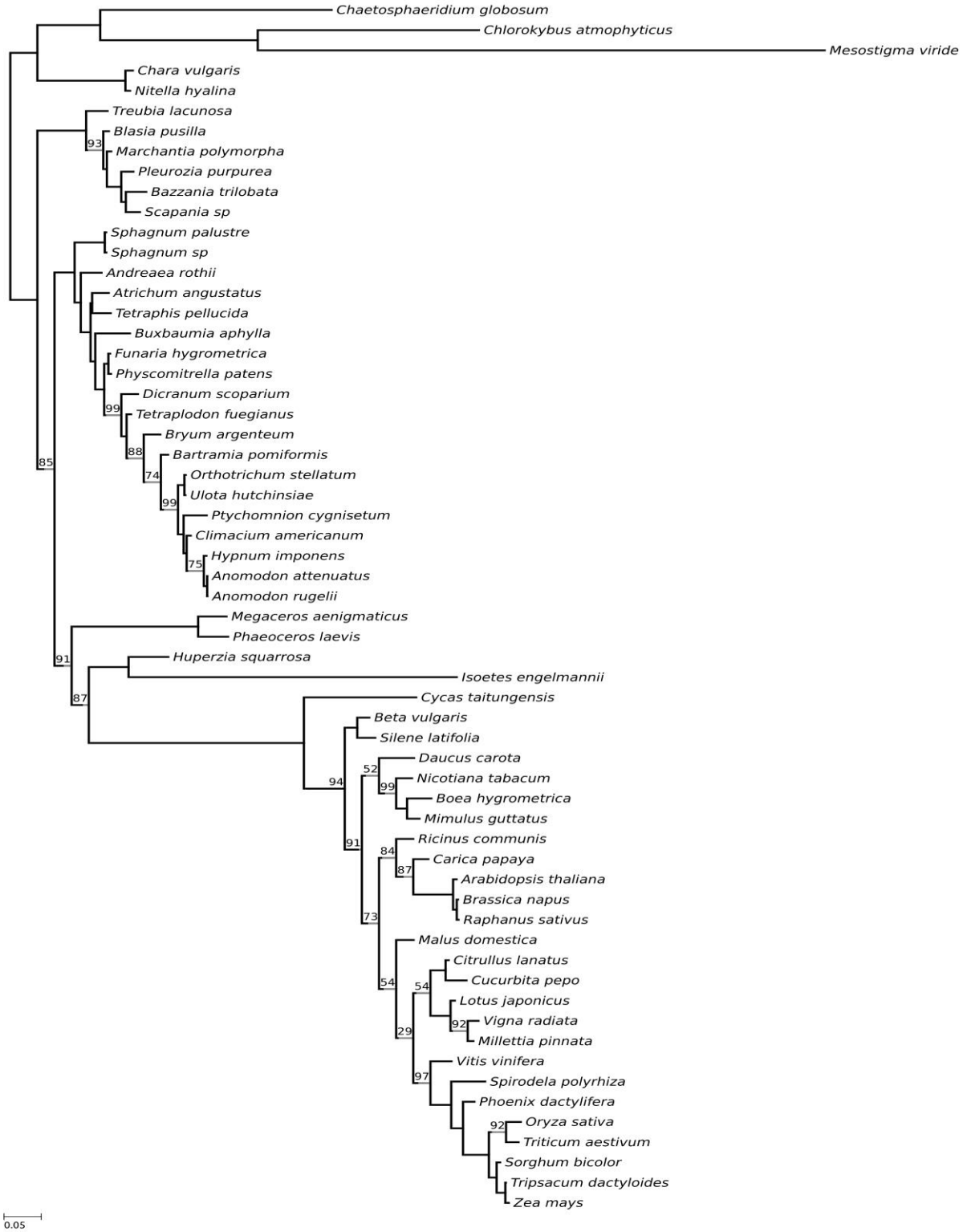


Figure A20 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -112777. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

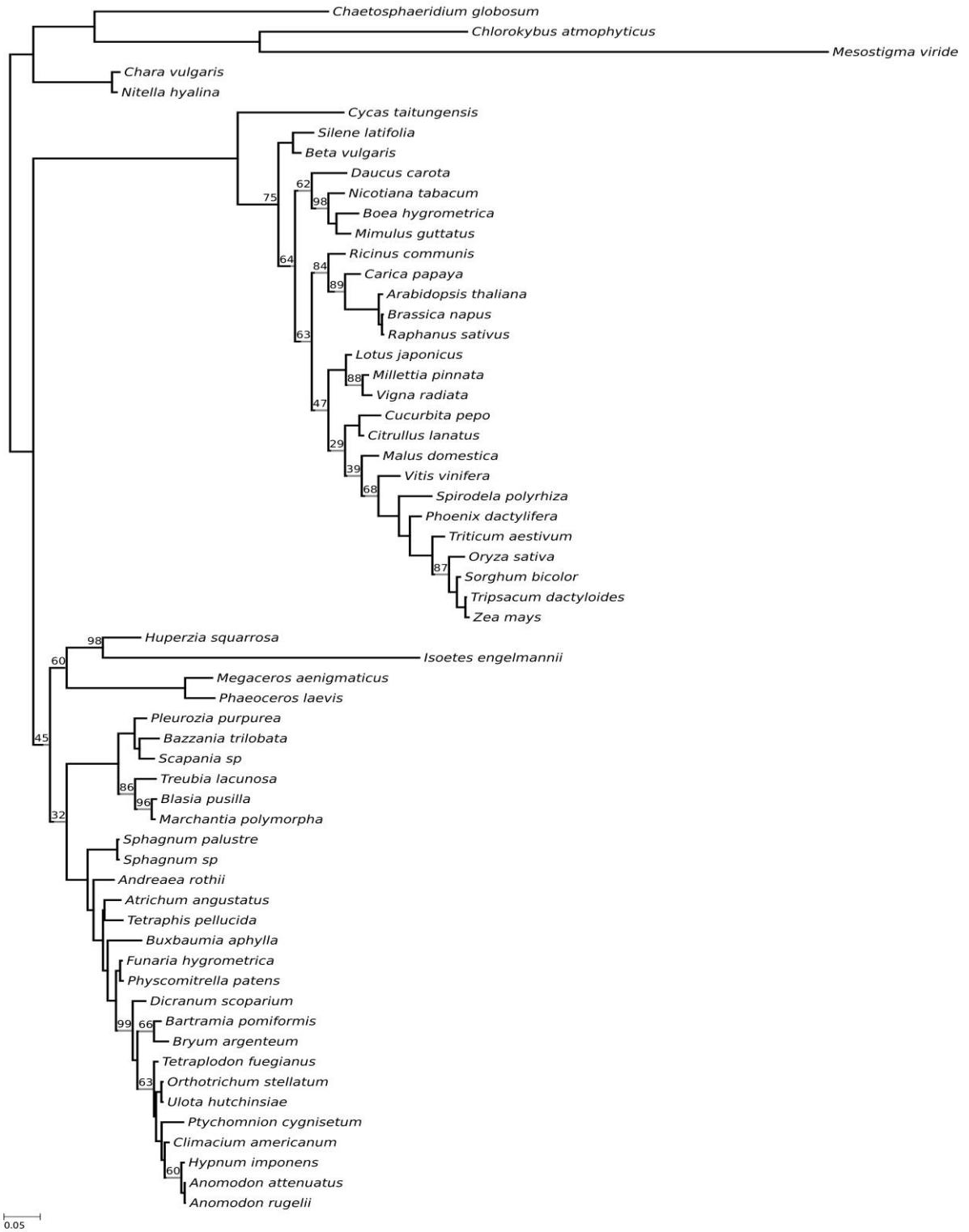


Figure A21 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -95825. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

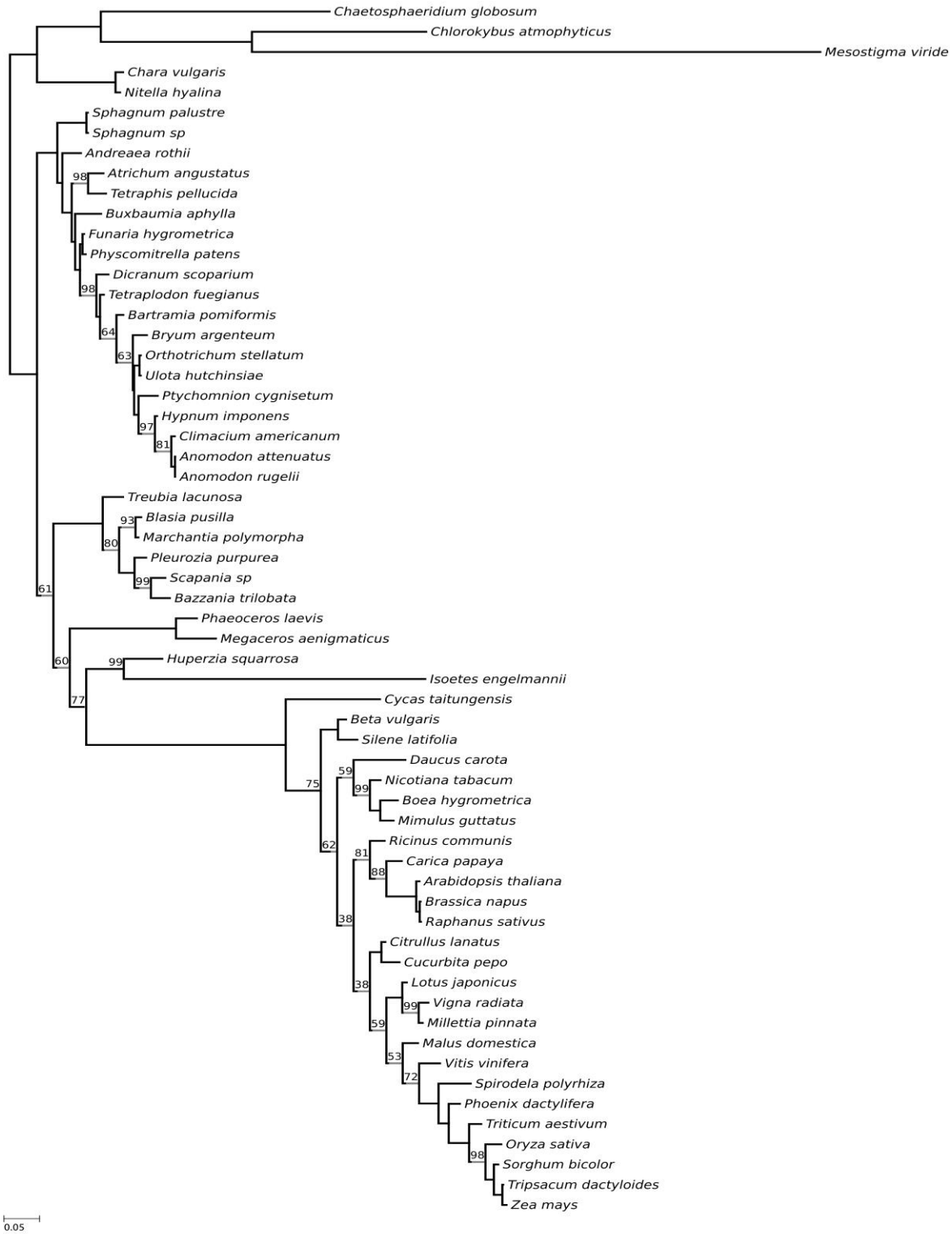


Figure A22 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -108389. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

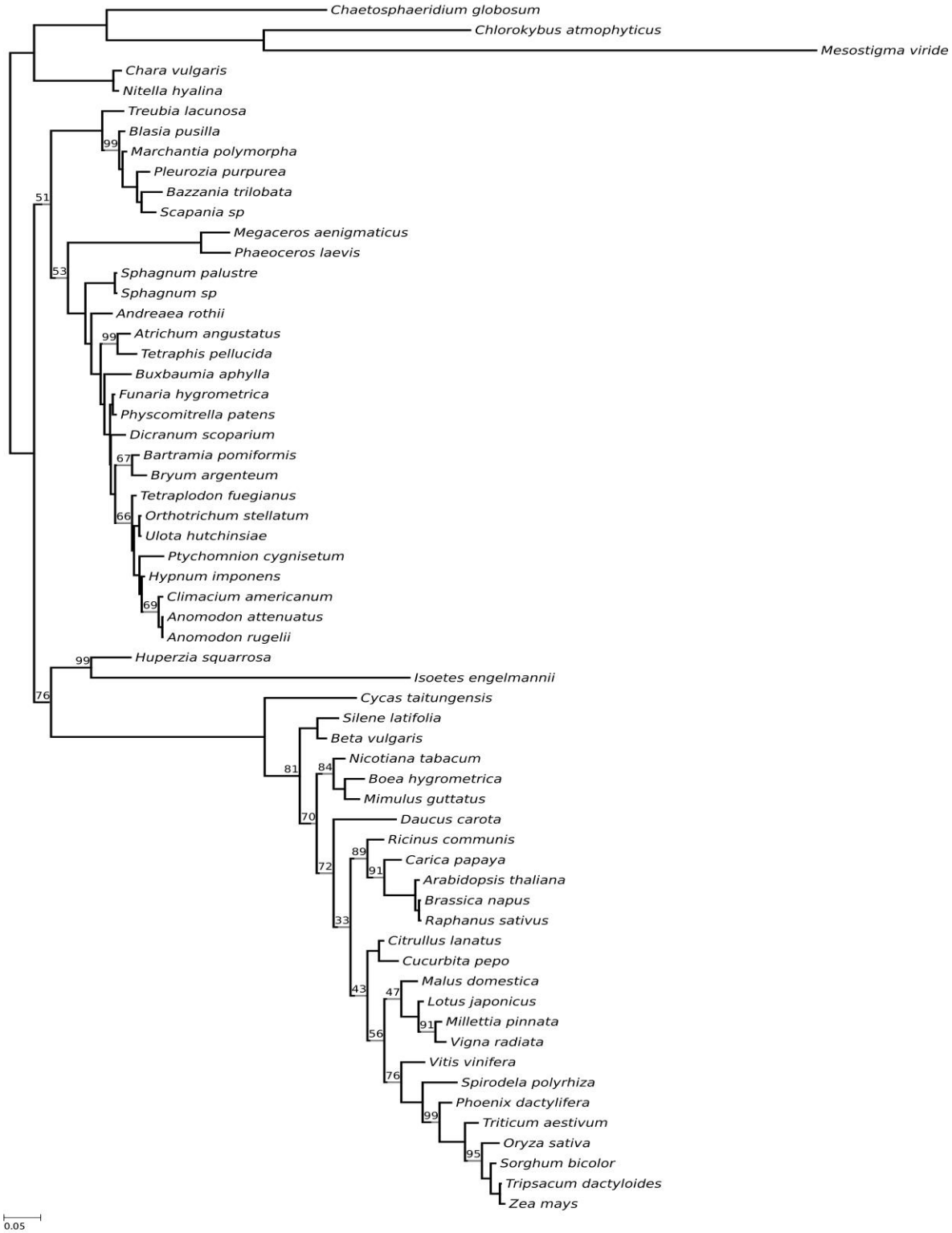


Figure A23 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -105176. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

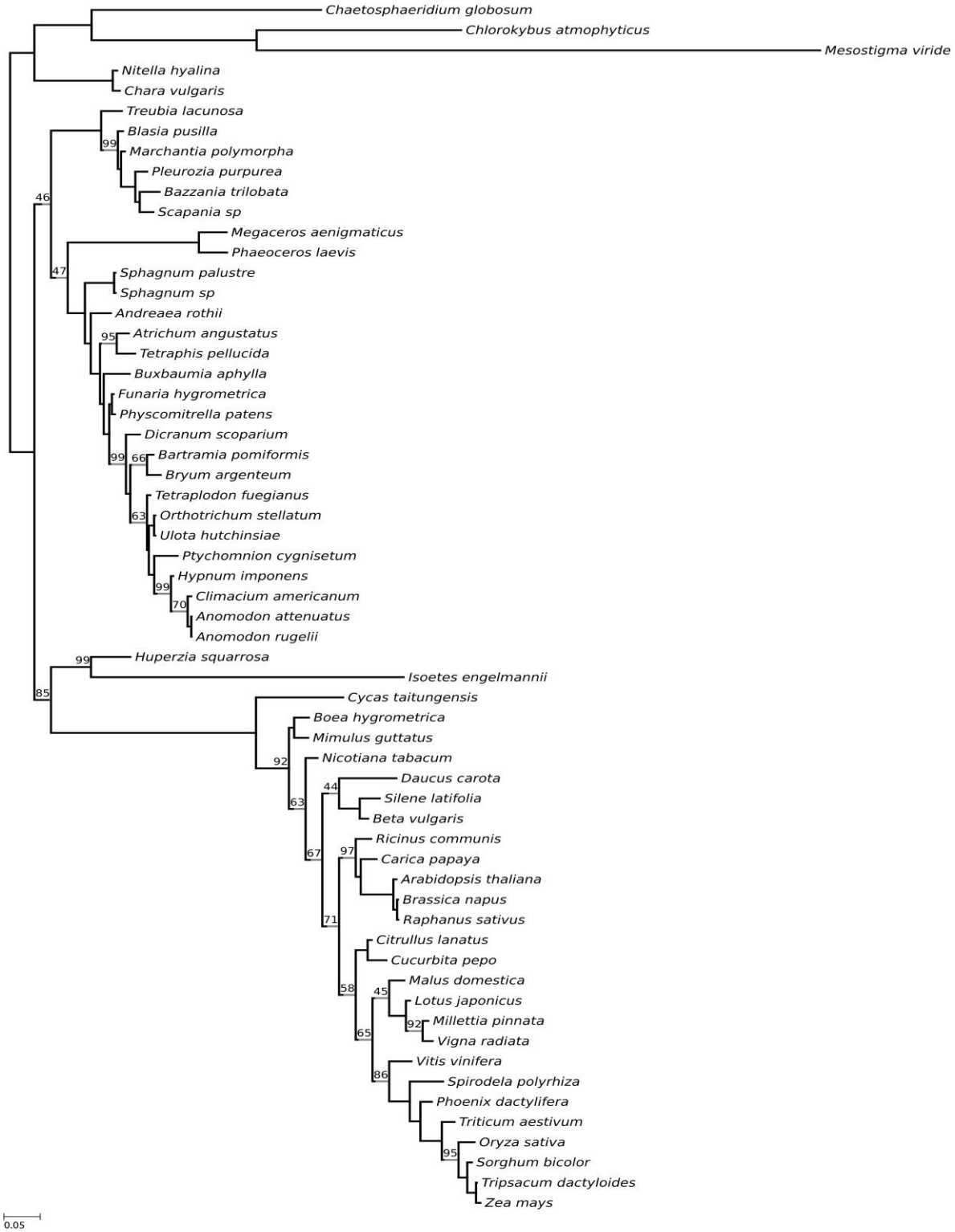


Figure A24 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -86206. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

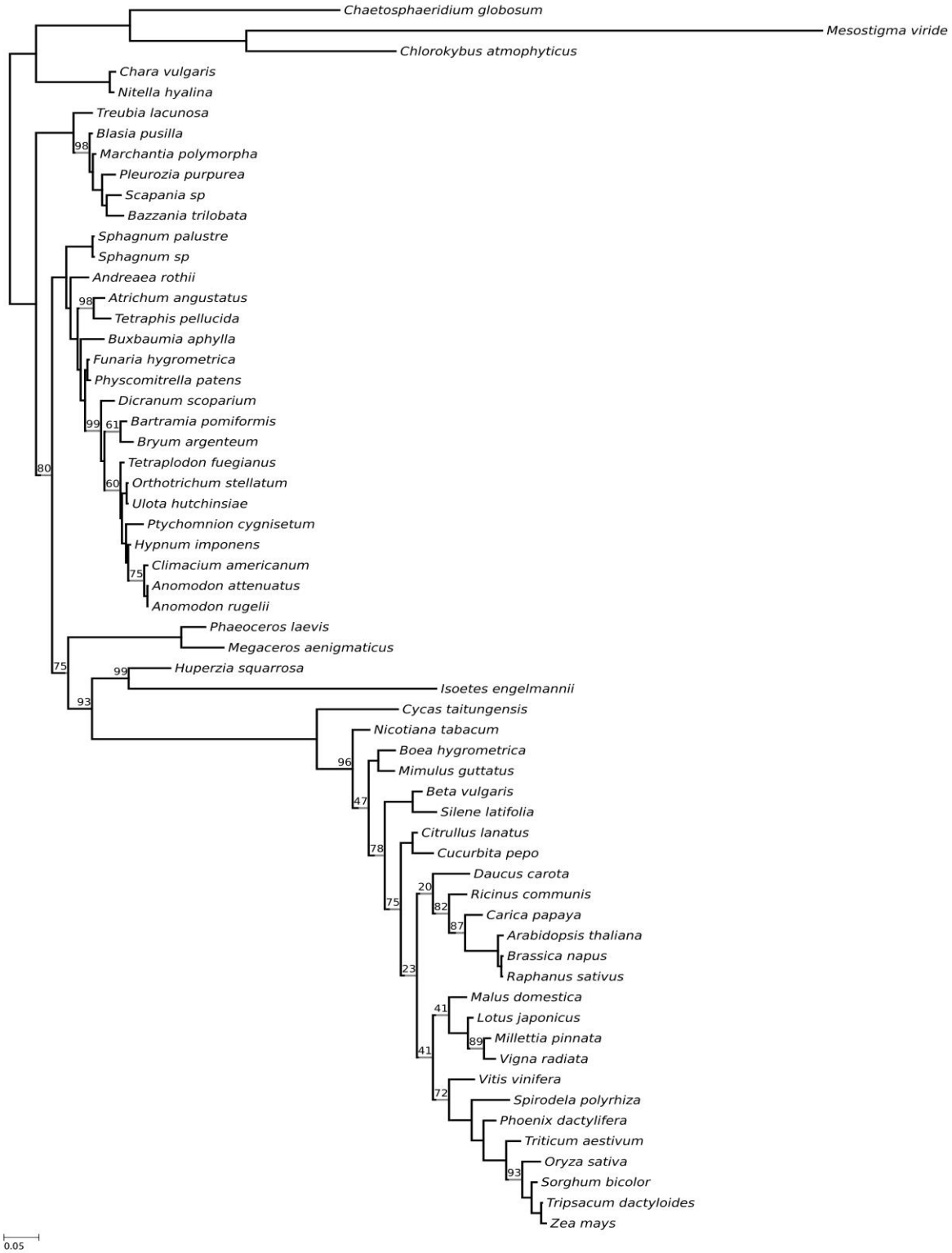


Figure A25 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al*., 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -77508. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

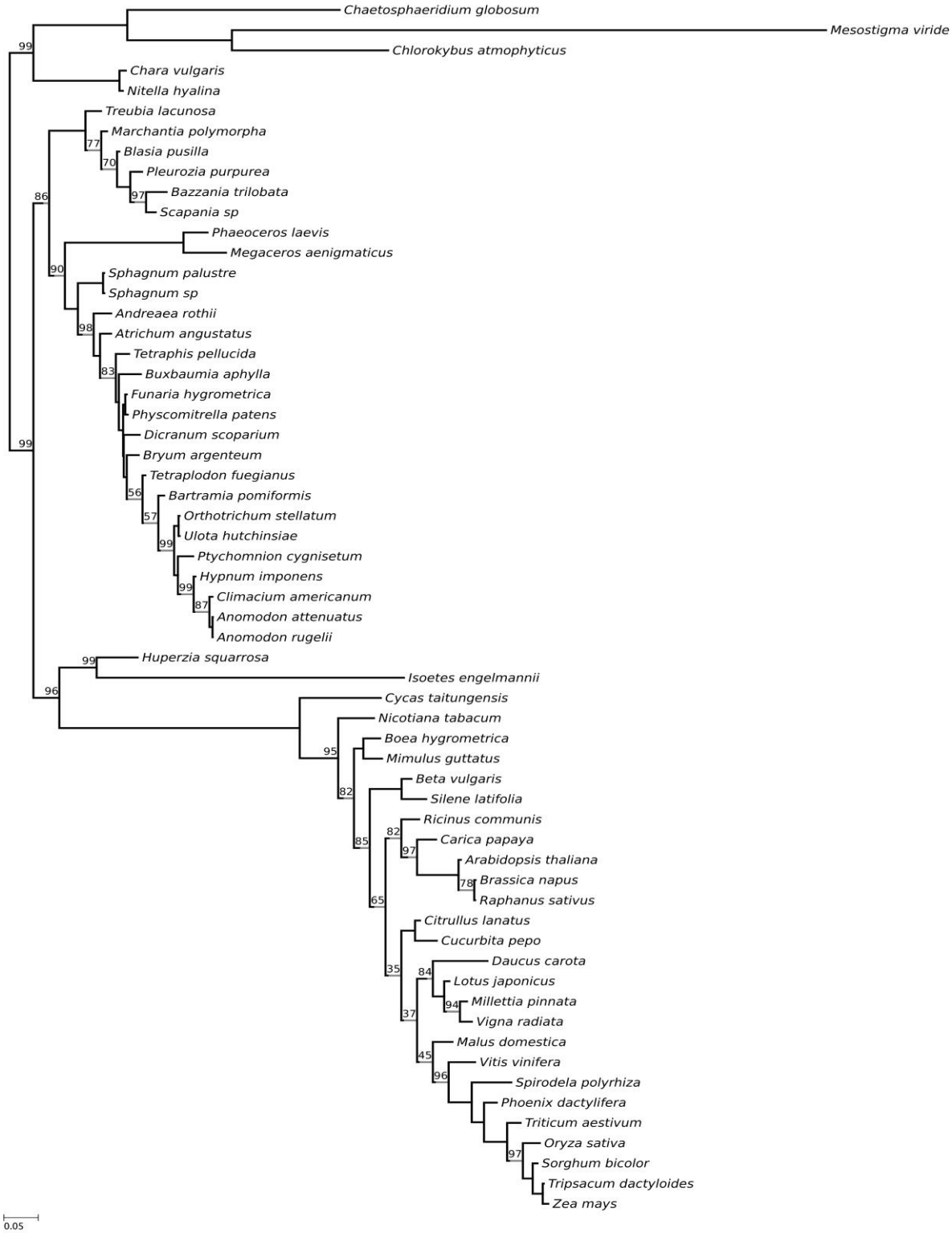


Figure A26 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from nine data partitions (derived from the Liu *et al.*, 2014 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -163665. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

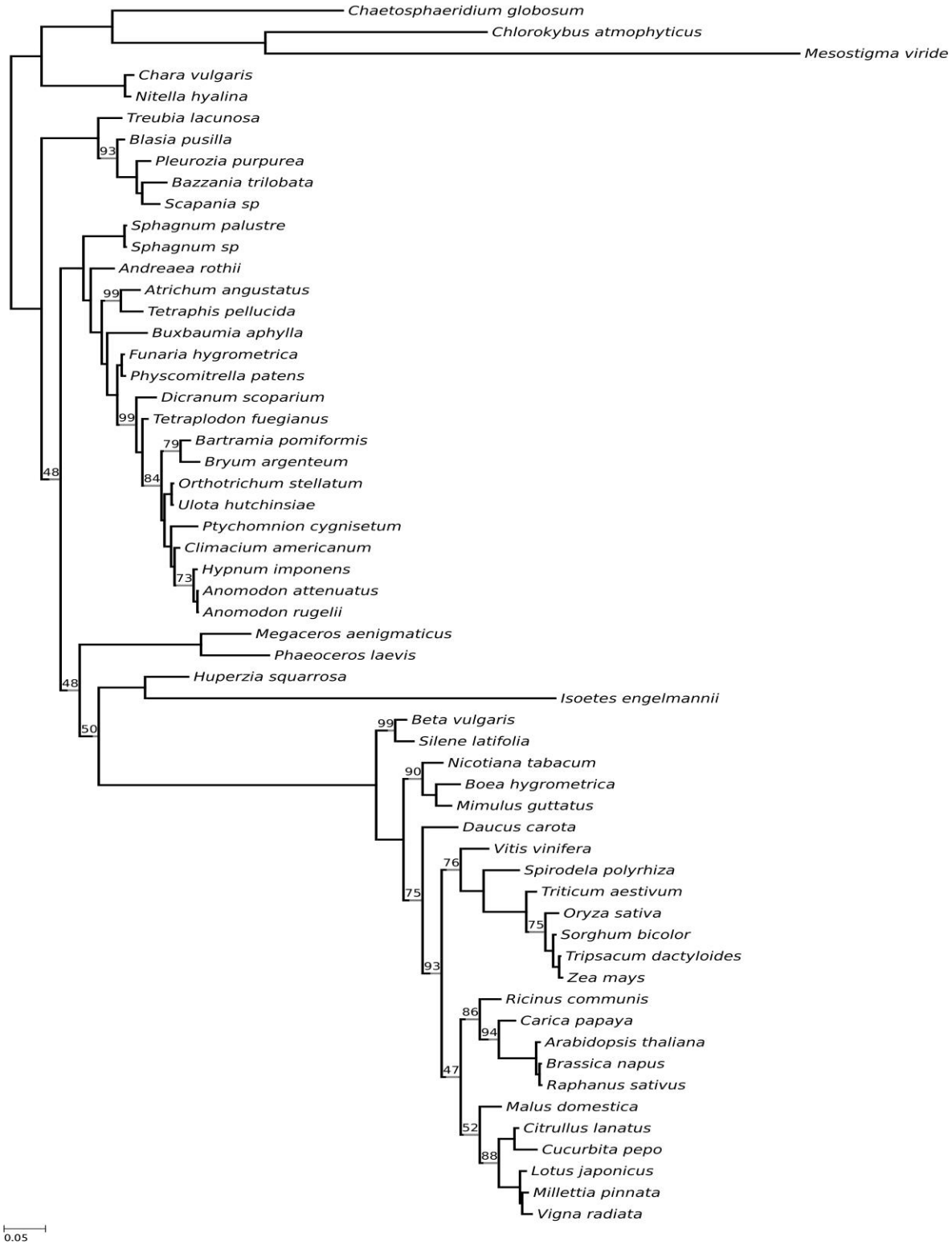


Figure A27 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from the entire concatenated data set according to the matched-pairs test of internal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -155803. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

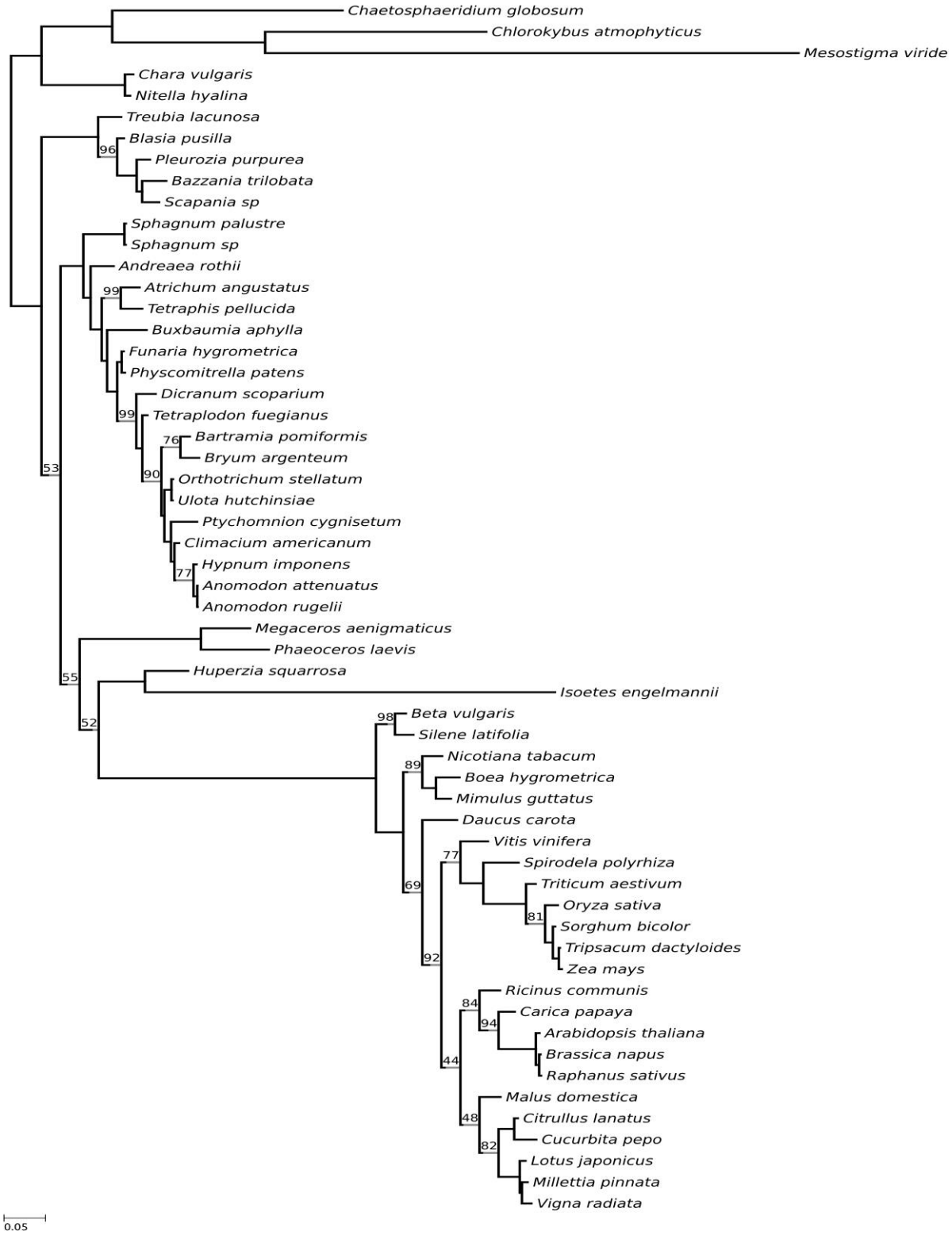


Figure A28 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set). Tree-heterogeneous sequences were filtered out from the entire concatenated data set according to the matched-pairs test of internal symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -155802. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

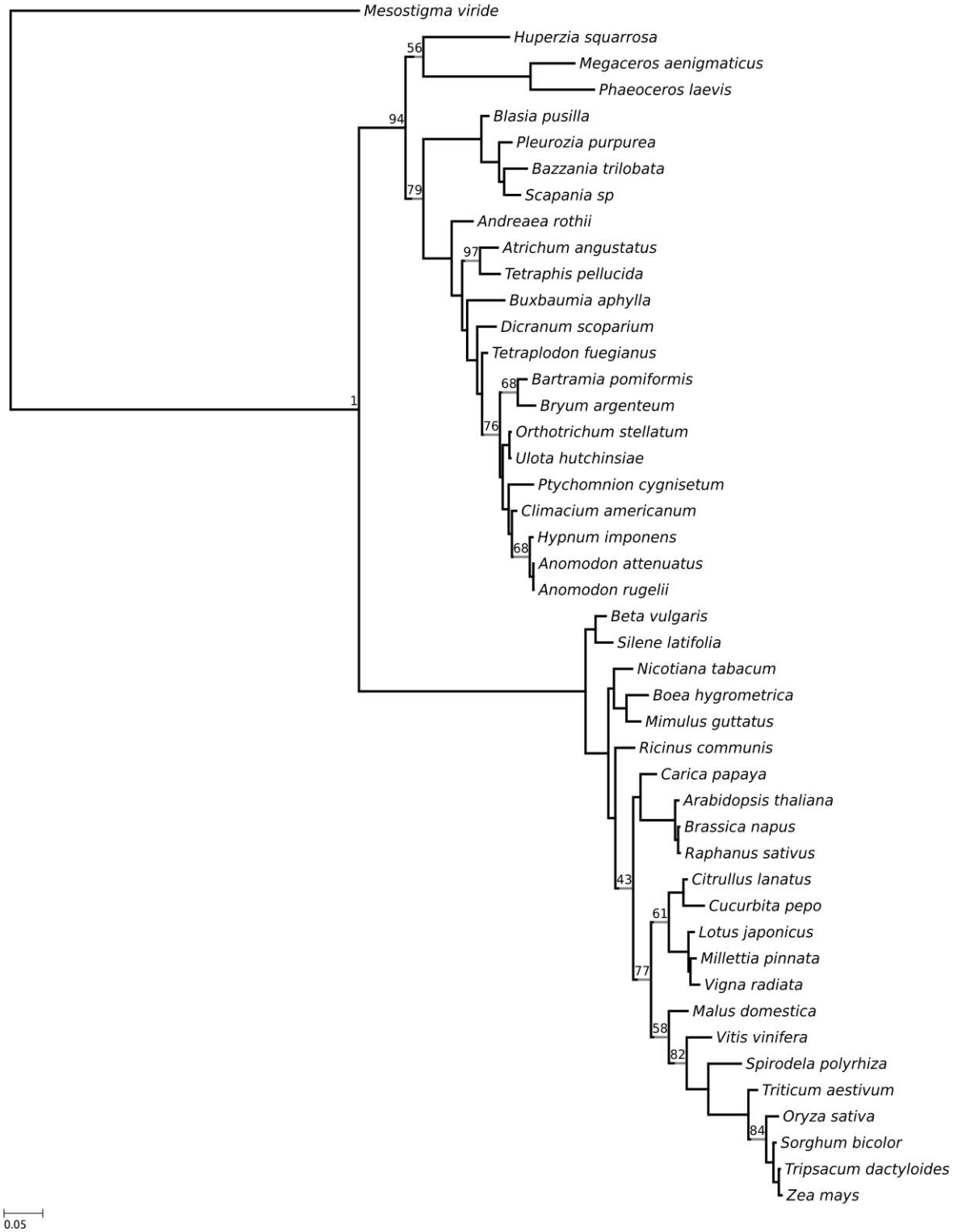


Figure A29 - Optimal maximum-likelihood tree comprising 60 taxa reconstructed from 41 concatenated proteins (derived from the Liu *et al*., 2014 data set Tree-heterogeneous sequences were filtered out from the entire concatenated data set according to the matched-pairs test of internal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -120802. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A30 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (Strassert *et al.*, 2021 data set). Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6352127. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A31 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al*., 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6341715. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

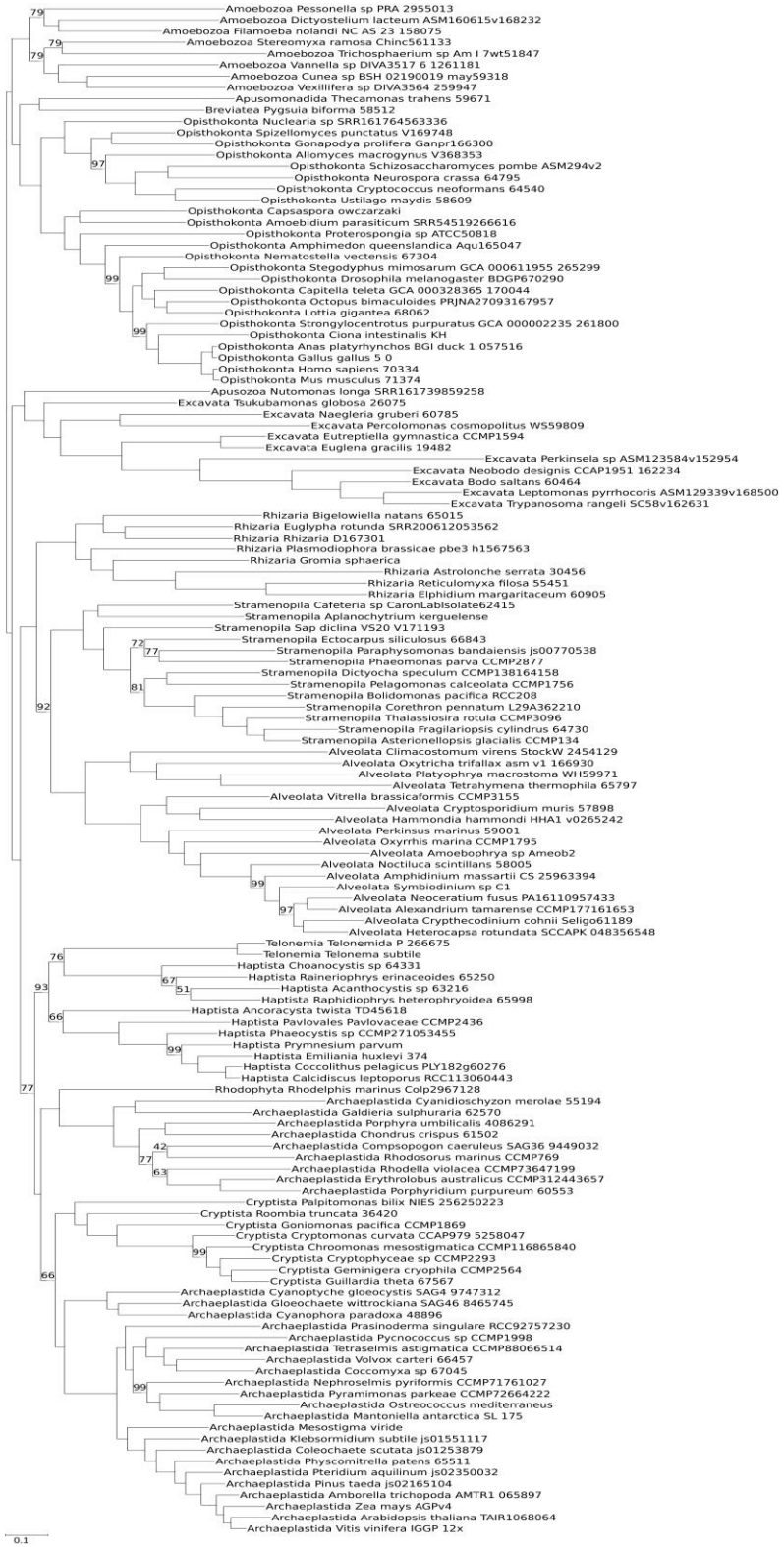


Figure A32 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al*., 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6333588. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A33 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al*., 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6290391. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

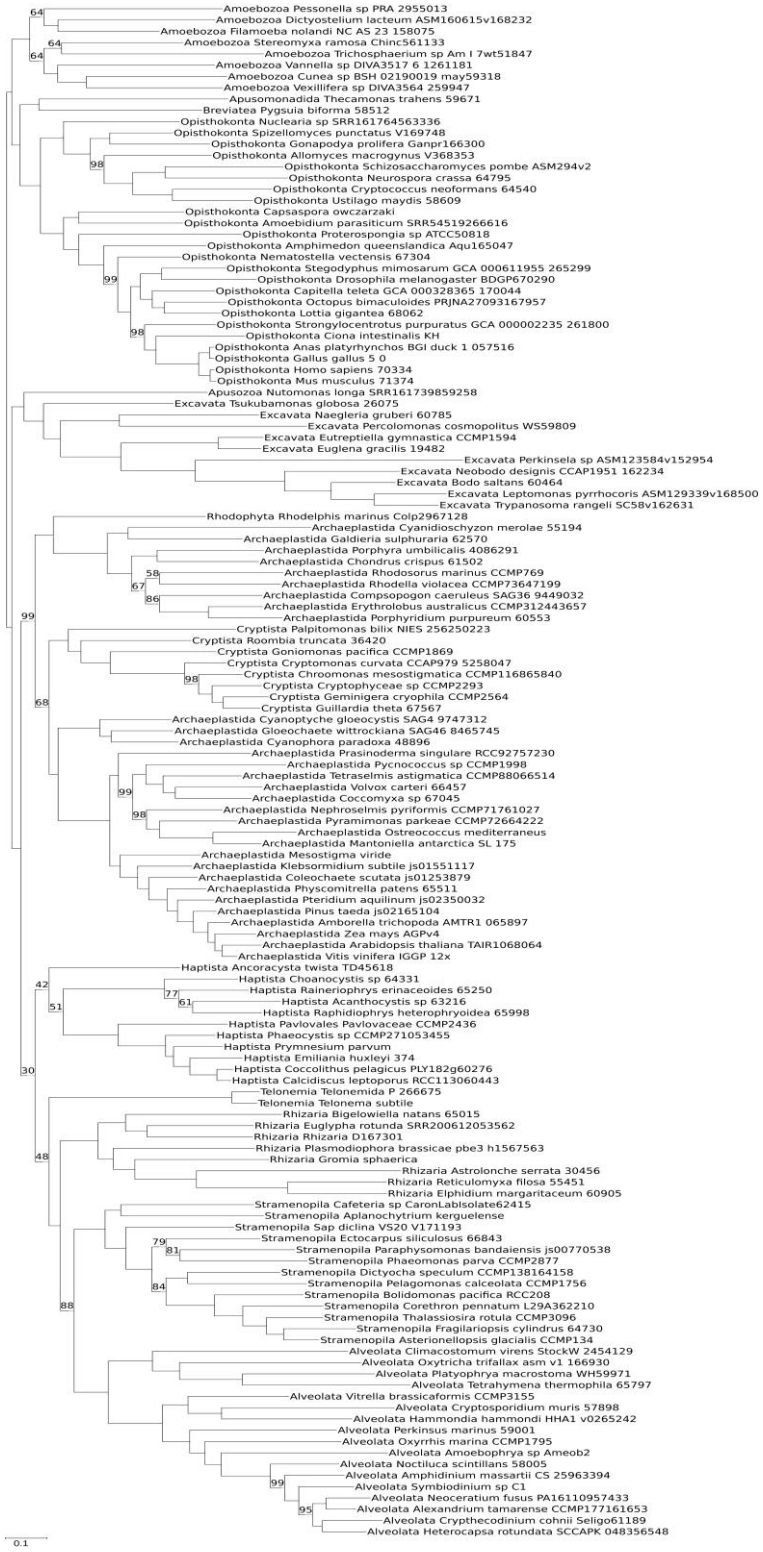


Figure A34 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al.*, 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6263637. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

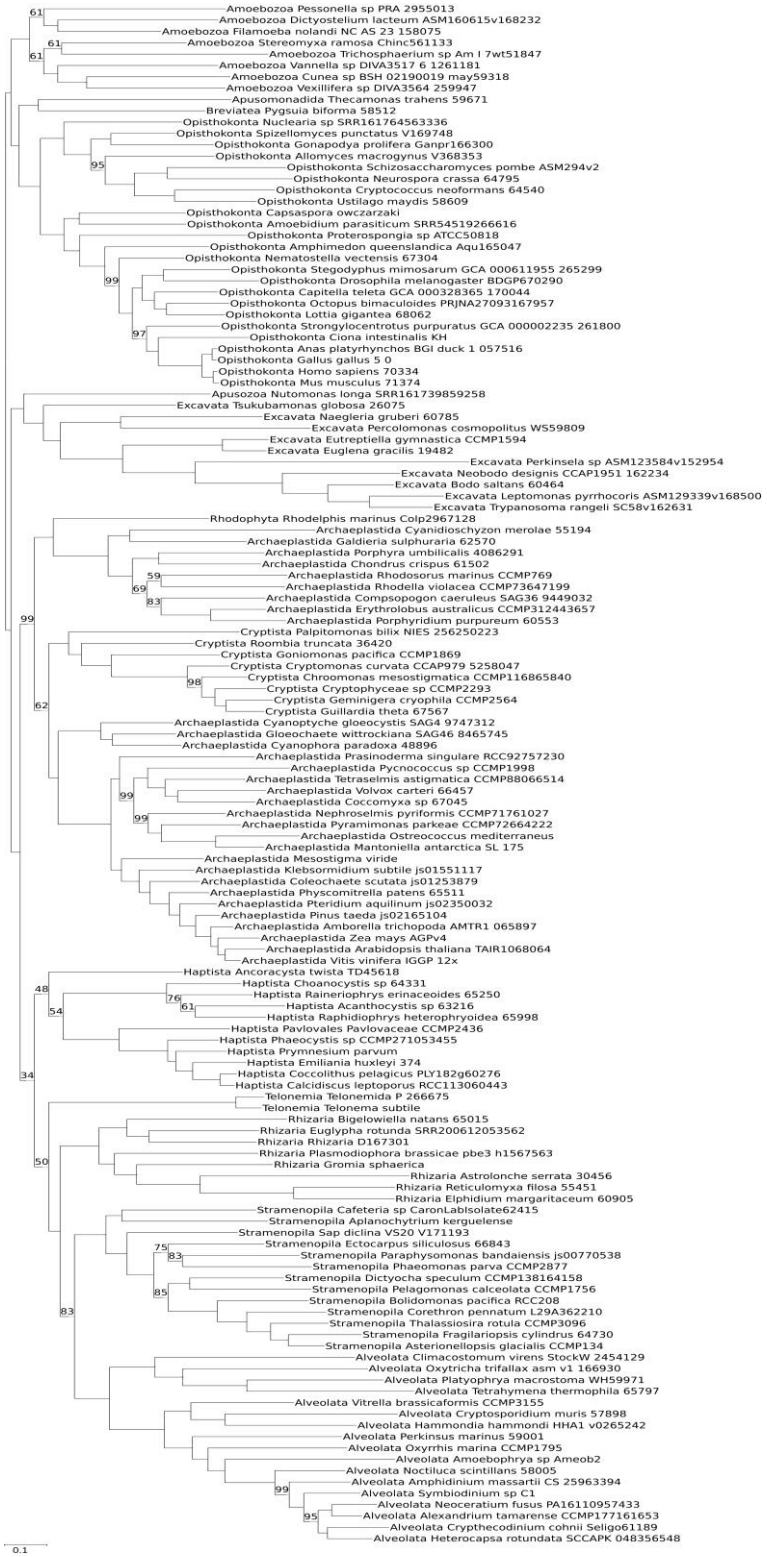


Figure A35 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al*., 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6187299. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

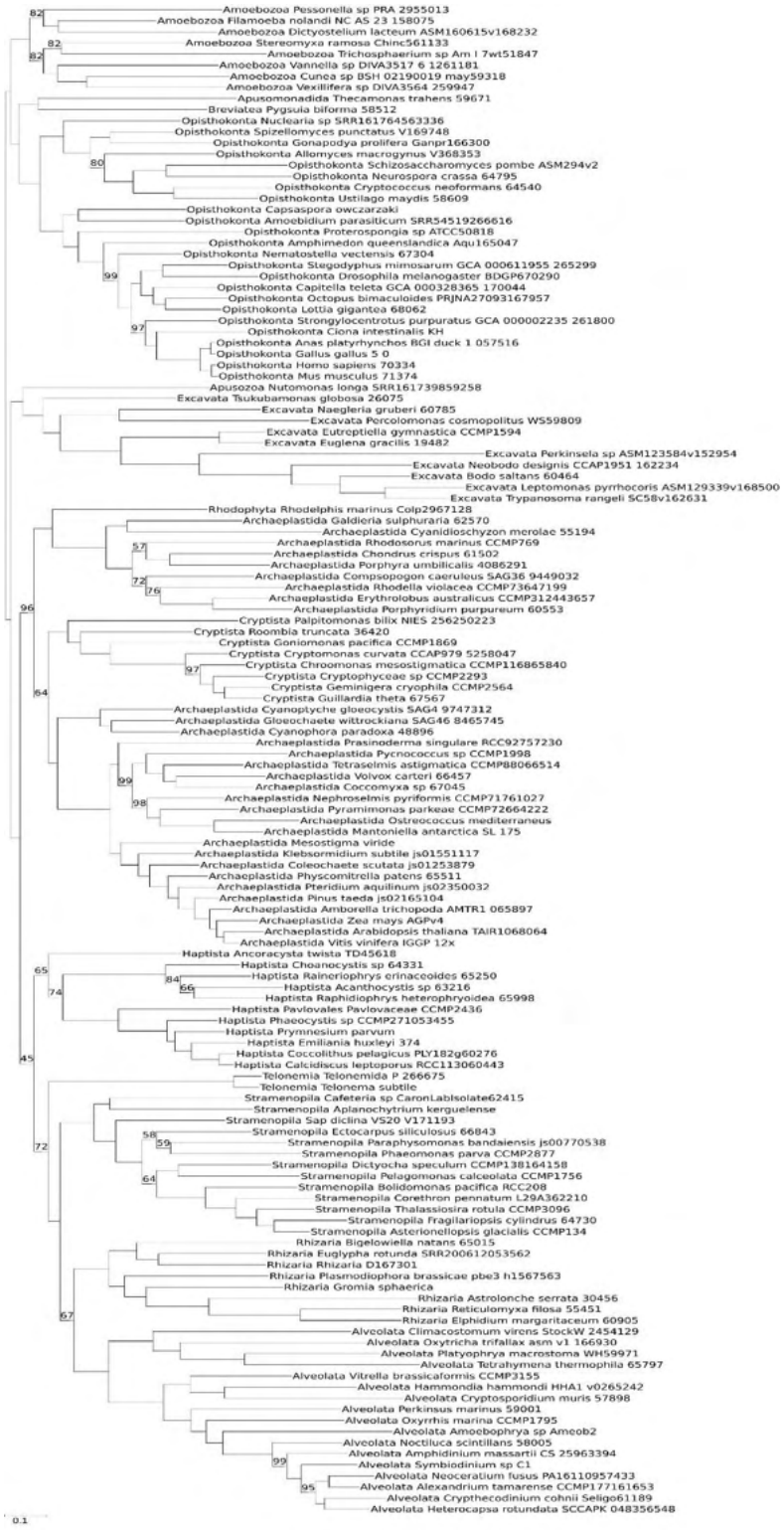


Figure A36 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6140112. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

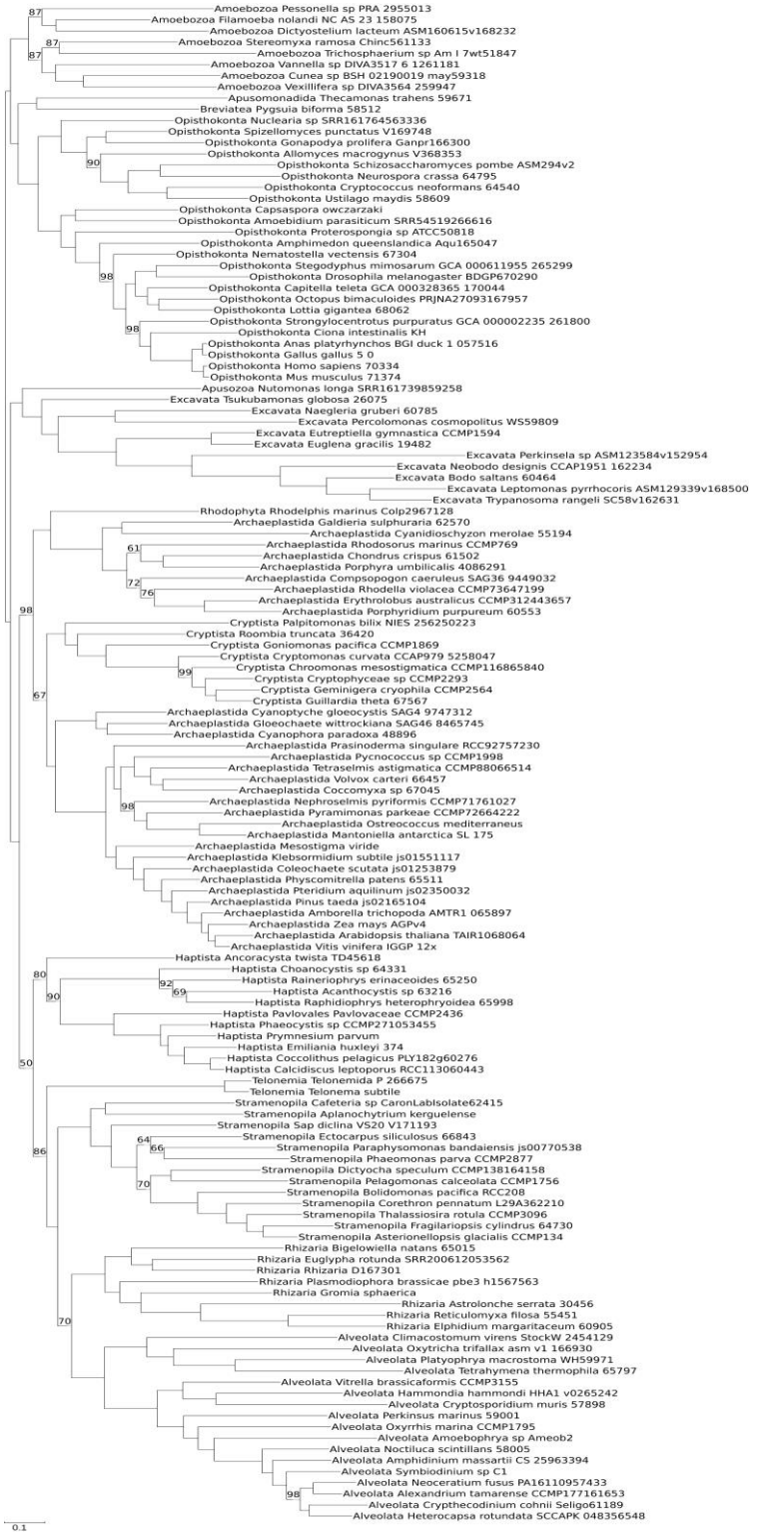


Figure A37 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6140096. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

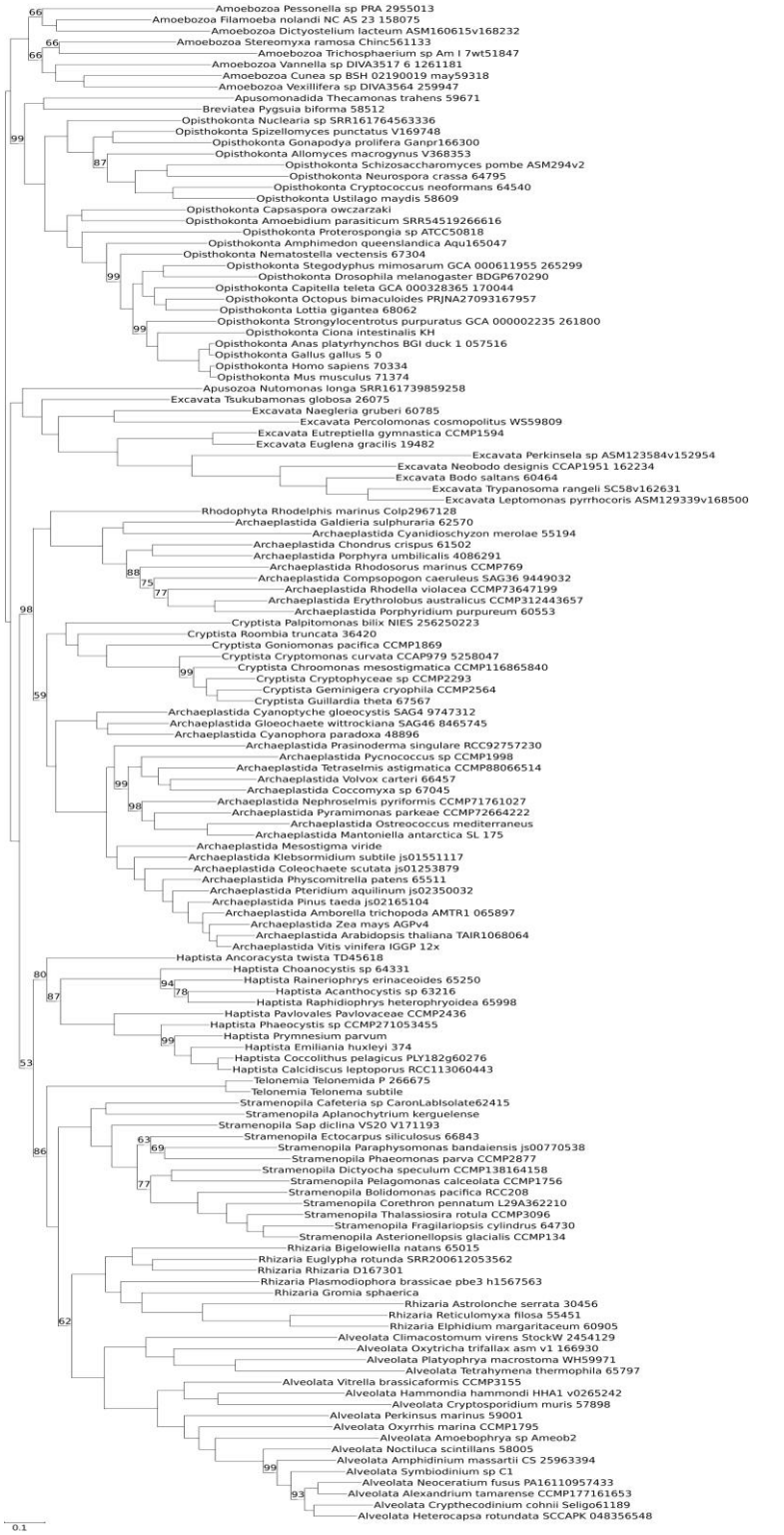


Figure A38 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6013630. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

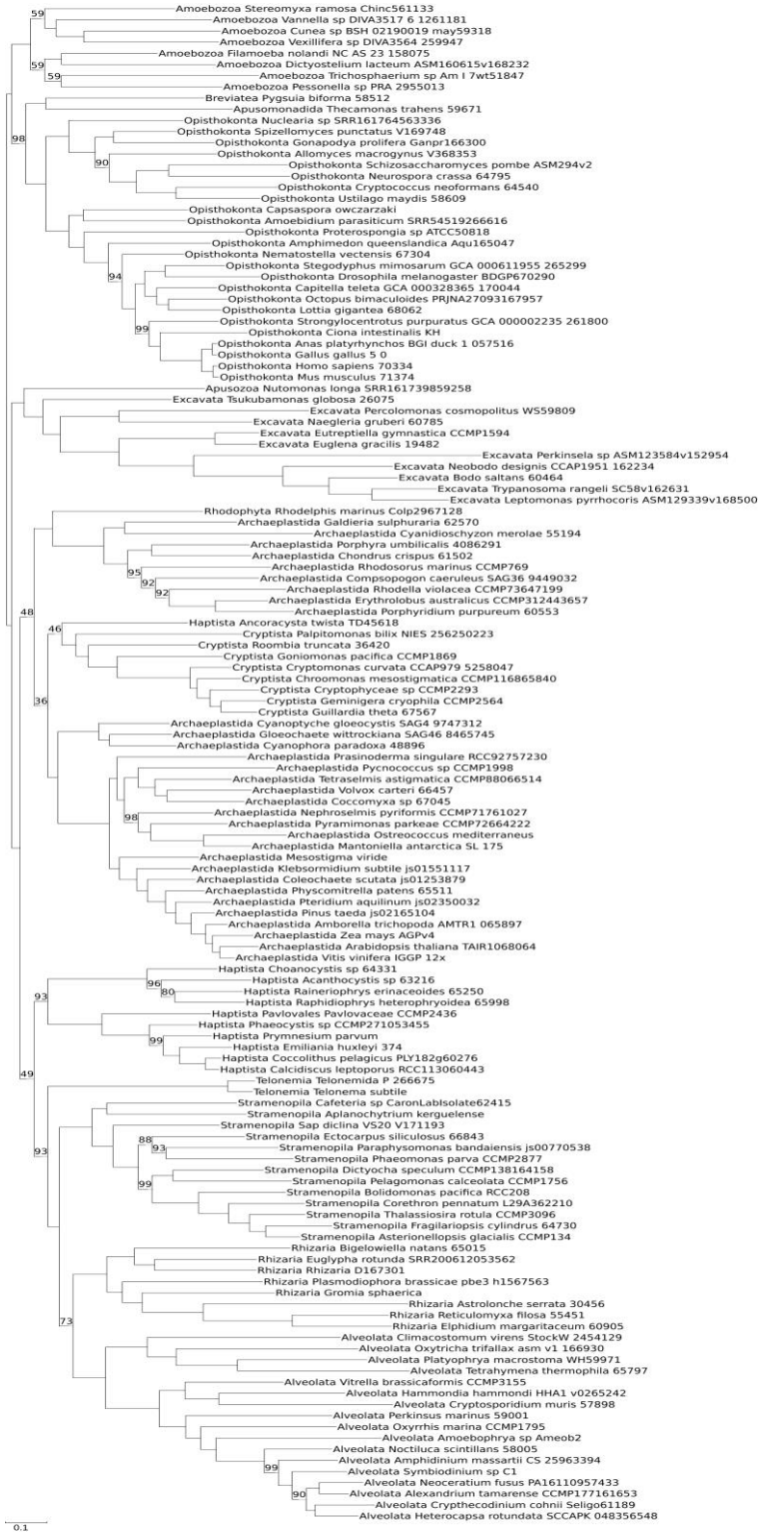


Figure A39 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -5728627. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A40 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -5574791. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A41 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -5559901. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

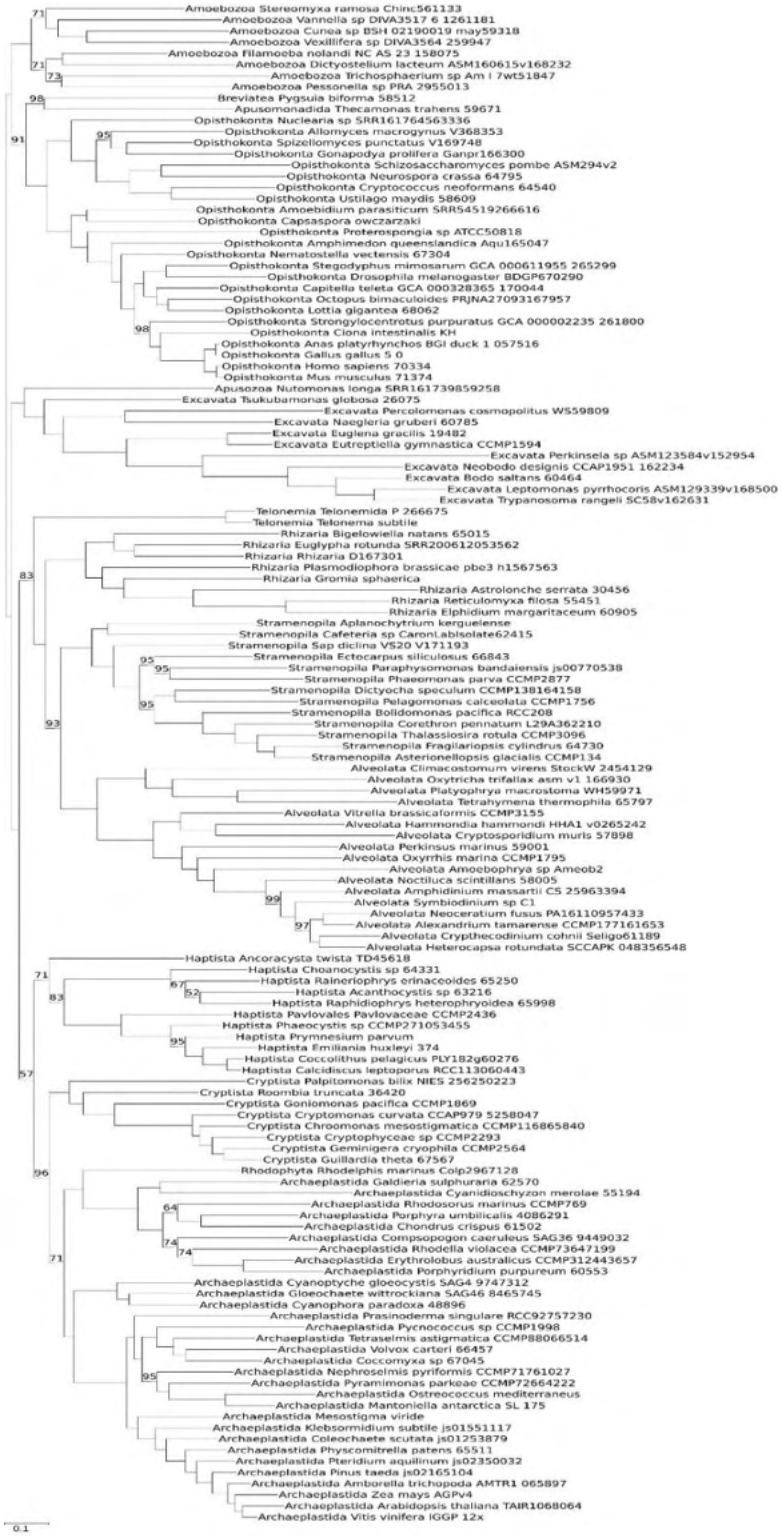


Figure A42 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -5025457. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

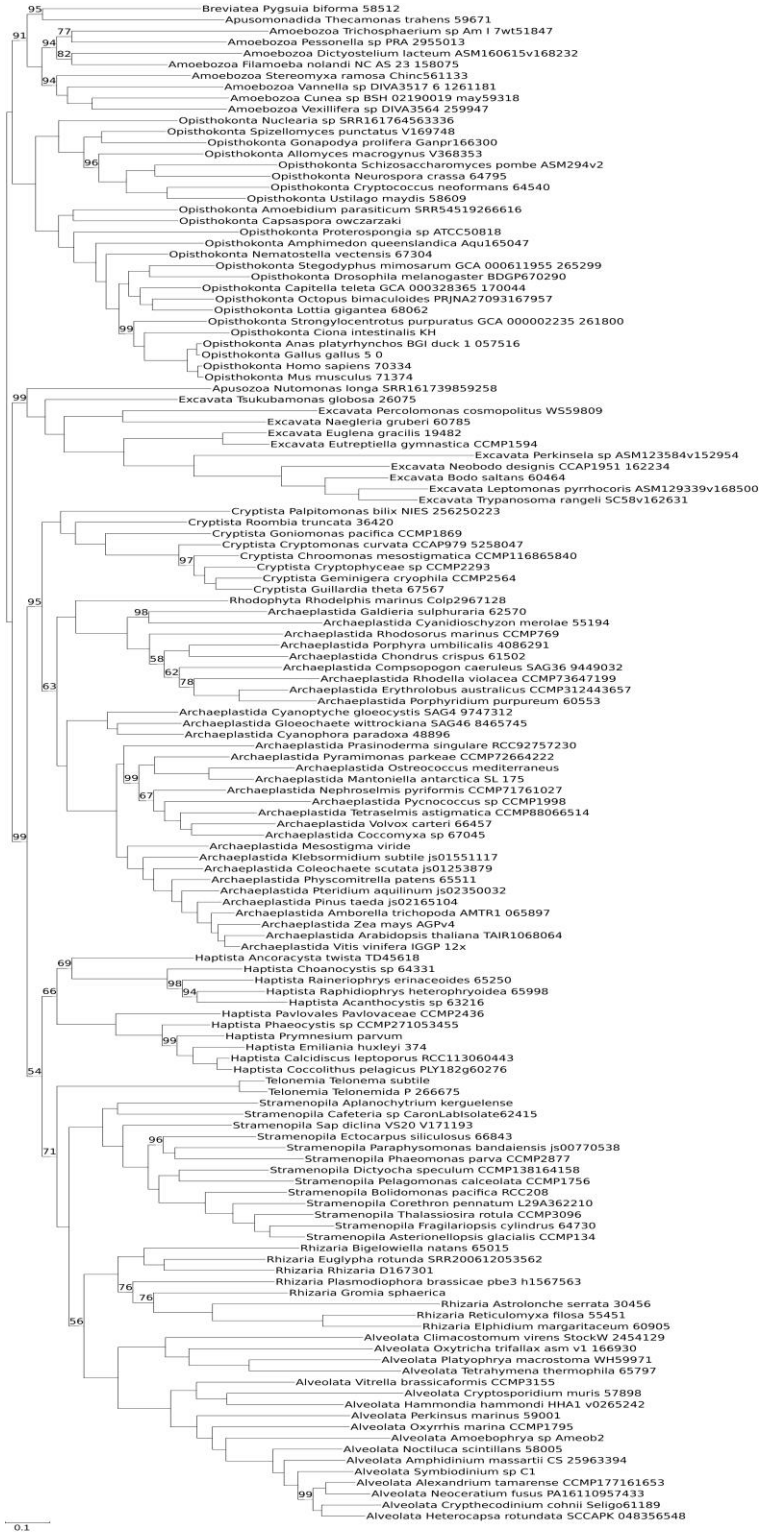


Figure A43 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 84 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -4252606. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

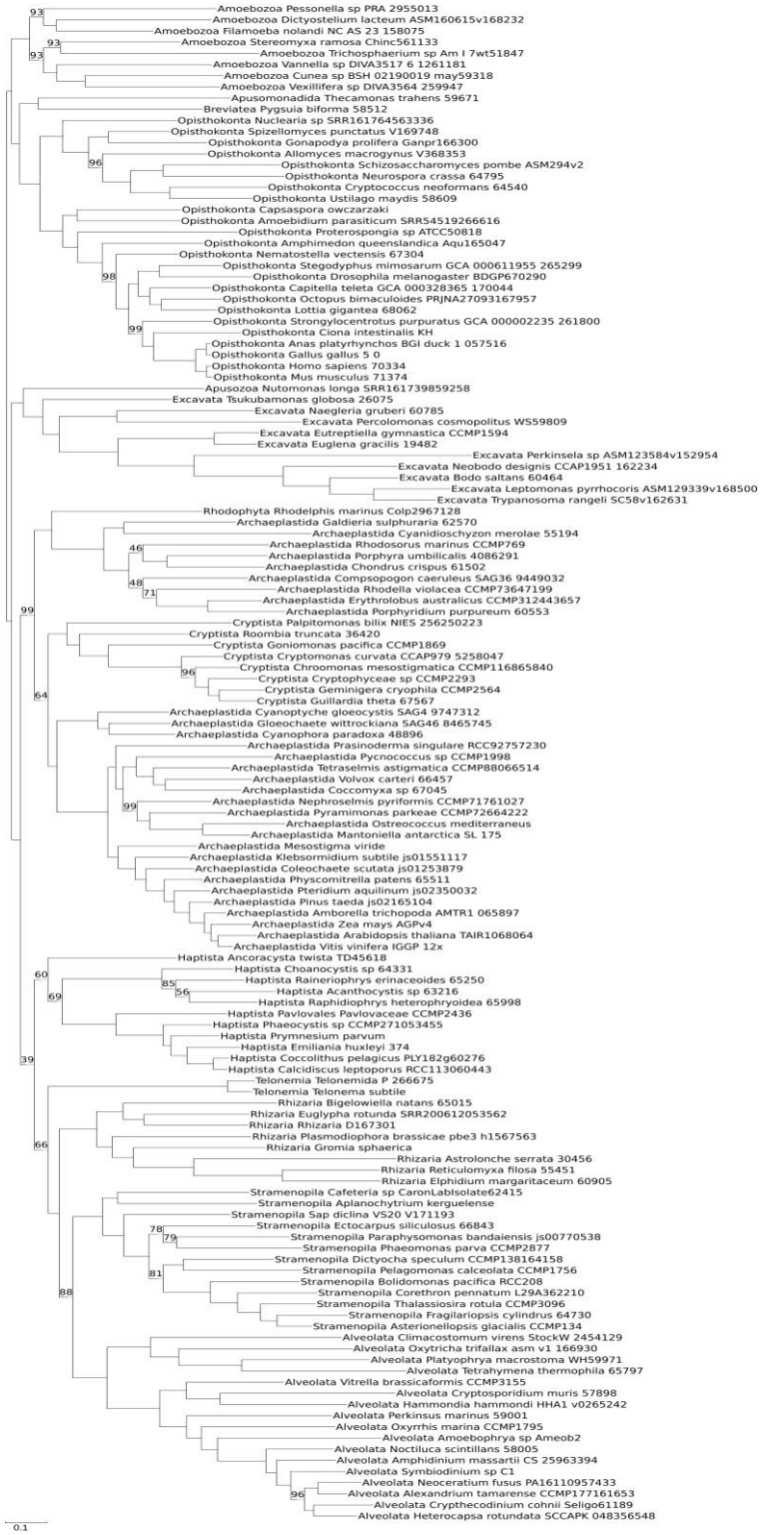


Figure A44 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 41 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6110648. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

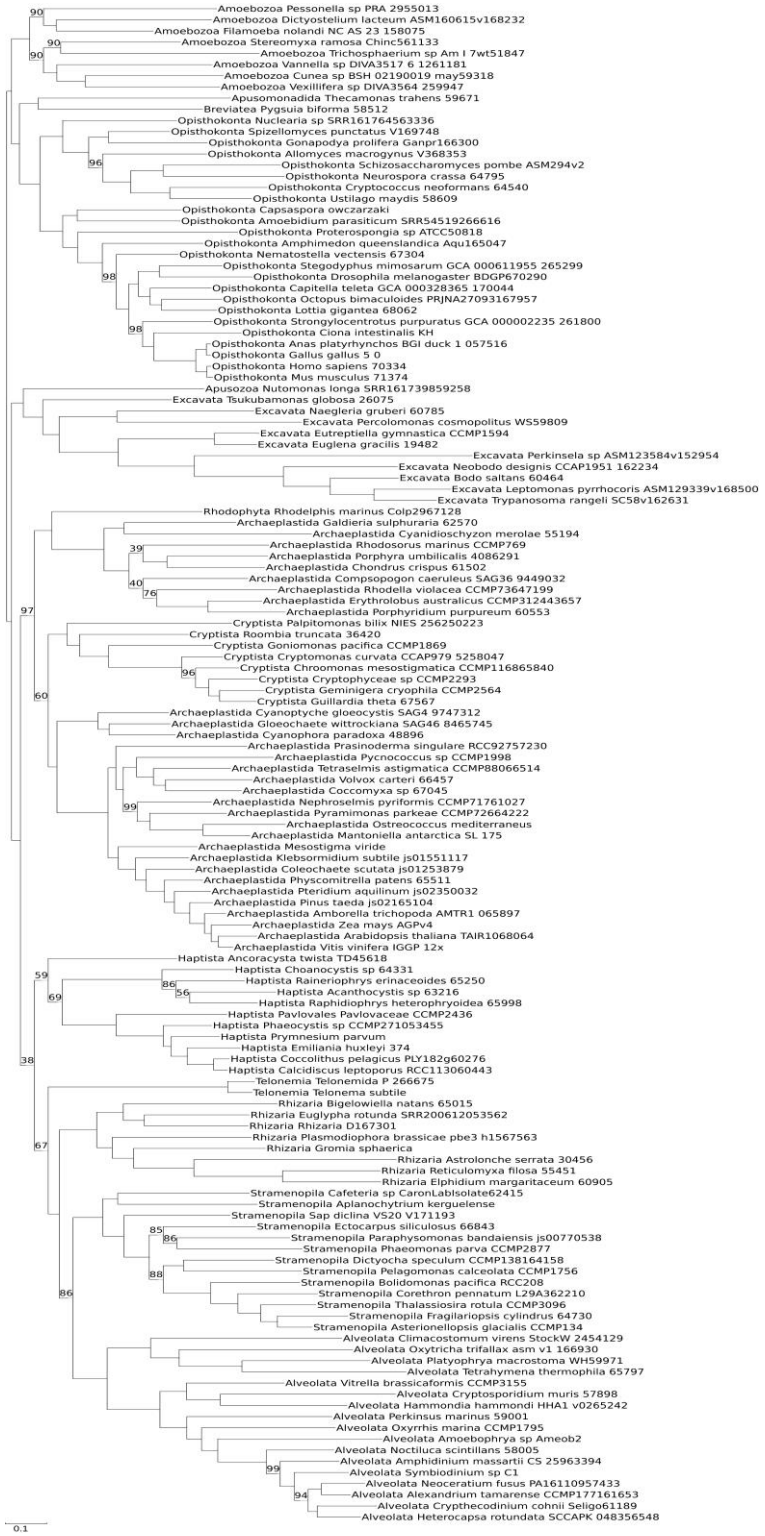


Figure A45 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 41 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6110632. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

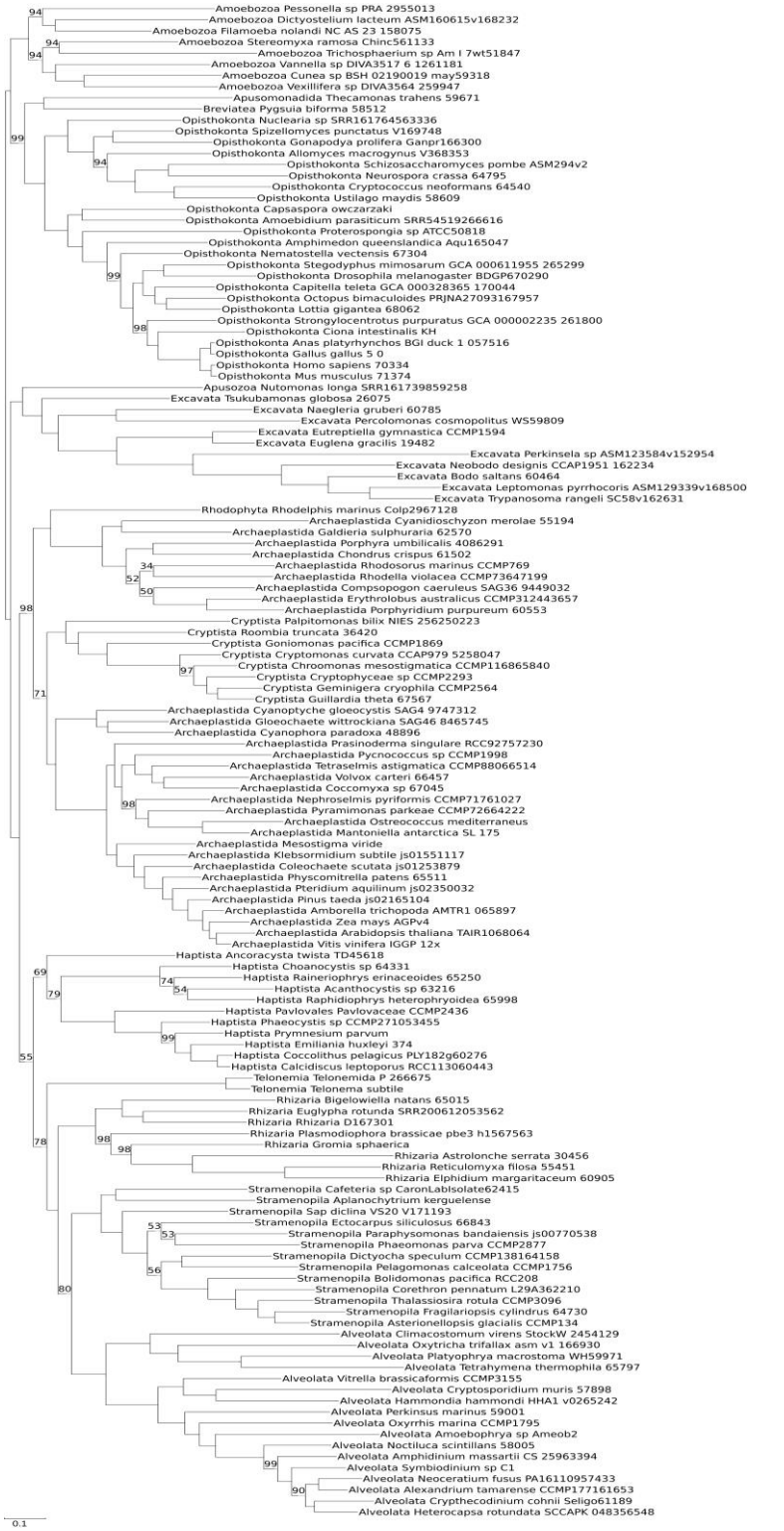


Figure A46 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 41 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from partition according to the matched-pairs test of symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6005780. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A47 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 41 data partitions (derived from the Strassert *et al*., 2021 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -5624277. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A48 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al.*, 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -5424637. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A49 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al.*, 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Holm p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -5418673. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A50 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al*., 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -4683674. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A51 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al*., 2021 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -3688287. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A52 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al*., 2021 data set). Tree-heterogeneous sequences were filtered out from the entire concatenated data set according to the matched-pairs test of internal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6237932. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

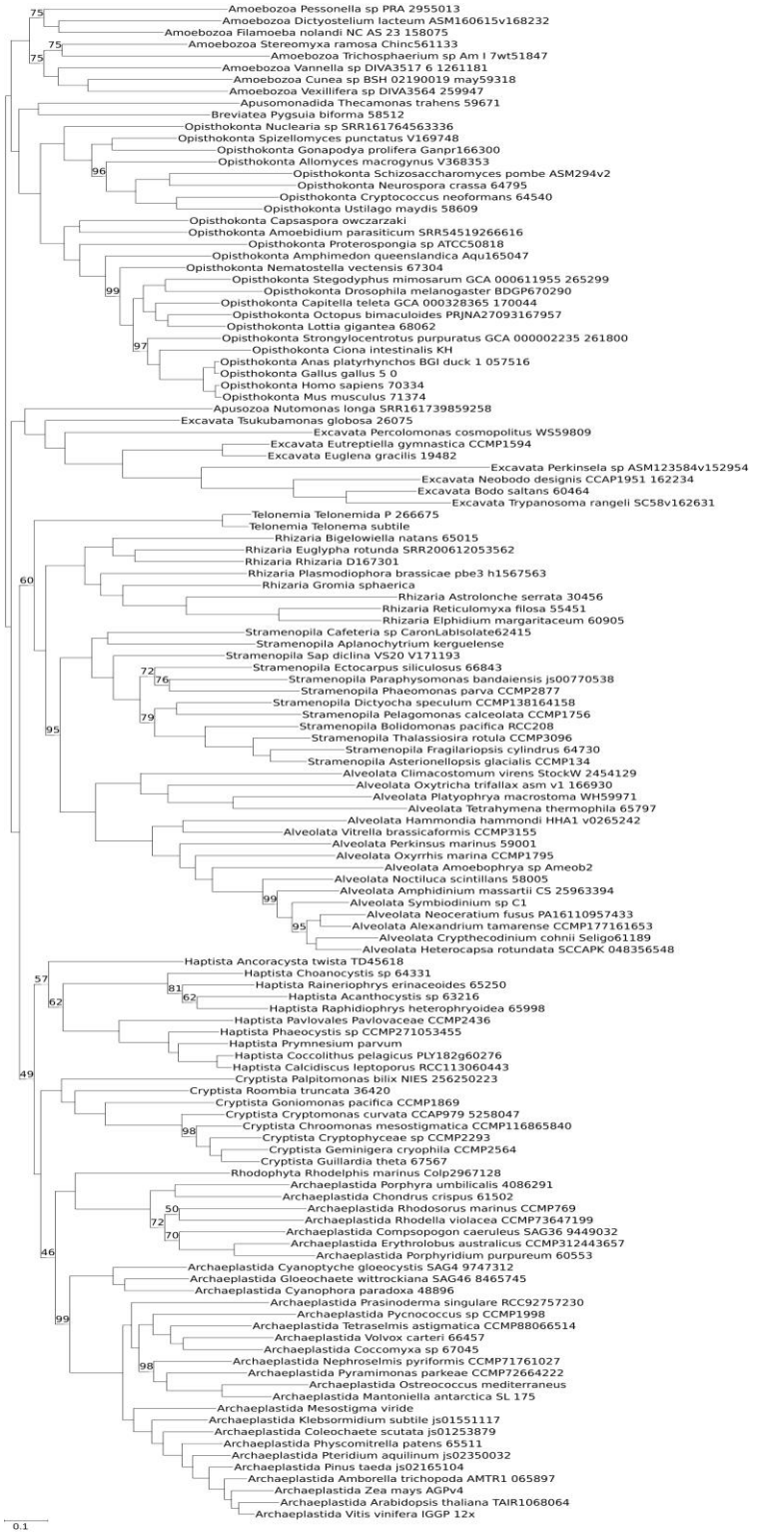


Figure A53 - Optimal maximum-likelihood tree comprising 136 taxa reconstructed from 320 concatenated proteins (derived from the Strassert *et al.*, 2021 data set). Tree-heterogeneous sequences were filtered out from the entire concatenated data set according to the matched-pairs test of internal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -6005116. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A54 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 209 concatenated proteins (Whelan et al., 2015 data set). Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -3154365. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A55 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 209 concatenated proteins (derived from the Whelan et al., 2015 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -3009644. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A56 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 209 concatenated proteins (derived from the Whelan et al., 2015 data set). Tree-heterogeneous sequences were filtered out from each protein alignment according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2872866. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A57 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 28 data partitions (derived from the Whelan *et al*., 2015 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2532046. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A58 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 28 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2361684. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

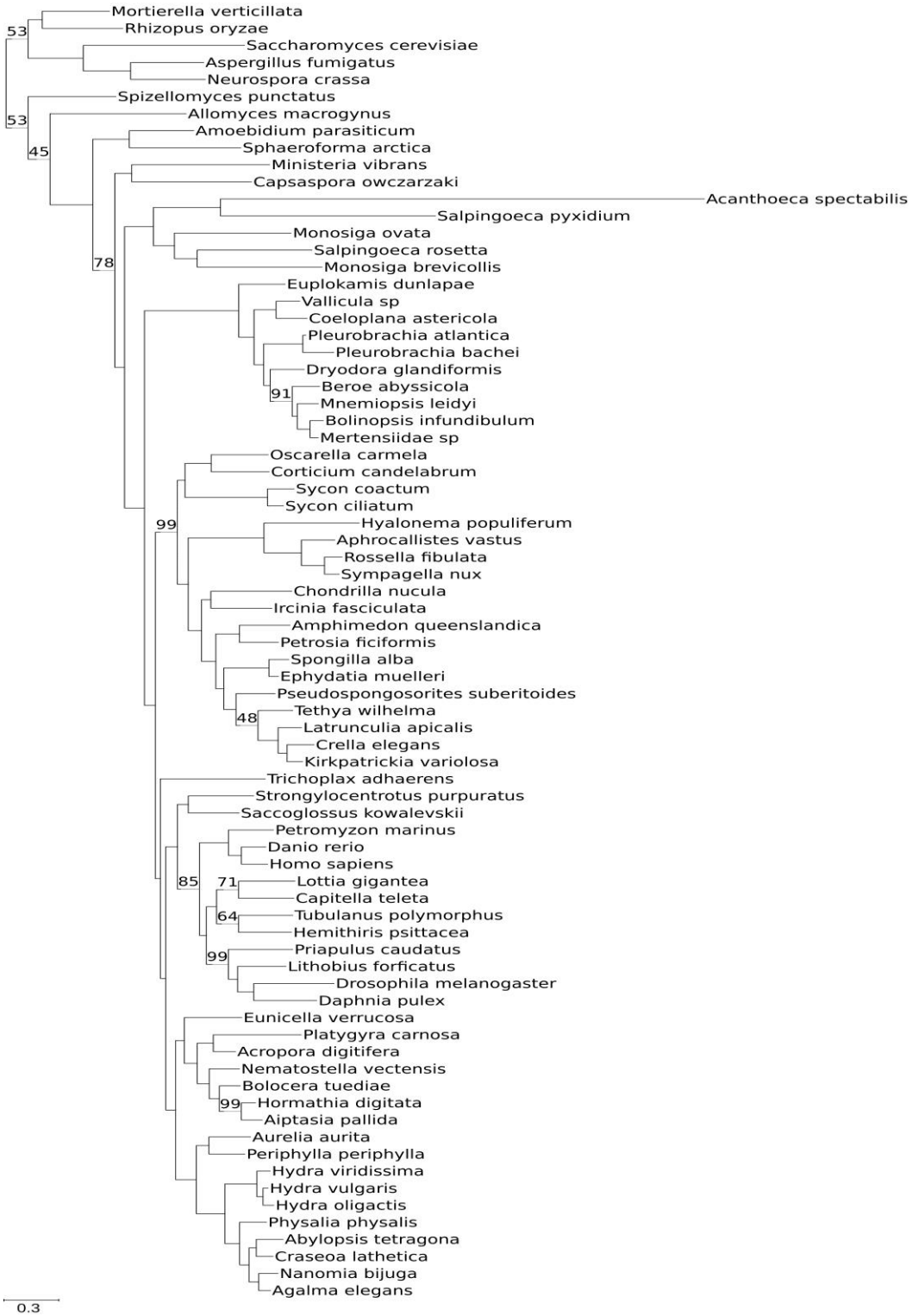


Figure A59 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 28 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2079561. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

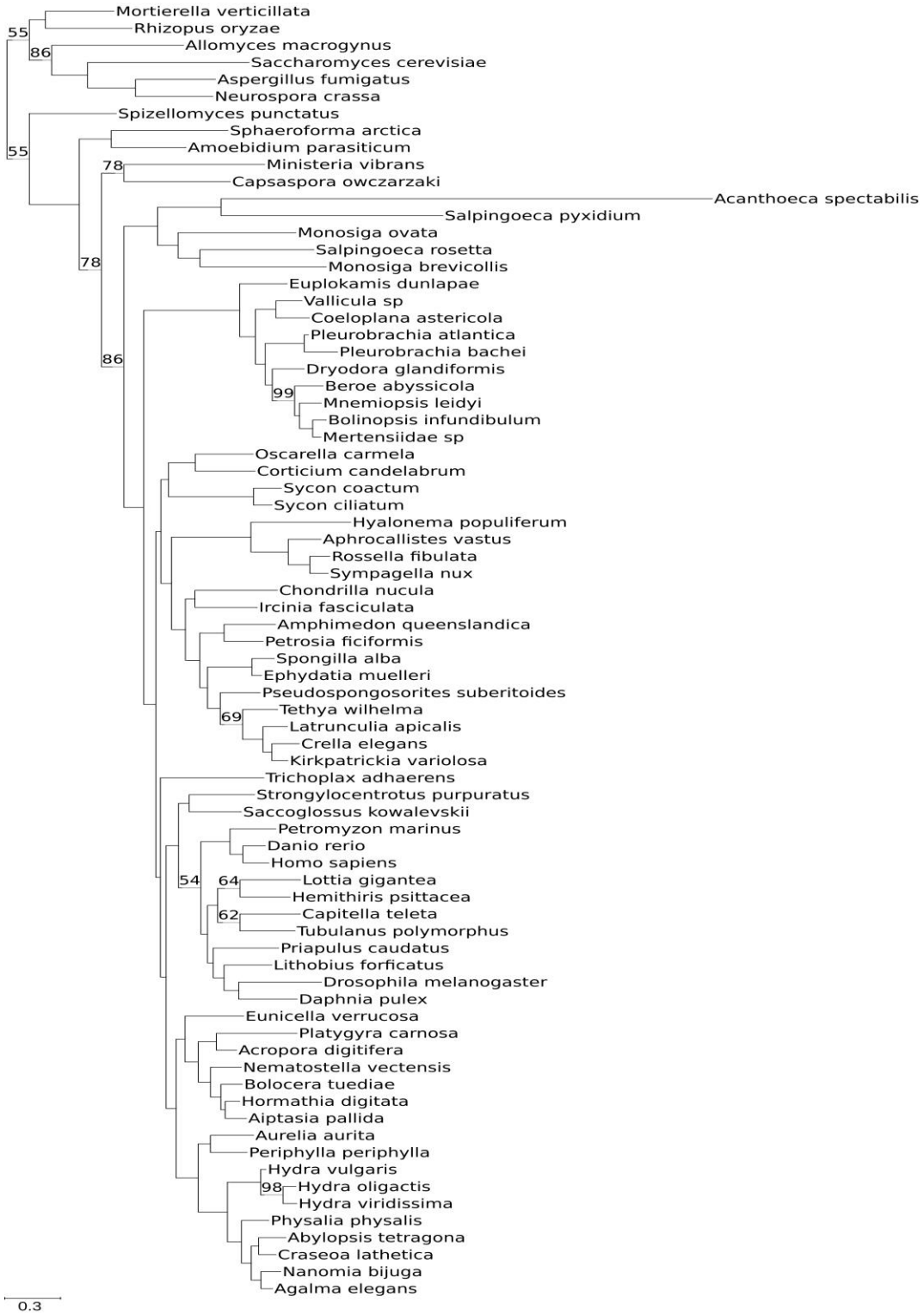


Figure A60 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 28 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1878936. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

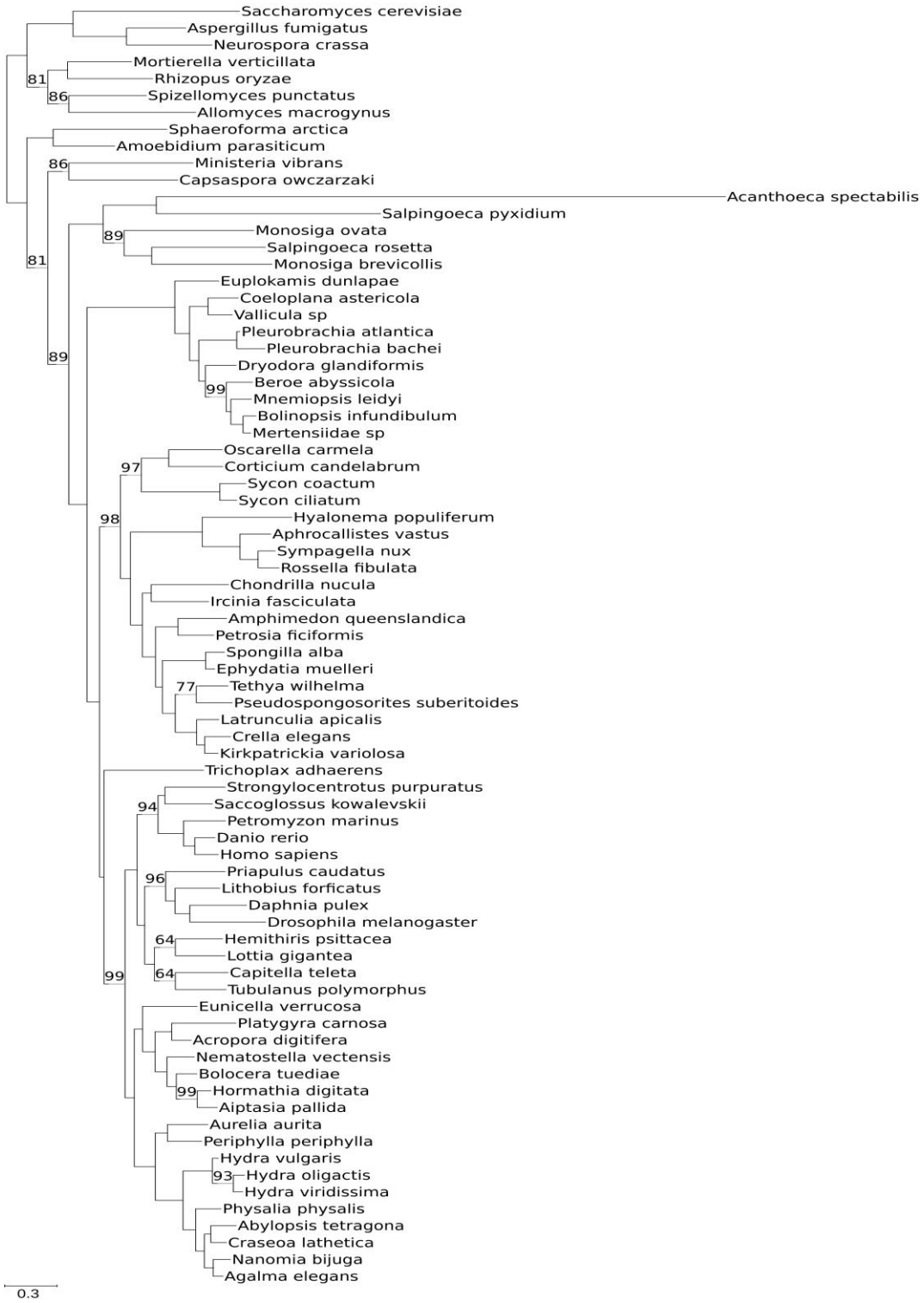


Figure A61 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 28 data partitions (derived from the Whelan *et al*., 2015 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1584491.Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

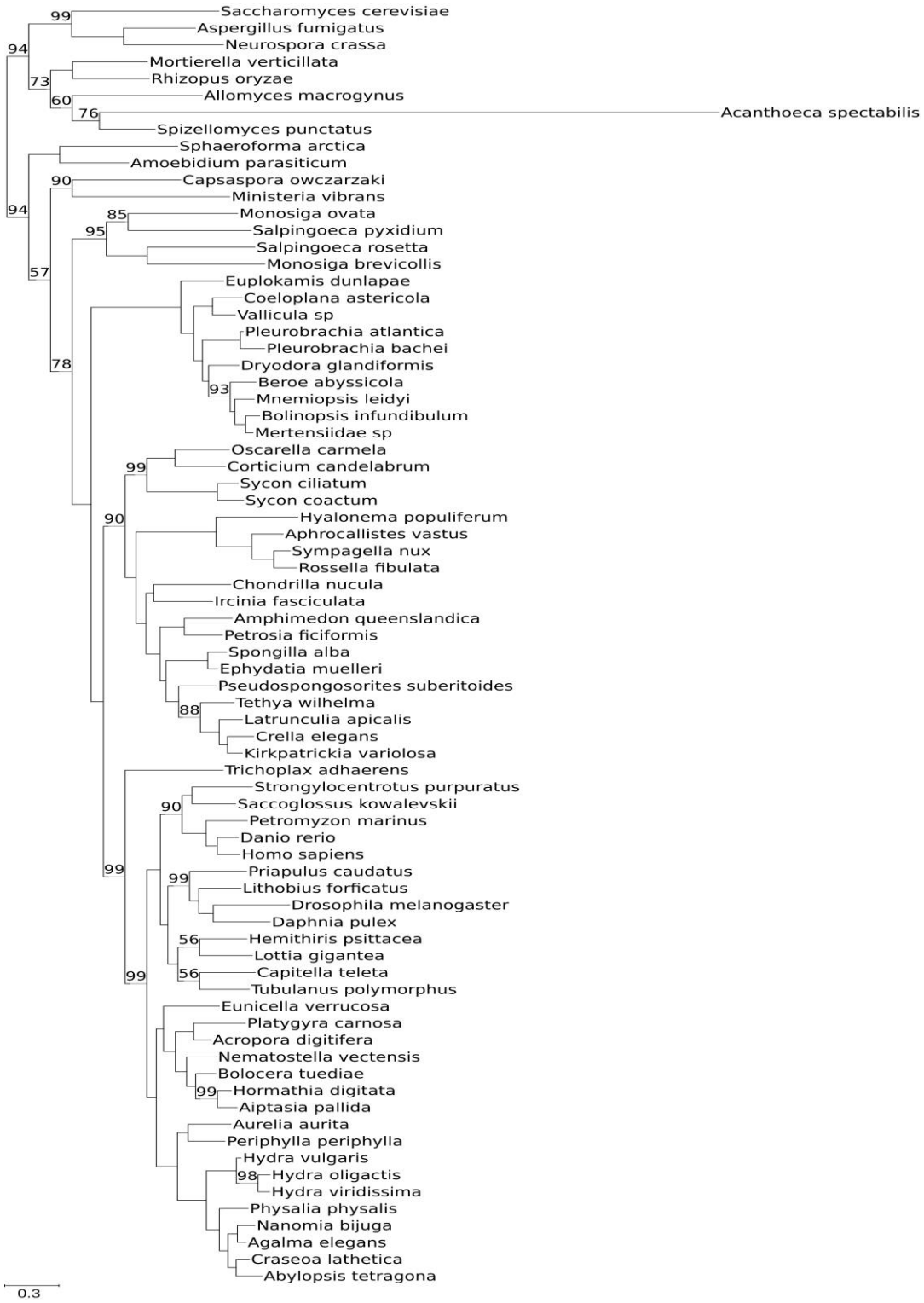


Figure A62 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 28 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the K-means algorithm. Tree-heterogeneous sequences were filtered out from each partition according to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1179076. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A63 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 35 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition to the matched-pairs test of symmetry coupled with the Bonferroni p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2932527. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

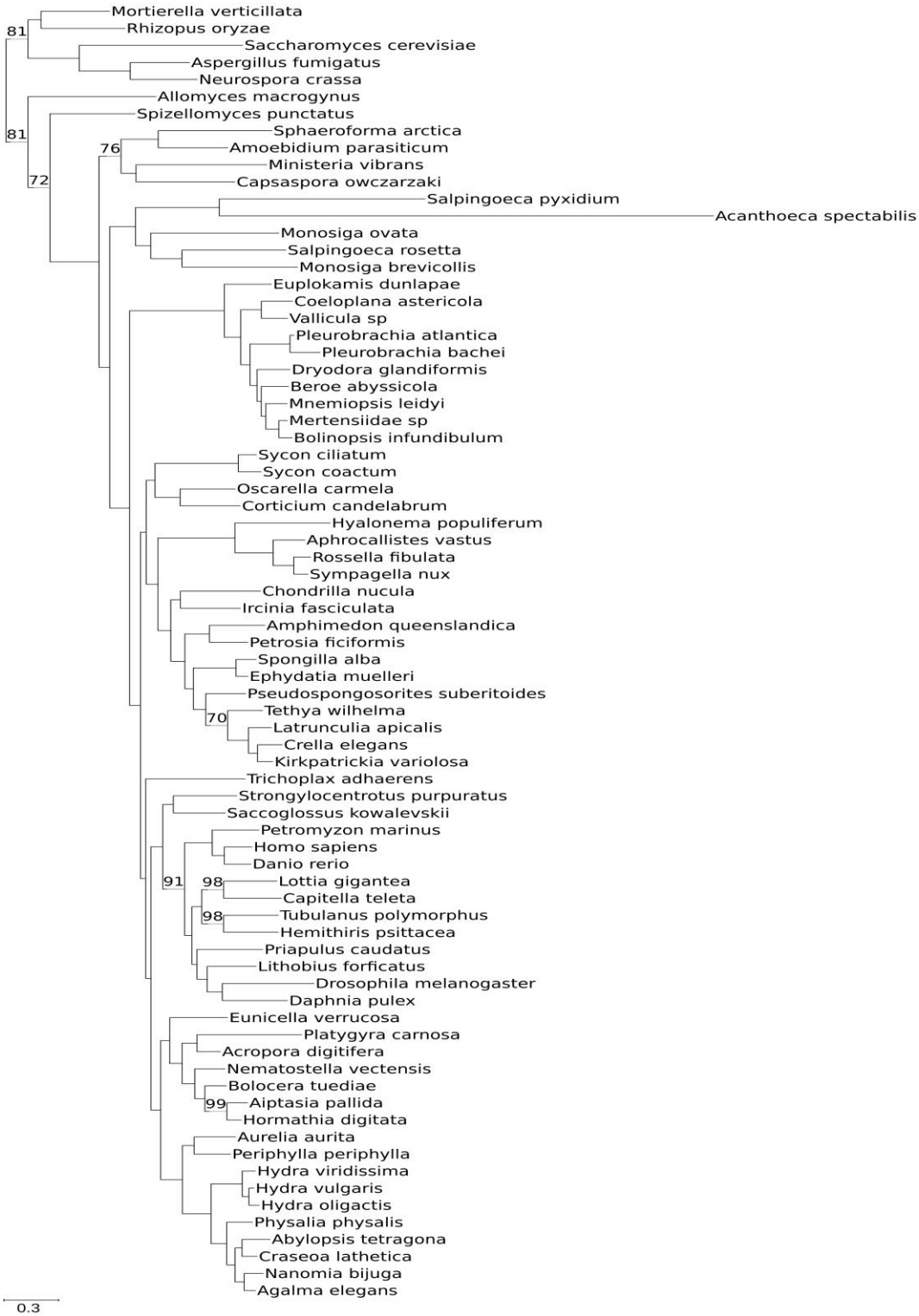


Figure A64 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 35 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition to the matched-pairs test of symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2882923. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A65 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 35 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition to the matched-pairs test of symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2708062. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

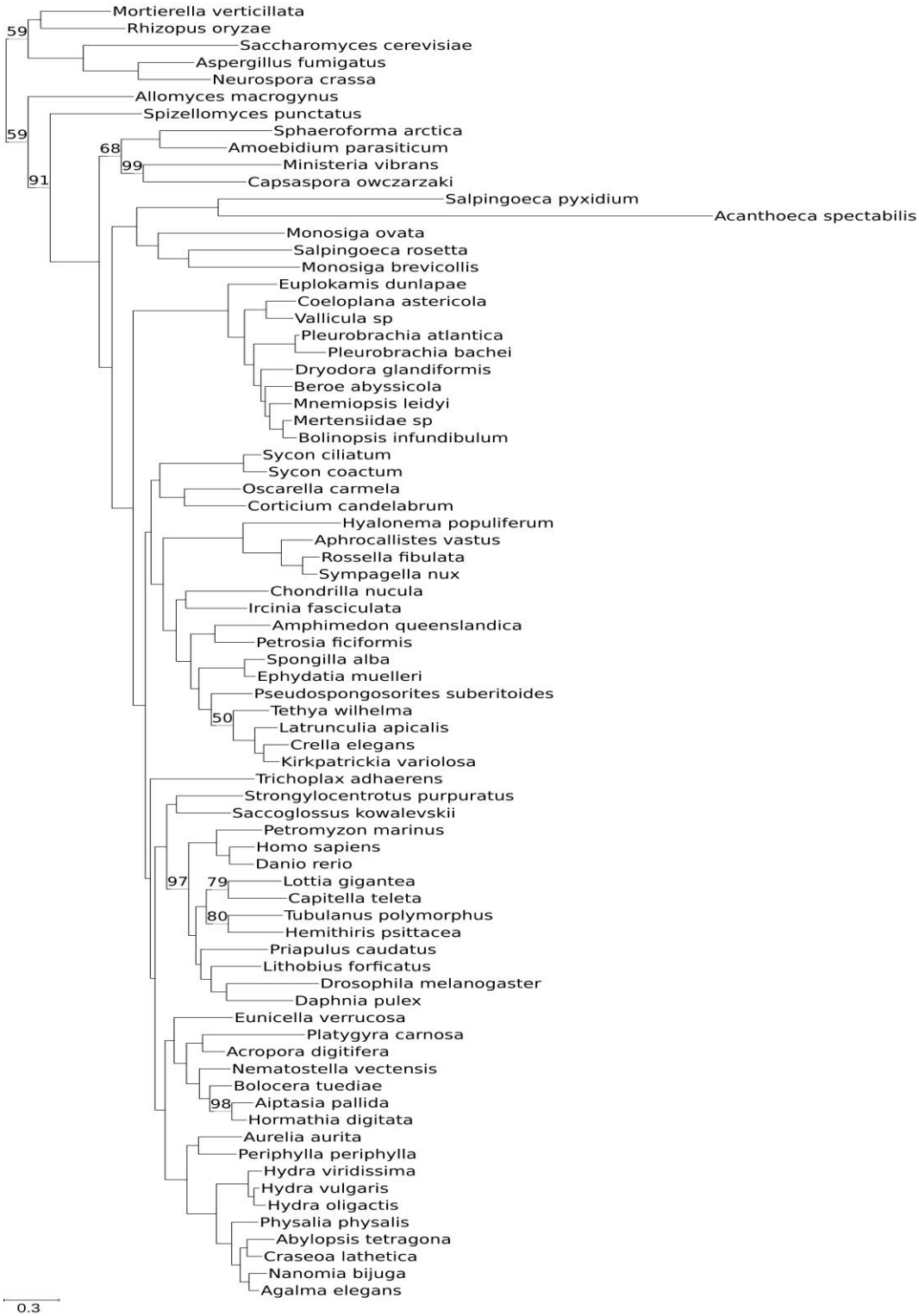


Figure A66 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 35 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition to the matched-pairs test of marginal symmetry coupled with the Bonferroni p-value adjustment procedure Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2371704. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A67 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 35 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition to the matched-pairs test of marginal symmetry coupled with the Benjamini-Yekutieli p-value adjustment procedure Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2095021. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Figure A68 - Optimal maximum-likelihood tree comprising 76 taxa reconstructed from 35 data partitions (derived from the Whelan *et al.*, 2015 data set). Data partitions were defined using the ModelFinder. Tree-heterogeneous sequences were filtered out from each partition to the matched-pairs test of marginal symmetry coupled with the Benjamini-Hochberg p-value adjustment procedure Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1696977. Support values at nodes are maximum-likelihood ultrafast bootstraps calculated from 10,000 replicates. Support values from nodes fully supported are omitted.

Chapter IV

Identifying the closest living algal relatives of land plants using data-specific amino acid substitution models to analyse each of the three plant genomes

Identifying the closest living algal relatives of land plants using data-specific amino acid substitution models to analyse each of the three plant genomes

Abstract

Charophyte green algae comprise six main groups and gave rise to land plants approximately 470 Mya ago, however the precise sister-group lineage to land plants has been difficult to determine. In this study, data from the three genomic compartments, the nucleus, chloroplast, and mitochondria, were used to infer the evolutionary history of charophyte algae and land plants (i.e. Streptophyta). Amino-acid sequence data were analysed using data-specific substitution models and the influence of systematic bias was investigated. Among-lineage heterogeneity of the substitution process was assessed for each single protein alignment using the matched-pairs tests of symmetry and the χ^2 test for compositional homogeneity and phylogenies inferred using site-based and tree-based homogeneous and heterogeneous substitution models. Data were also partitioned and analysed according to criteria such as site-specific rates and the position of the sites in the tertiary protein structure. Phylogenies derived from nuclear and chloroplast data recovered Zygnematophyceae as the sister-group to land plants, while the mitochondrial analyses recovered Charophyceae as the sister-group to land plants. Further mitochondrial data analyses using partitioned or reduced data indicated a second phylogenetic signal favouring Zygnematophyceae as the sister-group to land plants. The matched-pairs tests of symmetry and the χ^2 test for compositional homogeneity analyses indicated the mitochondrial data as the most compositionally tree-heterogeneous. Exclusion of tree-heterogeneous sequences from the mitochondrial data sets did not affect the relationship of Charophyceae to land plants but did recover the Setaphyta clade. The same methodology applied to chloroplast data improved the support for the bryophytes monophyly. The congruence between analyses of both nuclear and chloroplast data strongly supporting the Zygnematophyceae as the sister-group of land plants suggests that this relationship is most likely correctly reconstructed. However, there appears to be two distinct phylogenetic signals in the mitochondrial data supporting either Charophyceae and Zynematophyceae as the sister group to land plants. Analyses using tree-heterogeneous substitution models suggest that this result is not due to systematic bias caused by poor model-fit with respect to among-lineage variation: the cause of these differing signals, be they biological or due to systematic error, remain unknown.

4.1 Introduction

The phylum Streptophyta comprises the land plants and their green algal ancestors the charophyte algae, with land plants having diverged from charophytes approximately 470 Mya ago (Lewis and McCourt, 2004; Morris *et al.*, 2018; Puttick *et al.*, 2018). The identity of the closest-living extant algal relative of land plants has, however, been vigorously debated in recent decades (e.g., Karol *et al*., 2001; Turmel *et al*., 2007; Wodniok *et al.*, 2011; Wickett *et al.*, 2014; Leebens-Mack *et al*., 2019). Charophytes form a paraphyletic grade with six distinct lineages, namely, the Mesostigmatophyceae, Chlorokybophyceae, Klebsormidiophyceae, Charophyceae, Coleochaetophyceae, and Zygnematophyceae. The Mesostigmatophyceae and Chlorokybophyceae, plus *Spirotaenia* spp., diverged first, are species-poor, and the simplest charophyte algae. *Mesostigma viride* Lauterborn (Mesostigmatophyceae) is a freshwater asymmetric, biflagellate, unicell with scales, and is the only charophyte with a motile vegetative stage. Morphological and molecular studies place it as the earliest-diverging charophyte (Melkonian, 1989; Karol *et al*., 2001; Nedelcu *et al*., 2006; Petersen *et al*., 2006; Lemieux *et al*., 2007; Finet *et al*., 2010; Liang *et al.*, 2020). *Chlorokybus atmophyticus* Geitler, and four recently described *Chlorokybus* species (Irisarri *et al*., 2021), compose the only genus of Chlorokybophyceae. *Chlorokybus* inhabits wet soils and is composed of mucilaginous packets of cells. Initial molecular phylogenetic analyses (Karol *et al.*, 2001; Cocquyt *et al*., 2010b; Finet *et al*., 2010, 2012), as well as the absence of a flagellum in the vegetative stage, indicate that *Chlorokybus* is not a sister lineage of *Mesostigma,* but diverged later. However, later phylogenetic studies have recovered the two lineages united and sister to the remaining streptophytes (Lemieux *et al.*, 2007; Timme *et al.*, 2012; de Vries *et al.*, 2018; Leebens-Mack *et al.*, 2019; Wang *et al.*, 2019). *Spirotaenia* spp. are unicellular algae that occur in freshwater environments and were initially placed within Zygnematophyceae due to the lack of flagella and sexual reproduction by conjugation. Nevertheless, phylogenetic analyses of small subunit ribosomal RNA (SSU rRNA) and *rbc*L gene sequences recovered these algae outside of Zygnematophyceae (Gontcharov and Melkonian, 2004). This was corroborated by later analyses (Wickett *et al.*, 2014; Leebens-Mack et al., 2019; Irisarri *et al*., 2021), which placed *Spirotaenia minuta* Thuret sister to *Chlorokybus.* This lineage, together with Mesostigmatophyceae and Chlorokybophyceae form one clade sister to the remaining charophytes, with the latter diverging as a grade. The earliest-diverging lineage within this grade is Klebsormidiophyceae (Karol *et al*., 2001; Turmel *et al*., 2002a; Cocquyt *et al*., 2010b; Finet *et al*., 2010; Wodniok *et al*., 2011; Hori *et al.*, 2014; Lemieux *et al.*, 2016). Their members are terrestrial and freshwater algae and form

unbranched filaments or sarcinoid packets. The genera *Klebsormidium* and *Interfilum* have been placed as sister-groups, while *Entransia* has been recovered as the earliest-diverging lineage or with *Hormidiella* (Sluiman *et al.,* 2008). The genera *Streptosarcina* (recently proposed*;* Mikhailyuk *et al.,* 2018) was placed sister to *Hormidiella.*

The remaining three classes of charophytes, namely, Charophyceae, Coleochaetophyceae, and Zygnematophyceae, have greater structural complexity than other charophycean algae and display several ultrastructural and biochemical features also found in land plants (McCourt *et al.,* 2004). These features include two subapically inserted flagella with a single multilayered structure in motile cells, a persistent mitotic spindle, asymmetrical cell division, a cytokinetic phragmoplast, the presence of plasmodesmata, the presence of unique enzymes such as Cu/Zn superoxide dismutase, and the shared duplication of the GapA/GapB gene (Mattox and Stewart, 1984; de Jesus *et al*., 1989; Graham and Kaneko, 1991; Graham *et al.,* 2000; Petersen *et al.,* 2006; Leliaert *et al.,* 2012). The Charophyceae is morphologically complex macroscopic organisms with one order, Charales, and several genera. Most species have shoots ranging from centimetres to meters in length, with apical growth, and oogamus and clonal reproduction. Coleochaetophyceae comprises the order Coleochaetales and the genera *Coleochaete* and *Chaetosphaeridium*. Their members are microscopic algae, typically with branched filaments, oogamous sexual reproduction, and asexual reproduction by the formation of zoospores.

Zygnematophyceae is highly diverse and form the most speciose algal clade (4,378 species currently in AlgaeBase; Guiry and Guiry, 2019) within the Streptophyta. These algae are filamentous and unicellular with reproduction by conjugation, while flagellate stages, plasmodesmata, and apical growth are absent. Zygnematophyceae include two main orders, Zygnematales and Desmidiales, separated based on their morphology, cell wall ornamentation, and structure (Mix, 1972; Gerrath, 2003), and supported by phylogenetic analyses (Bhattacharya *et al.,* 1994; McCourt *et al*., 2000; Gontcharov *et al.,* 2008). The Zygnematales was resolved as paraphyletic (McCourt *et al*., 2000; Gontcharov *et al*., 2003), with some members (*Netrium* and *Roya*) showing greater affinity to Desmidiales than Zygnematales, while the Desmidiales was recovered as monophyletic. The taxonomy of the Zygnematophyceae has been controversial (e.g., reviewed in Gontcharov *et al.,* 2008) at the level of genus due to the wide variability of the morphological characters, and where the apparent uniformity of morphological characters within a genus is not necessarily an indication of monophyly. Consequently, some genera have been found to be paraphyletic, namely *Penium,* or polyphyletic, such as *Mesotaenium, Cylindrocystis, Netrium, Cosmarium,*

Staurastrum, and *Staurodesmus*, in phylogenetic analyses of SSU rRNA and *rbc*L genes (McCourt *et al*., 2000; Gontcharov *et al*., 2003, 2008; Hall *et al.*, 2008). Likewise, the circumscription of subclasses and orders has been debated. Recently, it was proposed that the class Zygnematophyceae includes two subclasses, Zygnematophycidae and Spirogloeophycidae (Cheng *et al*., 2019). The latter group includes *Spirogloea muscicola* (De Bary) Melkonian, the earliest-diverging taxon within Zygnematophyceae. A system of five orders was also proposed based on phylogenetic transcriptomic analyses, namely, Spirogloeales, Serritaeniales, Zygnematales, Spirogyrales, and Desmidiales (Hess *et al.*, 2022). The novel order Serritaeniales includes the *Mougeotiopsis calospora* Palla, which together with *M. endlicherianum* form a clade diverging after Spirogloeales and sister to the remaining Zygnematophyceae.

Several phylogenetic studies of chloroplast and nuclear sequence data have reconstructed Zygnematophyceae as the closest extant group of algae to land plants (Turmel *et al*., 2002, 2006, 2007; Palmer *et al*., 2004; Wodniok *et al*., 2011; Timme *et al*., 2012; Zhong, Xi *et al*., 2013; Zhong, Liu *et al*., 2013; Civan *et al*., 2014; Ruhfel *et al*., 2014; Wickett *et al*., 2014; Lemieux *et al*., 2016; Gitzendanner *et al*., 2018; Cheng *et al*., 2019; Orton *et al*., 2020; Hess *et al.*, 2022). Alternatively, analyses of nuclear data have recovered Zygnematophyceae and Coleochaetophyceae united as a clade and sister to land plants (Wodniok *et al*., 2011; Finet *et al*., 2010, 2012; Laurin-Lemay *et al*., 2012). In contrast, the analyses of combined data sets comprising chloroplast, mitochondrial, and nuclear sequences (Karol *et al*., 2001; Qiu *et al*., 2006), or mitochondrial sequences alone (Turmel *et al*., 2007, 2013; Orton *et al*., 2020) identified Charophyceae as the lineage most closely related to land plants. However, mitochondrial analyses of gene content and gene order, and amino-acid sequence analysis with the removal of fast-evolving sites, disfavoured Charophyceae as sister-group of land plants (Turmel *et al*., 2013). Likewise, nucleotide data analyses (concatenated loci and multispecies coalescent analyses) recovering Charophyceae sister to land plants (Wickett *et al.*, 2014) were poorly supported and suggested as a putative artefact of compositional tree-heterogeneity.

An important cause of incongruence in phylogenetic studies lies in methodological issues. These problems can be roughly split into misassigned data, random error, and model misspecification. The first includes the misidentification of genes as orthologous or wrongly assigned sequences due to contaminants, for instance. The random or stochastic error arises due to the use of a limited sample size, leading to a deviation between a population parameter and an estimate of that parameter (Swofford *et al.*, 1996). It emerges therefore due to the

finite nature of the data, and decrease as the data length approaches infinity. The systematic error occurs when the population parameter deviates from the estimated value of that population parameter due to incorrect assumptions in the inference method (Swofford *et al.*, 1996). Unlike random error, systematic error persists and may even amplify when the data size increases if it is not appropriately modelled by the evolutionary model. Indeed, model misspecification is currently likely the most debated cause of phylogenetic incongruence. Maximum-likelihood (ML) and Bayesian inference (BI) methods require the use of explicit models of the evolutionary process. An amino-acid substitution exchange rate model is typically represented by a 20x20 instantaneous rate matrix and by a vector of the 20 composition frequencies (long-term substitution rate to an amino acid). In a typical phylogenetic analysis, the former matrix remains fixed, while the composition frequencies are optimised from the study data. The substitution exchange rate matrix is typically selected from a set of available commonly-used empirical models. Nevertheless, given the extent of biological and genomic diversity, it seems unreasonable to assume that the empirical models can adequately accommodate all existing sequence data. Indeed, data-specific amino-acid substitution models calculated from simulated and empirical data have been shown to have a better fit to the data and to infer more accurate trees than empirical models (Chapter II; Brazão *et al*., 2023). Additionally, the availability of time-efficient software to calculate sufficiently accurate data-specific amino-acid substitution models is no longer a limitation.

Phylogenetic analysis methods often assume a site-homogeneous model, that is, the same substitution exchange rate matrix and composition frequencies for the entire data set. However, some sites are highly conserved due to strong evolutionary pressures such as functional or structural constraints, whereas other sites evolve rapidly, such as those evolved in protein-protein interactions or those that are located on the surface of the protein (Fraser and Hirsh, 2004; Goldman *et al*., 1998; Koonin *et al*., 2002; Bloom *et al*., 2006; Lynch, 2010). The partitioning of data enables different substitution models to be applied to each partition and thereby more accurately model heterogeneous substitution processes among sites. Partitions can be defined as those encompassing genes, codon positions, or defined by other criteria. For example, phylogenetic analyses of sites partitioned by their relative solvent accessibility (i.e., distinguishing sites on the surface of proteins from sites that are effectively ‘buried’ inside the protein) has shown protein structural constraints to be a source of conflicting phylogenetic signals (Pandey and Braun, 2019). By contrast, site-heterogeneous models do not require *a priori* definition of data partitions. For instance, the mixture CAT model assumes the existence of distinct classes (profiles) of sites based on their similar

equilibrium frequencies (Lartillot *et al*., 2004), thereby accommodating compositional heterogeneity among sites.

The most widely used amino-acid substitution models, including those described above, seek to model the evolutionary process of change among sites of the protein by assuming that the process is stationary, reversible, and homogeneous over time. Stationarity indicates that the marginal site frequencies are constant over time, while reversibility assumes that the evolutionary process is undirected such that the probability of change between two sites is equal in either direction (Bryant *et al*., 2004; Jayaswal *et al*., 2005; Ababneh *et al*., 2006a). Process homogeneity implies that instantaneous substitution rates are constant over time and therefore the same on each branch of the phylogeny. Empirical data often violate these assumptions and the deeper the phylogeny (the older the common ancestor of the taxa in question), the more likely lineage-specific evolutionary processes are to be present. Systematic biases caused by failure to account for time-heterogeneous processes in data can cause correct tree inference to fail (Foster and Hickey, 1999; Jermiin *et al*., 2004; Cox, 2018). Pairwise symmetry tests, namely Bowker's matched-pairs test of symmetry (MPTS; Bowker, 1948; Tavaré, 1986; Ababneh *et al*., 2006b), Stuart's matched-pairs test of marginal symmetry (MPTMS; Stuart, 1955) and Ababneh's matched-pairs test of internal symmetry (MPTIS; Ababneh *et al*., 2006b) assess the stationarity and homogeneity (among lineages) of the evolutionary processes. In each test, a divergence matrix of the aligned amino-acid sequences is computed, and p-distances are then calculated using a χ^2 test with different degrees of freedom. MPTS assesses the assumption of symmetry by computing the distance between the divergence matrix and its transpose, testing the stationarity and homogeneity of the aligned sequences. The MPTMS is a stationary test and assesses site equality between each sequence pair by computing the difference between two sequence frequencies and its variance-covariance matrix. When the variance-covariance matrix is not invertible, MPTMS is not applicable. The MPTIS assesses the homogeneity of the evolutionary processes, according to the difference between MPTS and MPTMS. The χ^2 test for compositional homogeneity also detects non-stationary evolutionary processes (Foster, 2004). This test assesses if the composition of a data partition fits an empirical homogeneous compositional model using a null distribution derived from simulations (the test computes a contingency table of taxon compositions against the mean compositions). Jermiin *et al*. (2020) propose that the assessment of these assumptions should be included in the standard phylogenetic procedure, and when rejected, the upstream (alignment construction) and downstream analyses (tree inference) should be adjusted to accommodate their implications. For instance, tree-

heterogeneity models can be used when the assumption of symmetry (assessed by MPTS) is rejected. Non-homogeneous rate processes can be modelled by using separate substitution models for branches (node-discrete rate heterogeneity (NDRH/NDRH2) model; Foster 2009), whereas non-stationarity composition processes can be modelled using separate composition vectors for each branch (node-discrete composition heterogeneity model (NDCH/NDCH2; Foster 2004; Williams *et al.,* 2020). These models have proven to be effective in phylogenetic analyses of land plants where NDCH analyses of mitochondrial data revealed the influence of compositional bias in the early land plant relationships and incongruence among nuclear and chloroplast data analyses (Liu *et al*., 2014). Recently, mitochondrial data analyses using the NDCH2 recovered the Setaphyta clade, as has been shown in analyses of chloroplast and nuclear data (Sousa *et al*., 2020). The evolutionary process is not uniform and often heterogeneous among land plant taxa. Consequently, the inadequate modelling of the patterns of molecular sequence evolution (i.e., model misspecification) in the data, be it across the tree or among sites, can be a source of systematic bias.

The use of highly parameterised models, such as NDCH2 and CAT, enables the accommodation of compositional tree-heterogeneity in the data; however, other strategies look to decrease data complexity by filtering out the heterogeneous data. Data partitions where stationarity and homogeneity assumptions are rejected can be excluded from further analyses (Naser-Khdour *et al*., 2019); or the tree-heterogeneous sequences can be removed, if not relevant to the phylogenetic issue under study (Jermiin *et al*., 2020). Other approaches target the so-called ‘rogue’ taxon sequences (e.g., Aberer *et al*., 2013). Rogue sequences are taxa without a strong and consistent position in the phylogeny because of ambiguous or poor phylogenetic signals (Wilkinson, 1996), and their removal improves the tree inference accuracy (Aberer *et al*., 2013). Decreasing data complexity can also be accomplished using other methods, such as site recoding strategies (e.g., Susko and Roger, 2007) or removing “noisy” sites according to their observed variability (OV; Goremykin *et al*., 2010). These methods primarily focus on addressing the ‘saturation’ of sites by the multiple substitutions, which due to the loss or reduction of the phylogenetic signal can promote taxon sequences to group based on convergently evolving character states. The OV statistic measures the variability of the substitution rate of sites, allowing for the sorting of sites and subsequent exclusion of the fastest evolving sites, which are putatively the most substitutionally saturated.

Population biological processes can also bias the inference of the species tree, namely, ancestral species polymorphism, introgression or horizontal gene transfer, and gene

duplication after gene losses. Differential sorting of polymorphisms in an ancestral population can cause the gene genealogies to differ among the species (Maddison and Knowles 2006). The sorting of polymorphisms into different descendant lineages of an ancestor is a process known as incomplete lineage sorting (ILS). This discordance between gene trees and species trees can be accounted for by using a multispecies coalescent model, where the probability of genes coalescing is modelled according to the population size (Degnan and Rosenberg, 2006, 2009). Conflicting gene trees explained by ILS are defined as hemiplasy (Avise and Robinson, 2008). Introgression, which can sometimes be hard to distinguish from ILS, is the transfer of genes between species followed by backcrossing, while the process of hybridisation is defined as cross-breeding between different species (Fitch, 1970; Doyle, 1997; Galtier and Daubin, 2008). Gene duplication and gene loss can also lead to gene tree incongruence, but here due to a methodological artefact where paralogous genes are misidentified and analysed as if they are orthologous genes.

In this study, relationships among the Streptophyta and the origin of land plants were investigated using nuclear and organellar data sets that were analysed using data-specific amino-acid substitution models rather than commonly-used empirical models. Amino acids evolve more slowly than the underlying nucleotide data and are deemed more appropriate data to reconstruct phylogenetic relationships of organisms that are as old as land plants (~470MYA) (Cox *et al*., 2014). Phylogenies were inferred under site-homogeneous, site-heterogeneous, tree-heterogeneous, and multispecies coalescent models. The sequence data were also split into partitions according to site-specific evolutionary rates and relative solvent accessibility profiles. Tree-heterogeneous processes were assessed using pairwise symmetry tests and the χ^2 test of compositional homogeneity.

4.2 Methods

Data set assembly

The data sets assembled from nuclear, chloroplast, and mitochondrial data comprised 63 taxa and 409 proteins, 71 taxa and 84 proteins, and 64 taxa and 40 proteins, respectively (Appendix Tables A1, A2, and A3). Each data set included 12 land plant taxa representing the six major lineages, namely bryophytes (hornworts, liverworts, and mosses) and tracheophytes (lycopods, ferns, and seed plants). Because the land plants are well established as a monophyletic group and as the relationships among land plant lineages are not the focus of this study, the number of land plants was kept low thereby decreasing the computation burden while still sampling the main lineages within the clade. The number of included

Zygnematophyceae taxa ranges from 37 to 44, Charophyceae from three to four, and the Coleochaetophyceae included four members. The outgroup included from seven to eight taxa, representing the Klebsormidiophyceae, Chlorokybophyceae, and Mesostigmatophyceae lineages.

Nuclear charophyte and land plant sequences were selected from 410 amino-acid seed alignments (unmasked version alignments; Leebens-Mack *et al.,* 2019). In addition, nuclear genome assemblies of *Clorokybus atmophyticus, Klebsormidium nitens* (Kutzing) Lokhorst, *Mesotaenium endlicherianum* Nägeli, *Mesostigma viride, Spirogloea muscicola, and transcriptome assemblies of Coleochaete orbicularis* Pringsheim, *Nitella mirabilis var. mirabilis* (O.Nordstedt ex J.Groves) R.D.Wood, *and Spirogyra pratensis* Transeau were obtained from NCBI GenBank (Bethesda, USA). The genome sequence assemblies were translated to six-frames as individual open reading frames using esl-translate (Easel miniapps, HMMER vers. 3.3; Eddy, 2011) when protein data were not available. Protein coding sequences were retrieved using HMMsearch with an E-value threshold of 1e-30. HMMsearch (HMMER) identified target sequences by searching protein profiles against the assembled genome or transcriptome sequences. Protein profile models were estimated from the masked alignments of Leebens-Mack *et al.,* (2019) using HMMbuild (HMMER). The resulting protein sequence matches were added to the initial seed alignments and re-aligned using Muscle (vers. 3.8.1551; Edgar, 2004), and assessed with Gblocks (vers. 0.91; Castresana, 2000; Talavera and Castresana, 2007) allowing half gap positions. Alignments were manually inspected to remove misaligned portions or poorly aligned positions. Single-protein trees were constructed for data curation (described below) to identify obvious rogue sequences that were wrongly placed, such as those of contaminants or paralogues. A single protein alignment was discarded because of the absence of charophyte taxa, and the final combined data set comprised 63 taxa, 409 proteins, and 88,394 amino acid sites, with 31.2% of amino acids missing (gaps) from the data set.

Initial chloroplast and mitochondrial protein alignments were assembled using 31 chloroplast and 23 mitochondrial genomes obtained from NCBI GenBank. The protein-coding nucleotide sequences were retrieved from the organelle genomes and aligned and translated using TranslatorX (vers. 1.1; Abascal *et al.,* 2010) with Gblocks allowing half gap positions. Protein profiles were generated from these alignments and used to search and retrieve protein sequences from the transcriptomes and assemblies as described above. The protein sequence matches were added to the initial organellar protein alignments and processed as described above. The chloroplast data set was assembled from 84 proteins,

comprising 17,057 sites (33.7% missing data), while the mitochondrial data set was assembled from 40 proteins comprising 8,008 sites (36.4% missing data). The data products (protein alignments, calculated substitution models, and trees), novel scripts and a machine actionable RO-Crate metadata specification are available from GitHub: https://github.com/joaobrazao/Applying-data-specific-substitution-models-and-mitigating-the-effects-of-among-lineage-heterogeneity.git

The chloroplast genome of the chlorophyte *Geminella terricola* was included in the chloroplast data set because it is wrongly classified as a Klebsormidiophyceae algae, *Interfilum terricola* (GenBank acession: NC_025542)

Estimation of data-specific amino-acid substitution models

The best-fitting commonly-used empirical model for each combined data set was determined using ModelFinder (Kalyaanamoorthy *et al.,* 2017) implemented in IQ-TREE (vers. 1.6.16; Nguyen *et al.,* 2015). Data-specific amino-acid substitution models were estimated from the nuclear, chloroplast, and mitochondrial data sets using ML with a general time-reversible (GTR_{data}) model and four discrete-gamma categories of among-site rate variation ($+\Gamma_4$). Best-fitting empirical models were then compared with data-specific substitution models using the Bayesian information criteria (BIC).

Site-homogeneous composition model analyses of single-protein alignments and combined data sets

Single-protein alignments are often short (<1000 amino-acids) and lack sufficient variation to estimate an accurate amino-acid substitution model. Indeed, substitution models estimated from protein sequences of only 400 sites have been shown to be less accurate than when estimated from longer alignments (Chapter II; Brazão *et al.,* 2023). The means of the length of the single-protein alignments were 216, 203, and 200 sites from the nuclear, chloroplast, and mitochondrial data. Therefore, single-protein alignments were analysed using data-specific models estimated from the combined data sets. Optimal ML trees were inferred from the individual proteins in IQ-TREE using the $GTR_{data}+ \Gamma_4$ and optimised composition frequencies ($+F_{est}$). Bootstrap support (BS) for each node was calculated using 300 replicates: nodes with bootstrap support values <70% were considered poorly supported, while nodes with >80% BS were considered well-supported. Single-protein trees were then analysed according to the gene concordance factor (gCF; Minh *et al.,* 2020a) using the optimal ML inferred from the concatenated data sets, as implemented in IQ-TREE 2 (vers. 2.2.2.6; Minh *et al.,* 2020b). The gCF describes the proportion of inferred single-protein trees that contain

the same branch as the reference tree (e.g. species tree). It measures thus the underlying variance in support of that branch at the gene-level.

Optimal ML trees were inferred from the combined data sets using a $GTR_{data}+\Gamma_4+F_{est}$ using IQ-TREE, and node support was calculated using 300 BS replicates. The BI analyses of the combined data sets using a site-homogeneous model were conducted in MrBayes (vers. 3.2.7a; Ronquist *et al.,* 2012) using fixed exchange rates from the $GTR_{data}+ \Gamma_4$, and composition frequencies modelled by a Dirichlet distribution. MrBayes uses Metropolis coupling and 4 chains, three heated and one 'cold' chain per run. Four independent Bayesian Markov Chain Monte Carlo (MCMC) analyses were run for two and four million generations for the nuclear and organellar data sets respectively. Topological convergence and the number of burnin samples were assessed according to the average standard deviation of split frequencies and by plotting and inspecting the likelihood traces over time. The model parameters were assessed according to the effective sample size (<100 indicates poor mixing behaviour) and the potential scale reduction factor (~1 is recommended). The latter measures were computed using the 'sump' command from MrBayes. After discarding burnin samples, marginal likelihoods were estimated according to the Newton and Raftery (1994) method, as implemented in P4 (vers. 1.3; Foster, 2004), and a majority rule 50% consensus tree was computed from the samples.

Site-heterogeneous composition model analyses of the combined data sets

The nuclear, chloroplast, and mitochondrial data sets were analysed under the site-heterogeneous composition model CAT (Lartillot *et al*., 2004) in Phylobayes (v.1.8mpi; Lartillot *et al.,* 2009) using a $GTR_{data}+ \Gamma_4$ model. The nuclear data set was reduced to 33 taxa and 64 proteins in order to reduce the computational burden. Taxa were selected to include representatives across all major clades and include those taxa with the least amount of missing data (missing data totals 11%). CAT is a Bayesian mixture model that assumes the existence of distinct classes of sites based on amino-acid equilibrium frequencies. Four independent MCMC chains were run in parallel for at least 10,000 generations. Convergence between chains was assessed according to the maximum and mean difference discrepancy across all bipartitions using the bpcomp program (Phylobayes). The convergence of the continuous model parameters was assessed using the tracecomp program (Phylobayes). Marginal likelihoods and majority rule consensus trees were computed as described above.

Estimating rates of site-specific substitutions and removing the fastest-evolving sites

Site-specific evolutionary rates were inferred from the nuclear, chloroplast, and mitochondrial combined data sets using IQ-TREE and the model $GTR_{emp}+ \Gamma_4+F_{emp}$. Site rates were assigned according to the among-site rate variation (ASRV) as the mean value over four discrete-gamma categories weighted by the posterior probability of the site occurring in one of the categories using an empirical Bayesian approach as implemented in IQ-TREE (Mayrose *et al*., 2004). Fast-evolving sites were also determined using the Observed Variability method (OV; Goremykin *et al.*, 2010). OV is calculated for each site as a score of the number of pairwise character-states, where mismatches are scored as 1 and matches as 0. The mean of all comparisons for a given site is used as the site variability score. The effect on the tree topologies of the fastest-evolving sites was assessed by incrementally discarding the fastest sites and re-inferring the optimal ML tree. Sites were removed in segments of 10% until a maximum of 50%. Optimal ML trees were inferred using a $GTR_{data}+ \Gamma_4+F_{est}$ and the node support was calculated using 300 BS replicates. The topological disagreements with the optimal tree inferred from the complete concatenated data set were assessed using the normalised Robinson-Foulds tree metrics (nRF; Robinson and Foulds, 1981; Kupczok et al., 2008).

Data partitioning based on relative solvent accessibility

Sites from the nuclear, chloroplast, and mitochondrial data sets were assigned to their category of solvent accessibility (Pandey and Braun, 2019) using the ACCpro tool from the SCRATCH 1D suite (Cheng *et al*., 2005). ACCpro assigns each amino-acid residue as 'exposed' when the relative solvent accessibility is equal or greater than 25%, or as 'buried' when the relative solvent accessibility is lower than 25%. Structural classifications were determined using weighted consensus sequences of each alignment and then extrapolated to the whole alignment (Pandey and Braun, 2019). Therefore, two partitions comprising the separated buried and exposed sites were assembled for each data set (nuclear, chloroplast, and mitochondrial). The buried-sites partitions and the exposed-sites partitions were then analysed individually using a $GTR_{data}+\Gamma_4+F_{est}$. BS support for each node was calculated using 300 replicates. Topological disagreements with the optimal tree inferred from the complete data set were assessed using the nRF metrics and vi inspection of each tree.

Detection of lineage-specific substitution processes

The extent of time-heterogeneous rates in each of the 409 nuclear, 84 chloroplast, and 40 mitochondrial proteins was assessed using the matched-pairs tests of symmetry, namely,

MPTS, MPTMS, and MPTIS, as implemented in P4. These tests assess a hypothesis of substitution process symmetry of the data, and if the assumption of symmetry is statistically rejected, non-stationary composition and/or non-homogeneous rate processes are supported in the data.

Because the matched-pairs tests of symmetry include multiple simultaneous comparisons, the usual p-value threshold of 5% can lead to false positives. This process is known as the family-wise error rate, and it is usually counteracted by implementing a more stringent threshold than the commonly-used 5% or by adjusting the calculated p-values. Therefore, p-values were adjusted using the false discovery rate Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995; see also Chapter 3), as implemented in the Python package Statsmodels (vers. 0.12.2; with an alpha of 5% and default parameters). The taxon sequences were assigned as tree-heterogeneous sequences when lineage-specific evolutionary processes were detected, i.e., the resulting p-values of the pairwise comparisons were below the 5% threshold. For each single-protein alignment, symmetry test, and p-value adjustment method, the tree-heterogeneous taxa were discarded. The failed pairwise comparisons were ranked from the lowest p-value. Then, the taxon within the lowest p-value comparison with the highest number of failed comparisons was discarded (as well as all p-values associated with this taxon, and therefore, the remaining pairwise comparisons, including the tree-heterogeneous taxon, were no longer considered). This process was reiterated until no more failed pairwise comparisons remained. If two taxa within the same failed pairwise comparison had an equal number of total failed comparisons, both were discarded. After this elimination procedure, the remaining taxa in the single-protein alignments were concatenated (for each of the nuclear, chloroplast, and mitochondrial data, symmetry test, and p-value adjustment method), and optimal ML trees were inferred using a $GTR_{data}+\Gamma_4+F_{est}$ model with node support calculated using 300 BS replicates.

The influence of time-heterogeneous composition, namely non-stationary rates, was also analysed using the χ^2 test for compositional homogeneity for each single protein, as implemented in P4. The null distribution is derived from simulations using a specified tree and model under ML. Each protein was tested using a data-specific exchange rate matrix estimated from the respective combined data set (nuclear, mitochondrial, or chloroplast) and using the corresponding single-protein ML tree. A p-value rejected at significance level indicated that the model composition did not fit the taxon sequence and therefore that tree-heterogeneous sequence evolved under non-stationary conditions. The p-value threshold was defined as 5% and the calculated p-values were to the Benjamini-Hochberg correction

methods. The identified tree-heterogeneous sequences were removed from each individual protein alignment. Optimal ML trees were then estimated from the latter alignments concatenated, as described before.

Tree-heterogeneous composition model analyses of combined data sets

Bayesian analyses using MCMC were computed using the NDCH2 model with each of the nuclear, chloroplast, and mitochondrial data sets. The nuclear data set was reduced to 33 taxa and 64 proteins (11% missing data), and the chloroplast data set to 33 taxa (maintaining all 84 proteins; 5% missing data) in order to reduce the computational burden (the NDCH2 model is highly parameterised with a discrete 20x20 amino acid substitution model applied to each node of the tree). The data sets were reduced according to the same criteria described above for the nuclear site-homogeneous analysis. The NDCH2 parameter values were constrained by a sampled Dirichlet prior on how much the composition vectors may differ from the empirical composition. The exchange rates were fixed using the values from the estimated GTR_{data} models (nuclear, chloroplast, and mitochondrial), and the among-site rate variation modelled by Γ_4. Four replicates were run for 3 million generations for each data set, and the model fit to the data was assessed by posterior predictive simulations of the χ^2 statistic of composition homogeneity, where p-values >= 5% indicate adequate model-fit. Convergence between MCMC chains was assessed by calculating the average standard deviation of split frequencies from each pair of checkpoints of two runs. Marginal likelihoods and majority rule consensus trees were computed as described above.

Multi-species coalescent species tree inference

The nuclear single-protein ML trees were used to compute a coalescent-based species tree in ASTRAL (vers. 5.7.3; Zhang et al., 2018). Posterior probabilities for each branch were computed as the transformation of quartets percentage in gene trees that agree or not with a branch in the inferred species tree. Additionally, coalescent-based species trees were computed using ML trees inferred from single-protein alignments after discarding the tree-heterogeneous sequences, according to the MPTMS.

4.3 Results

Data-specific substitution models calculated from the concatenated protein alignments of each genome fitted the data better than the best-fitting commonly-used models determined using the Modelfinder (Table 4.1). The empirical substitution model LG (Le and Gascuel, 2008) was determined as the best-fitting model for the nuclear data, while the cpREV (Adachi

et al., 2000) model was the best-fitting model for the chloroplast and mitochondrial data sets. That the chloroplast cpREV model was found to be the best-fitting model for the mitochondrial data is a consequence of the absence of a mitochondrion-specific empirical model in Modelfinder. The data-specific models better-fitted the data than the empirical LG and cpREV models by 18,512, 7,423, and 4,841 BIC score units calculated from the nuclear, chloroplast, and mitochondrial data sets respectively (Table 4.1). The optimal ML trees inferred using the data-specific models were 0.5-1.3 substitutions per site longer than the LG and cpREV derived trees (7%, 12%, and 6% longer nuclear, chloroplast, and mitochondrial trees respectively), indicating that the commonly used models underestimated the number of substitutions, although they resulted in the same optimal ML topologies as those inferred using data-specific models (Appendix Figs. A1-A3).

Table 4.1 - Bayesian information criterion scores of the GTR_{data} and best-fitting commonly-used empirical models and lengths (substitutions per site) of the optimal maximum-likelihood trees. The improvement over the commonly-used empirical models is noted in parenthesis. The LG was determined as the best-fitting model for the nuclear data set using ModelFinder, while the cpREV was similarly determined for the chloroplast and mitochondrial data sets. The optimal maximum-likelihood trees were inferred with a discrete gamma-distribution of among-site rate variation, discretised with 4 categories, and using optimised composition frequencies.

Data set	Bayesian information criterion		Optimal ML tree length	
	GTR_{data} (Δ best-fitting empirical model)	Best-fitting empirical model	GTR_{data} (Δ best-fitting model)	Best-fitting empirical model
Nuclear	6877096 (18512)	6895608	18.9 (1.3)	17.6
Chloroplast	765973 (7423)	773397	9 (1.1)	7.9
Mitochondrial	357662 (4841)	362503	8.5 (0.5)	8.0

The topologies inferred from single protein alignments recovered the monophyly of land plants in more than half of the optimal ML trees in the nuclear, chloroplast, and mitochondrial data analyses (54%-65%; tree descriptions for all result trees are provided in Nexus notation (Maddison *et al.,* 1997) in the Supplemental Information). Bryophytes were only resolved as monophyletic in 8% to 15% of the trees, and tracheophytes were monophyletic in 12% to 43% of trees. Between 81% and 90% of the inferred trees resolved Charophyceae algae as a monophyletic group, while the monophyly of Coleochaetophyceae

was observed in 25% to 52% of the trees. Zygnematophyceae was recovered as monophyletic in 12% to 25%. The topology placing Zygnematophyceae as sister-group to land plants was more prevalent in the nuclear (13%) and chloroplast (26%) data analyses than the alternatives placing Charophyceae (11%) or Coleochaetophyceae (8%). However, the proportion of inferred single-protein trees that recovered Zygnematophyceae and land plants within the same clade was low, with a gCF of 9% and 16% for the nuclear and chloroplast data, respectively (Appendix Figs. A1 and A2). Moreover, the mean bootstrap values computed from the nuclear and chloroplast trees were 38 and 39%. Notably, only the nuclear protein '6713' and the chloroplast proteins NdhD and RPOC2 demonstrated support for a clade comprising land plants and charophytes as sister-groups. The analysis of the nuclear protein '6713' proteins recovered a clade composed of Zygnematophyceae and Coleochaetophyceae (BS = 47%) sister-group to land plants (BS = 67%), forming a clade moderately supported (BS = 74%). The tree resulting from the analysis of the NdhD alignment recovered a clade composed of Zygnematophyceae and land plants (BS = 88%). Analyses of the RPOC2 recovered a clade comprising land plants and Coleochaetophyceae (BS = 77%). BI analyses of this alignment using site- and tree-heterogeneous models, CAT and NDCH2, recovered the same topology as the ML analyses, though not well-supported (PP = 0.83-0.84). With respect to the analyses of the mitochondrial proteins, the Charophyceae was recovered as sister-group to land plants in 11 single-protein analyses, while Zygnematophyceae was only recovered in nine. The gCF for the clade that included Charophyceae and land plants calculated from the mitochondrial proteins was 35% (Appendix Fig. A3). However, this clade received weak support in single-protein trees (BS = 13-68%). By contrast, the tree derived from the mitochondrial CcmB protein recovered a clade including land plants and Zygnematophyceae with moderate support (BS = 75%). Nevertheless, the gCF values and bootstrap supports (mean = 47) were higher than those derived from the nuclear and chloroplast data analyses, demonstrating that although the phylogenetic signal is weak in single-protein alignments, likely due to stochastic error, the mitochondrial proteins showed a stronger signal than nuclear and chloroplast data.

The BI (log marginal likelihood, *LML* = -3437794; Fig. 4.1) and the optimal ML tree search (log likelihood, *L* = -3437734; Appendix Fig. A1) analyses of the combined nuclear data using a site-homogeneous data-specific model ($GTR_{data}+\Gamma_4+F_{est}$) inferred the same topology, recovering Zygnematophyceae as the sister-group to land plants with maximum support. The main groups, that is, land plants, Zygnematophyceae, Charophyceae, and Coleochaetophyceae were all recovered as monophyletic (PP = 1.00; BS = 100). Land plants

comprised both the tracheophytes and bryophytes as monophyletic groups. The tracheophytes (PP= 1.0; BS = 100%) included lycopods sister to a clade uniting the ferns and seed plants, whereas bryophytes (PP= 1.0; BS = 92%) included Setaphyta (mosses and liverworts; PP= 1.0; BS = 100%) grouped with hornworts. Within Zygnematophyceae, *S. muscicola* was recovered sister to a clade consisting of the Zygnematales and Desmidiales (PP = 0.84; BS = 100%). The order Zygnematales was recovered as a paraphyletic grade with the Desmidiales having evolved from within their diversity. A well-supported clade (PP = 0.97; BS = 100%) included the zynematalean genera *Mesotaenium* and *Cylindrocystis* (both recovered as polyphyletic) *Mougeotia sp.* and *Zygnemopsis sp.* as the sister-group to the remaining Zynematales and Desmidiales (PP = 0.85; BS = 100%). The remaining Zynematales included three subgroups diverging in the following order: the genus *Spirogyra*, *Netrium digitus* (Brébisson ex Ralfs) Itzigsohn & Rothe united with *Nucleotaenium eifelense* A.A.Gontcharov and M.Melkonian, and *Roya obtusa* (Brébisson) West & G.S.West 1896 united with *Planotaenium ohtanii* Gontcharov & Melkonian 2010. Their placements received strong to maximum support using ML, but weak support from Bayesian inference (PP < 0.95) in most nodes, except the latter clade, *R. obtusa* untied with *P. ohtanii*, which were recovered as the closest Zygnematales algae to the monophyletic Desmidiales (PP = 1.00; BS = 100%). Relationships within the Desmidiales was well-supported but with little relation to taxonomic classification with, for example, the genus *Cosmarium* being highly polyphyletic, with taxa being resolved in four distinct positions in the tree, and the genera *Staurodesmus* and *Penium* being recovered as paraphyletic. The Zygnematophyceae together with land plants were recovered as the sister-group to Coleochaetophyceae, with this combined clade being sister to Charophyceae; all relationships were maximally supported. The Coleochaetophyceae included *Chaetosphaeridium globosum* (Nordstedt) Klebahn sister to the genus *Coleochaete*, while Charophyceae comprised *Nitella* and *Chara* as sister genera, again with maximum support.

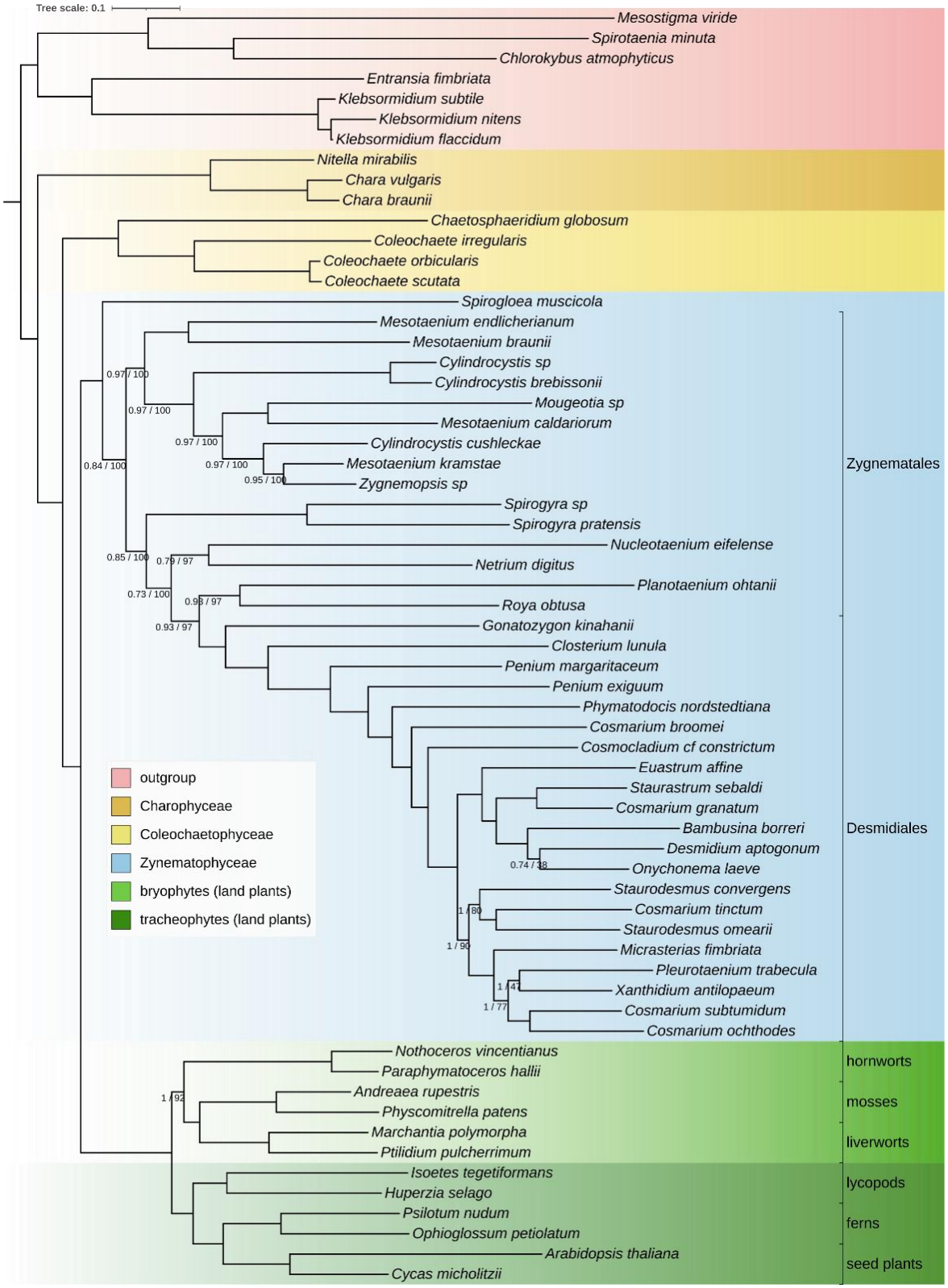


Figure 4.1 - Phylogeny comprising 64 taxa inferred from 409 concatenated nuclear proteins. A majority-rule consensus tree of 10,000 trees was obtained from the posterior distribution of an MCMC analysis with site-homogeneous composition and a data-specific substitution rate matrix ($GTR_{data}+\Gamma_4+F_{est}$). Marginal likelihood L = -3437794. Support values at nodes are Bayesian posterior probabilities and ML bootstrap calculated from 300 replicates using the same model. Support values from nodes fully supported are omitted. The scale bar corresponds to 0.1 substitutions per site.

The MCMC analysis of the reduced nuclear data set (33 taxa and 64 proteins) using the composition site-heterogeneous CAT model recovered Zygnematophyceae sister-group to land plants (PP = 1.0; *LML* = -361946; Appendix Fig. A4). All other nodes were statistically well-supported and congruent with the site-homogeneous trees in the monophyly of the main groups and in most relationships, except that regarding hornworts, which were recovered sister to tracheophytes.

The optimal ML trees inferred from the nuclear data after the incremental removal of the fast-evolving sites assigned according to the ASRV method recovered Zygnematophyceae as sister to land plants as did the site-homogeneous analyses of the original data set (Appendix Figs. A5-A9). The resulting trees had the same topology or very similar to that inferred from the original data set (nRF <= 3%). In each inference, the clade consisting of Zygnematophyceae and land plants, received maximum support, as well as the clade formed by tracheophytes, bryophytes, Zygnematophyceae, Charophyceae, and Coleochaetophyceae. The branch supports improved overall; particularly notable was the increased support for the monophyly of the bryophytes (hornworts as sister to Setaphyta) which rose from 92% in the analysis of the entire data set to 100% BS. The analyses of the reduced data sets using the OV criterion method also inferred Zygnematophyceae as sister-group to land plants with maximum support. Despite the ASRV and OV method trees and the optimal tree inferred from the original data set having a nRF distance varying between 2% and 22%, they were overall congruent with respect to the main groups. The tree rearrangements affected only the placement of taxa within the Desmidiales and Zygnematales (Appendix Figs. A10-A14). In both analyses the Desmidiales was derived from a paraphyletic Zygnematales.

The classification of sites in the nuclear proteins as being either ‘buried’ or ‘exposed’ identified 49,524 and 38,870 sites respectively. The exposed-sites partition contained fewer constant sites (12%) than the buried-sites partition (23%). Optimal ML tree searches with each site category modelled separately in a combined analysis recovered Zygnematophyceae as the sister-group to land plants with maximum BS support (Appendix Figs. A15 and A16). Most phylogenetic relationships were congruent with the analyses of the single-partitioned data (nRF = 2-3%), except for the placement of hornworts inferred from the exposed-sites partition which were recovered as sister to the remaining land plants (BS = 90%; Appendix Fig. A16).

BI (*LML* = -3126678; Appendix Fig. A17) and ML (*L* = -382212; Appendix Fig. A2) analyses of the chloroplast data recovered Zygnematophyceae (PP = 1.0) as sister-group to land plants (PP = 1.0). The main groups, including Zygnematophyceae plus land plants, the

Charophyceae and Coleochaetophyceae, received maximum support, and were overall congruent with those inferred using nuclear data. However, the tracheophyte and bryophyte clades were only supported by 90% and 78% BS respectively in the ML analyses. In addition, the lycopods as sister to the remaining tracheophytes had a support of 89% BS. Within Zygnematophyceae, *S. muscicola* was recovered as sister-group to the remaining members (PP = 1.00; BS = 71%). However, in contrast to nuclear data analyses, the zynematalean clade comprising *M. endlicherianum* and *M. braunii* De Bary were resolved as diverging early from the rest of the Zygnematales and as sister to the paraphyletic grade comprising the remaining Zygnematales and Desmidiales (PP = 1.00; BS = 100%). With respect to Desmidiales, the genera *Staurodesmus* and *Staurastrum* were recovered as polyphyletic (the two genera had more than one representative in the chloroplast data set, in contrast to nuclear and mitochondrial data). The topology inferred using ML showed some nodes poorly supported, and disagreements with the Bayesian analyses within Zygnematophyceae, with taxa in the Desmidiales showing the most notable differences. These results are likely due to the missing data in these taxa which reach the 50% in the chloroplast data set (12% more missing data than in nuclear data).

The composition site-heterogeneous analysis of the combined chloroplast data using the CAT model (*LML* = -320781; Appendix Fig. A18) also recovered Zygnematophyceae sister to land plants and the monophyly of the main groups with maximum support. The resulting tree was overall congruent with the site-homogeneous analyses, except for the placement of some Zygnematales, Desmidiales, and the hornworts. Hornworts were recovered as sister lineage to tracheophytes (PP = 1.00); the same relationship inferred from the nuclear data using the CAT model (Appendix Fig. A4).

The optimal ML trees inferred from the chloroplast data after the incremental removal of the fastest-evolving sites using the ASRV and OV methods did not change the placement of the Zygnematophyceae as sister-group to land plants. Nevertheless, the support for the clade comprising Zygnematophyceae and land plants decreased considerably after removal of 30% of the fastest-evolving sites assigned according to the ASRV criterion (Fig. 4.2A). Similarly, the support for the monophyly of the other main groups decreased as well, although with less severity. The trees inferred from the reduced data sets (10% to 50% removal of sites) exhibited differences of 13%, 18%, 32%, 34%, and 62% compared to the tree derived from the single-partitioned data set, according to the nRF metrics (Appendix Figs. A19-A23). The analyses of the data sets reduced according to OV criterion demonstrated a decline in the support for the clade comprising Zygnematophyceae sister to land plants, starting after the

removal of the first 10% of sites (Fig. 4.2A). The remained support values across the inferred trees also decreased (Appendix Figs. A24-A28). These tree rearrangements had nRF values of 32%, 35%, 46%, 66%, and 62% when trees were compared with that inferred from the entire data set. In the two set of analyses, using the ASRV and OV criteria, differing resolutions of taxa within the Zygnematophyceae clade comprised the majority of the tree rearrangements. However, none of the analyses resolved a monophyletic Zygnematales. The placement of the lycopods was also ambiguous, being recovered either sister to bryophytes or to hornworts. Nevertheless, none of the topological conflicts were statistically well supported, regardless of the number of fast-evolving sites that were removed. In addition, comparing the two methods, the topological rearrangements and node support variation were more severe in the trees derived from the data sets reduced using the OV criterion.

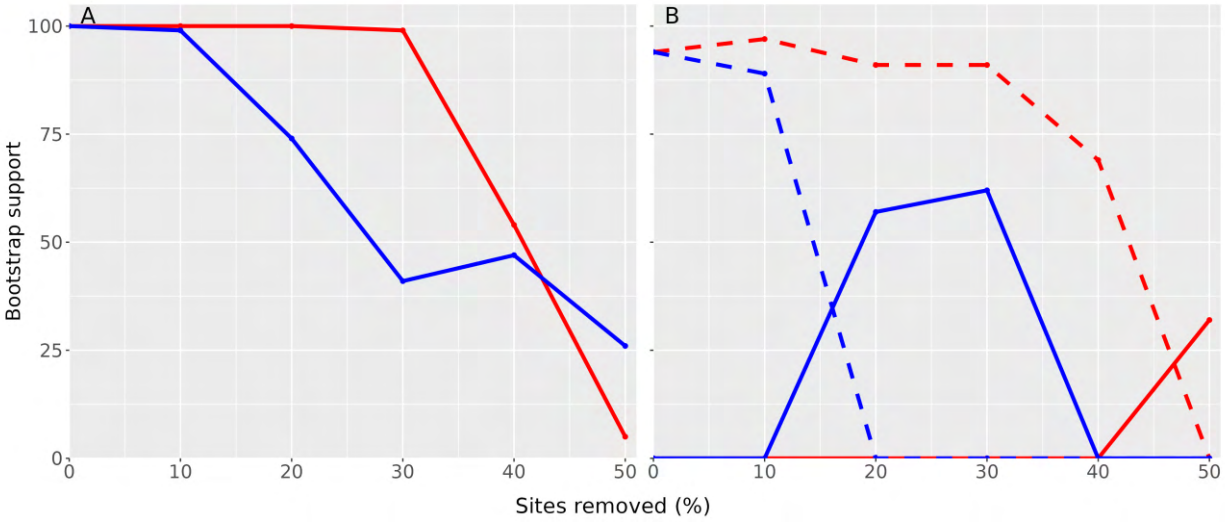


Figure 4.2 - Resolution and support for the clade composed of land plants and the closest charophyte relative. Bootstrap support variation of the chloroplast data (A) and the mitochondrial data (B) after incremental removal of the fastest-evolving sites according to the among-site rate variation (red line) and observed variability (blue line) criteria. The sister-group relationship of Zygnematophyceae and land plants is described by solid lines, while the resolution of Charophyceae as the sister group to land plants is described by dashed lines. Maximum-likelihood analyses were computed using a $GTR_{data}+\Gamma_4+F_{est}$ with 300 bootstrap replicates.

The chloroplast buried- and exposed-sites partitions had 10,630 and 6,423 sites, with 41% and 26% of those sites assigned as constant, respectively. The optimal ML trees inferred of each of these partitions recovered Zygnematophyceae as the sister-group to land plants (BS >= 97%; Appendix Figs. A29 and A30), and the monophyly of the main groups well-supported. The nRF distance to the optimal ML tree inferred from the entire data were of 24% and 15% regarding the trees derived from the buried- and exposed-sites partitions; with the tree rearrangements mainly affecting the placement of taxa within Zygnematophyceae. For

instance, in the optimal tree search using the buried-sites partition the genus *Spirogyra* showed a greater affinity for the earliest-diverging Zygnematales (diverging after *M. endlicherianum* and sister to the remaining Zygnematales; BS = 76%; Appendix Fig. A29) than with taxa that diverged later as observed in the analyses of the entire data set. This placement is congruent with the optimal ML tree inferred from the data set reduced by 20% according to the ASRV criterion (Appendix Figs. A20). Nevertheless, the inferred trees also showed topological rearrangements in land plants. The tree inferred from the buried-sites partition recovered ferns sister-group to a clade uniting lycopods and seed plants, relationships those weakly supported (BS = 56%). By contrast, the exposed-sites partition inferred a well-supported group with lycopods sister to a clade comprising ferns and seed plants (BS = 98%; Appendix Fig. A30); the same topology as that which resulted from the analysis of the entire data set. However, in the same analyses hornworts were sister to the tracheophytes (BS = 94%), although the hornworts plus tracheophyte clade was weakly supported (BS = 49%).

The phylogenetic analyses of the mitochondrial data set using a compositional site-homogeneous model ($GTR_{data}+\Gamma_4+F_{est}$) recovered a clade composed of Charophyceae and land plants (PP = 1.00; BS =94%; Fig. 4.3) using BI (*LML* = -178262) and ML (*L* = -178179, Appendix Fig. A3). The Charophyceae was recovered sister-group to land plants with maximum support. The clades Charophyceae and Zygnematophyceae was well-supported, while the Coleochaetophyceae inferred using ML received moderate support (BS = 80%). Within land plants, bryophytes were recovered as a paraphyletic grade, with hornworts grouped with tracheophytes (PP = 0.91; BS = 59%), and this clade grouped with mosses (PP = 0.99; BS = 64%). The liverworts were recovered as sister-group to the remaining land plant lineages. The tracheophytes included lycopods sister to a clade uniting with ferns and seed plants, with all nodes well-supported. Zygnematophyceae was recovered sister-group to the clade comprising Charophyceae and land plants, and together were recovered sister to Coleochaetophyceae. Within the Zygnematophyceae, *S. muscicola* was recovered as sister to all other Zygnematophyceae algae (PP = 1.00; BS = 94%), a position congruent with the nuclear and chloroplast analyses. Zygnematales was recovered as a paraphyletic grade with the Desmidiales monophyletic and emerging from within their diversity. The order Zygnematales included the polyphyletic genera *Mesotaenium an*d *Cylindocystis,* and together with *Mougeotia sp.* and *Zygnemopsis sp*., formed a clade (PP = 0.99; BS = 58%) sister to the remaining Zynematales and Desmidiales (PP = 0.99; BS = 78%); a topology similar to those inferred using nuclear data. However, in contrast to the nuclear and chloroplast data analyses,

the genus *Spirogyra* was nested within the former clade, namely sister to *M. braunii* (PP = 0.88; BS = 49%). Despite the order Desmidiales exhibiting some topological disagreements with the nuclear data analyses, the overall relationships among the major groups were congruent between the analyses. These specific incongruences could be due to the higher percentage of missing data (46%) within taxa of the Desmidiales which was 8% more than in the nuclear sequences.

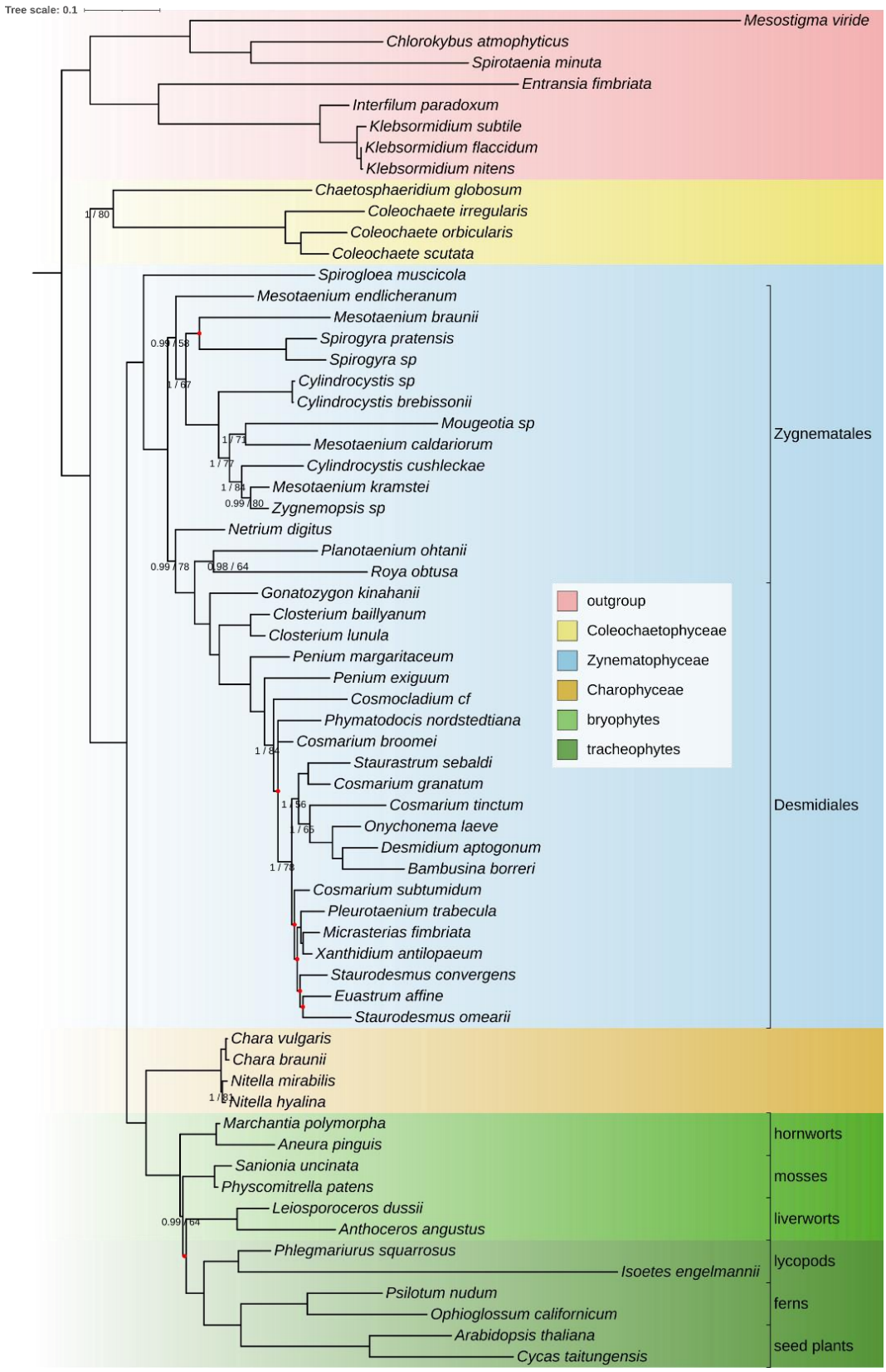


Figure 4.3 - Phylogeny inferred from 40 mitochondrial concatenated proteins, comprising 64 taxa. A majority-rule consensus tree of 10,000 trees was obtained from the posterior distribution of an MCMC analysis with site-homogeneous composition and an amino-acid exchange rate matrix estimated from the data ($GTR_{data}+\Gamma_4+F_{est}$). Marginal likelihood L = -178262. Node supports PP = 1.0 and BS >= 85 are omitted. Nodes poorly supported (PP < 0.95 and BS < 70%) are marked in red. The scale bar corresponds to 0.1

The analyses of the mitochondrial data set using the composition site-heterogeneous CAT model (*LML* = -152814; Appendix Fig. A31) differed from the composition site-homogeneous model analyses in the position of *M. endlicherianum*, which diverged after *S. muscicola*, and sister to the remaining Zygnematophyceae algae (PP = 0.71); this position *M. endlicherianum* being congruent with the chloroplast data analyses. The remaining few topological rearrangements were among taxa within the order Desmidiales.

In the analyses of the mitochondrial data using the data sets reduced according to the ASRV criterion, Zygnematophyceae was grouped with land plants when 50% of the sites were removed, however the resulting clade was poorly supported (BS = 32%; Appendix Fig. A36; Fig. 4.2B). The differences computed using the nRF between the trees inferred from the data sets reduced from 10% to 50% with the ML tree derived from the complete data was 3%, 11%, 13%, 25%, and 43%, with most tree rearrangements affecting poorly supported relationships among taxa in the Desmidiales (Appendix Figs. A32-A36). The main groups were well-supported in the resulting trees, except the clade Zygnematophyceae inferred from the two shortest data sets (BS < 70%). In addition, the Setaphyta were recovered in the analyses using the data set decreased by 20% and 30% of sites. The optimal ML trees inferred from the reduced data sets according to the OV criterion recovered Zygnematophyceae as most closely related to land plants when 20% and 30% of the fastest-evolving sites were removed (BS = 58% and 62%, respectively; Appendix Figs. A38 and A39; Fig. 4.2B). The trees inferred from the data sets reduced from 10% to 50% of sites differed from the tree inferred from the original data set in 13%, 21%, 33%, 48%, and 66% according to the nRF metrics (Appendix Figs. A37– A41). The trees inferred from the two shortest data sets (40 and 50% removed sites) had most nodes poorly supported and many unexpected placements of taxa. Nevertheless, these same analyses recovered Setaphyta, as did all trees inferred from the data sets reduced using the OV criterion.

The mitochondrial buried- and exposed-sites partitions were 5,739 and 2,269 sites long, with 23% and 18% of the sites assigned as constant, respectively. The optimal ML tree inferred from buried-sites partition recovered a clade consisting of Zygnematophyceae and land plants, albeit weakly supported (BS = 56; Appendix Fig. A42). The Zygnematophyceae received moderate support, while the remaining main groups were well-supported. The nRF distance of the optimal tree computed against the tree derived from the entire data set was of 16%. Besides the change in the sister group to land plants, the resulting topological rearrangements were among taxa within the order Zygnematales and the order Desmidiales. For example. The *M. endlicherianum* was recovered diverging after *S. muscicola* and sister to

the remaining Zygnematophyceae algae (BS = 30%), as was similarly found by the CAT model analysis. The optimal ML tree inferred from the exposed-sites partition recovered a clade (BS = 96%) with Charophyceae as most closely related to the land plants (Appendix Fig. A43). The nRF distances computed between this latter tree and the one inferred from the original data set indicated a disagreement of 23%, mostly due to tree rearrangements among poorly supported taxa within Zygnematales and Desmidiales. Nevertheless, the topological differences also included the paraphyly of Coleochaetophyceae with *Chaetosphaeridium globosum* diverging after the genus Coleochaete and sister to remaining ingroup taxa (BS = 90%), and a clade uniting hornworts and mosses (BS = 63%). The CAT model analyses of the buried-sites partition inferred Zygnematophyceae as the most closely related to land plants (PP = 1.0; *LML* = -97934; Appendix Fig. A44), while the exposed sites analysis recovered Charophyceae (PP = 1.0; *LML* = -55930; Appendix Fig. A45). Similarly, the NDCH2 analyses recovered Zygnematophyceae and Charophyceae as most closely related to land plants when inferred from the buried- and exposed-sites partitions respectively (Appendix Figs. A45 and A46). In the four independent analyses of the buried sites, the resulting X^2 p-values indicated the acceptance of the model (>= 0.98). However, the clade comprising Zygnematophyceae and land plants as sister-groups was only well-supported in two of four runs (PP > 0.95). With respect to relationships among the bryophyte, the run with the highest likelihood recovered the Setaphyta clade fully supported (*LML* = -112504; Appendix Fig. A46) while other two runs recovered the same topology not well-supported statistically, while a fourth run recovered it unresolved. The NDCH2 analyses of the exposed-sites partition recovered Charophyceae as most closely related to land plants, and hornworts sister to mosses, with all clades well-supported in the four runs (*LML* = -59373; X^2 p-value = 0.55; Appendix Fig. A47).

In the MPTS and MPTMS analyses the rate of rejection of the null hypothesis was higher for the mitochondrial data (1.4 - 1.8% p-values rejected; Appendix Table A4) than in the chloroplast and nuclear data, with the latter showing the lowest rate (=<0.1 % p-values rejected). By contrast, the MPTIS assumption of rate homogeneous was only rejected in the nuclear data (<0.1%). Overall, the results of the matched-pairs tests indicate greater compositional tree-heterogeneity in mitochondrial data than in chloroplast and nuclear data. The tree-heterogeneous sequences identified in the nuclear data according to the MPTS and MPTIS were less than 0.1% of all sequences, and were identified in the same protein (Fig. 4.4; Appendix Table A4). The MPTMS analyses identified 0.8% composition-heterogeneous

sequences among 49 protein partitions. In the analyses of the chloroplast data, the MPTS identified 0.2% heterogeneous sequences (in two proteins), and 1.5% (in seven proteins) heterogeneous sequences using the MPTMS. The same tests using the mitochondrial data identified 1.7% (10 proteins) and 3% (13 proteins) heterogeneous sequences, respectively. Analyses conducted after the removal of the tree-heterogeneous sequences inferred the same relationship between charophytes and land plants as the analyses of the initial combined data sets, regardless of the symmetry test used. Nevertheless, the analyses of the mitochondrial data set derived after removing the composition-heterogeneous sequences using the MPTMS recovered the Setaphyta clade (BS = 41%; Appendix Fig. A48), whereas the analysis of the chloroplast data using the same methodology (Appendix Fig. A49) showed an improved support for the bryophytes clade (BS = 87%, contrasting with a node support of 78% BS of Appendix Fig. A2). Rejection of the assumption of stationarity, as relative percentage differences among tests, were also higher in the mitochondrial data when analysed using the

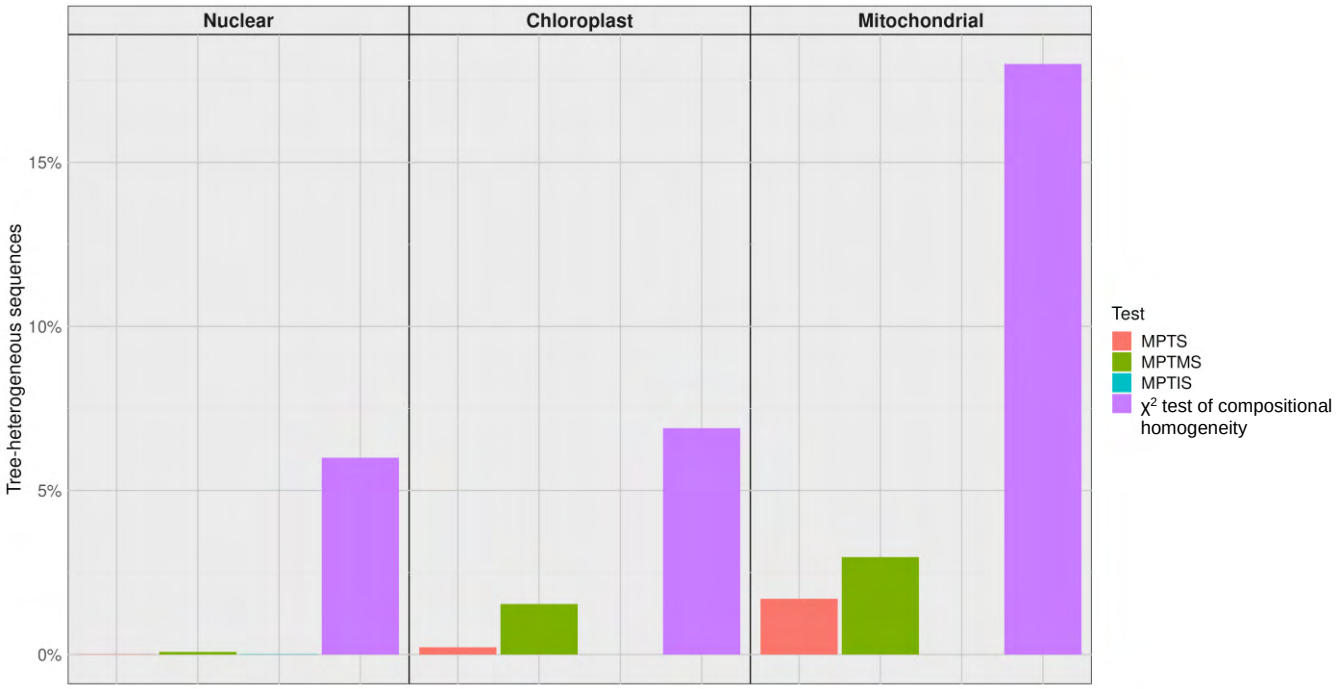


Figure 4.4 - Tree-heterogeneous sequences (%) identified in the nuclear, chloroplast, and mitochondrial single-protein partitions. Sequences were identified using the matched-pairs test of symmetry (MPTS), matched-pairs test of marginal symmetry (MPTMS), matched-pairs test of internal symmetry (MPTIS), and χ^2 test of compositional homogeneity.

χ^2 test for compositional homogeneity among lineages, while the nuclear data showed the lowest rate of rejected p-values (Appendix Table A4). Composition-heterogeneous sequences were identified in 200 nuclear proteins (6.0% of all sequences), 43 chloroplast proteins (6.9%), and 25 mitochondrial proteins (18.0%) (Appendix Table A4; Fig. 4.4). The major relationships between charophyte groups and land plants inferred after removal of the composition-heterogeneous sequences were congruent with those trees inferred from the complete data sets. However, the support for the clade uniting land plants and Charophyceae inferred from the mitochondrial data was lower than in the tree derived from the complete data set analyses (BS = 80%; Appendix Fig. A50). Additionally, the Setaphyta clade (BS = 46%) was recovered in the same tree and not in the analyses of the complete data set. Furthermore, the tree inferred from the chloroplast data after removal of the composition-heterogeneous sequences exhibited a higher support for the monophyly of bryophytes (BS = 87%; Appendix Fig. A51) compared to the analyses of the complete data (BS = 78%).

The MCMC analysis of the combined nuclear data set (reduced to 33 taxa and 64 proteins) using the non-stationary NDCH2 model recovered Zygnematophyceae sister to land plants with all nodes fully supported and congruent with the site-homogeneous analyses (*LML* = -417773; Appendix Fig. A52). Despite the NDHC2 model not fitting the data adequately according to the posterior predictive simulations of χ^2 (X^2 p-value = 0.02), the four independent runs converged with respect to the topology. Similar NDCH2 analyses of the reduced 33 taxon chloroplast data also recovered Zygnematophyceae as the sister group to land plants (*LML* = -322044; Appendix Fig. A53), and model composition was a good statistical fit to the data (X^2 p-value = 0.1). All nodes of the resulting chloroplast NDCH2 analysis tree were fully supported, and the consensus topology was fully congruent with the site-homogeneous tree. In contrast to the NDCH2 analyses of nuclear and chloroplast data, that of the combined mitochondrial proteins recovered Charophyceae as the sister-group to land plants with maximum support (best run *LML* = -174725; Appendix Fig. A54). Although the resulting topologies did not fully converge between the four independent MCMC run replicates (average standard deviation of split frequencies = 0.6), the placement of Charophyceae was congruent across the four analyses. The χ^2 posterior predictive test indicated that the NDCH2 model was an adequate fit to the data (X^2 p-value >= 0.4). The most notable difference compared to the site-homogeneous analyses was within bryophytes, where hornworts were recovered sister to mosses, though this clade was only supported in the run with the highest *LML* (PP = 1.00).

The multispecies coalescent analyses of the nuclear data set recovered a clade composed of Zygnematophyceae and land plants (PP = 0.99; Appendix Fig. A55), as was found with the ML and BI analyses of the same data. The Desmidiales and their closest Zygnematales had some nodes with low support (PP = 0.43-0.93) and some topological disagreements with former analyses. The land plants received a strong to maximum support (PP = 0.99-1.00) in most nodes, except the clade uniting bryophytes (PP = 0.96). The same methodology conducted using the optimal ML trees inferred from single-protein alignments with the composition-sequences (identified using the MPTMS) removed recovered the same topology.

4.4 Discussion

Nuclear, chloroplast, and mitochondrial phylogenies of the Streptophyta were inferred using data-specific amino-acid substitution rate models which were a better-fit to the data than the best-fitting empirical models. The data-specific models have the substitution parameters optimised for the study data and, therefore, are more likely to estimate accurate phylogenies and be less prone to systematic biases that result from using poorer-fitting models. Nevertheless, using only one substitution model, be it data-specific or not, for the entire data set assumes homogeneity of rates among sites and among lineages. Therefore, because possible systematic biases are not known *a priori*, data-specific models were extended with additional parameters to form more complex models, or combined with data-driven strategies, such as partitioning and data exclusion (sequence sub-setting and removal of among-lineage heterogeneous sequences). These methods aim to improve the model fit to the data and to accommodate heterogeneity that could bias the inferred trees. For instance, in the NDCH2 analyses of the mitochondrial data the p-value of the posterior predictive simulations of the χ^2 test was higher for the buried-sites partition than for analysis of the full data. Notably, partitioned analyses of the buried sites also led to topological rearrangements resulting in the emergence of expected clades such as the Setaphyta. These results demonstrate that the model-fit improved when the buried-sites were partitioned and analysed separately, highlighting the advantage of selecting appropriate data for analysis.

The analyses of both the nuclear data set and the chloroplast data set recovered Zygnematophyceae as the closest living algae of land plants as found in previous studies that analysed nuclear data (e.g., Wickett et al., 2014; Leebens-Mack et al., 2019) and chloroplast data (e.g., Lemieux *et al*., 2016; Gitzendanner et al., 2018). However, this result contrast with other studies where analyses of nuclear data recovered a clade comprising Zygnematophyceae

and Coleochaetophyceae as the sister-group sister to land plants probably due to limited taxon sampling (Wodniok *et al*., 2011; Finet *et al*., 2010, 2012; Laurin-Lemay *et al*., 2012). A broad taxon-sampling in the chloroplast analyses, along with the use of data-specific models, likely contribute to finding maximum support for the divergence between Zygnematophyceae and land plants in a ML framework. Previous studies had a robust, but not maximum, support for the same clade but used empirical substitution models, and included fewer taxa (e.g. Lemieux *et al*., 2016). The analyses conducted here also incorporated the use of the composition tree-heterogeneous NDCH2 model, contrasting with many previous studies which only used tree-homogeneous models for the analyses of the combined data (e.g., Gitzendanner *et al.*, 2018; Leebens-Mack *et al.*, 2019; Orton *et al.*, 2020; Hess et al., 2022). The trees computed from the nuclear and chloroplast data sets using the NDCH2 and the analyses of the data sets with composition-heterogeneous sequences removed did not indicate any influence of among-lineage composition heterogeneity in the emergence of land plants. This result contrasts with previous analysis of nucleotide data using all codon positions, which indicated among-lineage heterogeneity as source of systematic bias, as the analyses recovered a clade composed by Charophyceae as sister-group to land plants, albeit weakly supported (Wickett *et al.*, 2014). The nuclear and chloroplast data analyses using site-heterogeneous models, as well as partitioning and data exclusion analyses, also recovered Zygnematophyceae as the sister group to land plants (e.g., Fig. 4.1; Appendix Fig A4, A17, A18, A52, A53). However, most nuclear and chloroplast single-protein trees failed to demonstrate a statistically well-supported relationship between charophytes and land plants, likely due to the lack of information typical in short single-protein partitions (i.e. stochastic error), and thereby indicating the absence of significant incongruence between proteins. The multispecies coalescent analysis of the nuclear data indicated the absence of strong conflict between individual protein trees and the species tree indicating that ILS had little influence on the divergence between charophytes and land plants. Together these analyses demonstrate that the Zygnematophyceae is the most likely sister group to land plants based on phylogenetic signal derived from the nuclear and chloroplast data.

In contrast to the nuclear and chloroplast data, the analyses of the combined mitochondrial data set inferred Charophyceae as the closest charophyte relative to land plants, regardless of whether the data were analysed using site-homogeneous or more complex heterogeneous models (see also Turmel *et al.*, 2013). The congruence between trees resulting from the site-homogeneous, CAT, and NDCH2 model analyses indicated a likely absence of compositional bias among lineages and sites affecting the identity of the sister group of plants.

Moreover, strong support for the sister-group relationship between land plants and Charophyceae in these analyses reinforces the robustness of the inferred phylogeny and suggests that mitochondrial data may carry a different phylogenetic signal from the nuclear and chloroplast data. However, trees inferred from the buried-sites partition - the more slowly evolving sites - and some single protein alignments of the mitochondrial data recovered Zygnematophyceae as the sister-group to land plants. These trees did not show a full reconciliation with the nuclear and chloroplast topologies, as the resulting topologies recovered Charophyceae as the sister-group to the clade composed of Zygnematophyceae and land plants, rather than Coleochaetophyceae. Nevertheless, the observation that at least some partitions of the mitochondrial data are congruent with the well-supported results placing Zygnematophyceae as the sister-group to land plants suggests that the incongruent mitochondrial data placing Charophyceae more closely related to land plants may be subject to other systematic biases not accounted for by the heterogeneous models used in this study. Alternatively, the two phylogenetic signals present in the mitochondrial data could be both real, and perhaps due to evolutionary processes such as horizontal gene transfer (HGT). The transfer of genetic material between streptophyte taxa is a notable feature of their mitochondrial genomes (Richardson and Palmer, 2007; Mower *et al.*, 2012), especially among spermatophytes (e.g., Bergthorsson *et al.*, 2003; 2004; Woloszynska e*t al.*, 2004). Examples of HGT among the green algae have also been recorded, such as in the chlorophyte *Nephroselmis olivacea* (Turmel *et al*., 1999a). Hence despite HGT being present in green plant taxa (Viridiplantae), there has not been any recorded instance of HGT among charophyte mitochondrial genomes. It is notable that in the analyses of the mitochondrial data 11 single-protein trees recovered Charophyceae sister to land plants, while nine placed Zygnematophyceae as the sister-group, indicating that relatively large scale HGT would need to have occurred. These results appear to deviate from most recorded cases of HGT in the mitochondrial genome which usually involve just a single or few genes transfer (Bergthorsson *et al.*, 2003; Won and Renner, 2003; Woloszynska *et al.*, 2004; Davis *et al*., 2005; Kitazaki et al., 2011; Sanchez-Puerta *et al.*, 2011; but see also Bergthorsson *et al.*, 2004). Together the evidence might suggest that while HGT is a possible cause of the conflicting phylogenetic signals within the mitochondrial genome, because of the extent of the conflict among proteins it seems more likely to be a result of unaccounted for systematic bias in the analyses.

The taxon sampling of Zygnematophyceae (36 to 42 taxa) was the largest of any major group of Streptophytes among the data sets analysed here. Previously analyses of nuclear data indicated *S. muscicola* as the earliest-diverging Zygnematophyceae algae using ML and site-

heterogeneous models (Cheng *et al.*, 2019; Hess *et al.*, 2022). The analyses conducted here using BI with site- and tree-heterogeneous models were congruent with the previous findings. Moreover, the analyses of the organellar data recovered the same relationship. Therefore, the robust placement of *S. muscicola* across the three data sets supports its classification within its own taxonomic order (Cheng *et al.*, 2019). By contrast, the position of the clade comprising *M. endlicherianum* and *M. braunii* was not congruent between the analyses presented here, resulting in two main hypothesis: the two taxa form a clade sister to the remaining Zygnematales (e.g., Appendix Fig. A17), or this clade together with other Zygnematales, such as other *Mesotaenium spp.* and the genera *Cylindrocystis*, form a well-supported group within the Zygnematales grade (e.g., Fig. 4.1). The first hypothesis is supported by the analyses of the chloroplast data, with moderate to maximum support, while the nuclear data analyses support the second hypothesis with strong to maximum support. The analyses of the mitochondrial data did not recover the two *Mesotaenium* spp. as a clade, but some of the analyses recovered *M. endlicherianum* sister to the remaining Zygnematales as did some of the analyses of the chloroplast data. As a result of analyses similar to the chloroplast analyses presented here, Hess *et al.*, (2022) proposed that these two taxa be included in their own order (Serritaeniales S.Hess & J.de Vries ord. Nov.). The topological conflict with the trees derived from the nuclear data with regard to the placement of the two *Mesotaenium s*p. suggests that other biases that those investigated here are the cause of the incongruence. This disagreement may be due either to unaccounted for systematic bias in the analyses of either the nuclear data or the organellar data. Although it is usually advisable to use amino-acid sequences for the study of deep phylogenetic relationships, in this case it is possible that nucleotide sequences might be more appropriate for analyses of these shallower relationships.

The analyses conducted here identified phylogenetic conflict between methods and data sets regarding the relationships of bryophytes. Hornworts were not recovered as sister to Setaphyta in the CAT model analyses and exposed-sites partition analyses of the nuclear and chloroplast data and were instead placed sister to embryophytes or to all other land plants. In addition, when the hornworts were identified as sister to Setaphyta (i.e. a monophyletic bryophyte clade) in site-homogeneous analyses the sister-group relationship did not receive maximum support unlike most land plant nodes. These results alone suggest that the non-monophyly of the bryophytes is the correct species tree and that their monophyly in some analyses results from compositional bias among sites. However, after discarding the fastest-evolving sites according to the ASRV criterion the support for the clade uniting hornworts and Setaphyta increased in nuclear and chloroplast data analyses. Exposed sites are expected to

evolve more rapidly than the buried sites which was in part corroborated by the presence of fewer constant sites in the exposed-sites partitions. Consequently, exposed sites are more prone to having multiple substitutions that can obscure the phylogenetic signal. It is possible therefore that topologies derived from the exposed-sites partitions in both the nuclear and chloroplast data analyses were biased by unreliable, substitutional saturated, sites. Moreover, the phylogenies inferred from the buried-sites partitions recovered bryophytes monophyletic with maximum support. In addition, previously published analyses of nucleotide sequences have demonstrated that fast-evolving synonymous substitutions biased the inference of the phylogenetic relationships of hornworts (Sousa *et al.,* 2019). The analyses of both nuclear and chloroplast data sets demonstrated improved support for the placement of hornworts when composition-heterogeneous sequences were removed or analysed using the NDCH2 model (Appendix Fig. A52 and A53). These results also corroborate previous studies where compositional heterogeneity among lineages were shown to affect the inference of Bryophyta monophyletic (Sousa *et al.,* 2019, 2020). Consequently, it is likely that the non-monophyly of bryophytes found with the CAT model analyses of the nuclear and chloroplast data is due to non-stationary composition among lineages.

In contrast to nuclear and chloroplast data analyses, the majority of the analyses of the mitochondrial data did not recover a monophyletic Setaphyta. However, this clade was recovered when composition-heterogeneous sequences identified using the MPTMS or the χ^2 test of compositional homogeneity among lineages were removed. Moreover, the NDCH2 model analyses of the buried-sites partition and sequence sub-setting analyses using OV criterion recovered the Setaphyta as monophyletic. Overall, these results agree with the analyses of Sousa *et al.,* (2020) which support for Setaphyta when inferred from mitochondrial data using the model NDCH2, indicating that this clade is affected by among-lineage composition heterogeneity. In addition, the sub-setting analyses suggest that site variability as measured by the OV criterion may be related to compositional heterogeneity among lineages, and in such cases, the removal of sites with the highest OV score reduces compositional heterogeneity.

Among the three genomes, data from the mitochondrial genome was shown to have the most compositional heterogeneity as determined by the MPTS, MPTMS, and χ^2 test. The χ^2 test and the MPTMS identified the most composition-heterogeneous sequences (3% and 18% respectively) in the mitochondrial data which, after removal from the analyses, resulted in trees that recovered the Setaphyta clade. These results seemingly demonstrate therefore the efficacy of the MPTMS and χ^2 test to correctly identify composition-heterogeneous

sequences. By contrast, the MPTIS appears to be inefficient and lacks ability to identify rate-heterogeneous sequences, as demonstrated previously using simulated data (Chapter III).

4.5 Conclusions

Overall, the resulting nuclear and chloroplast phylogenies showed strong agreement which suggests the absence of discordant phylogenetic signals, be they compositional, structural, or substitution-related regarding the position of Zygnematophyceae as the sister-group to land plants. Analyses of mitochondrial data, on the other hand, show two distinct phylogenetic signals supporting either the Zygnematophyceae or the Charophyceae as the most closely related algal lineage to land plants. This result was robust to potential systematic biases due to either among-lineage composition or rate heterogeneity. The conflict detected in these data likely consist of either distinct phylogenetic signals due to biological processes such HGT or, alternatively, the currently available models are not able to account for all known processes affecting phylogenetic reconstruction in these data. Nuclear and chloroplast data analyses also showed the effect of fast-evolving sites, exposed sites, and compositional heterogeneity in the position and support for the placement of hornworts, and hence, the monophyly of the bryophytes. Similarly, mitochondrial analyses indicated the occurrence of systematic bias in the inference bryophytes relationships, in particular those that affected the monophyly of Setaphyta due compositional heterogeneity among lineages. Reconciling part of the incongruence in the phylogenetic inferences of all three genomes was possible by using better-fitting models and selecting the most appropriate data.

4.6 References

Ababneh, F., Jermiin, L. S., Ma, C., & Robinson, J. (2006a). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, *22*(10), 1225–1231. https://doi.org/10.1093/bioinformatics/btl064

Ababneh, F., Jermiin, L. S., & Robinson, J. (2006b). Generation of the exact distribution and simulation of matched nucleotide sequences on a phylogenetic tree. *Journal of Mathematical Modelling and Algorithms*, *5*(3), 291–308. https://doi.org/10.1007/s10852-005-9017-y

Abascal, F., Zardoya, R., & Telford, M. J. (2010). *TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations*. https://doi.org/10.1093/nar/gkq291

Aberer, A. J., Krompass, D., & Stamatakis, A. (2013). Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Systematic Biology*, *62*(1), 162–166. https://doi.org/10.1093/SYSBIO/SYS078

Adachi, J., Waddell, P. J., Martin, W., & Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution*, *50*(4), 348–358. https://doi.org/10.1007/s002399910038

Avise, J. C., & Robinson, T. J. (2008). Hemiplasy: A new term in the lexicon of phylogenetics. In *Systematic Biology* (Vol. 57, Issue 3, pp. 503–507). Oxford Academic. https://doi.org/10.1080/10635150802164587

Baños, H., Susko, E., Roger, A. (2023). Is Over-parameterization a Problem for Profile Mixture Models?, *Systematic Biology*, syad063, https://doi.org/10.1093/sysbio/syad063

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bergthorsson, U., Richardson, A. O., Young, G. J., Goertzen, L. R., & Palmer, J. D. (2004). Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm Amborella. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51), 17747–17752. https://doi.org/10.1073/pnas.0408336102

Bergthorsson, U., Adams, K. L., Thomason, B., & Palmer, J. D. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, 424(6945), 197–201. https://doi.org/10.1038/nature01743

Bhattacharya, D., Surek, B., Rusing, M., Damberger, S., & Melkonian, M. (1994). Group I introns are inherited through common ancestry in the nuclear-encoded rRNA of Zygnematales (Charophyceae). *Proceedings of the National Academy of Sciences of the United States of America, 91*(21), 9916. https://doi.org/10.1073/PNAS.91.21.9916

Bloom, J. D., Drummond, D. A., Arnold, F. H., & Wilke, C. O. (2006). Structural Determinants of the Rate of Protein Evolution in Yeast. *Molecular Biology and Evolution, 23*(9), 1751–1761. https://doi.org/10.1093/MOLBEV/MSL040

Bowker, A. H. (1948). A Test for Symmetry in Contingency Tables. *Journal of the American Statistical Association*, *43*(244), 572–574. https://doi.org/10.1080/01621459.1948.10483284

Brazão, J. M., Foster, P. G., & Cox, C. J. (2023). Data-specific substitution models improve protein-based phylogenetics. *PeerJ*, *11*, e15716. https://doi.org/10.7717/peerj.15716

Bryant, D., Galtier, N., & Poursat, M.-A. (2004). Likelihood calculation in molecular phylogenetics. *Mathemathics of Evolution and Phylogeny*, Oxford; New York: Oxford University Press. P 33–62. https://doi.org/10.1093/oso/9780198566106.003.0002

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. http://www.ncbi.nlm.nih.gov/pubmed/10742046

Cheng, J., Randall, A. Z., Sweredoski, M. J., & Baldi, P. (2005). SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Research*, *33*(SUPPL. 2). https://doi.org/10.1093/nar/gki396

Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W., Li, X., Xu, Y., Zhang, Y., Wittek, S., Reder, T., Günther, G., Gontcharov, A., Wang, S., Li, L., Liu, X., Wang, J., Yang, H. Melkonian, M., *et al* (2019). Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. *Cell*, *179*(5), 1057-1067.e14. https://doi.org/10.1016/j.cell.2019.10.019

Civan, P., Foster, P. G., Embley, M. T., Séneca, A., & Cox, C. J. (2014). Analyses of charophyte chloroplast genomes help characterize theancestral chloroplastgenomeof land plants. *Genome Biology and Evolution*, *6*(4), 897–911. https://doi.org/10.1093/gbe/evu061

Cocquyt, E., Verbruggen, H., Leliaert, F., & De Clerck, O. (2010). *Evolution and Cytological Diversification of the Green Seaweeds (Ulvophyceae)*. https://doi.org/10.1093/molbev/msq091

Cox, C. J. (2018). Land Plant Molecular Phylogenetics: A Review with Comments on Evaluating Incongruence Among Phylogenies. *Critical Reviews in Plant Sciences*, *37*(2–3), 113–127. https://doi.org/10.1080/07352689.2018.1482443

Cox, C. J., Li, B., Foster, P. G., Embley, T. M., & Civáň, P. (2014). Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Systematic Biology*, *63*(2), 272–279. https://doi.org/10.1093/sysbio/syt109

Cox, C. J., & Foster, P. G. (2013). A 20-state empirical amino-acid substitution model for green plant chloroplasts. *Molecular Phylogenetics and Evolution*, *68*(2), 218–220. https://doi.org/10.1016/j.ympev.2013.03.030

Davis, C. C., Anderson, W. R., & Wurdack, K. J. (2005). Gene transfer from a parasitic flowering plant to a fern. Proceedings of the Royal Society B*: Biological Sciences*, 272(1578), 2237–2242. https://doi.org/10.1098/rspb.2005.3226

De Vries, J., Curtis, B. A., Gould, S. B., & Archibald, J. M. (2018). Embryophyte stress signaling evolved in the algal progenitors of land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(15), E3471–E3480. https://doi.org/10.1073/pnas.1719230115

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of Species Trees with Their Most Likely Gene Trees. *PLOS Genetics*, *2*(5), e68. https://doi.org/10.1371/JOURNAL.PGEN.0020068

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, *24*(6), 332–340. https://doi.org/10.1016/J.TREE.2009.01.009

Doyle, J. J. (1997). Trees within trees: genes and species, molecules and morphology. *Systematic Biology*, 46(3), 537–553. https://doi.org/10.1093/SYSBIO/46.3.53

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113. https://doi.org/10.1186/1471-2105-5-113

Finet, C., Timme, R. E., Delwiche, C. F., & Marlétaz, F. (2010). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, *20*(24), 2217–2222. https://doi.org/10.1016/j.cub.2010.11.035

Finet, C., Timme, R. E., Delwiche, C. F., & Marlétaz, F. (2012). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, *22*(15), 1456–1457. https://doi.org/10.1016/j.cub.2012.07.021

Fitch, WM. (1970) Distinguishing homologous from analogous proteins. Syst Zool. Jun;19(2):99-113. PMID: 5449325.

Foster, P. G. (2004). modelling compositional heterogeneity. *Systematic Biology*, *53*(3), 485–495. https://doi.org/10.1080/10635150490445779

Foster, P. G., & Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, *48*(3), 284–290. https://doi.org/10.1007/PL00006471

Fraser, H. B., & Hirsh, A. E. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evolutionary Biology*, *4*(1), 1–5. https://doi.org/10.1186/1471-2148-4-13

Galtier, N., & Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1512), 4023–4029. https://doi.org/10.1098/RSTB.2008.0144

Gerrath, J. F. (2003). Conjugating green algae and desmids, in Freshwater Algae of North America: Ecology and Classification, eds J. D. Wehr and R. G. Sheath (San Diego: Academic Press), 353–381. https://doi.org/10.1016/B978-012741550-5/50010-6

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., & Soltis, D. E. (2018). Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany*, *105*(3), 291–301. https://doi.org/10.1002/ajb2.1048

Goldman, N., Thorne, J. L., & Jones, D. T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, *149*(1), 445. https://doi.org/10.1093/GENETICS/149.1.445

Gontcharov, A. A., & Melkonian, M. (2010). Molecular phylogeny and revision of the genus Netrium (Zygnematophyceae, Streptophyta): Nucleotaenium gen. nov. *Journal of Phycology*, *46*(2), 346–362. https://doi.org/10.1111/j.1529-8817.2010.00814.x

Goremykin, V. V, Nikiforova, S. V., & Bininda-Emonds, O. R. P. (2010). Automated removal of noisy data in phylogenomic analyses. *Journal of Molecular Evolution*, *71*(5–6), 319–331. https://doi.org/10.1007/s00239-010-9398-z

Gontcharov, A. A. (2008). Phylogeny and classification of Zygnematophyceae (Streptophyta): current state of affairs. *Fottea*, *8*(2), 87–104. https://doi.org/10.5507/fot.2008.004

Gontcharov, A. A., & Melkonian, M. (2004). Unusual position of the genus Spirotaenia (Zygnematophyceae) among streptophytes revealed by SSU rDNA and rbcL sequence comparisons. *Phycologia*, 43(1), 105–113. https://doi.org/10.2216/i0031-8884-43-1-105.1

Gontcharov, A. A., Marin, B., & Melkonian, M. (2003). Molecular phylogeny of conjugating green algae (Zygnemophyceae, Streptophyta) inferred from SSU rDNA sequence comparisons. *Journal of Molecular Evolution*, 56(1), 89–104. https://doi.org/10.1007/s00239-002-2383-4

Graham, L. E., Cook, M. E., & Busse, J. S. (2000). The origin of plants: Body plan changes contributing to a major evolutionary radiation. *Proceedings of the National Academy of Sciences of the United States of America*, 97(9), 4535–4540. https://doi.org/10.1073/pnas.97.9.4535

Graham, L. E., Kaneko, Y. and Renzaglia, K. (1991) 'Subcellular structures of relevance to the origin of land plants (embryophytes) from green algae', *Critical Reviews in Plant Sciences*, 10(4), pp. 323–342. doi: 10.1080/07352689109382315.

Guiry, M.D. and Guiry, G.M. (2019) AlgaeBase. World-Wide Electronic Publication, National University of Ireland, Galway.

Hall, J. D., Karol, K. G., McCourt, R. M., & Delwiche, C. F. (2008). Phylogeny of the conjugating green algae based on chloroplast and mitochondrial nucleotide sequence data. *Journal of Phycology*, *44*(2), 467–477. https://doi.org/10.1111/j.1529-8817.2008.00485.x

Hess, S., Williams, S. K., Busch, A., Irisarri, I., Delwiche, C. F., de Vries, S., Darienko, T., Roger, A. J., Archibald, J. M., Buschmann, H., von Schwartzenberg, K., & de Vries, J. (2022).

A phylogenomically informed five-order system for the closest relatives of land plants. *Current Biology*, 32(20), 4473-4482.e7. https://doi.org/10.1016/j.cub.2022.08.022

Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., Moriyama, T., Ikeuchi, M., Watanabe, M., Wada, H., Kobayashi, K., Saito, M., Masuda, T., Sasaki-Sekimoto, Y., Mashiguchi, K., … Ohta, H. (2014). Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, *5*(May), 2–6. https://doi.org/10.1038/ncomms4978

Irisarri, I., Darienko, T., Pröschold, T., Fürst-Jansen, J. M. R., Jamy, M., & De Vries, J. (2021). Unexpected cryptic species among streptophyte algae most distant to land plants. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1963). https://doi.org/10.1098/RSPB.2021.2168

Jayaswal, V., Jermiin, L. S., & Robinson, J. (2005). Estimation of Phylogeny Using a General Markov Model. *Evolutionary Bioinformatics*, *1*, 117693430500100. https://doi.org/10.1177/117693430500100005

Jermiin, L. S., Lovell, D. R., Misof, B., & Foster, P. G. (2020). Detecting and Visualising the Impact of Heterogeneous Evolutionary Processes on Phylogenetic Estimates. *BioRxiv*. https://doi.org/10.1101/828996

Jermiin, L. S., Catullo, R. A., & Holland, B. R. (2020). A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics. *NAR Genomics and Bioinformatics*, *2*(2), 1–14. https://doi.org/10.1093/nargab/lqaa041

Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J., & Larkum, A. W. D. (2004). The Biasing Effect of Compositional Heterogeneity on Phylogenetic Estimates May be Underestimated, *Systematic Biology*, Volume 53, Issue 4, Pages 638–643, https://doi.org/10.1080/10635150490468648

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*, 587. http://dx.doi.org/10.1038/nmeth.4285

Karol, K. G., McCourt, R. M., Cimino, M. T., & Delwiche, C. F. (2001). The closest living relatives of land plants. *Science*, *294*(5550), 2351–2353. https://doi.org/10.1126/science.1065156

Kitazaki, K., Kubo, T., Kagami, H., Matsumoto, T., Fujita, A., Matsuhira, H., Matsunaga, M., & Mikami, T. (2011). A horizontally transferred tRNACys gene in the sugar beet mitochondrial genome: Evidence that the gene is present in diverse angiosperms and its transcript is aminoacylated. *Plant Journal*, 68(2), 262–272. https://doi.org/10.1111/j.1365-313X.2011.04684.

Koonin, E. V., Wolf, Y. I., & Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature*, *420*(6912), 218–223. https://doi.org/10.1038/nature01256

Kupczok, A., Haeseler, A. Von, & Klaere, S. (2008). An exact algorithm for the geodesic distance between phylogenetic trees. *Journal of Computational Biology*, 15(6), 577–591. https://doi.org/10.1089/cmb.2008.0068

Lartillot, N., Lepage, T., & Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *BIOINFORMATICS APPLICATIONS NOTE*, *25*(17), 2286–2288. https://doi.org/10.1093/bioinformatics/btp368

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. https://doi.org/10.1093/molbev/msh112

Laurin-Lemay, S., Brinkmann, H., & Philippe, H. (2012). Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology*, *22*(15), R593–R594. https://doi.org/10.1016/j.cub.2012.06.013

Le, Q. S., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S. A., Soltis, D. E., Soltis, P. S., Stevenson, D. W., Ullrich, K. K., Wickett, N. J., DeGironimo, L., … Wong, G. K. S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, *574*(7780), 679–685. https://doi.org/10.1038/s41586-019-1693-2

Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F., & De Clerck, O. (2012). Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, *31*(1), 1–46. https://doi.org/10.1080/07352689.2011.615705

Lemieux, C., Otis, C., & Turmel, M. (2007). A clade uniting the green algae Mesostigma viride and Chlorokybus atmophyticus represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biology*, *5*, 2. https://doi.org/10.1186/1741-7007-5-2

Lemieux, C., Otis, C., & Turmel, M. (2016). Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Frontiers in Plant Science*, *7*(MAY2016). https://doi.org/10.3389/fpls.2016.00697

Lewis, L. A., & McCourt, R. M. (2004). Green algae and the origin of land plants. *American Journal of Botany*, *91*(10), 1535–1556. https://doi.org/10.3732/ajb.91.10.1535

Liang, Z., Geng, Y., Ji, C., Du, H., Wong, C. E., Zhang, Q., Zhang, Y., Zhang, P., Riaz, A., Chachar, S., Ding, Y., Wen, J., Wu, Y., Wang, M., Zheng, H., Wu, Y., Demko, V., Shen, L., Han, X., … Yu, H. (2020). Mesostigma viride Genome and Transcriptome Provide Insights into the Origin and Evolution of Streptophyta. *Advanced Science*, *7*(1). https://doi.org/10.1002/advs.201901850

Liu, Y., Cox, C. J., Wang, W., & Goffinet, B. (2014). Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Systematic Biology*, *63*(6), 862–878. https://doi.org/10.1093/sysbio/syu049

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–352. https://doi.org/10.1016/J.TIG.2010.05.003

Maddison, D. R., Swofford, D. L., Maddiso,n W. P. (1997). NEXUS: an extensible file format for systematic information. *Syst Biol*. 46(4):590-621. https://doi.org/10.2307/2413497

Maddison, W., & Knowles, L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30. https://doi.org/10.1080/10635150500354928

Mayrose, I., Graur, D., Ben-Tal, N., & Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Molecular Biology and Evolution*, *21*(9), 1781–1791. https://doi.org/10.1093/molbev/msh194

McCourt, R. M., Delwiche, C. F., & Karol, K. G. (2004). Charophyte algae and land plant origins. In Trends in Ecology and Evolution (Vol. 19, Issue 12, pp. 661–666*). Elsevier Current Trends*. https://doi.org/10.1016/j.tree.2004.09.013

McCourt, R. M., Karol, K. G., Bell, J., Helm-Bychowski, K. M., Grajewska, A., Wojciechowski, M. F., & Hoshaw, R. W. (2000). Phylogeny of the conjugating green algae (Zygnemophyceae) based on rbcL sequences. *Journal of Phycology*, *36*(4), 747–758. https://doi.org/10.1046/j.1529-8817.2000.99106.x

Melkonian, M. (1989). Flagellar apparatus ultrastructure inMesostigma viride (Prasinophyceae). *Plant Systematics and Evolution*, *164*(1), 93–122. https://doi.org/10.1007/BF00940432

Mikhailyuk, T., Lukešová, A., Glaser, K., Holzinger, A., Obwegeser, S., Nyporko, S., Friedl, T., & Karsten, U. (2018). New Taxa of Streptophyte Algae (Streptophyta) from Terrestrial Habitats Revealed Using an Integrative Approach. *Protist*, *169*(3), 406–431. https://doi.org/10.1016/j.protis.2018.03.002

Minh, B. Q., Dang, C. C., Vinh, L. S., & Lanfear, R. (2021). QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. Systematic Biology, 70(5), 1046–1060. https://doi.org/10.1093/sysbio/syab010

Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New Methods to Calculate Concordance Factors for Phylogenomic data sets. *Molecular Biology and Evolution*, *37*(9), 2727–2733. https://doi.org/10.1093/molbev/msaa106

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mix, M. (1972). Die Feinstruktur der Zellwände bei Mesotaeniaceae und Gonatozygaceae mit einer vergleichenden Betrachtung der verschiedenen Wandtypen der Conjugatophyceae und über deren systematischen Wert. *Archiv Für Mikrobiologie*, *81*(3), 197–220. https://doi.org/10.1007/BF00412239

Morris, J. L., Puttick, M. N., Clark, J. W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Yang, Z., Schneider, H., & Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115(10), E2274–E2283. https://doi.org/10.1073/pnas.1719588115

Mower, J. P., Jain, K., & Hepburn, N. J. (2012). The Role of Horizontal Transfer in Shaping the Plant Mitochondrial Genome. *In Advances in Botanical Research* (Vol. 63). Elsevier. https://doi.org/10.1016/B978-0-12-394279-1.00003-X

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., Lanfear, R., & Bryant, D. (2019). The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biology and Evolution, 11*(12), 3341–3352. https://doi.org/10.1093/gbe/evz193

Nedelcu, A. M., Borza, T., & Lee, R. W. (2006). A land plant-specific multigene family in the unicellular Mesostigma argues for its close relationship to Streptophyta. *Molecular Biology and Evolution, 23*(5), 1011–1015. https://doi.org/10.1093/MOLBEV/MSJ108

Newton, M.A., Raftery, A.E. (1994). Approximate BI with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society* B 56:3–48.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Orton, L. M., Fitzek, E., Fitzek, E., Feng, X., Grayburn, W. S., Mower, J. P., Mower, J. P., Liu, K., Liu, K., Zhang, C., Zhang, C., Duvall, M. R., & Yin, Y. (2020). Zygnema circumcarinatum UTEX 1559 chloroplast and mitochondrial genomes provide insight into land plant evolution. *Journal of Experimental Botany*, *71*(11), 3361–3373. https://doi.org/10.1093/JXB/ERAA149

Palmer, J. D., Soltis, D. E., & Chase, M. W. (2004). the Plant Tree of Life: an Overview Oand Some Points of View. *DNA Sequence*, *91*(10), 1437–1445. https://doi.org/10.3732/ajb.91.10.1437

Pandey, A., & Braun, E. L. (2019). Phylogenetic Analyses of Sites in Different Protein Structural Environments Result in Distinct Placements of the Metazoan Root. *Biology* 9(4):64. https://doi.org/10.3390/biology9040064

Petersen, J., Teich, R., Becker, B., Cerff, R., & Brinkmann, H. (2006). The GapA/B gene duplication marks the origin of streptophyta (charophytes and land plants). *Molecular Biology and Evolution*, *23*(6), 1109–1118. https://doi.org/10.1093/molbev/msj123

Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Schneider, H., Pisani, D., & Donoghue, P. C. J. (2018). The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology*, *28*(5), 733-745.e2. https://doi.org/10.1016/j.cub.2018.01.063

Qiu, Y. L., Li, L., Wang, B., Chen, Z., Knoop, V., Groth-Malonek, M., Dombrovska, O., Lee, J., Kent, L., Rest, J., Estabrook, G. F., Hendry, T. A., Taylor, D. W., Testa, C. M., Ambros, M., Crandall-Stotler, B., Duff, R. J., Stech, M., Frey, W., … Davis, C. C. (2006). The deepest divergences in land plants inferred from phylogenomic evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15511–15516. https://doi.org/10.1073/pnas.0603335103

Richardson, A. O., & Palmer, J. D. (2007). Horizontal gene transfer in plants. *Journal of Experimental Botany*, 58(1), 1–9. https://doi.org/10.1093/jxb/erl148

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1–2), 131–147. https://doi.org/10.1016/0025-5564(81)90043-2.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/SYSBIO/SYS029

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., & Burleigh, J. G. (2014). From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*, *14*(1), 1–27. https://doi.org/10.1186/1471-2148-14-23

Sanchez-Puerta, M. V., Abbona, C. C., Zhuo, S., Tepe, E. J., Bohs, L., Olmstead, R. G., & Palmer, J. D. (2011). Multiple recent horizontal transfers of the cox1 intron in Solanaceae and extended co-conversion of flanking exons. *BMC Evolutionary Biology*, 11(1), 1–15. https://doi.org/10.1186/1471-2148-11-277

Sluiman, H. J., Guihal, C., & Mudimu, O. (2008). Assessing phylogenetic affinities and species delimitations in Klebsormidiales (Streptophyta): Nuclear-encoded rDNA phylogenies

and its secondary structure models in Klebsormidium, Hormidiella, and Entransia. *Journal of Phycology*, *44*(1), 183–195. https://doi.org/10.1111/j.1529-8817.2007.00442.x

Sousa, F., Civáň, P., Brazão, J., Foster, P. G., & Cox, C. J. (2020). The mitochondrial phylogeny of land plants shows support for Setaphyta under composition-heterogeneous substitution models. *PeerJ*, *2020*(4), e8995. https://doi.org/10.7717/peerj.8995

Sousa, F., Civáň, P., Foster, P. G., & Cox, C. J. (2020). The Chloroplast Land Plant Phylogeny: Analyses Employing Better-Fitting Tree- and Site-Heterogeneous Composition Models. *Frontiers in Plant Science*, *11*(July), 1–10. https://doi.org/10.3389/fpls.2020.01062

Sousa, F., Foster, P. G., Donoghue, P. C. J., Schneider, H., & Cox, C. J. (2019). Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist*, *222*(1), 565–575. https://doi.org/10.1111/nph.15587

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, *42*(3–4), 412–416. https://doi.org/10.1093/biomet/42.3-4.412

Susko, E., & Roger, A. J. (2007). On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Molecular Biology and Evolution*, *24*(9), 2139–2150. https://doi.org/10.1093/MOLBEV/MSM144

Swofford, D. L. (1996). Phylogenetic inference. *Molecular Systematics, 2nd Ed.*, 407–514.

Talavera, G., & Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology*, *56*(4), 564–577. https://doi.org/10.1080/10635150701472164

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17, 57-86.

Timme, R. E., Bachvaroff, T. R., & Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE*, *7*(1). https://doi.org/10.1371/journal.pone.0029696

Turmel, M., Lemieux, C., Burger, G., Lang, B. F., Otis, C., Plante, I., & Gray, M. W. (1999). The complete mitochondrial DNA sequences of Nephroselmis olivacea and Pedinomonas minor: Two radically different evolutionary patterns within green algae. *Plant Cell*, 11(9), 1717–1729. https://doi.org/10.1105/tpc.11.9.1717

Turmel, M., Ehara, M., Otis, C., & Lemieux, C. (2002). Phylogenetic relationships among streptophytes as inferred from chloroplast small and large subunit rRNA gene sequences. *Journal of Phycology*, *38*(2), 364–375. https://doi.org/10.1046/j.1529-8817.2002.01163.x

Turmel, M., Otis, C., & Lemieux, C. (2006). The chloroplast genome sequence of Chara vulgaris sheds new light into the closest green algal relatives of land plants. *Molecular Biology and Evolution*, *23*(6), 1324–1338. https://doi.org/10.1093/molbev/msk018

Turmel, M., Otis, C., & Lemieux, C. (2007). An unexpectedly large and loosely packed mitochondrial genome in the charophycean green alga Chlorokybus atmophyticus. *BMC Genomics*, *8*(1), 137. https://doi.org/10.1186/1471-2164-8-137

Turmel, M., Otis, C., & Lemieux, C. (2013). Tracing the evolution of streptophyte algae and their mitochondrial genome. *Genome Biology and Evolution*, *5*(10), 1817–1835. https://doi.org/10.1093/gbe/evt135

Wang, H. C., Susko, E., & Roger, A. J. (2019). The Relative Importance of modelling Site Pattern Heterogeneity Versus Partition-Wise Heterotachy in Phylogenomic Inference. *Systematic Biology*, *68*(6), 1003–1019. https://doi.org/10.1093/sysbio/syz021

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., … Leebens-Mack, J. H. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, *111*(45), E4859–E4868. https://doi.org/10.1073/pnas.1323926111

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., & Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology and Evolution*, 4(1), 138–147. https://doi.org/10.1038/s41559-019-1040-x

Wilkinson, M. (1996). Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution*, *13*(3), 437–444. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A025604

Wodniok, S., Brinkmann, H., Glöckner, G., Heidel, A. J., Philippe, H., Melkonian, M., & Becker, B. (2011). Origin of land plants: Do conjugating green algae hold the key? *BMC Evolutionary Biology*, *11*(1), 104. https://doi.org/10.1186/1471-2148-11-104

Woloszynska, M., Bocer, T., Mackiewicz, P., & Janska, H. (2004). A fragment of chloroplast DNA was transferred horizontally, probably from non-eudicots, to mitochondrial genome of Phaseolus. Plant Molecular Biology, 56(5), 811–820. https://doi.org/10.1007/s11103-004-5183-

Won, H., & Rennert, S. S. (2003). Horizontal gene transfer from flowering plants to Gnetum. *Proceedings of the National Academy of Sciences of the United States of America*, 100(19), 10824–10829. https://doi.org/10.1073/pnas.1833775100

Zhang, S. Q., Che, L. H., Li, Y., Dan, L., Pang, H., Ślipiński, A., & Zhang, P. (2018). Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nature Communications*, *9*(1), 1–11. https://doi.org/10.1038/s41467-017-02644-4

Zhong, B., Liu, L., Yan, Z., & Penny, D. (2013). Origin of land plants using the multispecies coalescent model. In *Trends in Plant Science* (Vol. 18, Issue 9, pp. 492–495). Elsevier Current Trends. https://doi.org/10.1016/j.tplants.2013.04.009

Zhong, B., Xi, Z., Goremykin, V. V, Fong, R., Mclenachan, P. A., Novis, P. M., Davis, C. C., & Penny, D. (2013). Streptophyte Algae and the Origin of Land Plants Revisited Using Heterogeneous Models with Three New Algal chloroplast Genomes. *Molecular Biology and Evolution*, 31, Issue 1, Pages 177–183, https://doi.org/10.1093/molbev/mst200

4.7 Appendix

Table A1: Taxon samples included in the nuclear data analyses. For each species, the corresponding NCBI accession number or seed alignment source, number of proteins, and missing characters are shown.

Taxon names	**Accession**	**No. of proteins**	**Missing characters (%)**
Andreaea rupestris	Leebens-Mack et al., 2019	367	14.5
Arabidopsis thaliana	Leebens-Mack et al., 2019	385	6.4
Bambusina borreri	Leebens-Mack et al., 2019	277	44.2
Chaetosphaeridium globosum	Leebens-Mack et al., 2019	261	44.5
Chara braunii	GCA_003427395.1	324	23.1
Chara vulgaris	Leebens-Mack et al., 2019	184	62.5
Chlorokybus atmophyticus	GCA_009103225.1	386	16.2
Closterium lunula	Leebens-Mack et al., 2019	200	62.2
Coleochaete irregularis	Leebens-Mack et al., 2019	351	12.1
Coleochaete orbicularis	SRR1594679	391	30.3
Coleochaete scutata	Leebens-Mack et al., 2019	307	7.5
Cosmarium broomei	Leebens-Mack et al., 2019	199	66.1
Cosmarium granatum	Leebens-Mack et al., 2019	237	47.9
Cosmarium ochthodes	Leebens-Mack et al., 2019	256	32.4
Cosmarium subtumidum	Leebens-Mack et al., 2019	321	51.3
Cosmarium tinctum	Leebens-Mack et al., 2019	303	26.0
Cosmocladium cf	Leebens-Mack et al., 2019	297	30.1
Cycas micholitzii	Leebens-Mack et al., 2019	355	24.5
Cylindrocystis brebissonii	Leebens-Mack et al., 2019	364	15.0
Cylindrocystis cushleckae	Leebens-Mack et al., 2019	370	11.8
Cylindrocystis sp.	Leebens-Mack et al., 2019	356	18.0
Desmidium aptogonum	Leebens-Mack et al.,	268	42.3

	2019		
Entransia fimbriata	Leebens-Mack et al., 2019	358	19.3
Euastrum affine	Leebens-Mack et al., 2019	271	48.5
Gonatozygon kinahanii	Leebens-Mack et al., 2019	335	25.6
Huperzia selago	Leebens-Mack et al., 2019	374	14.2
Isoetes tegetiformans	Leebens-Mack et al., 2019	380	8.4
Klebsormidium flaccidum	Leebens-Mack et al., 2019	392	3.8
Klebsormidium nitens	GCA_000708835.1	354	48.5
Klebsormidium subtile	Leebens-Mack et al., 2019	245	65.6
Marchantia polymorpha	Leebens-Mack et al., 2019	331	25.5
Mesostigma viride	GCA_009746045.1	350	41.1
Mesotaenium braunii	Leebens-Mack et al., 2019	343	24.6
Mesotaenium caldariorum	Leebens-Mack et al., 2019	346	18.3
Mesotaenium endlicherianum	GCA_009602735.1	395	13.3
Mesotaenium kramstae	Leebens-Mack et al., 2019	370	14.0
Micrasterias fimbriata	Leebens-Mack et al., 2019	297	34.1
Mougeotia sp.	Leebens-Mack et al., 2019	339	20.2
Netrium digitus	Leebens-Mack et al., 2019	289	36.6
Nitella mirabilis	SRR486217	382	15.9
Nothoceros vincentianus	Leebens-Mack et al., 2019	326	23.7
Nucleotaenium eifelense	Leebens-Mack et al., 2019	174	69.2
Onychonema laeve	Leebens-Mack et al., 2019	313	28.4
Ophioglossum_petiolatum	Leebens-Mack et al., 2019	348	27.8
Paraphymatoceros hallii	Leebens-Mack et al., 2019	273	28.1
Penium exiguum	Leebens-Mack et al., 2019	310	28.2
Penium margaritaceum	Leebens-Mack et al., 2019	211	57.4

Phymatodocis nordstedtiana	Leebens-Mack et al., 2019	330	25.1
Physcomitrella pattens	Leebens-Mack et al., 2019	340	18.8
Planotaenium ohtanii	Leebens-Mack et al., 2019	295	30.6
Pleurotaenium trabecula	Leebens-Mack et al., 2019	262	46.4
Psilotum_nudum	Leebens-Mack et al., 2019	331	28.2
Ptilidium pulcherrimum	Leebens-Mack et al., 2019	346	26.9
Roya obtusa	Leebens-Mack et al., 2019	312	29.4
Spirogloea muscicola	GCA_009602725.1	357	56.4
Spirogyra pratensis	SRR1594156	385	16.4
Spirogyra sp.	Leebens-Mack et al., 2019	204	56.9
Spirotaenia minuta	Leebens-Mack et al., 2019	224	52.1
Staurastrum sebaldi	Leebens-Mack et al., 2019	287	40.4
Staurodesmus convergens	Leebens-Mack et al., 2019	322	23.6
Staurodesmus omearii	Leebens-Mack et al., 2019	323	24.4
Xanthidium antilopaeum	Leebens-Mack et al., 2019	276	46.7
Zygnemopsis sp.	Leebens-Mack et al., 2019	361	15.0

Table A2: Taxon samples included in the chloroplast data analyses. For each species, the corresponding NCBI accession number, number of proteins, and missing characters are shown.

Taxon name	**Accession**	**No. of proteins**	**Missing characters (%)**
Angiopteris evecta	DQ821119.1	80	0.8
Anthoceros angustus	AB086179.1	78	3.5
Arabidopsis thaliana	NC_000932.1	74	9.1
Bambusina borreri	ERS1830161	53	40.8
Chaetosphaeridium globosum	NC_004115.1	80	0.0
Chara braunii	AP018555	78	1.8
Chara vulgaris	NC_008097.1	78	1.7
Chlorokybus atmophyticus	NC_008822.1	78	2.4
Closterium baillyanum	NC_030314.1	80	0.0
Closterium lunula	ERS1830159	32	70.0
Coleochaete irregularis	ERS368235	9	86.1
Coleochaete orbicularis	SRR1594679	24	74.6
Coleochaete scutata	NC_030358.1	72	6.9
Cosmarium botrytis	NC_030357.1	80	0.0
Cosmarium broomei	ERS1830162	48	57.3
Cosmarium granatum	ERS1830163	68	28.1
Cosmarium ochthodes	ERS368244	36	69.9
Cosmarium subtumidum	ERS1830165	45	55.2
Cosmarium tinctum	ERS1830166	40	70.0
Cosmocladium cf	ERS3670391	46	71.1
Cycas taitungensis	NC_009618.1	78	1.6
Cylindrocystis brebissonii	ERS1830179	78	1.7
Cylindrocystis cushleckae	ERS368241	2	98.3
Cylindrocystis sp.	ERS3670392	46	63.7
Desmidium aptogonum	ERS1830160	67	33.7
Entransia fimbriata	NC_030313.1	74	3.8
Euastrum affine	ERS1830167	54	41.2
Gonatozygon kinahanii	ERS1830183	8	92.3
Huperzia lucidula	NC_006861.1	80	0.3
Interfilum paradoxum	ERS1830152	43	57.1

Interfilum terricola/ Geminella terricola	NC_025542	62	23.0
Isoetes flaccida	NC_014675.1	76	3.8
Klebsormidium flaccidum	NC_024167.1	70	7.1
Klebsormidium nitens	DF238762.1.J	69	8.0
Klebsormidium subtile	ERS368238	5	93.3
Leiosporoceros dussii	NC_039750.1	79	1.4
Marchantia polymorpha	NC_001319.1	77	2.2
Mesostigma viride	NC_002186.1	77	4.3
Mesotaenium braunii	ERS1830176	20	78.1
Mesotaenium caldariorum	ERS1830177	4	96.9
Mesotaenium endlicherianum	NC_024169.1	79	1.2
Mesotaenium kramstei	ERS1830178	27	69.5
Micrasterias fimbriata	ERS1830168	7	88.4
Mougeotia sp.	ERS368242	20	81.1
Netrium digitus	NC_030356.1	80	0.1
Nitella hyalina	KX306884.1	77	3.4
Nitella mirabilis	SRR486217	69	20.2
Nucleotaenium eifelense	ERS1830180	2	66.1
Onychonema laeve	ERS1830169	47	57.4
Penium exiguum	ERS1830182	69	29.6
Penium margaritaceum	ERS368250	55	40.8
Phymatodocis nordstedtiana	ERS1830170	9	85.9
Physcomitrella patens	NC_005087.2	77	3.0
Planotaenium ohtanii	ERS1830181	32	62.0
Pleurotaenium trabecula	ERS1830171	67	22.9
Psilotum nudum	AP004638	75	8.1
Ptilidium pulcherrimum	HM222519.1	76	8.2
Roya anglica	NC_024168.1	79	0.2
Sanionia uncinata	NC_025668.1	76	3.6
Spirogloea muscicola	GCA_009602725.1	68	9.1
Spirogyra maxima	NC_030355.1	80	0.1
Spirogyra pratensis	SRR1594156	69	10.5
Spirogyra sp.	ERS368243	71	13.8

Spirotaenia minuta	ERS368249	34	63.8
Staurastrum punctulatum	NC_008116.1	80	0.0
Staurastrum sebaldi	ERS1830172	64	41.0
Staurodesmus convergens	ERS1830173	6	93.2
Staurodesmus omearii	ERS1830174	34	69.6
Xanthidium antilopaeum	ERS1830175	46	58.5
Zygnema circumcarinatum	NC_008117.1	79	0.6
Zygnemopsis sp	ERS3670390	43	21.4

Table A3: Taxon samples included in the mitochondrial data analyses. For each species, the corresponding NCBI accession number, number of proteins, and missing characters are shown.

Taxon name	**Accession**	**No. of proteins**	**Missing characters (%)**
Aneura pinguis	NC_026901	35	12.2
Anthoceros angustus	NC_037476	20	35.79
Arabidopsis thaliana	NC_037304	27	19.13
Bambusina borreri	ERS1830161	17	54.9
Chaetosphaeridium globosum	NC_004118.1	35	10.99
Chara braunii	AP018556.1	38	9.95
Chara vulgaris	NC_005255.1	37	4.85
Chlorokybus atmophyticus	NC_009630.1	37	6.31
Closterium baillyanum	NC_022860.1	38	3.1
Closterium lunula	ERS1830159	34	18.36
Coleochaete irregularis	ERS368235	13	76.57
Coleochaete orbicularis	SRR1594679	12	81.38
Coleochaete scutata	NC_045180	35	9.68
Cosmarium broomei	ERS1830162	14	80.77
Cosmarium granatum	ERS1830163	29	35.29
Cosmarium subtumidum	ERS1830165	17	53.81
Cosmarium tinctum	ERS1830166	27	34.77
Cosmocladium cf	ERS3670391	17	66.07
Cycas taitungensis	NC_010303	36	5.34
Cylindrocystis brebissonii	ERS1830179	11	74.98
Cylindrocystis cushleckae	ERS368241	21	61.8
Cylindrocystis sp.	ERS3670392	17	63.01
Desmidium	ERS1830160	25	44.22

aptogonum			
Entransia fimbriata	NC_022861.1	33	12.44
Euastrum affine	ERS1830167	29	24.09
Gonatozygon kinahanii	ERS1830183	23	52.21
Interfilum paradoxum	KP165386.1	29	24.14
Isoetes engelmannii	DF238763.1	23	28.9
Klebsormidium flaccidum	NC_039751	34	9.07
Klebsormidium nitens	NC_001660	31	14.27
Klebsormidium subtile	NC_037508	24	35.64
Leiosporoceros dussii	NC_008240.1	23	30.71
Marchantia polymorpha	ERS1830176	38	6.17
Mesostigma viride	ERS1830177	35	13.47
Mesotaenium braunii	ERS1830176	7	88.99
Mesotaenium caldariorum	ERS1830177	27	35.59
Mesotaenium endlicheranum	ERS368245	37	11.99
Mesotaenium kramstei	ERS1830178	11	75.12
Micrasterias fimbriata	ERS1830168	22	38.45
Mougeotia sp.	ERS368242	24	42.21
Netrium digitus	ERS368247	6	89.69
Nitella hyalina	NC_017598.1	37	6.62
Nitella mirabilis	SRR486217	31	29.73
Onychonema laeve	ERS1830169	13	73.3
Ophioglossum californicum	NC_030900	36	8.44
Penium exiguum	ERS1830182	20	50.51
Penium margaritaceum	ERS368250	19	64.3
Phlegmariurus squarrosus	NC_017755	34	13.87

Phymatodocis nordstedtiana	ERS1830170	18	65.48
Physcomitrella patens	NC_007945	38	2.52
Planotaenium ohtanii	ERS1830181	31	27.3
Pleurotaenium trabecula	ERS1830171	27	28.97
Psilotum nudum	NC_030952	24	40.2
Roya obtusa	NC_022863.1	38	3.18
Sanionia uncinata	NC_027974	38	2.34
Spirogloea muscicola	GCA_009602725.1	31	34.53
Spirogyra pratensis	SRR1594156	33	15
Spirogyra sp.	ERS368243	16	53.51
Spirotaenia minuta	ERS368249	3	92.02
Staurastrum sebaldi	ERS1830172	23	57.48
Staurodesmus convergens	ERS1830173	21	52.46
Staurodesmus omearii	ERS1830174	18	54.77
Xanthidium antilopaeum	ERS1830175	31	21.99
Zygnemopsis sp	ERS3670390	29	35.7

Table A4: P-values lower than 5% and tree-heterogeneous sequences. Percentages of p-values < 5% and tree-heterogeneous sequences and number of proteins that evolved under tree-heterogeneous processes across the nuclear, chloroplast, and mitochondrial data sets. P-values are derived from the matched-pairs tests of symmetry and χ^2 test for compositional homogeneity coupled with the Benjamini-Hochberg and Bonferroni correction procedures and without correction.

Data	**P-values < 0.05 (%)** **(Tree-heterogeneous sequences %; Proteins N)**			
	MPTS	**MPTMS**	**MPTIS**	**χ^2 test of Compositional homogeneity**
Nuclear	<0.1 (<0.1; 1)	0.1 (0.8; 49)	<0.1 (<0.1; 1)	6.0 (6.0; 200)
Chloroplast	<0.1 (0.2; 2)	0.6 (1.5; 7)	0 0	6.9 (6.9; 43)
Mitochondrial	1.8 (1.7; 10)	1.4 (3.0; 13)	0 0	18.0 (18.0; 25)

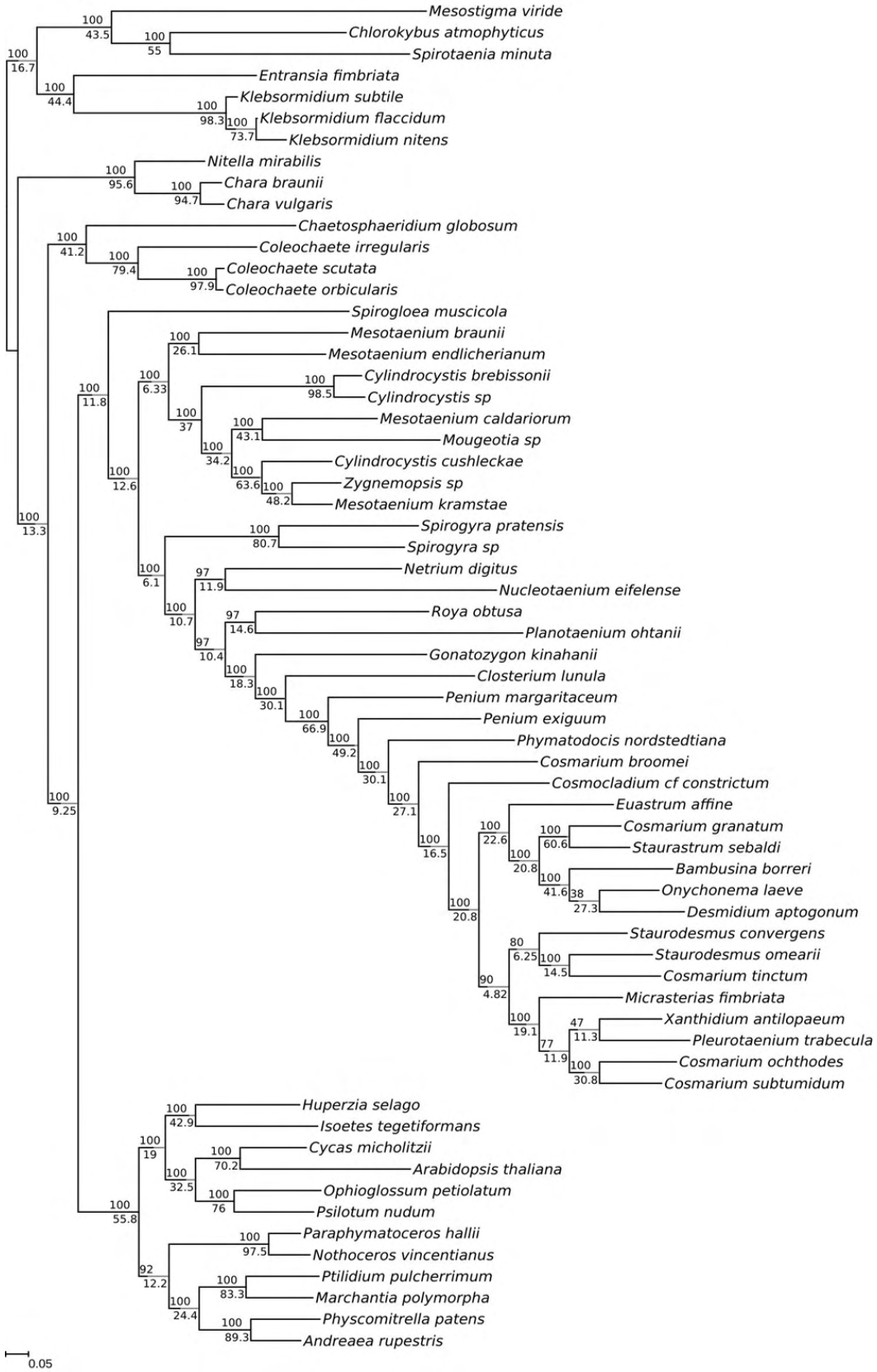


Figure A1 - Optimal maximum-likelihood tree comprising 63 taxa reconstructed from 409 concatenated nuclear proteins. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -3437734. Support values at nodes (top) are maximum-likelihood bootstraps calculated from 300 replicates using the same model, while gene concordance factors, calculated using the single-protein trees, are denoted below.

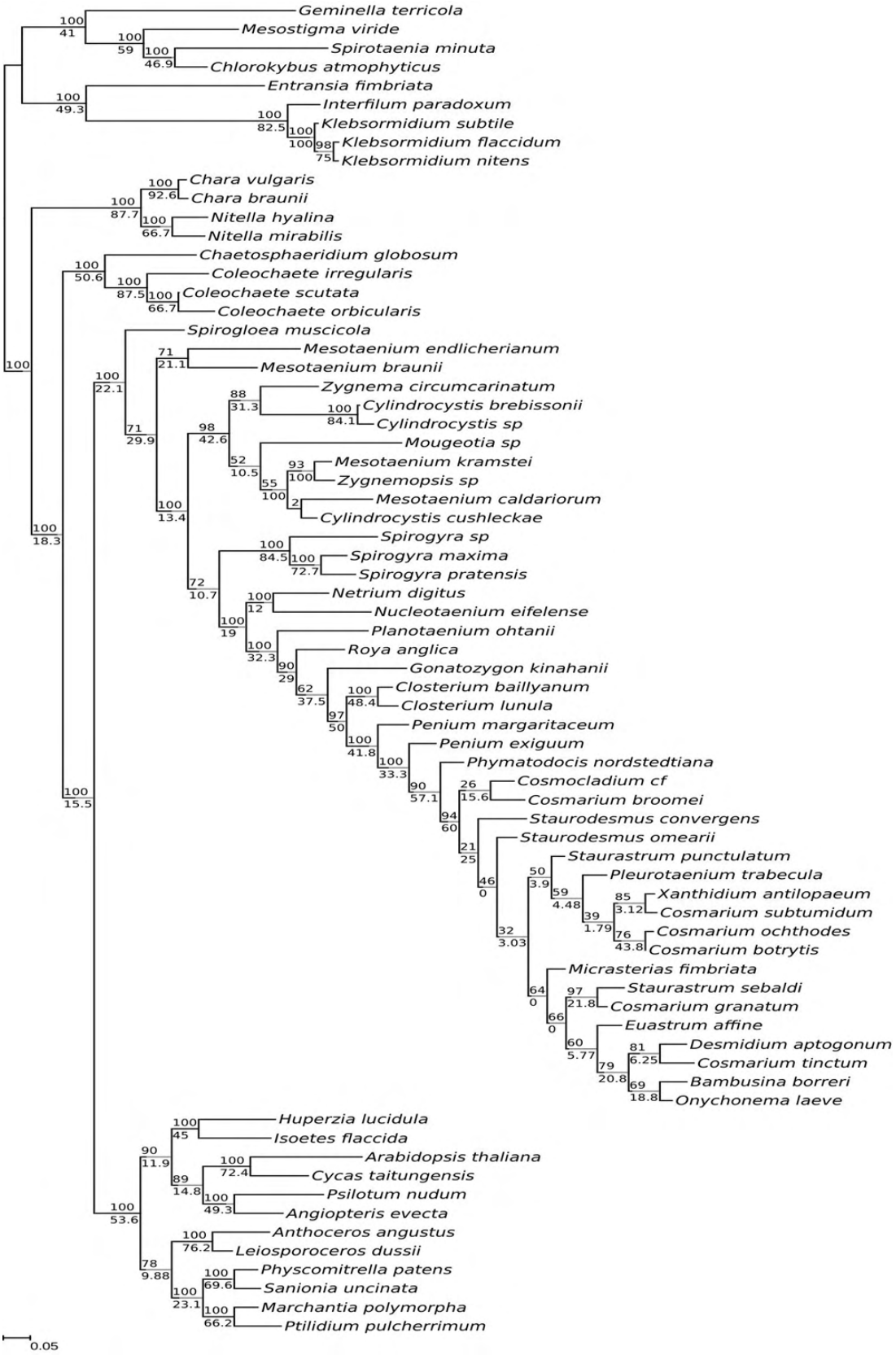


Figure A2 - Optimal maximum-likelihood tree comprising 71 taxa reconstructed from 84 concatenated chloroplast proteins. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -382212. Support values at nodes (top) are maximum-likelihood bootstraps calculated from 300 replicates using the same model, while gene concordance factors, calculated using the single-protein trees, are denoted below.

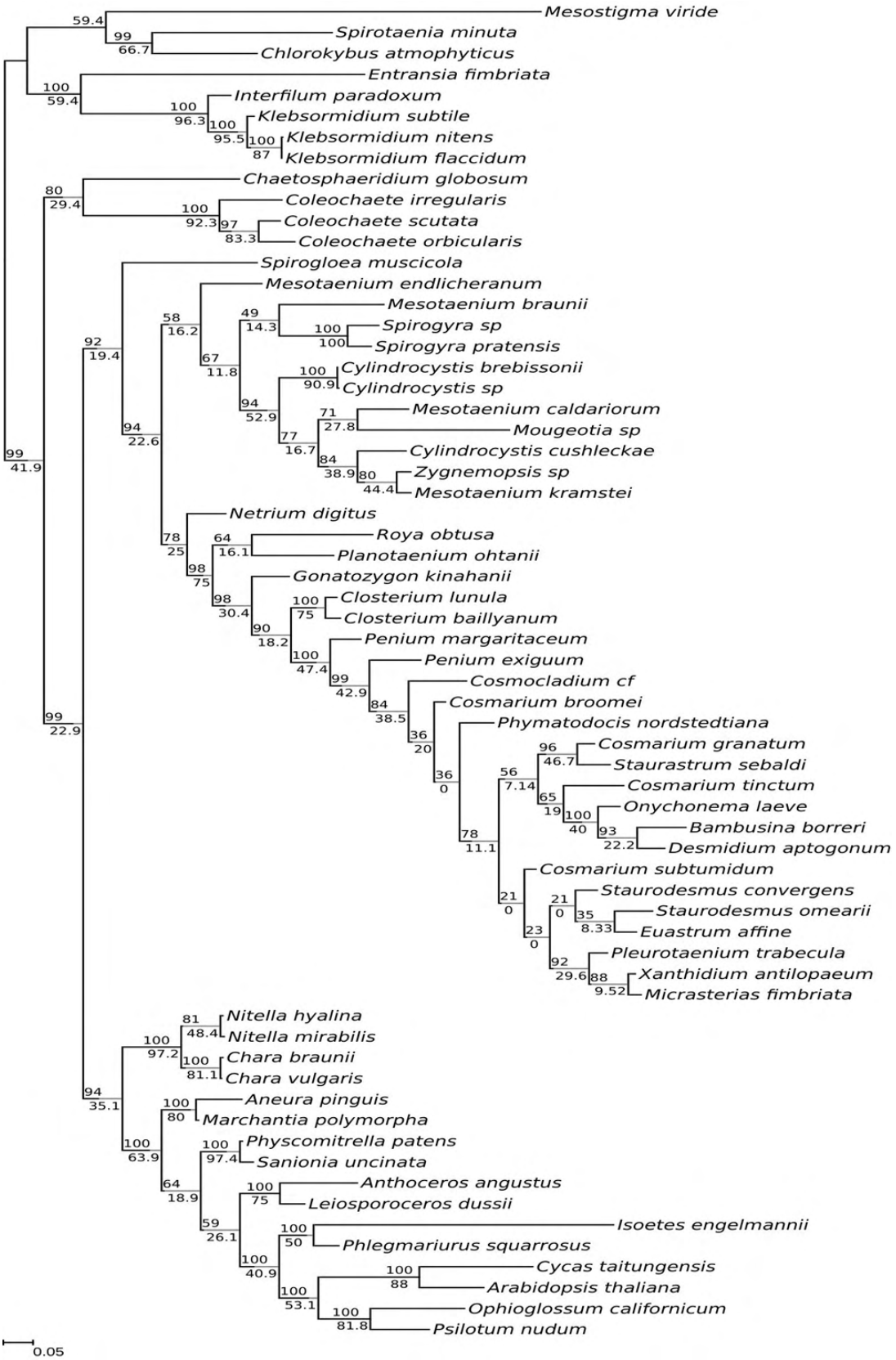


Figure A3 - Optimal maximum-likelihood tree comprising 64 taxa reconstructed from 40 concatenated mitochondrial proteins. Tree was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -178179. Support values at nodes (top) are maximum-likelihood bootstraps calculated from 300 replicates using the same model, while gene concordance factors, calculated using the single-protein trees, are denoted below.

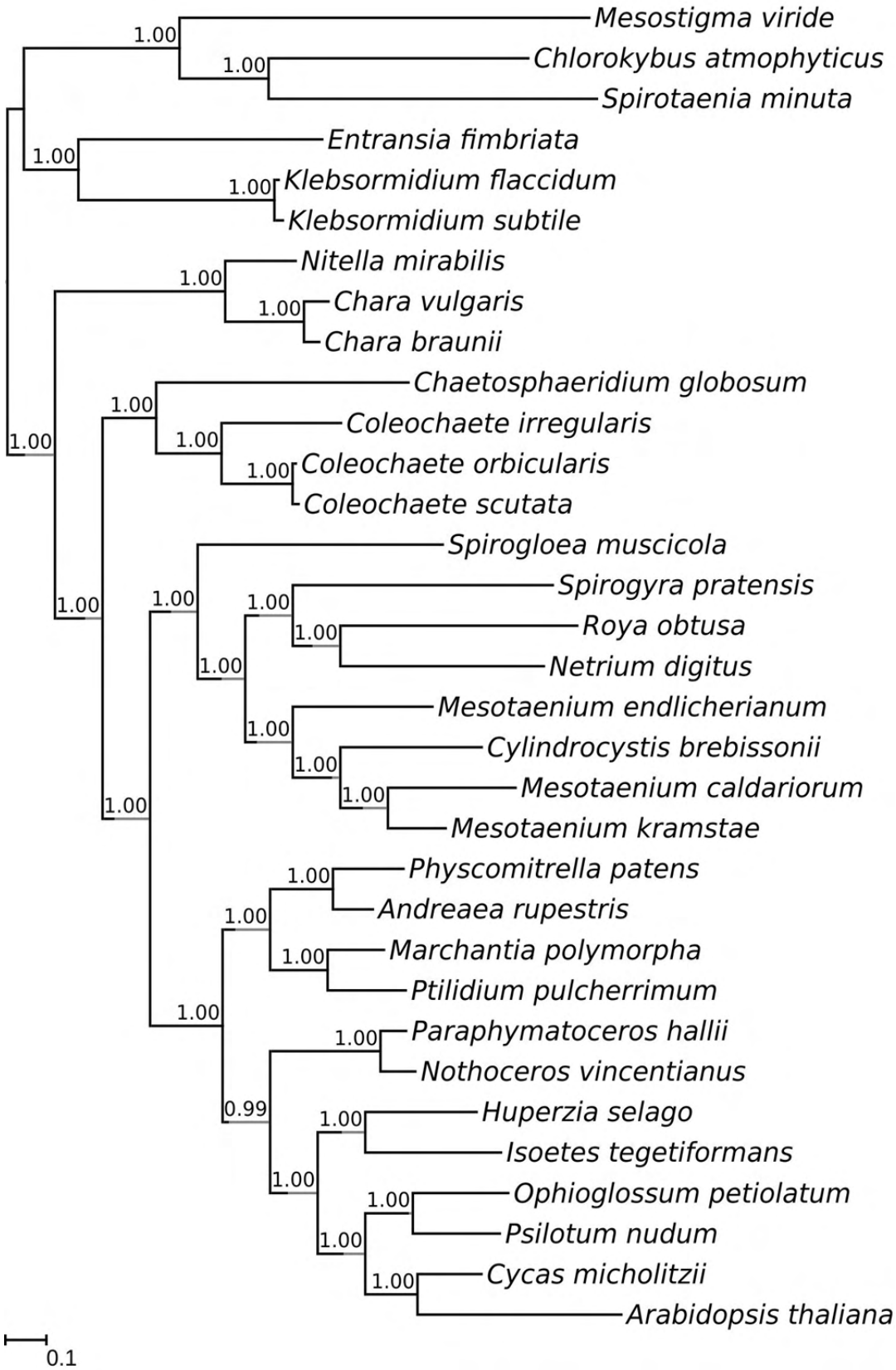


Figure A4 - Phylogeny comprising 33 taxa inferred from 64 concatenated nuclear proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of a MCMC analysis using the CAT model and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{CAT}$). Marginal likelihood, *LML* = -361946. Node support values are Bayesian posterior probabilities.

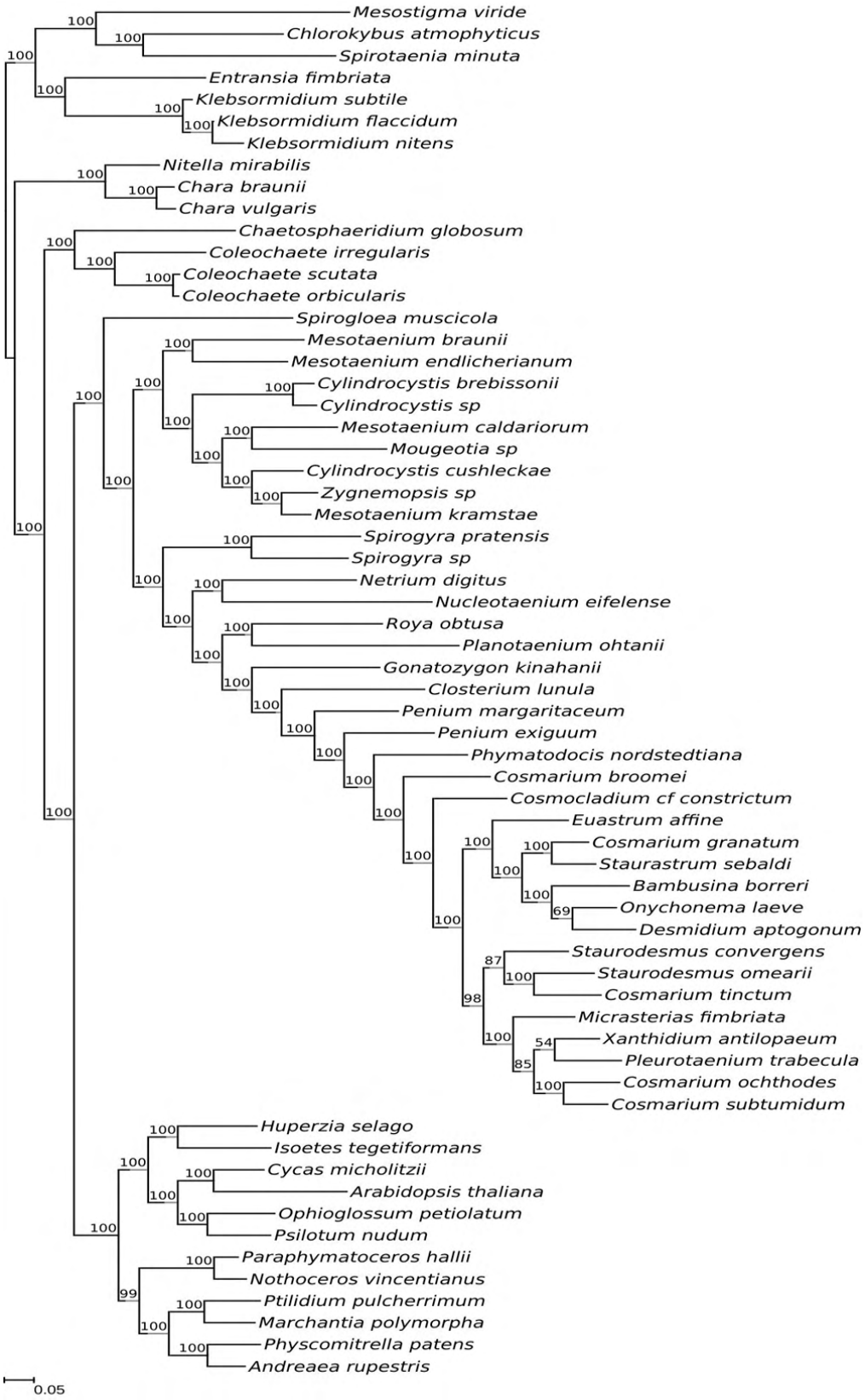


Figure A5 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 10% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2609234. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

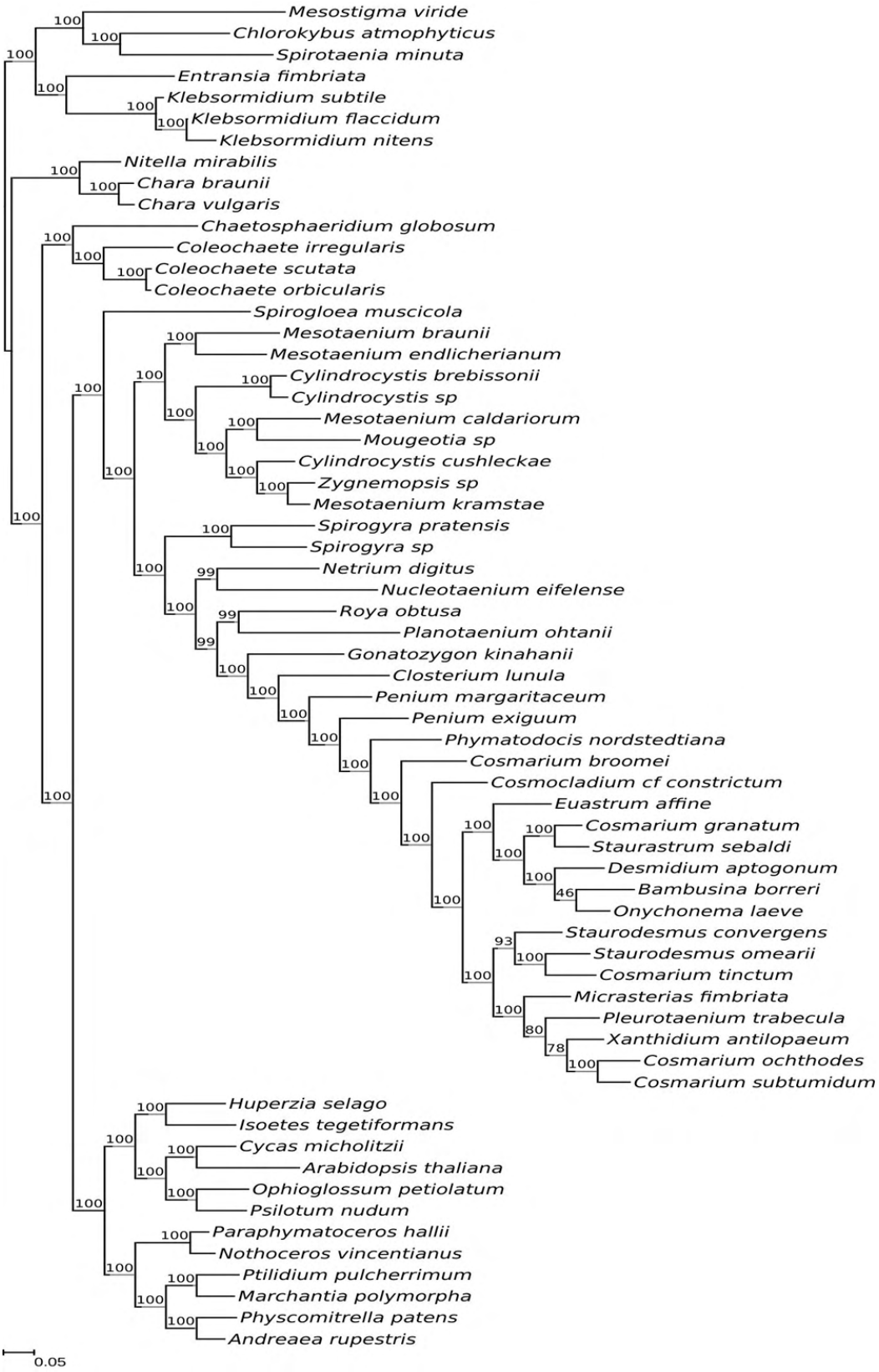


Figure A6 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 20% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2012555. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

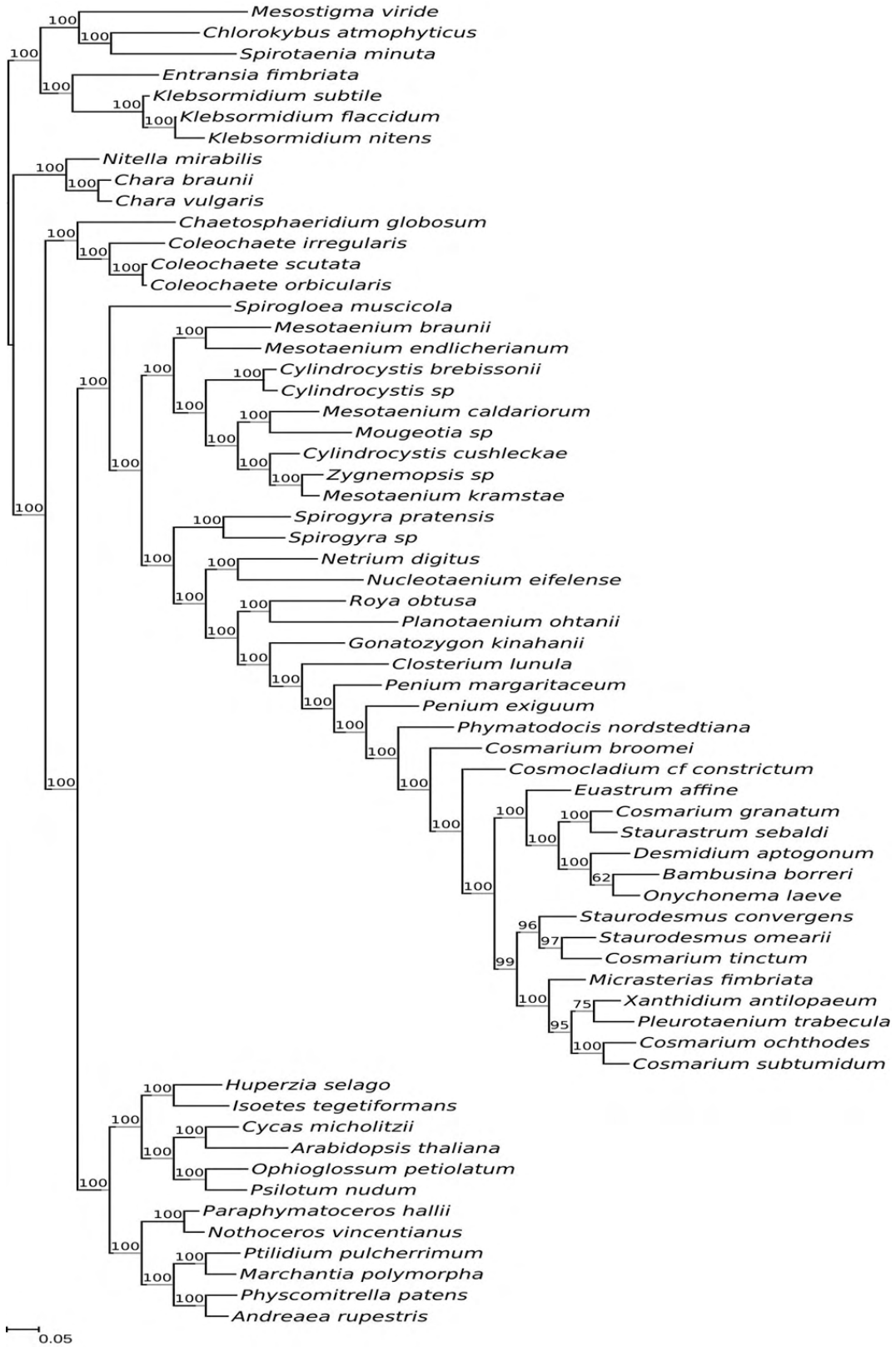


Figure A7 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 30% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1507447. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

Figure A8 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 40% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1007909. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

Figure A9 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 50% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -672970. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

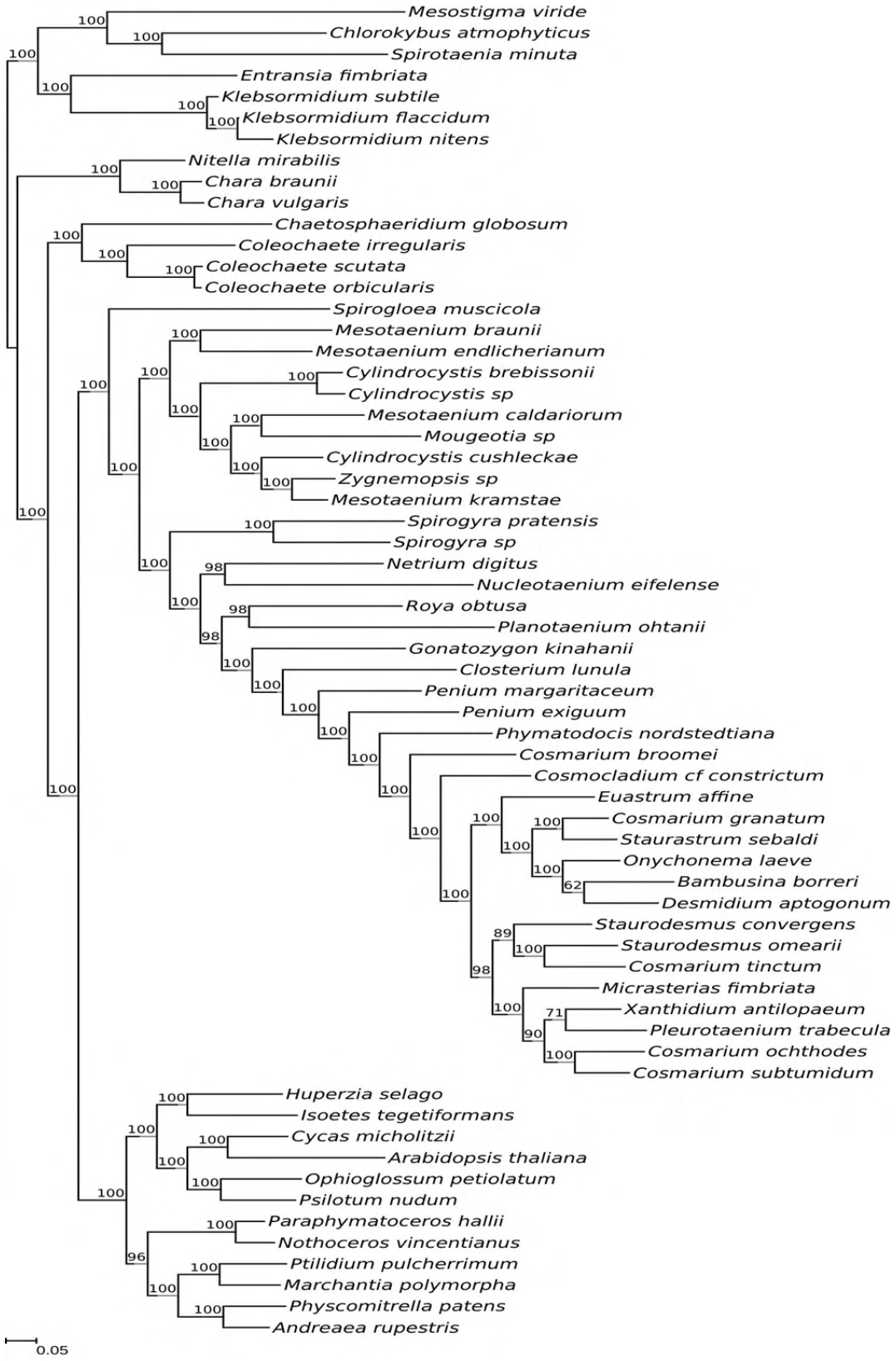


Figure A10 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 10% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2666368. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

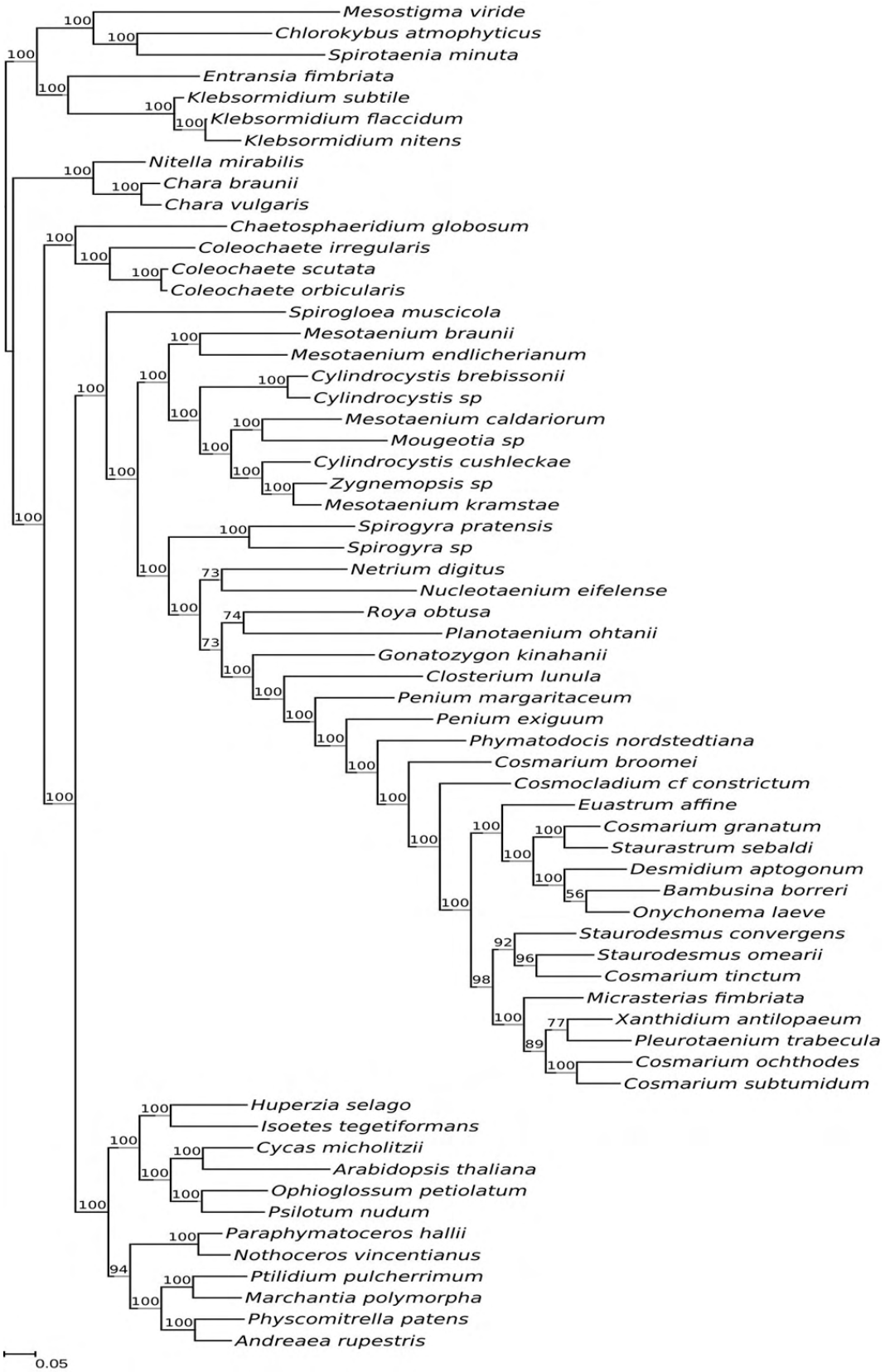


Figure A11 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 20% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -2015067. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

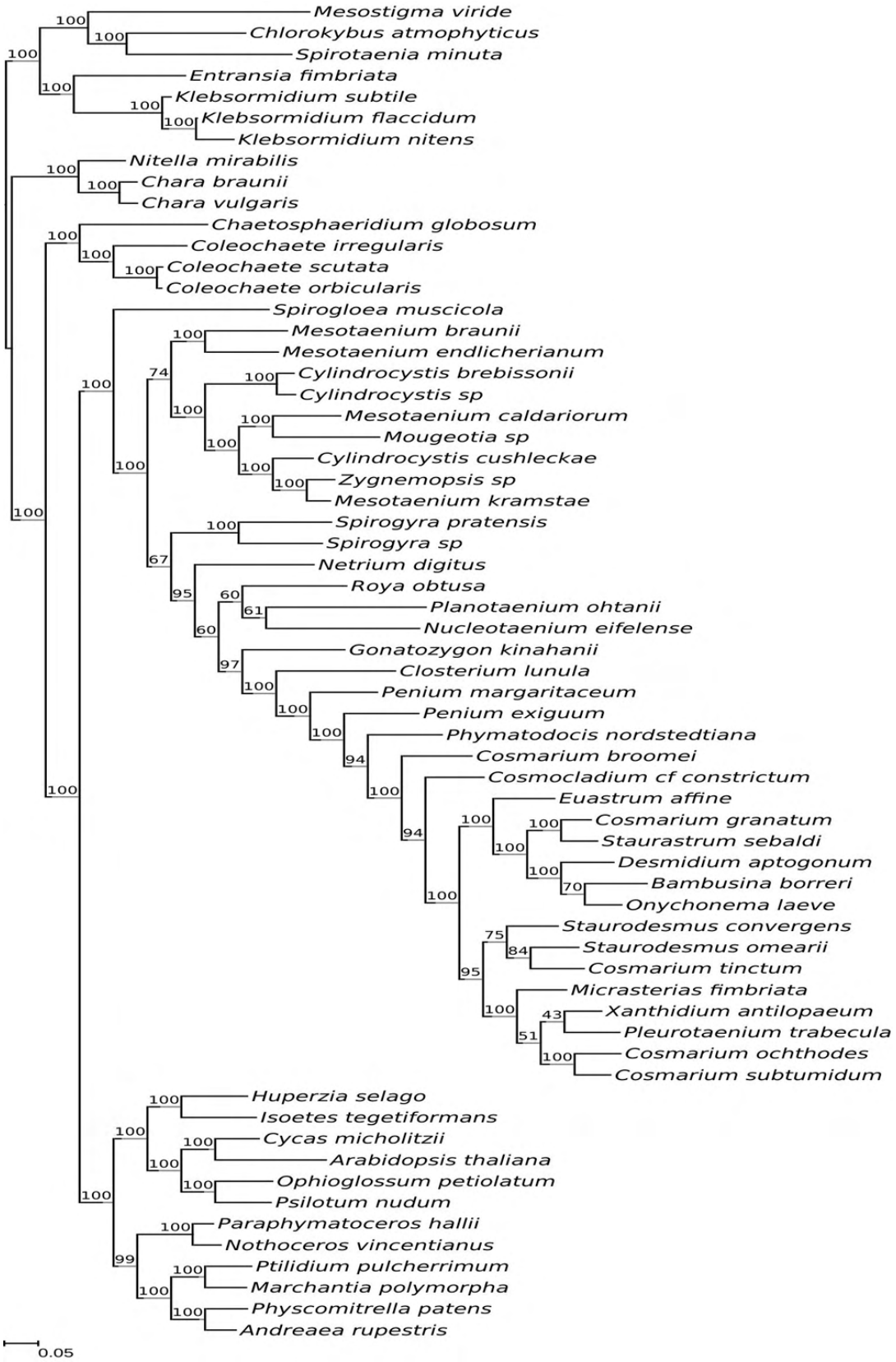


Figure A12 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 30% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1483727 Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

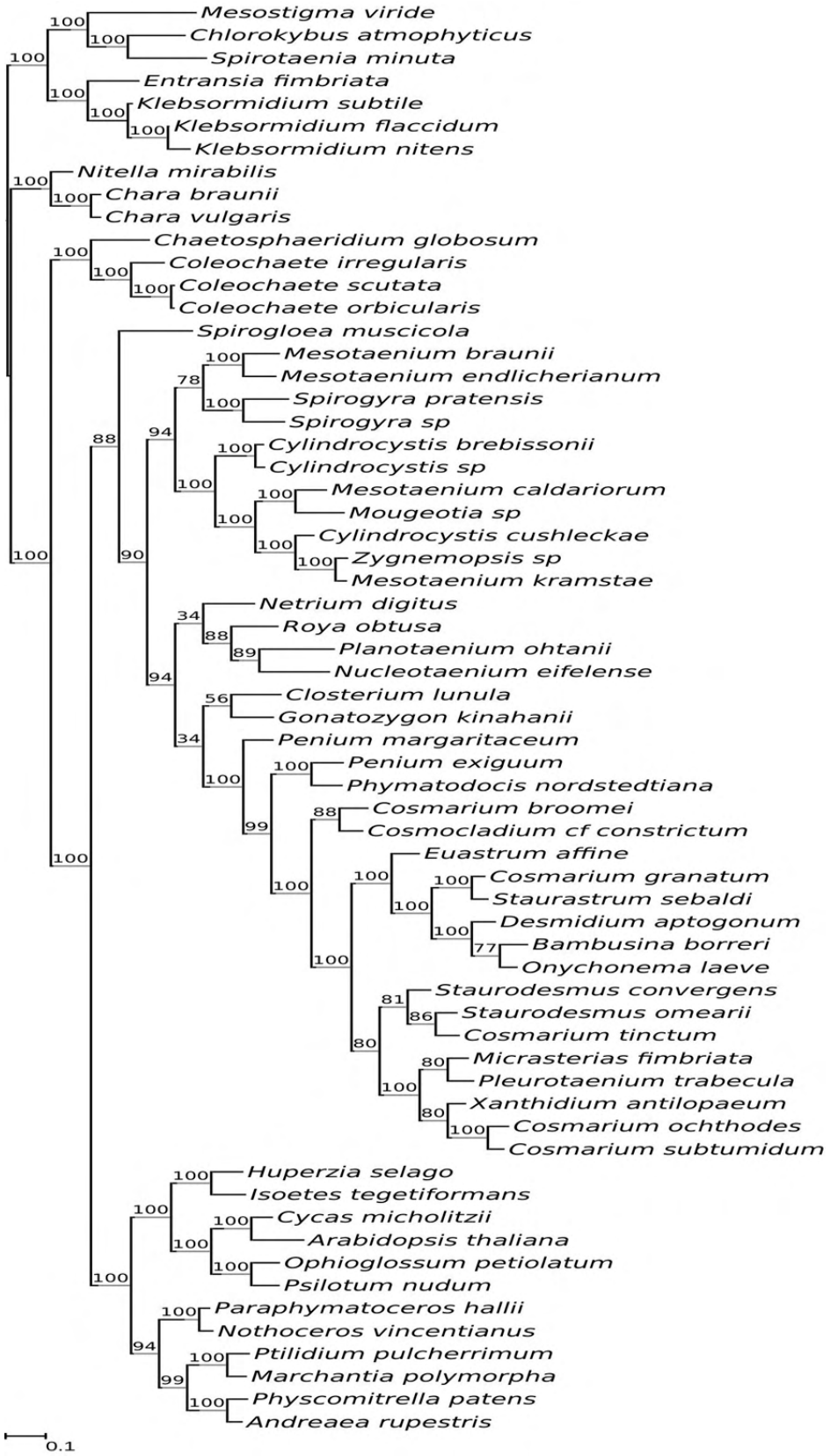


Figure A13 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 40% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1044837. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

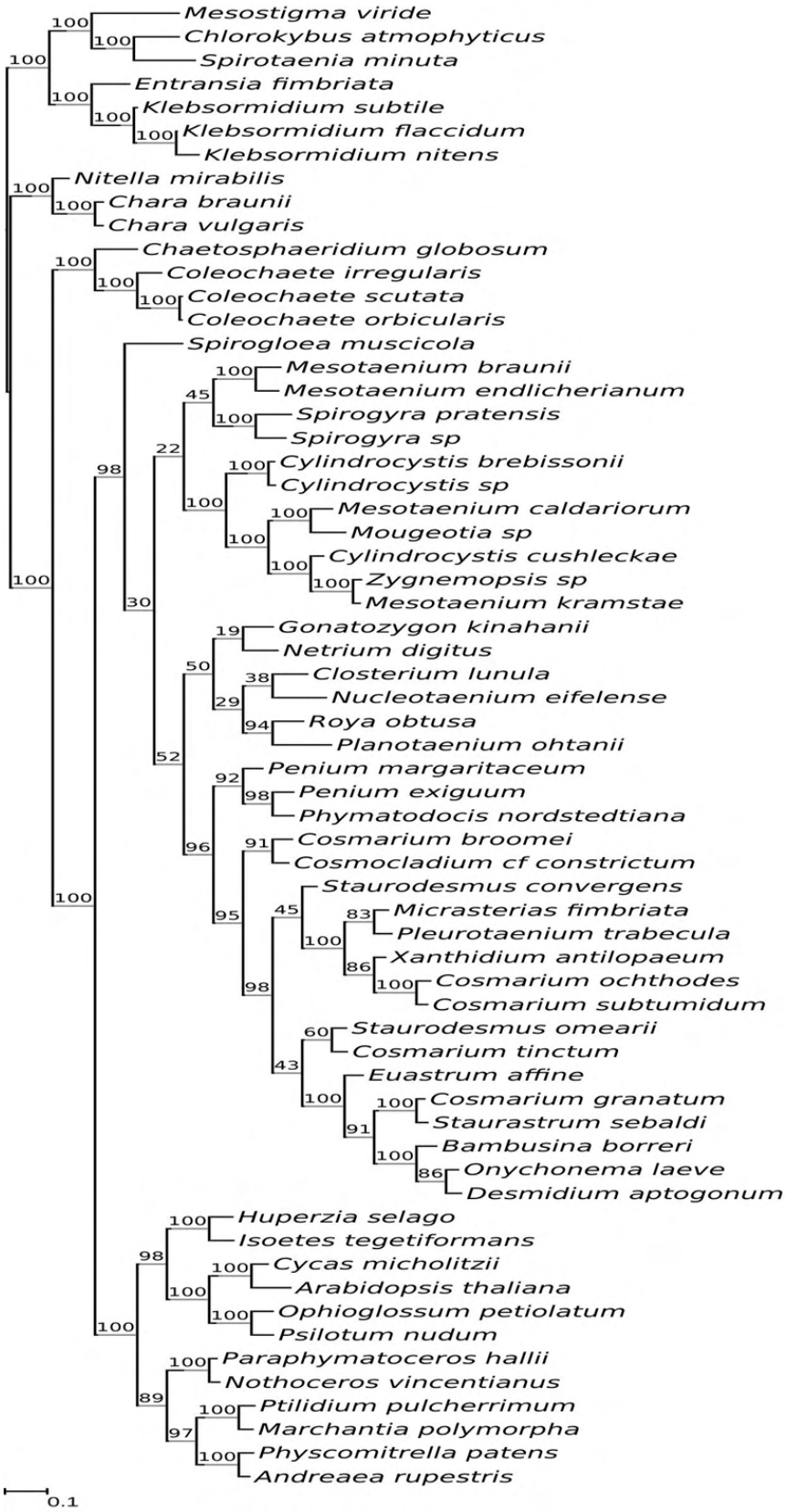


Figure A14 - Optimal maximum-likelihood tree reconstructed from 409 concatenated nuclear proteins after removal 50% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 63 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -673386. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

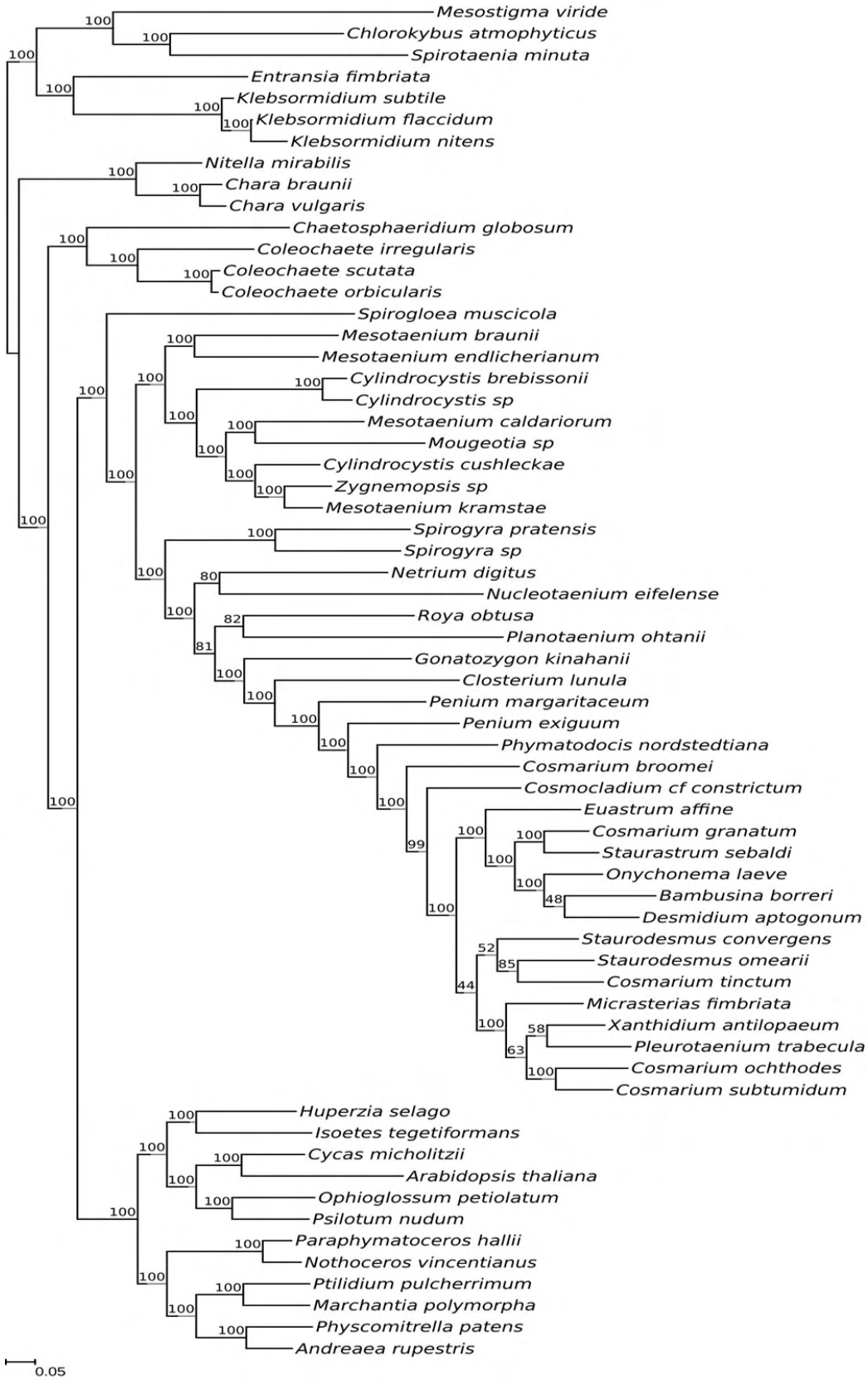


Figure A15 - Optimal maximum-likelihood tree comprising 409 taxa reconstructed from buried-sites partition of the 409 concatenated nuclear proteins. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1518530. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

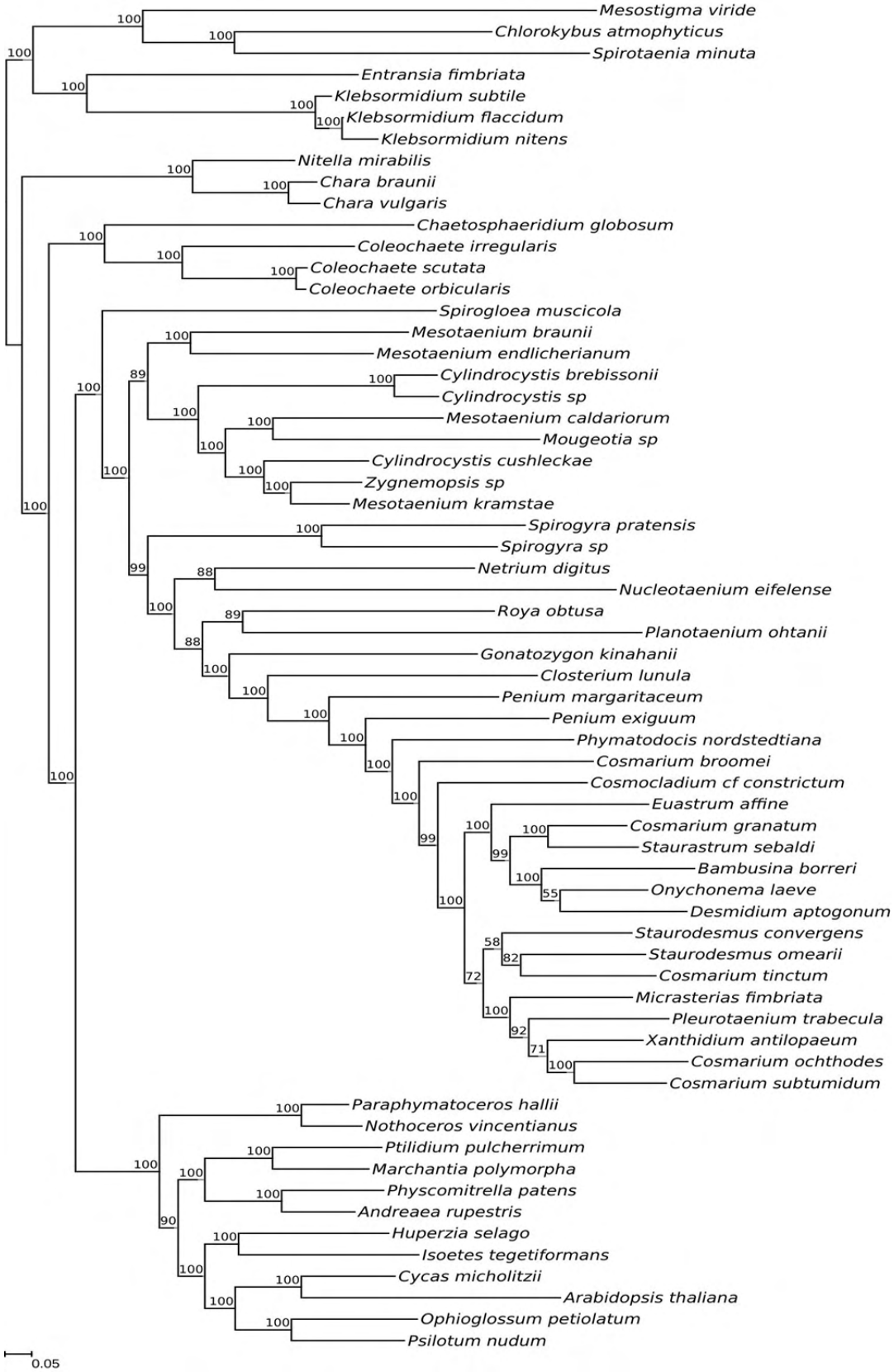


Figure A16 - Optimal maximum-likelihood tree comprising 409 taxa reconstructed from exposed-sites partition of the 409 concatenated nuclear proteins. Tree was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -1884393. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

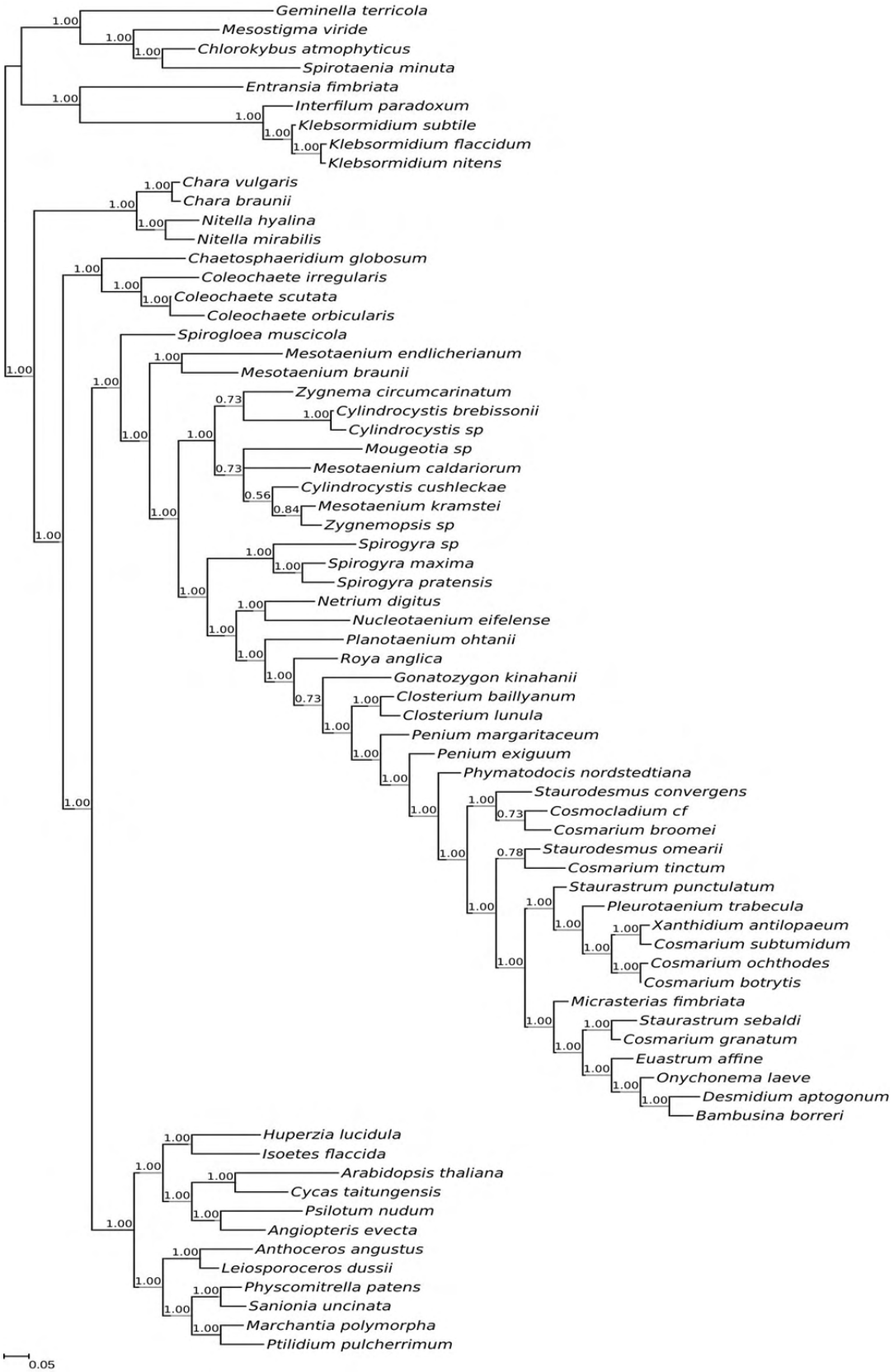


Figure A17 - Phylogeny comprising 71 taxa inferred from 84 concatenated chloroplast proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of a MCMC analysis with site-homogeneous composition and data-specific substitution rates model ($GTR_{data}+\Gamma_4+F_{est}$). Marginal likelihood, *LML* = -312668. Node support values are Bayesian posterior probabilities.

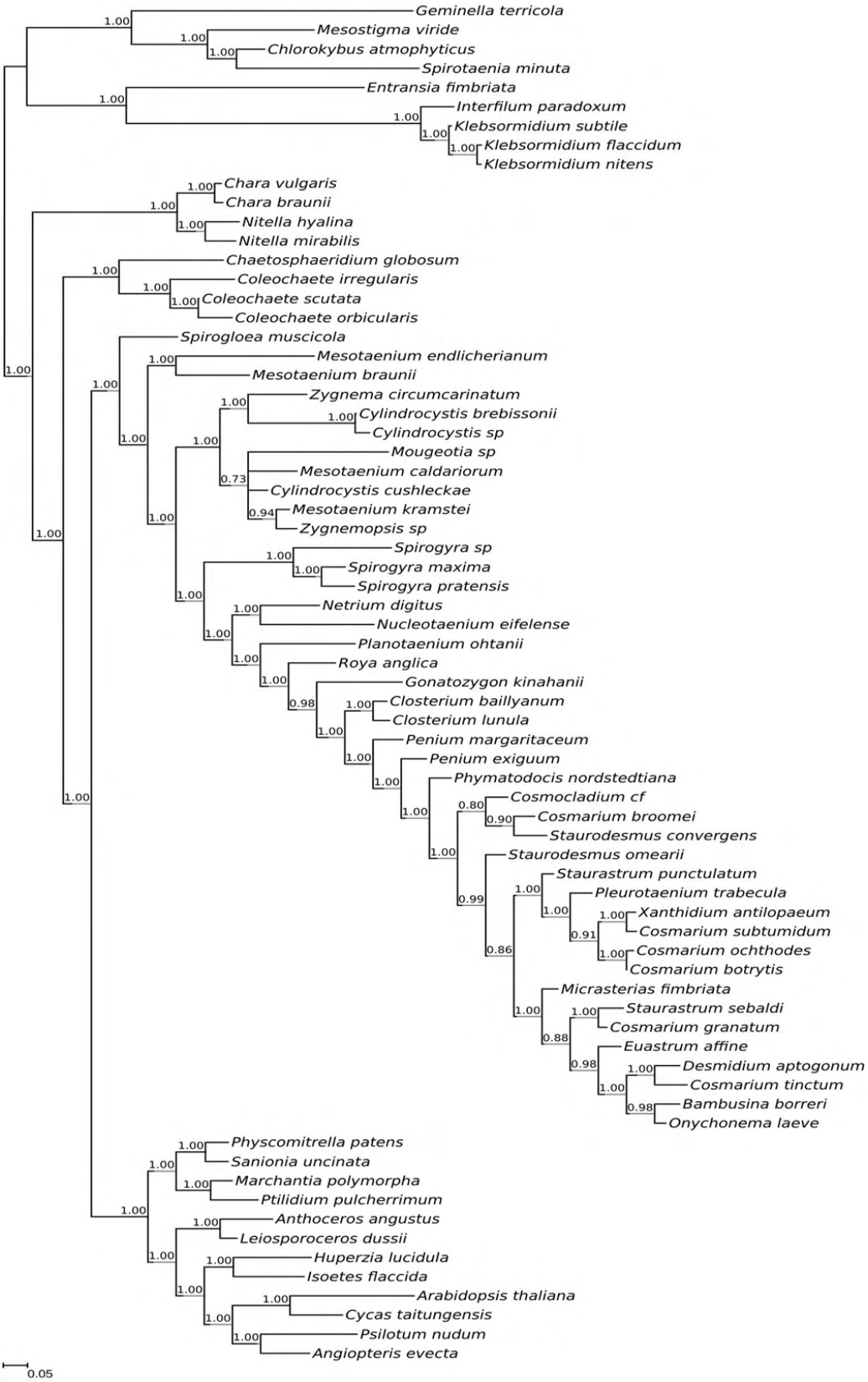


Figure A18 - Phylogeny comprising 71 taxa inferred from 84 concatenated chloroplast proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of a MCMC analysis using the CAT model and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{cat}$). Marginal likelihood, *LML* = -320781. Node support values are Bayesian posterior probabilities.

Figure A19 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 10% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -266630 Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

Figure A20 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 20% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -183614 Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

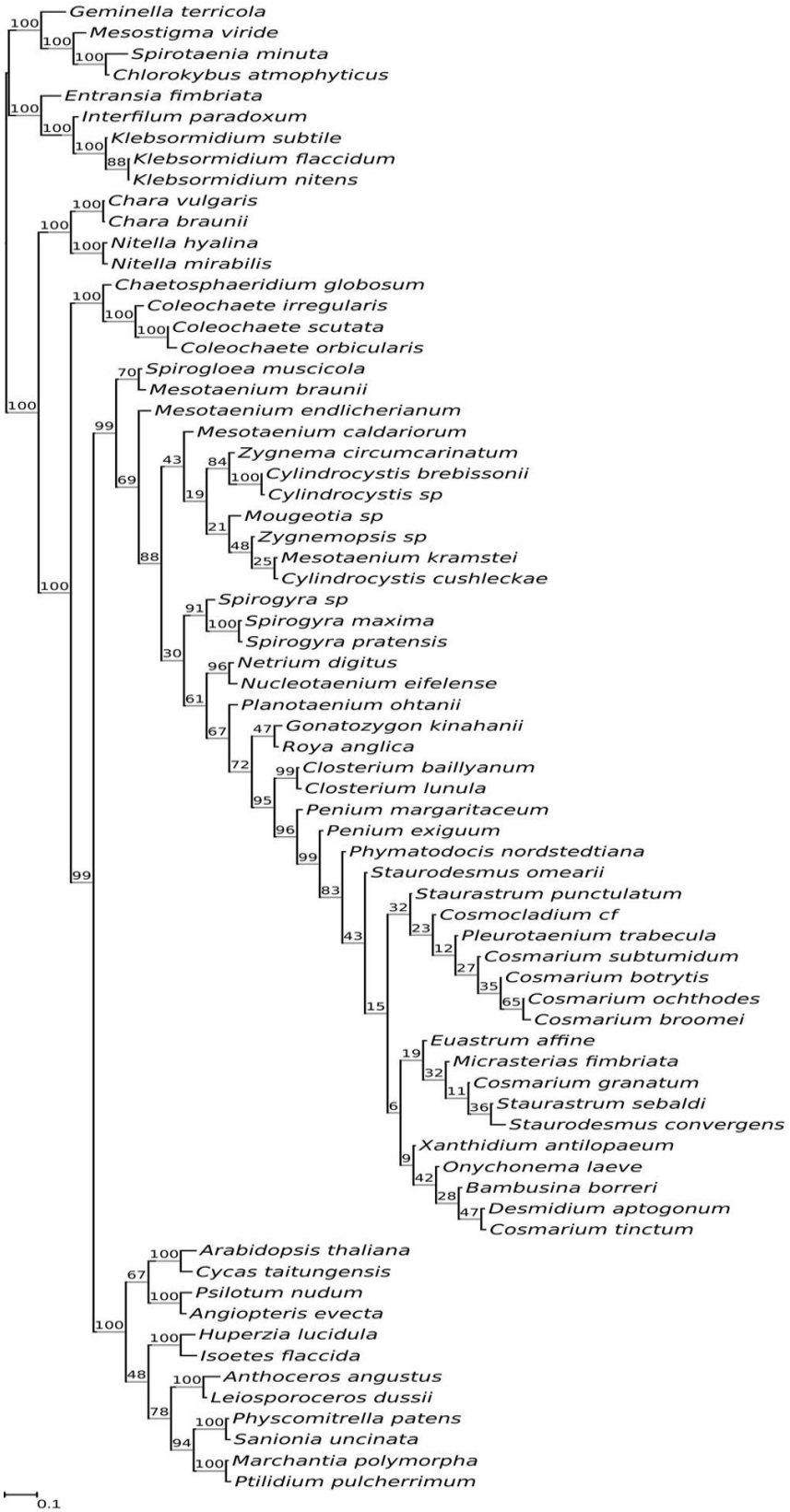


Figure A21 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 30% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -122749. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

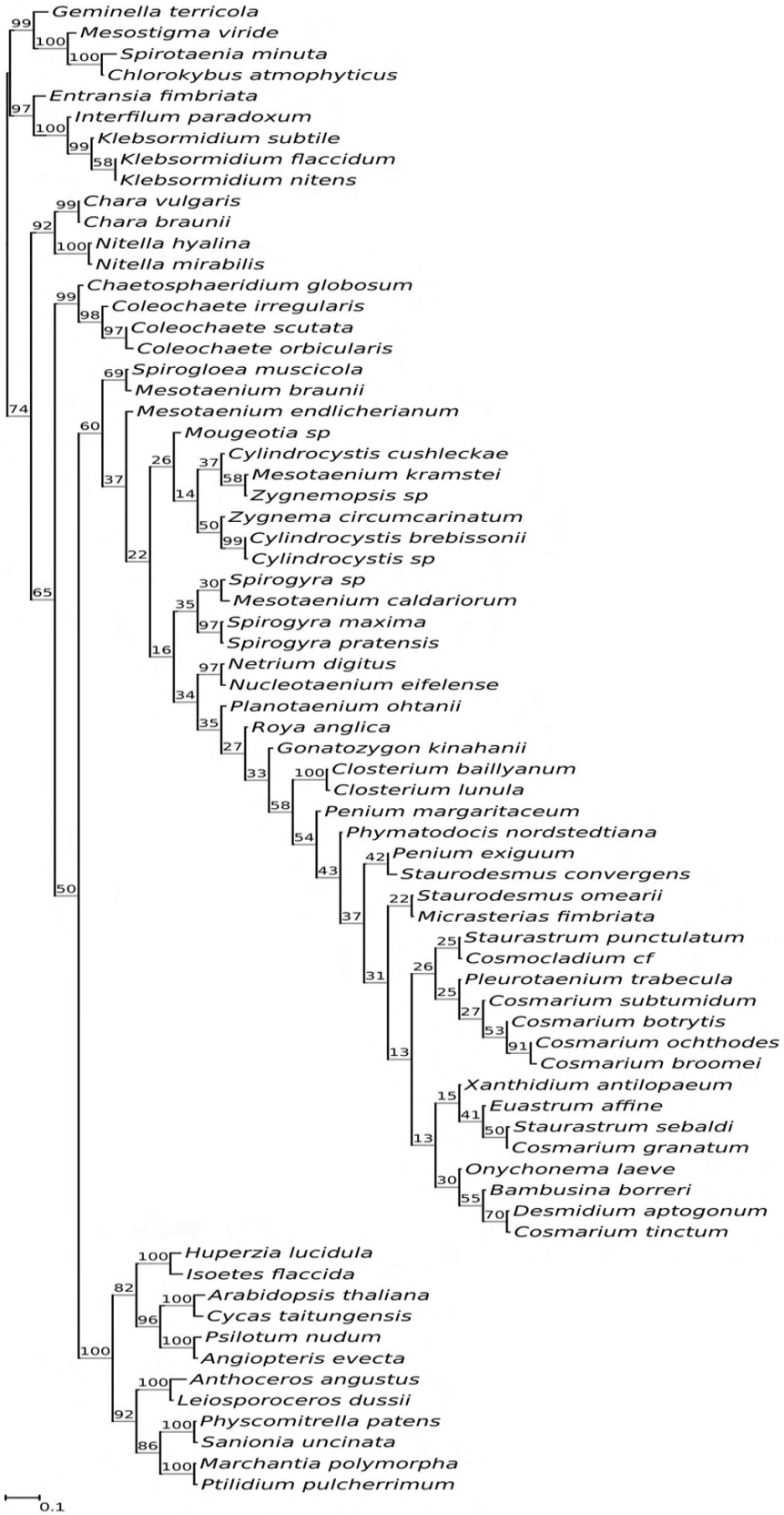


Figure A22 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 40% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -77842. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

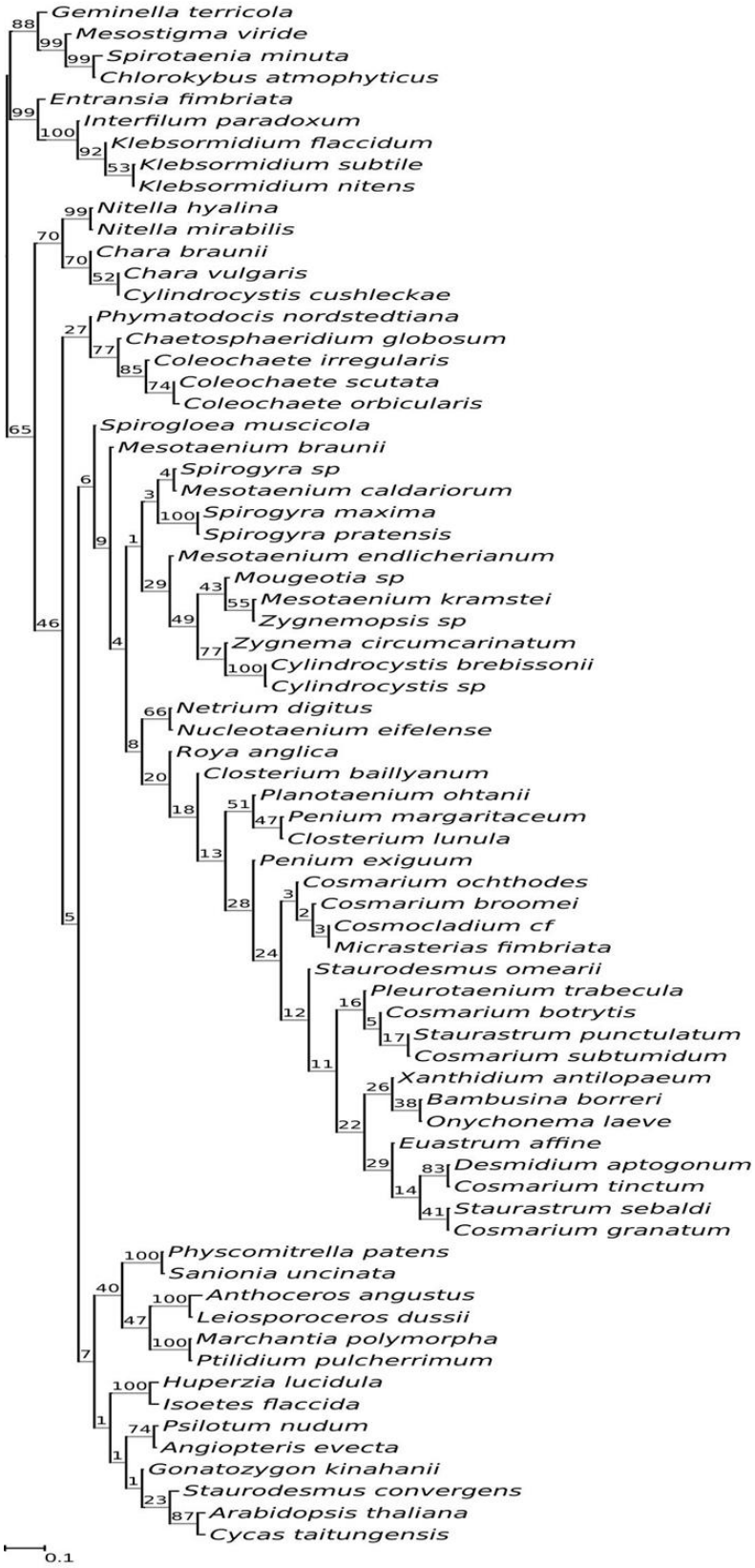


Figure A23 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 50% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -47277. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

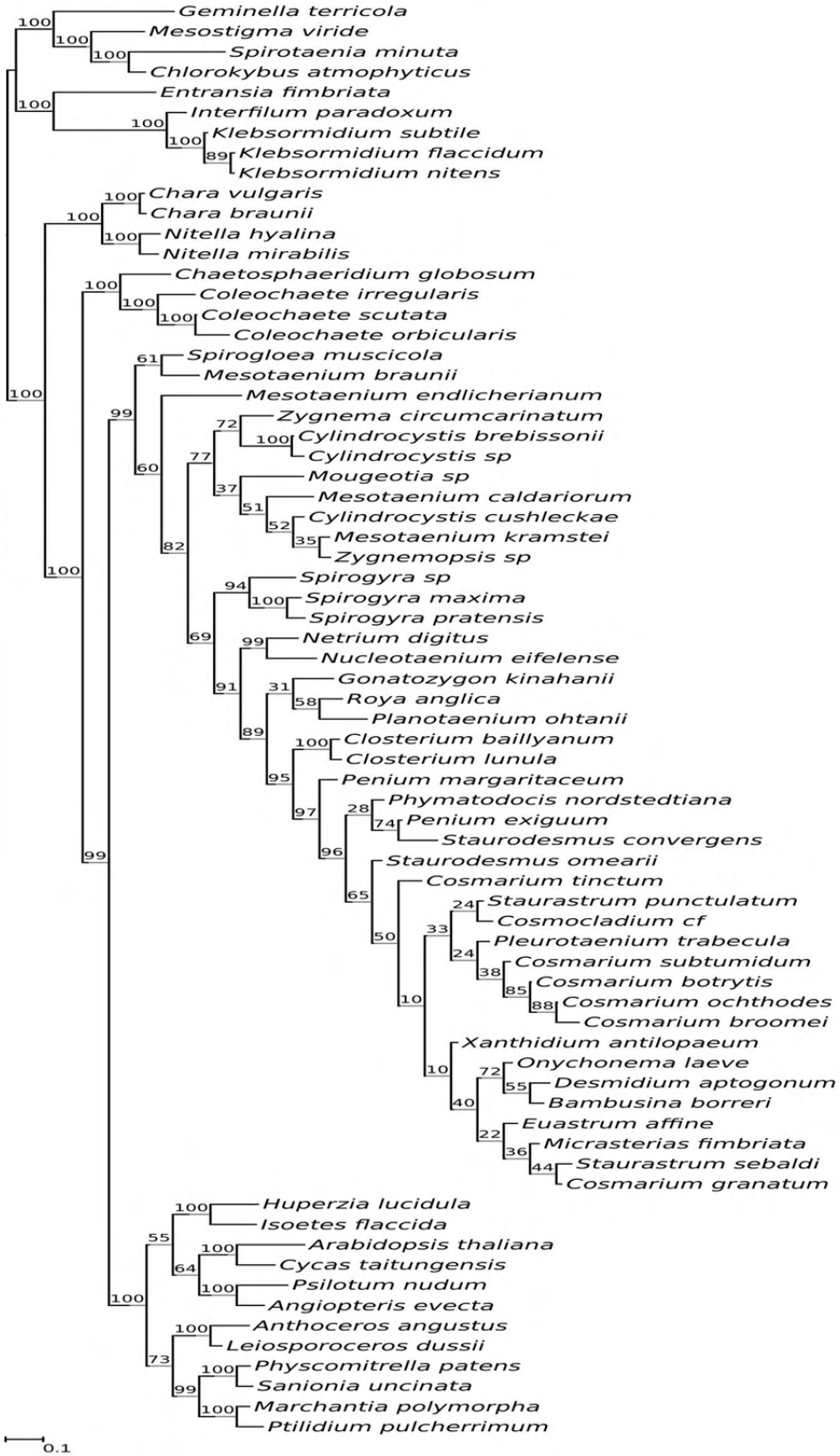


Figure A24 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 10% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -280743 Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

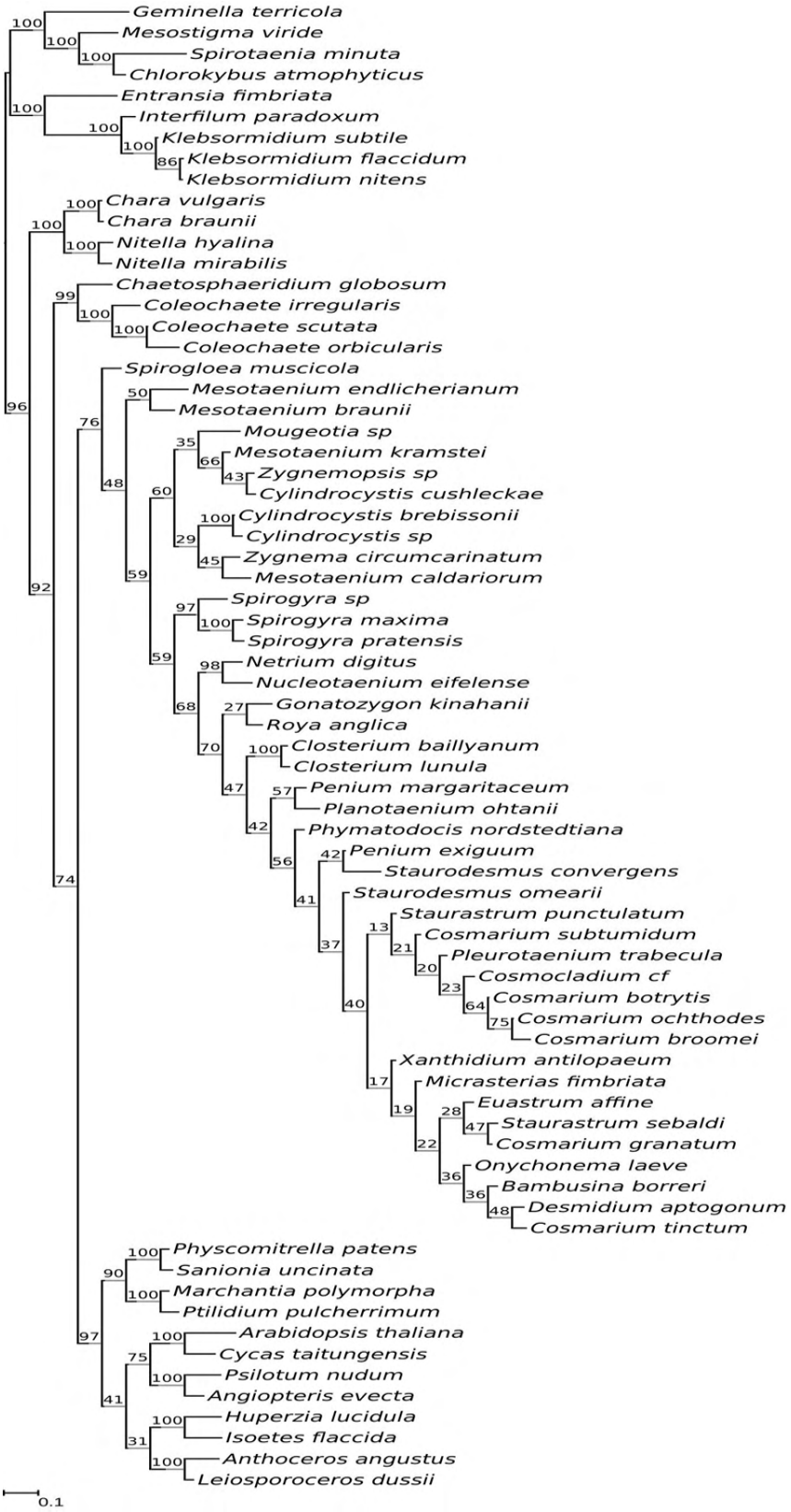


Figure A25 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 20% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -200708 Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

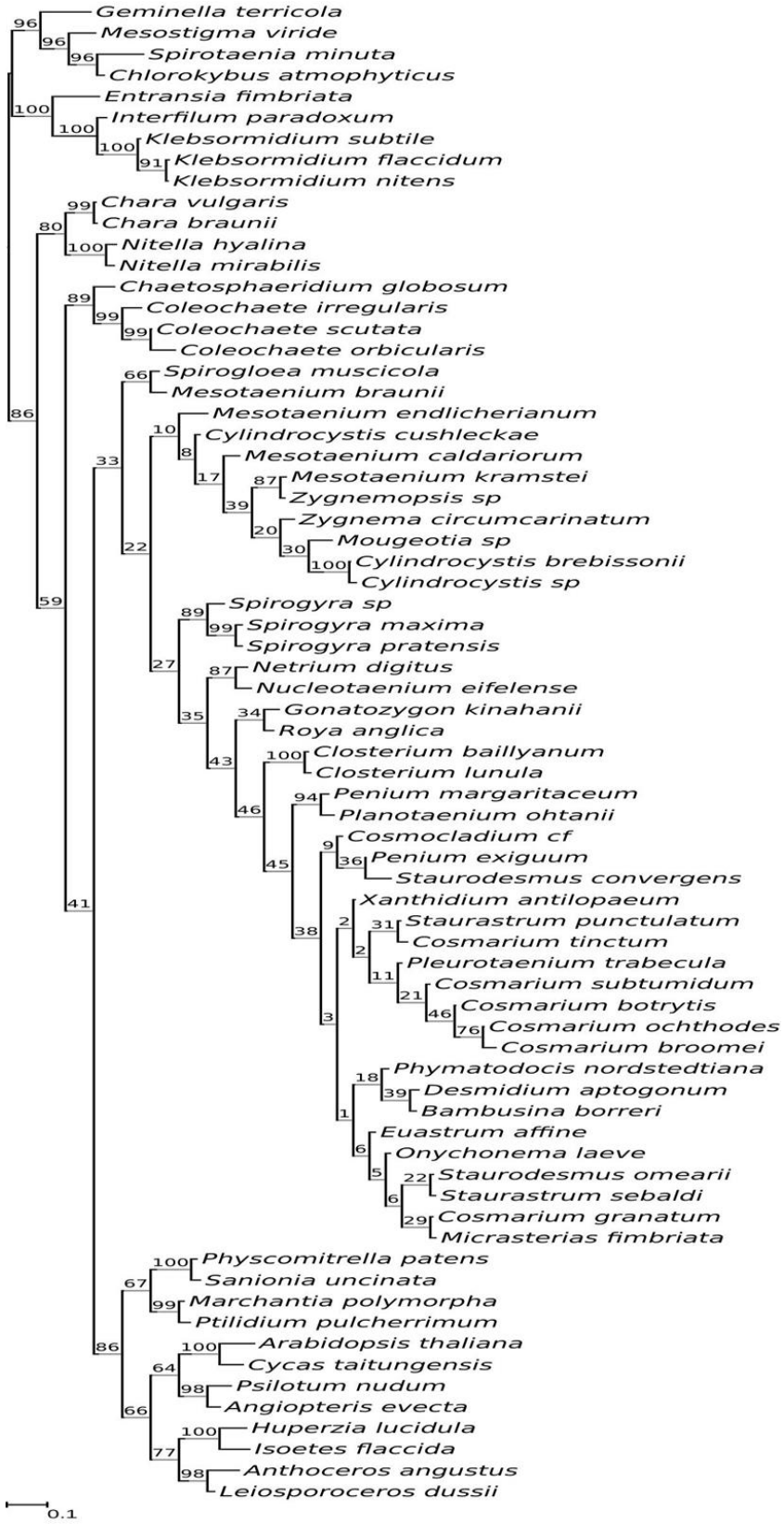


Figure A26 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 30% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -133170. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

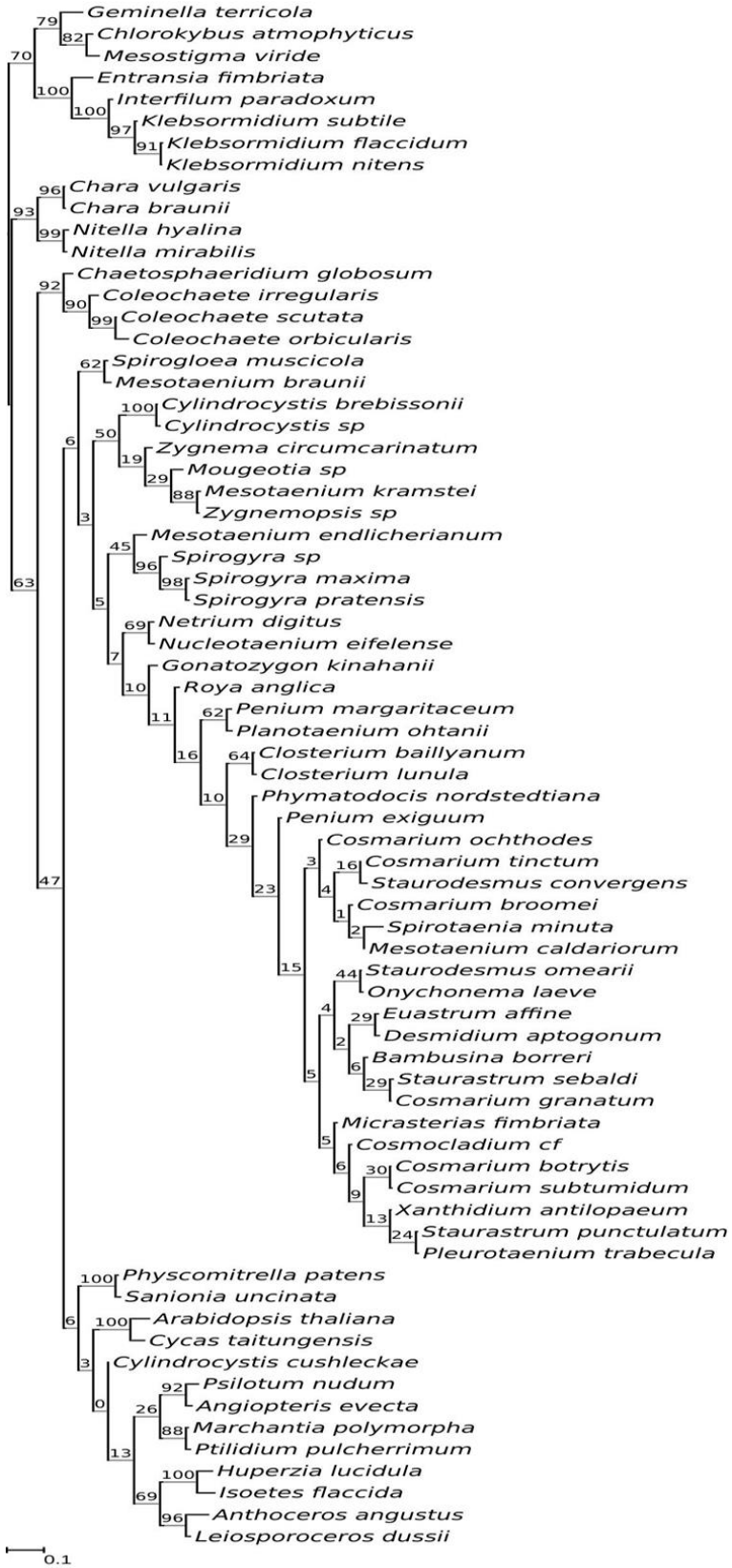


Figure A27 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 40% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -84504. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

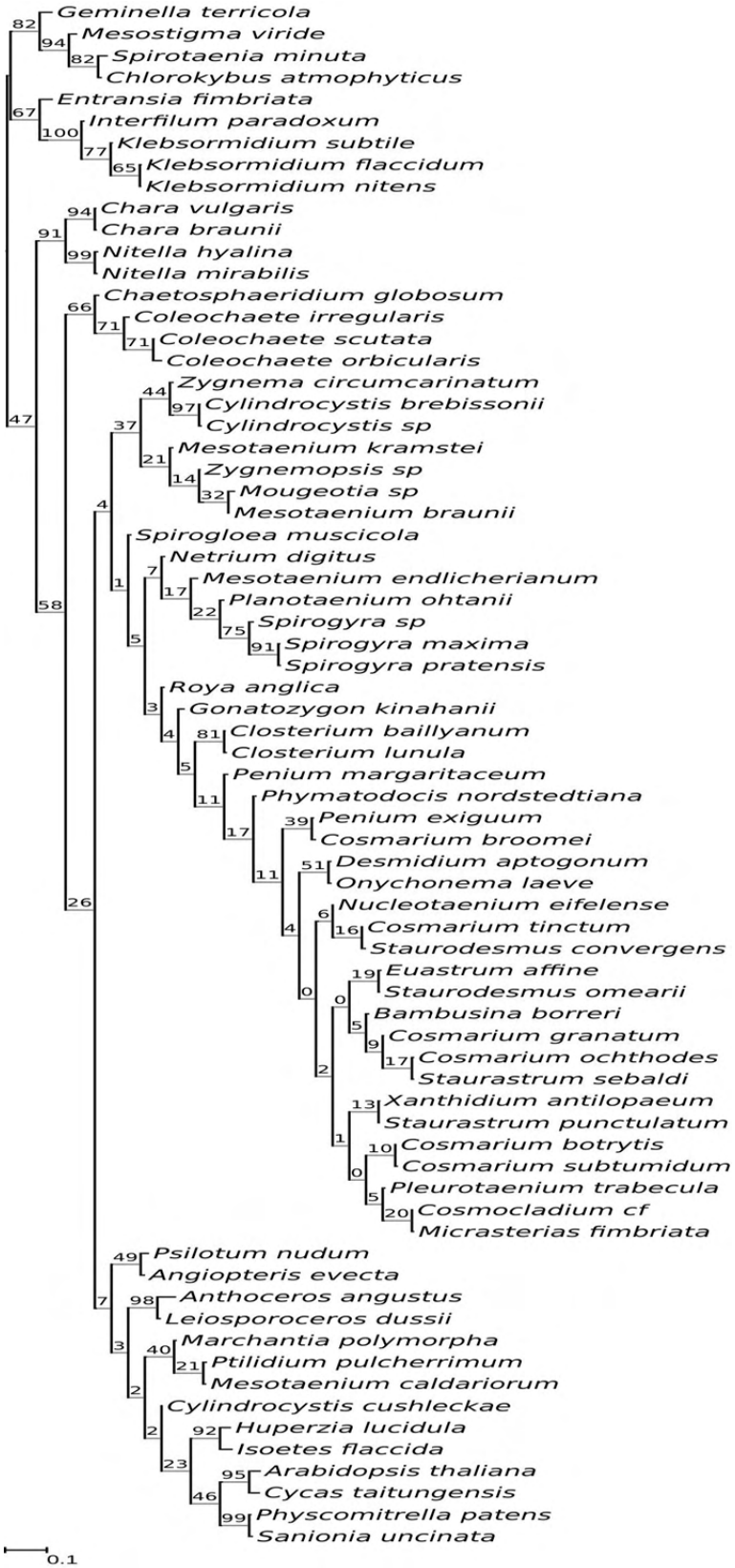


Figure A28 - Optimal maximum-likelihood tree reconstructed from 84 concatenated chloroplast proteins after removal 50% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 71 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -50985. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

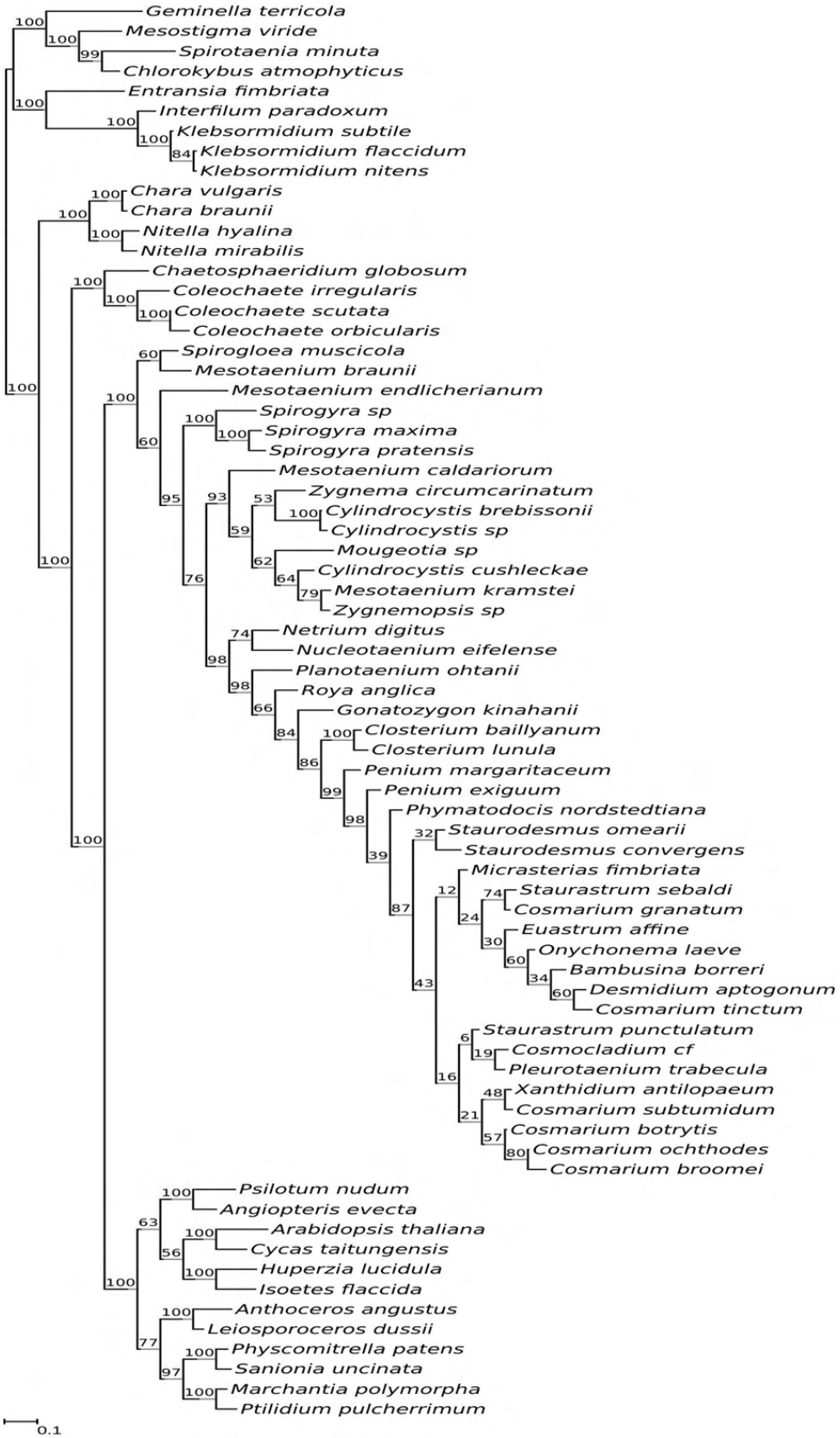


Figure A29 - Optimal maximum-likelihood tree comprising 71 taxa reconstructed from buried-sites partition of the 84 concatenated chloroplast proteins. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -195145. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

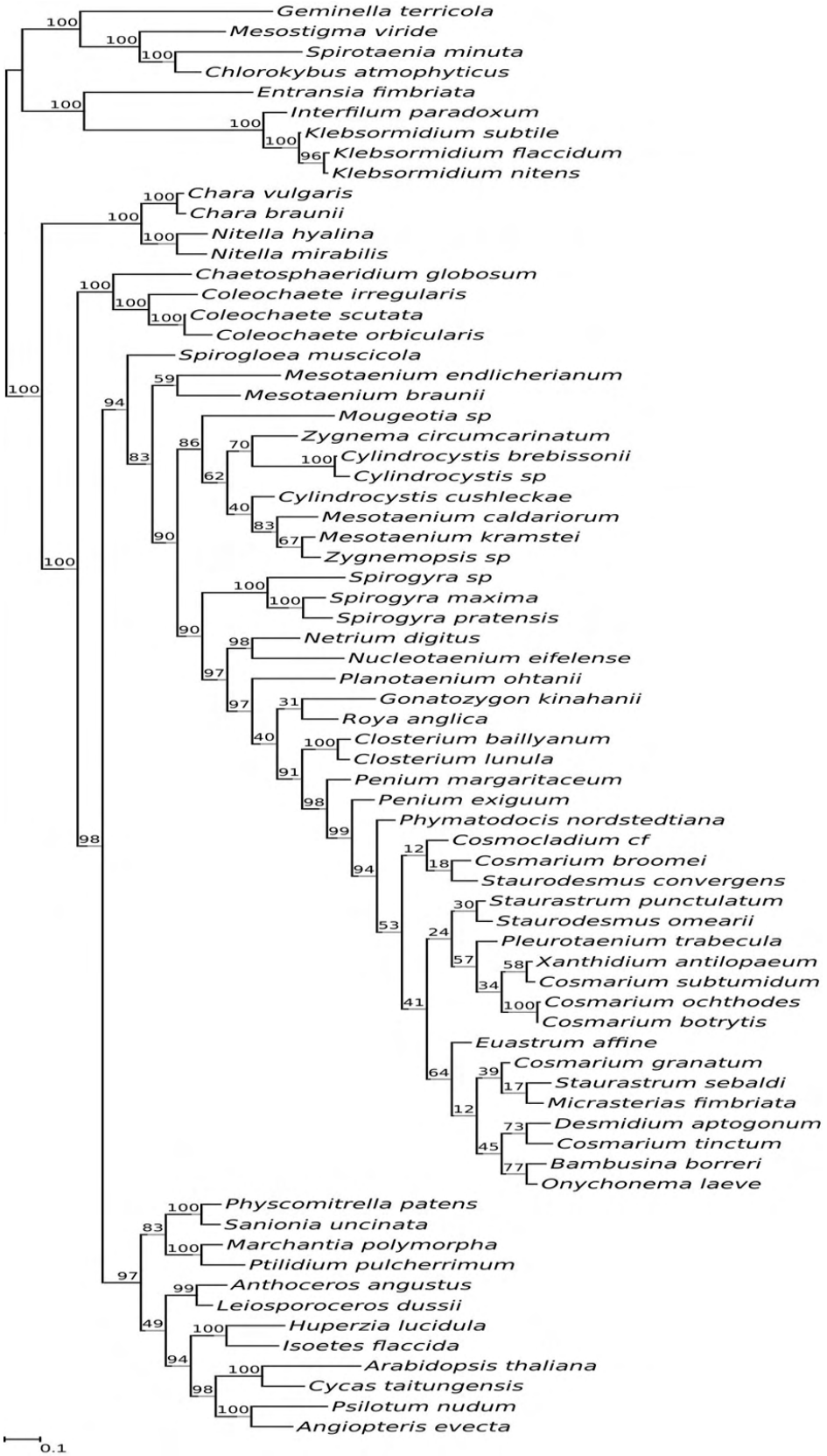


Figure A30 - Optimal maximum-likelihood tree comprising 71 taxa reconstructed from exposed-sites partition of the 84 concatenated chloroplast proteins. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -183404. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

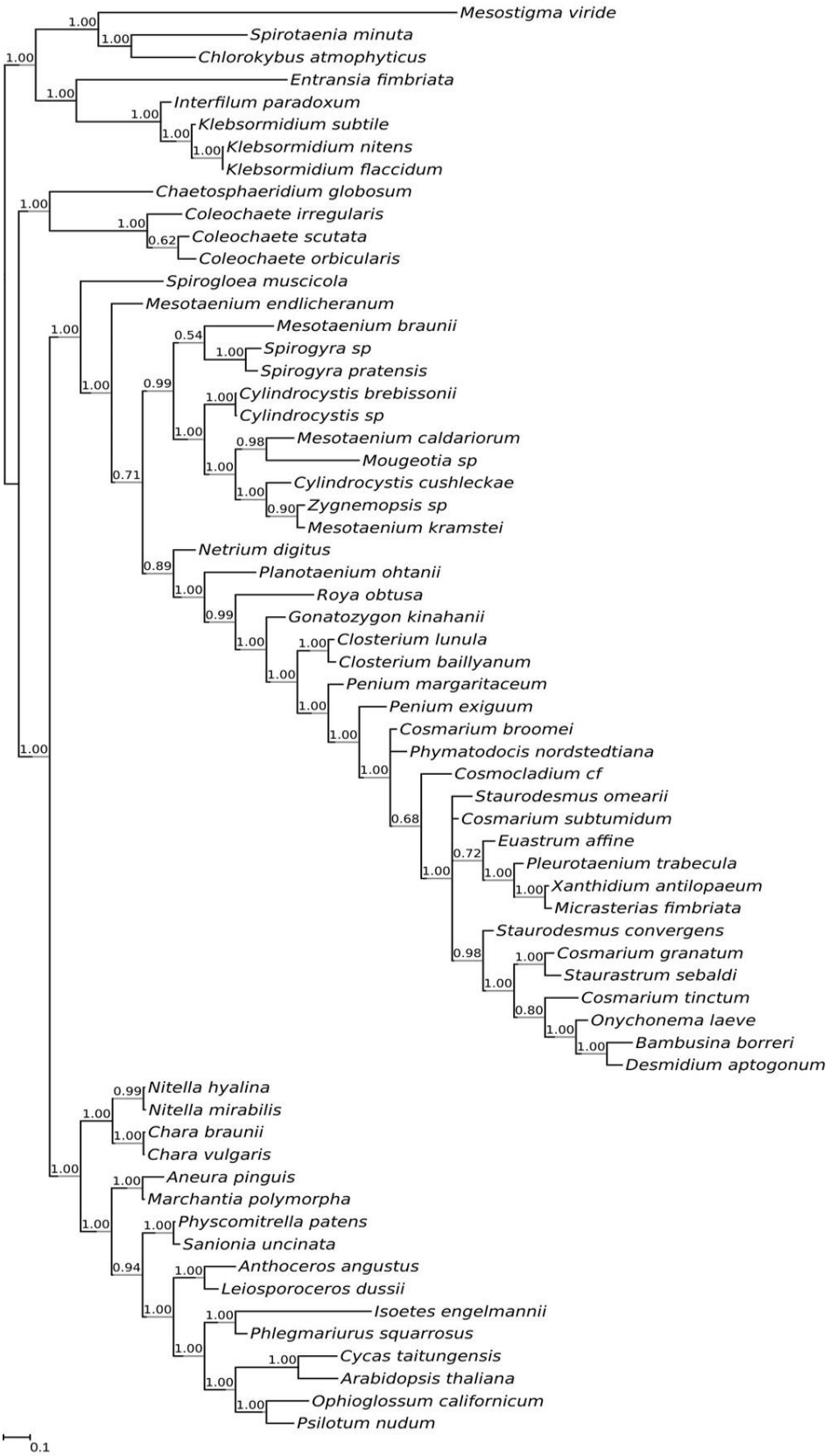


Figure A31 - Phylogeny comprising 40 taxa inferred from 64 concatenated mitochondrial proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of a MCMC analysis using the CAT model and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{CAT}$). Marginal likelihood, *LML* = -152815. Node support values are Bayesian posterior probabilities.

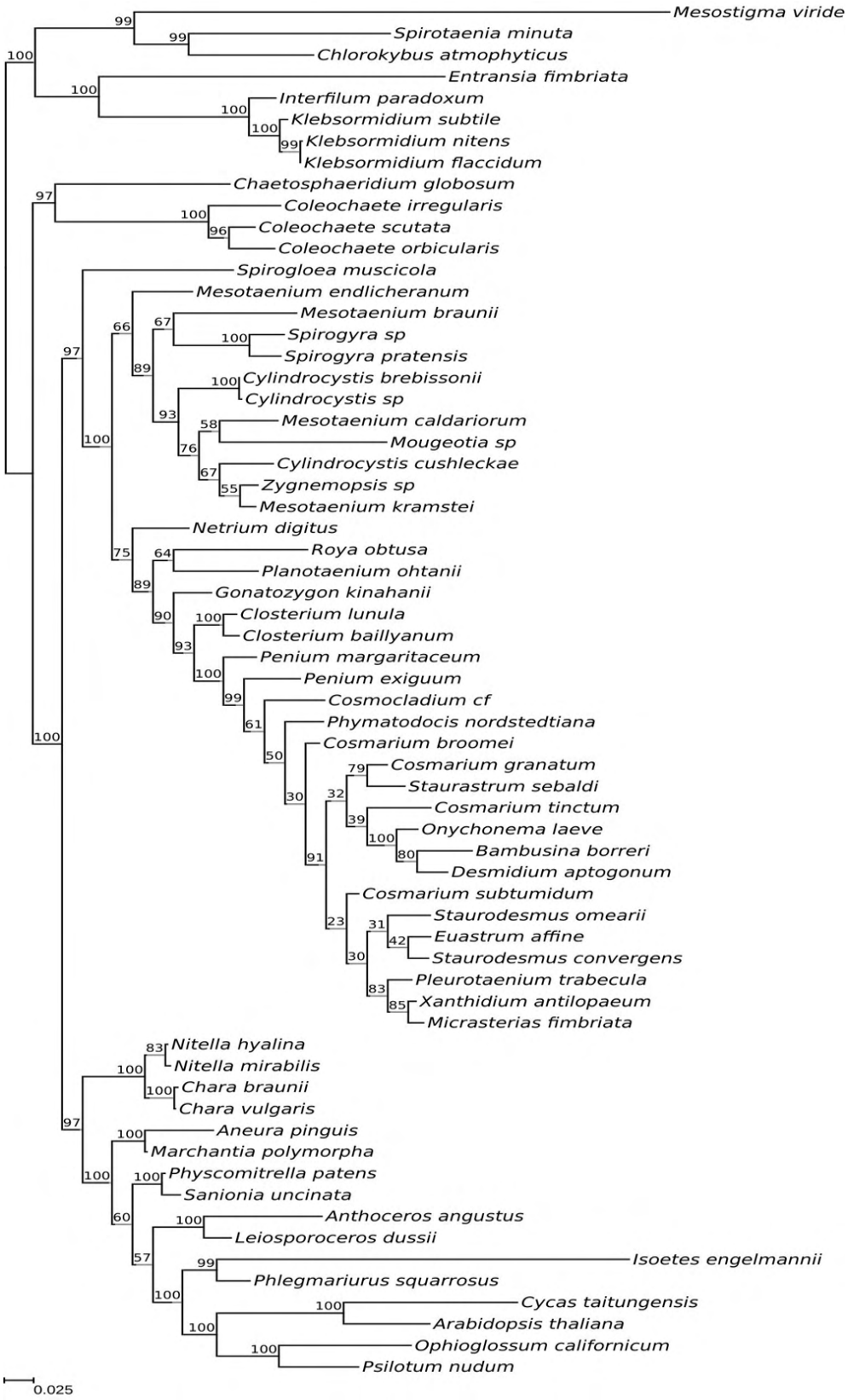


Figure A32 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 10% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -136071 Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

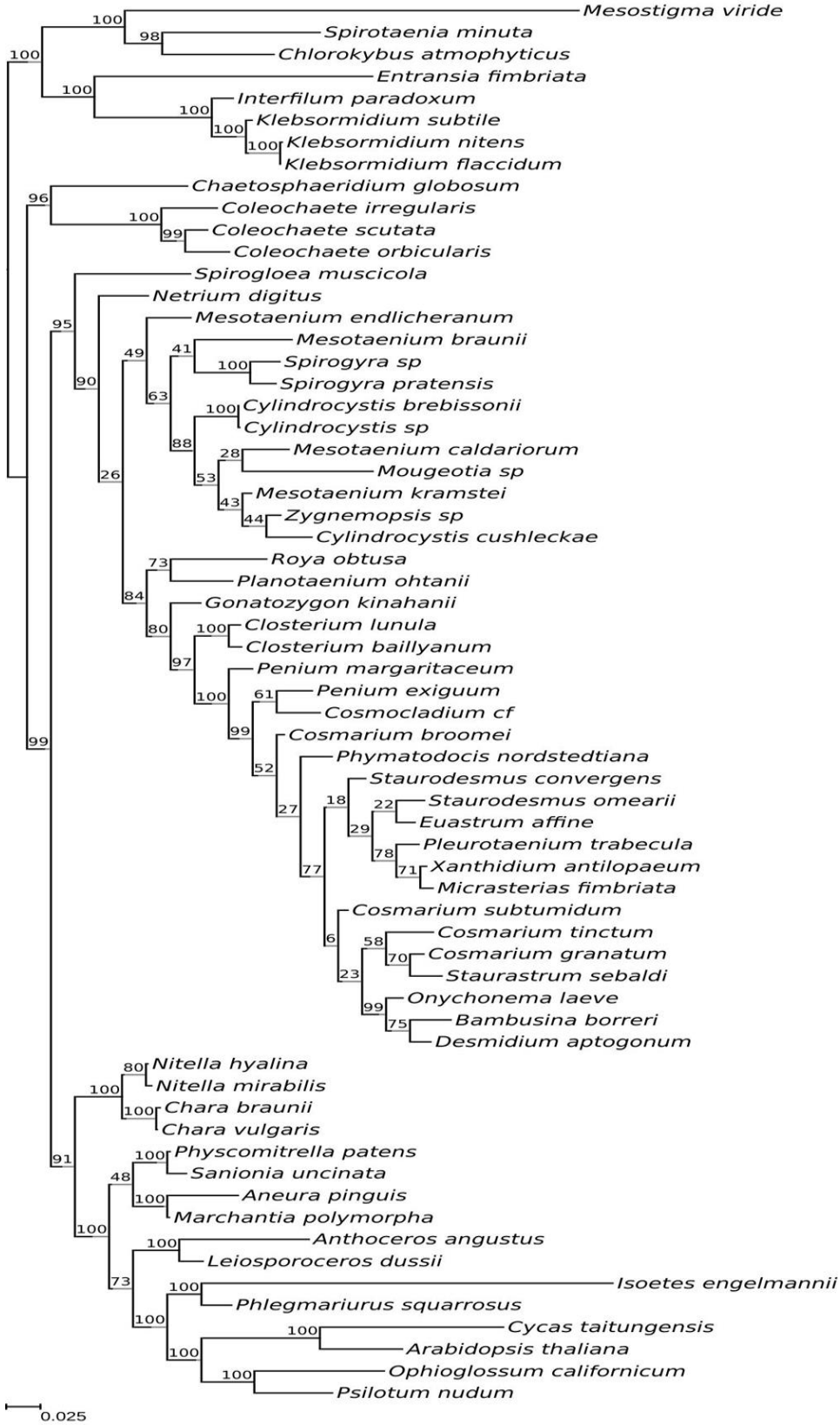


Figure A33 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 20% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -104102. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

Figure A34 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 30% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -76031. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

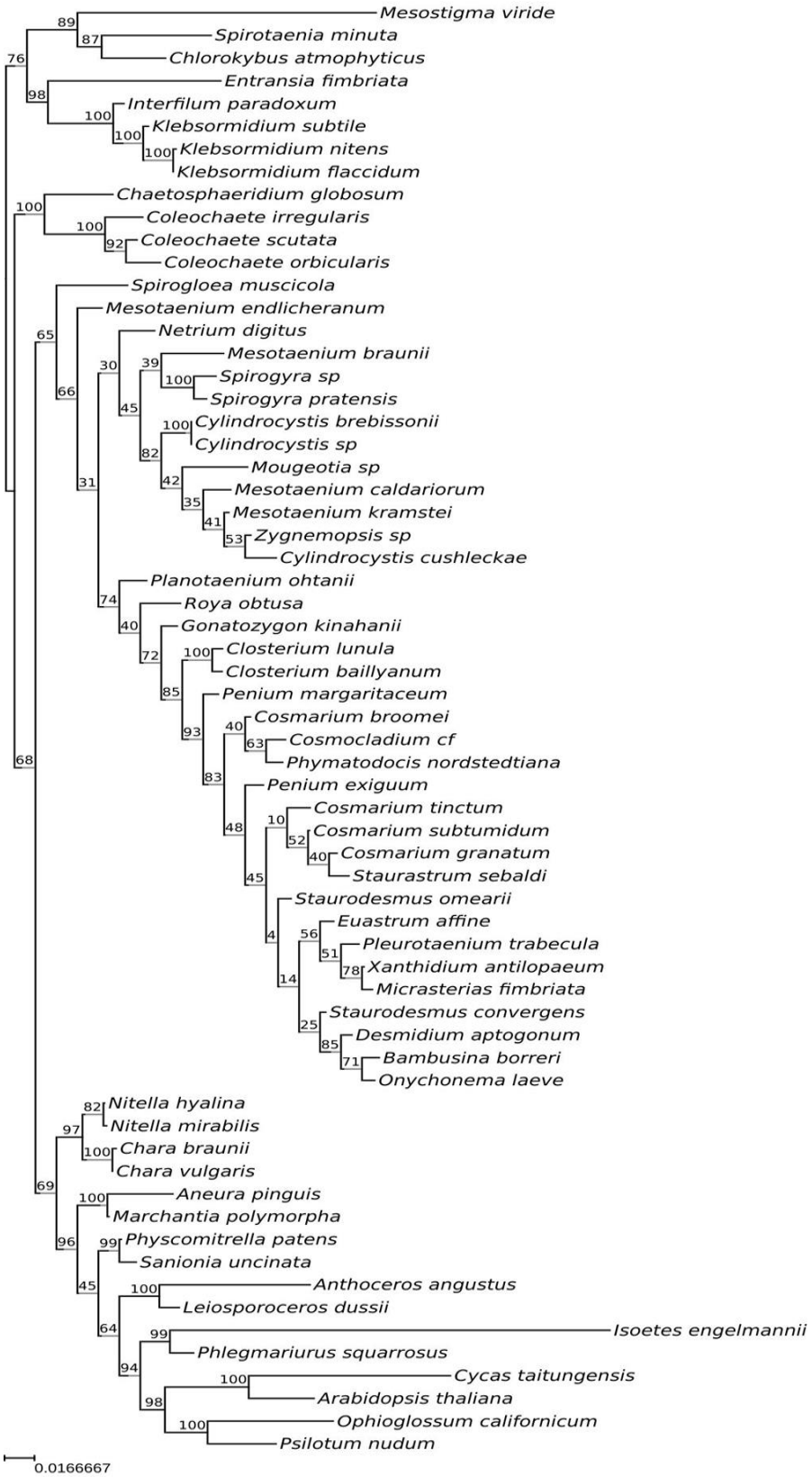


Figure A35 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 40% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -53126. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

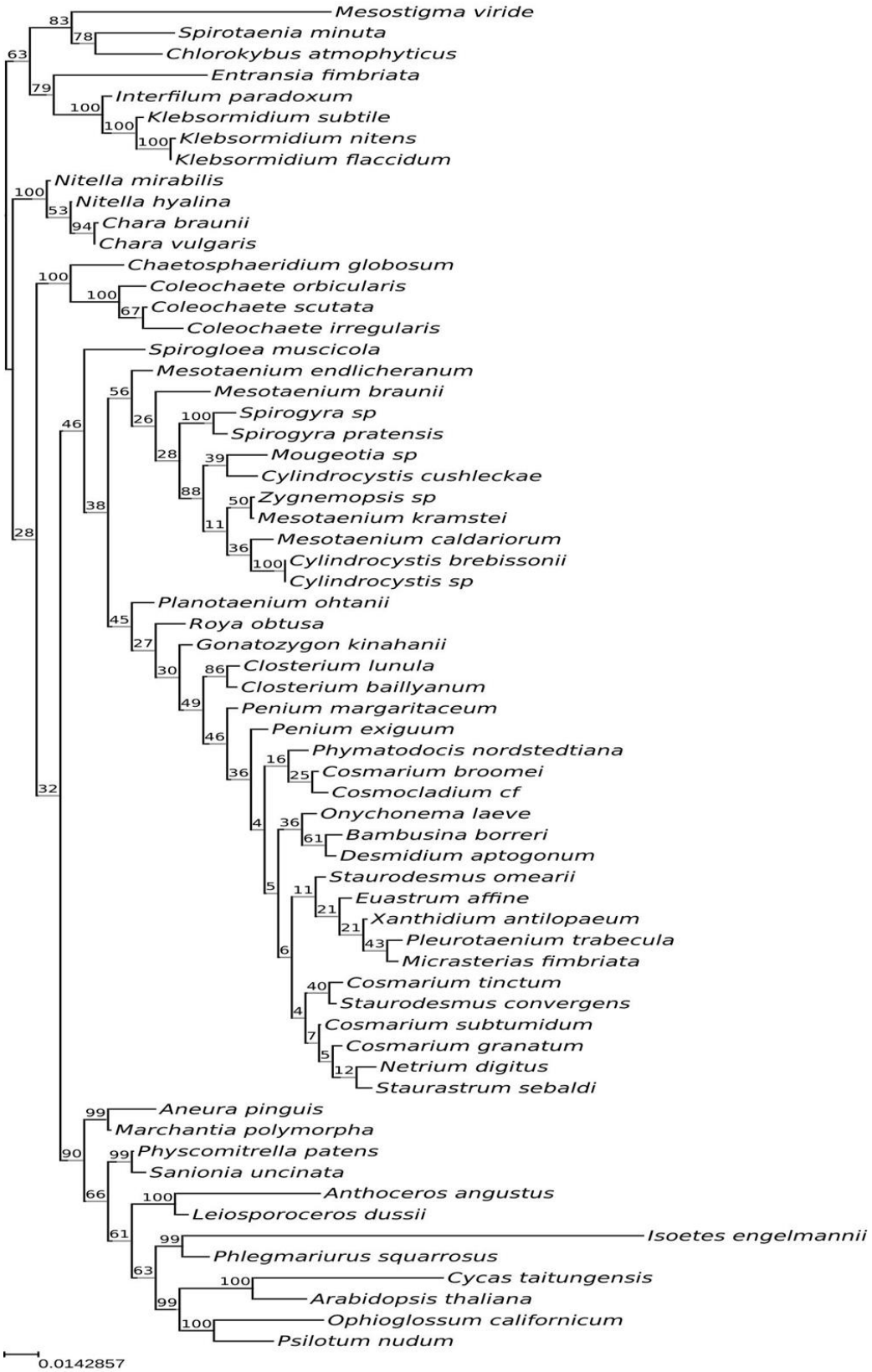


Figure A36 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 50% of the fastest-evolving sites determined according to the gamma-distributed among-site rate variation. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -35079. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

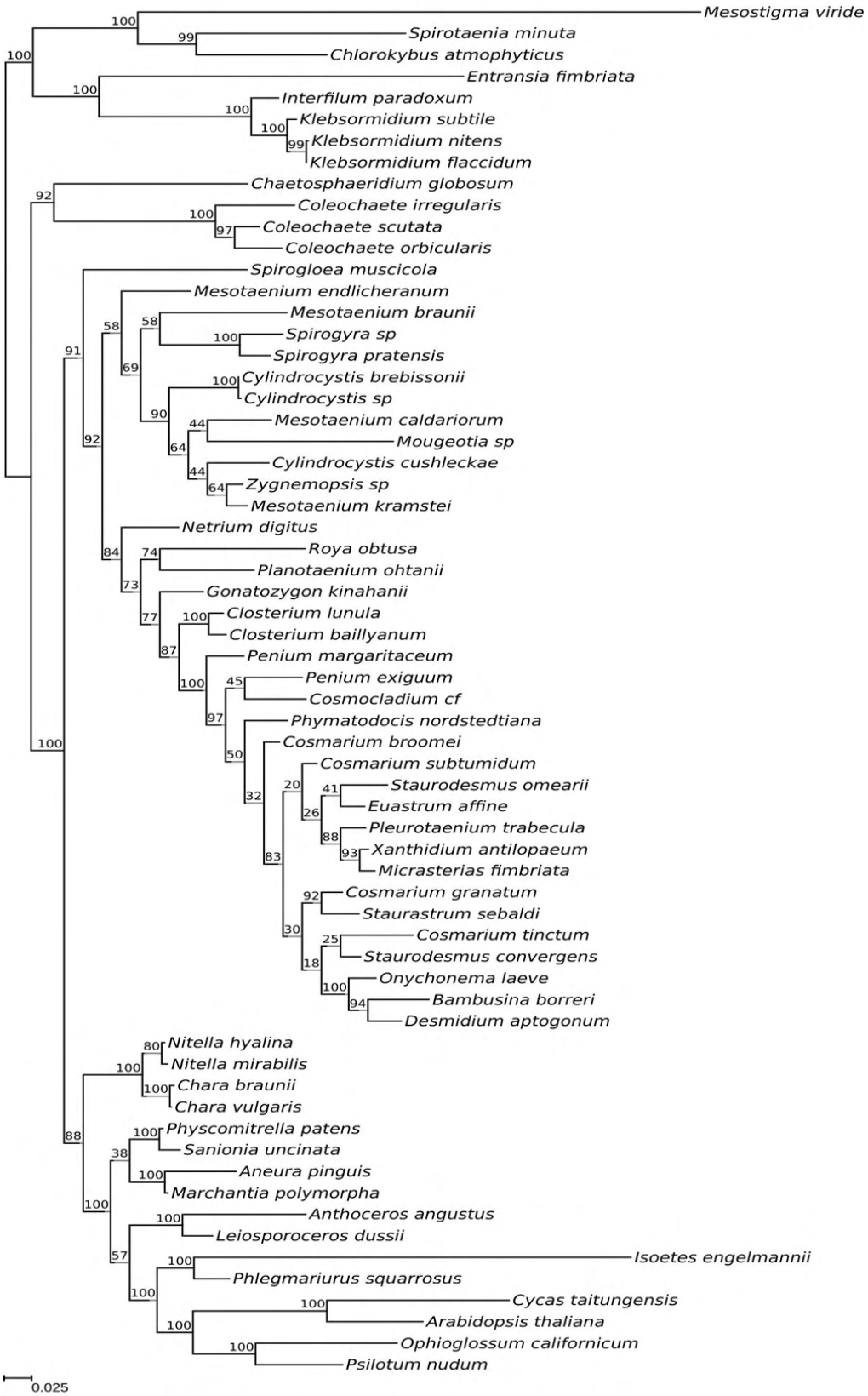


Figure A37 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 10% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -139030 Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

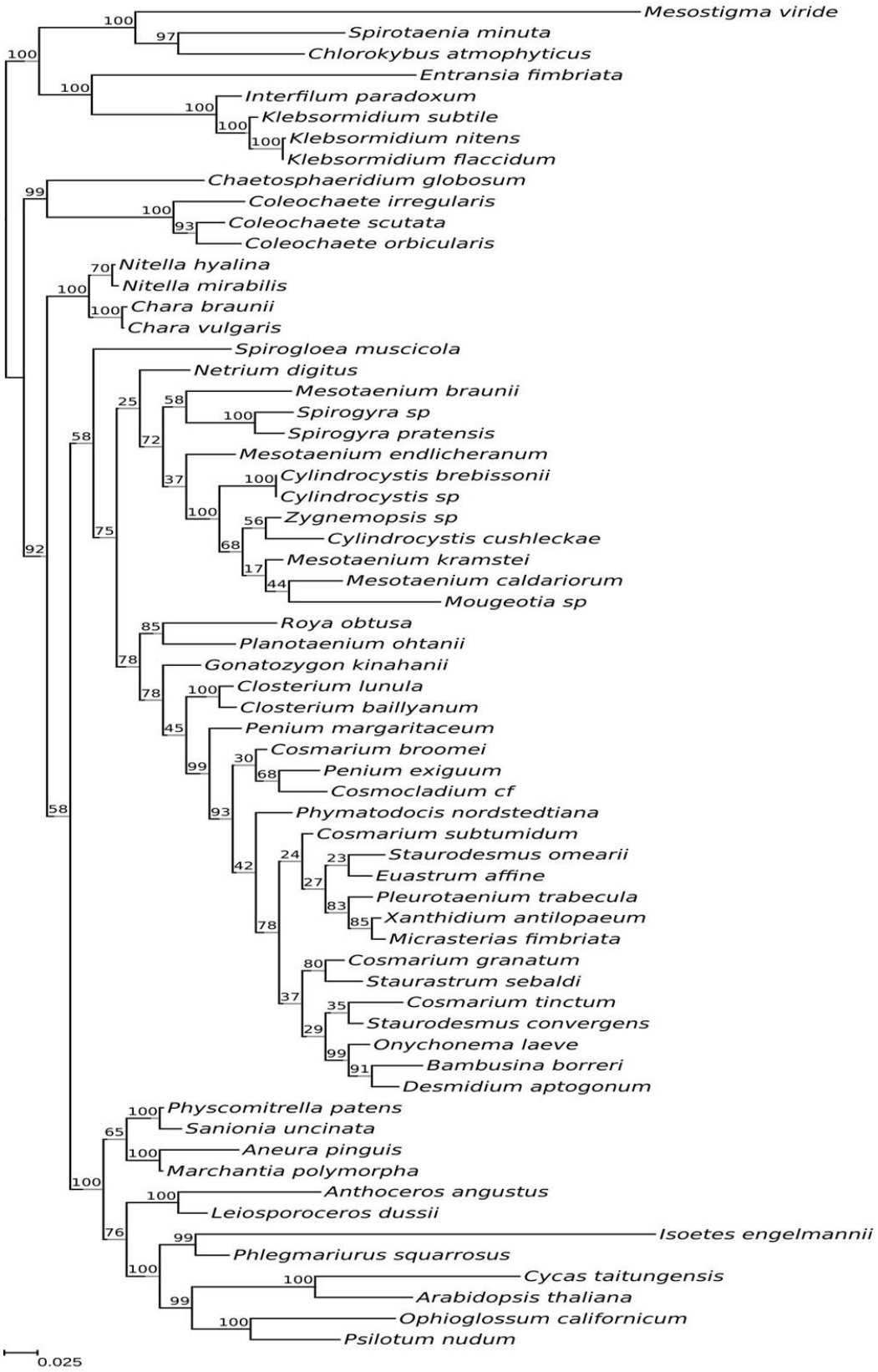


Figure A38 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 20% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -106208. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

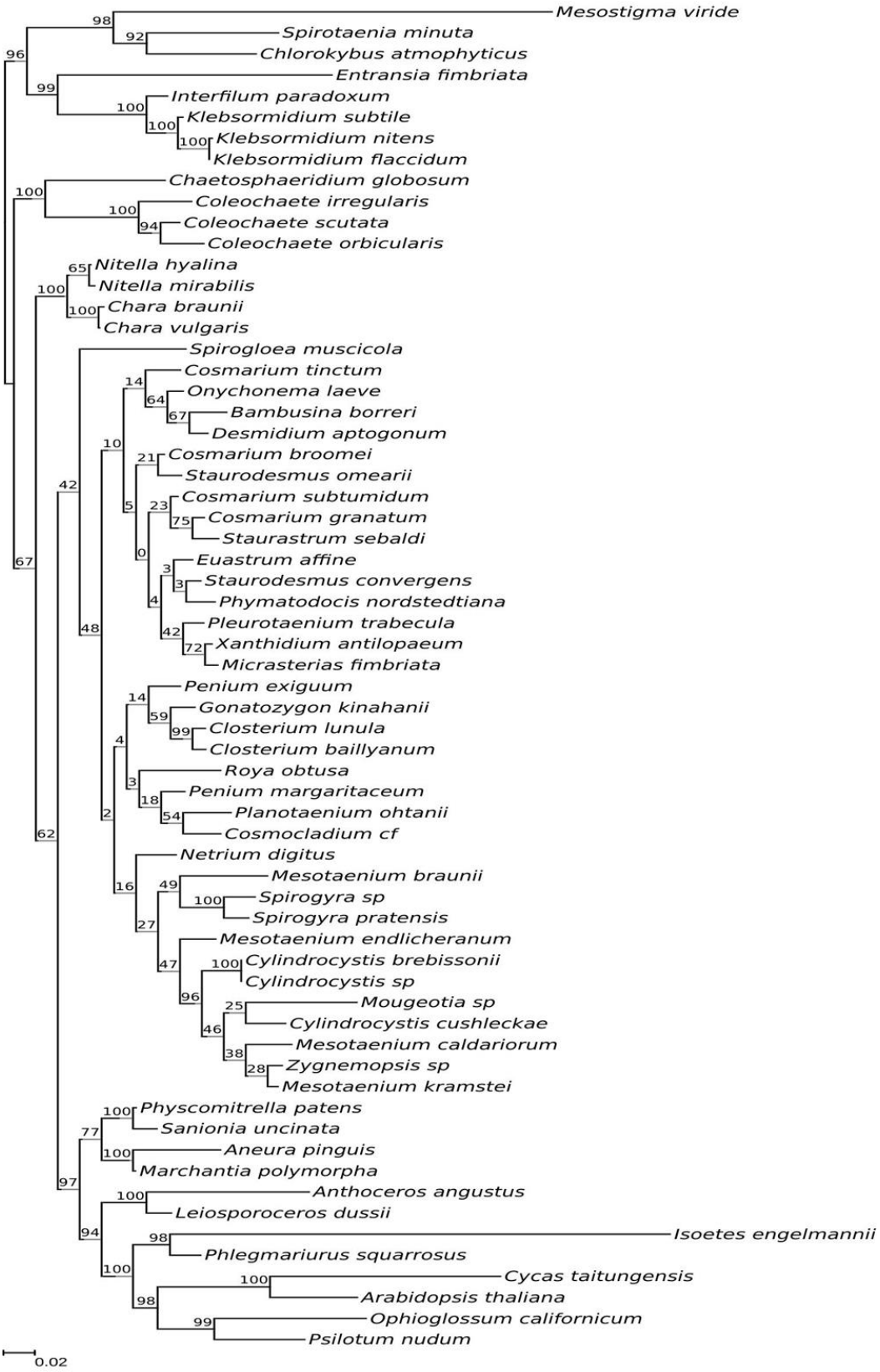


Figure A39 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 30% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -76901. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

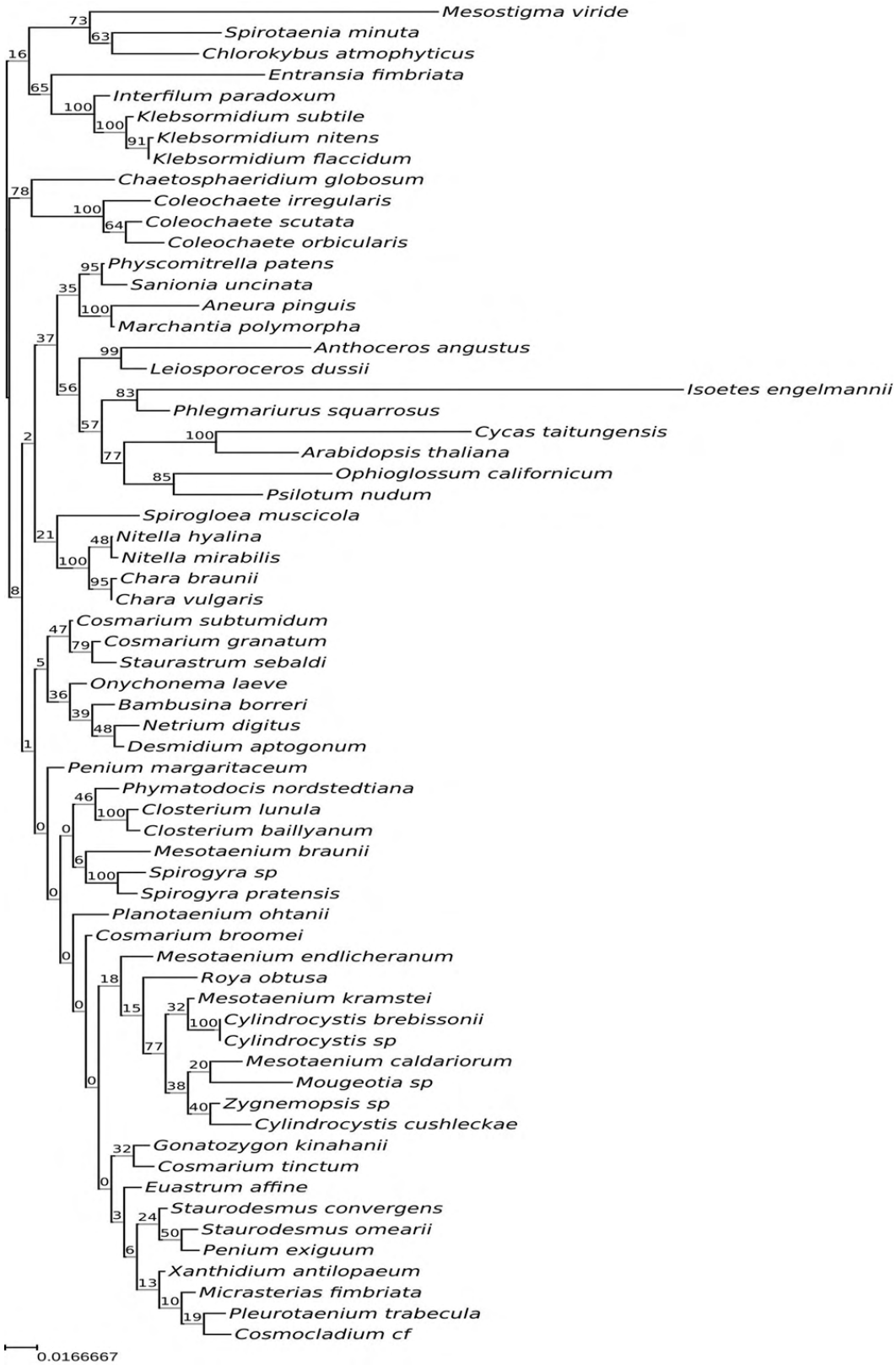


Figure A40 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 40% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -52561. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

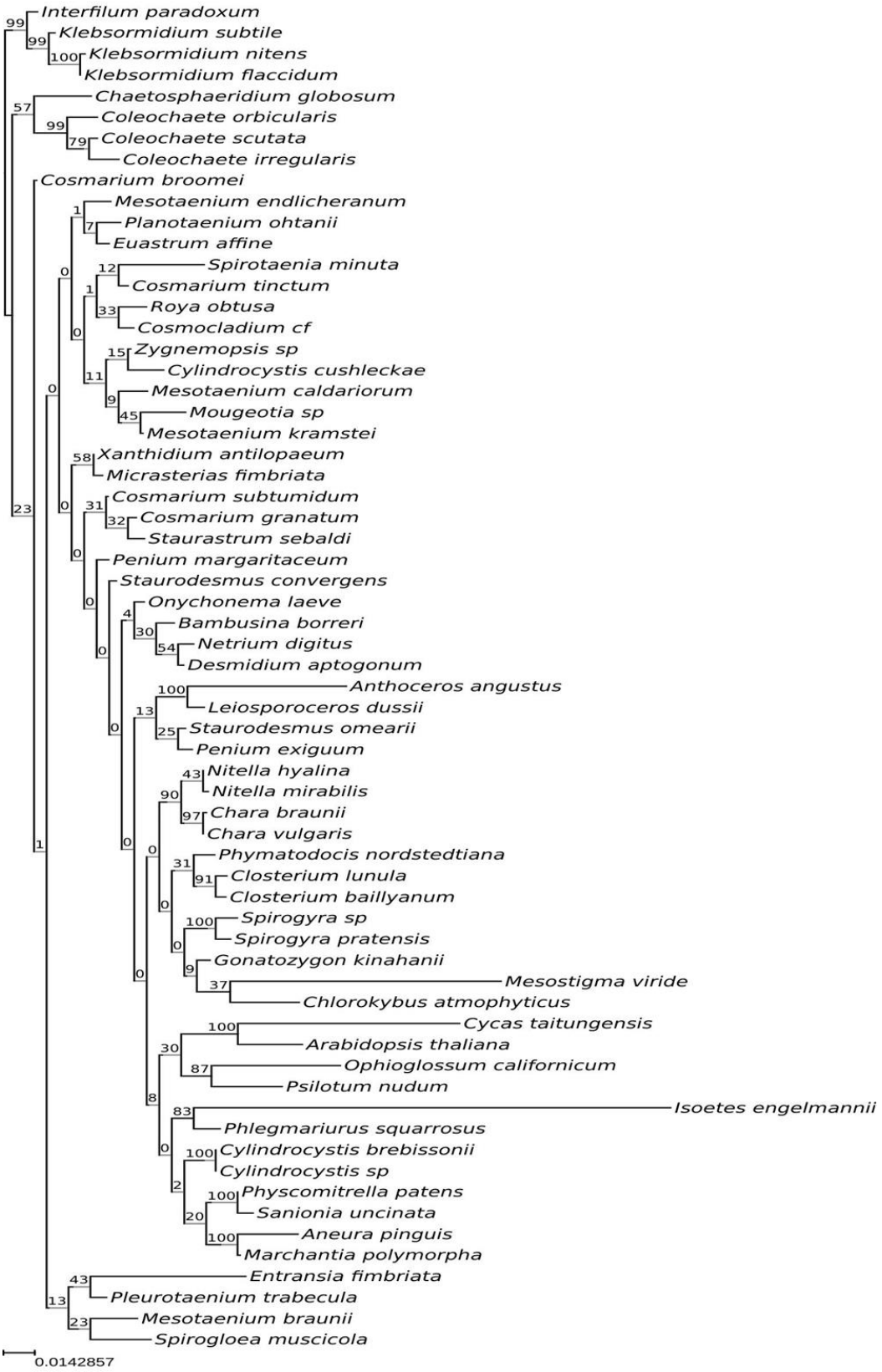


Figure A41 - Optimal maximum-likelihood tree reconstructed from 64 concatenated mitochondrial proteins after removal 50% of the fastest-evolving sites determined according to the observed variability score. Phylogeny comprising 40 taxa inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -34933. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

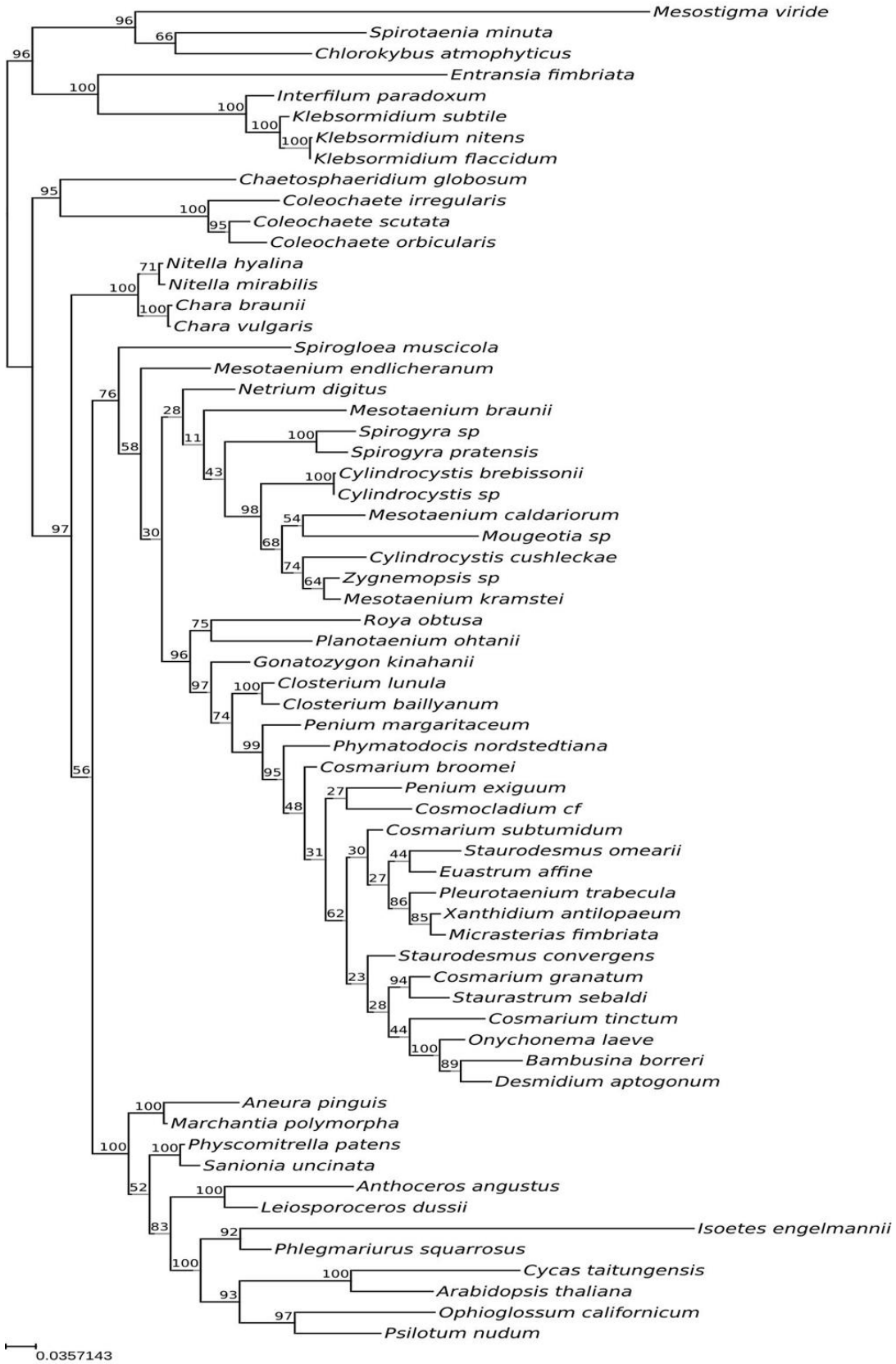


Figure A42 - Optimal maximum-likelihood tree comprising 40 taxa reconstructed from buried-sites partition of the 64 concatenated mitochondrial proteins. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -115910. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

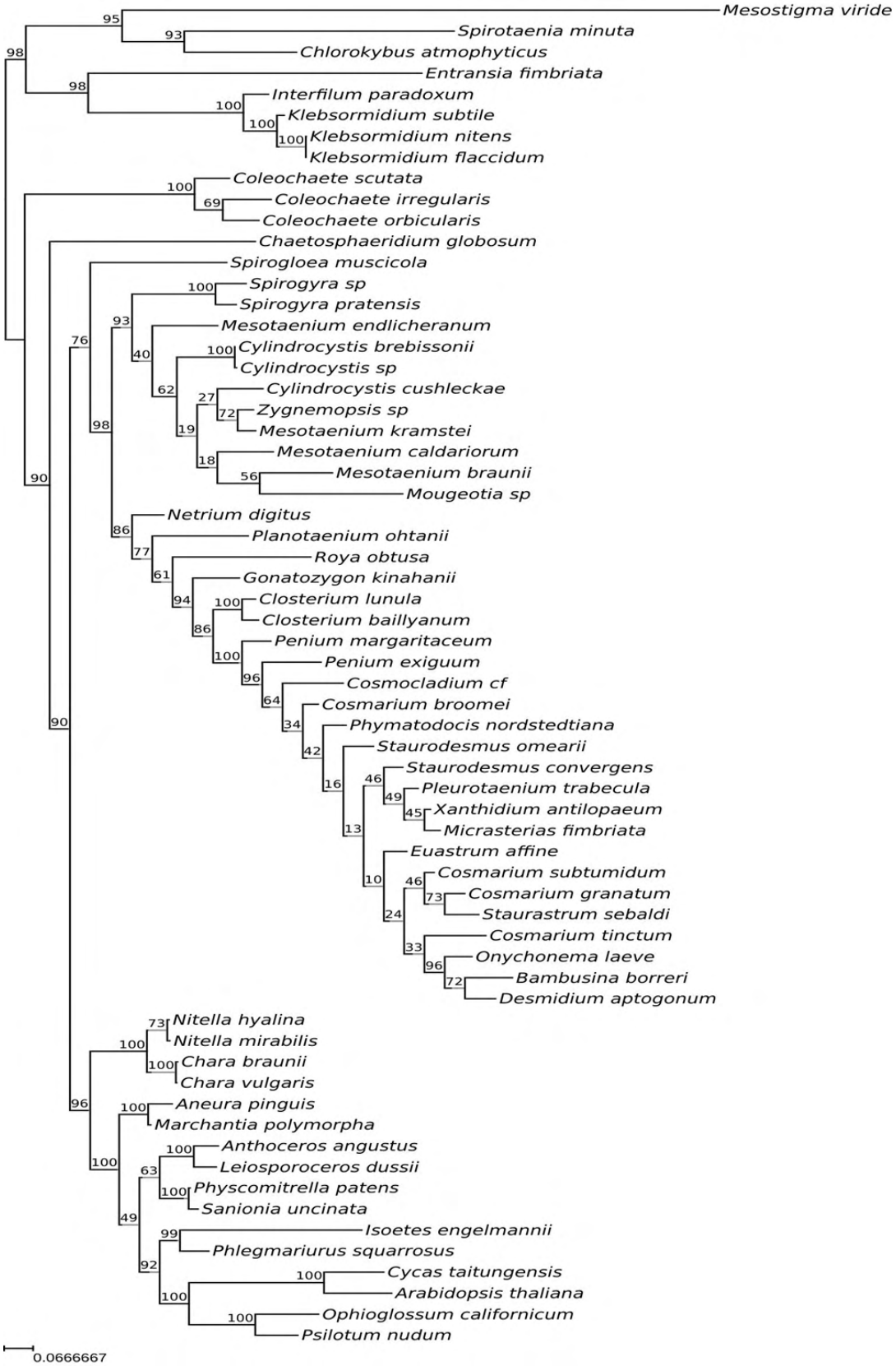


Figure A43 - Optimal maximum-likelihood tree comprising 40 taxa reconstructed from exposed-sites partition of the 64 concatenated mitochondrial proteins. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -59,799. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

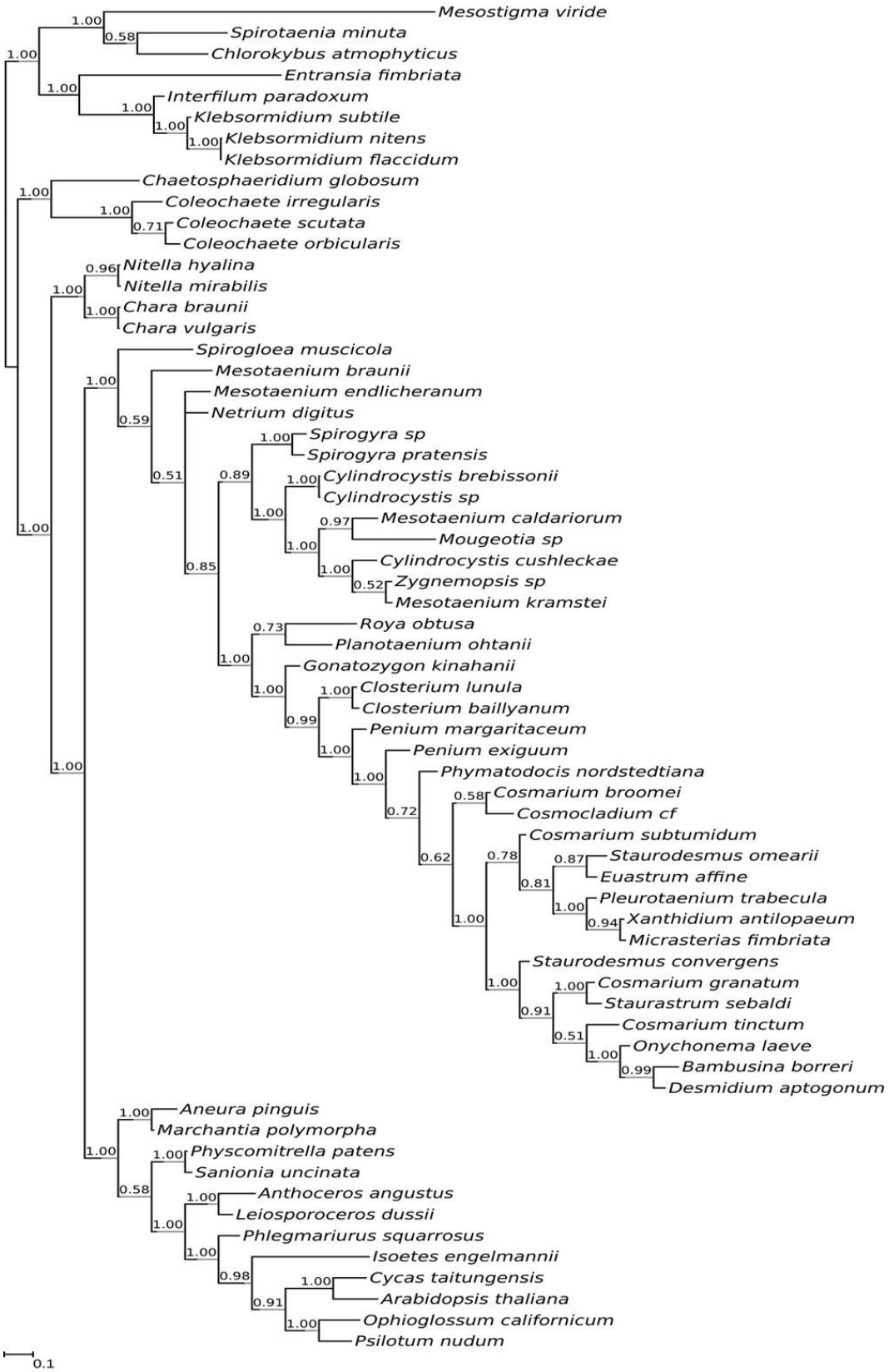


Figure A44 - Phylogeny comprising 40 taxa inferred from the buried-sites partition of the 64 concatenated mitochondrial proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of a composition-heterogeneous MCMC analysis using the CAT model and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{CAT}$). Marginal likelihood, *LML* = -97934. Node support values are Bayesian posterior probabilities.

Figure A45 - Phylogeny comprising 40 taxa inferred from the exposed-sites partition of the 64 concatenated mitochondrial proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of a composition-heterogeneous MCMC analysis using the CAT model and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{CAT}$). Marginal likelihood, *LML* = -55930. Node support values are Bayesian posterior probabilities. Node support values are Bayesian posterior probabilities.

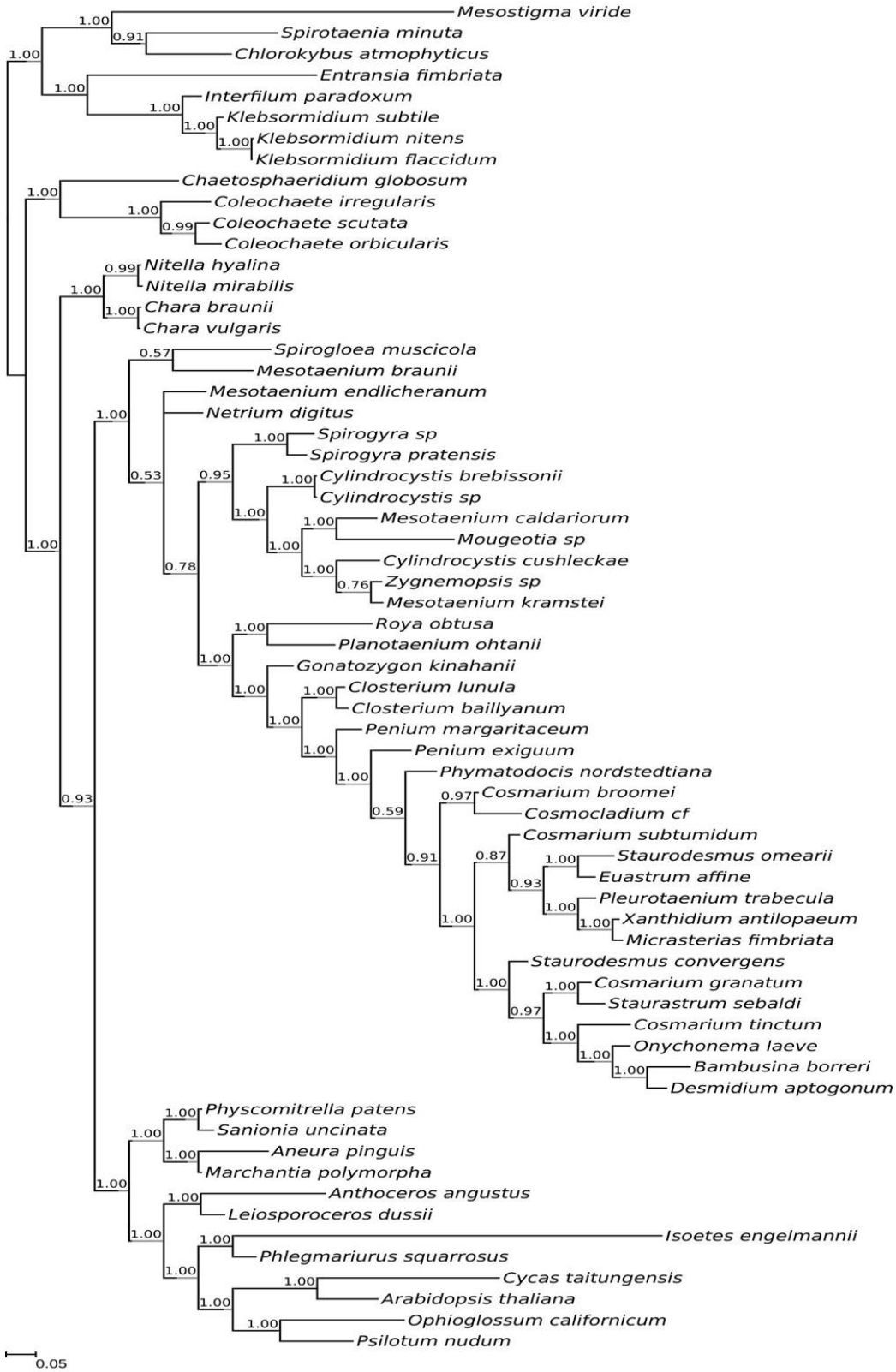


Figure A46 - Phylogeny comprising 64 taxa inferred from the buried-sites partition of the 40 concatenated mitochondrial proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of the best marginal likelihood run using the composition tree-heterogeneous model, NDCH2, and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{NDCH2}$). The model was a good statistical fit to the data **(X^2 *p-value* = *0.99*).** Marginal likelihood, *LML* = -112504. Node support values are Bayesian posterior probabilities.

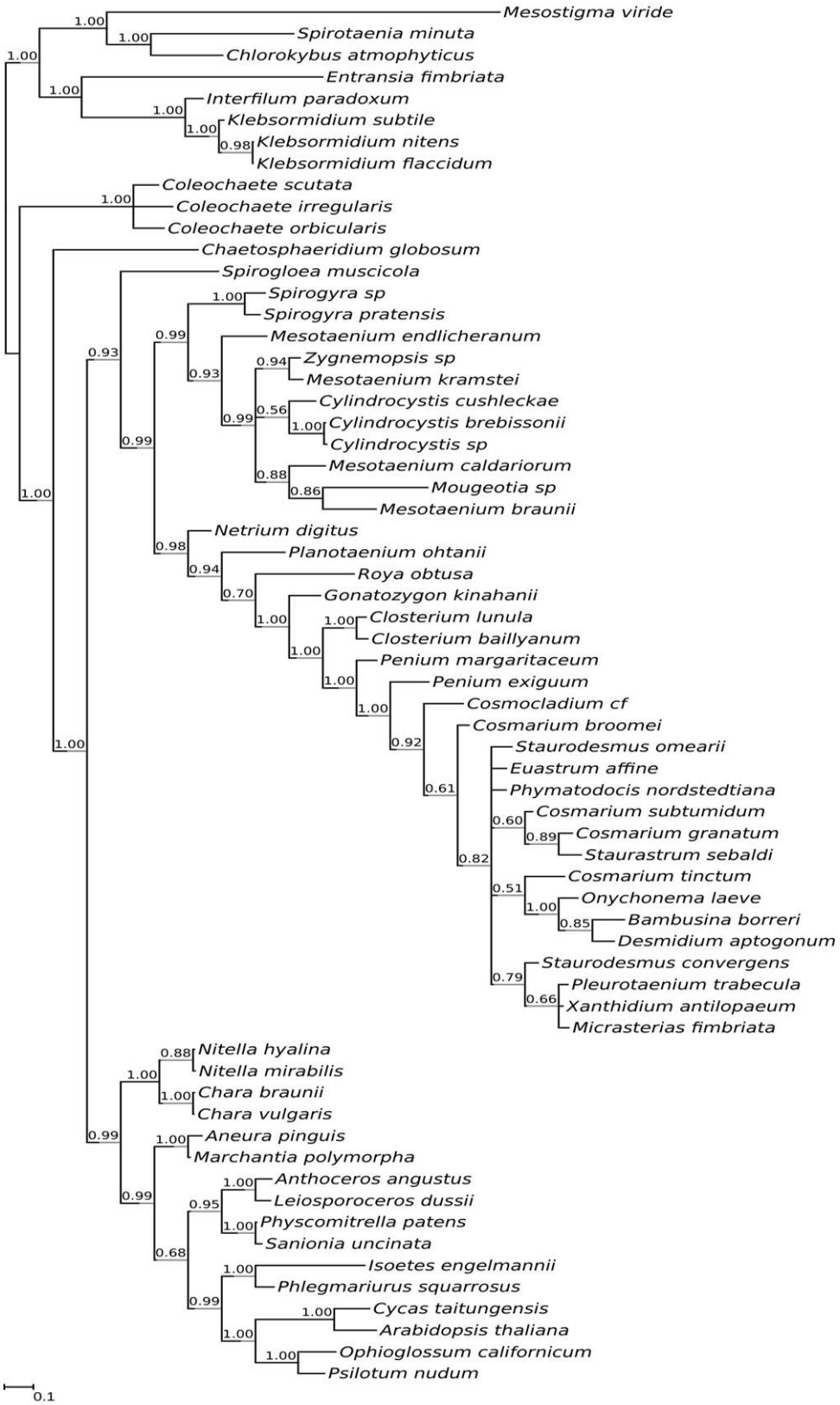


Figure A47 - Phylogeny comprising 64 taxa inferred from the exposed-sites partition of the 40 concatenated mitochondrial proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of four independent MCMC runs using the composition tree-heterogeneous model, NDCH2, and data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{NDCH2}$). The model was a good statistical fit to the data ($\boldsymbol{X^2}$ ***p-value = 0.55***). Marginal likelihood, *LML* = -59373. Node support values are Bayesian posterior probabilities.

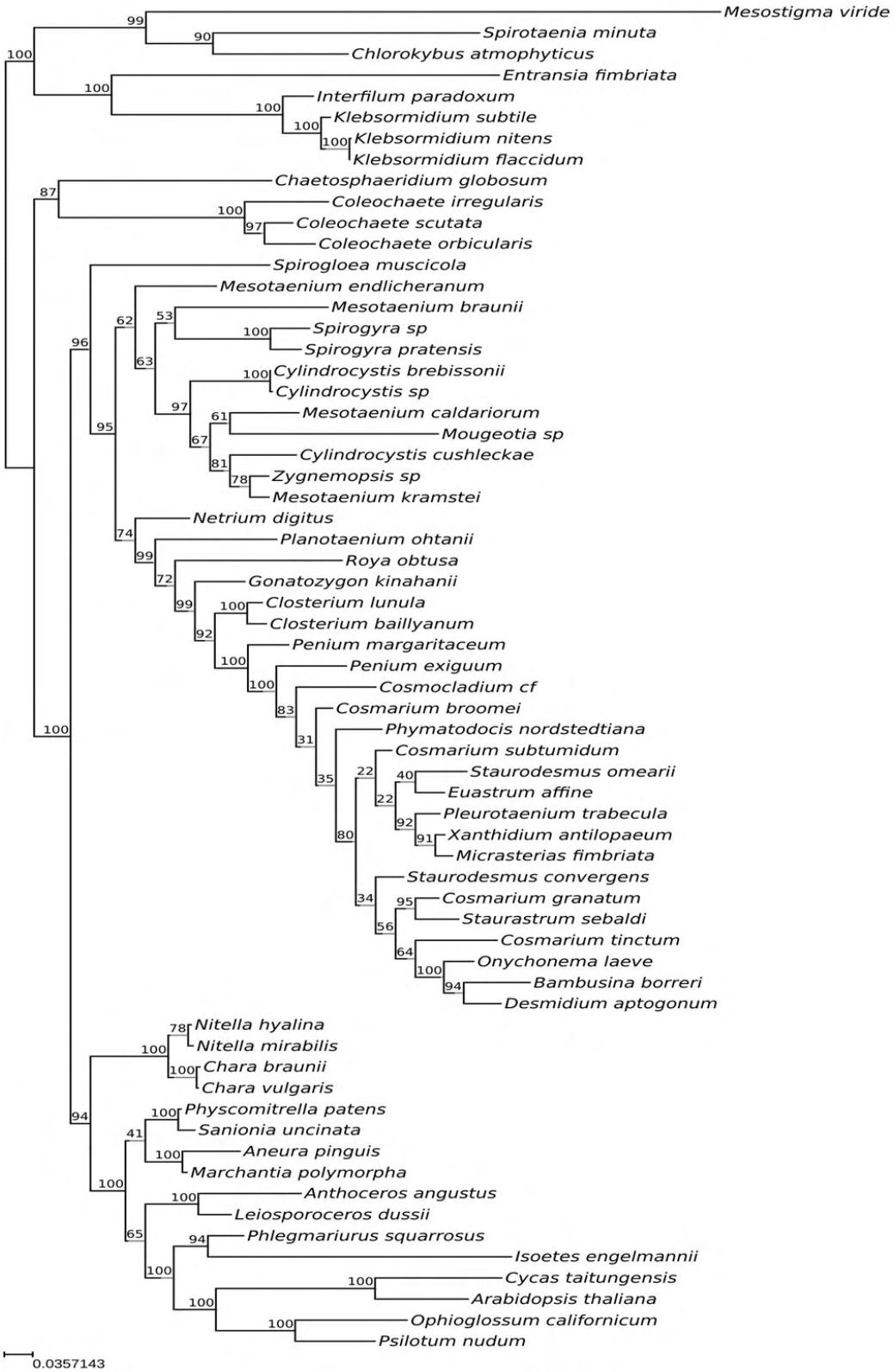


Figure A48 - Optimal maximum-likelihood tree comprising 64 taxa reconstructed from 40 concatenated mitochondrial proteins, after removal of the composition-heterogeneous sequences. Heterogeneous sequences were identified using the MPTMS coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -162286. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

Figure A49 - Optimal maximum-likelihood tree comprising 71 taxa reconstructed from 84 concatenated chloroplast proteins, after removal of the composition-heterogeneous sequences. Heterogeneous sequences were identified using the MPTMS coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -357298. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

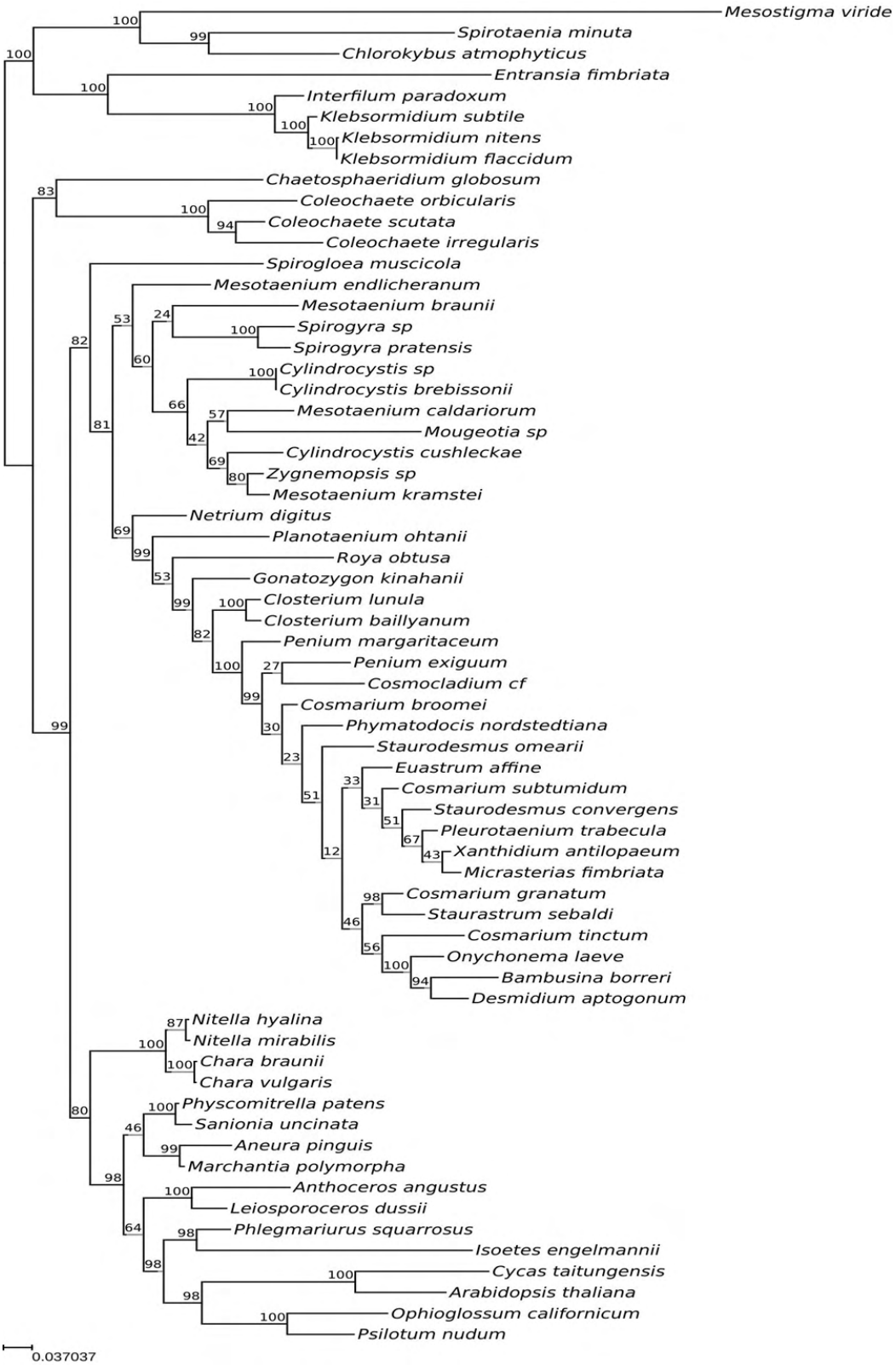


Figure A50 - Optimal maximum-likelihood tree comprising 64 taxa reconstructed from 40 concatenated mitochondrial proteins, after removal of the composition-heterogeneous sequences. Heterogeneous sequences were identified using the χ^2 test of compositional homogeneity coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -150009. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

Figure A51 - Optimal maximum-likelihood tree comprising 71 taxa reconstructed from 84 concatenated chloroplast proteins, after removal of the composition-heterogeneous sequences. Heterogeneous sequences were identified using the χ^2 test of compositional homogeneity coupled with the Benjamini-Hochberg p-value adjustment procedure. Phylogeny was inferred using a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{est}$). Log likelihood, L = -352991. Support values at nodes are maximum-likelihood bootstraps calculated from 300 replicates using the same model.

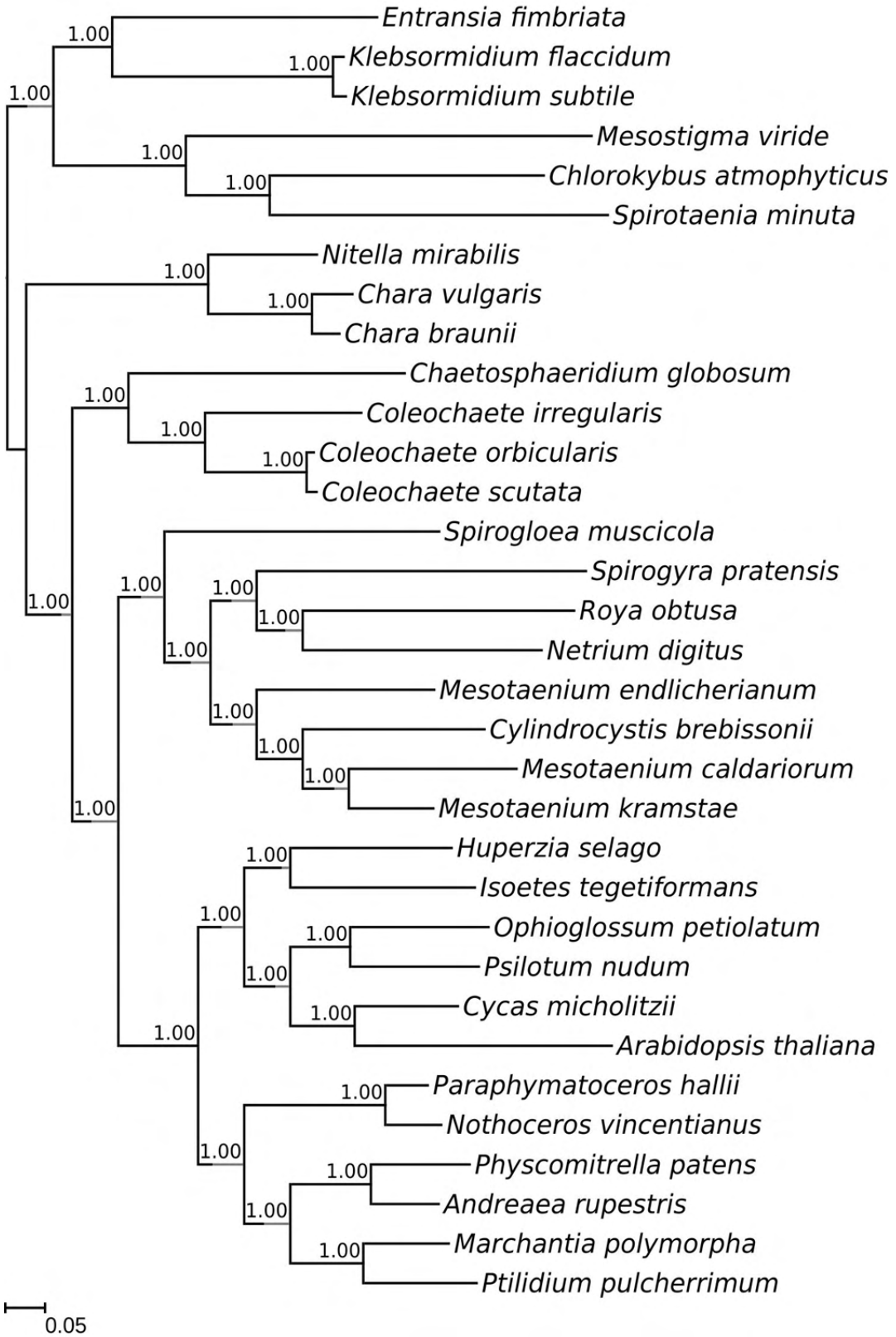


Figure A52 - Phylogeny comprising 33 taxa inferred from 64 concatenated nuclear proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of four independent MCMC runs using the composition tree-heterogeneous model, NDCH2, and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{NDCH2}$). The model was not a good statistical fit to the data (X^2 p-value = 0.02). Marginal likelihood, *LML* = -417773. Node support values are Bayesian posterior probabilities.

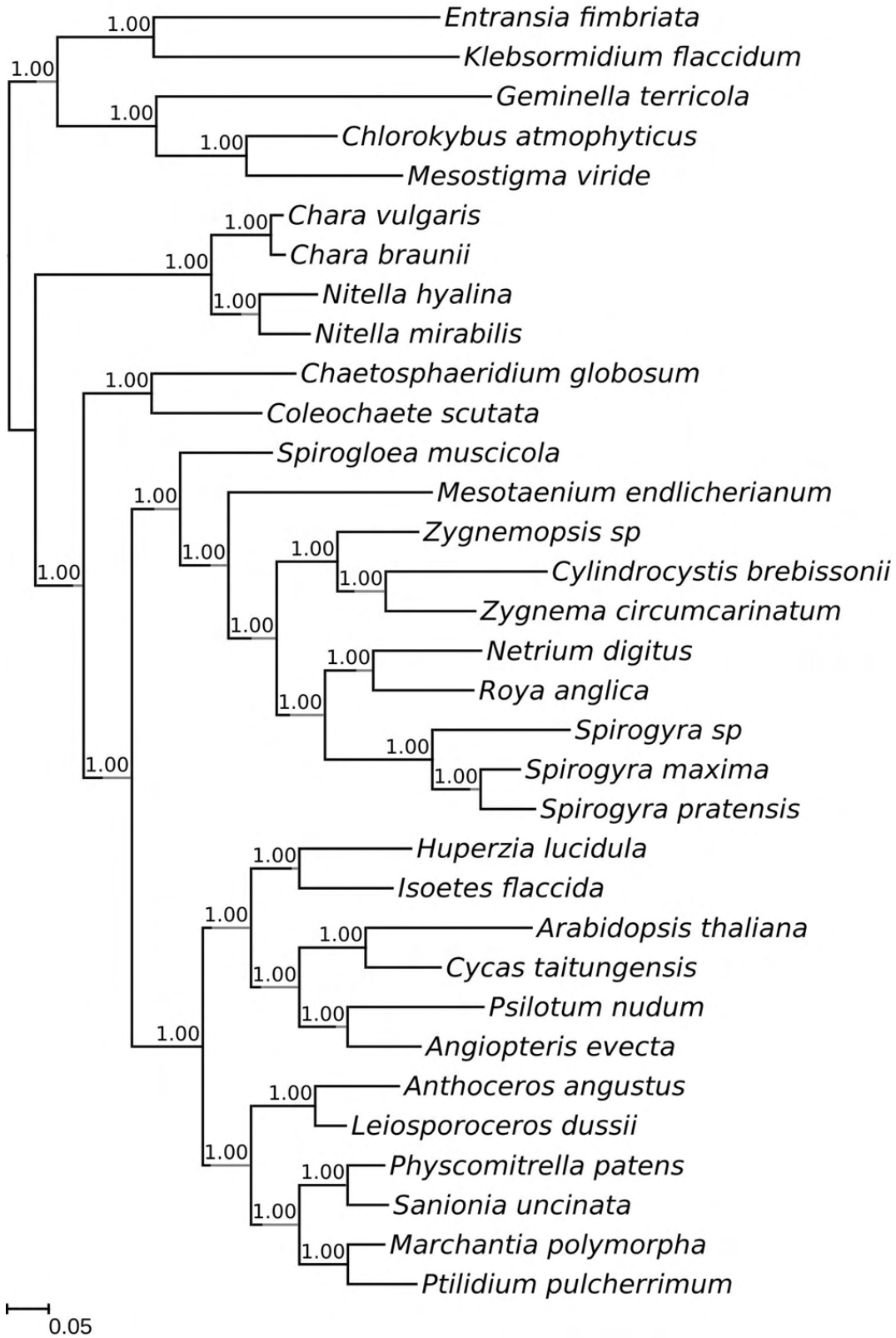


Figure A53 - Phylogeny comprising 33 taxa inferred from 84 concatenated chloroplast proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of four independent MCMC runs using the composition tree-heterogeneous model, NDCH2, and a estimated data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{NDCH2}$). The model was a good statistical fit to the data (X^2 p-value = 0.1). Marginal likelihood, *LML* = -322044. Node support values are Bayesian posterior probabilities.

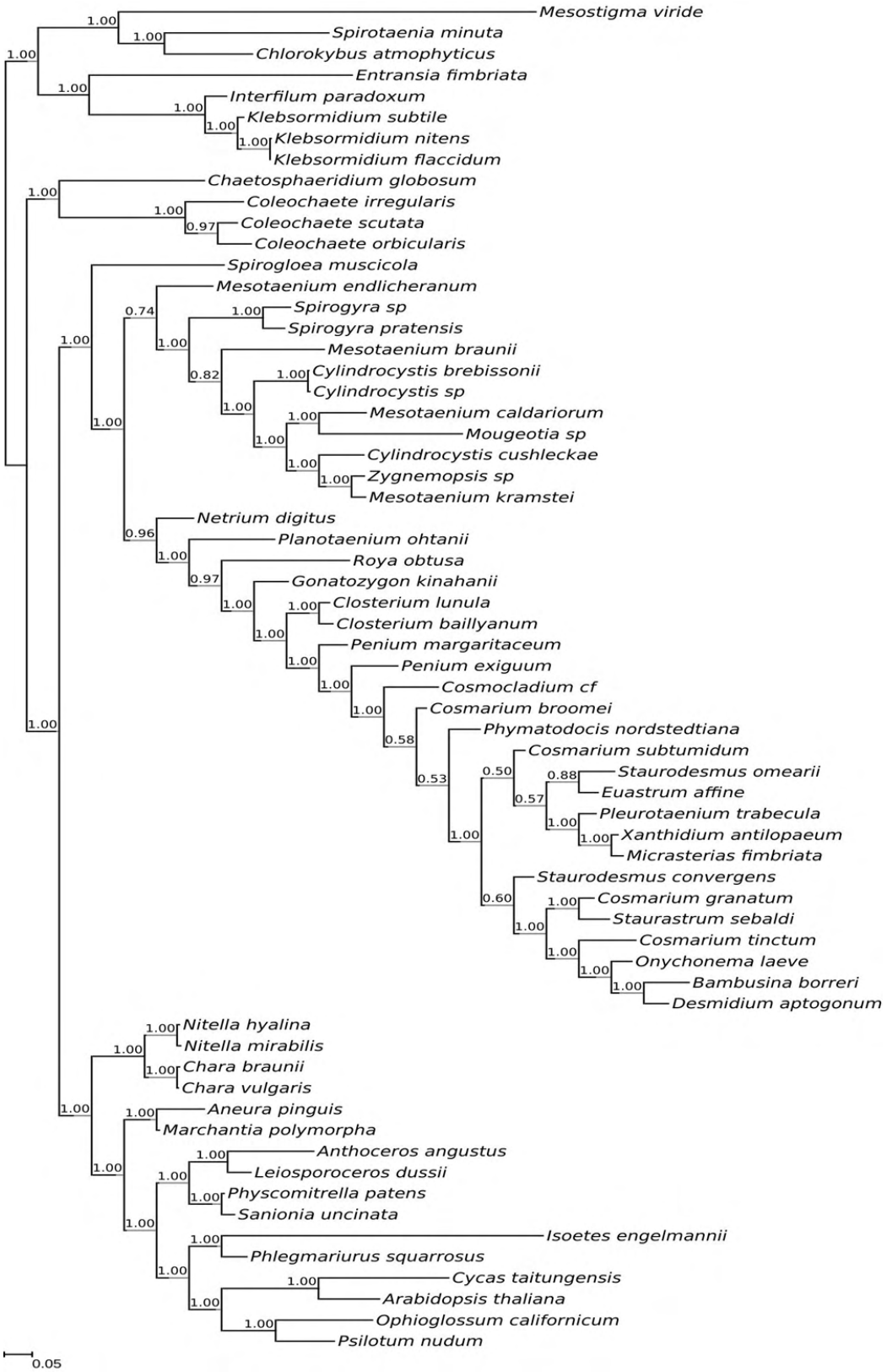


Figure A54 - Phylogeny comprising 64 taxa inferred from 40 concatenated mitochondrial proteins. Majority-rule consensus tree of 10,000 trees obtained from the posterior distribution of the best marginal likelihood run using the composition tree-heterogeneous model, NDCH2, and a data-specific substitution rate model ($GTR_{data}+\Gamma_4+F_{NDCH2}$). The model was a good statistical fit to the data (X^2 p-value = 0.5). Marginal likelihood, *LML* = -174725. Node support values are Bayesian posterior probabilities.

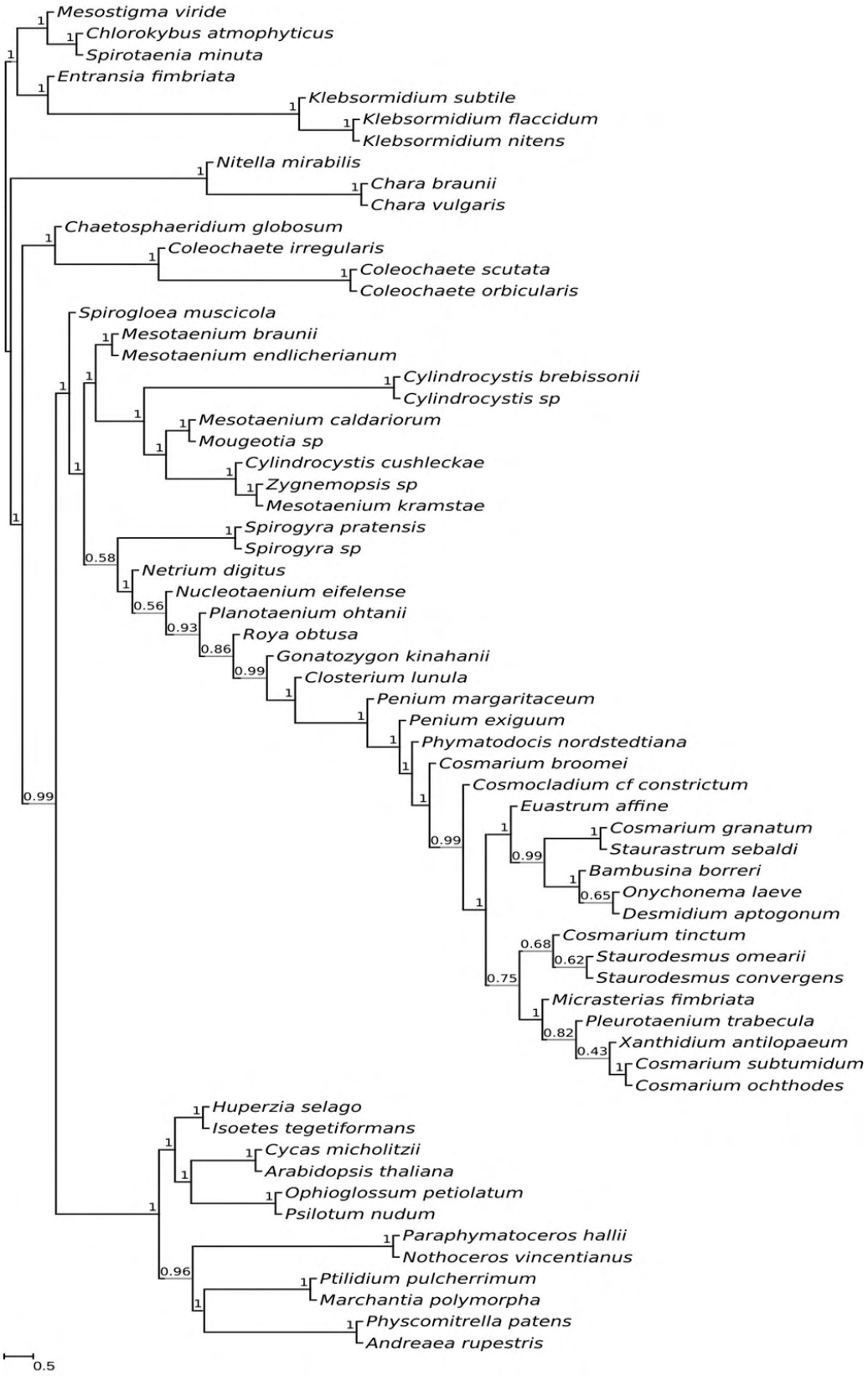


Figure A55 - Phylogeny comprising 64 taxa inferred from 409 nuclear proteins. Multispecies coalescent tree estimated from the optimal maximum-likelihood trees inferred from the 409 nuclear proteins. Node supports were given by posterior probabilities computed from the transformation of quartets percentage in gene trees that agree or not with a branch.

Chapter V

General Discussion

5.1 General Discussion

The main aim of this thesis was to investigate the use of better-fitting amino-acid substitution models that are estimated directly from the study data and to evaluate strategies to mitigate the effects of systematic bias in phylogenetic analyses. Software tools used to calculate data-specific substitution models were evaluated with the aim of identifying those tools that offer the best trade-off between model accuracy and the cost in time to calculate the model. Furthermore, procedures for assessing the substitution process heterogeneity among lineages were explored, and a novel methodology to identify and reduce bias caused by among-lineage heterogeneity was proposed. All analyses were conducted using protein sequence data, with the ultimate objective of reconstructing the relationships among Streptophyta, and specifically, identifying charophyte lineage most closely related to land plants. Protein sequences were chosen because they are usually better suited to reconstructing deep phylogenies (land plant evolved approximately 470MYA) due to their higher number of character states (20 amino acids) and slower rates of evolution than nucleotide sequences (Lemieux *et al.*, 2016; Cox, 2018; Puttick e*t al.*, 2018).

Most phylogenetic protein sequence studies rely on pre-computed empirical models (e.g., LG, WAG, cpREV), with researchers selecting the model for analysis with the best fit from among a set of commonly-used empirical models. The best-fitting empirical model is expected to be a good fit, or at least a good enough fit, to the data to be able to calculate an accurate phylogenetic tree. However, unless one uses methods for assessing the absolute fit (which is not that common), the adequacy of the model for analysis of the data cannot be known, such that even the best-fitting model can have a poor fit to the data. In **Chapter II**, phylogenies derived from simulated data using data-specific substitution models had a better log-likelihood and longer branches than those trees inferred using the best-fitting empirical models. This indicated that data-specific models better-fit the data and infer branch lengths more accurately than the empirical models which tend to underestimate the number of substitutions. Although the simulation procedure was relatively simple (e.g., it was site- and tree-homogeneous), possibly limiting the applicability of these findings to complex real data, the analyses of empirical data sets in **Chapter II** and **IV** confirm that data-specific models fit better the data than empirical models, corroborating the simulation results. Additionally, some of the trees resulting from the re-analysis of data from published studies using better-fitting data-specific models showed topological rearrangements when compared to the original published trees. Furthermore, as shown in **Chapter II,** more sophisticated models, for

instance, those that are partitioned or use empirical mixture models (LG4X; Le *et al*., 2012), fail to compensate for the use of inadequately fitting, albeit best-fitting, empirical models compared to the analysis of single-partitioned data using data-specific models, which showed a better fit to the data.

If the method of choosing a best-fitting empirical model is done using a particular software tool (e.g. ModelFinder), the empirical models implemented in the tool may be inadequate for the data under analysis. For example, in the analyses of the streptophyte mitochondrial data (**Chapter IV**), the selected best-fitting model was the chloroplast cpREV model because all mitochondrial models implemented in ModelFinder were metazoan-derived and did not fit the streptophyte mitochondrial data well. A potentially much better-fitting model is the stmtREV model estimated from streptophyte mitochondrial data, but this model is not among those tested by ModelFinder. Nevertheless, stmtREV was estimated using a data set dominated by land plant taxa with a limited representation of charophyte algae potentially making it a suboptimal solution compared to a data-specific model. While the lack of amino-acid substitution models has driven the estimation of new models, such as stmtREV, more comprehensive metazoan-derived mitochondrial models (Vinh *et al*., 2017), the chloroplast gcpREV model (streptophyte-derived; Cox and Foster, 2013), and nuclear models (Q.bird, Q.plant, Q.yeast, Q.insect, Q.mammal; Minh *et al*., 2021) from the analyses presented here it appears more straightforward and more efficient to calculate a data-specific model. Given the superior fit of data-specific models in the analyses of simulated and empirical data, whether nuclear, organellar or across several taxonomic groups, it is unlikely that the use of an empirical model can be justified. This holds regardless of how well a published empirical model might fit the data, as data-specific models are likely always to exhibit a better fit or, at minimum, a similar fit in cases where the data are 'very close' to being accurately modelled by one of the commonly-used empirical models. The analyses of short alignments, 400 sites, indicated that models calculated from these data sets are less accurate than those estimated from longer alignments, but they did not perform worse than empirical models. Moreover, the time required to select a model from available candidates can exceed the time needed to compute a data-specific model. Indeed, all ML based solutions were reasonably fast, with IQ-TREE showing the best trade-off between the model accuracy and speed calculation. The results of the Bayesian inference solution, implemented in P4, had the longest runtime. However, the fit of these Bayesian-estimated models improved significantly with longer alignments. Perhaps, if calculated from longer alignments (>8,000 sites long), the resulting models would produce the most accurate trees and would have the best fit. Nevertheless, even

if it is possible to calculate better models using this method, the time cost is prohibitive without a relevant gain compared with the ML methods.

As to be expected, the substitution rates estimated for a data-specific model are more accurate (better-fitting) than those of an empirical model, as they are optimised for the data under analysis. Consequently, data-specific models have a greater robustness against biases caused by using poorer-fitting models. Nevertheless, the benefits of employing data-specific models are limited as they address only the substitution rates in a tree-homogeneous and rate reversible context, while other sources of systematic bias are not addressed (Yang and Roberts, 1995; Foster and Hickey, 1999; Foster, 2004; Jermiin *et al.*, 2004; Blanquart and Lartillot, 2006). In **Chapter III**, the accuracy of methods for assessing compositional and rates heterogeneity among lineages was explored. The underlying motivation for these analyses rests on the assumption that identifying and excluding tree-heterogeneous sequences should reduce their distorting effects on tree inference. In addition, after removing heterogeneous sequences it is anticipated that the remaining data are tree-homogeneous and can be accurately analysed using a tree-homogeneous model without violating the model's assumptions of stationarity and among-lineage homogeneity and thereby improving the overall model fit and the accuracy of the inferred tree. The matched-pairs tests enable identification of composition- and/or rate-heterogeneous sequences, whereas the χ^2 test for compositional homogeneity among lineages can identify composition-heterogeneous sequences. The matched-pairs test of marginal symmetry (MPTMS) exhibited greater statistical power than the matched-pairs test of symmetry (MPTS) and the matched-pairs test of internal symmetry (MPTIS). The χ^2 test identified more composition-heterogeneous sequences than the matched-pairs tests in the analyses of the empirical data in **Chapter IV** therefore suggesting it had a greater statistical power than the matched pairs tests. However, this test was not evaluated using simulated data and consequently the magnitude of the rate of Type I and Type II errors was not assessed. The phylogenies inferred from the filtered data sets in **Chapter III** and **IV** where composition-heterogeneous sequences were removed using the MPTMS and the χ^2 test resulted in topological rearrangements. Such topological rearrangements included the Setaphya clade, bryophytes as monophyletic or a paraphyletic group with the lycophytes sister to hornworts, and the monophyly of the Archaeplastida. These topological rearrangements could be seen as indicative of the efficacy of the approach taken given that they are congruent with current phylogenetic understanding and presumed most likely to be correct.

By contrast, the MPTS had systematically lower statistical power than the MPTMS and the χ^2 tests and consistently identified less composition-heterogeneous sequences. In addition, MPTS had more false positives in some analyses and the subsequent inference analyses of the filtered data sets did not recover an improved topology in most cases. The analyses of composition- and rate-heterogeneous data result in similar comparisons, with MPTS performing poorly compared with the MPTMS. Indeed, the analyses presented in **Chapter III** demonstrated that rate-heterogeneous sequences are often difficult to identify using the MPTS or the MPTIS and potentially diminishing an advantage that MPTS could have over the MPTMS and χ^2 test, that is, identifying rate-heterogeneous sequences. Overall, these results appear to align with previous findings suggesting that MPTMS exhibits greater statistical power than MPTS, and the MPTIS the lowest (Naser-Khdour *et al.,* 2019; Jermiin e*t al.,* 2020).

Sites or loci often evolve under different constraints and therefore in analyses of among-lineage heterogeneity it is recommended to conduct the matched-pairs tests on data partitions that have evolved under the same evolutionary constraints (Jermiin *et al.,* 2008). This data partitioning enables a more precise characterisation of the conditions under which the sites evolve. However, selecting the best partitioning criteria implies a trade-off between the number of partitions and their size. More partitions but smaller size can make the tests inefficient, while larger partitions increase the statistical power although it may not guarantee a reliable characterisation of the heterogeneity among lineages if for instance sites that evolve at different rates are analysed together (Jermiin *et al.,* 2008). Partitioning the data according to individual proteins was effective for the data analysed here, in particular when using the MPTMS, since the subsequent analyses of the mitochondrial data filtered of composition-heterogeneous sequences showed topological rearrangements identified as likely to be correct (**Chapter III and IV**). However, single-protein partitions may be too small for analyses conducted using the MPTS, since the power of this test was lower than that in the MPTMS analyses, resulting in fewer identified composition-heterogeneous sequences and an absence of topological improvements. When using different the K-means and ModelFinder partitioning criteria, trees inferred from the data sets filtered according to the MPTMS, and one data set filtered using the MPTS, also exhibited topological rearrangements identified as likely correct. These results demonstrate that these partitioning criteria are suitable for the MPTMS and MPTS. Indeed, the MPTS appears to be more effective when using these partitioning criteria. In the analyses of single-partitioned data the MPTMS and MPTS demonstrated increased sensitivity since many sequences were identified as heterogeneous.

Treating multiple partitions, especially those corresponding to different loci, as a single partition inevitably joins loci together that evolved under distinct constraints thereby resulting in the identification of many more sequences as heterogeneous. However, despite the MPTIS showing an overall low sensitivity and consistently identifying few or no rate-heterogeneous sequences, those identified and removed from the single-partitioned data resulted in the inference of trees with topological improvements (**Chapter III**). One criticism of matched-pairs tests is their high statistical power enables the detection of small violations that may not significantly impact phylogenetic analyses (Yang, 2014). Perhaps, in the case of the MPTIS, unlike the MPTS and the MPTMS, due to its low statistical power only sequences with strong rate-heterogeneity that can impact phylogenetic analyses are identified. Moreover, at a time when data availability is often not a concern, excluding falsely identified tree-heterogeneous sequences from the analysis may not be as detrimental to the result if there are still sufficient data to obtain a robust phylogeny, which is the case when using the MPTIS. Nevertheless, the interaction of different loci can lead to complex outcomes, and therefore their interpretations should be approached carefully.

The matched-pairs tests and the χ^2 test used for identifying taxa that evolve under among-lineage heterogeneity fell within a statistical multi-dimensional problem where the statistical significance (p-values) for each pair-wise comparison had to be adjusted. Analyses of simulated and empirical data in **Chapter III** using the Benjamini p-value correction methods, particularly the Benjamini-Hochberg, revealed more false positives but identified more among-lineage heterogeneous sequences than the Bonferroni and Holm methods, thereby exhibiting greater statistical power. Furthermore, analyses of the data sets filtered using the Benjamini-Hochberg method more often led to topological improvements regarding the relationships among bryophytes and Archaeplastida than other p-value correction methods. Consequently, the Benjamini-Hochberg p-value correction method is recommended when using matched-pairs tests.

Contrary to traditional taxonomy of the land plants, the analyses presented here tend to support the monophyly of the bryophytes with the clade Setaphyta uniting the mosses and liverworts to the exclusion of hornworts. These phylogenetic inferences were obtained after filtering tree-heterogeneous sequences (**Chapter III and IV)** confirming the effect of systematic bias in the analyses of the mitochondrial data sets when among-lineage heterogeneity is ignored. Compositional heterogeneity among lineages in mitochondrial data has been previously shown to bias the inference of the relationships among the bryophytes (Liu *et al.*, 2014; Sousa *et al.*, 2019). Moreover, albeit less pronounced, the analyses of

nuclear and chloroplast data in **Chapter IV** also indicated the effect of among-lineage heterogeneity on the inference of bryophyte relationships most notably through variations of node support. Conversely, the absence of changes in the relationships between land plants and charophyte green algae in the trees resulting from three genomic data sets indicates that the resolution of either Zynematophyceae or Charophyceae as sister-group to land plants is unlikely due to a systematic bias caused by among-lineage heterogeneity (**Chapter IV**). The sister-group relationship between Zygnematophyceae and land plants found in nuclear and chloroplast data sets analyses is congruent with the previous analyses of the nuclear and chloroplast data (e.g., Timme *et al*., 2012; Wickett *et al*., 2014; Lemieux *et al*., 2016; Cheng *et al*., 2019; Leebens-Mack *et al*., 2019). The analyses of the mitochondrial data, especially of the combined data set, recovered Charophyceae as the sister-group to land plants without indication of a biased result. This topology was congruent with analyses of the mitochondrial data within previous studies (e.g., Turmel *et al*., 2007, 2013; Orton *et al*., 2020). However, additional analyses performed in **Chapter IV** revealed conflicting signals across mitochondrial data partitions. Trees inferred from more the slowly evolving sites (i.e. buried-sites partition) and from some single-protein alignments recovered Zygnematophyceae as the sister-group to land plants. Nevertheless, CAT model analyses of the combined data set resulted in the inference of Charophyceae as the sister group of land plants, confirming that the former result was not due to site-heterogeneous composition. If the incongruence with the nuclear and chloroplast analyses found in mitochondrial analyses is due to systematic bias, then the results imply that other biases than those examined in this study are likely contributing factors to the placement of Charophyceae as the sister-group to land plants. On the other hand, if the sister-group relationship of Charophyceae to land plants is a real phylogenetic signal then the mitochondrial genome is evidently chimeric and not congruent with the species tree as a whole. It is possible that the two conflicting signals do indeed result from biological processes and might be indicative of horizontal gene transfer.

Data-specific substitution models had a better fit to the data than empirical models, increasing the accuracy of the inferred protein-based phylogenies and reliableness of the topology and branch lengths. Furthermore, with the availability of time-efficient software capable of calculating accurate data-specific models, there is no longer any justification for relying on empirical models in phylogenetic inference analyses. Indeed, the time needed to calculate a model may be shorter than the time required to choose a model from among an assortment of empirical models. In addition, the best-fitting empirical models may, or may not, fit the data well. Data-specific models combined with extended models or with data

partitioning or data exclusion strategies helped to reduce systematic bias. A particular systematic bias analysed here was that caused by among-lineage heterogeneity, which was shown to affect the relationships among Archaeplastida, land plants, and perhaps the inference of the sister-group of Chordata. The MPTMS combined the Benjamini-Hochberg p-value correction method was able to identify composition-heterogeneous sequences that were biasing these relationships in standard phylogenetic analyses. By contrast, there was no indication that the identification of the charophyte group most closely related to land plants was biased by among-lineage heterogeneity, or other systematic biases caused by differing substitution rates due to the structural position of amino-acid residues in the protein, or the substitution rate among amino-acid sites in general. However, assuming incongruence of the mitochondrial data analyses which placed Charophyceae as the sister-group to land plants, rather than Zygnematophyceae, suggests that current models are not able to account for all evolutionary substitution processes affecting the divergence of land plants from charophytes.

Obtaining the same phylogenetic tree from the analysis of the same data using several different methods may indicate that the result is a correct representation of the species tree. This congruence among analyses may be due to the data being decisive and the inference easy. However, congruence amongst analyses may be because all the results are distorted in the same way due to a specific systematic bias. The emergence of competing hypotheses is usually also an indication that systematic bias inherent in the data. Consequently, both known and unknown properties of the data should be considered and addressed as potential causes of systematic bias if possible. Even if reconstructing a robust phylogeny seems straightforward, one can have greater confidence that the solution is accurate after conducting a comprehensive investigation of potential systematic biases. The analyses conducted in this work illustrate the importance of applying strategies to mitigate the effect of systematic bias inherited in the data. The accuracy of the inferred topologies was improved by using better-fitting data-specific substitution models and methods to remove tree-heterogeneous data or more accurately model variation in substitutions among sites.

General references

Ababneh, F., Jermiin, L. S., & Robinson, J. (2006). Generation of the exact distribution and simulation of matched nucleotide sequences on a phylogenetic tree. *Journal of Mathematical Modelling and Algorithms*, *5*(3), 291–308. https://doi.org/10.1007/s10852-005-9017-y

Aberer, A. J., Krompass, D., & Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, *62*(1), 162–166. https://doi.org/10.1093/sysbio/sys078

Adachi, J., & Hasegawa, M. (1996). Model of Amino Acid Substitution in Proteins Encoded by Mitochondrial DNA. *Journal of Molecular Evolution*, *42*, 459–468. https://doi.org/10.1007/BF02498640

Adachi, J., Waddell, P. J., Martin, W., & Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution*, *50*(4), 348–358. https://doi.org/10.1007/s002399910038

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267-281 in *Second Annual Symposium on Information Theory* (B. N. Petrov, and F. Csaki, eds.). Akademi Kiado, Budapest.

Bayes, T., & Price, null. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418. https://doi.org/10.1098/rstl.1763.0053

Benjamini, Y, & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29* (4), 1165–1188. http://www.jstor.org/stable/2674075

Benjamini, Yoav, & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bishop, M. J., & Friday, A. E. (1985). Evolutionary trees from nucleic acid and protein sequences. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *226*(1244), 271–302. https://doi.org/10.1098/rspb.1985.0096

Bishop, M. J., & Friday, A. E. (1987). Tetrapod relationships: the molecular evidence. *Molecules and Morphology in Evolution: Conflict or Compromise*, 123–139. https://doi.org/10.1016/s0169-5347(00)89004-7

Blanquart, S., & Lartillot, N. (2008). A Site- and Time-Heterogeneous Model of Amino Acid Replacement. *Molecular Biology and Evolution*, *25*(5), 842–858. https://doi.org/10.1093/molbev/msn018

Blanquart, S., & Lartillot, N. (2006). A Bayesian compound stochastic process for modelling nonstationary and nonhomogeneous sequence evolution. *Molecular Biology and Evolution*, *23*(11), 2058–2071. https://doi.org/10.1093/molbev/msl091

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, *19*(7), 1171–1180. https://doi.org/10.1093/oxfordjournals.molbev.a004175

Bollback, J. P. (2005). Posterior Mapping and Posterior Predictive Distributions. *Statistical Methods in Molecular Evolution*. https://doi.org/10.1007/0-387-27733-1_16

Boussau, B. (2013). Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic Biology*. https://doi.org/10.1093/sysbio/syt054

Boussau, B., & Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. *Systematic Biology*, *55*(5), 756–768. https://doi.org/10.1080/10635150600975218

Bowker, A. H. (1948). A Test for Symmetry in Contingency Tables. *Journal of the American Statistical Association*, *43*(244), 572–574. https://doi.org/10.1080/01621459.1948.10483284

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building, in Launer, R. L.; Wilkinson, G. N. (eds.); pp. 201–236. *Robustness in Statistics*, Academic Press. https://doi.org/10.1016/B978-0-12-438150-6.50018-2

Brandley, M. C., Schmitz, A., & Reeder, T. W. (2005). Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Systematic Biology*, *54*(3), 373–390. https://doi.org/10.1080/10635150590946808

Bravo, G. A., Antonelli, A., Bacon, C. D., Bartoszek, K., Blom, M. P. K., Huynh, S., Jones, G., Knowles, L. L., Lamichhaney, S., & Marcussen, T. (2019). Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. *PeerJ*, 7:e6399 https://doi.org/10.7717/peerj.6399

Brinkmann, H. (1999). Archaea sister group of Bacteria? Indications from tree reconstruction artefacts in ancient phylogenies. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a026166

Brinkmann, Henner, & Philippe, H. (2008). Animal phylogeny and large-scale sequencing: progress and pitfalls. *Journal of Systematics and Evolution*, *46*(3), 274. https://doi.org/10.3724/SP.J.1002.2008.08038

Bruno, W. J. (1996). modelling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution*, *13*, 1368–1374. https://doi.org/10.1093/oxfordjournals.molbev.a025583

Bryant, D, & Hahn, M. W. (2020). *The Concatenation Question.* (pp. 3.4: 1–3.4: 23). No commercial publisher—Authors open access book. https://inria.hal.science/PGE

Bryant, D., Galtier, N., & Poursat, M.-A. (2004). Likelihood calculation in molecular phylogenetics. *Mathemathics of Evolution and Phylogeny*, 33–62. http://books.google.com/books?hl=en&lr=&id=VjA8ThtLs7IC&oi=fnd&pg=PA33&dq=Likelihood+calculation+in+molecular+phylogenetics&ots=YqS6_OiCOv&sig=1YR4WtsmGYsSGVgXWFlGIkmkjJ8

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., & Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, *530*(7588), 89–93. https://doi.org/10.1038/nature16520

Cao, Y., Adachi, J., Janke, A., Paabo, S., & Hasegawa, M. (1994). Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. *Journal of Molecular Evolution*, *39*, N/A. https://doi.org/10.1007/bf00173421

Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W., Li, X., Xu, Y., Zhang, Y., Wittek, S., Reder, T., Günther, G., Gontcharov, A., Wang, S., Li, L., Liu, X., Wang, J., Yang, H., Melkonian, M., *et al*. (2019). Genomes of Subaerial Zygnematophyceae Provide Insights

into Land Plant Evolution. *Cell*, *179*(5), 1057-1067.e14. https://doi.org/10.1016/j.cell.2019.10.019

Cox, C. J. (2018). Land Plant Molecular Phylogenetics: A Review with Comments on Evaluating Incongruence Among Phylogenies. *Critical Reviews in Plant Sciences*, *37*(2–3), 113–127. https://doi.org/10.1080/07352689.2018.1482443

Cox, C. J., & Foster, P. G. (2013). A 20-state empirical amino-acid substitution model for green plant chloroplasts. *Molecular Phylogenetics and Evolution*, *68*(2), 218–220. https://doi.org/10.1016/j.ympev.2013.03.030

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., & Embley, T. M. (2008). The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences*, *105*(51), 20356–20361. https://doi.org/10.1073/pnas.0810647105

Crooks, G. E., & Brenner, S. E. (2005). An alternative model of amino acid replacement. *Bioinformatics*, *21*(7), 975–980. https://doi.org/10.1093/bioinformatics/bti109

Crotty, S. M., Minh, B. Q., Bean, N. G., Holland, B. R., Tuke, J., Jermiin, L. S., & Haeseler, A. Von. (2020). GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. *Systematic Biology*, *69*(2), 249–264. https://doi.org/10.1093/sysbio/syz051

Cummins, C. A., & 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*, *60*(6), 833–844. https://doi.org/10.1093/sysbio/syr064

Dang, C. C., Le, V. S., Gascuel, O., Hazes, B., & Le, Q. S. (2014). FastMG: a simple, fast, and accurate maximum likelihood procedure to estimate amino acid replacement rate matrices from large data sets. *BMC Bioinformatics*, *15*, 341. https://doi.org/10.1186/1471-2105-15-341

Dayhoff, M. O., Schwartz, R. M. R., & Orcutt, B. C. B. (1978). A Model of Evolutionary Change in Proteins. *Stat.Wisc.Edu*, 234–236. https://doi.org/10.1.1.145.4315

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, *24*(6), 332–340. https://doi.org/10.1016/J.TREE.2009.01.009

Delsuc, F. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg1603

Delsuc, F. (2008). Additional molecular support for the new chordate phylogeny. *Genesis*. https://doi.org/10.1002/dvg.20450

Doyle, J. J. (1997). Trees within trees: genes and species, molecules and morphology. *Systematic Biology*, *46*(3), 537–553. https://doi.org/10.1093/SYSBIO/46.3.537

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*(7188), 745–749. https://doi.org/10.1038/nature06614

Eck, R. V., & Dayhoff, M. O. (1967). Atlas of Protein Sequence and Structure, 1966. *Systematic Zoology*, *16*(3), 262. https://doi.org/10.2307/2412074

Eck, R. V, & Dayhoff, M. O. (1966). Inference from protein sequence comparisons. *Atlas of Protein Sequence and Structure*, 161–202.

Farris, J. S. (1974). Formal definitions of paraphyly and polyphyly. *Systematic Zoology*, 23(4), 548–554. http://dx.doi.org/10.2307/2412474

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Zoology*, *22*, 240. https://doi.org/10.2307/2412304

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*. https://doi.org/10.1007/bf01734359

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*. https://doi.org/10.1086/284325

Felsenstein J, Felenstein J. 2004. Inferring phylogenies: Sinauer associates Sunderland, MA

Feuda, R., Pisani, D., Rota-Stabelli, O., Lartillot, N., Pett, W., Dohrmann, M., Wörheide, G., & Philippe, H. (2017). Improved modelling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. *Current Biology*, *27*(24), 3864-3870.e4. https://doi.org/10.1016/j.cub.2017.11.008

Finet, C., Timme, R. E., Delwiche, C. F., & Marlétaz, F. (2010). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology*, *20*(24), 2217–2222. https://doi.org/10.1016/j.cub.2010.11.035

Finet, C., Timme, R. E., Delwiche, C. F., & Marlétaz, F. (2012). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Current Biology* (Vol. 22, Issue 15, pp. 1456–1457). *Cell Press*. https://doi.org/10.1016/j.cub.2012.07.021

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604), 309–368.

Fitch, W. M., & 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(5), 579–593. https://doi.org/10.1007/BF00486096

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Zoology*, *19*, 99. https://doi.org/10.2307/2412448

Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology*, *20*, 406. https://doi.org/10.2307/2412116

Fleming, J. F., Valero-Gracia, A., & Struck, T. H. (2023). Identifying and addressing methodological incongruence in phylogenomics: A review. In *Evolutionary Applications*. John Wiley and Sons Inc. https://doi.org/10.1111/eva.13565

Fleming, J. F., & Struck, T. H. (2023). nRCFV: a new, dataset-size-independent metric to quantify compositional heterogeneity in nucleotide and amino acid datasets. *BMC Bioinformatics*, 24(1), 145. https://doi.org/10.1186/s12859-023-05270-8

Fletcher, R. (2000). Practical Methods of Optimization. *No Journal Available*, *N/A*, N/A. https://doi.org/10.1002/9781118723203

Flouri, T. (2018). Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msy147

Foster, P. G., Schrempf, D., Szöllősi, G. J., Williams, T. A., Cox, C. J., & Embley, T. M. (2023). Recoding Amino Acids to a Reduced Alphabet may Increase or Decrease Phylogenetic Accuracy. *Systematic Biology*, *72*(3), 723–737. https://doi.org/10.1093/sysbio/syac042

Foster, P. G., Cox, C. J., & Martin 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*(1527), 2197–2207. https://doi.org/10.1098/rstb.2009.0034

Foster, P. G. (2004). modelling Compositional Heterogeneity. *Systematic Biology*, *53*, 485–495. https://doi.org/10.1080/10635150490445779

Foster, P. G., & Hickey, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, *48*(3), 284–290. https://doi.org/10.1007/PL00006471

Galtier, N, & Gouy, M. (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, *15*, 871–879. https://doi.org/10.1093/oxfordjournals.molbev.a025991

Galtier, N, Tourasse, N., & Gouy, M. (1999). A Nonhyperthermophilic Common Ancestor to Extant Life Forms. *Science*, *283*, 220–221. https://doi.org/10.1126/science.283.5399.220

Galtier, Nicolas. (2001). Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model. *Molecular Biology and Evolution*, *18*, 866–873. https://doi.org/10.1093/oxfordjournals.molbev.a003868

Galtier, Nicolas, & Daubin, V. (2008). Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1512), 4023–4029. https://doi.org/10.1098/RSTB.2008.0144

Gaston, D., Susko, E., & Roger, A. J. (2011). A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics*, *27*(19), 2655–2663. https://doi.org/10.1093/bioinformatics/btr470

Gatesy, J., & Springer, M. S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Molecular Phylogenetics and Evolution, 80, 231–266. https://doi.org/https://doi.org/10.1016/j.ympev.2014.08.013

Gaut, B. S., & Lewis, P. O. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution*, *12*(1), 152–162. https://doi.org/10.1093/oxfordjournals.molbev.a040183

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis Jumping Rules. In *Bayesian Statistics 5: Vol. N/A* (pp. 599–608). Oxford University Press. https://doi.org/10.1093/oso/9780198523567.003.0038

Giacomelli, M., Rossi, M. E., Lozano-Fernandez, J., Feuda, R., & Pisani, D. (2022). Resolving tricky nodes in the tree of life through amino acid recoding. *IScience*, *25*(12), 105594. https://doi.org/10.1016/j.isci.2022.105594

Gill, P. E., Murray, W., & Wright, M. H. (1981). Practical Optimization. *Academic Press, London*.

Goldman, N. (1993a). Simple diagnostic statistical tests of models for DNA substitution. *Journal of Molecular Evolution*, *37*(6), 650–661. https://doi.org/10.1007/BF00182751

Goldman, N. (1993b). Simple diagnostic statistical tests of models for DNA substitution. *Journal of Molecular Evolution*, *37*, N/A. https://doi.org/10.1007/bf00182751

Goldman, N. (1998). Phylogenetic information and experimental design in molecular systematics. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *265*, 1779–1786. https://doi.org/10.1098/rspb.1998.0502

Goremykin, V. V, Nikiforova, S. V., & Bininda-Emonds, O. R. P. (2010). Automated removal of noisy data in phylogenomic analyses. *Journal of Molecular Evolution*, *71*(5–6), 319–331. https://doi.org/10.1007/s00239-010-9398-z

Groussin, M., Boussau, B., & Gouy, M. (2013). A Branch-Heterogeneous Model of Protein Evolution for Efficient Inference of Ancestral Sequences. *Systematic Biology*. https://doi.org/10.1093/sysbio/syt016

Gu, X. (2001). Maximum-Likelihood Approach for Gene Family Evolution Under Functional Divergence. *Molecular Biology and Evolution*, *18*, 453–464. https://doi.org/10.1093/oxfordjournals.molbev.a003824

Guindon, S., & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, *52*, 696–704. https://doi.org/10.1080/10635150390235520

Hall, B. K. (2007). Homoplasy and homology: Dichotomy or continuum? Journal of Human Evolution, 52(5), 473–479. https://doi.org/https://doi.org/10.1016/j.jhevol.2006.11.010

Hartigan, J. A. (1973). Minimum Mutation Fits to a Given Tree. *Biometrics*, *29*, 53. https://doi.org/10.2307/2529676

Hasegawa, M., Kishino, H., & Yano, T. aki. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, *22*(2), 160–174. https://doi.org/10.1007/BF02101694/METRICS

Heath, T. A., Hedtke, S. M., & Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution*, *46*(3), 239. https://doi.org/10.3724/SP.J.1002.2008.08016

Hennig, W. (1966). Phylogenetic systematics. University of Illinois Press. http://dx.doi.org/10.1146/annurev.en.10.010165.000525

Heled, J., & Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, *27*(3), 570–580.

Hernandez, A. M., & Ryan, J. F. (n.d.). *Six-state amino acid recoding is not an effective strategy to offset the effects of compositional heterogeneity and saturation in phylogenetic analyses*. https://doi.org/10.1101/729103

Ho, J. W. K., Adams, C. E., Lew, J. Bin, Matthews, T. J., Ng, C. C., Shahabi-Sirjani, A., Tan, L. H., Zhao, Y., Easteal, S., Wilson, S. R., & Jermiin, L. S. (2006). SeqVis: Visualization of compositional heterogeneity in large alignments of nucleotides. *Bioinformatics*, *22*(17), 2162–2163. https://doi.org/10.1093/BIOINFORMATICS/BTL283

Ho, S. Y. W., & Lanfear, R. (2010). Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. *Mitochondrial DNA*, *21*(3–4), 138–146. https://doi.org/10.3109/19401736.2010.494727

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. http://www.jstor.org/stable/4615733

Huelsenbeck, J. P. (2002). Testing a Covariotide Model of DNA Substitution. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a004128

Hrdy, I., Hirt, R. P., Dolezal, P., Bardonová, L., Foster, P. G., Tachezy, J., & Martin Embley, T. (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature*, 432(7017), 618–622. https://doi.org/10.1038/nature03149

Huelsenbeck, J. P., & Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. In *Annual Review of Ecology and Systematics* (Vol. 28, pp. 437–466). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA. https://doi.org/10.1146/annurev.ecolsys.28.1.437

Huelsenbeck, J. P., & Rannala, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science*, *276*(5310), 227–232. https://doi.org/10.1126/science.276.5310.227

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*, 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. In *Science* (Vol. 294, Issue 5550, pp. 2310–2314). American Association for the Advancement of Science. https://doi.org/10.1126/science.1065889

Jayaswal, V., Jermiin, L. S., Poladian, L., & Robinson, J. (2011). Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. *Systematic Biology*, *60*(1), 74–86. https://doi.org/10.1093/sysbio/syq076

Jeffroy, O. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*. https://doi.org/10.1016/j.tig.2006.02.003

Jermiin, L. S., & Ho, J. (2009). *Chapter 4 SeqVis : A Tool for Detecting Compositional Heterogeneity*. *February*. https://doi.org/10.1007/978-1-59745-251-9

Jermiin, L. S., Ho, S. Y. W., Ababneh, F., Robinson, J., & Larkum, A. W. D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology*, 53(4), 638–643. https://doi.org/10.1080/10635150490468648

Jermiin, L. S. (2017). Identifying Optimal Models of Evolution. *Methods in molecular biology* (Vols. 347–369, Issue November). https://doi.org/10.1007/978-1-4939-6622-6

Jin, L., & Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution*, *7*(1), 82–102. https://doi.org/10.1093/oxfordjournals.molbev.a040588

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*, 275–282. https://doi.org/10.1093/bioinformatics/8.3.275

Jukes, T.H., & Cantor, C.R. (1969). Evolution of Protein Molecules. *Mammalian Protein Metabolism, N/A*, 21–132. https://doi.org/10.1016/b978-1-4832-3211-9.50009-7

Kainer, D., & Lanfear, R. (2015). The effects of partitioning on phylogenetic inference. *Molecular Biology and Evolution*, *32*(6), 1611–1627. https://doi.org/10.1093/molbev/msv026

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

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. https://doi.org/10.1007/bf01731581

Hasegawa, H. K. M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29, 170–179. https://doi.org/10.1007/bf02100115

Kocot, K. M. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*. https://doi.org/10.1038/nature10382

Koshi, J M, & Goldstein, R. A. (1996). Correlating structure-dependent mutation matrices with physical-chemical properties. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 488–499. https://doi.org/10.1002/(SICI)1097-0134(199703)27:3<336::AID-PROT2>3.0.CO;2-B

Koshi, Jeffrey M, & Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins: Structure, Function, and Bioinformatics*, *32*(3), 289–295. http://dx.doi.org/10.1002/(SICI)1097-0134(19980815)32:3<289::AID-PROT4>3.0.CO;2-D

Kosiol, C. (2004). A new criterion and method for amino acid classification. *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2003.12.010

Kumar, S., & Gadagkar, S. R. (2001). Disparity index: A simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics*, *158*(3), 1321–1327. https://doi.org/10.1093/genetics/158.3.1321

Lanave, C., Preparata, G., Sacone, C., & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, *20*(1), 86–93. https://doi.org/10.1007/BF02101990

Lanave, C., Tommasi, S., Preparata, G., & Saccone, C. (1986). Transition and transversion rate in the evolution of animal mitochondrial DNA. *Biosystems*, *19*(4), 273–283. https://doi.org/10.1016/0303-2647(86)90004-3

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701. https://doi.org/10.1093/molbev/mss020

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2017). Partitionfinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, *34*(3), 772–773. https://doi.org/10.1093/molbev/msw260

Larget, B., & Simon, D. L. (1999). Markov Chasin Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. Molecular Biology and Evolution, 16(6), 750. https://doi.org/10.1093/oxfordjournals.molbev.a026160

Lartillot, N., & Philippe, H. (2006). Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology*, 55(2), 195–207. https://doi.org/10.1080/10635150500433722

Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, *7*(SUPPL. 1), 1–14. https://doi.org/10.1186/1471-2148-7-S1-S4/FIGURES/5

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, *21*(6), 1095–1109. https://doi.org/10.1093/molbev/msh112

Le, Q. S., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Le, Q. S., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics*, *24*(20), 2317–2323. https://doi.org/10.1093/bioinformatics/btn445

Le, Q. S., Lartillot, N., & Gascuel, O. (2008). Phylogenetic mixture models for proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1512), 3965–3976. https://doi.org/10.1098/rstb.2008.0180

Le, S. Q., Dang, C. C., & Gascuel, O. (2012). modelling protein evolution with several amino acid replacement matrices depending on site rates. *Molecular Biology and Evolution*, *29*(10), 2921–2936. https://doi.org/10.1093/molbev/mss112

Le, S. Q., & Gascuel, O. (2010). Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Systematic Biology*, *59*(3), 277–287. https://doi.org/10.1093/sysbio/syq002

Le, V.S., Dang, C.C. & Le, Q.S. (2017). Improved mitochondrial amino acid substitution models for metazoan evolutionary studies. *BMC Evol Biol* 17, 136. https://doi.org/10.1186/s12862-017-0987-y

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S. A., Soltis, D. E., Soltis, P. S., Stevenson, D. W., Ullrich, K. K., Wickett, N. J., DeGironimo, L., … Wong, G. K. S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, *574*(7780), 679–685. https://doi.org/10.1038/s41586-019-1693-2

Lemieux, C., Otis, C., & Turmel, M. (2016). Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Frontiers in Plant Science*, *7*(MAY2016). https://doi.org/10.3389/fpls.2016.00697

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*, 99–121.

Lewis, L. A., & McCourt, R. M. (2004). Green algae and the origin of land plants. *American Journal of Botany*, *91*(10), 1535–1556. https://doi.org/10.3732/ajb.91.10.1535

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, *10*(1), 1–18. https://doi.org/10.1186/1471-2148-10-302

Liu, Y., Cox, C. J., Wang, W., & Goffinet, B. (2014). Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Systematic Biology*, *63*(6), 862–878. https://doi.org/10.1093/sysbio/syu049

Lopez, P., Casane, D., & Philippe, H. (2002). Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution*, *19*(1), 1–7. https://doi.org/10.1093/oxfordjournals.molbev.a003973

Lozano-Fernandez, J. (2022). A Practical Guide to Design and Assess a Phylogenomic Study. *Genome Biology and Evolution*, *14*(9), 1–21. https://doi.org/10.1093/gbe/evac129

Luo, H. (2015). Evolutionary origin of a streamlined marine bacterioplankton lineage. *The ISME Journal*, *9*(6), 1423–1433. https://doi.org/10.1038/ismej.2014.227

Maddison, W., & Knowles, L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*(1), 21–30. https://doi.org/10.1080/10635150500354928

Mau, B., & Newton, M. A. (1997). Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*, *6*(1), 122–131. https://doi.org/10.1080/10618600.1997.10474731

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Mirarab, S. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu462

Morgan, C., Foster, P. G., Webb, A. E., Pisani, D., McInerney, J. O., & O'Connell, M. J. (2013). Heterogeneous models place the root of the placental mammal phylogeny. *Molecular Biology and Evolution*, *30*(9), 2145–2156. https://doi.org/10.1093/molbev/mst117

Muñoz-Gómez, S. A., Hess, S., Burger, G., Lang, B. F., Susko, E., Slamovits, C. H., & Roger, A. J. (2019). An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. *Elife*, *8*, e42535. https://doi.org/10.7554/eLife.42535

Nascimento, F. F., Reis, M. Dos, & Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. In *Nature Ecology and Evolution* (Vol. 1, Issue 10, pp. 1446–1454). Springer US. https://doi.org/10.1038/s41559-017-0280-x

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., Lanfear, R., & Bryant, D. (2019). The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biology and Evolution*, *11*(12), 3341–3352. https://doi.org/10.1093/gbe/evz193

Nesnidal, M. P. (2013). New phylogenomic data support the monophyly of Lophophorata and an Ectoproct-Phoronid clade and indicate that Polyzoa and Kryptrochozoa are caused by systematic bias. *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-13-253

Nguyen, B. T., Shuval, K., & Yaroch, A. L. (2015). Nguyen et al., Respond. *American Journal of Public Health*, *105*(10), e2–e2. https://doi.org/10.2105/AJPH.2015.302827

Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W. E. G., Nickel, M., Schierwater, B., Vacelet, J., Wiens, M., & Wörheide, G. (2013). Deep metazoan phylogeny: When different genes tell different stories. *Molecular Phylogenetics and Evolution*, *67*(1), 223–233. https://doi.org/10.1016/j.ympev.2013.01.010

Nylander, J. A. A. (2004). Bayesian Phylogenetic Analysis of Combined Data. *Systematic Biology*. https://doi.org/10.1080/10635150490264699

Olsen, G. J., Matsuda, H., Hagstrom, R., & Overbeek, R. (1994). fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Bioinformatics*, *10*(1), 41–48. https://doi.org/10.1093/bioinformatics/10.1.41

Pamilo, P., & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, *5*(5), 568–583. https://doi.org/10.1093/oxfordjournals.molbev.a040517

Pandey, A., & Braun, E. L. (2019). *Phylogenetic Analyses of Sites in Different Protein Structural Environments Result in Distinct Placements of the Metazoan Root*. *October*. https://doi.org/10.20944/preprints201910.0302.v1

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*(8), 706–712. https://doi.org/10.1016/j.cub.2009.02.052

Philippe, H., & Lopez, P. (2001). On the conservation of protein sequences in evolution. *Trends in Biochemical Sciences*, *26*(7), 414–416. https://doi.org/10.1016/s0968-0004(01)01877-1

Phillips, M. J., & Penny, D. (2003). The root of the mammalian tree inferred from whole mitochondrial genomes. *Molecular Phylogenetics and Evolution*, 28(2), 171–185. https://doi.org/10.1016/S1055-7903(03)00057-5

Phillips, M. J., Delsuc, F., & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, 21(7), 1455–1458. https://doi.org/10.1093/molbev/msh137

Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., Wörheide, G., Philippe, H., Pisani, D., Dohrmann, M., Wörheide, G., Pett, W., & Rota-Stabelli, O. (2015). Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the National Academy of Sciences*, *112*(50), 15402–15407. https://doi.org/10.1073/pnas.1518127112

Posada, D. (2008). jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msn083

Posada, David, & Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics*, *14*(9), 817–818. https://doi.org/10.1093/bioinformatics/14.9.817

Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Schneider, H., Pisani, D., & Donoghue, P. C. J. (2018). The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology*, *28*(5), 733-745.e2. https://doi.org/10.1016/j.cub.2018.01.063

Rannala, B., & Yang, Z. (1996). Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference. *Journal of Molecular Evolution*. https://doi.org/10.1007/pl00006090

Rannala, B., Zhu, T., & Yang, Z. (2012). Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Molecular Biology and Evolution*, *29*(1), 325–335.

Posada, D., & Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics*, 14(9), 817–818. https://doi.org/10.1093/bioinformatics/14.9.817

Rannala, B. &Yang, Z. (2003). Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. *Genetics*. https://doi.org/10.1093/genetics/164.4.1645

Redmond, A. K., & McLysaght, A. (2021). Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-22074-7

Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A., & Thomson, R. C. (2018). Variation Across Mitochondrial Gene Trees Provides Evidence for Systematic Error: How Much Gene Tree Variation Is Biological? *Systematic Biology*, *67*(5), 847–860. https://doi.org/10.1093/sysbio/syy013

Rokas, A., Williams, B. I., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, *425*(6960), 798–804. https://doi.org/10.1038/nature02053

Rokas, A. & Carroll, S.B. (2005). More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msi121

Rzhetsky, A. (1995). Tests of applicability of several substitution models for DNA sequence data. *Molecular Biology and Evolution*. https://doi.org/10.1093/oxfordjournals.molbev.a040182

Schrempf, D., Lartillot, N., & Szöllosi, G. (2020). Scalable empirical mixture models that account for across-site compositional heterogeneity. *Molecular Biology and Evolution*, *37*(12), 3616–3631. https://doi.org/10.1093/molbev/msaa145

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461–464. http://www.jstor.org/stable/2958889

Seo, T. K., & Thorne, J. L. (2018). Information criteria for comparing partition schemes. *Systematic Biology*, *67*(4), 616–632. https://doi.org/10.1093/sysbio/syx097

Shapiro, B., Rambaut, A., & Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, *23*(1), 7–9. https://doi.org/10.1093/molbev/msj021

Shen, X., Steenwyk, J. L., & Rokas, A. (2021). Dissecting incongruence between concatenation-and quartet-based approaches in phylogenomic data. *Systematic Biology*, *70*(5), 997–1014. https://doi.org/10.1093/sysbio/syab011

Shi, C., & Yang, Z. (2018). Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Molecular Biology and Evolution*, *35*(1), 159–179. https://doi.org/10.1093/molbev/msx277

Smith, A. B. (1994). Rooting molecular trees: problems and strategies. Biological Journal of the Linnean Society, 51(3), 279–292. https://doi.org/10.1111/j.1095-8312.1994.tb00962.x

Sousa, F., Foster, P. G., Donoghue, P. C. J., Schneider, H., & Cox, C. J. (2019). Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytologist*, 222(1), 565–575. https://doi.org/10.1111/nph.15587

Sousa, F., Civáň, P., Brazão, J., Foster, P. G., & Cox, C. J. (2020). The mitochondrial phylogeny of land plants shows support for Setaphyta under composition-heterogeneous substitution models. *PeerJ*, *2020*(4), e8995. https://doi.org/10.7717/peerj.8995

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl446

Steel, M. (2005). Should phylogenetic models be trying to 'fit an elephant'? *Trends in Genetics*. https://doi.org/10.1016/j.tig.2005.04.001

Strassert, J. F. H., Irisarri, I., Williams, T. A., & Burki, F. (2021). A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nature Communications, 12(1), 1–13. https://doi.org/10.1038/s41467-021-22044-z

Struck, T. H. (2011). Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. *BMC Evolutionary Biology*. https://doi.org/10.1186/1471-2148-11-369

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, *42*(3–4), 412–416. https://doi.org/10.1093/biomet/42.3-4.412

Susko, E., & Roger, A. J. (2007). On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Molecular Biology and Evolution*, *24*(9), 2139–2150. https://doi.org/10.1093/MOLBEV/MSM144

Susko, E., Lincker, L., & Roger, A. J. (2018). Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models. Molecular Biology and Evolution, 35(5), 1266–1283. https://doi.org/10.1093/molbev/msy026

Susko, E., & Roger, A. J. (2020). On the Use of Information Criteria for Model Selection in Phylogenetics. *Molecular Biology and Evolution*, *37*(2), 549–562. https://doi.org/10.1093/molbev/msz228

Swofford, D. L. (1996). Phylogenetic inference. *Molecular Systematics, 2nd Ed.*, 407–514.

Szöllosi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., & Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6), 901–912. https://doi.org/10.1093/sysbio/syt054

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, *10*(3), 512–526. http://mbe.oxfordjournals.org/content/10/3/512.long

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequence. *Lecture of Mathematics for Life Science*, *17*, 57.

Thorne, J. L., Goldman, N., Jones, D. T. (1996). Combining protein evolution and secondary structure. Molecular Biology and Evolution, 13, 666–673. https://doi.org/10.1093/oxfordjournals.molbev.a025627

Timme, R. E., Bachvaroff, T. R., & Delwiche, C. F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE*, *7*(1). https://doi.org/10.1371/journal.pone.0029696

Uzzell, T., & Corbin, K. W. (1971). Fitting Discrete Probability Distributions to Evolutionary Events: The probability of fixing nucleotide substitutions varies over codons of a cistron. *Science*, *172*(3988), 1089–1096. https://doi.org/10.1126/science.172.3988.1089

Viklund, J., Ettema, T. J. G., & Andersson, S. G. E. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Molecular Biology and Evolution*, *29*(2), 599–615. https://doi.org/10.1093/molbev/msr203

Vinh, L. S. (2004). IQPNNI: Moving Fast Through Tree Space and Stopping in Time. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msh176

Wake, D. B., Wake, M. H., & Specht, C. D. (2011). Homoplasy: From Detecting Pattern to Determining Process and Mechanism of Evolution. *Science*, 331(6020), 1032–1035. https://doi.org/10.1126/science.1188545

Wang, D., Wu, Y. W., Shih, A. C. C., Wu, C. S., Wang, Y. N., & Chaw, S. M. (2007). Transfer of Chloroplast Genomic DNA to Mitochondrial Genome Occurred At Least 300 MYA. *Molecular Biology and Evolution*, *24*(9), 2040–2048. https://doi.org/10.1093/MOLBEV/MSM133

Wang, H. C., Li, K., Susko, E., & Roger, A. J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. *BMC Evolutionary Biology*, *8*(1), 1–13. https://doi.org/10.1186/1471-2148-8-331

Wang, H. C., Minh, B. Q., Susko, E., & Roger, A. J. (2018). modelling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology*, *67*(2), 216–235. https://doi.org/10.1093/sysbio/syx068

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Whelan, N. V., Kocot, K. M., Moroz, L. L., & Halanych, K. M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(18), 5773–5778. https://doi.org/10.1073/pnas.1503453112

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W. Leebens-Mack, J. H., *et al.* (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, *111*(45), E4859–E4868. https://doi.org/10.1073/pnas.1323926111

Wilkinson, M. (1996). Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution*, *13*(3), 437–444. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A025604

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., & Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology and Evolution*, *4*(1), 138–147. https://doi.org/10.1038/s41559-019-1040-x

Williams, T. A., Schrempf, D., Szöllősi, G. J., Cox, C. J., Foster, P. G., & Embley, T. M. (2021). Inferring the deep past from molecular data. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evab067

Woese, C. R., Achenbach, L., Rouviere, P., & Mandelco, L. (1991). Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglohus fulgidus in light of certain composition-induced artefacts. *Systematic and Applied Microbiology*, *14*(4), 364–371. https://doi.org/10.1016/S0723-2020(11)80311-5

Wu, J., & Susko, E. (2009). General heterotachy and distance method adjustments. *Molecular Biology and Evolution*, *26*(12), 2689–2697. https://doi.org/10.1093/molbev/msp184

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*. https://doi.org/10.1007/bf00178256

Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution 1994 39:3*, *39*(3), 306–314. https://doi.org/10.1007/BF00160154

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, *139*(2), 993–1005. https://doi.org/10.1093/genetics/139.2.993

Yang, Z. (1996a). Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution*, *42*, 294–307. https://doi.org/10.1007/BF02198856

Yang, Z. (1996b). Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution*, 42(5), 587–596. https://doi.org/10.1007/BF02352289

Yang, Z. (1996c). Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution*, 42(5), 587–596. https://doi.org/10.1007/BF02352289

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, *13*(5), 555–556. https://doi.org/10.1093/bioinformatics/13.5.555

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Yang, Z., Goldman, N., & Friday, A. (1995). Maximum Likelihood Trees from DNA Sequences: A Peculiar Statistical Estimation Problem. *Systematic Biology*, *44*, 384. https://doi.org/10.2307/2413599

Yang, Z., Nielsen, R., & Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, *15*, 1600–1611. https://doi.org/10.1093/oxfordjournals.molbev.a025888

Yang, Z., & Roberts, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, 12(3), 451–458. https://doi.org/10.1093/oxfordjournals.molbev.a040220

Zhang, C., Scornavacca, C., Molloy, E. K., & Mirarab, S. (2020). ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution*, *37*(11), 3292–3307.https://doi.org/10.1093/molbev/msaa139

Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*. https://doi.org/10.1007/bf00160155

Zou, L., Susko, E., Field, C., & Roger, A. J. (2012). Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the barry-hartigan model. *Systematic Biology*, *61*(6), 927–940. https://doi.org/10.1093/sysbio/sys046

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics, 19(S6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhou, Y., Rodrigue, N., Lartillot, N., & Philippe, H. (2007). Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol Biol 7, 206 (2007). https://doi.org/10.1186/1471-2148-7-206