Phylogenomics and Plant Evolution

A thesis presented in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy
in Genetics

at Massey University, Manawatu
New Zealand

Bojian Zhong

2013

Abstract

Phylogenomics, the study of evolutionary relationships among groups of organisms using genome-scale data, is central to our understanding of the evolution of life. While large amounts of data are available and methodological developments are increasing at a fast pace, there are basic problems that are overlooked in phylogenomic analyses of molecular sequences, which may impede the accuracy and reliability of tree reconstruction. These problems include: How can we detect the non-phylogenetic signals from genomic data? How can we offer a better statistical fitness between the evolutionary model and data? How can we improve the phylogenetic inference using sophisticated and realistic models? How can we accurately infer the species trees? How can we quantitatively confirm the evolutionary theory? With these goals, this thesis concentrates on phylogenomics of land plants (and their origin) and evolution in general.

- Resolving the phylogenetic position of Gnetales. We show that non-time reversible properties of positions in the chloroplast genomes of Gnetales mislead phylogenetic reconstruction, and highlight that the goodness of fit between substitution model and data should be taken into account when performing phylogenomic analyses.

- Resolving the origin of land plants: 1). The multispecies coalescent model is applied to estimate the species tree of origin of land plants, and it is proved to be able to estimate accurate and congruent species tree in the presence of ancient incomplete lineage sorting from nuclear genes. 2). The chloroplast phylogenomic analyses are conducted using sophisticated and realistic evolutionary models that can account for site-heterogeneity and compositional heterogeneity. These chloroplast phylogenomic results confirm the previous nuclear data analyses.

- We develop a statistical test and demonstrate that evolutionary theory could be tested by convergence of molecular data. It also indicates that the reality of evolution can be tested using standard methods and tools.

Acknowledgements

I would like to thank my supervisors, Professor David Penny and Professor Peter Lockhart for your support, your encouragement and your patience. I would not have had the opportunity to make such an achievement without you.

I also wish to thank my colleagues: Dr. Tim White, who always helps me figure out computing problems; P.A. (Trish) McLenachan and Richard Fong, who have done a lot of experimental work for different projects; Dr. Bennet J. McComish, Dr. Oliver Deusch and Dr. Patrick Biggs, who have provided bioinformatic assistance.

I express my warm thanks to my other co-authors:
Dr. Vadim V. Goremykin (Istituto Agrario San Michele all'Adige Research Center, San Michele all'Adige, Italy)
Dr. Liang Liu (Department of Statistics and Institute of Bioinformatics, University of Georgia, USA)
Dr. Zhenxiang Xi (Department of Organismic and Evolutionary Biology, Harvard University Herbaria, USA)
Dr. Philip M. Novis (Allan Herbarium, Landcare Research, Lincoln 7640, New Zealand)

I thank my friends and my former flatmates: Jingjing, Wesley, Jing Pang, Yanfei, Jian Han, Chuchu, Qinghao, Jinlin, and Hao Zhang who have made New Zealand as my "new home" with a wonderful time.

I thank AWC administrators Wendy and Joy for all your assistance with financial support, and thank to all (former) members in our lab, for their company and discussions.

Lastly, my deepest appreciation to my parents, for their love and support throughout my life.

CONTENTS

Chapter 1. Introduction

About 2.45 billion years ago, Cyanobacteria became oxygen evolving. This event revolutionized the atmosphere of the Earth, making it aerobic for the first time. Evolution of aerobic organisms began at the prokaryotic level and over the next billion years, at the eukaryotic level. Most of these eukaryotic organisms acquired by symbiosis an aerobic prokaryote, a rickettsial proteobacterium, which became the mitochondria of most eukaryotic organism. A subset of these eukaryotic protists acquired a photosynthetic cyanobacterium, which became the chloroplast. The lineages of photosynthetic protists are today known as algae (except for the apicomplexans which remains as an apicoplast), and another group the "fungal" oomycetes that are thought to have derived from photosynthetic protists.

Later, three independent evolutionary lines lead out of the protists into the current multicellular forms on Earth:
i) Multicellular plants (the land plants). It also has to be conceded that multicellularity evolved in other eukaryotic algae (Rhodophyceae, Phaeophyceae and some other Chlorophyceae), but these groups never migrated significantly onto the land.
ii) Multicellular animals.
iii) True Fungi (not including Oomycetes which derive from photosynthetic protists).

With the development of sequencing technology, molecular data have been used successfully for inferring many aspects of plant evolution, and three notable examples are summarized here:
(1) The origins and evolutionary relationships of Plastids: It has been a controversial issue about the single or multiple origins of plastids (e.g., Rodriguez-Ezpeleta et al. 2005; Reyes-Prieto and Bhattacharya 2007). Recent research has focused on the number of endosymbiotic events involved and controversy over the secondary evolution of red plastids (e.g., Keeling 2009; Baurain et al. 2010; Keeling 2013).
(2) The origin of land plants: Analyses of both morphological and molecular data have established land plants evolved from within streptophyte algae. However uncertainty still exists over which streptophyte algae are most closely related to the land plants. The recent phylogenomic analyses of both chloroplast and nuclear genome data have suggested that either Coleochaetales alone (Turmel et al. 2009), or Zygnematales alone

(Wodniok et al. 2011; Timme et al. 2012), or Coleochaetales and Zygnematales combined (Finet et al. 2012; Laurin-Lemay et al. 2012) are sister to land plants.
(3) The origin of seed plants: The phylogenetic position of Gnetales has remained one of the most contentious issues in seed plant systematics. Currently, three hypotheses have received some support from molecular analyses, i.e., Gnetales as sister to conifers (e.g., Chaw et al. 1997); or as sister to Pinaceae within conifers (e.g., Chaw et al. 2000; Hajibabaei et al. 2006; Wu et al. 2007); or as sister to Cupressophyta (non-Pinaceae conifers) within conifers (e.g., Nickrent et al. 2000; Doyle 2006).

The above examples all represent cases where great progress has been made in understanding plant evolution, but also where controversy still surrounds phylogenetic inferences. Understanding the controversy requires understanding the nature of the data studied and also its reliability. There are now many classes of data - nuclear, chloroplast and mitochondrial sequences, together with indels, gene order (synteny) and retrotransposons, and we would certainly expect to see some basic agreement between these different classes. The different classes of data may have different properties, which we expect to be able to better resolve phylogenetic divergences at different time scales. However, it is certainly to be expected that deeper divergences will become increasingly difficult to address as we go further back in time, because Markov models for sequence evolution are expected to saturate and lose information at the most ancient divergences (Mossel and Steel 2004). This qualification has also been reported for difficulties in reconstructing all-inclusive phylogenetic trees with bacterial divergences (Meyer et al. 1986). And, at shorter times there are other potentially misleading processes happening with real populations; these processes are real and will lead to errors in phylogenetic inference. We can summarise most of the potential problems into three main categories:

a) Sampling errors
b) Systematic errors from too simple assumptions about physical models (Markov models)
c) Systematic errors from overly simple assumptions about biological models (particularly speciation models)

The first class of problem (sampling error) is an important difficulty when only relatively short sequences are available, but it has become less of a problem in the past

decade. Essentially, I consider it a lesser issue at present, and will focus more on the second and third problems. But it is still an important factor when comparing individual genes or proteins, and there is a predicted, and verified, effect of shorter genes showing more ambiguity (White et al. 2013).

The issue of the simplicity/complexity of Markov models is relatively well studied. The simplest Markov model, the Jukes-Cantor (Jukes and Cantor 1969), treats all mutations as having an equal chance of occurring; there is only a single parameter for nucleotide (or amino acid) changes. In practice, a series of more complex models are possible, and Table 1 shows the range of popular nucleotide substitution models that have been developed, including allowing sites to have different rates (including allowing invariant sites). Again, it has been well recognised that many models are too simplistic, and that they are subject to errors. Two important sources are "long branch attraction" (LBA) artefact, and differences in sequence composition.

Table 1. The standard models of nucleotide substitution

Markov models of nucleotide substitution	Number of parameters for substitution-rate matrix	Base frequency	Reference
JC69	1 (equal rate of changing into any other nucleotide from one nucleotide)	Equal	Jukes and Cantor 1969
K80	2 (different transition and transversion rates)	Equal	Kimura 1980
F81	4 (equilibrium frequencies of four nucleotide)	Unequal	Felsenstein 1981
HKY85	8 (transition/transversion rates and equilibrium frequencies)	Unequal	Hasegawa et al. 1985
GTR	12 (substitution rates and equilibrium frequencies)	Unequal	Tavaré 1986; Yang 1994
Markov model+gamma	# +1 (rate variation by assuming that rate *r* for any site is a random variable drawn from a gamma distribution)	/ (depend on the Markov model)	Yang 1996
Markov model+I	# +1 (invariable site with rate *r* = 0)	/	Hasegawa et al. 1985
Markov model+gamma+I	# +2	/	/

It is well recognised that Markov models for nucleotide substitution do have their limits. Firstly, it was initially assumed that virtually all "errors/mutations" resulted from errors in copying and replicating DNA, and that there are also significant errors during the repair of double-stranded breaks and during recombination (Romiguier et al. 2013). Secondly Mossel and Steel (2004), for example, show that eventually the power of Markov models to recover a tree will fall off exponentially at very deep times, and that additional taxa will only improve the power of Markov models linearly. We therefore have an exponential decay at deeper times, faced off against a linear improvement with increasing the number of taxa, and there is no doubt that the exponential decay will dominate in the longer term.

There appears to be less agreement on the complexity of biological factors that may affect phylogenetic inference, but there does appear to be an increasing recognition that they are important. There are several effects here, starting with **lineage sorting**. It is well accepted within evolutionary studies (even if not among biologists as a whole) that there is a continuum from individuals, populations, races, varieties, sibling species, species, species complexes, subgenera, genera, etc. Along this continuum we expect introgression and hybridization to be quite normal, even if these two processes decrease at deeper divergences. Similarly, it may be difficult to identify genuine identical copies of genes. It is now known that copy number variation (CNV) is normal within populations (Redon et al. 2006), and there will always be some uncertainty about when a particular copy of a gene arose. These two properties are illustrated in Figure 1a and 1b.

Then there is **natural selection** at the level of the genes, that can cause real problems for phylogenetic analysis. As stated in the earlier section under Markov models, it has been generally assumed that most, if not all, mutations were 'neutral', and that genetic drift was the dominant effect. In practice, we know very little about the factors of natural selection that might be operating in related lineages. If the mutational process is "random", in that it is occurring all the time and is not related to any needs of the organism, then there is no surprise if related lineages independently happen upon similar mutations that are advantageous (see Figure 1c). For example, convergent molecular evolution can occur in the mitochondrial genome (Castoe et al. 2009) and in chloroplast genes (Zhong et al. 2010), due to parallel substitutions in distantly related

taxa. Certainly, even 3-D structures can also arise by convergent evolution (Barber and Elbe 2013). This convergence phenomenon is particularly difficult to take into account using conventional phylogenetic models. But certainly, it has been recognised for over 30 years (Penny et al. 1982) that different genes do not give identical trees, even if the trees are statistically extremely similar. As such, we need to evaluate all the sources of potential errors that could mislead phylogenetic inference as we go deeper in time (Salichos and Rokas 2013). As Delsuc et al. (2005) suggested "the congruence of results obtained from various datasets and/or various methods is the key validation of evolutionary inferences" (see also White et al. 2013). Ideally we expect congruence from the different classes of data. Figure 1 shows three biological cases that can mislead phylogenetic inference.

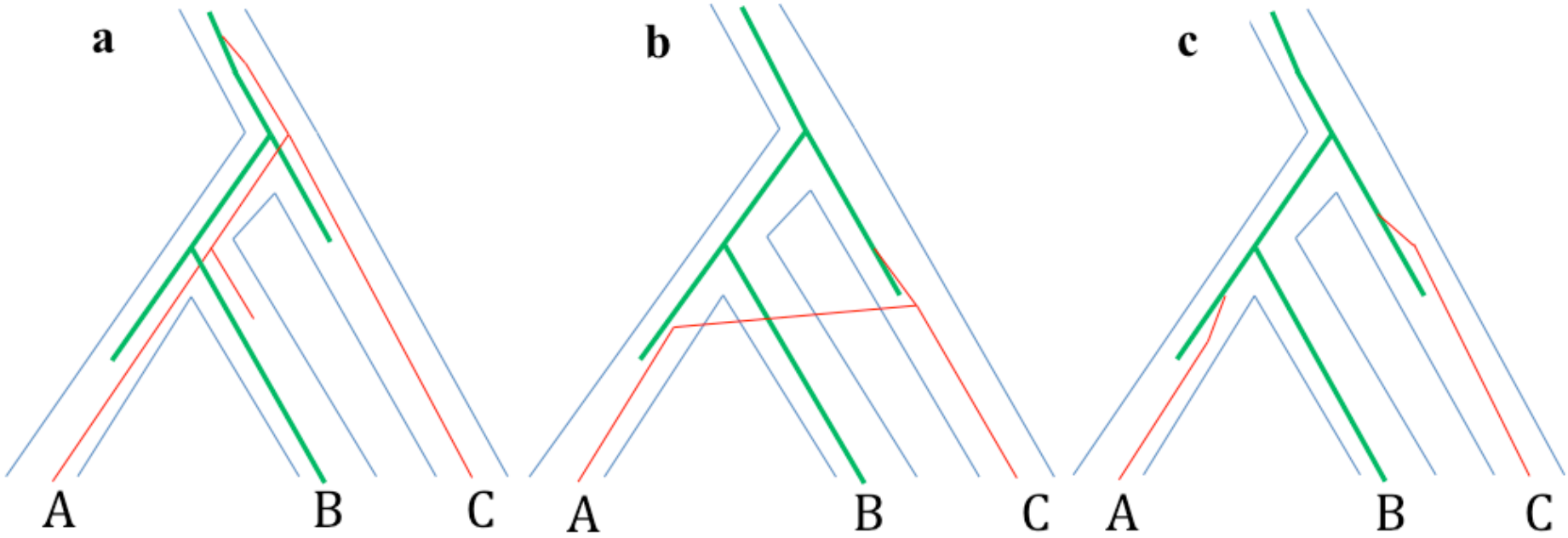


Figure 1. Three classes of biological mechanisms that can mislead phylogenetic inference – shown by the lineages A, B and C not agreeing with the underlying tree, which is ((A,B),C). (a): **Lineage sorting or ambiguous orthology** (e.g. from copy number variation, or polyploidy). Under lineage sorting we expect variation of alleles in a population, but that this will eventually lead to fixation of one allele. Ambiguous orthology can arise from copy number variation when different copies become the dominant form in different species or populations. (b): **Introgression, hybridisation and lateral gene transfer** (though the latter is usually deeper in time). (c): **Independent mutations**, either from sampling error or from systematic biases in mutations. The figure does not include natural selection for the same mutation on different lineages (possibly included under Fig. 1c).

Reconstructing the evolutionary relationships between species (such as reconstructing the "Tree of Life") has attracted many contributions since the times in which few molecular sequences of proteins were available (Zuckerkandl and Pauling 1965). However, when using only a few genes and taxa to resolve species relationships, sampling errors can affect the phylogenetic inference, resulting in reasonable (but not 100% correct) relationships, or poor resolution (Delsuc et al. 2005). Because the cost of next-generation sequencing has declined dramatically during the last ten years, the use of genomic data to infer evolutionary relationships (phylogenomics), has become a standard means for phylogenetic inferences, first from using complete mitochondrial or chloroplast genomes, and now from using whole genomes (e.g. Smith et al. 2011; Salichos and Rokas 2013). The increased availability of complete genome sequences has enabled researchers to use large numbers of sequences for phylogenetic and molecular evolutionary studies. The main advantage of phylogenomics is the use of large datasets to significantly reduce sampling biases. Increasingly, the focus is not just on the actual phylogeny, but also on the processes that are occurring during evolution.

However, while more data can reduce sampling errors in phylogenetic analysis, it can also exacerbate the problem of systematic errors. For example, it has been reported that the approach based on large-scale genomic data does result in significant incongruence among methods and character sets (Jeffroy et al. 2006), and simply adding more sequences may be insufficient for resolving difficult phylogenetic questions (such as rapid speciation events, and assuming that there is just one "true" phylogeny) because non-phylogenetic signals derived from model violation might increase and compete with "genuine" signals (Philippe et al. 2011). There are several factors contributing the systematic errors, including faster substitution rates in nonadjacent phylogenetic lineages (Felsenstein 1978), limited taxon sampling due to extinction or the limited availability of some taxa (Hendy and Penny 1989), and properties of sequences not well described by stationary time reversible models. The latter include base compositional heterogeneity (Lockhart et al.1994, Foster 2004; Jermiin et al. 2004) and lineage-specific changes in evolutionary constraint that can alter the proportion of variable sites in homologs (Lockhart and Steel 2005). Thus, detecting and mitigating the impact of systematic errors in large-scale genomic data is now an important issue in phylogenomic study. There are at least four approaches that have been developed to address the problem of systematic errors:

(1) Using a larger number of taxa, which can break long branches and better detect multiple substitutions – though this is at the cost of a very large increase in the possible number of trees.

(2) Identification and removal of non-phylogenetic sites, which fit the substitution model poorly and may have a significant impact on phylogenetic inference.

(3) Use of more realistic/complex models of sequence evolution especially by accommodating heterogeneity across sites and over time in the evolutionary process.

(4) Use of analytical methods that account for phylogenetic information from independent gene trees (species tree methods)

I will mainly review the approaches (2), (3) and (4) in the following sections as these approaches have been well discussed and developed in the past years. I am also well aware of the network principle and methodology that provide an alternative tool for presenting evolutionary relationship among species (Bapteste et al. 2013). Evolutionary networks are particularly useful to investigate hybridization in plants, or lateral gene transfer events in viruses and prokaryotes (though this is not a focus here).

1. Using a larger number of taxa

As mentioned earlier, this will not be a focus of the present analysis but is considered in both chapters 2 and 3, where it probably helps to reduce LBA artefacts.

2. Identification of non-phylogenetic signals

Because non-phylogenetic signals are one important source of systematic errors, there are several explicit methods focusing on identification and removal of these sites. Brinkmann and Philippe (1999) initially reported a parsimony method, called the Slow-Fast method (S-F), which requires a pre-defined tree and divides the complete data set into monophyletic groups. For each group the numbers of substitutions for each site were estimated by parsimony, and then summed over all groups, giving an estimated number of changes for each site. Then, the groups with the fast-evolving sites are removed from the data, and only the slow-evolving sites are used for subsequent phylogenetic analyses. The assumption is that fastest-evolving sites contain the non-phylogenetic signals. This method is implemented in the Slow-Faster program (Kostka et al. 2008), which removes the fast-evolving sites given a dataset and a threshold.

Additionally, Lopez et al. (1999) applied another similar method, called H-P method (H for Hennigian matrix, P for phenetic matrix), but the H-P method assumes a covarion model (rather than a rates across sites model as in the S-F method) and allows different rates among different sites, which is modelled by a gamma distribution. However, both S-F and H-P methods are tree-dependent methods, thus their effectiveness depends on the prior tree topology in which rates are assigned to positions.

In contrast, compatibility methods (Le Quesne 1969; Pisani 2004; Pisani et al. 2006) are tree-independent, thereby reducing the misidentification of fast-evolving sites based on erroneous tree topologies. These methods are based on the assumption that the highly incompatible characters of the dataset include homoplasious sites, and could be deleted according to compatibility measurement. Alternatively, Goremykin et al. (2010) recently described a tree-independent method, which measures the "observed variability" (OV) of each position in an alignment as:

$$OV = \Sigma\,(1..k)\{d_{ij}\}/k$$

Here k is the number of pair-wise comparisons for a given position and d_{ij} is the score of character variability in each pair-wise comparison (mismatches are scored as 1 and matches as 0). The alignment was ordered from the most highly variable sites to the most conserved sites, and a series of alignments was generated by successively shortening the original alignment. For each shortening step, two data partitions were obtained: 1) the shortened alignment containing the most conserved sites (partition ''A'') and 2) an alignment containing the most variable sites (partition ''B''). After model fitting for each partition the maximum likelihood (ML) distance and uncorrected *p* distance were calculated using PAUP* (Swofford 2002). At each shortening step, two Pearson correlation analyses of pairwise distances were conducted:

1) correlation of the ML and uncorrected *p* distances for partition "B".

2) correlation of the ML distances for partition "A" and "B".

The stopping point for site removal was determined as the point at which the two correlations showed a significant improvement. This OV-sorting method has been found to be effective in concentrating saturated positions toward the most varied end of the sorted alignment. Recent analyses using whole chloroplast genome data (Zhong et al. 2011; Parks et al. 2012; Goremykin et al. 2013) have suggested that accuracy can be improved by the removal of non-phylogenetic sites (e.g., the most variable sites) from datasets using the OV-sorting method. As illustrated in Zhong et al. (2011), the OV-

sorting method identifies and provides a basis for removing sites from a concatenated alignment that have a poor fit to phylogenetic model assumptions. Although this criterion does not remove all model-violating sites from the data, it has been shown to remove sites that have significant effect in misleading tree building. In particular, it appears very useful for reducing the LBA artefact and compositional heterogeneity in phylogenetic reconstruction. (The method does combine all the data, rather than analysing it by genes.)

3. Development of realistic/complex models of sequence evolution

As Box and Draper (1987) famously claimed, "all models are wrong, but some are useful". Although models should be biologically sound and realistic, development of more accurate and realistic models of evolution to detect multiple substitutions in a phylogenetic context is the top priority.

Rate heterogeneity among sites is a natural feature of the evolution of DNA sequences. Yang (1996) introduced a discrete gamma distribution to account for among-site rate variation (see Table 1), which has been widely used to improve the robustness of phylogenetic inference. In the following years, many substitution models were developed (Table 1) allowing more robust estimations to be made. However, most early approaches used a single homogeneous model of the replacement process to characterize all sites. The homogeneous model may produce phylogenetic artefacts due to the poor fit of the non-phylogenetic sites. Moreover, the artefacts may be exacerbated by using a large number of genes in phylogenomic studies (Delsuc et al. 2005). Some major achievements in accommodating site heterogeneity have been reported: 1) the site-heterogeneous category (CAT) Bayesian model (Lartillot and Philippe 2004), which can adjust for site-specific frequencies and allows the equilibrium frequencies to vary among sites; 2) the likelihood-based "mixture" model (Pagel and Meade 2004). The "mixture" model not only considers the different profile (i.e. equilibrium base frequencies) across sites but also assigns the different substitution rate matrix to different alignment sites. Both CAT and "mixture" models allow the relaxation of the assumption of homogeneity among positions of the alignment, but need a large number of parameters to describe a wide variety of evolutionary/mutation processes.

Another problem is that most substitution models assume the base composition is stationary (i.e. that base composition does not change over time). Violation of this model assumption is well known to lead to inaccurate tree reconstruction, and compositional heterogeneity (Lanave et al 1984; Lockhart et al. 1994; Foster 2004; Jermiin et al. 2004) is a common problem in this respect. Thus use of non-homogeneous and non-stationary models that account for this variability in evolution can help minimize compositional biases and hence should improve phylogenetic reconstructions (Galtier and Gouy 1998; Herbeck et al. 2005). There are two popular models that account for compositional heterogeneity. First, Galtier and Gouy (1998) introduced a non-stationary non-homogeneous model of evolution in phylogenetic inference as implemented in nhPhyML program (Boussau and Gouy 2006), and this model specifies the different GC content of different lineages in a likelihood framework and allows the base composition to vary between lineages. Second, the time-heterogeneous CAT-BP model (Blanquart and Lartillot 2008) can account for compositional heterogeneity between lineages by introducing breakpoints along the branches, and allows the stationary frequencies to vary in different parts of the tree. Again, these models allow/require more parameters.

A further underestimated but critical problem of model of evolution is the heterogeneity of rate at sites over time, a character known as heterotachy (Lopez et al. 2002). Unfortunately most evolutionary models assume that the rate of nucleotide replacement is stationary between lineages. It is difficult to address this problem probably because the number of free parameters would be very large, and it is impossible to give different rate replacement to each site of all species – we would just then be allowing too many parameters (a problem with parsimony). But several models have been developed and may provide a good start for resolving this problem, although they simplify the heterotachy process. 1) **Covarion models** that allow for evolutionary change over time at individual sites, but still assume the same rate of distribution (Tuffley and Steel 1998, Galtier 2001; Huelsenbeck 2002; Wang et al. 2007); 2) **Rate-shift models** that allow different rates of distribution in different subtrees (Susko et al. 2002); 3) **Mixture of branch lengths models** where different proportions of sites have their own set of branch lengths (Zhou et al. 2007; Kolaczkowski and Thornton 2008; Pagel and Meade 2008). These empirical studies show that ignoring the heterotachy property leads to erroneous estimates of tree topology.

4. Species tree estimation

Species are, by most definitions, evolving lineages that comprise many genes, each found in many individuals, so the species tree may be defined by reconciling conflicts among individual gene trees (Edwards 2009). In other words, the species tree might be distinct from the gene trees.

The concatenation method combines different genes into a single supermatrix and generates a single "supergene" tree that is considered to be equivalent to the species tree. This method is suggested to give more accurate trees than a consensus approach that summarizes congruence among individual gene trees (Gadagkar et al. 2005). But the problem of the concatenation method is that it assumes all genes have the same (or at least similar) phylogenies (de Queiroz and Gatesy 2007; William and Ballard 1996). However it has become clear that individual gene trees appear to be conflict with one another and gene tree heterogeneity is ubiquitous (Maddison 1997; Carstens and Knowles 2007). Thus the concatenation method can yield misleading inferences of species relationships in the presence of a high level of gene tree heterogeneity (Kubatko and Degnan 2007; Mossel and Vigoda 2005). Moreover, as more data are used, concatenation methods may lead to an "incorrect" tree with increasing support due to short branches in some of the gene tree topologies (Kubatko and Degnan 2007). Nevertheless, it is still expected that the concatenated tree will (statistically) be close to the "correct" tree.

There are many reasons for gene tree heterogeneity and gene trees vs species trees conflict, including horizontal gene transfer (HGT), gene duplication, and incomplete lineage sorting (Figure 1a and 1b). Horizontal gene transfer (also known as lateral gene transfer) is well known and common across the tree of life, for example, eukaryotic genomes contain a large amount of genes from bacteria and viruses (Keeling and Palmer 2008). Detecting HGT events may be inaccurate as current methodology relies on the estimated species tree beforehand and on the absence of other mechanisms relating to gene tree discordance (Rasmussen and Kellis 2007; 2011). Gene duplication is also common in all three domains of life (bacteria, archaebacteria and eukaryotes) (Zhang 2003), and may adversely affect phylogenetic accuracy if it is not recognized. But if duplicated genes can be identified correctly, then they can provide additional

genomic markers for phylogenetic analyses (e.g. Page and Charleston 1997; Sanderson and McMahon 2007).

The currently best-studied factor as to why gene trees are distinct is incomplete lineage sorting, i.e., the failure of two or more lineages in a population to coalesce, leading to the possibility that at least one of the lineages coalesces first with a lineage from a less-closely related population (Degnan and Rosenberg 2009). The probability of inferring the wrong species tree due to incomplete lineage sorting has been calculated theoretically for four individual species (Tajima 1983), and later Pamilo and Nei (1988) confirmed that incomplete lineage sorting is a general case, and proposed that adding more gene sequences will still provide the correct relationship.

The multispecies coalescent model is designed to approximate variation in a species tree topology derived from incomplete lineage sorting. The multispecies coalescence method chooses ancestors from the population backward through time for multiple sequences but places some constraints on how recently the coalescences occur. The multispecies coalescent model describes individual gene trees as the outcomes of a stochastic process (coalescence process) along the lineages of the species tree. Because gene trees are allowed to vary in the multispecies coalescent model, coalescent methods can consistently estimate species trees in spite of the presence of heterogeneous gene trees (Edwards 2009; Liu et al. 2009a; McCormack et al. 2009). Some recent studies have also shown that the multispecies coalescent model is able to produce accurate and congruent species trees in the presence of ancient incomplete lineage sorting (Degnan and Rosenberg 2009; Song et al. 2012; Zhong et al. 2013).

A number of phylogenetic programs have been developed for species tree reconstruction, such as *BEAST (Heled and Drummond 2010), BEST (Liu 2008) and STEM (Kubatko et al. 2009), but the disadvantage of these programs is that estimating large number of parameters (population sizes, divergence times and tree topologies) is computationally expensive. Thus, some fast-computational programs (MP-EST, STAR, NJst) have been written for estimating a species tree using large data sets. The MP-EST (Maximum Pseudo-likelihood Estimation of the Species Tree) method uses the frequencies of gene trees of triplets of taxa to estimate the topology and branch length from a collection of rooted gene trees (Liu et al. 2010). The STAR (Species Tree

estimation using Average Ranks of coalescence) method uses the average ranks of gene coalescence times to compute the topological distances in the set of gene trees (Liu et al. 2009b). The NJst method (Liu and Yu 2011) uses un-rooted neighbour joining (NJ) trees built from a distance matrix to infer species-tree phylogenies. MP-EST and STAR methods are partially parametric methods, which only use part of the information (i.e. topology of gene trees without branch length) in the data. So these programs can quickly reconstruct species trees from large-scale genomic data, and have computational advantages. Furthermore, all three methods (MP-EST, STAR, NJst) are relatively robust in the presence of a small amount of horizontal gene transfer as some short coalescence times due to horizontal gene transfer do not have major impact on these methods when the number of genes is large (Song et al. 2012). Table 2 lists some commonly used methods for estimating species tree.

Table 2. Some commonly used species tree reconstruction methods. (Updated and modified from Table 1 of Edwards, 2009).

Method (References)	Methodological basis	Input	Estimate species tree branch lengths
Deep coalescence (Maddison and Knowles 2006)	Parsimony	Gene trees	No
BUCKy (Larget et al. 2010)	Likelihood/Coalescent	Gene trees	No
BEST (Liu 2008)	Bayesian	Sequences	Yes
STAR (Liu et al. 2009b)	Ranks of pairwise coalescence times	Gene trees	No
STEM (Kubatko et al. 2009)	Likelihood	Gene trees	Yes
MP-EST (Liu et al. 2010)	Likelihood	Gene trees	Yes
*BEAST (Heled and Drummond 2010)	Bayesian	Sequences	Yes
NJst (Liu and Yu 2011)	Distance	Gene trees	No

Overview of the thesis

This thesis addresses several topics in phylogenetics and evolution in general, using large amounts of genomic data. The thesis mainly focuses on plant evolution (particularly land plants and their origin), and specific questions are covered in Chapters 2 and 3. Briefly, Chapter 2 investigates the phylogenetic position of Gnetales, which is one of the most contentious issues in seed plant systematics. It had been assumed that Gnetales were a group distinct from the gymnosperms. Recent molecular phylogenies now place them within the gymnosperms, but with which group of gymnosperms (Pinaceae or Cupressophyta) has been uncertain. In Chapter 3, the origin of land plants is revisited using a large number of nuclear genes and the multispecies coalescent model, followed by additional phylogenomic analyses incorporating three new algal chloroplast genomes and heterogeneous models. Then in Chapter 4, from a theoretical evolution aspect, a quantitative test of ancestral convergence (i.e., sequences of homologous proteins from different species converge as we go further and further back in time) is developed and tested for a range of datasets that have diverged at deeper times. This is followed by a short summary chapter (Chapter 5) that looks towards future directions. Three appendices are attached, of published papers I have contributed to during the course of my study.

References

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

Barber, M.F and Elbe, N.C. (2013) Mimicry all the way down. *Nature*. 501: 38-39.

Baurain, D., Brinkmann, H., Petersen, J., Rodriguez-Ezpeleta, N., Stechmann, A., Demoulin, V., Roger, A.J., Burger, G., Lang, B.F., and Philippe, H. (2010). Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol. Biol. Evol*. 27: 1698-1709.

Blanquart, S. and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* 25: 842-858.

Brinkmann, H. and Philippe, H. (1999). Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol*. 16: 817-825.

Box, G.E.P. and Draper, N.R. (1987). Empirical model-building and response surfaces. John Wiler&Sons, New York, NY, USA.

Boussau, B. and Gouy, M. (2006). Efficient likelihood computations with non-reversible models of evolution. *Syst. Biol.* 55: 756-68.

Carstens, B.C and Knowles, L.L. (2007). Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from melanoplus grasshoppers. *Syst. Biol.* 56: 400-411.

Castoe, T.A., de Koning, A.P., Kim, H.M., Gu, W., Noonan, B.P., Naylor, G., Jiang, Z.J., Parkinson, C.L., and Pollock, D.D. (2009). Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc. Natl Acad Sci. USA*. 106: 8986-8991.

Chaw, S.M., Zharkikh, A., Sung, H.M., Lau, T.C., and Li, W.H. (1997). Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. *Mol. Biol. Evol*. 14: 56-68.

Chaw, S.M., Parkinson, C.L., Cheng, Y., Vincent, T.M., and Palmer, J.D. (2000). Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl Acad Sci. USA*. 97: 4086-4091.

de Queiroz, A. and Gatesy, J. (2007). The supermatrix approach to systematics. *Trends Ecol. Evol*. 22: 34-41.

Degnan, J. and Rosenberg, N. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol*. 24: 332-340.

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

Doyle, J.A. (2006). Seed ferns and the origin of angiosperms. *J. Torrey Bot. Soc*. 133: 169-209.

Edwards, S.V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*. 63: 1-19.

Felsenstein, J. (1978). Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27: 401-410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17: 368-376.

Finet, C., Timme, R.E., Delwiche, C.F., and Marlétaz, F. (2012). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr. Biol*. 22: 1456-1457.

Foster, P.G. (2004). Modeling compositional heterogeneity. *Syst. Biol.* 53: 485-495.

Gadagkar, S.R., Rosenberg, M.S., and Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zoolog. B. Mol. Dev. Evol.* 304: 64-74.

Galtier, N., and M. Gouy. (1998). Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol*. 15: 871-879.

Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol*. 18: 866-873.

Goremykin ,V.V., Nikoforova, S.V., and Bininda-Emonds, O.P.P. (2010). Automated removal of noisy data in phylogenomic analyses. *J. Mol. Evol*. 71: 319-331.

Goremykin, V.V., Nikiforova, S.V., Biggs, P.J., Zhong. B., De Lange, P., Martin, W., Woetzel, S., Atherton, R.A., McLenachan, T., and Lockhart, P.J. (2013). The evolutionary root of flowering plants. *Syst. Biol*. 62: 51-62.

Hajibabaei, M., Xia, J., and Drouin, G. (2006). Seed plant phylogeny: gnetophytes are derived conifers and a sister group to Pinaceae. *Mol. Phylogenet. Evol*. 40: 208-217.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol*. 22: 160-174.

Heled, J., and Drummond, A.J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol*. 27: 570-580.

Hendy, M., and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Syst. Zool*. 38: 297-309.

Herbeck, J. T., Degnan, P. H., and Wernegreen, J. J. (2005). Nonhomogeneous model of

sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). *Mol. Biol. Evol*. 22: 520-532.

Huelsenbeck, J.P. (2002). Testing a covariotide model of DNA substitution. *Mol. Biol. Evol*. 19: 698-707.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: The beginning of incongruence? *Trends Genet*. 22: 225-231.

Jermiin, L. S., Ho, S.Y.W., Ababneh, F., Robinson, J., and Larkum, A.W.D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst. Biol*. 53: 638-643.

Jukes, T. H. and C. R. Cantor. (1969). Evolution of protein molecules. In H. N. Munro, ed., Mammalian Protein Metabolism, pp. 21-132, Academic Press, New York.

Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet*. 9: 605–618.

Keeling, P.J. (2009). Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J. Eukaryot. Microbiol*. 56: 1-8.

Keeling, P.J. (2013). The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu. Rev. Plant Biol*. 64: 583-607.

Kimura, M. A. (1980). Simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol*. 16: 111-120.

Kolaczkowski, B. and Thornton, J.W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol. Biol. Evol*. 25: 1054-1066.

Kostka, M., Uzlikova, M., Cepicka, I., and Flegr J. (2008). SlowFaster, a userfriendly program for slow-fast analysis and its application on phylogeny of Blastocystis. *BMC Bioinformatics*. 9: 341.

Kubatko, L.S. and Degnan, J.H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol*. 56: 17-24.

Kubatko, L.S., Carstens, B.C. and Knowles, L.L. (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*. 25: 971-973.

Lanave, C., Preparata, G., Sacone, C., and Serio G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol*. 20: 86-93.

Larget, B.R., Kotha, S.K., Dewey, C.N., and Ane, C. (2010). BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*. 26: 2910-2911.

Lartillot, N., and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol*. 21: 1095-1109.

Laurin-Lemay S, Brinkmann H, and Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Curr. Biol*. 22: R593-R594.

Le Quesne W.J. (1969). A method of selection of characters in numerical taxonomy. *Syst. Zool*. 18: 201-205.

Liu, L. (2008). BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*. 24: 2542-2543.

Liu, L., Yu, L.,Kubatko, L.,Pearl, D.K.,and Edwards, S.V. (2009a). Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol*. 53: 320-328.

Liu, L., Yu, L., Pearl, D.K., and Edwards, S.V. (2009b). Estimating species phylogenies using coalescence times among sequences. *Syst. Biol*. 58: 468-477.

Liu, L., Yu, L., and Edwards, S.V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol*. 10: 302.

Liu, L., and Yu, L. (2011). Estimating species trees from unrooted gene trees. *Syst. Biol*. 60: 661-667.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol*. 11: 605-612.

Lockhart, P. J. and Steel, M. A. (2005). A tale of two processes. *Syst. Biol*. 54: 948-951.

Lopez, P., Forterre, P., and Philippe, H. (1999). The root of the tree of life in the light of the covarion model. *J. Mol. Evol*. 49: 496-508.

Lopez, P., Casane, D., and Philippe, H. (2002). Heterotachy, an Important Process of Protein Evolution. *Mol. Biol. Evol*. 19: 1-7.

Maddison, W.P. (1997). Gene trees in species trees. *Syst. Biol*. 46: 523-536.

Maddison, W.P. and Knowles, L.L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*. 55: 21-30.

McCormack, J.E., Huang, H., and Knowles, L.L. (2009). Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Syst. Biol*. 58: 501-508.

Meyer, T. E., Cusanovich, M. A., and Kamen, M.D. (1986). Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. *Proc. Natl Acad Sci. USA*. 83: 217-220.

Mossel, E. and Steel, M. (2004). A phase transition for a random cluster model on phylogenetic trees. *Math. BioSci*. 187: 189-203.

Mossel, E. and Vigoda, E. (2005). Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*. 309: 2207-2209.

Nickrent, D.L., Parkinson, C.L., Palmer, J.D., and Duff, R.J. (2000). Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol*. 17: 1885-1895.

Page, R. and Charleston, M. A. (1997). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol*. 7: 231-240.

Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol*. 53: 571-581.

Pagel, M. and Meade, A. (2008). Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos. Phil. Trans. R. Soc. B*. 363: 3955-3964.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol*. 5: 568-583.

Parks, M., Cronn, R., and Liston, A. (2012). Separating the wheat from the chaff: Mitigating the effects of noise in a plastome phylogenomic data set from Pinus L. (Pinaceae). *BMC Evol. Biol*. 12: 100.

Penny, D., Foulds, L.R., and Hendy, M.D. (1982). Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*. 297: 197-200.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9: e1000602.

Pisani, D. (2004). Identifying and removing fast evolving sites using compatibility analysis: an example from the arthropoda. *Syst. Biol*. 53: 978-989.

Pisani, D., Mohun, M.M., Harris, S., McIterney, J.O., and Wilkinson, M. (2006). Molecular evidence for dim-light vision in the last common ancestor of the vertebrates. *Curr. Biol*. 16: 318-319.

Rasmussen, M.D. and Kellis, M. (2007). Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res*. 17: 1932-1942.

Rasmussen, M.D. and Kellis, M. (2011). A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol*. 28: 273-290.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H.,

Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. *Nature*. 444: 444-454.

Reyes-Prieto, A., and Bhattacharya, D. (2007). Phylogeny of nuclear encoded plastid-targeted proteins supports an early divergence of glaucophytes within Plantae. *Mol. Biol. Evol*. 24: 2358-2361.

Rodriguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Loffelhardt, W., Bohnert, H.J., Philippe, H., and Lang, B.F. (2005). Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. *Curr. Biol*. 15: 1325-1330.

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., and Douzery, E.J.P. (2013) Less is more in Mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol*. 30: 2134-2144.

Salichos, L. and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*. 497: 327-331.

Sanderson, M. J. and McMahon, M. M. (2007). Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol. Biol*. 7: S3.

Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C.S., Rouse, G.W., Giribet, G., and Dunn, C.W. (2011). Resolving the evolutionary relationships of mollusks with phylogenomic tools. *Nature*. 480: 364-367.

Song, S., Liu, L., Edwards, S.V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl Acad Sci. USA*. 109: 14942-14947.

Susko, E., Inagaki, Y., Field, C., Holder, M.E., and Roger, A.J. (2002). Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Mol. Biol. Evol*. 19: 1514-1523.

Swofford, D.L. (2002). PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*. 105: 437-460.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci*. 17: 57-86.

Timme, R.E., Bachvaroff, T.R., and Delwiche, C.F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One*. 7: e29696.

Tuffley, C. and Steel, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci*. 147: 63-91.

Turmel, M., Gagnon, M.C., O'Kelly, C.J., Otis, C., and Lemieux, C. (2009). The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol. Biol. Evol*. 26: 631-648.

Wang, H.C., Spencer, M., Susko, E., and Roger, A.J. (2007). Testing for covarion-like evolution in protein sequences. *Mol. Biol. Evol*. 24: 294-305.

White, T., Zhong, B., Penny, D. (2013). Beyond reasonable doubt: evolution from DNA sequences. *PLoS One*. 8: e69924.

William, J. and Ballard, O. (1996) Combining data in phylogenetic analysis. *Trends Ecol. Evol*. 11: 334.

Wodniok, S., Brinkmann, H., Glöckner, G., Heidel, A.J., Philippe, H., Melkonian, M., and Becker, B. (2011). Origin of land plants: Do conjugating green algae hold the key? *BMC Evol. Biol*. 11: 104.

Wu, C.S., Wang, Y.N., Liu, S.M., and Chaw, S.M. (2007). Chloroplast genome (cpDNA) of Cycas taitungensis and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. *Mol. Biol. Evol*. 24: 1366-1379.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol*. 39: 105-111.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol*. 11: 367-372.

Zhang, J. (2003). Evolution by gene duplication-an update. *Trends Ecol. Evol*. 18: 292-298.

Zhong, B., Yonezawa, T., Zhong, Y., and Hasegawa, M. (2010). The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol. Biol. Evol*. 27: 2855-2863.

Zhong, B., Deusch, O., Goremykin, V.V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S.V., and Lockhart, P.J. (2011). Systematic error in seed plant phylogenomics. *Genome Biol. Evol*. 3: 1340-1348.

Zhong, B., Liu, L., Yang, Z., and Penny, D. (2013). Origin of land plants using the multispecies coalescent model. *Trends Plant Sci*. 18: 492-495.

Zhou, Y., Rodrigue, N., Lartillot, N., and Philippe, H. (2007). Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evol. Biol*. 7: 206.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol*. 8: 357-366.

Chapter 2.

Zhong, B., Deusch, O., Goremykin, V.V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S.V., and Lockhart, P.J. (2011). Systematic error in seed plant phylogenomics. ***Genome Biology and Evolution***. 3: 1340−1348. (As the Corresponding author)

The phylogenetic position of Gnetales, a small group of gymnosperms comprised three genera (*Ephedra*, *Gnetum*, and *Welwitschia*), is one of the most controversial issues for the seed plant phylogeny. It had been assumed that Gnetales were a group distinct from the gymnosperms. Recent molecular phylogenies place them within the gymnosperms. To accurately elucidate the position of Gnetales, I firstly sequenced three new chloroplast genomes of southern hemisphere conifers (*Halocarpus kirkii*, *Podocarpus totara*, and *Agathis australis*) to reduce the sampling errors, and comprehensively evaluated the impact of different systematic errors (including fast-evolving sites, compositional heterogeneity and heterotachy phenomenon) for phylogenetic inference. In this study, we found that systematic errors arising from lineage-specific differences have an important impact of phylogenetic reconstruction. When these systematic errors are removed prior to tree building, the sister group relationship between the Gnetales and the Pinaceae (within Gymnosperm) is strongly favored.

I was responsible for sequencing and assembling three chloroplast genomes, and for all the data analyses. I was primarily responsible for the writing of the manuscript. All authors contributed to the final manuscript.

GBE

Genome Biology and Evolution

SMBE

Systematic Error in Seed Plant Phylogenomics

Bojian Zhong[1,2,*], Oliver Deusch[1], Vadim V. Goremykin[3], David Penny[1], Patrick J. Biggs[4], Robin A. Atherton[1], Svetlana V. Nikiforova[3], and Peter James Lockhart[1,5]

[1]Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand

[2]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

[3]Istituto Agrario San Michele all'Adige Research Center, San Michele all'Adige, Italy

[4]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

[5]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

*Corresponding author: E-mail: bjzhong@gmail.com.

Accepted: 6 October 2011

Abstract

Resolving the closest relatives of Gnetales has been an enigmatic problem in seed plant phylogeny. The problem is known to be difficult because of the extent of divergence between this diverse group of gymnosperms and their closest phylogenetic relatives. Here, we investigate the evolutionary properties of conifer chloroplast DNA sequences. To improve taxon sampling of Cupressophyta (non-Pinaceae conifers), we report sequences from three new chloroplast (cp) genomes of Southern Hemisphere conifers. We have applied a site pattern sorting criterion to study compositional heterogeneity, heterotachy, and the fit of conifer chloroplast genome sequences to a general time reversible + G substitution model. We show that non-time reversible properties of aligned sequence positions in the chloroplast genomes of Gnetales mislead phylogenetic reconstruction of these seed plants. When 2,250 of the most varied sites in our concatenated alignment are excluded, phylogenetic analyses favor a close evolutionary relationship between the Gnetales and Pinaceae—the Gnepine hypothesis. Our analytical protocol provides a useful approach for evaluating the robustness of phylogenomic inferences. Our findings highlight the importance of goodness of fit between substitution model and data for understanding seed plant phylogeny.

Key words: compositional heterogeneity, heterotachy, Gnetales, systematic error.

Introduction

Gnetales are a morphologically and ecologically diverse group of Gymnosperms, united as a monophyletic group based on special features of their cytology. Initially, they were thought to be the nearest relatives of flowering plants (angiosperms) based on the morphological similarities (the "Anthophyte" hypothesis) (Crane 1985). However, all recent molecular work has separated Gnetales away from the angiosperms and instead placed them with or within conifers. Some analyses have placed them as sister group to conifers (the "Gnetifer" hypothesis, Chaw et al. 1997), others close to Pinaceae (the "Gnepine" hypothesis, Bowe et al. 2000; Chaw et al. 2000; Finet et al. 2010; Zhong et al. 2010), and others within conifers but close to Cupressophyta (non-Pinaceae conifers; the "Gnecup" hypothesis, Nickrent et al. 2000; Doyle 2006). These alternative hypotheses are illustrated in figure 1*A*.

It has been reported that Gnetales have a faster substitution rate of sequence evolution than other gymnosperms, which could potentially cause a "long-branch attraction" (LBA) artifact in phylogenetic reconstruction (Zhong et al. 2010). The effects of LBA are well understood, even though the significance of contributing causes is often difficult to determine. These can include faster substitution rates in nonadjacent phylogenetic lineages (Felsenstein 1978), poor taxon sampling due to extinction or limited availability of some taxa (Hendy and Penny 1989), and properties of sequences not well described by stationary time reversible models. The latter include base compositional heterogeneity (Foster 2004; Jermiin et al. 2004) and lineage-specific changes in evolutionary constraint that can alter the proportion of variable sites in homologs (Lockhart and Steel 2005).

To improve taxonomic sampling of the Cupressophyta, we determined sequences for 52 genes from the chloroplast

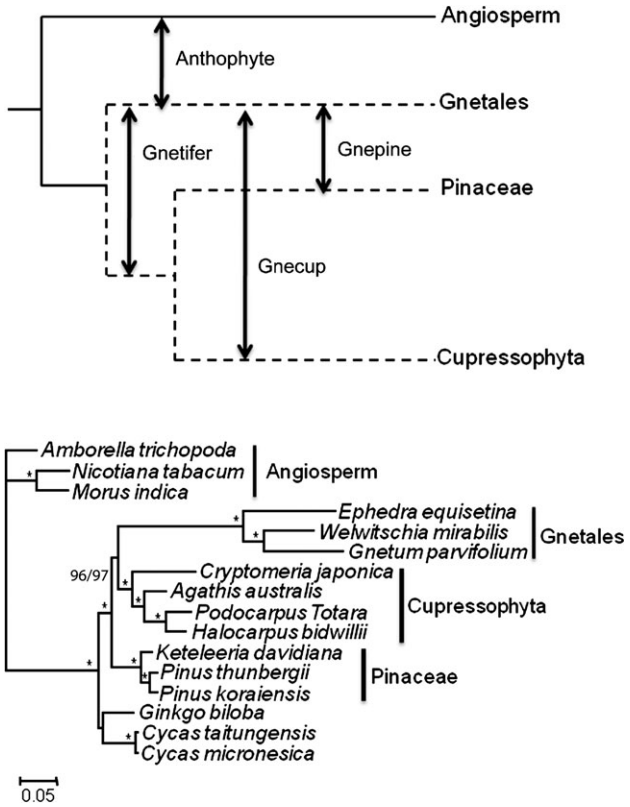


Fig. 1.—(*A*) Four major hypotheses for phylogenetic relationships involving Gnetales. (*B*) Optimal PhyML tree (GTR + G substitution model) reconstructed from all codon positions. The same topology is obtained using 1st + 2nd position sites. Bootstrap support for Gnecup hypothesis is 96% for all sites and 97% for 1st + 2nd position sites.

DNA (cpDNA) genomes of *Halocarpus kirkii*, *Podocarpus totara*, and *Agathis australis* using Illumina GAII sequencing. In phylogenetic analyses of concatenated seed plant chloroplast genome sequences, we demonstrate that sites exhibiting greatest character state variation are not well described by a time reversible substitution model. We show that this data property significantly impacts on the reconstruction accuracy of seed plant phylogeny.

Materials and Methods

Sample Collection and DNA Sequences

Tissue for Cupressophyta (*H. kirkii*, *P. totara*, and *A. australis*) was obtained with permission from the living collection at Massey University, Palmerston North. Chloroplasts were isolated and enriched DNA sequenced using the protocols described in Atherton et al. (2010). Short reads were filtered for the longest contigous subsequences below 0.05 error probability using DynamicTrim (Cox et al. 2010). Filtered reads were assembled with Velvet (Zerbino and Birney 2008) and a k-mer range from 23 to 63. Contigs were further assembled using the Geneious assembler (Drummond et al. 2011). Initial annotations for protein-coding genes were carried out using DOGMA (Wyman et al. 2004). Annotations were manually refined by comparison with genes of more closely related species.

We retrieved 13 cp genomes from the NCBI database, including the three genera of Gnetales, one Cupressophyta conifer (*Cryptomeria japonica*), three representatives of Pinaceae conifers (*Pinus thunbergii*, *Pinus koraiensis*, and *Keteleeria davidiana*), and three species from the Cycads/Ginkgo group, with three angiosperms representing the outgroup. GenBank accession numbers for gene sequences used and determined in the present study are listed in supplementary table S1 (Supplementary Material online). Fifty-two protein-coding genes were first aligned as proteins using MUSCLE (Edgar, 2004). Gaps were excluded from these alignments so that only blocks of ungapped residues bounded by similar or identical amino acids were used in phylogenetic analyses. Se-Al v2.0all (Rambaut 2002) was used to edit the underlying DNA sequences into the amino acid alignments. These alignments were then concatenated using Geneious v5.4 (Drummond et al. 2011). This approach produced an alignment of 33289 ungapped positions (not divisible by three as some gaps occur in Genbank sequences).

Sorting Sites Based on Character State Variation

The positions in our concatenated alignments were sorted based on their character state variation. As we demonstrate, this facilitated the study of systematic error in these data. Several methods have been suggested for ordering sites (e.g., discussed in Hansmann and Martin 2000; Goremykin et al. 2010). We used the method of observed variability (OV) sorting as described in Goremykin et al. (2010), which previously has been found to be efficient in concentrating saturated positions toward the most varied end of the sorted alignment. The alignment was ordered from the most highly varied sites to the most conserved sites, and a series of alignments was generated by successively shortening the OV-sorted alignment in steps of 250 sites. For each shortening step, two data partitions were obtained: 1) the shortened alignment containing the most conserved sites (partition "A") and 2) an alignment containing the more varied sites (partition "B"). After model fitting for each partition data, the maximum likelihood (ML) distance and uncorrected *p* distance were calculated using PAUP* (Swofford 2002). Two Pearson correlation analyses of pairwise distances were conducted at each shortening step: 1) correlation of the ML and uncorrected *p* distances for partition B and 2) correlation of the ML distances for partition A and B. The stopping point for site removal was determined as the point at which the two correlations showed a significant improvement (Goremykin et al. 2010).

Data Model Fit

We used MISFITS (Nguyen et al. 2011) to determine the occurrence of site patterns in our sorted alignment that were unexpected under a general time reversible (GTR) + G model using three alternative Gnetales phylogenetic trees incorporated as part of the evolutionary model. That is, given a GTR + G substitution model and weighted tree, the expected pattern likelihood vector was computed. For each entry in the vector, a simultaneous $\alpha = 95\%$ Gold confidence region was calculated. Sequence positions in the alignment indicating unexpected patterns were recorded. We also successively shortened our alignment by 250 positions and compared the log-likelihood scores for our OV-sorted alignment (partition A) to log-likelihood scores for identical length partitions jackknife resampled from the complete 33289 position alignment. PhyML 3.0 (Guindon et al. 2010) was used for log-likelihood calculations. Seqboot, implemented in the Phylip3.6 package (Felsenstein 2004), was used for jackknife resampling. *Z*-scores were calculated by subtracting the log-likelihood score on the original data from the mean log-likelihood score for the psuedoreplicate data sets and dividing by the standard deviation (SD) of mean scores.

Compositional Heterogeneity

MEGA5.0 (Tamura et al. 2011) was used to calculate the average nucleotide composition of 1) all codon sites, 1st + 2nd codon sites, and 3rd codon sites, and 2) intervals of increasing length (250 bp) beginning from the most varied end of the OV-sorted alignment. The SD of mean nucleotide frequencies was plotted to visualize compositional heterogeneity among taxa.

Phylogenetic Analyses

ML trees were built assuming a GTR + G model implemented in PhyML 3.0 (Guindon et al. 2010). The relative length of branches and extent of heterotachy (lineage-specific differences in evolutionary rate) in these trees was visualized using SplitsTree 4.0 (Huson and Bryant 2006).

Results

Effect of Improved Taxon Sampling

In ML analyses of all codon positions and 1st + 2nd sites, inclusion of the newly determined sequences from three Cupressophyta genomes halved the length of the internal branch subtending Gnetales and Cupressophyta when compared with phylogenetic reconstructions made without these taxa. Inclusion of sequences from these additional genomes did not change the topology. In the trees with additional taxa, the Gnecup hypothesis (fig. 1*B*) was strongly supported (96% and 97% bootstrap support for all positions and 1st + 2nd sites, respectively). However as we show below, support for this hypothesis was also strongly dependent on the inclusion of sites in the data that showed a poor fit to the GTR + G substitution model.

The Impact of Site Removal

We used the OV sorting criterion of Goremykin et al. (2010) to rank site patterns from most varied to least varied. Blocks of columns in steps of 250 sites were then removed sequentially. This produced a series of shortened alignments. ML trees under a GTR + G model were reconstructed for each partition, and the bootstrap support for alternative hypotheses was measured for each partition. This analysis was made for all sites, 1st + 2nd codon position sites, and 3rd codon position sites. Figure 2*A* (all sites) shows that the Gnecup hypothesis was favored only while the 2000 most varied positions were included in the analysis. After these sites were removed, the Gnepine hypothesis became favored until 3,250 sites were removed. After this point, alternative hypotheses were unresolved. With 1st and 2nd codon position data alone, the Gnepine hypothesis was favored after removal of 750 sites and before removal of 1,250 sites (fig. 2*B*). With 3rd codon position data, the Anthophyte hypothesis was initially weakly supported, but this support decreased as sites were removed (fig. 2*C*).

Data Model Fit

To help understand the impact of site removal, we investigated the fit of site patterns to three alternative evolutionary models (Gnecup, Gnepine, and Gnetifer trees) that assumed an optimal GTR + G substitution model. Using MISFITS (Nguyen et al. 2011), we computed the overrepresented and underrepresented site patterns in the OV-sorted data. For the Gnepine hypothesis, we observed that 46%of the sites not fitting the evolutionary model occurred within the 2250 most varied positions (i.e., in 7% of the total alignment length; 15% of all variable sites). About 3.1% (691/22193) of the 1st + 2nd position sites and 15.2% (1687/11096) of the 3rd position sites do not fit the Gnepine tree. A similar poor fit was also obtained for tree topologies that supported the Gnetifer and Gnecup hypotheses (fig. 3), suggesting that in the most varied positions of the OV-sorted alignment, misspecification was a general property of the GTR + G substitution model and not specific to any one hypothesis of evolutionary relationship.

To further evaluate the impact of the most varied positions on data model fit with our three tree models, we also compared the log-likelihood scores for the sequentially shorted (partition A) data sets, with scores for identical length data sets comprised of jackknife resampled site patterns taken from the original 33289 position alignment. The results from this analysis corroborated those obtained with MISFITS in identifying an extremely poor data model fit for sites at the most varied end of the OV-sorted alignment (supplementary fig. S1, Supplementary Material online).

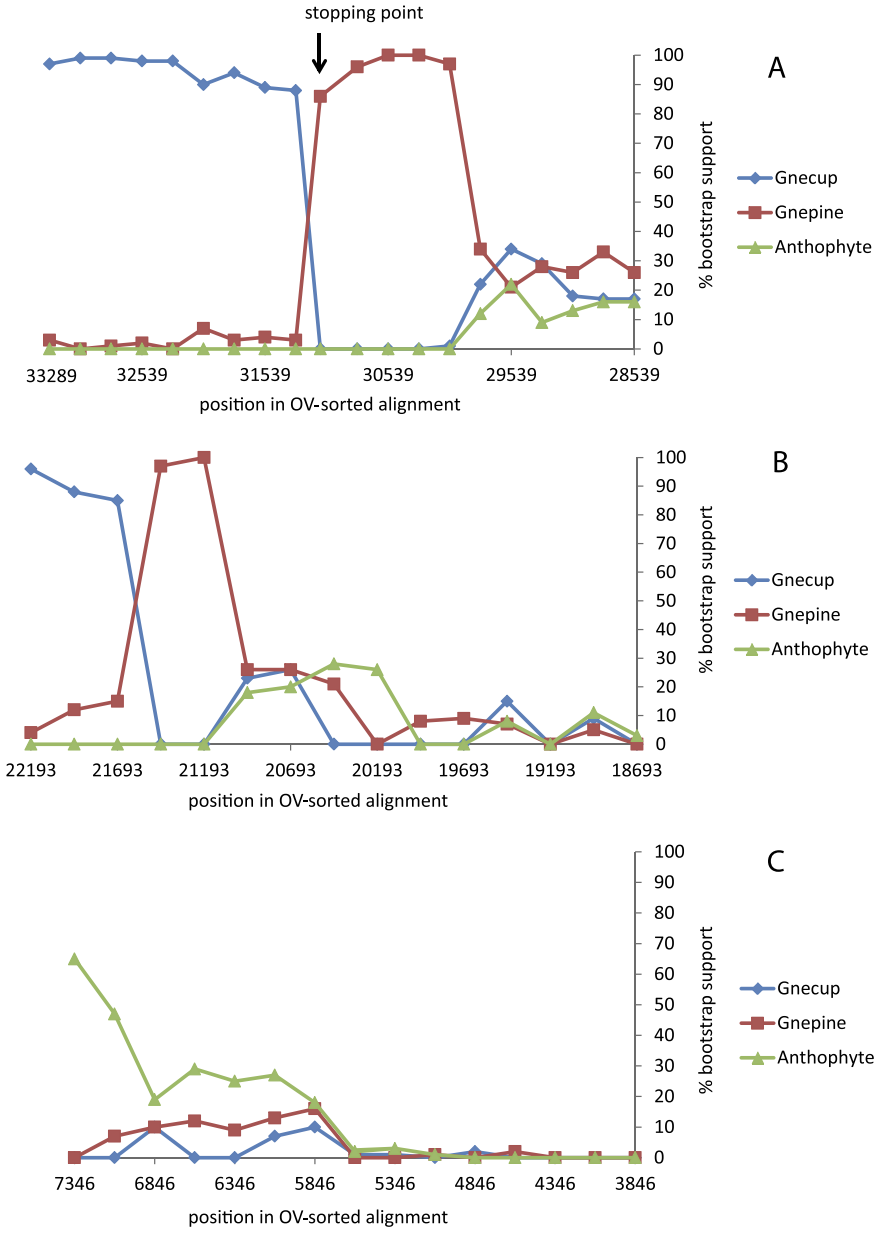


Fig. 2.—Bootstrap support in optimal PhyML trees for three alternative relationships as intervals of 250 bases were successively removed from the OV-sorted alignment. (*A*) all sites, (*B*) 1st + 2nd codon positions, and (*C*) 3rd codon positions.

Compositional Heterogeneity

Figure 4 shows the SD of individual base frequencies from mean (stationary) estimates for intervals increasing in length by 250 bases sampled from the most varied end of the OV-sorted alignment. While the average nucleotide compositional frequencies of all sites, 1st + 2nd sites, and 3rd sites are relatively homogeneous (Results not shown), the most varied OV-sorted sites in the alignment exhibit significant compositional heterogeneity. This decreases incrementally toward the more conserved positions of the OV-sorted alignment.

Heterotachy

Optimal PhyML trees (GTR + G substitution model) were reconstructed for sampling intervals that increased in length by 250 bases from the most varied end of the OV-sorted alignment. The relative length of the Gnetales internal branch separating Gnetales from other species in the 16 taxon data set for each sampling interval is shown in figure 5*A*. The relative length of the branches subtending the Cupressophyta, Pinaceae, and angiosperms in the 13 taxon data set is shown in figure 5*B*. A striking feature of the 16 taxon trees is that the branch leading to the Gnetales

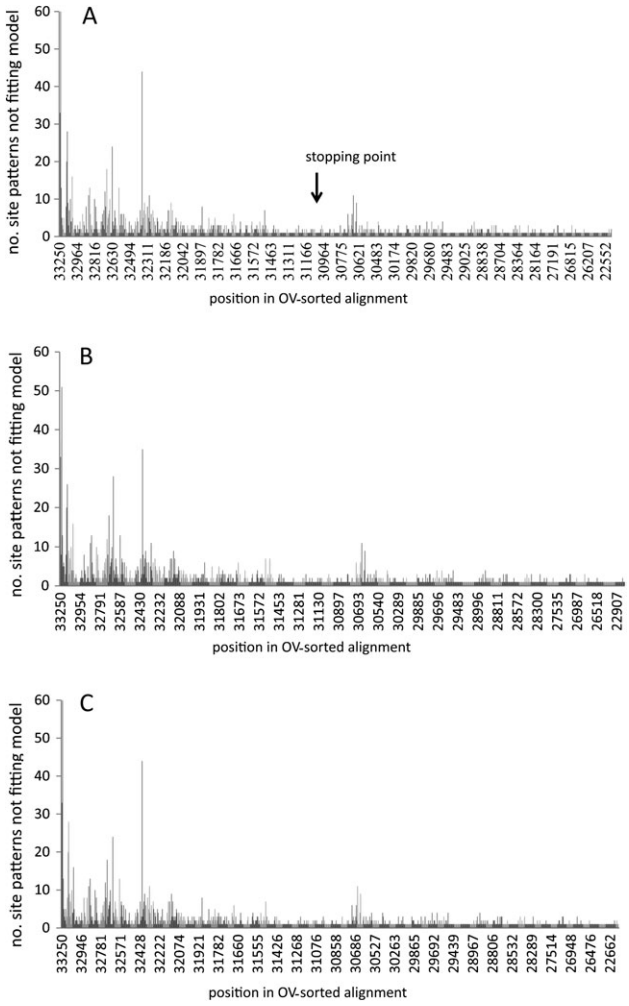


Fig. 3.—Histogram indicating consecutive misfitting site patterns under the (*A*) GTR + G + Gnepine, (*B*) GTR + G + Gnetifer, and (*C*) GTR + G + Gnecup evolutionary model. The height of each histogram indicates the number of unexpected site patterns.

lineage is disproportionately much longer than branches subtending other seed plant lineages (more than 60× longer over the first 1750 bases and between 10×–5× between 2000 and 2500 bases) at the most varied end of the OV-sorted alignment (fig. 5). This extreme branch length difference is a feature of both the 1st + 2nd codon position and 3rd codon position data (not shown).

Removal of Most Varied Sites from the Alignment

We used the stopping criterion of Goremykin et al. (2010) to make an assessment of the number of most varied sites that should be excluded prior to tree building. This criterion considers the alignment partitions created by the sequential shortening steps described previously and compares 1) ML distances for the conserved (A) and the variable (B) bipartition and 2) *p* distances and ML distances for the B partition. The authors have suggested that the removal of

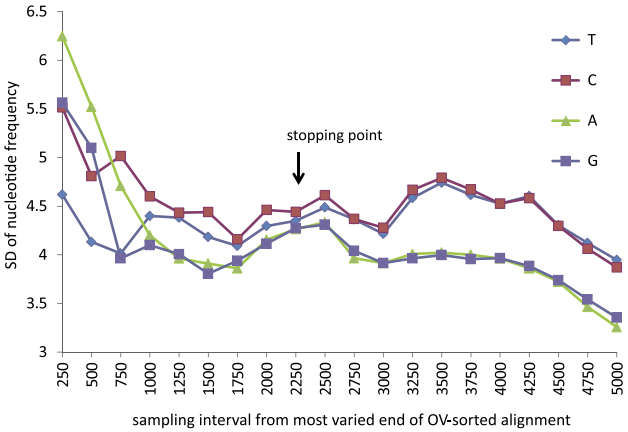


Fig. 4.—Plot indicating nucleotide compositional heterogeneity within intervals sampled from the most varied end of the OV-sorted alignment. Subsequent intervals increased in length by 250 bases per interval.

variable positions should be continued at least until the very end of the sharp rise in Pearson correlation values in either analysis. The stopping criterion identifies the point where the substitution properties of most varied sites (partition B) become more similar to those of the more conserved sites in the alignment (partition A), and where corrected and uncorrected distances for the variable B partition begin to show a strong positive correlation. As such it provides a means to objectively decide a cutoff point for excluding from tree building sites that exhibit site saturation and or model misspecification. Figure 6 indicates change in the correlation coefficient (*r*) and similarity of distances estimates as sites are removed. A sharp rise in (*r*) occurs when 2,000 sites have been removed and it ceases with removal of 2,250 sites in the correlation of *p* distances and ML distances estimated from B partitions. Reference to figure 5 shows that this is accompanied by reduction of heterotachy associated with the Gnetales lineage. It also marks the transition zone for bootstrap support of the Gnecup and Gnepine hypotheses. The Gnepine hypothesis is strongly favored after removal of 2,250 sites (position 31039). It continues to be favored until 3,250 sites are removed when the PhyML trees become unresolved.

Discussion

Most phylogenetic methods assume that DNA sequences have evolved under stationary, reversible, and homogeneous conditions. Violation of this model assumption is well known to lead to inaccurate tree reconstruction (e.g., Lanave et al 1984; Lockhart et al. 1994; Foster 2004; Jermiin et al. 2004; Delsuc et al. 2005; Lockhart and Steel 2005). Our MISFIT analyses indicate a poor fit between the most varied nucleotide sites in the Gnetales chloroplast concatenated data set and a GTR + G model—one of the more

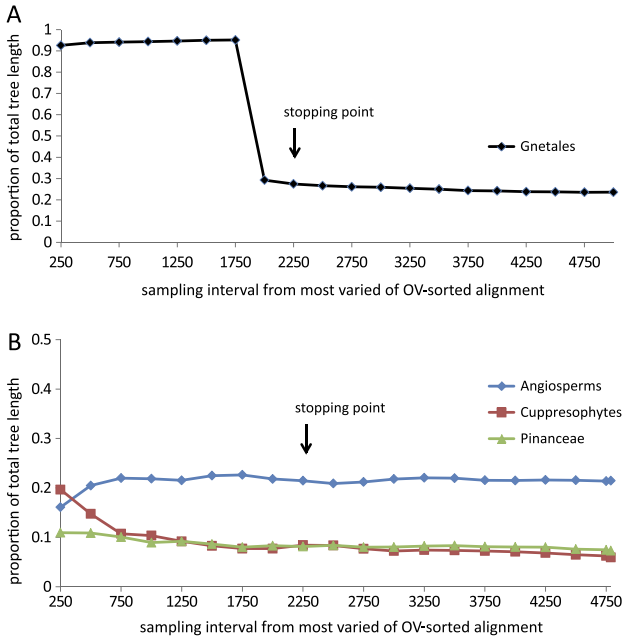


Fig. 5.—Relative length of internal branch leading to (*A*) Gnetales in a 16 taxon data set; (*B*) non-Pinaceae, Pinaceae, and Angiosperms in a 13 taxon data set (this second data set excluded Gnetales). The branch lengths are shown as a proportion of total tree length. Optimal PhyML trees were reconstructed for the same sampling intervals as used in figure 4.

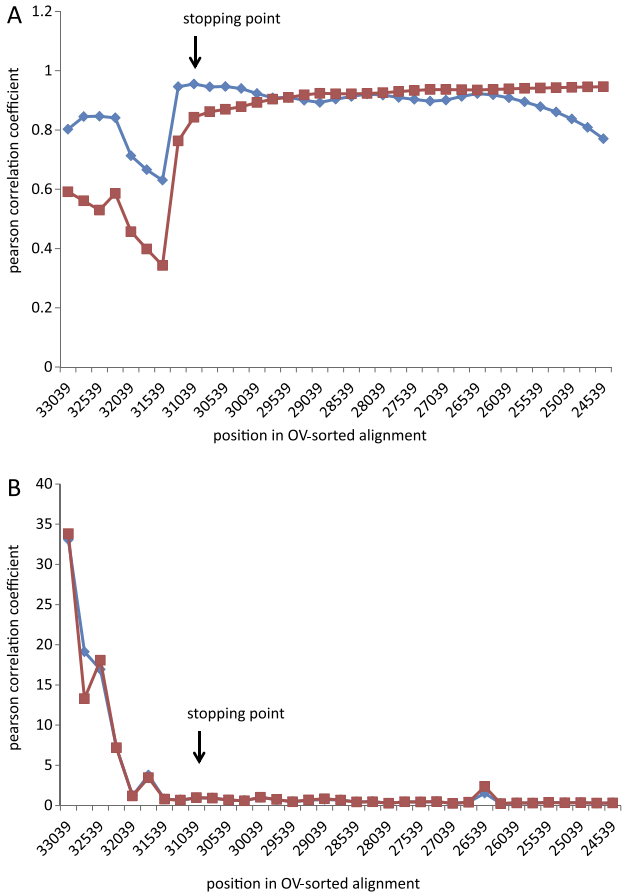


Fig. 6.—(*A*) Pearson correlation analyses. The blue dotted line indicates the Pearson correlation coefficients (*r*) of ML distances for (the more conserved) partition "A" and (less conserved) partition "B". The red dotted line represents *r* value of uncorrected *p* distances and ML distances for partition B. The *r* values begin to increase sharply at the eighth shortening step (31289 position remained). (*B*) Mean deviations of ML distances from *p* distances for B partitions. The red dotted line shows deviations between *p* distances and ML distances calculated using the best-fitting ML model as determined by ModelTest (Posada and Crandall 1998) using the *Akaike information criterion* (the neighbor joining tree was used to estimate ML model parameters). The blue dotted line indicates the deviation between *p* distances and ML distances calculated as above but using an ML tree to fit model parameters.

general models of substitution currently used in phylogenetic reconstruction. Although more complex mixture models exist (e.g., such as the CAT model, Lartillot and Philippe 2004), like GTR + G, they also assume a stationary distribution of base frequencies and have the expectation for a constant proportion of variable sites in all sequences.

Deviation from compositional homogeneity occurs in the most varied regions of the OV-sorted alignment. However, this heterogeneity extends past the OV sorting stopping point and shows no obvious relationship to it. Thus, compositional homogeneity appears an insufficient explanation for the significant increase in value of the Pearson statistic after removal of 2,000 sites and an insufficient explanation for the extent of poor model fit observed in the most varied part of the OV-sorted alignment.

More important for explaining the sharp rise in the Pearson statistic is the extent of substitution rate difference inferred for the Gnetales lineage across the sampling intervals at the most varied end of the OV-sorted alignment. This property of the aligned data causes high variance in ML distance estimation between Gnetales and other species when estimates are made from B partitions. This property of the sorted data explains much of the Pearson coefficient behavior in the correlation analyses. By the final shortening step, at 2250 bases, the relative length of the internal branch separating Gnetales shows approximately 60× reduction in length. This reduction is accompanied by a rapid change in the bootstrap support for the Gnepine hypothesis.

The extreme branch length differences between Gnetales and other lineages for sites at the most varied end of the OV-sorted alignment suggests an issue with alignment of some amino acid positions, despite a conservative approach being used in generating the sequence alignments in the present study. To investigate this further, we also aligned seed plant DNA sequences using the approach of Goremykin et al. (2010) and excluded regions of low sequence similarity (analyses not shown). Working with these alignments, we

also obtained very similar results and conclusions regarding heterotachy, compositional heterogeneity, misfit analyses, and bootstrap support. Thus, we conclude that heterotachy is a strong feature of the data and is not a feature of a specific alignment method.

Very recently, a similar study has been undertaken to that reported here. Wu et al. (2011) have determined chloroplast genome sequences for five Cupressophytes and a cycad. They also studied the phylogenetic placement of Gnetales with respect to other seed plants. Our general conclusions are similar to theirs—phylogenetic reconstruction of Gnetales in seed plant phylogeny is misled by non-time reversible properties of aligned chloroplast sequences. From their sampling of taxa, Wu et al. (2011) obtain stronger evidence than we do for lineage-specific change in the Cupressophyta that parallels Gnetales. Our studies also differ in that these authors did not evaluate the relative contribution of compositional heterogeneity and heterotachy in causing problems for tree building. Our analyses suggest that heterotachy is a more significant cause of systematic error in the seed plant sequences analyzed. As we have discussed below, our analyses also suggest that removal of sites rather than individual genes provides a better strategy for dealing with this problem.

Wu et al. (2011) divided chloroplast sequences into L (low heterotachy) and H (high heterotachy) genes and provide evidence that only phylogenetic inference from genes in the L data set is reliable. The H data set contains genes involved in translation including the rpo genes, which previously have been shown to exhibit nonconservative substitutions, indels, and increased proportions of variable sites in green algae (Lockhart et al. 2006). Our analyses indicate that while heterotachy is most pronounced in genes of the H data set, a significant level of heterotachy also occurs in the L data set for conifers that we have studied (not shown). There is also a significant amount of useful phylogenetic information in the H genes, as indicated from our results that favor the Gnepine hypothesis. This conclusion is based on an analysis of 31,039 sites, whereas that of Wu et al. (2011) is based on 21945 DNA positions (7,315 amino acids in the L data set). In general, we suspect that it will be more phylogenetically informative to remove model violating sites rather than genes prior to phylogenetic analyses.

Wu et al. (2011) suggest that the example of Gnetales follows the classic LBA scenario of Felsenstein (1978), wherein there is LBA between Gnetales and Cupressophyta. However, it is important to note that while similar, the LBA scenario for seed plants is likely to differ from this. The properties of seed plant sequences better fit the LBA scenario described by Lockhart and Steel (2005) in which proportions of variable sites change in a lineage-specific fashion, and where parallel changes occur (Zhong et al. 2010) because of similar proportions and convergent patterns of variable sites (modeled in Gruenheit et al. 2008). The significance of the difference in scenarios is important because current methods of tree building do not model lineage-specific change the proportion of variable sites in homologues (Lockhart and Steel 2005; Lockhart et al. 2006; Gruenheit et al. 2008; Shavit Grievink et al. 2008). Although it is possible to model changes in proportions of variable sites using branch length mixtures, these can be complex under some scenarios and thus problematic to identify (Matsen and Steel 2007; Gruenheit et al. 2008; Lartillot et al. 2009). Furthermore, Wu et al. (2011) observe that a mixture branch lengths model was unsuccessful in alleviating LBA with the H data set.

Conclusions

Observations of a poor fit between fast-evolving sites and time reversible models such as the GTR + G model of sequence evolution are not novel (e.g., Sullivan et al. 1995; Goremykin et al. 2004). However, the significance of having a poor fit becomes much more obvious in analysis of concatenated sequences. In the present study, systematic error arising from lineage-specific differences in evolutionary constraint dominates phylogenetic signal and misleads phylogenetic reconstruction. When systematic error contributing to most of the model misfit is removed prior to tree building, our analyses favor the Gnepine hypothesis for seed plant phylogeny (Bowe et al. 2000; Chaw et al. 2000; Finet et al. 2010; Zhong et al. 2010; Soltis et al. 2011; Wu et al. 2011).

We studied site removal in the context of substitution model misspecification and the stopping criterion of Goremykin et al. (2010). With respect to this, our study provides more insight into the performance of this method. Our results indicate that use of the stopping criterion also removes sites that provide a poor fit to tree-building assumptions. Although this criterion does not remove all model violating sites from data, it removes sites that significantly impact on phylogenetic estimates and thus sites most important for misleading tree building. Thus, it provides a useful tool to guide phylogenomic analyses.

Wu et al. (2011) note that improved taxon sampling was insufficient to overcome LBA between Curessophytes and Gnetales. We also obtained this result. However, we wish to be more positive about the contribution that improving taxon sampling of conifers will make to phylogenetic reconstruction of seed plant phylogeny. In our study, addition of sequences from three Cupressophytes reduced the length of the internal branch leading to Gnetales and Cupressophytes 2-fold, even if it was not sufficient to change the topology. Together with international efforts currently underway to sequence and analyze conifer genomes, we believe that analytical approaches such as those used here will be essential for evaluating and mitigating the impact of systematic error in large-scale phylogenomic data sets for seed plants.

Supplementary Material

Supplementary table S1, figure S1, and data matrix concatenated gapped alignment are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals. org/).

Acknowledgments

This work was financially supported by the Allan Wilson Centre, Massey University, the New Zealand Royal Society, and the Deutscher Akademischer Austausch Dienst. We thank Jennifer Tate, the anonymous reviewers, and the associate editor for helpful comments.

Literature Cited

Atherton RA, et al. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods. 6:22.

Bowe LM, Coat G, dePamphilis CW. 2000. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc Natl Acad Sci U S A. 97:4092–4097.

Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci U S A. 97:4086–4091.

Chaw SM, Zharkikh A, Sung HM, Lau TC, Li WH. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol Biol Evol. 14:56–68.

Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics. 11:485.

Crane PR. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. Ann Mo Bot Gard. 72:716–793.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

Doyle JA. 2006. Seed ferns and the origin of angiosperms. J Torrey Bot Soc. 133:169–209.

Drummond AJ, et al. 2011. Geneious v5.4. Auckland (New Zealand): Biomatters, Ltd. [cited 2011 Aug 3]. Available from: http://www.geneious.com/.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Felsenstein J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool. 27:401–410.

Felsenstein J. 2004. PHYLIP (phylogeny inference package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Finet C, Timme RE, Delwiche CF, Marletaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr Biol. 20:2217–2222.

Foster PG. 2004. Modeling compositional heterogeneity. Syst Biol. 53:485–495.

Goremykin VV, Hirsch-Ernst KI, Woelfl S, Hellwig FH. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 21:1445–1454.

Goremykin VV, Nikoforova SV, Bininda-Emonds OPP. 2010. Automated removal of noisy data in phylogenomic analyses. J Mol Evol. 71:319–331.

Gruenheit N, Lockhart PJ, Steel M, Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol. 25:1512–1520.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hansmann S, Martin WT. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int J Syst Evol Microbiol. 50:1655–1663.

Hendy M, Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst Zool. 38:297–309.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst Biol. 53:638–643.

Lanave C, Preparata G, Sacone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol. 20:86–93.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for calculating evolutionary substitution rates. Mol Biol Evol. 21:1095–1109.

Lockhart PJ, et al. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.

Lockhart PJ, Steel MA. 2005. A tale of two processes. Syst Biol. 54:948–951.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol. 11:605–612.

Matsen FA, Steel MA. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst Biol. 56:767–775.

Nguyen MAT, Klaere S, von Haeseler A. 2011. MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. Mol Biol Evol. 28:143–152.

Nickrent DL, Parkinson CL, Palmer JD, DuV RJ. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. Mol Biol Evol. 17:1885–1895.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14:817–818.

Rambaut A. 2002. Se-Al. Sequence alignment editor v2.0a11. Edinburgh (UK): Andrew Rambaut. [cited 2011 Aug 15] Available from: http://tree.bio.ed.ac.uk/software/seal/.

Shavit Grievink L, Penny D, Hendy MD, Holland BR. 2008. Lineage SpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. BMC Evol Biol. 8(1):317.

Soltis DE, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot. 98:704–730.

Sullivan J, Holsinger KE, Simon C. 1995. Among-site variation and phylogenetic analysis of 12s rRNA in sigmodontine rodents. Mol Biol Evol. 12:988–1001.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

Wu C-S, Wang Y-N, Hsu C-Y, Lin C-P, Chaw S-M. Forthcoming 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol.

Wu C-S, Wang Y-N, Hsu C-Y, Lin C-P, Chaw S-M. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and Cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol. Advance Access published September 19, 2011, doi:10.1093/gbe/evr095.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Zhong BJ, Yonezawa T, Zhong Y, Hasegawa M. 2010. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol. 27:2855–2863.

Associate editor: Martin Embley

Supplementary Information

Fig. S1. Plot indicating improvement in fit between data and evolutionary model (GTR + G and either Gnepine, Gnecup or Gnetifer tree) as the OV sorted alignment was progressively shortened at its most varied end by 500 base intervals. Z scores have been calculated as log likelihood score for the evolutionary model on the OV sorted alignment minus the mean log likelihood score for the evolutionary model on 100 data sets (jackknife resampled positions)/standard deviation of the mean scores. Most improvement in score occurs with removal of the first 1000 positions (comparison of two regressions i) data points 1-5 and ii) 6-12 shows that the initial slope is steeper p=<0.0001).

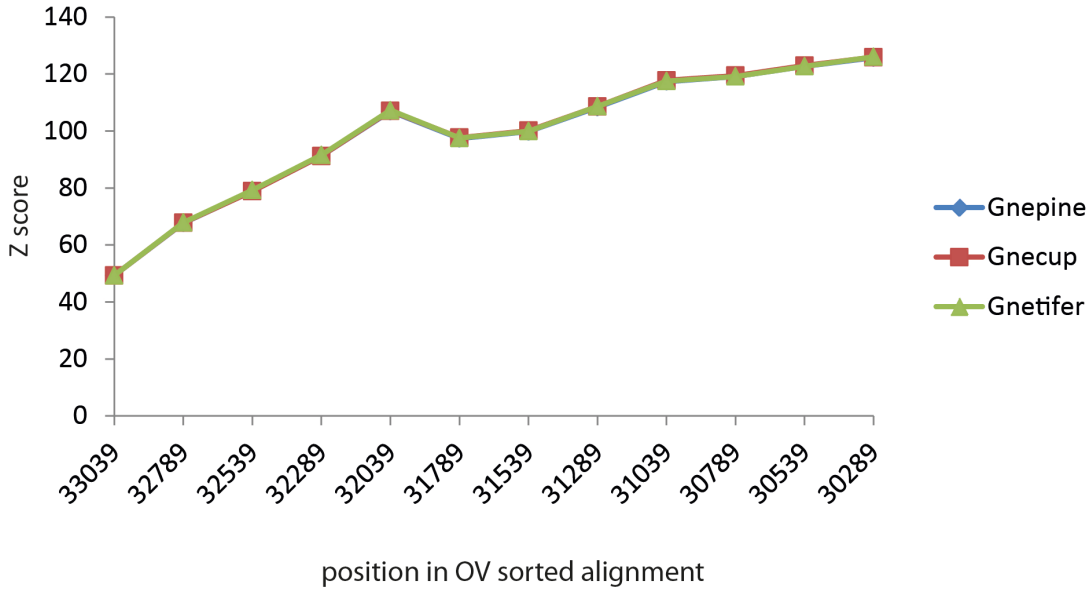


Table S1. Accession numbers and references for the 16 taxa used in this study

Taxon	Accession number	Reference
Angiosperms		
Nicotianatabacum	NC_001879	Shinozaki et al. (1986)
Morusindica	NC_008359	Ravi et al. (2006)
Amborellatrichopoda	NC_005086	Goremykin et al. (2003)
Gymnosperms		
Cycasmicronesica	EU016802–EU016882	Jansen et al. (2007)
Cycastaitungensis	NC_009618	Wu et al. (2007)
Ginkgo biloba	DQ069337-DQ069702	Leebens-Mack et al. (2005)
Pinusthunbergii	NC_001631	Wakasugi et al. (1994)
Pinuskoraiensis	NC_004677	Noh et al. 2003
Keteleeriadavidiana	AP010820	Wu et al. (2009)
Gnetumparvifolium	AP009569	Wu et al. (2009)
Welwitschia mirabilis	EU342371	McCoy et al. (2008)
Ephedra equisetina	AP010819	Wu et al. (2009)
Cryptomeria japonica	AP009377	Hirao et al. (2008)
Halocarpus sp.	JN627246-JN627297	this study
Podocarpus totara	JN627350-JN627401	this study
Agathis australis	JN627298-JN627349	this study

Table References

Goremykin VV, KI Hirsch-Ernst, S Wolfl, FH Hellwig. 2003. Analysis of the Amborella trichopoda chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol Biol Evol. 20:1499-1505.

Hirao T, A Watanabe, M Kurita, T Kondo, K Takata. 2008. Complete nucleotide sequence of the Cryptomeria japonica D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. BMC Plant Biol. 8:70.

Jansen RK, Z Cai , LA Raubeson , H Daniell , CW dePamphilis , J Leebens-Mack , K F Müller et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA. 104:19369-19374.

Leebens-Mack J, LA Raubeson, L Cui, JV Kuehl, MH Fourcade, TW Chumley , JL Boore , RK Jansen, CW dePamphilis. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol. 22:1948-1963.

McCoy SR, JV Kuehl, JL Boore, LA Raubeson. 2008. The complete plastid genome sequence of Welwitschia mirabilis: an unusually compact plastome with accelerated divergence rates. BMC Evol Biol. 8:130.

Noh EW, JS Lee, YI Choi, MS Han, YS Yi, SU Han. 2003. Complete nucleotide sequence of Pinus koraiensis. Direct Submission to GenBank. Accession No. NC_00467.

Ravi V, JP Khurana, AK Tyagi, P Khurana. 2006. The chloroplast genome of mulberry: complete nucleotide sequence, gene organization, and comparative analysis. Tree Genet. Genomes. 3:49-59.

Shinozaki K, M Ohme, M Tanaka et al. (23 co-authors). 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. **5**:2043–2049.

Wakasugi T, J Tsudzuki, S Ito, K Nakashima, T Tsudzuki, and M Sugiura. 1994. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine Pinus thunbergii. Proc Natl Acad Sci USA. 91:9794-9798.

Wu CS, YN Wang, SM Liu, SM Chaw. 2007. Chloroplast genome (cpDNA) of Cycas taitungensis and 56 cp protein-coding genes of Gnetum parvifolium: Insights into cpDNA evolution and phylogeny of extant seed plants. Mol Biol Evol. 24:1366-1379.

Wu CS, YT Lai, CP Lin, YN Wang, SM Chaw. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. Mol Phylogenet Evol. 52:115-124.

Chapter 3.

Zhong, B., Liu, L., Yang, Z., and Penny, D. (2013). Origin of land plants using the multispecies coalescent model. ***Trends in Plant Science***. 18: 492-495. (As the Corresponding author)

Zhong, B., Xi, Z., Goremykin, V.V., Fong, R., McLenachan, P.A., Novis, P., and Penny, D. (2014). Origin of land plants revisited using heterogeneous models and three new algal chloroplast genomes. ***Molecular Biology and Evolution.*** 31: 177-183. (As the Corresponding author)

The origin of land plants is a fundamental topic in plant evolutionary biology. Analyses of both morphological and molecular data have established land plants as a monophyletic group that evolved from within streptophyte algae, but which group of streptophyte algae is most closely related to the land plants is still uncertain. Thanks to advances of next-generation sequencing technology, massive nuclear data have been sequenced from land plants and green algae. In the first manuscript of this chapter, I collected 289 nuclear genes including 32 taxa, and applied the multispecies coalescent model and the concatenated method to infer the origin of land plants. Because the topological congruence from various datasets is the key validation for phylogenomic inference, I also used chloroplast genome data to address the origin of land plants. For the second manuscript of this chapter, I reported three new chloroplast genomes from streptophyte algae: *Coleochaetae orbicularis* (Coleochaetales), *Nitella hookeri* (Charales), and *Spirogyra communis* (Zygnematales). Adding these new genomes strongly reduced any possibility of "long branch attraction" artifact. Furthermore, the chloroplast phylogenomic analyses were conducted using the various evolutionary models that account for site heterogeneity and compositional heterogeneity.

For the first manuscript, I was responsible for collecting nuclear genes, estimating species trees using the coalescent model and the concatenation method. For the second manuscript, I was responsible for sequencing and assembling the chloroplast genomes, and for all phylogenetic analyses. I was primarily responsible for the writing of both manuscripts. All authors contributed to the final manuscripts.

Cell PRESS

Origin of land plants using the multispecies coalescent model

Bojian Zhong[1,2*], Liang Liu[3*], Zhen Yan[3], and David Penny[1]

[1] Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand
[2] Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand
[3] Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30606, USA

The origin of land plants is a fundamental topic in plant evolutionary biology. Despite the crucial importance for knowing the closest lineages of land plants, this question has not been fully answered yet. Using recently available nuclear sequences from streptophyte algae, the multispecies coalescent model produces a congruent phylogeny that is robust to different data sets, in contrast to the conflicting phylogenies produced by the concatenation method. Using phylogenomic data and the coalescent model, in this opinion article we postulate that the Zygnematales are the closest lineages of land plants. We suggest that the coalescent model can accommodate gene tree heterogeneity in deep-level phylogenies and can be potentially used to resolve other deep species phylogenies.

The enigmatic origin of land plants

The colonization of the terrestrial habitat by the ancestors of land plants (embryophytes) approximately 500–450 Mya [1] is a key event in evolution and has led to important environmental changes on earth, including a gradual increase in oxygen concentration [2] and development of the terrestrial ecosystems [3]. It is widely accepted that land plants evolved from streptophyte algae (a diverse group of green, fresh water algae), which consist of five orders: Chlorokybales, Klebsormidiales, Charales (stoneworts), Coleochaetales, and Zygnematales [4]. Knowing the closest lineages of land plants would allow insights into both the transition from multicellular green algae to terrestrial environments and the morphological character evolution within streptophyte algae.

Although genome-scale data from green algal lineages have been used to infer the origin of land plants, the closest sister group of land plants remains uncertain. Initially, Charales were placed as the sister group to the land plants using four genes with broad sampling of taxa [5]. Some recent studies based on multi-gene concatenation analyses of both plastid and nuclear data indicated that Charales are not the sister group of land plants, but three related relationships were suggested: (i) Zygnematales [6–9], (ii) Coleochaetales [10,11], or (iii) Zygnematales and Coleochaetales combined [12,13] as closest to land plants. Thus, the origin of land plants remained unresolved even though these studies applied dense taxon sampling and complex phylogenetic models to analyze the multi-gene data. One of the challenges facing the multi-gene phylogenetic analysis is the observation of a tremendous amount of variation among individual gene trees. In the face of highly incongruent gene trees, simply concatenating sequences across genes and building a phylogenetic tree from concatenated sequences (i.e., concatenation methods) may produce an inconsistent estimate of the species tree [14]. The multispecies coalescent model describes individual gene trees as the outcomes of a stochastic process (coalescence process) along the lineages of the species tree. Because gene trees are allowed to vary in the multispecies coalescent model, coalescent methods can consistently estimate species trees in spite of the presence of heterogeneous gene trees [15–17]. In this opinion article, we collect available nuclear data to investigate the origin of land plants by using both coalescent methods and concatenation methods, and propose that the multispecies coalescent model is beneficial to accurately and consistently reconstruct species phylogenies for the origin of land plants.

Ancient rapid radiation may confound the origin of land plants

Elucidating the phylogenetic history of ancient rapid radiations is increasingly difficult as the time between divergences becomes shorter [18]. A rapid radiation at the time of terrestrial colonization by the descendants of streptophyte algae [19,20] implies that the internal branches among the streptophyte algae are relatively short [13] and incomplete lineage sorting (ILS) is likely to occur [21] and thereby impede the resolution of the origin of land plants. Because the multispecies coalescent model is able to accommodate gene tree heterogeneity due to ILS [14,22–24], it was suggested that the use of coalescent methods may help understand the origin of land plants [13].

Congruent phylogeny from coalescent phylogenetic analyses

As a start toward resolving these deep lineages, we collected 289 nuclear genes from 32 taxa in the green lineages, taking advantage of the available EST (expressed sequence tag) sequences and genomic data from streptophyte algae [6,7,12,13]. Species trees were estimated from the rooted gene trees using the Maximum Pseudo-likelihood Estimation of the Species Tree (MP-EST) method developed under the coalescent model [25]. The pseudo-likelihood function of

Corresponding author: Zhong, B. (bjzhong@gmail.com, b.zhong@massey.ac.nz)
* These authors contributed equally to the article.

1360-1385/$ – see front matter
 http://dx.doi.org/10.1016/j.tplants.2013.04.009

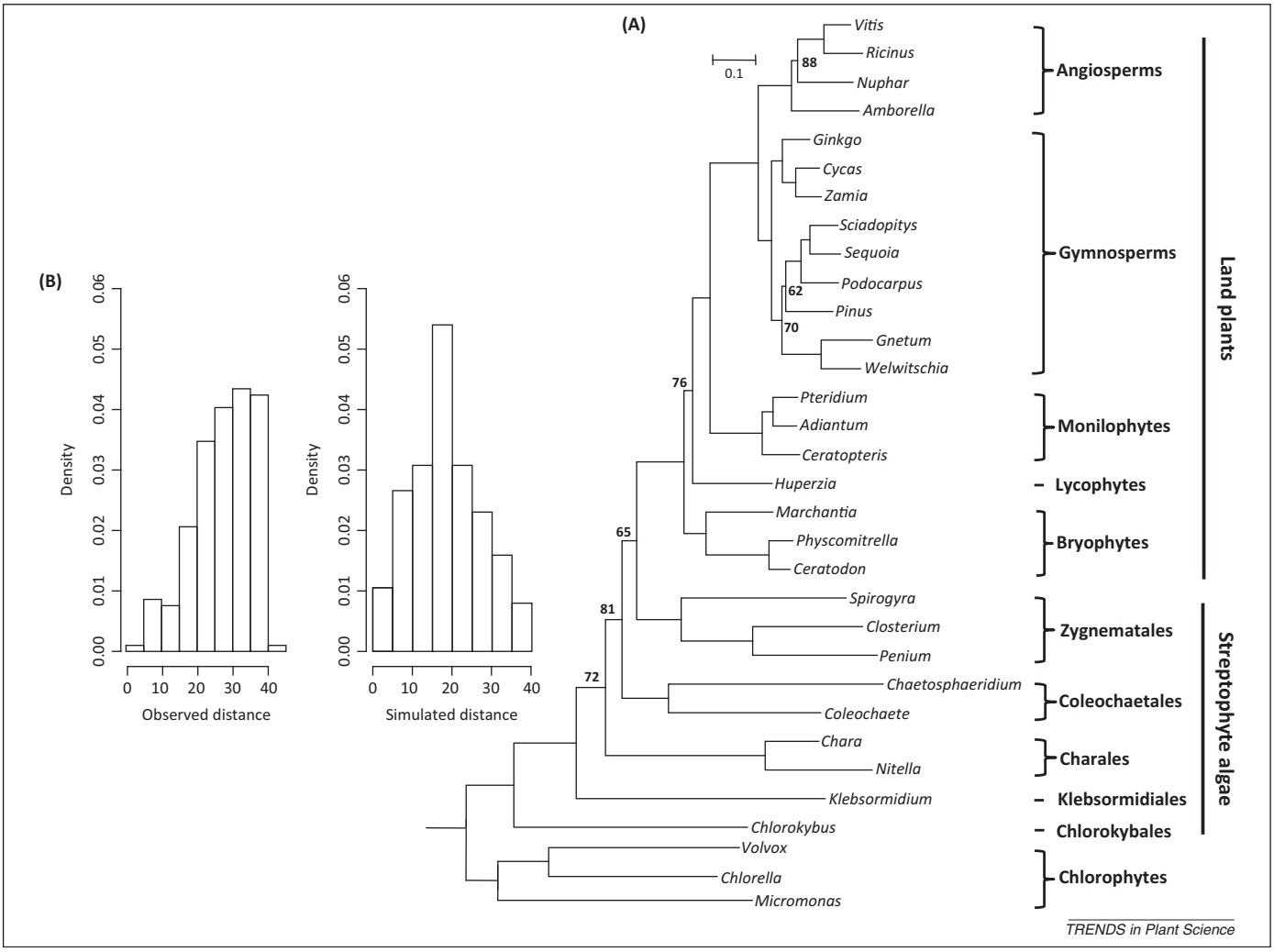


Figure 1. **(A)** The MP-EST tree for 184 genes. The numbers on the tree indicate bootstrap support values, and nodes with bootstrap values of >90% are not shown. The branch lengths (in mutation units) were estimated by fitting the concatenated 184 genes data to the MP-EST topology. Chlorophytes were used as outgroup in each case (although the outgroup is not shown on Figures S1–S11 in the supplementary material online). **(B)** Histograms of the observed and simulated distances. The average of the observed distance is 27.2, whereas the average of the simulated distance is 18.5. The ratio 18.5/27.2 = 68% is used to measure the proportion of gene tree variation (i.e., distance) that can be explained by the coalescent model. Abbreviation: MP-EST, Maximum Pseudo-likelihood Estimation of the Species Tree.

the species tree is the product of probabilities of gene tree triplets given the species tree. Because the probabilities of gene tree triplets are determined by the ratio of species divergence times (τ) and effective population sizes (θ), branch lengths of the species tree estimated by the MP-EST method are in coalescent units (i.e., τ/θ). The MP-EST tree for 289 genes supports Zygnematales as the sister group of land plants (Figure 1), but with a low bootstrap percentage (41%) (Figure S1 in the supplementary material online). We then calculated average bootstrap support values (BSVs) for the 289 individual gene trees and found that 105 gene trees were poorly supported with an average bootstrap percentage of <50%. It has been suggested that a high level of individual gene tree uncertainty can reduce the phylogenetic signal for species tree estimation [23]. We thus reconstructed a MP-EST tree using 184 genes with an average BSV of >50% (Figure 1A). The bootstrap support for (Zygnematales, land plants) on the MP-EST tree increased to 65% when the MP-EST tree was built from a set of relatively highly supported gene trees (184 gene trees). In addition, all MP-EST trees for different data sets (Table 1) consistently support the relationship (Zygnematales, land plants), although the bootstrap support varies across data sets (Table 1; Figure S1 in the supplementary material online). This indicates that the origin of land plants estimated by MP-EST is congruent across different genetic markers.

Testing the extent of gene tree heterogeneity

Gene tree heterogeneity is a major challenge for phylogenetic analyses using multi-genes or phylogenomic data [26]. To evaluate the extent of gene tree heterogeneity, we calculated the average of pairwise (symmetric) distances [27] among 184 gene trees. We observed that 182 gene trees have distinct topologies. To estimate how well the multispecies coalescent model can explain gene tree heterogeneity observed in the data, we simulated gene trees under the coalescent model. The simulated gene trees were then used to estimate gene tree variation expected under the coalescent model (see the supplementary material online). The result showed that the coalescent model could account for 68% of gene tree variation observed for 184 gene trees (Figure 1B), leaving only less than one-third of gene tree heterogeneity currently unaccounted for.

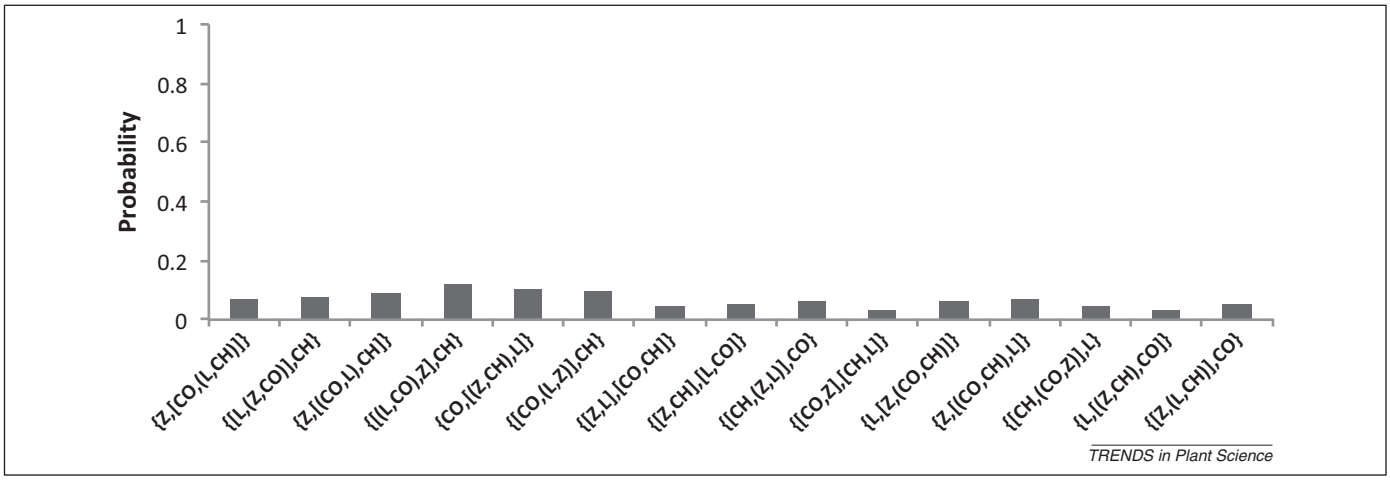


Figure 2. The distribution of 15 possible relationships of Zygnematales, Charales, Coleochaetales, and land plants among gene trees. Abbreviations: CH, Charales; CO, Coleochaetales; L, land plants; Z, Zygnematales.

We also calculated the distribution of 15 possible relationships (topologies) among four groups (Zygnematales, Coleochaetales, Charales, and land plants) in the estimated gene trees. The 15 topologies are almost evenly distributed with the maximum probability 0.118 (average probability of the 15 topologies ~0.067), suggesting a significant level of gene tree heterogeneity among these four groups (Figure 2).

Testing alternative hypotheses on the origin of land plants

The BSVs for four hypothetical relationships of the origin of land plants were calculated based on MP-EST trees constructed from different data sets (Table 1). All hypothetical relationships were weakly supported by the MP-EST analysis for 78 ribosomal genes, indicating that ribosomal genes may not have enough phylogenetic signal to resolve such deep phylogeny. Moreover, the MP-EST analyses based on different data sets consistently give higher BSVs for (Zygnematales, land plant) and lower BSVs for three alternative relationships – (Charales, land plant), [(Zygnematales, Coleochaetales), land plant], and (Coleochaetales, land plant). The highest bootstrap support for (Charales, land plant) is only 15%, whereas the highest bootstrap support for [(Zygnematales, Coleochaetales), land plant] and (Coleochaetales, land plant) is 26% and 28%, respectively. Thus, the relationship (Zygnematales, land plant) is the best hypothesis for the different data sets.

We also performed a likelihood ratio test on the three alternative relationships using the data set of 184 genes. The test significantly rejected these three relationships at α level of 0.05, thus leaving the relationship (Zygnematales, land plant) as the only supported hypothesis (Table 2; details of the likelihood ratio test are described in the supplementary material online).

Incongruent phylogenies from concatenation phylogenetic analyses

Concatenation methods, which assume that all genes have the same or similar phylogenies [28,29], have been widely used in phylogenomic studies. However, concatenation methods can yield misleading inferences of species relationships in the presence of a high level of gene tree heterogeneity [14,30]. The concatenation analyses for different data sets produced different relationships regarding the origin of land plants. The relationship (Zygnematales, land plants) was supported by the concatenation trees for 289 genes, 184 genes, and 211 non-ribosomal genes; the relationship [(Zygnematales, Coleochaetales), land plant] was supported by the concatenation trees for 78 ribosomal genes and 119 genes; and the relationship (Coleochaetales, land plant) was supported by the concatenation trees for 42 genes (Figures S2–S7 in the supplementary material online). In comparison with the MP-EST analyses, the concatenation analyses for different data sets are incapable of producing a consistent estimate of the origin of land plants.

Table 1. The bootstrap support values of four relationships on the MP-EST trees built from different data sets

Topology data	(Z, L)	(CO, L)	[(Z, CO), L]	(CH, L)
289 genes	41%	24%	8%	7%
184 genes[a]	65%	19%	6%	8%
119 genes[b]	41%	28%	12%	10%
42 genes[c]	61%	25%	6%	7%
78 ribosomal genes	28%	25%	26%	15%
211 non-ribosomal genes	44%	10%	6%	8%

Abbreviations: CH, Charales; CO, Coleochaetales; L, land plants; Z, Zygnematales.

[a]Genes with average bootstrap support values (BSVs) of >50% selected from 289 genes.

[b]Genes from Laurin-Lemay *et al.* [13] with the lowest number of missing sites.

[c]Genes with an average BSV of >50% selected from 119 genes.

Table 2. The likelihood ratio test for three alternative hypotheses on the origin of land plants

Hypotheses	Test statistic	Critical value	*P*
[(Z, CO), L]	73.44	70.54	<0.05
(CO, L)	73.32	67.65	<0.05
(CH, L)	171.4	14.25	<0.05

Abbreviations: CH, Charales; CO, Coleochaetales; L, land plants; Z, Zygnematales. The test statistic is the ratio of the maximum likelihood scores under H_1 and H_0. The critical values are obtained by a parametric bootstrap technique. A detailed explanation of the likelihood ratio test can be found in the supplementary material online.

Concluding remarks and future outlook

Owing to the likely rapid radiation of land plant ancestors and streptophyte algae, ancestral polymorphism caused by incomplete lineage sorting in the ancestral populations is retained in the deep phylogeny. The multispecies coalescent model has been proved to be able to produce accurate and congruent species trees in the presence of ancient incomplete lineage sorting [14,21,25]. In our study, a coalescent method (MP-EST) was applied to the multi-locus nuclear data in order to resolve the origin of land plants. The coalescent analyses across different data sets consistently suggest that the closest relatives of land plants are the Zygnematales, which is the most likely relationship on the MP-EST trees. Meanwhile, the other three alternative relationships [(Zygnematales, Coleochaetales), land plant], (Coleochaetales, land plant), and (Charales, land plants) suggested by previous studies are rejected by a likelihood ratio test based on the data set of 184 genes. Yet the bootstrap support for the most likely relationship (Zygnematales, land plants) is still only 65%. We anticipate that adding more taxa and more genes (i.e., additional nuclear genes, chloroplast and mitochondrial genomes) will help produce a well-supported phylogenetic relationship and make a more conclusive inference on the origin of land plants.

Acknowledgments

We thank Dr Tim White for computing assistance and Dr Peter Lockhart for helpful discussion. We thank three anonymous reviewers and editor for their constructive comments on the manuscript. This research is supported by the National Science Foundation (DMS-1222745 to L.L.), and an Allan Wilson Centre for Molecular Ecology and Evolution Doctoral Scholarship (to B.Z.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found at http://dx.doi.org/doi:10.1016/j.tplants.2013.04.009.

References

1 Gensel, P.G. (2008) The earliest land plants. *Annu. Rev. Ecol. Evol. Syst.* 39, 459–477
2 Berner, R.A. *et al.* (2003) Phanerozoic atmospheric oxygen. *Annu. Rev. Earth Planet. Sci.* 31, 105–134
3 Kenrick, P. and Crane, P.R. (1997) The origin and early evolution of plants on land. *Nature* 389, 33–39
4 Qiu, Y.L. and Palmer, J.D. (1999) Phylogeny of early land plants: insights from genes and genomes. *Trends Plant Sci.* 4, 26–30
5 Karol, K.G. *et al.* (2001) The closest living relatives of land plants. *Science* 294, 2351–2353
6 Timme, R.E. *et al.* (2012) Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE* 7, e29696
7 Wodniok, S. *et al.* (2011) Origin of land plants: do conjugating green algae hold the key. *BMC Evol. Biol.* 11, 104
8 Turmel, M. *et al.* (2006) The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. *Mol. Biol. Evol.* 23, 1324–1338
9 Chang, Y. and Graham, S.W. (2011) Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. *Am. J. Bot.* 98, 839–849
10 Turmel, M. *et al.* (2009) The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol. Biol. Evol.* 26, 631–648
11 Turmel, M. *et al.* (2009) The chloroplast genomes of the green algae *Pedinomonas minor*, *Parachlorella kessleri*, and *Oocystis solitatia* reveal a shared ancestry between the Pedinomonadales and Chlorellales. *Mol. Biol. Evol.* 26, 2317–2331
12 Finet, C. *et al.* (2012) Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr. Biol.* 22, 1456–1457
13 Laurin-Lemay, S. *et al.* (2012) Origin of land plants revisited in the light of sequence contamination and missing data. *Curr. Biol.* 22, R593–R594
14 Kubatko, L.S. and Degnan, J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24
15 Edwards, S. (2009) Is a new and general theory of molecular systematics emerging? *Evolution* 63, 1–19
16 Liu, L. *et al.* (2009) Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* 53, 320–328
17 McCormack, J. *et al.* (2009) Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Syst. Biol.* 58, 501–508
18 Whitfield, J.B. and Lockhart, P.J. (2007) Deciphering ancient rapid radiations. *Trends Ecol. Evol.* 22, 258–265
19 Stebbins, G.L. and Hill, G.J.C. (1980) Did multicellular plants invade the land. *Am. Nat.* 115, 342–353
20 McCourt, R.M. (1995) Green algal phylogeny. *Trends Ecol. Evol.* 10, 159–163
21 Degnan, J. and Rosenberg, N. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340
22 McCormack, J.E. *et al.* (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.* 22, 746–754
23 Song, S. *et al.* (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14942–14947
24 Oliver, J.C. (2013) Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* http://dx.doi.org/10.1111/evo.12047
25 Liu, L. *et al.* (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* 10, 302
26 Castresana, J. (2007) Topological variation in single-gene phylogenetic trees. *Genome Biol.* 8, 216
27 Robinson, D.F. and Foulds, L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147
28 de Queiroz, A. and Gatesy, J. (2007) The supermatrix approach to systematics. *Trends Ecol. Evol.* 22, 34–41
29 William, J. and Ballard, O. (1996) Combining data in phylogenetic analysis. *Trends Ecol. Evol.* 11, 334
30 Mossel, E. and Vigoda, E. (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309, 2207–2209

Supplementary Material

Origin of land plants using the multispecies coalescent model

Bojian Zhong[1,2,a], Liang Liu[3,a], Zhen Yan[3] and David Penny[1]

[1]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand. [2]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand. [3]Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30606, USA. [a]These authors contributed equally to this work.

Corresponding author: Zhong, B. (bjzhong@gmail.com; b.zhong@massey.ac.nz)

Data collection

The 289 nuclear genes (including 78 ribosomal genes) were collected from four publications [S1-S4] . The following five subsets were used for the analyses: (1) 184 genes (selected from 289 genes) with average bootstrap support values (BSV) of > 50%. (2) 119 genes from Laurin-Lemay et al [S4] with the lowest number of missing sites. (3) 42 genes (selected from 119 genes) with BSV of > 50%. (4) 78 ribosomal genes. (5) 211 non-ribosomal genes.

Phylogenetic inferences

The 289 individual gene trees and concatenated gene trees were inferred by RAxML [S5] and PhyML [S6] under the LG+GAMMA model [S7] which was selected as the best-fitting substitution model using ProtTest 3.0 [S8]. Species trees were estimated from the rooted gene trees using the Maximum Pseudo-likelihood Estimation of the Species Tree (MP-EST) method [S9]. The branch lengths in the MP-EST tree are in coalescent units. To estimate branch lengths in mutation units, the concatenated sequences of 184 genes were fit to the MP-EST topology by PhyML. The MP-EST trees based on different data sets were shown on Figure S1. The ML gene trees using concatenation methods were shown on Figure S2-S7. Note that the outgroup Chlorophytes were not shown on Figure S1-S11.

Gene tree heterogeneity

To evaluate the level of gene tree variation, we calculated the distribution of pairwise distances [S10] among estimated gene trees. We then built a MP-EST tree from 184

gene trees (see Figure 1A), and calculated the observed gene tree variation (denoted by O), which is equal to the average distance between 184 gene trees and the MP-EST tree. Moreover, 184,000 gene trees were generated from the MP-EST tree under the coalescent model, and the average distance between simulated gene trees and MP-EST tree were calculated as the expected gene tree variation (denoted by E). Last, we calculated the ratio of expected variation and observed variation, i.e., E/O (Figure 1B). The observed gene tree distances were compared with the distribution of the expected gene tree distances. If the observed distance of a gene tree is significantly larger than the expected gene tree distances, the gene tree is identified as an "outlier" gene tree. With this criterion, we identified 8 outlier gene trees (p-value < 0.05). The MP-EST tree was re-estimated by excluding 8 outliers from 184 genes (Figure S8), and it is congruent with the phylogeny based on 184 genes.

We calculated the Robinson-Foulds [S10] distances among 184 gene trees. The RF distance is expressed as the percentage of the distinct partitions between two trees. The average distance among 184 gene trees is 0.42. We further found a set of 114 gene trees (whose average distance is 0.367) and a set of 47 gene trees (whose average distance is 0.319). The MP-EST trees estimated by the two subsets of 184 gene trees (Figure S9 and Figure S10) consistently support Zygnmatales as the sister group of land plants. We calculated the distribution of 15 possible relationships of the four groups (Zygnematales, Coleochaetales, Charales, and land plants) across 184 gene trees. Because each of the four groups includes multiple species, we randomly chose one species to represent the group. We excluded 24 genes that have only three groups (due to missing data). The distribution of the relationships among the four groups was calculated based on 160 genes.

Amino acid compositional heterogeneity

The amino acid compositional differences among different taxa may violate standard homogeneous phylogenetic models, leading to inaccurate gene trees [S11, S12]. We used TREE-PUZZLE [S13] to identify 24 genes with significant amino acid bias from the data of 184 genes. The MP-EST tree built by removing these genes (Figure S11) is congruent with the MP-EST tree for 184 genes (Figure 1).

Testing the alternative hypotheses on the origin of land plants

The three alternative hypotheses on the origin of land plants were tested separately by a likelihood ratio test. As the same likelihood ratio test was conducted for all three hypotheses, we here use the hypothesis (Coleochaetales, land plants) as an example to demonstrate the test. The null hypothesis is that the origin of land plants is Coleochaetales and thus the relationship (Coleochaetales, land plants) is a true relationship in the tree. The test statistic of the likelihood ratio test is

$$T = \log(L_1) - \log(L_0),$$

where L_1 is the maximum likelihood score under H_1 and L_0 is the maximum likelihood score under H_0. Under the null hypothesis, the topology of the tree is constrained to have the relationship (Coleochaetales, land plants), whereas under the alternative hypothesis the topology of the tree is a free parameter. Thus the maximum likelihood score under H_1 is the likelihood score of the unconstrained MP-EST tree. The maximum likelihood score under H_0 is the likelihood score of the constrained MP-EST tree. The likelihood scores of the constrained and unconstrained MP-EST trees are calculated in the phylogenetic program MP-EST [S9]. The likelihood ratio test statistic for the relationship (Coleochaetales, land plants) is 73.32. Moreover, we use a parametric bootstrap technique to approximate the null distribution of test statistic T. We first simulated 100 bootstrap samples of gene trees from the constrained MP-EST tree (under H_0). Each bootstrap sample contains 184 gene trees. Because some taxa are missing in the original data, the missing taxa were removed from the simulated gene trees. Then we calculated the likelihood ratio test statistic for each bootstrap sample. Finally, 100 likelihood ratio test statistic are ordered and the 95% quantile of 100 test statistics is the critical value at α level of 0.05, which is 67.65. The observed test statistic $T = 73.32$ is greater than 67.65, indicating that p-value < 0.05. Thus we reject the null hypothesis that Coleochaetales is the origin of land plants. Similarly, the other two hypotheses were rejected at α level of 0.05.

Supplementary references

S1. Timme, R.E. *et al*. (2012) Broad phylogenomic sampling and the sister lineage of land plants. *PLoS ONE* 7, e29696

S2. Wodniok, S. *et al*. (2011) Origin of land plants: do conjugating green algae hold the key. *BMC Evol. Biol*. 11, 104

S3. Finet, C. *et al*. (2012) Multigene Phylogeny of the Green Lineage Reveals the Origin and Diversification of Land Plants. *Curr. Biol*. 22, 1456-1457

S4. Laurin-Lemay, S. *et al*. (2012) Origin of land plants revisited in the light of sequence contamination and missing data. *Curr. Biol*. 22, R593-R594

S5. Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688-2690

S6. Guindon, S. *et al*. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst*. *Biol*. 59, 307-321

S7. Le, S. Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol*. 25, 1307-1320

S8. Darriba D. *et al*. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164-1165

S9. Liu, L. *et al*. (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol*. 10, 302

S10. Robinson, D. F. and Foulds, L. R. (1981) Comparison of phylogenetic trees. *Math. Biosci*. 53, 131-147

S11. Foster, P. G. (2004) Modeling compositional heterogeneity. *Syst. Biol*. 53, 485–495

S12. Zhong, B. *et al*. (2010) The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol. Biol. Evol*. 27, 2855–2863

S13. Schmidt, H.A. *et al*. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18, 502–504

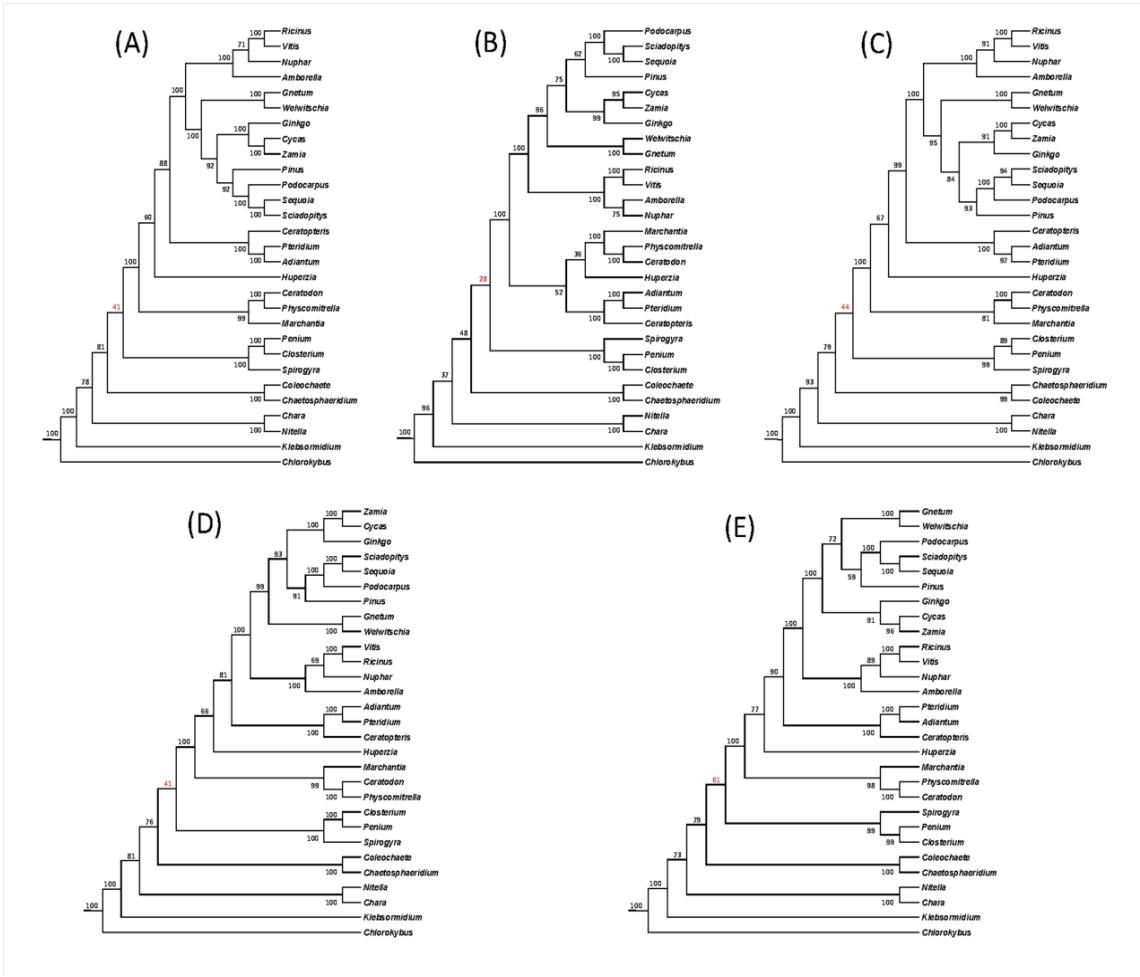


Figure S1. MP-EST trees based on different data sets. **(A)** 289 genes. **(B)** 78 ribosomal genes. **(C)** 211 non-ribosomal genes. **(D)** 119 genes (with the lowest number of missing data). **(E)** 42 genes (BSV > 50%) from 119 genes.

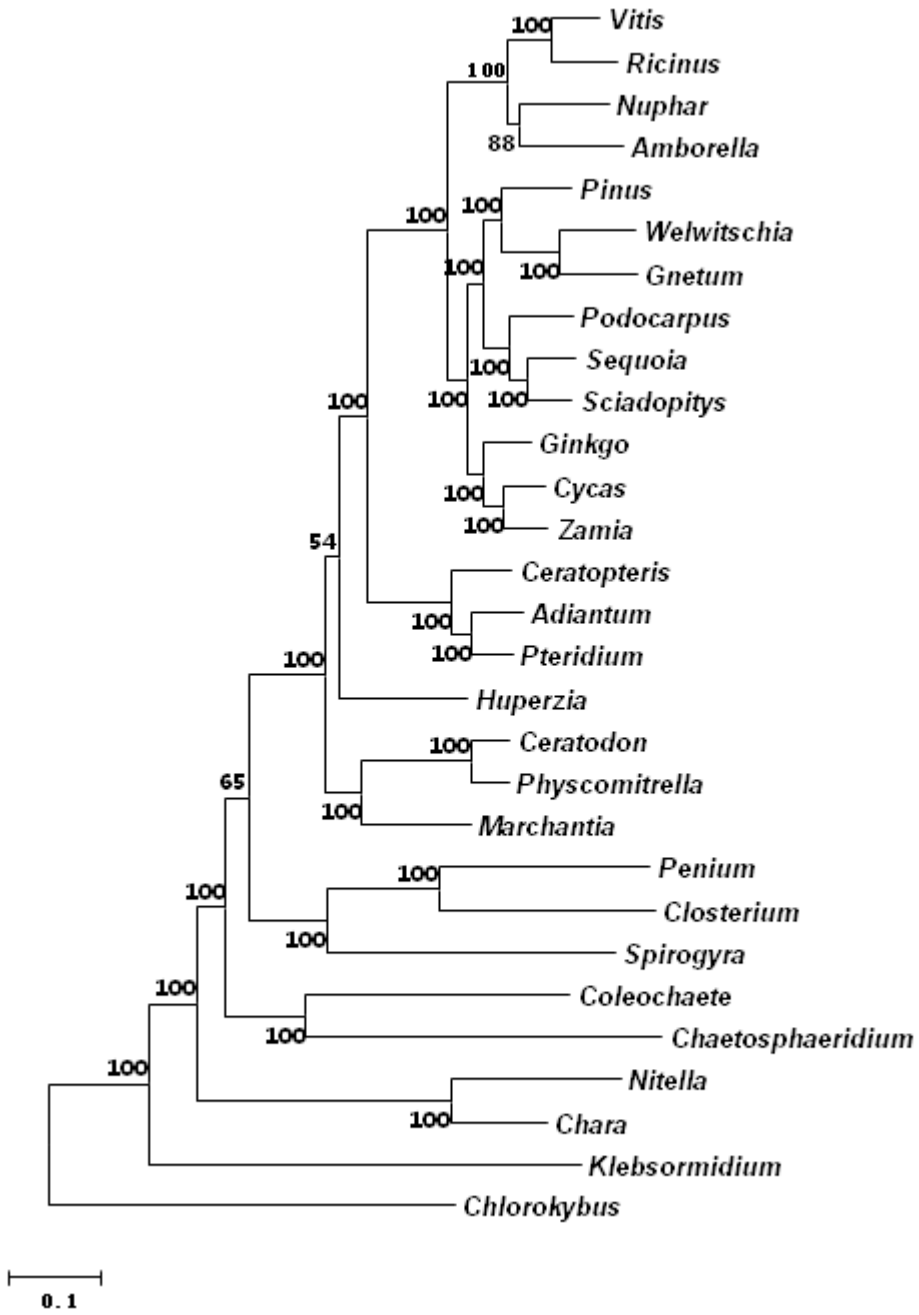


Figure S2. ML concatenation tree of the green lineage estimated from 289 genes with LG+GAMMA model. The numbers on the tree indicate bootstrap support values.

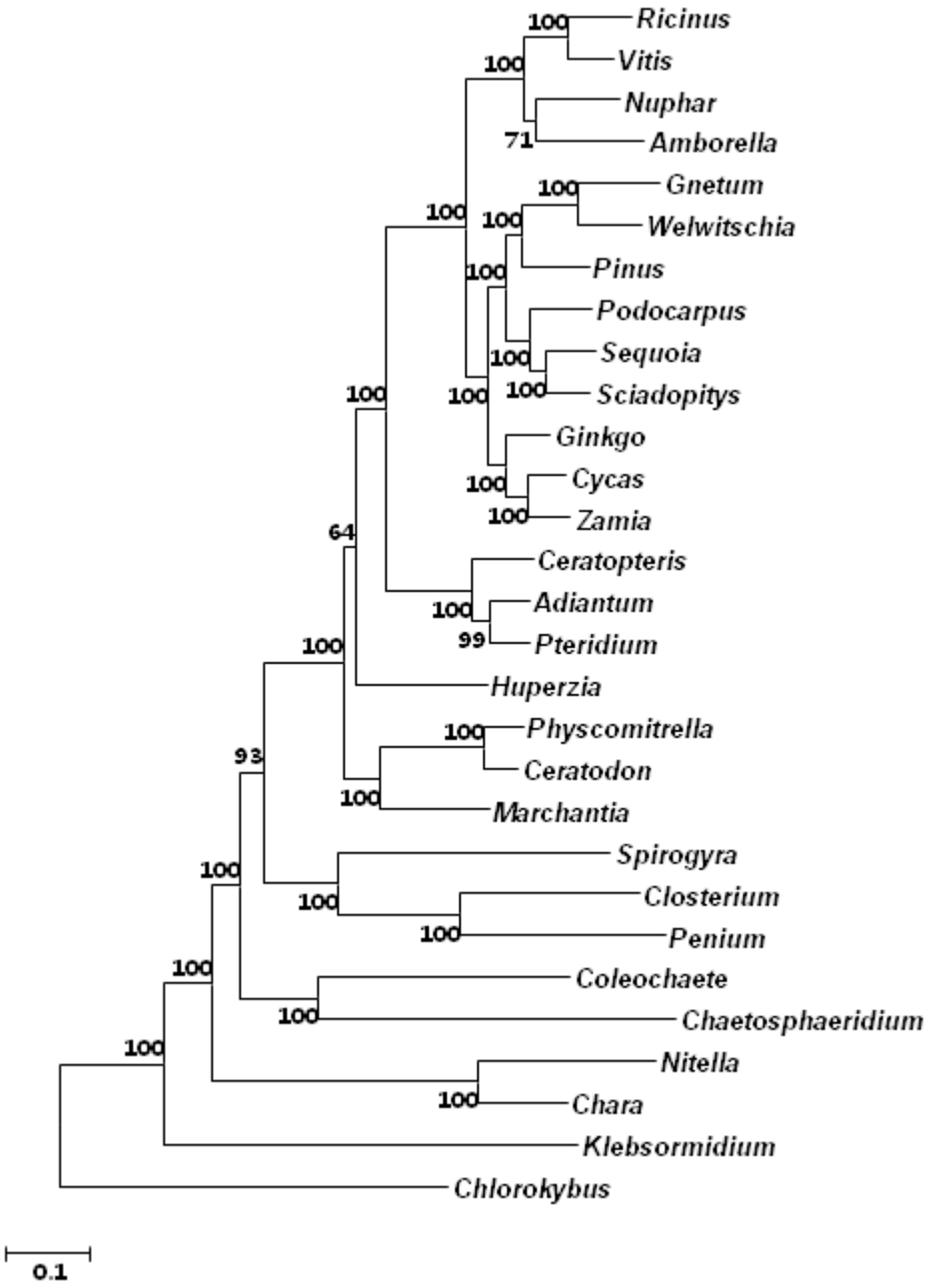


Figure S3. ML concatenation tree of the green lineage estimated from 184 genes with LG+GAMMA model. The numbers on the tree indicate bootstrap support values.

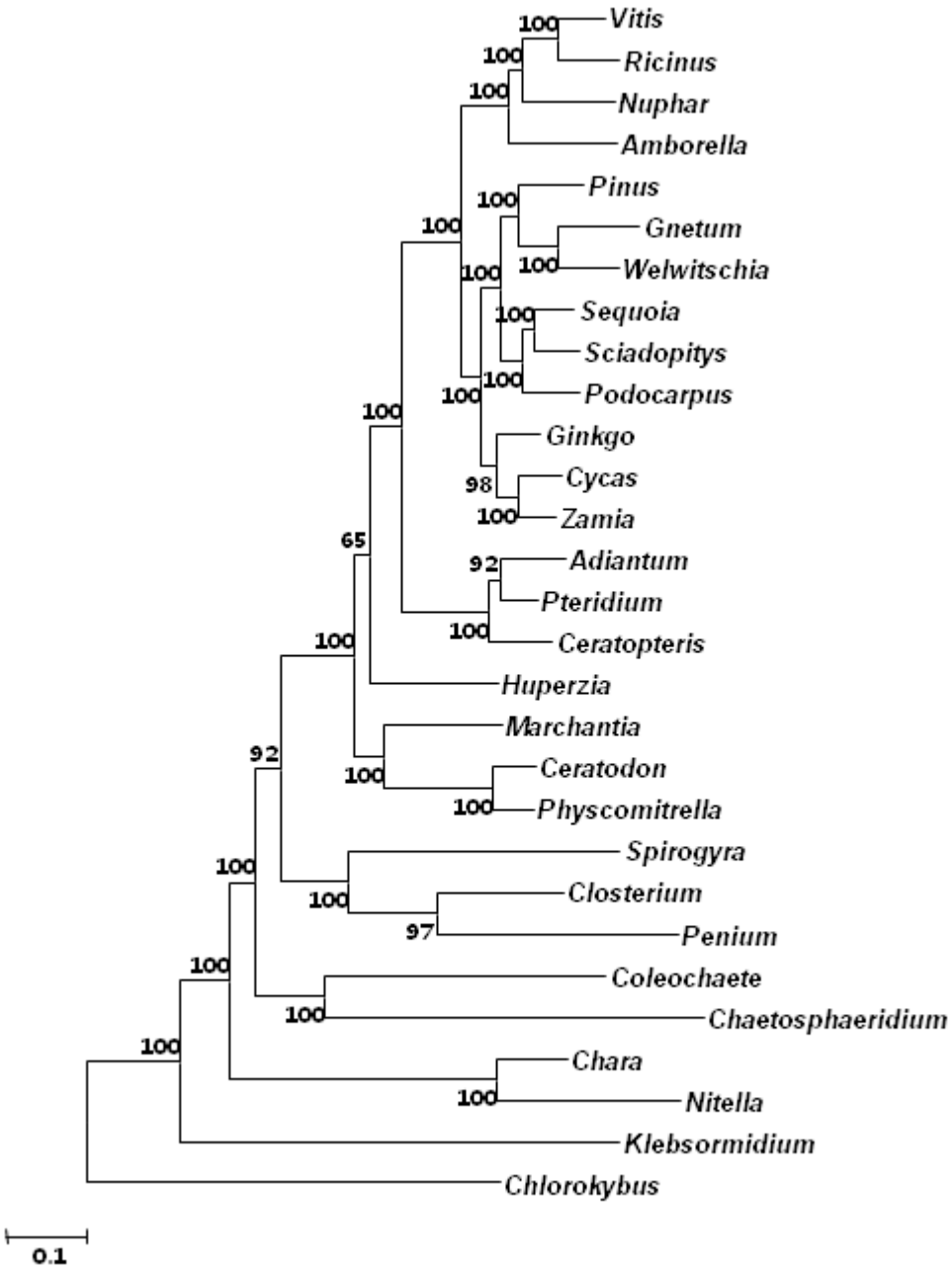


Figure S4. ML concatenation tree of the green lineage estimated from 211 non-ribosomal genes with LG+GAMMA model. The numbers on the tree indicate bootstrap support values.

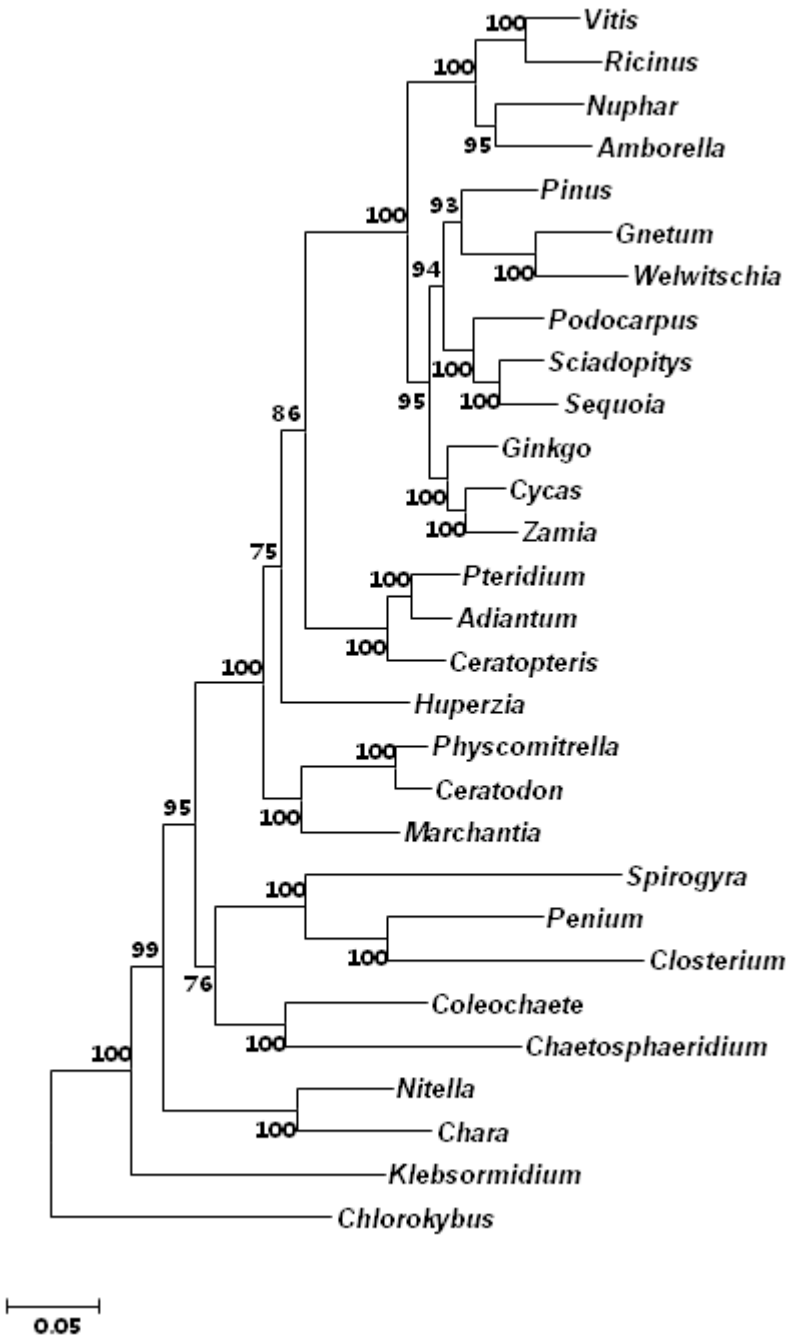


Figure S5. ML concatenation tree of the green lineage estimated from 78 ribosomal genes with LG+GAMMA model. The numbers on the tree indicate bootstrap support values.

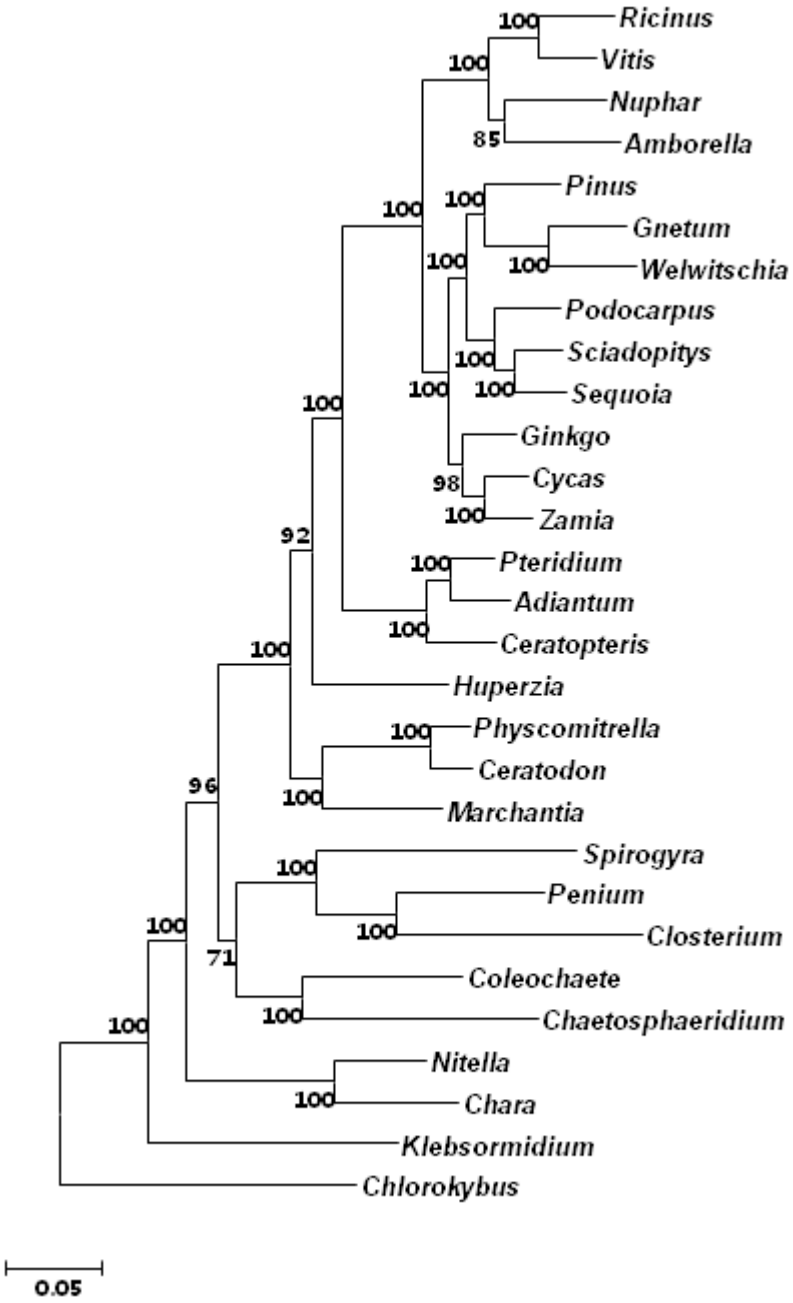


Figure S6. ML concatenation tree of the green lineage estimated from 119 genes with LG+GAMMA model. The numbers on the tree indicate bootstrap support values.

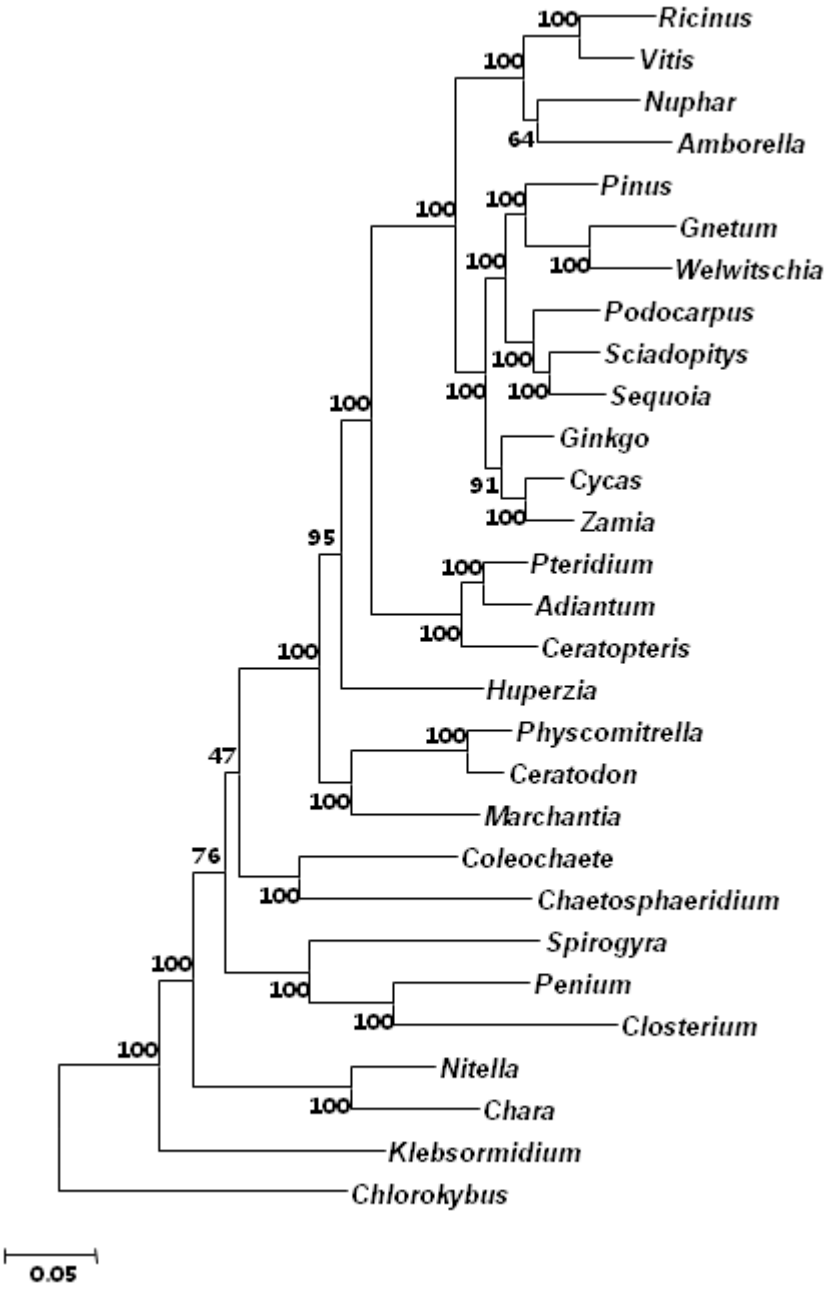


Figure S7. ML concatenation tree of the green lineage estimated from 42 genes (whose average bootstrap value of > 50% among 119 genes). The numbers on the tree indicate bootstrap support values.

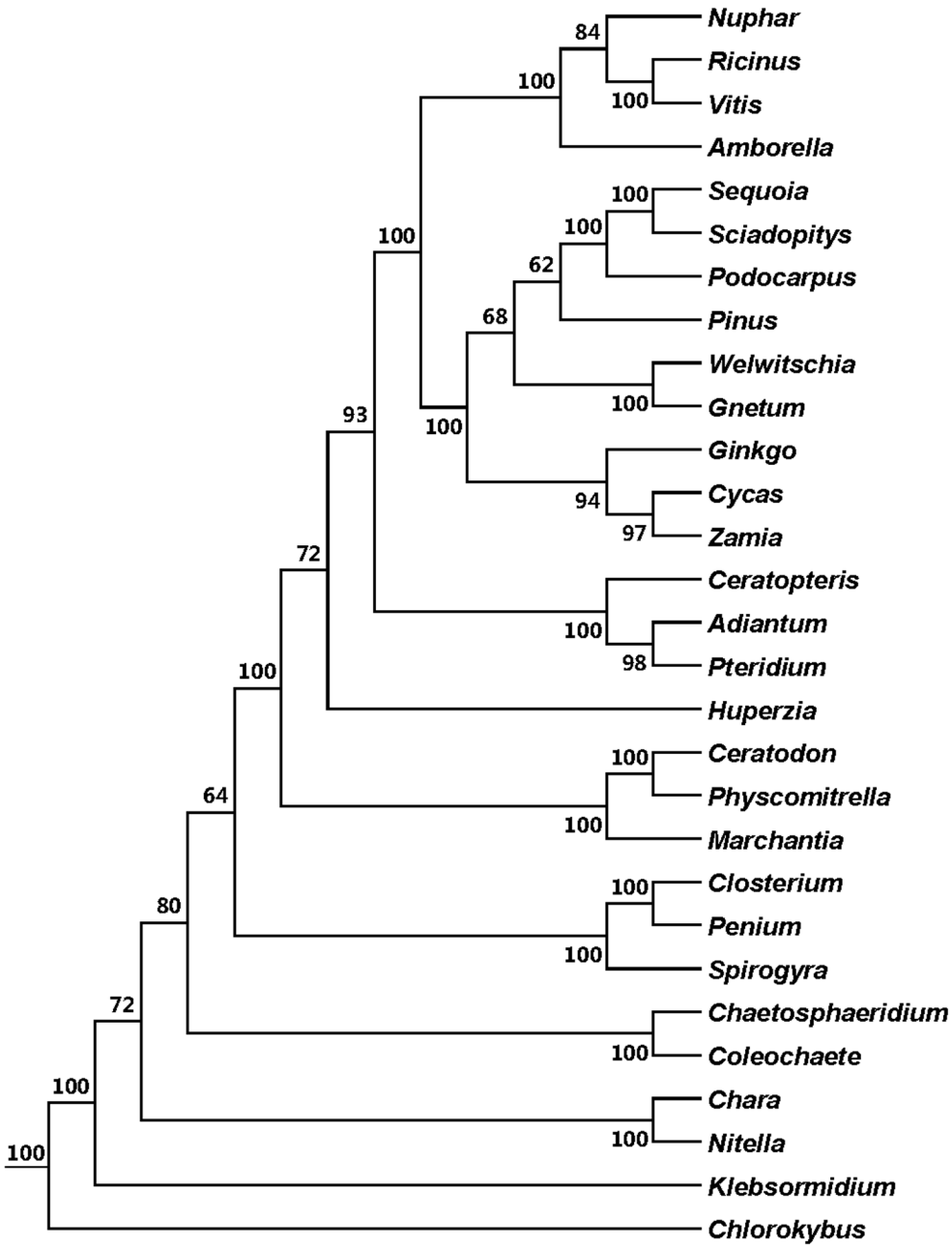


Figure S8. MP-EST tree for 176 genes. The MP-EST tree was constructed by removing the 8 outlier genes from the data of 184 genes. The numbers on the tree indicate bootstrap support values.

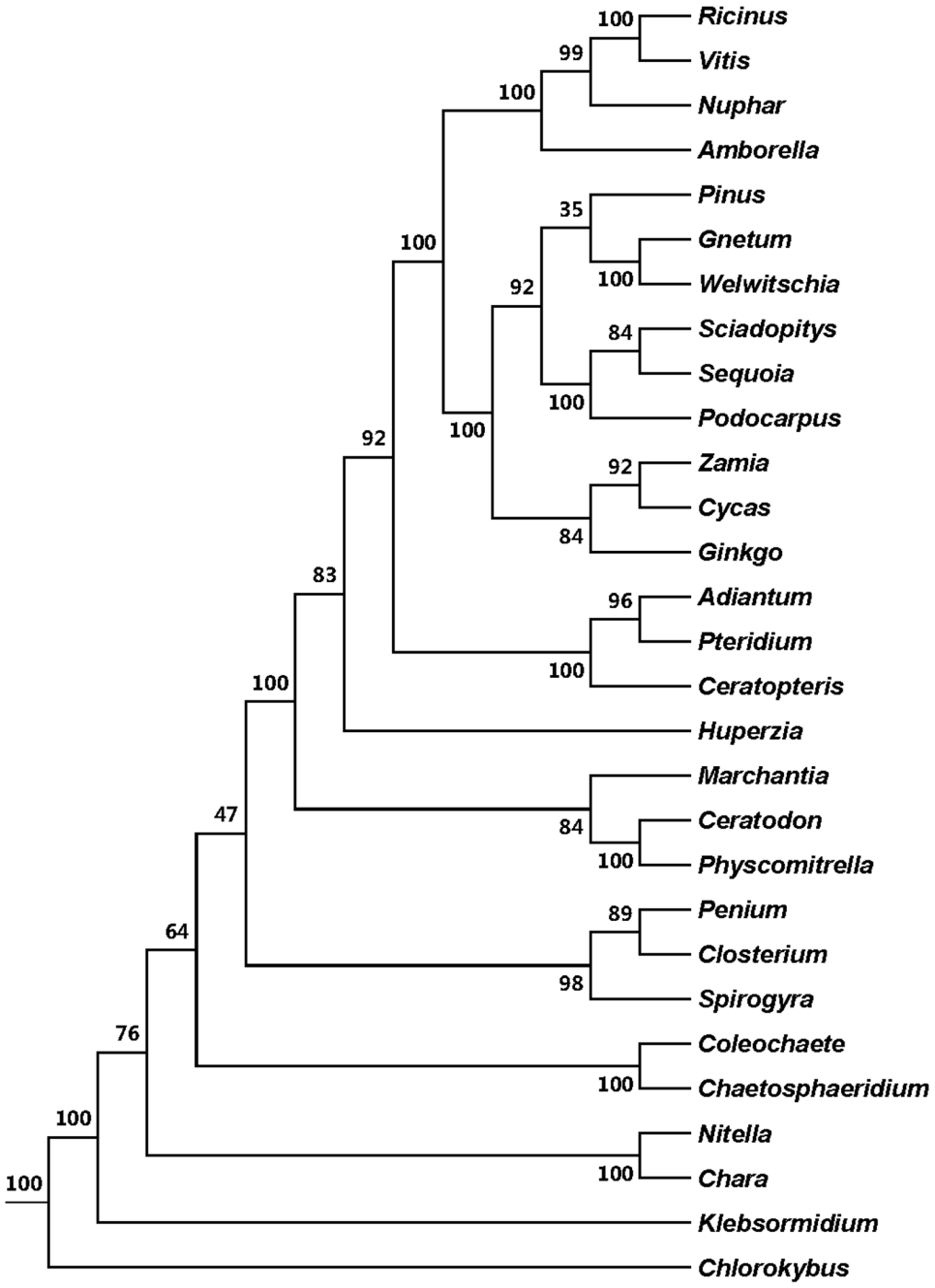


Figure S9. MP-EST tree for 114 genes. The numbers on the tree indicate bootstrap support values.

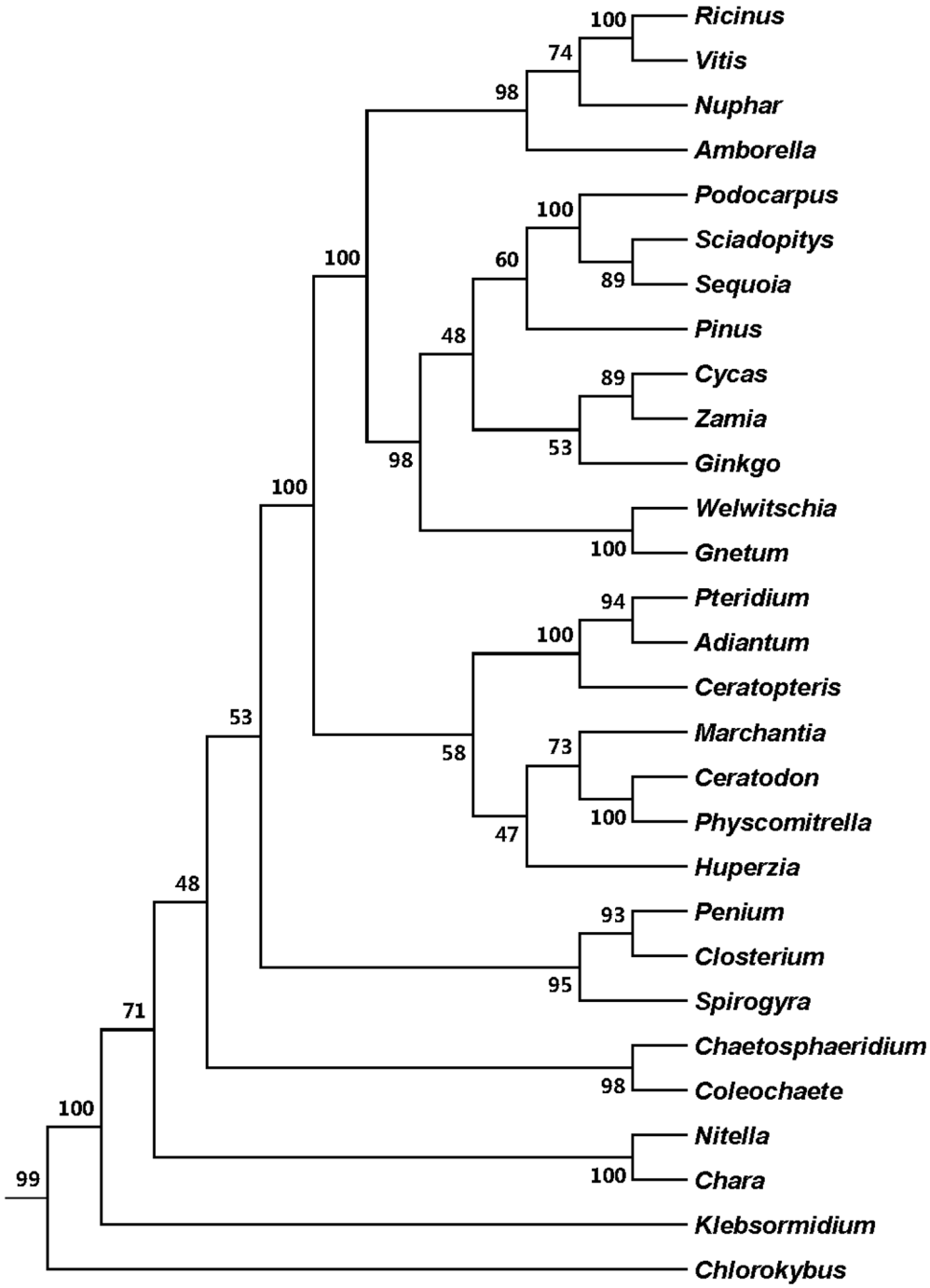


Figure S10. MP-EST tree for 47 genes. The numbers on the tree indicate bootstrap support values.

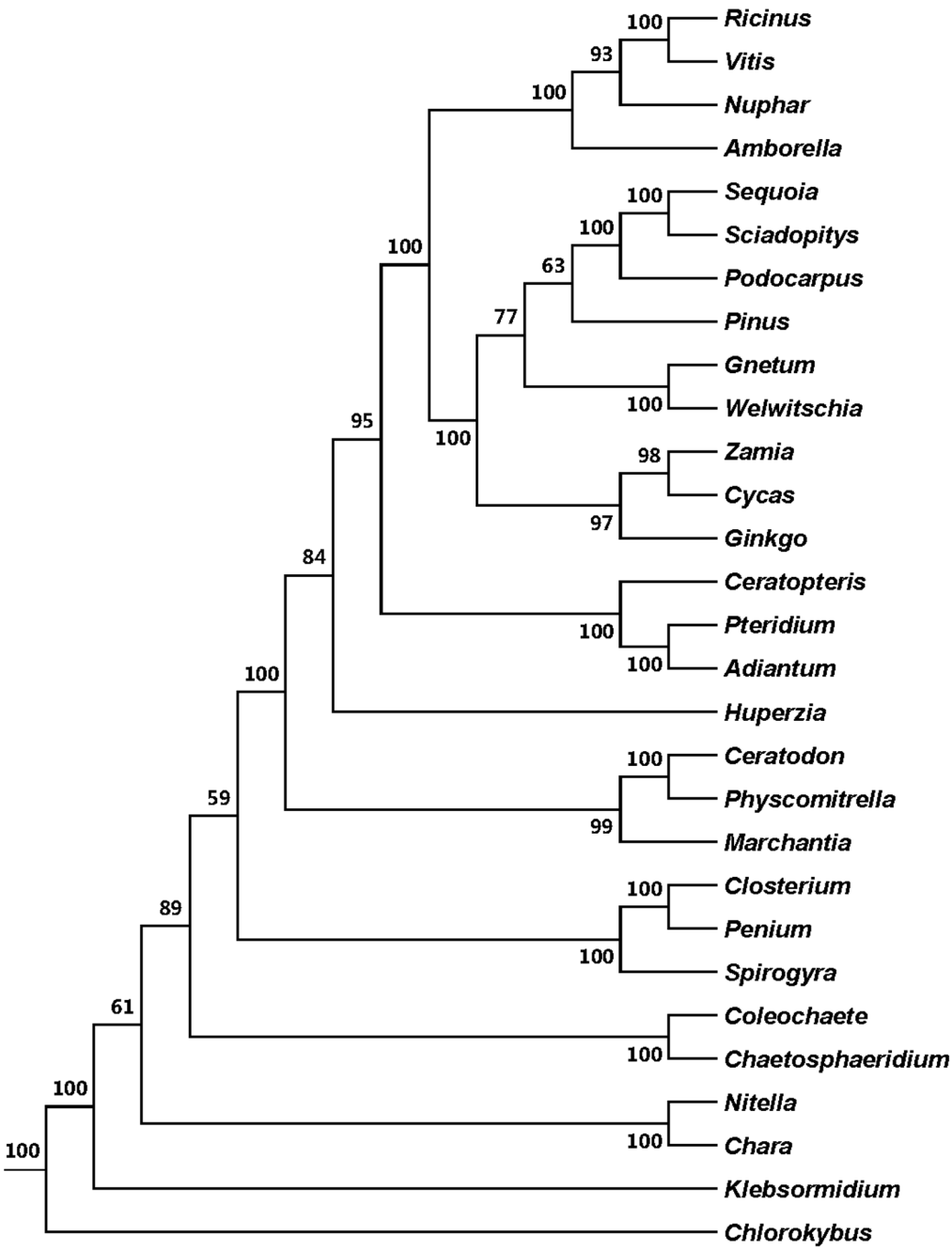


Figure S11. MP-EST tree for 160 genes. The MP-EST tree was constructed by removing the 24 genes with significant amino acid bias from the data of 184 genes The numbers on the tree indicate bootstrap support values.

Letter

(Discoveries)

Streptophyte algae and the origin of land plants revisited using heterogeneous models with three new algal chloroplast genomes

Bojian Zhong[1,*], Zhenxiang Xi[2], Vadim V. Goremykin[3], Richard Fong[1], Patricia A. Mclenachan[1], Philip M. Novis[4], Charles C. Davis[2], and David Penny[1]

[1] Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[2] Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Cambridge, Massachusetts, United States of America

[3] Istituto Agrario San Michele all'Adige Research Center, San Michele all'Adige, Italy

[4]Allan Herbarium, Landcare Research, Lincoln 7640, New Zealand

[*]Corresponding author: E-mail: bjzhong@gmail.com

Abstract

The phylogenetic branching order of the green algal groups that gave rise to land plants remains uncertain despite its fundamental importance to understanding plant evolution. Previous studies have demonstrated that land plants evolved from streptophyte algae, but different lineages of streptophytes have been suggested to be the sister group of land plants. To better understand the evolutionary history of land plants, and to determine the potential effects of "long-branch attraction" in phylogenetic reconstruction, we analysed a chloroplast genome dataset including three new chloroplast genomes from streptophyte algae: *Coleochaetae orbicularis* (Coleochaetales), *Nitella hookeri* (Charales), and *Spirogyra communis* (Zygnematales). We further applied a site pattern sorting method together with site- and time-heterogeneous models to investigate the branching order among streptophytes and land plants. Our chloroplast phylogenomic analyses are similar to nuclear data in placing Zygnematales alone, or a clade consisting of Coleochaetales plus Zygnematales, as the closest living relatives of land plants.

Key words: phylogenomics, chloroplast genomes, land plants, streptophyte algae, heterogeneous models

The relationship between green algae and land plants remains uncertain despite its importance to understanding plant evolution. Analyses of both morphological and molecular data have established land plants as a monophyletic group that evolved within streptophyte algae (also known as the charophycean algae). To better understand the colonization of the terrestrial habitat and the evolution of cellular complexity, it is critical to establish which streptophyte groups are most closely related to land plants.

An early study using four-gene markers from three genomic compartments indicated that Charales were sister to land plants (Karol et al. 2001). In contrast, recent phylogenomic analyses of both chloroplast and nuclear genome data indicated that (1) Coleochaetales alone (Turmel et al. 2009a; Turmel et al. 2009b), (2) Zygnematales alone (Turmel et al. 2006; Chang and Graham 2011; Wodniok et al. 2011; Timme et al. 2012; Zhong et al. 2013), or (3) Coleochaetales and Zygnematales combined (Finet et al. 2012; Laurin-Lemay et al. 2012) are sister to land plants. One likely explanation for this phylogenetic uncertainty is that Coleochaetales, Zygnematales, and land plants appear to have diverged rapidly during their early evolution (Stebbins and Hill 1980).

Based on their cytological characteristics, Charales (such as *Chara* and *Nitella*) are large and multicellular, but coenocytic, algae with thousands of nuclei per cell (Grant and Borowitzka 1984). In contrast, Coleochaetales (such as *Coleochaet*e and *Chaetosphaeridium*) and Zygnematales (such as *Zygnema* and *Spirogyra*) are multicellular organisms that are divided into much smaller cells, each with a single nucleus. In this cytological sense, Coleochaetales or Zygnematales may represent more appropriate sisters to land plants.

Chloroplast genomic data have been proven very useful for helping resolve plant phylogeny (e.g., Jansen et al. 2007; Moore et al. 2007; Zhong et al. 2010; Parks et al. 2012; Wu et al. 2013). In terms of sequenced chloroplast genomes of streptophyte algae, however, there is only one genome currently available for each of the Charales (*Chara vulgaris*) and Coleochaetales (*Chaetosphaeridium globosum*). The paucity of taxon sampling within the most deeply diverged regions of the green plant phylogeny is especially problematic and may lead to long-branch attraction (LBA) artefacts (Hendy and Penny 1989). To ameliorate the problem of LBA we sequenced three chloroplast genomes from streptophyte algae using next-generation sequencing technology: *Coleochaetae orbicularis* (Coleochaetales), *Nitella hookeri* (Charales), and *Spirogyra communis* (Zygnematales). We then analysed these data simultaneously in a larger chloroplast genome dataset, which includes 72 protein-coding genes (45,879 aligned nucleotide sites) common to 30 land plants and streptophyte algae.

It has been demonstrated that fast-evolving sites represent a challenge for phylogenetic inference because they are likely to experience multiple changes that tend to mask informative phylogenetic signals (Delsuc et al. 2005). Recent studies reported that the accuracy of chloroplast phylogenomic analyses could be improved by either removing the most rapidly evolving sites (which contain misleading phylogenetic information), or by using site-heterogeneous models (Zhong et al, 2010, 2011; Goremykin et al, 2013). As reported in Zhong et al. (2011) and Goremykin et al. (2013), the OV-sorting method (Goremykin et al. 2010) identifies not only the most rapidly evolving sites within a dataset, but also those sites that have a poor fit to model assumptions. We implemented this method to sort the 45,879 sites in our concatenated matrix from most variable to least variable. We then successively removed the most variable sites in increments of 500. The optimal break point for site removal was determined at site 36,879 (i.e., 9,000 sites were removed from the full matrix), where

we identified significant improvement in the two Pearson correlations (see Fig.1 and Materials and Methods).

For the fully concatenated (45,879 aligned sites) and reduced OV-sorted (36,879 aligned sites) matrices, we first used the site-homogeneous GTRGAMMA model with the *a posterior* partitioning strategy implemented by Xi et al. (2012) to infer our maximum likelihood (ML) phylogeny. Here, our ML analyses strongly support the monophyly of land plants (100 bootstrap percentage [BP]) and strongly place Zygnematales as sister to land plants (97 BP and 96 BP for the fully concatenated and reduced OV-sorted matrices, respectively; Fig.2a).

Site-homogeneous models assume that a single Markov process of substitution can be applied for all sites and at all times, yet many biological sequences cannot be adequately described using a single replacement matrix. In contrast, site-heterogeneous models introduce different categories by regrouping sites with similar profiles of stationary frequencies, and are thus more effective at minimizing LBA artefacts (Lartillot and Philippe 2008; Philippe et al. 2011; Kayal et al. 2013). Therefore, we applied two site-heterogeneous mixture models (Lartillot and Philippe 2004; Pagel and Meade 2004) to infer phylogenetic relationships in a Bayesian framework. Similar to our ML phylogeny using the homogeneous model, both site-heterogeneous models support the relationship (Coleochaetales, (Zygnematales, land plants)) using the fully concatenated matrix with 1.0 posterior probability (PP) (Fig. 3a and Table S1). Interestingly, using the OV-sorted matrix under the PhyloBayes analysis, a slightly different relationship ((Zygnematales, Coleochaetales), land plants) was also strongly supported (0.96 PP; Fig. 3b).

Most current phylogenetic methods (e.g., homogeneous and site-heterogeneous models) assume that base composition is stationary over time. Violation of this model assumption, however, may lead to inaccurate tree reconstruction. For example, compositional heterogeneity is a well-known problem in this respect (Lockhart et al. 1994; Foster 2004; Jermiin et al. 2004); and recent study has suggested that compositional shifts of plastid proteins, which could lead to such compositional heterogeneity, might allow streptophyte algae to better deal with environmental stresses on land (Jobson and Qiu 2011). To evaluate if stationarity of composition was violated, we performed posterior predictive tests (Lartillot et al. 2009) for the fully concatenated and reduced OV-sorted matrices. This statistical test indicated that the violation of compositional homogeneity occurs in both these matrices (Z-scores are 5.98 and 5.38,

respectively; see Table 1). Thus, compositional heterogeneity could potentially influence our phylogenetic inference on the origin of land plants. To examine the effect of compositional heterogeneity, we implemented two nonhomogeneous nonstationary (time-heterogeneous) models of DNA sequence evolution in our ML and Bayesian analyses (Galtier and Gouy 1998; Blanquart and Lartillot 2008). When taking into account the compositional heterogeneity using time-heterogeneous models, our results for both the full and OV-sorted matrices consistently supported the relationship (Coleochaetales, (Zygnematales, land plants)) (Table S1 and Figs. 3c and 3d).

To further evaluate the impact of rapidly evolving sites for estimating branching order among streptophytes and land plants, we produced a series of shortened alignments by sequentially removing fast-evolving sites in 500 increments using the OV-sorting method (i.e., the number of total sites ranges from 38,379 to 32,879; see Table 1 and Table S1). It is striking that the alternative relationship ((Coleochaetales, Zygnematales), land plants) is recovered for 34,879 and 32,879 matrices in all analyses (Table1 and S1; Figs. S1 and S2). In addition, the relationship (Coleochaetales, (Zygnematales, land plants)) was rejected at p=0.05 for five matrices (i.e., 36,379, 34,879, 34,379, 33,879, and 33,379 aligned sites) using the approximately unbiased (AU) test (Shimodaira, 2002) (Table S1). It is noteworthy that none of our analyses here recovered Charales or Coleochaetales alone as the sister group to land plants. Moreover, these two alternative hypotheses were rejected at p=0.05 using the AU test for the 45,879 and 36,879 matrices.

By removing the most rapidly evolving sites and using site- and time-heterogeneous models that reduce systematic errors (e.g., LBA and compositional heterogeneity), our plastid genomic data indicates that Charales or Coleochaetales alone are not the sister group to land plants. Instead, Zygnematales, or a clade containing Coleochaetales plus Zygnematales, appear to be the closest living relatives of land plants. This result is also in agreement with previous nuclear data analyses (Finet et al. 2012; Laurin-Lemay et al. 2012; Timme et al. 2012; Wodniok et al. 2011; Zhong et al. 2013). It is likely important in this context that it was not the reduced-celled "coenocytic" lineage of the Charales that gave rise to land plants. Nevertheless, it is also important to understand the reasons for some green algae (e.g. Charalaes, Zygnematales and Dasycladales) becoming larger during evolution, and it may be a key for unlocking the origin of land plants.

Materials and Methods

DNA sequencing and data assembly

Nitella hookeri was collected from Wanganui, New Zealand. *Spirogyra communis* was cultured on BG11 medium (Rippka et al. 1979) from material collected from the Styx River at the Spencerville Road Bridge, Christchurch, New Zealand. *Coleochaete orbicularis* samples were ordered from the Culture Collection of Algae at The University of Texas at Austin (http://web.biosci.utexas.edu/utex) and grown on Modified Bold 3N media. Total genomic DNA (~50 ng) from all three algae were extracted using the Qiagen Plant DNeasy kit according to the manufacturer's protocols, and then sequenced using Illumina MiSeq platform with 100-bp paired-end reads. The short reads were filtered with the error probability < 0.05, and were then assembled using Velvet (Zerbino and Birney 2008). The contigs were further assembled and compared to complete chloroplast genomes available using Geneious software version 5.6 (www.geneious.com). Protein-coding genes were annotated using DOGMA (Wyman et al. 2004) with manual correction. Each protein-coding gene from 30 taxa was aligned using MUSCLE (Edgar 2004), and trimmed to exclude poorly aligned positions using Gblocks (Castresana 2000) with default settings. These alignments were concatenated to generate a matrix of 45,879 sites.

Removal of most rapidly evolving sites

The OV-sorting method (Goremykin et al. 2010) was used to rank the full concatenated alignment from the most to least variable sites based on the measurement of "observed variability" (OV) of each alignment position. The most variable sites were then successively removed from the original matrix, in increments of 500. For each step, two data partitions were obtained: 1) "A" partition that consists of all positions except the most variable 500, 1000, …, 9000 sites and 2) "B" partition that contains the most variable 500, 1000, …, 9000 sites. After model fitting was applied to each partition using ModelTest (Posada and Crandall 1998), the ML distances for the "A" and "B" partitions were calculated, as well as the uncorrected *p*-distances for each "B" partition using PAUP* (Swofford 2002). Two Pearson correlation analyses of pairwise distances were conducted at each step: 1) correlation of the ML distances for "A" and "B" partitions and 2) correlation of the ML and uncorrected *p*-distances for "B" partitions. The stopping point for site removal was determined as the point at which the two correlations showed marked improvement (Goremykin et al. 2010) (see Fig.1).

Phylogenetic analyses

ML analyses were performed using RAxML (Stamatakis 2006) with the site-homogeneous GTRGAMMA model and the *a posteriori* data partitioning strategy, which has recently been shown to produce better ML trees than the commonly used *a priori* approaches (e.g., partitioning by gene or by codon position; Xi et al. 2012).

Two site-heterogeneous Bayesian analyses were implemented using 1) PhyloBayes (Lartillot et al. 2009) under the CAT-GTR model (Lartillot and Philippe 2004) that accounts for across-site heterogeneities, and 2) BayesPhylogenies (Pagel and Meade 2004) under a "reversible-jump" mixture model (Pagel and Meade 2008) that fits more than one model of sequence evolution to the data. Two independent MCMC analyses were run for 5,000 cycles in PhyloBayes and 10 million generations in BayesPhylogenies. Convergence was checked based on time-series plots of the likelihood scores using Tracer (http://tree.bio.ed.ac.uk/software/tracer/).

The posterior predictive test was used to measure compositional heterogeneity in PhyloBayes. Two nonhomogeneous nonstationary models that account for compositional heterogeneity were applied in both ML and Bayesian analyses, i.e., 1) the nonstationary and nonhomogeneous model of DNA sequence evolution (Galtier and Gouy 1998) as implemented in nhPhyML (Boussau and Gouy 2006) that specifies different GC contents for each lineage in a likelihood framework, and 2) the CAT-BP model (Blanquart and Lartillot 2008) in nhPhyloBayes that considers compositional heterogeneity between lineages by introducing breakpoints along the branches.

The AU test (Shimodaira 2002) was conducted in scaleboot (Shimodaira 2008), with the site log-likelihood scores estimated in RAxML using the *a posteriori* partitioning strategy.

Table 1. Phylogenetic analyses using Bayesian (PhyloBayes) and maximum likelihood (RAxML and nhPhyML) estimations

Data sets	Phylobayes	PP	Z-score	*P* Value	RAxML	BP	nhPhyML	BP
45,879 (full data)	(CO, (Z, L))	1.00	5.98	0.000	(CO, (Z, L))	99	(CO, (Z, L))	100
38,379	(CO, (Z, L))	0.82	5.33	0.003	(CO, (Z, L))	98	(CO, (Z, L))	96
37,879	((Z, CO), L)	0.78	5.70	0.000	(CO, (Z, L))	73	(CO, (Z, L))	93
37,379	((Z, CO), L)	0.91	4.20	0.007	(CO, (Z, L))	94	(CO, (Z, L))	84
36,879 (OV-sorted data)	((Z, CO), L)	0.96	5.38	0.000	(CO, (Z, L))	96	(CO, (Z, L))	81
36,379	((Z, CO), L)	0.90	5.28	0.007	((Z, CO), L)	100	(CO, (Z, L))	84
35,879	((Z, CO), L)	0.98	5.15	0.003	(CO, (Z, L))	94	(CO, (Z, L))	74
35,379	((Z, CO), L)	0.99	4.45	0.013	((Z, CO), L)	52	(CO, (Z, L))	60
34,879	((Z, CO), L)	1.00	5.85	0.000	((Z, CO), L)	99	((Z, CO), L)	54
34,379	((Z, CO), L)	0.99	5.30	0.000	((Z, CO), L)	100	(CO, (Z, L))	62
33,879	((Z, CO), L)	1.00	6.81	0.000	((Z, CO), L)	100	(CO, (Z, L))	53
33,379	((Z, CO), L)	1.00	6.39	0.000	((Z, CO), L)	100	(CO, (Z, L))	42
32,879	((Z, CO), L)	0.96	5.76	0.000	((Z, CO), L)	94	((Z, CO), L)	55

Abbreviations: CO = Coleochaetales, L = Land Plants, Z = Zygnematales, PP = Bayesian Posterior Probability, BP = Maximum Likelihood Bootstrap Percentage. The PP and BP values supporting Zygnematales closest to land plants are shown for (CO, (Z, L)) phylogeny, and both values supporting monophyletic relationship of Coleochaetales and Zygnematales are shown for ((Z, CO), L) phylogeny.

Fig.1. Pearson correlation results. The blue line indicates Pearson correlation coefficients (*r*) of maximum likelihood (ML) distances calculated from partitions "A" (more conserved) and "B" (less conserved). The red line indicates *r* values of uncorrected *p*-distances and ML distances for "B" partitions. The *r* values begin to increase dramatically at 36,879 sites remaining.

Fig. 2. Maximum likelihood trees using the homogeneous model (GTRGAMMA) with *a posteriori* partitioning strategy based on the full (45,879 aligned sites) and reduced OV-sorted (36,879 aligned sites) matrices. Numbers on the tree indicate bootstrap percentage (BP) and nodes with 100 BP are not marked.

Fig. 3. Phylogenetic trees using the site-heterogeneous model (i.e., the CAT model in PhyloBayes) and time-heterogeneous model (nhPhyML) based on the full (45,879 aligned sites) and OV-sorted (36,879 aligned sites) matrices. Numbers on the tree indicate the Bayesian posterior probability (PP) from PhyloBayes and the maximum likelihood bootstrap percentage (BP) from nhPhyML, and nodes with 100 BP or 1.0 PP are not marked.

References

Blanquart S, Lartillot N. 2008. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol*. 25:842-858.

Boussau B, Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol*. 55:756-768.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540-552.

Chang Y, Graham SW. 2011. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. *Am J Bot*. 98:839-849.

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

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792-1797.

Finet C, Timme RE, Delwiche CF, Marlétaz F. 2012. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr. Biol*. 22:1456-1457.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol*. 53:485-495.

Galtier N, Gouy M. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol*. 15:871-879.

Goremykin VV, Nikoforova SV, Bininda-Emonds OPP. 2010. Automated removal of noisy data in phylogenomic analyses. *J Mol Evol*. 71:319-331.

Goremykin VV, Nikiforova SV, Biggs PJ. Zhong B, De Lange P, Martin W, Woetzel S, Atherton RA, McLenachan PA, Lockhart PJ. 2013. The evolutionary root of flowering plants. *Syst Biol*. 62: 51-62.

Grant BR, Borowitzka MA. 1984. The chloroplasts of giant-celled and coenocytic algae: biochemistry and structure. *Bot Rev*. 50:267-307.

Hendy M, Penny D. 1989. A framework for the quantitative study of evolutionary trees. *Syst Zool*. 38:297-309.

Jansen RK, Cai Z, Raubeson LA, et al. (16 co-authors). 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci USA*. 104:19369-19374.

Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol*. 53:638-643.

Jobson RW, Qiu Y. 2011. Amino acid compositional shifts during streptophyte transitions to terrestrial habitats. *J Mol Evol*. 72: 204-214.

Karol KG, McCourt RM, Cimino MT, Delwiche CF. 2001. The closest living relatives of land plants. *Science*. 294:2351-2353.

Kayal E, Roure B, Philippe H, Collins AG, Lavrov DV. 2013. Cnidarian phylogenetic relationships as revealed by mitogenomics. *BMC Evol Biol*. 13:5.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across- site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 21:1095-1109.

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc Lond B Biol Sci*. 363:1463-1472.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*. 25:2286-2288.

Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Curr Biol*. 22:R593-R594.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*. 11:605-612.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci USA*. 104:19363-19268.

Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol*. 53:571-581.

Pagel M, Meade, A. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos Trans R Soc Lond B Biol Sci*. 363: 3955-3964.

Parks M, Cronn RC, Liston A. 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from Pinus L. (Pinaceae). *BMC Evol Biol*. 12:100.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9:e1000602.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics*. 14:817-818.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol*. 111:1–61.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*. 51:492-508.

Shimodaira H. 2008. Testing regions with nonsmooth boundaries via multiscale bootstrap. *J. Stat. Plann. Inference*. 138:1227-1241.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688-2690.

Stebbins GL, Hill GJC. 1980. Did multicellular plants invade the land? *Am. Nat.* 115: 342-353.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Timme RE, Bachvaroff TR, Delwiche CF. 2012. Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One*. 7: e29696.

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. *Mol Biol Evol*. 23:1324-1338.

Turmel M, Gagnon MC, O'Kelly CJ, Otis C, Lemieux C. 2009a. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol Biol Evol*. 26:631-48.

Turmel M, Otis C, Lemieux C. 2009b. The chloroplast genomes of the green algae *Pedinomonas minor*, *Parachlorella kessleri*, and *Oocystis solitaria* reveal a shared ancestry between the Pedinomonadales and Chlorellales. *Mol Biol Evol*. 26:2317-2331.

Wodniok S, Brinkmann H, Glöckner G, Heidel AJ, Philippe H, Melkonian M, Becker B. 2011. Origin of land plants: Do conjugating green algae hold the key? *BMC Evol Biol*. 11: 104.

Wu CS, Chaw SM, Huang YY. 2013. Chloroplast phylogenomics indicates that *Ginkgo biloba* is sister to cycads. *Genome Biol Evol*. 5:243-254.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*. 20:3252-3255.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, Davis CC. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc Natl Acad Sci USA*. 109:17519-17524.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18:821-829.

Zhong B, Yonezawa T, Zhong Y, Hasegawa M. 2010. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol Biol Evol*. 27:2855-2863.

Zhong B, Deusch O, Goremykin VV, Penny D, Biggs PJ, Atherton RA, Nikiforova SV, Lockhart PJ. 2011. Systematic error in seed plant phylogenomics. *Genome Biol Evol*. 3:1340-1348.

Zhong B, Liu L, Yan Z, Penny D. 2013. Origin of land plants using the multispecies coalescent model. *Trends Plant Sci*. 10.1016/j.tplants.2013.04.009

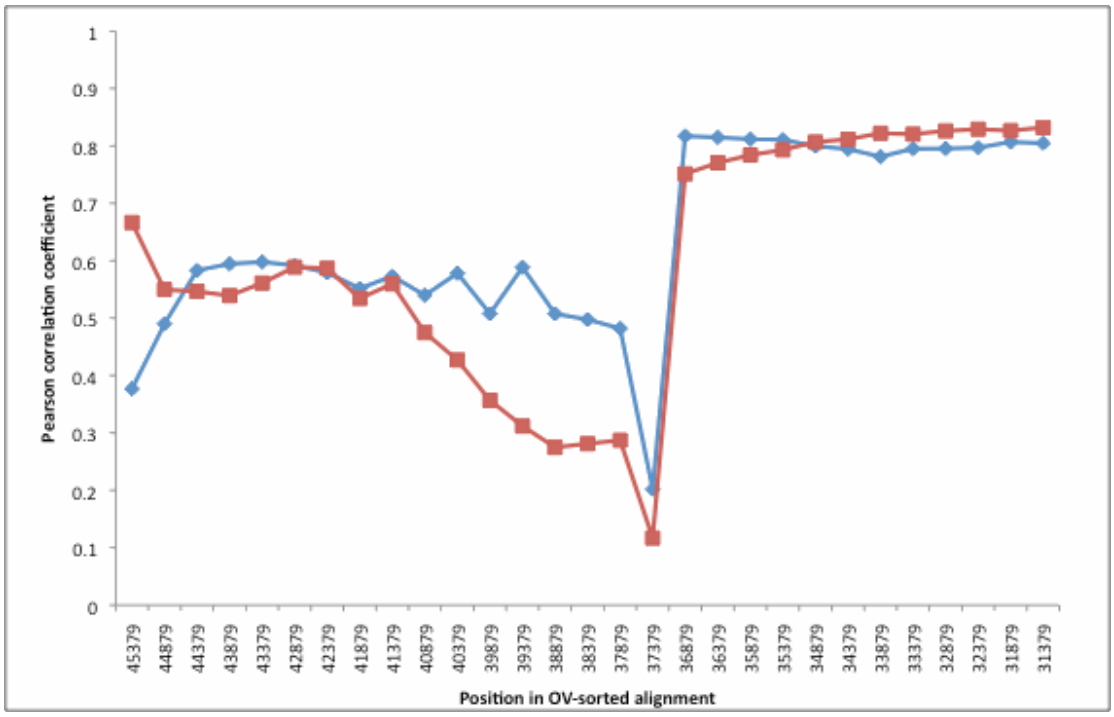


Fig.1. Pearson correlation results. The blue line indicates Pearson correlation coefficients (*r*) of maximum likelihood (ML) distances calculated from partitions "A" (more conserved) and "B" (less conserved). The red line indicates *r* values of uncorrected *p*-distances and ML distances for "B" partitions. The *r* values begin to increase dramatically at 36,879 sites remaining.

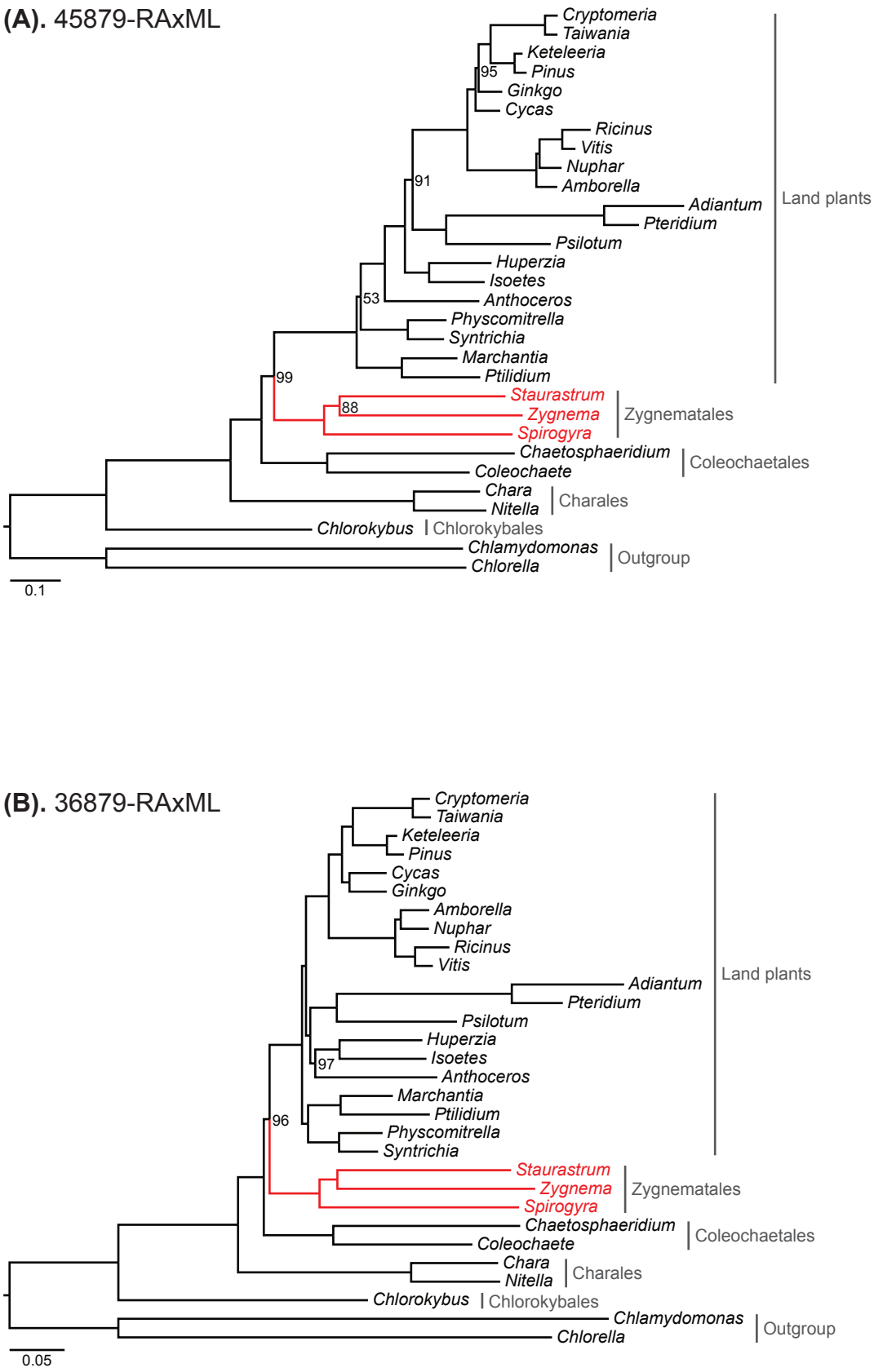


Fig. 2. Maximum likelihood trees using the homogeneous model (GTRGAMMA) with *a posteriori* partitioning strategy based on the full (45,879 aligned sites) and reduced OV-sorted (36,879 aligned sites) matrices. Numbers on the tree indicate bootstrap percentage (BP) and nodes with 100 BP are not marked.

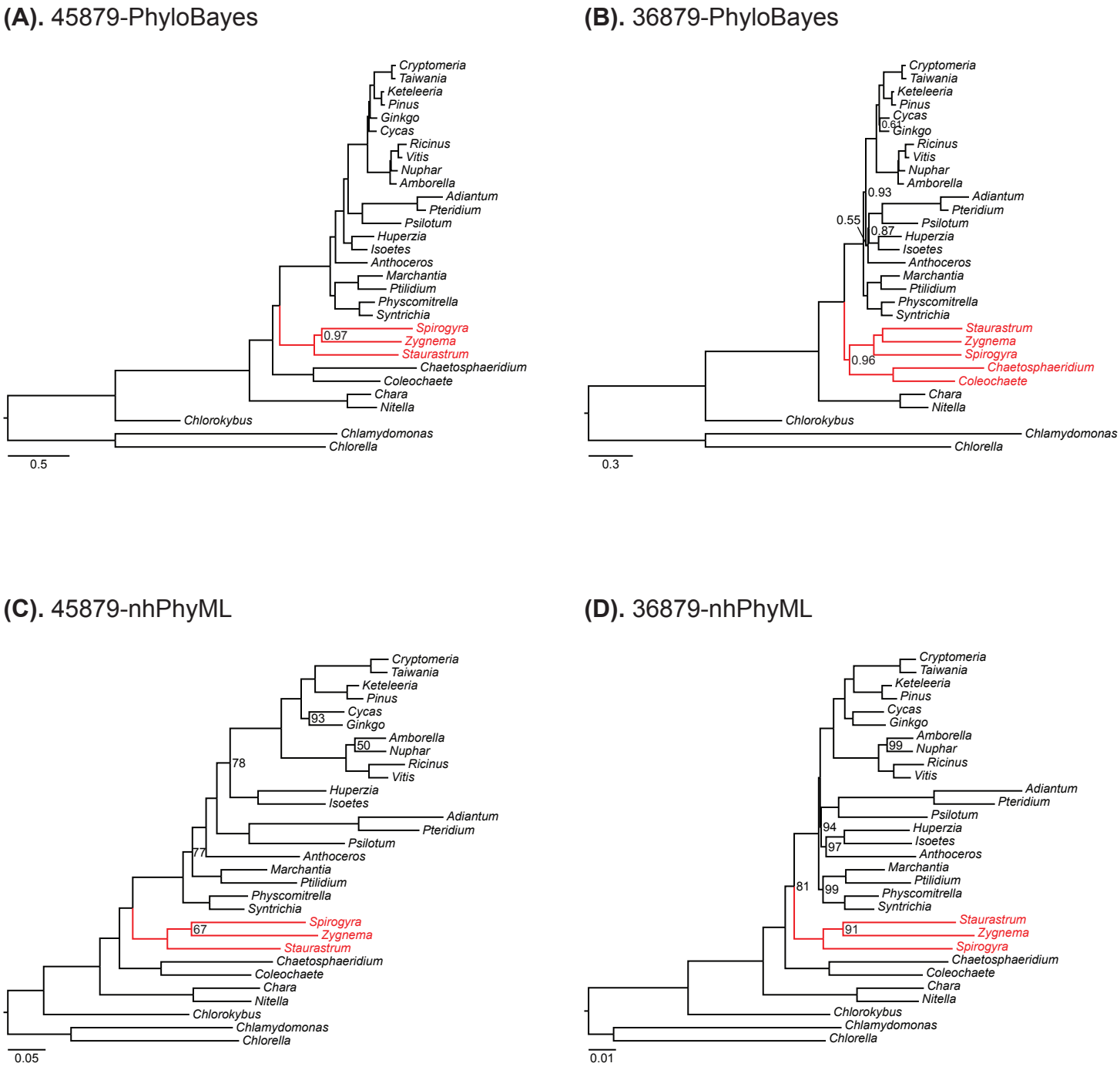


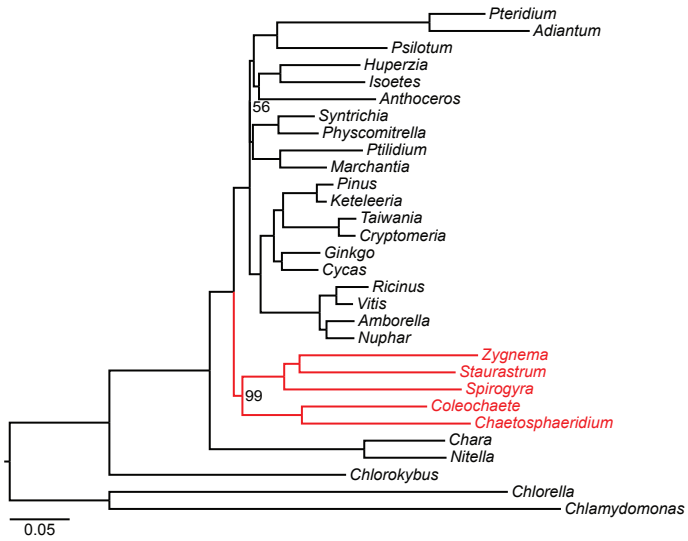
Fig. 3. Phylogenetic trees using the site-heterogeneous model (i.e., the CAT model in PhyloBayes) and time-heterogeneous model (nhPhyML) based on the full (45,879 aligned sites) and OV-sorted (36,879 aligned sites) matrices. Numbers on the tree indicate the Bayesian posterior probability (PP) from PhyloBayes and the maximum likelihood bootstrap percentage (BP) from nhPhyML, and nodes with 100 BP or 1.0 PP are not marked.

Table S1. Phylogenetic analyses using Bayesian heterogeneous models (BayesPhylogenies and nh-PhyloBayes).

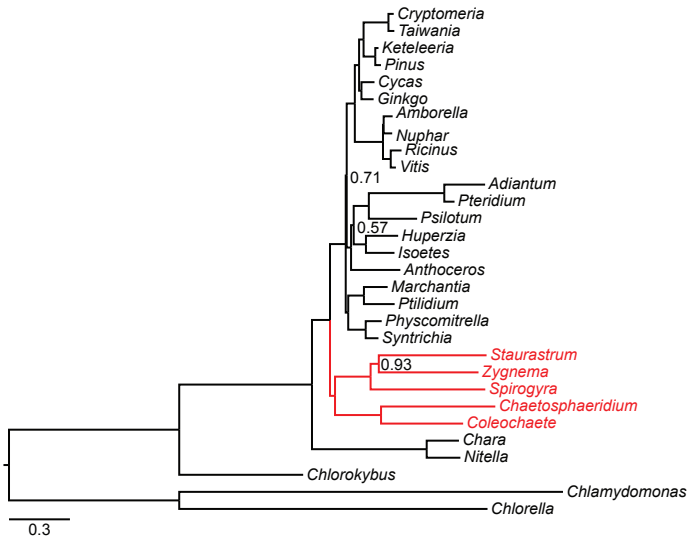
Data sets	BayesPhylogenies	PP	nh-PhyloBayes	PP	AU Test	*p*-value
45,879 (full data)	(CO, (Z, L))	1.00	(CO, (Z, L))	1.00	(CO, (Z, L))	0.06%
38,379	(CO, (Z, L))	1.00	-	-	(CO, (Z, L))	0.54%
37,879	(CO, (Z, L))	1.00	-	-	(CO, (Z, L))	NS
37,379	(CO, (Z, L))	1.00	-	-	(CO, (Z, L))	0.86%
36,879 (OV-sorted data)	(CO, (Z, L))	0.99	(CO, (Z, L))	1.00	(CO, (Z, L))	4.42%
36,379	((Z, CO), L)	0.86	-	-	((Z, CO), L)	0.64%
35,879	(CO, (Z, L))	0.96	-	-	(CO, (Z, L))	4.53%
35,379	(CO, (Z, L))	0.88	-	-	((Z, CO), L)	NS
34,879	((Z, CO), L)	1.00	((Z, CO), L)	1.00	((Z, CO), L)	0.20%
34,379	((Z, CO), L)	0.95	-	-	((Z, CO), L)	0.72%
33,879	((Z, CO), L)	0.99	-	-	((Z, CO), L)	0.49%
33,379	((Z, CO), L)	1.00	((Z, CO), L)	1.00	((Z, CO), L)	0.05%
32,879	((Z, CO), L)	0.83	((Z, CO), L)	1.00	((Z, CO), L)	NS

Abbreviations: CO = Coleochaetales, L = Land Plants, Z = Zygnematales, PP = Bayesian Posterior Probability, NS = Not Significant. Only five datasets are applied for nh-PhyloBayes analyses due to expensive computation. The PP values supporting Zygnematales closest to land plants are shown for (CO, (Z, L)) phylogeny, and PP values supporting monophyletic relationship of Coleochaetales and Zygnematales are shown for ((Z, CO), L) phylogeny.

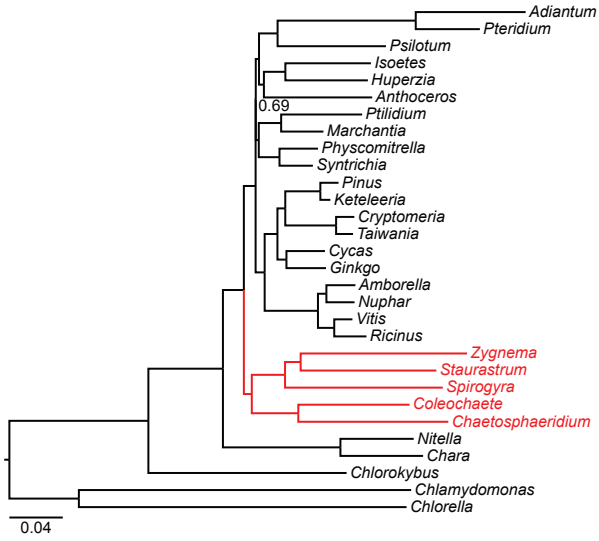
(A). 34879-RAxML



(B). 34879-PhyloBayes



(C). 34879-BayesPhylogenies



(D). 34879-nhPhyML

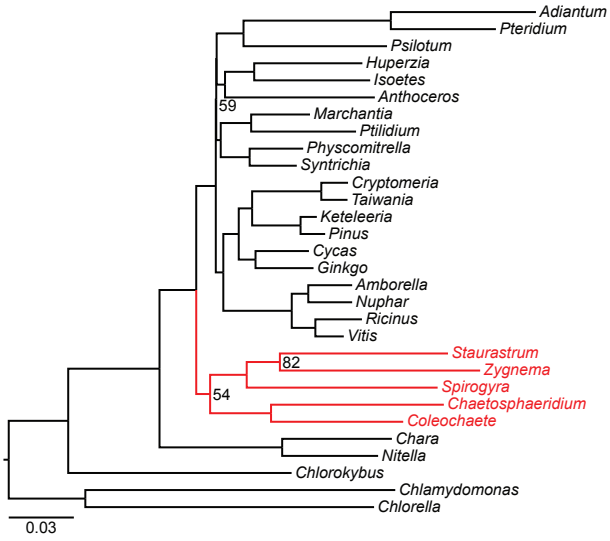
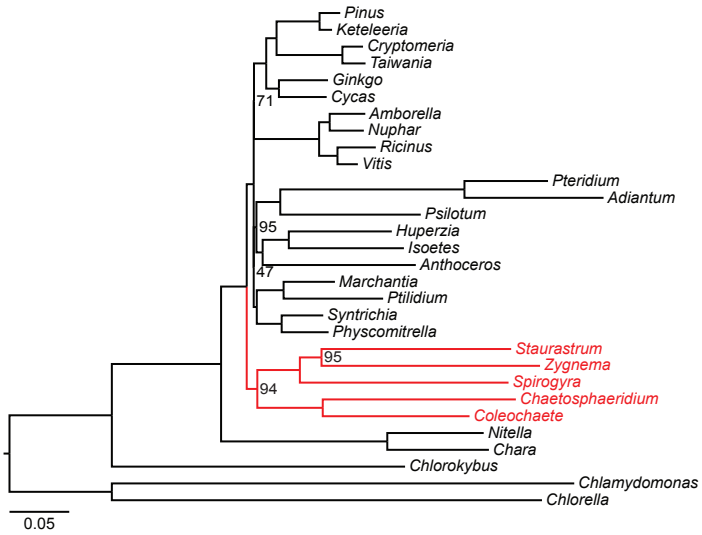
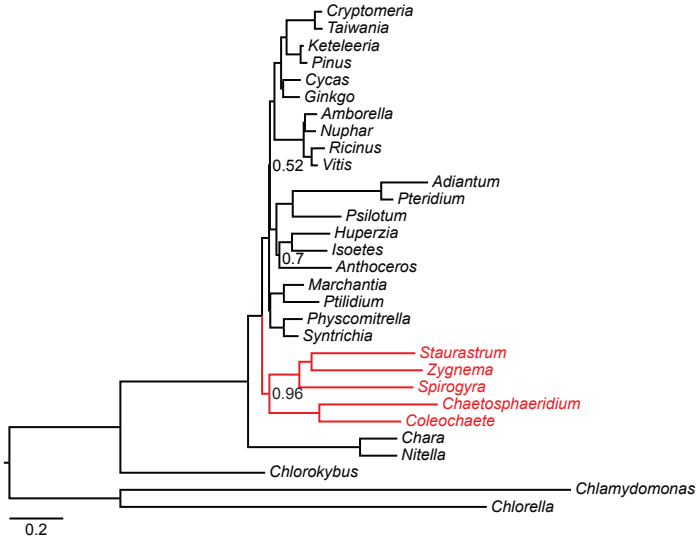


Figure S1. Phylogenetic trees using homogeneous and heterogeneous models based on the 34,879 matrices. Numbers on the tree indicate the maximum likelihood bootstrap percentage (BP) and Bayesian posterior probability (PP) , and nodes with 100 BP or 1.0 PP are not marked.

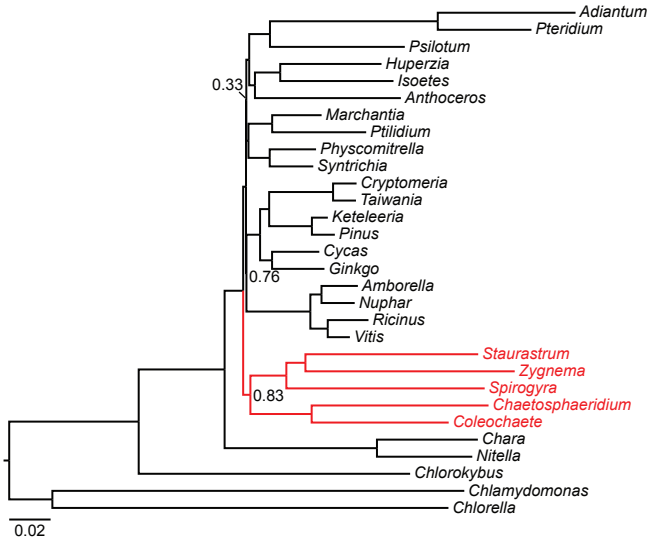
(A). 32879-RAxML



(B). 32879-PhyloBayes



(C). 32879-BayesPhylogenies



(D). 32879-nhPhyML

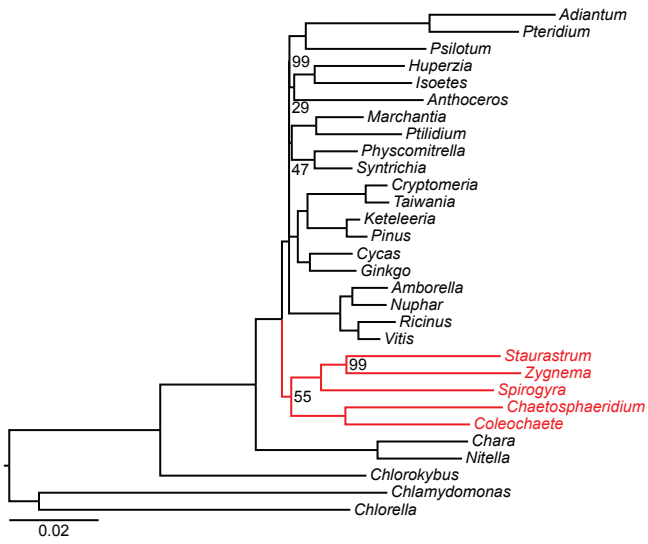


Figure S2. Phylogenetic trees using homogeneous and heterogeneous models based on the 32,879 matrices. Numbers on the tree indicate the maximum likelihood bootstrap percentage (BP) and Bayesian posterior probability (PP) , and nodes with 100 BP or 1.0 PP are not marked.

Chapter 4.

White, W.T.J*., **Zhong, B***., Penny, D. (2013). Beyond reasonable doubt: evolution from DNA sequences. ***PLoS One***. 8: e69924. (***equal contribution**)

There are several areas of science where there is still strong resistance to basic scientific conclusions: anthropogenic climate change, the reality of long-term evolution, and the use of stem cells in medical research are three best-known examples. The issues are not just of academic significance because the beliefs of some individuals have serious political consequences affecting innocent people. Thus we still require stronger but simple and direct tests of our main scientific hypotheses, even if the main conclusions appear obvious to researchers. In this chapter we developed a simple and direct quantitative test of a prediction of common ancestry. The inferred ancestral sequences of proteins should converge as we go further and further back in time. This ancestral convergence was demonstrated quantitatively and continuously using different types of datasets as we trace further back in time.

I was responsible for collecting all the empirical datasets (including chloroplast genome, mitochondrial genome, and nuclear data from plants and animals at increasing level of divergence), and for reconstructing the ancestral sequences of proteins. I also greatly contributed to the development of statistical test and the writing of the manuscript. All authors contributed to the final manuscript.

OPEN ACCESS Freely available online

PLOS ONE

Beyond Reasonable Doubt: Evolution from DNA Sequences

W. Timothy J. White[1¤◔], Bojian Zhong[1◔], David Penny[1]*

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

Abstract

We demonstrate quantitatively that, as predicted by evolutionary theory, sequences of homologous proteins from different species converge as we go further and further back in time. The converse, a non-evolutionary model can be expressed as probabilities, and the test works for chloroplast, nuclear and mitochondrial sequences, as well as for sequences that diverged at different time depths. Even on our conservative test, the probability that chance could produce the observed levels of ancestral convergence for just one of the eight datasets of 51 proteins is $\approx 1\times10^{-19}$ and combined over 8 datasets is $\approx 1\times10^{-132}$. By comparison, there are about 10^{80} protons in the universe, hence the probability that the sequences could have been produced by a process involving unrelated ancestral sequences is about 10^{50} lower than picking, among all protons, the same proton at random twice in a row. A non-evolutionary control model shows no convergence, and only a small number of parameters are required to account for the observations. It is time that that researchers insisted that doubters put up testable alternatives to evolution.

Citation: White WTJ, Zhong B, Penny D (2013) Beyond Reasonable Doubt: Evolution from DNA Sequences. PLoS ONE 8(8): e69924. doi:10.1371/journal.pone.0069924

Editor: Keith A. Crandall, George Washington University, United States of America

Received December 14, 2012; **Accepted** June 12, 2013; **Published** August 8, 2013

Funding: The work was supported (in part) by the N Z Marsden Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: d.penny@massey.ac.nz

¤ Current address: Institut für Informatik, Friedrich Schiller Universität, Jena, Germany

◔ These authors contributed equally to this work.

Introduction

There are some areas of science where there is still strong resistance to basic scientific conclusions: anthropogenic climate change [1], the reality of long term evolution [2] http://www.dissentfromdarwin.org, the origin of life, and the safety and efficacy of vaccination programs [3] are well-known examples. Thus we still require strong quantitative tests of our main scientific hypotheses, even if the conclusions appear obvious to most researchers. In the case of evolution, a strong prediction of Darwin's 'descent with modification' [4] is that, as we go further and further back in time, the sequences for a given protein should become increasingly similar – we call this either 'ancestral convergence' or 'reverse convergence'. The prediction from evolutionary theory is that DNA or protein sequences carrying out the same basic functions in different organisms are generally inherited from a common ancestor – in this sense they are fully homologous proteins (or orthologs) [5]. We must be able to measure this convergence and test it quantitatively. In practice, although the information comes primarily from DNA sequences, we convert them to protein sequences for the tests. As we see later, we currently cannot yet find any other hypothesis that leads inevitably to the same prediction without an explosive increase in the number of parameters.

It is basic to science that we have never tested all possible hypotheses; consequently we never obtain final and absolute knowledge about any aspect of the universe. Nevertheless, the scientific method provides us with the best form of knowledge that humans can attain, and ensures that we use the most thoroughly tested understanding at any time [6]. This Popperian framework allows both Bayesian and frequentist approaches to be used, dependent on what is appropriate for the questions being tested.

We use a non-evolutionary null model and develop a quantitative test of ancestral convergence, and apply it to a range of datasets that have diverged at deeper and deeper times. As a control we show that unrelated proteins do not show convergence. Furthermore, an excessive number of free parameters are required to account for the observed convergence by other processes. This clearly does not 'prove' that yet unknown models are impossible, but the theory of evolution leads to extremely strong predictions, and so the onus is now on others to propose testable alternatives.

Materials and Methods

We develop a statistical test for quantifying convergence that consists of eight simple steps For Step 1 we take two subgroups of taxa X and Y (see Figure 1) that on independent evidence have non-overlapping subtrees; that is, they are natural subgroups (or clades). For example, with chloroplast sequences, we select subgroups based on nuclear and/or mitochondrial data [7,8], and only later check that the subgroups are also supported by the chloroplast sequences. For each subgroup we independently align the sequences (Step 2); infer a subtree (Step 3); and infer the ancestral sequences a_x and a_y for the deepest nodes of each subtree (Step 4). For this step we use PAML [9], which is a well-established method that is robust to small changes in the tree [10]. Our test is

conservative in that ancestral sequences are estimated independently: information from subgroup X is not used to estimate the ancestral sequence for subgroup Y, nor vice versa. We used the cpREV model [11] for inferring chloroplast trees, the WAG model [12] for nuclear proteins, and the mtREV24 model [13] for animal mitochondria. We obviously can never know whether these are the best possible models for estimating convergence, but any better models are predicted to show even greater convergence.

The program MUSCLE [14] is used for calculating alignment scores, see details later. For Step 5, the pairwise alignment score $s(a_x,a_y)$ is then calculated between the inferred ancestral sequences a_x and a_y (we call this the 'ancestral score'), with higher values showing that ancestral sequences are more similar (Table 1). In Step 6 we then calculate the alignment score $s(i,j)$ for all pairs of sequences (with just one sequence from each of the two subgroups). From the resulting distribution of between-subgroup scores (see Figure 2) we calculate (Step 7) the probability p of observing scores at least as high as the ancestral score under the null model, which we now describe.

Our null model can be considered in the following way - that the taxa in subgroup X are descended from an unknown number $1<=r_X<=|X|$ of root sequences, the taxa in subgroup Y are descended from an unknown number $1<=r_Y<=|Y|$ of root sequences, and that the r_X+r_Y root sequences are all independent from each other. This allows, at one end of the spectrum, the possibility that all $|X|+|Y|$ taxa were independently created, and at the other end of the spectrum, the possibility that all taxa in one subgroup are descended from a single common ancestor *for that subgroup*, which was created independently of the single common ancestor for the other subgroup. In other words, this null model imposes no requirements on the presence or absence of internal (within-subgroup) evolution of the two subgroups of taxa; the only constraint is that there is no evolutionary link between the two subgroups. That is, neither subgroup contains taxa derived from the other, nor from a common ancestor.

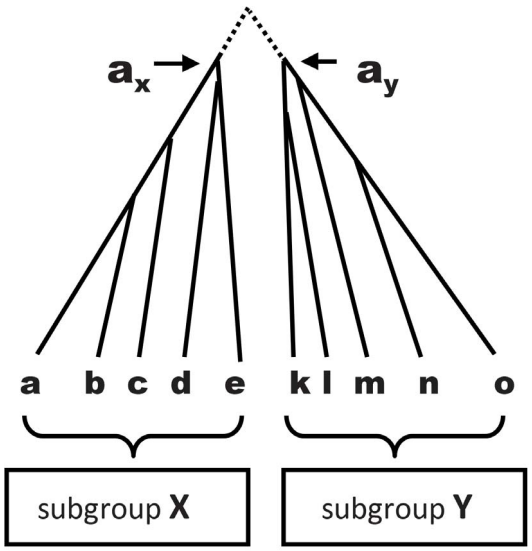


Figure 1. We use two natural subgroups (X and Y), independently align the sequences for the species in each subgroup, independently determine the optimal tree for each subgroup, independently infer the ancestral sequences a_x and a_y on the optimal subtrees (in practice the sequence at the nearest node to the root of the subtree is estimated), and finally measure the pairwise alignment score between the ancestral sequences, $s(a_x,a_y)$. Separately, we measure the alignment score between each pair of sequences ($s_{(i,j)}$) with one member in each of the two subsets, for example, $s_{(a,k)}$, $s_{(a,l)}$, $s_{(a,m)}$, and so on.
doi:10.1371/journal.pone.0069924.g001

That the numbers r_X and r_Y are not specified helps generalise the null model because a tree built on all taxa in X using any statistically consistent method will necessarily contain r_X long edges from some "central" node to the subtrees containing the taxa. (Sampling error will in general cause these long edges to be connected to the central node by one or more short edges, rather than being a "pure" star tree, but these edges can be made arbitrarily short by using enough characters.).

Thus our non-evolutionary null model predicts that the similarity between the ancestral sequences is equal to the similarity between the extant sequences (that was calculated above in Step 7). When evolution from a common ancestor has occurred, the ancestral sequences will be significantly more similar than that predicted by the null model, and the null model will be rejected. (Some implications of the choice of null model are discussed further in the Discussion section.) For Step 8, additional power is achieved by using independent tests on different genes and combining the resulting p values [15] into a single value that represents the probability of observing data as, or more, extreme than that actually observed. Our test is again conservative: when handling between-group pairwise alignment scores equal to the ancestral score we consider these to be larger than the ancestral score (see Figure 3).

At this point we mention the possibility that two sister taxa could have been (mis-)placed in different subgroups. Although this does not fit within our null model, the only effect is to increase the measured between-subgroups average similarity, making it *harder* for the measured ancestral similarity to exceed it. Thus this model violation cannot induce a false positive (i.e. a claim that evolution is present when it is not) – only a false negative could occur. In any case, we aim to avoid these false negatives by selecting subgroups using external data.

It is important that our test can reject ancestral convergence with a control generated by a non-evolutionary process. This control differs only in that specific property for which we are testing: shared ancestry of homologous proteins for the subgroups X and Y. For this reason, each of our control datasets is a pair of subgroups of taxa X and Y as before, but in which the sequences used for subgroup X come from a different gene than those used for subgroup Y. This corresponds to (i.e. could be generated by) the "archetypes, followed by degeneration" model favoured by some pre-Darwinian biologists, discussed later. We do not expect to see convergence between, say, the ancestor of the monocot atpA gene and the ancestor of the eudicot psbF gene.

We measure the similarity of two sequences by the pairwise alignment score calculated using the MUSCLE alignment program [14] with default scoring parameters. The alignment score is the sum of the per-site scores, which are found from a pre-specified table that records the score for every possible combination of two amino acids, or one amino acid and a gap (see Table 1). The freedom in placing gaps means that different alignments of two given sequences are possible; the job of an alignment program such as MUSCLE is to find a high-scoring (ideally the highest-scoring) alignment. Note that setting the scores of all equal pairs of amino acids to 0 and all other scores to -1 will cause an alignment algorithm to recover an alignment having the fewest possible insertions, deletions and substitutions, and the number of these events (the *Levenshtein edit distance*) will be equal to the negative of the alignment score. This *edit* distance is a useful measure of similarity between strings that are not constrained to have the same length, but more biologically realistic alignments can be

Table 1. Calculation of alignment score for the inferred ancestral monocot and eudicot sequences of the psbK gene.

1									**10**										**20**										**30**							
M	L	N	I	L	N	L	I	C	I	C	L	N	S	A	P	Y	S	S	S	F	F	C	A	K	–	–	P	A	Y					
M	L	N	I	I	S	L	–	I	C	L	N	S	A	L	H	S	S	S	F	F	F	A	K	L	P	E	A	Y						
1.010	0.771	0.840	0.759	0.376	0.328	0.771	−0.729	0.000	0.759	3.949	0.771	0.840	0.535	0.563	0.134	0.454	0.535	0.535	0.535	1.521	1.521	0.125	0.563	0.682	−0.725	0.000	0.207	0.563	1.877					
31										**40**									**50**										**60**					
A	V	F	N	P	I	V	D	F	M	P	V	I	P	V	L	F	F	L	L	A	F	V	W	Q	A	A	V	S	F	R			
A	F	F	N	P	I	V	D	F	M	P	V	I	P	V	L	F	F	L	L	A	L	V	W	Q	A	A	V	S	F	R			
0.563	0.228	1.521	0.840	1.812	0.759	0.664	1.047	1.521	1.010	1.812	0.664	0.759	1.812	0.664	0.771	1.521	1.521	0.771	0.771	0.563	0.376	0.664	7.849	0.717	0.563	0.563	0.664	0.535	1.521	1.072			

The ancestral sequences for the monocots, and the eudicots, are inferred independently, and then the alignment scores calculated in the program MUSCLE. The individual column scores depend on the frequencies and properties of the two amino acids; higher scores are given for pairs that are similar (readily substitutable) or specific (more readily substitutable for each other than for other amino acids). The column scores are summed to produce the alignment score.

doi:10.1371/journal.pone.0069924.t001

recovered by reducing the penalty for mutations between amino acids having similar codons or similar physical properties (e.g. size, hydrophobicity) as these mutations are more likely to occur or to survive into subsequent generations. Alignment quality is also improved by reducing the penalty incurred by multiple contiguous gap characters. MUSCLE's default scoring parameters have been empirically tuned to work well with most protein datasets, and as such MUSCLE's pairwise alignment score is a good measure of overall protein sequence similarity.

Evolution is a stochastic process that involves reversals and parallel changes – for example, if the change val → ile is effectively neutral at a site, and has already occurred, then it is always possible that the reverse mutation (ile → val) will occur. For such reasons, the ancestral sequence actually inferred depends on a stochastic process, so although we do not expect the relation $s(a_x,a_y)>s_{(i,j)}$ to hold in every case we predict reliability increases as sequences become longer (Figure 4). This effect of sequence length is important support for the stochastic process of evolution, but that is not the primary focus here.

Fisher's method [15] combines *p*-values from multiple independent tests of the same null hypothesis into a single *p*-value. We use it to combine the results of individual gene tests. Briefly, if the null hypothesis is true, then the *p*-value obtained from a test will be uniformly distributed between 0 and 1; taking the log and multiplying by −2 produces a quantity that is X^2-distributed with 2 degrees of freedom. Thus the following statistic,

$$X^2 = -2\sum_{i=1}^{k} \ln p_i$$

will be X^2-distributed with $2k$ degrees of freedom. Once this statistic is calculated, a one-sided test can be used to extract a *p*-value from it, representing the probability of observing *k p*-values as low as those that were observed, assuming the null hypothesis (namely that the *k* original null hypotheses are correct).

The standard evolutionary model is relatively simple, and explains the basic tree-like structure of the sequences. To infer an ancestral sequence, the simplest models require only ≈190

Figure 2. Cumulative frequency plots comparing the alignment score for the ancestral sequences ($s(a_x,a_y)$, small circle) with the alignment scores of all pairs of proteins, $s(i,j)$. The example is the monocot/eudicot chloroplast dataset and for the short protein psbF (**2A**), a longer protein atpA (**2B**), and the 51 concatenated genes (**2C**). The x-axis shows the alignment score, which increases with the length of the protein(s), and is largest for the 51 concatenated proteins. There are 1056 $s(i,j)$ scores between pairs of 24 monocots and 44 eudicots, and the y-axis indicates where the $s(a_x,a_y)$ fits as a proportion of this number. For some short proteins in particular, multiple $s(i,j)$ values equal the ancestral score $s(a_x,a_y)$, and in this case our test conservatively places the ancestral score below the rest (as in psbF in Fig 2A).

doi:10.1371/journal.pone.0069924.g002

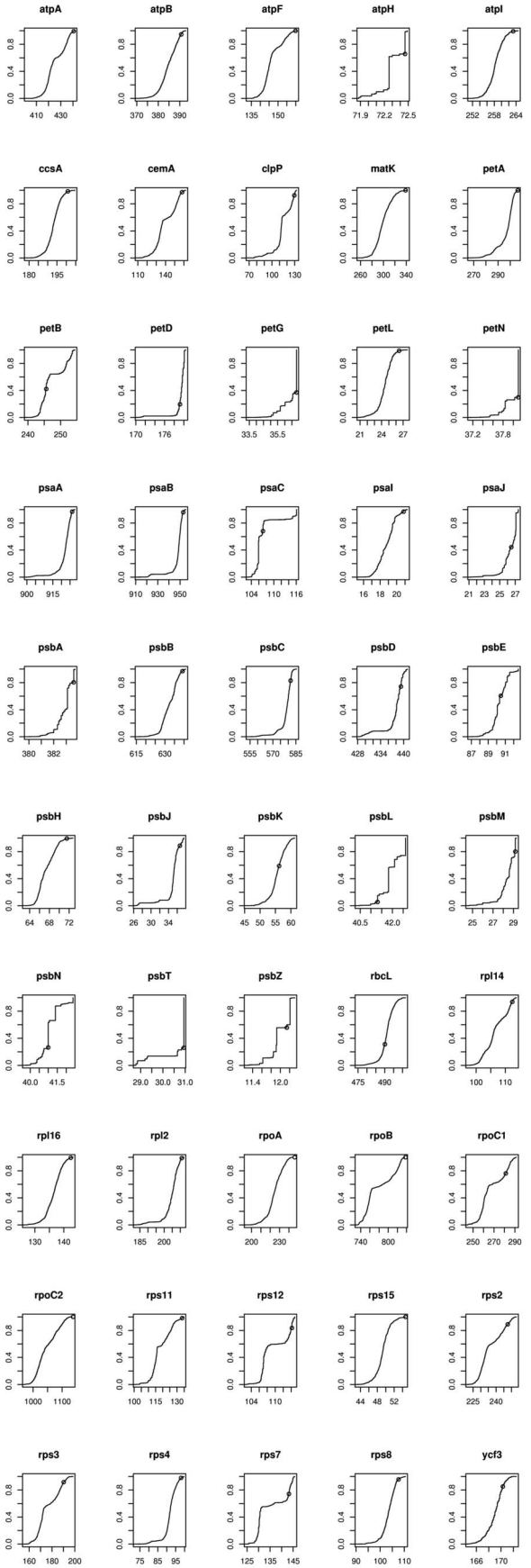


Figure 3. Cumulative distribution plots of the between-groups alignment scores for 50 of the 51 chloroplast proteins of monocots and eudicots (the plot for psbF is Fig 2A). The ancestral alignment score ($s(a_X,a_Y)$) is indicated by a small circle on each plot. There are 1056 comparisons (24 monocots×44 eudicots) for each protein. The y-axis is the same for each gene, but the x-axis is strongly dependent on the length of the protein (see also Figure 4).
doi:10.1371/journal.pone.0069924.g003

parameters (a 20×20 symmetric matrix for the probability of changes between pairs of amino acids, less 20 because each row has to sum to 1). Then there is one additional parameter for each edge (branch) of the tree (there are $2n$-3 edges for a binary tree, where n is the number of taxa). We could add one parameter for a probability of splitting of lineages, a second for an overall rate of change, a third for the distribution of rates across sites (e.g. for a Gamma distribution of rates), and a fourth for the proportion of invariable sites. Nevertheless as we later show, there are orders of magnitude fewer parameters required for a general evolutionary model than for a minimal 'design' model, and scientifically, we select the simpler model.

Genuine Subgroups

It is important to demonstrate that the two subgroups or clades (X and Y) are genuine, and we do this for each of the subgroups in Table 2 in two ways. Firstly, the two subgroups are determined by other data – for example by nuclear or by mitochondrial DNA sequences for the plant chloroplast data. Secondly, for each of the eight pairs of datasets in Table 2 we later combine the two datasets, and confirm that the same two subgroups are still found – for example, the monocots and eudicots. This independent selection of the two subgroups is necessary because if, for example, we formed one subgroup by randomly selecting half the monocots and half the eudicot sequences, and used the other taxa to form the second subgroup, then we could artefactually get similar ancestors. So both tests (selecting subgroups from independent data, and later showing that the subgroups are recovered with the data used) are important in demonstrating that the subgroups X and Y are natural.

Estimating the Root of the Two Subtrees

There are several ways of estimating the root of the two subtrees, but in practice it appears to make little difference which of several methods we use. In the chloroplast example, the root of each subtree can be inferred from nuclear or mitochondrial DNA sequences (not chloroplast), and so is independent of the chloroplast data we use. This gives the position of the root in each subtree from prior information; alternatively they can be independently estimated by 'midpoint rooting'. This can be done either by selecting the midpoint of the longest path, or the internal branch with the longest average of paths passing through it [16]. In practice, we take the node closest to the mid-point because we are estimating nodal sequences. There does appear to be an acceleration of the rate of evolution in the grasses [17], but, again in practice, this appeared to have little effect. The sequence of the root of the two subtrees appears to be quite robust.

Note that we could quite separately make an independent test for the similarity of evolutionary trees, by comparing the likelihood of chloroplast, nuclear, and mitochondrial datasets giving such highly similar trees. (Here we are only concerned, for example, about the ancestral sequences of the monocot/eudicot split – not the similarity of the trees as a whole.) Instead of computing alignment scores between pairs of sequences, maximum likelihood distances could in principle be computed for different-length

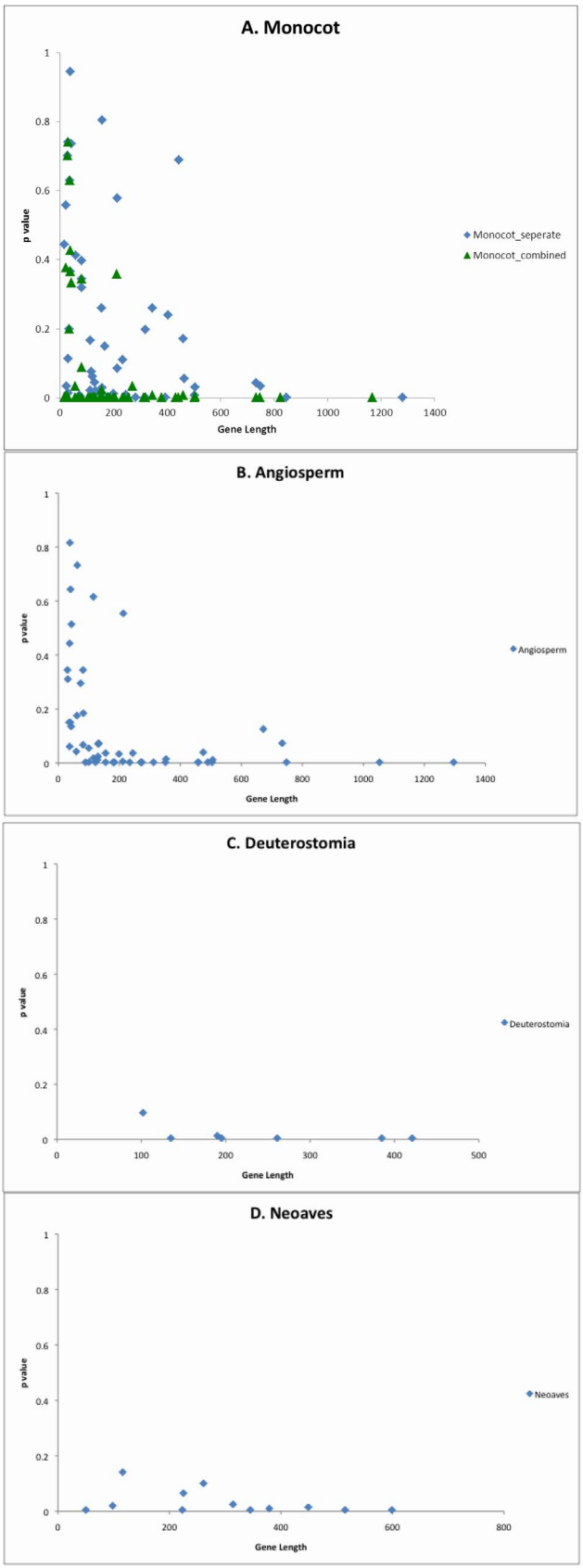


Figure 4. Protein length versus proportion of pairwise alignment scores higher than the ancestral score, for 4 datasets. Because of the possibility of slightly different gene lengths just one of the two datasets is used for illustration. As expected, longer proteins show convergence more strongly. Chloroplast results (4A and 4B) are for 51 chloroplast genes for divergences of monocots and for angiosperms (flowering plants). There are 7 nuclear proteins for 4C and 12 for the mitochondrial data in 4D.
doi:10.1371/journal.pone.0069924.g004

sequences by using models of evolutionary change that allow for insertions and deletions, such as the TKF model – however software for computing these distances is apparently not currently available.

We start with chloroplast genomes because they have more than 50 protein genes (allowing both individual and combined tests); although there is some loss of genes from chloroplasts, there are no basic problems identifying homologous genes. In addition, there are several datasets at increasing levels of divergence, e.g.

monocots versus eudicots (both within flowering plants);

flowering plants angiospermsversus gymnosperms,

seed plants versus ferns and fern allies;

Streptophytes (land plants plus some green algae) versus Chlorophytes (most green algae).

These plant subsets have a wide range in their inferred divergence times; from about 125 to over 700 million years before the present [18,19].

The tests are repeated on nuclear encoded sequences from animals, and then on avian mitochondrial genomes. With nuclear encoded proteins we test convergence for seven genes and for two groups for the deeper animal divergences, ranging from around 600–700 Mya [20]. The first test is for Vertebrata plus Urochordata versus Echinoderms plus Hemichordates. The second is for Deuterostomes versus Lophotrochozoa. For the nuclear datasets the 7 genes used are aldolase, methionine adenosyltransferase, ATP synthase beta chain, catalase, elongation factor 1 alpha, triosephosphate isomerase, and phosphofructokinase.

For mitochondrial sequences, we use a dataset from birds, using 12 protein-coding genes, and focus on two tests – firstly Neoaves [21] (most birds) versus Galloanseriforms (chickens and ducks), and secondly these two groups combined (neognaths) versus paleognaths (ratites and tinamous) [22]. Their estimated divergence times are around 80 and 100 million years ago, respectively [21,22]. For the mitochondrial dataset, the 12 genes are ATP6, ATP8, COX1, COX2, COX3, Cytb, ND1, ND2, ND3, ND4, ND4L, and ND5.

Results

Our primary results are very clear and are shown in Table 2. Our first example uses chloroplast genomes from 44 eudicotyledonous and 24 monocotyledonous flowering plants (monocots include grasses, palms and lilies). Combining results for all 51 genes gives a p value for our non-evolutionary null model of $\approx 2\times10^{-19}$, shown in the top row of results in Table 2. This eudicot/monocot subdivision can be derived independently from either nuclear or mitochondrial DNA sequences [8], and so is independent of the chloroplast information. The 51 chloroplast proteins common to all lineages total 11,414 amino acids in length with an average length per protein of ~225 amino acids (see Table 3). However, the proteins vary in length from 16 amino acids (psbZ) to 1168 amino acids (rpoC2). Across all 51 genes, on average 22% of pairwise scores were at least as high as the ancestral score, but this is mostly caused by a small number of shorter genes with relatively low ancestral scores (see Figure 3). Results are shown for each of the proteins in Table 3 and Figure 3.

Figure 4 shows a correlation between protein sequence length and convergence, certainly consistent with a stochastic mechanism.

Our standard approach infers the ancestral sequences on the two subtrees independently. If we follow the more usual method and jointly infer the ancestral sequences on a single tree using the combined monocot and eudicot data we get, as expected, an even higher alignment score between the two ancestral sequences a_X and a_Y, that is, the ancestral sequences are even more similar. Part of the reason for this is that we are using more information when inferring the ancestral sequences. Combining probabilities for all genes using Fisher's method as before, we find that the probability of observing such high ancestral scores for the 51 chloroplast proteins under our non-evolutionary null model is 1.51×10^{-57} (compared with $\approx2\times10^{-19}$, see the top row of Table 2) - our test is thus very conservative.

It is a fundamental prediction from evolutionary theory that convergence should continue at deeper times, and this is strongly supported as shown by the first four rows of results in Table 2, which use chloroplast genomes from deeper and deeper divergence times (column 4). This eliminates one simple model that allowed creation of 'archetypes' and limited evolution thereafter (see later discussion). Similarly, we find ancestral convergence with nuclear encoded sequences from vertebrates and invertebrates, and also with mitochondrial genomes from birds. Thus we have used chloroplast, nuclear, and mitochondrial DNA sequences, and from a wide variety of species. The times of divergence of the different datasets are estimated to vary from 80–700 millions of years ago (Mya) [18–22]. If we combine all 8 tests we get a p value of $\approx 2\times10^{-132}$, and this is shown in the second to bottom row of Table 2.

The last two columns in Table 2 are control values where we compare the inferred ancestral sequence of one protein against the inferred ancestral sequence of a different protein. As expected, there is no tendency for these separate proteins to converge to similar sequences, making them good and effective controls. Indeed, the combined p value on the eight control datasets is $p = 0.93172$: indicating that the inferred ancestral scores are consistently below the average between-subgroups alignment score – again, our test is conservative.

The analyses establish that some form of ancestral convergence is occurring, and it is essential to explain the continued convergence as we go back to more distantly related organisms. Of course, such analyses by themselves cannot establish the mechanisms of evolutionary change (though the results are fully consistent with a stochastic mechanism, see also later).

Discussion

Our test is based on the expectation that, under evolution, the ancestral sequence of one natural group of taxa will be more similar to the ancestral sequence of a second natural group of taxa, than to any sequence from the first group will be to any sequence from the second. In contrast, a variety of proposed non-evolutionary models either do not make this prediction, or require so many parameters that they cannot be said to make any testable predictions at all.

The basic results in Table 2 are overwhelming evidence that some form of ancestral convergence is occurring, and continues at deeper and deeper times. Individual tests have probabilities from 10^{-6} (for small numbers of genes) to 10^{-44} (for the larger number of genes in chloroplasts). Equally important, non-homologous controls show no tendency to converge (Table 2) – it is only homologous proteins that show ancestral convergence.

Table 2. Summary of X^2 and p values for the different datasets.

Data type	Group X	Group Y	divergence times	X^2	d.f.	p (X^2)	X^2 (control)	p (X^2) (control)
c/plast - 51	Eudicot (44)	Monocot (24)	~125mya	289.0582	102	1.94E-19	93.68962	0.70943
c/plast - 51	Angiosperm (25)	Gymnosperm (13)	~305mya	363.5268	104	1.23E-29	85.64681	0.90482
c/plast - 51	Seed plant (38)	Fern (7)	~390mya	457.1184	102	1.69E-44	100.4507	0.52486
c/plast - 51	Streptophyta (52)	Chlorophyta (6)	~700mya	300.1617	94	2.23E-23	90.98169	0.56897
Nuclear - 7	Vertebrata+Urochordata (9)	Echinoderms+Hemichords(10)	~600mya	54.50034	14	1.05E-6	9.642275	0.78784
Nuclear - 7	Deuterostomia (19)	Lophotrochozoa (12)	~670mya	67.63153	14	5.17E-9	12.66025	0.55343
mitochond-12	Neoaves (22)	Galloanserae (9)	~80mya	99.55765	24	3.57E-11	16.03990	0.88663
mitochond-12	Neognath (31)	Palaeognath (12)	~100mya	102.5291	24	1.11E-11	23.69508	0.47914
combined	all 8 pairs of datasets			1631.555	478	2.59E-132	432.8063	0.93172
joint tree	Eudicot (44)	Monocot (24)	~125mya	539.3154	102	1.51E-57		

The numbers of genes for the subgroups are indicated after the data type, and the number of taxa are indicated in parentheses '()'following the group name. The divergence times are minimum estimates from fossils and molecular data. Columns 5–7 relate to the probability that convergence could have arisen by chance, the last two columns are from controls where convergence is not expected. The penultimate row gives the combined values for the 8 datasets. The final row gives the results for the first example where combined information from both subsets (eudicots and monocots) is used for estimating the ancestral sequence of both subgroups; this again indicates that our test is very conservative.

doi:10.1371/journal.pone.0069924.t002

Table 3. Results for the 51 genes for the monocot/eudicot chloroplast dataset, and with the ancestral sequences (a_x and a_y) inferred independently.

Gene	Gene length (amino acids)	Ancestral alignment score	Number of higher alignment scores	Proportion of higher alignment scores	Chi-squared term
atpA	503	440.6	8	0.0076	9.7656
atpB	433	390.7	58	0.0549	5.8036
atpF	178	159.9	2	0.0019	12.5382
atpH	81	72.5	364	0.3447	2.1302
atpI	241	263.2	10	0.0095	9.3193
ccsA	150	200.8	16	0.0152	8.3793
cemA	155	158.0	30	0.0284	7.1221
clpP	186	129.5	80	0.0758	5.1604
matK	380	338.8	1	0.0009	13.9245
petA	313	306.2	1	0.0009	13.9245
petB	213	245.6	611	0.5786	1.0943
petD	157	179.2	850	0.8049	0.4340
petG	37	36.8	665	0.6297	0.9249
petL	30	26.4	15	0.0142	8.5084
petN	29	38.1	741	0.7017	0.7085
psaA	748	926.2	35	0.0331	6.8138
psaB	733	952.9	45	0.0426	6.3112
psaC	81	107.1	338	0.3201	2.2784
psaI	24	20.8	35	0.0331	6.8138
psaJ	22	26.5	589	0.5578	1.1676
psbA	319	383.6	208	0.1970	3.2494
psbB	506	639.2	33	0.0313	6.9315
psbC	460	581.6	181	0.1714	3.5275
psbD	345	439.3	274	0.2595	2.6982
psbE	80	90.5	419	0.3968	1.8488
psbF	39	43.8	387	0.3665	2.0076
psbH	71	71.5	7	0.0066	10.0327
psbJ	30	36.6	119	0.1127	4.3662
psbK	57	56.2	436	0.4129	1.7692
psbL	38	41.3	998	0.9451	0.1130
psbM	34	29.2	211	0.1998	3.2208
psbN	43	41.0	778	0.7367	0.6110
psbT	30	30.9	782	0.7405	0.6008
psbZ	16	12.1	469	0.4441	1.6233
rbcL	443	490.1	728	0.6894	0.7439
rpl14	122	112.4	66	0.0625	5.5452
rpl16	126	142.4	8	0.0076	9.7656
rpl2	201	210.9	13	0.0123	8.7946
rpoA	256	246.2	0	0.0009	13.9245
rpoB	824	837.7	0	0.0009	13.9245
rpoC1	270	281.3	253	0.2396	2.8577
rpoC2	1168	1138.0	0	0.0009	13.9245
rps11	131	133.6	20	0.0189	7.9330
rps12	113	114.2	175	0.1657	3.5949
rps15	57	54.7	2	0.0019	12.5382
rps2	234	247.6	117	0.1108	4.4001
rps3	207	190.2	90	0.0852	4.9249
rps4	105	98.0	22	0.0208	7.7424

Table 3. Cont.

Gene	Gene length (amino acids)	Ancestral alignment score	Number of higher alignment scores	Proportion of higher alignment scores	Chi-squared term
rps7	155	143.2	274	0.2595	2.6982
rps8	125	107.6	47	0.0445	6.2242
ycf3	157	170.3	158	0.1496	3.7993
Av/Sum	224.6	239.9	230.8	0.2186	289.0582

doi:10.1371/journal.pone.0069924.t003

It is always possible to 'design' much more complex models where a separate decision by some unknown agent chooses/selects each amino acid change. With so many parameters the model is able to precisely mimic evolution (or indeed any other model); it has no discernible ''signature'' of its own. A minimum number of parameters for such a complex model can be determined by constructing, on the complete set of sequences of both subgroups, a variation on a maximum parsimony tree that allows single-residue insertions and deletions in addition to single-residue substitutions. This tree is an example of a Steiner tree [23] – a tree of minimal total edge length that connects a given set of points in a metric space, allowing for the introduction of new intermediate (ancestral) points as required. In this case, the distance between two sequences is given by the edit distance (the minimum number of single-residue insertions, deletions and substitutions ['edits'] required to transform one sequence into the other).

The length of this tree is then by definition the minimum possible number of separate decisions, or equivalently free parameters, that a hypothetical external agent requires in order to produce the complete set of sequences, given any one of the sequences as a starting point. A lower bound for the length of a Steiner tree is given by half the length of a minimum spanning tree, which is a tree that connects all given points without introducing additional points [24]; minimum spanning trees can be computed efficiently. For the eudicot/monocot example, a minimum spanning tree requires 36,473 mutations to connect all 68 sequences, implying that we would need at least 36,473/2 = 18,237 free choices, each a separate parameter. Any suggestion that a model with such a huge number of parameters 'explains' the data is of course a serious violation of the scientific principle of selecting the simplest model.

Early (pre-Darwinian) biologists suggested several ideas as to the relationship of modern organisms, but a relevant one here is the 'archetype' model [25] that suggested that a number of 'forms' were originally created within high-level groups. For mammals say, one 'form' would have been a giant cat, which then independently evolved (or degenerated) into lions, tigers, leopards, panthers, cheetahs, etc. In our examples, this is tested (and eliminated) by demonstrating that successively deeper datasets continue to show ancestral convergence. In other words, we do not see a set of 'archetype' species originating at just one point in time – there is continuity in the evolutionary process. A qualification is that at the very deepest times we expect that information will be lost - this is a property of the Markov models used [26]. However, similar tests could be done at deeper times using measures of similarity of 3D structure. Indeed, it is important to note that modern methods of molecular biology now allow ancestral sequences to be synthesized, and the properties of the protein products of the 'ancestral' genes can be tested [27]. So there is now no doubt that these ancestral sequences do meet the functional requirements of the ancestors.

Perhaps, it is important for scientists to emphasize that by any scientific standard, evolution is simply inevitable. Good examples of continued evolution are RNA viruses, such as the influenza viruses; they just keep evolving from year to year – evolution in real time. New anti-viral immunisations are prepared for each northern hemisphere winter and for each southern hemisphere winter, and so on. Certainly, DNA-based organisms evolve more slowly; they have a lower mutation rate. But the inevitability of the process is there. ''Stop the World, I want to get off'', was the title of a 1960s musical. We can neither stop the world (and get off), nor can we stop evolution. In the viral case there can also be recombination between RNA genes e.g. influenza, [28] or between sections of the genome (hepB) [29] – these recombinations are the equivalent of macroevolution (ref [28], Chap 5). We have already tested (and rejected) some 'non-standard' models for influenza evolution [30]. However, each gene (or section of the gene) should still converge, even if there is lateral gene transfer. Even though the fidelity of DNA copying is extraordinary - around 1 error in 10^9–10^{10} nucleotides copied [31], no known organism can copy its DNA with absolute accuracy – thus there is always genetic diversity in natural populations.

So our conclusions are perhaps three-fold. Firstly we have provided a strong quantitative test rejecting a non-evolutionary model that amino acid sequences do not become more similar as we go back in time. Secondly, we have raised the problems of the number of parameters required of some alternatives, and finally we shift the requirement onto doubters to provide testable alternatives. On this third aspect, there does appear to also be a similar reaction from climate change advocates on placing responsibilities onto doubters [32]. Other aspects of evolution have been tested [33–35] and further aspects of evolution could be tested, perhaps especially the 'random' nature of mutations that occur without regard for any 'need' of the organism, but this is outside the scope of the present work. Indeed, there has always been excellent support for evolution from fossils and comparative morphology, and molecular data enables this to be quantitative. We can say that, as yet, no features of genomes have yet been found that are not understandable by 'causes now in operation' [4].

From the scientific point of view, there is no doubt that evolution has occurred, and there really were a continuous set of intermediates connecting individuals, populations, varieties, species, genera, families, etc. Nevertheless, as scientists we need to ensure that we have good quantitative tests available of all our favoured models. Given our results, we suggest that researchers need to be more assertive that evolution has both occurred, and continues to occur. It is essential that any person who does not accept the continuity of evolution puts forward alternative testable models. As we tell our first year undergraduates, 'belief is the curse of the thinking class'.

Aligned datasets are available upon request from the authors. Alignments were carried out by BJZ, trees for each subset were inferred by WTJW, and DP designed the original study. All authors contributed to the final manuscript.

Author Contributions

Conceived and designed the experiments: DP. Performed the experiments: WTJW BZ. Analyzed the data: WTJW BZ DP. Contributed reagents/materials/analysis tools: WTJW BZ. Wrote the paper: DP.

References

1. Rapley C (2012) Time to raft up. Nature 488: 583–585.
2. Garwood R (2012) Reach out to defend evolution. Nature 485: 291.
3. Roberts L (2012) Fighting polio in Pakistan. Science 337: 517–521.
4. Penny D (2011) Darwin's theory of descent with modification, versus the biblical Tree of Life. PLoS Biol 9: e1001096.
5. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19: 99–113.
6. Popper KR (1981) Conjectures and refutations: the growth of scientific knowledge. (Routledge and Kegan Paul, London).
7. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, et al. (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 99: 12246–12251.
8. Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, et al. (2011) Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot 98: 704–730.
9. Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
10. Hanson-Smith V, Kolaczkowski B, Thornton JW (2010) Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol Biol Evol 27: 1988–1999.
11. Adachi J, Waddell PJ, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50: 348–358.
12. 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: 691–699.
13. Adachi J, Hasegawa M (1996) MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comp Sci Monogr Inst Statist Math 28: 1–150.
14. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32: 1792–1797.
15. Fisher RA (1932) Statistical Methods for Research Workers. 4th edition, (Oliver and Boyd, London).
16. Penny D, Steel MA, Waddell PJ, Hendy MD (1995) Improved analyses of human mtDNA sequences support a recent African origin for *Homo sapiens*. Mol Biol Evol 12: 863–882.
17. Zhong B, Yonezawa T, Zhong Y, Hasegawa M (2009) Episodic evolution and adaptation of chloroplast genomes in ancestral grasses. PLoS ONE 4 e5297 DOI: 10.1371/journal.pone.0005297.
18. Clarke JT, Warnock RCM, Donoghue PCJ (2011) Establishing a time-scale for plant evolution. New Phytol 192: 266–301.
19. Leliaert F, Verbruggen H, Zechman FW (2011) Into the deep: new discoveries at the base of the green plant phylogeny. BioEssays 33: 683–692.
20. Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, et al. (2011) The Cambrian conundrum: early divergence and later ecological success in the early history of animals. Science 334: 1091–1097.
21. Pacheco MA, Battistuzzi FU, Lentino M, Aguilar RF, Kumar S, et al. (2011) Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. Mol Biol Evol 28: 1927–1942.
22. Phillips MJ, Gibb GC, Crimp EA, Penny D (2010) Tinamous and moa flock together: mitochondrial genome sequence analysis reveals independent losses of flight among Ratites. Syst Biol 59: 90–107.
23. Hwang FK, Richards DS (1992) Steiner tree problems. Networks 22: 55–89.
24. Gilbert EN, Pollak HO (1968) Steiner minimal trees. SIAM J. Appl. Math. 16: 1–29.
25. Penny D, Hendy MD, Poole AM (2003) Testing fundamental evolutionary hypotheses. J. Theoret Biol 223: 377–385.
26. Mossel E, Steel M (2004) A phase transition for a random cluster model on phylogenetic trees. Math BioSci 187: 189–203.
27. Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW (2012) Evolution of increased complexity in a molecular machine. Nature 481: 360–364.
28. Holmes EC (2009) The evolution and emergence of RNA viruses. (Oxford Univ. Press, Oxford).
29. Harrison GLA, Lemey P, Hurles M, Moyes C, Horn S, et al. (2011) Genomic analysis of Hepatitis B virus reveals antigen state and genotype as sources of evolutionary rate variation. Viruses 3: 83–101.
30. Henderson IM, Hendy MD, Penny D (1989) Influenza viruses, comets, and the science of evolutionary trees. J. Theor Biol 140: 289–303.
31. Herr AJ, Williams LN, Preston BD (2011). Antimutator variants of DNA polymerases. Critical Rev Biochem Molec Biol 46: 548–570.
32. Wei T, Yang S, Moore JC, Shi P, Cui X, et al. (2012) Developed and developing world responsibilities for historical climate change and CO_2 mitigation. Proc Natl Acad Sci USA 109: 12911–12915.
33. Penny D, Foulds LR, Hendy MD (1982) Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 297: 197–200.
34. Theobald DL (2010) A formal test of the theory of universal common ancestry. Nature 465: 219–222.
35. Theobald D (2011) On universal common ancestry, sequence similarity, and phylogenetic structure: The sins of P-values and the virtues of Bayesian evidence. Biol Direct 6: 60.

Chapter 5. Summary and Future directions

Summary

Next-generation sequencing techniques have changed the prospects for molecular evolution. Today there are vast amounts of all types of molecular data available, and it is feasible to obtain more data at a reasonable cost. For example, target enrichment techniques have been recently reported to easily capture homologous genes for multiple species (Faircloth et al. 2012, Lemmon et al. 2012). In the field of phylogenomics which is the use of genomic data to establish and clarify evolutionary relationships, more data indeed is essential to accurately estimate phylogenetic trees (e.g. reducing sampling errors by increasing the amount of information).

However, if evolutionary models do not describe the biological properties of the data, then tree building can go wrong. Worst of all perhaps, while the use of more data could reduce sampling errors, it simultaneously makes systematic errors more apparent. Thus, not all phylogenetic problems can be easily resolved with genome-scale analyses, and more attention must be given to systematic errors when large datasets are used for phylogenetic inference.

In my PhD study, I worked both on plant phylogenomics and on evolution in general. With the next-generation sequencing data, I focused on evaluating and mitigating the systematic errors, applying more sophisticated models to improve phylogenetic inference, and developing a statistical test to provide confirmatory evidence for evolution.

The first project addressed the phylogenetic position of the Gnetales. The position of the Gnetales among the seed plants has been one of the most contentious problems in seed plant phylogeny (Chaw et al. 2000; Donoghue and Doyle 2000; Burleigh and Mathews 2004; Mathews 2009). There was no agreement on the position of the Gnetales within the gymnosperms, and it reported that the Gnetales close to Cupressophyta (non-Pinaceae conifers) may be problematic due to the “long-branch attraction” (LBA) artifact (Zhong et al. 2010). To alleviate the LBA artifact, we sequenced three new chloroplast genomes of southern hemisphere conifers

(*Halocarpus kirkii*, *Podocarpus totara*, and *Agathis australis*) to increase the taxonomic sampling of the Cupressophyta, and applied a site-pattern sorting criterion (Goremykin et al. 2010) to study compositional heterogeneity, heterotachy, and the fit of genome sequences to substitution models. This was done to identify potential systematic artifacts in the data. The results of this study are novel because they show that non-time reversible properties of sites in the chloroplast genomes of Gnetales could mislead phylogenetic reconstruction, and highlight that the goodness of fit between substitution models and the data is crucial to the improvement of phylogenetic inferences. I was also able to pinpoint that there was strong rate heterogeneity in the data, and that because other studies did not identify and correct for this bias, their results were affected by tree reconstruction artifacts (e.g. Chaw et al. 1997; Nickrent et al. 2000; Doyle 2006). Finally, the results support a sister group relationship between the Gnetales and the Pinaceae within Gymnosperms, helping resolve the problem of the Gnetales relationships.

The second topic addressed the origin of land plants. Knowing the closest living relatives of land plants is a prerequisite to understanding the transition from aquatic multicellular green algae to the invasion of terrestrial environments. Firstly, I applied the multispecies coalescent model (Liu et al. 2010) to investigate the origin of land plants using a large number of nuclear genes. In this study, the coalescent analyses across different data sets consistently suggested that the closest relatives of land plants are the Zygnematales. I also showed that the multispecies coalescent model could accommodate gene tree heterogeneity in deep-level phylogenies. In contrast, concatenation methods can yield misleading inferences of species relationships in the presence of a high level of gene tree heterogeneity for the origin of land plants.

Given that congruence of results from multiple independent lines of evidence is a key approach for the validation of phylogenetic estimation, I organized the sequencing of three new chloroplast genomes from streptophyte algae: *Coleochaetae orbicularis* (Coleochaetales), *Nitella hookeri* (Charales), and *Spirogyra communis* (Zygnematales), and used a site-pattern sorting method (Goremykin et al. 2010) and site- and time-heterogeneous models (Pagel and Meade 2004; Boussau and Gouy 2006; Blanquart and Lartillot 2008; Lartillot et al. 2009) to address the branching order among streptophyte algae and land plants. The chloroplast phylogenomic results

strongly rule out earlier hypotheses placing Charales (stoneworts) or Coleochaetales as sister group to land plants. Instead, Zygnematales alone, or a clade consisting of Zygnematales and Coleochaetales, is more likely the closest living relatives of land plants. This result is significant as it substantially confirms the previous nuclear data analyses (Finet et al. 2012; Laurin-Lemay et al. 2012; Timme et al. 2012; Wodniok et al. 2011; Zhong et al. 2013), and indicates that more realistic models have a better fit to the data with more confidence and better infer the phylogenetic trees.

For the third topic as part of my PhD, I worked on a project in which we tested evolutionary theory quantitatively using molecular data in the context of the debate between Evolution and Intelligent Design. Using different types of molecular data, we devised a new sequence-based statistical test to confirm a major prediction of evolution by Darwin's "descent with modification" (Penny 2011): as we go further and further back in time, the ancestral sequences for a given functional protein should converge and become increasingly similar. This is a theoretical paper with potentially overarching implications as it shows that the reality of biological evolution can be tested using standard tools, such as ancestral sequence reconstruction. It is always important to be able to quantitatively test hypotheses, and this is an important example of such a test in action.

In molecular phylogenetics, Markov models have been traditionally used to describe substitutions among DNA or protein sequences and therefore to reconstruct phylogenetic trees and understand evolutionary events. When selecting the "best" model for specific data, there is always a balance between the over-simplified models and over-fitted models. Over-simplified models often describe the evolutionary property with few parameters, but can lead to biased conclusions. In contrast, evolutionary models that use too many parameters may over-fit the data, which also result in errors for estimating a large number of parameters. So it is important to evaluate whether or not the data are well and adequately explained by the evolutionary models, and to identify the mis-fitting parts in the data as well. In addition, it is likely that more than one model can fit the data because they could well describe different evolutionary properties (i.e. substitution rate heterogeneity, or nucleotide compositional heterogeneity). Indeed, there are many complex evolutionary models that incorporate heterogeneity of the substitution process across sites and/or over time,

and these models are generally shown to fit the data better. We anticipate that a goodness-of-fit test between models and data should become a standard step in phylogenetic analyses, and that the use of more complex (well-fitted) models will significantly improve the accuracy of phylogenetic inference. Further, with the increase of genomic data, gene tree vs species tree incongruence is becoming even more obvious, implying that biological factors may lead to "incorrect" gene trees and blur the tree-like relationships. Such biological mechanisms may include incomplete lineage sorting, introgression/hybridisation, lateral gene transfer, gene duplication/loss, and natural selection. Beyond the tree-like inference, using phylogenetic network reconstruction to describe the evolutionary history is an alternative approach to visually present these potential biological events, and improve understanding of the differences between gene trees.

Future directions

Mossel and Steel (2004) have shown that under standard Markov models, primary sequences may lose their information at the very deepest times. In this case, sequences are highly divergent and methods designed to identify distant homologs based on similarity of extant sequences (e.g. hidden Markov models) may fail to conclusively prove or disprove patterns of homology. There are several ideas that may help remedy the difficulty of identifying the homologs with highly divergent sequences, and improve deep phylogenetic inferences.

(1). Using an approach based on the similarity of the three-dimensional (3D) structure of the protein is beneficial to identify homologs with low sequence similarity (Daly et al. 2013), although this approach may also identify sequences that are the results of convergent evolution.

(2). Use of a statistical test where the dissimilarity among extant sequences is compared against that observed among reconstructed ancestral sequences. This idea is based on our work (White et al. 2013, shown in Chapter 4), which shows that when going back in time the similarity of divergent proteins that shared a common ancestor is expected to increase. In contrast, for non-homologous sequences that independently converged on the same three-dimensional structure, similarity should not increase.

(3). "Alignment-free" methods. Because multiple sequence alignment (MSA) of highly divergent homologous sequences is known to be problematic, the "alignment-free" methods have been reported to address this issue, which estimate trees without

performing MSA. For example, Nelesen et al. (2012) presented a method for phylogenetic estimation from unaligned sequence using a divide-and-conquer approach, in which MSA was only conducted on subsets of closely related sequences. Recently, Chan and Ragan (2013) implemented another approach based on short sub-sequences (*k*-mers), which is fully independent of MSA.
These methods indeed need to further testing for robustness and accuracy.
(4). Gene content and gene order. Unlike classical sequence-based approaches, methods that are based on gene content and gene order do not depend on multiple sequence alignment. Gene content information might present good phylogenetic markers for deeper divergences. In this approach, genomic data are scanned and the presence or absence of genes/gene families in the data is summarized to generate the data matrix. The matrix can be analysed either using distance analyses (Blair et al. 2005; Vishnoi et al. 2010), parsimony method (Mirkin et al. 2003) or maximum-likelihood estimation (Huson and Steel 2004). Gene order is also recognized as a valuable phylogenetic character to reconstruct phylogenies (Belda et al. 2005). Genes are rearranged in the genome by evolutionary events (such as inversion and transposition), and analyses of these rare events have the potential to resolve ancient phylogenetic relationships (e.g. Belda et al. 2005; Luo et al. 2008).
(5). Rare genomic changes. Classical "rare genomic changes" include insertions and deletions (indels), retroposon (SINE and LINE), gene fission and gene fusion events. These characters have become an increasingly popular approach to estimate deep evolutionary relationships (Rokas and Holland 2000; Rogozin et al. 2009; Chernikova et al. 2011). It is premature to expect that rare genomic changes will definitively resolve all phylogenetic questions, but these characters are ideal as additional independent markers to test phylogenetic hypotheses. However, more efficient algorithms still need to be investigated to evaluate the accuracy of this approach.

In general, there has been little theoretical work on the ability of methods to recover deeper divergences, and this is a major gap in our knowledge at present. So although we cannot say that it is impossible to recover very deep phylogeny accurately, neither has it been shown that we can. In the future, we need to better understand deeper and deeper phylogeny beyond the limit of Markov models that were applied to primary sequences. We are now living in very exciting times, and the power of phylogenomics can be combined and integrated with many other aspects of biology to be able to study

a wide range of questions. Perhaps this has started with our recent studies (Zhong et al. 2013; 2014) on the closest relationship of streptophyte algae, which appear to be the "multicellular" lineage of algae (rather than the "coenocytic" lineage of the Charales) that led to the "multicellular" land plants. Here, we are using "multicellular" to refer to a single nucleus per small cell. Further, it would be very interesting to study the evolution of cell structure of streptophyte algae. Another direction is to understand the principle/mechanism behind mutation rate variation, such as the relationship between generations for a range of genome size, population size and mutation rates. Also Gibb et al. (2013) and Cutter (2013) are both good examples for integrating ecological and biogeographical questions with phylogenetic theory. Let the future begin – now!

References

Belda, E., Moya, A., and Silva, F.J. (2005). Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. *Mol. Biol. Evol*. 22: 1456-1467.

Blanquart, S., and Lartillot, N. (2008). A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol*. 25: 842-858.

Blair, J.E., Shah, P., and Hedges, S.B. (2005). Evolutionary sequence analysis of complete eukaryote genomes. *BMC Bioinformatics*. 6: 53.

Boussau, B., and Gouy, M. (2006). Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol*. 55: 756-768.

Burleigh, J.G., and Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *Am. J. Bot*. 91: 1599-1613.

Chan, C.X., and Ragan, M.A. (2013). Next-generation phylogenomics. *Biol. Direct*. 8: 3.

Chaw, S.M., Zharkikh, A., Sung, H.M., Lau, T.C., and Li, W.H. (1997). Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. *Mol. Biol. Evol*. 14: 56-68.

Chaw, S.M., Parkinson, C.L., Cheng, Y., Vincent, T.M., and Palmer, J.D. (2000). Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl Acad. Sci USA*. 97: 4086-4091.

Chernikova, D., Motamedi, S., Csuros, M., Koonin, E., and Rogozin, I. (2011). A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct*. 6: 26.

Cutter, A.D. (2013). Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. *Mol. Phylogenet. Evol*. 69: 1172-1185.

Daly, T.K., Sutherland-Smith, A.J., and Penny, D. (2013). Beyond BLASTing: Tertiary and quaternary structure analysis helps identify major vault proteins. *Genome Biol. Evol*. 5: 217-232.

Donoghue, M.J., and Doyle, J.A. (2000). Seed plant phylogeny: demise of the anthophyte hypothesis. *Curr. Biol*. 10: R106-R109.

Doyle, J.A. (2006). Seed ferns and the origin of angiosperms. *J. Torrey Bot. Soc*. 133: 169-209.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., and Glenn, T.C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol*. 61: 717-726.

Finet, C., Timme, R.E., Delwiche, C.F., and Marlétaz, F. (2012). Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr. Biol*. 22: 1456-1457.

Gibb, G.C., Kennedy, M., and Penny, D. (2013). Beyond phylogeny: pelecaniform and ciconiiform birds, and long-term niche stability. *Mol. Phylogenet. Evol*. 68: 229-238.

Goremykin, V.V., Nikoforova, S.V., and Bininda-Emonds, O.P.P. (2010). Automated removal of noisy data in phylogenomic analyses. *J. Mol. Evol*. 71: 319-331.

Huson, D.H., and Steel, M. (2004). Phylogenetic trees based on gene content. *Bioinformatics*. 20: 2044-2049.

Lartillot, N., Lepage, T., and Blanquart, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*. 25: 2286-2288.

Laurin-Lemay, S., Brinkmann, H., and Philippe, H. (2012). Origin of land plants revisited in the light of sequence contamination and missing data. *Curr. Biol*. 22: R593-R594.

Lemmon, A.R., Emme, S.A., and Lemmon, E.M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol*. 61: 727-744.

Liu, L., Yu, L., and Edwards, S.V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol*. 10: 302.

Luo, H., Shi, J., Arndt, W., Tang, J., and Friedman, R. (2008). Gene order phylogeny of the genus *Prochlorococcus*. *PLoS One*. 3: e3837.

Mathews, S. (2009). Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. *Am. J. Bot*. 96: 228-236.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y., and Koonin, E.V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol*. 3: 2.

Mossel, E., and Steel, M. (2004). A phase transition for a random cluster model on phylogenetic trees. *Math. BioSci*. 187: 189-203.

Nelesen, S., Liu, K., Wang, L.S., Linder, C.R., and Warnow, T. (2012). DACTAL: divide-and-conquer trees (almost) without alignments. *Bioinformatics*. 28: i274-i282.

Nickrent, D.L., Parkinson, C.L., Palmer, J.D., and Duff, R.J. (2000). Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol*. 17: 1885-1895.

Pagel, M., and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol*. 53: 571-581.

Penny, D. (2011). Darwin's theory of descent with modification, versus the biblical Tree of Life. *PLoS Biol*. 9: e1001096.

Rogozin, I.B., Basu, M.K., Csuros, M., and Koonin, E.V. (2009). Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome Biol. Evol*. 25: 99-113.

Rokas, A., and Holland, P.W. (2000). Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol*. 15: 454-459.

Timme, R.E., Bachvaroff, T.R., and Delwiche, C.F. (2012). Broad phylogenomic sampling and the sister lineage of land plants. *PLoS One*. 7: e29696.

Vishnoi, A., Roy, R., Prasad, H.K., and Bhattacharya, A. (2010). Anchor-based whole genome phylogeny (ABWGP): a tool for inferring evolutionary relationship among closely related microorganisms. *PLoS One*. 5: e14159.

White, T., Zhong, B., and Penny, D. (2013). Beyond reasonable doubt: evolution from DNA sequences. *PLoS One*. 8: e69924.

Wodniok, S., Brinkmann, H., Glöckner, G., Heidel, A.J., Philippe, H., Melkonian, M., and Becker, B. (2011). Origin of land plants: Do conjugating green algae hold the key? *BMC Evol. Biol*. 11: 104.

Zhong, B., Yonezawa, T., Zhong, Y., and Hasegawa, M. (2010). The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol. Biol. Evol*. 27: 2855-2863.

Zhong, B., Liu, L.,Yang, Z., and Penny, D. (2013). Origin of land plants using the multispecies coalescent model. *Trends Plant Sci*. 18: 492-495.

Zhong, B., Xi, Z., Goremykin, V.V., Fong, R., McLenachan, P.A., Novis, P., and Penny, D. (2014). Origin of land plants revisited using heterogeneous models and three new algal chloroplast genomes. *Mol. Biol. Evol*. 31: 177-183.

Appendix 1.

Goremykin,V.V., Nikiforova, S.V., Biggs, P.J. **Zhong, B.**, DeLange, P., Martin,W., Woetzel, S., Atherton, R.A., McLenachan, T., and Lockhart P.J. (2013). The evolutionary root of flowering plants. ***Systematic Biology***. 62: 51-62.

In this paper the *Trithuria inconspicua* chloroplast genome was sequenced, and the evolutionary root of flowering plants is thoroughly investigated by the removal of different systematic errors. After improving the fit between the sequence data and substitution models, the phylogenetic analyses supported that *Trithuria*, Nymphaeaceae, and *Amborella* are the surviving relatives of the most basal lineage of flowering plants.

I was responsible for investigating base compositional heterogeneity, and conducting the goodness-of-fit analyses. I also contributed to the writing of the manuscript.

Syst. Biol. 62(1):50–61, 2013

DOI:10.1093/sysbio/sys070
Advance Access publication on July 31, 2012

The Evolutionary Root of Flowering Plants

VADIM V. GOREMYKIN[1]*, SVETLANA V. NIKIFOROVA[1], PATRICK J. BIGGS[2], BOJIAN ZHONG[3], PETER DELANGE[4], WILLIAM MARTIN[5], STEFAN WOETZEL[6], ROBIN A. ATHERTON[3], PATRICIA A. MCLENACHAN[3], AND PETER J. LOCKHART[3,7]

[1]*IASMA Research Center, Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy;* [2]*Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand;* [3]*Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand;* [4]*Department of Conservation, Auckland Conservancy, New Zealand;* [5]*Institut für Botanik III Heinrich-Heine-Universität, Germany;* [6]*Max Planck Institute for Plant Breeding Research, Cologne, Germany; and* [7]*Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand*
**Correspondence to be sent to: IASMA Research Center, Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy;*
E-mail: Vadim.Goremykin@iasma.it.

Received 14 December 2011; reviews returned 21 February 2012; accepted 26 July 2012
Associate Editor: Lars Jermiin

Abstract.—Correct rooting of the angiosperm radiation is both challenging and necessary for understanding the origins and evolution of physiological and phenotypic traits in flowering plants. The problem is known to be difficult due to the large genetic distance separating flowering plants from other seed plants and the sparse taxon sampling among basal angiosperms. Here, we provide further evidence for concern over substitution model misspecification in analyses of chloroplast DNA sequences. We show that support for *Amborella* as the sole representative of the most basal angiosperm lineage is founded on sequence site patterns poorly described by time-reversible substitution models. Improving the fit between sequence data and substitution model identifies *Trithuria*, Nymphaeaceae, and *Amborella* as surviving relatives of the most basal lineage of flowering plants. This finding indicates that aquatic and herbaceous species dominate the earliest extant lineage of flowering plants. [*Trithuria inconspicua*; chloroplast genome; angiosperm origins; heterotachy; base compositional heterogeneity; data model fit.]

Although there is increasing consensus about many relationships among major lineages of flowering plants (Soltis et al. 2011) and convergence toward more similar dates for the origin of angiosperms (Jiao et al. 2011; Sun et al. 2011), determining the root of the angiosperm phylogeny has been more problematic. This difficulty is not unique to the study of angiosperms; reconstructing basal relationships in species radiations is known to be hard (Shavit et al. 2007; Graham and Iles 2009). Not only can the shape of the true underlying phylogeny make it difficult to accurately reconstruct gene trees (Whitfield and Lockhart 2007), even correct gene trees can be incongruent with the underlying species phylogeny (Degnan and Rosenberg 2009).

In phylogenetic studies of chloroplast DNA (cpDNA), nuclear DNA (nuDNA), and mitochondrial DNA (mtDNA), *Amborella* has often been recovered as the sole survivor of the first lineage to diverge from that leading to all the other extant flowering plants (Mathews and Donoghue 1999; Qiu et al. 1999; Soltis and Soltis 2004; Stefanović et al. 2004; Leebens-Mack et al. 2005; Jansen et al. 2007; Saarela et al. 2007; Graham and Iles 2009; Soltis et al. 2011). However, a closer relationship between *Amborella* and aquatic angiosperm species has been reported in analyses of mitochondrial and nuclear DNA (Qiu et al. 2010; Jiao et al. 2011; Soltis et al. 2011;) as well as in model-based analyses of chloroplast genes that typically exclude or reduce the impact of third codon positions (Barkman et al. 2000; Wu et al. 2007). Opinion has been divided over how to treat third codon positions in cpDNA. Although inclusion of these sites might improve phylogenetic resolution between some taxa (Zanis et al. 2002; Stefanović et al. 2004; Leebens-Mack et al. 2005), they also exhibit evidence of a decayed historical signal (due to multiple substitutions at the same site) between some taxa (Goremykin et al. 2003; Chaw et al. 2004). Analyses of short independent nuclear markers have not provided improved phylogenetic resolution, suggesting instead alternative relationships among basal angiosperms (e.g., Mathews and Donoghue 1999; Jiao et al. 2011; Soltis et al. 2011). This finding is perhaps not unexpected given the short internal branches typically reconstructed for angiosperm phylogenies (e.g., see Martin et al. 2005).

We have previously suggested that a poor fit between commonly used phylogenetic models and sequence data contributes to uncertainty concerning relationships among early diverging lineages of flowering plants (Lockhart and Penny 2005; Martin et al. 2005). Here, we provide further evidence for this hypothesis in a study of the substitution properties of concatenated chloroplast genome sequences, and in particular of sites in the alignment that are most varied. These sites, often called "fast sites," show the greatest character state variation as well as evidence of multiple substitutions. Numerous methods for sorting, identifying, and removing fast sites have been suggested, and the impact of removal of the fastest evolving sites on phylogenetic reconstruction is well known (e.g., Brinkmann and Philippe 1999; Hirt et al. 1999; Lopez et al. 1999; Ruiz-Trillo et al. 1999; Hansmann and Martin 2000; Burleigh and Mathews 2004; Pisani 2004). Less well appreciated is the observation that the sorting of sites based on character state variation or compatibility criteria allows the properties of sites that impact on tree building to be more easily studied (Sperling et al. 2009). We have examined the compositional heterogeneity of fast sites and the fit of concatenated chloroplast sequences to

the GTR+I+Γ substitution model commonly used in angiosperm phylogeny studies. We address the problem of identifying which of the fast sites to exclude from the phylogenetic data by applying the GNB criterion (named after the inventors: Goremykin et al. 2010) to the concatenated alignment after the sites in this alignment had been reordered according to their observed variability (OV; see "Materials and Methods"). This criterion has been suggested as suitable for identifying sites most affected by multiple substitutions in a multiple sequence alignment. Here, we examine the properties of the fast sites identified under the GNB criterion and the contribution of these sites to topological distortion in phylogenetic trees reconstructed for angiosperm and conifer sequences. To obtain optimal phylogenetic estimates, we employed the CAT+covarion model, which was consistently identified in our cross-validation analyses as the best-fitting model to our original data and to data partitions generated in the "noise reduction" protocol of Goremykin et al. (2010). This substitution model better accommodates a restricted substitution profile across sites and describes spatial heterogeneity of substitutions in terms of simple covarion models (Ane et al. 2005).

To improve taxon sampling at the base of the angiosperm radiation, we also sequenced the chloroplast genome of *Trithuria inconspicua*, a species from a genus of minute aquatic herbs, which recently has been found to be closely related to Nymphaeaceae (Saarela et al. 2007). Our findings highlight the importance of the fit between model and data when evaluating relationships among basal angiosperms.

Materials and Methods

Sequencing of the Chloroplast Genome of Trithuria inconspicua

Trithuria inconspicua was collected from the Kai Iwi Lakes (Lakes Waikare and Taharoa), Northland, North Island, New Zealand, and sent by courier to Massey University, Palmerston North. Voucher specimens have been deposited at the Auckland War Memorial Museum Herbarium AK (see AK 308938, AK 320388). Enriched cpDNA was sequenced on an Illumina GAII platform as described in Atherton et al. (2010). Contigs were assembled using Velvet version 0.7.60 (Zerbino and Birney 2008) and odd kmer values ranging from 25 to 61. Because the copy number of cpDNA was higher than that for the nuDNA (though not a higher absolute amount), coverage cutoffs of 10, 20, 40, and 80 were applied during the assembly of contigs. Staden 2.0.0b7 (http://staden.sourceforge.net/) was used to join the contigs generated by Velvet. Nine gaps remained after the assembly; 8 gaps were closed by designing primers to flanking regions and sequencing the missing parts using standard ABI3730 sequencing protocols (Massey Genome Service http://genome.massey.ac.nz/).

Taxon Selection and Multiple Sequence Alignment

Protein-coding sequences of 61 genes common to 31 chloroplast genomes from angiosperms and gymnosperms were downloaded from GenBank. NAD dehydrogenase genes were not included in analyses as these are absent from the cpDNAs of gnetophytes and conifers (Wakasugi et al. 1994; Braukmann et al. 2009). In our taxon sampling, we included representatives of all available basal angiosperm lineages but not all crown group angiosperm species for which chloroplast genomes have been determined. This taxon selection retained species most important for inferring relationships among basal angiosperms and reduced computation time for model-fitting and tree-building analyses on a 16-core Linux server. Eudicots were represented by 6 basal species. We excluded grasses, known to be subtended by a very long branch in previous analyses (Goremykin et al. 2005), keeping all other monocots.

As concern over alignment procedures remains an important practical consideration for phylogenomic analyses (Philippe et al. 2011), multiple sequence alignments were generated using 2 alignment protocols in the present study. The first protocol, used as a basis for figures shown in this article, uses the same principles described in Goremykin et al. (2004). This alignment protocol provides a rapid and reliable method of aligning similar gene sequences and for producing data sets comprising first and second codon positions and all 3 codon positions. With this approach, gene sequences were sorted into 61 Fasta files, each containing orthologs. For each file, first and second codon positions were aligned using the program MUSCLE (Edgar 2004). Alignments for sequences that included all 3 codon positions were also generated by the same script. The resulting 122 alignment files were each manually edited, such that regions of low similarity between the ingroup and outgroup sequences were discarded. Individual gene alignment files were concatenated using Geneious v5.5.4. (Drummond et al. 2010) to produce: (i) a gapped alignment of 40 553 positions in length, provided as a supplementary material (Supplementary File S1) and (ii) an alignment of first and the second codon positions 25 246 positions in length (Supplementary File S2). An OV sorted (see below) version of the 40 553 pos. long alignment has been provided as a Supplementary File S3.

A MUSCLE alignment of translated nucleotide sequences from 56 individual Fasta files was also generated and used to confirm results of phylogenetic analyses obtained using the first alignment protocol. This second alignment approach used the same principles as previously implemented for obtaining conservative alignments between anciently diverged sequences (Lockhart et al. 1996). With this method, we imported each Fasta file into MEGA 5.0 (Tamura et al. 2011), translated the sequences, and then aligned them with MUSCLE (default options). We concatenated these aligned files using Geneious v5.5.4. (Drummond et al. 2010) and then imported the concatenated file into Se-Al.

v2.0a11. (Rambaut 2002). Site patterns adjacent to indels were then removed if they did not contain amino acids with similar physical/chemical properties as specified in Se-Al. Finally, the columns with gaps were removed and the sequences back-translated. This alignment protocol produced a much shorter concatenated alignment than did the first method (31 674 ungapped positions). This alignment has been provided as a supplementary material (Supplementary File S4).

OV Sorting and "Noise Reduction"

Site patterns in our concatenated alignments were reordered according to their OV scores and data partitions identified for tree building using the GNB criterion (Goremykin et al. 2010). Previously, this approach was found effective in the recovery of benchmark clades of mammalian phylogeny, and more effective than other methods in identifying fast-evolving sites that cause long branch attraction (LBA) artifacts (Goremykin et al. 2010).

OV sorting involves calculating a sum-of-pairs mismatch score for each site in the full alignment (including positions with gaps) and then ordering the sites according to the OV scores (Goremykin et al. 2010). This produces an alignment with the most conserved (least varied) site patterns at one end, and the least conserved (most varied) positions at the other end. We refer to this alignment as the OV alignment. The OV alignment was generated using the script Sorter.pl. This script also splits the OV alignment into several bipartitions of sites. Each bipartition contains an "A" partition, which includes site patterns from the conserved end of the alignment, and a "B" partition, which includes site patterns from the least conserved end of the OV alignment. In the present study, the bipartition of sites into partitions A and B occurred at position $i \times 250$ (where $i = 1, 2, 3, \ldots$) upstream from the most varied end of the OV alignment. The incremental increase in interval length of 250 sites for the B partition is an arbitrary size previously found suitable for monitoring change in the properties of the ordered sites at the most varied end of the OV alignment. Once the bipartitions are formed, the script Sorter.pl calls ModelTest (Posada and Crandall 1998) to identify an optimal time-reversible substitution model for each of the A and B partitions using a 2-step procedure (for further details, see Goremykin et al. 2010). The script then calls PAUP* (Swofford 2002; Unix v. 4.0b10) to calculate a matrix of maximum likelihood (ML) distances for the A and B partitions. A matrix of *p*-distances (number of sites with observed differences/total number of sites) is also calculated for each B partition.

Sorter.pl also calculates the average of the ML-distances minus the average of the *p*-distances and reports this mean deviation of the ML- and *p*-distances for the B partitions, and Pearson correlation coefficient values (r) between these estimates (Goremykin et al. 2010). Dissimilarity between relative ranking of ML- and *p*-distances calculated from the B partitions occurs if distance estimates are not similar between taxa. Stochastic error associated with the short sequence length of the initial B partitions will cause such dissimilarity, as will substitution model violations and saturation with multiple substitutions. By monitoring the r values as the length of the B partition is increased, it is possible to identify a point of transition with respect to the similarity of the distances compared. As the relative ranking of absolute distance values within 2 groups of distance estimates (*p*- and ML-distances) becomes similar, there is a dramatic rise in the value of r.

In addition to comparing the ML- and *p*-distances for B partitions, the script Sorter.pl also compares optimal ML-distances for the A and B partitions. Deviation is again measured in terms of r. As with the ML- and *p*-distance comparison for the B partition, a dramatic rise in r occurs when the distances become proportional, and their ranking becomes similar. The comparison identifies the relative length of the A and B partitions, at which point the evolutionary properties of the B partition become similar to those of the A partition.

Goremykin et al. (2010) have suggested that the site stripping process should cease when there is a dramatic increase in the value of r in both correlation analyses. At this point, positions added from the conserved A partition to the variable B partition clearly start to mask the nonphylogenetic signal associated with the most varied positions in the B partitions. Here, we also report that the topological distortion induced by the presence of B partition sites is also greatly reduced at this point. Model misspecification contributed by compositional heterogeneity, as we also show, still persists beyond this point. However, this has little impact on the relative ranking of distances in B partitions. Thus, further character removal is not justified on the basis of the GNB criterion.

As demonstrated in Zhong et al. (2011), the GNB criterion also identifies and provides a basis for removing sites from a concatenated alignment that have a poor fit to phylogenetic model assumptions. Although this criterion does not remove all model-violating sites from the data, it has been shown to remove sites that significantly impact on phylogenetic estimates, and thus sites that have significant effect in misleading tree building. In particular, it appears very useful for reducing LBA artifacts in phylogenetic reconstruction. This was demonstrated in reanalysis of mitochondrial DNA sequences, which previously and consistently had yielded a rodent polyphyly artifact (Goremykin et al. 2010) and also in recent analyses of chloroplast sequences from Gnetales and other seed plants (Zhong et al. 2011).

To study the relationship between changes in r and branch length support in reconstructed trees, splits can be calculated for individual A and B partitions. We calculated NeighborNet (NNET: Bryant and Moulton 2004) splits from the optimal ML-distances obtained for each B partition generated during the noise reduction protocol. These were calculated using SplitsTree 4.0 (Huson and Bryant 2006). Of particular interest are the

splits that separate outgroup and ingroup taxa as these are relevant for the question of rooting the angiosperm radiation. In the present study, we plotted the relative size of the splits separating: (i) angiosperms from gymnosperms and (ii) Gnetales from other species. Such a "heterotachy plot," as it was referred to in Zhong et al. (2011), allows visualization of the relationship between B-partition distances and any topological distortion (Lockhart et al. 1996; Bruno and Halpern 1999) of reconstructed trees due to including the most varied sites of the OV alignment when tree building.

Base Composition Heterogeneity

Base compositional heterogeneity (Lockhart et al. 1992; Jermiin et al. 2004) was examined over the most varied end of the OV alignment. To investigate this, intervals of sites with the same length (360 jacknife resampled ungapped positions; 3 replicates for each interval) were sampled from nonoverlapping locations at the most varied end of the OV alignment (between 0 and 500 sites, 500 and 1000 sites, 1000 and 1500 sites, … , 9500 and 10 000 sites). We examined each of these sets of sites using Bowker's matched-pair symmetry test (Ababneh et al. 2006), as implemented in Seq-Vis (Ho et al. 2006). We used Seqboot from the PHYLIP v3.69 (Felsenstein 2004) package for jackknife resampling of sites (sampling without replacement) and SeqVis v1.5 (Ho et al. 2006) for the symmetry test. The smallest interval from which sites were resampled was the first interval: 0–500 sites (these 500 gapped positions contained 380 ungapped positions).

Goodness of Fit Analyses

We used MISFITS (Nguyen et al. 2011) and Tree-Puzzle-5.2 (Schmidt et al. 2002) to identify those site patterns in the OV alignment whose observed frequencies were unexpected under a GTR+I+Γ substitution model. This model was identified as the best-fitting model to the OV alignment among all models that assumed a single matrix of base frequencies. This was also the case for the increasingly short A partitions according to a double-fitting procedure that employed an Akaike information criterion (AIC) (described in Goremykin et al. 2010). The fit of the GTR+I+Γ model to chloroplast data sets is also of significant interest as this model has been commonly used in phylogenetic analyses of basal angiosperms (e.g., Barkman et al. 2000; Zanis et al. 2002; Stefanović et al. 2004; Leebens-Mack et al. 2005; Saarela et al. 2007; Wu et al. 2007; Graham and Iles 2009; Qiu et al. 2010; Jiao et al. 2011; Soltis et al. 2011). Although there are computational issues with coestimation of the I+Γ parameter values (e.g., see "Discussion" in Yang 2006), this model has been found to have higher reconstruction accuracy than GTR+Γ models in more biologically realistic simulations (Gruenheit et al. 2008). The impact that deletion of sites from the most varied end of the OV alignment had on the fit of this substitution model was also studied at different shortening steps. Log-likelihood scores for the evolutionary model obtained for the increasingly short A partitions were also compared with the log-likelihood scores for equal length partitions that were jackknife resampled from the complete OV alignment. We used Seqboot for jackknife resampling and PhyML 3.0 (Guindon et al. 2010) for calculating log-likelihood scores.

Substitution Model Selection for A Partitions

The optimal substitution model was determined for the A partition data sets using cross-validation as implemented in PhyloBayes 3.2e (Lartillot and Philippe 2004). To determine the length of time needed for convergence of posterior probabilities, we initially ran PhyloBayes on a 16-core Linux server for at least 2 weeks with alignments of the first and second codon positions, and of all 3 codon positions, choosing between 6 substitution models for each input file: The "classical" GTR+Γ, GTR+Γ+covarion, GTR+Γ+covext, GTR+Γ+CAT, GTR+Γ+CAT+covarion, and GTR+Γ+CAT+covext (Lartillot and Philippe 2004). Here, "CAT" refers to the site-heterogeneous mixture model of Lartillot and Philippe (2004), "covarion" to the covarion model of Tuffley and Steel (1998), and "covext" to a variant of the Tuffley and Steel model that allows for variation in rate heterogeneity across sites. We assumed a four-category discrete Γ distribution in modeling rate-heterogeneity across sites. From these initial 12 runs, we determined that 200 cycles were sufficient for convergence on our Linux server. Since cross-validation is multistaged and computationally demanding, we wrote a script Cross.pl, which initiates parallel multiple PhyloBayes and cross-validation runs. This script first invokes PhyloBayes, lets it run for 1000 cycles under the abovementioned models, and builds consensus trees discarding the first 200 cycles as burn-in. Then the script invokes the PhyloBayes program cvrep to randomly sample 10 learning and 10 test data partitions from each alignment, so that each learning data partition has 90% of the input alignment length and each test partition has 10% of the input alignment length. The script then calls PhyloBayes and performs Markov chain Monte Carlo sampling for 200 cycles in parallel for the learning sets created by the PhyloBayes program cvrep. Subsequently, the script initiates the Phylo Bayes program readcv in parallel for all data replicates and computes a cross-validation score (i.e., calculates the likelihood under the test set, averaged over the posterior distribution of the learning set) discarding a burn-in of 50 sampling points and taking every point thereafter. Finally, the script invokes the PhyloBayes program sumcv to compute summary statistics. Using the AIC for the double-fitting procedure, the GTR+I+Γ model was selected as the best-fitting model among those with one matrix of base frequencies for the OV alignment and its next 20 shortened subsets.

Tree Building

Phylogenetic reconstructions were performed using PhyloBayes and the PAUP*-embedded scripts in Sorter.pl (Goremykin et al. 2010). RAxML (Stamatakis et al. 2005) was also used to reanalyze a recently published data set of chloroplast, mitochondrial, and nuclear genes (Soltis et al. 2011).

Availability of scripts and of Trithuria chloroplast genome sequence.—Scripts not already publically available and used in this study have been provided as supplementary material. The sequence for the *Trithuria inconspicua* chloroplast genome determined in this study has been deposited with EMBL (Accession no. HE963749).

Results

Alignments

Two alignments were obtained using different approaches in the present study. Despite differences in their lengths, both methods produced very similar alignments. This can be visualized by comparing split networks that display the NNET split systems (*p*-distances) for each alignment (Supplementary File S5). Similar analytical results were obtained for both alignments. The figures shown in subsequent sections were based on the alignment method of Goremykin et al. (2003).

GNB Analyses

A significant improvement in r occurred after 8 steps: 2000 sites (Fig. 1a); that is, once the 2000 most varied sites were included in the B partition, *p*-distances and ML-distances for the B partition had become highly correlated. Similarly, at this shortening step ML-distances for A and B partitions also became highly correlated (Fig 1b), indicating similar evolutionary distances for both partitions, and suggesting a point had been reached at which further removal of sites from the A partition was no longer justified. Most significantly, the distance between outgroup and ingroup taxa reduced dramatically by the eighth sampling step. This was visualized in Figure 2, which shows the relative length of outgroup splits in the NNET split system for the taxon set. The extreme branch length separating the outgroup and ingroup sequences is a property of the 2000 sites at the most varied end of the OV alignment.

Compositional Heterogeneity

It has been previously observed that compositional heterogeneity and the rate of substitutions of sites are tightly correlated (Rodriguez-Ezpeleta et al. 2007). Our analyses provide some support for this observation. Figure 3 indicates that compositional heterogeneity is a feature of the most varied end of the OV alignment. In particular, it indicates the number of pairs failing a matched-pairs test of symmetry at $P < 0.00005$ when these are calculated on identical length partitions (360 sites each) sampled within 500-bp nonoverlapping gaped intervals at the most varied end of the OV alignment. The plot suggests that heterogeneity in composition is most significant over the first 3000–3500 most varied positions of this alignment. This heterogeneity is most significant between angiosperm and outgroup sequences and among outgroups sequences (values for individual pairs not shown). It extends past the stopping point identified by the GNB method. Hence, although compositional heterogeneity is likely to contribute to the extreme branch length difference between ingroup and outgroup sequences, it does not appear to explain the extreme branch length differences over the first 2000 most varied positions in the OV alignment.

Fit of Data to a GTR+I+Γ Substitution Model

The effect of removal of the most varied sites on the fit of the aligned data to a GTR+I+Γ substitution model was investigated. Table 1 reports log-likelihood scores for 2 tree models (*Amborella* most basal; *Amborella*+*Trithuria*+Nymphaeaceae most basal) on A partitions generated by the script Sorter.pl. These scores were compared against the log-likelihood scores for data sets identical in length to the shortened A partitions that were jackknife resampled from the OV alignment. They were always significantly better than the scores for the randomly resampled data, indicating that the sites removed by OV noise reduction significantly contribute to the poor fit between the evolutionary models and the aligned sequence data.

Assuming the same evolution models as examined in Table 1, MISFITS and Tree-Puzzle were used to the identify site patterns whose relative frequencies are over- and underrepresented in the OV alignment. Figure 4 plots the position of unexpected site patterns in the ungapped OV alignment. The height of each bar in the histogram indicates the number of consecutive sites at which unexpected site patterns occur. The most varied end of the OV alignment is identified as containing many site patterns that contribute to the poor fit of the GTR+I+Γ substitution model.

Tree Building

Phylogenetic trees were built from the OV alignment for the different length A partitions generated by the Sorter.pl script. This was done both for a CAT+GTR+Γ+covext model and for a GTR+I+Γ model. The former was found under cross-validation to be optimal for: (i) the full-length OV alignment, (ii) the alignment of the first and the second codon positions, and (iii) the alignment of the most conserved 38 553 positions in the OV alignment. The optimal tree reconstructed with a CAT+GTR+Γ+covext model on the A partition at the GNB stopping point is

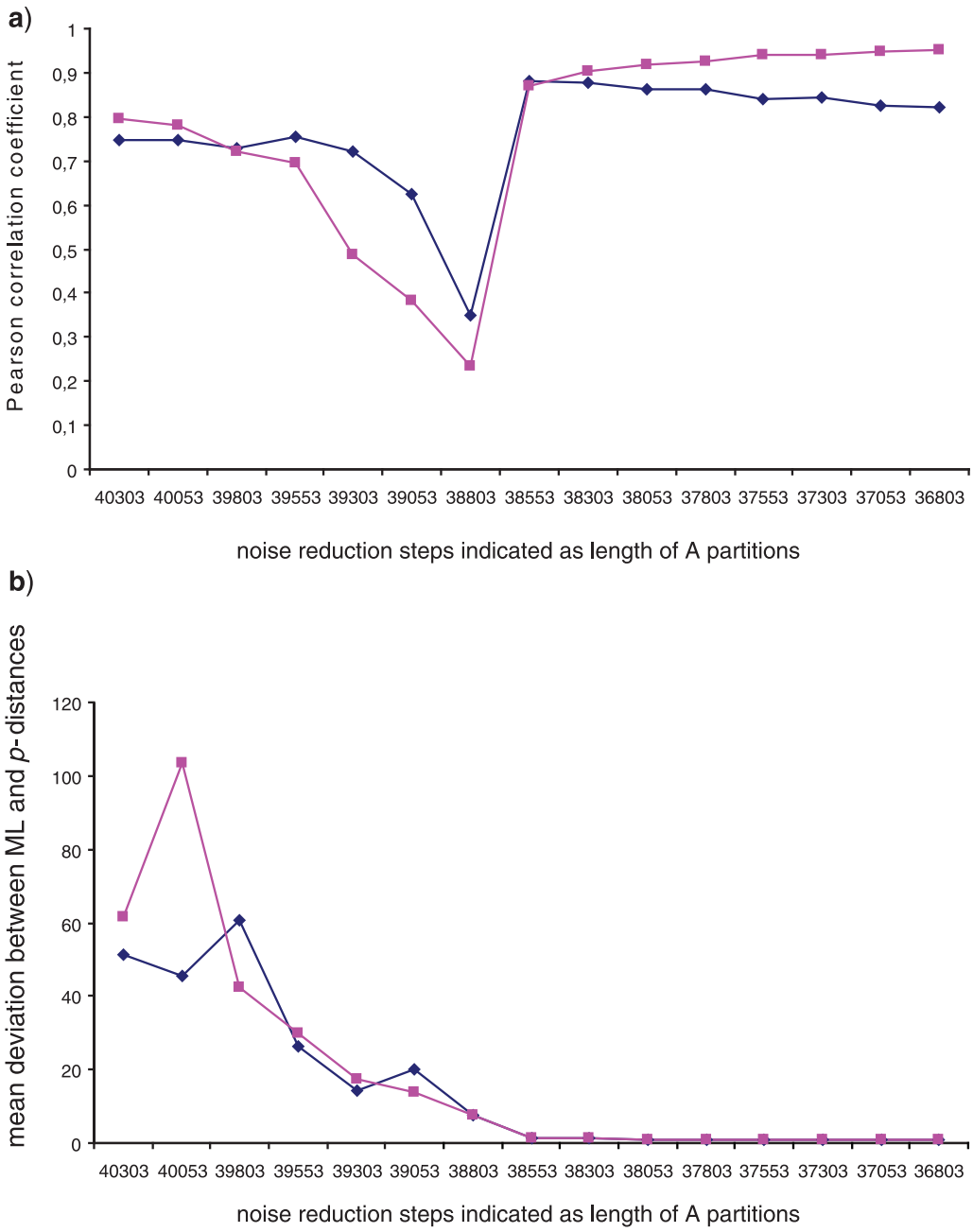


FIGURE 1. a) Plot showing results of the correlation analyses. The blue line indicates Pearson correlation coefficient values (r) obtained for pair-wise comparisons of ML-distances calculated from the A and B partitions whose combined length was 40 553 gapped positions in the OV alignment. The pink line indicates r values obtained for pair-wise comparisons of p-distances and ML-distances calculated for B partitions, discarded at each shortening step. At the 8th shortening step, when the A partition is 38 553 gapped positions in length, it passes both correlation tests (Goremykin et al. 2010). b) Plot showing mean deviation between ML- and p-distances calculated for B partitions at each shortening step. In calculating ML-distances, the best-fitting ML model for each partition length was first determined under an AIC using ModelTest (Posada and Crandall 1998). The pink line indicates results from analyses using a Neighbor-Joining tree to fit ML model parameters. The blue line indicates results obtained when an ML tree is used to fit substitution model parameters. This ML tree was computed using settings of the best-fitted model determined by the standard ModelTest procedure employing AIC.

shown in Figure 5. This tree indicates the same relationships among basal angiosperms as does the GTR+I+Γ tree reconstructed on the A partition at the GNB stopping point. Both reconstructions identify a lineage comprising *Amborella*+*Trithuria*+ Nymphaeaceae as most basal in the angiosperm radiation. Figure 6a indicates relationships inferred when a CAT+GTR+Γ+covext model is used to analyze the full-length (40 553 site) concatenated data set. With this data set, *Amborella* is inferred to be the most basal lineage in the radiation of angiosperms. Trees built from the alignment of first and second codon positions using

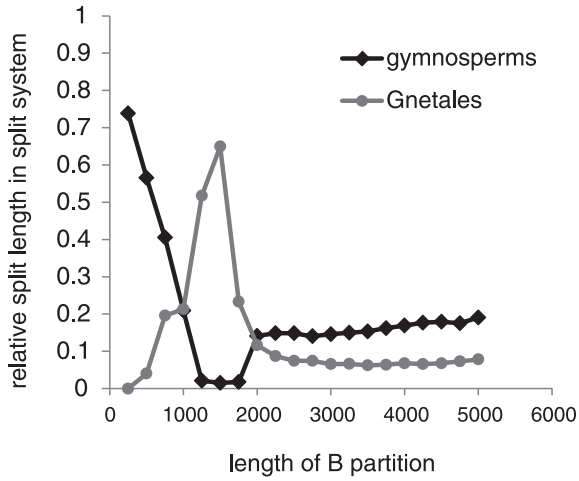


FIGURE 2. Plot showing the relative size of NNET splits separating: (i) angiosperms from gymnosperms and (ii) Gnetales from other taxa. The NNET splits were calculated from the optimal distances estimated for each B partition formed at the most varied end of the OV alignment.

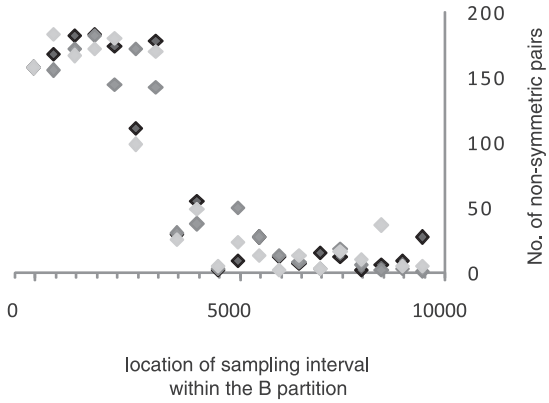


FIGURE 3. The number of pairwise distances (645 comparisons) failing a matched-pairs test of symmetry at $P < 0.00005$ was determined for equal length, nonoverlapping intervals at the most varied end of the OV alignment. For these estimates, we analyzed only ungapped sites (360 positions: 3 replicates per estimate) randomly sampled without replacement from 500-bp nonoverlapping gapped partitions at the most varied end of the OV alignment ("C" partitions in Goremykin et al. 2010).

the best-fitting CAT+GTR+Γ+covext model (Fig. 6b) show *Amborella*+*Trithuria*+Nymphaeaceae as the most basal lineage. Substitution models rejected in cross-validation supported the tree with the most basal branch subtending *Trithuria*+Nymphaeaceae (e.g., GTR+Γ model, Fig. 6c) based on the first and second position data set.

The support for relationships among basal angiosperms under a CAT+GTR+Γ+covext covarion model was also investigated after each shortening step of 250 positions in the alignment of all codon positions. The results are shown in Figure 7. These indicate that (i) support for *Amborella* joining with the outgroup occurs only when the most varied positions of the alignment are included, (ii) the grouping of *Amborella*+*Trithuria*+Nymphaeaceae is strongly favored as the most basal lineage after removal of 1750 sites and remains supported until 2500 sites are removed, and (iii) a basal grouping of *Amborella*+*Trithuria*+Nymphaeaceae+*Illicium* becomes favored after removal of 2750 sites. Note that under the CAT+GTR+Γ+covext model, support for *Amborella*+*Trithuria*+Nymphaeaceae as a most basal clade is realized prior to the GNB stopping point, which might indicate a better fit of this substitution model to the data.

DISCUSSION

Our findings reported here, and those in recent analyses of other seed plants (Zhong et al. 2011), reemphasize the importance of considering the fit of time-reversible models to the fast-evolving sites in sequence alignments (Sullivan et al. 1995). We show that site sorting can facilitate studies of the substitution properties of concatenated gene alignments and help to identify site patterns relevant to substitution model misspecification and potential tree-building artifacts. The sites providing most support for the *Amborella* most basal hypothesis are characterized by poor fit between model and data and by evolutionary properties that induce extreme topological distortion in reconstructed trees. The GNB stopping criterion removes many of these sites (38% of the removed sites did not fit an *Amborella* basal+GTR+I+Γ model; 39% of the removed sites did not fit an *Amborella*+*Trithuria*+Nymphaeaceae basal+GTR+I+Γ model).

In the present study, when sites causing topological distortion were removed, reconstruction under the optimal CAT model and GTR+I+Γ model favors a tree indicating *Amborella*+*Trithuria*+Nymphaeaceae as the most basal hypothesis. Although compositional heterogeneity will contribute to topological distortion when time-reversible Markov models are used in analysis of the data, our heterotachy and matched-pairs test of symmetry plots suggest that compositional heterogeneity is alone insufficient to explain the different topologies obtained during tree building with different A partitions. In general, the impact of compositional heterogeneity needs to be evaluated in the context of the extent of divergence between sequences exhibiting this heterogeneity (Jermiin et al. 2004) and the spatial pattern of sites free to vary in the sequences (Lockhart et al. 2006).

We propose that our analyses and observations provide a basis for understanding the discrepancy among recent findings from phylogenetic analyses of cpDNA and mtDNA concerning the rooting of the angiosperm phylogeny. Our reconstructed phylogeny (Fig. 5) obtained after exclusion of a large number of model-violating sites is consistent with that recently obtained in analyses of nuclear EST amino acid sequences that also implemented a CAT model. In this case, although *Trithuria* was not available for study,

TABLE 1. Data-model fit after removal of 500, 1000, 1500, and 2000 sites

Tree model	*Amborella*+Nymphaeaceae+*Ttithuria* most basal			
Number of sites retained	40053	39553	39053	38553
Mean log-likelihood values from jackknife samples	−332368.96	−328151.34	−323995.76	−319864.78
Log-likelihood values of shortened OV alignment	−321728.05	−307453.85	−294354.23	−282625.00
SD of jackknife samples	181.75	259.90	318.15	365.44
z-score	58.55	79.64	93.17	101.91
Tree model	***Amborella* most basal**			
Number of sites retained	40053	39553	39053	38553
Mean log-likelihood values from jackknife samples	−332328.03	−328110.72	−323955.46	−319825.17
Log-likelihood values of shortened OV alignment	−321708.11	−307439.72	−294347.58	−282630.57
SD of jackknife samples	181.64	260.30	318.54	365.80
z-score	58.47	79.41	92.95	101.68

The *z*-score is the difference between the mean log-likelihood value from jackknife samples and the log-likelihood value of the shortened OV alignment (equivalent length A partition). This difference is expressed in terms of number of standard deviations (SD) calculated for the jackknife samples. The improvement in data-model fit obtained by excluding sites at the most varied end of the OV alignment was always significant at $P < 0.001$ (no score for any jacknife sample was better than the score generated by noise reduction).

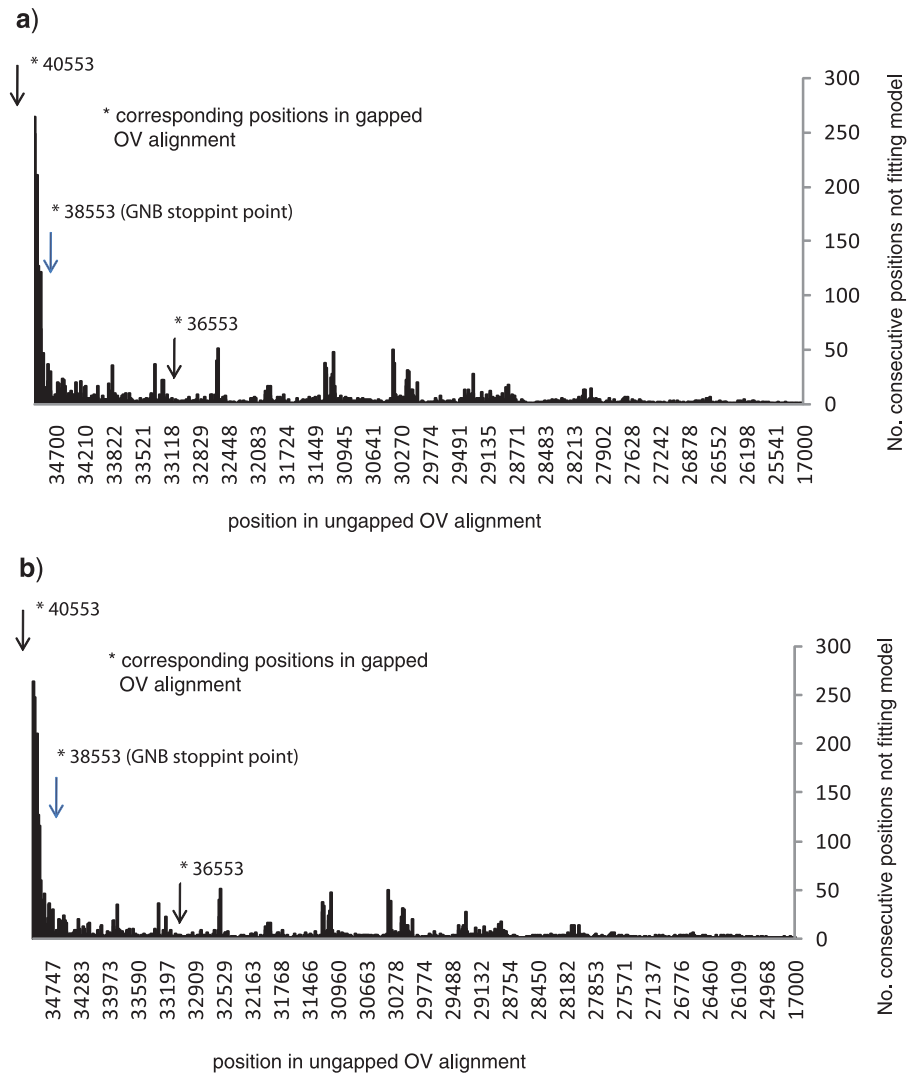


FIGURE 4. a) Histogram showing positions of sites in the OV alignment that contain site patterns unexpected under a GTR+I+Γ substitution model and *Amborella*+*Trithuria*+Nymphaeaceae hypothesis. b) Histogram showing positions of sites in the OV alignment that contain site patterns unexpected under a GTR+I+Γ substitution model and *Amborella* most basal hypothesis. A feature of both graphs is that relatively few sites fit either model at the most varied end of the OV alignment. Both ungapped positions and gapped positions (∗) have been indicated on the figure.

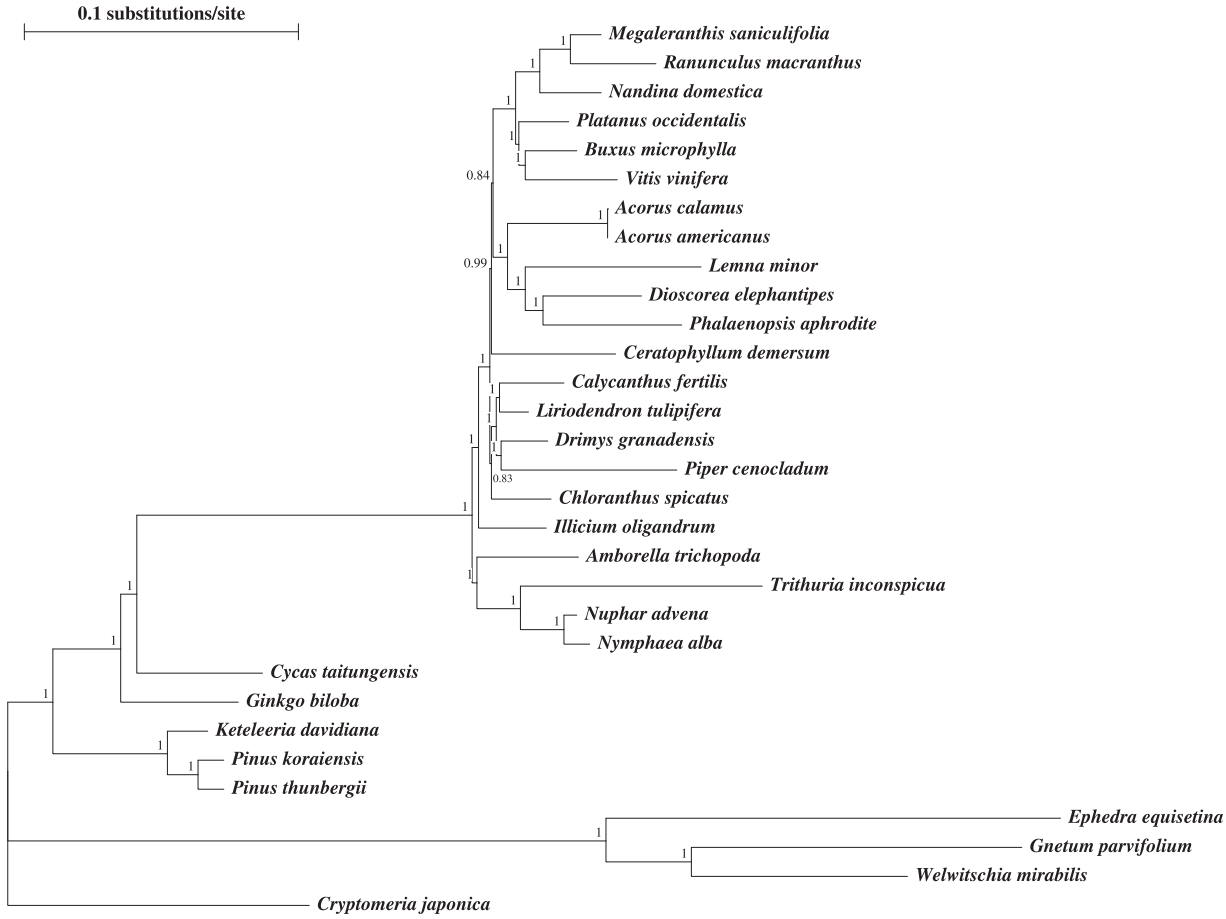


FIGURE 5. Tree reconstructed from Bayesian analysis and best-fitting substitution model (CAT+GTR+Γ+covext model) for the conserved A partition (38 553 sites) identified by the GNB criterion.

Amborella and *Nuphar* were inferred to be sister taxa (Finet et al. 2010). Our reconstruction is also congruent with recent analyses of 4 slowly evolving mitochondrial genes (Qiu et al. 2010).

Our phylogenetic reconstruction differs from that obtained in a recent and well-sampled ML-based phylogenetic analyses for 17 concatenated nuclear, mitochondrial, and chloroplast genes (Soltis et al. 2011). This study reported *Amborella* as most basal. Reanalyzing these data with a GTR+I+Γ model and RaxML, we were unable to confirm this finding. Rather, we inferred a phylogenetic tree wherein a clade comprising *Amborella*, *Trithuria*, and Nymphaeaceae received 94% nonparametric bootstrap support (results not shown). Whether this result indicates a shortcoming of the heuristic search with RaxML or a more accurate reconstruction of angiosperm phylogeny from this joint data matrix requires further investigation.

We conclude that analyses of available sequence data do not support the earliest angiosperms being woody and terrestrial. Evidence from phylogenetic analyses of concatenated chloroplast genes appears equally consistent with some of the earliest species being herbaceous and aquatic. Further tests of this hypothesis are needed. We suggest that our analytical protocol provides a valuable approach, and one that is potentially useful for other questions currently being investigated with phylogenomic data sets.

SUPPLEMENTARY MATERIAL

Data files and/or other supplementary information related to this paper have been deposited at Dryad under doi:10.5061/dryad.vs49s (www.datadryad.org).

FUNDING

PJL thanks the Allan Wilson Centre and New Zealand Royal Society (James Cook Fellowsship scheme and Marsden Fund) for financial support. WM also thanks the Deutsche Foschungsgemeinschaft (DFG) for funding.

ACKNOWLEDGEMENTS

We thank Dr. Lars Jermiin and the reviewers for their constructive and helpful suggestions.

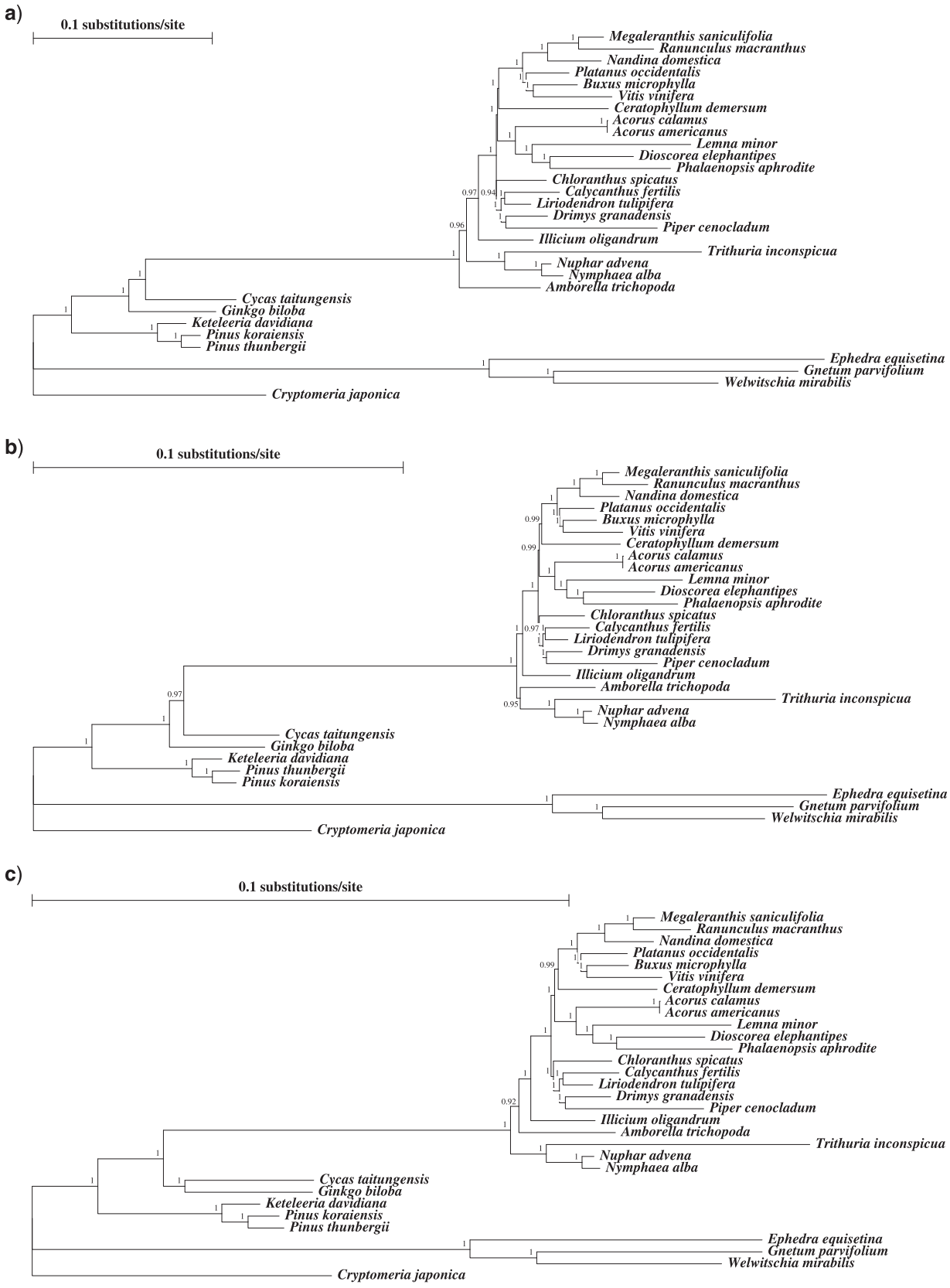


FIGURE 6. a) Tree reconstructed from Bayesian analysis and best-fitting substitution model (CAT+GTR+Γ+covext model) for the full-length (40 553 site) concatenated data set. b) Tree built from the alignment of the first and the second codon positions employing best-fitting CAT+GTR+Γ+covext model. c) Tree built from the alignment of the first and the second codon positions employing the GTR+Γ model.

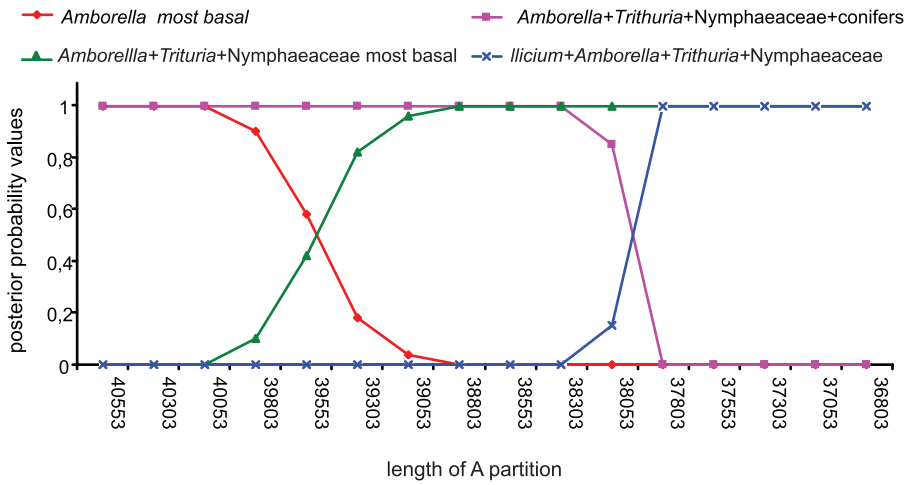


FIGURE 7. Posterior probability support for alternative hypotheses of relationship as sites are removed from the most varied end of the OV alignment computed under the best-fitting substitution model (CAT+GTR+Γ+covext). Similar inferences were obtained with taxon subsets that excluded the most compositionally heterogeneous sequences.

REFERENCES

Ababneh F., Jermiin L.S., Ma C., Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225–1231.

Ane C., Burleigh J.G., McMahon M.M., Sanderson M.J. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol. Biol. Evol. 22:914–924.

Atherton R.A., McComish B.J., Shepherd L.D., Berry L.A., Albert N.W., Lockhart P.J. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods. 6:22.

Barkman T.J., Chenery G., McNeal J.R., Lyons-Weiler J., Ellisens W.J., Moore G., Wolfe A.D., dePamphilis C.W. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc. Natl. Acad. Sci. U.S.A. 97:13166–13171.

Braukmann T.W., Kuzmina M., Stefanović S. 2009. Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. Curr. Genet. 55:323–337.

Brinkmann H., Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Bruno W.J., Halpern A.L. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16:564–566.

Bryant D., Moulton V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Burleigh J.G., Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am. J. Bot. 91:1599–1613.

Chaw S.-M., Chang C.-C., Chen H.-L., Li W.-H. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. 58:424–441.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Drummond A.J., Ashton B., Buxton S., Cheung M., Cooper A., Heled J., Kearse M., Moir R., Stones-Havas S., Sturrock S., Thierer T., Wilson A. 2010. Geneious v5.1, Available from: URL http://www.geneious.com.

Edgar R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Felsenstein J. 2004. PHYLIP (Phylogeny Inference Package). Version 3.6. Distributed by the author. Seattle: Department of Genome Sciences, University of Washington.

Finet C., Timme R.E., Delwiche C.F., Marletaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr. Biol. 20:2217–2222.

Goremykin V., Holland B., Hirsch-Ernst K., Hellwig F. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol. Biol. Evol. 22:1813–1822.

Goremykin V.V., Hirsch-Ernst K.I., Woelfl S., Hellwig F.H. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal Angiosperm. Mol. Biol. Evol. 20:1499–1505.

Goremykin V.V., Hirsch-Ernst K.I., Woelfl S., Hellwig F. 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol. Biol. Evol. 21:1445–1454.

Goremykin V.V., Nikiforova S.V., Bininda-Emonds O.R.P. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol. 71:319–331.

Graham S.W., Iles W.J.D. 2009. Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny. Am. J. Bot. 96:216–227.

Gruenheit N., Lockhart P.J., Steel M.A., Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites Mol. Biol. Evol. 25:1512–1520.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Hansmann S., Martin W.T. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int. J. Syst. Evol. Microbiol. 50: 1655–1663.

Hirt R.P., Logsdon J.M., Healy B., Dorey M.W., Doolittle W.F., Embley T.M. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. U.S.A. 96:580–585.

Ho J.W.K., Adams C.E., Lew J.B., 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 22:2162–2163.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Jansen R.K., Cai Z., Raubeson L.A., Daniell H., dePamphilis C.W., Leebens-Mack J., Mueller K.F., Guisinger-Bellian M., Haberle R.C., Hansen A.K., Chumley T.W., Lee S.-B., Peery R., McNeal J.R., Kuehl J.V., Boore J.L. 2007. Analysis of 81 genes from 64 plastid genomes

resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. U.S.A. 104:19369–19374.
Jermiin L., 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. Syst. Biol. 53:638–643.
Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.
Lartillot N., Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.
Leebens-Mack J., Raubeson L.A., Cui L., Kuehl J.V., Fourcade M.H., Chumley T.W., Boore J.L., Jansen R.K., dePamphilis C.W. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol. Biol. Evol. 22:1948–1963.
Lockhart P. J., Beanland T. J., Howe C. J., Larkum A.W.D. 1992. Sequence of *Prochloron didemni* atpBE and the inference of chloroplast origin. Proc. Natl. Acad. Sci. U.S.A. 89:2742–2746.
Lockhart P.J., Larkum A.W.D Steel M.A., Waddell P.J., Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis.Proc. Natl. Acad. Sci. U.S.A. 93:1930–1934.
Lockhart P.J., Novis P., Milligan B.G., Riden J., Rambaut A., Larkum A.W.D. 2006. Heterotachy and tree building: a case study with plastids and Eubacteria Mol. Biol. Evol. 23:40–45.
Lockhart P.J., Penny D. 2005. The place of *Amborella* within the radiation of angiosperms. Trends Plant Sci. 10:201–202.
Lopez P., Forterre P., Philippe H. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49:496–508.
Martin W.T., Deusch O., Stawski N., Gruenheit N., Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10: 203–205.
Mathews S., Donoghue M.J. 1999. The root of Angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286:947–950.
Nguyen M.A.T., Klaere S., von Haeseler A. 2011. MISFITS: evaluating the goodness of fit between a phylogenetic model and an alignment. Mol. Biol. Evol. 28:143–152.
Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Woerheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst. Biol. 53:978–989.
Posada D., Crandall K.A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.
Qiu Y.-L., Lee J., Bernasconi-Quadroni F., Soltis D.E., Soltis P.S., Zanis M., Zimmer E.A., Chen Z., Savolainen V., Chase M.W. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404–407.
Qiu Y.-L., Wang B., Xue J.-Y., Hendry T.A., Li R.-Q., Brown J. W., Liu Y., Hudson G.T., Chen Z.-D. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. J. Syst. Evol. 48: 391–425.
Rambaut A. 2002. Se-Al. Sequence Alignment Editor v2.0a11. Available from: URL http://evolve.zoo.ox.ac.uk.
Rodriguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang B.F., Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.
Ruiz-Trillo I., Riutort M., Littlewood D.T., Herniou E.A., Baguna J. 1999. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. Science 283:1919–1923.
Saarela J.M., Rai H.S., Doyle J.A., Endress P.K., Mathews S., Marchant A.D., Briggs B.G., Graham S.W. 2007. Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. Nature 446:5–8.
Schmidt H.A., Strimmer K., Vingron M., von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
Shavit L., Penny D., Hendy M.D., Holland B.R. 2007. The problem of rooting rapid radiations. Mol. Biol. Evol. 24:2400–2411.
Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.L., Hilu K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J., Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98:704–730.
Soltis D.E., Soltis P.E. 2004. *Amborella* not a basal angiosperm? Not so fast. Am. J. Bot. 91:997–1001.
Sperling E.A., Peterson K.J., Pisani D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of eumetazoa. Mol. Biol. Evol. 26:2261–2274.
Stamatakis A., Ludwig T., Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.
Stefanović S., Rice D.W., Palmer J.D. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol. Biol. 4:35.
Sullivan J., Holsinger K.E., Simon C. 1995. Among-site variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. Mol. Biol. Evol. 12:988–1001.
Sun G., Dilcher D.L., Wang H., Chen Z. 2011. A eudicot from the Early Cretaceous of China. Nature 471:625–628.
Swofford D.L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland (MA): Sinauer Associates.
Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.
Tuffley C., Steel M.A. 1998. Modelling the covarion hypothesis of nucleotide substitution. Math. BioSci. 147:63–91.
Wakasugi T., Tsudzukit J., Itot S., Nakashimat K., Tsudzuki T. 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc. Natl. Acad. Sci. U.S.A. 91:9794–9798.
Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.
Wu C.-S., Wang Y.-N., Liu S.-M., Chaw S.-M. 2007. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. Mol. Biol. Evol. 24: 1366–1379.
Yang Z. 2006. Computational Molecular Evolution. Oxford University Press, Oxford, England.
Zanis M.J., Soltis D.E., Soltis P.S., Mathews S., Donoghue M.J. 2002. The root of the angiosperms revisited. Proc. Natl. Acad. Sci. U.S.A. 99:6848–6853.
Zerbino D.R., Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829.
Zhong B., Deusch O., Goremykin V.V., Penny D., Biggs P.J., Atherton R.A., Nikiforova S.V., Lockhart P.J. 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3: 1340–1348.

Appendix 2.

Zhong, B., Fong, R., McLenachan, P.A., and Penny, D. Phylogenetic analysis of two monilophyte chloroplasts and decelerated evolution linked to the generation time in tree ferns. (in preparation)

In this study the chloroplast genomes of tree fern *Dicksonia squarrosa* and the "fern ally" *Tmesipteris elongata* were sequenced. The phylogenetic inferences and divergence time estimation were conducted and compared by using the original and most conserved datasets. We found that fast-evolving sites mislead the phylogenetic inference of position of Lycophytes, and also impact the divergence time estimation at deeper lineages.

I was responsible for sequencing and assembling two new chloroplast genomes, conducting all the phylogenetic analyses and divergence time estimation. I was also responsible for writing a complete draft of the paper.

Phylogenetic analysis of two monilophyte chloroplasts and decelerated evolution linked to the generation time in tree ferns

Bojian Zhong[1], Richard Fong[1], Lesley J Collins[2], Patricia A. McLenachan[1] and David Penny[1]

[1] Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

[2] Faculty of Health Sciences, Universal College of Learning, Palmerston North, New Zealand

Abstract

We report the chloroplast genomes of a tree fern (*Dicksonia squarrosa*) and a 'fern ally' (*Tmesipteris elongata*), and show that the monilophytes (ferns and relatives) phylogeny is basically as expected. The *Tmesipteris* comes with *Psilotum* as a fern ally, and the tree fern shows the major reduction in the rate of evolution. Thus there has been a major slowdown in the rate of mutation in both families of tree ferns, and we suggest this is related to a generation time effect. If there is a long time period between generations then this is probably incompatible with a high mutation rate because otherwise nearly every propagule would probably have several lethal mutations. This effect will be especially strong in organisms that have large numbers of cell divisions between generations. This shows the necessity of going beyond phylogeny and integrating its study with other properties of organisms. Such integration appears to be a future activity of phylogenies.

Keywords: *Tmesipteris*, *Dicksonia*, ferns and fern allies, chloroplast genomes, generation time effect.

We address two types of questions in this study: the phylogeny of some early land plants, and some of the biological reasons for the observed differences in mutation rates. Firstly the taxa. *Tmesipteris* (or 'hanging fork fern') was to help test the possibility that the more widespread *Psilotum* was misplaced because of 'long branch attraction' (Hendy and Penny, 1989). *Tmesipteris* and *Psilotum* are both interesting plants in that *Psilotum* superficially resembles certain extinct early vascular plants, such as the rhyniophytes and the trimerophyte genus *Psilophyton* (Bierhorst 1977). The unusual features of *Psilotum* that suggest an affinity with early vascular plants include dichotomously branching sporophytes, aerial stems arising from horizontal rhizomes, a simple vascular cylinder, homosporous and terminal eusporangia

and a lack of roots. However, recent results have tended to place the Psilotales (which includes *Tmesipteris*) closer to the ferns (e.g. Qiu and Palmer 1999). However, this left *Psilotum* as a 'long branch' in the tree, and these are well-known to be problematic. *Tmesipteris* only grows in New Caledonia, New Zealand, and parts of eastern Australia, and so it is difficult for some researchers to obtain it for sequencing. *Dicksonia* was chosen to test whether it would also show the slowdown in rates (see Korall et al. 2010) that was known for the chloroplast genes of the other main family of tree-ferns (Cyatheaceae, that includes the *Alsophila* genus, and whose full chloroplast genome is already available).

Next we turn to the question of evolutionary rates. Early in molecular evolution studies, researchers were surprised at the relatively equal rate of molecular evolution, for example, between vertebrates, fungi and plants: there did not appear to be the expected correlation between mutation rates in diversified and non-diversified lineages (see, for example, Kimura and Ohta 1974). This observation, together with the much higher than expected genetic diversity within populations, led to the development of the neutral theory of molecular evolution (Kimura and Ohta 1974) where mutations that were neutral tended to outnumber those that were advantageous. As a side issue perhaps, this did lead to the concept of a 'molecular clock', with a relatively constant rate of DNA evolution in different eukaryote groups.

However, more recently there has been considerable interest in the actual variation in rates, and the observation of lineage-specific rate heterogeneity has been well characterized within fungi (Lumbsch et al. 2008), mammals (Goldie et al. 2011), seed plants (Smith and Donoghue 2008; Xiang et al. 2008; Bromham et al. 2013) and ferns (Soltis et al. 2002; Schneider et al. 2004; Korall et al. 2010; Rothfels et al. 2012). We still lack a good biological understanding of factors that might affect this observed variation in rates. There are at least three general types of explanation that might affect rate, the first is a general increase (or decrease) in mutation rate; the second is a change in the number of sites 'free to vary' (that is, a change in selection pressures); and the third is variations in the mechanisms that might, for example, lead to double-stranded breaks and subsequent repair. This last aspect of the heterogeneity has probably made it difficult for resolving aspects of the placental mammal phylogeny (Romiguier et al. 2013) because the location where genetic recombination occurs (and increasing the number of double-stranded breaks, which are more error prone during their correction) appears to keep changing within Placentals. It is important (essential) to understand the biological principles for the observed variation in rates of molecular evolution

in different groups (e.g. Lanfear et al. 2014). We should be able to make *predictions* about what we expect.

The heterogeneous pattern of among-lineage rate variation has presented a significant challenge for estimating divergence times. Various substitution models that relax the assumption of the strict molecular clock have been developed to account for rate heterogeneity between lineages in molecular phylogenetics (e.g. Thorne et al. 1998; Sanderson 2002; Drummond et al. 2006). It has been reported that fast-evolving sites are one important source of systematic errors in molecular phylogenetics, and accuracy of phylogenetic analyses can be improved by removal of the most variable sites regardless of the mechanism of mutation (e.g. Goremykin et al. 2010; Zhong et al. 2011, 2014; Parks et al. 2012). However, few studies have evaluated whether the fast-evolving sites could affect the accuracy of the estimation of divergence time even though using the relaxed clock models. To test the impact of divergence time estimation based on different sites with different evolutionary rates, and to investigate the relation between generation for a range of genome sizes and mutation rate, we designed an empirical study using the chloroplast genomes of land plants, which include two newly sequenced species (a tree fern and a fern ally) to give a total of 28 chloroplast genomes.

Materials and Methods.

The tree fern *Dicksonia squarrosa* and the 'fern ally' *Tmesipteris elongata* were collected and sourced from Palmerston North, New Zealand. The *Dicksonia* sample was a cultivated plant from Palmerston North, and the *Tmesipteris* sample was growing in the Kahuterawa valley, inland from Palmerston North. Total genomic DNA (~50 ng) from each was extracted using the Qiagen Plant DNeasy kit according to the manufacturer's protocols, and then sequenced using Illumina GAIIx sequencing platform with 100-bp paired-end reads. The short reads were filtered with the error probability < 0.05, and were then assembled using Velvet (Zerbino and Birney 2008). The contigs were further assembled using Geneious software version 5.6 (www.geneious.com). Protein-coding genes were annotated using DOGMA (Wyman et al. 2004) with manual correction. Each protein-coding gene from 28 taxa was aligned using MUSCLE (Edgar 2004), and trimmed to exclude poorly aligned positions using Gblocks (Castresana 2000) with default settings. These alignments were concatenated to generate a matrix of 34,386 sites.

The OV-sorting method (Goremykin et al. 2010) was used to rank the original concatenated alignment from the most to least variable sites based on the measurement of

"observed variability" (OV) of each alignment position. The most variable sites were then successively removed from the original matrix, in increments of 250 sites. The stopping point for site removal was determined as the point at which the two correlations showed significant improvement (see Goremykin et al. 2010, 2013 for details of the method).

The divergence times were estimated using the Bayesian software BEAST version 1.7.2 (Drummond and Rambaut 2007). The optimal substitution model was selected using ModelTest (Posada and Crandall 1998). Rate heterogeneity among lineages was modelled using a UCLN relaxed clock (Drummond et al. 2006). Samples from the posterior distribution were drawn every 2000 steps over 100,000,000 steps of a single chain, with the first 10% of samples discarded as burn-in. Four independent MCMC chains were run. Convergence was checked based on time-series plots of the likelihood scores using Tracer program (Drummond and Rambaut 2007). Eight fossil-based calibrations were utilized for molecular dating analyses. The root age was set at 449-1024 Ma (Clarke et al. 2011). The other internal fossil calibrations were representatives of the oldest known clades to provide minimum age constraints (see Table 1).

For the rates of evolution we used a Python script to estimate the number of mutations that were expected to occur for a set number of genes and for a given mutation rate. In order to make the calculation in a reasonable amount of time, the mutation rate and numbers of genes were scaled to keep the same proportion. In practice, the genomes started with 1000 genes, each 1000 nucleotides long an error rate in copying DNA of about 10^{-9} per errors per nucleotide – this is a realistic rate for eukaryotes (Drake 1999). The number of mutations was recorded, and if there were more than 10 mutations in a gene that was considered lethal on that gene. Alternatively, if two mutations occurred at the same amino acid site (a double hit) this was also taken as a lethal mutation – and led to a 'dead (non-functional) gene'. The mutation rate was the same for all genes. In general photosynthetic organisms seem to have a higher number of genes, often having 30,000 genes (Raven et al. 2013).

The average 'generation time' for tree ferns does not appear to be accurately known (nor for many organisms) and so an estimate of about 100 years for tree ferns being actively reproductive was used, based on results in Ash (1987), Shepherd and Cook (1988), and Large and Braggins (2004). In some cases an estimate of 200 years was available, but we limited it to, on average, 100 year generation time. However, as we point out, the number of cell divisions per generation is probably a critical factor.

Results

For this study we had 34,386 aligned sites, and identified 3,250 rapidly evolving sites using the OV-sorting method (see Figure 1). Thus the reduced OV-sorted data is 31,136 aligned sites.

The first step was to reconstruct the phylogeny. Zhong et al. (2011) and Goremykin et al. (2013) have reported that the OV-sorting method is effective in identifying the fastest evolving sites, and phylogenetic inference is significantly improved after their removal. We then used the GTRGAMMA model to infer maximum likelihood (ML) phylogeny. The ML analyses with RAxML based on original and OV-sorted data produced both well-supported phylogenetic trees (Fig.2 and Fig.3). All major groups (e.g. Seed plants, Monilophytes and Lycophytes) are monophyletic with high bootstrap support (BP), and the 'fern ally' *Tmesipteris elongata* is, as expected, the sister group to *Psilotum nudum*, and they are basal to the ferns. The tree fern clade (i.e. *Dicksonia squarrosa* and *Alsophila spinulosa*) is strongly supported as monophyletic (BP=100), and there is a major rate deceleration occurred along both tree ferns (notably shorter branches within the tree fern clade). Thus the slowdown in rates does occur in both tree ferns. This slowdown is in marked contrast to the other ferns where there is rate acceleration, especially among the more advanced (derived) ferns.

The only difference between the two trees (Figs 2 and 3) was the position of Lycophytes. For the original data, Lycophytes as sister to seed plants was weakly supported (BP=64%, Fig. 2). In contrast, after removing faster-evolving sites, the phylogenetic tree supported Lycophytes close to [Seed plant + Monilophytes] (Fig. 3), and for which previous studies have given similar results (Clarke et al. 2011; Pryer et al. 2001; Rai and Graham. 2010). This confirms that these fast-evolving sites may mislead the phylogenetic inference of position of Lycophytes, thus, we used the phylogenetic tree based on OV-sorted data for further divergence time estimation.

To evaluate the impact of divergence time estimation of the fast-evolving sites, we estimated the divergence times using the original and OV-sorted data sets with eight fossil records. We found that age estimates from most nodes did not vary substantially between original and OV-sorted data (see Table 1). For instance, relaxed molecular clock analyses using original data and OV-sorted data yielded the similar mean estimates as 136.6 and 150.1 million years before present (Ma) for crown angiosperms (node 1 in table 1), and the estimated age of crown Tracheophyta (node 20) is 445.7 and 428.9Ma, respectively. However, for some deeper nodes (e.g. node 25, 26 and 27), the mean ages and confidence intervals reduced considerably with OV-sorted data compared with the original data. This does require more investigation to clarify such variation because estimates of divergence time are an

important aspect of molecular evolution. Consequently, these results are encouraging in that they appear to give more realistic estimates for the deeper divergences.

Results with numbers of mutations between generations, in Fig. 4, show the expected number of mutations based on both the mutation rate and the generation time. At the longer times there are predicted to be many more mutations in the offspring, and many of these are potentially lethal. We only count half the genes, in that we allow some 'lethal' mutations to have occurred in leaf tissue, and any such cells will simply stop at that point. Basically, it is the next generation that is important here, many genes will only be expressed earlier in developmental, and in root tissue – so they will not have been selected against during stem and leaf growth. In practice, there will be an effect from cells being diploid, but that is not expected to alter the basic result in the longer term.

Discussion

The phylogenetic aspects covered here appear largely to be as expected for the deeper phylogeny of Land Plants. *Tmesipteris* comes strongly with *Psilotum*, and the two families of tree ferns also strongly come together. Again the OV sorting method appears to work well, and the method does not appear to make many major changes to the times of divergence estimates, but the more recent divergences for the deepest nodes on Figure 3 do warrant further testing of the OV (and other) methods. So whilst the phylogeny is relatively stable, the further testing of the times of divergence is needed.

The other aspect that needs further testing is that it appears that larger organisms (with more cell divisions and longer generation times) tend to have lower mutation rates, possibly in order to limit the number of mutations between a parent and its offspring. If there are very large numbers of mutations between parents and offspring, then almost certainly some of these mutations would be detrimental, and this appears to place a limit on the mutation rate of organisms with a long generation time. It appears that the slow-down of mutation rates affects both the nuclear and chloroplast genes (Rothfels and Schuettpelz, 2013). There has been considerable effort into testing some of the reasons behind variation in rates between different lineages. Lanfear et al. (2013) pointed out that for many trees, about a fifth of the rate variation can be explained by slower rates of mutation in taller trees. It may well be that the generation time effect is at least a partial explanation for why there appears, on a geological time scale, to be continued turnover of large organisms. However, it is always going to be important to take the number of cell divisions between generations, for most vertebrates (including humans) there are special germ-line cells set aside that may have fewer cell

divisions that many of the somatic cells – this is discussed in Kong et al. (2012) where the differing contribution of males and females (for humans) to the numbers of mutations per generation is discussed.

Furthermore, Lynch and Abegg (2010) reported that larger populations can acquire much faster useful combinations of mutations than can smaller populations (such as might be found with larger individuals). Consequently, smaller organisms (with larger population sizes) are better able to continue evolving, particularly gaining complex new features. It would be important to determine the number of mutations between generations for a range of genome sizes, mutation rates, and population sizes, and the apparent slowdown in mutation rates among the tree ferns could be an important test of this hypothesis. We do need further tests on whether very long generation times are associated with lower mutation rates in other organisms – we predict that a high mutation rate and a long generation time are incompatible (see also Thomas et al, 2010). It is also important to understand the mechanisms involved in the lower mutation rates (as observed in tree ferns). Is higher accuracy (lower mutation rates) associated with slower copying of DNA (and therefore more time for checking of potential errors during copying)? Or is it some intrinsic mechanism that is independent of the rate of DNA copying? All aspects, the phylogeny of Land Plants, the use of the OV method for time estimates, and the effect of life cycle and generation time, certainly warrant continued study. But the numbers of mutations between parents and offspring is certainly an interesting factor that merits additional study.

Acknowledgements

We thank Patrick Brownsey and Leon Perrie, both from Te Papa Museum, Wellington, for information about generation times of tree ferns.

References

Ash, J. (1987). Demography of *Cyathea hornei* (Cyatheacee), a tropical fern form Fiji. *Austr. J Bot*. 35: 331-341.

Bierhorst, D.W. (1977). The systematic position of *Psilotum* and *Tmesipteris*. *Brittonia.* 29: 3-13.

Bromham, L., Cowman, P.F., Lanfear, R. (2013). Parasitic plants have increased rates of molecular evolution across all three genomes. *BMC Evol. Biol*. 13:126.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol*. 17:540-552.

Clarke, J.T., Warnock, R.C.M., Donoghue, P.C.J. (2011). Establishing a time-scale for plant evolution. *New Phytol*. 192:266-301.

Drake, J.W. (1999). The distribution of rates of spontaneous mutation over viruses, prokaryotes and eukaryotes. *Ann NY Acad. Sci*. 870: 100-107.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A.(2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology.* 4:e88.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol*. 7:214.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res*. 32:1792-1797.

Goldie, X., Lanfear, R., Bromham, L. (2011) Diversification and the rate of molecular evolution: no evidence of a link in mammals *BMC Evol. Biol*. 11:286.

Goremykin, V.V., Nikoforova, S.V., and Bininda-Emonds, O.P.P. (2010). Automated removal of noisy data in phylogenomic analyses. *J. Mol. Evol*. 71:319-331.

Goremykin, V.V., Nikiforova, S.V., Biggs, P.J., Zhong. B., De Lange, P., Martin, W., Woetzel, S., Atherton, R.A., McLenachan, T., and Lockhart, P.J. (2013). The evolutionary root of flowering plants. *Syst. Biol*. 62: 51-62.

Hendy, M.D. and Penny, D. (1989) A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38: 297-309.

Kimura, M. and Ohta, T. (1974). On some principles governing molecular evolution. *Proc. Natl Acad. Sci. USA*. 71: 2848-2852.

Kong, A et al. (22 authors) (2012) Rate of de novo mutations and the importance of father's age to disease risk *Nature*. 488: 471-475.

Korall, P, Schuettpelz, E, Pryer, K.M. (2010) Abrupt deceleration of molecular evolution

linked to the origin of arborescence in ferns. *Evolution* 64:2786-2792.

Lanfear, R., Ho, S.Y.W., Davies, T. J., Moles, A.T.A.L., Swenson, N.G., Warmann, L., Zanne, A.E., and Allen, A.P. (2013). Taller plants have lower rates of molecular evolution. *Nature Comm*. 4:1-29.

Lanfear, R., Kokko, H., Eyre-Walker, A. (2014) Population size and the rate of evolution. *Trends Ecol. Evol*. 29: 33-41

Large, M.F. and Braggins, J.E. (2004) *Tree ferns*. Portland, Oregon, Timber Press.

Lumbsch, H., A. L. Hipp, P. K. Divakar, O. Blanco and A. Crespo. (2008). Accelerated evolutionary rates in tropical and oceanic parmelioid lichens (Ascomycota). *BMC Evol. Biol.* 8:257.

Lynch, M., and Abegg, A. (2010). The rate of establishment of complex adaptations. *Mol. Biol. Evol*. 27:1404-1414.

Parks, M., Cronn, R., and Liston, A. (2012). Separating the wheat from the chaff: Mitigating the effects of noise in a plastome phylogenomic data set from Pinus L. (Pinaceae). *BMC Evol. Biol*. 12: 100.

Posada, D., and Crandall, K.A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics*. 14:817-818.

Pryer, K.M., Schneider, H., Smith, A.R., Cranfill, R., Wolf, P.G., Hunt, J.S., and Sipes, S.D. (2001). Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature*. 409:618-622.

Qiu, Y. and Palmer, J. (1999) Phylogeny of early land plants: insights from genes and genomes. *Trends Plant Sci*. 4: 26-30.

Rai, H.S., and Graham, S.W., 2010. Utility of a large, multigene plastid data set in inferring higher-order relationships in ferns and relatives (monilophytes). *Am. J. Bot*. 97: 1444-1456.

Raven, J.A., Beardall, J , Larkum, A.W.D., Sanchez-Baracaldo, P (2013) Interactions of photosynthesis with genome size and function. *Phil. Trans. R. Soc. B-Biol. Sci*. 368: 20120264.

Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N., Douzery, E. (2013) Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol*. 30: 2134-2144.

Rothfels, C.J., Larsson, A., Kuo, L.Y., Korall, P., Chiou W.L., and Pryer, K.M. (2012). Overcoming deep roots, fast rates, and short internodes to resolve the ancient rapid radiation of eupolypod II ferns. *Syst. Biol*. 61:490-509.

Rothfels, C.J., and Schuettpelz, E. (2013) Accelerated rate of molecular evolution for vittarioid ferns is strong and not driven by selection. *Syst Biol*. 63: 31-54

Sanderson, M.J. (2002). Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol*. 19:101-109.

Schneider, H., Schuettpelz, E., Pryer, K.M., Cranfill, R., Magallon, S., and Lupia, R. (2004). Ferns diversified in the shadow of angiosperms. *Nature*. 428:553-557.

Shepherd, W., and Cook, W.C. (1988) *The Botanic Garden, Wellington: A New Zealand History, 1840-1987*. Wellinton Millwood Press

Smith, S.A. and Donoghue, M.J. (2008). Rates of molecular evolution are linked to life history in flowering plants. *Science*. 322:86-89.

Soltis, P.S., Soltis, D.E., Savolainen, V., Crane, P.R., and Barraclough, T.G. (2002). Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. *Proc. Natl Acad. Sci. USA*. 99: 4430-4435.

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

Swofford, D.L. (2002). *PAUP*. Phylogenetic analysis using parsimony (*and other methods)*. Version 4. Sunderland (MA): Sinauer Associates.

Thomas, JA, Welch, JJ, Lanfear, R, Bromham, L. (2010) A generation time effect on the rate of molecular evolution in invertebrates. *Mol. Biol. Evol*. 27: 1173-1180.

Thorne, J.L., Kishino, H., and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 51:1647-1657.

Wyman, S.K., Jansen, R.K., and Boore, J.L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*. 20:3252-3255.

Xiang, Q.-Y., Thorne, J.L., Seo, T.-K., Zhang, W., Thomas, D.T., and Ricklefs, R.E. (2008). Rates of nucleotide substitution in Cornaceae (Cornales)--Pattern of variation and underlying causal factors. *Mol. Phylog. Evol*. 49:327-342.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18:821-829.

Zhong, B., Deusch, O., Goremykin, V.V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S.V., and Lockhart, P.J. (2011). Systematic error in seed plant phylogenomics. *Genome Biol. Evol*. 3:1340-1348.

Zhong, B., Xi, Z., Goremykin,V.Z., Fong, R., McLenachan, P.A., Novis, P.M., Davis, C.C., D. Penny. (2014) Origin of land plants revisited using heterogeneous models and three new chloroplast genomes. *Mol. Biol. Evol*. 31, 177-183.

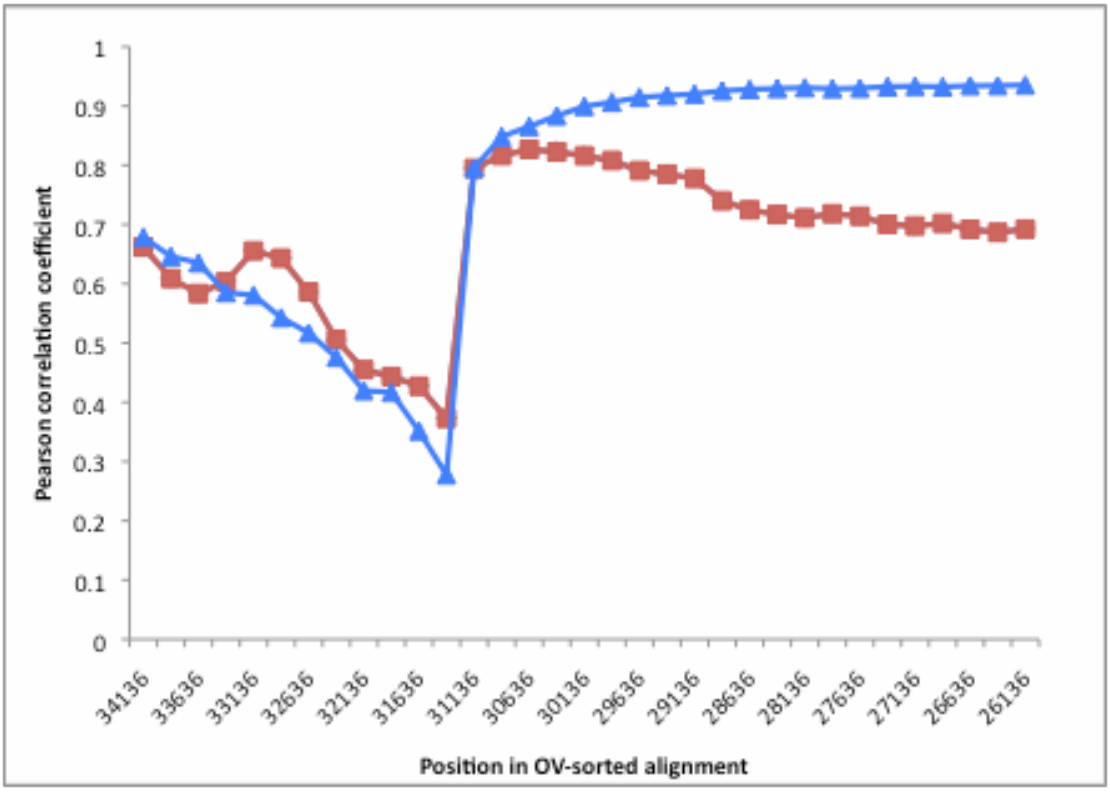


Fig.1. Pearson correlation results. The blue line indicates the Pearson correlation coefficient (*r*) of the ML distance calculated from "A" (more conserved) and "B" (less conserved) partitions. The red line indicates the *r* value of uncorrected *p*-distances and ML distances for B partitions. The *r* values begin to increase significantly at 31,136 sites remaining and this is taken to indicate that the assumed model of nucleotide evolution is beginning to fit the data well.

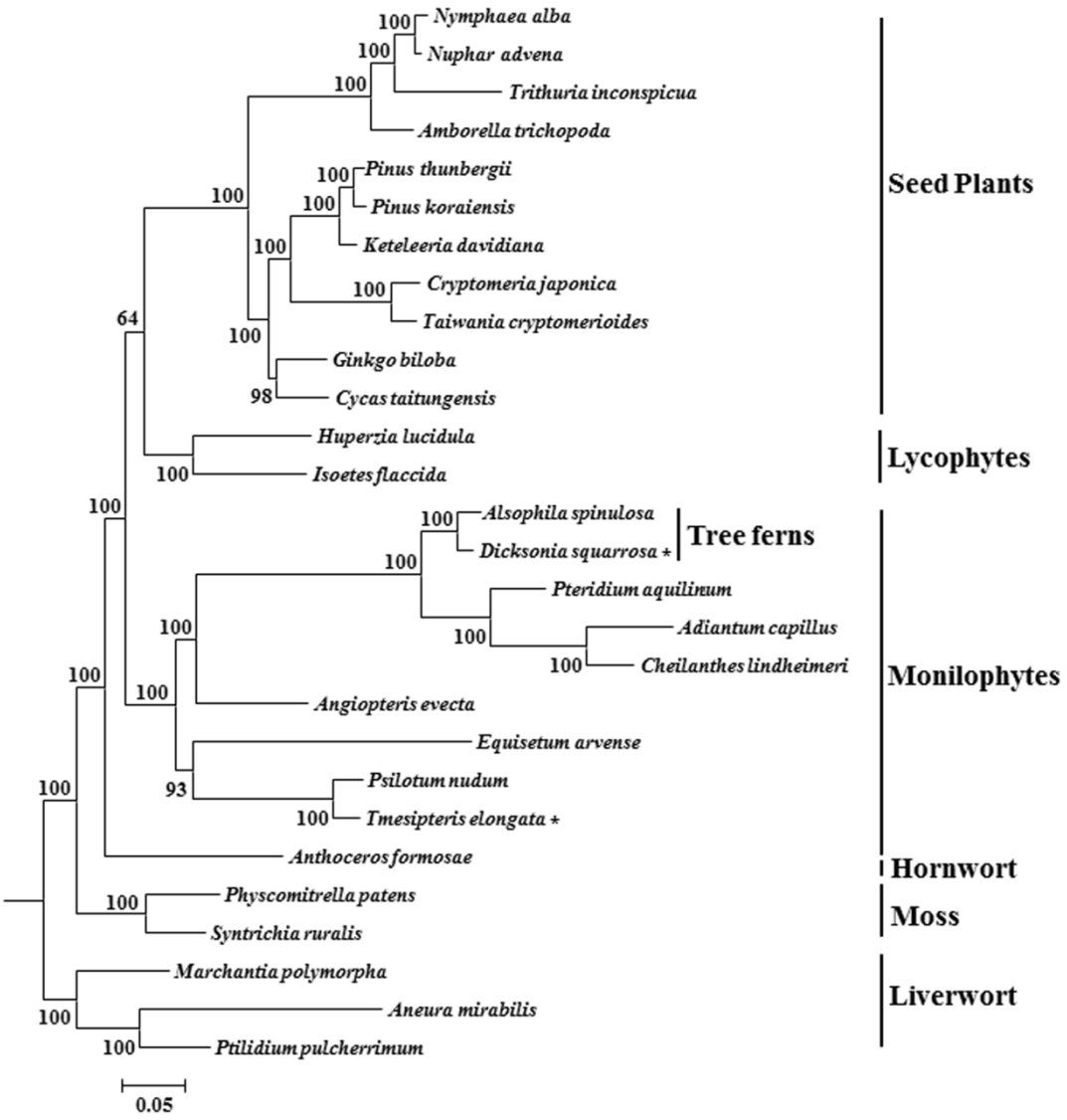


Fig.2. Maximum likelihood (ML) tree of land plants based on the original data (34,386 sites). Bootstrap support values are indicated along the branches. The two newly sequenced genomes are indicated as *. In this tree the Lycophytes are adjacent to the seed plants with weak bootstrap support (BP=64%).

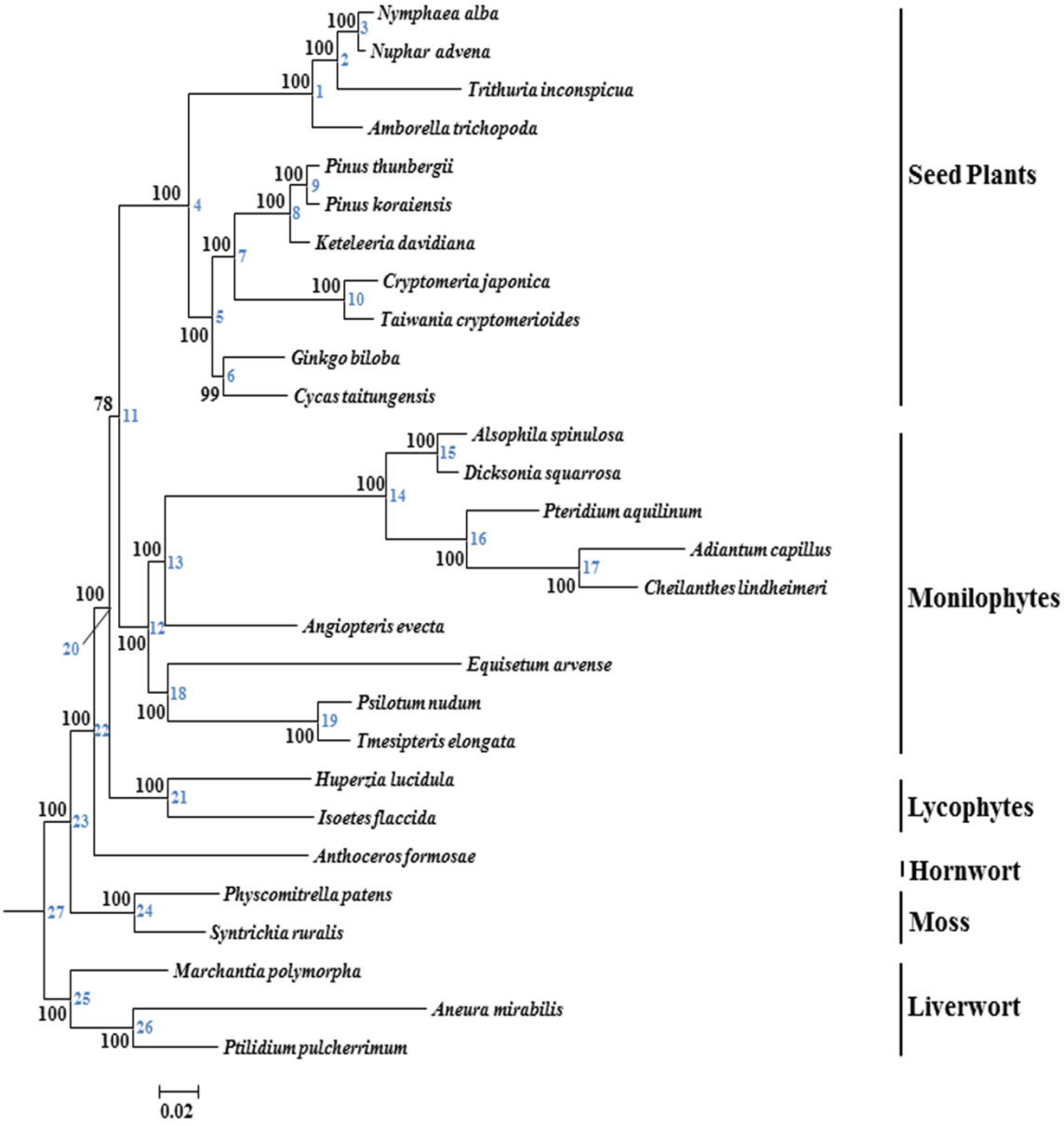


Fig.3. Maximum likelihood (ML) tree of land plants based on the OV-sorted matrix (31,136 sites). Bootstrap support values are indicated along the branches and node numbers are marked as blue. This ML tree is the same as the Fig.1 except that the Lycophytes are now adjacent to the seed plants plus Monilophytes, and with 100% bootstrap support.

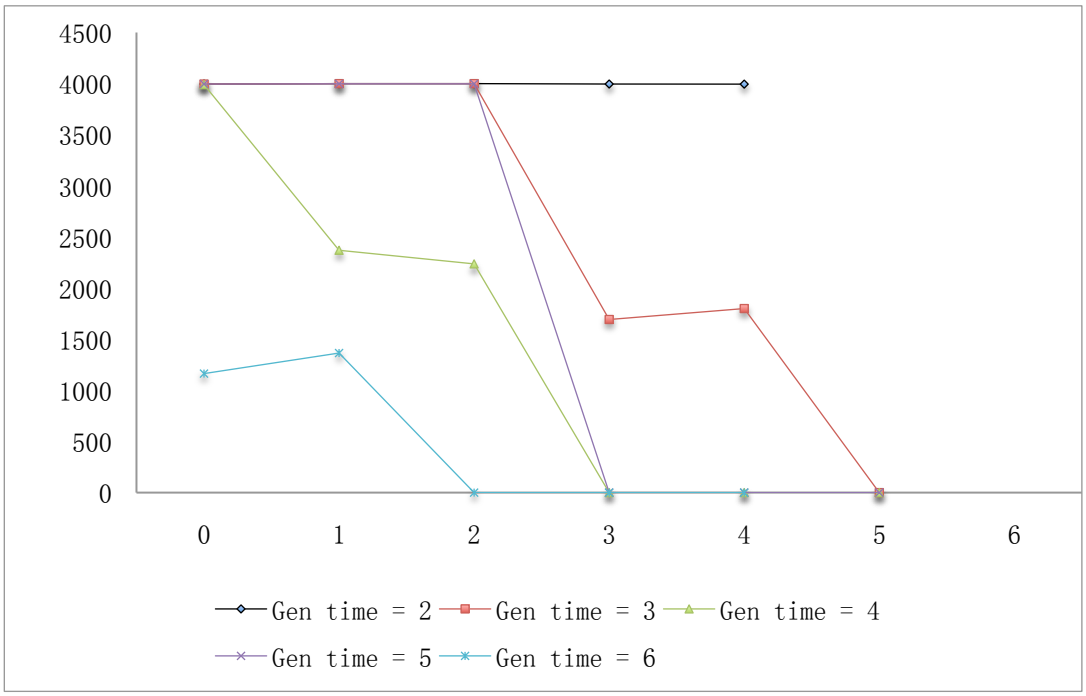


Fig.4. Numbers of mutations at different mutation rates and generation times. As the generation time increases, particularly with more cell divisions per generation, the chance of successful offspring (with no lethal mutations) decreases considerably. (x axis is relative generation time, and y axis is the population after 50 years).

Table 1. Summary of age estimates for all nodes using both the original and the reduced OV-sorted matrices

	mean estimates (in Ma)		95% credibility intervals		
Data / Node	Full matrix (34,386 sites)	OV-sorted matrix (31,136 sites)	Full matrix	OV-sorted matrix	Fossil calibrations (Ma)
1	136.6	150.1	67.0-208.7	75.6-242.7	
2	85.8	97.7	38.4-140.1	42.8-176.6	
3	22.9	28.2	4.1-47.7	3.8-60.6	
4	315.8	317.5	306.2-333.2	306.2-339.4	>306.2[&]
5	225.0	224.1	168.5-287.8	165.4-285.3	
6	187.0	163.3	108.8-262.2	75.9-257.7	
7	160.9	161.4	147.0-187.8	147.0-187.3	>147.0[&]
8	57.3	62.6	22.7-99.7	21.6-108.6	
9	23.8	26.9	5.3-47.7	5.7-54.9	
10	55.1	58.4	12.2-98.8	16.5-106.6	
11	413.1	404.8	388.2-447.9	388.2-429.2	>388.2[&]
12	366.2	368.5	354.0-388.2	354.0-390.7	>354.0[#]
13	327.8	336.7	280.8-365.4	291.5-378.8	
14	221.9	228.8	179.8-264.1	187.5-270.1	
15	165.7	168.0	159.0-180.6	159.0-185.3	>159.0[#]
16	144.8	154.3	91.1-201.1	93.1-217.0	
17	73.1	76.5	36.0-116.5	34.5-122.0	
18	296.1	296.2	203.6-364.9	189.9-370.3	
19	69.1	72.3	18.5-147.7	14.7-147.1	
20	445.7	428.9	403.3-492.9	400.1-463.5	
21	387.7	386.3	377.4-406.9	377.4-403.0	>377.4[+]
22	483.4	454.4	423.0-553.2	413.6-501.5	
23	534.7	487.6	450.9-629.1	435.6-550.8	>420.4[&]
24	190.3	178.3	51.3-353.1	37.5-364.7	
25	677.3	375.8	277.4-1030.3	172.5-569.7	
26	468.5	228.3	130.4-758.6	99.1-397.8	
27	775.5	529.8	502.4-1024.0	449.0-629.5	449-1024[&]

Age calibrations; [&] Clarke et al. 2011; [#] Schneider et al. 2004; [+] Soltis et al. 2002. Node numbers are shown on Figure 3.

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bojian Zhong

Name/Title of Principal Supervisor: Prof. David Penny

Name of Published Research Output and full reference:

Zhong, B., Deusch, O., Goremykin, V.V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S.V., Lockhart, P.J. (2011). Systematic error in seed plant phylogenomics. Genome Biology and Evolution. 3:1340–1348. (As the Corresponding author)

In which Chapter is the Published Work: 2

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  Bojian Zhong was responsible for sequencing and assembling three new chloroplast genomes, and for conducting all the phylogenetic analyses. Bojian was also primarily responsible for writing a complete draft of the paper.

Bojian Zhong
Digitally signed by Bojian Zhong
DN: cn=Bojian Zhong, o=Massey University, ou=Institute of Fundamental Sciences, email=b.zhong@massey.ac.nz, c=NZ
Date: 2013.09.16 16:11:38 +12'00'

Candidate's Signature

16/09/13

Date

David Penny
Digitally signed by David Penny
DN: cn=David Penny, o=Massey University, ou=Institute of Fundamental Science, email=D.Penny@massey.ac.nz, c=NZ
Date: 2013.09.16 15:33:31 +12'00'

Principal Supervisor's signature

16/09/13

Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bojian Zhong

Name/Title of Principal Supervisor: Prof. David Penny

Name of Published Research Output and full reference:

Zhong, B., Liu, L., Yang, Z., and Penny, D. (2013). Origin of land plants using the multispecies coalescent model. Trends in Plant Science. 18: 492-495. (As the Corresponding author)

In which Chapter is the Published Work: 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  Bojian Zhong was responsible for collecting nuclear genes, estimating species trees using coalescent model and concatenation method, and contributing some statistical analyses. Bojian was also primarily responsible for writing a complete draft of the paper.

Bojian Zhong
Digitally signed by Bojian Zhong
DN: cn=Bojian Zhong, o=Massey University, ou=Institute of Fundamental Sciences, email=b.zhong@massey.ac.nz, c=NZ
Date: 2013.09.16 16:14:53 +12'00'

Candidate's Signature

16/09/13

Date

David Penny
Digitally signed by David Penny
DN: cn=David Penny, o=Massey University, ou=Institute of Fundamental Science, email=D.Penny@massey.ac.nz, c=NZ
Date: 2013.09.16 15:32:07 +12'00'

Principal Supervisor's signature

16/09/13

Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bojian Zhong

Name/Title of Principal Supervisor: Prof. David Penny

Name of Published Research Output and full reference:

Zhong, B., Xi, Z., Goremykin, V.V., Fong, R., McLenachan, P.A., Novis, P., and Penny, D. (2014). Origin of land plants revisited using heterogeneous models and three new algal chloroplast genomes. Molecular Biology and Evolution. 31: 177-183. (As the Corresponding author)

In which Chapter is the Published Work: 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  Bojian Zhong was responsible for sequencing and assembling three new chloroplast genomes, and for conducting all the phylogenetic analyses. Bojian was also primarily responsible for writing a complete draft of the paper.

Bojian Zhong
Digitally signed by Bojian Zhong
DN: cn=Bojian Zhong, o=Massey University, ou=Institute of Fundamental Sciences, email=b.zhong@massey.ac.nz, c=NZ
Date: 2013.09.16 16:15:53 +12'00'

Candidate's Signature

16/09/13

Date

David Penny
Digitally signed by David Penny
DN: cn=David Penny, o=Massey University, ou=Institute of Fundamental Science, email=D.Penny@massey.ac.nz, c=NZ
Date: 2013.09.16 15:34:43 +12'00'

Principal Supervisor's signature

16/09/13

Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bojian Zhong

Name/Title of Principal Supervisor: Prof. David Penny

Name of Published Research Output and full reference:

White, T*., Zhong, B*., Penny, D. (2013). Beyond reasonable doubt: evolution from DNA sequences. PLoS ONE. 8(8): e69924. (*equal contribution)

In which Chapter is the Published Work: 4

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  Bojian Zhong was responsible for collecting all the empirical datasets, including chloroplast genome, mitochondrial genome, and nuclear data from plants and animals at increasing level of divergence. Bojian also greatly contributed to the development of statistical test and the writing of the manuscript.

Bojian Zhong — Digitally signed by Bojian Zhong, DN: cn=Bojian Zhong, o=Massey University, ou=Institute of Fundamental Sciences, email=b.zhong@massey.ac.nz, c=NZ, Date: 2013.09.16 16:16:42 +12'00'

Candidate's Signature — 16/09/13 — Date

David Penny — Digitally signed by David Penny, DN: cn=David Penny, o=Massey University, ou=Institute of Fundamental Science, email=D.Penny@massey.ac.nz, c=NZ, Date: 2013.09.16 15:35:32 +12'00'

Principal Supervisor's signature — 16/09/13 — Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bojian Zhong

Name/Title of Principal Supervisor: Prof. David Penny

Name of Published Research Output and full reference:

Goremykin,V.V., Nikiforova, S.V., Biggs, P.J. Zhong, B., DeLange, P., Martin,W., Woetzel, S., Atherton, R.A., McLenachan, T., Lockhart P.J. (2013). The evolutionary root of flowering plants. Systematic Biology. 62: 51-62.

In which Chapter is the Published Work: Appendix 1

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  Bojian Zhong was responsible for conducting compositional heterogeneity analyses, and goodness of fit analyses. Bojian also contributed to the writing of the manuscript.

Bojian Zhong
Digitally signed by Bojian Zhong
DN: cn=Bojian Zhong, o=Massey University, ou=Institute of Fundamental Sciences, email=b.zhong@massey.ac.nz, c=NZ
Date: 2013.09.16 16:17:31 +12'00'

Candidate's Signature

16/09/13
Date

David Penny
Digitally signed by David Penny
DN: cn=David Penny, o=Massey University, ou=Institute of Fundamental Science, email=D.Penny@massey.ac.nz, c=NZ
Date: 2013.09.16 15:37:08 +12'00'

Principal Supervisor's signature

16/09/13
Date

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Bojian Zhong

Name/Title of Principal Supervisor: Prof. David Penny

Name of Published Research Output and full reference:

Zhong, B., Fong, R., McLenachan, P.A., and Penny, D. Phylogenetic analysis of two monilophyte chloroplasts and decelerated evolution linked to the generation time in tree ferns. (In preparation)

In which Chapter is the Published Work: Appendix 2

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  Bojian Zhong was responsible for sequencing and assembling two new chloroplast genomes, conducting all the phylogenetic analyses and divergence time estimation. Bojian was also primarily responsible for writing a complete draft of the paper.

Bojian Zhong
Digitally signed by Bojian Zhong
DN: cn=Bojian Zhong, o=Massey University, ou=Institute of Fundamental Sciences, email=b.zhong@massey.ac.nz, c=NZ
Date: 2013.09.16 16:26:58 +12'00'

Candidate's Signature

16/09/13

Date

David Penny
Digitally signed by David Penny
DN: cn=David Penny, o=Massey University, ou=Institute of Fundamental Science, email=D.Penny@massey.ac.nz, c=NZ
Date: 2013.09.16 15:38:05 +12'00'

Principal Supervisor's signature

16/09/13

Date