

Review

Not peer-reviewed version

Phylogeny of Land Plants: a Case Study of Deep-Time Reconstructions

[Yin-Long Qiu](#) * and [Brent D. Mishler](#)

Posted Date: 2 July 2024

doi: 10.20944/preprints202407.0135.v1

Keywords: land plants; phylogeny; morphology; sequence characters; genome structural characters; congruence



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Review

Phylogeny of Land Plants: A Case Study of Deep-Time Reconstructions

Yin-Long Qiu ¹ and Brent D. Mishler ²

¹ The Herbarium and Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, USA

² University and Jepson Herbaria and Department of Integrative Biology, University of California, Berkeley, USA; bmishler@berkeley.edu

* Correspondence: ylqiu@umich.edu

Abstract: Relationships of the three bryophyte lineages (liverworts, mosses, and hornworts) to vascular plants is a pivotal question in the study of origin and evolution of land plants. In recent decades, this question has been subject to intense phylogenetic analyses using morphological characters, multigene DNA sequences, and genome structural characters. A tentative consensus reached ten years ago suggested that bryophytes are a paraphyletic group, with liverworts being sister to all other land plants and hornworts being sister to vascular plants. However, several more recent nuclear phylogenomic studies have concluded that bryophytes represent a monophyletic group that is sister to vascular plants. A discussion is presented here on strengths and weaknesses of different types of characters (morphological traits, nucleotide sequences, and genome structural arrangements), and their suitability for resolving deep phylogenetic relationships. Moreover, several criteria for credible phylogenetic reconstruction are proposed. Strong statistical support for reconstructed relationships should be derived from high quality, independent characters selected for suitability to the particular question being addressed. The relationships inferred in a study should be congruent with those from as many other lines of phylogenetic evidence as possible. Any incongruities should be explicable by well-understood biological mechanisms. It is concluded that the relationships of the three bryophyte lineages to vascular plants should currently be viewed as unresolved. This is a difficult phylogenetic problem; the land plants underwent a rapid radiation a long time ago. Yet, further exploration of analytical methods and careful choice of characters should lead to eventual elucidation of diversification patterns among early land plants.

Keywords: land plants; phylogeny; morphology; sequence characters; genome structural characters; congruence

1. Theoretical Considerations in Deep Phylogenetic Analyses

While this review is focused on early land plant relationships, it is necessary to begin by pointing out that this case is but one example of a large set of similar, very difficult cases in phylogenetics. Scattered around throughout the tree of life are places where rapid radiations that occurred a long time ago are represented by only a few terminal lineages in the present. These extant lineages have been pruned by extinction, resulting in a situation where a number of long, naked phylogenetic branches trace back to very short connecting branches. These sorts of cases are referred to as "deep" phylogenetic inference, following a distinction made earlier [1] – the worst case scenario for phylogeny reconstruction. Much of the relevant phylogenetic signal, i.e., characters that changed state on the short, deep connecting branches, which can thus serve as evidence of relationships, has been erased over time. In addition, false phylogenetic signal has been generated independently on the long branches leading to the present, homoplastic matches potentially resulting in "long branch attraction (LBA)" [2].

When approaching a problem of deep phylogenetic inference it is important to think clearly about the task at hand, and what tools are needed to address that task. It is abundantly clear that there are four main clades of extant land plants: liverworts, mosses, hornworts, and vascular plants. There is uncontroversial evidence for the monophyly of each of these clades taken separately. There

is also uncontroversial evidence for their monophyly taken together (as a major branch of the tree of life known as the land plants or embryophytes). However, the relationships among those four clades are maximally controversial. Seven of the 15 possible relationships among the four lineages have received statistically significant support in published analyses of morphological, single gene, multigene, genome structural or genome-scale data (Figure 1, Table 1), with some data supporting only partially resolved relationships (Figure 2, Table 1). It is abundantly clear that remnants of an ancient radiation are being studied; whatever the shared lineages were among these four extant clades, they existed for short periods of time, a very long time ago, so much of the true signal has decayed, and the potential for false signal prone to LBA is maximal.

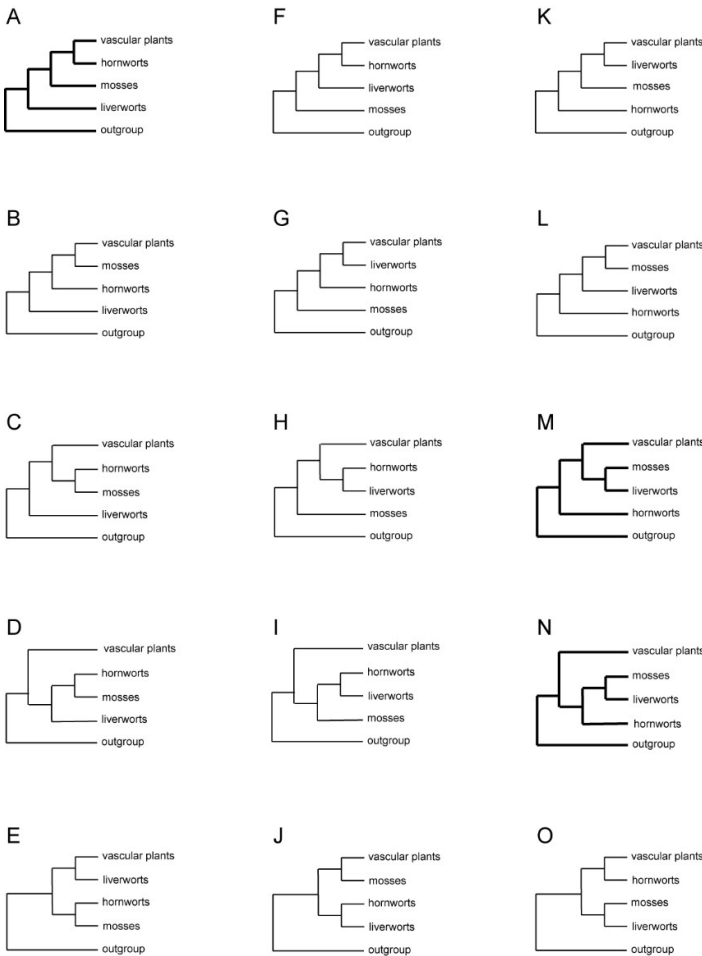


Figure 1. All 15 possible arrangements of three bryophyte lineages, vascular plants, and the outgroup. Three topologies commonly seen in studies of morphological and molecular data are shown in thick lines.

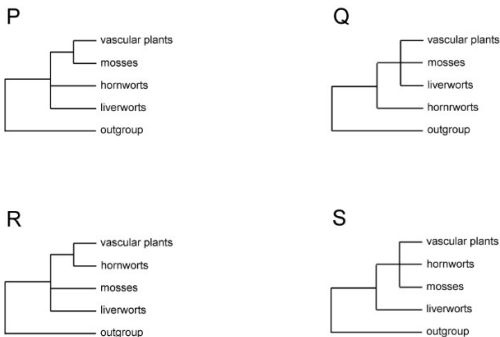


Figure 2. Four partially resolved topologies of relationships among three bryophyte lineages, vascular plants, and the outgroup, as seen in studies of some morphological and molecular data.

Table 1. Relationships of bryophyte lineages and vascular plants as inferred in phylogenetic studies reviewed in this article*.

Tree Topology	Data Source	Characters	OTUs**	Method(s) of Analysis	Reference
			Outgroup Liverworts Mosses Hornworts Vascular Plants		
Figure 1A	morphology	17 characters	O: 2 L: 1 M: 1 H: 1 V: 14	MP	[4]
	chloroplast, one DNA region: rDNA ITS	nucleotide: all positions	O: 2 L: 14 M: 20 H: 4 V: 7	ML, MP & NJ	[17]
	chloroplast, 67 genes	<u>nucleotide: all positions</u>	O: 6 L: 1 M: 1 H: 1 V: 27	ML & MP	[23]
		<u>nucleotide: 1st-3rd positions</u>			
		<u>nucleotide: 2nd-3rd positions</u>		MP	
		<u>nucleotide: 3rd positions</u>			
	multigene (cp-atpB, cp-rbcL, cp-SSU cp-LSU, mt-LSU, and nu-18S)	nucleotide: all positions	O: 9 L: 47 M: 43 H: 6 V: 88	ML & MP	[23]
	chloroplast, 73 genes	<u>nucleotide: all positions</u>	O: 1 L: 1 M: 1 H: 1 V: 16	ML & MP	[25]
		nucleotide: 1 st -2 nd positions			
	chloroplast, 17 genes and associated noncoding regions	nucleotide: all positions	O: 5 L: 6 M: 14 H: 3 V: 15	ML & MP	[30]
	multigene (cp-atpB, cp-rbcL, mt-atp1, mt-nad5, and nu18S)	nucleotide: all positions	O: 23 L: 57 M: 56 H: 5 V: 492	ML	[31]
chloroplast, 51 genes	nucleotide: 1 st -2 nd positions	O: 4 L: 1 M: 1 H: 1 V: 13	MP	[32]	
chloroplast, 78 genes	<u>nucleotide: all positions</u> nucleotide: RY-coded sequences	O: 32 L: 3 M: 2	ML	[34]	

Figure 1B			H: 1 V: 322	ML (RAxML)	[35]
	chloroplast, 72 genes	nucleotide: all positions, the 45879 matrix	O: 10 L: 2 M: 2 H: 1 V: 15		
	chloroplast, 72 genes	nucleotide: all positions	O: 82 L: 6 M: 8 H: 2 V: 3556		
	mitochondrion, 40 genes	nucleotide: all positions (the 9013 & 7210 matrices)	O: 13 L: 3 M: 2 H: 2 V: 5		
	mitochondrion, 41 genes	nucleotide: all positions under a composition heterogeneous model	O: 5 L: 6 M: 19	BI <hr/> ML	[40]
		amino acid: under a composition homogeneous model	H: 2 V: 28		
	mitochondrion, 36 genes	nucleotide: codon-degenerate data under a homogeneous composition model amino acid, under a homogeneous composition model	O: 5 L: 4 M: 5 H: 2 V: 10	BI	[41]
	chloroplast, 47 intron positions, presence/absence, inversion, translocation, or duplication of gene(s)	genome structural characters	O: 2 L: 1 M: 2 H: 1 V: 22		
	nucleus, 18560 orthogroups	gene duplication data	O: 0 L: 1 M: 4 H: 2 V: 17	STRIDE	[52]
	morphology	37 characters	O: 2 L: 1 M: 1 H: 1 V: 1		
	morphology	113 characters	O: 4 L: 7 M: 5 H: 2 V: 9	MP	[6]

Figure 1F	chloroplast, 72 genes	nucleotide: all positions, the 45879 matrix	O: 10 L: 2 M: 2 H: 1 V: 15	ML (nhPhyML)	[35]
	chloroplast, 83 genes	nucleotide: all positions	O: 13 L: 2 M: 2 H: 2 V: 11	ML	[37]
	mitochondrion, 41 genes	nucleotide: all positions under a homogeneous composition model	O: 5 L: 6 M: 19 H: 2 V: 28	ML	[40]
	mitochondrion, 36 genes	nucleotide: all positions under a homogeneous composition model	O: 5 L: 4 M: 5 H: 2 V: 10	BI	[41]
	nucleus, 18560 orthogroups	gene duplication data	O: 0 L: 1 M: 4 H: 2 V: 17	ALE	[52]
Figure 1G	mitochondrion, 36 genes	nucleotide: all positions under a heterogeneous composition model	O: 5 L: 4 M: 5 H: 2 V: 10	BI	[41]
Figure 1M	morphology	125 characters	O: 1 L: 7 M: 5 H: 3 V: 7	MP	[9]
	nucleus, 18S	nucleotide: all positions	O: 3 L: 7 M: 9 H: 2 V: 6	MP	[11]
	nucleus, 18S	nucleotide: all positions	O: 7 L: 8 M: 20 H: 2 V: 31	MP	[12]
	multigene (cp-SSU, mt-SSU, nu-SSU, cp-rbcL)	nucleotide: all positions		ML	[29]
		nucleotide: all positions with rate variation accounted for	O: 2 L: 2 M: 3 H: 2 V: 21	ML	
		nucleotide: RY-coded sequences for cp-rbcL 3 rd positions		MP	

Figure 1N	mitochondrion, 36 genes	nucleotide, codon-degenerate data under a heterogeneous composition model amino acid under heterogeneous composition model	O: 5 L: 4 M: 5 H: 2 V: 10	BI	[41]
	nucleus, 674 genes	nucleotide, 1 st -2 nd positions	O: 22 L: 6 M: 11 H: 2 V: 62	ML	[44]
	nucleus, 100 genes	nucleotide: all positions	O: 6 L: 4 M: 4 H: 2 V: 10	ML	[48]
	nucleus, 410 genes	nucleotide: all positions	O: 235 L: 22 M: 42 H: 10 V: 869	ML	[49]
	nucleus, 18560 orthogroups	gene duplication data	O: 0 L: 1 M: 4 H: 2 V: 17	ALE	[52]
	morphology	72 characters (spermatogenesis)	O: 1 L: 4 M: 3 H: 2 V: 12	MP	[9]
	chloroplast, 51 genes	amino acid	O: 4 L: 1 M: 1 H: 1 V: 13	ML	[32]
	chloroplast, 83 genes	amino acid	O: 13 L: 2 M: 2 H: 2 V: 11	BI	[37]
	chloroplast, 57 genes	nucleotide: 1 st -2 nd positions (LogDet correction for compositional bias) amino acid (paralinear correction for compositional bias)	O: 1 L: 1 M: 1 H: 1 V: 14	NJ	[38]
	chloroplast, 72 genes	nucleotide: 1 st -2 nd positions	O: 82 L: 6 M: 8	ML	[55]

		H: 2 V: 3556		
nucleus, 424 genes	gene trees	O: 22 L: 6 M: 11 H: 2 V: 62	AA	[44]
chloroplast, 78 genes	amino acid	O: 275 L: 28 M: 46 H: 11 V: 1519	ML	[101]
nucleus, 852 genes	gene trees	O: 22 L: 6 M: 11 H: 2 V: 62	BSI	[46]
nucleus, 100 genes	nucleotide: codon-degenerate data	O: 6 L: 4 M: 4 H: 2 V: 10	ML	[48]
	amino acid		ML BI	
nucleus, 410 gene families	gene trees	O: 235 L: 22 M: 42 H: 10 V: 869	AA	[49]
nucleus, 151 orthologs	amino acid	O: 19 L: 23 M: 20 H: 7 V: 93	ML	[50]
nucleus, 1440 genes	nucleotide: codon degenerate data	O: 20 L: 18 M: 44 H: 9 V: 32	BI & ML	[51]
	amino acid			
nucleus, 160 genes	gene trees		AA	[52]
	amino acid	O: 23 L: 24 M: 20 H: 9 V: 101	BI & ML	
nucleus, 18560 orthogroups	gene duplication data	O: 0 L: 1 M: 4 H: 2 V: 17	STRIDE	
			ALE	
chloroplast, 67 genes	nucleotide: 1 st positions	O: 6 L: 1 M: 1 H: 1 V: 27	ML	[23]
	nucleotide: 1 st -2 nd positions		ML & MP	

Figure 10

	chloroplast, 51 genes	nucleotide: fourfold degenerate sites (LogDet correction for compositional bias)	O: 4 L: 1 M: 1 H: 1 V: 13	NJ	[32]
	chloroplast, 49 genes	nucleotide: all positions	O: 6 L: 1 M: 2 H: 1 V: 33	BI & ML	[33]
	chloroplast, 78 genes	<div>nucleotide: 1st-2nd positions</div> <div>amino acid</div>	O: 32 L: 3 M: 2 H: 1 V: 322	ML	[34]
	chloroplast, 72 genes	nucleotide: the 45879 matrix	O: 10 L: 2 M: 2 H: 1 V: 15	BI (PhyloBayes)	[35]
	chloroplast, 88 genes	<div>nucleotide: all positions</div> <div>amino acid</div>	O: 18 L: 1 M: 2 H: 1 V: 6	BI & ML	[36]
	nucleus, 142 genes	gene trees	O: 22 L: 6 M: 11 H: 2 V: 62	BSI	[46]
	Figure 2P	morphology	O: 2 L: 3 M: 5 H: 2 V: 3	MP	[7]
Figure 2Q	multigene (nu-18S & mt-SSU)	nucleotide: all positions	O: 4 L: 2 M: 5 H: 4 V: 11	MP	[9]
	mitochondrion, SSU	nucleotide: all positions	O: 1 L: 2 M: 4 H: 3 V: 10	MP	[19]
Figure 2R	chloroplast, rbcL	nucleotide: all positions	O: 2 L: 25 M: 4 H: 3 V: 6	ML & MP	[15]
	chloroplast, 67 genes	<div>nucleotide: 1st positions</div> <div>nucleotide: 3rd positions</div>	O: 6 L: 1	<div>MP</div> <div>ML</div>	[23]

			M: 1 H: 1 V: 27		
	chloroplast, 73 genes	amino acid	O: 1 L: 1 M: 1 H: 1 V: 16	ML & MP	[25]
	chloroplast, 83 genes	nucleotide: codon-degenerate data	O: 13 L: 2 M: 2 H: 2 V: 11	ML	[37]
	mitochondrion, 72 genes	nucleotide: all positions (the 9013 & 7210 matrices)	O: 13 L: 3 M: 2 H: 2 V: 5	ML	[39]
Figure 2S	mitochondrion, 28 intron positions	genome structural characters	O: 3 L: 2 M: 2 H: 3 V: 6	MP	[23]

* Tree topologies referred to are shown in Figures 1 and 2. For reconstructed relationships among bryophytes lineages and vascular plants, they are all supported with at least a bootstrap value of 50%, a decay index of 3, or a posterior probability of 0.8; exceptions were made in a few cases for morphological results. Unorthodox topologies produced by chloroplast genome amino acid data are not included here, such as monophyletic bryophytes embedded among pteridophytes (sister to lycophytes) and the first split in land plants being between seed plants and the rest (bryophytes, lycophytes and ferns) [55] or between monilophytes and the rest (seed plants + lycophytes-bryophytes) [37]. Abbreviations: AA – ASTRAL analysis, BI – Bayesian inference, BSI – Bayesian supertree inference, ML – maximum likelihood analysis, MP – maximum parsimony analysis, NJ – neighbor-joining analysis; cp – chloroplast, mt – mitochondrial, nu – nuclear; RY – purine (R) and pyrimidine (Y) recoded nucleotide sequence. ** The operational taxonomic units (OTUs) are usually species but sometimes are infra- or supra-specific units as indicated in original references.

In approaching this problem it is thus necessary to pay careful attention to finding evidence that is likely to preserve traces of the actual branching pattern. Additional, judicious taxon sampling, always a good idea when possible, does not work in these cases (for molecular data at least), given the patterns of extinction that have generated these long, naked branches. Deep phylogenetic inferences are not going to be solved simply by adding more and more characters, the misguided trend in recent "phylogenomic" analyses. If one is not careful, adding more rapidly evolving genes to such an analysis is simply adding more LBA and potentially making one increasingly confident in the wrong relationships. Smaller, more carefully curated datasets are needed instead. As discussed earlier [1,3], clocklike markers are particularly undesirable in these deep analyses, since they keep on ticking and remove the signal for the deep branches. Also undesirable are markers with a high rate of change relative to the short intervals of deep time we are interested in capturing, for example, most nuclear and some organellar genes. These markers not only tend to lose the true signal, and more importantly they gain false signal leading to LBA. Instead, the way forward is selecting the right characters for the task at hand: those that have changed slowly and episodically (to preserve true phylogenetic signal unobscured by further changes), ideally with many character states (to lower the chance of parallel changes causing false signal).

2. Previous Studies on Early Land Plant Relationships

2.1. Morphological and Early Molecular Studies

Formal phylogenetic analyses of land plants focusing on relationships among the major lineages began with cladistic analyses of mostly morphological characters in the early 1980s [4,5]. The three bryophyte lineages were found to be paraphyletic to vascular plants, but how they were related to each other and to vascular plants differed in these seminal studies. In one study, liverworts were sister to all other land plants, and mosses were sister to hornworts plus vascular plants [4] (Figure 1A, Table 1). In the other study, liverworts were sister to the rest of land plants, with hornworts sister to mosses plus vascular plants [5] (Figure 1B). Two subsequent cladistic studies of morphological data sets reached the same or similar conclusion as the latter study [6,7] (Figures 1B and 2P). However, two analyses of spermatogenesis characters both uncovered a topology showing bryophyte monophyly with hornworts sister to a clade of liverworts and mosses, but with bryophytes either nested in [8] or sister to vascular plants [9] (Figure 1N). Furthermore, when some of the spermatogenesis characters were combined with developmental and morphological characters in a parsimony analysis, bryophytes were found to be paraphyletic to vascular plants, with hornworts being sister to all other land plants and a clade or grade of liverworts-mosses sister to vascular plants [9] (Figure 1M).

As molecular systematics started to develop in the early 1990s, analyses of DNA sequences of mostly single genes, from all three plant genomic compartments, dominated the first wave of phylogenetic reconstruction of early land plants: nuclear 18S and 26S rDNAs [6,10], 18S [11–13], chloroplast *rbcL* [14–16], rDNA ITS [17], mitochondrial *cox3* [18], and SSU rDNA [19]. While these genes were chosen for their putatively slow evolutionary rate – a feature critically important for reconstructing ancient phylogenetic branching events, taxon sampling schemes, including outgroup choices, varied widely due to different goals of these studies (Table 1). Sometimes, molecular data were analyzed in combination with morphological characters [6,9]. As a result, a dazzling array of relationships among the four major lineages of land plants were obtained, mostly with low to moderate statistical support. Three general patterns emerged from these studies. First, bryophytes were paraphyletic to vascular plants in most analyses [6,9–12,15,17–19]. Second, hornworts were sister to either all other land plants in analyses of nuclear 18S rDNA [11,12] (Figure 1M, Table 1) and mitochondrial SSU rDNA [19] (Figure 2Q), or just vascular plants as seen in analyses of chloroplast *rbcL* [15] (Figure 2R). Third, liverworts were sister to all other land plants in analyses of chloroplast rDNA ITS [17] (Figures 1A).

Because DNA sequences have only four (or five if missing information is counted) character states in evolution, they can suffer from a problem where homoplastic mutations cause LBA [2] when taxon sampling is sparse (which was the case for most single gene analyses in the 1990s) or sampled genes evolved too fast relative to the depth of reconstruction. To avoid this predicament, slow-changing structural characters such as intron gains [20] or large DNA inversions [21] in organellar genomes have been explored for their phylogenetic utility. A survey of group II introns in the mitochondrial genome of over 350 species of land plants found that three introns in *cox2* and *nad1* genes were present in all major lineages of land plants except liverworts, and that they were also absent in the outgroup taxa green and red algae. These data were interpreted as evidence of independent gains of the three introns in the early stage of land plant evolution, supporting liverworts as the sister lineage to the rest of land plants as well as bryophytes being paraphyletic to vascular plants [22]. These conclusions were reinforced later by an expanded analysis of absence/presence of 28 group II introns in 13 land plant and three related green alga mitochondrial genomes [23] (Figure 2S, Table 1). Moreover, three versions of a parsimony analysis of chloroplast genome structural characters, the first with 40 characters from 19 taxa, the second with 42 characters from 21 taxa, and the third with 47 characters from 28 taxa, all reached the same conclusions [24–26], with further evidence on hornworts being sister to vascular plants (Figure 1A, Table 1).

As automated sequencing became available in the late 1990s, and theoretical simulation and empirical exploration also showed that longer sequences were needed to reconstruct complex phylogenies [27,28], multiple genes began to be used for unravelling difficult phylogenetic patterns. In a study analyzing chloroplast *rbcL* and SSU rDNA from all three plant genomic compartments using both parsimony and maximum likelihood methods, bryophytes were found to be paraphyletic to vascular plants, with hornworts being sister to all other land plants and liverworts and mosses

together forming a clade sister to vascular plants [29] (Figure 1M). This study took special care to deal with site-to-site rate variation in sequence evolution as well as the third codon position saturation of the protein-coding *rbcL*. The reconstructed relationships received moderate to high statistical support in most of the analyses. However, strength of the conclusions was weakened by sparse taxon sampling used in the study, especially in bryophytes, as only thirty taxa were used to cover the entire land plants and green algae, and merely seven bryophyte taxa were sampled (Table 1). Another multi-gene study, sampling chloroplast *atpB*, *rbcL*, SSU and LSU rDNAs, mitochondrial LSU rDNA, and nuclear 18S rDNA, made a significant improvement in taxon sampling by including 193 taxa of land plants and green algae, with 96 of them covering the full diversity of bryophytes [23] (Table 1). In both parsimony and maximum likelihood analyses, bryophytes were shown to be paraphyletic to vascular plants, with liverworts being sister to all other land plants and hornworts being sister to vascular plants (Figure 1A). As described above, analyses of absence/presence of 28 group II introns in 13 land plants and green algae were also included in this study, which supported bryophyte paraphyly and liverworts being sister to all other land plants with relatively high bootstrap values [23]. In a third multi-gene study, which sampled 17 chloroplast genes and associated non-coding regions from 43 taxa of land plants and green algae, the same relationships among three bryophyte lineages and vascular plants as in the last study were reconstructed using both parsimony and maximum likelihood methods, all with high statistical support [30].

Finally, an analysis of five genes, chloroplast *atpB* and *rbcL*, mitochondrial *atp1* and *nad5*, and nuclear 18S rDNA, was carried out using a comprehensive matrix with intensive and unbiased family level sampling across land plants, including 633 land plant taxa plus 23 green alga species as outgroups [31] (Table 1). Each of the five genes was analyzed individually using a maximum likelihood method, as were concatenated matrices of the two genes from the chloroplast and mitochondrial genomes respectively. These analyses aimed to ensure that all problematic sequences were removed and that effects of any gene- and genome-specific sequence evolutionary phenomena, such as RNA editing and base-composition bias, could be revealed. The five-gene matrix was then analyzed with the knowledge of single gene data set limitations. In none of the single or multi-gene analyses did bryophytes form a monophyletic group. The final analysis of the concatenated five-gene matrix showed that bryophytes were paraphyletic to vascular plants, with strong support for liverworts as sister to all other land plants [31] (Figure 1A).

2.2. Large-Scale Phylogenomic Studies

Advancement of both sequencing technology and informatics in the early 2000s ushered in the era of phylogenomics. The chloroplast genome, thanks to its small size and conservative structural evolution, was the first explored among three plant cellular compartments for reconstructing land plant phylogeny. Bryophytes were shown to be paraphyletic to vascular plants by most studies that analyzed all or most of the genes in the chloroplast genome [23,25,32–37] (Table 1), with liverworts and mosses always placed near each other as either a clade or a grade, and hornworts being sister to vascular plants, depending on which portions of the data were analyzed with which methods (Figures 1A, 1F, 1M, 1N, 1O, & 2R). Statistical support was generally high for the resolved relationships. However, two of these studies [32,37] plus a third one [38] found topologies in which bryophytes formed a monophyletic group, being either sister to [32,37,38] or nested among vascular plants [37]. Two things worth noting are that the number of bryophyte species used in all these studies was very small, only one to at most three from each of the three lineages, and that all genes were included regardless of their suitability (due to evolutionary rate) to this deep reconstruction problem (Table 1).

In phylogenomic analyses of mitochondrial genome sequences, bryophytes were also found to be paraphyletic to vascular plants, with either liverworts sister to the rest of land plants and hornworts being sister to vascular plants [39,40] (Figure 1A), or hornworts sister to the rest of land plants and liverworts and mosses together forming a clade sister to vascular plants [41] (Figure 1M). Most of these relationships received high support. Taxon sampling was sparse in two studies [39,41] but relatively extensive in the third one [40] (Table 1). Again, all genes were included with no concern as to their suitability for the questions at hand.

In all these organellar phylogenomic studies, various treatments of data were used in an attempt to moderate potential distorting effects of some evolutionary forces on reconstructing the underlying

organismal phylogeny. The most common approaches were the removal of the third codon positions in nucleotide sequences and use of amino acid sequences of protein-coding genes when nucleotide substitution saturation was suspected to cause erosion of phylogenetic signal [23,25,32,34,36–41]. A less common and more nuanced way of dealing with this saturation problem was the removal of fast-evolving codons or amino acid positions entirely after they were ranked by certain criteria [35,39]. Base compositional bias in nucleotide sequence evolution had also been considered [32,38] and a method of data transformation, LogDet [42,43], was used to correct the problem. Deletion of taxa that had excessively long branches was done to try to alleviate the LBA problem [33]. Finally, different methods of phylogenetic reconstruction were used, ranging from the commonly used maximum likelihood and parsimony methods, to less frequently used Bayesian and neighbor-joining analyses. Model-fitting was performed in two studies when model-based methods were used [37,41]. All these practices seemed to be sensible according to some presumably understood molecular evolutionary mechanisms. Nevertheless, whether they have led to successful reconstruction of the phylogenies from the given data is difficult to conclude (see below). Many of these practices clearly did not work, because many different relationships were reconstructed when at most one of them could be correct.

Two of the chloroplast phylogenomic studies discussed above were performed in conjunction with analyses of a second matrix of different types of characters, i.e., genome structural characters such as intron gains and gene order changes from the same [25] or a different genomic compartment [23]. This kind of analyses was designed to overcome some potential problems associated with sequence characters, with their limited number of character states. In both studies, the results from analyses of the two types of data were congruent, showing bryophytes as a paraphyletic group to vascular plants with liverworts being sister to all other land plants (Figures 1A & 2S, Table 1).

The scale of phylogenomics and its impact on studies of relationships among major land plant lineages could really be felt with the increase of sequenced nuclear genomes and transcriptomes, as the number of genes that were analyzed increased exponentially in comparison to all other molecular systematic studies. In the first such study, a whopping 852 nuclear genes from 103 species of land plants and green algae were analyzed in 69 analyses with different permutations of data matrix and phylogenetic method, including supermatrix, supertree, and coalescent-based approaches, maximum likelihood and Bayesian methods, partitioned and unpartitioned analyses, and amino acid versus DNA alignments [44]. Among various results obtained, two different topologies stood out. One, from a maximum likelihood analysis of concatenated alignments of the first and second codon positions of 674 genes from 103 species, showed hornworts being sister to all other land plants and liverworts and mosses together forming a clade sister to vascular plants (Figure 1M, Table 1). The other, in a coalescent-based tree from an ASTRAL analysis [45] of 424 gene trees obtained through analyzing the first and second codon position alignments, showed all three bryophyte lineages forming a monophyletic group sister to vascular plants, with hornworts sister to liverworts plus mosses (Figure 1N, Table 1). Both these two topologies had 100% bootstrap support for nearly all relevant relationships. The data from this study were re-analyzed later by two groups of authors using different methods. In one study [46], the bryophyte monophyly topology (Figure 1N) was recovered with strong support when gene trees of 852 single copy orthologs were analyzed using the Bayesian supertree inference [47]. However, a different topology, in which a clade of liverworts and mosses was sister to the clade of hornworts and vascular plants (Figure 1O), was obtained when gene trees of 148 orthologs with the least heterogeneous base composition were analyzed with the same method, also with strong support. In the other study [48], while re-analyses of the data obtained similar results to the original study [44], a reduced matrix was made using some putatively stringent criteria to remove some genes and taxa so as to minimize effects of composition heterogeneity, substitution saturation, missing data, and %GC deviation. This matrix consisted of only 100 genes and 26 taxa. In both maximum likelihood and Bayesian analyses of the amino acid version of this reduced matrix, as well as a maximum likelihood analysis of the codon-degenerate re-coding nucleotide version of the matrix, bryophytes were found to form a monophyletic group sister to vascular plants (Figure 1N, Table 1). On the other hand, in a maximum likelihood analysis of the nucleotide version of the matrix, with all positions included and unaltered, bryophytes were found to be paraphyletic to vascular plants, with hornworts sister to all other land plants (Figure 1M, Table 1). Almost all of the relationships had perfect bootstrap or Bayesian probability support [48].

In a second major nuclear phylogenomic study from the 1000 Plant Transcriptome Project (1KP Project), an even more impressive amount of data was analyzed [49]. A tree was obtained from an ASTRAL analysis of 410 single-copy nuclear gene families from 1090 species of green plants plus 63 species of outgroup taxa, showing bryophytes together as a monophyletic group sister to vascular plants, with hornworts being sister to liverworts plus mosses [49] (Figure 1N, Table 1). Some of the data from this study were again re-analyzed by other authors. In one study [50], a matrix consisted of concatenation of 151 orthologs conserved across 162 green plant genomes and transcriptomes was assembled from the 1KP Project data source. A maximum likelihood analysis of this matrix showed bryophytes as a monophyletic group sister to vascular plants, with hornworts being sister to liverworts plus mosses (Figure 1N, Table 1). These relationships all received 100% bootstrap support [50]. In the second study, 1440 genes from 120 streptophyte species were gathered from the 1KP Project as well as other public data sources and the authors' own sequencing project [51]. Maximum likelihood and Bayesian analyses of amino acid sequences recovered bryophyte monophyly with strong support, as did a coalescent species tree from an ASTRAL analysis [51] (Figure 1N, Table 1). A third study, using data from both 1KP Project and other public sources, took two different and complementary approaches to investigate the rooting issue of land plant phylogeny [52]. One consisted of conventional supermatrix and supertree analyses of amino acid sequences of 160 single-copy gene families from 154 land plant and 23 alga species. When the tree was rooted using the algae as the outgroup, bryophytes were monophyletic and sister to vascular plants with high support across all analyses (Figure 1N, Table 1). The other approach was to infer root placement on a data set of 24 high-quality embryophyte genomes without the inclusion of an algal outgroup, using two methods: ALE (amalgamated likelihood estimation) [53] and STRIDE (species tree root inference from gene duplication events) [54]. The former identified three possible roots (the first diverging lineage) of the land plant phylogeny: hornworts, mosses, and all bryophytes (Table 1). The latter assigned 0.2%, 39.9% and 59.8% probability to hornworts, all bryophytes, and liverworts separately as the root in three possible scenarios (Table 1). Paradoxically, those authors concluded that bryophytes together represented the root of land plant phylogeny despite these variable results. Clearly, rooting of the land plant phylogeny remains one of the most difficult issues. LBA involving the long branch to the outgroup is just as problematic as LBA among the four lineages of interest in the ingroup.

Thus, despite rapidly increasing amounts of raw data, no conclusion seems to be in sight for resolution of this recalcitrant problem. More raw data is not necessarily the answer; in fact as indicated in the first section, adding more and more inappropriate data may make things worse. Several studies have also made efforts to fine-tune the data before phylogenetic analyses, for chloroplast [32,34,37,38], mitochondrial [40,41], and nuclear phylogenomic matrices [44,46,48]. The dilemma is how a particular data treatment and its consequential effect on the result can be justified without reference to other information. One particular issue that has received much attention is base composition heterogeneity in DNA sequences. When two lineages share similar base compositions, it is impossible to know whether they are due to shared common ancestry or convergence. Hence, using models that only accept base composition homogeneity in analyses could prematurely discard phylogenetic signal in the case of common ancestry-derived similar base compositions. It is thus not surprising to see when one phylogenomic data set is analyzed by sub-data sets and/or different methods, different and conflicting topologies are produced [23,32,35,37,40,41,44,46,48,55].

We are clearly not making progress towards a consensus about this important question over time. Thus, we need to step back and think more clearly about what sort of data is needed to make progress. Not all raw nucleotide sequences should be regarded as useful data to use at this level, and other types of information such as genome structural characters and morphology should be explored and included in analyses as much as possible. In the following section, several types of characters are discussed that have been used in phylogenetic studies over the last several decades, and their strengths and weaknesses are evaluated.

3. Properties of Different Types of Characters for Reconstructing Deep Relationships of Land Plants

Following on from the discussion in the initial section of this paper, we will focus here on the strengths and weaknesses of different categories of characters for "deep" reconstruction problems, as

distinguished from "shallow" reconstruction problems [1]. The former are characterized by long extant branches, which have been much pruned by extinction, connected in the deep past by short branches of primary phylogenetic interest; the latter are characterized by relatively recent, similar-length branches of primary phylogenetic interest, little pruned by extinction. These of course occupy the extremes of a spectrum, but our focus is on the extremes in these discussions given that land plant relationships are clearly at the deep end of the spectrum. The types of characters suited to that extreme are quite different from those suited to the shallow end (e.g. an analysis of a plant radiation over the last few million years where most of the lineages are still around to be sampled).

3.1. Morphology

Morphological data can often be quite difficult to score objectively for use in shallow phylogenetic studies, given that differences between very closely related taxa can be hard to quantify. Thus, they have tended to fall out of favor in such studies as compared to DNA sequences, and at most are mapped onto the tree built from molecular data. However, in deep phylogenetic studies some morphological characters are easy to score objectively, if they are highly conserved. For example stomates are quite easy to compare among the major groups of land plants [56]. While we do not have good evolutionary models for such complex characters, and they are relatively few in number as compared to nucleotide characters, which could both be considered drawbacks, morphological characters have strengths that more than compensate.

Morphological characters have more states (e.g., there are many more ways to modify stomates than there are possible point mutations at a given site) -- that alone reduces the possibility of LBA since the more possible states, the less likely for a homoplastic match to occur at random [57]. Homology can be easier to hypothesize in morphology, given its 3-dimensional positional information plus ontogeny, as compared to the one-dimensional positional information present for use in DNA sequence alignments. Furthermore, morphological characters tend to show episodic evolution, i.e., major changes followed by periods of stasis. A clock-like marker is ideal for a shallow phylogenetic study, but the ideal character for a deep phylogenetic study is a broken clock, which quit ticking at some point in the past and retains its state until the present [3]. Finally, having morphological characters in the matrix can allow the addition of fossil taxa to the matrix, if those characters are observable in them, which is a potential way to break up long branches and reduce LBA.

Some have objected to morphological characters because they are more likely to be subject to selection than DNA sequence data, but this is an unfair and illogical comparison -- certainly genes that are conserved enough to be useful in deep phylogenetic studies must be under tremendous selection pressure as well. Having functional ribosomes and a working RuBisCO enzyme is certainly more important to fitness of a plant than a certain shape of stomata. Convergence at the molecular level just needs to be considered as seriously as with morphology. Conflicting topologies among different datasets is not only due to LBA, lineage sorting, or horizontal transfer (the three processes most often invoked), it can just as well be due to selectively-driven convergence (a process that is completely ignored by current analytical methods). Our ability to find the true history in the face of all these confounding processes depends on the inclusion of as many distinctly different data types as possible. Each type may have its issues, but at least if the issues with one data type are independent of the issues of other data types, there is hope of finding a common historical pattern shared by all. When morphology is left out, an important category of distinct data is left out.

3.2. Nucleotide Sequences

Nucleotide sequences represent the most abundant characters for phylogenetic reconstruction. They include coding and non-coding regions that make up the genome and are directly inherited (unlike phenotypic features). For protein coding genes, their amino acid sequences are often used in analysis to minimize the effect of long branch attraction caused by the limited number of character states in nucleotide sequences. Both DNA and protein sequences have the advantage of clearly defined states. Well understood chemistry of the four nucleotides and 20 amino acids also allows development of evolutionary models relatively easily. In recent years, automated sequencing has made acquisition of this colossal body of potential phylogenetic information within the reach of systematists.

The massive number of sequence characters, however, poses difficulties for sorting out phylogenetic signals from noise generated by LBA, convergence, lineage sorting, and reticulate evolution, as a genome has been shaped by many forces over evolutionary time. One manifestation of this complexity is evolutionary rate heterogeneity in different genes and at different positions in DNA and protein sequences [58,59] as well as across lineages of organisms [60]. Thus, although evolution at DNA and protein levels may seem to be easier to model than at other levels, it is no simple task to develop models to fit real data. Specifically, several factors may contribute to the data-model fit challenge. First, a vast majority of nuclear genes evolve much faster than chloroplast and mitochondrial genes used in earlier years of molecular systematics. This rate increase complicates an already difficult rate heterogeneity problem. Second, most phylogenomic data sets are much larger than multigene data sets in character number but not in taxon number, causing under-sampling of intermediate states for faster evolving characters. Third, because most sequenced genomes are from model organisms or economically important species, taxon sampling in phylogenomic studies tends to skew densely toward certain groups yet sparsely in phylogenetically critically positioned taxa, resulting in highly uneven representation of diversity in the study group. As a result, many minor insignificant but system-wide factors that generate only negligible random errors in single to multigene matrices can exert their cumulative influence to produce systematic errors leading to statistically well supported but incorrect results in phylogenomic studies [61].

Recent studies have shown that algorithms employing models that incorporate higher-level genome and proteome organizational information, such as protein secondary structures [62,63] and amino acid solvent accessibility [64], tend to outperform those that use simple models in dealing with systematic errors in phylogenomic data. One study of animal phylogeny found that by using a site-heterogeneous infinite mixture model CAT-GTR (which is able to adapt to the complexity actually present in the data) [63], and recoding conventional amino acid sequences into partitioned phylogenomic data by masking biochemically similar and/or highly exchangeable amino acids, analyses of two data sets that contain 89 genes from 62 taxa and 117 genes from 76 taxa both placed sponges as the sister lineage to all other animals, as most previous studies had shown [65]. The competing hypothesis advocated by two more simplistic phylogenomic studies, which placed comb jellies as the sister lineage of all other animals [66,67], turned out to be a result of LBA. In another study [68], which analyzed a chloroplast data set of 61 proteins from 24 land plants [69], both *Amborella* and *Nymphaea* were identified as members of a clade sister to all other angiosperms when three models that considered distinct structural and functional constraints of protein evolution were used (JTT+C20+F+gamma, JTT+PMSF2+gamma, and CAT+GTR+gamma). In contrast, *Amborella* alone was found to occupy such a position when a conventional model (JTT+F+gamma) was used [68]. The former result is consistent with the consensus that has emerged from analyses of slow-evolving mitochondrial genes [70] and nuclear phylogenomic data with a coalescent model [71], whereas the latter likely represents an artifact generated by LBA. Both of these cases involve lineages stemming from ancient radiations, and demonstrate that when analyzing large phylogenomic matrices to unravel phylogenetic patterns at highly compressed deep nodes, models incorporating higher-level structural features of protein sequences may offer better resolution than conventional DNA sequences.

Another manifestation of complexity of the nuclear genome is that a vast majority of nuclear genes exist in families of varying sizes, which arose from repeated genome duplications throughout the history of eukaryotes [72]. These genes have duplicated copies that coalesce back in time at different points. Functional divergence and selective retention/loss of the copies after duplication [73] produce genes that fail to meet orthology and evolutionary rate constancy criteria of ideal phylogenetic markers. Nevertheless, theoreticians have developed coalescence conceptual frameworks that allow the process of gene duplication and species/lineages cladogenesis to be modeled and reconciled [74–77]. Consequently, gene copies that are not strictly orthologous may provide information for species tree reconstruction. As mentioned above, use of a coalescent method in a phylogenomic analysis uncovered the presumably correct position of *Amborella* in the angiosperm phylogeny [71]. However, it has been found that a protocol of “statistical binning”, which seeks to overcome gene tree estimation error by concatenating loci of different coalescent histories into longer multi-locus supergenes, is operating under an assumption/model that is often

violated, as >92% of supergenes comprise discordant loci [78]. In an ideal situation, the supergene tree set should be an accurate estimate of the true underlying gene tree distribution. In this regard, it is not certain whether the bryophyte monophyly topology obtained by two phylogenomic studies of land plant phylogeny in coalescent model-based analyses [44,46] was caused by this model violation. More theoretical clarification on data matrix parameters is needed before any firm conclusions can be drawn on conditions under which coalescent models can or cannot perform optimally to reconcile the discrepancy between gene trees and species trees. What should be added here is that organellar genes generally do not have this homology-orthology confounding problem, as they are all single-copy genes that date back to at least the eubacterial ancestors of both chloroplasts and mitochondria.

Model violation has been reported to be rather prevalent in phylogenetic studies [78,79]. This phenomenon may be caused by more higher-level structural organization features of the genome/proteome than codons and exons/protein secondary structures, e.g., protein tertiary structures and interaction networks [80,81], or paralogy-orthology [82] mixing of many genes in the nuclear genome. Further challenges lie ahead to develop more realistic models that can handle large complex phylogenomic data sets as more taxa and genes become available for analysis.

In light of the crucial role played by models in analyses of large phylogenomic matrices, it may be helpful to examine how model choices were made in several land plant phylogenomic studies reviewed earlier, so as to understand why they almost all produced the bryophyte monophyly topology. Three studies used the CAT-GTR model, but because it was computationally costly [68,83], adjustment was made to cut taxon and/or character numbers [44,46,48]. This compromise between model and data effectively reduced the matrix size considerably and might have undermined the merit of genome-wide extensive character sampling of phylogenomic studies. Further, no protein secondary structural constraints were introduced together with the model in their analyses. Four other studies used less sophisticated models, GTR+gamma for nucleotide data [49] and JTT, JTTF, and JTTDCMUT [49], LG+C60+G+F [50,52], and LG+C20+F+R5 [51] for amino acid sequences, presumably because they represented computationally implementable choices. Only one study employed the CAT-GTR model in a Bayesian analysis of a large amino acid sequence matrix of 160 genes and 177 taxa, but again without adding protein secondary structural constraints to the model [52]. These model choices and implementation modification raise a question on accuracy of the phylogenetic relationships reconstructed in these studies.

As to how the orthology assumption was followed in these land plant phylogenomic studies, it is worth noting that the data analyzed by two original and several follow-up studies were largely from organ-specific transcriptomes [44,46,49–52]. It is questionable whether sequences of many so-called “single-copy” genes are truly orthologous across such a broad time span, because during colonization of the land and subsequent diversification into both terrestrial and aquatic niches plants went through many rounds of whole genome duplication followed by loss of some copies and retention of tissue-specific copies [73,84]. This gene sampling strategy again casts doubt on validity of the conclusions reached from these studies.

From the above discussion, it appears that although phylogenomics has the potential to locate all historical information present in genomes, there is also a huge amount of confounding variation to be sorted out, especially for these deep reconstruction issues. It will be some time before its power can be fully realized to resolve many deep phylogenetic nodes, as the current models and computational resources still cannot match the complexity and ever-increasing sizes of phylogenomic data sets. Therefore, it is time to re-think the practice of uncritically constructing a mammoth-sized super-matrix and relying on supercomputers to deliver “correct” results. The time-tested strategy of constructing a smaller but better-curated data set, by carefully and rigorously selecting orthologous characters/genes as well as taxa that are suitable to the questions attempted, still should be a critical part of a good study design.

Several specific recommendations may be worth consideration. First, for all sequenced genomes, only those that represent diversity as evenly as possible spanning the study group need to be included. Extra taxon sampling should be concentrated on lineages that (based on previous work) are thought to branch relatively closely to the deep nodes of interest in a particular study; adding many lineages nested in recent clades will not particularly help. Second, among all single-copy genes, hopefully with only orthologous sequences, that can be collected from sequenced

genomes, different rate categories can be established even at the alignment stage. Slow-evolving genes can form a core data set, and fast-evolving genes should be either removed or at least analyzed separately so that they do not unnecessarily burden the analysis with LBA. Since chloroplast and mitochondrial genes are generally slow-evolving and have been used extensively in the early stage of plant molecular systematics, it is always advisable to include them in analyses (no matter how many nuclear genes are included) so that comparative analyses can be conducted to monitor performance of the genes from three cellular genome compartments. Third, it may be a good idea to divide the genes according to their functions, preferably after the rate categories are established, as this kind of data partition can help to identify function-specific selective forces that might have distorted phylogenetic signals. For example, in the study of early land plant phylogeny, it is possible that genes related to gametophyte development might have undergone convergent evolution due to the similar life cycle of the three bryophyte lineages and thus produced the bryophyte monophyly topology. Rate comparison of these genes and those that support the bryophyte-paraphyly topology might identify the cause of phylogenetic incongruence. Fourth, different models should be applied to genes of different rate categories to identify the best data-model fit sets for analyses, even though some may perform better than others in some aspect such as running time.

Recently, an investigation of the thorny rooting issue of the eukaryote phylogeny showed that when eight different models were applied in analyses of a matrix of 183 eukaryotic proteins of archaeal ancestry from 185 taxa, all analyses were all able to obtain the presumably correct result, placing four excavate taxa, Parabasalia, Fornicata, Preaxostyla, and Discoba as serial sister lineages to other eukaryotes. Nevertheless, one analysis with a protein structure partition model, which recognized six categories of buried and exposed helices, sheets and loops of secondary structures, took only a fraction of run time used by analyses with seven other common models [83]. Obtaining the same result using different models also satisfies one of the gold criteria of scientific studies, reproducibility [85], hence increasing the likelihood that the result represents reconstruction of a historical divergence event. On the other hand, if some models are found to fit certain matrices better than others, and when different results are produced, the rate and functional characteristics of the matrices should offer explanations on which topology represents the historical diversification pattern and which topologies are analytical artifacts. Explanation of homoplasy is one of the fundamental goals of systematics and deepens understanding of evolution. In this regard, it is worth noting that several land plant phylogenomic studies obtained multiple conflicting topologies of bryophyte relationships and could not provide clear evidence to support a firm conclusion because of lack of diagnostic information on the matrices as all data were amalgamated into one super-matrix [44,46,51].

3.3. Genome Structural Characters

Genome structural characters refer to non-point mutations in genomes, such as gene order or syntenic block changes, insertions/deletions in genes or on chromosomes, gene/genome duplication events, intron gains/losses, and *cis*- to *trans*-splicing changes of group I or II introns. In the early years of molecular systematics, thanks to their rarity these mutations were used as special characters to resolve phylogenetic issues using the maximum parsimony principle [20–22,86]. As more genomes were sequenced, such characters became more abundant and thus could form a separate matrix for reconstructing phylogeny of a group in parallel to sequence characters [23,25,36,39]. Mitochondrial and chloroplast genomes, being fundamentally prokaryotic in nature with extremely low levels of recombination in most eukaryotes, have been the major sources of such characters for both plants [23,24,36,39] and animals [86,87]. The nuclear genome, being much larger and also having more rearrangements of gene order than its organellar counterparts, has not yet been broadly amenable to such type of analysis. However, recent development in bioinformatics has revealed a large number of microsyntenic characters in angiosperms through genome-wide screening of gene order at a fine scale [88]. Further, inference of the ancestral timing of genome duplication by examining distribution patterns of duplicated genes among sequenced nuclear genomes has made it possible to determine relationships among early land plant lineages without having to use an outgroup [52]. Finally, the Archaea sister group of Eukaryota has also been resolved using shared presence of several eukaryote signature protein genes (actin and tubulin, archaeal cell division

proteins related to the eukaryotic Endosomal Sorting Complexes Required for Transport (ESCRT)-III complex, and several information-processing proteins involved in transcription and translation) [89]. Hence, the nuclear genome is likely to become a rich source of such structural characters for resolving difficult deep issues in the tree of life as more species are fully sequenced.

One major strength of genome structural characters is their conservative nature of evolution, which would tend to result in low levels of homoplasy in the sort of deep analyses being discussed in this review. This merit is partly derived from a pre-screening process during the matrix assembly, where fast-evolving changes in genomes are either excluded or simply unrecognized when their frequency and extent of changes are high and substantial. A second strength of these characters is their complexity – there are diverse mechanisms for genome structural changes from recombination to intron-splicing – yet having a simple mode of transformation, having merely two states – presence or absence. Some complex changes in genome structure can be broken into multiple characters, with coding made simpler. Thus, the clarity in character and character state definition inherently leads to low homoplasy levels in the matrices of this type of characters, as long as their rate of change is slow. They are also less prone to being affected by LBA in comparison to nucleotide and amino acid sequences, which often suffer from this problem due to compositional bias in ancient lineages involved in water-land environment transition [32,37–39,41]. Third, genome structural characters are usually independent of each other, represent a different kind of genetic variation than point mutations, and are sampled across the entire genome, presumably randomly. Thus, they satisfy the independence criterion of ideal phylogenetic characters [90,91] better than sequence characters from a gene or a gene-network. Taken together, these features of genome structural characters make them a class of unique and distinct characters that can contribute to reconstructing organismal phylogenies together with DNA sequences and morphological characters.

The major weakness of genome structural characters, that they are few in numbers, comes from one of their strengths, their rarity. Many characters of this category go undetected until variants from newly sequenced genes or genomes are exposed and their phylogenetic informativeness is manifested. A certain amount of prior knowledge on gene and genome evolution is required before the matrix assembly. To exploit the large numbers of cryptic microsyntenic characters in the nuclear genome, as shown in the recent study of angiosperm phylogeny [88], sophisticated niche bioinformatic tools are a pre-requisite. Hence, genome structural characters, despite having a unique role to play in phylogenetic reconstruction, currently remain as a secondary source of information to the massive number of sequence characters, but hopefully their use will increase in importance in the future.

In the study of phylogeny of early land plants, genome structural characters have played a distinctive and prominent role in resolving some highly contentious issues. As reviewed earlier, such characters from both mitochondrial and chloroplast genomes were discovered and when analyzed supported the paraphyly of bryophytes, with liverworts being sister to all other land plants (Figures 1A & 2Q) [22–26]. Most recently, two analyses of a data set of 24 high-quality embryophyte nuclear genomes were conducted to infer root placement on the land plant phylogeny without including an algal outgroup, one using the method ALE (amalgamated likelihood estimation) [53] and the other with STRIDE (species tree root inference from gene duplication events) [54]. The former identified three possible roots of the land plant phylogeny: hornworts (Figure 1M), mosses (Figure 1F), and all bryophytes (Figure 1N), and the latter assigned 0.2%, 39.9% and 59.8% probability to hornworts (Figure 1M), all bryophytes (Figure 1N), and liverworts (Figure 1A) separately as the root [52]. Despite some ambiguity, it seemed that there was signal in these nuclear gene duplication events that could shed light on the relationships of early land plants. Moreover, the microsyntenic characters amassed in the nuclear genome of angiosperms recently [88] should be able to be extended to all streptophytes, as all key lineages of green algae and land plants have had representative genomes sequenced. Ultimately, genome structural characters, given their unique characteristics, should be able to provide a distinct perspective, complementary to that of sequence characters and morphology, on diversification patterns of early land plants, as the five long branches separating green algae, each of the three bryophyte lineages, and vascular plants will continue to haunt phylogenetic algorithms for analyzing sequence data [92].

4. Criteria for Successful Phylogenetic Reconstruction

From the trees of life reconstructed by naturalists in the 19th century, such as those by Ernest Haeckel [93], relationships among organisms have been among the primary targets of intellectual pursuits in biological sciences [94]. The cladistic conceptual revolution [95] and the development of molecular systematics over the last several decades greatly accelerated the pace of such reconstruction. At the beginning of molecular phylogenetics, what constituted the proper kind and number of markers to retrace an accurate organismal evolutionary history was vigorously debated [90,96,97]. As the quantity and quality of analyzed information increased, it became evident that congruence among trees inferred from different types of data using various methods serves as the best criterion of successful phylogenetic reconstruction [91,98,99].

Paradoxically, recent phylogenomic studies have largely ignored this congruence criterion; they seem to rely on internal statistical support such as bootstrap values and Bayesian posterior probabilities as the sole indicator of phylogenetic analysis quality. It should be stressed that these statistical values merely measure the *precision* of an analysis, which gauges the fit between a hypothesis and the analyzed data [100]. In reality, it is *accuracy* that is desired in a study, which reflects the relationship between a hypothesis and the underlying universe [100]. Several factors may affect phylogenetic inference, including processes causing confounding signal (codon usage bias, horizontal gene transfer, incomplete lineage sorting, and convergence) and data selection issues causing homoplasy that can lead to LBA (insufficient or biased data sampling, and inappropriately high evolutionary rates of the markers included). In phylogenomic studies, these factors or processes can lead to statistically well-supported but inaccurate results due to massive sizes of the data sets [61].

To provide a thorough and objective assessment of the quality of a phylogenetic study, several criteria are proposed here. It is important to examine the characters influencing strong statistical support for a particular relationship, bringing transparency to an otherwise rather opaque process that generated the number. Firstly, they should have low evolutionary rates/homoplasy levels, be free of compositional and codon usage bias, and have no heavy RNA editing in the case of organellar genes. If some of these mechanisms appear to be involved in generating the results, they should be reported so that the “synapomorphies” can be refuted or reaffirmed when the hypothetical relationship is evaluated against other evidence. This process is what Willi Hennig called reciprocal illumination in phylogenetic analysis [95]. Secondly, the characters should be from diverse sources, i.e., the sites are from genes encoding different functions. Character independence is one of the prime standards for good phylogenetic markers [90,91]. If informative sites are distributed in many different types of genes, they satisfy the character independence criterion better than otherwise. In addition, the results of an analysis based on sequences from the nucleus, should be compared with those from two other compartments, i.e., chloroplast and mitochondrion. Furthermore,, the results from a sequence analysis should always be checked against those from genomic structural characters (which may come from one, two or all three cellular genomic compartments) and those inferred from morphology. Thirdly, it is crucial in nuclear phylogenomic studies that all sequences of a gene are orthologous to one another. Given the common occurrence of duplications in the nuclear genome, many “single-copy” genes have actually experienced duplications, with only one copy retained due to functional divergence and differential extinction [73], and sampled sequences may not be strictly orthologous to each other in representing the species where they are from. Lastly, when maximum likelihood and Bayesian analysis methods are used, models used must be carefully chosen in an attempt to minimize suspected problems. However, because of the colossal sizes of most phylogenomic data sets, shortcuts are often implemented through different parameter choices, which may compromise performance of the software, especially on challenging parts of a tree. Thus, it is crucial that verification be conducted after analysis to ensure that no serious methodological errors were introduced.

To summarize, the primary goal of a phylogenetic study is to seek reconstruction of congruent relationships among taxa from as many data types as possible, and equally important is the task of explaining any incongruities through well-understood biological mechanisms. Through this dual-goal pursuit, a comprehensive understanding of the evolution of biological patterns and processes can be reached. While this goal may not be immediately achievable in some situations, it is crucial to recognize that an insufficiently tested phylogenetic hypothesis should never be regarded as the

true historical relationships among investigated taxa. The phylogeny represents the central organizing scheme of all biological information [94], and only through rigorous falsification and corroboration tests can a set of hypothetic relationships attain the status of serving as the most efficient guide to further study the genome, phenome and their interactions with the environment.

References

1. Mishler, B.D. Deep phylogenetic relationships among "plants" and their implications for classification. *Taxon* **2000**, *49*, 661-683.
2. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* **1978**, *27*, 401-410.
3. Mishler, B.D. The logic of the data matrix in phylogenetic analysis. In *Parsimony, Phylogeny, and Genomics*, Albert, V.A., Ed.; Oxford University Press: 2005; pp. 57-70.
4. Parenti, L.R. A phylogenetic analysis of the land plants. *Biol. J. Linn. Soc.* **1980**, *13*, 225-242.
5. Mishler, B.D.; Churchill, S.P. A cladistic approach to the phylogeny of the bryophytes. *Brittonia* **1984**, *36*, 406-424.
6. Mishler, B.D.; Lewis, L.A.; Buchheim, M.A.; Renzaglia, K.S.; Garbary, D.J.; Delwiche, C.F.; Zechman, F.W.; Kantz, T.S.; Chapman, R.L. Phylogenetic relationships of the "green-algae" and "bryophytes". *Ann Missouri Bot Gard* **1994**, *81*, 451-483.
7. Kenrick, P.; Crane, P.R. *The Origin and Early Diversification of Land Plants: A Cladistic Study*; Smithsonian Institution Press: Washington, D. C., 1997; p. 441.
8. Garbary, D.J.; Renzaglia, K.S.; Duckett, J.G. The phylogeny of land plants - a cladistic analysis based on male gametogenesis. *Plant Syst. Evol.* **1993**, *188*, 237-269.
9. Renzaglia, K.S.; Duff, R.J.; Nickrent, D.L.; Garbary, D.J. Vegetative and reproductive innovations of early land plants: implications for a unified phylogeny. *Philos. Trans. R. Soc. London Ser. B* **2000**, *355*, 769-793.
10. Chapman, R.L.; Buchheim, M.A. Green algae and the evolution of land plants: inferences from nuclear-encoded rRNA gene sequences. *BioSyst.* **1992**, *28*, 127-137.
11. Hedderson, T.A.; Chapman, R.L.; Rootes, W.L. Phylogenetic relationships of bryophytes inferred from nuclear-encoded rRNA gene sequences. *Plant Syst. Evol.* **1996**, *200*, 213-224.
12. Hedderson, T.A.; Chapman, R.; Cox, C.J. Bryophytes and the origins and diversification of land plants: new evidence from molecules. In *Bryology for the Twenty-first Century*, Bates, J.W., Ashton, N.W., Duckett, J.G., Eds.; W.S. Maney & Son Ltd.: Leeds, U.K., 1998; pp. 65-77.
13. Soltis, P.S.; Soltis, D.E.; Wolf, P.G.; Nickrent, D.L.; Chaw, S.; Chapman, R.L. The phylogeny of land plants inferred from 18S rDNA sequences: Pushing the limits of rDNA signal? *Mol. Biol. Evol.* **1999**, *16*, 1774-1784.
14. Manhart, J.R. Phylogenetic analysis of green plant *rbcl* sequences. *Mol. Phylogen. Evol.* **1994**, *3*, 114-127.
15. Lewis, L.A.; Mishler, B.D.; Vilgalys, R. Phylogenetic relationships of the liverworts (Hepaticae), a basal embryophyte lineage, inferred from nucleotide sequence data of the chloroplast gene *rbcl*. *Mol. Phylogen. Evol.* **1997**, *7*, 377-393.
16. Kallersjo, M.; Farris, J.S.; Chase, M.W.; Bremer, B.; Fay, M.F.; Humphries, C.J.; Petersen, G.; Seberg, O.; Bremer, K. Simultaneous parsimony jackknife analysis of 2538 *rbcl* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Syst. Evol.* **1998**, *213*, 259-287.
17. Samigullin, T.K.; Yacentyuk, S.P.; Degtyaryeva, G.V.; Valiehoroman, K.M.; Bobrova, V.K.; Capesius, I.; Martin, W.M.; Troitsky, A.V.; Filin, V.R.; Antonov, A.S. Paraphyly of bryophytes and close relationship of hornworts and vascular plants inferred from analysis of chloroplast rDNA ITS (cpITS) sequences. *Arctoa* **2002**, *11*, 31-43.
18. Malek, O.; Lattig, K.; Hiesel, R.; Brennicke, A.; Knoop, V. RNA editing in bryophytes and a molecular phylogeny of land plants. *EMBO J.* **1996**, *15*, 1403-1411.
19. Duff, R.J.; Nickrent, D.L. Phylogenetic relationships of land plants using mitochondrial small-subunit rDNA sequences. *Am. J. Bot.* **1999**, *86*, 372-386.
20. Manhart, J.R.; Palmer, J.D. The gain of two chloroplast transfer-RNA introns marks the green algal ancestors of land plants. *Nature* **1990**, *345*, 268-270.
21. Raubeson, L.A.; Jansen, R.K. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science* **1992**, *255*, 1697-1699.
22. Qiu, Y.-L.; Cho, Y.R.; Cox, J.C.; Palmer, J.D. The gain of three mitochondrial introns identifies liverworts as the earliest land plants. *Nature* **1998**, *394*, 671-674.
23. Qiu, Y.-L.; Li, L.B.; Wang, B.; Chen, Z.D.; Knoop, V.; Groth-Malonek, M.; Dombrowska, O.; Lee, J.; Kent, L.; Rest, J.; et al. The deepest divergences in land plants inferred from phylogenomic evidence. *Proc. Natl. Acad. Sci., USA* **2006**, *103*, 15511-15516.

24. Kelch, D.G.; Driskell, A.; Mishler, B.D. Inferring phylogeny using genomic characters: a case study using land plant plastomes. In *Molecular Systematics of Bryophytes*, Goffinet, B., Hollowell, V., Magill, R., Eds.; Missouri Botanical Garden Press: St. Louis, 2004; pp. 3-11.
25. Wolf, P.G.; Karol, K.G.; Mandoli, D.F.; Kuehl, J.; Arumuganathan, K.; Ellis, M.W.; Mishler, B.D.; Kelch, D.G.; Olmstead, R.G.; Boore, J.L. The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). *Gene* **2005**, *350*, 117-128.
26. Mishler, B.D.; Kelch, D.G. Phylogenomics and early land plant evolution. In *Bryophyte Biology*, 2nd ed.; Goffinet, B., Shaw, A.J., Eds.; Cambridge University Press: Cambridge, 2009; pp. 173-197.
27. Soltis, D.E.; Soltis, P.S.; Mort, M.E.; Chase, M.W.; Savolainen, V.; Hoot, S.B.; Morton, C.M. Inferring complex phylogenies using parsimony: An empirical approach using three large DNA data sets for angiosperms. *Syst. Biol.* **1998**, *47*, 32-42.
28. Hillis, D.M. Inferring complex phylogenies. *Nature* **1996**, *383*, 130-131.
29. Nickrent, D.L.; Parkinson, C.L.; Palmer, J.D.; Duff, R.J. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **2000**, *17*, 1885-1895.
30. Chang, Y.; Graham, S.W. Inferring the higher-order phylogeny of mosses (Bryophyta) and relatives using a large, multigene plastid data set. *Am. J. Bot.* **2011**, *98*, 839-849.
31. Qiu, Y.-L.; Taylor, A.B.; Fine, I.M. A molecular temporal evolutionary framework of land plants and the age of angiosperms. *Ann Missouri Bot Gard* **2024**, in press.
32. Nishiyama, T.; Wolf, P.G.; Kugita, M.; Sinclair, R.B.; Sugita, M.; Sugiura, C.; Wakasugi, T.; Yamada, K.; Yoshinaga, K.; Yamaguchi, K.; et al. Chloroplast phylogeny indicates that bryophytes are monophyletic. *Mol. Biol. Evol.* **2004**, *21*, 1813-1819.
33. Karol, K.G.; Arumuganathan, K.; Boore, J.L.; Duffy, A.M.; Everett, K.D.E.; Hall, J.D.; Hansen, S.K.; Kuehl, J.V.; Mandoli, D.F.; Mishler, B.D.; et al. Complete plastome sequences of *Equisetum arvense* and *Isoetes flaccida*: implications for phylogeny and plastid genome evolution of early land plant lineages. *BMC Evol. Biol.* **2010**, *10*, 321.
34. Ruhfel, B.R.; Gitzendanner, M.A.; Soltis, P.S.; Soltis, D.E.; Burleigh, J.G. From algae to angiosperms—inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* **2014**, *14*, 23.
35. Zhong, B.; Xi, Z.; Goremykin, V.V.; Fong, R.; Mclenachan, P.A.; Novis, P.M.; Davis, C.C.; Penny, D. Streptophyte algae and the origin of land plants revisited using heterogeneous models with three new algal chloroplast genomes. *Mol. Biol. Evol.* **2014**, *31*, 177-183.
36. Lemieux, C.; Otis, C.; Turmel, M. Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Frontiers in Plant Science* **2016**, *7*, 679.
37. Sousa, F.; Civan, P.; Foster, P.G.; Cox, C.J. The chloroplast land plant phylogeny: analyses employing better-fitting tree- and site-heterogeneous composition models. *Frontiers in Plant Science* **2020**, *11*, 1062.
38. Goremykin, V.V.; Hellwig, F.H. Evidence for the most basal split in land plants dividing bryophyte and tracheophyte lineages. *Plant Syst. Evol.* **2005**, *254*, 93-103.
39. Turmel, M.; Otis, C.; Lemieux, C. Tracing the evolution of streptophyte algae and their mitochondrial genome. *Genome Biol. Evol.* **2013**, *5*, 1817-1835.
40. Liu, Y.; Cox, C.J.; Wang, W.; Goffinet, B. Mitochondrial phylogenomics of early land plants: mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Syst. Biol.* **2014**, *63*, 862-878.
41. Sousa, F.; Civan, P.; Brazao, J.; Foster, P.G.; Cox, C.J. The mitochondrial phylogeny of land plants shows support for Setaphyta under composition-heterogeneous substitution models. *PeerJ* **2020**, *8*, e8995.
42. Lake, J.A. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc. Natl. Acad. Sci., USA* **1994**, *91*, 1455-1459.
43. Lockhart, P.J.; Steel, M.A.; Hendy, M.D.; Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **1994**, *11*, 605-612.
44. Wickett, N.J.; Mirarab, S.; Nguyen, N.; Warnow, T.; Carpenter, E.; Matasci, N.; Ayyampalayam, S.; Barker, M.S.; Burleigh, J.G.; Gitzendanner, M.A.; et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci., USA* **2014**, Published Online, E4859-E4868.
45. Mirarab, S.; Reaz, R.; Bayzid, M.S.; Zimmermann, T.; Swenson, M.S.; Warnow, T. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **2014**, *30*, i541-i548.
46. Puttick, M.N.; Morris, J.L.; Williams, T.A.; Cox, C.J.; Edwards, D.; P., K.; Pressel, S.; Wellman, C.H.; Schneider, H.; Pisani, D.; et al. The Interrelationships of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* **2018**, *28*, 733-745.
47. Steel, M.; Rodrigo, A. Maximum likelihood supertrees. *Syst. Biol.* **2008**, *57*, 243-250.
48. Sousa, F.d.; Foster, P.G.; Donoghue, P.C.J.; Schneider, H.; Cox, C.J. Nuclear protein phylogenies support the monophyly of the three bryophyte groups (Bryophyta Schimp.). *New Phytol.* **2019**, *222*, 565-575.

49. OneThousandPlantTranscriptomesInitiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **2019**, *574*, 679-685.
50. Harris, B.J.; Harrison, C.J.; Hetherington, A.M.; Williams, T.A. Phylogenomic evidence for the monophyly of bryophytes and the reductive evolution of stomata. *Curr. Biol.* **2020**, *30*, 2001-2012.
51. Su, D.; Yang, L.; Shi, X.; Ma, X.; Zhou, X.; Hedges, S.B.; Zhong, B. Large-scale phylogenomic analyses reveal the monophyly of bryophytes and Neoproterozoic origin of land plants. *Mol. Biol. Evol.* **2021**, *38*, 3332-3344.
52. Harris, B.J.; Clark, J.W.; Schrempf, D.; Szollosi, G.J.; Donoghue, P.C.J.; Hetherington, A.M.; Williams, T.A. Divergent evolutionary trajectories of bryophytes and tracheophytes from a complex common ancestor of land plants. *Nature Ecology & Evolution* **2022**, *6*, 1634-1643.
53. Szöllosi, G.J.; Tannier, E.; Lartillot, N.; Daubin, V. Lateral gene transfer from the dead. *Syst. Biol.* **2013**, *62*, 386-397.
54. Emms, D.M.; Kelly, S. STRIDE: species tree root inference from gene duplication events. *Mol. Biol. Evol.* **2017**, *34*, 3267-3278.
55. Yang, T.; Sahu, S.K.; Yang, L.; Liu, Y.; Mu, W.; Liu, X.; Strube, M.L.; Liu, H.; Zhong, B. Comparative analyses of 3,654 plastid genomes unravel insights into evolutionary dynamics and phylogenetic discordance of green plants. *Frontiers in Plant Science* **2022**, *13*, 808156.
56. Sack, F.D. The development and structure of stomata. In *Stomatal Function*, Zieger, E., Farquhar, G.D., Cowan, I.R., Eds.; Stanford University Press: Stanford, California, 1987; pp. 59-89.
57. Mishler, B.D. Cladistic analysis of molecular and morphological data. *Am. J. Phys. Anthropol.* **1994**, *94*, 143-156.
58. Yang, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **1993**, *10*, 1396-1401.
59. Pagel, M.; Meade, A. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* **2004**, *53*, 571-581.
60. Blanquart, S.; Lartillot, N. A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.* **2008**, *25*, 842-858.
61. Philippe, H.; de Vienne, D.M.; Ranwez, V.; Roure, B.; Baurain, D.; Delsuc, F. Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy* **2017**, *283*, 1-25.
62. Thorne, J.L.; Goldman, N.; Jones, D.T. Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **1996**, *13*, 666-673.
63. Lartillot, N.; Philippe, H. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **2004**, *21*, 1095-1109.
64. Goldman, N.; Thorne, J.L.; Jones, D.T. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **1998**, *149*, 445-458.
65. Redmond, A.K.; McLysaght, A. Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nature Communications* **2021**, *12*, 1783.
66. Dunn, C.W.; Hejnol, A.; Matus, D.Q.; Pang, K.; Browne, W.E.; Smith, S.A.; Seaver, E.; Rouse, G.W.; Obst, M.; Edgecombe, G.D.; et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **2008**, *452*, 745-749.
67. Laumer, C.E.; Fernández, R.; Lemer, S.; Combosch, D.; Kocot, K.M.; Riesgo, A.; Andrade, S.C.S.; Sterrer, W.; Sørensen, M.V.; Giribet, G. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B* **2019**, *286*, 20190831.
68. Wang, H.-C.; Minh, B.Q.; Susko, E.; Roger, A.J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **2018**, *67*, 216-235.
69. Leebens-Mack, J.; Raubeson, L.A.; Cui, L.Y.; Kuehl, J.V.; Fourcade, M.H.; Chumley, T.W.; Boore, J.L.; Jansen, R.K.; dePamphilis, C.W. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the felsenstein zone. *Mol. Biol. Evol.* **2005**, *22*, 1948-1963.
70. Qiu, Y.-L.; Li, L.; Wang, B.; Xue, J.-Y.; Hendry, T.A.; Li, R.; Brown, J.W.; Liu, Y.; Hudson, G.T.; Chen, Z. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *J. Syst. Evol.* **2010**, *48*, 391-425.
71. Xi, Z.; Liu, L.; Rest, J.S.; Davis, C.C. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Syst. Biol.* **2014**, *63*, 919-932.
72. Ohno, S. *Evolution by gene duplication.*; Springer-Verlag: New York, 1970.
73. Wendel, J.F. Genome evolution in polyploids. *Plant Mol. Biol.* **2000**, *42*, 225-249.
74. Kingman, J.F.C. On the genealogy of large populations. *Journal of Applied Probability* **1982**, *19*, 27-43.
75. Hudson, R.R. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology*, Futuyma, D.J., Antonovics, J.D., Eds.; Oxford University Press: New York, 1990; pp. 1-44.
76. Maddison, W.P. Gene trees in species trees. *Syst. Biol.* **1997**, *46*, 523-536.

77. Rannala, B.; Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **2003**, *164*, 1645–1656.
78. Adams, R.H.; Castoe, T.A. Statistical binning leads to profound model violation due to gene tree error incurred by trying to avoid gene tree error. *Mol. Phylog. Evol.* **2019**, *134*, 164–171.
79. Naser-Khdour, S.; Minh, B.Q.; Zhang, W.; Stone, E.A.; Lanfear, R. The prevalence and impact of model violations in phylogenetic analysis. *Genome Biol. Evol.* **2019**, *11*, 3341–3352.
80. Zitnik, M.; Susic, R.; Feldman, M.W.; Leskovec, J. Evolution of resilience in protein interactomes across the tree of life. *Proc. Natl. Acad. Sci., USA* **2019**, *116*, 4426–4433.
81. ArabidopsisInteractomeMappingConsortium. Evidence for network evolution in an Arabidopsis interactome map. *Science* **2011**, *333*, 601–607.
82. Fitch, W.M. Distinguishing homologous from analogous proteins. *Syst Zool* **1970**, *19*, 99–113.
83. Al Jewari, C.; Baldauf, S.L. An excavate root for the eukaryote tree of life. *Science Advances* **2023**, *9*, eade4973.
84. Qiao, X.; Zhang, S.; Paterson, A.H. Pervasive genome duplications across the plant tree of life and their links to major evolutionary innovations and transitions. *Computational and Structural Biotechnology Journal* **2023**, *20*, 3248–3256.
85. Popper, K. *The Logic of Scientific Discovery*; Routledge Classics: London & New York, 2002.
86. Boore, J.L.; Lavrov, D.V.; Brown, W.M. Gene translocation links insects and crustaceans. *Nature* **1998**, *392*, 667–668.
87. Macey, J.R.; Schulte, J.A.I.; Larson, A. Evolution and phylogenetic information content of mitochondrial genomic structural features illustrated with acrodont lizards. *Syst. Biol.* **2000**, *49*, 257–277.
88. Zhao, T.; Zwaenepoel, A.; Xue, J.-Y.; Kao, S.-M.; Li, Z.; Schranz, M.E.; Van de Peer, Y. Whole-genome microsynteny-based phylogeny of angiosperms. *Nature Communications* **2021**, *12*, 3498.
89. Spang, A.; Saw, J.H.; Jørgensen, S.L.; Zaremba-Niedzwiedzka, K.; Martijn, J.; Lind, A.E.; van Eijk, R.; Schleper, C.; Guy, L.; Ettema, T.J.G. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **2015**, *521*, 173–179.
90. Doyle, J.J. Gene trees and species trees - molecular systematics as one-character taxonomy. *Syst. Bot.* **1992**, *17*, 144–163.
91. Miyamoto, M.M.; Fitch, W.M. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* **1995**, *44*, 64–76.
92. Bell, D.; Lin, Q.; Gerelle, W.K.; Joya, S.; Chang, Y.; Taylor, Z.N.; Rothfels, C.J.; Larsson, A.; Villarreal, J.C.; Li, F.-W.; et al. Organellomic data sets confirm a cryptic consensus on (unrooted) land-plant relationships and provide new insights into bryophyte molecular evolution. *Am. J. Bot.* **2019**, *107*, 1–25.
93. Haeckel, E. *Generelle Morphologie der Organismen. Zweiter Band: Allgemeine Entwicklungsgeschichte*; Georg Reimer: Berlin, 1866.
94. Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*; The Belknap Press of Harvard University Press: Cambridge, Massachusetts, 1982; p. 974.
95. Hennig, W. *Phylogenetic Systematics*; University of Illinois Press: Urbana, 1966.
96. Pamilo, P.; Nei, M. Relationships between gene trees and species trees. *Mol. Biol. Evol.* **1988**, *5*, 568–583.
97. Atchley, W.R.; Fitch, W.M. Gene trees and the origins of inbred mice. *Science* **1991**, *254*, 554–558.
98. Leigh, J.W.; Lapointe, F.-J.; Lopez, P.; Baptiste, E. Evaluating phylogenetic congruence in the post-genomic era. *Genome Biol. Evol.* **2011**, *3*, 571–587.
99. Keating, J.N.; Garwood, R.J.; Sansom, R.S. Phylogenetic congruence, conflict and concision between molecular and morphological data. *BMC Ecology and Evolution* **2023**, *23*, 30.
100. Swofford, D.L.; Olsen, O.J.; Waddell, P.J.; Hillis, D.M. Phylogenetic inference. In *Molecular Systematics*, 2nd ed.; Hillis, D.M., Moritz, C., Mable, B.K., Eds.; Sinauer: Sunderland, Massachusetts, 1996; pp. 407–514.
101. Gitzendanner, M.A.; Soltis, P.S.; Wong, G.K.-S.; Ruhfel, B.R.; Soltis, D.E. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *Am. J. Bot.* **2018**, *105*, 291–301.