Opinion TRENDS in Plant Science Vol.9 No.10 October 2004

Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics

Douglas E. Soltis[1], Victor A. Albert[2], Vincent Savolainen[3], Khidir Hilu[4], Yin-Long Qiu[5], Mark W. Chase[3], James S. Farris[6], Saša Stefanović[7], Danny W. Rice[7], Jeffrey D. Palmer[7] and Pamela S. Soltis[8]

[1]Department of Botany and the Genetics Institute, University of Florida, Gainesville, FL 32611, USA
[2]The Natural History Museums and Botanical Garden, University of Oslo, NO-0318 Oslo, Norway
[3]Molecular Systematics Section, Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, UK TW9 3DS
[4]Department of Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[5]Department of Ecology and Evolutionary Biology, University Of Michigan, Ann Arbor, MI 48109, USA
[6]Molekylärsystematiska Laborataoriet, Naturhistoriska Riksmuseet, Box 50007, S 104 05 Stockholm, Sweden
[7]Department of Biology, Indiana University, Bloomington, IN 47405, USA
[8]Florida Museum of Natural History and the Genetics Institute, University of Florida, Gainesville, FL 32611, USA

As systematists grapple with assembling the Tree of Life, recent studies have encouraged a genomic-scale approach, obtaining DNA sequence data for entire nuclear, plastid or mitochondrial genomes for a few exemplar taxa. Some have proclaimed that this comparative genomic strategy heralds the end of incongruence in phylogeny reconstruction. Although we applaud the use of many genes to resolve phylogenetic patterns, there is a significant caveat. In spite of, or even because of, the abundant data per taxon, whole-genome sequencing for a few exemplars can provide completely resolved and strongly supported, but incorrect, evolutionary reconstructions. We provide a conspicuous example that includes *Amborella*, the putative sister of all other extant angiosperms, highlighting the limits of phylogenetics when whole genomes are used but taxon sampling is poor.

Molecular data (primarily DNA sequence data) have prompted a revolution in the reconstruction of the phylogenetic history of organisms. During just the past decade, the amount of sequence data available for building phylogenetic trees has increased ~20-fold [1], and will continue to increase rapidly given the immense technical progress in DNA sequencing. As a result, systematists are now faced with the actual possibility, as well as the daunting challenge, of reconstructing the Tree of Life [2] (a special volume of *American Journal of Botany* comprising articles about the Tree of Life will be published in 2004). The challenges facing systematists in assembling the Tree of Life are many and include: (i) the selection of methods for phylogeny reconstruction (tree building) [e.g. PARSIMONY (see Glossary), maximum likelihood, Bayesian]; (ii) the choice between assembling supertrees based on separate phylogenetic analyses and supermatrices of concatenated sequences; and (iii) the trade-off between genomics-based approaches, in which entire genomes (nuclear, plastid or mitochondrial) are sequenced for a small suite of exemplar species, and dense taxon sampling for fewer genes. The first two issues dealing with the challenges of reconstructing large phylogenetic trees have been recently reviewed [1]. Here, we address the third challenge, the potential pitfalls of using genomics-based approaches to reconstruct the Tree of Life. Although genome sequences are viewed as a phylogenetic panacea by some, the consequences of sampling biological diversity too sparsely can be dire.

Corresponding author: Douglas E. Soltis (dsoltis@botany.ufl.edu).

Glossary

Branch lengths: the number of character-state changes that have occurred between two consecutive nodes in a phylogenetic tree.
Homoplasy: similarity caused by parallelism or reversal of character states.
Incongruence: different genes or different data sets do not yield identical phylogenies.
Long-branch attraction: refers to a lineage that has experienced so much evolution that its character states can become virtually randomized with respect to neighboring nodes (i.e. multiple substitutions appear as an inevitable source of false similarity, or homoplasy). As a result, unrelated taxa with long branches can attract each other, appearing erroneously as closest relatives in a phylogenetic analysis.
Parsimony: a method of phylogeny reconstruction (tree building) that relies on Ockham's razor (the simplest explanation is preferred). This approach selects the tree or trees that minimize the amount of change (the number of steps).
Synapomorphies: shared derived character states.
Synonymous sites: genomic locations at which a nucleotide change in a protein-coding gene does not result in a change in the amino acid encoded.
Templeton test: a parsimony-based test of competing tree topologies [6].

www.sciencedirect.com 1360-1385/$ - see front matter doi:10.1016/j.tplants.2004.08.008

Genomics and the end of incongruence

The application of a recent genomics-based approach to the study of angiosperm phylogeny illustrates the problems that can result when too few taxa are sampled for many genes. Vadim Goremykin *et al.* [3] sequenced the entire plastid genome of *Amborella trichopoda*, the only member of the family Amborellaceae, a taxon of crucial importance because it had been identified in a series of molecular phylogenetic investigations as the probable sister to all other extant angiosperms (see below). Goremykin *et al.* [3] phylogenetically analyzed a dataset that included 61 plastid genes (45 kbp of the plastid genome) from *Amborella* and 12 other plant species. In contrast to many previous studies, Goremykin *et al.* [3] found that the three monocot plastid genomes available (all from grasses, which represent only a small subset of monocot diversity) were sisters, in virtually all analyses, to all other angiosperms included, with strong internal support (100% bootstrap value). *Amborella* instead appeared with strong support as sister to *Calycanthus* (allspice, Calycanthaceae), leading Goremykin *et al.* to conclude that the studies that placed *Amborella* as sister to all other living flowering plants were incorrect. Goremykin *et al.* [4] recently added the plastid sequence of Nymphaeaceae to their dataset of 13 taxa. They again obtained 'monocot basal' topologies that they interpreted as reinforcing their criticism of published angiosperm topologies in which *Amborella* and Nymphaeaceae are sisters to all other extant flowering plants.

In another recent article, Antonis Rokas *et al.* [5] used 106 genes in a phylogenetic analysis of eight yeast species. They noted that individual gene trees were often incongruent, showing different relationships among species, whereas the combined gene dataset for these eight species yielded a single, well-supported tree. Furthermore, all alternative topologies resulting from single-gene analyses were rejected by the TEMPLETON TEST. Rokas *et al.* [5] concluded 'that analyses based on a single or a small number of genes provide insufficient evidence for establishing or refuting phylogenetic hypotheses.' However, the appropriateness and assumptions of the Templeton test [6] remain the subject of debate [7,8]. Furthermore, genome-based approaches to phylogeny reconstruction have produced highly supported but incorrect topologies (e.g. Refs [9,10]). For example, analyses of mitochondrial DNA genome sequences in chordates (using a few taxa) resulted in a well-supported but incorrect topology that conflicts with widely held views of relationships (Figure 1). Rokas *et al.* [5] also compared their topology with that obtained from a study that looked at 75 yeast species but only eight commonly sequenced genes [11], and noted topological differences, with higher support for relationships observed in the Rokas *et al.* [5] tree. Rokas *et al.* [5] stated that '…the unreliability of single-gene datasets (or large datasets composed of linked genes such as genes for the mitochondrial genome) stems from each gene being shaped by a unique set of functional constraints through evolution' (see also Refs [9,10,12]). Rokas *et al.* [5] stated further that 'it is only through the analysis of a larger amount of sequence data that confidence in the proposed phylogenetic construction can be obtained.' In their

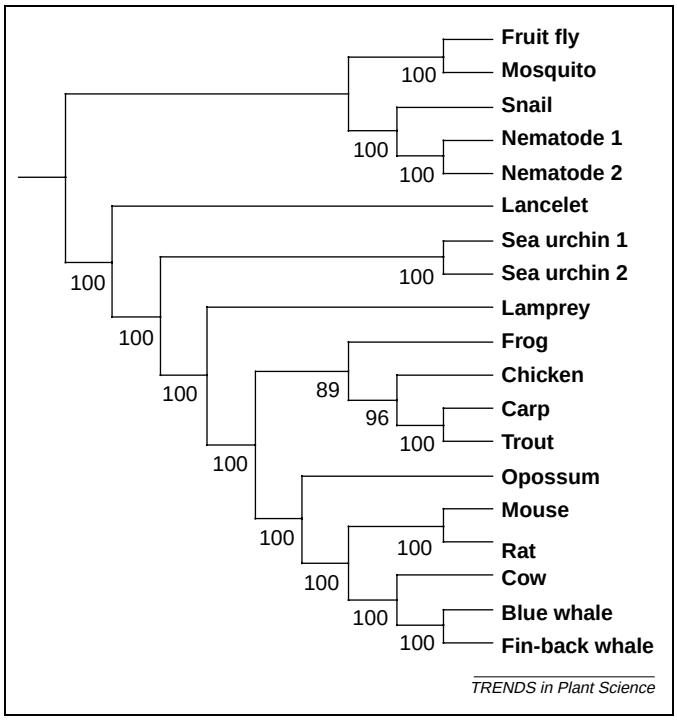

Figure 1. Complete mitochondrial genome-based approach to phylogeny reconstruction led to a single highly supported, but incorrect, tree for animals (bootstrap values below branches). Notice that bony fishes (carp and trout) are embedded within amphibians (frog) and birds (chicken) and that sea urchins are placed within the chordates *sensu lato* (from the lineage of lancelet onwards to the whales). Reproduced, with permission, from Ref. [10].

analysis of eight yeast species, combined datasets of 20 genes 'are sufficient to provide strong (>95%) [bootstrap] support for the species tree'. However, Rokas *et al.* equated getting a single tree with getting the correct tree and also believed that high bootstrap values are invariably indicators of a correct topology. We address both of these misconceptions below.

In an accompanying editorial to Rokas *et al.* [5], Henry Gee [13] proclaimed that this genomic-scale approach of sequencing numerous genes heralds the end of INCONGRUENCE in phylogeny reconstruction. We similarly applaud the use of many genes to resolve phylogenetic patterns and have, along with others, advocated this approach for several years (e.g. Refs [14–22]). However, a crucial caveat not noted by either Rokas *et al.* [5] or Gee [13] is that the number and choice of taxa are at least as crucial for phylogeny estimation as the number of characters (here, nucleotides). With too few and/or the 'wrong' taxa and many base pairs, a strongly supported but incorrect topology can be recovered. Taxa with particularly large amounts and/or biased patterns of nucleotide change can be grouped together because parallelisms and reversals are misinterpreted as evidence for phylogenetic history [23,24] and, the greater the amount of sequence data, the stronger the evidence for an incorrect topology.

With high support for their tree, researchers can become confident in incorrect topologies. Although the bootstrap method for assessing confidence was originally suggested to provide confidence intervals for tree branches (i.e. how well the data at hand represent an underlying

universe of data) [25], it is now well recognized that this resampling method and others, such as the jack-knife [26], provide, at best, a representation of confidence given only the data at hand [27]; even random data can yield high bootstrap support. Furthermore, bootstrap values decrease as the number of taxa increases [28], making it much more likely that high bootstrap support values will be obtained if few taxa are analyzed. Finally, bootstrap values >50% can be overestimates of accuracy when rates of change among taxa are unequal [29], as in the *Amborella* study [3]; no BRANCH LENGTHS were given by Rokas *et al.* [5]. High bootstrap support therefore does not necessarily signify 'the truth'. Likewise, the recovery of a single perfectly resolved tree does not of itself indicate that this topology is closer to 'the truth' than many equally optimal trees obtained in another analysis. Thus, the two criteria heralded as evidence of 'ending incongruence' – high bootstrap support and a single, completely resolved tree – must be evaluated carefully, along with the adequacy of taxon sampling.

The importance of adding taxa is not a new concept. This point has been made for more than a decade, well before whole genomes were sequenced for comparative studies. In an early example based on a single plastid gene, Mark Chase *et al.* [30] showed that individual taxa could be grossly misplaced in the topology when only a small sampling of taxa was analyzed. A series of empirical and simulation studies has reiterated the importance of sampling many taxa, as well as of characters [15,31–34]. However, what constitutes 'adequate' taxon sampling is not always a straightforward issue. We recommend sampling across much of the major morphological diversity encompassed by a group; for example, for taxon sampling of all angiosperms, representatives of most families would be desirable. However, the flaws in taxon sampling that we are pointing to are so egregious that they are obvious. One cannot hope to represent the diversity of major groups of flowering plants (~250 000 species) with 10 or 11 species, three of which are grasses [3,4].

Utility of third codon positions

One often-overlooked reason for the importance of adequate taxon sampling involves nucleotides at third positions of codons. Third codon positions are often excluded from phylogenetic analyses because of presumed difficulties with parallel and back-mutations at this position (owing to the degeneracy of the genetic code, such changes usually do not affect translation into a protein). However, Mari Källersjö *et al.* [35], among others, have demonstrated that nucleotide sites in third positions can enhance tree-resolving power, even though they exhibit more parallelisms and reversals than the other two positions. These authors also demonstrated empirically that the performance of nucleotides in third positions increased as the number of taxa sampled increased, at least in terms of initial similarities retained as hierarchically informative in most-parsimonious trees [35,36]. Further support for the performance of third positions was also found in vertebrate cytochrome *b* genes [37,38]. One simple reason for this trend is that, as taxon number increases, so too can the range of variation (g) seen within a single site, particularly for sites that change more rapidly than others, which makes it easy to recover their historical signal in the tree-building process. This can be readily appreciated from Figure 2; sites with high g have greater potential to support more than one large group in large trees without added cost against parsimony, and multiple smaller groups with added homoplasy but minimal character-state conflict (V.A. Albert *et al.*, unpublished) [36]. In other words, greater variation optimized onto trees with more taxa can take the form of SYNAPOMORPHIES for more larger and smaller groups, and this is exactly what has been observed in empirical studies. For example, Källersjö *et al.* [35] point out in reference to the 2538-taxon *rbcL* analysis of Källersjö *et al.* [39] that, analyzed by themselves, third positions resolve 1327 supported groups with an average jack-knife value of 85%, whereas the first two positions together resolve only 431 groups, with an average value of only 75%. The groups recovered by third positions are also well supported by the full data and are spread over the tree, including both older and younger taxa. By contrast, the first two positions fail, for example, to recognize either land or flowering plants as monophyletic groups. Similar results were obtained with a *matK* dataset for angiosperms [40], in which overall internal support for a tree based on third codon positions was higher than for trees produced from analyzing the first or second positions.

Amborella: a cautionary tale

Returning to the genomics-based study of Goremykin *et al.* [3], these investigators challenged a series of molecular phylogenetic analyses [14–17,19–21,41–43] regarding basal angiosperm relationships (Figure 3a). However, two separate analyses ([44], S. Stefanović, D.W. Rice and J.D. Palmer, unpublished) demonstrate a crippling taxon-sampling flaw in the Goremykin *et al.* study. Soltis and Soltis [44] reduced a large, combined three-gene (*rbcL*, *atpB*, 18S rDNA) dataset for 560 angiosperms [14] that had yielded the '*Amborella*-basal' topology (i.e. *Amborella* sister to all other extant angiosperms) to the few placeholders used by Goremykin *et al.* [3]. Phylogenetic analysis of this greatly reduced dataset [44] yielded the Goremykin *et al.* [3] topology ('monocots basal'; *Amborella* sister to *Calycanthus*; Figure 3a,c). Remarkably, however, the simple addition to this small three-gene dataset of only one or more additional monocots [such as an orchid (*Oncidium*)] again resulted in the '*Amborella*-basal' topology (Figure 3d) that Soltis *et al.* [14], and many other researchers, had earlier recovered (e.g. [16,17,19,20,41–43]).

S. Stefanović, D.W. Rice and J.D. Palmer (unpublished) addressed the Goremykin *et al.* [3] result using a different approach. These investigators obtained the nearly complete plastid sequence for the early diverging monocot *Acorus* and added it to the Goremykin *et al.* [3] dataset. With the substitution of just this one taxon for grasses, the '*Amborella*-basal' topology was recovered with 93–100% bootstrap support, instead of the 'monocots-basal' topology found with the original dataset of Goremykin *et al.* [3]. S. Stefanović, D.W. Rice and J.D. Palmer (unpublished) also showed that reanalysis of the same dataset analyzed

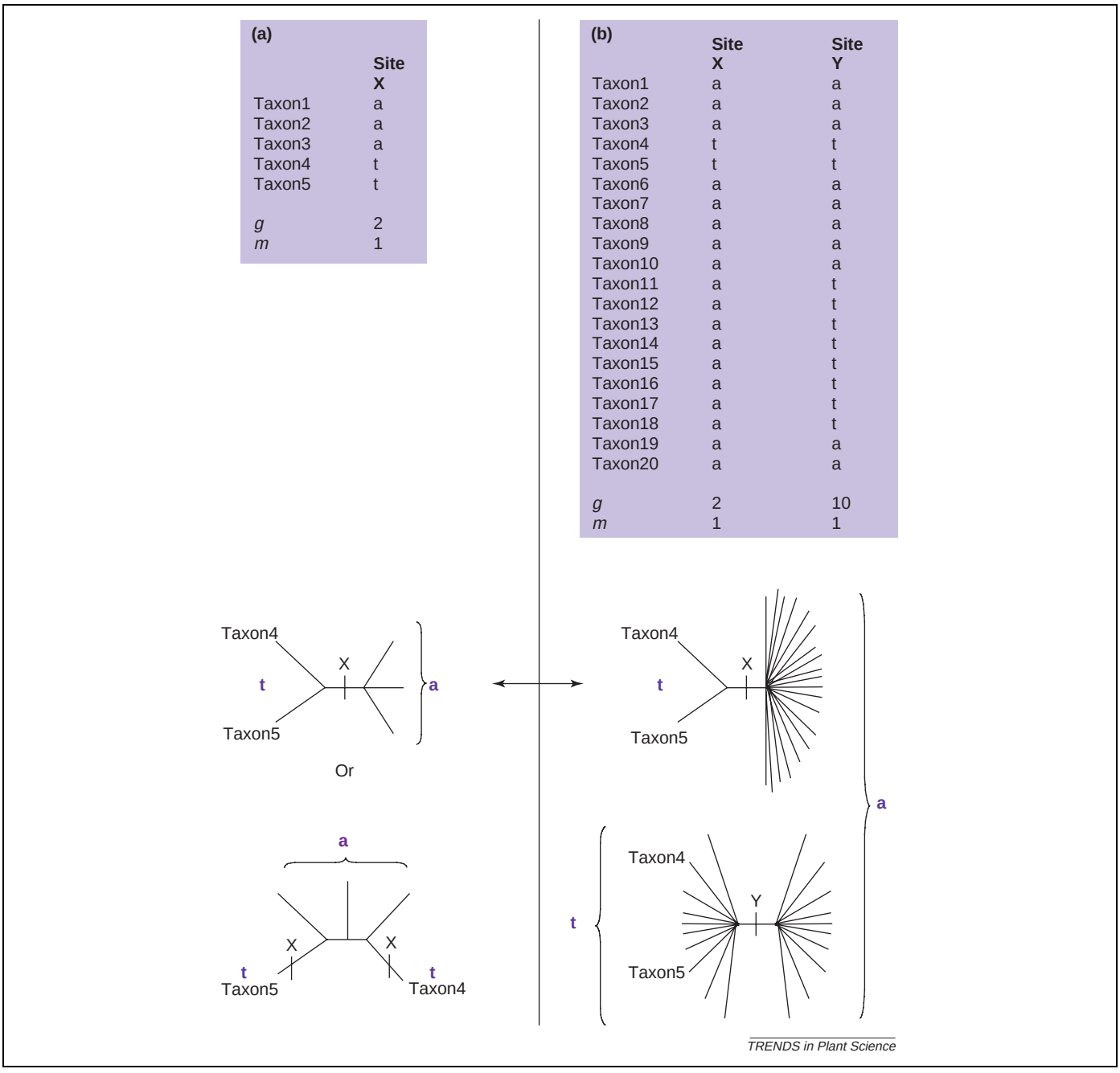

Figure 2. Sites that display a greater range of nucleotide variation have greater potential to support more than one large group in large trees without added cost against parsimony. **(a)** An imaginary site, X, from a sampling of five taxa, two of which are different from the others (taxon4 and taxon5 have t, the rest have a). The minimum number of changes possible in this character (*m*) is 1, and the maximum number (*g*) is 2. Only one group can be supported by site X, one in which taxon4 and taxon5 form a group (in which $m=1$; top tree). All other trees for this site, such as the bottom tree, require an extra change (homoplasy). **(b)** Sampling has now increased to 20 taxa; two imaginary sites are considered, X and Y. In X, only taxon4 and taxon5 were found to be t, with the rest being a. This example can be considered to be analogous to a slowly evolving site. In Y, greater sampling revealed more variation, ten taxa with t and ten with a. Site X still has the same potential tree-resolving power as in example **(a)** above, even though the taxon number is increased. This is because one large group could be identified, that excluding taxon4 and taxon5 (top tree), with *m* and *g* remaining the same. However, site Y with its greater range of variation (analogous to a more rapidly evolving site), now has $g=10$; as such, a tree is possible that has $m=1$ and supports two groups of ten taxa each (bottom tree). Of course, other trees (not shown) could also be possible, such as splitting the group containing taxon4+taxon5, but these would require extra changes (again, homoplasy). As sampling grows larger and larger, and if *g* continues to rise, such homoplastic changes can become more and more capable of independently supporting groups themselves (V.A. Albert *et al.*, unpublished). Trees are shown in abstract form (i.e. with branch tips for only taxon4 and taxon5 labeled). Nucleotide states are shown in purple bold for taxa 4 and 5 as well as collectively for abstracted branch tips. Changes in sites X and Y are shown mapped onto branches with cross-slashes where a nucleotide change occur.

by Goremykin *et al.* [3] using a maximum likelihood model that should be relatively insensitive to LONG-BRANCH ATTRACTION (i.e. with γ-distributed categories), recovered '*Amborella* basal' trees (Figure 3b). Thus, these studies illustrate how inadequate taxon sampling greatly exacerbates the potential to recover phylogenetic artefacts caused by long-branch attraction [23] and how crucial the choice of optimality criteria and model assumption can be with datasets that have many characters but few taxa [45].

Inadequate taxon sampling

Inadequate taxon sampling can have two interacting dimensions – too few taxa and the 'wrong' taxa as

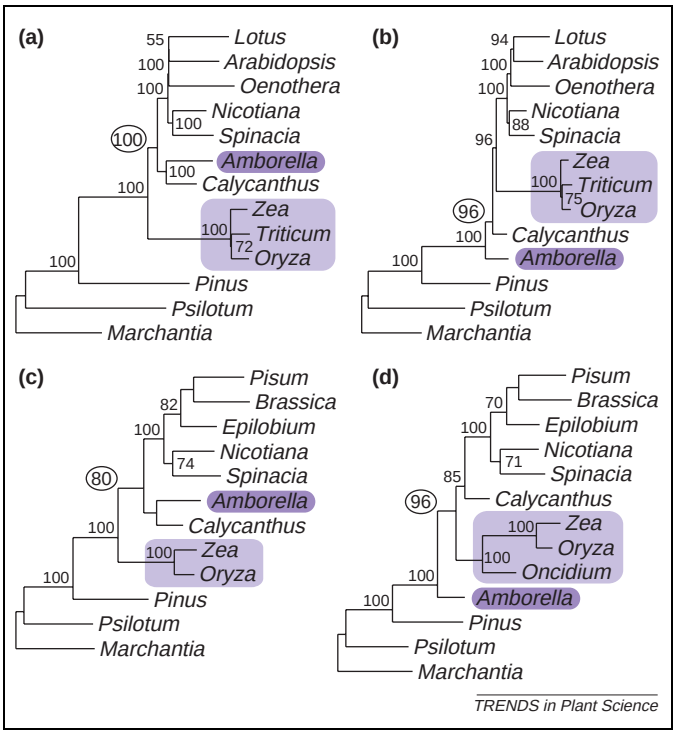

Figure 3. Comparison of tree obtained by Goremykin *et al.* [3] **(a)** with those retrieved in the reanalyses of S. Stefanović, D.W. Rice and J.D. Palmer (unpublished) **(b)** and Soltis and Soltis [44] **(c,d)**. Bootstrap values above 50% are shown in each tree. **(a)** Neighbor-joining tree based on first and second codon positions from 61 plastid protein genes. **(b)** Maximum likelihood tree (using four γ-distributed rate categories) obtained upon analysis of the same dataset analyzed in Goremykin *et al.* [3]. **(c)** Single most parsimonious tree obtained in a phylogenetic analysis of three genes (*rbcL*, *atpB*, 18S rDNA; data from Ref. [14]) with taxa removed from this 567-taxon dataset to approximate the taxon sampling of Goremykin *et al.* [3]. **(d)** Single most parsimonious tree obtained when the orchid *Oncidium* was added as an additional monocot exemplar to the dataset in **(c)** to break up the long branch leading to the grasses.

exemplars – both of which can negatively impact phylogeny reconstruction. Monocots were represented in Goremykin *et al.* [3,4] by only three grasses, members of the monocots well known to have high rates of plastid nucleotide change [18,19,46]. As a result of this high rate of change, these taxa have long branch lengths, which can lead to erroneous phylogenetic trees because of long-branch attraction. The analyses of Soltis and Soltis [44] and S. Stefanović, D.W. Rice and J.D. Palmer (unpublished) clearly showed that, in the absence of other monocots, the Goremykin *et al.* data lead to a spurious rooting of angiosperms, reflecting attraction between the long branch leading to grasses and the long branch separating angiosperms from the outgroups (e.g. Figure 3a). That it takes only a single other monocot, either in addition to or in place of grasses, to recover the *Amborella*-basal topology emphasizes the sensitivity of analyses with few taxa to artefacts stemming from systematic bias in one or more lineages chosen for inclusion. Increased taxon sampling often provides a substantial buffer against such artefacts, allowing divergent genomes that might otherwise bias the phylogenetic tree to be included and more appropriately placed.

Goremykin *et al.* [3] excluded third positions from their analyses because most of the 61 chloroplast genes they analyzed were 'very divergent' at SYNONYMOUS SITES. The K_s values (rate of synonymous base substitutions) for most genes in their comparisons between *Pinus* and angiosperms were between 0.50 and 1.50 substitutions per site, which they apparently feared could lead to 'misleading' phylogenetic trees. However, the results of Soltis and Soltis [44] and S. Stefanović, D.W. Rice and J.D. Palmer (unpublished) indicate that third positions are probably not contributing 'excessive' HOMOPLASY and yielding incorrect trees with this dataset, in spite of the small number of taxa. This is consistent with our discussion above on third positions, as well as with several other studies that have recognized the phylogenetic utility of third positions in organellar genes [12,35,37–40,47]. Rapidly evolving positions or genes do not necessarily impede phylogeny reconstruction; indeed, they can enhance it [40], as long as the use of 'rapidly evolving genes' is balanced by dense, judicious taxon sampling. This is not to say that third positions must always be included in phylogenetic analyses. For example, plastid third positions might be excluded when divergences are significantly greater than in Goremykin *et al.* [3] (e.g. in analyses that span all plastid evolution) [48]. Rapidly evolving third-position nucleotides might even be problematic in inferring certain aspects of seed plant phylogeny, in which sampling is intrinsically scanty (because of extinction) and there are several long, unbreakable branches [12,21,22,49].

The results for *Amborella* summarized above are another clear lesson about the importance of taxon sampling, particularly in this age of genomics, in which many assume that having plenty of base pairs will solve most phylogenetic problems – 'ending incongruence' [13]. Many nucleotides for a small sampling of taxa will indeed end incongruence but, as the examples of *Amborella* and animals illustrate (Figures 1,3), an incorrect, yet strongly supported, topology might be recovered.

Sequencing many genes (e.g. entire organellar genomes) will place constraints (in terms of both time and money) on the number of taxa studied. Because research funds are limited, all investigators must make choices between the number of genes sequenced and number of taxa sampled. However, taxon sampling is crucial and should not be ignored in this genomics era in which sequencing becomes ever more rapid and inexpensive. In this regard, genomic regions that provide sufficient signal without compromising taxon representation are essential for accurate (and cost effective) assessment of evolutionary histories.

Another concern regarding whole organellar genome sequencing involves functional constraints (i.e. functional requirements such as chemical properties, charge and hydrophobicity) in these genomes. In studies using entire mitochondrial genome sequences in the assessment of animal phylogeny, Gavin Naylor and Wesley Brown [9,10] and David Pollock *et al.* [50] found that highly erroneous topologies could be recovered with strong support (Figure 1), which could be attributed in part to the strong functional constraints inherent in the rapidly evolving mitochondrial genomes of animals [9,10,12,50]. Naylor and Brown [10] argued that systematic zoology will benefit from an increased understanding of the functional and structural constraints acting on the mitochondrial genome

of animals 'to improve phylogenetic inference using large datasets'.

Conclusions

We agree that 'ending incongruence' might be possible with genomic-scale data, but only in the context of broad taxon sampling. David Hillis *et al.* [34] summarized our view, which is also the view of many phylogenetic systematists based on both empirical data and simulations. 'If one is interested in inferring the evolutionary history of life, a much broader sample of taxa (perhaps sequenced for far less than full genomes) will result in a much more accurate estimate of phylogeny than will complete genomes of only a small number of taxa.' As scientists assemble the Tree of Life, perhaps we need to rethink the strategies behind some ongoing projects. Some funded initiatives are primarily or exclusively using whole organellar genome sequencing for a small number of taxa. Our example of *Amborella* indicates that such a strategy can be seriously flawed and could easily result in a strongly supported, but incorrect, tree. Evolutionary biologists and other scientists using trees could easily become overconfident in trees that are incorrect. This is not to say that whole genome sequencing is not needed, but it might be more prudent in many initiatives to sequence fewer base pairs but add more species to the analysis. As an alternative strategy, in addition to some whole genome sequencing, a significant 'targeted sequencing' component could be added in which many additional taxa are sequenced for a subset of genes that seem to be ideally suited for reconstructing phylogeny in that particular group at a given level of inference. That is, add a targeted sequencing approach to a genome sequencing strategy. In addition, taxa must be chosen judiciously to span morphological and lineage diversity; it is much better to err on the side of too many taxa sampled than too few. To facilitate broader taxon sampling, systematists might need to make DNAs widely available for as many species as possible, perhaps establishing DNA banks at one or several institutions where DNAs could be stored, maintained and curated, much as plant and animal specimens are maintained in herbaria and museums. The systematics community must be careful that we are not blinded by genomics; complete genome data by themselves are not the panacea for phylogeny reconstruction.

References

1 Sanderson, M.J. and Driskell, A.C. (2003) The challenge of constructing large phylogenetic trees. *Trends Plant Sci.* 8, 374–379
2 Cracraft, J. and Donoghue, M., eds (2004). *Assembling the Tree of Life*, Oxford University Press
3 Goremykin, V.V. *et al*. (2003) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol. Biol. Evol.* 20, 1499–1505
4 Goremykin, V.V. *et al*. (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
5 Rokas, A. *et al*. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804
6 Templeton, A.R. (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* 37, 221–244
7 Shimodaira, H. and Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116
8 Goldmann, N. *et al*. (2000) Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* 49, 652–670
9 Naylor, G.J.P. and Brown, W.M. (1997) Structural biology and phylogenetic estimation. *Nature* 388, 527–528
10 Naylor, G.J.P. and Brown, W.M. (1998) *Amphioxus* mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* 47, 61–76
11 Kurzman, C.P. and Robnett, C.J. (2003) Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEM Yeast Res.* 3, 417–432
12 Savolainen, V. *et al*. (2002) Phylogeny reconstruction and functional constraints in organellar genomes: plastid versus animal mitochondrion. *Syst. Biol.* 51, 638–647
13 Gee, H. (2003) Ending incongruence. *Nature* 425, 782
14 Soltis, P.S. *et al*. (1999) Angiosperm phylogeny inferred from multiple genes as a research tool for comparative biology. *Nature* 402, 402–404
15 Soltis, D.E. *et al*. (1998) Inferring complex phylogenies using parsimony: an empirical approach using three large DNA datasets for angiosperms. *Syst. Biol.* 47, 32–42
16 Parkinson, C.L. *et al*. (1999) Multigene analyses identify the three earliest lineages of extant flowering plants. *Curr. Biol.* 9, 1485–1488
17 Qiu, Y-L. *et al*. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402, 404–407
18 Chase, M.W. *et al*. (2000) Higher-level systematics of the monocotyledons: an assessment of current knowledge and a new classification. In *Monocots: Systematics and Evolution* (Wilson, K.L. and Morrison, D.A., eds), pp. 3–16, CSIRO Publishing, Collingwood, Victoria, Australia
19 Graham, S. and Olmstead, R.G. (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.* 87, 1712–1730
20 Barkman, T.J. *et al*. (2000) Independent and combined analysis of sequences from all three genomic compartments converge to the root of flowering plant phylogeny. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13166–13171
21 Magallón, S. and Sanderson, M.J. (2002) Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. *Am. J. Bot.* 89, 1991–2006
22 Burleigh, J.G. and Mathews, S. Phylogenetic signal from nucleotide data from seed plants: implications for resolving the seed plant tree of life. *Am. J. Bot.* (in press)
23 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410
24 Steel, M.A. *et al*. (1993) Confidence in evolutionary trees from biological sequence data. *Nature* 364, 440–442
25 Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791
26 Farris, J.S. *et al*. (1996) Parsimony jackknifing outperforms neighbor-joining. *Cladistics* 12, 99–124
27 Soltis, P.S. and Soltis, D.E. (2003) Applying the bootstrap in phylogeny reconstruction. *Stat. Sci.* 18, 256–267
28 Sanderson, M.J. and Wojciechowski, M.F. (2000) Improved bootstrap confidence limits in large scale phylogenies, with an example from Neo-Astragalus (Leguminosae). *Syst. Biol.* 49, 671–685
29 Hillis, D.M. and Bull, J.J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192
30 Chase, M.W. *et al*. (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Ann. MO Bot. Gard.* 80, 526–580
31 Graybeal, A. (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* 47, 9–17
32 Hillis, D.M. (1998) Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* 47, 3–8
33 Zwickl, D.J. and Hillis, D.M. (2002) Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51, 588–598
34 Hillis, D.M. *et al*. (2003) Is sparse taxon sampling a problem for phylogenetic inference. *Syst. Biol.* 52, 124–126
35 Källersjö, M. *et al*. (1999) Homoplasy increases phylogenetic structure. *Cladistics* 15, 91–93
36 Savolainen, V. and Chase, M.W. (2003) A decade of progress in plant molecular phylogenetics. *Trends Genet.* 19, 717–724

37 Yoder, A.D. *et al*. (1996) Molecular evolutionary dynamics of cytochrome *b* in strepsirrhine primates: the phylogenetic significance of third position transversions. *Mol. Biol. Evol.* 13, 1339–1350
38 Björklund, M. (1999) Are third positions really that bad? A test using vertebrate cytochrome *b*. *Cladistics* 15, 191–197
39 Källersjö, M. *et al*. (1998) 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.* 213, 259–287
40 Hilu, K.W. *et al*. (2003) Inference of angiosperm phylogeny based on *matK* sequence information. *Am. J. Bot.* 90, 1758–1776
41 Mathews, S. and Donoghue, M. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286, 947–950
42 Savolainen, V. *et al*. (2000) Phylogenetics of flowering plants based upon a combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.* 49, 306–362
43 Zanis, M.J. *et al*. (2002) The root of the angiosperms revisited. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6848–6853
44 Soltis, D.E. and Soltis, P.S. (2004) *Amborella* not a 'basal angiosperm'? Not so fast. *Am. J. Bot.* 91, 997–1001
45 Phillips, M.J. *et al*. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* 21, 1455–1458
46 Gaut, B.S. *et al*. (1992) Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. *J. Mol. Evol.* 35, 292–303
47 Olmstead, R.G. *et al*. (1998) Patterns of sequence evolution and implications for inferring phylogeny using nucleotide sequence data. In *Molecular Systematics of Plants II: DNA Sequencing* (Soltis, D.E. *et al*., eds), pp. 164–187, Kluwer
48 Delwiche, C.F. *et al*. (1995) Phylogenetic analysis of *tufA* sequences indicates a cyanobacterial origin of all plastids. *Mol. Phylogenet. Evol.* 4, 110–128
49 Soltis, D.E. *et al*. (2002) Phylogeny of seed plants based on evidence from eight genes. *Am. J. Bot.* 89, 1670–1681
50 Pollock, D.D. *et al*. (2000) A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* 17, 1776–1788