Review TRENDS in Ecology and Evolution Vol.22 No.5



# Deciphering ancient rapid radiations

James B. Whitfield[1] and Peter J. Lockhart[2]

[1] Department of Entomology, 320 Morrill Hall, 505 S. Goodwin Ave., University of Illinois, Urbana, IL 61801, USA
[2] Allan Wilson Centre for Molecular Ecology and Evolution, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

**A deeper phylogenetic understanding of ancient patterns of diversification would contribute to solving many problems in evolutionary biology, yet many of these phylogenies remain poorly resolved. Ancient rapid radiations pose a major challenge for phylogenetic analysis for two main reasons. First, the pattern to be deciphered, the order of divergence among lineages, tends to be supported by small amounts of data. Second, the time since divergence is large and, thus, the potential for misinterpreting phylogenetic information is great. Here, we review the underlying causes of difficulty in determining the branching patterns of diversification in ancient rapid radiations, and review novel data exploration tools that can facilitate understanding of these radiations.**

## Introduction

Phylogenies are crucial to our understanding and explanation of the origin and evolution of the major adaptations and lineages of organisms on Earth. Yet, despite recent advances in the availability of data and methodologies for investigating the historical relationships among organisms, many major phylogenetic patterns remain poorly resolved. Rapid evolutionary radiations have been proposed to explain poorly resolved phylogenies in many groups of organisms, including aphids, black flies, bees, birds, turtles, mammals and higher plants (Table 1).

Classic tales of radiation include the Cambrian explosion of animal phyla, the Cretaceous origin of angiosperms, the diversification of birds and mammals, and controversies over the origins of highly social behavior in the bees. Identifying which of these cases of diversification represent genuine rapid radiations, and which do not, has broad ramifications in evolutionary biology. However, deciphering patterns of ancient radiation is inherently problematic, often requiring an array of data sources and analytical techniques at the cutting edge of current knowledge. Why is the evolutionary history of these radiations so much more difficult to reconstruct than it is for many other phylogenetic questions?

## Some phylogenies are harder to estimate than others

As the methodology of phylogeny estimation has become more sophisticated, diverse and efficient, our ability to unravel successfully the phylogenetic histories of many groups of organisms has improved. With the advent of molecular systematics, the evolutionary relationships of many organisms (e.g. placental mammals [1,2]), are now thought to be understood with a high degree of confidence. This is despite considerable controversy remaining over the advantages and limitations of various phylogenetic methods [3,4]. The reason for this confidence results from the accumulation of independent data sets whose analysis (irrespective of the method used) tends to converge on the same estimations of phylogenetic history.

However, elucidating the evolutionary history of some relationships remains difficult. These instances often concern situations where different methods provide evidence for different phylogenies. These cases typically involve the 'long-branch attraction' problem (see Glossary) [5–8]. Strategies for recognizing and correcting the problem have received considerable recent attention [9–11], but are

### Glossary

**Consensus network**: a phylogenetic network that displays the splits found in source trees that have the same (completely overlapping) set of taxa; thresholds can be implemented to display all the splits in all trees, or only the splits that occur in x% of source trees; a network version of a consensus tree.
**Hadamard transform**: a discrete Fourier transform that enables easy reversible translation between distances and site patterns (partition frequencies); given a tree with branch lengths, one can calculate the expected frequency of data site patterns that perfectly fit that tree.
**Lento plot**: a bar graph that displays, in ranked order, the data support and conflict for each split (both contained in the optimal phylogenetic tree and in alternative trees with any support).
**Likelihood map**: a triangular diagram, calculated using maximum likelihood from all sets of quartets of taxa, depicting the distribution of support (frequency of data patterns) for all internal branches in a phylogeny.
**Lineage sorting**: discordance between the timing of mutations or origins of new alleles and splitting of taxon lineages, leading to differences in alleles among taxa that do not reflect the history of taxon splitting.
**Long-branch attraction**: the tendency of phylogenetic methods to group long but (in reality) nonadjacent branches; a problem originally associated with the method of parsimony, but subsequently recognized as a problem for all methods where there is substitution model misspecification owing to any of several systematic biases in the data.
**Parametric bootstrapping**: a tree-testing approach in which simulated data sets are generated based on some evolutionary model, and then analyzed to obtain an expected distribution of resulting optimal trees from empirical data evolving in a manner hypothesized to be similar to the model.
**Phylogenetic network**: any nonbifurcating (anastomizing) graphical representation of relationships suggested by the data; the extra internal branches can represent conflicting data patterns in a data set or conflicting splits from different sources.
**Polytomy**: a node in a phylogenetic tree that subtends more than two descendant branches; termed a soft polytomy when it is due to inadequate data, or a hard polytomy when it reflects a true simultaneous divergence of more than two daughter lineages.
**Power analysis**: a statistical analysis to determine the quantity of data that is needed to distinguish the null from alternative hypotheses using a given test.
**Split**: a partition of the taxa in a phylogenetic tree into two groups, supported by a character or data pattern; a split corresponds to a supported branch or internode in a tree.
**Supernetwork**: a phylogenetic network incorporating the splits found in a series of source trees that contain overlapping, but not necessarily identical, sets of taxa; a network version of a supertree.

*Corresponding author:* Whitfield, J.B. (jwhitfie@life.uiuc.edu).
Available online 14 February 2007.

 0169-5347/$ – see front matter  doi:10.1016/j.tree.2007.01.012

**Table 1. A selection of the more convincing cases for ancient rapid radiation**

| Taxon | Approximate age of radiation[a,b] | Data applied to resolution | Refs |
|---|---|---|---|
| Early placental mammals | 100–110 mya | 22 genes (19 nuclear, 3 mtDNA) | [1,2] |
| Avian orders | Late Cretaceous | 12S, 16S, 18S rRNA | [41] |
| | | *ZENK*, *c-myc*, *RAG-1* | [42] |
| | | Complete mtDNA genomes | [43] |
| Cryptodiran turtles | 90–120 mya | Morphology, fossils, *cyt b*, 12S | [44] |
| Surgeonfish (Acanthuridae) | Eocene | mtDNA sequences from 12S, 16S, *t-Pro*, control region | [45] |
| Corbiculate bees | >80 mya | Morphology, DNA sequences from 16S, 28S, *cyt b*, LW *opsin*, *EF-1alpha* | [13,46,47] |
| Microgastrine parasitoid wasp genera | 40–50 mya | Morphology, sequences from 16S, COI, 28S | [28,48] |
| Black flies (Simuliidae) | Cretaceous | DNA sequences from 12S, 28S, *EF-1alpha*, dopa decarboxylase, *PEPCK* | [49] |
| Modern aphid lineages | Late Cretaceous | Fossils, DNA sequences from 12S and 16S | [50] |
| Major metazoan phyla | Late Precambrian | Fossils, nuclear and mtDNA sequences | Many studies; e.g. [37,38,51] |
| Basal angiosperm lineages | Late Jurassic–Early Cretaceous | *rbc*L, *atp*B, 18S rRNA | [52] |
| | | Complete chloroplast genomes | [53] |
| | | Duplicate phytochrome genes | [14,24] |
| Major Saxifragales lineages | Late Cretaceous | Chloroplast *atpB*, *mat*K, *rbc*L, nuclear 18S, 26S | [31] |
| Families of Lamiales | Eocene? | *mat*K, *ndh*F, *rbc*L | [54] |
| Major Brassicaceae lineages | 15–30 mya | *Adh*, *chs*, *mat*K, *trn*L-F, ITS | [55,56] |

[a]Estimated ages of the radiations do not all correspond to times just after known mass extinctions.
[b]Mya, millions of years ago.

complicated by the fact that the long-branch attraction problem can be produced by different underlying causes [12].

Where a group has radiated rapidly in the distant past, the long-branch attraction problem can arise as a result of the relatively large amount of divergence time separating extant taxa compared with the relatively small amount of divergence time that separates ancestors of distinct lineages (Figure 1). In this situation, there might have been little opportunity for evolutionary changes (molecular or morphological) to accumulate in these short internal time spans. Furthermore, such changes that have occurred might not be observable owing to subsequent lineage-specific changes that follow divergence. If so, a large amount of data and good estimates of the evolutionary divergences between all taxa are required to resolve relationships (e.g. Refs [6,13–16]).

## Difficulties underlying ancient rapid radiations

The biology of genomes and gene sequences also conspires to make the phylogenetic reconstruction of radiations difficult. Their evolutionary properties and the problems that they cause can be straightforward to diagnose, but when co-occurring, they can easily confound one another and complicate the resolution of rapid radiations.

As the time between divergences becomes shorter, it eventually enters the time span of lineage-sorting problems (Figure 2), where individual gene trees might not reflect species phylogeny because alternative ancestral alleles are still present in the different lineages [17,18]. For instance, in theory, the internode span of time (in generations of the organism) needs to be at least five times the effective population size for there to be a 95% chance that the gene tree is congruent with the species tree [17]. Thus, whereas

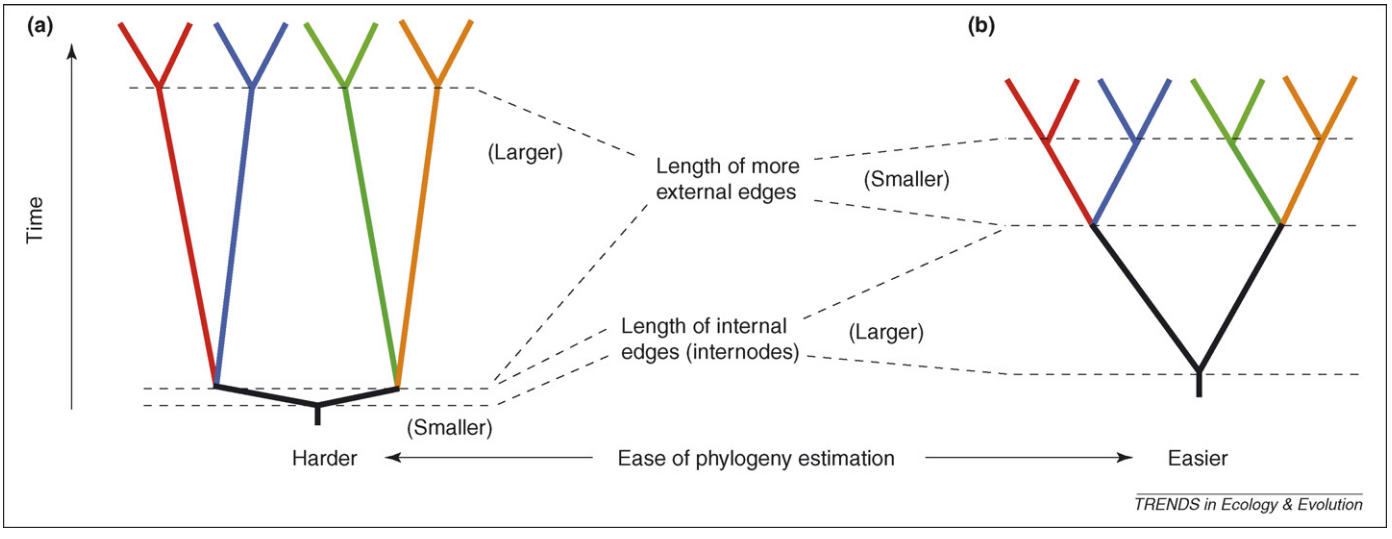


**Figure 1.** The key pattern of an ancient rapid radiation, where the deeper internal branches of the phylogeny are much shorter than the branch lengths between them and the tips, is shown in **(a)**. The pattern in **(b)** represents a relatively steady rate of diversification. The pattern in (a) is more difficult to estimate correctly, owing not only to the short timescales implied by the internal branches at the bottom of the figure, but also to the apparent long times since divergence, which provide opportunity for the loss or obfuscation of the phylogenetic signal.

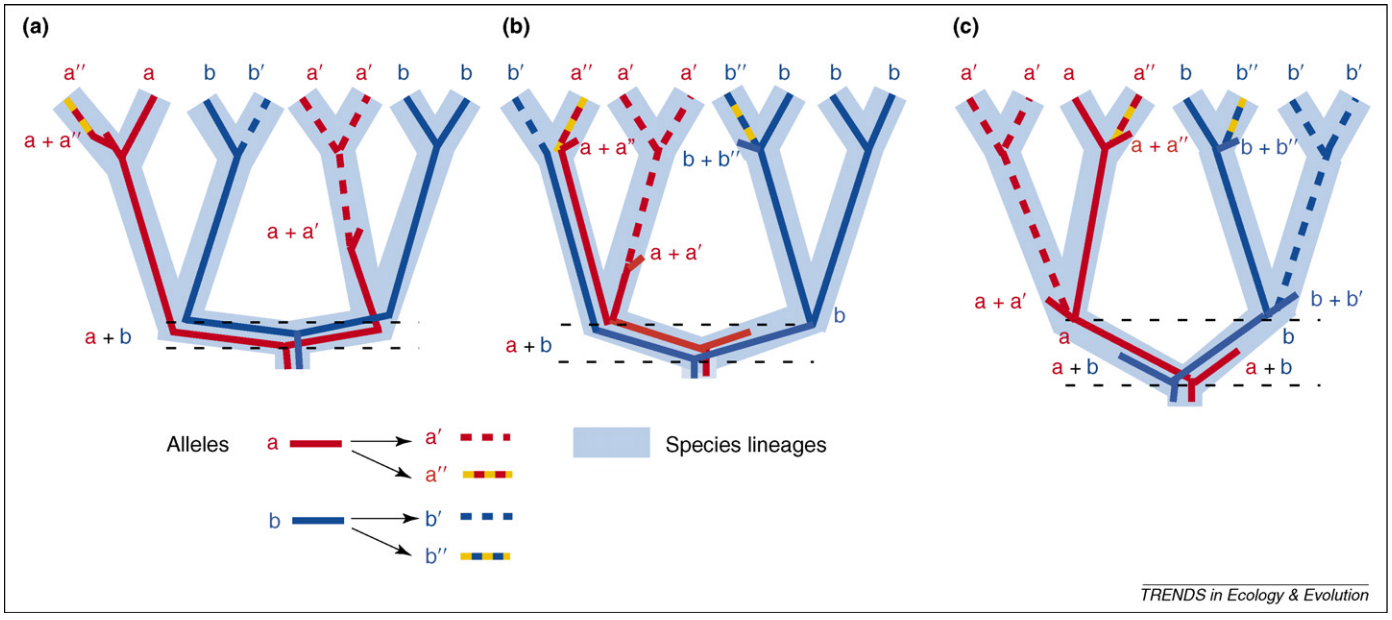


**Figure 2**. The problem of lineage sorting of alleles, illustrated using divergences in two original alleles (an ancestral polymorphism). Lineage-sorting problems are exacerbated when the time between divergences is short, as this time span provides little opportunity for the polymorphisms to disappear. As the time span between divergences increases, the chance also increases that one of the alleles might be lost before the second splitting event. Thus, the proportion of genes showing misleading lineage-sorting patterns will decrease. In this hypothetical example, alleles a and b are assumed to represent an ancestral polymorphism found within the species lineage (shaded pale blue). Allele a is replaced over time in some lineages by a′ or a″, and allele b is replaced over time by b′ or b″. Short stubs of allele lineage branches represent the brief persistence of the parent allele alongside the daughter allele.

phylogenetic analysis of higher taxa generally finds enough signal to overcome lineage sorting, this might not be the case when divergences are closely spaced, especially if population-level variation is not taken into account [19]. In these situations, different genes can suggest different relationships, and phylogenetic incongruence resulting from lineage sorting might not be straightforward to resolve in terms of the underlying species phylogeny. In respect of

## Box 1. Visualizing conflict in data using phylogenetic networks

Whereas some network methods are intended to model reticulate phylogenetic phenomena such as hybridization, intraspecific gene flow and recombination directly, others are intended to depict uncertainty owing to conflicts in the data. Several network construction methods have been developed for visualizing whether data support a tree-like arrangement of taxa, conflict in repeated patterns or are random 'noise'. Phylogenetic network methods display conflict with additional internal branches or internodes representing the alternative resolutions of the relationships (splits) among the taxa.

Exploratory network methods, such as split decomposition [57], median and median-joining networks [58], Neighbor-Net [59] and Qnet [60], were designed to visualize site pattern incompatibilities within molecular data sets in the form of splitsgraphs. The internal reticulate relationships in these graphs represent character conflicts, including those resulting from stochastic noise. Many of these methods feature ways to reduce the complexity of the diagrams by displaying only the strongest or most recurrent patterns of conflict.

A second set of exploratory network methods uses gene trees rather than site patterns as input. Reticulation in the resulting network diagrams depicts incongruence between the gene trees. Consensus networks [61–63] extend the concept of consensus trees to display the possible relationships among taxa indicated by the different source trees. Currently, the source trees must have identical sets of taxa, but methods are being developed to add in taxa found in only some source trees. Thresholds can also be set so that only relationships that occur in some specified proportion of source trees are displayed. Supernetworks (such as the Z-closure method; [64]) generalize the concept of supertrees. They enable visualization of relationships among taxa indicated by multiple source trees with overlapping, but not necessarily identical, sets of taxa ('partial trees'). Koch *et al.* [55] used supernetworks to describe the evolution of *trn*F pseudogenes within the rapid radiation of Brassicaceae (Figure I). Methods for filtering splits (e.g. only showing splits that recur in more than one source tree) have also been developed for supernetworks [65].

Many of these implicit phylogenetic network methods are available in Splitstree 4.5. Median and median-joining networks are also available in the Network package; QNet has its own software package and Neighbor-Net is available on its own or within SplitsTree (see Table I for a summary of the software mentioned here).

**Table I. Software available for analyses mentioned in article**

| Package | Function | Website |
|---|---|---|
| MrBayes | Bayesian phylogeny estimation | http://mrbayes.csit.fsu.edu/ |
| Network | Median networks, median-joining networks | http://www.fluxus-engineering.com/netwinfo.htm |
| PAML | Likelihood estimation and simulation | http://abacus.gene.ucl.ac.uk/software/paml.html |
| PAUP* | Phylogeny estimation (multiple methods) | http://paup.csit.fsu.edu |
| Qnet | Constructing networks from quartets | http://www.cmp.uea.ac.uk/~vlm/qnet |
| SeqGen | Simulating sequence evolution on trees | http://evolve.zps.ox.ac.uk/software.html?id=seqgen |
| Spectronet | Spectral analysis, median networks, etc. | http://awcmee.massey.ac.nz/downloads.htm |
| Spectrum | Spectral analysis | http://taxonomy.zoology.gla.ac.uk/%7Emac/spectrum/spectrum.html |
| SplitsTree | Variety of network and supernetwork methods | http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome_en.html |
| TreePuzzle | Likelihood maps, likelihood estimation using quartets | http://www.tree-puzzle.de/ |

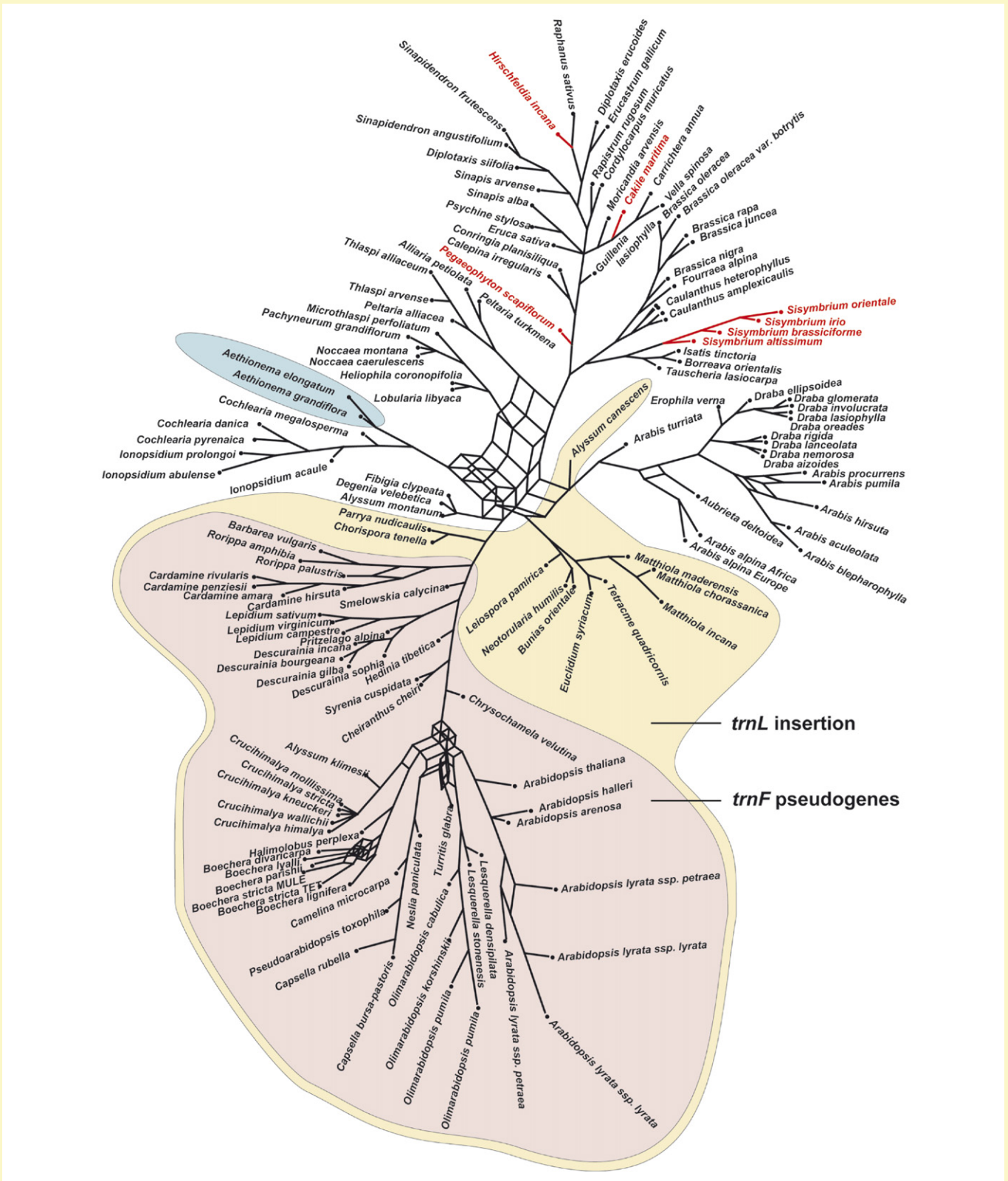


**Figure I**. Supernetwork of Brassicaceae sampled by Koch *et al.* [55], showing alternative resolutions of internal branches in the form of parallelograms, and with the nested *trn*F rearrangements highlighted by different colors. Red branches indicate the presence of alternative *trn*F pseudogenes; the blue–green shading highlights the outgroup taxa. As with combined-data phylogenetic analyses, the supernetwork enabled tree information from the analysis of five genes for which sequences were available from overlapping but not identical data sets to be combined. Unlike the results of typical phylogenetic analyses, however, the supernetwork can display alternative signals in the data simultaneously. Reproduced with permission from Ref. [55].

this potential problem, it has recently been pointed out that, with some topology–branch length combinations, the most probable individual gene trees can be incongruent with the underlying species tree [20]. It is not yet known how pervasive this is in empirical studies, but it highlights the fact that there might be parts of phylogenies for which the addition of more data will not result in convergence on the correct answer.

At the other end of the internode branch-length spectrum, long times since divergence (i.e. long external branches in the tree) provide significant opportunity for the loss of phylogenetic signal through substitutional saturation. In addition, lineage-specific deviations in substitution processes from the assumed substitution model (e.g. differences among lineages in preferred types and spatial patterns of substitution) can have a significant effect. Where external branches are long, and internal branches short (e.g. when basal internodes are closely spaced in time), even small deviations from the assumed nucleotide or amino acid substitution model can reduce reconstruction accuracy [21]. Theory suggests that, as the time since divergence becomes extremely long, it can be impossible to reconstruct phylogenetic history using DNA sequence data [22,23]. This conclusion also follows from some empirical results [24,25]. For extremely ancient divergences, other more rare genetic events might be required to resolve the sequence of divergences [26].

These difficulties in determining relationships among rapidly diverged lineages make it particularly difficult to place accurately the root of the tree, because there is a tendency for the outgroup to join the tree on one of the longer branches, even if its correct position is on a short central internode [6,13,16]. This finding is perhaps unsurprising given that the branch lengths to outgroup taxa tend to be longer than for ingroup taxa.

## Ancient rapid radiation or inadequate data?

Ancient rapid radiations lead to reconstructed phylogenies with low support for basal relationships. However, poor internal branch support in phylogenies can also be obtained by: (i) using molecular or morphological data that are not variable enough at the appropriate level; (ii) having data sets that strongly conflict with one another; (iii) applying inappropriate phylogenetic methods and substitution models; or (iv) not having enough data to solve the problem. So when does it appear that rapid radiation is the appropriate explanation for poor phylogenetic resolution?

First, a rapid radiation will tend to defy resolution using most types of data. By contrast, if a lack of resolution is not caused by truly short times between divergences, relationships should ultimately be resolvable using data sources with appropriate levels of variation for the target age of divergence, as long as the lineages have not accumulated a large quantity of conflicting data patterns in the time since the divergence. As Donoghue and Sanderson [27] pointed out, the ideal characters for solving the rapid radiation problem would be those that initially evolved as rapidly as the diverging lineages, but then evolved more slowly subsequently. Such characters are difficult to find; homoplasy-free rare changes might be the only option [26].

One can test to some extent for agreement on the reality of internal short branches by testing whether different data sets agree that the same branches are short (i.e. branch lengths are highly correlated), especially for short internal branches [28]. More global statistical tests for the existence of hard polytomies (real polytomies or those where internal branch lengths are statistically indistinguishable from zero even with sufficient data [29]), have also been under development. However, with any of these methods, there are statistical issues to be considered, such as the relative value of increasing data input from one source versus adding more independent inputs from other sources.

## Eliminating character conflict as an explanation

Short branch lengths (i.e. low support) in phylogenies can result from different characters or data sources providing support for conflicting trees, rather than from the absence of support. Because this kind of conflict among characters can stem from causes unrelated to the absence of sufficient time for support to accumulate in clades, it can be useful to first test whether short internal branches are due to character conflict. If they are, it is likely that additional data will more strongly support one of the competing phylogenetic resolutions and that a hard polytomy will not result. Conflicting but essentially equally parsimonious phylogenies can be visualized using network methods that keep track of alternative resolutions implied by the data (Box 1). In addition, several other graphical methods have been developed to visualize and quantify data conflict relative to the phylogenetic signal (Box 2).

## How much data will be required to resolve relationships?

If there has been simultaneous diversification of multiple taxa within a group, the phylogeny is more accurately represented by a polytomy rather than by a bifurcating tree [30]. Thus, we might expect in these cases that no amount of data will resolve a bifurcating tree. If instead the time between lineage splits is short but at least detectable, it is possible to estimate (i.e. extrapolate) how much additional data it would take to resolve the branching order.

Several statistical tests for rapid radiations (actual or near-hard polytomies) have been developed that differ in whether they use a simulation approach and whether they assume a discrete cut-off point for how much branch length is required before a hard polytomy is rejected. Two popular tests are forms of 'power analysis' in the sense that they test either how long a branch length must be to be considered different from zero, or how much data would be required to obtain such a branch length.

The first of these tests, from Fishbein *et al.* [31], uses resampling from the original data to simulate new data sets of increasing size, and then uses (parametric or nonparametric) bootstrap support as a measure of when enough branch length is obtained to obtain evidence of sequential diversifications. Whereas the original study focused upon the parametric bootstrap, the nonparametric bootstrap is likely to better capture evolutionary properties of 'real' data, in that it resamples from the original data rather than simulating new data using a model.

An alternative approach from Walsh *et al.* [32] first establishes what kind of interdivergence time interval one would consider significantly different from a hard polytomy, then how much sequence change would be expected to accumulate during this time period given the data. It then compares this amount of change to that observed in the data to determine how far along one is on the 'power curve' towards distinguishing successive cladogenesis.

Both tests approach the question of how much data would be needed to resolve the short internal branches if more data that are similar to the original data are available. They cannot accurately predict whether the near-polytomy would be resolved using new genes with novel patterns of variation, but they can be useful in providing a rough idea of how much data might be required.

Neither of the two power tests is implemented directly via a computer package. The Fishbein *et al.* [31] approach

**Box 2. Data exploration: obtaining graphical summaries of conflict versus signal**

A variety of graphical methods are available for examining conflicting signals in molecular sequence data. All have their strengths and weaknesses, but they can be useful for understanding data patterns and phylogenetic utility of sets of sequences. It can be valuable to explore whether a set of data contains a tree-like signal before it is analyzed using phylogeny estimation methods, or to select appropriate data sets for analysis. Lento plots ([66]; Figure Ia) display support and conflict in a data set as a series of bars (each representing a split) extending above (support) and below (conflict) a horizontal line. The height or depth of each bar corresponds to the proportion of data patterns that either support or conflict with that split. The main advantage of Lento plots is that they enable one to identify the amount of support and conflict for individual splits or putative clades. Lento plots can be obtained using either the Spectrum or Spectronet computer packages. A current disadvantage of these implementations is that the graphs are limited to relatively small numbers of taxa (currently ~20), as exact enumeration of splits is performed via the Hadamard transform. However, it is possible to obtain splits with larger taxon sets heuristically using Neighbor-Net before making a Lento Plot.

An alternative type of exploratory graph for visualizing conflict versus support in a data set is likelihood mapping [67]. In this case a triangular plot of probability vectors is produced for each quartet of taxa or sequences (Figure Ib). Points falling in the three corners of the triangle suggest strong support for a tree-like pattern in the data. The central triangle contains points that indicate a lack of support for any resolution (star-like phylogeny), whereas points in the peripheral sectors between the corners indicate uncertainty between support for two different resolutions of the quartet (conflict). Clean data, with a strong phylogenetic signal, should produce points falling mainly in the three corners, with a scattering of points elsewhere; data in which multiple superimposed substitutions at the same site have obscured the history of change will produce many points in the center along the three radii. Likelihood mapping results in easily interpretable visualizations of signal even for large data sets, but does have the weaknesses that internal edges (internodes) are not separated from external nodes (trivial splits for each individual taxon), and that quartets do not always fully capture the complexity of data patterns and, thus, might not be compatible with a single tree. Likelihood maps can be obtained using the TreePuzzle package.

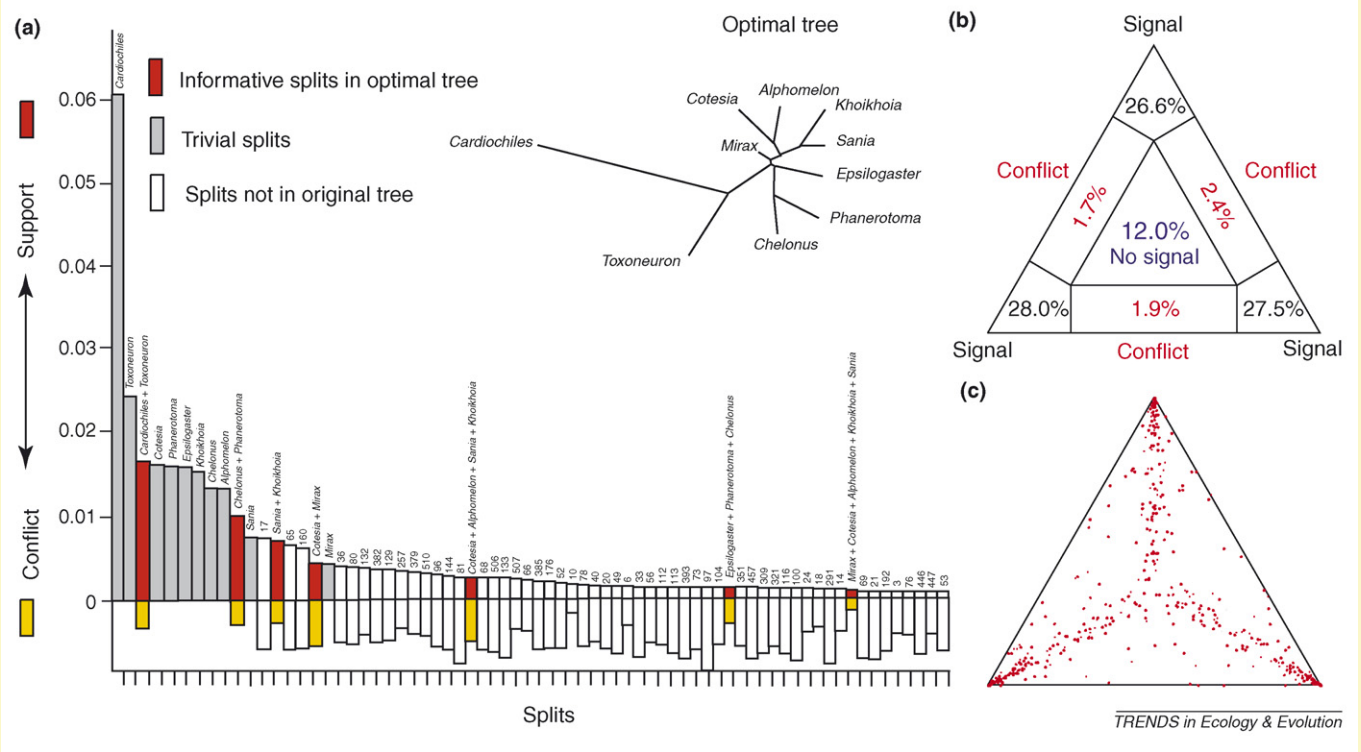


**Figure I.** Visualizing support versus conflict using Lento Plots and likelihood maps. **(a)** Lento plot [66] displaying conflict and support for the splits found in the optimal tree (shown) and alternative trees. The height of each bar represents the frequency of support for, or conflict with, that particular split among the data patterns. The plot color codes those splits occuring in the optimal estimated tree differently from those occurring in alternative trees. The plot also highlights trivial splits including only one taxon differently from phylogenetically information splits. **(b)** and **(c)** show likelihood maps [67] for the same set of data, categorizing amounts of conflict and support summarized from quartets of taxa (c) into sectors (b) of a triangular graph. Clear phylogenetic signals are represented by dots concentrating in the three corners of the triangle (representing the three resolutions of each quartet), whereas dots along the three central radii represent more random 'noise'; dots located near the edge between the corners represent conflicting signals. The data for the examples here come from a phylogenetic study of microgastrine wasp genera (and related subfamilies) based on seven genes [48], in which a relatively low phylogenetic signal was found.

can be accomplished by using a standard phylogeny estimation program such as either PAUP* [33] or MrBayes 3.1 [34] to obtain the optimal topology, branch lengths and substitution model, then SeqGen [35] or PAML [36]to simulate new bootstrap data sets based upon these and, finally, the original program to re-estimate the phylogeny. The Walsh *et al.* [32] test requires use of standard statistical packages in addition to estimation of the optimal topology–branch length–model combination for the original data.

These tests also require further testing with real data. The ultimate application of the results of such analyses are limited by the kinds of data available (e.g. how many genes are truly available with the appropriate levels of variation to resolve splits given the age of lineage divergence, and are the genes already used a good sample of these?). In addition, extrapolation from sampled genes does not take into account potential interactions among signals from any new genes that might be added.

## Outstanding questions

The unearthing of ancient rapid radiations has led to several questions that are still unresolved. For instance, are there ancient patterns of divergence that are impossible to resolve using molecular sequence data? Theoretical explorations [22,23] suggest that there are, but empirical studies are still largely data limited. If sequence data do not suffice, will rare genomic changes such as gene content, presence–absence of biochemical pathways, gene arrangements, intron and transposon positions, secondary structures, and so on [26] solve some deep phylogenetic problems that other data cannot?

In a broader evolutionary context, it would be of interest to examine to what extent ancient rapid radiations are correlated in time with one another. Following major extinctions, fossil evidence often records a rebounding of biotic diversity. Convincing molecular systematic studies of animal–plant co-divergence following such extinction events are still relatively limited, but are likely to be a main focus of future research.

Rokas and colleagues [37,38] have recently emphasized that there are some radiations (most prominently the 'Cambrian Explosion' of animal phyla) that might be unresolvable using even large amounts of DNA sequence data. These authors place an emphasis on the exploitation of homoplasy-free rare genomic changes, which are underexploited and show considerable promise for resolving ancient radiations. By contrast, other authors [39,40] have argued that effective use of DNA sequence data depends on the application of relatively accurate substitution models fitted to each data partition (gene, codon position, etc.). In other words, they suggest that one can make better use of the data one has, if the data are thoroughly explored.

## The value of data exploration

Modern phylogenetic analysis has become an increasingly complex task as the focus of study widened to encompass the full range of hierarchical questions. It is evident that many phylogenetic histories will require significant scientific effort to decipher, not necessarily because our methods are still poor but because the histories are truly challenging to recover. In some cases, we will exceed the limits of resolution of certain classes of data and many phylogenetic methods. It will be essential in these situations to maintain an open mind towards alternative ways of looking at data, and finding new ways to extract useful information from a variety of perspectives. In this context, it is encouraging to see so many new avenues of data interpretation being explored.

## Acknowledgements

The original motivation for this review came from molecular phylogenetic research on microgastrine braconid wasps funded originally by USDA NRI grant 9501893 and continued under NSF grant DEB 0316566. A Sabbatical Leave Grant to J.B.W. from the Allan Wilson Centre facilitated the writing of the review during February–April 2006, as well as valuable input from Sydney Cameron, Barbara Holland, Daniel Huson, David Penny and Mike Steel.

## References

1 Murphy, W.J. *et al.* (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 294, 2348–2351
2 Springer, M.S. *et al.* (2003) Placental mammal diversification and the Cretaceous–Tertiary boundary. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1056–1061
3 Kolaczkowski, B. and Thornton, J.W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431, 980–984
4 Steel, M. (2005) Should phylogenetic models be trying to 'fit an elephant'? *Trends Genet.* 21, 307–309
5 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410
6 Hendy, M.D. and Penny, D. (1989) A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38, 297–309
7 Penny, D. *et al.* (1991) Testing the theory of descent. In *Phylogenetic Analysis of DNA Sequences* (Miyamoto, M. and Cracraft, J., eds), pp. 155–183, Oxford University Press
8 Anderson, F.E. and Swofford, D.L. (2004) Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Mol. Phylogenet. Evol.* 33, 440–451
9 Kim, J. (1996) General inconsistency conditions for maximum parsimony: effects of branch length and increasing the number of taxa. *Syst. Biol.* 45, 363–374
10 Steel, M. (2001) Sufficient conditions for two tree reconstruction techniques to succeed on sufficiently long sequences. *SIAM J. Discr. Math.* 14, 36–48
11 Kennedy, M. *et al.* (2005) Untangling long branches: identifying conflicting phylogenetic signal using spectral analysis, Neighbor-Net, and consensus networks. *Syst. Biol.* 54, 620–633
12 Lockhart, P. and Steel, M. (2005) A tale of two processes. *Syst. Biol.* 54, 948–951
13 Lockhart, P.J. and Cameron, S.A. (2001) Trees for bees. *Trends Ecol. Evol.* 16, 84–88
14 Lockhart, P.J. and Penny, D. (2005) The place of *Amborella* within the radiation of angiosperms. *Trends Plant Sci.* 10, 201–202
15 Hillis, D.M. *et al.* (1994) Hobgoblin of phylogenetics? *Nature* 369, 363–364
16 Holland, B.R. *et al.* (2003) Outgroup misplacement and phylogenetic inaccuracy under a molecular clock-a simulation study. *Syst. Biol.* 52, 229–238
17 Nichols, R. (2001) Gene trees and species trees are not the same. *Trends Ecol. Evol.* 7, 358–364
18 Rosenberg, N.A. (2002) The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.* 61, 225–247
19 Maddison, W.P. and Knowles, L.L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55, 21–30
20 Degnan, J.H. and Rosenberg, N.A. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genetics* 2, e68 0762–0768
21 Jermiin, L. *et al.* (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst. Biol.* 53, 638–643

22 Mossel, E. and Steel, M. (2004) A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.* 187, 189–203
23 Mossel, E. and Steel, M. (2005) How much can evolved characters tell us about the tree that generated them? In *Mathematics of Evolution and Phylogeny* (Gascuel, O., ed.), pp. 384–412, Oxford University Press
24 Martin, W. *et al.* (2005) Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. *Trends Plant Sci.* 10, 203–209
25 Ho, S.Y.W. and Jermiin, L.S. (2004) Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.* 53, 623–637
26 Boore, J.L. (2006) The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol. Evol.* 21, 439–446
27 Donoghue, M.J. and Sanderson, M.J. (1992) The suitability of molecular and morphological evidence in reconstructing plant phylogeny. In *Molecular Systematics of Plants* (Soltis, P.S. *et al.*, eds), pp. 340–368, Chapman & Hall
28 Mardulyn, P. and Whitfield, J.B. (1999) Phylogenetic signal in the COI, 16S, and 28S genes for inferring relationships among genera of Microgastrinae (Hymenoptera: Braconidae): evidence of a high diversification rate in this group of parasitoids. *Mol. Phylo. Evol.* 12, 282–294
29 Maddison, W.P. (1989) Reconstructing character evolution on polytomous cladograms. *Cladistics* 5, 365–377
30 Hoelzer, G.A. and Melnick, D.J. (1994) Patterns of speciation and limits to phylogenetic resolution. *Trends Ecol. Evol.* 9, 104–107
31 Fishbein, M. *et al.* (2001) Phylogeny of Saxifragales (Angiosperms, Eudicots): analysis of a rapid, ancient radiation. *Syst. Biol.* 50, 817–847
32 Walsh, H.E. *et al.* (1999) Polytomies and the power of phylogenetic inference. *Evolution* 53, 932–937
33 Swofford, D.L. (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4,* Sinauer
34 Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574
35 Rambaut, A. and Grassly, N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* 13, 235–238
36 Yang, Z. (2004) *PAML: Phylogenetic Analysis by Maximum Likelihood*, (http://abacus.gene.ucl.ac.uk/software/paml.html)
37 Rokas, A. *et al.* (2005) Animal evolution and the molecular signature of radiations compressed in time. *Science* 310, 1933–1938
38 Rokas, A. and Carroll, S.B. (2006) Bushes in the tree of life. *PLoS Biol.* 4, 1899–1904
39 Baurain, D. *et al.* Lack of resolution in the animal phylogeny: closely spaced cladogenesis or undetected systematic errors? *Mol. Biol. Evol.* (in press)
40 Kelchner, S.A. and Thomas, M.A. (2007) Model use in phylogenetics: nine key questions. *Trends Ecol. Evol.* 22, 87–94
41 Cooper, A. and Penny, D. (1997) Mass survival of birds across the Cretaceous—Tertiary boundary: molecular evidence. *Science* 275, 1109–1113
42 Poe, S. and Chubb, A.L. (2004) Birds in a bush: five genes indicate explosive evolution of avian orders. *Evolution* 58, 404–415
43 Slack, K.E. *et al.* (2006) Early penguin fossils, plus mitochondrial genomes, calibrate avian evolution. *Mol. Biol. Evol.* 23, 1144–1155
44 Shaffer, H.B. *et al.* (1997) Tests of turtle phylogeny: molecular, morphological and paleontological approaches. *Syst. Biol.* 46, 235–268
45 Clements, K.D. (2002) Rapid evolutionary divergences in reef fishes of the family Acanthuridae (Perciformes: Teleostei). *Mol. Phylogenet. Evol.* 26, 190–201
46 Cameron, S.A. and Mardulyn, P. (2001) Multiple molecular data sets suggest independent origins of highly social behavior in bees (Hymenoptera: Apinae). *Syst. Biol.* 50, 194–214
47 Cameron, S.A. (2003) Data from the elongation factor-1 gene corroborates the phylogenetic pattern from other genes revealing common ancestry of bumble bees and stingless bees. *Seminario Mesoamericano sobre Abejas sin Aguijon, III* 132–136
48 Banks, J.C. and Whitfield, J.B. (2006) Dissecting the ancient rapid radiation of microgastrine wasp genera using additional nuclear genes. *Mol. Phylogenet. Evol.* 41, 690–703
49 Moulton, J.K. (2003) Can the current molecular arsenal adequately track rapid divergence events within Simuliidae (Diptera)? *Mol. Phylogenet. Evol.* 27, 45–57
50 Von Dohlen, C. and Moran, N.A. (2000) Molecular data support a rapid radiation of aphids in the Cretaceous and multiple origins of host alternation. *Biol. J. Linn. Soc.* 71, 689–717
51 Aris-Brosou, S. and Yang, Z. (2003) Bayesian models of episodic evolution support at late Precambrian explosive diversification of the Metazoa. *Mol. Biol. Evol.* 20, 1947–1954
52 Mathews, S. and Donoghue, M.J. (1999) The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286, 947–950
53 Soltis, D.E. *et al.* (2004) Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. *Trends Plant Sci.* 9, 477–483
54 Wortley, A.H. *et al.* (2005) How much data are needed to resolve a difficult phylogeny: case study in Lamiales. *Syst. Biol.* 54, 697–709
55 Koch, M.A. *et al.* (2007) Supernetwork identifies multiple events of plastid $trnF_{(GAA)}$ pseudogene evolution in the Brassicaceae. *Mol. Biol. Evol.* 24, 63–73
56 Al-Shehbaz, I.A. *et al.* (2006) Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant Syst. Evol.* 259, 89–120
57 Bandelt, H-J. and Dress, A.W. (1992) Split decomposition: a new useful approach to phytogenetic analysis of distance data. *Mol. Phylogenet. Evol.* 1, 242–252
58 Bandelt, H-J. *et al.* (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48
59 Bryant, D. and Moulton, V. (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21, 255–265
60 Grünewald, S. *et al.* QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. *Mol. Biol. Evol.* (in press)
61 Holland, B.R. and Moulton, V. (2003) Consensus networks: a method for visualizing incompatibilities in collections of trees. *Workshop on Algorithms in BioInformatics (WABI)* 2006, 165–176
62 Holland, B. *et al.* (2005) Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. *Syst. Biol.* 54, 66–76
63 Holland, B.R. *et al.* (2006) Improved consensus network techniques for genome-scale phylogeny. *Mol. Biol. Evol.* 23, 848–855
64 Huson, D.H. *et al.* (2004) Phylogenetic super-networks from partial trees. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 151–158
65 Huson, D.H. *et al.* (2006) Reducing distortion in phylogenetic networks. *Workshop on Algorithms in Bioinformatics (WABI)* 2006, 150–161
66 Lento, G.M. *et al.* (1995) Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Mol. Biol. Evol.* 12, 28–52
67 Strimmer, K. and von Haeseler, A. (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. U. S. A.* 94, 6815–6819