Review

Gene tree discordance, phylogenetic inference and the multispecies coalescent

James H. Degnan[1,2] and Noah A. Rosenberg[1,3,4]

[1] Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[2] Current address: Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
[3] Center for Computational Medicine and Biology, University of Michigan, Ann Arbor, MI 48109, USA
[4] Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA

The field of phylogenetics is entering a new era in which trees of historical relationships between species are increasingly inferred from multilocus and genomic data. A major challenge for incorporating such large amounts of data into inference of species trees is that conflicting genealogical histories often exist in different genes throughout the genome. Recent advances in genealogical modeling suggest that resolving close species relationships is not quite as simple as applying more data to the problem. Here we discuss the complexities of genealogical discordance and review the issues that new methods for multilocus species tree inference will need to address to account successfully for naturally occurring genomic variability in evolutionary histories.

The problem of gene tree discordance

Until recently, the state of the art for molecular phylogenetic studies typically involved (i) sequencing a gene in individual representatives of a collection of species; (ii) inferring a 'gene tree' (see Glossary) for the sequences; and (iii) declaring the gene tree to be the estimate of the tree of species relationships. With the increasing abundance of molecular data and the recognition that evolutionary trees from different genes often have conflicting branching patterns [1–8], it is becoming increasingly feasible to implement multilocus approaches to phylogenetic inference. Many of the first studies to examine the conflicting signal of different genes have found considerable discordance across gene trees: studies of hominids [9–11], pines [12], cichlids [13], finches [14], grasshoppers [15] and fruit flies [16] have all detected genealogical discordance so widespread that no single tree topology predominates. These examples highlight the issue of 'incomplete lineage sorting' (Box 1) and the need to account for gene tree discordance in phylogenomic studies.

Concurrent with the proliferation of empirical studies of gene tree discordance, new analytical and simulation tools have increasingly made it possible to investigate the magnitude of this discordance under probabilistic models of how genetic lineages evolve across species. This theoretical work also finds that high levels of discordance are often expected. Most strikingly, methods such as 'democratic vote' and concatenation can be more likely to result in an incorrect species tree as more data are added.

Here we describe how gene tree discordance can be predicted under a widely used evolutionary model, the coalescent, applied to multiple species. We also describe the conceptual basis for gene tree discordance and methods

Glossary

Ancestral polymorphism: the existence of more than one allele at a locus in an ancestral population; through incomplete lineage sorting, polymorphisms can persist through species divergences, resulting in misleading similarities of DNA sequences that do not necessarily reflect population relationships.
Anomalous gene tree (AGT): a gene tree topology that is more probable than the gene tree topology that matches the species tree topology.
Anomaly zone: for a given species tree topology, the set of branch lengths for which there is at least one AGT.
Coalescent event (or coalescence): the most recent common ancestral gene for a pair of gene lineages; coalescent events correspond to nodes on gene trees.
Coalescent history: for a given gene tree–species tree pair, a list specifying the ancestral populations of the species tree in which the gene tree coalescences occur. The set of coalescent histories compatible with a gene tree–species tree pair depends only on the topologies of the species tree and gene tree. A coalescent history can be compatible with more than one sequence of coalescences within a population.
Coalescent time unit: a unit of time normalized by population size. If T is the number of generations of a species tree branch, and N_e is the effective number of chromosomes in the population, then T/N_e is the length of the branch in coalescent time units. Thus, 1.0 coalescent time units corresponds to N_e generations, and a short branch can arise from a small number of generations, a large population size, or both.
Gene tree: a tree of ancestor–descendant relationships for a gene (or locus), where the same gene is sampled from several individuals. Nodes of a gene tree are coalescent events. We use 'gene tree' to refer only to a topology, but branch lengths can also be of interest. We use 'gene genealogy' to refer to a gene tree with branch lengths.
Incomplete lineage sorting: the failure of two or more lineages in a population to coalesce, leading to the possibility that at least one of the lineages first coalesces with a lineage from a less closely related population.
Monophyly: the condition in which the most recent ancestral copy of a set of lineages is not an ancestor of any lineages outside the set. We use this term to refer to gene lineages.
Multispecies coalescent: the coalescent model applied to gene trees in a species tree; this model is used to assemble separate coalescent processes occurring in populations connected by an evolutionary tree.
Pectinate: a branching pattern for a bifurcating tree in which each internal node has at least one branch connected to a tip of the tree, such as for the tree ((((AB)C)D)E).
Species tree: a tree of ancestor–descendant relationships for a set of populations. Branch lengths depend on time measured in number of generations and on effective population sizes. In our species tree diagrams, the height of a branch indicates time in generations, while the width of a branch is often drawn proportionally to N_e.

Corresponding authors: Degnan, J.H. (j.degnan@math.canterbury.ac.nz); Rosenberg, N.A. (rnoah@umich.edu).

0169-5347/$ – see front matter © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2009.01.009 Available online 21 March 2009

Box 1. Incomplete lineage sorting

'Lineage sorting' and 'incomplete lineage sorting' are used in several ways by different authors. Some authors (including us) use them primarily as descriptions of particular types of genealogical pattern. Other authors use them to describe a process that explains the gene tree discordance detected in genetic data, and require that genetic data be investigated before the terms apply. Still others describe 'lineage sorting' as 'complete' when polymorphism no longer exists at a locus in descendant populations [22,75]. The term 'hemiplasy' has been suggested [76] for gene tree incongruence specifically caused by incomplete lineage sorting when ancestral polymorphism is retained through speciation events.

An important insight from coalescent theory is that ancestry of lineages can be modeled independently of the process of mutation [18]. Thus, incongruent gene trees can occur even without ancestral polymorphism – or without any present-day polymorphism. Although detecting gene tree incongruence (or incomplete lineage sorting) does depend on the occurrence of mutations, detectability is conceptually distinct from whether incongruence (or incomplete lineage sorting) exists. Because gene trees are expected to sometimes disagree with the species tree independently of the existence of polymorphism, we suggest that 'incomplete lineage sorting' be used only to refer to failures of lineages in a population to coalesce. Whether such failures result in incongruent gene trees depends on coalescences in ancestral populations. With this definition, incongruence is not built into the concept of incomplete lineage sorting, and the usage parallels the way HGT, gene duplication, hybridization, recombination, natural selection and other phenomena are cited as potential causes of gene tree incongruence.

for obtaining gene tree probabilities given a species tree. We discuss implications of gene tree discordance and the 'multispecies coalescent' for experimental design, and review new approaches that allow for high levels of gene tree discordance when inferring species trees. Finally, we conclude with a proposed list of questions for framing future investigations of gene tree discordance, incomplete lineage sorting and multilocus phylogenetics.

The multispecies coalescent

Coalescent theory [1,2,17], which models genealogies within populations, can be used to investigate probabilities that gene trees have branching patterns (topologies) that differ from a species tree topology. The basic model, which we call the 'multispecies coalescent,' generalizes the Wright-Fisher model of genetic drift [18–20], applying it to multiple populations connected by an evolutionary tree.

The coalescent for a single population traces the ancestries of a subset of individual copies of a gene backward in time from the present. Figure 1a depicts a population shaded in blue with five (haploid) individuals, tracing the ancestries of three of the individuals back ten generations. The population is assumed to have constant size and nonoverlapping generations. Each gene is copied from a random 'parental' gene in the previous generation. The coalescent model approximates the process of choosing random parents backward in time when the population size is large relative to the number of sampled lineages [18–20].

In population genetics, the coalescent is typically applied to several individuals sampled from one population. In phylogenetics, individuals from the same population are usually assumed to be similar compared to the differences that exist among populations (or species) and, often, only one individual is sampled per population.

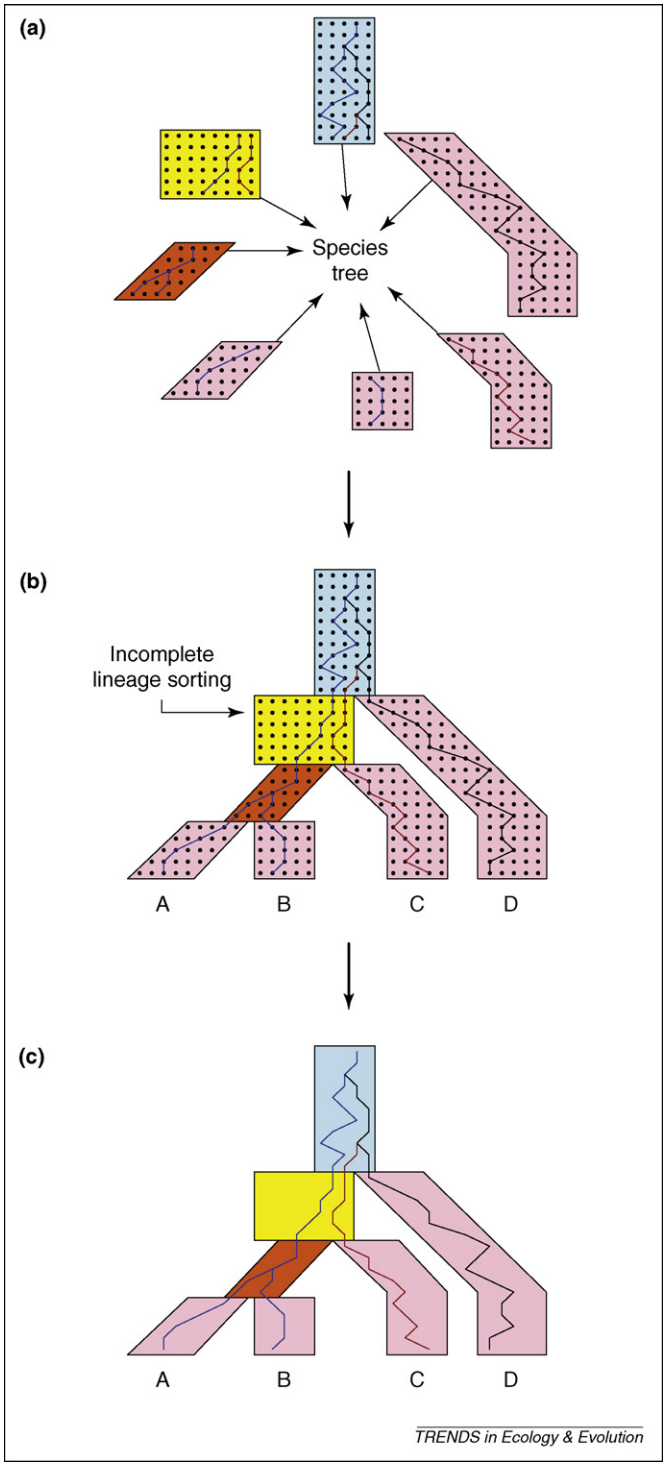

Figure 1. The multispecies coalescent. Each dot represents an individual gene copy, with each row representing one generation. Lines connect an individual gene copy to its ancestor in the previous generation, one row higher. The width of a population represents the population size, and the height represents time measured in generations. **(a)** The coalescent in several populations. The four populations shaded pink each have only one lineage (gene copy) sampled per species. **(b)** Populations arranged by evolutionary relationships. Because the lineage ancestral to the gene sampled from population C fails to coalesce in the population in yellow, this lineage can coalesce with the D lineage before coalescing with the lineage ancestral to the lineages sampled from populations A and B. Consequently, the gene tree topology is ((AB)(CD)), whereas the species tree topology is (((AB)C)D). **(c)** A gene tree in a species tree, obtained by ignoring individuals that are not ancestral to individuals in the sample.

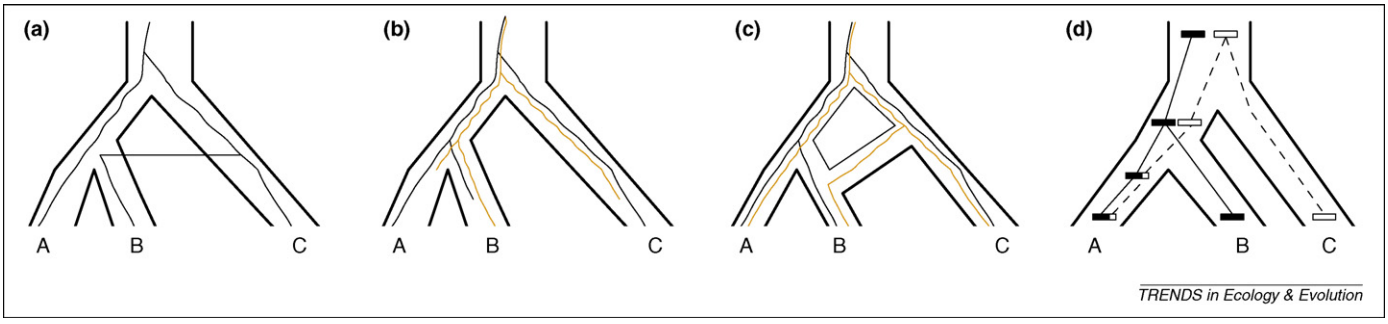

Figure 2. Sources of gene tree–species tree discordance other than incomplete lineage sorting. **(a)** HGT: a lineage jumps from the population ancestral to A and B to the population ancestral to C, leading to the gene tree (A(BC)). **(b)** Gene duplication and loss: through extinction of lineages, gene duplication can produce apparent relationships incongruent with the species tree. Even if paralogs are not lost, the sampling of lineages that are not true orthologs can cause lineages from A and C to appear more closely related to each other than either is to B. **(c)** Hybridization causes some genes sampled from species B to descend from the population ancestral to A and B, whereas others descend from the population ancestral to B and C. The two gene trees depicted in (c) are ((AB)C) (black) and (A(BC)) (orange). Hybridization affects whole genomes, whereas HGT typically affects only small DNA segments. **(d)** Recombination can lead to different histories for neighboring segments within a gene. For the DNA segment depicted in black, the gene tree is ((AB)C), but for the segment in white, the gene tree is ((AC)B).

However, the coalescent still applies because two or more lineages can coexist in the same ancestral population (Figure 1b,c). For studies of closely related populations, differences among genes from separate populations can be similar in magnitude to differences among genes within a population; consequently, multiple gene copies (alleles) per population are often sampled [21,22].

Considering multiple populations, the multispecies coalescent can be used to describe a probability distribution of random gene trees that evolve along the branches of a species tree [1,2,5,23–27]. Gene lineages from different species trace backward through time, finding common ancestors at rates specified by the model. Coalescences of gene lineages from separate species can only occur more anciently than the splitting times of the species to which they belong.

In its simplest form for a non-recombining locus, the multispecies coalescent inherits many of the assumptions of the Wright–Fisher model: constant effective population sizes (N_e) within (but not necessarily across) populations; neutral evolution for the loci modeled; no structure within populations; and random joining of lineages backward in time, so that all pairs of lineages in a population are equally likely to coalesce. It also accommodates multiple individuals (alleles or lineages) sampled per species [23,24,28–31].

Box 2. Coalescent time units

Branch lengths on species trees, measured in coalescent time units, depend on both the number of generations and N_e. Thus, a small number of generations need not produce a branch that is short in coalescent time units (Table I). For example, with 10 000 diploid individuals or $N_e = 20\,000$ chromosomes, if the length of time is $T = 100\,000$ generations, then the branch length is $T/N_e = 100\,000/20\,000 = 5.0$ coalescent time units. For the same number of generations, $N_e = 100\,000$ diploid individuals would imply a branch length of 0.5 coalescent time units.

Gene tree branch lengths are often measured in terms of the expected number of mutations. For diploids, branch lengths in coalescent time units can be converted into mutation units by multiplying by $\theta/2$, where $\theta = 2N_e\mu$ and μ is the mutation rate per site per generation. This computation works because $(\theta/2)\cdot T/N_e = \mu T$, the expected number of mutations that occur in T generations. (If $2N_e$ is used as the effective population size, $\theta = 4N_e\mu$ and $(\theta/2)\cdot T/(2N_e) = \mu T$.) For example, if $\theta = 0.01$, 0.5 coalescent time units corresponds to $(0.01)(0.5/2) = 0.0025$ mutation units. This corresponds to an expected 2.5 mutations per 1000 sites along this branch. Mutation units can be converted into coalescent time units by dividing by $\theta/2$.

What branch lengths on species trees occur in real data? For the species tree (((HC)G)O) for human, gorilla, chimpanzee and orangutan, using an estimated time from the gorilla divergence to the split between humans and chimps of 1.2 million years, and $N_e/2 = 24\,600$ individuals (= 49 200 for the number of autosomal gene copies) and a generation time of 20 years [30], this value corresponds to 1 200 000/60 000 generations and, therefore, to 60 000/49 200 $\approx$ 1.2 coalescent time units. A similar calculation yields ~4.2 coalescent time units separating the branch leading to orangutans from the most recent common ancestor of humans, chimpanzees and gorillas. Shorter coalescent branch lengths can occur with larger population sizes and faster population divergences. Passerina buntings have been estimated to have N_e near 1 000 000 individuals and intervals between speciation events as small as ~100 000 generations [63], suggesting branches as short as 0.05 coalescent time units.

Probabilities of gene tree topologies (online Supplementary Box S1) given species trees with branch lengths can be calculated using the program COAL [25] by enumerating coalescent histories [77,78]. Using the species tree (((HC)G)O) and branch lengths based on Ref. [30] yields probabilities of 0.79 for the gene tree (((HC)G)O) and 0.099 for each of the gene trees (((HG)C)O) and (((CG)H)O). These values agree closely with a genome-wide analysis using ~12 000 genes [11].

Table I. Coalescent time units for different combinations of N_e and number of generations

Number of generations	N_e				
	10 000	50 000	100 000	500 000	1 000 000
10 000	1	0.2	0.1	0.02	0.01
50 000	5	1	0.5	0.1	0.05
100 000	10	2	1	0.2	0.1
500 000	50	10	5	1	0.5
1 000 000	100	20	10	2	1

The multispecies coalescent is perhaps the simplest model available for making quantitative predictions about probabilities of gene trees, and it generalizes a standard model used for within-species population-genetic data [18–20,32,33]. When exact predictions are difficult, gene trees can be easily simulated under the model. Additionally, the multispecies coalescent can serve as a baseline for investigating diverse causes of gene tree discordance (Figure 2). The model has also been extended to include within-species migration [34–36], hybridization [37], horizontal gene transfer (HGT) between species [38] and recombination [27,39,40]. This flexibility makes the coalescent particularly useful for multispecies studies and provides a natural model for gene tree discordance.

Conceptual basis for discordance

Given enough time measured in coalescent time units (Box 2), lineages within a population coalesce with high probability. After $\sim 5N_e$ generations along species tree branches, where N_e is the effective number of chromosomes, lineages are likely to have coalesced within each population, and monophyly of lineages (and, therefore, congruence between gene trees and the species tree) is probable [3,25,29,41,42]. With shorter branches, multiple gene lineages tend to persist into deeper portions of the species tree. Coalescences can then occur between lineages that are not from the most closely related species, resulting in discordant gene trees: lineages do not necessarily 'sort' by species when they are coalescing, and 'incomplete lineage sorting' becomes probable (Figure 1b).

Although incomplete lineage sorting is typical of shallow species trees, where taxa are closely related and the root of the tree is recent, it can also occur in deep phylogenies. For some combinations of branching patterns and branch lengths, lineages are likely to sort in a way that violates monophyly of lineages for a species deep in the tree [21,43]. A disagreement between the gene and species tree topologies can get 'stuck' deep in the past, leading to discordance in the present. This phenomenon requires some short branches, possibly only one, deep in the tree.

Good candidates for ancient incomplete lineage sorting are ancient rapid radiations [44], in which short ancient species tree branches are likely to be common. Potential examples include the early period in bird evolution [45], the radiation of South American rodents [46] and the more recent radiations of *Drosophila* [16] and cichlids [13]. Lineage sorting has also been cited as a possible explanation for gene tree conflict in deeper phylogenies, such as in the most ancient splits within the mammals [47]; however, in such cases, divergence time estimates can be too uncertain to be confident that incomplete lineage sorting is likely. Short branches can also be less likely in deep phylogenies:

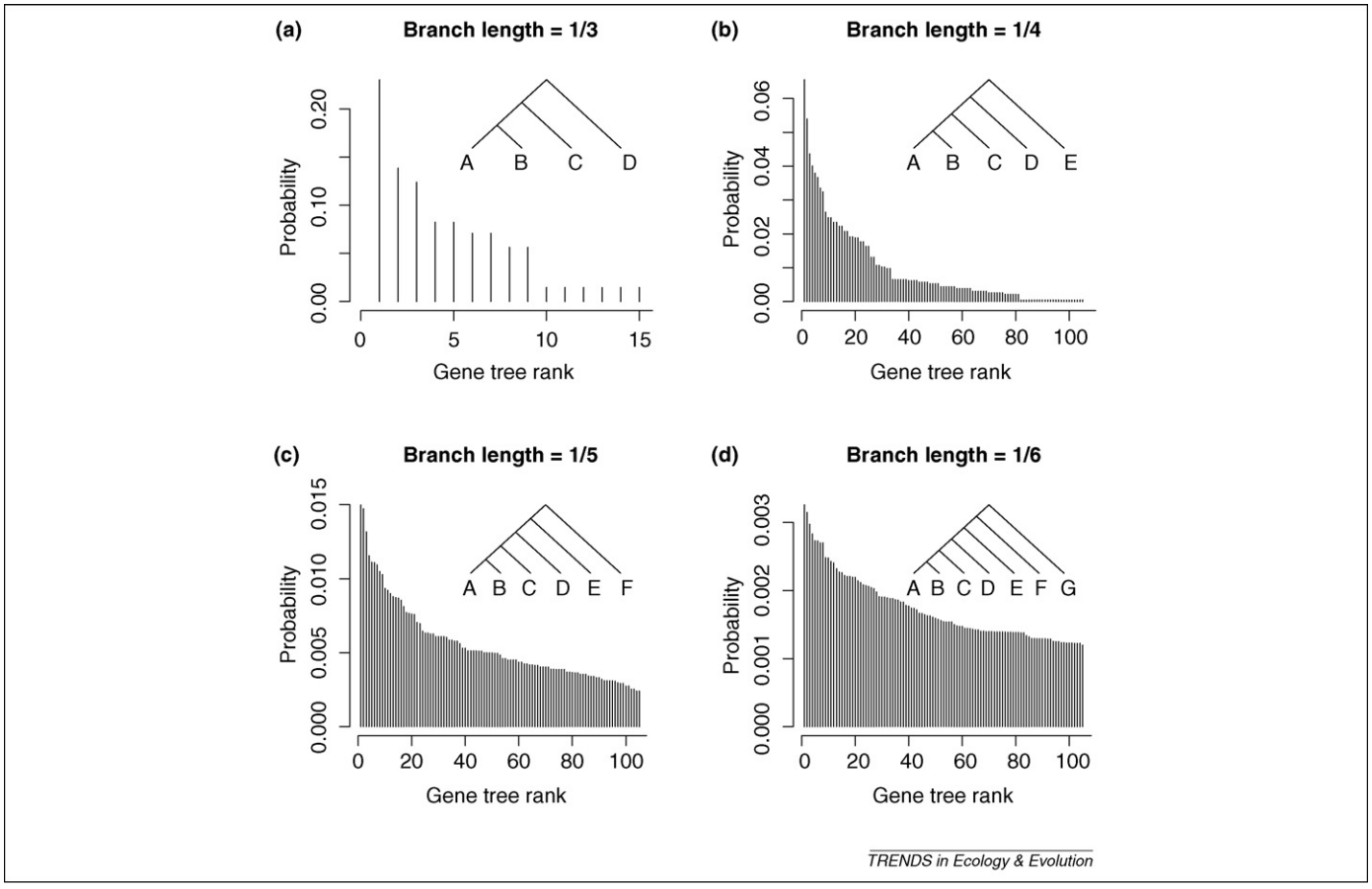

Figure 3. Gene tree distributions for pectinate species trees. The species tree is shown above each distribution. The total tree depth is fixed at 1.0 coalescent time units, including external branches, although only internal branch lengths are used to calculate gene tree probabilities when one lineage is sampled per species. **(a,b)** For four and five taxa, the most probable gene tree matches the species tree. **(c,d)** For six and seven taxa, the most probable gene tree is an AGT. For each plot, the gene tree topologies are ranked by their probabilities. Thus, in (a) and (b), the leftmost gene tree probabilities correspond to the (((AB)C)D) and ((((AB)C)D)E) topologies, respectively. In (c), the leftmost gene tree probability corresponds to ((((AB)C)D)(EF)). In (d), the most probable gene tree is (((((AB)C)D)E)(FG)), and the matching gene tree is the sixth most probable tree. For (c) and (d), only the 105 most probable gene trees are shown.

sampled taxa can be more distantly related than for shallower phylogenies, and extinction can lengthen branches deep in the tree, reducing the likelihood of incomplete lineage sorting.

In molecular data, gene tree discordance owing to incomplete lineage sorting is generally detected by analysis of segregating sites in aligned DNA sequences. However, we emphasize that the multispecies coalescent examines the underlying discordance of gene and species trees separately from mutation models used during data analysis that can also cause inferred gene trees to disagree with the species tree. Thus, even correctly inferred gene trees do not necessarily match the species tree. It is therefore useful to know the properties of underlying gene trees independently of difficulties inherent in inferring these trees from molecular data.

Gene tree probabilities

Probability calculations for properties of gene trees given a species tree are important for understanding the magnitude of genealogical discordance, for predicting the behavior of phylogenetic algorithms and for assessing the fit of the multispecies coalescent. Such computations rely on the concept of coalescent histories, which for a given gene tree and species tree topology represent the sequences of species tree branches on which gene tree coalescences can occur (online Supplementary Box S1). By considering all possible gene tree topologies for a given species tree with specified branch lengths, we can compute a full probability distribution of gene trees (online Supplementary Box S1). Each species tree topology with a set of branch lengths has a characteristic gene tree probability distribution; thus, the species tree with branch lengths can be considered a parameter for the gene tree distribution [25]. For pectinate species trees, Figure 3 shows these gene tree distributions for different numbers of taxa when the total tree depth is 1.0 coalescent time units. Holding tree depth constant, sampling more taxa increases the discordance, leading to lower gene tree probabilities and less peaked distributions.

The symmetries in gene tree distributions can facilitate the use of gene trees for testing the coalescent model and estimating species tree branch lengths (Box 3). For example, if the species tree has topology (((AB)C)D), then the probabilities of gene trees (((BC)A)D) and (((AC)B)D) are identical. A study of great apes [11] found that among 11 945 gene trees with high posterior probability, 76.6% supported the ((human,chimp),gorilla) relationship, whereas 11.5% and 11.4% supported the ((chimp,gorilla),human) and ((human,gorilla),chimp) relationships, respectively. These results are potentially compatible with the multispecies coalescent when there is a long separation between the split of orangutan (which has the role of species 'D') and the divergence of the other great apes, but a short interval between the separation of gorillas and the human–chimpanzee split.

One surprising property of gene tree distributions is that the most probable gene tree topology need not match the species tree topology. For example, in the six- and seven-taxon distributions in Figure 3, the most probable gene trees are ((((AB)C)D)(EF)) and (((((AB)C)D)E)(FG)), respectively, which have different topologies from the (pectinate) species trees. We have termed gene trees that are more probable than the gene tree that matches the species tree 'anomalous gene trees' (AGTs) and, for a given species tree topology, we call the region of branch length space that gives rise to AGTs the 'anomaly zone' [26]. An unexpected result is that for all species tree topologies with five or more taxa, and for pectinate topologies with four taxa, there exist choices of branch lengths for which AGTs occur.

The existence of AGTs implies that the most commonly observed gene tree in a genome-wide collection might not match the species tree. The problem of AGTs is not expected to diminish as the number of taxa increases. For example, when the internal branches have equal length, the maximum value of the shared branch length that still yields an AGT increases from 0.1568 coalescent time units (Box 2) for four taxa to 0.1934 coalescent time units for five taxa [48]; thus, with more taxa, branches can become longer while remaining in the anomaly zone.

Box 3. Testing the multispecies coalescent

The multispecies coalescent predicts certain distributions of gene tree frequencies. Only specific distributions are compatible with any particular species tree topology. For example, for three species, the most probable gene tree is expected to match the species tree, whereas the two non-matching topologies are expected to be equally frequent [4,20]. Processes such as natural selection, non-independence of loci, ancestral population subdivision [79,80] and hybridization can cause gene tree distributions to differ from the distribution expected under the multispecies coalescent. Although compatibility with the multispecies coalescent does not rule out the possibility that factors other than incomplete lineage sorting contribute to gene tree conflict, gene tree patterns can be used in a goodness-of-fit test for the multispecies coalescent.

A study of 30 loci in three in-group Australian grassfinch species found 16 gene trees with topology ((*acuticauda*,*hecki*),*cincta*), seven gene trees with topology ((*acuticauda*,*cincta*),*hecki*) and five gene trees with topology ((*cincta*,*hecki*),*acuticauda*) [14]. Are these data compatible with the multispecies coalescent? One way to test for such compatibility is to determine whether a species tree exists that could be consistent with these data. Because the ((a,h),c) gene tree is the most frequent and there are only three taxa, it has the highest likelihood of matching the species tree. Assuming that the species tree has topology ((a,h),c), the probability that a gene tree has the topology ((a,h),c) is $1 - (2/3)e^{-t}$, and gene tree topologies ((a,c),h) and ((c,h),a) both have probability $e^{-t}/3$ [4,20]. Using these probabilities, and ignoring two loci with unresolved estimated gene trees, the ML value for t is ~0.442 [20]. Using this value for t and the assumed species tree ((a,h),c), we can compute the expected number of times each topology would occur in a sample of 28 gene trees. These values are 16.002 for ((a,h),c) and 5.999 for ((a,c),h) and ((c,h),a). A chi-square test can be used to assess goodness of fit by comparing the observed and expected numbers of gene trees for each topology:

$$X^2 = \sum_i \frac{(Observed_i - Ex\ pected_i)^2}{Ex\ pected_i}$$

$$= \frac{(16 - 16.002)^2}{16.002} + \frac{(7 - 5.999)^2}{5.999} + \frac{(5 - 5.999)^2}{5.999} = 0.333.$$

The probability of observing X^2 this large or larger (the P value) is ~0.56, so the data are compatible with the multispecies coalescent. The test uses one degree of freedom, because only one free parameter (the species tree internal branch length) determines all gene tree probabilities. A species tree topology with n taxa has $n - 2$ parameters (internal branch lengths) that determine the gene tree distribution when one individual is sampled per species [25].

AGTs are more likely when at least some short branches occur in the species tree, such as in a rapid species radiation [44] or in a sample of closely related populations. Although it is currently unknown how often AGTs arise, it is sensible to use species tree inference procedures that perform well when they do occur; thus, scenarios in the anomaly zone can provide a useful set of parameter values for testing new methods for species tree inference.

Species tree inference

Discordant gene trees contain information about features of the species tree, such as its topology, divergence times and population sizes. Conflicting gene trees therefore provide a basis for inferring species trees using procedures that do not simply equate the estimated species tree with a single estimated gene tree. A desirable property for methods that estimate species trees is statistical consistency: an estimator should converge on the true species tree as more individuals, longer DNA sequences or more genes are added. An algorithm should further be computationally tractable and should produce reasonable estimates with data of feasible size. Existing methods exhibit these features in varying degrees.

Consensus and concatenation

Perhaps the most straightforward method of inferring species trees from multilocus data is the 'democratic vote' procedure, in which the most commonly occurring gene tree topology is used as the estimate of the species tree. Under the multispecies coalescent, this method is statistically consistent for three-taxon trees [9,10,49]. However, it can converge on an incorrect estimate when four or more taxa are present and an AGT exists, and it can be sensitive to sampling variation for small numbers of loci. Because the democratic vote procedure can produce misleading results, inferring species trees from multilocus data requires a more nuanced approach than simply increasing the number of loci. Two popular perspectives are the approaches of separate and combined analysis, represented by consensus methods [32,50,51] and concatenation of sequences [10,52,53]. Consensus and concatenation are attractive because they can reuse existing software. However, they do not explicitly model relationships between gene trees and species trees.

Consensus methods construct a tree that summarizes input trees defined on the same set of taxa (supertree methods are used if the input trees have overlapping but nonidentical sets of taxa [54]). Many consensus algorithms exist [50], some of which have favorable theoretical properties when applied to separate gene trees [55]. Rooted triple consensus [56] (approximately) constructs the tree that is most compatible with the most frequently occurring relationships for taxa taken in groups of three. Although the most frequently occurring gene tree considered on all taxa can be misleading, rooted triple consensus is motivated by the fact that the most frequently occurring three-taxon trees over all loci are expected to match the relationships in the species tree for the same taxa (there are no three-taxon AGTs) [55].

The concatenation approach, in which all sampled genes are concatenated for each taxon and are then analyzed as a single 'supergene,' assumes that all the data have evolved according to a single evolutionary tree, possibly under different mutation rates and models for different sites. When recombination occurs in a genome, decoupling the evolutionary histories of different loci, this assumption is violated. As a result, concatenation ignores the occurrence of different evolutionary histories at different loci, potentially leading to overconfident support for incorrect species trees [57–60]. Although consensus methods do not have this same limitation in theory, a simulation-based comparison [51] found concatenation to be more accurate than a consensus method, but sometimes with misleadingly high bootstrap support. Such limitations have motivated the need for new species tree inference approaches in the presence of gene tree discordance.

New approaches

One new method of inferring species trees involves minimizing the number of deep coalescent events [7,28]. In this approach, coalescence between two lineages is called 'deep' if it occurs more anciently than the most recent ancestral population from which the lineages were sampled. The inferred species tree is the one that minimizes the number of deep coalescences needed for the species tree to be compatible with each gene tree. This approach can also handle the sampling of multiple individuals per species, a strategy that, for closely related species and fixed effort, can be more informative than sampling more genes [28].

A second method is maximum likelihood (ML), in which a species tree likelihood is obtained by conditioning on the gene trees at each locus and summing over all possible sets of gene trees [6,7]. The ML species tree can then be obtained by searching over species trees, computing the likelihood by summing over all possible gene genealogies (gene tree topologies with coalescent times) for each species tree. However, this method is computationally intensive and has only been partially developed [61], although a pruning algorithm for species tree likelihoods that accounts for gene tree variation provides a substantial computational improvement [62]. Approximations to this type of approach have also been implemented using probabilities of gene tree topologies [15,63].

ML and Bayesian methods can incorporate branch lengths and uncertainty in estimated gene genealogies. A Bayesian approach using a density for gene genealogies [30], coded in the program BEST [31,64], simultaneously estimates the species tree along with gene trees and performs well in cases where concatenation performs poorly [58]. 'Bayesian concordance factors' [65] estimate the degree of conflict in a set of gene trees without assuming that a particular mechanism, such as the coalescent, explains the discordance. These two Bayesian methods take into account statistical dependency between genes.

One species tree inference method proven to be statistically consistent is the 'GLASS tree' approach [66] (also called the 'maximum tree' [64]). This method updates a single-locus method [23], which uses the minimum coalescent times taken over all pairs of individuals between two species, extending this strategy by also taking the minimum over multiple loci. The species tree topology is then implied by the minimum divergence times. A limitation of

this method is that its estimated divergence times are biased to be more ancient than actual divergence times, although the estimates asymptotically approach the true values. In practice, two difficulties with the method are: (i) for closely related species, lack of sequence divergence between two individuals leads to estimated coalescent times of 0 generations and, therefore, to unresolved trees; and (ii) different loci can have different mutation rates or can be non-clocklike, requiring coalescent times to be rescaled so that they can be combined to estimate a single tree.

Although diverse strategies for species tree inference are now becoming available, the relative performance of these methods given a high degree of gene tree discordance has yet to be investigated in detail, including in cases for which simpler methods, such as consensus and concatenation, perform poorly. In addition, issues such as robustness to violations of assumptions and taxon sampling in the species tree context have yet to be investigated.

Taxon sampling for species trees

Phylogenetic researchers have long been aware that the choice of taxa analyzed can impact the accuracy of tree estimates. Methods such as parsimony can be misled by 'long branch attraction,' in which species at tips of long branches are erroneously estimated as closely related [67]. Sampling more taxa can break long branches and can often produce improved phylogenetic inferences [68,69], although the opposite is sometimes true [70,71]. Additional taxa can introduce new long branches [70], and it was observed that when there was no gene tree conflict among 106 gene trees inferred from five taxa in a study of yeast [52], adding a distant outgroup caused conflict among the five taxa [71].

Issues of taxon sampling, concerning the choice of taxa for inclusion in phylogenetic studies, have been considered primarily for gene trees. Species trees, however, introduce new complications. Taxon sampling affects both gene tree branch lengths and species tree branch lengths. For a fixed total species tree depth, sampling taxa more densely shrinks some branches (Figure 3), making gene tree discordance more likely. Furthermore, because different gene trees can occur at different loci, the effect of taxon sampling can be locus dependent; thus, taxon sampling might break long branches for some loci but not for others. As past work on taxon sampling has focused on inferring gene trees, the effects of taxon sampling on various methods of species tree inference remain unexplored.

Conclusions

Conflicts between gene trees estimated at different loci have sometimes been seen as obstacles for inferring phylogenies. However, we suggest that gene tree conflict provides an opportunity to obtain information regarding the processes that have shaped organismal genomes. Researchers have used conflicting gene genealogies to infer ancestral population parameters such as population size and divergence times [30,72], and to examine species divergence processes [11,36]. It is only recently, however, that population-genetic and phylogenetic perspectives are being integrated in the effort to improve methods for inferring species trees.

With the increasing abundance of genomic data, it is important that phylogenetic methods take into account many loci and, therefore, many gene trees. Conflicting topologies are likely to become the norm, and the amount of gene tree discordance expected by chance under a simple neutral model can now be predicted analytically or by simulation. New ways of understanding gene trees will assist in modeling multiple sources of gene tree conflict simultaneously [37,38], or in distinguishing sources of conflict, such as in deciding whether discordance is due to hybridization or incomplete lineage sorting [73,74], and in judging whether discordance is more frequent than expected under a null model.

Long-standing issues about inferring species trees can now be reexamined in a new light, including problems with combining data sources, effects of taxon sampling and statistical consistency of phylogenetic estimators. Opportunities also exist for modeling, such as in relaxing the assumptions of the multispecies coalescent. The outstanding questions detailed in Box 4 could provide a useful framework for future research on gene tree discordance in phylogenetics.

In many cases, the answers to the questions posed in Box 4 will depend on the species under consideration.

Box 4. Outstanding questions

(i) Which species tree estimators from multilocus data are statistically consistent, even when there are AGTs? Among consistent algorithms, which offer the fastest convergence to the species tree?

(ii) Do computationally tractable ML algorithms exist that consistently infer the species tree while accounting for variation among gene trees?

(iii) What are the effects of taxon sampling for methods of inferring species trees? Do improvements in gene tree estimation owing to increased taxon sampling lead to improvements in species tree estimation?

(iv) What is the computational complexity of the evaluation of gene tree probabilities? For a given number of taxa, which gene tree–species tree combination maximizes the number of coalescent histories, and what is this maximum? If the gene tree matches the species tree, which topologies minimize and maximize the number of coalescent histories?

(v) Is there a way of computing gene tree probabilities that does not depend linearly on the number of coalescent histories?

(vi) For data sets with high levels of gene tree conflict, how can researchers determine whether an AGT is likely? How often do AGTs arise in real data sets?

(vii) How sensitive are predictions under the multispecies coalescent to violations of assumptions? What outcomes are expected in cases with ancestral population structure or high levels of intragenic recombination?

(viii) How much discordance in real data sets can be attributed to incomplete lineage sorting, hybridization, gene duplication, HGT, natural selection, recombination and sampling error? What are the best ways of distinguishing sources of discordance?

(ix) How does heterogeneity in evolutionary processes interact with gene tree discordance in phylogenetic inference? To what extent do difficulties such as heterogeneity in sequence evolution compound the problems of gene tree discordance?

(x) How should tradeoffs among sampling longer sequences, more genes and more individuals per species affect the design of multilocus phylogenetic studies?

However, as the focus of molecular phylogenetics moves from gene tree inference to multilocus inference of species trees, it will be important to determine the features of underlying biological processes, experimental designs and computational methods that give rise to the best estimates of species phylogenies.

Acknowledgements

We thank M. DeGiorgio, S. Edwards, M. Slatkin and two anonymous reviewers for comments. This work was supported by grants from the National Science Foundation (DEB-0716904), the Burroughs Wellcome Foundation and the Alfred P. Sloan Foundation.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tree.2009.01.009.

References

1 Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460
2 Hudson, R.R. (1983) Testing the constant-rate neutral allele model with protein sequence data. *Evolution Int. J. Org. Evolution* 37, 203–217
3 Neigel, J.E. and Avise, J.C. (1986) Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In *Evolutionary Processes and Theory* (Karlin, S. and Nevo, E., eds), pp. 515–534, Academic Press
4 Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press
5 Pamilo, P. and Nei, M. (1988) Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583
6 Felsenstein, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22, 521–565
7 Maddison, W.P. (1997) Gene trees in species trees. *Syst. Biol.* 46, 523–536
8 Nichols, R. (2001) Gene trees and species trees are not the same. *Trends Ecol. Evol.* 16, 358–364
9 Satta, Y. *et al.* (2000) DNA archives and our nearest relative: the trichotomy problem revisited. *Mol. Phylogenet. Evol.* 14, 259–275
10 Chen, F-C. and Li, W-H. (2001) Genomic divergences between human and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68, 444–456
11 Ebersberger, I. *et al.* (2007) Mapping human genetic ancestry. *Mol. Biol. Evol.* 24, 2266–2276
12 Syring, J. *et al.* (2007) Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Syst. Biol.* 56, 163–181
13 Takahashi, K. *et al.* (2001) Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. *Mol. Biol. Evol.* 18, 2057–2066
14 Jennings, W.B. and Edwards, S.V. (2005) Speciational history of Australian grassfinches (*Poephila*) inferred from thirty gene trees. *Evolution Int. J. Org. Evolution* 59, 2033–2047
15 Carstens, B.C. and Knowles, L.L. (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Syst. Biol.* 56, 400–411
16 Pollard, D.A. *et al.* (2006) Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet.* 2, e173
17 Kingman, J.F.C. (1982) On the genealogy of large populations. *J. Appl. Probab.* 19A, 27–43
18 Nordborg, M. (2001) Coalescent theory. In *Handbook of Statistical Genetics* (Balding, D.J. *et al.*, eds), pp. 179–212, Wiley
19 Hein, J. *et al.* (2005) *Gene Genealogies, Variation and Evolution.* Oxford University Press
20 Wakeley, J. (2009) *Coalescent Theory.* Roberts
21 Avise, J.C. (2000) *Phylogeography.* Harvard University Press
22 Funk, D.J. and Omland, K.E. (2003) Species-level paraphyly and polyphyly: frequency, causes and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* 34, 397–423
23 Takahata, N. (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122, 957–966
24 Rosenberg, N.A. (2002) The probability of topological concordance of gene trees and species trees. *Theor. Popul. Biol.* 61, 225–247
25 Degnan, J.H. and Salter, L.A. (2005) Gene tree distributions under the coalescent process. *Evolution Int. J. Org. Evolution* 59, 24–37
26 Degnan, J.H. and Rosenberg, N.A. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2, 762–768
27 Slatkin, M. and Pollack, J.L. (2006) The concordance of gene trees and species trees at two linked loci. *Genetics* 172, 1979–1984
28 Maddison, W.P. and Knowles, L.L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55, 21–30
29 Rosenberg, N.A. (2003) The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution Int. J. Org. Evolution* 57, 1465–1477
30 Rannala, B. and Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164, 1645–1656
31 Liu, L. *et al.* (2008) Estimating species trees using multiple-allele DNA sequence data. *Evolution Int. J. Org. Evolution* 62, 2080–2091
32 Felsenstein, J. (2004) *Inferring Phylogenies.* Sinauer
33 Ewens, W.J. (2004) *Mathematical Population Genetics.* (2nd edn), Springer
34 Wakeley, J. (2000) The effects of subdivision on the genetic divergence of populations and species. *Evolution Int. J. Org. Evolution* 54, 1092–1101
35 Hey, J. and Machado, C.A. (2003) The study of structured populations – new hope for a difficult and divided science. *Nat. Rev. Genet.* 4, 535–543
36 Innan, H. and Watanabe, H. (2006) The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Mol. Biol. Evol.* 23, 1040–1047
37 Meng, C. and Kubatko, L.S. (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Pop. Biol.* 75, 35–45
38 Than, C. *et al.* (2006) Identifiability issues in phylogeny-based detection of horizontal gene transfer. In *RECOMB-CG 2006, LNBI 4205* (Bourque, G. and El-Mabrouk, N., eds), pp. 215–229, Springer
39 Hobolth, A. *et al.* (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* 3, e7
40 Wiuf, C. *et al.* (2004) The probability and chromosomal extent of *trans*-specific polymorphism. *Genetics* 168, 2363–2372
41 Hudson, R.R. and Coyne, J.A. (2002) Mathematical consequences of the genealogical species concept. *Evolution Int. J. Org. Evolution* 56, 1557–1565
42 Hudson, R.R. and Turelli, M. (2003) Stochasticity overrules the 'three-times' rule: genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution Int. J. Org. Evolution* 57, 182–190
43 Edwards, S.V. *et al.* (2005) Phylogenetics of modern birds in the era of genomics. *Proc. R. Soc. Lond. B Biol. Sci.* 272, 979–992
44 Whitfield, J.B. and Lockhart, P.J. (2007) Deciphering ancient rapid radiations. *Trends Ecol. Evol.* 22, 258–265
45 Poe, S. and Chubb, A.L. (2004) Birds in a bush: five genes indicate explosive evolution of avian orders. *Evolution Int. J. Org. Evolution* 58, 404–415
46 Lessa, E.P. and Cook, J.A. (1998) The molecular phylogenetics of tuco-tucos (genus *Ctenomys*, Rodentia: Octodontidae) suggests an early burst of speciation. *Mol. Phylogenet. Evol.* 9, 88–99
47 Murphy, W.J. *et al.* (2007) Using genomic data to unravel the root of the placental mammal phylogeny. *Genome Res.* 17, 413–421
48 Rosenberg, N.A. and Tao, R. (2008) Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst. Biol.* 57, 131–140
49 Ruvolo, M. (1997) Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Mol. Biol. Evol.* 14, 248–265
50 Bryant, D. (2003) A classification of consensus methods for phylogenetics. In *BioConsensus* (Janowitz, M. *et al.*, eds), pp. 163–183, American Mathematical Society
51 Gadagkar, S.R. *et al.* (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zool.* 304B, 64–74

52 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804
53 de Quieroz, A. and Gatesy, J. (2007) The supermatrix approach to systematics. *Trends Ecol. Evol.* 22, 34–41
54 Bininda-Emonds, O.R.P. (2004) The evolution of supertrees. *Trends Ecol. Evol.* 19, 315–322
55 Degnan, J.H. *et al.* Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol*. (in press)
56 Ewing, G.B. *et al.* (2008) Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.* 8, 118
57 Kubatko, L.S. and Degnan, J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24
58 Edwards, S.V. *et al.* (2007) High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 5936–5941
59 Mossel, E. and Vigoda, E. (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* 309, 2207–2209
60 Kolaczkowski, B. and Thornton, J. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431, 980–984
61 Nielsen, R. (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor. Popul. Biol.* 53, 143–151
62 RoyChoudhury, A. *et al.* (2008) A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* 180, 1095–1105
63 Carling, M.D. and Brumfield, R.T. (2008) Integrating phylogenetic and population genetic analyses of multiple loci to test species divergence hypotheses in Passerina buntings. *Genetics* 178, 363–377
64 Liu, L. and Pearl, D.K. (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56, 504–514
65 Ané, C. *et al.* (2007) Bayesian estimation of concordance factors. *Mol. Biol. Evol.* 24, 412–426
66 Mossel, E. and Roch, S. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE Comp. Biol. Bioinform.* (in press)
67 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410
68 Hendy, M.D. and Penny, D. (1989) A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38, 297–309
69 Hedtke, S.M. *et al.* (2006) Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* 55, 522–529
70 Poe, S. and Swofford, D.L. (1999) Taxon sampling revisited. *Nature* 398, 299–300
71 Gatesy, J. *et al.* (2007) How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Syst. Biol.* 56, 355–363
72 Wall, J.D. (2003) Estimating ancestral population sizes and divergence times. *Genetics* 163, 395–404
73 Buckley, T.R. *et al.* (2006) Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada* Dugdale). *Syst. Biol.* 55, 411–425
74 Holland, B.R. *et al.* (2008) Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol. Biol.* 8, 202
75 Masta, S.E. and Maddison, W.P. (2002) Sexual selection driving diversification in jumping spiders. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4442–4447
76 Avise, J.C. and Robinson, T.J. (2008) Hemisplasy: a new term in the lexicon of phylogenetics. *Syst. Biol.* 57, 503–507
77 Rosenberg, N.A. (2007) Counting coalescent histories. *J. Comput. Biol.* 14, 360–377
78 Than, C. *et al.* (2007) Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J. Comput. Biol.* 14, 517–535
79 Wakeley, J. (2003) Inferences about the structure and history of populations: coalescents and intraspecific phylogeography. In *The Evolution of Population Biology* (Singh, R. and Uyenoyama, M., eds), pp. 193–215, Cambridge University Press
80 Slatkin, M. and Pollack, J.L. (2008) Subdivision in an ancestral species creates asymmetry in gene trees. *Mol. Biol. Evol.* 25, 2241–2246

Forthcoming Conferences

Are you organizing a conference, workshop or meeting that would be of interest to *TREE* readers? If so, please e-mail the details to us at TREE@elsevier.com and we will feature it in our Forthcoming Conference filler.

1–5 September 2009
2nd European Congress of Conservation Biology: Conservation biology and beyond: from science to practice
Prague, Czech Republic
http://www.eccb2009.org

8–12 September 2009
7th Cold Spring Harbor meeting on Microbial Pathogenesis and Host Response
Cold Spring Harbor, NY, USA
http://meetings.cshl.edu/meetings/host09.shtml

8–10 September 2009
BES Annual Meeting 2009
Hatfield, UK
http://www.britishecologicalsociety.org/meetings/current_future_meetings/2009_annual_meeting/index.php

23–26 September 2009
69th Society of Vertebrate Paleontology Annual Meeting
Bristol, UK
http://www.vertpaleo.org/meetings/index.cfm

27–30 October 2009
9th Cold Spring Harbor Laboratory/Wellcome Trust conference on Genome Informatics
Cold Spring Harbor, NY, USA
http://meetings.cshl.edu/meetings/info09.shtml

8–12 February 2010
Island Invasives: Eradication and Management
Auckland, New Zealand
http://www.cbb.org.nz/conferences.asp