Is a New and General Theory of Molecular Systematics Emerging?

Submitted as a Commentary to *Evolution*

Is a new and general theory of molecular systematics emerging?

Scott V. Edwards

Museum of Comparative Zoology and Department of Organismic & Evolutionary Biology,

Harvard University, 26 Oxford Street, Cambridge, MA 02138 USA

RUNNING HEAD: THE CASE FOR SPECIES TREES

Key words: polytomy, fossil, macroevolution, genome, phylogeography, Neanderthal

Four figures; one table

Contact for corresponding author:

Scott Edwards
Museum of Comparative Zoology
Harvard University
26 Oxford Street
Cambridge, MA 02138 USA
sedwards@fas.harvard.edu
Ph: 617-384-8082
FAX: 617-495-5667

ABSTRACT

The advent and maturation of algorithms for estimating species trees - phylogenetic trees whose OTUs are lineages, populations and species, as opposed to genes - represents an exciting confluence of phylogenetics, phylogeography, and population genetics, and ushers in a new generation of concepts and challenges for the molecular systematist. In this essay I argue that to better deal with the large multilocus data sets brought on by phylogenomics, and to better align the fields of phylogeography and phylogenetics, we should embrace the primacy of species trees, not only as a new and useful practical tool for systematics, but as a long standing conceptual goal of systematics that, largely due to the lack of appropriate computational tools, has been eclipsed in the past few decades. I suggest that phylogenies as gene trees are a 'local optimum' for systematics, and review recent advances that will bring us to the broader optimum inherent in species trees. In addition to adopting new methods of phylogenetic analysis (and ideally reserving the term ‘phylogeny’ for species trees rather than gene trees), the new paradigm suggests shifts in a number of practices, such as sampling data to maximize not only the number of accumulated sites but also the number of independently segregating genes; routinely using coalescent or other models in computer simulations to allow gene tree heterogeneity; and understanding better the role of concatenation in influencing confidence in phylogenies. By building on the foundation laid by concepts of gene trees and coalescent theory, and by taking cues from recent trends in multilocus phylogeography, molecular systematics stands to be enriched. Many of the challenges and lessons learned for estimating gene trees will carry over to the challenge of estimating species trees, although adopting the species tree paradigm will clarify many issues (such as the nature of polytomies and the star tree paradox), raise conceptually new challenges, or provide new answers to old questions.

Introduction

The title of this essay is borrowed from one of the famous essays written by Stephen Jay Gould, "Is a new and general theory of evolution emerging?", published in *Paleobiology* in 1980 (Gould 1980). Gould was speculating as to whether the constellation of observations and trends from the fossil record and developmental biology, collectively known as 'macroevolution', might constitute a genuinely new set of phenomena, a set that had not been covered adequately by the reigning paradigm of Darwinian microevolution. Of course whether one answers Gould's question in the positive or negative depends on one's perspective; although Gould and others would not have raised the question unless one could answer 'yes', many evolutionary biologists have argued that the quantitative framework provided by microevolution can adequately account for the observations of punctuation, stasis and apparent saltation that had suggested a new paradigm to some (Charlesworth et al. 1982; Smith 1983; Estes and Arnold 2007). Yet there is a pervasive feeling that the paradigms laid down by the Modern Synthesis still may not adequately capture the plethora of phenomena ushered in by modern evolutionary biology (Erwin 2000; Pigliucci 2007). Although the paradigm that I question is more limited in scope than Gould's, in a similar spirit I raise the question of whether molecular phylogenetics is experiencing an important conceptual shift, one that may affect the daily practice of phylogeny building as well as the relationship between systematics and other evolutionary disciplines. The developments I will review are indeed new in a practical sense, yet they mark a return to the goals and concepts that have been in the back of systematists minds for many decades (Felsenstein 1981; Takahata 1989; Neigel and Avise 1986; Avise 1994; Maddison 1997; Yang 1997). Put simply, the response to Joe Felsenstein's oft quoted complaint that "Systematists and evolutionary geneticists don't often talk to each other" (Felsenstein 1988: 445) is, I think, finally maturing and reaching

fruition, and thus it is an opportune time to reflect on this new interdisciplinary dialogue and to forecast what might lay ahead.

What is phylogeny, and how do we infer it from sequence data?

One of my favorite essays in systematics, with one of my favorite essay titles, is the paper by Rod Page and the late Joe Slowinski, innocently entitled "How do we infer species phylogenies from sequence data?" (Slowinski and Page 1999). In it they argued cogently for a distinction between gene and species trees and outlined ways to estimate the latter. As obvious as the answer to these questions may seem to some, they worth raising again, if only to reiterate the an answer so simple that we sometimes overlook it. Phylogeny is the history of *species and populations*. It records the branching pattern of evolving *lineages* through time. One of the grand missions of systematics is to reconstruct and provide details on the great Tree of Life. As difficult as it may be for modern methodologies to reconstruct this history, and as fraught with reticulations, hybridization events, horizontal gene transfer and other mechanisms that cloud the picture of organismal history, it is important to reiterate that, at the level of populations and species, there is only one such history, even when reticulate. With species and populations as the focus, there is no heterogeneity in this demographic history, because the history has happened only once.

In pursuit of the goal of reconstructing the history of life, the core approaches of phylogenetic systematics have evolved into a suite of methodologies that focus on amassing character data to build these trees (Nei and Kumar 2000; Felsenstein 2003; Delsuc et al. 2006), and the leading role of DNA sequences in providing these characters over the past several decades has helped to invigorate systematics and to provide many fruitful intersections with other biological disciplines, such as genomics and molecular ecology. Yet the use of DNA

sequences has also led to challenges in the translation of histories of DNA sequence diversity - phylogenetic trees of genes and alleles - into the currency that is surely still the major focus of systematics - phylogenetic trees of species and populations, or species trees. Ultimately these challenges arise because genes and species are different entities, assuming different levels in the biological hierarchy (Avise and Wollenberg 1997; Doyle 1997; Maddison 1997; Avise 2000). The diversity recovered in our surveys of DNA sequence evolution within and between species are ultimately an indirect and incomplete window into the history of species, precisely because species are by most definitions evolving lineages that comprise many genes, each found in many individuals. The fact that species comprise a higher level of biological organization than do genes ensures that the program of systematics will be incomplete until phylogenetic methods make a clear distinction between gene trees and species trees and explicit reference to the phylogenetic relationships of species within which genes are embedded. The overwhelming dominance of molecular data in systematics and phylogenomics makes the development of methods for estimating species trees a key, if not the key, task for the years ahead.

Causes of gene tree heterogeneity and the ubiquity of coalescent effects

The causes of gene tree heterogeneity and of gene tree-species tree conflicts are by now well known to molecular systematists and nicely summarized, for example, in Maddison's 1997 review (Maddison 1997). The three primary causes – gene duplication, horizontal gene transfer and deep coalescence – have varying levels of importance depending on the taxa and genes under study. Horizontal gene transfer is a well-known and common cause of discordance in the microbial world – so much so that some microbial phylogeneticists have questioned whether a coherent Tree of Life exists for microbes (e.g., Bapteste et al. 2005; Doolittle and Bapteste 2007). Gene duplication is in some taxa also common and widespread, and can subvert

phylogenetic analysis if it is not recognized; alternatively gene duplication can provide a rich source of information for phylogenetic analysis (e.g., Page and Charleston 1997; Sanderson and McMahon 2007; Rasmussen and Kellis 2007).

Deep coalescence, the third major cause of gene tree heterogeneity and gene tree-species tree conflicts, is distinct in so far as its occurrence is, in principle, much more widespread, depending not on specific, molecular events that occur only in some lineages (and whose consequences can be avoided by appropriate gene sampling), but on the intrinsic properties of every population. The root cause of deep coalescence is the rate of genetic drift – deep coalescence will be more prevalent when the rate is low (due to large populations) compared to the length of internodes in the species tree. Thus deep coalescence is in principle detectable in any taxonomic group, and for any gene, whether in an organelle or the nucleus, provided that the branch lengths in the underlying species tree are sufficiently short as measured in coalescent units (Maddison 1997 ; Hudson and Turelli 2003). (Coalescent units are calculated as the ratio of the length of internodes in the species tree as measured in generations over the effective population size, as measured in individuals, of ancestral species during those internodes.) Thus deep coalescence knows no taxonomic or gene bias, as do the other two phenomena, and thus holds a special place in the triumvirate of causes of gene tree heterogeneity. Indeed, empirical examples of deep coalescence or incomplete lineage sorting are now routine and taxonomically ubiquitous (for recent examples, see Satta et al. 2000; Jennings and Edwards 2005; Patterson et al. 2006; Pollard et al. 2006; Hobolth et al. 2007; Wong et al. 2007).

Yet there is a fourth and even more widespread cause of gene tree heterogeneity, if not gene tree/species tree 'conflict', one that influences branch lengths only but causes heterogeneity

nonetheless. I suggest *branch length heterogeneity* due to the coalescent process as a useful additional source of gene tree heterogeneity (Fig. 1). Branch length heterogeneity specifically

[Figure 1 about here]

highlights the heterogeneity of branch lengths of gene trees in situations when all gene trees are topologically identical, such as will occur when the underlying species tree has branch lengths that are long in coalescent units; by contrast, deep coalescence emphasizes the heterogeneity of gene tree *topologies*. Branch length heterogeneity is a useful concept for systematists because it highlights the fact that, even when gene trees are topologically identical – a situation in which most systematists would feel comfortable in combining data through concatenation and other traditional means – there can be significant and detectable heterogeneity in branch lengths, such that the gene trees are for practical purposes still heterogeneous. Such collections of trees that vary only in branch length have become drawn the attention of systematists because of the related issue of heterotachy (Kolaczkowski and Thornton 2004, 2008) and also because they can generate unexpected phylogenetic signals in DNA sequence data sets (Kolaczkowski and Thornton 2004; Matsen and Steel 2007). In fact, branch length heterogeneity and deep coalescence are ends of a continuum, and the latter is really an expression of the former in the limit as gene tree topologies begin to depart from the species tree. But branch length heterogeneity is indeed ubiquitous, due to the finiteness of all populations. It will occur to varying extents in all taxa, genes and contexts, even in situations in which deep coalescence is not occurring. Thus, branch length heterogeneity from gene to gene is probably the most common of all causes of gene tree heterogeneity. In addition, branch length heterogeneity could be a potent source of phylogenetic inconsistency in real data sets; like deep coalescence, it essentially introduces mixtures of gene trees into data sets, a situation that is known to mislead

phylogenetic analysis (Mossel and Vigoda 2005). Masten and Steel (2007) have recently shown that DNA sequences simulated from mixtures of topologically identical but branch-length variable trees can in some cases mimic signals from a topologically different tree very well; thus the problem of branch length heterogeneity is in principle a serious one for empiricists, although whether this is the case empirically is not known.

The extent of variation in branch lengths due to branch length heterogeneity will be a function of the effective population size of ancestral populations, scaling with its square (θ is a measure of effective population size as measured in DNA substitutions; $\theta = 4N\mu$, where N is the effective population size and μ the mutation rate, and can be easily calculated, at least for extant populations, from DNA sequence data). We expect that branch lengths will vary from gene to gene sometimes by hundreds of thousands of years, if not millions of years (Lynch and Jarrell 1993; Edwards and Beerli 2000), if estimates of θ from extant natural populations are any guide. Sequence simulation packages such as SeqGen (Rambaut 2007) and other approaches implicitly assume that the underlying gene trees are identical in topology and branch lengths and, when multiple loci are simulated, the sequence data sets record only the mutational variance accumulated within the specified phylogenetic trees. (Rarely are multiple different trees used with DNA sequence simulators such as SeqGen to incorporate both mutational and coalescent variance). These simulated data sets will differ from the more realistic situation, embodied in packages such as MCMCcoal (Yang 2002), the coalescent module in Mesquite (Maddison 1997; Maddison and Maddison 2008) and serialsimcoal (Laval and Excoffier 2004; Anderson et al. 2005), in which, at minimum, branch lengths differ subtly among loci due to branch length heterogeneity. Differences between DNA data sets with and without coalescent variance will

vary depending on the effective population sizes in the species tree used in the simulations and the extent of gene tree heterogeneity, although the precise pattern of differences

[Figure 2 about here]

is not known (Fig. 2). We might expect that DNA sequence data sets produced under models with only branch length heterogeneity, such as in computer simulations, will deviate from simulations on single trees less so than will data sets produced under deep coalescence; Fig. 2 suggests that deep coalescence and single gene tree simulations can indeed produce very different distributions of site patterns in DNA sequences. Models traditionally used to simulate DNA sequence data for phylogenetic purposes essentially assume that the population sizes in the species tree are zero, and thus ignore the contribution of coalescent variance to molding the variation and signal present in DNA sequence data sets (Carstens et al. 2005).

Origin and consequences of the concatenation paradigm

The current paradigm under which molecular phylogenetics operates - one characterized by the accumulation of many genes which are then concatenated into large supermatrices before analysis - arose in part from a need to amass larger data sets, and in part from debates in the early 1990s spurred by Arnold Kluge's call for 'total evidence' - a philosophical mandate to include all available information into phylogenetic analyses (Kluge and Ag 1989; Kluge 2004). Concatenation - the practice of combining different genes or data partitions in to a single supermatrix and analyzing this matrix such that all genes conform to the same topology - provided a convenient means of implementing Kluge's call. It soon became clear, however, that although total evidence might have substantial philosophical justification, the practice could clash with some of the practical nuances of molecular systematics and with the growing appreciation of heterogeneity in gene trees, which grew separately from observations of the

behavior of gene trees in natural populations (Wilson et al. 1985; Avise et al. 1987; Doyle 1992; Avise 1994). There were generally two practical arguments against total evidence. The first was the demonstration that in computer simulations, DNA sequences evolving under substantially different substitution rates and patterns could give erroneous results when analyzed with currently available software and models of phylogenetic reconstruction (Bull et al. 1993). This first concern has largely been addressed in the past decade with the development of efficient likelihood and Bayesian algorithms permitting different data partitions to evolve under different models (e.g., Nylander et al. 2004). Although it is widely appreciated that the most commonly employed models of DNA substitution do not adequately describe the complexities of DNA sequence evolution (not only because of their simplicity but also because of their frequent reliance on the assumption of stationarity through time), applying different models to different genes or data partitions is well known to dramatically improve phylogenetic inference. Variation in substitution patterns was sometimes considered a benefit to phylogenetic analysis, provided that it was not too great. For example, combining many genes encompassing both fast and slow rates of evolution was suggested as a better means of improving phylogenetic analysis as compared with using genes having similar rates (Cummings et al. 1995; Otto et al. 1996).

The second argument against total evidence was the suggestion that different genes should not be combined if they can be shown to have different topological histories. By the early 1990s, gene tree heterogeneity had been observed frequently in real data sets. Yet in roughly the same time it has taken systematics to embrace sophisticated mixture models of the substitution variation across partitions, the challenge of heterogeneity in gene trees has not received commensurate attention. For example, Felsenstein's recent survey of phylogenetic methods (Felsenstein 2003) contains only a single chapter on species trees and the potential variability of

their constituent gene trees (chapter 28), and only recently have a few phylogenetic methods incorporating gene tree heterogeneity, with the ability to analyze large data sets, been available. This relative inattention to dealing with gene tree heterogeneity - even in the knowledge that such heterogeneity does not necessarily conflict with the unique species history in question – was, I think, partly due to the perceived success of the concatenation approach in delivering high confidence in phylogenetic trees, and the suggestion that more genes could improve this resolution. However, as I describe next, it was not so much the multiplicity of genes that was deemed responsible for the success of combining information via concatenation, but rather the multiplicity of characters or sites.

By now it is routine for phylogenetics and phylogenomics projects to amass multiple genes, sometimes hundreds of them, in pursuit of phylogenetic rigor. Yet the current justification for collecting multiple genes is, I suggest, somewhat out of sync with their real service in phylogenetics. When asked why collecting multiple genes is useful in phylogenetic analysis, many systematists might answer "In order to capture a diversity of mutation rates, so as to resolve deep and shallow branches in the tree." (This answer is partly a legacy of the influential paper by Cummings et al. (1995), which specifically prescribed sampling many, short (mitochondrial) genes with varying mutation rates, rather than a few longer genes.) In addition, our imaginary systematist would probably prefer to sample widely throughout the genome, rather than from one chromosomal segment, even if one could assure her that a single segment contained as much mutation rate variation as genes spread across the genome. Yet, as attractive as this protocol sounds, the demand for genomically widespread markers – tantamount to demanding a measure of genealogical independence among markers due to recombination between them – does not reconcile easily with the concatenation or supermatrix approaches that

have become the norm in phylogenomics, because these approaches do not allow for genealogical independence of different genes. I therefore argue that the motivation for sampling many markers in modern phylogenomics is not due to an explicit desire to sample many (phylo)genetically independent markers, but rather to sample many *sites*, perhaps with varying rates; and that the goal of sampling many genes is favored only in so far as it might bring some measure of rate heterogeneity among loci that might resolve both deep and shallow nodes in phylogenetic trees. Missing from this justification for sampling many genes is any reference to the possibility that the sample of gene *trees* will increase with the sample of genes, and will thereby better portray the statistical tendencies of genomes and populations that comprise the biological levels above the sampled entities. Few systematists today would say they prefer sampling many genes "So as to obtain a diversity of gene trees." It is this answer, however, that underlies the sampling properties of the species tree approach.

Gene tree phylogenetics: a local optimum

The molecular biology revolution drastically changed phylogenetics in key ways. In addition to the obvious advances allowing collection of vastly more characters for phylogenetic analysis, the revolution in restriction enzymes, and eventually rapid DNA sequencing via PCR, allowed researchers to collect molecular data that could be directly analyzed phylogenetically (Avise 1994). For example, allozymes proved extremely useful in advancing phylogenetics and biogeography, yet the molecular data themselves – alleles and allele frequencies – could not easily be analyzed phylogenetically without first transforming the data in some way, such as by estimating a genetic distance. By contrast, data from restriction enzymes, or DNA or protein sequences, are easily and almost effortlessly amenable to phylogenetic analysis, because they

come to the researcher already in the form of a character matrix (Hillis et al. 1993; Swofford et al. 1996).

Producing phylogenies directly from gene sequences essentially in one step, without additional transformations, is now the dominant mode of phylogenetic analysis and indeed it has advanced the field enormously. Nonetheless, I suggest that the very success of this paradigm and the ease with which phylogenies could be produced directly from DNA matrices led to a comfort zone in phylogenetics. If we can imagine systematic methods themselves as a likelihood surface, I suggest that the current paradigm is a local optimum in that surface, an optimum that is useful but ultimately incomplete in so far as it has failed to model the potential for gene tree/species tree discordance even cursorily (Fig. 3).

[Figure 3 about here]

Recent phylogenomic analyses have begun to enshrine the concatenation paradigm by amassing hundreds of genes to unravel the Tree of Life (e.g., Rokas et al. 2003; Delsuc et al. 2006; Dunn et al. 2008; reviewed in Delsuc et al. 2005). At the same time, other recent results are beginning to question the suitability of concatenation for all data types and time scales, in particular those from DNA sequences sampled from rapidly diverging clades (Edwards et al. 2007; Kubatko and Degnan 2007). For example, Degnan and Rosenberg (2006) showed that for any species tree of 5 or more taxa, there exist branch lengths in species trees for which gene trees that do not match the species tree are more common than gene trees matching the species tree - so called anomalous gene trees. In such situations, or even slightly outside of this zone, phylogenetic analysis of concatenated sequences can positively mislead inference of species relationships (Kubatko and Degnan 2007); by contrast, some of the new species tree approaches appear promising in or near the anomaly zone (e.g., Edwards et al. 2007; Liu and Edwards 2008;

Liu et al. 2008c). Although the parameter space of species trees that produce anomalous gene tree topologies is probably not large (we do know yet of any empirical examples of this phenomenon), it stands to reason that concatenation will under many circumstances be a worse approximation of the underlying diversity of gene trees than will approaches that allow for gene tree heterogeneity, because we know, as stated above, that gene trees will always differ from one another subtly, even when topologically congruent. What few statistical comparisons that have been done suggest that species tree approaches that allow for gene tree heterogeneity are significantly better explanations of multilocus sequence data than is concatenation, even in situations in which gene tree heterogeneity is moderate (Liu and Pearl 2007; Edwards et al. 2007; Belfiore et al. 2008). The cost, however, of the species tree approach can sometimes be substantially increasing the number of parameters to be estimated. For example, in addition to the usual nucleotide substitution parameters for each gene (partition), species tree analysis can involve parameters for the relative mutation rates of different genes, branch length and tree length parameters for each gene, as well as branch lengths, effective population sizes and topologies of the species tree.

Concatenation, phylogenetic confidence and polytomies

Concatenation has many implications beyond whether recovered tree topologies are correct or incorrect. As stated before, in all likelihood the topologies generated by concatenation are reasonable approximations of reality, and in many cases it is not concatenation per se that might derail a phylogenetic analysis but some other detail, such as specification of the substitution model, inhomogeneous base compositions, vagaries of the molecular clock, etc. Yet a serious and still unanswered question is whether concatenation itself can strongly influence confidence values, if not the topologies, of phylogenetic trees. Many papers have been devoted

recently to understanding the type I and II error rates of phylogenetic inference methods, most recently with Bayesian inference, and several researchers have suggested that confidence values on branches can be strongly overestimated under a variety of circumstances frequently encountered in routine data analysis. Some researchers, particularly those working with simulations, have often attributed this overconfidence to the inference method as encoded in various software programs (Suzuki et al. 2002; Misawa and Nei 2003; Simmons et al. 2004), or to misspecifications of the model of evolution (Huelsenbeck et al. 2002; Yang and Rannala 2005; reviewed in Alfaro and Holder 2006). In such cases, DNA sequences are indeed simulated on trees that lack coalescent variance, and so such a conclusion may be reasonable. Yet the source of the often high posterior probability values seen in empirical trees has a less obvious explanation. Misspecifications of the substitution model may often be to blame, but concatenation itself – a type of model misspecification, given the coalescent process - represents a major unexplored source of such overconfidence. An example of this is illustrated by the extreme case in which a polytomy in a species tree is used as a model to generate gene trees, DNA sequences, and to reconstruct the phylogeny from these simulation. Despite the polytomy in the species tree, we expect the gene trees generated by this species tree to be dichotomous except in extreme circumstances (Slowinski 2001). (As discussed below, I believe that species trees clarify many aspects of polytomies, and associated concepts such as the 'star tree paradox' (Lewis et al. 2005; Kolaczkowski and Thornton 2006), that have been confused in the literature due to a gene tree perspective.) Fig. 4 shows how we can expect three distinct dichotomous gene

[Figure 4 about here]

trees from a single polytomous species tree. In this situation, whereas species tree analysis gives a reasonable estimate of confidence in the species tree, providing fairly even support for all three

constituent trees underlying the species tree polytomy, concatenation unrealistically places high confidence on one or another gene tree (depending on the details of the replicate), to the exclusion of the remaining two trees. Thus, because something approximating a coalescent process generates DNA sequences in nature, yet we analyze them as if coalescence did not exist, it's worth exploring this source of misestimation further, and the brief example in figure 4 is by no means the last word. A separate but important issue is the fact that, until recently, most explorations of phylogenetic accuracy and overcredibility of phylogenetic methods have been performed on gene trees, not species trees, and it is unclear to what extent these conclusions will translate to the higher level embodied in species trees (e.g., Douady et al. 2003; Taylor and Piel 2004).

I suggest, as have others (Slowinksi 2001), that species trees are the more relevant entity when discussing polytomies (e.g., Braun and Kimball 2001), or related concepts such as the 'star tree paradox' (Lewis et al. 2005; Kolaczkowski and Thornton 2006). (The star tree paradox is the finding that posterior probabilities of trees can be grossly overestimated when the true tree is a polytomy but when polytomies are not visited frequently or at all during the MCMC run)., Nonetheless, polytomies in gene trees have remained the focus of discussion and theoretical attention (Walsh et al. 1999; see Slowinski 2001 for an excellent review; Lewis et al. 2005; Steel and Matsen 2007). Polytomies in species trees are of real relevance to systematics and biogeography, and likely exist in nature, whereas polytomies in gene trees are expected to be rare on biological grounds, and in any case are not a necessary consequence of polytomies in species trees (Slowinksi 2001; Fig. 4). For these reasons I suggest that studying the behavior of DNA sequences generated by polytomous gene trees will be less productive than studying the types of gene trees generated from polytomous species trees, and the sequences that arise from them.

A brief history of species trees

Species trees are, of course, synonymous with phylogeny and the Tree of Life. Methodologically, species trees can be defined as any phylogenetic approach that distinguishes gene trees or genetic variation from species trees, and explicitly estimates the latter. Species trees by this definition need not be derived from DNA sequence data, but they often involve a model of gene tree evolution – a model distinct from that for nucleotide substitution - that serves as a basis for evaluating the likelihood of the collected data under various candidate species trees. This model can explicitly capture biological processes, such as the coalescent process, or it can capture trends in gene tree heterogeneity without specifically modeling coalescence (e.g., Steel and Rodrigo 2008).

Species trees as distinct from gene trees are not a new idea. As early as the 1960s, Cavalli-Sforza (Cavalli-Sforza 1964), and in the late 1970s Joe Felsenstein (Felsenstein 1981), were applying simple drift models to tables of allele frequencies and using these models to evaluate competing hypotheses of population and species relatedness. In the 1980s John Avise brought species trees as distinct from gene trees to the forefront of the burgeoning field of phylogeography (Neigel and Avise 1986; Avise et al. 1987). Species trees are a realization of Doyle's characterization of gene trees as characters (Doyle 1992, 1997), or Maddison's (1997) 'cloudogram'. Nonetheless, as I suggest in this section, the concept of species trees appears newer than it is in part because the use of DNA sequence data mass-produced a closely related entity, the gene tree, that systematists must now distinguish from it. The concept also appears new from a practical standpoint, since, until now, there have been few means to directly incorporate gene stochasticity into the phylogenetic analysis of moderately sized data sets with workable software (Table 1). Statistical methods for dealing with gene tree heterogeneity and

coalescent stochasticity have already been in the mainstream of related fields, such as phylogeography and historical demography, for a number of years, as evidenced by a battery of software focused near the species level that deals with multilocus data. Examples of such software include MIGRATE, LAMARC, BEAST, IM and other methods that treat the gene tree as a statistical quantity with associated errors in estimation and as a means for estimating parameters at the population level (Wakeley and Hey 1997; Yang 1997; Drummond and Rambaut 2003; Beerli 2006; Kuhner 2006). By estimating population parameters above the level of the gene, these models make reference to the species history in which gene histories are embedded, and indeed go so far as to integrate out gene trees as nuisance parameters. Hey and Machado (2003) captured the distinctive properties of this new view of phylogeography, as well as the spirit of the debates that accompanied the transition in perspective.

In stark contrast to the situation in phylogeography, phylogenetic inference itself still largely retains its focus on gene trees – if not philosophically then operationally, at the UNIX prompt or GUI menu. The thought of integrating out the gene trees from a phylogenetic analysis would likely seem paradoxical to practitioners of the current paradigm, and for this reason again, species trees may appear to be a new concept. Table 1 summarizes a number of approaches to estimating species trees that have been developed over the years, many in the last five years. All of these approaches make explicit the distinction between the underlying genetic variation – whether manifested as allele frequencies or as gene trees - and the species tree that is the object of estimation. Table 1 does not necessarily include all methods for combining data from multilocus data sets – for example, consensus trees, majority rule trees, supertrees and supermatrices have been suggested as ways of combining data from multiple genes (de Queiroz 1993; de Queiroz et al. 1995; Wiens 1998; Steel et al. 2000; Gadagkar et al. 2005; Holland et al.

2005; Holland et al. 2006). Although I do include some recent evaluations of these approaches for estimating the species tree under a coalescent model (Degnan et al. 2008), I do not consider these methods true species tree methods because they do not specifically acknowledge an overarching species tree in which gene trees are embedded, any sort of correlation among gene trees, or a model connecting the two, other than simply calling the consensus tree or supertree the species tree (for an exception see Steel and Rodrigo 2008). A complete review of species tree methods is beyond the scope of this Commentary (see Degnan and Rosenberg 2008 and Brito and Edwards 2008 for an introduction), but the following overview may be helpful.

Methods for inferring species trees have adopted likelihood or parametric statistical or model-free approaches and have proved useful with varying degrees of success. For example, some of the most statistically robust methods are challenging to implement and are not generally available to empiricists (Nielsen 1998; Nielsen et al. 1998; chapter 28 of Felsenstein 2003). Other approaches, such as likelihood methods (Pamilo and Nei 1988; Wu 1991; Hudson 1992; Chen and Li 2001; Waddell et al. 2001, 2002) are generally not applicable to more than three species. Recent parsimony methods for inferring species trees, such as methods minimizing deep coalescence, appear promising, particularly given their implementation in powerful software packages such as Mesquite (Maddison and Knowles 2006). Likelihood approaches, such as direct evaluation and comparison of species trees via the likelihood of gene trees in the data (Seo et al. 2005; Carstens and Knowles 2007; Seo 2008), or constructing supertrees from gene trees via a summary likelihood function (Steel and Rodrigo 2008), also appear promising. Recently Liu and colleagues have proposed a promising Bayesian method (Liu and Pearl 2007; Liu et al. 2008b), as well as several parametric methods (Liu et al. 2008a), for estimating species trees, the latter of which is quick to compute on very large data sets. All of these methods assume

a model that allows gene tree heterogeneity, and yet these methods each estimate a single species tree, and in some cases can handle multiple alleles per species (Maddision and Knowles 2006; Liu et al. 2008b). They are distinct from traditional methods of phylogenetic analysis in so far as there is no assumption that the estimated gene tree is isomorphic with the species tree; instead, they perform additional computation, whether calculation of likelihoods or summary statistics, on the collected gene trees to derive a species tree.

What's in a name?

It is a legitimate question to ask, as a colleague of mine did recently, whether species trees have any validity if in fact the definition of species themselves are still in limbo (as they are likely to be for a long time). This colleague suggested that the term 'population tree' is better suited to the new paradigm, because it avoids the issue of species validity (notwithstanding the problem of defining populations in nature). I would be happy with this terminology, but defining it this way might seem to exonerate those working at higher taxonomic levels, for whom population processes are minor concerns. Phylogeneticists working on the higher level questions tend not to concern themselves with populations, or their genetics. For this reason, 'population trees' might become appropriated solely by phylogeographers and those working near the species level. This would be unfortunate, since gene tree heterogeneity and the species tree problem in principle affects all levels of phylogeny, even if the extent of deep coalescence or branch length heterogeneity is less among higher taxa or sparsely sampled clades. For this reason I suggest we simply exercise a verbal substitution and reserve the term 'phylogeny' to refer to species trees. Phylogenies as they have been built in the last few decades would then be called gene trees, which is generally what they are, *sensu stricto*.

The logic of the species tree approach

From what little we know at this time, the species tree approach appears to derive its power from the accumulated signal of many gene trees, or independently segregating SNPs, each with their own 'tree' or bipartition. As such the approach leaves open the possibility that the collected DNA sequences may contain site patterns that are not directly mappable on to the resulting phylogeny. Complex signals and hidden support have been observed in combined and concatenated molecular data sets and have been suggested to arise from 'discrepant patterns of homoplasy' (Gatesy and Baker 2005). Yet, notwithstanding these complex interactions among characters, ultimately there can be no site patterns in a concatenated data sets that are not present in the original partitions. By contrast, species tree approaches explicitly conduct additional computation on trees from individual partitions; the end result can sometimes derive from signals that are not specifically encoded in the site patterns of the original partitions (Fig 2). A good illustration of this is the fact that species trees correctly estimated from gene trees in or near the anomaly zone differ from the most common gene tree, and by inference, from the signal in the most common site pattern in constituent partitions of the data (Edwards et al. 2007; Liu and Edwards 2008; Liu et al. 2008c). The additional signal not found in the original sequence data comes from the likelihood function of gene trees given a species tree (Maddison 1997; chapter 28 of Felsenstein 2003; Liu and Pearl 2007). This likelihood is distinct from the likelihood function modeling nucleotide substitution and its function is to provide probabilities of gene trees given a species tree. Such likelihoods have appeared in several forms recently and provide a solid foundation for developing new species tree methods (Rannala and Yang 2003; Degnan and Salter 2005; Steel and Rodrigo 2008).

Species trees: confidence and missing data

Although it is too early to tell clearly, I predict that statistical confidence in species trees when estimated with new multilocus approaches will in general be less than when estimated via concatenation, particularly when analyzing data sets of long-diverged clades, such a Orders of mammals or birds. I suggest this prediction even though we know that in some instances the species tree approach is more efficient at extracting information from DNA sequences than is the concatenation approach, such as the example from yeast (Edwards et al. 2007). This prediction stems from consideration of how signal is propagated in supermatrix and species tree approaches, and from a recent multilocus study on turtles that suggested that the effect of missing data was much stronger for species tree approaches than for concatenation approaches (Thomson et al. 2008).

It stands to reason that species tree approaches will be more sensitive to missing data than will supermatrix approaches because, in species tree approaches, a missing gene for a given taxon means that that taxon's genealogy is unknown for that particular gene (although it could probably be estimated for that gene based on the information from other genes). By contrast, in supermatrix approaches, a missing gene for a given taxon can easily be compensated for by other genes for that taxon, although the ease of compensation will no doubt vary. Hence there is may be less of a penalty for missing data in supermatrix approaches (although I confess my argument at this stage is not airtight). In the turtle study, the phylogeny of concatenated genes based on a data set in which nearly a third of the taxon-by-gene matrix had empty cells nonetheless had high confidence, with most branches achieving high posterior probability (Thomson et al. 2008). Similar claims of high confidence from vastly undersampled supermatrices have been made for other taxa as well (Driskell et al. 2004). Both the statistical inference issues – species trees are, after all, a different and more complex entity to estimate than gene trees – as well as the effects

of missing data may conspire to prove species trees in general harder to estimate than trees obtained by concatenation. Thus we may have to work harder to estimate species trees. This no doubt could be frustrating – after all, the community has become comfortable with the levels of confidence delivered under the current paradigm. But on the other hand, this extra effort may be telling us something about species trees and their ease of inference from genetic data. Concatenation also suffers from the problem of data 'swamping', in which one or a few partitions provides essentially all of the signal in a particular study, even in molecules-only analyses (Kluge 1983; Hillis 1987; Baker et al. 1998). I predict that the contribution to phylogenetic signal will be more evenly distributed among genes in species trees approaches, because in the end, each partition is only one gene, and extra signal comes from a each gene independently as well as from additional sites within any one gene. Of course, low confidence in species trees could also be the result of violations of the model assumed, such as when gene tree discordance is generated not just by coalescent phenomena but by horizontal gene transfer, intragenic gene conversion, paralogous genes or other processes. In general, as we begin to compare the relative merits of species trees and the concatenation approach, we should bear in mind that the two are different entities; although not exactly apples and oranges, they are nonetheless distinct statistical quantities that are correlated with one another and yet will behave differently with regard to signal maximization.

The future: simulations, sampling, species and SNPs

The species tree paradigm suggests a number of new directions that will impact future research. I choose three areas in particular - simulation practices, data sampling, and species delimitation - to complement the list of specific research questions outlined in Degnan and Rosenberg's recent review on related subjects (Degnan and Rosenberg 2008). Firstly, I suggest

that simulations of DNA sequences should from now on be conducted in a coalescent context, even if the simulated sequences are to be analyzed by traditional phylogenetic approaches. By this I mean DNA sequences should be simulated with a specific species tree in mind on which gene trees evolve, rather than through the traditional approach, which simply simulates DNA sequences on a static phylogenetic tree. For example, several simulation packages, such as MCMCcoal (Yang 2002), serialsimcoal (Laval and Excoffier 2004; Anderson et al. 2005) or Mesquite (Maddison and Maddison 2008) can simulate DNA sequences generated from gene trees that are in turn generated from explicitly specified species trees. By contrast, the most popular DNA sequence simulation packages, such as SeqGen, assume no coalescent stochasticity. The suggestion on simulation practices is independent of whether or not to concatenate. But simulating from coalescent gene trees would be an easy way to better approximate reality in ways that we do not now. One of course will be left with the choice of whether to simulate from long, thin species trees, which will generate a series of nearly identical gene trees (both in topology and branch lengths) or to simulate from short, fat species trees, which will generate substantial gene tree heterogeneity, and by extension, heterogeneity in phylogenetic signal of the underlying DNA sequences. This choice could essentially offer a 'way out' for those researchers who are reluctant to adopt the species tree approach; simulating from long, thin species trees and then concatenating these sequences prior to analysis is tantamount to the current approach to simulation, since there could be few if any signals emanating from the DNA sequences that are not easily ascribed to the shape of the species tree generating them. For some clades there will be population genetic information on the values of θ for extant populations; these could be used as a guide to assign lineage widths to species trees used in simulations (Edwards and Beerli 2000).

Second, I suggest that the new species tree paradigm will influence how we sample genomic data for phylogenetic analysis, and how confident we are of the results. As discussed above, for most purposes, sampling multiple genes for phylogenetic analysis has had as its most important consequence the accumulation of many sites for phylogenetic analysis. By contrast, the species tree approach places high value not only on the total number of sites, but also on the total number of independently segregating genes. I suggest, as have others (Maddison 1997; Avise 2000), that phylogenies are population phenomena and that the parameters of species trees and the means for estimating them from genetic data qualitatively are in the same class as recent models for estimating phylogeographic and demographic parameters within species, such as genetic diversity, rates of gene flow or population divergence times. These phylogeographic methods derive their statistical power from combining the information from many genes while still treating gene trees as independent of each other conditional on the demographic history being estimated. Recent theoretical and empirical analyses have demonstrated the dependence of statistical confidence in phylogeographic parameter estimation on the number of sampled loci (Jennings and Edwards 2005; Felsenstein 2006; Lee and Edwards in press); in many cases, the number of sampled loci appears to be more important in reducing variance of parameter estimates than the total number of base pairs (Carling and Brumfield 2007; Janes et al. 2008). In the same way, simulations have shown that confidence in species trees is also critically dependent on the number of sampled loci, although the contribution of the number of sites per locus to statistical confidence is still not known (Edwards et al. 2007; Liu et al. 2008b). The number of alleles sampled per species has also been shown to be an important variable determining phylogenetic accuracy and confidence (Maddison and Knowles 2006). Fortunately, many recent phylogenetics and phylogenomics data sets have already focused heavily on

sampling multiple loci, making extension to a species tree approach easier. Still, we don't yet know the optimal allocation of effort towards characterizing loci, individuals, and sequence length for phylogenetic analysis, if resources for a given project are limited.

Species and population delimitation will become fundamental to constructing species trees (O'Meara B 2008). This suggestion comes from the fact that another key assumption, at least in this first generation of species tree approaches, is lack of gene flow between OTUs. Lack of gene flow or other mechanisms of lateral genetic transfer go a long way towards satisfying the assumptions of many species tree approaches. (The impact of gene flow on species tree inference is likely to be substantial, yet in many ways no more severe than for gene tree inference; in both cases, care is required in interpreting the resulting tree). For this reason, a critical step in species tree analysis will be defining OTUs in such a way that this assumption is met.

In fact, species trees are often compatible with a number of prominent species concepts, particularly those that emphasize reproductive isolation, genetic cohesion and lineage isolation. For example, the 'metapopulation lineage species concept' proposed by DeQueiroz (2005) views species as sets of wholly or partially interbreeding units and subsumes many of the positive aspects of multiple species concepts. The growing appreciation of the multidimensionality of species and the variation in their embedded gene lineages (even by Willi Hennig, in his famous and frequently re-published diagram of gene lineages in two diverging populations in his *Phylogenetic Systematics*) makes such species concepts attractive and increasingly compatible with multilocus DNA sequence data sets that are becoming the norm. By contrast, species tree approaches are less compatible with species concepts that focus on diagnosibility via monophyly of gene trees. Although such monophyly is often criticized as a useful criterion for recognizing

species, particularly with mitochondrial DNA, such a criterion is nonetheless used quite regularly (Zink 2006; Zink and Barrowclough 2008). Other species concepts based on multilocus genealogical distinctiveness, such as the genealogical species concept or Avise and Ball's (1990) genealogical concordance concept, in which ~95% of gene lineages should be monophyletic under good species, are less useful in a species tree context, because the very nature of species trees acknowledges the possibility of distinct species despite rampant and ongoing incomplete lineage sorting (Edwards et al. 2005). In my view gene tree monophyly should be abandoned as a criterion for species, because, in addition to its conflation of patterns and criteria for diagnosibility at the level of genes and species, it can easily split biodiversity far too narrowly, or lump taxa far too liberally, depending on a variety of accidents of population genetics, including allelic sampling, natural selection, founder effects and other vagaries of population history (Rosenberg 2003, 2007).

A final issue that will be important to watch as species tree approaches diversify is the issue of recombination. Most species tree approaches (Table 1) have the tacit assumption that recombination is absent within genetic segments, but complete between such segments. This assumption allows each gene tree to be conditionally independent of the other trees, yet the signal of each gene to be internally consistent. There has been surprisingly little interest in studying the effects of recombination on phylogenetic analysis, in part because recombination can only occur among alleles in the same population; for this reason it is thought that recombination within diverging lineages that are not exchanging genes with other such lineages is unlikely to strongly affect higher level phylogenetics; no information is exchanged between species. Yet under the species tree paradigm, recombination within loci, or lack of recombination between loci (linkage) is likely to have important effects, and these should be

quantified; theory suggests that even small amounts of recombination between loci can quickly render their histories independent of one another in a species tree context (Slatkin and Pollack 2006). For these reasons individual unlinked SNPs may emerge as an important type of character to estimating phylogenies (species trees), and we are beginning to see efforts in this area (RoyChoudhury et al. 2008). Individual SNPs are a relief for those who worry about recombination within loci (since there is no recombination within a single SNP) and they can be collected rapidly on very large scales, as recent genome projects have shown. Again, phylogeographic methods might help show the way, as there are several methods tailored for within-species variation that extract useful information on population parameters from linked or unlinked SNPs (Falush et al. 2003, 2008; Pritchard Beerli 2006; Kuhner 2006). Some recent phylogeographic approaches that incorporate recombination within loci into the model for multlocus data appear promising (Kuhner 2006; Becquet and Przeworski 2007).

Conclusion – The Relevance of Species Trees

John Avise encapsulated the relationship between gene and species trees well in 1994: "Gene trees and species trees are equally 'real' phenomena, merely reflecting different aspects of the same phylogenetic process. Thus, occasional discrepancies between the two need not be viewed with consternation as sources of "error" in phylogeny estimation. When a species tree is of primary interest, gene trees can assist in understanding the population demographies underlying the speciation process" (pp. 133 and 138 in Avise 1994). This essay is in part meant to re-emphasize Avise' perspective and to remind readers that species trees are in fact the 'primary interest' of systematics.

My essay is not meant to champion any particular new software or statistical approach; but my polemic against concatenation and supermatrix approaches has no doubt been

emboldened by the recent success of a new generation of species tree approaches in a wide variety of phylogenetic situations (Table 1). Despite the advent of these new and often promising approaches, there is still a great need for additional models and methods that can efficiently analyze the very large phylogenomics data sets that are becoming the norm. Thus my essay is instead meant to champion a perspective on phylogenetics that has had many conceptual ancestors, yet is still in need of new models by theoreticians and experimentation by empiricists. The call for embracing species trees does not derive from the success of particular methods in a slightly wider region of tree space (such as the anomaly zone) than traditional methods. Nor does it derive from a failure of concatenation approaches to deliver reasonable trees, although I have suggested several ways in which concatenation can mislead. Rather, a heightened focus on species trees arises from an awareness of the near ubiquity of gene tree heterogeneity (whether in topology or branch lengths); from a consideration of the basic goals of systematics, whose focus is on trees of *species* and *lineages*; and from the fact that we can now act on these goals given the availability of at least a few computationally feasible methods. In one sense the transition could be construed as trivial; after all, species tree approaches really represent just a different way of combining data in phylogenetic analysis. On the other hand, the array of new approaches that have already appeared and the renewed focus on lineages and populations that they provide allow us to state in hindsight that systematics has been overly 'gene centric', at least since entering the PCR era. This gene centrism has been an extremely valuable way station as many other issues with the analysis of DNA sequence data have been sorted out. I suggest, however, that the field has now matured enough that we can move on to the next phase in which species and populations regain their rightful place as the primary focus in phylogenetic analysis.

Species tree approaches will of course open up a plethora of new debates and challenges for the field, both for higher level systematics and for phylogenetic analysis near the species level. For example, virtually any debate that has already taken place in the modern era of molecular systematics, can and will take place with species trees as the new focus. Such debates include issues on the molecular clock, taxon sampling, phylogenetic bias, rooting, incorporating fossil data, merging morphological and molecular data, and ways of achieving high levels of confidence. And yet in some cases, the consensus of the community may settle on an answer different from that proffered during the gene tree era of systematics. After all, the statistical quantities of species trees – topologies, branch lengths, times of divergence – are different from those for gene trees. For example, is more taxa or more sequence better for estimation of species trees? This question has for the most part received the answer of 'more taxa' or perhaps in some cases 'both' (Graybeal 1998; Pollock et al. 2002; Zwickl and Hillis 2002; Hedtke et al. 2006), but we have already seen that it might have a wholly new answer of 'more genes' in the case of species trees. Another example where species trees will usher in a new dialogue is the nature and sources polytomies (discussed above), a debate that I feel has been fraught with confusion precisely because the community has failed to adequately distinguish polytomies in gene trees versus polytomies in species trees. Ways of treating polymorphic characters in phylogenetic analysis, as well as optimal sampling of species for phylogenetic analysis may also benefit from clearly distinguishing between gene and species trees (Wiens 1999; Geuten et al. 2007). We can look forward to a more seamless integration of phylogeography and phylogenetics, two fields that have been divided in the recent past due to methodological and conceptual differences (Hey and Machado 2003; Brito and Edwards 2008). I suspect that species tree approaches, along with the new and awesome power of modern sequencing and computational methods, will play an

important role in creating a uniform methodological platform on which the diversity of genetic patterns emanating from diverse genomes can be interpreted and compared. They should be celebrated as a return to the genuine focus of systematics and will play an important role in helping build the Tree of Life, perhaps even facilitating the completion of this goal and a move beyond a focus on pattern to considerations of evolutionary mechanism and process.

Acknowledgements

I thank *Evolution* Editor Mark Rausher for inviting me to write a commentary and for his patience during writing. I thank my collaborators, L. Liu and D. Pearl for inspiring many of the ideas in this essay, as well as B. Rannala, J. Wakeley, J. Fensenstein, L. Knowles, D. Baum, N. Rosenberg, B. Carstens and R. Nielsen for helpful discussion over the years that has helped clarify my thoughts. L. Liu performed the simulations in Figures 2 and 4. I received very helpful comments on the manuscript from B. O'Meara, G. Spellman, B. Arbogast, C. Marshall and T. Near. B. O'Meara, N. Rosenberg, A. RoyChoudhury and J. Degnan provided helpful discussion and extensive comments on Table 1. Thanks to N. Rosenberg, J. Degnan and A. RoyChoudhury for providing preprints of as yet unpublished material. This work was supported by NSF grant 0743616 with D. Pearl.

Literature Cited

Alfaro, M. E., and M. T. Holder. 2006. The posterior and the prior in Bayesian phylogenetics. Ann. Rev. Ecol. Evol. Syst. 37: 19-42.

Anderson, C. N. K., U. Ramakrishnan, Y. L. Chan, and E. A. Hadly. 2005. Serial SimCoal: A population genetics model for data from multiple populations and points in time. Bioinformatics 21: 1733-1734.

Ané, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24: 412-426.

Avise, J. C. 1994. Molecular markers, Natural History and Evolution. Chapman and Hall, New York.

Avise, J. C. 2000. Phylogeography: The History and Formation of Species. Harvard University Press, Cambridge, MA.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Ann. Rev. Ecol. Syst. 18: 489-522.

Avise, J. C., and R. M. J. Ball. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Oxford Surveys in Evolutionary Biology 7: 45-67.

Avise, J. C., and K. Wollenberg. 1997. Phylogenetics and the origin of species. Proc. Natl. Acad. Sci. USA 94: 7748-7755.

Baker, R. H., X. B. Yu, and R. DeSalle. 1998. Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. Mol. Phyl. Evol. 9: 427-436.

Bapteste, E., E. Susko, J. Leigh, D. MacLeod, R. L. Charlebois, and W. F. Doolittle. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol. Biol. 5: 33.

Becquet, C., and M. Przeworski. 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res. 17:1505–1519.

Beerli, P. 2006. Comparison of Bayesian and maximum likelihood inference of population genetic parameters. Bioinformatics 22: 341-345.

Belfiore, N. M., L. Liu, and C. Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia : Geomyidae). Syst. Biol. 57:294-310.

Braun, E. L., and R. T. Kimball. 2001. Polytomies, the power of phylogenetic inference, and the stochastic nature of molecular evolution: A comment on Walsh et al. (1999). Evolution 55: 1261-1263.

Brito, P., and S. Edwards. 2008. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica. doi: 10.1007/s10709-008-9293-3

Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42: 384-397.

Carling, M. D., and R. T. Brumfield. 2007. Gene sampling strategies for multi-locus population estimates of genetic diversity (theta). PLoS One 2: e160.

Carstens, B., C., J. D. Degenhardt, A. L. Stevenson, and J. Sullivan. 2005. Accounting for coalescent stochsticity in testing phylogeographical hypotheses: modelling Pleistocene population structure in the Idaho giant salamander *Dicamptodon aterrimus*. Mol. Ecol. 14: 255-265.

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

Cavalli-Sforza, L. L. 1964. Population structure and human evolution. Proc. R. Soc. Lond. B

164: 362-379.

Charlesworth, B., R. Lande, and M. Slatkin. 1982. A neo-Darwinian commentary on macroevolution. Evolution 36: 474-498.

Chen, F. C., and W. H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am. J. Hum. Genet. 68: 444-456.

Cummings, M. P., S. P. Otto, and J. Wakeley. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12: 814-822.

de Queiroz, A. 1993. For consensus (sometimes). Syst. Biol. 42: 368-372.

de Queiroz, A., M. J. Donoghue, and J. Kim. 1995. Separate versus combined analysis of phylogenetic evidence. Ann. Rev. Ecol. Syst. 26: 657-681.

de Queiroz, K. 2005. Ernst Mayr and the modern concept of species. Proc. Natl. Acad. Sci. (USA) 102: 6600-6607.

Degnan, J. H., M. DeGiorgio, D. Bryant, and N. A. Rosenberg. 2008. Coalescent consequences for consensus cladograms. Syst. Biol. in press.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics 2: 762-768.

Degnan, J. H., and N. A. Rosenberg. 2008. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. Trends in Ecology & Evolution in press.

Degnan, J. H., and L. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59: 24-37.

Delsuc, F., H. Brinkmann, D. Chourrout, and H. Philippe. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439: 965-968.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6: 361-375.

Doolittle, W. F., and E. Bapteste. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc. Natl. Acad. Sci. U S A 104: 2043-2049.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20: 248-254.

Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one character taxononmy. Systematic Botany 17: 144-163.

Doyle, J. J. 1997. Trees within trees: genes and species, molecules and morphology. Syst. Biol. 46: 537-553.

Driskell, A. C., C. Ane, J. G. Burleigh, M. M. McMahon, C. O'Meara B, and M. J. Sanderson. 2004. Prospects for building the tree of life from large sequence databases. Science 306: 1172-1174.

Drummond, A. J., and A. Rambaut. 2003. BEAST v1.0.

Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452: 745-U745.

Edwards, S. V., and P. Beerli. 2000. Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54: 1839-1854.

Edwards, S. V., S. B. Kingan, J. D. Calkins, C. N. Balakrishnan, W. B. Jennings, W. J. Swanson,

and M. D. Sorenson. 2005. Speciation in birds: genes, geography, and sexual selection. Proc. Natl. Acad. Sci. (USA) 102, supp.1: 6550-6557.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. (USA) 104: 5936-5941.

Erwin, D. H. 2000. Macroevolution is more than repeated rounds of microevolution. Evolution & Development 2: 78-84.

Estes, S., and S. J. Arnold. 2007. Resolving the paradox of stasis: Models with stabilizing selection explain evolutionary divergence on all timescales. Am. Nat. 169: 227-244.

Ewing, G. B., I. Ebersberger, H. A. Schmidt, and A. von Haeseler. 2008. Rooted triple consensus and anomalous gene trees. BMC Evol Biol 8: 118.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567-1587.

—. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Molecular Ecology Notes 7:574-578.

Felsenstein, J. 1981. Evolutionary trees from gene-frequencies and quantitative characters - Finding Maximum-Likelihood Estimates. Evolution 35: 1229-1242.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. Ann. Rev. Genet. 22: 521-565.

Felsenstein, J. 2003. Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, MA.

Felsenstein, J. 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol. Biol. Evol. 23: 691-700.

Gadagkar, S. R., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from

multiple genes: concatenated sequence tree versus consensus gene tree. Journal of Experimental Zoology, Part B: Molecular and Developmental Evolution 304: 64-74.

Gatesy, J., and R. H. Baker. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst. Biol. 54: 483-492.

Geuten, K., T. Massingham, P. Darius, E. Smets, and N. Goldman. 2007. Experimental design criteria in phylogenetics: Where to add taxa. Syst. Biol. 56:609-622.

Gould, S. J. 1980. Is a new and general theory of evolution emerging? Paleobiology 6: 119-130.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problems? Syst. Biol. 47: 9-17.

Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55: 522-529.

Hey, J., and C. A. Machado. 2003. The study of structured populations - new hope for a difficult and divided science. Nature Reviews Genetics 4: 535-543.

Hillis, D. M. 1987. Molecular Versus Morphological Approaches to Systematics. Ann. Rev. Ecol. Syst. 18: 23-42.

Hillis, D. M., M. W. Allard, and M. M. Miyamoto. 1993. Analysis of DNA sequence data: phylogenetic inference. Meth. Enzymol. 224: 456-487.

Hobolth, A., O. F. Christensen, T. Mailund, and M. H. Schierup. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet 3: e7.

Holland, B., F. Delsuc, and V. Moulton. 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.

Holland, B. R., L. S. Jermiin, and V. Moulton. 2006. Improved consensus network techniques for genome-scale phylogeny. Mol. Biol. Evol. 23: 848-855.

Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics 131: 509-512.

Hudson, R. R., and M. Turelli. 2003. Stochasticity overrules the "three-times rule": Genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. Evolution 57: 182-190.

Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51: 673-688.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

Janes, D. E., T. Ezaz, J. A. Marshall Graves, and S. V. Edwards. 2008. Recombination and nucleotide diversity in the sex chromosomal pseudoautosomal region of the Emu, *Dromaius novaehollandiae*. J. Hered. doi:10.1093/jhered/esn065

Jennings, W. B., and S. V. Edwards. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from 30 gene trees. Evolution 59: 2033-2047.

Kluge, and Ag. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst Zool 38: 7-25.

Kluge, A. G. 1983. Cladistics and the classification of great apes. Pp. 151-177 *in* R. L. Ciochan, and R. S. Coruccini, eds. New Interpretations of Ape and Human Ancestry. Plenum, New York.

Kluge, A. G. 2004. On total evidence: for the record. Cladistics 20: 205-207.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and

likelihood phylogenetics when evolution is heterogeneous. Nature 431: 980-984.

Kolaczkowski, B., and J. W. Thornton. 2006. Is there a star tree paradox? Mol. Biol. Evol. 23: 1819-1823.

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

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

Kuhner, M. K. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22: 768-770.

Laval, G., and L. Excoffier. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20: 2485-2487.

Lee, J. Y., and S. V. Edwards. 2008. Divergence across Australia's Carpentarian barrier: statistical phylogeography of the Red-backed Fairy Wren (*Malurus melanocephalus*). Evolution, in press.

Lewis, P. O., M. T. Holder, and K. E. Holsinger. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54: 241-253.

Liu, L., and S. V. Edwards. 2008. Phylogenetic analysis in the anomaly zone. Manuscript, Cambridge, MA.

Liu, L., L. S. Kubatko, D. K. Pearl, and S. V. Edwards. 2008a. Coalescent methods for estimating multilocus phylogenetic trees. submitted.

Liu, L., and D. K. Pearl. 2006. Species trees from gene trees: reconstructing posterior distributions of a species phylogeny using estimated gene tree distributions. Pp. 24.

Mathematical Biosciences Institute Technical Report #53. Ohio State University, Columbus.

Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56: 504-514.

Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008b. Estimating species trees using multiple-allele DNA sequence data. Evolution: 2080-2091.

Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards. 2008c. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. in review.

Lynch, M., and P. E. Jarrell. 1993. A method for calibrating molecular clocks and its application to animal mitochondrial DNA. Genetics 135: 1197-1208.

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

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

Maddison, W. P. and D.R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. Version 2.5 http://mesquiteproject.org

Matsen, F. A., and M. Steel. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst. Biol. 56: 767-775.

Misawa, K., and M. Nei. 2003. Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. J. Mol. Evol. 57: S290-S296.

Mossel, E., and S. Roch. 2007. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. Available at http://arxiv.org/abs/0710.0262.

Mossel, E., and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of

trees. Science 309: 2207-2209.

Nei, M., and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Neigel, J. E., and J. C. Avise. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. Pp. 515-534 *in* S. Karlin, and E. Nevo, eds. Evolutionary Processes and Theory. Academic Press, New York.

Nielsen, R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor. Pop. Biol. 53: 143-151.

Nielsen, R., J. L. Mountain, J. P. Huelsenbeck, and M. Slatkin. 1998. Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. Evolution 52: 669-677.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53: 47-67.

O'Meara B, C. 2008. Using Trees: *Myrmecocystus* Phylogeny and Character Evolution and New Methods for Investigating Trait Evolution and Species Delimitation (PhD Dissertation). Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2008.2261.1>.

Otto, S. P., M. P. Cummings, and J. Wakeley. 1996. Inferring phylogenies from DNA sequence data: The effects of sampling. New Uses For New Phylogenies 349.

Page, R., and M. A. Charleston. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol. Phyl. Evol. in press.

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

Patterson, N., D. J. Richter, S. Gnerre, E. S. Lander, and D. Reich. 2006. Genetic evidence for

complex speciation of humans and chimpanzees. Nature 441: 1103-1108.

Pigliucci, M. 2007. Do we need an extended evolutionary synthesis? Evolution 61: 2743-2749.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in Drosophila: Evidence for incomplete lineage sorting. Plos Genetics 2: 1634-1647.

Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51: 664-671.

Rambaut, A. 2007. SeqGen.

Rannala, B., and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.

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

Rokas, A., B. Williams, N. King, and S. Carroll. 2003. Genome-scale approachese to resolving incrongruence in molecular phylogenies. Nature 425: 798-804.

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.

Rosenberg, N. A. 2007. Statistical tests for taxonomic distinctiveness from observations of monophyly. Evolution 61: 317-323.

RoyChoudhury, A., J. Felsenstein, and E. A. Thompson. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. Genetics 180.

Sanderson, M. J., and M. M. McMahon. 2007. Inferring angiosperm phylogeny from EST data

with widespread gene duplication. BMC Evol Biol 7 Suppl 1: S3.

Satta, Y., J. Klein, and N. Takahata. 2000. DNA archives and our nearest relative: The trichotomy problem revisited. Mol. Phyl. Evol. 14: 259-275.

Seo, T. K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25: 960-971.

Seo, T. K., H. Kishino, and J. L. Thorne. 2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. Proc. Natl. Acad. Sci. USA 102: 4436-4441.

Simmons, M. P., K. M. Pickett, and M. Miya. 2004. How meaningful are Bayesian support values? Mol. Biol. Evol. 21: 188-199.

Slatkin, M. and J. L. Pollack. 2006. The concordance of gene trees and species trees at two linked loci. Genetics 172: 1979-84.

Slowinski, J., and R. D. M. Page. 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48: 814-825.

Slowinski, J. B. 2001. Molecular polytomies. Mol Phylogenet Evol 19: 114-120.

Smith, J. M. 1983. The genetics of stasis and punctuation. Ann. Rev. Genet. 17: 11-25.

Steel, M., A. W. Dress, and S. Bocker. 2000. Simple but fundamental limitations on supertree and consensus tree methods. Syst. Biol. 49: 363-368.

Steel, M., and F. A. Matsen. 2007. The Bayesian "star paradox" persists for long finite sequences. Mol. Biol. Evol. 24: 1075-1079.

Steel, M., and A. Rodrigo. 2008. Maximum likelihood supertrees. Syst. Biol. 57: 243-250.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA 99: 16138-16143.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference *in* D. M. Hillis, C. Moritz, and B. K. Mable, eds. Molecular Systematics. Sinauer, Sunderland, MA.

Takahata, N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122: 957-966.

Taylor, D. J., and W. H. Piel. 2004. An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. Mol. Biol. Evol. 21: 1534-1537.

Thomson, R. C., A. M. Shedlock, S. V. Edwards, and H. B. Shaffer. 2008. Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. Mol. Phyl. Evol. in press.

Waddell, P. J., H. Kishino, and R. Ota. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Informatics 12: 141-154.

Waddell, P. J., H. Kishino, and R. Ota. 2002. Very fast algorithms for evaluating the stability of ML and Bayesian phylogenetic trees from sequence data. Genome Informatics 13: 82-92.

Wakeley, J., and J. Hey. 1997. Estimating ancestral population parameters. Genetics 145: 847-855.

Walsh, H. E., M. G. Kidd, T. Moum, and T. Friesen. 1999. Polytomies and the power of phylogenetic inference. Evolution 53: 932-937.

Wiens, J. J. 1998. Combining data sets with different phylogenetic histories. Syst. Biol. 47: 568-581.

Wiens, J. J. 1999. Polymorphism in systematics and comparative biology. Ann. Rev. Ecol. Syst. 30: 327-362.

Wilson, A. C., R. L. Cann, S. M. Carr, M. George, U. B. Gyllensten, K. M. Helm-Bychowski, R.

G. Higuchi, S. R. Palumbi, E. M. Prager, R. D. Sage, and M. Stoneking. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. Biol. J. Linn. Soc. 26: 375-400.

Wong, A., J. D. Jensen, J. E. Pool, and C. F. Aquadro. 2007. Phylogenetic incongruence in the *Drosophila melanogaster* species group. Mol. Phyl. Evol. 43: 1138-1150.

Wu, C. I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics 127: 429-435.

Yang, Z. 1997. On the estimation of ancestral population sizes of modern humans. Genetical Research 69: 111-116.

Yang, Z. 2002. MCMCcoal: Markov Chain Monte Carlo Coalescent Program, version 1.0. Pp. 8, Oxford.

Yang, Z., and B. Rannala. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54: 455-470.

Zink, R. M. 2006. Rigor and species concepts. Auk 123: 887-891.

Zink, R. M., and G. F. Barrowclough. 2008. Mitochondrial DNA under siege in avian phylogeography. Mol. Ecol. 17: 2107-2121.

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51: 588-598.

Table 1. Examples of methods for estimating species trees*.

Method (References)	Methodological basis	Data required	Accounts for stochastic variation or gene tree error?	Yields species tree branch lengths?	Yields effective population sizes?	Applicable to many loci?	Applicable to many taxa?
GENE TREE DISTRIBUTIONS							
PROBABILITY OF INCONGRUENCE (Pamilo and Nei 1988; Wu 1991; Hudson 1992; Chen and Li 2001; Waddell et al. 2002)	LIKELIHOOD/ COALESCENT	GENE TREES	NO	YES	YES	YES	NO
DEMOCRATIC VOTE (PAMILO AND NEI 1988; SATTA ET AL. 2000)	GENE TREE COUNTS	GENE TREES	NO	NO	NO	YES	NO
SINE METHOD (discordance) (Waddell et al. 2001)	LIKELIHOOD	BINARY CHARACTERS	NO	NO	YES	YES	NO
GENE TREE SHAPES OR CONFLICT MINIMIZATION							
GENETREE PARSIMONY (Page and Charleston 1997)	PARSIMONY	MULTIGENE FAMILY TREES	NO	NO	NO	YES	MODERATE
DEEP COALESCENCE (Maddison 1997; Maddison and Knowles 2006)	PARSIMONY	GENE TREES	NO	NO	NO	YES	YES
SPECIES TREES USING AVERAGE RANK OF COALESCENCE TIME (STAR) (LIU ET AL. 2008C)	RANKS OF PAIRWISE COALESCENCE TIMES	COALESCENCE TIMES/GENE TREES	VIA BOOTSTRAPPING	NO	NO	YES	YES
SPECIE TREES USING ESTIMATED AVERAGE COALESCENCE TIME (STEAC) (LIU ET AL. 2008C)	PAIRWISE COALESCENCE TIMES	COALESCENCE RANKS/GENE TREES	VIA BOOTSTRAPPING	NO	NO	YES	YES
MINIMUM DIVERGENCE (Takahata 1989); MAXIMUM TREE (LIU AND PEARL 2006); GLASS (MOSSEL AND ROCH 2007)	DIVERGENCE IN GENE TREES/ COALESCENT	GENE TREES	NO	YES (ASSUMING ULTRAMETRICITY)	NO	YES (MAXIMUM AND GLASS)	YES
JOINT INFERENCE OF SPECIES AND TREE (JIST) (O'MEARA B 2008)	LIKELIHOOD/ COALESCENT	GENE TREES	NO	NO	NO	YES	MODERATE
ALLELE FREQUENCIES, SNPS OR HAPLOTYPE CONFIGURATIONS							
DRIFT MODEL (FELSENSTEIN 1981)	LIKELIHOOD/ BROWNIAN MOTION	ALLELE FREQUENCIES	YES	YES	NO	YES	YES
INFINITE SITES MODEL (Nielsen 1998)	LIKELIHOOD	HAPLOTYPES	YES	YES	YES	YES	NO
FST METHOD (Nielsen et al. 1998)	LIKELIHOOD/ COALESCENT	ALLELE FREQUENCIES	YES	YES	YES	YES	NO
PRUNING	LIKELIHOOD/	SNPS	YES	YES	NO	YES	MODERATE

Algorithm(RoyChoudhury et al. 2008)	Coalescent						
Gene tree probabilities/likelihoods							
Gene tree probabilities (Carstens and Knowles 2007)	likelihood/ Coalescent	Gene Trees	Partially	No	No	Yes	No
Bayesian Estimation of Species Trees (BEST) (Liu and Pearl 2007; Liu et al. 2008b)	Bayesian	DNA sequences	Yes	Yes	Yes	Moderate	Moderate
Bayesian Concordance Factors (BCA) (Ane et al. 2007)	Bayesian	DNA sequences	Yes	No	No	Yes	Moderate
Sum And Average Criteria (Seo et al. 2005)	Likelihood	DNA sequences	Yes	No	No	Yes	Yes
Consensus And Supertree approaches							
Likelihood supertrees (Steel and Rodrigo 2008)	Likelihood	Gene trees	Yes	No	No	Yes	Yes
Rooted Triple Consensus (Degnan et al. 2008; Ewing et al. 2008)	Consensus	Gene Trees	No	No	No	Yes	Yes
Majority Rule Consensus/Greedy Consensus (Degnan et al. 2008)	Consensus	Gene Trees	No	No	No	Yes	Yes

*Modified and expanded from Table 1 of Brito and Edwards (2008). The table is not mean to be exhaustive (see text)

Figure legends

Figure 1. Distinction between deep coalescence and branch length heterogeneity as sources of gene tree heterogeneity and gene tree/species tree conflict. Example species trees are shown at the top, with constituent gene trees in the bottom row; taxa from which gene trees are sampled are given as A, B, C and D. Whereas deep coalescence emphasizes topological differences between gene and species trees, branch length heterogeneity emphasizes branch length differences between gene and species trees and variation among genes in branch length, without topological variation. Branch length heterogeneity is ubiquitous and will be important for impacting site distributions in DNA sequences when effective population sizes and species tree branch lengths are large enough to permit substantial variation in coalescence times without deep coalescence. Heterogeneity in branch lengths among constituent gene trees is indicated by the dashed lines in the lower right panel.

Figure 2. Example of the contribution of coalescent variance to the distribution of site patterns in DNA sequences. A species tree (top left and white bars in graphs) was used to simulate gene trees and DNA sequences of 500 bp using the Jukes Cantor model of DNA evolution using MCMCcoal (Yang 2002). A gene tree (top right and grey bars in graphs) with branch lengths and topology identical to the species tree, but without the effective population size parameter, was used to simulate gene trees and DNA sequences. 1000 gene trees were simulated from the species tree and one sequence per gene tree was then simulated; for the gene tree analysis, 1000 DNA sequences were simulated from the single gene tree. The frequency of two of the 15 possible site patterns for four taxa under the Jukes Cantor model was counted for each replicate.

The two graphs show the frequency of two sites, one consistent with the species tree and gene tree (XXYY) and one apparently in conflict with the species tree and gene tree (XYYX). The species tree used was ((((H:0.05 , C:0.05) : 0.0025 #0.1, G:0.0525):0.0025 #0.1 , O:0.055) #0.1); numbers after the pound sign indicate value of $\theta = 4N\mu$. Approximately 68% of the gene trees simulated with this species tree are concordant with the species tree. The value of θ used in the species tree simulations is admittedly high and primarily for illustrating the situation with a high mutation rate; but substantial tails to the distribution of site frequencies is achieved with species trees an order of magnitude shorter and thinner. The gene tree used was (((H:0.05 , C:0.05) : 0.0025, G:0.0525):0.0025, O:0.055). DNA sequences simulated from species trees are likely to contain a higher number of sites that "conflict" with the species tree, even though from a species tree perspective they are not really in conflict with it. For example, over 20% of the gene trees simulated from the species tree (n = 215) gave rise to DNA sequences in which greater than 10 percent of sites (> 5 sites) had pattern XYYX, which is naively in conflict with the true tree. But in fact, such sites are consistent with the species tree when a clear distinction between gene and species trees is made. By contrast, no sequences simulated from the single gene tree had this many sites with pattern XYYX, and less than 20% (n = 192) of sequences had more than one site of this type.

Figure 3. A fictitious likelihood plot illustrating the idea that gene trees represent a local optimum historically in the field of systematics. The plot also alludes to the greater explanatory power of species tree models over models without gene tree heterogeneity. I showed a plot similar to this at the symposium on species trees at the 2008 meetings of the Society for the Study of Evolution in Minneapolis, and received a number of chuckles from the audience, and it

is presented here in a similarly irreverent spirit. In fact, models that allow for gene tree heterogeneity do have significantly more explanatory power for those DNA data sets that have been tested than do concatenation or supermatrix models. However, the extra parameters of some species tree approaches are a disadvantage.

Figure 4. Illustration of the utility of the species tree approach as a framework for studying polytomies (top); the mixture of dichotomous gene trees that are expected to result from a polytomy in the species tree (middle); and the tendency for concatenation to excessively favor one particular topology when presented with a mixture of gene trees that, together, should cause lower confidence in any one topology (bottom). In two simulations, the polytmous species tree at the top was used to generate 30 gene trees, which in turn were used to generate DNA sequences under the Jukes Cantor model using MCMCcoal (Yang 2002). The three possible gene trees produced from a polytomous species tree are indicated in black, grey and dotted lines. These sequences were analyzed either with the method Bayesian Estimation of Species Trees (BEST, Liu and Pearl 2007, lower left corner; Liu et al. 2008b) or were concatenated and analysed using MrBayes (Huelsenbeck and Ronquist 2001). This procedure was repeated 10 times ('replicate'). The optimal distribution of posterior probabilities would be even at ~0.33 across all replicates and trees; given the finite nature of the simulation, the observed probabilities are expected to vary from this optimum somewhat. Whereas BEST achieves moderately even posterior probabilities across trees and replicates, concatenation produces strongly uneven probabilities that favor one tree or another depending on detail of each replicate. This unevenness is likely a consequences of concatenation, rather than any idiosyncracies in

MrBayes, and illustrates that concatenation itself can be a major source of overconfidence in phylogenetic trees (see text).

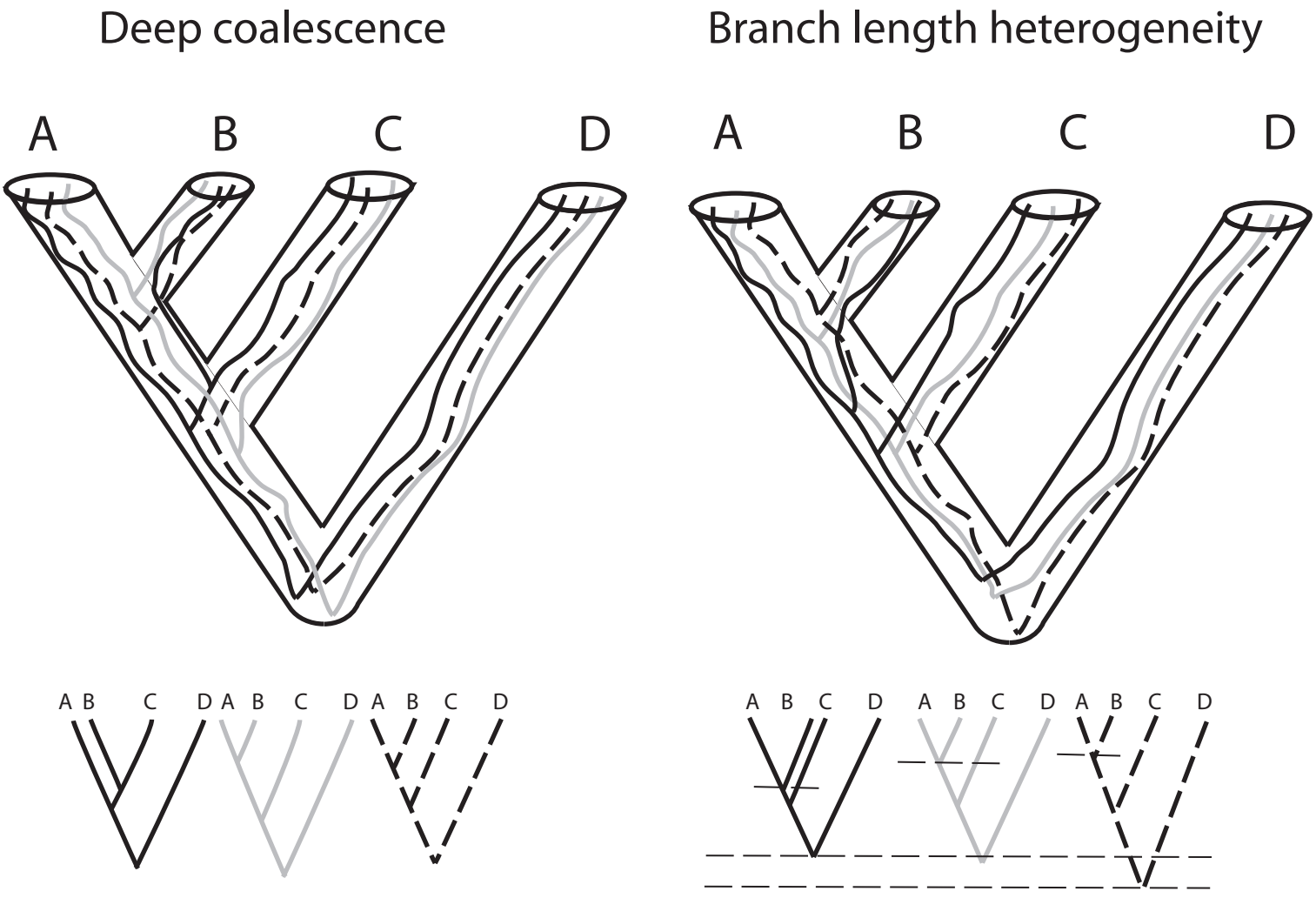
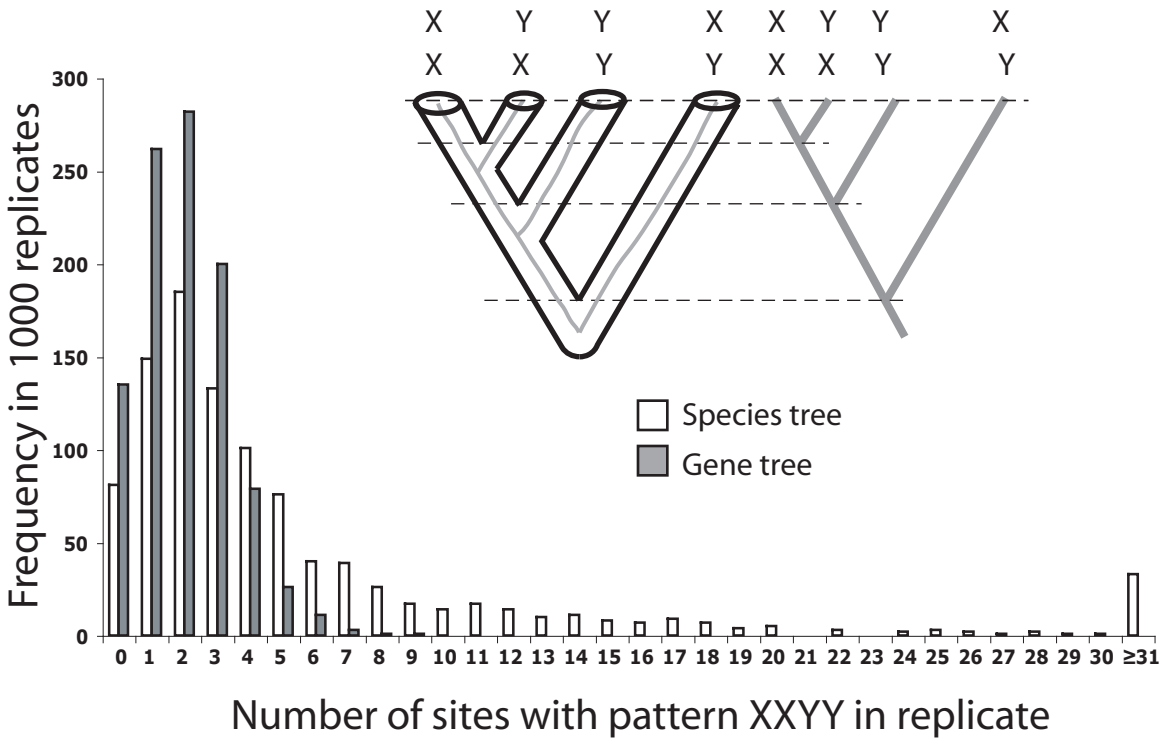

Deep coalescence
Branch length heterogeneity
A
B
C
D
A B C D A B C D A B C D
A B C D A B C D A B C D

Figure 1


X X Y X Y X X Y Y X
X X Y Y X X Y Y Y
Species tree
Gene tree
Frequency in 1000 replicates
0 50 100 150 200 250 300
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 ≥31
Number of sites with pattern XXYY in replicate

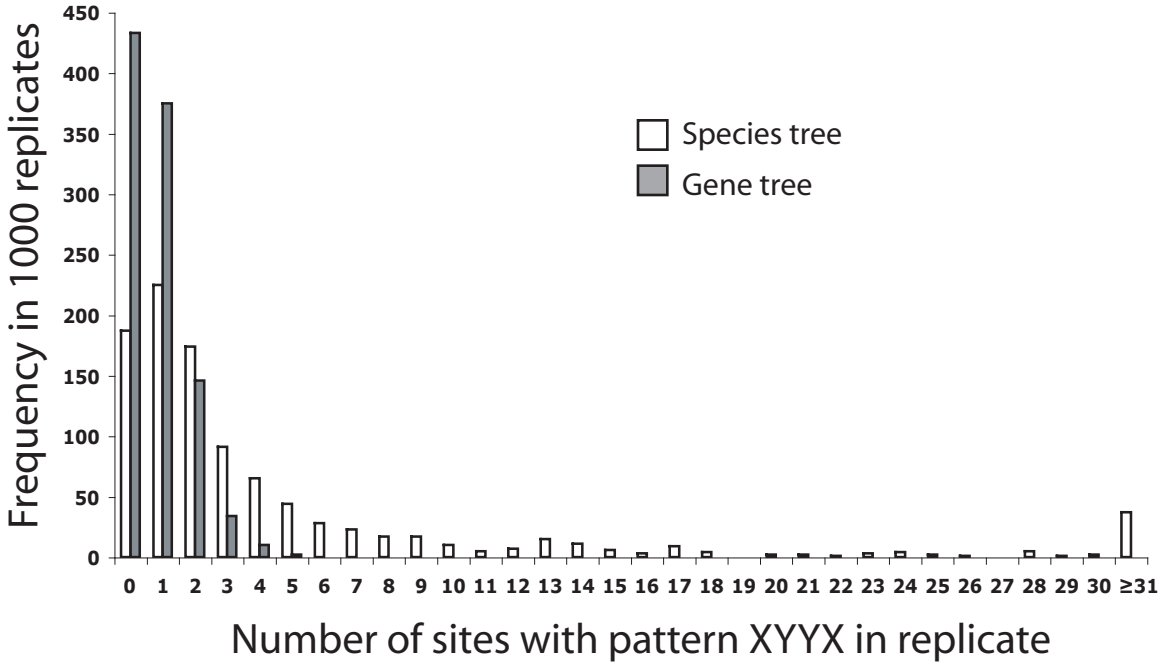

Species tree
Gene tree
Frequency in 1000 replicates
0 50 100 150 200 250 300 350 400 450
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 ≥31
Number of sites with pattern XYYX in replicate

Figure 2

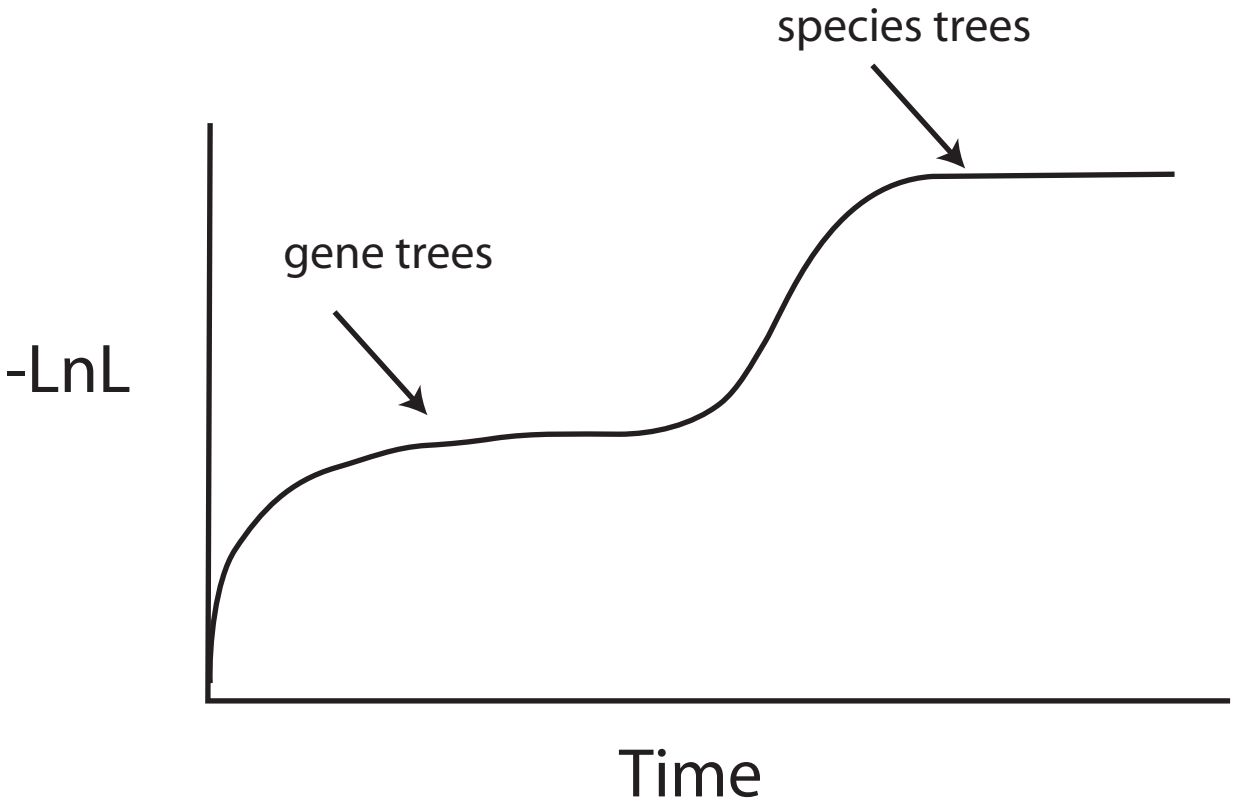
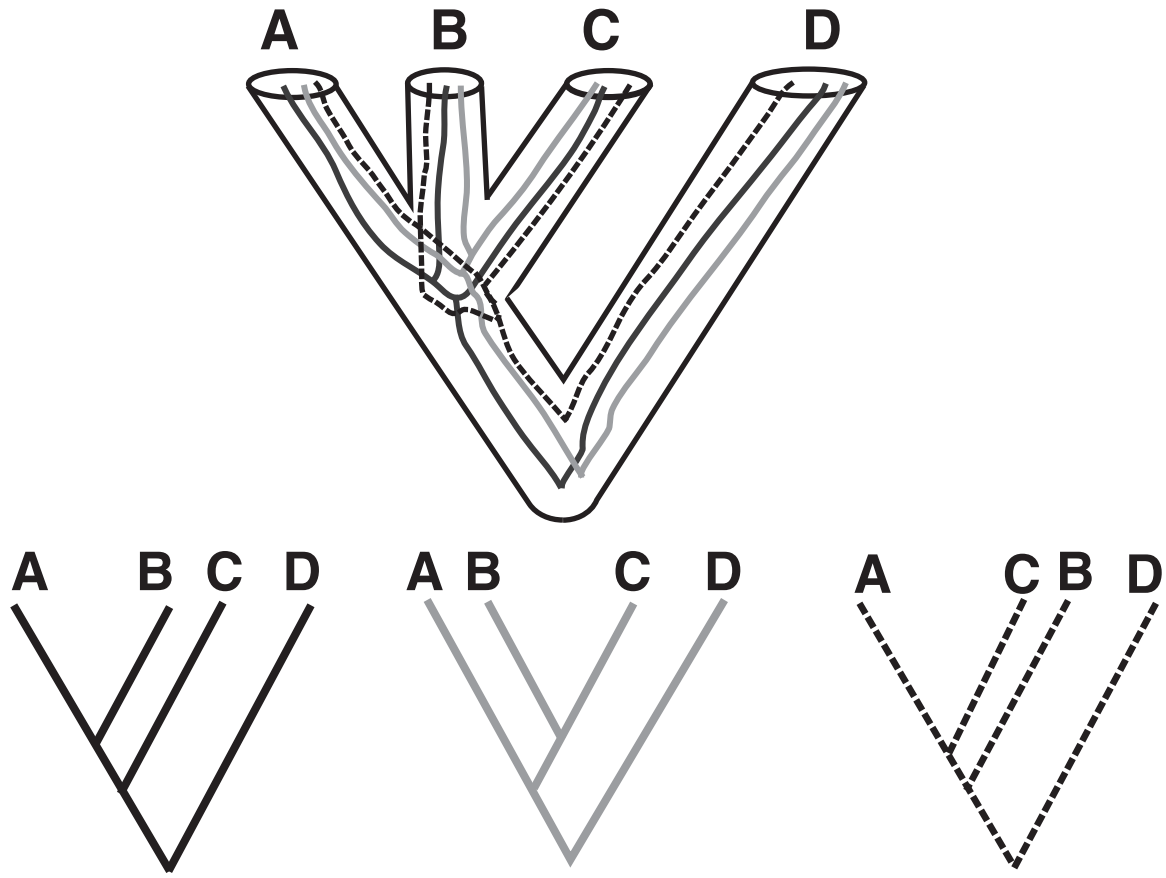


Figure 3

A


A B C D
A B C D
A B C D
A B C D
Gene tree 1
Gene tree 2
Gene tree 3

B

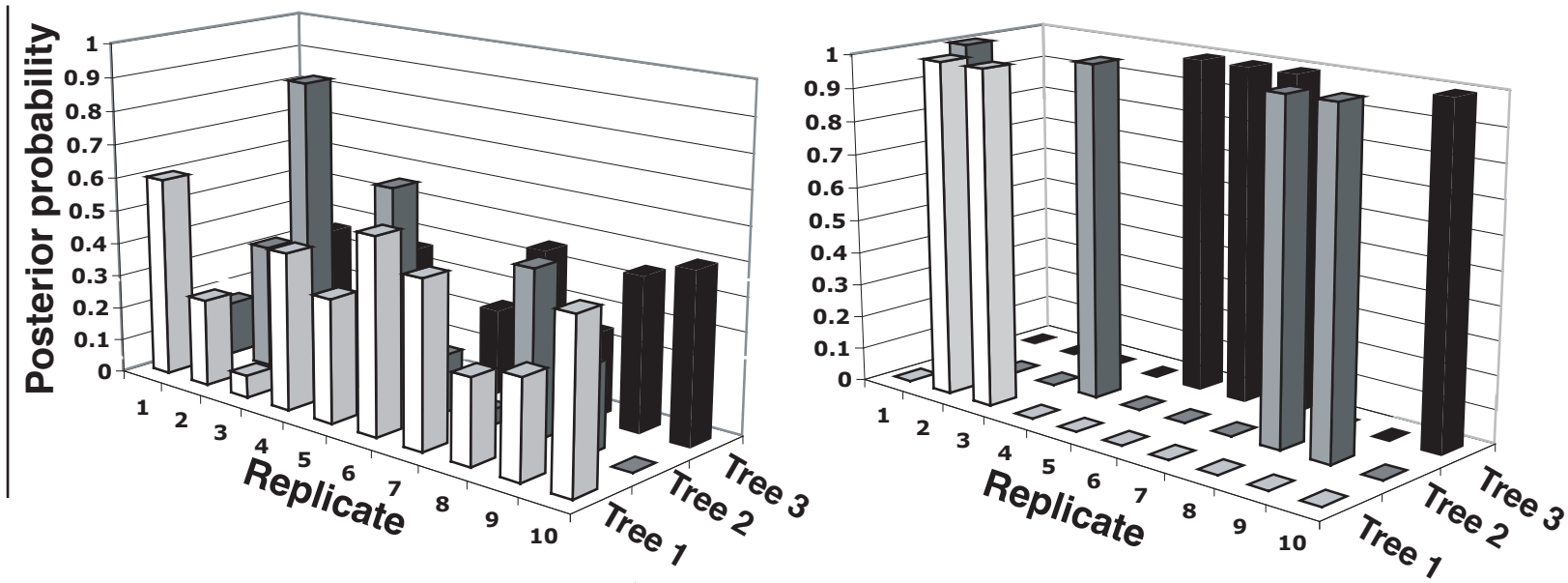

Posterior probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 2 3 4 5 6 7 8 9 10
Replicate
Tree 1
Tree 2
Tree 3

Figure 4