Species tree inference in the age of genomics

Nathan V. Whelan

Department of Biological Sciences, University of Alabama, USA

Abstract

Species trees are an essential tool in conservation and evolutionary biology. In phylogenomics, not only is data choice (*e.g.* using unlinked orthologs rather than paralogs) an important systematic consideration, but the choice of phylogenetic algorithm is also important. Since individual gene phylogenies can differ from the true species phylogeny, new methods have been proposed for species tree estimation using multiple unlinked genes. Improvements in genome sequencing technologies have increased the amount of data available to researchers and this has increased the utility of multi-locus species tree inference methods. The Bayesian methods BEST and *BEAST that incorporate a coalescent model to account for gene tree and species tree conflict offer promising advances in species tree inference directly from DNA sequences. Methods that infer species trees from gene trees rather than directly from sequence data such as STAR, STEAC, NJ_{st} and the likelihood method STEM have been recently developed as computationally efficient alternatives. Bayesian concordance analysis, which has been shown to perform well when horizontal gene transfer is the cause of gene tree and species tree conflict, is also discussed. Furthermore, methods for species delimitation including a non-parametric species tree inference method that does not require *a priori* species assignments can remove subjectivity from species delimitation. Here, I review the assumptions, required inputs, and the performance of these methods under simulation and in recent empirical studies. Researchers in many disciplines should understand the methods available for phylogenomic species tree inference in order to enhance evolutionary and conservation studies.

Introduction

Species level phylogenies are essential to answering many biological questions including the analysis of the speciation process,[1-2] species delimitation,[3-5] community ecology questions,[6-7] and basic species relationships.[8] Species level phylogenies are also used to define management units[9,10] which are of particular importance in the light of the current global biodiversity crisis.[11] Therefore, the accuracy and confidence put into species trees can have broad implications as these trees are one of the best tools available for species delimitation under the evolutionary species concept[2] and for modern comparative biology.[12] It is essential that researchers utilize phylogenetic methods and increasing amounts of genomic data in a biologically meaningful manner.

Before the development of PCR and first generation gene sequencing technology, researchers relied on finding homologous morphological characters that could be used to infer phylogeny.[2] Although such characters still have their use today in systematics (see review by Giribret[13]), there are limits to how many morphological characters are available for any given taxon. This is especially true for organisms such as bacteria, morphologically cryptic species, and for undistinguishable life history stages (*e.g.* insect larva). Sequence data have resulted in an explosion of phylogenies for a breadth of taxa, but the tree of life is far from assembled. It is evident that different genes can have differing evolutionary histories which can cause individual alleles within a species to be polyphyletic, or not share the same genealogy as the species.[14-15] Such gene tree conflict (Figure 1) can be caused by introgression, incomplete lineage sorting, and conflicting modes of selection and inheritance (*e.g.* mitochondrial *vs* nuclear genes;[15,16]). For this reason, there is a need to distinguish between gene trees and species trees.

The question of differentiating gene trees from species trees has been a fundamental issue for systematists since the advent of multi-loci analysis. Although high throughput sequencing has presented great opportunities for data collection and has allowed for the development of phylogenomics, this too presents easily overlooked issues including misleading data (*e.g.* bias introduced by ambiguous data and discordance between the most likely gene trees and species trees;[17,18-19]), gene choice (*e.g.* orthologs rather than paralogs;[20]), missing data[17,21-22] and adequate taxon sampling.[23-24] Phylogenetic inference algorithm choice is also of concern. New methods for inferring species trees with multiple loci (Table 1) have recently been developed and their properties, use, and accuracy warrant a thorough review. Here I review important considerations for character and taxon sampling in phylogenomics, the most recent and promising methodological advances in species tree inference and phylogenomics, and how the algorithms have performed in simulation and empirical studies to date.

Correspondence: Nathan V. Whelan, Department of Biological Sciences, University of Alabama, BOX 870345, Tuscaloosa, AL 35487 USA.
E-mail: nwhelan@crimson.ua.edu

Key words: species tree, coalescence, phylogenomics.

Acknowledgments: I would like to thank the Spring 2011 Genomics class at the University of Alabama and especially LK Reed for insight and comments on earlier drafts of this manuscript. Two anonymous reviewers also provided helpful comments which greatly improved this paper from previous versions.

Conflict of interest: the authors have no conflict of interest.

Received for publication: 17 August 2011.
Revision received: 20 September 2011.
Accepted for publication: 4 November 2011.

Trends in Evolutionary Biology 2011; 3:e5
doi:10.4081/eb.2011.e5

Character and taxon sampling

Many phylogenomic studies address hypotheses of deep evolution along the tree of life,[25-26] but questions about species and sub-species relationships are just beginning to be pursued under a phylogenomic framework.[3,19,27-29] The argument of whether increased taxon sampling or increased character sampling will enhance phylogenetic robustness has been a lively debate in the systematic literature for years. However, recent theoretical advances suggest that the answer likely depends on whether the lack of phylogenetic resolution exists at the base of the phylogeny or at the tips. According to Townsend and Lopez-Giraldez,[24] increased character sampling is ideal when resolving the tips of trees, which is more of an issue when looking at species level relationships rather than deep nodes. Furthermore, taxon and character sampling design in conjunction with how recently species level divergence took place can affect species tree inference accuracy.[30,31]

Although genomic technology can provide many genes for use in phylogenetic analyses, care must be taken to ensure orthologs rather than paralogs are used so as to not introduce homoplasy into an analysis. For groups with sequenced genomes, multiple databases exist for the identification of orthologs (see Altenhoff and Dessimoz[20] for an analysis of database search effectiveness). In model organisms such as *Drosophila*, full genomes for 12 species have been used to infer phylogenetic relationships.[32]

For less studied groups such as many macroinvertebrate taxa, genomic data are typically only available for a handful of taxa. For non-model organisms, BAC libraries (genome fragments sequenced with plasmid vectors) or shotgun sequencing approaches can provide a source of phylogenetically informative markers.[33] Thomson *et al.*[34] and Carstens and Dewey[3] showed BAC libraries are a reliable source of primers for traditional PCR and gene sequencing of phylogenetically informative markers. PCR primers developed from such genomic techniques do not appear to have the functional breadth of so called universal primers.[28]

Using a wide variety of unlinked orthologs and multiple individuals per species is important to capture differing gene evolutionary histories within the genomes of any given taxon.[31] However, Townsend *et al.*[19] found that it may be better to sequence a smaller number of fast evolving loci rather than many fast and/or slowly evolving loci using the species tree inference methods BEST and *BEAST (discussed below); this is in notable contrast to the standard phylogenomics paradigm of using as much sequence data as possible.

One issue with phylogenomic studiesis that many include missing data.[25,26,35] The importance and effects of missing data on phylogenetic studies has been a point of lively debate for years, but exactly how such missing data affects species tree inference methods discussed below has yet to be studied.[36] Modern phylogenomic studies commonly have missing data, and for this reason an in-depth analysis of how missing data affects species tree inference algorithms should be pursued with vigor.

Species tree estimation algorithms

Traditional phylogenetic algorithms such as Maximum Parsimony (MP), Maximum Likelihood (ML) and Bayesian Inference (BI) are widely used and thoroughly tested methods for phylogenetic gene tree inference. A major criticism of these methods is that they may only capture gene history and not the true species history. In fact, these methods can be positively misleading when applied to multilocus data giving statistically supported but incorrect phylogenies. This can occur because genes from different genomic regions can have genealogies that are statistically more likely than the true species tree.[18] This problem may be exacerbated when the focus of a phylogeny is species level relationships, since full coalescence is unlikely to have occurred for all genes at shallow evolutionary time scales.[15,37]

Species tree estimation methods that utilize multi-locus data have been advanced as practical methods for inferring species trees from gene trees, but these methods have not been as extensively vetted as traditional methods of gene tree inference. Coalescent based methods (*e.g.* BEST, *BEAST, and STEM) all have the underlying assumption that the only source of incongruence between gene trees is the result of incomplete lineage sorting; this assumption may not always be realistic and other non-coalescent model based methods should be considered. Numerous methods have been developed for species tree inference (Table 1), but this review focuses on a subset of the most used and promising methods for estimating species trees within a phylogenomic paradigm.

The Bayesian methods, BEST (Bayesian Estimation of Species Trees)[38] and *BEAST,[39] use multiple genetic markers, a model of molecular evolution (*e.g.* general time reversible model for DNA or WAG for protein data), and the multispecies coalescent model[40] to infer a species tree. The user provides sequence data as the input, models of molecular evolution for each loci, standard priors for Bayesian phylogenetic analyses, and species designations. BEST also requires a prior probability value for population size (θ) and a defined outgroup. These methods utilize a Markov chain Monte Carlo (MCMC) algorithm where gene trees and the species tree are updated in each iteration[39,41] The species tree is estimated using the majority rule consensus tree of trees sampled after the algorithm reaches stationarity. BEST estimates branch lengths as the mean of the estimated posterior distribution, and *BEAST calculates branch lengths as either the mean or the median of the estimated posterior distribution depending on user input.[38,39] Support values in terms of clade posterior probabilities are given and both programs estimate divergence times.[38,39]

A likelihood method for species tree estimation that utilizes a coalescent model, STEM (Species Tree Estimation with Maximum Likelihood),[42] uses individual gene trees, calculated from any standard phylogenetic method, as input rather than DNA sequence data. STEM assumes these gene trees are inferred without error, which may be unrealistic. As in BEST and *BEAST, STEM assumes that gene tree discordance is produced only by incomplete lineage sorting. The user must provide θ which is constant throughout the analysis. This is in contrast to BEST where θ is estimated across the Bayesian algorithm based on the prior probability. A caveat of this method is that support values for the species tree are not given, but this could be addressed using bootstrap techniques.

The previously discussed methods assume that incomplete lineage sorting in the coalescent process is the cause of gene tree discordance, but in groups with potentially high levels of horizontal gene transfer this assumption may cause the methods to fail in finding an accurate species tree. The non-parametric Bayesian concordance analysis (BCA), as implemented in BUCKy,[43,44] does not make assumptions for the cause of gene tree incongruence. This method requires a sample of trees from the posterior probability of a standard BI analysis of each locus. The user is required to provide a prior value for gene tree discordance (α). The concordance tree output represents the species tree in that it consists of the clades found with the highest amount of genomic support as measured by concordance factors.[45] Nodal support is determined by a concordance factor which is the percentage of the genes included in the analysis that support the clade.

Additional species tree inference methods that are designed for efficient computational times based on coalescent theory summary statistics include species tree estimation using average ranks of coalescences (STAR;[46]), and species tree estimation using average coalescent times (STEAC;[46]). Liu and Yu[47] developed an additional method that uses average gene-tree internode distances and a neighbor joining algorithm to compute the species tree (NJ_{st}). All three methods were designed to be more computationally efficient than coalescent based methods that infer

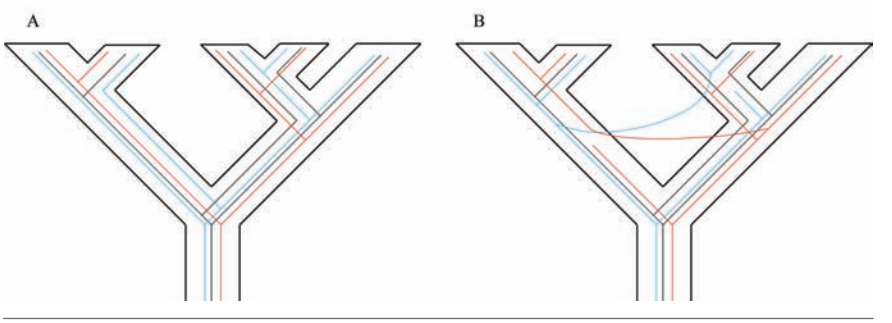

Figure 1. Gene tree vs species tree conflict. The large tree represents the species tree with three individual gene trees (black, red, blue) inside of each species tree. (A) Gene tree and species tree conflict due to incomplete lineage sorting. (B) Gene tree and species tree conflict due to horizontal gene transfer and/or introgression. Gene branches not reaching the tips of the tree represent haplotypes that were deleted from the species.

Table 1. Species tree inference algorithms and select properties of each method.

Method	Theoretical background	Inputs	Outputs	Available program	Ref.
BEST	Bayesian phylogenetics and coalescent theory	DNA sequences or protein sequences; standard Bayesian priors and a population size (θ) prior; defined outgroup	Species tree with branch lengths, nodal support expressed as posterior probabilities, and divergence time estimates	BEST: http://www.stat.osu.edu/~dkp/BEST/downloads/	41
*BEAST	Bayesian phylogenetics and coalescent theory	DNA sequences or protein sequences; standard Bayesian priors	Species tree with branch lengths, nodal support expressed as posterior probabilities, and divergence time estimates	Incorporated into BEAST: http://beast.bio.ed.ac.uk/Main_Page	39
STEM	Maximum likelihood and coalescent theory	Rooted gene tree(s) for each locus; population size (θ)	Species tree with branch lengths but without measures of nodal support	STEM: http://www.stat.osu.edu/~lkubatko/software/STEM/	42
BCA	Bayesian concordance	Trees from the posterior distribution of a standard Bayesian analysis for each locus; prior for gene tree discordance (α)	A concordance and population tree without branch lengths and with nodal support expressed as a concordance factor	BUCKy: http://www.stat.wisc.edu/~ane/bucky/	44
STAR	Coalescent theory	Rooted gene tree(s) for each locus	Species tree without branch lengths and nodal support measured by bootstrap analysis	R module PHYBASE: http://code.google.com/p/phybase/downloads/list	46
STEAC	Coalescent theory	Rooted gene tree(s) for each locus	Species tree without branch lengths and nodal support measured by bootstrap analysis	R module PHYBASE: http://code.google.com/p/phybase/downloads/list	46
NJ_{st}	Neighbor joining distance method	Unrooted gene tree(s) for each locus	Species tree without branch lengths and nodal support measured by bootstrap analysis	R module PHYBASE: http://code.google.com/p/phybase/downloads/list	47
Maximum Tree (MT)	Coalescent theory	Rooted gene tree(s) for each locus	Species tree without branch lengths and nodal support measured by bootstrap analysis	R module PHYBASE: http://code.google.com/p/phybase/downloads/list	48
Maximum Pseudo-likelihood for Estimating Species Trees (MP-EST)	Coalescent theory	Rooted gene trees(s) for each locus	Species tree with branch lengths and nodal support measured by bootstrap analysis	MP-EST: http://code.google.com/p/mp-est/	49
AUGIST	Gene tree uncertainty	Unrooted gene trees for each locus	Consensus species tree without branch lengths and with consensus percentage as nodal support values	Mesquite and the AUGIST module: http://www.lycaenid.org/augist/	50
Shallowest Coalescences (SC)	Coalescent theory	DNA sequences	Species tree without branch lengths or nodal support values	Mesquite and the Coalescence module: http://mesquiteproject.org	51
Minimizing Deep Coalescences (MDC)	Coalescent theory	Unrooted gene tree for each locus	Consensus species tree without branch lengths and with consensus percentage as nodal support values	Mesquite and the Coalescence module: http://mesquiteproject.org	15
Gene tree parsimony	Coalescent theory	Unrooted gene trees for each locus	Species tree without branch lengths or measures of nodal support	iGTP: http://genome.cs.iastate.edu/CBL/iGTP/	52
O'Meara method	Gene tree parsimony and coalescent theory	Gene tree for each locus without *a priori* species designations	Species tree with a posteriori species delimitation and branch lengths, but without measures of nodal support	Brownie: http://www.brianomeara.info/brownie	53

species trees directly from sequence data. STAR and STEAC require rooted gene trees (*i.e.* with a defined outgroup) calculated with any standard phylogenetic method for each locus as the user input, whereas NJ_{st} can use unrooted gene trees as the input. One shortfall of these methods is that branch lengths are not calculated. The outputs for these methods are species trees, and nodal support can be assessed for STAR, STEAC, and NJst using a bootstrap analysis described by Liu *et al.*[46] and Liu and Yu.[47]

O'Meara[53] developed a non-parametric method of species tree inference (referred herein as the O'Meara method) based on gene tree parsimony that does not require species designation to gene sequences *a priori*. Such a method is ideal in that subjectivity of species delimitation is removed. Theoretically, this model attempts to find the species tree that minimizes gene tree conflict on interspecific branches while also minimizing excess genetic structure across the tree.[53] The output tree assigns species designations and has branch lengths, but no clade support values are given.

The methods discussed above are by no means an exhaustive list of all species tree inference methods proposed. (Table 1 shows additional methods.) However, they do represent what appear to be the most promising methods based on recent comparisons using both simulated and empirical data. Many of the above methods have been used in applied inferences of species trees using original data.[3-4,27,36,45,54,55]

Algorithm performance

All of the above methods must grapple with standard issues concerning molecular phylogenetic analyses. Heuristic methods for DNA alignments[56-58] are commonly used, but alignment uncertainty is still possible.[58-59] Talavera and Castresana[60] showed that the removal of ambiguously aligned positions can actually increase phylogenetic signal when sequences are not too short. The accurate alignment of homologous DNA sequence positions is essential to any comparative genomic analysis (see review by Kumar and Filipski[61]) and, therefore, species tree inference. Researchers should be careful to ensure the most accurate alignment possible in any phylogenomic study.

In addition to DNA alignment, the choice of the model of molecular evolution utilized for each locus used in the above methods should be considered (see review by Sullivan and Joyce[62]). Poor model choice can lead to over- or under-parameterization of data[63] and a model testing analysis (*e.g.* jModelTest;[64] or MrModerltest)[65] should be performed for each locus whether gene trees or DNA sequences are the input for any given species tree inference algorithm. Furthermore, when choosing a species tree inference method, researchers should decide whether methods that use the coalescent model,[40] which assumes no horizontal gene transfer or introgression, are appropriate for their data.

Many of the simulation tests that have been carried out to explore the function and utility of traditional Bayesian phylogenetic analysis are likely to apply to BEST, *BEAST, and BCA as they all incorporate traditional Bayesian phylogenetic analysis into their algorithms to certain degrees. Simulation studies have expressed concerns about evolution rate priors[17] and the use of uninformative priors in Bayesian analyses[66-67] and, therefore, the choice of priors in any Bayesian framework should be carefully considered. Furthermore, whether posterior probabilities are a reliable source of clade confidence and if alternative measures would be preferable[68] should be considered whenever estimating species trees under a Bayesian framework.

BEST has been shown to outperform BCA and STEM[36] when the assumptions of the coalescent model are not violated, but not the newer *BEAST method. Liu *et al.*[69] analyzed yeast (*Saccharomyces*) and Manakins (*Manacus*) under the BEST framework. In the yeast analysis, the authors conclude that multiple well supported gene trees may not be enough for a well supported species tree (although the BEST algorithm did give the accepted species tree). BEST also gave a similar species tree to a previously published Manakin phylogeny.[69] This was in the light of probable introgression events in the data set,[69] which is a violation of the assumption of the coalescent model that no gene flow occurs after speciation. An empirical study of Hawaiian flowering plants in the genus *Schiedea* found well supported but different inferred with BEST and traditional concatenation approaches[70] indicating some or all methods were positively misleading compared to the *true* tree. This finding may be a result of reticulate evolution in *Schiedea* and a violation of the coalescent model's primary assumption.[70]

Recent studies suggest that at least under certain conditions, BCA and *BEAST may be better alternatives to BEST. Chung and Ané[45] showed BCA, as implemented by BUCKy[43,44] outperformed BEST when horizontal gene transfer was the primary cause of gene tree discordance, and it seems that this conclusion may be extrapolated for *BEAST and STEM as the presence of horizontal gene transfer violates the primary assumption of the coalescent model. Furthermore, *BEAST was more accurate under simulation than BEST in direct comparisons with simulated data.[39] In a study on orioles (*Icetrus*), Jacobsen and Omland[27] found BEST to be far too computationally demanding for practical use, whereas *BEAST performed well with identical data, although incongruence between *BEAST, BCA and traditional methods was not well explained.

Huang *et al.*[54] found that the accuracy of STEM decreases in data sets of recently diverged species, and this is corroborated by the finding of Leaché and Rannala[36] that, as the number of substitutions per site from the root to the tip of the tree (T) decreases, so does the accuracy of STEM for simulated data. BEST does not appear to suffer from this problem.[36] STEM will also inherently be less accurate if the supplied gene trees are themselves less accurate.[36] Kabatko *et al.*[71] found *BEAST and STEM gave similar relationships of *Sistrurus* rattlesnakes, but that BEST required too much computational time for thorough comparisons to the STEM and *BEAST trees.

Liu *et al.*[46] found STAR outperformed STEAC in most scenarios except when substitution rates between loci did not vary, and Liu and Yu[47] found STAR and NJ_{st} to perform equally as well in estimating the true species tree with simulated data. All three methods have been shown to require many more loci (50-100+;[46,47]) to achieve accurate results than those used in empirical *BEAST, BEST and STEM analyses.[3,24,27,71] No analysis has dealt with how these methods perform when gene tree incongruence is not entirely due to incomplete lineage sorting. Under simulation and in empirical analyses, BEST outperformed both STAR and NJ_{st}[47] Given this, and together with the findings of other studies mentioned above, STAR, STEAC and NJ_{st} are likely only appropriate when other methods are too computationally demanding for the data set of interest.

Although the non-parametric method of O'Meara[53] has not been analyzed to the extent of methods discussed above, it has great potential for genomic environmental sampling since species designations are not required *a priori*. This method was not always accurate when tested with a variety of simulations, and the computational time demanded when species assignments are not fixed may be prohibitive with large datasets.[53] In many cases, this method may not be the most effective since many taxonomists have information, such as localities of specimens and morphology, that make *a priori* designation of species or subspecies appropriate.

Species delimitation with genomic data

All of the methods discussed above are appropriate in a phylogenomic paradigm as they require many unlinked loci, and species trees are integral to the application of a phylogenetic species concept for species delimitation. As phylogenomics continues to shift away from only studies of deep divergence to species level relationships, the use of species tree inference methods, and other methods to be

developed in the future, will enhance species delimitation, which is integral to the study of biodiversity and conservation biology.

Yang and Rannala[5] developed two Bayesian methods implemented in the program bpp for species delimitation. The first method uses a reversible jump MCMC algorithm and the other uses a constant MCMC algorithm which is designed to be more computationally efficient. Both utilize a guide species tree to test species boundaries under the biological species concept. Although a species tree (potentially inferred with any of the above algorithms) must be provided as a guide tree, these methods largely remove subjectivity in species delimitation. SpedeSTEM,[72] an extension of STEM, can also be used to indentify evolutionary distinct lineages with a STEM species tree and, therefore, help species delimitation. Furthermore, the O'Meara species tree inference method[53] does not require species identification *a priori* to tree inference, but rather infers them with species tree estimation. Although computational problems with the O'Meara method are noted above, future development of *a posteriori* species delimitation methods with little subjectivity that also allow for horizontal ggene transfer could become invaluable to microbial field biologists and others interested in speciation of understudied groups.

In poorly studied taxa, such as many invertebrate metazoans, there is a need for BAC library development or shotgun sequencing projects to provide data for phylogenomic analyses. Since they are understudied, these taxa often have the least understood species boundaries and would benefit the most from species tree estimation methods that can help in species delimitation. I argue that major phylogenetic research programs should aim to develop shotgun sequencing libraries for primer development in traditionally understudied taxa so species trees can be accurately resolved and used for species delimitation. Since management decisions are based on recognized species boundaries, there is a need to apply new phylogenomic species tree and species delimitation methods to understudied groups, particularly given the global biodiversity crisis.

Conclusions

Resolving the tree of life is a major goal in evolutionary biology. With high throughput sequencing, the achievment of nearly full and accurate phylogenetic resolution of the tree of life appears possible in the future. However, the advancements of sequencing technology far outpaced the development of species tree algorithms specifically designed for multi-locus data. It is still unclear which species tree inference method is the *best* for inferring the true species tree and this may depend on whether a pure coalescent process is the only, or at least primary cause, of gene tree discordance. The development of a well tested algorithm that allows for both incomplete lineage sorting and horizontal gene transfer will certainly help in determining true species level relationships. Advancements towards removing subjectivity from the species delimitation discussed above are also a step in the right direction. Testing is still needed in areas concerning missing data and how species tree algorithms could handle non-DNA sequence data (*e.g.* proteins and gene order) combined with traditional sequence data.

References

1. Barracloough TG, Volger AP. Detecting the geographical pattern of species from species-level phylogenies. American Naturalist 2000;155:419-34.
2. Wiley EO. Phylogenetics: the theory and practice of phylogenetic systematics. Wiley, 1981, p. 439.
3. Carstens BC, Dewey TA. Species delimitation using a combined coalescent and information-theoretic approach: an example from North American Myotis bats. Systematic Biology 2010;58:400-14.
4. Leaché AD, Fujita MK. Bayesian species delimitation in West African forest geckos. Proceeding of the Royal Society B 2010; 277:3071-7.
5. Yang Z, Rannala B. Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences. 2010;107:9264-9.
6. Kembel SW, Hubbell SP. The phylogenetic structure of a neotropical forest tree community. Ecology 2006;Supplement:S86-S99.
7. Vamosi SM, Heard SB, Vamosi JC, Webb CO. Emerging patterns in the comparative analysis of phylogenetic community structure. Molecular Ecology 2009;18:572-92.
8. Agnarsson I, Kuntner M, May-Collado LJ. Dogs, cats, and kin: a molecular species-level phylogeny of Carnivora. Molecular Phylogenetics and Evolution 2010;54:726-45.
9. Douglas ME, Douglas MR, Schuett GW, et al. Conservation phylogenetics of helodermatid lizards using multiple molecular markers and a supertree approach. Mole-cular Phylogenetics and Evolution 2010;55: 153-67.
10. Perez KE, Minton RL. Practical applications for systematics and taxonomy in North American freshwater gastropod conservation. Journal of the North American Benthological Society 2008;27:471-83.
11. Laurance WF. Have we overstated the tropical biodiversity crisis? Trends in Ecology and Evolution 2007;22:65-70.
12. Felsenstein J. Phylogenies and the comparative method. American Naturalist 1985;125:1-15.
13. Giribet G. A new dimension in combining data? The use of morphology and phylogenomic data in metazoan systematics. Acta Zoologica 2009;91:11-9.
14. Funk DJ, Omland KE. Species-level paraphyly and polyphyly: frequency, causes, and consequences with insights from animal mitochondrial DNA. Annual Review of Ecology, Evolution and Systematics 2003; 34:397-423.
15. Maddison WP. Gene trees in species trees. Systematic Biology. 1997;3:523-36.
16. Edwards SV. Natural selection and phylogenetic analysis. Proceedings of the National Academy of Sciences 2009;106: 8799-800.
17. Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology 2009;58: 130-45.
18. Degnan JH, Rosenberg NA. Discordance of species trees with their most likely gene trees. PLoS Genetics 2006;2:e68.
19. Townsend TM, Mulcahy DG, Noonna BP, et al. Phylogeny of iguanian lizards inferred from 29 nuclear loci, and a comparison of concattenated and species-tree approaches for an ancient, rapid radiation. Molecular Phylogenetics and Evolution 2011;61:263-80.
20. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Computational Biology 2009;5:e1000262.
21. de la Torre-Bárcena JE, Kolokotronis S-O, Lee EK, et al. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. PLoS ONE 2009;4:e5764.
22. Sanderson MJ, McMahon MM, Steel M. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evolutionary Biology 2010;10:1-13.
23. Pollock DD, Zwickl DJ, Jimmy A M, Hillis DM. Increased taxon sampling is advantageous for phylogenetic inference. Systematic Biology 2002;51:664-71.
24. Townsend JP, Lopez-Giraldez F. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Systematic Biology 2010;59:446-57.
25. Dunn CW, Hejnol A, Matus DQ, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 2008;452:745-9.
26. Burleigh JG, Bansal MS, Eulenstein O, et al. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Systematic Biology 2011;60:117-25.
27. Jacobsen F, Omland KE. Species tree infer-

ence in a recent radiation of orioles (Genus Icterus): Multiple markers and methods reveal cytonuclear discordance in the northern oriole group. Molecular Phylogenetics and Evolution 2011;61:460-69.
28. Horvath JE, Weisrock DW, Embry SL, et al. Development and application of a phylogenomic toolkit: resolving the evolutionary history of Madagascar's lemurs. Genome Research. 2008;18:489-99.
29. Knowles LL, Kubatko LS, eds. Estimating species trees: practical and theoretical aspects. Hoboken, New Jersey, Wiley-Blackwell, 2010.
30. Knowles LL. Sampling strategies for species tree estimation. In: LL Knowles, LS Kubatko, editors. Estimating species trees: pracitical and theoretical aspects. Hoboken, New Jersey, Wiley-Blackwell, 2010, p. 163-73.
31. McCormack JE, Huang H, Knowles LL. Maximum likelihood estimates of species trees: how accracy of phylogenetic inference depends upon divergence history and sampling design. Systematic Biology 2009;58:501-8.
32. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. Nature 2007;450: 203-18.
33. Shedlock AM, James DE, Edwards SV. Amniote phylogenomics: testing evolutionary hypotheses with BAC library scanning and targeted clone analysis of large-scale DNA sequences from reptiles. In: WJ Murphy WJ, ed. Methods in molecular biology: phylogenomics. Totowa, NJ, Humana Press Inc., 2008.
34. Thomson RC, Shedlock AM, Edwards SV, Bradeley S. Developing markers for multilocus phylogenetics in non-model organisms: a test case with turtles. Molecular Phylogenetics and Evolution 2008;49:514-25.
35. Parfrey LW, Grant J, Tekle YI, et al. Broadly sampled multigene analyses yield a well-resolved Eukaryotic tree of life. Systematic Biology 2010;59:518-33.
36. Leaché AD, Rannala B. The accuracy of species tree estimation under simulation: a comparison of methods. Systematic Biology 2011;60:126-37.
37. Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biolo-gy 2009;56:17-24.
38. Liu L, Pearl DK. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Systematic Biology 2007;56:504-14.
39. Heled J, Drummond AJ. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 2010;27: 570-80.
40. Rannala B, Yang Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 2003;164: 1645-56.
41. Liu L. BEST: bayesian estimation of species trees under the coalescent model. Bioinformatics 2008;24:2542-3.
42. Kubatko LS, Carstens BC, Knowles LL. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 2009;25:971-3.
43. Larget B, Kotha SK, Dewey CN, Ané C. BUCKy: gene tree / species tree reconciliation with the Bayesian concordance analysis. Bioinformatics. 2010;26:2910-1.
44. Ané C, Larget B, Baum DA, et al. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 2007;24:412-26.
45. Chung Y, Ané C. Comparing two bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. Systematic Biology 2011;60:261-75.
46. Liu L, Yu L, Pearl DK, Edwards SV. Estimating species phylogenies using coalescence times among sequences. Systematic Biology 2009;58:468-77.
47. Liu L, Yu L. Estimating species trees from unrooted gene trees. Systematic Biology 2011;60:661-7.
48. Liu L, Yu L, Pearl DK. Maximum tree: a consistent estimator of the species tree. Journal of Mathematical Biology 2010;60: 96-106.
49. Liu J, Yu L, Edwards SV. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology 2010;10: 302.
50. Oliver JC. AUGIST: inferring species trees while accommodating gene tree uncertainty. Bioinformatics 2008;24:2932-3.
51. Maddison WP, Knowles LL. Inferring phylogeny despite incomplete lineage sorting. Systematic Biology 2006;55:21-30.
52. Chaudhary R, Bansal MS, Wehe A, et al. iGTP: a software package for large-scale gene tree parsimony analysis. BMC Bioninformatics 2010;11:574.
53. O'Meara BC. New heuristic methods for joint species delimitation and species tree inference. Systematic Biology 2010;59:59-73.
54. Huang H, He Q, Kubatko LS, Knowles LL. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Systematic Biology 2010;59:573-83.
55. Keck BP, Near TJ. A young clade repeating an old pattern: diversity in Nothonotus darters (Teleostei:Percidae) endemic to the Cumberland River. Molecular Ecology 2010;19:5030-42.
56. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947-8.
57. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004;32:1792-7.
58. Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 2008;320:1632-5.
59. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science 2008;319:473-6.
60. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 2007;56:564-77.
61. Kumar S, Filipski A. Multiple sequnece alignment: in pursuit of homologous DNA positions. Genome Research 2007;17:127-35.
62. Sullivan J, Joyce P. Model selection in phylogenetics. Annual Review of Ecology, Evolution and Systematics 2005;36:445-66.
63. Kelchner SA, Thomas MA. Model use in phylogenetics: nine key questions. Trends in Ecology and Evolution 20007;22:87-94.
64. Posada D. jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 2008;25:1253-6.
65. Nylander JA. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University, 2004.
66. Pickett KM, Randle CP. Strange bayes indeed: uniform topological priors imply non-uniform clade priors. Molecular Phylogenetics and Evolution 2005;34:203-11.
67. Randle CP, Pickett KM. Are nonuniform clade priors important in Bayesian phylogenetic analysis? A response to Brandley et al. Systematic Biology 2006;55:147-51.
68. Randle CP, Pickett KM. The conflation of ignorance and knowledge in the inference of clade posteriors. Cladistics 2010;26: 550-9.
69. Liu L, Pearl DK, Brumfield R, Edwards SV. Estimating species trees using multiple allele data. Evolution 2008;62:2080-91.
70. Willyard A, Wallace LE, Wagner WL, et al. Estimating the species tree for Hawain Schiedea (Caryophyllaceae) from multiple loci in the presence of reticulate evolution. Molecular Phylogenetics and Evolution 2011;60:29-48.
71. Kubatko LS, Gibbs HL, Bloomquist EW. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in Sistrurus rattlesnakes. Systematic Biology 2011.
72. Ence DD, Carstens BC. SpedeSTEM: a rapid and accurate method for species delimitation. Molecular Ecology Resources 2011;11:473-80.