Chapter 3
An Introduction to Supertree Construction (and Partitioned Phylogenetic Analyses) with a View Toward the Distinction Between Gene Trees and Species Trees

Olaf R. P. Bininda-Emonds

Abstract The dominant approach to the analysis of phylogenomic data is the concatenation of the individual gene data sets into a giant supermatrix that is analyzed en masse. Nevertheless, there remain compelling arguments for a partitioned approach in which individual partitions (usually genes) are instead analyzed separately and the resulting trees are combined to yield the final phylogeny. For instance, it has been argued that this supertree framework, which remains controversial, can better account for natural evolutionary processes like horizontal gene transfer and incomplete lineage sorting that can cause the gene trees, although accurate for the evolutionary history of the genes, to differ from the species tree. In this chapter, I review the different methods of supertree construction (broadly defined), including newer model-based methods based on a multispecies coalescent model. In so doing, I elaborate on some of their strengths and weaknesses relative to one another as well as provide a rough guide to performing a supertree analysis before addressing criticisms of the supertree approach in general. In the end, however, rather than dogmatically advocating supertree construction and partitioned analyses in general, I instead argue that a combined, "global congruence" approach in which data sets are analyzed under both a supermatrix (unpartitioned) and supertree (partitioned) framework represents the best strategy in our attempts to uncover the Tree of Life.

3.1 Introduction

A comparatively recent, but nevertheless fundamental insight within the field of comparative biology was the realization that it could only be done properly within a phylogenetic context (Felsenstein 1985b; see also Chap. 1), thereby disentangling

O. R. P. Bininda-Emonds (✉)
AG Systematics and Evolutionary Biology, IBU—Faculty V,
Carl von Ossietzky Universität Oldenburg, Carl von Ossietzky Strasse 9–11,
26111 Oldenburg, Germany
e-mail: olaf.bininda@uni-oldenburg.de

L. Z. Garamszegi (ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, DOI: 10.1007/978-3-662-43550-2_3,

the similarity between species that arises via natural selection and convergent evolution versus that from their shared evolutionary history ("phylogenetic inertia"; sensu Harvey and Pagel 1991). Applying this form of phylogenetic correction thereby also acts as a statistical "fix" for any effects from other unmeasured variables. More recently, the use of well-resolved phylogenetic trees have helped to provide valuable insights into speciation and extinction rates (including their correlates and variation between and within groups), models of trait evolution, and community phylogenetics, among other fields. The key to performing all such analyses, naturally, is a reliable estimate of the phylogenetic history of the focal taxa, where great strides have been made within the last 25 years due in large part to the molecular revolution.

Despite many claims to the contrary, molecular phylogenetics has generally not uprooted our picture of the Tree of Life (Hillis 1987; Asher and Müller 2012) and many taxa have escaped the molecular revolution fairly unscathed (e.g., mammals or insects as a group). Furthermore, support for many phylogenetic hypotheses supposedly rooted on molecular data can also be found from morphological data. For instance, the molecular hypothesis that whales nest within even-toed ungulates rather than form the sister group to it was actually proposed at least as early as Beddard (1900) based on anatomical evidence (although admittedly largely ignored since then). Moreover, a recent study (Lee and Camens 2009) showed that many morphological data sets also contain substantial HIDDEN SUPPORT (see Box 3.1 for this and all other glossary entries as indicated in small caps) for otherwise conflicting molecular hypotheses of mammal phylogenetic relationships. Nevertheless, what the molecular revolution has unquestioningly provided is a plentiful, universal data source (i.e., DNA sequence data) that is becoming increasingly easy to tap into. Indeed, the advent and cost-effectiveness of next-generation sequencing means that DNA sequence data are often no longer limiting for phylogenetic purposes and are arguably becoming computationally, rather than financially prohibitive! A clear example here is the 1KITE project (http://www.1kite.org), with its goal of obtaining the entire transcriptomes of 1000 insect species covering all known orders, an amount of sequence data that would have been unthinkable a decade ago.

Box 3.1 List of abbreviations

HGT	Horizontal gene transfer
ILS	Incomplete lineage sorting
MLC	Multilocus coalescent model
MRP	Matrix representation with parsimony
OG	Outgroup
STK	Supertree Tool Kit

Accordingly, methodological discussions in molecular phylogenetics have long since shifted from issues of data quantity (e.g., if a limited number of taxa or characters is more detrimental with respect to accuracy; Graybeal 1998) to the best way to analyze the sequence data that are now so abundant. In this regard, the de facto standard is the total evidence or supermatrix approach, in which all the sequence data are concatenated into a single matrix and analyzed en masse. Proponents of this approach have championed it using both philosophical and methodological arguments. In the former case, the principle of "total evidence" is invoked in that the method uses all available data (sensu Kluge 1989). (Theoretically, nonsequence data that can also be accommodated in a matrix format (e.g., morphological characters) can also be included in the analysis; however, this is the exception rather than the rule.) In the latter case, simultaneous analysis facilitates the phenomenon of SIGNAL ENHANCEMENT (sensu de Queiroz et al. 1995)/HIDDEN SUPPORT (sensu Gatesy et al. 1999), whereby the concatenated data set might present a novel solution compared to the individual data partitions through the combination of the latter and effective upweighting of their consistent secondary signals. Importantly, analytical possibilities within a supermatrix framework have also kept pace, with analyses of over 50,000 taxa under a likelihood framework now being possible (e.g., Smith et al. 2011), including the possibility to apply separate models of evolution for each individual partition, even for disparate data types (e.g., DNA, amino-acid, and morphological data) (Stamatakis in press), thereby assuring a more optimal analysis of each data type or partition.

However, even within the possibilities offered by individual Bremer support analyses for each partition (partitioned Bremer support; Baker and DeSalle 1997) to visualize conflict among the different data partitions, the supermatrix approach tends to neglect that different genes often have different evolutionary histories and ones that can differ from that of the species. This fundamental gene tree/species tree conflict has been recognized since at least Maddison (1997) and derives from two main causes. The first problem is that individual genes essentially represent a statistical sample of the entire population (i.e., the genome) and so are subject to normal sampling artifacts. Thus, small genes might not possess a sufficient sample size in terms of the number of base pairs they contain to provide an accurate or well-resolved solution. Compounding this problem is that because DNA consists of only four nucleotides, it is subject to convergent evolution that typically confounds phylogenetic analyses as either SATURATION and/or LONG-BRANCH ATTRACTION (see Bergsten 2005) for fast-evolving genes along long branches. By contrast, extremely short branches, as are typically found in adaptive radiations, are also problematic because of the insufficient time to generate substitutions that provide evidence of the order of speciation events. Indeed, because such substitutions are more likely to derive from fast-evolving sites because of the short-time interval, this evidence is also more likely to disappear with time through subsequent substitutions at the same site and saturation.

Together, the above issues represent the normal, stochastic variation associated with any population estimate, possibly confounded by biases in the method of phylogeny reconstruction (e.g., long-branch attraction). A second, less appreciated

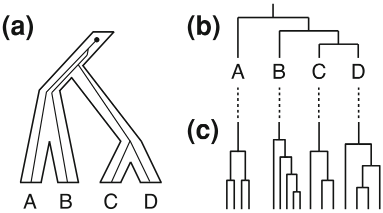


Fig. 3.1 The traditional representation of incomplete lineage sorting (ILS). **a** The species tree is represented by the outline and contains a gene tree (*thin lines*). In this case, the gene tree conflicts with the species tree and gives a wrong estimate of it (**b**). This problem will not disappear with time given that the terminal taxa of (**b**) comprise the common ancestors of those in (**c**). ILS is more prevalent during rapid speciation in large populations, when the time to coalescence for the gene tree is less than that for speciation

problem is that the evolutionary history of a gene can truly differ from that of the species due to any number of natural processes such as horizontal gene transfer (HGT), recombination or incomplete lineage sorting (ILS, also known as deep coalescence; see Fig. 3.1). Although these phenomena were believed originally to be relatively rare and/or confined to otherwise difficult taxa like microbes, there is a growing realization that both HGT and ILS might indeed be both more common and widespread than we had thought previously. Indeed, given the right set of evolutionary conditions (e.g., rapidity of speciation events), the number of gene trees that conflict with the species tree can outnumber the number that agree with it, even for species trees containing as few as five species (Degnan and Rosenberg 2006). Both phenomena are also particularly insidious because they imply that our gene trees are accurate even though they misrepresent the species tree! Again, recent, rapid speciation events are particularly problematic, especially when they occur sequentially (Rosenberg 2013) and in large populations with low rates of genetic drift because the speciation rate exceeds the coalescent rates of the different genes in this so-called anomaly zone (Degnan and Rosenberg 2006), thereby facilitating ILS (Steel and Rodrigo 2008; Edwards 2009). Even more worrisome is that the misleading effects of both HGT and ILS do not necessarily disappear with time. In the case of ILS, because the daughter species arising from the speciation event represent the common ancestors of future higher-level clades (e.g., the "orders" within mammals), a misleading species tree from the past can translate into a misleading ordinal tree in the future (see Fig. 3.1c).

As something of an aside, the same artifacts can potentially arise in the absence of ILS through the process of speciation itself, which can create paraphyletic daughter species. A classic example is the origin of the polar bear (*Ursus maritimus*). Although recent studies based on nDNA markers now indicate it to be an ancient lineage forming the sister group to brown bears (*Ursus arctos*) (Hailer et al. 2012; Miller et al. 2012), it was believed until very recently that this species, based on studies of mtDNA, arose from isolated populations of brown bear from the Admiralty and Baranof Islands of the Alexander Archipelago of southeastern

Alaska about 150,000 years ago (e.g., Lindqvist et al. 2010). Were the latter scenario indeed true (which is undoubtedly generally the case for many other species), then some individuals/populations of brown bear are more closely related to polar bears than to other members of their own species, which, depending on the pattern of future speciation and extinction in the brown-bear lineage, could give rise to ILD-like knock-on effects.

As such, there have been recent attempts to move away from a pure supermatrix approach to ones that can potentially better accommodate such instances of "gene-tree heterogeneity" (sensu Edwards 2009) by focusing on the gene tree as the fundamental unit of analysis rather than individual nucleotides or amino acids. In so doing, it is recognized that although gene trees and species trees are closely related evolutionarily, they nevertheless derive from distinct evolutionary processes (Liu et al. 2010). In essence, these arguments are merely the latest thoughts in a long-standing debate as to whether it is more desirable to automatically combine data or to perform some form of partitioned analysis (see Chippindale and Wiens 1994).

Against this backdrop, the goal of this chapter is to outline and describe two such frameworks for partitioned analyses: the now "traditional" supertree approach and the more recent multilocus coalescent (MLC) model that explicitly builds on coalescent and population genetic theory to derive a species tree from a set of potentially conflicting gene trees. Both approaches are united in having their analytical focus at the level of a set of input trees and, although this was never advanced originally as a justification for traditional supertrees, thereby possess the potential to account for any gene-tree heterogeneity. More controversially, both for expediency and because of the unquestionably strong parallels between the two frameworks, I will refer to them collectively as "supertrees".

This chapter is structured as follows. First, I initially provide a short historical perspective of the supertree framework before providing a summary of both traditional supertree methods and the newer methods based on the MLC model (both summarized in Table 3.1). In so doing, I hope to show the similarities between these two "classes" of methods as well as to point out that MLC-based methods are not the only supertree methods to include an explicit evolutionary model. Second, I briefly address previous criticisms of the supertree framework, especially in relation to the supermatrix framework. However, here and throughout, apart from general comparisons to the supertree framework, I will largely refrain from discussing details of the mechanics of a supermatrix analysis given the overwhelming prevalence of (and therefore likely familiarity with) this technique. An excellent summary of general phylogenetic tree building, which forms the backbone of the supermatrix framework (as well as the derivation of individual gene trees), is also provided in Chap. 2 of the volume. Finally, I provide a rough guide as to how to perform a supertree/partitioned data analysis. Given the vast array of supertree methods available, this guide is purposely agnostic in the sense that it does not advocate any one method, but concentrates instead on the various issues that must be considered at each step in the process.

Table 3.1 A summary of the supertree methods listed in the main text, including some brief notes on their properties (where known) and implementations

Method	Notes	Implementation
Average consensus	A method that can explicitly account for branch-length information among the set of source trees when constructing the supertree.	CLANN (http://bioinf.nuim.ie/clann/) (Creevey and McInerney 2005)
Gene-tree parsimony	Derives the supertree based on the method of reconciled trees and therefore explicitly accounts for biological processes like gene duplication and loss.	DupTree (http://genome.cs.iastate.edu/CBL/DupTree/) (Wehe et al. 2008) iGTP (http://genome.cs.iastate.edu/CBL/iGTP/) (Chaudhary et al. 2010)
Matrix representation	A class of methods whereby the topology of a source tree is encoded as a matrix using additive binary coding. The resulting matrix derived from all source trees can be analyzed using virtually any optimization criterion (e.g., MP, NJ, ML, BI, or compatibility).	Among others: CLANN Rainbow (http://genome.cs.iastate.edu/CBL/download/) (Chen et al. 2004) SuperMRP.pl (http://www.uni-oldenburg.de/ibu/systematik-evolutionsbiologie/programme/) (Bininda-Emonds et al. 2005) SuperTree (http://www2.unil.ch/phylo/bioinformatics/supertree) (Salamin et al. 2002) Supertree Tool Kit (http://sourceforge.net/projects/stck) (Davis and Hill 2010)
MRF (matrix representation with flipping)	A matrix representation method with a specific implementation. The optimization criterion of “flipping” involves finding the fewest number of 0 → 1 or 1 → 0 changes in the matrix required to produce a matrix without conflict.	HeuristicMRF2 (http://genome.cs.iastate.edu/CBL/download/); Rainbow
MinCutSupertree	Fast, polynomial time method that preserves nestings among the source trees and where the supertree displays each source tree. Because it preserves nestings and not clades, it is akin to Adams consensus and the supertree cannot necessarily be interpreted phylogenetically.	Supertree (http://darwin.zoology.gla.ac.uk/%7Erpage/supertree) (Page 2002)
Modified MinCutSupertree	Modification to the previous method to achieve greater resolution.	Supertree Rainbow

(continued)

Table 3.1 (continued)

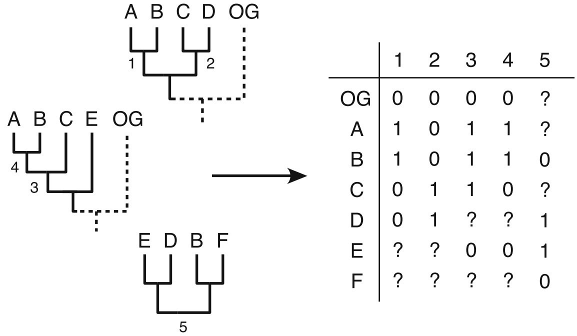
Method	Notes	Implementation
MultiLevelSupertree	Accounts for both horizontal and vertical overlap among the set of source trees and so can account for nested taxa among the set of source trees.	MLS (http://www.atgc-montpellier.fr/supertree/mls/) (Berry et al. 2013)
Multilocus coalescent models	A class of methods that rely on coalescent theory to obtain the supertree from the set of gene trees and so explicitly account for biological processes such as ILS. Some methods can account for branch-length information within the gene trees.	MP-EST (https://code.google.com/p/mp-est/) (Liu et al. 2010) Phybase (https://code.google.com/p/phybase/) (Liu and Yu 2010)
PhySIC and PhySIC_IST	Derive a supertree showing relationships that do not contradict any of those on the source trees and are induced by them (in both cases, singly or jointly).	Webservers available at: http://www.atgc-montpellier.fr/physic/ http://www.atgc-montpellier.fr/physic_ist/ (Ranwez et al. 2007; Scornavacca et al. 2008)
Quartet puzzling	Examines all possible quartets to determine the most likely one for any set of four taxa. The latter are then essentially combined into the final tree. Used for the analysis of DNA data, but the principle could be applied in a supertree framework.	TREE-PUZZLE (http://www.tree-puzzle.de) (Schmidt et al. 2002)
Quartet supertrees	Breaks source trees down into all their possible quartets and builds a supertree from the latter based on their frequencies across the set of source trees.	Quartet suite (http://genome.cs.iastate.edu/CBL/download/) (Piaggio-Talice et al. 2004)
SuperFine	A meta-method that can theoretically be used to speed up (“boost”) any existing supertree method.	SuperFine (http://www.cs.utexas.edu/∼phylo/software/superfine/) (Swenson et al. 2012)

All would be used in Step 3 of the guide presented in Sect. 3.4.3 of the text.
Note that many of the implementations have not been recently updated and so might not run on the latest operating systems

3.2 The Supertree Framework

Supertrees are essentially as old as systematics itself, where our vision of the Tree of Life as a whole was essentially patched together from many smaller subtrees, often using a form of taxonomic substitution. In this, the terminal taxa in a higher-level tree were simply substituted for the nested tree showing the relationships within that taxon. Thus, for a tree of the vertebrate classes, the taxon Mammalia could be replaced by an ordinal-level tree of this group, for example, and so on. Although this technique is still in use today to provide us with some picture of the Tree of Life as a whole, it is distinctly limited in that it requires us to choose some "best" tree at each level and so make a subjective judgment among the many, possibly conflicting options.

A more objective foundation for supertrees essentially dates to 1986, when the mathematician Allan Gordon proposed a generalization of the well-known strict consensus method that could be applied to a set of trees that differed in the terminal taxa they contained (Gordon 1986). For various largely methodological reasons, the solution was largely unworkable and/or ignored (see Bininda-Emonds 2004b), and it was only in 1992 that the next breakthrough was achieved by Baum (1992) and Ragan (1992), who independently described the method now known as matrix representation with parsimony (MRP). Building on the one-to-one correspondence between a tree and its binary equivalent in matrix form ("matrix representation"; Ponstein 1966) (see Fig. 3.2), Baum and Ragan each hit upon the idea of concatenating the individual matrix representations of a set of source trees



	1	2	3	4	5
OG	0	0	0	0	?
A	1	0	1	1	?
B	1	0	1	1	0
C	0	1	1	0	?
D	0	1	?	?	1
E	?	?	0	0	1
F	?	?	?	?	0

Fig. 3.2 Matrix representation of a set of three gene trees. Using additive binary coding, the nodes of any given gene tree can be represented in matrix format in turn. For a focal node, terminal taxa that are descended from that node receive *1*, taxa that are not but are present on the gene tree receive *0*, and all other taxa receive? (e.g., character *2* which represents node *2*). To root the analysis, a fictitious outgroup (OG) comprising all *0*s is added to the base of each gene tree. If a distinction is made between rooted and unrooted gene trees, the OG can also receives? for unrooted source trees (character *5*; see Bininda-Emonds et al. 2005). For any single tree, there is a one-to-one correspondence between it and its matrix representation. To derive the supertree that best fits to the set of gene trees, the entire matrix is then analyzed using any desired optimization criterion (typically parsimony) after which the OG is subsequently removed

into a single (super)matrix and then analyzing the latter with parsimony to derive a "supertree". The potential of the method was quickly realized by Purvis (1995a), who combined numerous estimates of primate phylogeny taken from the literature to derive the first, complete species-level evolutionary tree for the group based on an objective, robust methodology. From there, the field exploded, both in terms of the supertrees themselves that were being generated as well as the supertree methods used to obtain them. A now highly outdated list on both counts can be found in Bininda-Emonds (2004a) and many new trees and methods have been developed since then. It nevertheless remains that MRP is by far and away the most popular of the supertree methods.

3.2.1 Traditional Supertrees

With the growing number of methods (see Table 3.1), supertrees are becoming increasingly difficult to summarize meaningfully as a group as well as to categorize with respect to their methodologies. The supertree framework has historically been seen as a generalization of that for consensus trees, which requires identical taxon sets among the source trees. However, the analogy only holds so far in that most supertree methods do not have clear consensus equivalents (including MRP) and many popular consensus methods did not have a corresponding supertree one until comparatively recently (e.g., majority-rule supertrees; Cotton and Wilkinson 2007). In the end, perhaps, a supertree is now best summarized as the summary tree derived from a set of source trees that need not have identical taxon sets. Under this definition, supertrees remain a generalization of consensus trees, but can extend beyond this as well. An alternative, but not mutually exclusive, interpretation is that the summary tree obtained from a supertree analysis represents the "best fit" to the set of source trees according to some objective function (Thorley and Wilkinson 2003; Bruen and Bryant 2008). In most cases (including MRP), this objective function is unknown, but some supertree methods have been developed explicitly with an objective function in mind, including majority-rule (minimizes partition metric; Cotton and Wilkinson 2007) and maximum-likelihood supertrees (minimizes error function among source trees; Steel and Rodrigo 2008) as well as MinCutSupertree (minimizes sum of triplet distances; Wilkinson et al. 2004).

A clear subcategory of supertrees are those that, like MRP, rely on an explicit intermediate step of building a matrix of pseudo-characters, with each pseudo-character representing a node on a particular source tree. In a sense, the combined matrix functions as a table of the bipartition frequencies among the set of source trees, where a bipartition splits an unrooted tree into two taxon sets (e.g., for the bipartition AB|CD, removing a branch on the tree will result in two subtrees, one with taxa A and B and the other with taxa C and D). The matrix can then be analyzed using any preferred optimization criterion. Although parsimony remains by far the method of choice here (as in MRP), other suggested methods include compatibility (Ross and Rodrigo 2004), flipping (Chen et al. 2003), Bayesian inference of the bipartition frequencies (Ronquist et al. 2004), or, most recently, maximum likelihood with a two-state/parsimony character model (Nguyen et al. 2012). A variant on these methods is the average consensus method (Lapointe and Cucumel 1997), whereby the matrix to be analyzed consists of the sum of the path-length distances between all pairs of taxa among a set of gene trees, with some estimate of these distances for pairs of taxa that do not co-occur on any single tree (Lapointe and Levasseur 2004).

A long-standing critique of traditional supertree methods is that, although arguably reasonably accurate empirically and in simulation, most are not based on any explicit model of biological evolution (Liu et al. 2010) and so are better classed as nonparametric methods (Liu et al. 2009a) and/or the properties of most remain uncharacterized (see Wilkinson et al. 2004). Indeed, MRP represents the poster child that has attracted the most attention in this regard. Despite its long-standing popularity and use, the objective function of MRP remains unknown and the little that is known about its properties is worrisome. For example, it was known almost from the outset that MRP, like many other supertree methods, gives more weight to larger source trees (i.e., is not "sizeless"; Purvis 1995b; Wilkinson et al. 2004) (although it actually favors larger subtrees rather than trees as a whole; see Bininda-Emonds and Bryant 1998); other potentially undesirable properties are summarized in Wilkinson et al. (2004). Nevertheless, the fact that MRP shows reasonable accuracy in practice and can even outperform equivalent supermatrix analyses in simulation (Bininda-Emonds and Sanderson 2001) show that its deficiencies are either not that severe and/or only arise in extreme cases.

That being said, a few supertree methods have been designed explicitly to fulfill some properties identified by Steel et al. (2000) and Wilkinson et al. (2004) as being desirable when combining trees ("desiderata"; sensu Wilkinson et al. 2004). In some ways, this can be viewed as part of the objective function of these methods. For instance, MinCutSupertree (Semple and Steel 2000) and its derivative modified MinCutSupertree (Page 2002) output supertrees that can be found in polynomial time, preserve nestings and binary trees found among all source trees, display all input trees if the latter are compatible, and are independent of the input order of the trees (Semple and Steel 2000). However, by preserving nestings, rather than clades, among the sets of source trees (Semple and Steel 2000), the MinCutSupertree tends to resemble the Adams consensus (Adams 1972, 1986) of the source trees, meaning that the result cannot always be interpreted phylogenetically. For instance, MinCutSupertree will only preserve the nesting information that A and B are a part of a larger cluster ABCD, without any statement as to the relationship between A and B themselves. Thus, even if A and B form sister taxa in the resulting supertree it cannot automatically be assumed that they do indeed form a clade (because MinCutSupertree does not preserve clades).

Other examples of supertree methods designed to meet certain properties a priori are PhySIC (Ranwez et al. 2007) and PhySIC_IST (Scornavacca et al. 2008), which ensure that the resulting supertree displays all relationships that are induced by and are not contradicted by the set of source trees, either alone or in combination. The latter method builds on the former by removing highly conflicting source trees in the hopes of obtaining a better resolved supertree. Together, both methods are perhaps a direct answer to methods like MRP, which theoretically can output relationships that are contradicted by every source tree (see Bininda-Emonds and Bryant 1998) although this appears to be extremely rare in practice (Bininda-Emonds 2003).

3.2.2 Multilocus Coalescent "Supertrees"

A more recent advance toward supertree methods based on evolutionarily sound models—and on the potential distinction between gene trees and the species tree in particular—is the MLC model, which builds theoretically on Rannala and Yang's (2003) characterization of the likelihood function of the species tree under a multispecies coalescent via two probability distributions (Liu et al. 2009a). The first, $f(\mathbf{D}|\mathbf{G})$, describes the probability of deriving a particular gene tree ($\mathbf{G}$) given a set of sequence data ($\mathbf{D}$) and represents the same likelihood function used routinely in molecular phylogenetics. The second, $f(\mathbf{G}|S)$, describes the probability of observing a gene tree given a particular species tree (S) and derives from the multispecies coalescent. Essentially, for a species tree with well-defined clades separated by long branches (i.e., divergence times), the majority of gene trees will resemble the species tree and gene-tree heterogeneity will be low. However, when the species tree contains one or more regions with short branch lengths, the probability for gene-tree heterogeneity in these anomaly zones increases and many more, different gene trees are expected.

Practical implementations of the MLC model, however, are more indirectly related to these probability distributions. Indeed, at least one procedure has been termed as a "maximum pseudo-likelihood approach" by its authors (Liu et al. 2010). Given a set of gene trees (essentially component $f(\mathbf{D}|\mathbf{G})$ from above), one form of the MLC model derives a distance matrix between all pairs of terminal taxa based on their coalescent events across the set of gene trees. For any given cell, the distance value is given either by (1) the minimum number of ranks (nodes) across the set of gene trees until the taxa share a common ancestor/coalesce (GLASS distance; Mossel and Roch 2007), (2) the average number of ranks until they do so (STAR distance; Liu et al. 2009b) or (3) the average coalescence time (STEAC distance; Liu et al. 2009b). Thus, whereas the first two distances only account for topological information within the gene trees (and are therefore only "partially parametric"; Song et al. 2012: 14943), the last can incorporate branch-length information directly when it is present. Finally, the distance matrix is analyzed via a distance method like NJ to derive the species tree (component $f(\mathbf{G}|S)$ from above). By contrast, a second implementation of the MLC model, MP-EST (Liu et al. 2010) derives the frequencies of all triplets of taxa from the set of gene trees (together with path-length information) to obtain the topology and branch lengths of the species tree in a pseudo-likelihood framework, again representing a "partially parametric" method. Both sets of methods appear to perform well under conditions of gene-tree heterogeneity where equivalent supermatrix analyses become statistically inconsistent (Wu et al. 2013).

Although it has not been recognized to date, the two implementations of the MLC model have clear connections to existing traditional supertree methods. For example, the distance-based MLC methods bear strong resemblances with the average consensus supertree method in that both explicitly incorporate branch-length information from the gene trees, even if only indirectly in the form of ranks. Likewise, MP-EST shows similarities with quartet puzzling (Strimmer and von Haeseler 1996) or quartet supertrees (Piaggio-Talice et al. 2004), albeit with MP-EST requiring rooted gene trees (and hence using triplets) instead of the unrooted framework (and thus quartets) employed by the latter two methods. (Quartet puzzling also proceeds directly from the DNA sequence data without explicit regard to gene trees. However, it could be modified from this supermatrix format to work in a gene-tree context.) More generally, the explicit use of an underlying biological model also characterizes the gene-tree parsimony method (Cotton and Page 2004), which has been recognized as a supertree method and uses reconciled trees (Goodman et al. 1979; Page 1994) to account for possible discrepancies between the gene trees and the species tree as a result of processes including HGT and gene duplication and loss.

Nevertheless, by being explicitly couched within a coalescent framework and building on the likelihood function of Rannala and Yang (2003), the MLC methods differ from most other supertree methods in being based on explicit biological models and phenomena. For instance, the MLC model assumes, among other things (see Liu et al. 2009a), constant population sizes through time, random mating, no gene flow, and no HGT, and thus the predominance of ILS as the cause of gene-tree heterogeneity. Although many of these assumptions are unrealistic, the MLC methods are apparently robust to minor rates of HGT and could, in theory, be easily expanded to account for both this and gene flow (Liu et al. 2009a).

MLC-based methods also possess a distinct advantage in that they are very fast compared to most other supertree methods (except for polynomial-time methods like MinCutSupertree) once the input trees have been calculated. As shown by Liu et al. (2010), runtimes are on the order of seconds for problem sizes of 80 gene trees each comprising 20 taxa, both for STAR-based analyses as well as those using ML-EST, albeit with the latter being demonstrably slower. The speed accrues either from the use of NJ as an optimization criterion or the pseudo-likelihood framework compared to the NP-complete algorithms (e.g., parsimony or likelihood) typically used by traditional supertree methods. However, even in the latter case, tremendous speed gains have been achieved by implementing supertrees in a divide-and-conquer framework, in which the supertrees represent more of a search strategy than the end product of the analysis (see Bininda-Emonds 2010). Here, the general idea is to take a large, computational demanding problem (e.g., a large multigene data set of thousands of taxa) and to break it down into many smaller, overlapping data sets that are more tractable because of their small size. The resulting trees from the latter data sets are then combined as a supertree, which can then be further resolved on the basis of the entire data set (Roshan et al. 2004). This general strategy, which also underlies quartet puzzling, has most recently been implemented in SuperFine (Swenson et al. 2012), a so-called meta-method designed to boost the speed of existing supertree methods like MRP. Indeed, the method does appear to deliver more optimal supertrees in a reduced amount of time compared to nonboosted analyses (Swenson et al. 2012; Nguyen et al. 2012), but still at best only on a par in terms of speed and accuracy with equivalent supermatrix analyses (Swenson et al. 2012; Nguyen et al. 2012). In this, the problem with the divide-and-conquer approach appears to lie with the final resolving step, which is based on the full data set and is therefore subject to the same size-based tractability problems (Bininda-Emonds 2010).

3.2.3 Accounting for Vertical Taxonomic Overlap

A feature shared by all the above methods is that they essentially only account for horizontal overlap among the gene trees (i.e., among the terminal taxa). As such, the terminal taxa must all occur at the same taxonomic level (e.g., species in the case of gene trees) or minimally cannot be nested within one another. Thus, the case where a source tree possessed the terminal taxon Mammalia and another possessed *Homo sapiens* would result in a supertree where these two taxa would, at best, be sister groups, despite the latter clearly nesting within the former. Recalling to some degree the process of taxonomic substitution characterizing informal supertree methods, MultiLevelSupertree (Berry et al. 2013) is able to simultaneously account for both horizontal and vertical overlap among the source trees, the latter representing the nested, higher-level relationships implicit among the set of source trees. Moreover, the program is also able to infer the latter from information among the source trees themselves, such that it is not necessary to provide a reference taxonomy providing the nested sets of relationships. Although MultiLevelSupertree would appear to be of use when combining source trees out of the literature, this traditional use of supertrees is rapidly falling by the wayside and it is not clear if its ability to also accommodate vertical overlap will provide any benefit for gene trees based on DNA sequence data, which normally all have species as terminal taxa.

3.3 Criticisms of Supertrees

Even when couched within the context of explicitly accommodating gene-tree heterogeneity, the supertree framework has been highly criticized and remains controversial (e.g., see the exchange between Gatesy and Springer (2013) and Wu et al. (2013) for MLC-based methods). The primary areas of criticism include (1) the potential for duplication of data between the source partitions, (2) the black-box nature of most supertree methods and MRP in particular, and (3) the fact that the methods are a form of meta-analysis and thus one step removed from the primary character data.

Data duplication does indeed represent a potential problem area within a supertree framework as was elegantly shown by Gatesy et al. (2002) for the supertree analysis of mammalian families by Liu et al. (2001). For instance, the same genes (if not the same sequences) are often used for separate phylogenetic analyses, often in combination with other genes. A cogent example here is cytochrome b, which represents by far the most widely sampled gene for mammals to date and one that is often used for phylogenetic analyses within the group. As such, it often comprises part of the data set underlying different phylogenetic trees for mammals, meaning that these trees are nonindependent of one another. Thus, constructing a supertree for mammals by simply collecting and combining all published trees for the group means that cytochrome b would have an unduly greater influence on the end result compared to other genes and sources of character data.

Indeed, many early supertree studies ran afoul of this problem before it was so forcefully pointed out by Gatesy et al. (2002). Fortunately, data duplication is a largely historical problem that can be mitigated today by more careful selection of the source trees and/or by complicated weighting schemes designed to address it (e.g., Nyakatura and Bininda-Emonds 2012). More generally, this criticism is largely obsolete when supertrees are used in an explicit gene-tree framework, where each gene tree is present only once within the data set. Even so, it should be remembered that even the subdivision of the genome into individual genes is to some extent subjective, with our concept of "genes" having become increasingly blurred with increased knowledge of the tremendous degree of complexity underlying the genome (e.g., via recombination, exon shuffling, HGT, and alternative splicing, among other processes). Instead, of note here are newly developed methods like PartitionFinder (Lanfear et al. 2012), which use data-driven, information-theoretic metrics to more objectively reveal partitions within a data set (within the bounds of a set of a priori user-defined partitions). However, it remains to be seen how well these partitions match up with those expected under a gene-tree heterogeneity scenario largely driven by ILS (i.e., classic gene trees). The finding that individual genes are composed of several partitions (e.g., according to codon position in protein-coding genes or stems vs. loops in rDNA genes) would not be problematic, but instead serve to improve our estimate of the individual gene trees. By contrast, the sharing of partitions across genes might force us to rethink our notion of gene trees entirely.

The remaining two criticisms of supertrees are to some degree linked and mirror that of Liu et al. (2010) in claiming that traditional supertree methods do not resolve conflict among the source trees with respect to explicit evolutionary events (Gatesy and Springer 2004). However, this is no longer the case and several supertree methods, such as gene-tree parsimony and the MLC-based methods, now exist that fulfill this criterion. It is important to remember, however, that the supermatrix and supertree methods do operate at different hierarchical levels (DNA sequence data vs. gene trees, respectively; Bininda-Emonds 2004c) such that each will be accommodating different sets of evolutionary events (e.g., character-state transformations vs. HGT or ILS, respectively). Moreover, through their focus at the level of the gene tree, only methods like gene-tree parsimony and the MLC-based methods have the potential to account for processes like ILS, which, when frequent enough, have been demonstrated in simulation to impact on the accuracy of supermatrix methods to the point of them being statistically inconsistent (Wu et al. 2013).

3.4 A Primer to Supertree Construction

The following represents a rough guide to the process of creating a supertree and is also illustrated in Fig. 3.3. It takes its form both from my own experiences and from their formalization and extension in the excellent Supertree Tool Kit (STK) of Davis and Hill (2010). Given the huge variety of supertree methods and choices available, the guide is not intended to be either exhaustive or dogmatic. Other, often unnamed, variations on this framework are conceivable and should be explored and not excluded a priori. A simple, worked example to be used as a jumping-off point can be found in the OPM.

3.4.1 Step 1: Obtaining the Source Trees

Much of the previous discussion has centered on the concept of gene trees, with the implication that they have been obtained directly via phylogenetic analysis of primary molecular sequence data by the researcher. These data can derive either from de novo sequences generated by the researcher and/or from online resources such as GenBank. Indeed, in the latter case, numerous phylogenetic pipelines now exist (see Bininda-Emonds 2011) for the express purpose of mining GenBank and other similar resources for homologous sequence data.

However, gene trees represent only one source of data potentially available under a supertree framework. Because the raw data of a supertree analysis is a phylogenetic tree, any statement of phylogenetic relationship that can be expressed as a bifurcating tree can be included in the analysis. It was this very principle that underlay the earliest empirical supertree studies in which source trees were mined from the literature and either encoded directly in matrix format or as nexus-formatted tree statements for later processing. The online archiving of phylogenetic trees through resources like TreeBASE (www.treebase.org; Sanderson et al. 1994) merely represents the modern and more convenient extension of the traditional paper-based sifting of the literature. Although the inclusion of literature data is quickly falling out of favor in the era of molecular phylogenetics, it remains that it provides access to more of the global phylogenetic database and data that would be otherwise difficult to include in a supermatrix framework. The latter includes not only older molecular data such as DNA–DNA hybridization or isozymes,

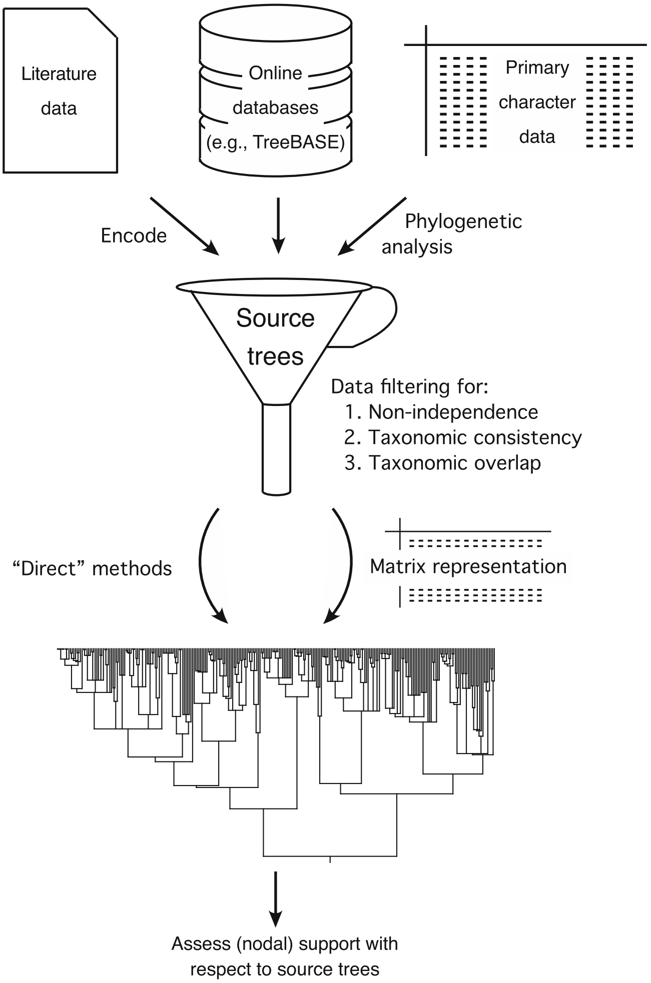


Fig. 3.3 Flow diagram illustrating the general framework of a supertree analysis. Particularly crucial is the middle, filtering step, which acts as a measure of quality control for the source trees derived from any or all of the literature, online databases, or primary character data. Thereafter, any supertree method of choice can be applied to the filtered set of source trees. Adapted from Bininda-Emonds et al. (2004) and Davis and Hill (2010)

but also morphology and evidence from rare genomic changes (Rokas and Holland 2000), the signals of both of which threaten to be swamped by the much more numerous molecular sequence data. Thus, in some respects, a supertree framework can better accommodate the principle of total evidence (i.e., using as much data as possible) than can a supermatrix framework (Bininda-Emonds et al. 2003). That being said, the inclusion of literature data does harbor particular difficulties that are addressed in the next step of the process.

3.4.2 *Step 2: Filtering the Source Trees for Data-Quality Assurance*

This crucial step was inspired initially by the paper of Gatesy et al. (2002), who elegantly documented several weaknesses with respect to data quality in several previously published empirical supertree studies (see above). Although Gatesy et al. (2002) took extended aim at the supertree framework in general, it remains that their paper is essentially about quality assurance in phylogenetic analysis in general and not for any method or framework in particular.

That being said, a supertree framework, especially one that incorporates literature data can present special problems in this regard, again because of the disconnect between the primary character data and the source trees that provide the raw data for the supertree analysis. Because the latter are often mute with respect to the former, a greater potential for the duplication at the level of the primary character data exists within a supertree framework (see above). However, as pointed out above, careful diligence, perhaps in combination with complicated weighting schemes to account for any data duplication, will often be sufficient to ameliorate this potential problem. Within a pure gene-tree framework, this problem is unlikely to occur or, at least, will have the same impact as on the equivalent supermatrix studies that could be performed on the fundamental data set. The issue of data quality, and which source trees to actually include in the analysis, is more thorny with arguably no correct answer. Whereas some investigators will be comfortable including taxonomies as source trees (ignoring the fact that it might be based in part on data also used in other source trees and so represent a case of data duplication), others will reject this possibility categorically. As with any scientific study, it is important in such cases to be open and to make the data available for other researchers to replicate the study under their preferred set of conditions. A sensitivity analysis can also be envisaged, whereby source trees of arguably lower quality are either downweighted or removed from the analysis to ascertain what their impact on the supertree topology is.

In a subsequent step, it is important to ensure consistency among the taxonomic labels among the set of source trees, especially for those methods that only account for horizontal overlap among the terminal taxa. Although MultiLevelSupertree, by also accounting for vertical overlap among the source trees, can avoid problems of nested taxa, a check for taxonomic consistency is necessary here as well to ensure that the same taxa are not present as different synonyms (e.g., Mammalia vs. just "mammals") in different source trees. The issue of taxonomic consistency was first raised by Bininda-Emonds et al. (2004), who also present different, general solutions to the overall problem, which can be implemented either through synonoTree.pl (Bininda-Emonds et al. 2004) or the STK (Davis and Hill 2010). Finally, although this general problem will be more rare in a pure gene-tree context, it can also be relevant here (e.g., GenBank often indexes sequence information separately for a species and its subspecies).

A final check, and one that is neatly implemented in the STK, is to assess the degree of taxonomic overlap among the set of source trees. Minimally, a supertree analysis requires that each source tree overlaps with at least one other by two terminal taxa. (This requirement is loosened through the vertical overlap recognized by MultiLevelSupertree.) Where such overlap does not occur, the resulting supertree should be completely unresolved (within the bounds of the optimization criterion used) because the taxa from one nonoverlapping group can cluster equally optimally with those from another group. This problem is easily ameliorated by either removing nonoverlapping source trees (or running separate supertree analyses for nonoverlapping sets of trees) or by including a "seed tree" (sensu Bininda-Emonds and Sanderson 2001) that contains most if not all the taxa among the set of source trees and so provides a scaffold for the analysis. The seed tree is often derived from a taxonomy, with the poor resolution of such entities meaning that the seed tree provides minimal clustering information of its own. The impact of the seed tree, the use of which is controversial, can be minimized further by downweighting it within the analysis compared to the other primary source trees.

3.4.3 Step 3: Obtaining the Supertree

This represents the most obvious and direct of the four steps in performing a supertree analysis. However, as the previous Sect. 3.2 makes clear, the sheer and still growing variety of supertree methods (see also Table 3.1) can make the selection of the final method difficult. MRP remains by far the method of choice; however, this seems to obtain more for historical considerations rather than the method being demonstrably superior to any alternatives. Therefore, perhaps merely for reasons of comparability with other supertree studies, an MRP analysis is to be recommended. Nevertheless, other methods should also be explored, either because of their arguably better accuracy and/or because of their more desirable properties or objective functions.

It is in this third step where weighting is employed, not only to account for potential data duplication, but also for potential differences in the robustness/ quality of the different source trees. (Early attempts to employ weighting to counteract the apparent size bias of MRP (Ronquist 1996) were ultimately unsuccessful because MRP does not favor larger source trees per se, but overlapping parts of those source trees (Bininda-Emonds and Bryant 1998), making any weighting scheme impossible to implement with large numbers of source trees with different degrees of overlap among them.) A simple solution here is to simply replicate entire source trees proportional to some measure of their inferred quality.

When weighting for source-tree "quality", however, it is important to recognize that phylogenetic relationships within any given tree can also differ in support, with some clades being comparatively better supported than others. In a default supertree analysis, where only the topology of the source trees is used, this information is completely lost, an early criticism of the supertree framework as a meta-analysis (e.g., Gatesy and Springer 2004; but see above). A simple solution in this regard for matrix representation-based methods at least is to weight each pseudo-character by the inferred support for the node in the source tree that it encodes [e.g., according to its nonparametric bootstrap frequency (Felsenstein 1985a) or Bremer support (Bremer 1988)]. Indeed, although this form of weighting still cannot account for hidden support among the primary data partitions, simulation studies have shown that doing so improves the accuracy of MRP supertrees, often to the point where the supertree analysis slightly outperforms an equivalent supermatrix analysis (Bininda-Emonds and Sanderson 2001). Important here, however, is to ensure that the weighting schemes are comparable among the source trees (e.g., not a combination of bootstrap frequencies and Bremer support values); however, this should not be a problem for gene trees generated de novo from public databases like GenBank. For other supertree methods where there is no direct connection with the individual clades on a given tree, some form of clade duplication proportional to inferred support can also be envisaged.

Finally, it is important to realize that because all supertree methods ignore the data underlying the source trees, this third step essentially delivers a tree topology only. With the possible exception of the average consensus method, any branch lengths on the supertree are either essentially meaningless (e.g., MinCutSupertree or gene-tree parsimony) or cannot be interpreted phylogenetically (e.g., matrix representation methods). This is especially important to realize for MRP supertrees, where the natural temptation is to interpret the resulting branch lengths in terms of the number of synapomorphies supporting that branch. Although the MRP supertree is indeed derived from a parsimony analysis, there is no connection with the original data such that one cannot talk about shared derived characters per se. Instead, meaningful branch lengths for the supertrees have to be obtained by mapping the primary character data a posteriori onto the topology of the supertree (e.g., Song et al. 2012), possibly in combination with calibration data to obtain real divergence-time estimates (e.g., Nyakatura and Bininda-Emonds 2012).

3.4.4 Step 4: Assessing Support Within the Supertree

As pointed out by Purvis (1995b), the use of the nonparametric bootstrap to summarize the nodal support within a supertree was invalid because the inherent nonindependence of the additive binary coding (Farris et al. 1970) underlying matrix representation violates a key assumption of the bootstrap. Although this is correct, the real problem with the application of this and any other character-based support method (e.g., Bremer support) is that all fail to account for the fact that the raw data of a supertree analysis are the source trees and not the character data underlying them or even the pseudo-characters derived from them via matrix representation.

Although their development was somewhat delayed and nowhere near as well explored as the creation of new supertree methods, several supertree-specific support measures now exist. One class contains methods that are analogous to the nonparametric bootstrap for character data (Felsenstein 1985a), except that the source trees are instead resampled with replacement. This procedure has been implemented in the software package CLANN (Creevey and McInerney 2005), but obviously only applies to the supertree methods available within it. An implementation of this method, multilocus bootstrapping, is also available for MLC-based supertree methods (Liu et al. 2010). A variation on this basic scheme, stratified bootstrapping, builds the supertree in each replicate from a randomly chosen tree from the bootstrap profile of each gene tree (Burleigh et al. 2006). Although this point has not been examined, stratified bootstrapping might be able to account in a limited fashion for hidden support within the raw character data as far as it is expressed among the trees in the individual bootstrap profiles.

As with the normal bootstrap, a clear disadvantage of this method in general is its high computational load in that *n* replicates of the supertree analysis are essentially being performed. Although these searches can be simplified to save time (e.g., performing no branch swapping like in PAUP*'s (Swofford 2002) faststep bootstrap search), this solution invokes other problems because the individual bootstrap trees will not be as optimal, thereby potentially biasing the overall bootstrap frequencies in some unknown manner. Another potential problem with a supertree bootstrap analysis is that some bootstrap replicates might contain nonoverlapping sets of trees and/or might not contain the full taxon set present across all source trees, with this probability increasing as the degree of overlap among the set of source trees decreases. Again, such bootstrap replicates will obtain a completely unresolved supertree, thereby artificially decreasing the overall bootstrap frequencies. Although this scenario is also possible for an equivalent supermatrix analysis (i.e., character partitions that do not overlap with respect to their taxa), it is less likely given the larger number of characters compared to source trees (e.g., 10 partially nonoverlapping source trees might be obtained from 10,000 base pairs worth of sequence data). A potential solution here might be to include a seed tree in each bootstrap replicate, should one be present in the global analysis, to again provide a scaffold ensuring sufficient taxonomic overlap and complete taxon coverage.

Importantly, these supertree bootstrap methods not only provide an estimate of the differential support among the nodes within the supertree, but the profile of bootstrap supertrees is also useful for comparative analyses. Given that the results of the latter are dependent on the accuracy of the underlying phylogenetic tree, accounting for uncertainty/error in the latter is desirable such that the recent trend has been to perform comparative tests on a distribution of trees rather than on a single point estimate of the phylogeny (e.g., Arnold et al. 2010; Jetz et al. 2012; see also Chaps. 10–12). Typically, this distribution is obtained from a Markov chain Bayesian framework; however, there seems to be no reason why a profile of bootstrap trees cannot fulfill the same purpose.

A second class of support measures comprises those that directly quantify the degree of conflict between the supertree and the set of source trees. Examples here include the QS index (Bininda-Emonds 2003) and *V* (Wilkinson et al. 2005). Compared to the bootstrap, these methods are extremely rapid because both the supertree and the set of source trees have already been computed. Nevertheless, an inherent difficulty of the method is how to define support versus conflict in the case of missing taxa between the supertree and source tree (Bininda-Emonds 2003). For example, do source trees that contain either taxon A or taxon B, but not both, support or contradict a sister-group relationship between A and B specified by the supertree, or are they uninformative? Both the QS index and *V* take different approaches to this problem and it is unclear which, if either, is better.

Finally, although supertree methods like PhySIC and PhySIC_IST guarantee that no node on the supertree is contradicted by any of the source trees (Ranwez et al. 2007; Scornavacca et al. 2008), assessing the nodal support on these supertrees using either of the two classes of methods above is arguably still recommended. Key is that both PhySIC and PhySIC_IST do not assure that all source trees directly support a given supertree node, such that while all nodes are not contradicted, some might enjoy more absolute support than others.

3.5 Conclusions

As mentioned in the Introduction (Sect. 3.1), the molecular revolution has arguably been more revolutionary in terms of the massive amounts of phylogenetic data it has provided rather than in the novel hypotheses of phylogenetic relationships it has produced. The latter stability also extends to the gene tree/species tree dichotomy that forms the basis of this chapter, where the reality is that most phylogenetic methods and analytical frameworks seem to be pointing in the same general direction. Thus, the reassuring trend we see is one of growing congruence rather than increasing conflict. Problem areas do remain (e.g., the root of the placental mammals; Teeling and Hedges 2013), but have long been recognized as such, even within any single framework.

Nevertheless, as I have argued in the past (Bininda-Emonds 2004c), a supertree framework—including the MLC model—remains a valid and desirable complement (not alternative) to a pure concatenation-based supermatrix framework, which remains the de facto standard of (molecular) phylogenetics. This point has also been admitted to some extent by even the most vocal critics of supertrees, who minimally see the methodological need for supertrees in piecing together the entire Tree of Life (Gatesy and Springer 2004) and/or do not object to the supertree framework in general (Murphy et al. 2012). More generally, by focusing on different levels of the phylogenetic data set—gene trees versus individual nucleotides, respectively—both the supertree and supermatrix frameworks place slightly different analytical emphases on the same base data set and the use of both approaches in parallel potentially balances out their respective strengths and weaknesses. For example, whereas only a supermatrix framework can account for hidden support, supertrees are better able to account for gene-tree heterogeneity. Given these different foci, analyzing a data set using both frameworks (i.e., essentially parallel partitioned vs. unpartitioned analyses) will therefore provide us with greater confidence in those areas where their results are congruent and greater insight into the causes of any incongruence where they are not, an approach in agreement with the global-congruence framework of Lapointe et al. (1999). In this way, we will also be better able to establish the frequency of ILS among different taxonomic groups as well as its potential for leading supermatrix-based analyses astray. Moreover, the potential to expand the MLC model in particular to incorporate processes of gene flow and HGT (Liu et al. 2009a) should provide even greater information regarding their frequency and their effects on speciation and phylogenetic history.

Acknowledgments I thank László Zsolt Garamszegi for the invitation to contribute to this exciting project and his incredible patience in putting it all together. Thanks also go to Las and two anonymous reviewers for their comments that helped improve and focus my original thoughts.

Glossary

Hidden support (AKA signal enhancement) The phenomenon whereby consistent secondary signals among a set of data partitions can overrule their conflicting primary signals to yield a novel solution not to be found among any of the individual data sets. As a simplified example, take the case of two separate gene data sets, each with an aligned length of 1000 nucleotides. In the first data set, 60 % of the positions support a sister-group relationship between A and B (primary signal), whereas 40 % support the clustering of B and C (secondary signal). In the second data set, 60 % support A and C, whereas 40 % support B and C.

Separate analyses of each data set will yield conflicting results (AB vs. AC); however, when the data sets are combined, each of these solutions is now only supported by 30 % of the data. By contrast, the secondary signals supporting BC are now present among 40 % of the combined data and now form the primary signal. In other words, each separate data set possessed hidden support for BC that could combine and determine the overall solution upon the concatenation of the data sets. Because supertree analyses work with trees as their primary data source, these secondary signals in the raw character data are normally invisible and cannot be accounted for.

Long-branch attraction An artifact in the phylogenetic analysis of DNA sequence data that was first exposed by Felsenstein (1978) and is a result of saturation in such data. Felsenstein observed that taxa at the ends of very long branches that themselves were separated by a short intervening branch often clustered to form sister taxa in a maximum parsimony analysis. Optimization criteria that used an explicit model of evolution like maximum likelihood were more immune to this problem.

This artifactual attraction of the long branches arises because the taxa are characterized by high rates of molecular evolution (as indicated by the long branches) and concomitant large number of shared convergent changes that, through their high number, are falsely interpreted as evidence for shared common ancestry. It is now known that long-branch attraction is a general problem (i.e., it can affect nonmolecular data, although is far less likely to do so) and can occur even if the branches occur on distant parts of the tree (see Bergsten 2005).

Matrix representation

A long-standing mathematical principle (Ponstein 1966) showing that there is a one-to-one correspondence between a tree (a "directed acyclic graph") and its encoding as a binary matrix. Whereas additive binary coding (Farris et al. 1970) of the tree will derive the matrix, the tree can be recreated from the matrix via analysis of the latter using virtually any optimization criterion (see Fig. 3.2).

NP-complete

A class of nondeterministic polynomial (NP) time methods for which no efficient solution is known and for which the running time increases tremendously with the size of the problem. As such, heuristic rather than exact algorithms must be used beyond a certain problem size, meaning that there is no guarantee that the optimal solution has been found. In phylogenetics, classic examples of NP-complete algorithms include maximum parsimony and maximum likelihood.

Polynomial time

Polynomial time algorithms are said to be "fast" in the sense that they have an efficient solution that scales "reasonably" with the size of the problem. A cogent example here is neighbor joining (NJ), the running time of which scales no worse than the cube of the number of taxa (i.e., $O(n^3)$). This is in stark contrast to the NP-complete maximum parsimony and maximum-likelihood methods, where the running times scale super-exponentially with respect to the problem size.

Saturation

A phenomenon attributed primarily to DNA sequence data and which arises because of the limited character state space for such data (i.e., the four nucleotides A, C, G, and T). As such, the potential for homoplasy in the form of either convergence or back mutation is high (e.g., two completely random DNA sequences are expected to be 25 % similar). Saturation, however, can also occur, but is less likely, for both amino-acid and morphological character data.

In practice, saturation is visualized by the degree of divergence between two sequences leveling off or plateauing with time since their divergence because faster evolving sites have experienced multiple substitutions ("multiple hits") with the increased potential for homoplastic similarity. Another method is to examine for deviations from an expected transition: transversion ratio of 1:2 in neutral/silent sites, given the faster rate of evolution for transitions compared to transversions and, again, greater opportunity for multiple hits.

References

Adams EM III (1972) Consensus techniques and the comparison of taxonomic trees. Syst Zool 21:390–397

Adams EM III (1986) N-trees as nestings: complexity, similarity, and consensus. J Classif 3:299–317

Arnold CL, Matthews J, Nunn CL (2010) The 10k Trees website: a new online resource for primate phylogeny. Evol Anthropol 19:114–118

Asher RJ, Müller J (2012) Molecular tools in palaeobiology: divergence and mechanisms. In: Asher RJ, Müller J (eds) From clone to bone: the synergy of morphological and molecular tools in palaeobiology. Cambridge studies in morphology and molecules: new paradigms in evolutionary biology, vol 4. Cambridge University Press, Cambridge, pp 1–15

Baker RH, DeSalle R (1997) Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst Biol 46:654–673

Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41:3–10

Beddard FE (1900) A book of whales. G.P. Putnam's Sons, New York

Bergsten J (2005) A review of long-branch attraction. Cladistics 21(2):163–193

Berry V, Bininda-Emonds ORP, Semple C (2013) Amalgamating source trees with different taxonomic levels. Syst Biol 62(2):231–249

Bininda-Emonds ORP (2003) Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. Syst Biol 52(6):839–848

Bininda-Emonds ORP (2004a) The evolution of supertrees. Trends Ecol Evol 19(6):315–322

Bininda-Emonds ORP (2004b) New uses for old phylogenies: an introduction to the volume. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 3–14

Bininda-Emonds ORP (2004c) Trees versus characters and the supertree/supermatrix "paradox". Syst Biol 53(2):356–359

Bininda-Emonds ORP (2010) The future of supertrees: bridging the gap with supermatrices. Palaeodiversity 3(Suppl.):99–106

Bininda-Emonds ORP (2011) Inferring the Tree of Life: chopping a phylogenomic problem down to size? BMC Biol 9:59

Bininda-Emonds ORP, Beck RMD, Purvis A (2005) Getting to the roots of matrix representation. Syst Biol 54(4):668–672

Bininda-Emonds ORP, Bryant HN (1998) Properties of matrix representation with parsimony analyses. Syst Biol 47(3):497–508

Bininda-Emonds ORP, Jones KE, Price SA, Cardillo M, Grenyer R, Purvis A (2004) Garbage in, garbage out: data issues in supertree construction. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the Tree of Life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 267–280

Bininda-Emonds ORP, Jones KE, Price SA, Grenyer R, Cardillo M, Habib M, Purvis A, Gittleman JL (2003) Supertrees are a necessary not-so-evil: a comment on Gatesy et al. Syst Biol 52 (5):724–729

Bininda-Emonds ORP, Sanderson MJ (2001) Assessment of the accuracy of matrix representation with parsimony supertree construction. Syst Biol 50(4):565–579

Bremer K (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803

Bruen TC, Bryant D (2008) Parsimony via consensus. Syst Biol 57(2):251–256

Burleigh JG, Driskell AC, Sanderson MJ (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst Biol 55(3):426–440

Chaudhary R, Bansal MS, Wehe A, Fernandez-Baca D, Eulenstein O (2010) iGTP: a software package for large-scale gene tree parsimony analysis. BMC Bioinformatics 11:574

Chen D, Diao L, Eulenstein O, Fernández-Baca D, Sanderson MJ (2003) Flipping: a supertree construction method. In: Janowitz MF, Lapointe F-J, McMorris FR, Mirkin B, Roberts FS (eds) Bioconsensus, vol 61., DIMACS Series in discrete mathematics and theoretical computer scienceAmerican Mathematical Society, Providence, RI, pp 135–160

Chen D, Eulenstein O, Fernández-Baca D (2004) Rainbow: a toolbox for phylogenetic supertree construction and analysis. Bioinformatics 20(16):2872–2873

Chippindale PT, Wiens JJ (1994) Weighting, partitioning, and combining characters in phylogenetic analysis. Syst Biol 43:278–287

Cotton JA, Page RDM (2004) Tangled trees from multiple markers: reconciling conflict between phylogenies to build molecular supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 107–125

Cotton JA, Wilkinson M (2007) Majority-rule supertrees. Syst Biol 56(3):445–452

Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21(3):390–392

Davis KE, Hill J (2010) The supertree tool kit. BMC Res Notes 3:95

de Queiroz A, Donoghue MJ, Kim J (1995) Separate versus combined analysis of phylogenetic evidence. Annu Rev Ecol Syst 26:657–681

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. PLoS Genet 2(5):762–768

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? Evolution 63(1):1–19

Farris JS, Kluge AG, Eckhardt MJ (1970) A numerical approach to phylogenetic systematics. Syst Zool 19:172–191

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Felsenstein J (1985a) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1985b) Phylogenies and the comparative method. Am Nat 125:1–15

Gatesy J, Matthee C, DeSalle R, Hayashi C (2002) Resolution of a supertree/supermatrix paradox. Syst Biol 51(4):652–664

Gatesy J, O'Grady P, Baker RH (1999) Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 15(3):271–313

Gatesy J, Springer MS (2004) A critique of matrix representation with parsimony supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 369–388

Gatesy J, Springer MS (2013) Concatenation versus coalescence versus "concatalescence". Proc Natl Acad Sci U S A 110(13):E1179–E1179

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst Zool 28(2):132–163

Gordon AD (1986) Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. J Classif 3:31–39

Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol 47(1):9–17

Hailer F, Kutschera VE, Hallstrom BM, Klassert D, Fain SR, Leonard JA, Arnason U, Janke A (2012) Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science 336(6079):344–347. doi:10.1126/science.1216424

Harvey PH, Pagel MD (1991) The comparative method in evolutionary biology. Oxford University Press, Oxford

Hillis DM (1987) Molecular versus morphological approaches to systematics. Annu Rev Ecol Syst 18:23–42

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO (2012) The global diversity of birds in space and time. Nature 491:444–448

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

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29(6):1695–1701

Lapointe F-J, Cucumel G (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. Syst Biol 46(2):306–312

Lapointe F-J, Kirsch JAW, Hutcheon JM (1999) Total evidence, consensus, and bat phylogeny: a distance based approach. Mol Phylogenet Evol 11(1):55–66

Lapointe F-J, Levasseur C (2004) Everything you always wanted to know about the average consensus, and more. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 87–105

Lee MS, Camens AB (2009) Strong morphological support for the molecular evolutionary tree of placental mammals. J Evol Biol 22 (11):2243–2257. doi:JEB1843 [pii] 10.1111/j.1420-9101.2009.01843.x

Lindqvist C, Schuster SC, Sun Y, Talbot SL, Qi J, Ratan A, Tomsho LP, Kasson L, Zeyl E, Aars J, Miller W, Ingolfsson O, Bachmann L, Wiig O (2010) Complete mitochondrial genome of a Pleistocene jawbone unveils the origin of polar bear. Proc Natl Acad Sci U S A 107(11):5053–5057. doi:10.1073/pnas.0914266107

Liu F-GR, Miyamoto MM, Freire NP, Ong PQ, Tennant MR, Young TS, Gugel KF (2001) Molecular and morphological supertrees for eutherian (placental) mammals. Science 291:1786–1789

Liu L, Yu L (2010) Phybase: an R package for species tree analysis. Bioinformatics 26:962–963

Liu L, Yu LL, Kubatko L, Pearl DK, Edwards SV (2009a) Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol 53(1):320–328

Liu L, Yu LL, Pearl DK, Edwards SV (2009b) Estimating species phylogenies using coalescence times among sequences. Syst Biol 58(5):468–477

Liu LA, Yu LL, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol 10

Maddison WP (1997) Gene trees in species trees. Syst Biol 46(3):523–536

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC, Drautz DI, Wittekindt NE, Tomsho LP, Ibarra-Laclette E, Herrera-Estrella L, Peacock E, Farley S, Sage GK, Rode K, Obbard M, Montiel R, Bachmann L, Ingolfsson O, Aars J, Mailund T, Wiig O, Talbot SL, Lindqvist C (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci USA 109(36):E2382–E2390. doi:10.1073/pnas.1210506109

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

Murphy WJ, Janecka JE, Stadler T, Eizirik E, Ryder OA, Gatesy J, Meredith RW, Springer MS (2012) Response to comment on "impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification". Science 337(6090):34

Nguyen N, Mirarab S, Warnow T (2012) MRL and SuperFine plus MRL: new supertree methods. Algorithms Mol Biol 7(1):3

Nyakatura K, Bininda-Emonds ORP (2012) Updating the evolutionary history of Carnivora (mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol 10:12

Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst Biol 43(1):58–77

Page RDM (2002) Modified mincut supertrees. In: Guigó R, Gusfield D (eds) Proceedings of Algorithms in bioinformatics, second international workshop, WABI, Rome, Italy. Lecture Notes in computer science, vol 2452. Springer, Berlin, pp 537–552, 17–21 Sept 2002

Piaggio-Talice R, Burleigh JG, Eulenstein O (2004) Quartet supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 173–191

Ponstein J (1966) Matrices in graph and network theory. Van Gorcum, Assen, Netherlands

Purvis A (1995a) A composite estimate of primate phylogeny. Philos Trans R Soc Lond B 348:405–421

Purvis A (1995b) A modification to Baum and Ragan's method for combining phylogenetic trees. Syst Biol 44:251–255

Ragan MA (1992) Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol 1:53–58

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

Ranwez V, Berry V, Criscuolo A, Fabre PH, Guillemot S, Scornavacca C, Douzery EJ (2007) PhySIC: a veto supertree method with desirable properties. Syst Biol 56 (5):798–817. doi:782748826 [pii] 10.1080/10635150701639754

Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol 15(11):454–459

Ronquist F (1996) Matrix representation of trees, redundancy, and weighting. Syst Biol 45:247–253

Ronquist F, Huelsenbeck JP, Britton T (2004) Bayesian supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 193–224

Rosenberg NA (2013) Discordance of species trees with their most likely gene trees: a unifying principle. Mol Biol Evol 30(12):2709–2713. doi:10.1093/molbev/mst160

Roshan U, Moret BME, Williams TL, Warnow T (2004) Performance of supertree methods on various data set decompositions. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 301–328

Ross HA, Rodrigo AG (2004) An assessment of matrix representation with compatibility in supertree construction. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 35–63

Salamin N, Hodkinson TR, Savolainen V (2002) Building supertrees: an empirical assessment using the grass family (Poaceae). Syst Biol 51(1):136–150

Sanderson MJ, Donoghue MJ, Piel W, Eriksson T (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Am J Bot 81(6):183

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18(3):502–504

Scornavacca C, Berry V, Lefort V, Douzery EJ, Ranwez V (2008) PhySIC_IST: cleaning source trees to infer more informative supertrees. BMC Bioinformatics 9:413. doi:1471-2105-9-413 [pii] 10.1186/1471-2105-9-413

Semple C, Steel M (2000) A supertree method for rooted trees. Discrete Appl Math 105(1–3):147–158

Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ (2011) Understanding angiosperm diversification using small and large phylogenetic trees. Am J Bot 98(3):404–414

Song S, Liu L, Edwards SV, Wu SY (2012) Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci USA 109(37):14942–14947

Stamatakis A (in press) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. doi:10.1093/bioinformatics/btu033

Steel M, Dress AWM, Böcker S (2000) Simple but fundamental limitations on supertree and consensus tree methods. Syst Biol 49(2):363–368

Steel M, Rodrigo A (2008) Maximum likelihood supertrees. Syst Biol 57(2):243–250

Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Swenson MS, Suri R, Linder CR, Warnow T (2012) SuperFine: fast and accurate supertree estimation. Syst Biol 61(2):214–227

Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

Teeling EC, Hedges SB (2013) Making the impossible possible: rooting the tree of placental mammals. Mol Biol Evol 30(9):1999–2000

Thorley JL, Wilkinson M (2003) A view of supertree methods. In: Janowitz MF, Lapointe F-J, McMorris FR, Mirkin B, Roberts FS (eds) Bioconsensus, vol 61., DIMACS series in discrete mathematics and theoretical computer scienceAmerican Mathematical Society, Providence, RI, pp 185–193

Wehe A, Bansal MS, Burleigh JG, Eulenstein O (2008) DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24(13):1540–1541

Wilkinson M, Pisani D, Cotton JA, Corfe I (2005) Measuring support and finding unsupported relationships in supertrees. Syst Biol 54(5):823–831

Wilkinson M, Thorley JL, Pisani D, Lapointe F-J, McInerney JO (2004) Some desiderata for liberal supertrees. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: combining information to reveal the tree of life, computational biology, vol 4. Kluwer Academic, Dordrecht, the Netherlands, pp 227–246

Wu SY, Song S, Liu L, Edwards SV (2013) Reply to Gatesy and Springer: The multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. Proc Natl Acad Sci USA 110(13):E1180–E1180