Review

Computational approaches to species phylogeny inference and gene tree reconciliation

Luay Nakhleh[1,2]

[1] Department of Computer Science, Rice University, Houston, TX 77005, USA
[2] Department of Ecology and Evolutionary Biology, Rice University, Houston, TX 77005, USA

An intricate relation exists between gene trees and species phylogenies, due to evolutionary processes that act on the genes within and across the branches of the species phylogeny. From an analytical perspective, gene trees serve as character states for inferring accurate species phylogenies, and species phylogenies serve as a backdrop against which gene trees are contrasted for elucidating evolutionary processes and parameters. In a 1997 paper, Maddison discussed this relation, reviewed the signatures left by three major evolutionary processes on the gene trees, and surveyed parsimony and likelihood criteria for utilizing these signatures to elucidate computationally this relation. Here, I review progress that has been made in developing computational methods for analyses under these two criteria, and survey remaining challenges.

Multilocus analyses and evolutionary processes

Species phylogenies and gene trees have an intricate relation that stems from the evolutionary processes acting within, and sometimes across, species boundaries to shape the gene trees. Three major evolutionary processes are gene duplication, horizontal gene transfer (HGT), and hybridization. Gene duplication results in the creation of new copies of genes and, thus, has a central role in genome evolution [1]. Given that these copies acquire genetic differences, their evolutionary fates might differ and result in novel gene functions [2].

In asexual species, HGT shapes the genomic repertoire and imports new genes, sometimes of beneficial consequences, into the host genome [3,4]. HGT occurs mainly through one of three mechanisms: (i) transformation, which is the uptake of naked DNA from the environment; (ii) transduction, which is the transfer of genetic material through a plasmid or bacteriophage; and (iii) conjugation, which is the direct transfer of DNA between two cells.

In eukaryotes, the evolutionary histories of various groups of plants and animals have been shown to involve hybridization [5], which is the production of viable offspring from interspecific mating [6]. Two major outcomes of hybridization are introgression and hybrid speciation. Although some parts of the genetic material contributed to the offspring in interspecific mating is eliminated from the population in later generations, other parts are integrated into the genome, an event that is referred to as introgression. It is important to note that both HGT and introgression leave similar genomic signatures, although the former process occurs in asexual species, whereas the latter occurs in sexual species. In some cases, hybridization results in hybrid lineages that become reproductively isolated from the parental species, a phenomenon known as hybrid speciation. Figure 1 illustrates gene duplication, HGT, and hybridization in three-taxon scenarios.

Two of the main tasks of multilocus analyses are the inference of a species phylogeny and the evolutionary processes that acted upon the individual loci. Although species phylogeny inference used to be conducted almost exclusively based on a single gene sampled across species [7], it is becoming more common to use whole-genome data or, more generally, multiple loci. When gene trees have been inferred for the individual loci, the first task amounts to inferring the species phylogeny from these gene trees. The second task amounts to contrasting, or reconciling, the gene trees with the species phylogeny to elucidate the evolutionary processes that shaped the gene trees and their phenotypic consequences. Multilocus analyses provide power, in terms of phylogenetic signal, to solve both tasks with high accuracy, yet pose new modeling and computational challenges for phylogenetic inference that mostly stem from a phenomenon known as gene tree incongruence.

Phylogenetic incongruence: a signal, rather than a problem

As illustrated in Figure 1, each of the evolutionary processes operating on a gene leaves its signature on the gene tree. These processes alone do not necessarily result in signatures in the form of incongruence between gene trees and the species phylogeny. It is often the evolutionary fates of gene copies that result in such signatures. These evolutionary fates are determined by forces such as mutation, drift, and selection. For example, in Figure 1A, if the gene copies b_1, c_1, and a_2 are lost, the resulting gene tree differs from the species tree. In Figure 1B, if the HGT event

Corresponding author: Nakhleh, L. (nakhleh@rice.edu).

0169-5347/$ – see front matter
 http://dx.doi.org/10.1016/j.tree.2013.09.004

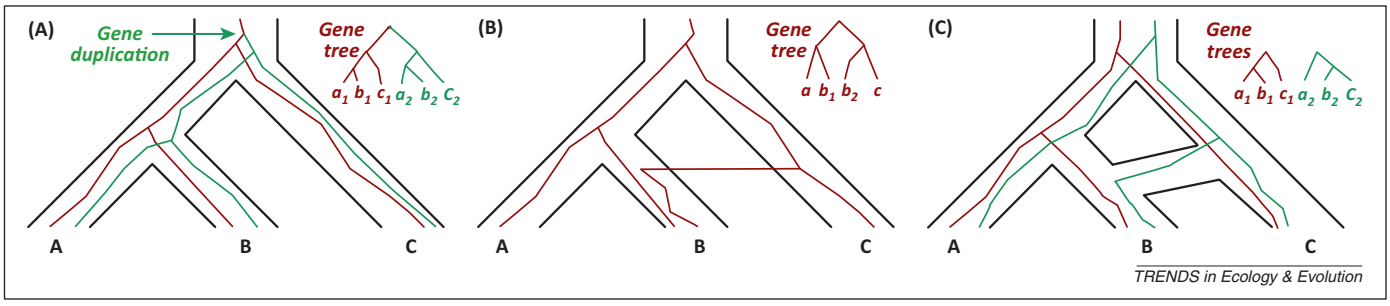

Figure 1. Evolutionary processes within and across species boundaries. **(A)** A gene duplication event at the most recent common ancestor (MRCA) of all three taxa, results in two copies (red and green) of a gene within the genome and, as the genome undergoes evolution, these copies evolve, diverge, and might have different fates. **(B)** In prokaryotic organisms, DNA containing genes might be transferred across species boundaries, for example, from C to B, resulting in a new gene copy. Furthermore, a similar signature might arise in cases of introgression in sexual species. **(C)** Hybridization between species A and C amounts to individuals from A and B mating and producing viable offspring such that the genetic material in individuals of B can be traced back to two parental species. The gene tree in each case is shown in the inset.

results in the displacement of the b_1 gene copy, then the resulting gene tree differs from the species tree. By contrast, if the horizontally transferred gene copy, b_2, is eventually lost, then the gene tree remains congruent with the species tree. In the case of hybridization, the scenario is dictated by the mode of the evolutionary process. In homoploid hybridization, the offspring has the same ploidy level, or number of chromosomes, as each of the parents in the two hybridizing species. In this case, hybridization is often followed by backcrossing, which is further mating between individuals from the hybrid population and either of the two parental populations. Repeated backcrossing, combined with drift and selection, results in unequal parental genomic contributions in the hybrid offspring and a distribution of differing gene tree topologies across the genomes. In (allo)polypoloid hybridization, the offspring gets the complete set of chromosomes from the parents and, thus, has a number of chromosomes that is double that of either of the two parents. Although backcrossing does not occur in cases of polyploid hybridization, drift and selection result in unequal parental genomic contributions in the hybrid offspring.

From an inference perspective, these signatures can then be utilized as phylogenetic signal to recover population parameters, evolutionary processes, and the species phylogeny itself [8]. However, it is important to keep in mind several issues that make the inference task challenging in practice. First, incomplete sampling of gene copies by the practitioner might give rise to artificial signatures that mislead or confound inference tasks. For example, if the practitioner samples only copes a_1, b_1, and c_1 in the scenario given in Figure 1A, the occurrence of a gene duplication event might not be recovered. Second, multiple occurrences of the same evolutionary process might cancel out or complicate the signature. For example, assuming displacement of gene copy b_1 by the HGT event in the scenario of Figure 1B, a subsequent HGT event from B to A, involving gene copy b_2 and the displacement of the original copy of the gene in A, results in a gene tree that is congruent with the species tree (in terms of topology, but not branch lengths). Third, the occurrence of an evolutionary process might not leave a signature on the gene tree topologies. For example, an HGT between two sister taxa does not result in incongruence between the gene and species trees. Fourth, the signature left by an evolutionary process might not be unique to that process [9]. For example, if gene copies b_1, c_1, and a_2 are lost in the scenarios of Figure 1A and C, and the gene copy b_1 is lost in the scenario of Figure 1B, then one ends up with the same gene tree topology in all three cases. Furthermore, as discussed above, HGT and introgression might give rise to identical genomic signatures, although they occur in different groups of species. It is crucial that these issues are kept in mind when applying inference methods, developing new ones, or interpreting the results thereof.

It is probably due to these issues, and others, that several genomic studies that are mainly aimed at obtaining the species phylogeny mask signatures by selecting few loci that satisfy stringent criteria so as to eliminate the possibility of incongruence and other studies have strived to do phylogenetic inference despite incongruence. By contrast, in this review, I take the position that incongruence is a powerful phylogenetic signal that is 'desirable, as it often illuminates previously poorly understood evolutionary phenomena' [9]. Fields such as molecular population genetics and phylogenetics have long relied on polymorphism and divergence at the sequence level as signal for inference and, in the postgenomic era, phylogenomics relies on phylogenetic incongruence as the major signal for inference. Therefore, phylogenetic incongruence should not be viewed as a problem to be masked or despite which inference should be made; rather, it should be viewed as a powerful character with a rich set of states to reconstruct and understand evolutionary phenomena, while accounting for the aforementioned issues.

For example, in 1979, Goodman *et al.* proposed a parsimony-based approach for fitting a gene tree onto a species tree to elucidate gene duplication and loss (DL) events from a set of globin sequences [10]. In 1997, Maddison proposed to count the minimum number of branch moves needed to convert the species tree into the gene tree, where branch moves do not violate the temporal constraints provided by the trees, as a proxy for the number of HGT or hybridization events [11]. Indeed, if these methods were applied to the scenarios in Figure 1, the true evolutionary events would be uncovered. Although these two approaches mainly reveal information about the evolutionary processes themselves, model-based approaches would help elucidate, in addition, knowledge about parameters such as population sizes, divergence times, duplication rates, etc. Furthermore, these reconciliation approaches can be turned into species phylogeny inference approaches by seeking a species phylogeny that, when all gene trees are reconciled

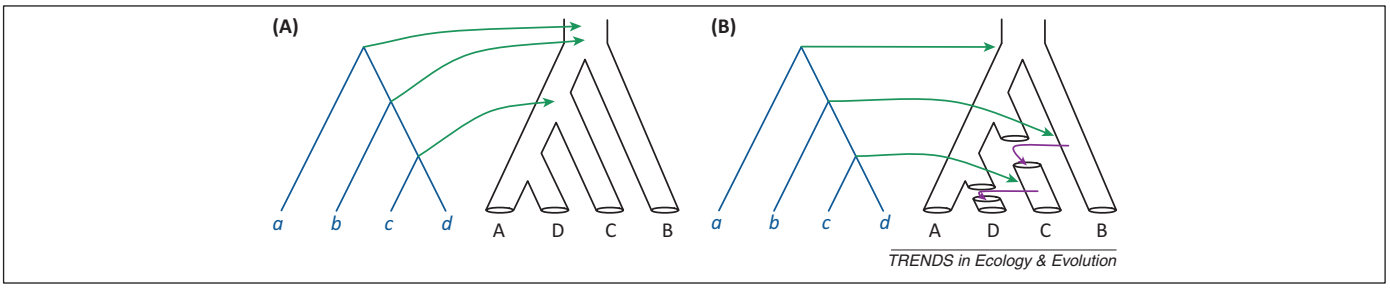

Figure 2. Fitting a gene tree onto a species tree. Gene trees are drawn with solid lines, and species trees are drawn with tubes. **(A)** In the case of gene duplication and loss (DL) [and incomplete lineage sorting (ILS)], each node x in the gene tree is mapped to (denoted by the green arrows) the most recent common ancestor (MRCA) of the species that contain gene copies descended from node x. **(B)** In the cases of horizontal gene transfer (HGT) and hybridization, a smallest set of branch moves (denoted by the purple arrows) that makes the species tree identical to the gene tree and do not violate 'a linear time order' is a parsimonious set of HGT or hybridization events that explain the difference between the species tree and gene tree.

with it, achieves some optimality score. In 1997, Maddison surveyed phylogenetic incongruence, and described parsimony and likelihood criteria for various reconciliation and inference problems [11]. Much progress has been made since 1997 on developing mathematical models and computational methods for these problems, and the goal of this review is to revisit the two criteria and provide an update on this progress. I use Maddison's article as an organizing principle for the remainder of this review.

Phylogenetic incongruence: Maddison's 1997 survey

In [11], Maddison discussed phylogenetic incongruence and the two computational problems of reconciliation and inference. The reconciliation problem seeks a fitting of a given gene tree within, or across, the branches of a given species tree assuming a source of incongruence. That is, every leaf in the gene tree is mapped to a leaf in the species phylogeny, and then internal nodes (which correspond to events of coalescence, duplication, HGT, etc.) in the gene tree are mapped to the branches of the species phylogeny. In this way, the reconciliation reveals the evolutionary processes that act on the gene and, when model-based approaches are also used, also reveals information about the timing of these processes, as well as parameters such as population sizes, duplication rates, etc. The inference problem seeks the species tree, given a collection of loci sampled from a set of species. In traditional phylogenetics, the inference problem amounts to estimating a phylogenetic tree from a molecular sequence alignment, often assuming only base-pair mutations. Analogously, in phylogenetic analyses involving multiple loci, the inference problem amounts to estimating a species phylogeny from a collection of gene trees, assuming some of the evolutionary processes discussed above.

For the reconciliation problem, Maddison discussed parsimony approaches for the cases where gene DL are both at play and when HGT is at play. In the case of DL, Goodman *et al.* had already proposed a parsimony-based approach for fitting a gene tree onto a species tree to minimize the number of duplications [10]. In this approach, a node x in the gene tree is mapped to the most recent common ancestor (MRCA) of the set of species that contain gene copies descended from node x (Figure 2A). In the case of HGT, Maddison proposed to count the minimum number of branch moves needed to convert the species tree into the gene tree, where branch moves do not violate the temporal constraints provided by the trees. This number would constitute a lower bound on the number of HGT events needed to explain the incongruence between the species tree and gene tree (Figure 2B).

Given the imbalance in the parental genetic contributions to hybrid offspring, parsimonious detections of hybridization events can be carried out in a similar fashion to that of HGT. In other words, although HGT and hybridization are different biological processes, their inference under parsimony is similar, and the same can be said of meiotic recombination. In addition to the aforementioned evolutionary processes, Maddison discussed the role that random genetic drift has in phylogenetic incongruence, a phenomenon that I now introduce here. Gene trees might disagree with each other, as well as with the species tree, due to random genetic drift acting within the populations, a phenomenon known as incomplete lineage sorting (ILS; Figure 3) [7,12]. Unlike gene DL, HGT, and hybridization, ILS does not introduce new genetic material into genomes; instead, it is a reflection of the inherent stochasticity associated with neutral evolution. Maddison proposed that the same mapping of gene tree nodes to species tree nodes as that used by [10] would result in a parsimonious reconciliation (one that minimizes the number of 'extra' gene lineages) assuming ILS as the source of incongruence. The reconciliations of the gene tree and species tree given in Figure 2 are shown under ILS, DL, and HGT in Figure 4.

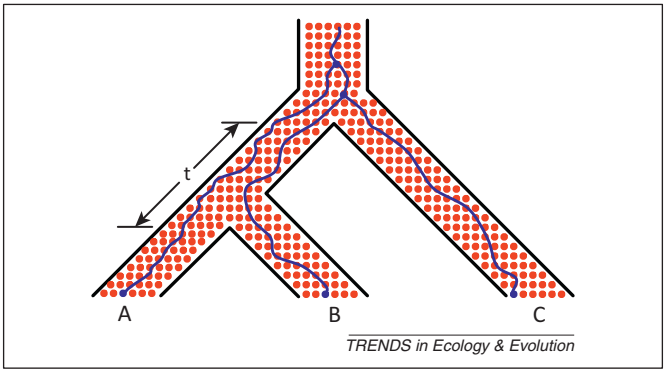

Figure 3. Incomplete lineage sorting. As the evolution of three sampled alleles (blue solid circles at the bottom) is traced backward in time, alleles from A and B might fail to coalesce in the ancestral population. This results in all three alleles entering the ancestral population of all three species, and the alleles from B and C coalescing first, by chance, giving rise to a gene tree that is incongruent with the species tree. The probability of this event happening in this scenario is a function of the branch length, *t*, as measured in coalescent units (one coalescent unit equals 2*N* generations, where *N* is the population size).

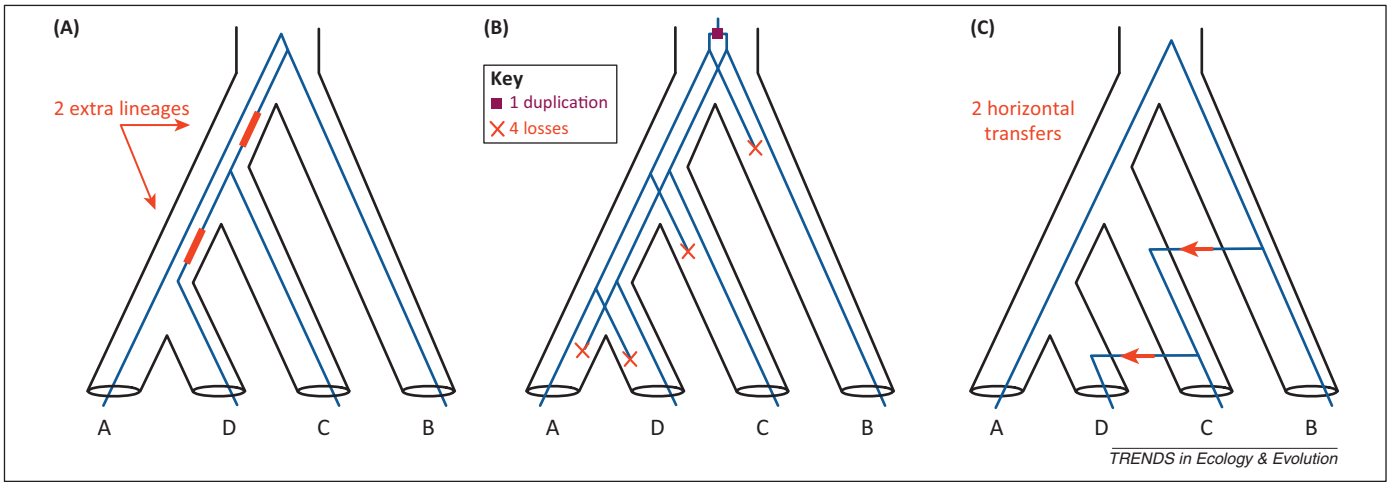

Figure 4. Reconciliation of a gene tree with a species tree. **(A)** Reconciliation assuming incomplete lineage sorting (ILS) results in two extra lineages, highlighted with thick red lines. **(B)** Reconciliation assuming gene duplication and loss (DL) results in a single duplication event and four losses. **(C)** Reconciliation assuming horizontal gene transfer (HGT) (or hybridization) results in two horizontal transfer events, highlighted with red arrows.

These parsimony-based approaches to reconciliation naturally give rise to three parsimony-based criteria for species tree inference: of all the possible species tree candidates, seek one that minimizes the total number of 'extra' gene lineages, duplication events, or HGT events, respectively, when all gene trees in the sample are reconciled with it. Maddison further proposed a maximum likelihood (ML) formulation for the inference problem. However, unlike the case of the parsimony formulations, he considered only deep coalescence (equivalently, ILS) in the case of ML, the reason being that the coalescent theory from population genetics had already provided a mechanism for computing the probability of a gene tree, whereas no similar theory existed for computing gene tree probabilities when DL, HGT, or hybridization were involved. The ML formulation proposed in [11] assumes a given collection of sequence alignments, each for a sampled locus, and seeks a tree that maximizes the probability of observing these alignments by accounting for mutations within each locus and incongruence across loci. In the next section, I review progress that has been made on Maddison's proposals for reconciliation and inference under the assumption of an individual evolutionary process being at play, and then discuss progress that has been made in the unification of processes within integrated frameworks.

Progress on methods that deal with individual processes

As Maddison mentioned, a reconciliation of a gene tree with a species tree under ILS or DL, using the mapping described above, is efficiently computable. Several algorithms have been introduced to compute reconciliations under ILS [13] and DL [14]. In terms of mathematical characterizations, Zhang [15] showed that, when the species tree and gene tree have exactly the same leaf set (i.e., exactly one gene copy from each species is used to infer the gene tree), then the number of extra lineages required to reconcile the trees assuming only ILS equals the number of losses minus twice the number of duplications required to reconcile the same trees assuming only DL. For example, the number of extra lineages in Figure 4A is 2, the number of losses and duplications in Figure 4B are 4 and 1 respectively, and, therefore, $2 = 4 - (2 \cdot 1)$. The formula relating the three quantities becomes slightly more involved when the two trees do not necessarily have the same leaf set [15]. For the inference problem, Maddison and Knowles [16] proposed a heuristic for searching for the species tree that minimizes the number of extra lineages assuming ILS is the sole cause of incongruence. Than and Nakhleh [13] later devised exact algorithms for the problem, including for cases where multiple alleles are sampled [17]. Than and Rosenberg proved that this parsimony criterion of minimizing the number of extra lineages is in fact statistically inconsistent (that is, inference under this criterion might converge on the wrong species tree, even as the number of gene trees used in the inference increases) [18]. Bayzid *et al.* devised exact algorithms for inferring a species tree that minimizes the number of DL [19].

As for HGT, the field has evolved rapidly so as to deal with complexities not discussed in [11]. The reconciliation problem assuming HGT is hard algorithmically [20–22] and, several methods for reconciling a pair of trees were devised (reviewed in [23]); these methods vary in the assumptions and restrictions that they make about the trees and reconciliations. Perhaps the issue that challenges Maddison's original proposal most is the concept of a species tree when HGT, or other reticulate evolutionary events, occur. Although a species tree in the case of ILS and DL can fit within its branches the evolutionary histories of all genes within the genomes under consideration, that structure would fail to capture adequately the evolutionary histories of genes that are exchanged horizontally. To accommodate reticulate evolutionary histories, phylogenetic networks were introduced as a model of evolutionary histories that capture both vertical and horizontal descent of genetic material [24–27] (Figure 5). A phylogenetic network extends the notion of phylogenetic trees by allowing for nodes with more than one parent–reticulation node. Assuming no ILS or DL, the evolutionary history of each gene in a set of species whose evolutionary history is given by a phylogenetic network N is captured by one of the trees displayed (or induced) by the phylogenetic network

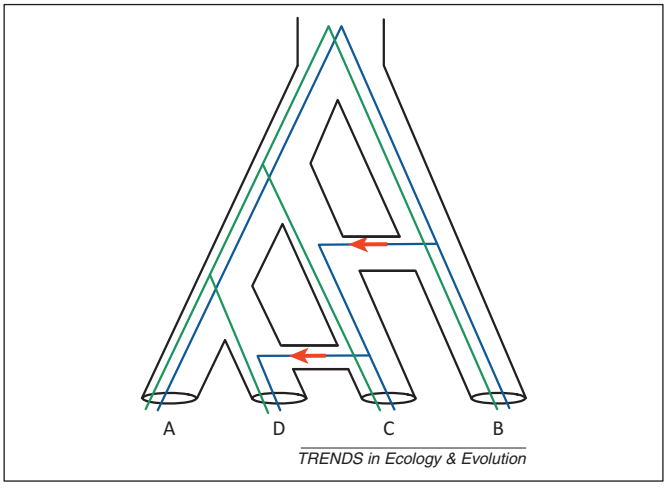

Figure 5. Gene trees within the branches of a phylogenetic network. The phylogenetic network, drawn with tubes, fits the evolutionary histories of all genes, including those that evolve vertically (e.g., the gene tree drawn with green lines) and those that involved horizontal transfer (e.g., the gene tree drawn with blue lines, and horizontal gene transfer or introgression events highlighted with red arrows).

N. A tree is induced by phylogenetic network N if it can be obtained by removing all but one of the parents for each of the reticulation nodes in the network. For example, the four trees induced by the network in Figure 5 are (((A,D),C),B), (A,(B,(C,D))), ((A,(C,D)),B), ((A,D),(B,C)). Reconciling a gene tree with a phylogenetic network, excluding ILS and DL, is related to testing whether the gene tree is one of the trees induced by the network, which has been shown to be computationally difficult [28].

Not only do phylogenetic networks provide a more adequate model compared with trees for capturing reticulate evolutionary histories, but they also enable Maddison's original proposal to be extended from reconciling a pair of trees to a collection of trees. Indeed, in current phylogenomic analyses, multiple loci are sequenced and multiple gene trees need to be reconciled. Maddison's proposal for reconciling a gene tree with a species tree does not carry over cleanly to a set of gene trees for the inference problem. By introducing the notion of a phylogenetic network, the parsimony version of the inference problem under HGT is now the need to find a phylogenetic network with the minimum number of reticulation nodes needed to display all of the gene trees. Several methods have been proposed recently for solving versions of this problem [29–35].

The progress on likelihood approaches for dealing with gene tree incongruence has been greater for ILS than for the other evolutionary processes, owing mainly to the mature theoretical foundations of the coalescent model that deal with ILS naturally. As discussed above, the ML formulation given in [11] was proposed in the context of ILS alone. Based on that formulation, the likelihood of a species tree is shown in Equation 1:

$$\prod_{loci} \sum_{gene\ trees} [\mathbf{P}(sequences|gene\,tree)\mathbf{P}(gene\,tree|species\,tree)] \quad [1]$$

Although Maddison used the summation over gene trees, this is to be treated as integration when branch lengths of the gene trees are also considered. The first probability of observing a set of (aligned) sequences given a gene tree depends on the model of sequence evolution and can be computed efficiently [36]. The second probability of observing a gene tree given a species tree is derived from coalescent theory, and methods have been devised for computing it when the gene tree is given only by its topology [37] or by its topology and branch lengths [38]. Likelihood methods have been proposed for inference based on this formulation [39,40]. Advances have been made recently in methods for computing the second probability when only DL is at play [41–43]. When only reticulation is at play and a parameterized species phylogenetic network is provided, computing the probability of a gene tree is straightforward [23].

Unifying processes and accounting for error

The fact that much progress has been made on methods that deal with each of the evolutionary processes individually is not to be construed as a statement that these processes do indeed occur in a mutually exclusive manner. As phylogenomic analyses grow in scope to involve more species, individuals, and loci, accounting simultaneously for multiple evolutionary processes becomes essential. Indeed, several studies have highlighted this issue in various groups of organism. For example, although introgression was hypothesized between Neanderthals and humans [44], this hypothesis was later dismissed in favor of ILS [45]. Simultaneous patterns of introgression and ILS were reported in 2012 in the genomes of the house mouse (*Mus musculus*) [46], the butterfly *Heliconius melpomene* [47], sunflower (*Helianthus* spp.) [48], and yeast [49]. Simultaneous patterns of ILS and DL were recently reported in a multilocus analysis of a group of fungi [50]. Furthermore, simultaneous patterns of DL and reticulation have also been reported [51]. Maddison [11] pointed to two challenges facing the development of a 'mixed method' that allows all three processes to occur: the algorithmic challenge of conducting reconciliation and inference under multiple processes, and the challenge associated with weighting the three different processes (e.g., should one HGT event be counted as equal to one duplication event?). Although the weighting relates mostly to parsimony approaches, its counterpart in a likelihood approach is setting the rates of the various processes (e.g., the rates of duplication, loss, etc.).

As discussed above, a phylogenetic network provides a more appropriate model of evolutionary relations compared with trees when reticulation is involved. However, a phylogenetic network not only accommodates HGT and hybridization, but treelike evolutionary processes, such as ILS and DL, can also be modeled within its branches. For example, Figure 6 illustrates how a phylogenetic network simultaneously models hybridization between species and ILS involving gene trees. It further illustrates the generality of the model in terms of accommodating multiple individuals sampled per species or population. Therefore, although Maddison did not discuss phylogenetic networks in his original survey, I take the position that, for the unification of all evolutionary processes, a species phylogeny in the form of a network is more appropriate than a tree. In fact, when recombination occurs within a locus,

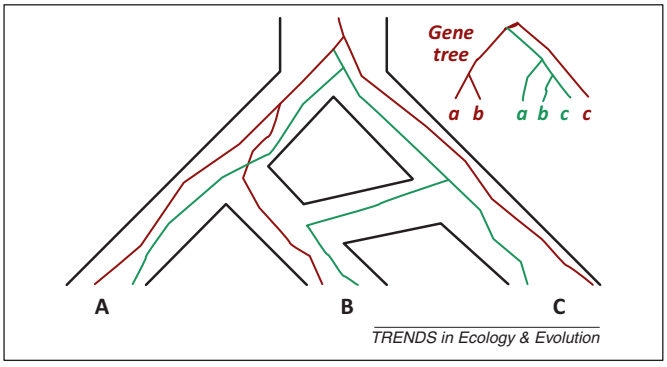

Figure 6. Simultaneous modeling of hybridization and incomplete lineage sorting (ILS) with a phylogenetic network. Two individuals are sampled per species, and there is a hybridization event that involves species B and C. Furthermore, ILS patterns complicate the gene genealogy, giving rise to the gene-tree topology shown in the inset. For example, the gene copies in green coalesce with the ancestral copy of the genes in red from A and B, before the latter one coalesces with the copy from C.

even the gene tree is better modeled using a network that is often referred to in the population genetics literature as an ancestral recombination graph [8]. This position is not to be interpreted as invoking reticulation in every analysis; rather, it advocates the development of mathematical models and computational methods that utilize the more general model, which is a network rather than a tree, and account for the possibility that in some, or perhaps most, cases the inferred network could be a special case that is a tree. The other approach of utilizing a tree as the topological model would exclude the possibility of a reticulate evolutionary history, merely by definition of the model used.

Progress in parsimonious reconciliation and inference methods that assume more than a single source of incongruence has also been made. Bansal *et al.* [52] recently introduced an efficient algorithm for reconciling a gene tree with a species tree assuming both DL and HGT. Yu *et al.* [53,54] introduced methods that assume both hybridization and ILS. In particular, the work in [54] provides algorithms for reconciliation as well as search heuristics that explore the space of phylogenetic networks to solve the inference problem. Recently, Stolzer *et al.* introduced a method for reconciling a gene tree with a species tree under DL, HGT, and ILS [55].

Whereas a natural way for integrating all evolutionary processes within a parsimony framework is to optimize a weighted sum of the numbers of events detected, a likelihood approach requires probabilistic models of these processes. Although the coalescent model has provided a natural framework for thinking about ILS, recent studies are beginning to shed light on how to model probabilistically processes including HGT [56–58] and DL [2,43,59]. For integrative likelihood approaches, a method for computing the probability of a gene tree given a species tree assuming both ILS and DL was given in [50], assuming DL and HGT was given in [60], and assuming ILS and HGT in a special case was given in [61]. Methods have also been developed for computing the probabilities of gene trees under hybridization and ILS in special, limited cases [53,62–64], and then for computing the probabilities in general cases [49]. Marcussen *et al.* [65] recently developed a method for inferring phylogenetic networks in the presence of ILS that is aimed at modeling polyploid hybridization.

A salient feature of all phylogenetic analyses, whether they involve a single locus or multiple loci, is the fact that gene trees are estimated from molecular sequences and, consequently, they are likely to be inaccurate. Maddison [11] wrote: 'I assume through most of this discussion that the true gene trees are known without error. Of course, there will be errors in practice, and these errors will mean that reconstructed gene trees and species trees will have additional sources of discord.' Indeed, Hahn [66] recently showed the effect of error in gene tree estimates on the computed reconciliations, and Yang and Warnow [67] showed that methods that explicitly account for error in the gene trees outperform others. Although incongruence caused by evolutionary processes provides a signal for inferring the processes themselves and the species phylogeny, incongruence due to error in the inferred gene trees is a confounding factor that must be accounted for carefully, because it produces topological signatures in the gene trees that can cancel out true evolutionary signals or masquerade as ones. One way to deal with error in the estimates of gene tree topologies is to contract all branches with low support (e.g., as measured by a bootstrap analysis, or posterior probabilities from a Bayesian analysis), and develop methods that can handle nonbinary, or multifurcating, trees. In the parsimony setup, a natural way to define the reconciliation of a nonbinary gene tree with a species tree is to find the refinement of the polytomies in the gene tree that results in the most parsimonious reconciliation over all possible reconciliations (Figure 7). Indeed, this refinement concept was used in [68–70] for reconciling nonbinary gene and species trees. Although the number of refinements is exponential in the degrees of the polytomies (the number of children of a node), Yu *et al.* recently devised polynomial-time, exact algorithms for finding the refinement that results in an optimal reconciliation under ILS [71,72]. Furthermore, the same ideas were

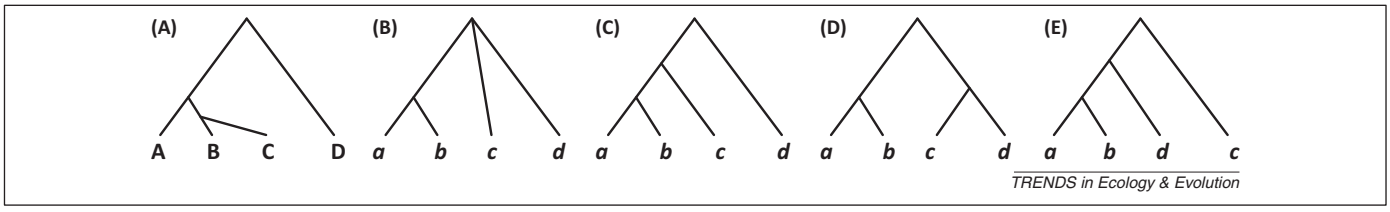

Figure 7. Parsimonious reconciliation of a nonbinary gene tree. **(A)** A binary species tree. **(B)** A nonbinary gene tree. **(C–E)** Three possible refinements of the nonbinary gene tree. Under parsimony, the refinement in (C) results in the best reconciliation with the species tree, because it results, assuming incomplete lineage sorting, gene duplication and loss, or horizontal gene transfer (HGT), in one extra lineage, one duplication and three losses, and one HGT, respectively.

extended to the problem of parsimonious reconciliation of a nonbinary gene tree with a phylogenetic network [54].

Under the likelihood approach, it is less clear how to deal with nonbinary gene trees. Should the gene tree be refined in a way that maximizes the probability of observing the (binary) gene tree (e.g., as implemented for inference under ILS in [73])? Or, should the probability of the nonbinary gene tree be computed as the average probability of all binary refinements? A different method to handle error in the gene trees is to directly make use of the support values. The major challenge facing this approach is in translating support values from gene tree branches to support values of reconciliations. Nonetheless, some heuristics were introduced recently based on this approach for reconciliation under HGT [74,75]. Furthermore, Yu *et al.* incorporated posterior probabilities in methods for reconciling a gene tree with a phylogenetic network under both likelihood [49] and parsimony [54]. Of course, methods that work directly from the sequence alignments of the multiple loci, rather than from estimated gene trees, account implicitly for error. The Bayesian methods of [76,77] for inference under ILS, and the parsimony and likelihood methods of [78,79] for inferring HGT events follow this approach.

Other approaches

Several approaches that do not fit within Maddison's parsimony and likelihood formulations have been proposed. Concatenation is an approach in which the sequences from multiple loci are concatenated, thus resulting in a 'super gene', from which a phylogenetic tree is inferred. For example, this approach was used in inferring a phylogenetic tree of a set of yeast species [80]. There are at least three issues with this approach. First, the approach is applicable to loci for which exactly one copy per species is sampled. However, even then, the phylogeny estimated from the concatenated alignment might be wrong [81]. Second, this approach yields, by definition, a phylogenetic tree. Therefore, it masks any signal of reticulate evolution if it exists. Third, the method does not allow for inference of the evolutionary processes. The democratic vote approach amounts to taking the gene tree with the highest frequency as a proxy for the species tree [82]. Applicability of this approach when a small sample of loci is used or when DL events are involved is questionable. Even when a large number of loci is used and each has exactly a single copy sampled per species, this method produces a misleading phylogeny in the 'anomaly zone' [7]. Finally, it is not clear how to interpret the gene tree with the highest frequency when the species evolutionary history is reticulate.

The majority-rule consensus is a third approach to produce a phylogenetic tree given a set of conflicting gene tree topologies. This approach often results in a phylogenetic tree with a low degree of resolution. Furthermore, the approach is not well defined for cases where the incongruence is due to DL. Similar to the previous two approaches, this approach always produces a tree, even when the evolutionary history is reticulate.

Mossel and Roch [83] recently introduced a distance-based method for inferring a species tree from pairwise distances computed from multiple loci. This method requires accurate estimates of the distances, and is applicable to neither DL nor reticulate evolutionary events. Bayesian approaches for inferring species trees under ILS were also recently introduced [77,84]. These methods differ from all the methods discussed above in that they work directly with sequence alignments and perform simultaneous inference of gene and species trees. They have been shown to produce good results, yet to be inefficient computationally. Furthermore, these Bayesian approaches currently do not handle DL or reticulate evolutionary events. Methods for inferring HGT events based on an assumed species tree and sequence alignments of genes were proposed based on the maximum parsimony and ML criteria [23]. These methods do not account for ILS or DL, and assume knowledge of an underlying species tree. Joly *et al.* showed how to use coalescent-based simulations to detect hybridization [85]; however, their approach was presented for a pair of species only. Last but not least, Holland *et al.* [85,86] demonstrated how to use consensus networks to detect hybridization in the presence of ILS.

Concluding remarks and future directions

In 1997, Wayne Maddison discussed the intricate relation between a species tree and the gene trees that grow within, and across, its branches. Furthermore, he discussed parsimony and likelihood approaches to reconciling a gene tree with a species tree, and for the inference of species trees from collections of gene trees. Solving these two tasks would shed light on central issues in evolutionary and molecular biology, including speciation, evolutionary processes acting within and across population, the evolution of morphological characters, and genotype–phenotype relations. Sixteen years later, the significance of understanding this relation cannot be overstated, given the ability to sequence hundreds of prokaryotic genomes in a day, and eukaryotic genomes over slightly longer timeframes. Indeed, in less than two decades, the computational biology and bioinformatics communities responded to Maddison's proposal by making significant inroads in establishing mathematical results and devising computational methods for detecting, resolving, and ameliorating incongruence that arises in phylogenomic studies. Still, more is needed in terms of mathematical and computational developments. Although models of incongruence and methods for reconciliation and inference have been developed, computational requirements are still a major bottleneck. Most, if not all, of the methods described above are limited to small- or medium-sized data sets. High-performance computing approaches will be needed if these methods are to be applied to thousands of loci and hundreds to thousands of taxa. Currently, these data sets are beyond the capabilities of existing tools.

Maddison explicitly stated in [11] that his formulations assumed no recombination within a locus. However, what happens if this assumption is violated? Recent work has accounted for recombination within phylogenetic networks [87]. A recent study showed that ignoring recombination within loci might not have a significant effect on the quality of the inferred species tree under ILS [88]. Similar studies do not exist for cases of DL and HGT. Nevertheless, more generally, the availability of whole-genome data allows for

defining gene trees as the genealogies built from nonrecombining regions, which include coding and noncoding DNA. However, potentially more challenging than recombination are the findings of rearrangements at the subgene level, such as gene fission and fusion, which seem to be ubiquitous in prokaryotic genomes [89] and even in eukaryotic genomes [90]. These findings not only complicate the species–gene evolutionary relations, but also raise broader questions about orthology, gene families, and the 'cloudiness' [11] of the species phylogeny.

Furthermore, Maddison assumed that loci are unlinked and, hence, the fact that gene trees can be reconciled with a species phylogeny independently. Indeed, all methods described above assume unlinked loci and it is currently incumbent upon the practitioner to sample loci from the genomes in such a way that ensures this assumption holds (or that violation thereof is minimal). However, to make full use of whole-genome data, models that incorporate linkage across loci, including functional linkage [91], must be devised, and methods for inference under such models must be developed. A mathematical model for two linked loci was introduced in [92]. More recently, approaches for modeling ILS while accounting for linkage across loci were introduced [93,94]. These approaches do not account for DL or reticulation at the species level (they do account for recombination). Additionally, these methods have been applied to three-species data sets and it would be challenging to achieve scalability of these methods to large data sets. This further emphasizes the need for high-performance computing approaches, because modeling dependence only makes the problem harder.

Although many methods have been developed for reconciliation, the relative performance of these methods is yet to be investigated thoroughly. This is especially important because the practitioner is faced with a wide array of methods that differ in terms of the assumptions they make and the computational resources they need. Although studies are beginning to emerge [67,88,95–97], more comprehensive studies are still needed. In particular, most performance studies focus on ILS, probably due to the fact that the coalescent theory provides a clean generative model for simulating synthetic data, whereas no such theory exists for DL or HGT. In addition, although some methods perform poorly under certain conditions, they might perform well under other conditions. Full characterization of conditions under which a method performs well would be of utmost help to practitioners. Most importantly, measures that reflect these characterizations from real data are needed. For example, the anomaly zone has been established for several methods [7,18]. However, the question that practitioners face is: do their data fall within an anomaly zone for a specific method?

Last but not least, when the evolutionary history is reticulate, it is more appropriate to speak of a phylogenetic, or species, network, rather than a species tree. In the population genetics literature, this issue has long been recognized, and ancestral recombination graphs (a class of phylogenetic networks) have been adopted for modeling genealogies that include recombination [98]. Using phylogenetic networks and, more generally, networks, might help uncover hypotheses that would be undetected otherwise [99]. The different flavors in which phylogenetic networks come might be confusing to the community of practitioners and, consequently, have limited their applicability. Recent monograms have been written to clarify the similarities, differences, and applications of the various types of network [23,26,27]. Developments to address the issues above should be applicable to phylogenetic networks.

Acknowledgments

I would like to acknowledge the three anonymous reviewers for extensive and thorough comments on the first revision of this manuscript, which helped improve it significantly in terms of content and organization. Furthermore, I acknowledge Siavash Mirarab, Noah Roseberg, and Tandy Warnow for comments on the first revision. This work was supported in part by National Science Foundation (NSF) grants DBI-1062463 and CCF-130217, grant R01LM009494 from the National Library of Medicine, an Alfred P. Sloan Research Fellowship, and a Guggenheim Fellowship to L.N. The contents are solely the responsibility of the author and do not necessarily represent the official views of the NSF, National Library of Medicine, the National Institutes of Health, the Alfred P. Sloan Foundation, or the John Simon Guggenheim Memorial Foundation.

References

1 Dittmar, K. and Liberles, D., eds (2010) *Evolution after Gene Duplication*, Wiley-Blackwell
2 Innan, H. and Kondrashov, F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108
3 Lerat, E. *et al.* (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* 3, 0807–0814
4 Boto, L. (2010) Horizontal gene transfer in evolution: facts and challenges. *Proc. R. Soc. B* 277, 819–827
5 Abbott, R. and Rieseberg, L. (2012) Hybrid speciation. *eLS* http://dx.doi.org/10.1002/9780470015902.a0001753.pub2
6 Baack, E. and Rieseberg, L. (2007) A genomic view of introgression and hybrid speciation. *Curr. Opin. Genet. Dev.* 17, 513–518
7 Degnan, J. and Rosenberg, N. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340
8 Siepel, A. (2009) Phylogenomics of primates and their ancestral populations. *Genome Res.* 19, 1929–1941
9 Wendel, J.F. and Doyle, J.J. (1998) Phylogenetic incongruence: window into genome history and molecular evolution. In *Molecular Systematics of Plants II* (Soltis, D. *et al.*, eds), pp. 265–296, Springer
10 Goodman, M. *et al.* (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* 28, 132–163
11 Maddison, W. (1997) Gene trees in species trees. *Syst. Biol.* 46, 523–536
12 Knowles, L. and Kubatko, L., eds (2010) *Estimating Species Trees: Practical and Theoretical Aspects*, Wiley–Blackwell
13 Than, C. and Nakhleh, L. (2009) Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.* 5, e1000501
14 Eulenstein, O. *et al.* (2010) Reconciling phylogenetic trees. In *Evolution after Gene Duplication* (Dittmar, K. and Liberles, D., eds), pp. 185–206, Wiley-Blackwell
15 Zhang, L. (2011) From gene trees to species trees II: species tree inference by minimizing deep coalescent events. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 1685–1691
16 Maddison, W.P. and Knowles, L.L. (2006) Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55, 21–30
17 Than, C. and Nakhleh, L. (2010) Inference of parsimonious species phylogenies from multi-locus data by minimizing deep coalescences. In *Estimating Species Trees: Practical and Theoretical Aspects* (Knowles, L. and Kubatko, L., eds), pp. 79–98, Wiley–Blackwell
18 Than, C.V. and Rosenberg, N.A. (2011) Consistency properties of species tree inference by minimizing deep coalescences. *J. Comput. Biol.* 17, 1–15
19 Bayzid, M. *et al.* (2013) Inferring optimal species trees under gene duplication and loss. *Pac. Symp. Biocomput.* 18, 250–261

20 Bordewich, M. and Semple, C. (2004) On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Comb.* 8, 409–423
21 Bordewich, M. and Semple, C. (2007) Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Appl. Math.* 155, 914–928
22 Humphries, P. *et al.* (2013) On the complexity of computing the temporal hybridization number for two phylogenies. *Discrete Appl. Math.* 161, 871–880
23 Nakhleh, L. (2010) Evolutionary phylogenetic networks: models and issues. In *The Problem Solving Handbook for Computational Biology and Bioinformatics* (Heath, L. and Ramakrishnan, N., eds), pp. 125–158, Springer
24 Morrison, D.A.D. (2005) Networks in phylogenetic analysis: new tools for population biology. *Int. J. Parasito.* 35, 567–582
25 Huson, D. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Bio. Evol.* 23, 254–267
26 Huson, D. *et al.* (2011) *Phylogenetic Networks: Concepts, Algorithms, and Applications*, Cambridge University Press
27 Morrison, D. (2011) *Introduction to Phylogenetic Networks*, RJR Productions
28 Kanj, I. *et al.* (2008) Seeing the trees and their branches in the network is hard. *Theor. Comput. Sci.* 401, 153–164
29 Huson, D. and Rupp, R. (2008) Summarizing multiple gene trees using cluster networks. *Lect. Notes Bioinform.* 5251, 296–305
30 Beiko, R. and Ragan, M. (2009) Untangling hybrid phylogenetic signals: horizontal gene transfer and artifacts of phylogenetic reconstruction. *Methods Mol. Biol.* 532, 241–256
31 van Iersel, L. *et al.* (2010) Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. *Bioinformatics* 26, i124–i131
32 Wu, Y. (2010) Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics* 26, 140–148
33 Park, H. *et al.* (2010) Algorithmic strategies for estimating the amount of reticulation from a collection of gene trees. In *Proceedings of the Ninth Annual International Conference on Computational Systems Biology*. pp. 114–123, Association for Computing Machinery
34 Park, H. and Nakhleh, L. (2012) MURPAR: a fast heuristic for inferring parsimonious phylogenetic networks from multiple gene trees. *Lect. Notes Bioinform.* 7292, 213–224
35 Wu, Y. (2013) An algorithm for constructing parsimonious hybridization networks with multiple phylogenetic trees. *Lect. Notes Comput. Sci.* 7821, 291–303
36 Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376
37 Degnan, J.H. and Salter, L.A. (2005) Gene tree distributions under the coalescent process. *Evolution* 59, 24–37
38 Rannala, B. and Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164, 1645–1656
39 Kubatko, L. *et al.* (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25, 971–973
40 Wu, Y. (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66, 763–775
41 Akerborg, O. *et al.* (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5714–5719
42 Górecki, P. *et al.* (2011) Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics* 12, S15
43 Konrad, A. *et al.* (2011) Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biol. Evol.* 3, 1197
44 Green, R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328, 710–722
45 Eriksson, A. and Manica, A. (2012) Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominids. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13956–13960
46 Staubach, F. *et al.* (2012) Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (*Mus musculus*). *PLoS Genet.* 8, e1002891
47 Consortium, T.H.G. (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487, 94–98
48 Moody, M. and Rieseberg, L. (2012) Sorting through the chaff, nDNA gene trees for phylogenetic inference and hybrid identification of annual sunflowers (*Helianthus* sect *Helianthus*). *Mol. Phylogenet. Evol.* 64, 145–155
49 Yu, Y. *et al.* (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* 8, e1002660
50 Rasmussen, M. and Kellis, M. (2012) Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* 22, 755–765
51 Kamneva, O.K. *et al.* (2012) Analysis of genome content evolution in PVC bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biol. Evol.* 4, 1375–1390
52 Bansal, M. *et al.* (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics* 28, i283–i291
53 Yu, Y. *et al.* (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.* 60, 138–149
54 Yu, Y. *et al.* (2013) Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Syst. Biol.* 62, 738–751
55 Stolzer, M. *et al.* (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28, i409–i415
56 Jain, R. *et al.* (2003) Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* 20, 1598–1602
57 Cohen, O. *et al.* (2011) The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* 28, 1481–1489
58 Stiller, J.W. (2011) Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. *BMC Evol. Biol.* 11, 259
59 Hughes, T. and Liberles, D.A. (2008) The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families. *Gene* 414, 85–94
60 Sjöstrand, J. *et al.* (2012) DLRS: gene tree evolution in light of a species tree. *Bioinformatics* 28, 2994–2995
61 Than, C. *et al.* (2007) Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J. Comput. Biol.* 14, 517–535
62 Meng, C. and Kubatko, L.S. (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Popul. Biol.* 75, 35–45
63 Kubatko, L.S. (2009) Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.* 58, 478–488
64 Jones, G. *et al.* (2013) Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst. Biol.* 62, 467–478
65 Marcussen, T. *et al.* (2012) Inferring species networks from gene trees in high-polyploid north American and Hawaiian violets (*Viola*, Violaceae). *Syst. Biol.* 61, 107–126
66 Hahn, M. (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* 8, R141
67 Yang, J. and Warnow, T. (2011) Fast and accurate methods for phylogenomic analyses. *BMC Bioinformatics* 12, S4
68 Berglund-Sonnhammer, A.C. *et al.* (2006) Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J. Mol. Evol.* 63, 240–250
69 Durand, D. *et al.* (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* 13, 320–335
70 Than, C. and Nakhleh, L. (2008) SPR-based tree reconciliation: non-binary trees and multiple solutions. *Ser. Adv. Bioinform. Comput. Biol.* 6, 251–260
71 Yu, Y. *et al.* (2011) Algorithms for MDC-based multi-locus phylogeny inference. *Lect. Notes Bioinform.* 6577, 531–545

72 Yu, Y. *et al.* (2011) Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. *J. Comput. Biol.* 18, 1543–1559
73 Than, C. *et al.* (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9, 322
74 Than, C. *et al.* (2008) Integrating sequence and topology for efficient and accurate detection of horizontal gene transfer. *Lect. Notes Bioinform.* 5267, 113–127
75 Park, H. *et al.* (2010) Bootstrap-based support of HGT inferred by maximum parsimony. *BMC Evol. Biol.* 10, 131
76 Ané, C. *et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24, 412–426
77 Heled, J. and Drummond, A. (2010) Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27, 570–580
78 Jin, G. *et al.* (2007) Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Mol. Biol. Evol.* 24, 324–337
79 Jin, G. *et al.* (2006) Maximum likelihood of phylogenetic networks. *Bioinformatics* 22, 2604–2611
80 Rokas, A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804
81 Kubatko, L. and Degnan, J. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56, 17–24
82 Wu, C.I. (1991) Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* 127, 429–435
83 Mossel, E. and Roch, S. (2010) Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7, 166–171
84 Liu, L. and Pearl, D.K. (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56, 504–514
85 Joly, S. *et al.* (2009) A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am. Nat.* 174, E54–E70
86 Holland, B. *et al.* (2008) Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol. Biol.* 8, 202
87 Arenas, M. *et al.* (2008) Characterization of reticulate networks based on the coalescent with recombination. *Mol. Biol. Evol.* 25, 2517–2520
88 Lanier, H. and Knowles, L. (2012) Is recombination a problem for species-tree analyses? *Syst. Biol.* 61, 691–701
89 Bapteste, E. *et al.* (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18266–18272
90 Wu, Y.C. *et al.* (2012) Evolution at the subgene level: domain rearrangements in the *Drosophila* phylogeny. *Mol. Biol. Evol.* 29, 689–705
91 Freeling, M. and Thomas, B.C. (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16, 805–814
92 Slatkin, M. and Pollack, J.L. (2006) The concordance of gene trees and species trees at two linked loci. *Genetics* 172, 1979–1984
93 Hobolth, A. *et al.* (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* 3, e7
94 Dutheil, J.Y. *et al.* (2009) Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183, 259–274
95 Huang, H. *et al.* (2010) Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.* 59, 573–583
96 Chung, Y. and Ané, C. (2011) Comparing two Bayesian methods for gene tree/species tree reconstruction: a simulation with incomplete lineage sorting and horizontal gene transfer. *Syst. Biol.* 60, 261–275
97 Knowles, L. *et al.* (2012) Full modeling versus summarizing gene-tree uncertainty: method choice and species-tree accuracy. *Mol. Phylogenet. Evol.* 65, 501–509
98 Griffiths, R. and Marjoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* 3, 479–502
99 Bapteste, E. *et al.* (2013) Networks: expanding evolutionary thinking. *Trends Genet.* 29, 439–441