*Syst. Biol.* 64(1):e42–e62, 2015

DOI:10.1093/sysbio/syu048
Advance Access publication July 28, 2014

# The Inference of Gene Trees with Species Trees

GERGELY J. SZÖLLŐSI[1], ERIC TANNIER[2,3,4], VINCENT DAUBIN[2,3], AND BASTIEN BOUSSAU[2,3,*]

[1]*ELTE-MTA "Lendület" Biophysics Research Group, Pázmány P. stny. 1A., 1117 Budapest, Hungary;* [2]*Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France;* [3]*Université de Lyon, F-69000 Lyon, France; and* [4]*Institut National de Recherche en Informatique et en Automatique Rhône-Alpes, F-38334 Montbonnot, France;*

**Correspondence to be sent to: Bastien Boussau, Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France; Université de Lyon, F-69000 Lyon, France; E-mail: bastien.boussau@univ-lyon1.fr.*

*Received 30 October 2013; reviews returned 11 July 2014; accepted 14 July 2014*
*Associate Editor: Tanja Stadler*

*Abstract*.—This article reviews the various models that have been used to describe the relationships between gene trees and species trees. Molecular phylogeny has focused mainly on improving models for the reconstruction of gene trees based on sequence alignments. Yet, most phylogeneticists seek to reveal the history of species. Although the histories of genes and species are tightly linked, they are seldom identical, because genes duplicate, are lost or horizontally transferred, and because alleles can coexist in populations for periods that may span several speciation events. Building models describing the relationship between gene and species trees can thus improve the reconstruction of gene trees when a species tree is known, and vice versa. Several approaches have been proposed to solve the problem in one direction or the other, but in general neither gene trees nor species trees are known. Only a few studies have attempted to jointly infer gene trees and species trees. These models account for gene duplication and loss, transfer or incomplete lineage sorting. Some of them consider several types of events together, but none exists currently that considers the full repertoire of processes that generate gene trees along the species tree. Simulations as well as empirical studies on genomic data show that combining gene tree–species tree models with models of sequence evolution improves gene tree reconstruction. In turn, these better gene trees provide a more reliable basis for studying genome evolution or reconstructing ancestral chromosomes and ancestral gene sequences. We predict that gene tree–species tree methods that can deal with genomic data sets will be instrumental to advancing our understanding of genomic evolution. [Algorithm; amalgamation; Bayesian inference; birth–death model; coalescent; dynamic programming; gene duplication; gene loss; gene transfer; gene tree; hybridization; maximum likelihood; phylogenetics; species tree.]

> *Toutes choses étant causées et causantes, aidées et aidantes, médiates et immédiates, et toutes s'entretenant par un lien naturel et insensible qui lie les plus éloignées et les plus différentes, je tiens impossible de connaître les parties sans connaître le tout non plus que de connaître le tout sans connaître particulièrement les parties.* (Pascal 1669).

During the last 50 years, phylogeny has become more and more based on molecular data, increasingly favoring homologous sequences over morphological characters. This approach has been extremely fruitful, producing constant improvement in the accuracy and resolution of phylogenetic reconstruction together with our understanding of evolutionary processes at the molecular level. However, we have known all along that we are barking up the wrong trees: with increasing sophistication in the models of sequence evolution, we have been reconstructing trees describing the history of fragments of genomic sequence, which we will liberally call "gene" in this review, but never the history of species. Gene trees are not species trees (Maddison 1997).

Each gene tree reflects a unique story, which is linked to species history, but often significantly differs from it (Szöllősi and Daubin 2012). Gene trees reflect the process of replication at a local level: a copy of a gene at a locus in the genome, for example, a protein coding gene, replicates and its copies are passed on from parent to offspring, generating branching points in the gene tree. Because each gene copy has a single ancestral copy, barring recombination, all gene trees would be identical. Recombination, however, breaks up the genomic history into a series of partially independent stories, that is into gene trees along the genomes of species.

Starting from an individual site in a genome up to the species level, a hierarchy of evolutionary processes generate genomic sequences. Individual sites evolve as a result of point mutations. The fate of individuals carrying each mutation is played out at the population level, and determines whether a mutation is fixed in the population as a substitution, or is ultimately lost. The birth and death of stretches of sequence, *e.g.* of single sites or even of entire genes, occurs as a result of insertions and deletions in individual genomes, the fate of which, similar to point mutations, is played out at the population level. The source of the inserted sequence differentiates between duplication events, wherein a sequence from the same genome is inserted, and lateral transfer events, wherein a sequence from an external source is inserted. Finally, species, that is, populations of genomes, evolve through speciation and extinction events.

As illustrated in Figure 1 each level of the hierarchy contributes to generating phylogenetic signal that can lead to differences between reconstructed gene trees. Segregating mutations that cross speciation events (a process called incomplete lineage sorting) leave

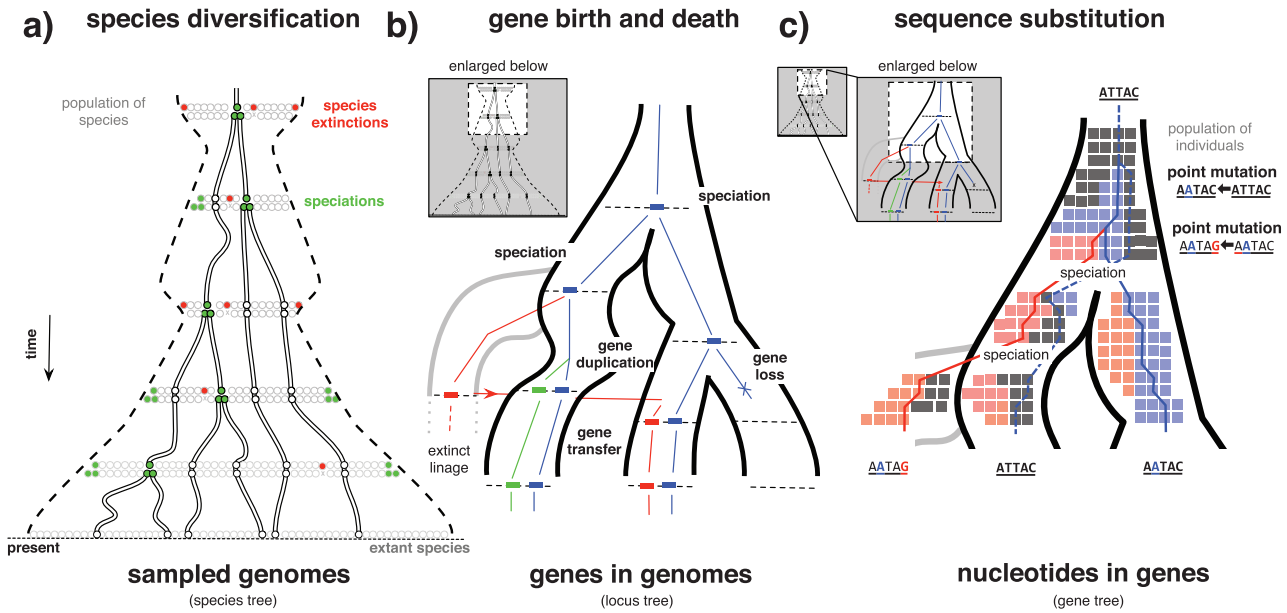


FIGURE 1. A hierarchy of evolutionary processes contribute to sequence evolution. a) Individual species (circles) and their genomes evolve among a population of species, according to a diversification process consisting of speciation (light gray, green online) and extinction (dark gray, red online) events. The variation in the number of species existing at any given time is indicated by the dashed contour. When attempting to infer the species tree typically only a fraction of existing species (gray and black circles on dashed line) are sampled (black circles). b) Inside each genome, each gene evolves according to gene duplication, loss, and transfer events. c) Individual sites evolve through point mutations. Processes at the gene and site level are played out at the population level, where changes fix or are lost.

topological signatures in gene trees (see Fig. 1c). Current estimates indicate that up to 30% of the sequence of the human genome is more closely related to Gorilla than to Chimpanzee due to this process (Scally et al. 2012). Duplication, transfer, and loss events (Fig. 1b) lead to large differences in both the size and phylogenetic distribution of families of homologous genes, and at the same time produce patent phylogenetic discord (Szöllősi and Daubin 2012). Finally, the species diversification processes influence lateral transfer, as most transfer events come from donor species that have gone extinct or have not been sampled (Szöllősi et al. 2013b).

When reconstructing a gene tree it is desirable to integrate these different types of information in order to maximize the amount of information used and to insure the consistency of our prediction. In a pioneering attempt to do so, Goodman et al. (1979) proposed to reconstruct the history of a gene by searching for the tree that minimizes the sum of the number of nucleotide substitutions, duplications, and losses. This parsimonious approach is clear and conceptually straightforward, but it raises the difficulty of determining the relative weights of events that are very different in nature. However, if we can construct a coherent and principled approach to combine these events, it becomes possible to envisage the reconstruction of a gene tree given a sequence alignment and a known species tree (here we consider sequence alignments, but unaligned sequences could also be considered if a model of insertion–deletion is also used to jointly infer the alignment and the gene tree). Moreover, given a set of gene alignments, it would allow us to reconstruct the species history that has most likely generated them. In a probabilistic framework, a model of sequence evolution and a model of gene family evolution accounting for duplication, loss, lateral transfer, and/or incomplete lineage sorting can be combined using the intuitive hypothesis of conditional independence (we then obtain the "Felsenstein equation" Felsenstein 1988). The probabilities of the two models can be multiplied under the following two assumptions: (i) the species tree is independent from the sequences conditional on the gene tree, that is, if we are given the gene tree we do not need additional information on the species tree to compute the probability of the sequences and, (ii) if we are given the gene tree we do not need the sequences to compute the probability of a species tree.

In addition, parsimony approaches often generate a large number of equivalent solutions and do not allow an efficient exploration or integration over solution spaces. Probabilistic models allow integrating or sampling over the very large number of possible scenarios. For example, it is possible to estimate the probability of a particular gene tree, which is not only the probability of the most likely scenario of sequence evolution and events of gene duplication, transfer, and loss, but the sum of the probabilities of all possible scenarios that could have generated this tree given the species tree, and the sum of the probabilities of all substitution histories consistent with the gene tree that could have generated the sequence alignment.

In the past 15 years, several such methods, which model consistently the dependence between gene trees and the species tree, have been developed and have

shown improved accuracy for inferring both gene trees and the species tree. In this review we present these methods, explain the assumptions they make, introduce how they work, and highlight some of the results obtained with them. We focus on probabilistic models, but discuss parsimony-based approaches in situations where probabilistic models have not been developed yet. We do not review methods that consider gene count or gene presence/absence information, as they altogether ignore sequence information once homology relationships have been defined (some of these methods do account for several processes however, for example, gene duplication, transfer, and loss (Csuros et al. 2006)).

## Modeling the Dependence between Gene Tree and Species Tree

A gene family can contain genes from different species at the same locus, or genes in a same genome at different loci. The processes known to contribute to gene family evolution include speciation and lineage sorting (ILS if incomplete), gene duplication and loss (DL), and gene transfer (T). Lineage sorting concerns genes from different genomes at the same locus, whereas duplications give rise to homologous genes on the same genome at different loci. Transfers can insert a gene at a new locus, or replace a homologous gene at its locus. Hybridization can be seen as a special type of transfer, affecting a large portion of the genome, and resulting in a gene replacement in the receiving species. Allopolyploidization is a particular type of hybridization, in which the two genomes keep cohabiting in subsequent generations. For each individual process, there are published models accounting for its effect, and recently some tend to integrate several of them. So far, no model has been published that deals with all processes together in a coherent statistical framework.

### *Gene Birth–Death Generates Gene Trees Along the Species Tree*

Irrespective of whether they deal with ILS, DL, or T, all models of gene family evolution can be seen as generating a tree inside a tree, that is a gene tree inside a species tree. In this respect, the models encountered in the literature dealing with processes of speciation and extinction (Rannala and Yang 1996; Morlon et al. 2009) are very similar to the models of gene family evolution. They both invoke birth–death processes that generate a rooted, time-like tree topology. Birth events correspond to bifurcations, death events correspond to the loss of a lineage. In diversification models, lineages correspond to species, in models of gene family evolution they correspond to genes. After the birth–death process arrives at the present, lineages with no descendant among extant species are usually pruned, and the remaining lineages constitute the generated tree topology. Models of gene family evolution, however, are constrained by the species tree, whereas species diversification models are not. The species tree constitutes a set of constraints corresponding to speciation events and branch lengths that control the birth–death process generating the gene family tree. A gene family evolution model is in essence a series of birth–death models fitted piecewise along the branches of a species tree (Fig. 2). The birth–death process generating the gene tree starts above the root of the species tree. Each time the birth–death process reaches a speciation node, two new processes are created in the children lineages. These processes can be identical, or can have different parameters. In general, for $n$ branches in the species tree, counting the branch above the root, there are $n$ independent birth–death processes. Of course, the parameters of these $n$ independent processes do not need to be independent: one can imagine that the birth parameter for instance evolves according to for example a Brownian motion process running along the species tree. Such a model would penalize large jumps in the birth parameter between neighboring branches of the species tree, but to our knowledge such an idea has not yet been implemented. Another source of dependence between processes is lateral gene transfer: a birth in a lineage may originate in another.

The above describes how a gene tree, complete with branch lengths in units of time, is generated along the clock-like branches of the species tree. In practice, to simplify the problem, we will see that several methods choose to consider only the topologies of gene trees, that is the branch length information is discarded. In this case, the mathematical machinery of the birth death model is not used to compute the probability of a specific dated scenario of observed birth and death events; instead it is used to compute the probability of a given succession of birth and death events (Degnan and Salter 2005; Wu 2011), or the probability of observing $k$ genes at the beginning of a branch of the species tree, and $l$ genes at the end of this same branch (Boussau et al. 2013). The choice of discarding branch length information in the gene tree in some cases simplifies the problem, because fewer processes need to be modeled. Potentially useful information, however, is discarded in the process.

### *The Coalescent Models Population-level Processes Along the Species Tree*

Coalescent models aimed at modeling the discordance between gene tree and species tree arising from population-level processes have enjoyed increasing popularity in the last 10 years. Here birth events correspond to the appearance of a new allele, and death events to the disappearance of an allele, without any change in the locus of the gene. At any given time in a species, for a given locus in the genome, there may be several alleles. These alleles have their own history, some alleles being more closely related than others. When speciation occurs, most alleles will be sorted randomly

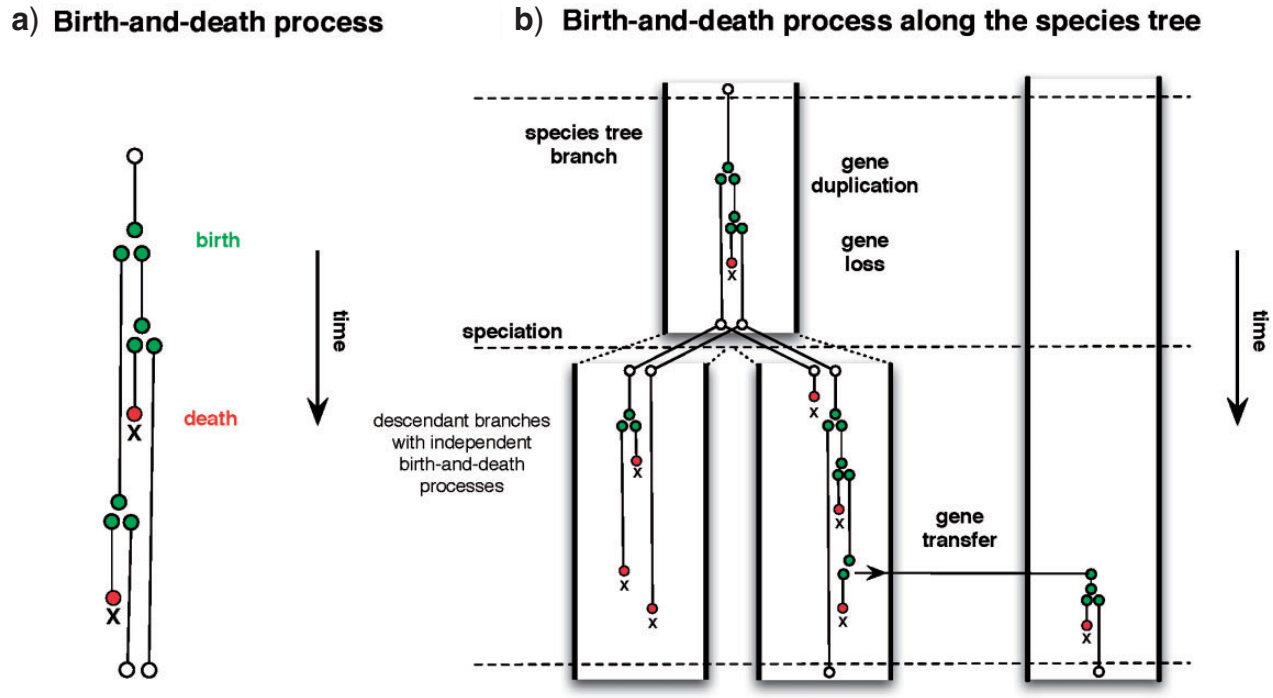


FIGURE 2. Birth–death processes for generating species trees and gene trees. Death events (species extinctions and gene loss) are in dark gray (red online), birth events (speciation, duplication and transfer events) are light gray (green online). a) Birth–death processes modeling speciation and extinction. b) Birth–death process modeling gene family evolution inside a species tree.

between the two incipient species: in some cases both species will receive copies of all alleles, in others, each will receive only a subset of the alleles present in the parent population. In all cases, the history reconstructed from the allele sequences will be the allele history, and not the species history. The allele history and the species history always differ in the timing of the bifurcation events: assuming no hybridization has occurred between the lineages, alleles have necessarily split before species split. They can also differ in their topology, especially if only a brief interval of time passes between successive speciation events, and/or the effective population size of the parent species is large (Rosenberg and Nordborg 2002). Given the coalescent model, the amount of discordance in topology and divergence times between the trees of several loci and a species tree can therefore be used to estimate effective population sizes along the species tree (Rannala and Yang 2003; Liu and Pearl 2007; Heled and Drummond 2008; Minin et al. 2008; Kubatko et al. 2009). Such a model where the population size is assumed to differ between the branches of the species tree has been called the multispecies coalescent (Rannala and Yang 2003), and is now widely used to infer species trees given several loci, with one allele per locus (Liu and Pearl 2007; Heled and Drummond 2008; Kubatko et al. 2009), or several alleles per locus (Liu et al. 2008; Heled and Drummond 2010), using branch lengths information in the gene trees (Liu and Pearl 2007; Heled and Drummond 2008), or instead relying solely on their topologies (Degnan and Salter 2005; Wu 2011), in the maximum likelihood (Kubatko et al. 2009; Wu 2011) or Bayesian frameworks (Liu and Pearl 2007; Heled and Drummond 2008; Liu et al. 2008; Heled and Drummond 2010; Minin et al. 2008). Hey (2010) has developed an extension of the multispecies coalescent that allows migration between branches of the species tree. The multispecies coalescent has also been used in conjunction with a Hidden Markov Model (HMM) running along a chromosome to infer recombination rates, locus-specific trees, divergence times and effective population sizes (Hobolth et al. 2007, 2011; Li and Durbin 2011; Mailund et al. 2011; Scally et al. 2012), given a species tree and a genomic alignment. In this model, gene trees can change along the alignment, and how often the gene trees change provides information on the rate of recombination.

One assumption of all the models we have presented so far is that sequences are assigned to species without ambiguity. However, for recently diverged populations, it is often unclear whether two distinct populations should be considered as coming from the same single species, or should define two distinct species. In a given sample, it can also be unclear how many species have really been sampled. To address such cases, Rannala and Yang have developed a model for delimiting species given a set of sequences for different loci and several individuals (Yang and Rannala 2010; Rannala and Yang 2013). They assume a multispecies coalescent model, and postulate that hybridization does not occur after speciation. Information coming from for example geographical distribution, morphology, or behavior can be introduced into the model thanks to a prior on the

expected number of species, but also through a guide tree, whose other purpose is to improve the efficiency of the MCMC algorithm used for inference. Under this model, they can analyze about 10 species for 100 sequences, with a finite number of loci. They apply this method to well studied data sets of asexual rotifers and fence lizards, and recover the species found by other means.

### *Models of Gene Duplication and Loss*

Models of gene duplication and loss usually ignore population-level processes (but see Rasmussen and Kellis (2012), discussed below) that drive the fixation or disappearance of an allele, and only consider events of gene duplication and loss that have fixed in the species. In this setting, birth events correspond to fixed gene duplications, and death events to fixed gene losses. Probabilistic models for gene duplication and loss were first proposed by Arvestad (2003), and further developed in subsequent papers by the same group (Arvestad et al. 2004; Akerborg et al. 2009; Sjöstrand et al. 2012) and by a few others (Dubb 2005; Rasmussen and Kellis 2007, 2010). The focus of these works was to infer gene trees given a fixed species tree, with clock-like branch lengths in units of time, and with fixed rates of gene duplication and gene loss over the entire tree. Combined with the birth–death model of gene evolution is a hierarchical model of the rate of sequence evolution, wherein the species tree provides dates, and each gene family is associated with one or several rates. Alternatively, Górecki et al. (2011); Górecki and Eulenstein (2013) developed another model for gene duplication, not based on a birth–death process, but based on a Poisson process for computing the probability of a parsimonious reconciliation of a gene tree topology against a species tree. The gene tree does not need to have branch lengths, but the species tree does. For this reason, and because it does not include a loss parameter, this model misses some of the realism of the birth death processes described above, but gains in speed.

More recently, we modified birth–death models to allow different duplication/loss parameters for each branch of a nondated species tree (Boussau et al. 2013). To speed-up computations, we took an approach similar to Górecki et al. (2011), and did not account for the branch lengths of gene trees in their reconciliation with the species tree. Instead we only reconciled topologies. However, because our hierarchical model includes a model of sequence evolution for joint inference of gene trees and species tree, we still needed to estimate branch lengths in the gene trees. To simplify the problem, we chose not to have a hierarchical model of rates of sequence evolution: rates for each gene family were considered to be entirely independent. This decreased the number of global parameters to estimate, but increased drastically the number of gene family-specific parameters to estimate.

### *Models of Lateral Gene Transfer*

Lateral gene transfer (LGT) corresponds to the incorporation in a genome of a gene coming from a different species. There is compelling evidence that this process has played an important role in the evolution of life, particularly in the domains of Bacteria and Archaea (Abby et al. 2012). Several models have been proposed to account for LGT in gene tree–species tree reconciliation. So far they all consider events that have fixed and ignore population-level processes. One key feature of these models is whether they consider or not the possibility of gene replacement. The recipient of a transfer may, or may not have a gene homologous to the incoming gene in its genome. If it has, the gene in the recipient can be either conserved or lost. Transfer-only models usually consider only gene replacement. Models that do not make this assumption are often coupled with duplications and therefore are presented in the subsequent section. The difference between the two is important in that gene replacement is not modeled by birth and death. Indeed, replacing a lineage by a gene coming from another breaks the independence assumption between lineages, which is at the basis of efficient computations in birth–death models. Thus, models of gene replacement differ from the more mainstream birth–death models.

The first attempt at modeling probabilistically lateral gene replacement explicitly was made by Suchard (2005), but a model for host–parasite cophylogeny developed by Huelsenbeck et al. (2000) could also be used to detect gene transfer between two loci. This model assumed that the parasite phylogeny differed from the host phylogeny through a Poisson-distributed number of host-switches, or replacements in our case. Transfer events could be placed uniformly among branches, or preferentially among branches close to each other in the host tree. Rates of evolution were independent in the host tree and the parasite tree, but sequence evolution was assumed to follow a strict clock model. Inference was conducted in a Bayesian framework employing MCMC, and resulted in a host tree distribution, a parasite tree distribution, and distributions of the number and rates of host switches.

Suchard (2005) proposed a model specifically designed to tackle gene replacement on a gene tree topology, discarding branch length information. Computing the probability of a gene tree given a species tree therefore did not involve mapping the gene tree onto the species tree. Instead, this method involved estimating how many Subtree Prune and Regraft (SPR) moves (Hordijk and Gascuel 2005) were necessary to explain the topological difference between a species tree and a gene tree (another type of move was also considered). The probability of a gene tree was then based on this number of moves, through a Poisson model. Using this model, Suchard (2005) could estimate the species tree that best explained a forest of over 140 gene trees. However, the approach was limited to trees with only 6–8 species because of its computational cost. In addition, this approach does not make sure that all transfers detected in a set of gene families

are time-consistent: in principle a gene can only be transferred to contemporaneous species, present in the sample analyzed or not, and certainly not to more ancient species. In Suchard (2005)'s approach, because the species tree topology is not anchored in time, time-consistent transfers and "back-to-the-future" type transfers are not distinguished.

Bloomquist and Suchard (2010) chose another road to modeling gene replacement, by drawing on models of population genetics. They considered the Ancestral Recombination Graph (ARG). An ARG is a type of rooted network that combines both vertical and transfer edges. Once an ARG is built, it can be used to generate dated gene trees, which correspond to tree-like paths obtained by selecting edges of the ARG. They aimed at reconstructing an ARG that represents all of the evolutionary histories of a set of distinct loci, some of which evolved along the species tree, and some of which underwent replacement events. They used MCMC to propose ARGs, adding or subtracting transfer edges through reversible jumps. Given an ARG, a gene tree is drawn for each locus under study; these gene trees can be totally independent of each other, or can incorporate spatial information, so that two neighboring genes on the genome are more likely to be transferred together than distant genes. This allows modeling different situations, such as single-gene conversion or homologous recombination. In addition, the sequences of different genes can evolve at different rates. In the end, the method builds dated gene trees, and a dated ARG in which vertices are annotated as vertical or transfer nodes, and in which edges involved in a transfer event can be annotated with the genes that are inferred to have been transferred.

### *Models that Combine the Above*

We know of five probabilistic models that combine some of the processes listed above: the DTLSR model of Tofigh (2009), the DLCoal model of Rasmussen and Kellis (2012), our ODT model (Szöllősi et al. 2012, 2013b), network-based models of hybridization with incomplete lineage sorting Than et al. (2007); Meng and Kubatko (2009); Kubatko et al. (2009); Yu et al. (2012), and network-based models of polyploidization (Jones et al. 2013). Finally we discuss Bucky (Ane et al. 2007), a method that does not assume any particular process to explain the difference between gene tree and species tree but instead clusters gene families according to how similar their evolutionary history is.

The DLCoal model combines a coalescent model with a model of gene duplication and loss. In doing so, it reintroduces population genetics concepts into the framework of DL models: for instance, it acknowledges that a new duplicate first has a very low frequency in a population, as it is present only in the individual where the gene duplication occurred. Under this model, the reconciliation of a gene tree with a species tree requires three objects: first, a dated species tree, with branches annotated with effective population sizes. Second, a dated locus tree, generated thanks to a birth–death process placing events of duplications and losses along the species tree branches, according to a duplication rate and a loss rate. The locus tree contains implicit information about chromosomal positions, as under this model chromosomal position changes at duplication nodes, but is generated according to the same birth–death process used in previous duplication-loss models (Arvestad 2003; Arvestad et al. 2004; Dubb 2005; Rasmussen and Kellis 2007; Akerborg et al. 2009; Rasmussen and Kellis 2010; Sjöstrand et al. 2012). Third, the gene tree (black trees in Fig. 3) is generated according to a coalescent process running along the locus tree (blue trees in Fig. 3). The DLCoal model makes the simplifying assumption, termed the "hemiplasy assumption", wherein all duplications and losses are considered to either always go extinct or never go extinct in all descendant lineages. This assumption allows

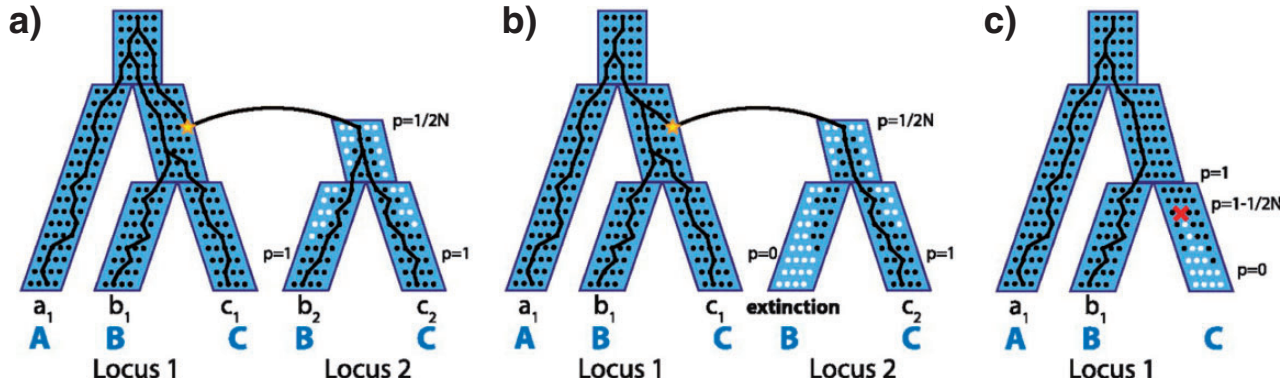


FIGURE 3. Duplication and loss events within a multispecies coalescent with the locus tree in gray (blue online) and the gene tree in black. a) A duplication occurs in one chromosome and creates a new locus, locus 2 in the genome. At locus 2, the Wright–Fisher model dictates how the frequency p of the daughter duplicate (black dots) competes with the null allele (white dots) until it eventually fixes (p = 1). A gene tree is therefore a traceback in this combined process. b) A new duplicate can undergo hemiplasy, and fixes in some lineages and goes extinct in others. c) Similar to duplication, a gene loss (deletion) starts in one chromosome and drifts until it fixes or goes extinct. Reproduced from Figure 1 in Rasmussen M.D., and Kellis M., Genome Res. 2012;22:755–765.

the separation of the duplication–loss process from the multilocus coalescent. Aside of this assumption, the model entails two successive tree-embedded birth–death processes and can be regarded as a multispecies colaescent process taking place along the locus tree which in turn is generated along the species tree by a DL process (cf. Fig. 3). Rasmussen and Kellis (2012) implemented this model into a program that can reconcile a gene tree with a species tree, given rates of duplication and loss, a dated species tree, and effective population sizes.

The DTLSR (Tofigh 2009; Tofigh et al. 2011; Sjöstrand 2013) and ODT models combine a model of gene transfer with a model of gene duplication and loss. Both are natural extensions of the DL models, and do not employ an additional object like the locus tree in the DLCoal model. Instead, the birth–death process is modified to accommodate two types of birth events, gene duplications and gene transfers. The ODT model has been applied to biological datasets, and has been extended to account for gene transfers that involve species that have gone extinct or are otherwise unrepresented in the tree under consideration (Szöllősi et al. 2013b). In this extension of the ODT model, the ex-ODT model, sampled species and their ancestors are assumed to come from a population of species, which at any given time contains many more species than are present in the sample. This population of species evolves according to the Moran process, but could evolve according to more complex processes (e.g., (Morlon et al. 2009)). Within this framework, a gene can be transferred from the genome of an ancestor of a sampled species to the genome of a species living at the same time, but which gave no descendant among the sampled species. Then this gene can evolve among the genomes of species that gave no sampled descendant, including through transfers, duplications, losses and speciations, and finally reintegrate genomes with sampled descendants.

Contrary to the DLCoal model, neither the ODT models or the DTLSR model account for population-level processes, and thus do not model allele fixation. However, the two model types could be easily mixed, for example the ex-ODT model could be used to generate a locus tree, which would then be used to generate a gene tree according to the coalescent model. This ex-DTLCoal model would then account for speciations, extinctions, duplications, losses, transfers, and incomplete lineage sorting in a hierarchical probabilistic model (cf. Fig. 4).

Meng and Kubatko (2009); Kubatko et al. (2009); Yu et al. (2012) propose a model for the detection of hybridization in the presence of incomplete lineage sorting, extending early efforts by Than et al. (2007). As we noted above, from a modeling point of view, hybridization may be seen as a high frequency of gene replacements between two lineages. There, a rooted phylogenetic network is used instead of a phylogenetic tree. In a hybrid genome, any gene is assumed to be coming from one of two parental genomes. Hybridization nodes are thus represented in such a network by nodes with two parents $A$ and $B$. A probability $\gamma$ indicates the proportion of genes coming from parent $A$, the rest coming from parent $B$. Under this model, evolution of alleles within the network is very similar to their evolution along a species tree. At nodes with a single parent, the usual multispecies coalescent model applies, with the length of the branch in coalescent units. At hybridization nodes, parent $A$ is chosen with probability $\gamma$, otherwise parent $B$ is chosen; then the usual multispecies coalescent applies with the chosen parent. The model therefore does not attempt to model the bottleneck that might occur during hybridization. Yu et al. (2012), building upon Meng and Kubatko (2009) who worked with a single hybridization event, provide formulas for computing the likelihood of a gene tree topology (without branch lengths) under this model, for networks with any number of hybridization nodes, which makes it possible in principle to carry out gene tree inference or species tree inference. In practice, they implemented this model in PhyloNet, so that different candidate species trees or networks can be compared according to their likelihood given a set of gene tree topologies, or a set of gene tree topology distributions. Kubatko et al. (2009) implemented an algorithm to search for the optimal network according to information criteria, with a model that considers branch lengths in the gene trees in addition to their topologies. In this implementation, the topology of the network is fixed, and the object of inference is the presence and number of hybridization nodes with the associated parameters $\gamma$.

Allopolyploidization has also been modeled as a specific case of hybridization (Jones et al. 2013). In this context, allopolyploidization occurs when two diploid individuals from different species mate, which results in the birth of a viable new tetraploid species. Jones et al. (2013) make the simplifying assumption that there is no recombination between the alleles inside the allopolyploid, and propose two models. In one model, any number of allopolyploidization events can be inferred, but evolution in the two genomes forming the allopolyploid is assumed to be independent, which disregards the fact that these genomes belong to the same species. In the second model, only one event of allopolyploidization can be inferred, but then the evolution of the two genomes in the allopolyploid is linked. In both cases, they use the multispecies coalescent model to describe the evolution of alleles along the species phylogeny or, in this case, the species phylogenetic network. They apply these models to simulated data sets as well as empirical data sets of less than 10 species, less than 10 genes, and up to three alleles per gene.

Another popular approach that deals with the variety of gene tree histories is Bucky (Ane et al. 2007). Bucky does not attempt to model the processes by which gene trees differ from species trees. Instead, it attempts to cluster gene families according to their histories. Bucky needs families of orthologous genes, which contain exactly 1 sequence per species. Once

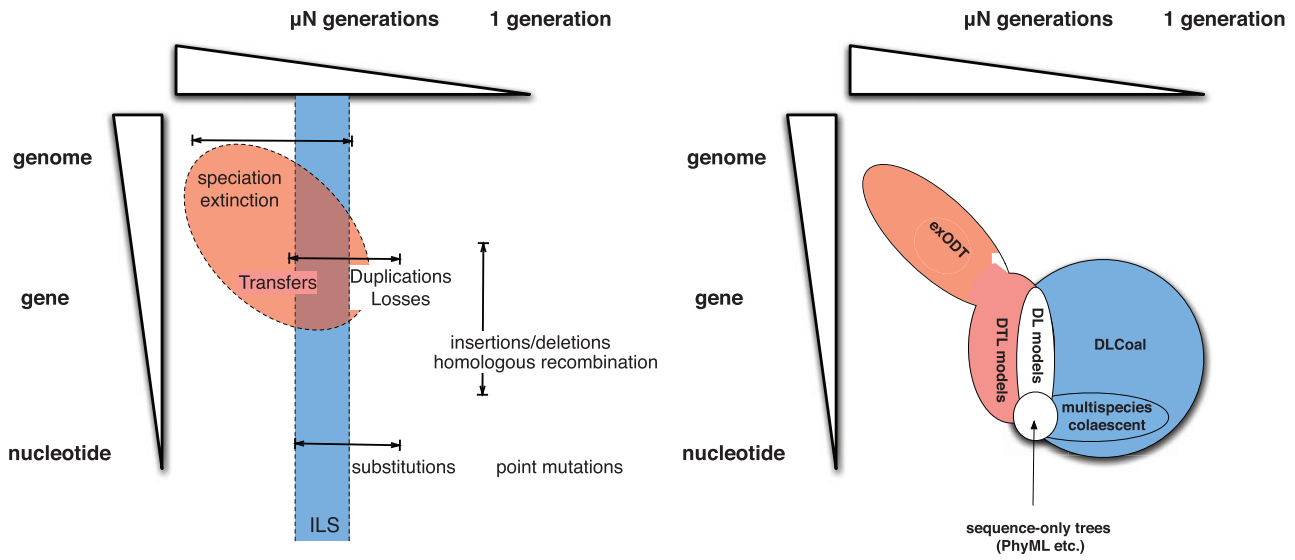


FIGURE 4. Left: the hierarchy of sequence evolutionary processes plotted along two dimensions: sequence length, from single nucleotides through genes to whole genomes; and time, from single generations through the neutral fixation time to large numbers of generations. Events occurring in single individuals, such as point mutations, insertions/deletions etc. are filtered by the population-level process of selection and drift with only a minority reaching eventual fixation. Speciation and extinction events affect entire genomes and require many generations. Incomplete lineage sorting (blue online) occurs when fixation time overlaps with speciation time. Transfer events can cross large phylogenetic distances and almost always involve evolution along extinct or unrepresented species and hence are affected by speciation dynamics. Right: distribution of models discussed in the text based on the time and length scales of the evolutionary process they model. Classic molecular phylogeny methods model only substitutions. DL models handle fixed duplication and loss events along with substitutions. The multispecies coalescent and related methods model explicitly the fixation of point mutations (blue online). DLCoal models the fixation of both point mutations and of gene-scale insertion/deletion events that lead to fixed duplication and loss events. DTL models (red online; ODT, DTLSR) extend DL models to fixed transfer events, but ignore speciation dynamics. The ex-ODT model combines speciation dynamics with DTL events to provide a more realistic model of transfer paths. A potential "ex-DTLCoal" model, as discussed in the main text would cover the area of all these models.

posterior distributions of phylogenetic trees have been built from these families, Bucky attempts to cluster the families by similarity of their evolutionary history. This clustering is done in a nonparametric Bayesian framework, which means that both the clustering and the number of clusters are estimated from the data. The end result may provide insight into the sources of the heterogeneity among gene histories, for instance in cases where neighboring genes are found to share the same history. Additionally, Bucky has also often been used to provide a candidate species tree, by gathering clade frequencies from all gene histories. One strength of Bucky is the fact that, provided orthology between genes has been well defined, it does not depend on a particular model of gene family evolution, and will work equally well in the presence of transfers or incomplete lineage sorting for instance. The corresponding drawback is that it does not return direct estimates of the species tree, or of rates of events such as duplications or transfers that might be of interest to students of molecular evolution.

### *Simulation and Inference*

One can see a phylogenetic pipeline as a series of statistical inferences, starting from raw sequences coming out of sequencing machines, and finishing with the inference of a species tree (Fig. 5). Necessary steps include sequencing error correction, assembly of reads into contigs and scaffolds, gene annotation, gene family clustering, alignment, and tree reconstruction. Most of these steps are done sequentially, so that later steps in the pipeline entirely disregard any estimate of uncertainty from the previous steps, and do not provide any feedback to these. Gene tree–species tree models take a step toward a more principled approach, by allowing communication between two steps of this pipeline, the construction of gene trees, and the construction of a species tree. Figure 5 places the above discussed models and associated phylogenetic software in the context of the complete phylogenetic inference pipeline. Gray nodes are considered known, and white nodes are inferred. This figure shows that a large diversity of inferential problems have been addressed, considering gene alignments, gene trees, species trees, or several of these as data. In this section, we review some of the methods and algorithms that have been used to address these inferential problems. We do not discuss methods aiming at reconstructing an alignment, and instead focus on gene tree–species tree methods. As a consequence, in the following we use "probability of an alignment" loosely to describe the probability coming from events of substitutions or jointly from events of substitutions and insertion–deletions. We present how data can be simulated, how the likelihood of a gene tree or of a species tree can be computed efficiently, and how good gene trees and species trees can be searched for.

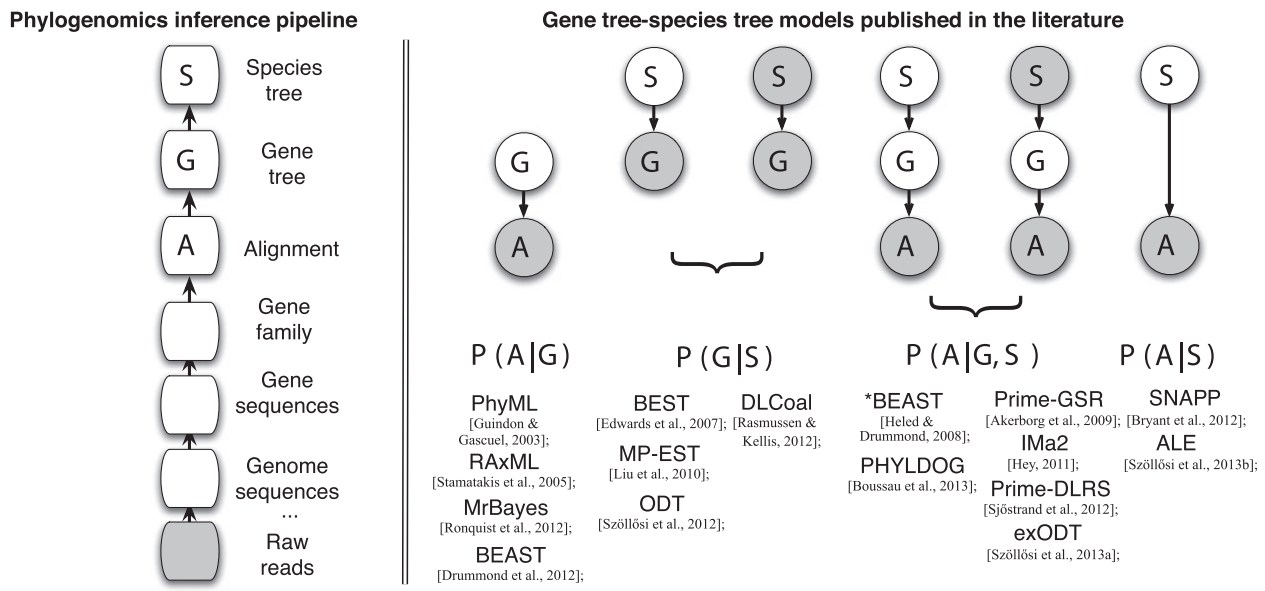


FIGURE 5. Gene tree–species tree models in the context of the phylogenomics inference pipeline. Left: the inference pipeline (some steps are not represented, such as sequencing error correction). Right: graphical representation of the inferential problem for a selection of the models and associated phylogenetic software discussed in the main text. The sequence of steps in the graphical model representations correspond to the hierarchical sequence of evolutionary process generating genomic sequences (cf. Fig. 1). The likelihood that must be computed is also shown. Graphical model conventions are observed: stochastic nodes, nodes corresponding to data considered as known are gray, and nodes whose states are inferred are in white. The models have been simplified, and parameters others than the gene tree and the species tree have not been represented.

*Simulating gene trees.*—Generating a simulated gene tree according to a model is straightforward, and usually involves traversing a species tree from its root to its leaves (although simulation under the coalescent is done from the leaves to the root). Assume there are $k$ lineages at time $t_{\text{begin}}$ (which at first is the beginning of a branch) on a branch $i$ which ends at $t_{\text{end}}$, with birth rate $\lambda_i$ and death rate $\mu_i$. Two waiting times $t_{\text{birth}}$ and $t_{\text{death}}$ are randomly drawn from exponential distributions, one with parameter $k*\lambda_i$, and the other with parameter $k*\mu_i$. If both are longer than $t_{\text{end}}-t_{\text{begin}}$, no event occurs along this branch. If $t_{\text{birth}}$ is the shortest, then a birth event occurs, and $t_{\text{begin}} \leftarrow t_{\text{birth}}$ and $k \leftarrow k+1$. If $t_{\text{death}}$ is the shortest, then a death event occurs, and $t_{\text{begin}} \leftarrow t_{\text{death}}$ and $k \leftarrow k-1$. In both cases, the lineage that has undergone the event is chosen uniformly. Then the process starts again. Lineages found at the end of a branch of the species tree are then given as input to the next birth–death processes, running along the descendant branches.

Lateral gene transfer can be regarded as simply a peculiar birth event, one which results in the birth of a gene copy in a branch of a species tree different from the species tree branch of the ancestral copy. Computationally, this introduces a dependency between species tree branches, requiring that all contemporaneous branches are considered together. To simulate gene trees, one can then consider different rates for different birth events, that is for duplications and transfers Szöllősi et al. (2012). Alternatively, one can consider replacement transfer, wherein, if a member of the homologous gene family is present in the recipient genome, it is replaced by the transferred gene. Computationally, this introduces a dependency between gene tree branches that prevents the use of algorithms that rely on the independence of gene lineages (see below), but simulations can be carried out in a straightforward manner (Galtier 2007). More problematically, however, no simulation method has been constructed to take into consideration the fact that, in the presence of transfer, gene trees record evolutionary paths along the complete species tree, including extinct and unsampled branches, and not only along the phylogeny of the species in which their descendants reside today. This is the case because, as first noted by (Maddison 1997) and later elaborated by Zhaxybayeva and Gogarten (2004); Fournier et al. (2009), although transfer events imply that the donor and receiver lineages existed at the same time, the donor lineage might have subsequently become extinct, or more generally, might not have been sampled. Brute force simulation of transfers along a "complete phylogeny" are expensive due to the large number of species that must be considered. There are at least two possible alternatives: (i) use instead parametric bootstrap-like methods described below or (ii) use approximations. One such approximation could be based on the assumption that the number of species represented in the species tree is much smaller that the total number of species Szöllősi et al. (2013b), similar to the assumptions of the coalescent.

A parametric bootstrap-like approach was used by (Szöllősi et al. 2013b) in the context of the ALE+ex-ODT

model (ALE standing for Amalgamated Likelihood Estimation) to produce simulated alignments based on a species tree and real alignments from 36 cyanobacteria. The approach consisted of first reconstructing the most probable gene trees according to the joint likelihood associated with duplication, transfer and loss rates given a fixed species tree and the gene family alignments. Second, the inferred gene trees were then used to simulate alignments. Third, these alignments were fed back into ALE+ex-ODT to assess its reconstruction accuracy, comparing both the reconstructed gene trees and the associated duplication, transfer and loss events to those used in the simulation. This approach has the advantage of circumventing potentially complex simulations while at the same time retaining otherwise hard to reproduce properties of biological datasets, such as the distribution of gene family sizes and the variation of evolutionary rates within and among gene families Szöllősi and Daubin (2012).

*Computing the conditional probability of a gene tree.*—By the joint conditional probability of a gene tree, we mean the probability of a gene tree given a gene alignment and a species tree. There are (at least) two components to the conditional probability of a gene tree. One component corresponds to the model of sequence evolution running along the gene tree; the other to the model of gene family evolution running along the species tree. In both cases, dynamic programming algorithms traversing the nodes of gene trees and species trees can often efficiently compute the relevant component of the probability.

Along a branch of a tree of a given length (the gene tree for sequence-based models or the species trees for gene family evolution models), probabilities of descendant states given ancestral states are computed by solving differential equations similar to other birth–death processes (for some processes, the solution can be obtained analytically, in others, numerical integration is necessary). Dynamic programming is then used to traverse branches of the tree in postorder. That is, branches are considered starting from their leaves up to the root. If a tree is unrooted, and all nodes need to be considered as potential roots, nodes need to be visited three times (although only two tree traversals are necessary, (Guindon and Gascuel 2003; Boussau and Gouy 2006)), according to the three possible directions for the root.

The probability of an alignment given a gene tree is computed along the gene tree alone, whereas the probability of a gene tree given a species tree is computed along both the gene tree and the species tree. In both cases, at a leaf, data can be used to fill a vector of probabilities. For sequence evolution, data correspond to the state found at the site under consideration (e.g., A, C, G, or T in a DNA sequence). For the model of gene family evolution, this corresponds to the presence, absence, or number of genes found in a given extant species. Then, at internal nodes, probability distribution vectors from the children nodes are used to compute the probability of a given subtree, according to the process considered in the branches. At the root, dynamic programming algorithms yield a probability for the entire tree.

The rough description above outlines the algorithm developed by Felsenstein (1981) to compute the probability of a multiple alignment of gene sequences given a gene tree and a model of sequence evolution. In this case, the differential equations corresponding to the Markovian process of sequence evolution can be solved analytically to obtain substitution probabilities along a branch of a given length. Computing the probability of a gene tree given a species tree is a bit more complicated, as it involves mapping the gene tree onto the species tree to compute the probability of presence of a gene tree node or branch at each node or branch of the species tree. This mapping is natural at the leaves: a gene from species $A$ is mapped onto leaf $A$ of the species tree. For internal nodes, the mapping can be helped by the consideration of node ages in models that consider that both the gene trees and the species tree are dated. This is typically the case with multispecies coalescent models (e.g., Rannala and Yang 2003). Such a method yields a single mapping between the nodes of a given gene tree and a given species tree, for given rates of sequence evolution. In a duplication and loss context, Akerborg et al. (2009) improved upon this approach by analytically integrating over the possible mappings as well as over rates of sequence evolution, again through dynamic programming. Their approach requires "slicing" the species tree by dropping extra nodes along the branches of the tree. These two methods yielding either a single mapping or integrating over all mappings in the context of dated trees have counterparts in the context of nondated trees. On one hand, Boussau et al. (2013) assumed the most parsimonious mapping between the nodes of the gene tree and the nodes of the species tree. This most parsimonious mapping is obtained with a single tree traversal (Zmasek and Eddy 2001). On the other hand, Szöllősi et al. (2012) took a similar approach to Akerborg et al. (2009) by integrating over all the possible mappings between the nodes of the gene tree and the nodes of the species tree, again through dynamic programming, but without considering dated gene trees. This allowed them to avoid using a model of rate evolution. However, it was necessary to order the nodes of the species tree, which has the effect of slicing it and adding new nodes, for correctly computing the probability of a gene tree given a species tree. As this inference includes transfer in addition to duplication and loss, numerical integration is necessary to solve the differential equations describing the birth and death process because gene lineages mapping to different branches of the species tree are dependent and no analytical solutions are available.

Usually such algorithms can achieve linear complexity in the number of genes for coalescent or DL models, but modeling transfers raises the complexity to the product between the number of nodes in the gene tree and the number of nodes in the species tree. Methods that require slicing the species tree (Akerborg et al.

2009; Tofigh 2009; Doyon et al. 2010; Szöllősi et al. 2012) introduce new nodes in the species tree and therefore are more expensive. However, in models with transfers, slicing the species tree is necessary as computing the probability of a gene tree given a species tree is provably difficult when the species tree nodes are not ordered, except if time-inconsistent transfers are allowed (Tofigh et al. 2011). It is also costly to model gene transfers that occur by replacing existing genes (like in SPR-like events of (Suchard 2005; Nakhleh et al. 2005; Beiko and Hamilton 2006; Bloomquist and Suchard 2010; Abby et al. 2012)). Indeed, this replacement introduces a dependency between separate gene tree lineages, which prevents the use of dynamic programming algorithms. In this context, the space of gene trees must be explored using SPR-like moves to compute the probability of a gene tree given a species tree. Note that current methods that slice the species tree (Tofigh et al. 2011; Szöllősi et al. 2012, 2013b) explicitly handle gene replacement transfers, but can only account for such events as a transfer event followed immediately by a loss in the receiving lineage.

For models of sequence evolution as well as models of gene family evolution, the same dynamic programming scheme can be used to make a variety of inferences. If we focus on models of gene family evolution, this scheme can be used to obtain a maximum parsimony reconciliation or the reconciliation that maximizes the probability of the gene tree given the species tree, to integrate over all reconciliations to compute the probability of a gene tree given a species tree, or to sample among reconciliations according to their probability. For a survey on available algorithms and software for reconciliations as of 3 years ago see (Doyon et al. 2011).

Combining the probability of a gene alignment given a gene tree with the probability of the gene tree given a species tree can be achieved by the multiplication of the two probabilities (Maddison 1997; Akerborg et al. 2009; Szöllősi et al. 2013b), assuming that the gene alignment is independent of the species tree conditional on the gene tree. The same assumption is at the heart of the model by Rasmussen and Kellis (2012) who combine probabilities from a multispecies coalescent model and a DL model, through the addition of an additional layer, the locus tree (see section "Modeling the dependence between gene tree and species tree"). Thanks to two conditional independence assumptions, the probability of the entire structure is obtained by the product of three probabilities.

Computing the probability of a gene tree given a phylogenetic network according to a multispecies coalescent model, as in Yu et al. (2012), requires specific algorithms in addition to usual ones, and the introduction of a new type of tree, the MUL tree. The MUL tree is a multilabeled tree generated from a phylogenetic network as follows: every hybridization node with its two parents $A$ and $B$ is removed from the tree, then duplicated, and finally one copy is attached to $A$ and the other to $B$. The MUL tree therefore contains several copies of subtrees coming from hybridization nodes, but it is no longer a network, as all nodes have a single parent; the computation of the probability of a gene tree then involves usual dynamic programming algorithms running along the MUL tree, with three complications. The first is that all possible mappings of the alleles to the duplicated subtrees must be considered—this creates a combinatorial factor that can substantially increase computing time. The second is that the multispecies coalescent process must be aware of the number of alleles evolving in a duplicated subtree when computing the propagation probability of an allele. The third is that the hybridization probabilities $\gamma$ must be taken into account (see the "Modeling the dependence between gene tree and species tree" section for more details about $\gamma$) to compute the probability of an allele trajectory along the MUL tree. MUL trees are also used by Jones et al. (2013), as an intermediate step to compute the probability of a phylogenetic network in their most accurate model that can handle a single allopolyploidization event, and as object of inference in their more flexible but less faithful model.

*Finding good gene trees.*—If we do not assume gene trees to be known, exploring the space of possible gene trees is usually achieved, similar to sequence evolution models, by either hill climbing maximum likelihood strategies (Vilella et al. 2008; Thomas 2010; Rasmussen and Kellis 2012; Boussau et al. 2013; Wu et al. 2013), or stochastic sampling of trees using Markov Chain Monte Carlo (MCMC) algorithms (Liu and Pearl 2007; Heled and Drummond 2008; Minin et al. 2008; Akerborg et al. 2009; Heled and Drummond 2010; Rasmussen and Kellis 2010).

These local searches are inspired by classic gene tree search algorithms (Guindon and Gascuel 2003), and the MCMC or hill climbing steps use, for example, random NNI (nearest neighbour interchanges) or SPR (subtree prune and regraft) moves. As such searches can be computationally intensive, SPRs are sometimes bounded (Boussau et al. 2013), or rearrangements are directed (Szöllősi et al. 2012). Devising good directed rearrangements is sometimes called gene tree correction, and entails for example decreasing the number of duplications in a duplication/loss most parsimonious reconciliation: such moves have some chance to increase the probability of a gene tree according to the model of gene family evolution. If at the same time they do not decrease the likelihood according to the model of sequence evolution by a large portion, such moves are accepted (Chang and Eulenstein 2006; Muffato et al. 2010; Chaudhary et al. 2012; Lafond et al. 2013). Reconciliations with polytomies in a gene tree (Lafond et al. 2012) can also be used to correct or construct good gene trees according to a sequence model and a species tree.

An alternative to local search in gene tree space is the amalgamation of reconciled gene trees from a sample of trees (David and Alm 2011; Szöllősi et al. 2013a). As illustrated in Figure 6 this approach consists in combining clades found in a sample of gene trees based

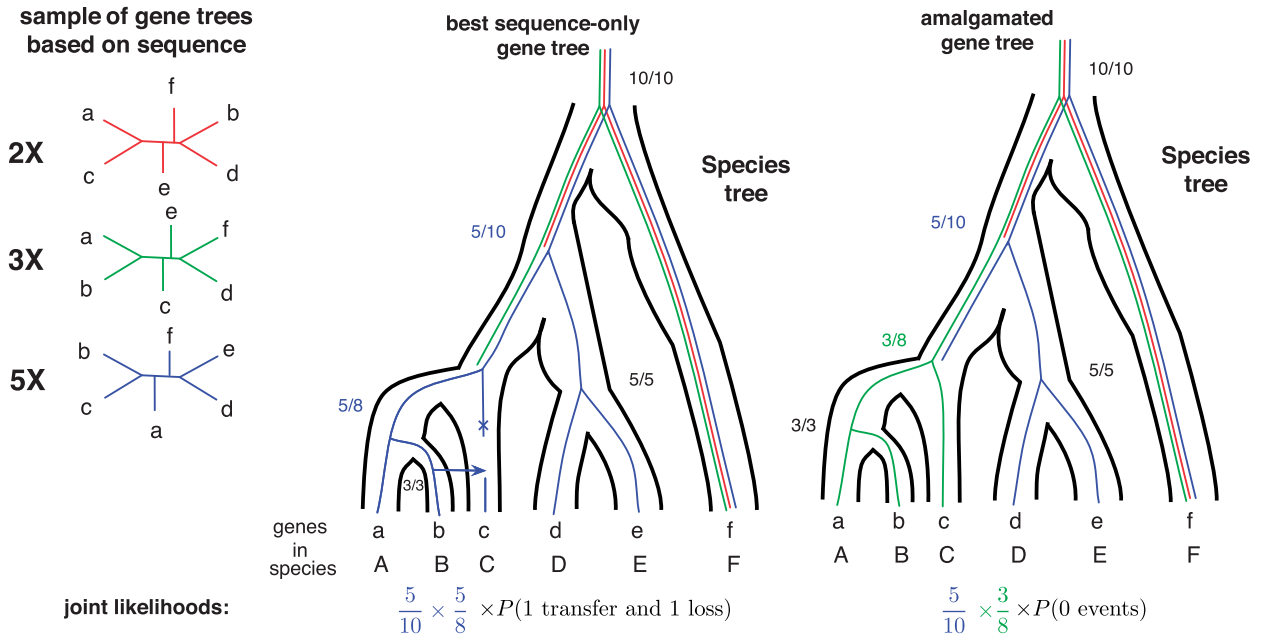


FIGURE 6. Based on gene trees sampled according to their posterior probability, conditional clade probabilities (CCP) can used to estimate the posterior probability of any tree that can be *amalgamated* (Höhna and Drummond 2012) from clades present in the sample. Conditional clade frequencies can be used to approximate CCPs and are computed as the fraction of times a particular split of a clade, for example (abc,de) is observed among all trees in which the containing clade, for example (abcde) is found. Estimates based on the sample of trees on the left are shown as fractions for two different gene trees that can be amalgamated. The estimate for a gene tree is given by the product of the frequencies. Amalgamated likelihood estimation (ALE (Szöllősi et al. 2013a)) is a probabilistic approach to exhaustively explore all reconciled gene trees that can be amalgamated as a combination of clades observed in a sample of gene trees. Based on the sample on the left, the tree with the highest posterior probability is the third tree (blue online). Reconciling it with the species tree requires 1 transfer and 1 loss event. It is, however, possible to combine clades present in the second (green online) and third (blue online) trees to produce a gene tree that is not present in the original sample but is identical to the species tree, that is it requires 0 events to draw it into the species tree. Depending on the relative probabilities of $P$(0 events) and $P$(1 transfer and 1 loss), the joint conditional probability may prefer the scenario without transfer.

on a model of sequence evolution (e.g., sampled using MCMC) in order to find the optimal gene tree according to both the model of sequence evolution and the model of gene family evolution. The probabilistic application of this method relies on the observation (Höhna and Drummond 2012; Larget 2013; Szöllősi et al. 2013a) that it is possible to efficiently and accurately approximate the probability of an alignment for a very large number of gene trees using conditional clade frequencies based on a much smaller sample of trees. Combined with a model of gene family evolution, this allows the construction of a gene tree that maximizes the product of the probability of the alignment given the gene tree and the probability of the gene tree given the species tree, or alternatively to sample reconciled gene trees according to the joint conditional probability (Szöllősi et al. 2013a). Here also a dynamic programming algorithm is used and the computation of a gene tree that maximizes the joint conditional probability is polynomial in the number of trees in the sample, which can be much faster than a local exploration, and can also be used as a starting point for local exploration.

*Finding a good species tree.*—One can also assume that the species tree is unknown, and search for it. Indeed integrative models of evolution should be able to retrieve the information about species evolution from the alignment better than averaging methods like concatenations or supertrees.

If we assume independence between genes, much like we usually assume independence between sites in most models of sequence evolution, a score for a species tree can be computed by adding (in parsimony) or multiplying (in probabilistic contexts) the joint conditional probabilities (scores in parsimony) for all gene trees. Optimization algorithms can then be used to search for the species tree with the best overall score or probability.

Methods for species tree inference under coalescent models (minimizing the number of ILS events or maximizing a joint conditional probability for a set of gene trees) are reviewed by Liu et al. (2009) and can be found on the online resource STRAW (Shaw et al. 2013). Fast approximations are given by distance or supertree methods (Than and Nakhleh 2009; Liu et al. 2010; Liu and Yu 2011; Yu et al. 2013), whereas Bayesian sampling is more precise but computationally intensive (Heled and Drummond 2010; Liu et al. 2008; Kubatko et al. 2009).

In a duplication and loss framework, Wehe et al. (2008); Bansal et al. (2010); Chaudhary et al. (2010) use the total number of duplications and losses as a global score and propose an efficient way to perform SPR (subtree prune and regraft) and tree bisection reconnection (TBR) moves on one candidate species tree to decrease the score. These

heuristics perform SPRs in a specific order, so that only a small portion of the mapping between gene trees and species trees needs to be recomputed, hence resulting in significant savings in computing time.

Alternatives to local searches are the search for exact solutions (Chang et al. 2011), or, inspired by coalescent models, supertree methods resembling the amalgamation of gene trees mentioned in the previous section, which seem to quickly provide good approximations of parsimonious species trees (Bayzid et al. 2013).

Considering transfers in a probabilistic framework, Szöllősi et al. (2012) explored time-ordered species trees, that is, species trees in which internal nodes are totally ordered. Topology search was performed by a directed local search guided by apparent highways of transfers: rearrangements are proposed in parts of the species tree that show the highest numbers of transfers in the hope of proposing rearrangements that reduce phylogenetic discord.

All of the above methods take as their input fixed gene trees. However, as we have several times recalled, good gene trees computed with the help of the (correct) species tree are substantially more accurate. Joint estimation has been achieved by Heled and Drummond (2010); Boussau et al. (2013), but with a very high computational cost. Improvements in the algorithms used would be very welcome.

## Impact on Systematics and Genome Evolution

The methods we described in the previous sections have shown repeatedly that they improve on methods that do not model gene family evolution for problems as diverse as species tree estimation, gene tree estimation, and the study of genomic evolution. In this section, we present some of their most salient results.

### *Learning about Species Relationships and History*

Coalescent models have been used extensively to investigate species trees (e.g., (Alström et al. 2011; Gray et al. 2011; Reid et al. 2012; Rocha et al. 2013)), because contrary to supertree or concatenation methods they should be robust against incomplete lineage sorting effects (Degnan and Rosenberg 2006). However, methods that jointly infer gene and species trees remain limited in their ability to handle large data sets: they cannot handle more than a few dozen species, and a few dozen loci. In their place, approximate methods have had to be used on genomic-scale data sets. As these take gene trees as input, they require much less computer resources, but suffer from the propagation of errors made during gene tree inference. Song et al. (2012) used MP-EST to reconstruct a well-resolved mammalian species tree based on 447 genes from 37 species. They found that the traditional technique, which consists of concatenating the alignments for several loci, was significantly less consistent when run on subsets of the data than MP-EST. However, another study by McCormack et al. (2013) found a highly unresolved tree when they used the same program on a data set of 416 Ultra Conserved Elements (UCEs) form 32 species of birds. They assumed that the small size of the UCEs led to unresolved and erroneous gene trees, which in turn caused MP-EST to estimate a highly unresolved species tree. This illustrates that methods that do not infer gene trees, but rather obtain them from external sources are by construction very sensitive to the quality of the input gene trees, and calls for more accurate and robust methods able to jointly infer gene and species trees in the coalescent framework, for large data sets.

Models of gene duplication and loss, or of gene transfer, have also been used to reconstruct species trees, although more sparingly than coalescent-based models. The construction of a species tree given many fixed gene trees has been performed many times under a parsimony DL framework, searching for the species tree that minimizes the total number of duplications or duplications and losses. In that context, phylogenies have been proposed for many clades (Slowinski et al. 1997; Page and Cotton 2002; Cotton and Page 2003; Than et al. 2008). Genome-wide scale was reached by Burleigh et al. (2011) who propose a plant phylogeny with 18,896 gene trees. So far, these models have mostly provided species phylogenies that were consistent with the literature (Suchard 2005; Szöllősi et al. 2012; Boussau et al. 2013). However, perhaps one of the most surprising benefits of modeling gene family evolution comes from modeling lateral gene transfer. Gene transfer is often described as a mere nuisance, which prevents phylogeneticists from obtaining well-resolved and easy-to-interpret species trees. According to this viewpoint, modeling gene transfer is useful because it provides a principled way to discriminate between vertical descent and lateral transfer: lateral transfers can then be discarded to focus on vertical descent and obtain a species phylogeny. However, gene transfer also provides additional information about ancestral species and their history.

For example, David and Alm (2011) infer the gene birth rate along a deep phylogeny of prokaryotes, and conclude that 25% of the genes in their data set were born in the Archean. They used a dated tree to infer transfers. But transfers can help date species trees, because gene transfers can only occur among contemporaneous species, and then be inherited by descendant species. A pattern of gene transfers therefore orients a species tree, from ancestral species that gave genes but did not receive many, to more recent ancestors that received genes from more ancient species.

Based on this idea, we ordered speciation events in Cyanobacteria using 8332 genes in 36 species (Szöllősi et al. 2012). We found that the information provided by transfer events supported a root different from the root obtained using outgroup species. However, outgroup sequences are usually very distant from Cyanobacteria, and the choice of the outgroup species affects the rooting

of the tree. In addition, we find that support for our unusual root comes from more than 200 transfer events. Overall, the information gained thanks to the use of a model of gene family evolution provides a new light into the order of speciation events in Cyanobacteria. It also provides a unique insight into genomic evolution in this clade, by providing an accurate reconstruction of ancestral gene contents. Because the ODT model infers events of gene transfers, duplications and losses, the number of genes present in ancestral genomes in each gene family is a natural outcome. Future analyses of ancestral gene contents based on models like ODT should provide windows into ancient metabolisms and phenotypes.

Another important process shaping species relationships is hybridization. Models that aim at inferring hybridization in the presence of incomplete lineage sorting have been used in several systems and have often found cases of hybrid speciations. Meng and Kubatko (2009) studied four genes in four species of cicadas from New Zealand to support an hypothesized hybrid origin for one species. Yu et al. (2012) and Bloomquist and Suchard (2010), using a Maximum Likelihood and a Bayesian approach, investigated 106 genes from yeast species (6 in Bloomquist and Suchard (2010), 5 in Yu et al. (2012)) and agreed about their inference of hybridization ancestral to two species. In addition, Bloomquist and Suchard (2010) studied 9 gene regions in spirochaete Bacteria, and confirmed previous results that one horizontal gene transfer happened in the history of these genes. Thanks to their integrative Bayesian method, they were able to date this event. Finally, Yu et al. (2012) studied more than 9000 genes in three *Drosophila* genomes and also detected hybridization ancestral to one of the three species, this time in disagreement with Pollard et al. (2006), whose analysis concluded that incomplete lineage sorting was enough to explain the pattern of incongruence in these genomes. Overall, these results show that network-based methods are powerful and can detect past hybridization events. Only Bloomquist and Suchard (2010)'s method can infer the network topology, but the other methods can be run on a set of topologies to compare their likelihoods.

### *Improving Gene Tree Reconstruction and Learning about Genome Evolution*

Beyond species tree reconstruction, coalescent models have also been used to investigate genomic evolution in closely related species. For instance, a method based on a Hidden Markov Model was used to estimate divergence times, effective population sizes and recombination rates in several species of primates. Insights include weaker selection operating on the X chromosome than expected (Hobolth et al. 2007), evidence for selection operating on genes (Hobolth et al. 2011), and a negative correlation between chromosome size and chromosome-specific ancestral effective population size (Mailund et al. 2011). The latter correlation is indicative of the power of recombination to increase effective population size: because, at each meiosis, each chromosome undergoes at least one recombination event, there are more recombination events per base on small chromosomes, which then increases the effective population sizes on small chromosomes.

The use of a species tree in addition to a gene alignment yields better gene trees than methods that only consider the gene alignment. Akerborg et al. (2009) studied a dataset of about 180 gene families in 17 yeast genomes with two methods, their own method that uses the sequence alignment and a species tree, and mrBayes, that only uses the alignment (Ronquist et al. 2012). Several of these yeast species descend from a species whose genome has been duplicated. As a consequence, all gene trees in the data set must show a duplication event in the branch containing this species. They found that their method detects a branch corresponding to a whole genome duplication in 66% of the gene families, when mrBayes only detects this branch in 35% of the cases. Rasmussen and Kellis (2010) obtained a similar result by comparing the inferred orthologs from gene trees obtained using 11 methods to orthologs inferred from synteny information, on a data set of 16 fungi. The seven methods that use the information provided by the species tree were found to outperform the four methods that only use the sequence alignments, agreeing with synteny in about 90% of the cases versus 60%, respectively. Other tests based on a measure of tree balance after a duplication, or based on simulated data all concurred that the information provided by the species tree and interpreted by DL models greatly improves phylogenetic reconstruction. More recently Mahmudi et al. (2013) sample gene trees and reconciliations in an MCMC framework under a DL model and infer duplication and loss rates on a vertebrate tree. Their conclusion is not only that sequence-based trees are often wrong, but also that most parsimonious reconciliations of good gene trees are often improbable. Reconciled gene trees have also been used to detect paralogs that originate from whole genome duplications in teleosts (Ouangraoua et al. 2011; Howe et al. 2013) or at the base of vertebrates (Makino and McLysaght 2010; Affeldt et al. 2013) and understand the causes of their maintenance or detect the current traces of these duplications and reconstruct ancestral genomes. They have also been used to study the evolution of metabolism in fungi (Eastwood et al. 2011; Floudas et al. 2012). These authors study fungi that digest wood: brown-rot fungi, which digest only cellulose, and white-rot fungi, which digest both cellulose and lignin, the most resistant component in wood. Focusing on a subset of enzymes, and reconciling their gene trees with the species tree, they find that brown-rot fungi are derived white-rot fungi that have lost several important genes. They also infer that white-rot fungi appeared concomitantly with the disappearance of coal deposits, and suggest that lignin decay pathways in white-rot fungi may have caused this disappearance.

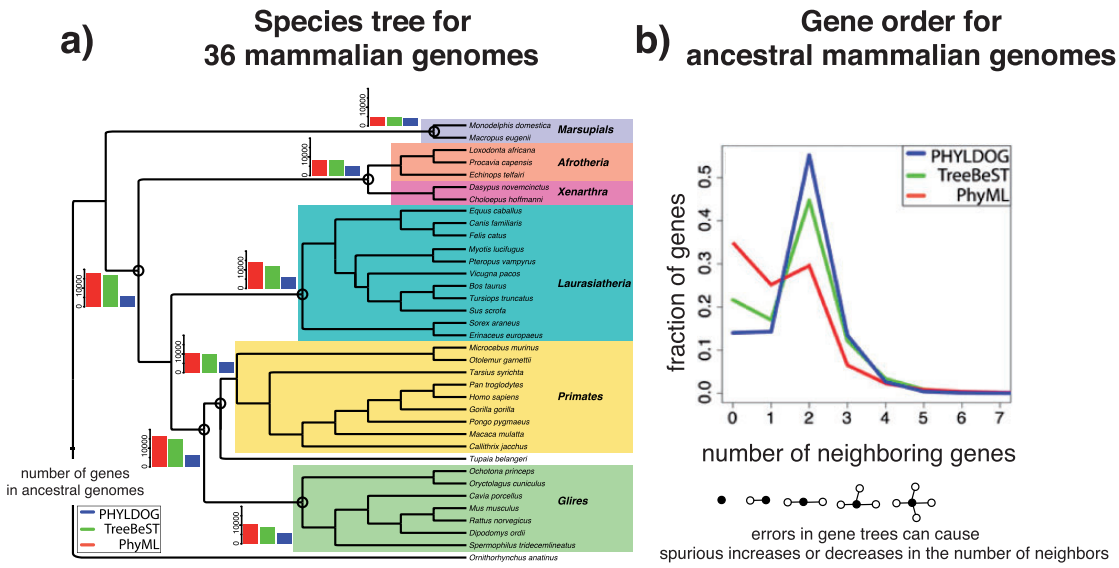


FIGURE 7. Left: species tree inferred by PHYLDOG, with ancestral genome contents reconstructed by different methods on selected nodes. Ancestral genomes reconstructed by PHYLDOG, in blue, have a size similar to that of extant genomes. Right: reconstruction of ancestral gene neighborhoods. Two genes are considered neighbors if there is no other gene between them in the dataset, so in a linear chromosome every gene (except two) should have two neighboring genes. So we expect from a method to recover exactly two neighbors for most ancestral genes. PHYLDOG is in blue. Figure reprinted with permission from Boussau et al. (2013).

Although gene tree reconstruction benefits from the input of an accurate species tree, it can probably suffer from the use of an erroneous species tree. Unfortunately, especially as new genomes get sequenced, accurate and well-resolved species trees are not always available. To address such situations and jointly infer the species tree and gene trees, we developed PHYLDOG (Boussau et al. 2013) and Fig. 7. We tested this program on both simulated and real data and found that both the species tree and the gene trees were reconstructed with good accuracy. Notably, as with the methods that use a given species tree to reconstruct gene trees, we found that our method improved upon methods that only use sequence alignments as input.

From a statistical point of view, joint reconstruction of the species tree and gene trees is the most principled approach, but it is computationally very challenging. In Szöllősi et al. (2013a), we used an approach that is almost as accurate as methods that explicitly infer gene trees while being easier to put in practice. We considered as our data a distribution of gene trees for each gene family instead of just a single gene tree. Using the ALE dynamic programming algorithm (see section "Simulation and inference"), this approach provides a fast but accurate approximation of the actual amount of phylogenetic information contained in the sequence alignment. We found that by not using a single gene tree as our data, the estimates of the number of gene transfers in the Cyanobacterial data set were 60% lower than when a single gene tree was used. This suggests that the claim that there have been too many transfers in Bacteria and Archaea for reconstructing the tree of life may have been a premature exaggeration. Szöllősi et al. (2013a) provides a method to reconstruct a gene tree given a species tree and rates of duplication, transfer, and loss, which can be given or all be inferred provided enough information is given to the program. Simulations and measures based on reconstructed ancestral genomes show that these gene trees are more accurate, but the biological relevance of how improved these trees are is perhaps best shown by ancestral sequence reconstruction. Groussin et al. (submitted) reconstructed sequences based on trees inferred through the Szöllősi et al. (2013a) approach, which uses the species tree and a distribution of gene trees, or through PhyML (Guindon et al. 2010), an accurate method that does not take the species tree into account. On simulations, this comparison showed that the ancestral sequences were much more accurate when based on the trees obtained with the help of the species tree. More strikingly, the in vitro resurrection of a protein belonging to the ancestor of Firmicutes, an ancient group of Bacteria, showed that the protein reconstructed based on the method using the species tree was thermodynamically more stable than the protein reconstructed from the alignment-only tree, and exhibited better enzymatic capabilities. As ancestral sequence resurrection is a popular and powerful approach (Gaucher et al. 2003; Thomson et al. 2005; Gaucher et al. 2008; Bridgham et al. 2009; Perez-Jimenez et al. 2011; Finnigan et al. 2012; Harms and Thornton 2013), methods using a model of gene family evolution could make an important contribution toward a better understanding of molecular evolution.

## FUTURE CHALLENGES

Methodological issues are still numerous and leave open wide research avenues, while at the same time the

potential of already available methods can be exploited on an increasingly large scale.

*Bypassing the Gene Tree in the Multispecies Coalescent*

The multispecies coalescent model describes the evolution of polymorphisms along a species phylogeny. Computing the likelihood of a gene alignment using this model requires summing over a large space of gene trees, given a species tree. This computational difficulty is a major hurdle to using this approach on large data sets, containing large numbers of species, and large numbers of gene families. Very recently, Bryant et al. (2012) and De Maio et al. (2013) came up with two elegant approaches to computing the likelihood of an alignment under the multispecies coalescent, by bypassing entirely the gene tree level, and instead analytically integrating over the space of possible allele histories. These models present the first methods to explicitly carry out the integral in the Felsenstein equation (Felsenstein 1988; Hey and Nielsen 2007). Bryant et al. (2012) consider biallelic data, and provide a model and an algorithm, called SNAPP, that can be used to reconstruct a species tree given an alignment of single nucleotide polymorphisms for instance. They develop a specific algorithm to address the fact that the coalescent process fundamentally functions from the tips of the species tree to its root, whereas the mutation process works forwards. They use this algorithm to reconstruct species trees with 69 individuals in 6 species of Digitalis plants. De Maio et al. (2013) instead propose a model for sequence data with $A,C,G,T$ data by using a substitution matrix over a larger state space than the usual $4\times4$ substitution matrices: it contains all 6 biallelic states $\{A,C\}$, $\{A,G\}$... with a range of frequencies. They focus on a specific model, where they consider a range of 10 possible frequencies per biallelic frequencies: for the state $\{A,C\}$, we therefore have the states $\{A10\%,C90\%\}$, $\{A20\%,C80\%\}$, ..., $\{A90\%,C10\%\}$. Two assumptions are made: first, no more than 2 alleles at a given site can be found at any time in a population, and second their frequencies are well approximated by the limited range included in the model. They construct transitions between states of this matrix from a population size parameter, selection coefficients, and mutation rates. The resulting instantaneous rate matrix is then exponentiated to provide a matrix of substitution probabilities. Overall, the matrix obtained with a range of 10 possible frequencies per biallelic state contains 58 states, that is about the same number of states as a codon substitution model. De Maio et al. (2013) use this model, with some further refinements to account for context-dependent mutations and strand-specificity on a large alignment of four species of primates and find evidence for a smaller ancestral population size in orangutans, and selection on splicing enhancers in exons.

Such analytical approaches seem very promising for combining coalescent models with duplication, loss and transfer models, as they bypass the problem of sampling allele histories. How they improve upon multispecies coalescent gene tree-species tree models is still an open question.

*More Integrative Models*

The integrative program of Goodman et al. (1979) is being progressively implemented. The probabilistic framework makes it possible to integrate sequence mutations with gene duplications and losses through the coalescent (Rasmussen and Kellis 2012), or to integrate duplications, losses, and transfers with substitutions (Szöllősi et al. 2012; Boussau et al. 2013; Szöllősi et al. 2013a,b). Rearrangements can be handled using parsimony if ILS is ignored (Bérard et al. 2012; Patterson et al. 2013).

A model and method to handle a union of all of these processes is currently missing. However, there are very good reasons for the integration of different levels of data analysis to continue. For instance, below the gene tree / species tree problem, is the inference of gene alignments. Only recently has the problem of joint inference of alignments and gene trees been considered seriously, with attempts to model the process of insertion/deletion in the evolution of sequences. Such approaches show dramatic improvements over phylogenetically unaware alignment methods (Redelings and Suchard 2005; Satija et al. 2009; Warnow 2013). However, they obviously need all the information necessary to have the best possible gene tree, for example a link to the species tree. Hence, it is probable that the integration of gene tree–species tree models and alignment methods should benefit the inference of alignments, gene trees and perhaps species trees.

Although a global model seems difficult to imagine presently, the entire pipeline of sequence data analysis, from sequencing error corrections to gene annotation and genome assembly is likely to benefit from probabilistic evolutionary models. The recognition of homologous sequences, the prediction of gene functions based on information from other organisms, and the proximity of genes on chromosomes all depend ultimately on the structure of the species tree and the possible events of substitution, duplication, loss, and lateral transfer that may have occurred in the history of genomes. There is currently no proposition of an integration of these processes on all levels of the pipeline described in Figure 5, but phylogenetically aware methods have proved very promising at many different steps of the process (Boussau and Daubin 2010) including on genome assembly (Husemann and Stoye 2010; Rajaraman et al. 2013).

*Algorithmics and Computing Time*

The score of a gene tree, especially if it is the combination of scores from several models, can be fairly costly to compute. Therefore, the exploration of trees is always time consuming. Already the inference of a gene tree that maximizes the probability of the alignment given the gene tree is provably hard. The joint inference,

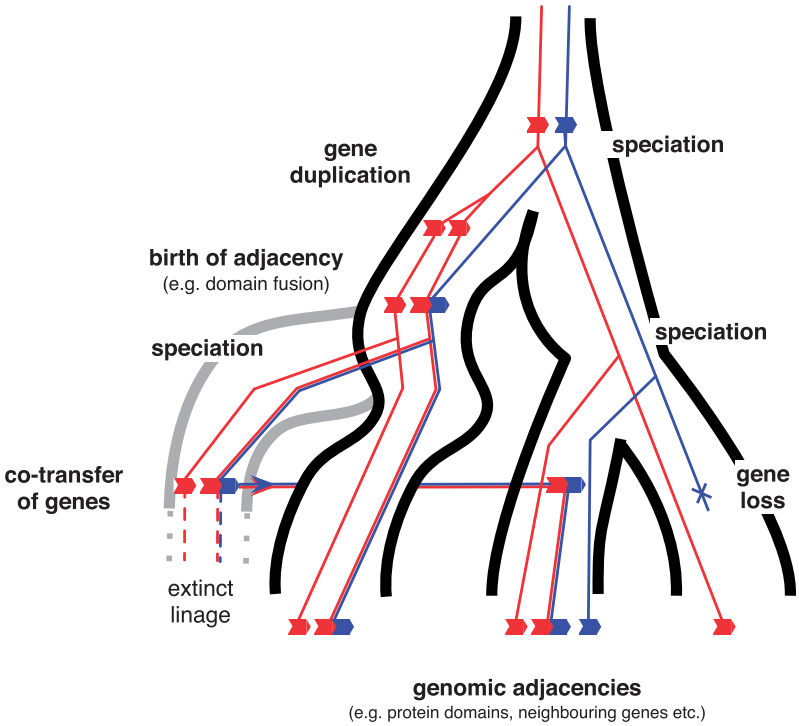


FIGURE 8. Evolutionary units below or above genes. Individual units (red and blue online) can be inside genes or genes that are neighbors along a chromosome or genes involved in a protein complex. Adjacencies are binary relations between genes, and evolve along a species phylogeny. Adjacencies can be gained or lost regardless of the birth and death of the units. When two units together undergo speciation, duplication, or transfer, adjacencies undergo the same events.

estimation of parameters, and exploration of dated or ordered species trees combine intractable problems. In practice, optimizing a gene tree can necessitate up to a few hours for very large families. As there can be thousands of gene families in a typical dataset, the computations even for a fixed species tree can take a long time. However, models of gene family evolution as well as sequence-based models all make the assumption that genes evolve independently from each other. This assumption can be questioned (see below) and is also broken by evolutionary parameters shared among gene families. But it allows a trivial parallelization by the data. All genes trees can be computed independently, given a common species tree. Hence, a species tree exploration is mainly constrained by the largest multigene families. A simple way to increase computational efficiency is to ignore these large families in a first step of species tree exploration. Large multigene families can be considered later, when a good species tree is found based on smaller gene families, or, in a sampling context, using importance sampling. However, such tricks can only help as long as the number of genomes under study is relatively small. For studying larger datasets, we will need to devise more efficient algorithms.

### *Reconstructing and Dating the Tree of Life*

A confusion between gene trees and species trees is arguably at the origin of the claim that Darwin was wrong when he evoked the image of a tree of life, because he failed to foresee the role of lateral gene transfer in microbial evolution (Doolittle 1999). The models and methods described above actually show that the plurality of gene histories can not only be overcome but more importantly provides additional information on the processes and patterns of species evolution. The phylogenies for a diversity of clades have been reconstructed with coalescent, DL or DTL models. In each case, the degree of conflict among gene trees can be interpreted in biological terms, such as divergence time and ancestral population size with the coalescent, or relative timing of speciation with LGT. There is a great hope that the development and use of these models will help resolve many issues that were left pending by traditional methods.

### *Beyond the Gene as an Evolutionary Unit*

Although we have adopted a liberal sense for "gene", in many of the studies we reported, a gene is a sequence coding for a protein or a functional RNA, and is considered as an evolutionary unit. However, within such genes, different parts may have different histories (Didelot et al. 2010; Wu et al. 2012). Alternatively, some genes may be associated throughout evolutionary times because their functions are interdependent or simply because they are close to each others in the genome. As such, they may be duplicated or transferred

together (Bansal et al. 2013; Patterson et al. 2013). Hence, the definition of evolutionary units is difficult, and fluctuates in time (Fig. 8). As we have shown, almost all existing models describe the reconciliation of one gene tree with one species tree, supposing its evolution is coherent and independent from other genes. Some genomic studies, however, allow genome-wide parameters like the rates of duplications and losses to vary across branches of the species tree (Boussau et al. 2013). This can be seen as a trick to model large-scale events like genome duplications without doing away with the independence of genes, which is computationally advantageous. But it fails to model more local rearrangements such as duplications of parts of a chromosome. These events could be informative for phylogeny, but models of genome rearrangements are often combinatorially so complex (Fertin et al. 2009) that they do not scale up well with the size and number of genomes (York et al. 2002; Darling et al. 2008; Miklós and Tannier 2010). Until now, their complexity has precluded a coupling with other models such as gene tree–species tree reconciliation. However, assuming neighborhoods between genes are independent, meaning that for any 3 genes *A*, *B*, *C* the neighborhood between genes *A* and *B* is independent of whether genes *A* and *C* are neighbors or not, it is possible to integrate rearrangements into DL (Bérard et al. 2012) or DTL (Patterson et al. 2013) models. Such approaches describe the evolution of neighborhoods (or any other relationship between genes, including functional ones) along pairs of reconciled gene trees, allowing one to reconstruct adjacencies in ancestral genomes and evolutionary events of duplication, loss, and transfer that have affected genomic fragments comprising several genes. Because such multiple events are frequent, it is likely that the parameters of duplication, transfer, and loss that are estimated in DL and DTL models are biased and it seems necessary to integrate models of neighborhood evolution with phylogenetic reconstruction into the reconstruction of genome histories.

There are also models for detecting breakpoints inside gene sequences using HMMs for instance (McGuire et al. 2000; Suchard et al. 2002; Martins et al. 2008; Boussau et al. 2009), or detecting breakpoints of phylogenetic discordance at a whole genome scale (Ané 2011), but so far these models have not been included in models of gene family evolution.

### *Keeping Up with the Pace of Data Acquisition*

Currently, genome sequencing is no longer a limiting step for comparative genomics. Instead, assembling gene families, gene alignments, gene trees, and a species tree are becoming increasingly problematic. In this context, methods using models of gene family evolution may offer an advantage because they effectively reduce the space of possible solutions to explore: given a species tree, the space of possible gene trees is limited compared with species tree unaware methods, and consequently, so is the space of possible alignments. Devising smart algorithms that make use of these reductions of complexity may provide fast yet accurate inferences for large-scale comparative genomics projects.

Another area where progress is needed is in the reuse of prior information. Currently, every time a new comparative genomics project is undertaken, or every time a database of homologous sequences is updated, many inference tasks need to be redone from scratch. The computations of gene families, alignments, trees, and species tree are usually done as if there was no prior information obtained from previous analyses. This is obviously a huge waste of useful information, as these computations are often very demanding. Future approaches to comparative genomics will need to be not only integrative, but also incremental. There is a clear need for new developments, and the Systematic Biology community is well equipped to undertake them.

## Conclusion

In the past 15 years, the relationship between gene trees and the species tree has been greatly clarified. This conceptual advance has been accompanied by methodological developments in models of gene family evolution and in the algorithms needed for statistical inference. These rely heavily on coalescent and birth–death processes and dynamic programming. In the next few years, these developments will go two seemingly incompatible ways: they will have to increase in complexity to more accurately model genome evolution, but they will also need to scale up as the sizes of data sets keep increasing. This tension presents exciting challenges.

## Funding

This project was supported by the French Agence Nationale de la Recherche (ANR) through Grant ANR-10-BINF-01-01 “Ancestrome.” G.J.S. was supported by the Marie Curie CIG 618438 “Genestory” and the Albert Szent-Györgyi Call-Home Researcher Scholarship A1-SZGYA-FOK-13-0005 supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/1-11-1-2012-0001 “National Excellence Program.”

## References

Abby S.S, Tannier E., Gouy M., Daubin V. 2012. Lateral gene transfer as a support for the tree of life. Proc. Nat. Acad. Sci. USA. 109:4962–4967.

Affeldt S., Singh P.P., Cascone I., Selimoglu R., Camonis J., Isambert H. 2013. Evolution and cancer: expansion of dangerous gene repertoire by whole genome duplications. Med. Sci. (Paris) 29:358–361.

Akerborg O., Sennblad B., Arvestad L., Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc. Nat. Acad. Sci. 106:5714–5719.

Alström P., Höhna S., Gelang M., Ericson P.G.P., Olsson U. 2011. Non-monophyly and intricate morphological evolution within the avian family Cettiidae revealed by multilocus analysis of a taxonomically densely sampled dataset. BMC Evol. Biol. 11:352.

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. Genome Biol. Evol. 3:246–258.
Ane C., Larget B., Baum D.A, Smith S.D, Rokas A., Ané C. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–426.
Arvestad L. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19:7i–15.
Arvestad L., Berglund A.-C.C., Lagergren J., Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. Proceedings of the Eighth Annual International Conference on Computational Molecular Biology - RECOMB '04, New York, New York, USA: ACM Press. p. 326–335.
Bansal M.S., Banay G., Harlow T.J., Gogarten J.P., Shamir R. 2013. Systematic inference of highways of horizontal gene transfer in prokaryotes. Bioinformatics 29:571–579.
Bansal M.S., Burleigh J.G., Eulenstein O. 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. BMC Bioinformatics. 13:S11.
Bayzid M.S., Mirarab S., Warnow T. 2013. Inferring optimal species trees under duplication and loss. Proceedings, Pacific Symposium on Biocomputing, vol. 18, p. 250–261.
Beiko R.G., Hamilton N. 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evol. Biol. 6:15.
Bérard S., Gallien C., Boussau B., Szöllősi G.J., Daubin V., Tannier E. 2012. Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics 28:i382–i388.
Bloomquist E.W., Suchard M.A. 2010. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. Syst. Biol. 59:27–41.
Boussau B., Daubin V. 2010. Genomes as documents of evolutionary history. Trends Ecol. Evol. 25:224–232.
Boussau B., Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756–768.
Boussau B., Guéguen L., Gouy M. 2009. A mixture model and a hidden markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. Evol. Bioinformatics 5:67–79.
Boussau B., Szöllősi G.J., Duret L., Gouy M., Tannier E., Daubin V. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23:323–330.
Bridgham J.T., Ortlund E.A., Thornton J.W. 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461:515–519.
Bryant D., Bouckaert R., Felsenstein J., Rosenberg N.A., RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.
Burleigh J.G., Bansal M.S., Eulenstein O., Hartmann S., Wehe A., Vision T.J. 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Syst. Biol. 60:117–125.
Chang W.C., Eulenstein O. 2006. Reconciling gene trees with apparent polytomies. In Chen D.Z. and Lee D.T., editors, Proceedings of the 12th Conference on Computing and Combinatorics (COCOON), vol. 4112 of Lecture Notes in Computer Science, p. 235–244.
Chang W.-C., Burleigh G.J., Fernández-Baca D.F., Eulenstein O. 2011. An ilp solution for the gene duplication problem. BMC Bioinformatics 12:S14.
Chaudhary R., Bansal M.S., Wehe A., Fernández-Baca D., Eulenstein O. 2010. igtp: a software package for large-scale gene tree parsimony analysis. BMC Bioinformatics 11:574.
Chaudhary R., Burleigh J.G., Eulenstein O. 2012. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. BMC Bioinformatics 13:S11.
Cotton J.A., Page R.D.M. 2003. Gene tree parsimony vs uninode coding for phylogenetic reconstruction. Mol. Phylogenet. Evol. 29:298–308.
Csuros M., Miklos I., Csűrös M., Miklós I. 2006. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. Res. Comput. Mol. Biol. q-bio.PE:206–220.
Darling A.E., Miklós I., Ragan M.A. 2008. Dynamics of genome rearrangement in bacterial populations. PLoS Genet. 4:e1000128.
David L.A., Alm E.J. 2011. Rapid evolutionary innovation during an archaean genetic expansion. Nature 469:93–96.
De Maio N., Schlötterer C., Kosiol C. 2013. Linking great Apes genome evolution across time scales using polymorphism-aware phylogenetic models. Mol. Biol. Evol. 30:2249–2262.
Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.
Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evol. 59:24–37.
Didelot X., Lawson D., Darling A., Falush D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. Genet. 186:1435–1449.
Doolittle W.F. 1999. Phylogenetic classification and the universal tree. Science (New York, NY) 284:2124–2129.
Doyon J.-P., Scornavacca C., Gorbunov K.Y., Szöllősi G.J., Ranwez V., Berry V. 2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In Tannier E., editor. Proceedings of RECOMB Comparative Genomics, LNBI, p. 93–108.
Doyon J.-P., Ranwez V., Daubin V., Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. Brief. Bioinform. 12:392–400.
Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with beauti and the beast 1.7. Mol. Biol. Evol. 29:1969–1973.
Dubb L. 2005. A likelihood model of gene family evolution. PhD thesis, University of Washington, Seattle.
Eastwood D.C., Floudas D., Binder M., Majcherczyk A., Schneider P., Aerts A., Asiegbu F.O., Baker S.E., Barry K., Bendiksby M., Blumentritt M., Coutinho P.M., Cullen D., de Vries R.P., Gathman. A, Goodell B., Henrissat B., Ihrmark K., Kauserud H., Kohler A., LaButti K., Lapidus A., Lavin J.L., Lee Y.H., Lindquist E., Lilly W., Lucas S., Morin E., Murat C., Oguiza J.A., Park J., Pisabarro A.G., Riley R., Rosling A., Salamov A., Schmidt O., Schmutz J., Skrede I., Stenlid J., Wiebenga A., Xie X., Kües U., Hibbett D.S., Hoffmeister D., Högberg N., Martin F., Grigoriev I.V., Watkinson S.C. 2011. The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science (New York, NY), 333:762–765.
Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Nat. Acad. Sci. USA. 104:5936–5941.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22:521–565.
Fertin G., Labarre A., Rusu I., Tannier E., Vialette S. 2009. Combinatorics of genome rearrangements, MIT Press. Cambridge, Massachusetts.
Finnigan G.C., Hanson-Smith V., Stevens T.H., Thornton J.W. 2012. Evolution of increased complexity in a molecular machine. Nature 481:360–364.
Floudas D., Binder M., Riley R., Barry K., Blanchette R.A., Henrissat B., Martínez A.T., Otillar R., Spatafora J.W., Yadav J.S., Aerts A., Benoit I., Boyd A., Carlson A., Copeland A., Coutinho P.M., de Vries R.P., Ferreira P., Findley K., Foster B., Gaskell J., Glotzer D., Górecki P., Heitman J., Hesse C., Hori C., Igarashi K., Jurgens J.A., Kallen N., Kersten P., Kohler A., Kües U., Kumar T.K., Kuo A., LaButti K., Larrondo L.F., Lindquist E., Ling A., Lombard V., Lucas S., Lundell T., Martin R., McLaughlin D.J., Morgenstern I., Morin E., Murat C., Nagy L.G., Nolan M., Ohm R.A., Patyshakuliyeva A., Rokas A., Ruiz-Dueñas F.J., Sabat G., Salamov A., Samejima M., Schmutz J., Slot J.C., St John F., Stenlid J., Sun H., Sun S., Syed K., Tsang A., Wiebenga A., Young D., Pisabarro A., Eastwood D.C., Martin F., Cullen D., Grigoriev I.V., Hibbett D.S. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science (New York, NY), 336:1715–1719.
Fournier G.P., Huang J., Gogarten J.P. 2009. Horizontal gene transfer from extinct and extant lineages: biological innovation and the coral of life. Philos. Trans. R. Soc. Lond. B Biol. Sci. 364:2229–2239.
Galtier N. 2007. A model of horizontal gene transfer and the bacterial phylogeny problem. Syst. Biol. 56:633–642.
Gaucher E.A., Govindarajan S., Ganesh O.K. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704–707.

Gaucher E.A., Thomson J.M., Burgan M.F., Benner S.A. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285–288.
Goodman M., Czelusniak J., Moore G.W., Romero-Herrera A.E., Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132–163.
Górecki P., Burleigh G.J., Eulenstein O. 2011. Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. BMC Bioinformatics 12:S15.
Górecki P., Eulenstein O. 2013. Drml: probabilistic modeling of gene duplications. J. Comput. Biol. 21:89–98.
Gray R.R., Tatem A.J., Johnson J.A., Alekseyenko A.V., Pybus O.G., Suchard M.A., Salemi M. 2011. Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant Staphylococcus aureus ST239 genome-wide data within a Bayesian framework. Mol. Biol. Evol. 28:1593–603.
Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.
Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.
Harms M.J., Thornton J.W. 2013. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat. Rev. Genet. 14:559–571.
Heled J., Drummond A.J. 2008. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8:289.
Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570.
Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27:905–920.
Hey J., Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc. Nat. Acad. Sci. USA. 104:2785–2790.
Hobolth A., Christensen O.F., Mailund T., Schierup M.H. 2007. Genomic relationships and speciation times of human, Chimpanzee, and Gorilla inferred from a coalescent Hidden Markov model. PLoS Genet. 3:e7.
Hobolth A., Dutheil J.Y., Hawks J., Schierup M.H., Mailund T. 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res. 21:349–356.
Höhna S., Drummond A. 2012. Guided tree topology proposals for Bayesian phylogenetic inference. Syst. Biol. 61:1–11.
Hordijk W., Gascuel O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics 21:4338–4347.
Howe K. et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503.
Huelsenbeck J.P., Rannala B., Larget B. 2000. A Bayesian framework for the analysis of cospeciation. Evol. Int. J. Organic Evol. 54:352–364.
Husemann P., Stoye J. 2010. Phylogenetic comparative assembly. Algorithms Mol. Biol. 5:3.
Jones G., Sagitov S., Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. Syst. Biol. 62:467–478.
Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.
Lafond M., Swenson K.M., El-mabrouk N. 2012. An optimal reconciliation algorithm for gene trees with polytomies. Algorithms Bioinformatics, p. 106–122.
Lafond M., Semeria M., Swenson K.M., Tannier E., El-Mabrouk N. 2013. Gene tree correction guided by orthology. BMC Bioinformatics. 15:S5.
Larget B. 2013. The estimation of tree posterior probabilities using conditional clade probability distributions. sysbio. oxfordjournals.org.
Li H., Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–496.
Liu L., 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.
Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080–2091.
Liu K., Raghavan S., Nelesen S., Linder C.R., Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science (New York, NY), 324:1561–1564.
Liu L., Yu L. 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60:661–667.
Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.
Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.
Mahmudi O., Sjöstrand J., Sennblad B., Lagergren J. 2013. Genome-wide probabilistic reconciliation analysis across vertebrates. BMC Bioinformatics 14:S10.
Mailund T., Dutheil J.Y., Hobolth A., Lunter G., Schierup M.H. 2011. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genetics 7:e1001319.
Makino T., McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Nat. Acad. Sci. USA. 107:9270–9274.
Martins L.O., Leal E., Kishino H. 2008. Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. PLoS One 3:e2651.
McCormack J.E., Harvey M.G., Faircloth B.C., Crawford N.G., Glenn T.C., Brumfield R.T. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PloS One, 8:e54848.
McGuire G., Wright F., Prentice M.J. 2000. A Bayesian model for detecting past recombination events in DNA multiple alignments. J. Comput. Biol. J. Comput. Mol. Cell Biol. 7:159–170.
Meng C., 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.
Miklós I., Tannier E. 2010. Bayesian sampling of genomic rearrangement scenarios via double cut and join. Bioinformatics. 26:3012–3019.
Minin V.N., Bloomquist E.W., Suchard M.A. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459–1471.
Morlon H., Potts M.D., Plotkin J.B. 2009. Inferring the dynamics of diversification: a coalescent approach. PLoS Biol. 8(9).
Muffato M., Louis A., Poisnel C.-E., Crollius H.R. 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics 26:1119.
Nakhleh L., Ruths D., Wang L.S. 2005. Riata-hgt: a fast and accurate heuristic for reconstructing horizontal gene transfer. In Wang L., editor. *Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05)*, 84–93. [LNCS #3595] Computing and Combinatorics.
Ouangraoua A., Tannier E., Chauve C. 2011. Reconstructing the architecture of the ancestral amniote genome. Bioinformatics. 27:2664–2671.
Page R.D.M., Cotton J.A. 2002. Vertebrate phylogenomics: reconciled trees and gene duplications. Pac. Symp. Biocomput. p. 536–547.
Pascal B. 1669. *Oeuvres: les pensées*. (ed. 1830 Librairie de Firmin Didot Frères). Rue Jacob, Paris.
Patterson M., Szöllősi G.J., Daubin V., Tannier E. 2013. Lateral gene transfer, rearrangement, reconciliation. BMC Bioinformatics 14:S4.
Perez-Jimenez R., Inglés-Prieto A., Zhao Z.-M., Sanchez-Romero I., Alegre-Cebollada J., Kosuri P., Garcia-Manyes S., Kappock T.J., Tanokura M., Holmgren A., Sanchez-Ruiz J.M., Gaucher E.A., Fernandez J.M. 2011. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat. Struct. Mol. Biol. 18:592–596.
Pollard D.A., Iyer V.N., Moses A.M., Eisen M.B. 2006. Widespread discordance of gene trees with species tree in drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.
Rajaraman A., Tannier E., Chauve C. 2013. Fpsac: fast phylogenetic scaffolding of ancient contigs. Bioinformatics. 29:2987–2994.

Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Rannala B., Yang Z. 2013. Improved reversible jump algorithms for Bayesian species delimitation. Genetics 194:245–253.
Rasmussen M.D., Kellis M. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Res. 17:1932–1942.
Rasmussen M.D., Kellis M. 2010. A Bayesian approach for fast and accurate gene tree reconstruction. Mol. Biol. Evol. 28:273–290.
Rasmussen M.D., Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. 22:755–765.
Redelings B., Suchard M. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst. Biol. 54:401–418.
Reid N., Demboski J.R., Sullivan J. 2012. Phylogeny estimation of the radiation of western North American chipmunks (Tamias) in the face of introgression using reproductive protein genes. Syst. Biol. 61:44–62.
Rocha S., Posada D., Harris D.J. 2013. Phylogeography and diversification history of the day-gecko genus Phelsuma in the Seychelles islands. BMC Evol. Biol. 13:3.
Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.
Rosenberg N.A., Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3:380–390.
Satija R., Novák A., Miklós I., Lyngsø R., Hein J. 2009. Bigfoot: Bayesian alignment and phylogenetic footprinting with mcmc. BMC. Evol. Biol. 9:217.
Scally A. et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature, 483:169–175.
Shaw T.I., Ruan Z., Glenn T.C., Liu L. 2013. Straw: species tree analysis web server. Nucleic Acids Res. 41(Web Server issue):W238–W241.
Sjöstrand J. 2013. *Reconciling gene family evolution and species evolution*. [PhD thesis], KTH, School of Computer Science and Communication.
Sjöstrand J., Sennblad B., Arvestad L., Lagergren J. 2012. DLRS: gene tree evolution in light of a species tree. Bioinformatics 28:2994–2995.
Slowinski J.B., Knight A., Rooney A.P. 1997. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (serpentes) based on the amino acid sequences of venom proteins. Mol. Phylogenet. Evol. 8:349–362.
Song S., Liu L., Edwards S., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Nat. Acad. Sci. 109:14942–14947.
Stamatakis A., Ludwig T., Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 21:456–463.
Suchard M.A. 2005. Stochastic models for horizontal gene transfer: taking a random walk through tree space. Genetics, 170:419–431.
Suchard M.A., Weiss R.E., Dorman K.S., Sinsheimer J.S. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. Syst. Biol. 51:715–728.
Szöllősi G.J., Daubin V. 2012. Modeling gene family evolution and reconciling phylogenetic discord. Methods Mol. Biol. 856:29–51.
Szöllősi G.J., Boussau B., Abby S.S., Tannier E., Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Nat. Acad. Sci. USA. 109:17513–17518.
Szöllősi G.J., Rosikiewicz W., Boussau B., Tannier E., Daubin V. 2013a. Efficient exploration of the space of reconciled gene trees. Syst. Biol. 62:901–912.
Szöllősi G.J., Tannier E., Lartillot N., Daubin V. 2013b. Lateral gene transfer from the dead. Syst. Biol. 62:386–397.
Than C., Sugino R., Innan H., Nakhleh L. 2008. Efficient inference of bacterial strain trees from genome-scale multilocus data. Bioinformatics 24:i123–i131.
Than C., Nakhleh L. 2009. Species tree inference by minimizing deep coalescences. PLoS Comput. Biol. 5:e1000501.
Than C., Ruths D., Innan H., Nakhleh L. 2007. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. J. Comput. Biol. 14:517–535.
Thomas P.D. 2010. Giga: a simple, efficient algorithm for gene tree inference in the genomic age. BMC Bioinformatics 11:312.
Thomson J.M., Gaucher E.A., Burgan M.F., De Kee D.W., Li T., Aris J.P., Benner S.A. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nat. Genet. 37:630.
Tofigh A. 2009. *Using trees to capture reticulate evolution: lateral gene transfers and cancer progression*. [PhD thesis], KTH, School of Computer Science and Communication.
Tofigh A., Hallett M., Lagergren J. 2011. Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans. Comput. Biol. Bioinformatics/IEEE, ACM 8:517–535.
Vilella A.J., Severin J., Ureta-Vidal A., Heng L., Durbin R., Birney E. 2008. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.
Warnow T. 2013. Large-scale multiple sequence alignment and phylogeny estimation. In Cedric C., editor. Models and algorithms for genome evolution. Springer.
Wehe A., Bansal M.S., Burleigh J.G., Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24:1540–1541.
Wu Y. 2011. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evol. 66:763–775.
Wu Y.-C., Rasmussen M.D., Kellis M. 2012. Evolution at the subgene level: domain rearrangements in the drosophila phylogeny. Mol. Biol. Evol. 29:689–705.
Wu Y.-C., Rasmussen M.D., Bansal M.S., Kellis M. 2013. Treefix: statistically informed gene tree error correction using species trees. Syst. Biol. 62:110–120.
Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Nat. Acad. Sci. USA. 107:9264–9269.
York T.L., Durrett R., Nielsen R. 2002. Bayesian estimation of the number of inversions in the history of two chromosomes. J. Comput. Biol. 9:805–818.
Yu Y., Degnan J.H., Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 8:e1002660.
Yu Y., Ristic N., Nakhleh L. 2013. Fast algorithms and heuristics for phylogenomics under ils and hybridization. BMC Bioinformatics, 14:S6.
Zhaxybayeva O., Gogarten J.P. 2004. Cladogenesis, coalescence and the evolution of the three domains of life. Trends Genet. 20:182–187.
Zmasek C.M., Eddy S.R. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 17:821–828.