A Bayesian Approach for Fast and Accurate Gene Tree Reconstruction

Matthew D. Rasmussen*,[1] and Manolis Kellis*,[1,2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
[2]Broad Institute of MIT and Harvard

***Corresponding author:** E-mail: manoli@mit.edu; rasmus@mit.edu.

Associate editor: Sudhir Kumar

Research article

Abstract

Recent sequencing and computing advances have enabled phylogenetic analyses to expand to both entire genomes and large clades, thus requiring more efficient and accurate methods designed specifically for the phylogenomic context. Here, we present SPIMAP, an efficient Bayesian method for reconstructing gene trees in the presence of a known species tree. We observe many improvements in reconstruction accuracy, achieved by modeling multiple aspects of evolution, including gene duplication and loss (DL) rates, speciation times, and correlated substitution rate variation across both species and loci. We have implemented and applied this method on two clades of fully sequenced species, 12 *Drosophila* and 16 fungal genomes as well as simulated phylogenies and find dramatic improvements in reconstruction accuracy as compared with the most popular existing methods, including those that take the species tree into account. We find that reconstruction inaccuracies of traditional phylogenetic methods overestimate the number of DL events by as much as 2–3-fold, whereas our method achieves significantly higher accuracy. We feel that the results and methods presented here will have many important implications for future investigations of gene evolution.

Key words: phylogenetics, gene tree, species tree, gene duplication and loss, reconciliation, Bayesian.

Introduction

Phylogenetic analysis has become an increasingly popular and fruitful approach for studying genomes (Hahn et al. 2005; Li et al. 2006; Hobolth et al. 2007; Wapinski et al. 2007; Butler et al. 2009). Methods for reconstructing phylogenies from sequence data have a long history (Felsenstein 1981; Saitou and Nei 1987; Rannala and Yang 1996) and new methods are continually developed to address a wide range of evolutionary questions. The question we approach in this work is the study of gene family evolution, namely how each family of genes has expanded and contracted over evolutionary time in a clade of related species. "Phylogenomics" has been proposed (Eisen 1998) as a systematic approach for studying gene families, where every gene family in several fully sequenced genomes is reconstructed and compared with a common species tree to infer orthologs, paralogs, and all evolutionary events, including gene duplications, losses, and horizontal transfers (Zmasek and Eddy 2002; Li et al. 2006; Huerta-Cepas et al. 2007; Wapinski et al. 2007; Butler et al. 2009; Vilella et al. 2009). However, as with any computational approach, the quality of the conclusions of phylogenomic studies are heavily dependent on the accuracy of the underlying methodologies. Accordingly, there has been much recent work on measuring and improving methods for phylogenetic reconstruction for both species trees and individual gene family trees. Advances have come from increased sequencing data for both additional taxa and loci as well as from new methods for leveraging that data.

For the problem of "species tree reconstruction," many advances have been made by combining data across loci either by concatenating multiple aligned loci into a "supermatrix" (Rokas et al. 2003; Ciccarelli et al. 2006), combining multiple gene trees into a"supertree" (Creevey and McInerney 2005), or by using a model for how such loci are correlated and coordinated in their evolution (Maddison and Knowles 2006; Liu and Pearl 2007). For example, in the BEST model (Liu and Pearl 2007), the correlated evolution of loci is captured by modeling a common species tree that constrains the evolution of each locus while still allowing some topological differences at each locus to occur via a coalescent process (Wakeley 2009). A probabilistic approach such as this allows one to use sequence alignments from multiple loci to estimate the posterior distribution of the species tree.

The problem of "gene tree reconstruction" also needs a similar strategy for exploiting abundant sequence data. Many recent efforts to reconstruct gene families in isolation (i.e., not accounting for their shared species tree or correlated evolution) have met many challenges. For example, the TreeFam project (Li et al. 2006) had found that automatic methods of reconstruction (such as maximum likelihood [ML], Felsenstein 1981; maximum a posteriori [MAP], Rannala and Yang 1996; neighbor joining [NJ], Saitou and Nei 1987; and parsimony, Felsenstein 2005) were not sufficiently accurate for systematic use and thus relied on human curators to adjust trees using additional information from the species tree, syntenic alignments, and the relevant literature. In a study by Hahn (2007), simulations were used to study how errors in gene tree reconstruction propagate into later inferences of gene duplication and loss (DL)

© The Author(s) 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/1044960
by Auburn University Libraries user
on 08 June 2018

events. In particular, the study showed that methods such as NJ frequently make reconstruction errors that lead to a biased inference of many erroneous duplications in ancestral lineages followed by numerous compensating losses in recent lineages.

In our own empirical work, we have found that the phylogenetic information available within a single locus is quite limited for most genes (Rasmussen and Kellis 2007). For example, in the recently sequenced 12 *Drosophila* species, we found that for alignments of orthologous genes, the inferred gene trees, regardless of the method used, have only a 38% chance of congruence with the species tree. For the 62% of alignments that supported an incongruent ML gene tree topology, only 5.7% did so with sufficient statistical significance ($P < 0.01$; SH test; Shimodaira and Hasegawa 1999). This along with several other measures of information content indicated that most loci lack enough information to confidently support one gene tree topology over the many other competing alternatives.

As we show below, the phylogenomic setting allows us to overcome the issue of limited information within individual loci by studying many gene families across the genome simultaneously. The additional information ultimately improves our ability to reconstruct gene trees but requires properly integrating information from both across species and genes while building upon several recent advances that we describe next.

Modeling Gene Trees and Species Trees

Our work fits within a growing body of literature addressing the simultaneous modeling of gene and species evolution. In one branch of this field, the primary concern is to model orthologous loci whose phylogeny may become incongruent with the species phylogeny due to incomplete lineage sorting (Maddison and Knowles 2006; Liu and Pearl 2007). In that case, gene trees are often modeled with the coalescent process (Wakeley 2009), which defines how topologies and branch lengths are distributed across loci (Rannala and Yang 2003), and has been used to reconstruct both gene trees (Hobolth et al. 2007; Dutheil et al. 2009) and species trees (Maddison and Knowles 2006; Liu and Pearl 2007), as well as many population related statistics, such as ancestral population sizes and recombination rates.

In another branch of the field, the loci of interest are those whose phylogeny is incongruent because of evolutionary events such gene duplication, loss, and horizontal transfer, and several models have been developed for each of these events. In the specific case of modeling DL, both probabilistic approaches (Arvestad et al. 2004; Gu and Zhang 2004; Hahn et al. 2005) and nonprobabilistic or parsimony-based methods have been developed (Goodman et al. 1979; Page 1994; Chen et al. 2000; Wapinski et al. 2007) to improve the reconstruction of either gene trees (Arvestad et al. 2004; Rasmussen and Kellis 2007; Wapinski et al. 2007) or species trees (Page and Charleston 1997). Our focus will be in this part of the field and specifically on the goal of the probabilistic reconstruction of gene trees in the context of a common and previously determined species tree.

For studying gene trees, Hahn et al. (2005) used the birth–death (BD) process to track changes in the number of paralogs in a gene family across a clade of species. Although it provides a way to look for significantly changing paralog copy counts, the method lacks a way of incorporating information from DNA or peptide sequences.

A method for incorporating such sequences was later developed by Wapinski et al. (2007) and was implemented in their SYNERGY gene tree reconstruction program. The method makes use of peptide sequences by combining a species-aware NJ algorithm along with an optimization for minimizing DLs while maximizing synteny (i.e., conserved gene order) between orthologs. However, this combination is ad hoc and nonprobabilistic, making it difficult to determine the best way to weigh conflicting information (Åkerborg et al. 2009). For example, in the cases where synteny information can be misleading, such as cases of gene conversions, SYNERGY shows significantly reduced reconstruction accuracy, suggesting that the primary sequence information is not sufficiently incorporated into the reconstruction (fig. 6).

A fully Bayesian model was proposed by Arvestad et al. (2004) that combined a model for gene DLs with sequence evolution. This was done by defining a prior for gene tree topologies and branch lengths using a BD process, which when combined with a sequence substitution model (e.g., JC69; Jukes and Cantor 1969) produced a Bayesian method for gene tree reconstruction and reconciliation. One disadvantage of this approach was the assumption of a clock model for substitution (i.e., constant substitution rates).

In 2007, we introduced a distance-based ML method for gene tree reconstruction that incorporates information from the species tree but avoids the clock model assumption (Rasmussen and Kellis 2007). Our model decomposes substitution rates into gene-specific and species-specific components, which was motivated by our observation of substitution rate correlations across the genomes of 12 *Drosophila* and 9 fungal species. By first learning parameters for gene- and species-specific rate distributions from genome-wide information and then using that model to reconstruct gene trees, SPIDIR showed significantly increased reconstruction accuracy compared with several other popular phylogenetic algorithms at the time. However, despite these improvements, the approach was distance based and thus did not fully utilize all the information available in sequence data.

Recently, Arvestad et al. (2004) have introduced PrIME-GSR, an extension of their previous work, which relaxes the clock assumption by using identical independent gamma distributions to model rate variation (Åkerborg et al. 2009), however, no species-specific rate variation is learned or modeled. In our evaluations (see Results), we find that modeling these rates can provide a significant benefit in gene tree reconstruction.

In summary, although much progress has been made in gene tree reconstruction, what remains missing is a principled, fast, and accurate method that incorporates all of

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

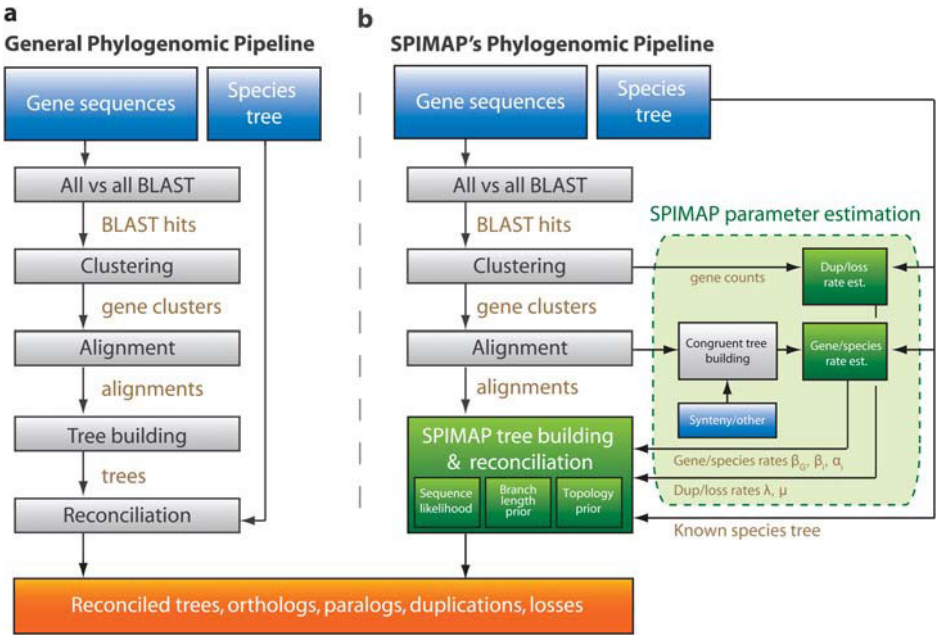


FIG. 1. Overview of the phylogenomic pipeline. (*a*) The typical phylogenomic pipeline consists of several common steps, although particular implementations may vary. The pipeline input is the set of all gene sequences across several species and the known species tree relating the species (blue boxes). Gene sequences are then compared across species and clustered according to their sequence similarity, resulting in a set of homologous gene families. A multiple sequence alignment is then constructed for each gene family, followed by phylogenetic reconstruction of each aligned family to produce gene trees. Each gene tree is then reconciled to the known species tree in order to infer orthologs, paralogs, and gene duplications and loss events, which are the pipeline outputs (orange box). (*b*) Our phylogenomic pipeline follows similar steps, except that SPIMAP includes a model parameter estimation step (dashed light green box) for DL rates (learned from the per-species gene counts in the gene families resulting from the clustering step), and gene- and species-specific substitution rates (learned from a subset of trusted orthologous alignments supported by synteny or other information and congruent to the species trees). These learned evolutionary parameters are then used in a joint tree building and reconciliation step (dark green box), specifically informing our topology prior (duplication/loss model) and our branch length prior (gene/species-specific substitution model). The joint step also enables us to use the known species tree and duplication/loss model to rapidly score topology proposals and speed up tree search in contrast to the traditional pipeline that only uses the known species tree in the reconciliation step.

these various models. In addition, freely available software is needed to facilitate further analyses in this field.

Here, we present SPIMAP, a Bayesian gene tree reconstruction method that incorporates within a unified framework models for gene DL, gene- and species-specific rate variations, and sequence substitution. We model gene DL using the BD process (Arvestad et al. 2004). Similar to the other methods, we do not attempt to model incomplete lineage sorting or horizontal transfers, although approaches for doing so in the future could be useful. We have implemented a relaxed clock, defined using the rate variation model we have previously developed (Rasmussen and Kellis 2007). A key distinction of our method is that we employ an empirical Bayes approach, where the parameters of the rate model are learned using a novel Expectation–Maximization (EM) training algorithm that incorporates sequence data across many loci. Once these parameters are estimated, we use them along with the species tree to reconstruct gene trees for thousands of sequence alignments from across the genome. Our method also achieves significant speed increases by using a novel tree search strategy derived from our gene tree topology prior. Lastly, we demonstrate the feasibility and increased performance of this method on several real and simulated data sets. The SPIMAP software is written in C++ and is available for download at http://compbio.mit.edu/spimap/.

Methods

Method Overview

The reconstruction of gene trees for every gene family in several genomes typically requires a computational pipeline similar to the one shown in figure 1*a*. Databases that have followed this general outline include TreeFam (Li et al. 2006), Ensembl (Vilella et al. 2009), and many others (Huerta-Cepas et al. 2007; Datta et al. 2009), whereas other methods such as SYNERGY (Wapinski et al. 2007) perform similar tasks but not necessarily as separate consecutive steps. The general pipeline goes as follows: The input (blue boxes in fig. 1*a*) consists of nucleotide or peptide sequences for all genes in all genomes under consideration as well as a species tree estimated prior to the pipeline computation using any method or information desired. Next, the sequences are compared with each other using a method such as an all-vs-all Blast search (Altschul et al. 1990) or HMMER (Eddy 2000). The Blast hits are then clustered using a method such as OrthoMCL (Li et al. 2003) or a method like that of Phylogenetically Inferred Groups (PHIGs)

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

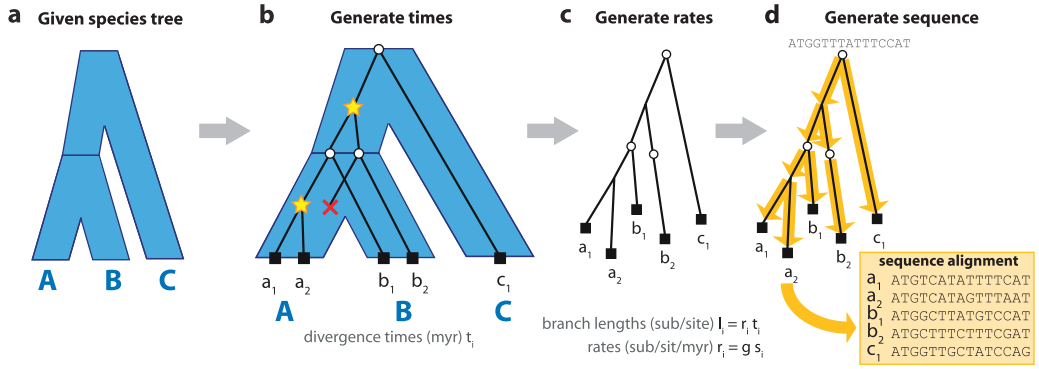


FIG. 2. SPIMAP's generative model. (*a*) First, the process begins with a given species tree S and divergence times. (*b*) Second, a gene tree T (black lines and labels) is evolved inside the species tree according to a DL model. The gene tree bifurcates either at speciation events (white circles located at species tree nodes) or at duplication events (stars located along species tree branches). Gene tree lineages can also terminate within a species tree branch at gene loss events (red "X"). (*c*) Third, substitution rates are generated according to our relaxed clock rates model of species-specific and gene-specific rates. (*d*) Lastly, sequences are evolved down the gene tree according to a continuous-time Markov process to produce a sequence alignment (yellow box) which is emitted from the process.

(Dehal and Boore 2006) in order to form clusters of highly similar genes that are likely to represent gene families. For each cluster, a multiple sequence alignment is then constructed (e.g., MUSCLE, Edgar 2004) followed by gene tree reconstruction using a phylogenetic algorithm (e.g., PhyML, Guindon and Gascuel 2003; BIONJ, Gascuel 1997; or MrBayes, Ronquist and Huelsenbeck 2003). Lastly, a "reconciliation" algorithm is used to compare each gene tree to the species tree in order to infer all DL events as well as all ortholog and paralog relationships. Reconciliation methods include maximum parsimony reconciliation (MPR) (Page 1994; Zmasek and Eddy 2001), RAP (Dufayard et al. 2005), and Notung (Chen et al. 2000), each of which take different approaches to inferring gene DL events in presence of possibly uncertain gene trees. The duplications, losses, orthology, paralogy, and the gene trees themselves typically constitute the outputs of a phylogenomic pipeline (orange box; fig. 1).

The pipeline we have constructed for SPIMAP follows the same general structure (fig. 1*b*). For clustering, we have implemented our own method (Butler et al. 2009) similar to that of PHIGs. For multiple sequence alignment, we have used the MUSCLE (Edgar 2004) program. In contrast to other methods, however, ours takes an Empirical Bayes approach by including a "training" step (dashed green box; fig. 1*b*) that supplies several species-level evolutionary parameters to SPIMAP's gene tree reconstruction step. In the training step, we estimate the average genome-wide gene DL rates $\theta_t = (\lambda, \mu)$ based on gene counts within each gene family cluster using a method similar to that of Hahn et al. (2005) (see Estimating DL Rate Parameters). We also estimate substitution rate parameters $\theta_b = (\alpha_G, \beta_G, \boldsymbol{\alpha}, \boldsymbol{\beta})$ based on a subset of the alignments using a novel EM method (see Estimating Substitution Rate Parameters). These parameters are then used in a combined gene tree reconstruction and reconciliation step (dark green box; fig. 1*b*) performed simultaneously within a single probabilistic model. From this model, we compute the MAP gene tree using a novel rapid gene tree search that incorporates information from the species tree and from DL rates. In the following sections, we will discuss how we compute the posterior probability of a gene tree and describe the details of our rapid tree search.

Gene Tree and Species Tree Definitions

We define a "gene family" as the set of all genes descended from a single gene in the last common ancestor of all species in consideration. We represent the rooted phylogenetic tree of n genes by a tree with topology $T = (V, E)$, which describes the set of nodes (vertices) $V(T)$ and a set of branches (edges) $E(T)$ of the tree. The leaves $L(T) \subset V(T)$ of a gene tree represent observed genes from extant species, whereas the internal nodes $I(T) = V(T) \setminus L(T)$ represent ancestral genes from ancestral species. We will use several functions to discuss how nodes are related to one another. For example, we use $\text{child}(v)$ to represent the set of children of v, $\text{left}(v)$ and $\text{right}(v)$ to represent the left and right children, and $\textit{parent}(v)$ to represent its parental node. For any node v, we use $b(v)$ to denote the branch $(v, \text{parent}(v))$ and $l(v)$ to be the length of that branch, measured in substitutions per site. Lastly, we use $\boldsymbol{l}$ to denote the vector of all branch lengths of a tree, namely $\boldsymbol{l} = (l(v_1), \ldots, l(v_{2n-2}))$. Thus, a "gene tree" is represented by the tuple $(T, \boldsymbol{l})$.

In addition, we will also consider a phylogeny S relating species, called a "species tree." The branch lengths $\boldsymbol{t}$ of S are expressed in units of time (e.g., millions of years) and are thus typically ultrametric. For a node $u \in V(S)$, we express its length as time $t(u)$. We will assume all trees are rooted and all nodes to have at most two children.

Each gene tree can be viewed as evolving "inside" the species tree (fig. 2*a*). A reconciliation R is a mapping from gene nodes to species nodes that defines the species to which each extant and ancestral gene belongs (Goodman et al. 1979) (fig. 3*a*). In this setting, a gene tree is "congruent" if R is an isomorphic mapping between T and S and "incongruent" otherwise. Also, all internal nodes of a gene tree represent either "gene duplication" or speciation events (represented as stars and white circles, respectively; fig. 2*b*).

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

Generative Model of Gene Family Evolution

In our generative model, gene trees are generated in three steps: given a species tree with specified topology and speciation times, 1) we first generate a gene tree topology and duplication times by repeated use of a BD process, 2) we then generate substitution rates from gene and species-specific distributions, and 3) lastly, we use these rates to generate molecular sequences according to a continuous-time Markov process (fig. 2).

The parameters of our model are $\theta = (S, \boldsymbol{t}, \theta_t, \theta_b)$, where S and $\boldsymbol{t}$ are the species tree topology and branch lengths, θ_t are the topology parameters λ and μ, and θ_b are the branch length parameters α_G, β_G, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$, the details of which are given below.

Generating Topology and Divergence Times

We use the gene DL model first developed by Arvestad et al. (2003), which is based on a repeated use of the BD process (Feller 1939) to define the topologies and branch lengths (in units of time) of a gene tree evolving inside a species tree (fig. 2*b*).

The BD process is a continuous-time process that generates a binary tree according to a constant rate λ of lineage bifurcation (which will represent gene duplication) and rate μ of lineage termination (representing gene loss). After running a BD process for a time t, all lineages that exist at time t are called "surviving," whereas all others are called "extinct." A node is "doomed" if it has no surviving descendants. The BD process has been used widely in phylogenetics (Arvestad et al. 2004; Gu and Zhang 2004; Hahn et al. 2005), although typically for defining priors for species trees (Rannala and Yang 1996).

The gene DL model is defined by repeatedly using the BD process to generate a gene tree. To initialize, we begin with a single gene node v reconciled to the root of S (i.e., $R(v) = \text{root}(S)$) and mark it as a speciation node. We then recursively apply the following: 1) For each speciation node v at the top of a species branch $b(u)$ of length $t(u)$, we generate a tree according to the BD process for $t(u)$ units of time. 2) For each newly created node w, we record its reconciliation as $R(w) = u$. 3) For each w that survives across that species branch, we mark it as an "extant gene" if u is a leaf species, otherwise mark it as a speciation. 4) We recursively apply steps 1–3 until all speciation nodes have been processed. 5) We mark all nodes in the gene tree not marked as extant genes or speciation as duplications. 6) As a postprocessing step, we prune all doomed lineages, namely lineages with no extant descendants.

Generating Substitution Rates

We use a relaxed clock model where substitution rates are allowed to vary between lineages (fig. 2*c*). Each branch has a length $l(v)$ (measured in substitutions/site) that is the product of a duration of time $t(v)$ and a substitution rate $r(v)$. The times are given by the DL model. The substitution rates indicate the number of substitutions per site per unit time and are described by a rates model. Previously (Rasmussen and Kellis 2007), we developed a rates model that captured the substitution rate $r(v)$ as the production of two components, a gene-specific rate and a species-specific rate. Here, we define these components with the following distributions:

(a) For each gene family j, the "gene-specific rate" g_j scales all rates in a tree. We represent the gene rate as a random variable G_j that is distributed across families as an inverse-gamma distribution with shape and scale parameters, α_G and β_G. Without loss of generality, we constrain G_j to have a mean value of one across all gene families (i.e., $\alpha_G = \beta_G + 1, \alpha_G > 1$). Thus, we have

$$P(G_j = g_j | \beta_G) = \text{InvGamma}(g_j | \alpha_G = \beta_G + 1, \beta_G).$$

(b) For each branch $b(v_k)$, the "species-specific rate" s_k defines a rate specific to that branch in the gene tree. It is represented by a random variable S_k and has a gamma distribution whose scale and shape parameters (α_i, β_i) depend on the species $u_i = R(v_k)$. This allows one to model rate accelerations and decelerations that are specific to a species u_i and exists across all genes of that species. Thus,

$$P(S_k = s_k | \alpha_i, \beta_i) = \text{Gamma}(s_k | \alpha_i, \beta_i), \quad \text{where } u_i = R(v_k). \tag{1}$$

We also assume that each S_k is independent of the others and of the gene rate G. Given these definitions for the substitution rate, we can then express the branch length $l(v_k)$ of a gene tree j as

$$l(v_k) = r(v_k) \times t(v_k) = g_j \times s_k \times t(v_k). \tag{2}$$

In total, our rate model has parameters $\theta_b = (\beta_G, \alpha, \beta)$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$, and m is the number of species branches $|E(S)|$.

Generating Sequence

After generating a gene tree with a topology, divergence times, and substitution rates, we finally evolve a molecular sequence down the tree using a continuous-time Markov chain to model sequence substitution. Specifically, we have implemented Hasegawa–Kishino–Yano (HKY; Hasegawa et al. 1985) to generating nucleotide sequences. The HKY process uses the branch lengths $l(v_k) = r(v_k)t(v_k)$ as parameters for sampling derived sequences. Only sequences on the leaves of the tree are emitted, whereas ancestral sequences are hidden (fig. 2*d*). In our current formulation, sequence insertion and deletion (indels) are not modeled. Instead, gaps in the sequence alignment are treated as missing data.

MAP Reconstruction of Gene Family Evolution

In our current implementation of the algorithm, we compute the MAP gene tree according to our model. Thus, we seek to calculate

$$\hat{\boldsymbol{l}}, \hat{T}, \hat{R} = \underset{\boldsymbol{l},T,R}{\text{argmax}}\ P(\boldsymbol{l}, T, R | \boldsymbol{D}, \theta) \tag{3}$$

$$= \underset{\boldsymbol{l},T,R}{\text{argmax}}\ P(\boldsymbol{D}|\boldsymbol{l}, T, R, \theta) P(\boldsymbol{l}|T, R, \theta) \times P(T, R|\theta)/P(\boldsymbol{D}|\theta) \tag{4}$$

$$= \underset{\boldsymbol{l},T,R}{\text{argmax}}\ P(\boldsymbol{D}|\boldsymbol{l}, T) P(\boldsymbol{l}|T, R, \theta) P(T, R|\theta). \tag{5}$$

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

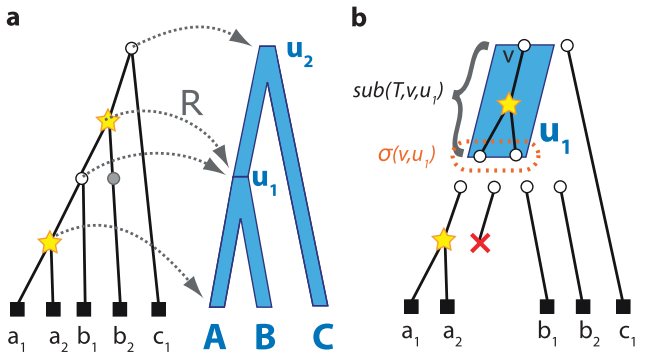


FIG. 3. Reconciliation and duplication subtrees. (*a*) A reconciliation R maps gene nodes to species nodes for both speciation events (white circles) and duplication events (stars). "Implied speciation nodes" (gray circle) are then inferred based on the reconciliation. (*b*) Our algorithm breaks the gene tree T into subtrees $\text{sub}(T, v, u_1)$, where the subtree root v is a speciation and the subtree leaves $\sigma(v, u_1)$ are the next speciation nodes below v that reconcile to species u_1.

The first term in equation (5) is the likelihood of a gene tree with branch lengths $\boldsymbol{l}$ and topology T given the sequence data $\boldsymbol{D}$. The probability is defined by the sequence evolution model (e.g., HKY) and can be computed efficiently using the pruning algorithm (Felsenstein 1981), which we have implemented for SPIMAP. Because this model only depends on the topology and branch lengths of the gene tree, the likelihood term is conditionally independent of the reconciliation R and parameters θ.

The prior of our model is factored into two terms: the prior of the topology and the prior of the branch lengths. The topology prior $P(T, R \mid \theta)$ is defined based on the DL model and it can be computed efficiently (see Computing the Topology Prior). We have also found that factoring out the topology prior from the branch lengths provides a unique advantage for fast tree search (see Rapid Tree Search).

Lastly, the branch length prior $P(\boldsymbol{l} \mid T, R, \theta)$ represents the probability of observing of gene tree branch lengths $\boldsymbol{l}$. This prior incorporates both divergence times of duplications in the BD process as well as the distribution of substitution rates. We present how to compute this term numerically (see Computing the Branch Length Prior).

Computing the Topology Prior

The topology prior $P(T, R \mid \theta)$ (from eq. 5) helps SPIMAP reconstruct gene trees that have plausible patterns of gene DL. For completeness, we describe how to compute this term.

According to the DL model introduced by Arvestad et al. (2004, 2009), the BD process is repeatedly used to generate the gene tree topology T as it evolves from the root of the species tree S to the leaves. Therefore, T can be viewed as a union of several subtrees, each of which was generated by one BD process. Because these processes are independent of one another, we can view the topology prior $P(T, R \mid \theta)$ of gene tree T as a product of the probabilities of the BD process generating each of the subtrees. Performing this factoring is the key step in computing the topology prior, but, there are two additional caveats to consider: 1) how to account for lineages in the gene tree that are hidden from observation due to extinction and 2) how to account for labeled and unlabeled nodes in the gene tree. By combining these ideas, we can compute the prior of a gene tree topology.

Factoring the Gene Tree

Given a gene tree topology T, we first decompose it into the subtrees that were generated from each individual BD process (fig. 3). We call each of these subtrees "duplication subtrees" because all of their internal nodes consist of duplication nodes. To identify these subtrees, first notice that each speciation node v is the root of two such subtrees. If v has reconciliation $R(v) = w$ and $w \in V(S)$, then the two subtrees perfectly reconcile within the child species branches $\text{left}(w)$ and $\text{right}(w)$. Also notice that the leaves of each duplication subtree are either speciation nodes or extant genes.

Some speciation nodes (e.g., the gray node in fig. 3*a*) may be initially hidden in a gene tree due to gene losses. We call such nodes "implcit speciation nodes" and they can be added to a gene tree by identifying gene tree branches that span multiple branches in the species tree (e.g., branch b_2 in fig. 3*a*). If a given gene tree T lacks implied speciation nodes, we can add them by locating each v and $w = \text{parent}(v)$, where $\text{parent}(R(v)) \neq R(w)$. Next, the edge (v, w) is replaced by a new speciation node x and two new edges (v, x) and (x, w) while setting $R(x) = \text{parent}(R(v))$. This procedure can applied repeatedly until all implied speciation nodes are identified.

When all speciation nodes are explicit, we can identify duplication subtrees by partitioning the gene tree at all speciation nodes $\text{spec}(T)$ (fig. 3). We denote a particular subtree as $\text{sub}(T, v, u)$, where $v \in \text{spec}(T)$ is the root of the subtree and $u \in \text{child}(R(v))$ is the species to which the leaves $L(\text{sub}(T, v, u))$ reconcile. The leaves are defined by the set

$$\sigma(v, u) = \{w : w \in \text{spec}(T) \cup L(T), R(w) = u, w \in V(T_v)\}, \quad (6)$$

where T_v is a subtree of T containing node v and all of its descendants.

For each duplication subtree, we can derive its probability from the BD process (Rannala and Yang 1996). First, for a BD process with a birth rate λ and death rate μ, the probability that one lineage will leave s survivors after time t is

$$p(s, t) = (\lambda/\mu)^s p(1, t)(p(0, t))^{s-1}, \quad (7)$$

where

$$p(0, t) = \frac{\mu(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}}$$

$$p(1, t) = \frac{(\lambda - \mu)^2 e^{-(\lambda-\mu)t}}{(\lambda - \mu e^{-(\lambda-\mu)t})^2}. \quad (8)$$

Second, for s survivors there are $\xi_s = s!(s-1)!/2^{s-1}$ equally likely "labled histories," which are leaf labeled

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

topologies whose internal nodes are order by their time. Thus, for a topology T with s leaves and $H(T)$ labeled histories, its probability is

$$P(T|t,\lambda,\mu) = \frac{H(T)}{\xi_s} p(s,t), \quad \text{where} \tag{9}$$

$$H(T) = \prod_{v \in I(T)} \binom{|I(T_{\text{right}(v)})| + |I(T_{\text{left}(v)})|}{|I(T_{\text{right}(v)})|}. \tag{10}$$

Doomed Lineages

In addition to factoring the tree, there are two caveats to consider. The first to consider is the possibility of lineages in the gene tree that are hidden from observation because they have gone extinct, that is, they leave no descendants in the leaves of the species tree. We call such lineages "doomed," and this extinction process must be accounted for in our topology prior.

Let $d(u)$ be the probability that a lineage starting at node u in the species tree will be doomed, that is, losses occur such that no descendants exist at the leaves of the species tree. This probability $d(u)$ is the product of the probability of extinction occurring in both the left and the right subtrees beneath node u. For a child branch $b(c)$, where $\text{parent}(c) = u$, we must consider two possibilities. Either the gene lineage goes extinct in $b(c)$ with probability $p(0, t(c))$ (eq. 8) or it survives and leaves i survivors, each of which themselves are doomed with probability $d(c)$. Thus, this probability can be expressed recursively as

$$d(u) = \begin{cases} \prod_{c \in \text{child}(u)} \sum_{i=0}^{\infty} p(i, t(c)) d(c)^i & \text{if } u \in I(S), \\ 0 & \text{if } u \in L(S). \end{cases} \tag{11}$$

The value $d(u)$ can be computed efficiently for each node u in the species tree S by dynamic programing following a postorder traversal of S.

Labeled and Unlabeled Nodes

The second caveat of the topology prior computation is distinguishing between labeled and unlabeled nodes within the gene tree. In equation (9), we give the probability of a BD process generating a labeled topology T. Each duplication subtree $\text{sub}(T, v, u)$ is generated by one BD process, however, only duplication subtrees with extant leaves (i.e., $L(\text{sub}(T, v, u)) \subseteq L(T)$) are labeled topologies. All other duplication subtrees have leaves that are speciation nodes and thus are unlabeled topologies.

To properly account for labeled and unlabeled nodes, we envision the DL model as a three step process. First, a gene tree T' is generated by repeated use of the BD process after which as all extant and speciation nodes are labeled. The probability of this tree is $P(T', R|\theta)$ and it can be computed by factoring T' into duplication subtrees, each of which has a known probability (eq. 9).

Second, a mapping U is applied to T' that removes all labels to produce an unlabeled gene tree T''. The probability $P(T'', R|\theta)$ is thus the sum of the probability of each T' that becomes T'' after removing labels,

$$P(T'', R|\theta) = \sum_{\{T' : T'' = U(T')\}} P(T', R|\theta). \tag{12}$$

We call two trees T_i' and T_j' equivalently labeled if $U(T_i') = U(T_j')$. Because equivalently labeled trees T_i' all have equal probability, the probability $P(T'', R|\theta)$ is simply the probability of T' times the number of equivalent labelings. The number of equivalent labelings is computed as a product of correction terms, one for each duplication subtree. Specifically, for each internal subtree T_2 (i.e., leaves are speciations nodes), we multiply by the term $N_2(T, T_2, R)$ and for each external subtree T_2 (i.e., leaves are extant genes), we multiply by $N_1(T_2, R)$. See supplementary section 2.2, Supplementary Material online for the definition of these terms.

In the third and final step, labels are added back to the leaves of T'' to create our desired leaf labeled gene tree topology T. Because each labeling is equally likely to be generated by this process, the probability $P(T, R|\theta)$ is $P(T'', R|\theta)$ divided by the number of ways to relabel T''. This final correction factor is $1/N_1(T, R)$ and is derived in supplementary section 2.2, Supplementary Material online.

The Full Topology Prior

Combining these ideas, we can compute the probability of a gene tree T being generated by the DL model as

$$P(T, R|S, \boldsymbol{t}, \lambda, \mu) = \frac{1}{N_1(T, R)} \times \prod_{v \in \text{spec}(T)} \prod_{u \in \text{child}(R(v))} g(v, u, \text{sub}(T, v, u)), \tag{13a}$$

$$g(v, u, T_2) = f(T, T_2, R) \sum_{i=0}^{\infty} \binom{|L(T_2)| + i}{i} \times p(|L(T_2)| + i, t(u)) d(u)^i, \tag{13b}$$

$$f(T, T_2, R) = \begin{cases} N_2(T, T_2, R) H(T_2) / \xi_{|L(T_2)|} & \text{if } L(T_2) \subseteq I(S), \\ N_1(T_2, R) H(T_2) / \xi_{|L(T_2)|} & \text{if } L(T_2) \subseteq L(S). \end{cases} \tag{13c}$$

The sum in equation (13) is a sum over how many doomed lineages i might have been present at node u. Within the sum, we find the probability that a BD process generates the survivors $L(T_2)$ that are present plus i hidden doomed lineages. The term $d(u)^i$ is the probability that those i lineages go extinct. The permutation term describes the number of ways to choose i doomed lineages from the total number of survivors $i + |L(T_2)|$.

Although this calculation involves an infinite sum, it can be computed analytically and the total computation of the topology prior takes at most $O(|V(T)||V(S)|)$ run time

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

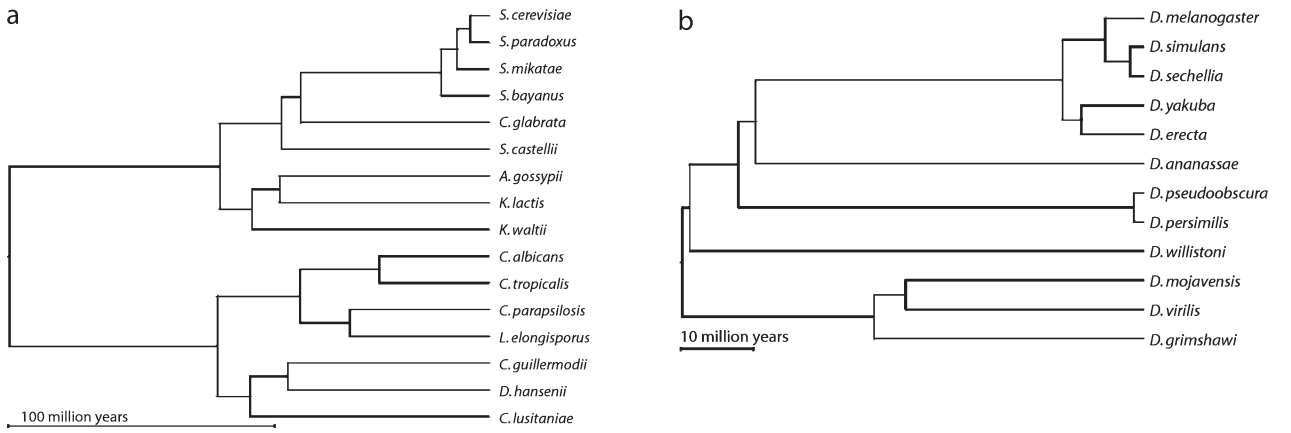


FIG. 4. Species and phylogenies used in evaluation. (*a*) Phylogeny of 16 fungal species used for the reconstruction pipelines of real and simulated evaluation data sets. The phylogeny was estimated in Butler et al. (2009) with divergence times estimated by the r8s program (Sanderson 2003) assuming 180 My (Massey et al. 2003) for the divergence depth. (*b*) The phylogeny of 12 *Drosophila* species used in our simulation evaluation. Phylogeny was estimated by Tamura et al. (2004).

(Arvestad et al. 2009). Currently, we only consider reconciliations R that are maximally parsimonious for DLs. This approximation is likely reasonable, as we find that the true reconciliation is the most parsimonious one in 98% of gene trees simulated using our species tree (fig. 4) and independently estimated DL rates (Hahn et al. 2005), agreeing with results from similar studies (Doyon et al. 2009).

Computing the Branch Length Prior

The final term in our model is the branch length prior $P(\boldsymbol{l}|T,R,\theta)$, which is the prior probability of the branch lengths $\boldsymbol{l}$ given the topology T, reconciliation R, and model parameters θ. This term helps SPIMAP choose gene trees that have branch lengths that are more reasonable given the time span implied by the reconciliation and our prior knowledge of the substitution rates.

We will explain the calculation of this term in a top-down fashion, breaking it into smaller parts until each part is defined. We begin by viewing the branch prior as a marginal over the gene rate g of the family in consideration

$$P(\boldsymbol{l}|T,R,\theta) = \int P(\boldsymbol{l}|g,T,R,\theta)P(g|\alpha_G,\beta_G)\mathrm{d}g. \quad (14)$$

Once conditioned on the gene rate g, many of the branch lengths of T become independent because we know their common scale factor g. However, those branches that surround a duplication node are still nonindependent because their lengths depend on the time of the duplication, which is unknown. However, if we partition T into a set of subtrees $\mathbb{T}$ by segmenting at each speciation node $v \in \mathrm{spec}(T)$ (without adding implied speciation nodes), each subtree $\tau \in \mathbb{T}$ will contain branch lengths that are independent of the other subtrees. In particular, each subtree τ is rooted by a speciation node, its leaves are either extant or speciation nodes, and all other internal nodes are duplication nodes. We refer to branch lengths for each subtree τ as $\boldsymbol{l}^\tau$, its divergence times as $\boldsymbol{t}^\tau$, and its substitution rates as $\boldsymbol{r}^\tau$. Thus, $\boldsymbol{l}^\tau = (l(w_1), l(w_2), \ldots, l(w_k))$ and $\boldsymbol{t}^\tau = (t(w_1), \ldots, t(w_k))$, where $w_1, w_2, \ldots, w_k$ are the nonroot nodes of subtree τ. Using this notation, we can continue to factor,

$$P(\boldsymbol{l}|g,T,R,\theta) = \prod_{\tau\in\mathbb{T}} P(\boldsymbol{l}^\tau|g,T,R,\theta). \quad (15)$$

The branch lengths within $\boldsymbol{l}^\tau$ are nonindependent because they depend on the duplication times. However, if we condition on the branch times $\boldsymbol{t}^\tau$, each branch length l_i^τ becomes a simple function of the branch rate r_i^τ because $l_i^\tau = t_i^\tau r_i^\tau$. Because we model all branch rates as being independent of one another, we can then finally factor the branch prior as a product of the probability of each branch length l_i^τ,

$$P(\boldsymbol{l}^\tau|g,T,R,\theta) = \int P(\boldsymbol{l}^\tau|\boldsymbol{t}^\tau,g,T,R,\theta)P(\boldsymbol{t}^\tau|g,T,R,\theta)\mathrm{d}\boldsymbol{t}^\tau, \quad (16)$$

$$\text{where } P(\boldsymbol{l}^\tau|\boldsymbol{t}^\tau,g,T,R,\theta) = \prod_i P(l_i^\tau|t_i^\tau,g,T,R,\theta), \quad (17)$$

and where $P(\boldsymbol{t}^\tau|g,T,R,\theta)$ describes the distribution of branch times in subtree τ which is defined by the BD process. We have integrated over the branch times $\boldsymbol{t}^\tau$ because they are unknown.

The last term to define is the distribution of a single branch length $l(v_i)$. In the simplest case (see the next section for a caveat), the distribution can be derived as follows:

$$\begin{aligned} l(v_i) &= g \times t(v_i) \times s(v_i) \sim g \times t(v_i) \\ &\quad \times \text{Gamma}\left(\alpha_{R(v_i)}, \beta_{R(v_i)}\right) \\ &= \text{Gamma}\left(\alpha_{R(v_i)}, \frac{\beta_{R(v_i)}}{g \times t(v_i)}\right), \end{aligned} \quad (18)$$

where, $s(v_k)$ is the species-specific rate for branch $b(v_k)$. In our implementation of computing the branch prior, we integrate over gene rates g (eq. 14) by approximating with a summation with equally probable gene rates. Also, the integral over times $\boldsymbol{t}^\tau$ (eq. 17) is performed with Monte Carlo by

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

sampling from $P(\boldsymbol{t}^{\tau}|g,T,R,\theta)$. The run time of this calculation is implemented to be linear with the size of the gene tree.

Handling Implied Speciation Nodes

One complexity not considered in equation (18) is the effect of implied speciation nodes. In such a case, we can have a branch length $l(v_i)$ that spans multiple species branches. For example, the branch b_2 in figure 3*a* spans the species B and u_1. Also note that the length of branch b_2 is the sum of two smaller branches: one within species branch B and one within species branch u_1. Thus, to complete our description of the branch prior, we must define the probability $P(l(v_i)|t(v_i),g,T,R,\theta)$ for branches that span multiple species.

To handle these cases, we introduce a topology T' that is defined as the topology T with implied speciation nodes added. Also let $\boldsymbol{l}'$ and $\boldsymbol{t}'$ be the length and time vectors of T', and R' be a reconciliation of T' to S. For each branch $b(v_i) = (v_i, w_i)$ in T, where w_i is the parent of v_i in T, there is a path $p = (v_i, \ldots, w_i)$ in T'. Let $p(v_i)$ be the set of all vertices in p excluding the top node w_i. Thus, the branch lengths and times in tree T can be expressed as sums of branch lengths and times in tree T',

$$l(v_i) = \sum_{v'_k \in p(v_i)} l(v'_k) \quad \text{and} \quad t(v_i) = \sum_{v'_k \in p(v_i)} t(v'_k). \tag{19}$$

The distribution of each $l(v'_k)$ is the same as the distribution given in equation (18) using R' as the reconciliation. To define the probability $P(l(v_i)|t(v_i),g,T,R,\theta)$, we note that $l(v_i)$ is simply the sum of independent gamma random variables, and methods exist to compute this probability efficiently (Moschopoulos 1985).

Branches Near the Root

If a gene branch contains the root, then it is still distributed by a sum of gamma distributions and thus can use the same methods developed here. For nodes that reconcile before the species tree root, we still treat them as being generated by a BD process in the basal branch of the species tree. We model the length T_0 of the basal branch as expontentially distributed with mean λ_0 and model the species-specific substitution rate as a gamma-distributioned random variable with mean and variance that is the average of the other species-specific rate distributions.

Rapid Tree Search

To compute the argmax in equation (5), we search over the space of possible gene tree topologies T, branch lengths $\boldsymbol{l}$, and reconciliations R using a hill climbing approach to find the MAP reconciled gene tree $(\hat{T}, \hat{\boldsymbol{l}}, \hat{R})$. We begin our search with an initial tree constructed using the NJ algorithm (Saitou and Nei 1987). We use subtree pruning and regrafting to propose additional topologies T. For each T, branch lengths $\boldsymbol{l}$ are proposed using numerical optimization (Newton–Raphson) of the likelihood term $P(\boldsymbol{D}|\boldsymbol{l},T)$.

One unique feature of our search is that we use the gene tree topology prior $P(T,R|\theta)$, a relatively fast computation compared with computing $P(\boldsymbol{D}|\boldsymbol{l},T)$ by 2–3 orders of magnitude to prescreen topology proposals for those that are likely to have high posterior probability. Given the best topology T thus far, we make $N \in [100, 1,000]$ unique rearrangements T_i and compute their topology prior $k_i = P(T_i, R_i|\theta)$, where R_i is the MPR. As our next proposal, we then choose a topology T_i from $T_1, \ldots, T_N$ with probability $p_i = \frac{c}{N} + \frac{(1-c)k_i}{\sum_i k_i}$, where parameter $c \in (0,1)$ defines a mixing between the weights k_i and the uniform distribution. In practice, we use $c = 0.2$.

We have found that this simple adjustment to our search strategy greatly increases the speed of finding the MAP gene tree (See Results and table 2).

Estimating Substitution Rate Parameters

As discussed previously, our substitution rate model is able to describe rate variation that occurs in both gene- and species-specific ways. In order to achieve this, it requires the estimation of several parameters $\theta_b = (\alpha_G, \beta_G, \boldsymbol{\alpha}, \boldsymbol{\beta})$. One unique approach in our method is that we estimate these parameters prior to reconstruction by analyzing substitution rates from multiple loci with known phylogenetic trees. This constitutes a "training step" in an empirical Bayes approach. Figure 1*b* illustrates how this estimation fits within the larger phylogenomic pipeline.

Currently for our training data set, we use trees of one-to-one orthologous gene alignments (e.g., syntenic orthologs or unambiguous best reciprocal Blast hits) where we can be reasonably confident that the gene tree topology is congruent to the species tree. Fixing the gene tree topology, we estimate the ML branch lengths for N trees with $M = |E(S)|$ branches each in order to construct a matrix $\boldsymbol{L}$ of branch lengths, such that l_{ij} is the length of the jth branch in the ith tree. We then use the $\boldsymbol{L}$ matrix along with a species tree S and its branch lengths $\boldsymbol{t}$ to estimate the parameters θ_b. Because the gene rates $\boldsymbol{g}$ of these trees are not known, we treat them as hidden data and use an EM algorithm to estimate our parameters.

The variables of the substitution rate training model are as follows. A gene tree will have a "gene rate" g, a vector of "species rates" $\boldsymbol{s}$ (measured in substitutions/site/unit time), and a vector of "branch lengths" $\boldsymbol{l}$ (measured in substitutions/site). Thus, for a single gene tree, we have the following variables:

$$g, \boldsymbol{l} = [l_1, \ldots, l_M]^T, \quad \boldsymbol{s} = [s_1, \ldots, s_M]^T,$$
$$\boldsymbol{t} = [t_1, \ldots, t_M]^T, \quad \text{with } l_i = g s_i t_i. \tag{20}$$

For a set of N gene trees indexed by j, we can describe them using the variables

$$\boldsymbol{g} = [g_1, \ldots, g_N]^T, \quad \boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N],$$
$$\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N], \quad \text{with } l_{ij} = g_j s_{ij} t_i. \tag{21}$$

We have designed this method to assume that $\boldsymbol{L}$ is directly observed and is given as input along with the divergence times $\boldsymbol{t}$. In contrast, the gene rates $\boldsymbol{g}$ and species rates $\boldsymbol{S}$ are not directly observed and have to be inferred from the model.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

As for the distribution of these variables, recall that g_j are independent and identically distributed (i.i.d.) by the inverse gamma $\text{InvGamma}(\alpha_G, \beta_G)$ and that s_{ij} are independently distributed by $\text{Gamma}(\alpha_i, \beta_i)$. Thus, the distribution of the branch length matrix $\boldsymbol{L}$ is

$$P(\boldsymbol{L}|\boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_G, \beta_G)$$
$$= \prod_j P(\boldsymbol{l}_j|\boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_G, \beta_G) \quad (22)$$
$$= \prod_j \int_0^\infty P(g_j|\alpha_G, \beta_G) P(\boldsymbol{l}_j|g_j, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \mathrm{d}g_j \quad (23)$$
$$= \prod_j \int_0^\infty \text{InvGamma}(g_j|\alpha_G, \beta_G) \times \prod_i \text{Gamma}\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) \mathrm{d}g_j. \quad (24)$$

In our EM algorithm, the branch length matrix $\boldsymbol{L}$ is the observed data and the gene rate vector $\boldsymbol{g}$ is the hidden data. In EM, the goal is to iteratively find better estimates of θ_b by maximizing this function

$$\theta_b{}^{k+1} = \underset{\theta_b}{\text{argmax}} \sum_j \int P(g_j|\boldsymbol{l}_j, \theta_b{}^k) \log P(\boldsymbol{l}_j, g_j|\theta_b) \mathrm{d}g_j. \quad (25)$$

For the derivation of equation (25), see supplementary section 2.3, Supplementary Material online. Conditioning on the hidden data allows us to find the next estimates of the gene- and species-specific rate parameters separately,

$$\beta_G^{k+1} = \underset{\beta_G}{\text{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^k) \log \text{InvGamma}(g_j|\beta_G) \mathrm{d}g_j \right] \quad (26)$$

$$\alpha_i^{k+1}, \beta_i^{k+1} = \underset{\alpha_i, \beta_i}{\text{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^k) \log \text{Gamma}\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) \mathrm{d}g_j \right]. \quad (27)$$

These expressions are computed using the Brent root finding algorithm for β_G and Broyden–Fletcher–Goldfarb–Shanno for α_i, β_i as implemented in the GNU scientific library. The gradients of these expressions are given in supplementary section 2.3, Supplementary Material online.

Computing the term $P(g_j|\boldsymbol{l}_j, \theta_b{}^k)$ (i.e., the probability of hidden data) constitutes the E-step. By exploiting conjugate priors (see supplementary section 2.3, Supplementary Material online), we have

$$P(g_j|\boldsymbol{l}_j, \boldsymbol{t}, \theta_b) = \text{InvGamma}\left(g_j|\alpha_G + \sum_i \alpha_i, \beta_G + \sum_i \frac{\beta_i l_{ij}}{t_i}\right). \quad (28)$$

We have currently implemented the EM algorithm such that $P(g_j|\boldsymbol{l}_j, \boldsymbol{t}, \theta_b)$ is discretized. Thus, the integrals in the argmax expressions (26) and (27) are approximated as sums. See supplementary figure S8, Supplementary Material online for an example of parameters learned from data sets of 12 flies and 16 fungi species.

Estimating DL Rate Parameters

We have also implemented a training procedure for estimating the genome-wide average duplication rate λ and loss rate μ. We use the algorithm of Hahn et al. (2005), which uses the gene counts in each gene family cluster (fig. 1*b*) to estimate λ and μ. However, unlike Hahn *et al.*, we do not require λ and μ to be equal. Examples of parameters estimated from data are given in supplementary table S1, Supplementary Material online.

Results

Phylogenomic Data Sets

To evaluate our approach for gene tree reconstruction, we have reconstructed gene trees for both real and simulated data sets. For our real data set, we have used 16 fungi species (fig. 4*a*) whose genomes have been sequenced to either draft or high coverage quality (Goffeau et al. 1996; Cliften et al. 2003; Kellis et al. 2003; Dietrich et al. 2004; Dujon et al. 2004; Jones et al. 2004; Kellis et al. 2004; Butler et al. 2009). For our simulated data sets, we simulated gene alignments that share many properties of real gene trees, by using a model with parameters estimated from real data sets. Thus, we have simulated gene trees that capture the properties of the 16 fungal genomes as well as 12 fully sequenced *Drosophila* genomes (Adams et al. 2000; Richards et al. 2005; Clark et al. 2007) (fig. 4*b*). By using both clades, we can evaluate the performance of phylogenetic methods across a variety of species tree topologies, divergence times, and gene DL rates.

For the species trees, we obtained the topologies and divergence times from several data sources. For the 16 fungi, we used the species phylogeny as constructed in Butler et al. (2009) and estimated time divergence using the r8s program (Sanderson 2003) with an estimate of 180 My (Massey et al. 2003) for the clade depth (fig. 4*a*). For the 12 flies, we used the same topology and divergence times as used in several recent studies (Tamura et al. 2004; Hahn et al. 2007) (fig. 4*b*).

Training SPIMAP's Model Parameters

To run SPIMAP in our evaluations, we applied our training algorithms to estimate the parameters of our gene family model. These parameters were also used to generate the simulated data sets. Here, we describe how we prepared the input data for our training procedure for both the 16 fungi and 12 *Drosophila* data sets. Our training procedure contains two methods: one to estimate our substitution rate parameters $\theta_b = (\beta_G, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and one to estimate our DL rates $\theta_t = (\lambda, \mu)$.

The first method (see Estimating Substitution Rate Parameters) estimates our substitution rate parameters from a

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

data set of one-to-one orthologous gene trees that are congruent to the species tree. To obtain such trees, we identified families that are highly likely to be one-to-one orthologous (i.e., one gene from each species in the clade). For the 16 fungi, we previously identified 739 confident one-to-one orthologous families (Butler et al. 2009). This was done by identifying synteny blocks containing at least three consecutive genes and spanning across the *Saccharomyces* or *Candidae* clades. Pairs of syntenic clusters with best reciprocal Blast hits spanning across the clades were merged, resulting in 739 families. For the 12 flies clade, we previously identified 5,154 one-to-one families where genes belong to a synteny block spanning all 12 species and contains at least three consecutive genes along each chromosome (Rasmussen and Kellis 2007). Next, for each one-to-one family, we made peptide multiple alignments using MUSCLE (Edgar 2004). Coding sequences were mapped onto the alignments to produce codon-aligned nucleotide alignments, substituting every amino acid with the corresponding codon and every gap with a triplet of gaps. PhyML v2.4.4 (Guindon and Gascuel 2003) was run on each nucleotide alignment using the HKY + Γ + I model and a fixed topology (congruent with the species tree), resulting in estimates for the branch lengths of each gene tree. Lastly, these branch lengths $\boldsymbol{L}$ were used in our EM method to estimate the model parameters (see supplementary figs. S8 and S9, Supplementary Material online).

The second method (see Estimating DL rate parameters) estimates our DL parameters from gene counts present within gene family clusters that contain DLs. For the 16 fungi, we used gene counts from gene families previously clustered (Butler et al. 2009) to estimate the gene DL rates $\lambda = 0.000732$, $\mu = 0.000859$ (events/gene/My). For the 12 *Drosophila* clade, we used the DL rates $\lambda = 0.0012$, $\mu = 0.0012$ that were previously estimated by Hahn et al. (2007) using a similar method as our fungi rate estimation, except that DL rates were assumed to be equal.

Reconstructing Gene Families from 16 Fungi

In our first evaluation, we analyzed the performance of SPIMAP versus several other popular phylogenetic programs on a data set of 16 fungi species. We have included four traditional "sequence-only" methods: PhyML v2.4.4 (Guindon and Gascuel 2003; ML), RAxML 7.0.4 (Stamatakis et al. 2005; ML), BIONJ (Gascuel 1997; NJ), and MrBayes v3.1.1 (Ronquist and Huelsenbeck 2003; Bayesian). We have also evaluated several other methods that use species-related information, which we call "species tree aware." These include our previous method SPIDIR (Rasmussen and Kellis 2007), SYNERGY (Wapinski et al. 2007), and PrIME-GSR (Åkerborg et al. 2009).

For our 16 fungi real data set, we downloaded coding sequences and peptides from the January 2009 update of fungi data set used by the SYNERGY method (Wapinski et al. 2007; Wapinski et al. 2009). By using this data as the input for all the other methods, we can compare against the trees constructed by SYNERGY (also downloaded from the January 2009 update). We focused the analysis on the same 16 species as used in Butler et al. (2009), which is a tree that also agrees with the one used by SYNERGY. We used the same gene clusters as defined by SYNERGY's trees, in effect using SYNERGY as the clustering step for the phylogenomic pipeline (fig 1*a*). Peptide alignments were made using MUSCLE (Edgar 2004), and coding sequences were mapped onto them to produce nucleotide alignments. In addition, from the nucleotide alignments, we also produced RY-encoded alignments, which only indicate whether a base is purine (R) or pyrimidine (Y). No other information from SYNERGY trees was made available to the other methods.

We used the following parameters for each of the methods. For PhyML and BIONJ, we used a HKY + Γ + I model of nucleotide substitution, whereas for RAxML, we used the GTRCAT model. We configured MrBayes with four chains, an automatic stop rule, a 25% burn-in, sampled every ten generations from a total of 10,000 generations, a 4×4 model for nucleotides, and enforced a binary tree. For methods that do not produce reconciled trees (i.e., PhyML, RAxML, MrBayes, and BIONJ), we have used MPR to infer DLs. For SPIDIR, we used DL penalties of 0.001 and an error cost of -600. For PrIME-GSR, we used 50,000 iterations, the Jones, Taylor, and Thorton model, gamma-distributed rates, and our own species tree (fig. 4). The tree search was initialized by an ML tree found by PhyML. We also ran PrIME-GSR with 1,000,000 iterations (as recommended by Åkerborg et al. 2009) but for only 500 trees randomly chosen from the data set in order to limit the computational run time. SPIMAP was executed with two settings: "long" (2,000 iterations with 1,000 prescreening iterations) and "short" (100 iterations with 1,000 prescreening iterations). For all other programs and options, defaults were used.

Although, a ground truth is not known for real data sets, we have used several informative metrics to assess the quality of gene trees, gene duplications, and losses inferred by these methods. Each of these metrics also illustrate different advantages and shortcomings of each method.

Recovering Syntenic Orthologs

The first metric we investigated was the ability to infer syntenic orthologs—pairs of genes that are highly likely to be orthologous given their surrounding conserved gene order. Although not all orthologous pairs are syntenic, synteny information does allow us to identify a conservative set of orthologous genes using a method independent of phylogenetics and thus provides a useful gold standard to test against. See supplementary section 2.1, Supplementary Material online for a description of our synteny determination method. When we construct trees on families that contain such genes, we expect a syntenic gene pair to appear within the reconstructed gene tree such that their most recent common ancestor is a speciation and thus are inferred as orthologs.

SPIMAP recovered syntenic orthologs with 96.5% sensitivity, followed by PrIME-GSR at 88.9% and PhyML at 64.1% (table 1). Because SYNERGY uses synteny as one of its inputs, this test alone cannot assess its accuracy, and indeed 99.2%

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

Table 1. Evaluation of Several Phylogenetic Programs on Gene Trees from 16 Fungi.

Program	Orthologs[a] (%)	Number of Orthologs[b]	Number of Dup[b]	Number of Loss[b]	Average run time[c]
SPIMAP (quick)[d]	96.2	550,800	5,541	10,884	1.0 min
SPIMAP (long)[d]	96.5	557,981	5,407	10,384	21.9 min
SPIMAP (i.i.d.)[e]	93.9	547,976	6,201	13,428	21.6 min
SPIDIR (quick)[d]	83.3	524,292	10,177	33,550	2.2 min
SYNERGY	99.2	595,289	4,604	8,179	—[f]
PrIME-GSR	88.9	527,153	7,951	21,099	53.1 min
PrIME-GSR (long)[d]	90.7	—	—	—	20.7 h
RAxML	63.8	463,020	21,485	65,392	18.4 s
MrBayes	63.9	460,510	21,307	65,238	43.2 s
PhyML	64.2	464,479	21,264	64,391	45.3 s
BIONJ	60.4	439,193	22,396	71,231	0.5 s

[a]Percentage of syntenic orthologs recovered.
[b]Number of pairwise orthologs, duplications, and losses inferred from trees.
[c]Average run time for reconstructing each gene tree.
[d]Both SPIMAP and PrIME-GSR were run with a few iterations (quick) of 100 and 50,000 and with many iterations (long) 2,000 and 1,000,000.
[e]SPIMAP was also run using a i.i.d. species-specific rate model.
[f]Because SYNERGY trees were downloaded, no run time was estimated.

of syntenic genes are orthologs in SYNERGY's trees. When given more iterations, PrIME-GSR's accuracy increases to 90.7% but computational time increases dramatically, 24-fold from 53 min to 20 h per gene tree. In contrast, SPIMAP achieved its accuracy of 96.5% in 29.1 min on average per tree and can achieve as much as 96.2% accuracy even when limited to an average run time of 1.0 min ("quick" mode). Also, SPIMAP achieves 96.3% ortholog accuracy when assessing the same 500 tree subset as PrIME-GSR's long mode. Note that the species tree aware programs (SPIMAP, SYNERGY, and PrIME-GSR) predict as much as 20% more ortholog pairs than the leading competing sequence-only program (PhyML).

For SPIMAP, performance was greater on RY-encoded alignments (96.5%) versus the full nucleotide alignments (92%, data not shown). This is likely due to that fact that the nucleotide alignments contained a gas chromatography (GC) bias that varies across species (supplementary table S1, Supplementary Material online), thus violating the stationarity assumption made in our implemented sequence evolution model (HKY). Reconstruction accuracy of PhyML and MrBayes was slightly diminished on RY-encoded alignments (63.0% and 61.1%, respectively), most likely due to the their lower information content. We also found that PrIME-GSR performs best on peptide data (88.9%), and that syntenic ortholog recovery decreased to 86.2% on nucleotide alignments using HKY (parameters estimated using PHYML) and 81.2% on RY-encoded alignments using JC69.

One important distinction between SPIMAP and PrIME-GSR is that SPIMAP models species-specific rate variation. To investigate the effect of this difference, we configured SPIMAP to learn an i.i.d. rates model similar to PrIME-GSR. For each branch, our modified training step estimated ($\alpha_i = 2.819, \beta_i = 663.0$) as the parameters for the i.i.d. gamma distributions. Reconstructing gene trees using these parameters, we found fewer syntenic orthologs (93.9%) and greater numbers of DLs.

Counting DL Events

Second, we evaluated the total numbers of DLs inferred across the clade (supplementary fig. S1, Supplementary Material online). Both SPIMAP and SYNERGY inferred at least 32% fewer duplications and 50% fewer losses than PrIME-GSR and the three sequence-only methods. The sequence-only methods, which do not use the species tree, infer many more events on nearly every branch, especially for short interior branches. The distribution of DL events that occur within each gene tree is illustrated in supplementary figure S2, Supplementary Material online. Interestingly, each of the other sequence-only methods inferred over four times as many gene duplication events and six times as many gene loss events as SPIMAP. For the sequence-only methods, duplications are more frequent near the root of the species tree and losses are more frequent near the leaves, a pattern suggesting that these events are erroneous (Hahn 2007).

Duplication Consistency Score

With our third metric, we sought to characterize the plausibility of the inferred duplications using the "duplication consistency score," introduced by Ensembl for evaluating their phylogenomic pipeline (Vilella et al. 2009). The consistency of a duplication node with children l and r is defined as $|A \cap B|/|A \cup B|$, where A and B are the set of species represented in descendants of l and r, respectively (see example in fig. 5*a*). The consistency score is designed to detect duplications that are wrongly inferred due to phylogenetic reconstruction errors because such false duplications are often followed by many compensating losses (Hahn 2007; Vilella et al. 2009) (i.e., low species overlap $|A \cap B|$). Figure 5 depicts the distribution for the duplication consistency score for each program. Both SPIMAP and SYNERGY showed similar consistency distributions that are heavily shifted toward 1 (47.8–49.0% and 4.2–17.2% of duplications with a score of 1 and 0, respectively; fig. 5). The sequence-only methods have many low scoring duplications (<11% and >70% with scores 1 and 0, respectively), an effect seen previously (Vilella et al. 2009). PrIME-GSR's distribution lies in between these extremes with 30.0% and 42.1% for scores 1 and 0, respectively. Lastly, the i.i.d. version of SPIMAP also scored lower than SPIMAP, inferring nearly

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

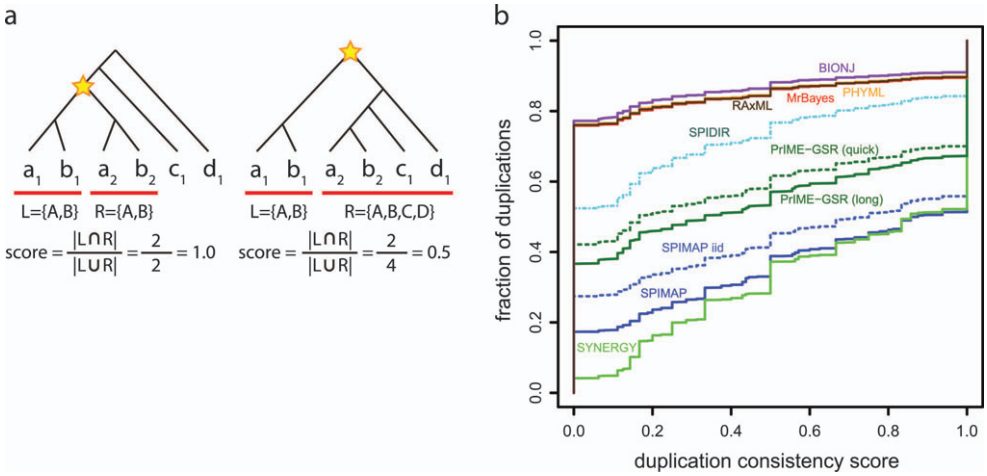


FIG. 5. The duplication consistency score for assessing phylogenetic methods. (*a*) Duplication consistency score computed on two example trees. For each duplication node (star), this score computes the number of species present in both the left and right subtrees divided by the total number of species descendant from the duplication node. Erroneous duplications show an increased rate of compensating losses and thus lower scores. (*b*) Cumulative distribution of duplication consistency scores for all duplications inferred in the 16 fungi data set by each method. SPIMAP (blue) and SYNERGY (green) perform best according to this metric, having the fewest duplications with low consistency scores. SPIMAP trained with an i.i.d. model similar to PrIME-GSR (dashed blue) infers duplications with overall lower consistency scores. These are followed by PrIME-GSR (dark green) and SPIDIR (dashed light blue) that show more moderate performance. Lastly, the four traditional methods implemented in the programs MrBayes, RAxML, PHYML, and BIONJ, all have similar and significantly lower score distributions.

twice the number of duplications with a consistency score of zero (fig. 5).

Recovering Gene Conversions

The fourth metric was specifically designed to test the case where species-level information is misleading, effectively testing the ability of species-aware methods to properly weigh species information against conflicting sequence information.

The fungal clade contains a whole-genome duplication (WGD) event, such that every gene simultaneously duplicated followed by many gene losses (Wolfe and Shields 1997; Kellis et al. 2004). Of the paralog pairs that are still present in the *Saccharomyces cerevisiae* genome, 37 of them have a K_s less than the average K_s between the *S. cerevisiae* and *S. bayanus* genomes of 1.05, indicating that these paralogs have undergone recent gene conversions near or after the speciation of the *S. cerevisiae* and *S. bayanus* lineages (Gao and Innan 2004) (see an example in fig. 6*a*). Also indicative of gene conversion (Noonan et al. 2004), these genes have a significantly elevated GC frequency of 42.0% in the third codon position, compared with a frequency of 37.9% for all *S. cerevisiae* genes ($P < 1.7 \times 10^{-09}$; Mann–Whitney *U*). Of these paralogs, SPIMAP infers 15 of them happening after the *S. bayanus* speciation and 31 after the *Candida glabrata* speciation (fig. 6*b*). In comparison, SYNERGY infers none of the paralogs duplicating after the *S. bayanus* speciation and only one after the *C. andidaglabrata* speciation. Instead 34 of the 37 paralogs are inferred as occurring on the branch containing the WGD, thus indicating that synteny information between *S. cerevisiae* and other postduplication species overrides sequence information in the vast majority of cases. For 33 families, the SPIMAP-constructed tree has a higher likelihood than the SYNERGY tree and for 22 families the likelihood is significantly higher ($P < 0.01$; SH test). In contrast, SYNERGY never has significantly higher likelihood.

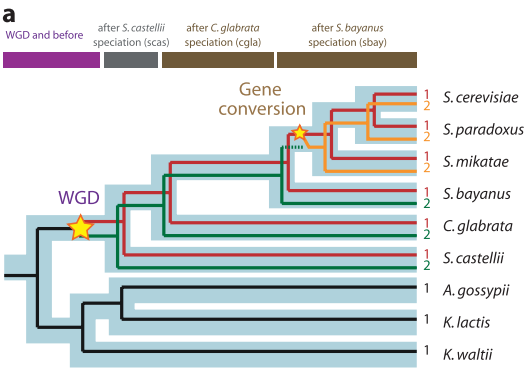
Together these four metrics applied to real gene trees from 16 fungi suggest that SPIMAP often outperforms both sequence-only and species-aware methods. From these trees, we observe what appears to be an over estimation of DL events by the other methods, an error that has been observed in previous empirical studies (Hahn 2007). To better understand how phylogenetic errors influence the accuracy of event inference, we turn now to simulated data.

Reconstructing Simulated Gene Trees

To test our method on a data set where the correct phylogeny is unambiguously known, we implemented a simulation program based on our model for gene family evolution. Our intent was to make the simulations realistic by capturing the same gene and species-specific rate variation as well as gene DL rates as seen in real gene trees. Thus, the same model parameters and species phylogeny were used as those estimated for both the 12 flies and 16 fungi clades.

For each clade, we simulated 1,000 gene trees and generated the corresponding nucleotide alignments (supplementary figs. S3 and S4, Supplementary Material online). Next, we reconstructed gene trees from these simulated alignments using SPIMAP, PrIME-GSR, and the other sequence-only phylogenetic methods. Note, SYNERGY was excluded from the analysis because synteny, which is one of its inputs, was not simulated. SPIMAP's substitution rate parameters were estimated on a simulated data set with no DLs (supplementary figs. S8 and S9, Supplementary Material online). Its DL parameters were trained from the gene counts of each

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018



b

Program	WGD	scas	cgla	sbay	% conversion recovered
SPIMAP	5	1	16	15	83.8%
SPIDIR	9	4	18	6	64.9%
SYNERGY	34	2	1	0	2.7%
PrIME-GSR	4	0	24	9	89.2%
MrBayes	5	0	21	11	86.5%
RAxML	6	0	22	9	83.8%
PHYML	5	0	19	13	86.5%
BIONJ	5	1	19	12	83.8%

FIG. 6. Inferred duplication times for recent *Saccharomyces cerevisiae* gene conversions. (*a*) Typical gene tree topology for 37 paralogous gene pairs originally arising from WGD and previously reported (Gao and Innan 2004) to have undergone gene conversion events (small star) near or after the speciation of *S. cerevisiae* and *S. bayanus*, such that one gene copy (green) is replaced by the other (red), followed by subsequent nucleotide divergence (orange). The correct inferred duplication of the two *S. cerevisiae* paralogs (red and orange lines, denoted 1 and 2) should occur within the time span indicated by the top brown bars. However, we expect methods that are heavily biased to follow the known species tree to incorrectly infer these events further up the tree. (*b*) We evaluated both traditional and species-aware methods in their ability to recover the correct trees in these cases and report the counts of where different gene conversion events are inferred for each method. We find that both SPIMAP and PrIME-GSR, as well as all traditional methods find the vast majority of these paralogs duplicates near or after *S. bayanus* speciation. However, SYNERGY incorrectly infers a WGD topology, most likely due to strong reliance in synteny information which is misleading in this case.

simulated data set (supplementary table S1, Supplementary Material online).

First, we measured topology accuracy across all the methods. SPIMAP outperforms the sequence-only programs by 7–29% on the simulated 12 flies data set and by 52–81% for the 16 fungi data set (fig. 7*a*). SPIMAP also showed a 3–8% accuracy increase over PrIME-GSR. To test whether lower reconstruction accuracy of the sequence-only methods was

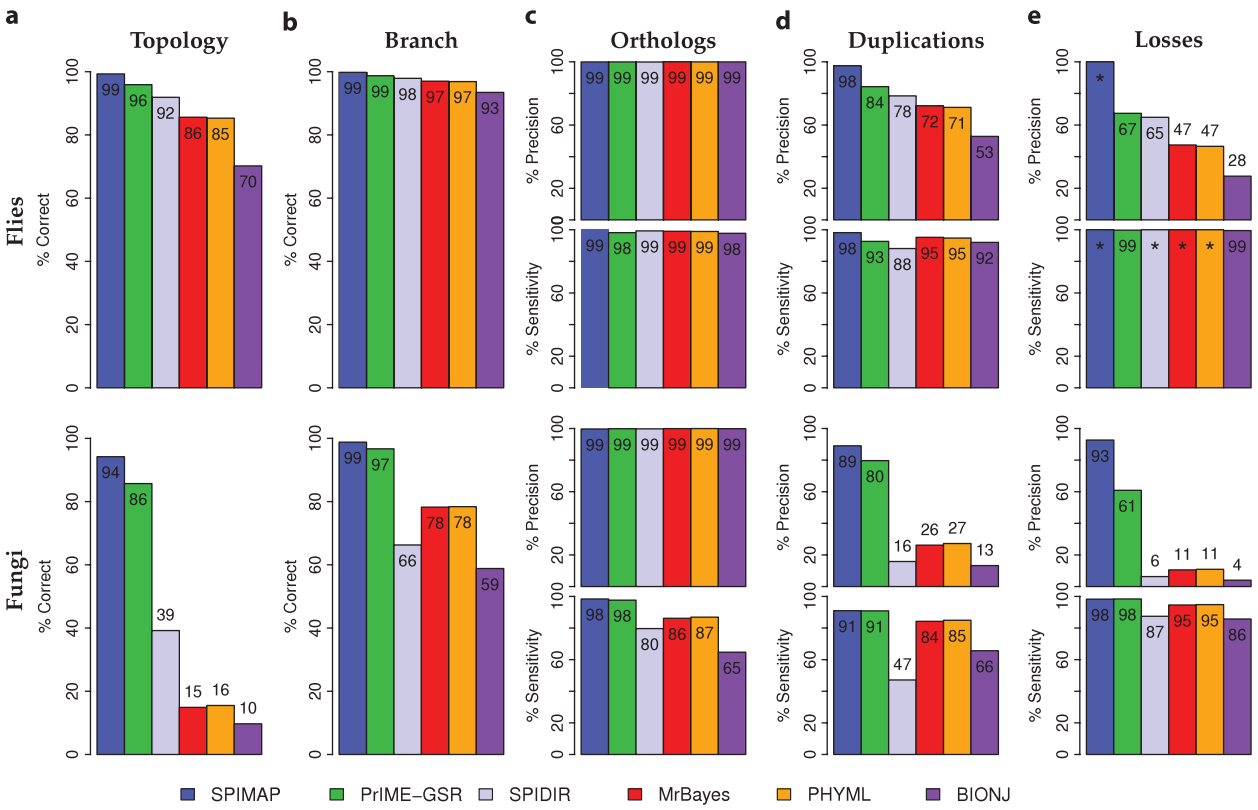


FIG. 7. Metrics of phylogenetic accuracy on simulated data sets for 12 *Drosophila* and 16 fungal species. (*a*) SPIMAP has a higher reconstruction accuracy for correctly inferring the full gene tree topology for both fly and fungal data sets. (*b*) The percent of accurately reconstructed branches is similar across methods for the 12 flies but a larger improvement is seen for SPIMAP on the larger and more diverse 16 fungi clade. (*c*) Despite topological and branch inaccuracies, pairwise ortholog detection is robust in both precision and sensitivity. (*d*, *e*) In contrast, DL inference is very sensitive to phylogenetic errors, especially in terms of precision. Stars indicate 100%.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

due to insufficient search, we performed an addition run of PhyML where the tree search was initialized with the correct tree. However, topology accuracy for these runs only increased from 15.5% to 16.7%, indicating errors due to insufficient search play only a minor role. Overall the accuracy improvement for SPIMAP is larger on the fungi data set, which has a more complex and divergent phylogeny.

Second, we assessed partial topology correctness using the percent of branches accurately reconstructed. For the flies, SPIMAP consistently performs better but by only a few percent (fig. 7*b*). However, for the fungi, SPIMAP again shows a larger accuracy improvement at 20–39% over the sequence-only methods.

Third, we looked at the percentage of orthologs inferred correctly, where we noticed a surprising trend. Although topologies and branches had high error rates for many methods, there was also a high percentage of correctly inferred ortholog pairs (fig. 7*c*). Upon closer inspection, we found that often when a branch is misplaced it only disrupts a small fraction of the pairwise orthologs. Thus, it appears that orthology discovery at the pairwise level is quite robust to phylogenetic errors. In addition, we noticed that false positive orthologs calls are rarely made, although false negatives are more frequent, especially on the fungal clade.

Fourth, we looked at the accuracy of inferring gene DLs, which is very important for studies interested in study the rate of such events. As opposed to the ortholog pairwise metric, we find that DLs are very sensitive to phylogenetic errors. Notice that although branch accuracy may be high for some programs and data sets, even a small number of errors can lead to dramatic overestimation DLs (fig. 7*c* and *d* and supplementary fig. S6*a* and S6*b*). In general, all programs are able to recover DL events for the flies and fungi data sets with similar sensitivity (fraction of true positives among all actual positives). Programs differ by less than 6% difference, with SPIDIR and BIONJ as outliers. However, in terms of precision (fraction of true positives among all predicted positives), SPIMAP has a dramatic improvement in event inference over sequence-only methods: 2127% and 4553% for the flies DL, respectively, and 58–69% and 75–80% for fungi DLs (with BIONJ as an outlier in each case). Compared with PrIME-GSR, SPIMAPshows a 9–12% increase in duplication precision and 32–33% increase in loss precision. The 2–3-fold over prediction of events by the sequence-only phylogenetic methods (fig. 7*c* and *d*) is an effect similar to that seen in the real data.

Lastly, we find that these results also hold when simulations are performed with unusually high DL rates at twice (2×) and four times (4×) the estimated true rates (1×). We performed simulations with five different settings 1×–1×, 2×–2×, 4×–4×, 4×–1×, 1×–4× for DL rates, respectively. We find that SPIMAP has increased performance for topology, branch, and event accuracy for all these rate settings (supplementary figs. S5*a* and S6*a*, Supplementary Material online). In addition, we found that SPIMAP was robust to errors in the DL rate parameters. When we reconstructed trees from the 1×–1× data set using DL rate parameters that were mis-specified to be four times faster than the true rates, topology accuracy was still 92.0% compared with 94.2% when using properly estimated parameters. Similarly, when reconstructing the 4×–4× data set using DL rate parameters that were one fourth the true rate, we obtained 71.1% topology accuracy compared with 69.2%.

Search Efficiency

In addition to evaluating reconstruction accuracy, we also evaluated reconstruction speed. Our goal with SPIMAP was to develop a method that is feasible enough to include in a phylogenomic pipeline containing thousands of trees and a variety of family sizes.

From the reconstruction of genes from our real data set (table 1), we found that SPIMAP has an average reconstruction time per tree (1.0 min) that is only slightly longer than that of PhyML (43.2 s). To investigate how our search strategy influences reconstruction run time, we generated a simulated data set of 500 gene families using 16 fungi species tree. For this simulation, we used i.i.d. species-specific rates ($\alpha_i = 2.819, \beta_i = 663.0$), no variation occurs in the gene rate, and the Jukes–Cantor model. We also used the same gene DL rates as estimated from real fungi gene families ($\lambda = 0.000732, \mu = 0.000859$). SPIMAP's substitution rate model was trained on a data set with the same parameters but no DLs. The parameters used by SPIMAP during reconstruction are given in supplementary figure S10, Supplementary Material online.

Although we have not implemented many optimizations for SPIMAP, our prescreening search strategy allows SPIMAP to compete with the highly optimized PhyML program (table 2). The RAxML program achieves significantly faster run times, but this occurs with a small decrease in accuracy for this data set. We believe our speed increase is because the gene family model, through the use of the species tree in the prior, produces a posterior distribution that is far more concentrated than the likelihood. Thus, many seemingly equivalent trees from a likelihood perspective are significantly different based on their priors and posteriors. In addition, our prescreening search strategy (see Rapid Tree Search) appears to greatly help in speeding up discovery of the MAP gene tree. For example, with no prescreening, SPIMAP achieves a topology accuracy of 32.4% with an average run time of 7.2 s. By using 100 prescreening iterations, accuracy increases to 84.8%, whereas run time only increases to 8.5 s. For comparison, PhyML achieves 26.0% topology accuracy in about 25.8 s on average.

Whether this prescreening strategy can scale to much larger trees with thousands of sequences (Stamatakis et al. 2005; Price et al. 2010), remains to be seen. However, the duplication subtree factoring within the topology prior may allow reuse of many computations between tree local rearrangements and could be combined with existing strategies for speeding up tree search.

SPIMAP is currently implemented as a MAP method, thus if branch support values are needed, bootstrapping will be required. Given the speed of our search, we can perform 100 bootstraps in about 11.1 minutes to achieve 86.4%

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

Table 2. Evaluation of Search Time for Several Phylogenetic Methods.

Program	Iterations[a]	Prescreens[b]	Bootstraps	Topology(%)	Branch(%)	Run time
RAxML	—	—	0	20.8	80.3	3.5 s
RAxML	—	—	100	22.8	1.1	39.8 s
PhyML	—	—	0	26.0	83.9	25.8 s
PhyML	—	—	100	26.0	83.9	13.9 min
SPIMAP	50	1	0	32.4	81.4	7.2 s
SPIMAP	100	1	0	50.8	87.1	12.7 s
SPIMAP	500	1	0	83.8	96.0	1.2 min
SPIMAP	1,000	1	0	88.6	97.5	2.0 min
SPIMAP	50	100	0	84.8	96.7	8.5 s
SPIMAP	1,000	100	0	90.8	98.1	2.3 min
SPIMAP	50	100	100	86.4	97.1	11.1 min

[a]Number of iterations used for each method.
[b]Number of prescreening iterations used for SPIMAP.

accuracy. This run time is comparable to 100 bootstraps of PhyML at 13.9 min and 26.0% accuracy. Thus, bootstrap analysis is quite feasible for SPIMAP, and the method should be efficient and practical enough for any pipeline that uses phylogenetic programs with run times on the order of PhyML's.

Lastly, we evaluated the influence of run time and family size on reconstruction accuracy. Using the same parameters above, we simulated more gene trees from the 16 fungal species tree and divided them into six classes based on the their number of extant genes: 5–9, 10–19, 20–29, 30–39, 40–49, and 50–59. Each size class was populated with 100 simulated trees and alignments. SPIMAP was run in two modes, one without bootstrapping (1,000 iterations and 100 prescreens) and one with 100 bootstraps (100 iterations and 100 prescreens). For the middle gene size class 20–29, SPIMAP achieved average run times of 5.3 and 50.4 min, respectively. For each data set, PrIME-GSR was also executed, using the same amount of time as SPIMAP, which required 7,300 iterations (quick mode) and 77,000 iterations (long mode) . We find that for smaller trees with 5–29 extant genes that both SPIMAP runs and PrIME-GSR's long mode achieve similar topology accuracy in the range of 80–100% (supplementary fig. S11, Supplementary Material online). However, for larger gene trees with 30–49 extant genes, as accuracy degrades for all methods, both modes of SPIMAP have a 20% increase in topology accuracy over PrIME-GSR. Improvements in inferring DL accuracy is also seen for the larger trees (>10% increase in duplication precision and >30% increase in loss precision).

Discussion

We present a novel probabilistic model and algorithm for gene tree reconstruction. Our approach uses a Bayesian framework to model sequence evolution, gene duplication, loss, and substitution rate variation, thus incorporating many disparate types of information in a principled way. This unified framework presents many advantages.

In contrast to previous gene tree reconstruction methods (Zmasek and Eddy 2002; Dufayard et al. 2005; Durand et al. 2006; Vilella et al. 2009), where a gene tree is reconciled only after full reconstruction by a method such as NJ or ML, our method finds a reconciliation and gene tree simultaneously. In addition, the parameters of our model are interpretable (e.g., substitutions rates and duplication/loss rates), and we have provided training algorithms for each one. This provides an advantage over a method like SYNERGY (Wapinski et al. 2007) that optimizes a parsimony-based cost function for several different events such duplications, loss, and syntenic relationships. Without a probabilistic basis, the weights of these costs and the behavior of their combination are more difficult to determine and analyze. Our study of gene conversions demonstrates more work is needed to understand how synteny information should be weighed against conflicting sources of information.

Our method models rate variation that is correlated across all branches of the tree (gene-specific rate) as well as rates specific to each species lineage (species-specific rates). We have found that when both these effects are modeled, the result is a more informative prior which leads to increased reconstruction accuracy (see the i.i.d. version of SPIMAP in table 1 and fig. 5). In contrast, PrIME-GSR uses identical and independent gamma distributions for rate variation which do not model species-specific rate variation. Thus, species with rate acceleration or decelerated across the genome will have branches that are consistently penalized by an i.i.d. rate prior. One complication for modeling species-specific rates is possibility of overparameterizing the model. We addressed this issue by learning the rate distributions prior to reconstruction from a data set of multiple orthologous gene trees. This learning step provides an advantage even when i.i.d. rates are used (see SPIMAP with i.i.d. and PrIME-GSR in table 1). By combining data across loci, the rate variation prior can be estimated more accurately than if the gene trees were considered in isolation.

The rate prior of our current work builds upon a previously developed method, SPIDIR (Rasmussen and Kellis 2007). We designed SPIDIR to be a distance-based likelihood method that exploits the rate variations we had observed in the 12 fly and 9 fungal genomes. Although the method proved effective, its reliance on pairwise distances did not fully utilize the available character information and it lacked an explicit model for DL rates. Indeed, we find in our latest comparison that SPIMAP has more consistent accuracy improvements than SPIDIR even for larger species trees

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

(16 fungi) and fast rates of DL (fig. 7 and supplementary figs. S5*a* and S6*a*, Supplementary Material online).

Thus far, we have only implemented a very simple model for sequence evolution within the SPIMAP program. Currently SPIMAP uses the HKY model, although providing additional models as well as modeling rate variation across sites may lead to additional improvements to reconstruction accuracy. However, we note that for the evaluations we present here, that modeling rate variation across sites did not contributed significantly to improved reconstructions. In fact, the PhyML program found very similar recovery rates for syntenic orthologs using both rate variation (64.2% recovery using four categories and an estimated α) or using none (64.9% recovery).

Lastly, we envision this method participating in a larger phylogenomic pipeline. We believe that within most clades of interest, there will be sufficient data for training our model. For example, in the 12 sequenced *Drosophila* species, about one-third of all genes are syntenic across all 12 species (Clark et al. 2007; Rasmussen and Kellis 2007) and thus can serve as a training set for our substitution rates model. Once a model is learned from simple gene families, it can then be applied to reconstruct gene families with more complicated histories of gene DL. Given these advances and many others to come, phylogenetics will likely play an ever increasing role in understanding the evolution and function of genomes.

Supplementary Material

Supplementary tables S1, sections 2.1–2.3, and figures S1–S11 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Michael F. Lin and the rest of the MIT Compbio group for helpful comments, feedback, and discussions. We thank M.R.'s thesis committee, Scott V. Edwards, Eric J. Alm, and Tommi S. Jaakkola for frutiful discussions at various stages of this work. This work was supported by National Science Foundation (NSF) CAREER award NSF 0644282 to M.K.

References

Adams MD, Celniker SE, Holt RA, et al. (195 co-authors). 2000. The genome sequence of Drosophila melanogaster. *Science* 287: 2185–2195.

Åkerborg O, Sennblad B, Arvestad L, Lagergren J. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A*. 106:5714–5719.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Arvestad L, Berglund AC, Lagergren J, Sennblad B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19(Suppl 1):i7–i15.

Arvestad L, Berglund A, 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; 2004 March 27-31; San Diego, CA. New York: Association for Computing Machinery (ACM). p. 326–335.

Arvestad L, Lagergren J, Sennblad B. 2009. The gene evolution model and computing its associated probabilities. J. ACM 56:1–44.

Butler G, Rasmussen MD, Lin MF, et al. (51 co-authors). 2009. Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature* 459:657–662.

Chen K, Durand D, Farach-Colton M. 2000. Notung: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol*. 7:429–447.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.

Clark AG, Eisen MB, Smith DR, et al. (417 co-authors). 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450:203–218.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. 2003. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science* 301:71–76.

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

Datta RS, Meacham C, Samad B, Neyer C, Sjlander K. 2009. Berkeley PHOG: phylofacts orthology group prediction web server. *Nucleic Acids Res*. 37:W84–W89.

Dehal PS, Boore JL. 2006. A phylogenomic gene cluster resource: the phylogenetically inferred groups (PHIGs) database. *BMC Bioinformatics* 7:201.

Dietrich FS, Voegeli S, Brachat S, et al. (14 co-authors). 2004. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* 304:304–307.

Doyon JP, Chauve C, Hamel S. 2009. Space of gene/species trees reconciliations and parsimonious models. *J Comput Biol*. 16:1399–1418.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21:2596–2603.

Dujon B, Sherman D, Fischer G, et al. (67 co-authors). 2004. Genome evolution in yeasts. *Nature* 430:35–44.

Durand D, Halldorsson BV, Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*. 13:320–335.

Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183:259–274.

Eddy SR. 2000. HMMER: profile hidden Markov models for biological sequence analysis [Internet]. Ashburn (VA): Janelia Farms Laboratory; 2000 [cited 2010 July 25]. Available from: http://hmmer.org.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*. 8:163–167.

Feller W. 1939. Die grundlagen der Volterraschen Theorie des kampfes ums dasein in Wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheo*. 5:11–40.

Felsenstein J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*. 17:368–376.

Felsenstein J. 2005. PHYLIP (phylogeny inference package). Version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Gao L-z, Innan H. 2004. Very low gene duplication rate in the yeast genome. *Science* 306:1367–1370.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14:685–695.

Goffeau A, Barrell BG, Bussey H, et al. (16 co-authors). 1996. Life with 6000 genes. *Science* 274:546, 563–546, 567.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018

Goodman M, Czelusniak J, Moore G, Romero-Herrera A, 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.

Gu X, Zhang H. 2004. Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol*. 21:1401–1408.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Hahn M. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol*. 8:R141.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*. 15:1153–1160.

Hahn MW, Han MV, Han S-G. 2007. Gene family evolution across 12 Drosophila genomes. *PLoS Genet*. 3:e197.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22:160–174.

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

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. 2007. The human phylome. *Genome Biol*. 8:R109.

Jones T, Federspiel NA, Chibana H, et al. (12 co-authors). 2004. The diploid genome sequence of candida albicans. *Proc Natl Acad Sci U S A*. 101:7329–7334.

Jukes T, Cantor C. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature* 428:617–624.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–254.

Li H, Coghlan A, Ruan J, et al. (15 co-authors). 2006. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 34:D572–D580.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*. 56:504–514.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*. 55:21–30.

Massey SE, Moura G, Beltro P, Almeida R, Garey JR, Tuite MF, Santos MAS. 2003. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. *Genome Res*. 13:544–557.

Moschopoulos P. 1985. The distribution of the sum of independent gamma random variables. *Ann Inst Stat Math*. 37:541–544.

Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res*. 14:354–366.

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

Page RD, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*. 7:231–240.

Price MN, Dehal PS, Arkin AP. 2010. Fasttree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

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.

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

Richards S, Liu Y, Bettencourt BR, et al. (52 co-authors). 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res*. 15:1–18.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.

Ronquist F, Huelsenbeck JP. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 4:406–425.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*. 16:1114–1116.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21:456–463.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Mol Biol Evol*. 21:36–44.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19:327–335.

Wakeley J. 2009. Coalescent theory: an introduction. Greenwood Village (CO): Roberts & Company Publishers.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2009. Synergy dataset January 2009 update. Cambrdige (MA): Harvard University; 2009 [cited 2010 July 25]. Available from: http://www.broadinstitute.org/regev/orthogroups/.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17:821–828.

Zmasek CM, Eddy SR. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*. 3:14.

Downloaded from https://academic.oup.com/mbe/article-abstract/28/1/273/980067
by Auburn University Libraries user
on 08 June 2018