

Fast and consistent estimation of species trees using supermatrix rooted triples

Michael DeGiorgio* and James H. Degnan†

*Center for Computational Medicine and Bioinformatics, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA; and †Department of Mathematics and Statistics, Private Bag 4800, University of Canterbury, Christchurch 8140 New Zealand

Running Head: Supermatrix rooted triples

Corresponding authors:

Michael DeGiorgio
Center for Computational Medicine and Bioinformatics
University of Michigan
2017 Palmer Commons
100 Washtenaw Avenue
Ann Arbor, MI 48109-2218, USA
Phone: +1 734 615 9551
Fax: +1 734 615 6553
E-mail: degiormi@umich.edu

James H. Degnan
Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch 8140
New Zealand
+64 (03) 364 2600
+64 (03) 364 2587
j.degnan@math.canterbury.ac.nz

Abstract

Concatenated sequence alignments are often used to infer species-level relationships. Previous studies have shown that analysis of concatenated data using maximum likelihood (ML) can produce misleading results when loci have differing gene tree topologies due to incomplete lineage sorting. Here, we develop a polynomial-time method that utilizes the modified mincut supertree algorithm to construct an estimated species tree from inferred rooted triples of concatenated alignments. We term this method SuperMatrix Rooted Triple (SMRT) and use the notation SMRT-ML when rooted triples are inferred by ML. We use simulations to investigate the performance of SMRT-ML under Jukes-Cantor and General Time-Reversible substitution models for four- and five-taxon species trees, and also apply the method to an empirical dataset of yeast genes. We find that SMRT-ML converges to the correct species tree in many cases in which maximum likelihood on the full concatenated dataset fails to do so. SMRT-ML can be conservative in that its output tree is often partially unresolved for problematic clades. We show analytically that when the species tree is clocklike and mutations occur under the Cavender-Farris-Neyman substitution model, as the number of genes increases, SMRT-ML is increasingly likely to infer the correct species tree even when the most likely gene tree does not match the species tree. SMRT-ML is therefore a computationally efficient and statistically consistent estimator of the species tree when gene trees are distributed according to the multispecies coalescent model.

Key words.—phylogenetics, phylogenomics, anomaly zone, anomalous gene tree, statistical consistency, lineage sorting.

Introduction

A species tree is a branching pattern representing the divergence of multiple species, whereas a gene tree depicts the evolutionary history of a single gene. Though only a single species tree exists, trees for different genes often have conflicting topologies. This discordance of gene trees with the species tree is due to processes such as gene duplication, horizontal gene transfer, and incomplete lineage sorting (Page and Charleston 1997; Maddison 1997; Than et al. 2007; Degnan and Rosenberg 2009).

When analyzing data from multiple loci, the most frequently occurring gene tree topology is sometimes used as an estimate of the species tree topology. For example, in a study of 30 loci, Jennings and Edwards (2005) used the gene tree that was inferred in 16 of 28 resolved topologies from three ingroup species of Australian grass finches as the species tree topology. However, even in the absence of complications such as hybridization (Buckley et al. 2006; Holland et al. 2008; Meng and Kubatko 2009) and population structure (Slatkin and Pollack 2008), this procedure is only justified for studies of three taxa. This is because the most likely three-taxon gene tree is expected to match the species tree topology when incomplete lineage sorting is modeled by the multispecies coalescent (Nei 1987; Pamilo and Nei 1988). However, when a species tree has four taxa and is asymmetric, or has five or more taxa, the most likely gene tree does not necessarily match the species tree (Degnan and Rosenberg 2006; Rosenberg and Tao 2008). Such anomalous gene trees (AGTs; Degnan and Rosenberg 2006) occur when the species tree falls into a particular space of branch lengths called the anomaly zone. Anomaly zones for four-taxon and five-taxon species trees are depicted in Figure 2 of Degnan and Rosenberg (2006) and Figures 3–5 of Rosenberg and Tao (2008), respectively.

The absence of AGTs for rooted three-taxon trees motivates the development of methods for inferring species trees using rooted triples, or three taxa at a time (Degnan and Rosenberg 2006), as has been described for rooted triple consensus methods (Ewing et al. 2008; Degnan et al. 2009) and supertree methods (Steel and Rodrigo 2008; Willson 2009). Supertree methods generalize consensus methods to the setting in which input gene trees have overlapping subsets of taxa that need not be identical (Bininda-Emonds 2004). Because a rooted tree is completely described by its set of rooted triples (Steel 1992), we can utilize a supertree method to construct the species tree from correctly inferred rooted triples.

Supertree and other phylogenetic methods can be applied to sets of concatenated alignments, or supermatrices, to infer a species tree. A concatenated alignment contains sequences of multiple loci linked together to create a single “supergene” (Rokas et al. 2003; de Queiroz and Gatesy 2007), thus increasing the size of the dataset. Though statistical power generally increases with the size of a dataset, the accuracy of concatenation is currently under debate. Rokas et al. (2003) reported that the application of phylogenetic inference methods to con-

catenated sequence alignments can yield a strongly supported inferred species tree. However, several studies (Kolaczkowski and Thornton 2004; Mossel and Vigoda 2005; Edwards et al. 2007; Kubatko and Degnan 2007) have also shown that inferring trees from concatenated data with maximum likelihood (ML) can perform poorly when sites are generated under different tree topologies and can produce bootstrap values that are misleadingly high (Gadagkar et al. 2005; Kubatko and Degnan 2007).

Here, we develop a divide-and-conquer approach (Cormen et al. 2001) called SuperMatrix Rooted Triple (SMRT), which is a polynomial-time algorithm that circumvents some of the weaknesses of concatenation by linking it with rooted triple and supertree methods. SMRT assembles rooted triples inferred from concatenated alignments into a species tree using a supertree algorithm such as modified mincut (MMC) (Page 2002). We compare SMRT in which rooted triples are inferred by maximum likelihood (SMRT-ML) to the method in which all taxa are analyzed simultaneously by applying ML to a supermatrix (SM-ML). In simulations that assume a molecular clock, SMRT-ML performs favorably on four- and five-taxon species trees both inside and outside the anomaly zone. Further, introducing two model violations—analysis under a molecular clock when gene trees are not clocklike and analysis under an incorrect substitution model—has little effect on the performance of SMRT-ML. We illustrate the SMRT-ML procedure using a yeast dataset frequently analyzed in phylogenetic studies (Rokas et al. 2003; Gatesy and Baker 2005; Edwards et al. 2007) and find that SMRT-ML recovers the same species tree as that found using either SM-ML or the software BEST (Liu 2008).

Assuming that incomplete lineage sorting is the source of discordance of gene trees with species trees and that there are no hybridization or horizontal gene transfer events, we prove that SM-ML is a statistically consistent estimator for three-taxon clocklike species trees when concatenated sequence alignments are generated from a coalescent distribution under a molecular clock and a binary substitution model (Neyman 1971). Under the same set of assumptions, we then prove in Theorem 3 that SMRT-ML is a statistically consistent estimator of a species trees. Therefore, our computationally efficient strategy is justified both theoretically and through simulations in the context of gene tree conflict due to incomplete lineage sorting. Although we assume here that rooted triples are inferred using a ML method, we stress that SMRT is a general approach that can utilize rooted triples that have been inferred from other methods such as parsimony and distance methods as well. We focus on triples inferred from ML because we compare our method to a method in which trees are inferred by ML from concatenated alignments.

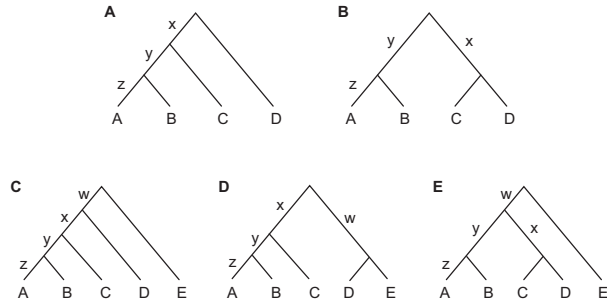


FIG. 1.—Four- and five-taxon clocklike species tree topologies. (A, B) Four-taxon species tree topologies with branch lengths x , y , and z . (C–E) Five-taxon species tree topologies with branch lengths w , x , y , and z . Branch lengths are in coalescent time units $t/(2N_e)$, where t is the time in generations and N_e is the effective population size. For all simulations, we let $z = 1$.

Methods

Supermatrix rooted triple (SMRT)

The SMRT approach takes a concatenated alignment of n taxa and breaks it into $\binom{n}{3}$ alignments, one for each set of three taxa. A rooted three-taxon tree is inferred for each alignment using any phylogenetic method by either assuming a molecular clock, or by including a known outgroup as a fourth taxon to root the tree. The species tree is then constructed by using the resulting rooted triples as input for a supertree algorithm. Here, we use MMC, which extends the mincut algorithm (Semple and Steel 2000). The mincut algorithm satisfies five desirable properties: (1) the order of the input set of trees does not affect the method; (2) relabeling the set of taxa of the input trees produces the same output tree on the relabeled set of taxa; (3) if there exists a tree that has each input tree as a subtree, then the output tree will display all of these trees; (4) any taxon that is in the input set of trees is also in the output tree; and (5) the method is polynomial in the number of distinct taxa (Semple and Steel 2000). Page (2002) created MMC by modifying the mincut method so that uncontradicted nestings are preserved in the output tree.

Simulation

We examined the performance of SMRT-ML using simulated sequence alignments. First we chose a species tree σ with topology $((((AB)C)D)$, $((AB)(CD))$, $((((AB)C)D)E)$, $((((AB)C)(DE))$, or $((((AB)(CD))E)$. Model species tree topologies are depicted in Figure 1. Branch lengths and probabilities for the matching gene tree topology and most probable nonmatching gene tree topologies are shown in Table 1. The branch lengths chosen for the species tree $((((AB)C)D)$ are the same as those used in Kubatko and Degnan (2007). One additional case was considered in which both internal branch lengths equal 0.1 coalescent units for the $((((AB)C)D)$ species tree. For each species tree and each simulation replicate, using COAL (Degnan and Salter 2005) conditional on σ , we simulated $m = 100, 200, 300,$

400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000, 5000, and 6000 independent (within and across each set) gene trees with branch lengths. Branch lengths were simulated in coalescent units, $t/(2N_e)$, where t is the number of generations, and N_e is the effective population size. We converted the branch lengths for each gene tree to mutation units by multiplying each length by $\theta/2$, where $\theta = 4N_e\mu$, and μ is the mutation rate per site per generation. As a consequence, all populations had equal values of θ . For each gene tree, we converted branch lengths to the expected number of mutations by multiplying them by $\theta/2$, where $\theta = 0.01$. We generated sequence alignments of length $L = 500$ nucleotides (nt) with Seq-Gen (Rambaut and Grassly 1997). These m independent alignments were concatenated to create single n -taxon alignments of length mL .

The concatenated alignments were then broken into all $\binom{n}{3}$ three-taxon alignments of length mL . We inferred rooted ML trees for the n -taxon alignment, as well as for all three-taxon alignments, employing an exhaustive search over all tree topologies from PAUP* (Swofford 2003). All three-taxon rooted trees were entered as input to the program `supertree` (Page 2002), which implements the MMC algorithm. Each time PAUP* was called, it returned $k \geq 1$ tree topologies tied for the most likely species tree. The count for each of the tied topologies was increased by $1/k$. We repeated this procedure, beginning with the simulation of gene trees, 300 times for each combination of species tree topology and number of loci. The count for each tree topology was averaged over all replicate simulations. Unless otherwise stated, the results are for data simulated under a Jukes-Cantor (JC) model and analyzed under ML assuming JC and a molecular clock. A schematic of the simulation procedure is provided in Figure 2.

Empirical example

SMRT-ML was applied to analyze a yeast dataset consisting of 106 genes spanning over 127,000 nt (Rokas et al. 2003). We used ML in PAUP* under a GTR + Γ + I model without a molecular clock on each of the $\binom{7}{3} = 35$ three-taxon subsets of the seven ingroup taxa, using the outgroup *C. albicans* to root the triples. In addition to the full concatenated alignment, we analyzed concatenated alignments of random subsets of $m = 10, 20, 30, 40, 50, 60, 70, 80, 90,$ and 100 genes. For each value of m , SMRT-ML was applied to 300 random subsets of m genes, and we reported the proportion of times that SMRT-ML returned either the presumed species tree or a tree with at least one false clade. Bootstrapping for SMRT-ML was performed by reading the concatenated sequence data in R (R Development Core Team 2008) and using the `sample` function to create 300 bootstrap eight-taxon alignments, SMRT-ML was applied to each bootstrap replicate in separate PAUP* runs.

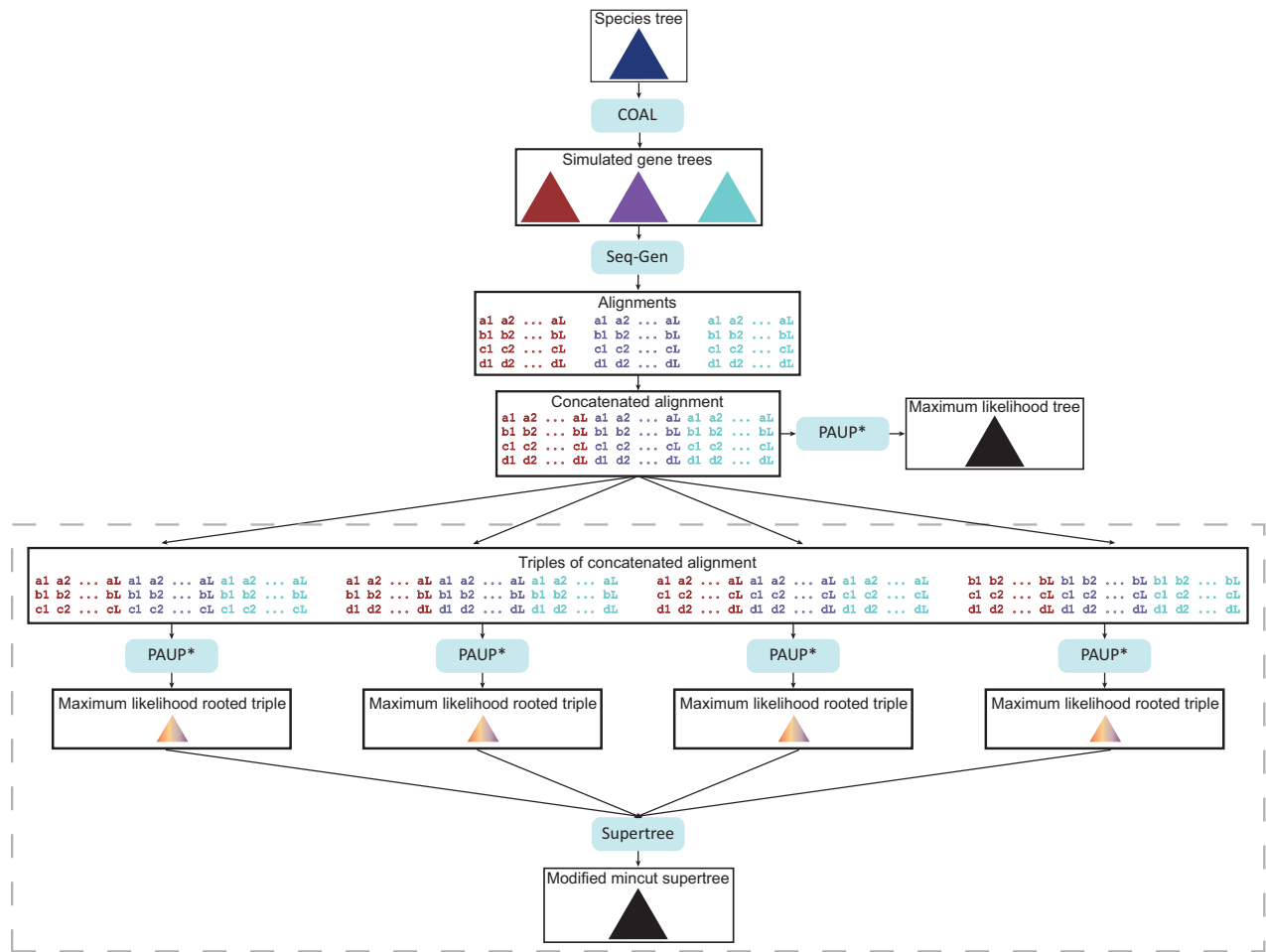


FIG. 2.—Schematic of our simulation procedure. First, an n -taxon species tree is chosen with branch lengths, which is fed through COAL (Degnan and Salter 2005) to produce a set of n -taxon gene trees simulated under this species tree. Seq-Gen (Rambaut and Grassly 1997) is then used to create alignments of n species based on the gene trees, which are linked to create a single concatenated alignment. The concatenated alignment is analyzed under maximum likelihood (SM-ML) with PAUP* (Swofford 2003) to infer a species tree. The concatenated alignment is also broken into all $\binom{n}{3}$ alignments of three species, which are then fed through PAUP* to infer a total of $\binom{n}{3}$ rooted triples. These rooted triples are used as input to supertree (Page 2002) to infer a species tree (SMRT-ML). The dashed gray box represents the part of the procedure that is SMRT-ML.

Table 1 Probabilities of concordant and most probable discordant gene trees and performance of SM-ML and SMRT-ML with 6000 loci

Species tree σ	Branch lengths (x,y) or (w,x,y) (see Figure 1)	Highest-prob. non-matching gene tree	Concordance prob.	Highest-prob. non-matching tree	SM-ML correct > 50%	SMRT-ML correct > 50%	Figures
(((AB)C)D)	(0.01, 2.0)	((AB)(CD))	0.30170	0.30039	NO	NO	3A,I
	(0.05, 1.0)	((AB)(CD))	0.25483	0.24116	NO	YES	3B,J
	(0.1, 1.0)	((AB)(CD))	0.27762	0.23099	YES	YES	3C,K
	(0.1568, 0.1568)	((AB)(CD))	0.13344	0.13349 ^a	YES	YES	3D,L
	(0.01, 1.0)	((AB)(CD))	0.23595	0.24948 ^a	NO	NO	3E,M
	(0.05, 0.05)	((AB)(CD))	0.07879	0.12079 ^a	NO	YES	3F,N
	(0.1, 0.05)	((AB)(CD))	0.08867	0.11901 ^a	NO	YES	3G,O
	(0.25, 0.01)	((AB)(CD))	0.10376	0.10511 ^a	YES	YES	3H,P
	(0.1, 0.1)	((AB)(CD))	0.10370	0.12792 ^a	NO	YES	S13B
((AB)(CD))	(0.01, 2.0)	(((AB)C)D), (((AB)D)C)	0.30929	0.29280	YES	NO	4A,I
	(0.05, 1.0)	(((AB)C)D), (((AB)D)C)	0.27612	0.21987	YES	YES	4B,J
	(0.1, 1.0)	(((AB)C)D), (((AB)D)C)	0.29946	0.20915	YES	YES	4C,K
	(0.1568, 0.1568)	(((AB)C)D), (((AB)D)C), (((CD)A)B), (((CD)B)A)	0.18497	0.08196	YES	YES	4D,L
	(0.01, 1.0)	(((AB)C)D), (((AB)D)C)	0.25659	0.22884	YES	NO	4E,M
	(0.05, 0.05)	((AC)(BD)), ((BC)(AD))	0.13384	0.10054	YES	YES	4F,N
	(0.1, 0.05)	((AC)(BD)), ((BC)(AD))	0.14516	0.09563	YES	YES	4G,O
	(0.25, 0.01)	(((CD)A)B), (((CD)B)A)	0.16346	0.11584	YES	NO	4H,P
((((AB)C)D)E)	(0.1, 0.1, 0.1)	(((AB)C)(DE))	0.02217	0.03321 ^a	NO	YES	5A,E
	(0.1, 1.0, 0.1)	(((AB)C)(DE))	0.09388	0.08158	NO	YES	5B,F
	(0.1, 0.1, 1.0)	(((AB)C)(DE))	0.07055	0.08941 ^a	NO	YES	5C,G
	(1.0, 0.1, 0.1)	(((AB)(CD))E)	0.06547	0.07705 ^a	YES	YES	5D,H
(((AB)C)(DE))	(0.1, 0.1, 0.1)	(((DE)C)(AB))	0.04002	0.03034	YES	YES	6A,E
	(0.1, 1.0, 0.1)	(((AC)B)(DE)), (((BC)A)(DE))	0.10506	0.07656	YES	YES	6B,F
	(0.1, 0.1, 1.0)	(((AB)D)(CD)), (((AB)E)(CD))	0.10970	0.06465	YES	YES	6C,G
	(1.0, 0.1, 0.1)	(((DE)C)(AB))	0.07781	0.08825 ^a	NO	YES	6D,H
(((AB)(CD))E)	(0.1, 0.1, 0.1)	(((AB)E)(CD)), (((CD)E)(AB))	0.02914	0.03626 ^a	YES	YES	7A,E
	(0.1, 1.0, 0.1)	(((CD)E)(AB))	0.07339	0.09065 ^a	NO	YES	7B,F
	(0.1, 0.1, 1.0)	(((AB)E)(CD))	0.07339	0.09065 ^a	YES	YES	7C,G
	(1.0, 0.1, 0.1)	(((AC)(BD))E), (((AD)(BC))E)	0.09591	0.05231	YES	YES	7D,H
((((AB)(CD))E)F)	see Figure S12A	(((AB)(CD))(EF))	0.01525	0.02343 ^a	NO	YES	S10B

^aThe most probable gene tree is an AGT.

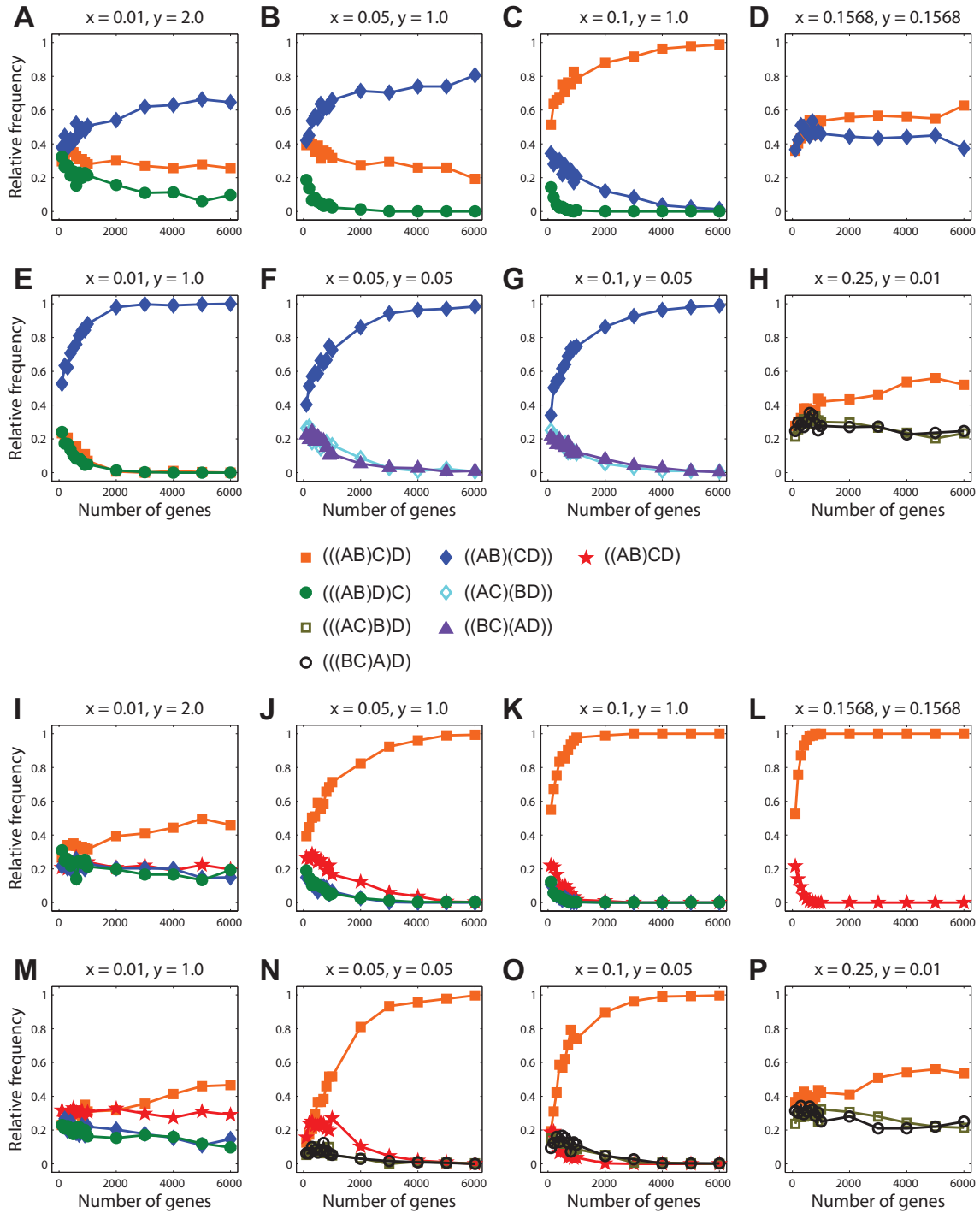


FIG. 3.—Results of simulations for the four-taxon tree $((AB)C)D$ (Figure 1A) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML (resimulated from Kubatko and Degnan (2007)). (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

Results for simulations

Four taxa

A four-taxon asymmetric species tree is depicted in Figure 1A. Figures 3A–H and Figures 3I–P display simulation results for this species tree for SM-ML and SMRT-ML, respectively. As shown by Kubatko and Degnan (2007) and replicated here, SM-ML is misleading in that increasing the number of loci can make it more likely to return an incorrect species tree. In contrast, SMRT-ML outperformed SM-ML on the (((AB)C)D) species tree for all branch lengths tried except for $(x, y) = (0.25, 0.01)$ (Figures 3H, P), where both methods performed poorly. For these branch lengths, using 6000 loci, SM-ML returned the species tree 52% of the time, SMRT-ML returned the species tree 54% of the time, and both methods returned each of the nonmatching trees ((AC)B)D and ((BC)A)D less than 25% of the time. For extremely small branch lengths of 0.01, the proportion of times that SMRT-ML recovers the species tree topology increases slowly (Figures 3I, M, P). However, the method does not appear to be misleading, suggesting that there is a trade-off between consistency and speed of convergence as was seen for consensus methods by Degnan et al. (2009). For these sets of branch lengths, the proportion of times that SM-ML returned the species tree either increased just as slowly (Figure 3H) or was misleading (Figures 3A, E). Even though SMRT-ML did not always infer the matching species tree when $x = 0.01$, it often inferred the partially unresolved tree ((AB)CD) (e.g., Figures 3I, M), which is not misleading for the species tree topology. On the other hand, for $(x, y) = (0.1, 1.0)$ (Figures 3C, K), both methods converged to the species tree with SMRT-ML converging more quickly than SM-ML; however, only SMRT-ML was increasingly likely to recover the species tree as loci were added for all branch lengths tried.

Simulation results on the four-taxon symmetric species tree (Figure 1B) are shown in Figures 4A–H (SM-ML) and Figures 4I–P (SMRT-ML). In contrast to what was observed for the asymmetric tree, for the symmetric tree SM-ML is not misleading and converges to the true species tree faster than SMRT-ML for each set of branch lengths tested. This observation is not surprising, given that no anomaly zone exists for the four-taxon symmetric species tree and that SM-ML simultaneously analyzes all available sequence data for the four taxa. However, one must also be careful in assuming that SM-ML will perform well outside of the anomaly zone because the anomaly zone has no obvious relationship to the problems encountered with concatenation. As with the case for the asymmetric tree, SMRT-ML tends to have a slow rate of convergence at extremely small branch lengths (Figures 4I, M, P). However, it is still not misleading and frequently returns either the ((AB)CD) or ((CD)AB) partially unresolved tree. Thus, although SMRT-ML can be slower to converge to the species tree for symmetric four-taxon trees, simulations for both symmetric and asymmetric four-taxon species trees suggest that SMRT-ML has the desirable property of not being misleading regardless

of the species tree topology or branch lengths.

Five taxa

Five-taxon trees are illustrated in Figures 1C–E. For these trees, SM-ML is misleading with certain branch lengths (Figures 5A–C and 6D). In contrast, SMRT-ML is not misleading under any parameters tested, attaining the correct tree 100% of the time with 6000 genes for all topologies and branch lengths tested.

Similarly to the results presented for four taxa, in cases where both SM-ML and SMRT-ML recover the species tree (given enough loci), the method that has faster convergence depends on the topology and branch lengths of the species tree. For the species tree (((AB)C)D)E, the only set of branch lengths tested for which SM-ML was not misleading was $(w, x, y) = (1.0, 0.1, 0.1)$, in which case SMRT-ML converged more quickly to the species tree than SM-ML. For these branch lengths, SMRT-ML recovered the species tree 94% of the time with 1000 loci, whereas SM-ML recovered the species tree 84% of the time. For the species tree (((AB)(CD))E), SM-ML showed slightly faster convergence to the species tree for two branch length combinations (Figures 7A, D). For example, with 1000 loci and branch lengths $(w, x, y) = (0.1, 0.1, 0.1)$, SM-ML and SMRT-ML recovered the species 93% and 91% of the time, respectively. However, for the same species tree topology with $(w, x, y) = (0.1, 0.1, 1.0)$, SMRT-ML appears to converge more quickly, with the species tree being estimated $\sim 89\%$ of the time with SMRT-ML using 1000 loci versus $\sim 60\%$ of the time with SM-ML. Furthermore, whereas SM-ML was never found to be misleading for four-taxon symmetric species trees, SM-ML can fail to converge to the species tree for every five-taxon tree shape. SMRT-ML converged to the species tree for all branch lengths tested on every five-taxon tree shape.

Model violations

To assess how SM-ML and SMRT-ML perform with violations of assumptions, we made gene trees non-clocklike by independently multiplying each branch by a value sampled from an exponential distribution with mean 1. The concatenated alignment generated by these gene trees was then analyzed assuming JC and a molecular clock.

Figure 8 shows that, for the (((AB)C)D) tree, both methods were fairly robust to violation of the molecular clock (when compared with Figure 3). The molecular clock violation slowed down the convergence to the species tree that was inferred with clocklike gene trees. For example, the species tree was inferred 98% of the time with 1000 genes under a molecular clock (Figure 3K), whereas it was inferred 80% of the time with 1000 genes and 96% of the time with 3000 genes when the molecular clock was violated (Figure 8K). This trend also held for the symmetric four-taxon species tree (Figure S1) and the three five-taxon species trees (Figures S2–

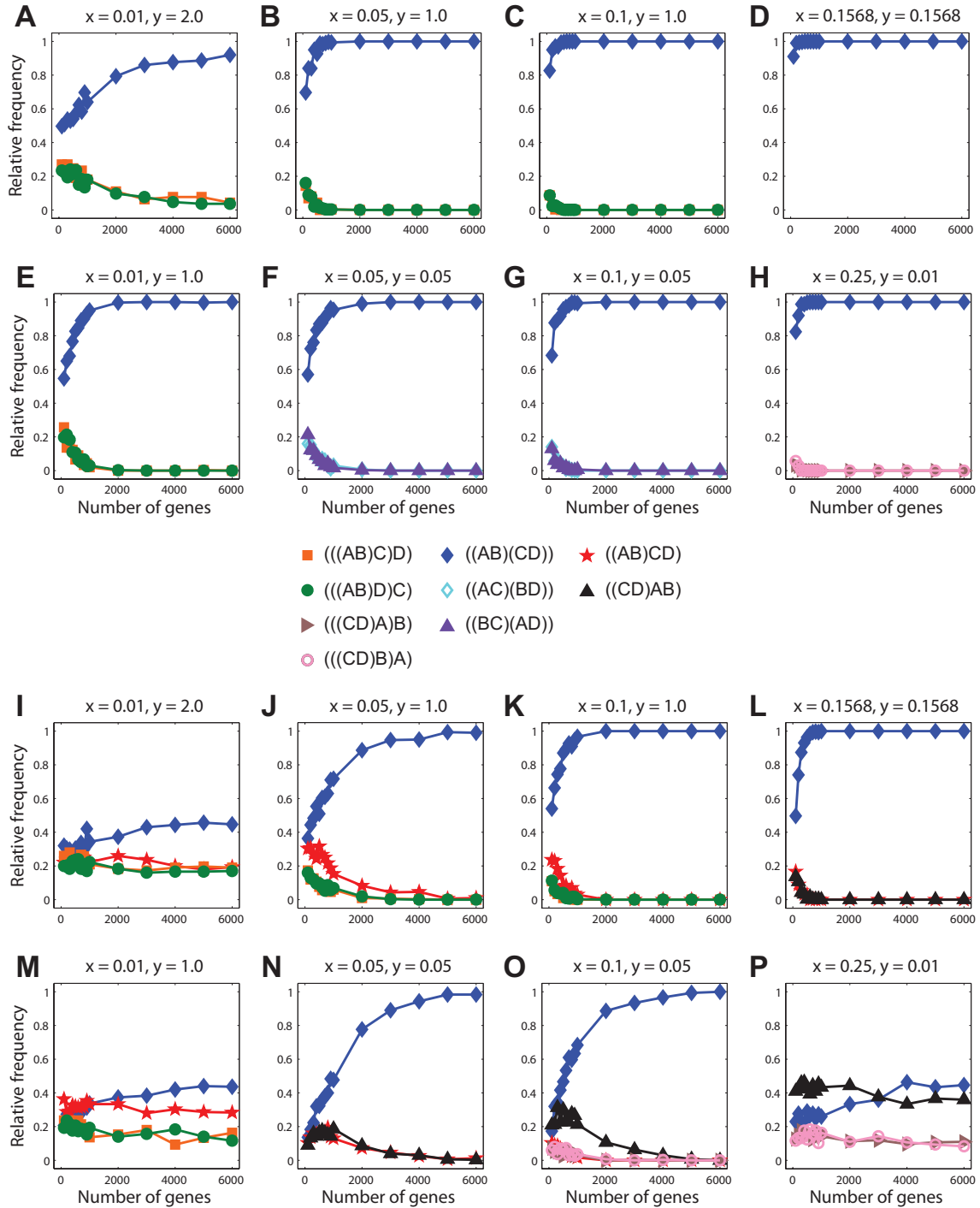


FIG. 4.—Results of simulations for the four-taxon tree ((AB)(CD)) (Figure 1B) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

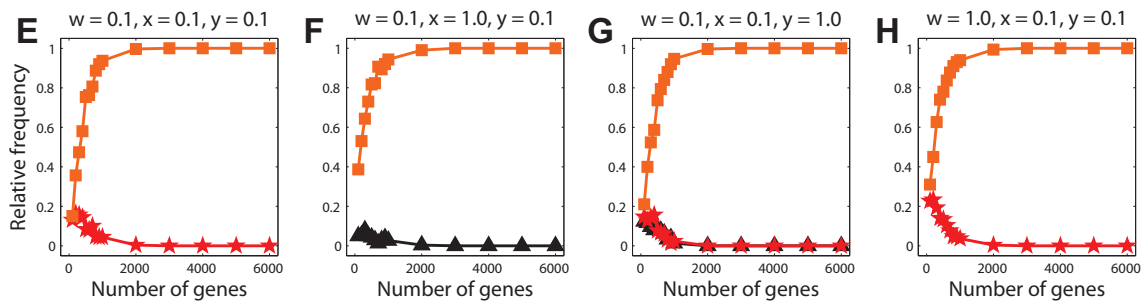
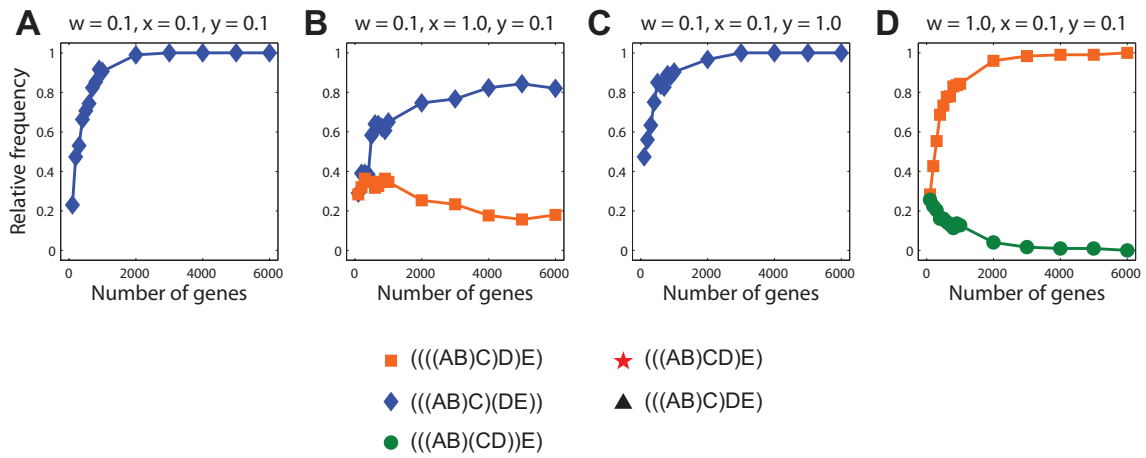


FIG. 5.—Results of simulations for the five-taxon tree $(((AB)C)D)E$ (Figure 1C) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-E) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations..

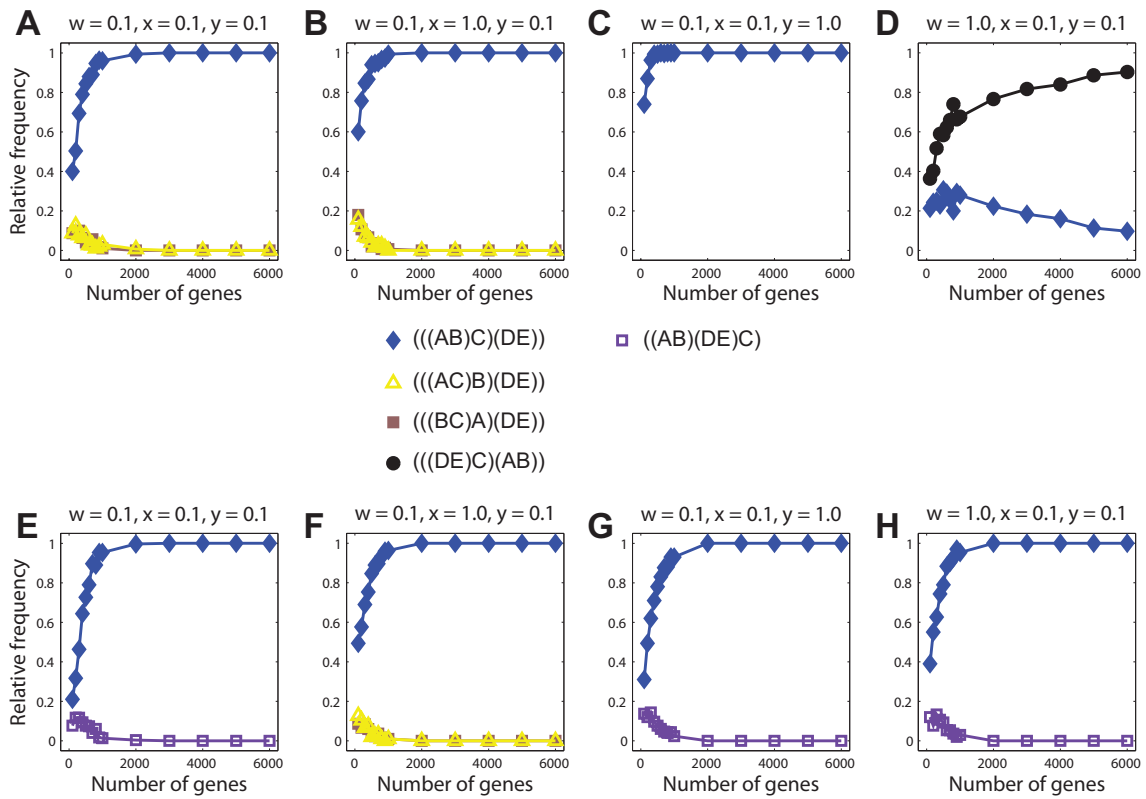


FIG. 6.—Results of simulations for the five-taxon tree $((AB)C)(DE)$ (Figure 1D) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-E) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

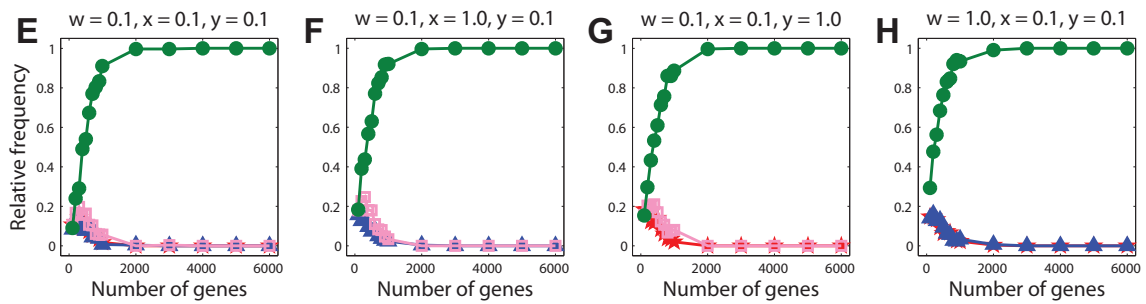
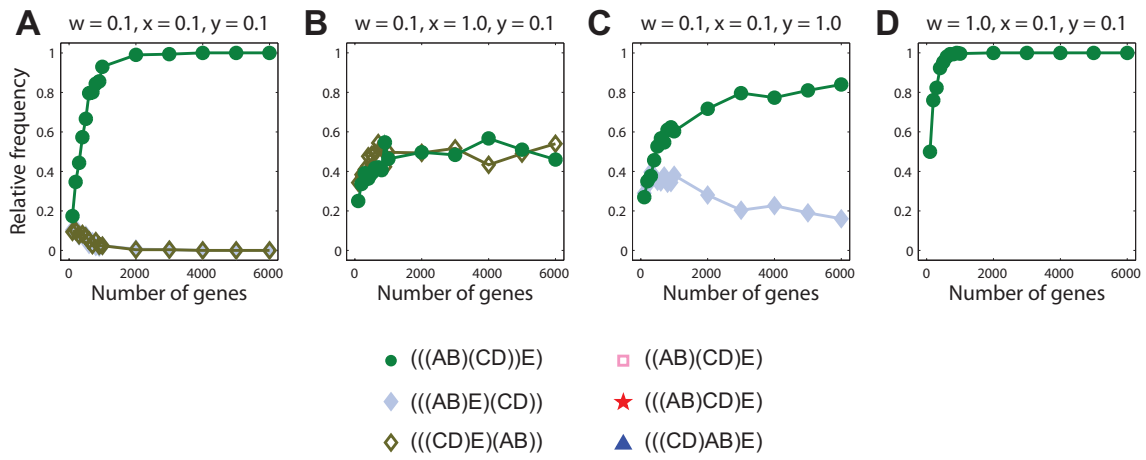


FIG. 7.—Results of simulations for the five-taxon tree $((AB)(CD))E$ (Figure 1E) generated under a Jukes-Cantor model with $\theta = 0.01$ and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-E) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

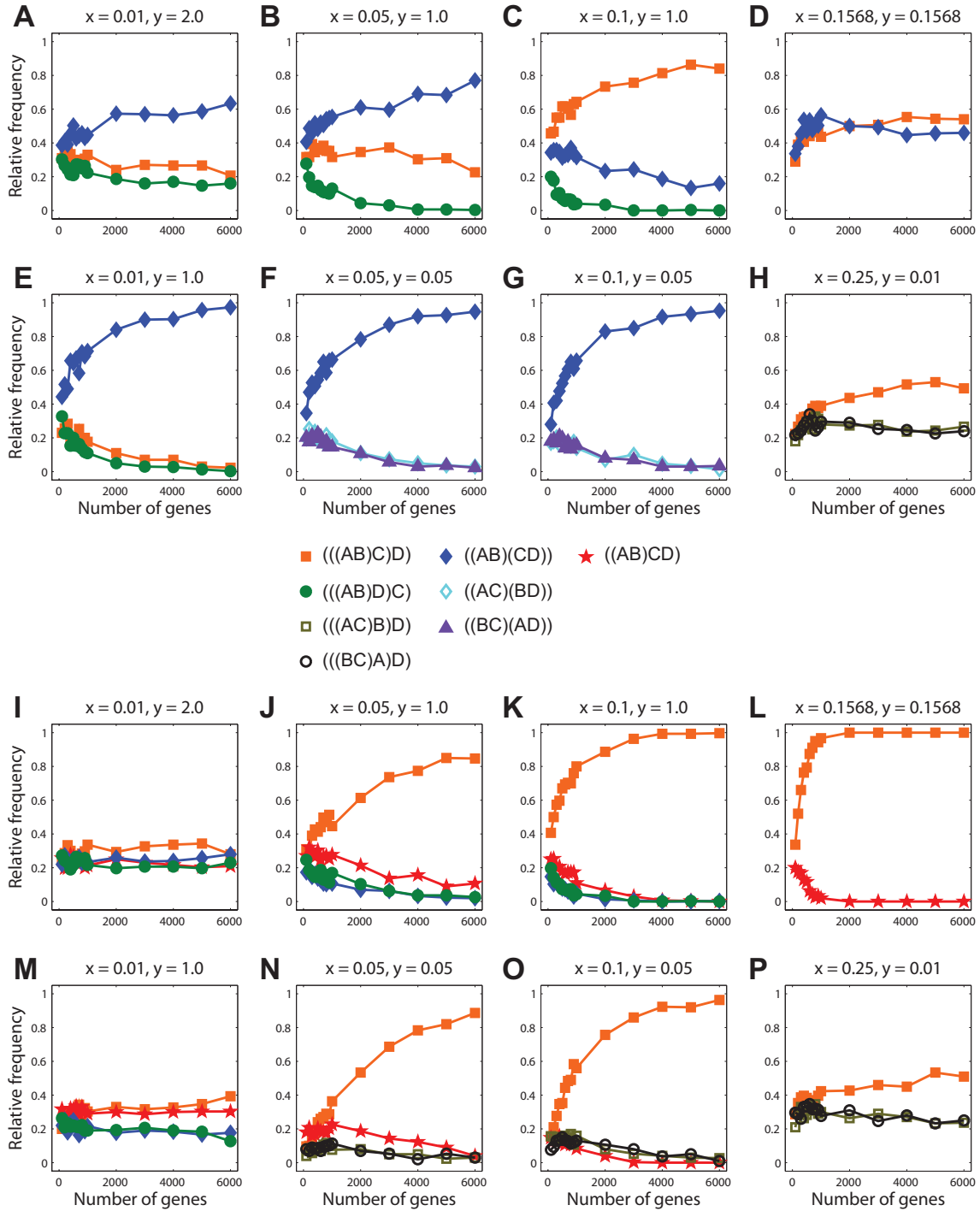


FIG. 8.—Results of simulations for the four-taxon tree $((AB)C)D$ (Figure 1A) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

4). Also, the violation of the molecular clock affected SMRT-ML more than SM-ML. For example, when the molecular clock is violated, it may require 2000 genes instead of 1000 genes to obtain the same fraction of correctly inferred trees (compare Figure 5E with Figure S2E, Figure 6E with Figure S3E, and Figure 7E with Figure S4E). From these results, we conclude that the performance of the two methods is only slightly influenced by the molecular clock violation.

We introduced a second model violation by generating sequence alignments under a complex substitution model (General Time-Reversible (GTR)) and then comparing SM-ML and SMRT-ML when trees were inferred assuming a simple substitution model (JC). As with the case of the molecular clock violation, the general patterns displayed by the two methods were not significantly altered (Figures S5–9). However, under this substitution model violation, SM-ML was more negatively affected than SMRT-ML. Based on simulations, SM-ML can converge more quickly to the wrong tree (compare Figures 3A,B with Figures S5A,B and Figure 7B with Figure S9B) and more slowly to the correct tree (compare Figure 3C with Figure S5C and Figure 7C with Figure S9C) compared to analysis under the correct model. Furthermore, this model violation can reverse the effect of adding more data. For example, when both branches of the four-taxon species tree (((AB)C)D) had lengths of 0.1568, SM-ML was increasingly likely to infer the correct tree when there was no model misspecification (Fig. 3D; 63% probability with 6000 genes), but decreasingly likely under model misspecification (Figure S5D; 26.7% chance of inferring the matching tree with 6000 genes). However, this model violation can also favorably influence SM-ML by causing a faster convergence to the correct tree (compare Figure 4A with Figure S6A and Figure 5B with Figure S7B).

In simulations, neither SM-ML nor SMRT-ML performed uniformly better than the other method for all possible species trees. Table 1 gives a summary of these results and notes whether each method recovered the species tree in more than 50% of simulations with 6000 loci of 500 nt each. A “NO” in the table indicates that either the method was likely to pick one of several trees (including the species tree) or converged to the wrong tree. Convergence to an incorrect tree only occurred for SM-ML. In cases where less than 50% probability of recovering the species tree was observed for SMRT-ML, SMRT-ML typically returned the species tree topology > 40% of the time and frequently returned some other tree, often a partially unresolved tree with no false positive clades. We note that a “NO” only occurred for SMRT-ML in the four-taxon cases where there was one extremely short branch length of 0.01 coalescent units, leading to a high probability of a partially unresolved tree. SM-ML had poorer performance as the number of taxa was increased even though branch lengths were less extreme than for most of the four-taxon simulations. SMRT-ML, however, had similar performance as the number of taxa increased.

Results for yeast data

Although the causes of gene tree conflict in the yeast dataset analyzed by Rokas et al. (2003) are unknown, the analysis of this dataset by several groups (e.g., Gatesy and Baker 2005; Edwards et al. 2007) makes it useful for comparing methods of inferring species trees. Rokas et al. (2003) reported that 20 concatenated genes were sufficient for maximum parsimony or ML to infer the same tree with high reliability. On the estimated species tree, the five taxa with the most difficult relationships to infer form the five-taxon subtree (((*S. cerevisiae*, *S. paradoxus*), *S. mikatae*), *S. kudriavzevii*), *S. bayanus*).

Using SMRT-ML on all 106 genes, we recovered the species tree found using SM-ML on the full data (i.e., the same tree that was reported as the estimated species tree in Rokas et al. (2003)). When a clock was assumed, SMRT-ML returned the species tree with the five-taxon subtree replaced by (((*S. cerevisiae*, *S. paradoxus*), *S. mikatae*), (*S. kudriavzevii*, *S. bayanus*)). The same result was produced by the program BEST (Liu 2008) analyzing the full data under a molecular clock; however, the molecular clock assumption is unreasonable because the data are not clocklike at most loci (Edwards et al. 2007).

To compare the efficiency of species tree estimation methods when methods agree on the full data, it is useful to consider subsets of the genes. For example, although Rokas et al. (2003) found that 20 randomly chosen genes were sufficient for SM-ML to estimate the species tree with high probability, Edwards et al. (2007) found that eight genes were sufficient using BEST. Because of the tradeoff between consistency and speed of convergence, we expect SMRT-ML to perform less efficiently than SM-ML for many cases when both methods have a high probability of returning the same tree, and this expectation is indeed what we found with the yeast data. The proportion of times SMRT-ML returned the species tree, inferred from all 106 loci, using random subsets of 20 loci was approximately 33%, with another 8% of cases returning a tree that was unresolved with respect to the taxa *S. kudriavzevii* and *S. bayanus* and the {*S. cerevisiae*, *S. paradoxus*, *S. mikatae*} clade. With 60 genes, the proportion of times that SMRT-ML returned the species tree increased to 59% (Figure S10). The SMRT-ML method was therefore increasingly likely to return the tree reported by Rokas et al. (2003) as the number of genes from this dataset was increased.

Using SMRT-ML on the full dataset of 106 genes, the bootstrap support for clades {*S. cerevisiae*, *S. paradoxus*} and {*S. cerevisiae*, *S. paradoxus*, *S. mikatae*} was 99% and 91%, respectively (as opposed to the 100% bootstrap support observed for the total concatenated dataset in Rokas et al. (2003)), while the clade {*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*} had 61% bootstrap support (Figure S11) (see Methods under the “Empirical example” section for how the bootstrap with SMRT-ML was performed). The clade {*S. bayanus*, *S. kudriavzevii*} occurred in 29% of bootstrap replicates. Thus, although SMRT-ML and SM-ML produced the same esti-

mated species tree for the 106-gene yeast dataset, SMRT-ML converged to this estimated tree more slowly than SM-ML. The speed of approach to this tree could be either a product of the tradeoff between consistency and speed of convergence sometimes observed for SMRT-ML, or misleadingly high bootstrap support for SM-ML (Gadagkar et al. 2005; Kubatko and Degnan 2007). The slower convergence of SMRT-ML compared to SM-ML observed for this dataset is not expected to generalize to all species trees since simulations found that there are also species trees for which SMRT-ML converges more quickly than SM-ML.

In both simulations and analysis of the yeast data, SMRT-ML was not misleading, in the sense of becoming increasingly less likely to infer an incorrect tree with more data, even in cases where SM-ML converged to the wrong tree. To see whether the observation that SMRT-ML was not misleading is expected to be true in general, we next assess the properties of SMRT-ML theoretically. We derive the probability that a site has pattern \mathbf{x} for a three-taxon species tree by averaging over gene genealogies under a simple substitution model. This result is then used to prove that SMRT-ML is statistically consistent when estimating species trees from coalescent mixtures of site patterns, at least in a simplified setting.

Theory

In this section, we begin by developing the probability distribution of site patterns under a Cavender-Farris-Neyman (CFN) substitution model given a clocklike three-taxon species tree. This substitution model assumes binary characters with equal rates of mutation between the characters. Assuming that incomplete lineage sorting is the source of discordance between gene trees and species trees and that the species tree has no hybridization or horizontal gene transfer events, we then show that the frequency of a certain site pattern in a concatenated alignment converges in probability to the probability of the site pattern (Lemma 1). From this result, we provide a proof that SM-ML is a consistent estimator of a clocklike three-taxon species tree (Lemma 2). Utilizing Lemma 2, we show in Theorem 3 that SMRT-ML is consistent for estimating clocklike species trees under the CFN model.

Consider a species tree with three taxa. Denote the true species tree by σ with speciation times ρ_0 and ρ_1 (see Figure 9). Denote the topology of the species tree as ((AB)C). The species tree is therefore written as

$$\sigma = ((A:\rho_1, B:\rho_1):\rho_0 - \rho_1, C:\rho_0),$$

which has clocklike branch lengths. Further, denote the topology of the gene tree that matches the species tree σ as τ_1 and denote the other gene tree topologies as the star tree $\tau_0 = (ABC)$ and the two discordant trees $\tau_2 = ((AC)B)$ and $\tau_3 = ((BC)A)$.

Random gene trees evolving along the species tree σ can take on any of the topologies τ_1 , τ_2 , or τ_3 . Define θ as

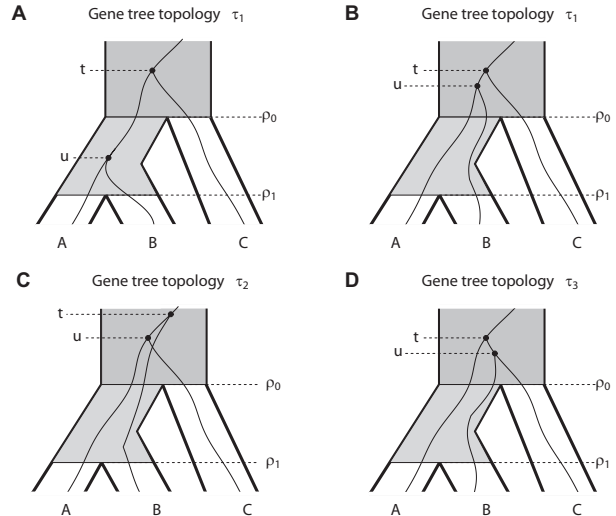


FIG. 9.—A three-taxon gene tree within a model species tree with notation used in the paper. In all cases, the species tree has the topology ((AB)C). Dots represent coalescent events. (A) and (B) depict the same gene tree topology with different coalescent histories. The gene tree in (C) has the ((AC)B) topology; the gene tree in (D) has the ((BC)A) topology.

the population mutation rate for each branch of the tree. For a random gene tree topology, we define t as the total length of the gene tree and u as the time from the present to the most recent coalescent event in mutation units.

Our goal is to determine the probability of a site pattern $\mathbf{x} = (x_1, x_2, x_3)$ under a CFN substitution model, where x_1 , x_2 , and x_3 are the characters at a site for species A, B, and C, respectively. If two species have the same character at a site, then they share the same letter. Therefore, the possible site patterns are xxx , xyx , xyx , and yxx . We note that only the xyx pattern supports the matching gene tree ((AB)C). We will show that when data are concatenated under a coalescent model and a ML tree is inferred from the concatenated data, the probability that the ML tree has topology τ_1 is higher than the probability of any other bifurcating tree topology. Further, we will show that this probability approaches 1 as the number of sites approaches infinity.

The probability of a site pattern given the species tree is obtained by conditioning on the gene genealogy and integrating over the joint density of the two coalescent times. The form of the joint density depends on whether both coalescent events occur more anciently than the root of the species tree σ or whether one coalescent event occurs more recently than the root of the species tree. This latter case only occurs when the gene tree matches the species tree. In this case (Figure 9A), define $g_\sigma(t, u, \tau_1)$ as the joint density for the coalescent times and gene tree topology τ_1 . Following Rannala and Yang (2003), this joint density is written as

$$g_\sigma(t, u, \tau_1) = \frac{4}{\theta^2} e^{\frac{2}{\theta}(\rho_0 + \rho_1 - t - u)}. \quad (1)$$

When all coalescent events occur more ancient than the root and the genealogy has topology τ_i (Figures 9B–D), the joint density of coalescent times and topology is

$$f_\sigma(t, u, \tau_i) = \frac{4}{\theta^2} e^{\frac{2}{\theta}(2\rho_0 + \rho_1 - t - 2u)}. \quad (2)$$

Note that when all coalescent events occur above the root, the form of the joint density for the gene tree topology and two coalescent times is the same for each of the three topologies. The probability of site pattern \mathbf{x} given that the species tree is σ is

$$P_\sigma(\mathbf{x}) = \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} P_\sigma(\mathbf{x}|t, u, \tau_1) g_\sigma(t, u, \tau_1) du dt + \sum_{i=1}^3 \int_{\rho_0}^{\infty} \int_{\rho_0}^t P_\sigma(\mathbf{x}|t, u, \tau_i) f_\sigma(t, u, \tau_i) du dt, \quad (3)$$

where $P_\sigma(\mathbf{x}|t, u, \tau_i)$ is the probability of the site pattern given the gene genealogy with topology τ_i and the branch lengths u and $t - u$. The first term in the probability is for the case that the gene tree matches the species tree and there is a coalescence between the A and B lineages more recent than the root of the species tree (Figure 9A). The second term is a summation corresponding to the three possible gene tree topologies when all coalescent events are more ancient than the root (Figures 9B–D).

For a CFN substitution model, Yang (2000) provided the probabilities of the site pattern \mathbf{x} conditional on the gene tree topology with branch length t and u as

$$\begin{aligned} P(\text{xxx}|t, u, \tau_i) &= \frac{1}{4} + \frac{1}{4}e^{-4u} + \frac{1}{2}e^{-4t} \\ P(\text{xxy}|t, u, \tau_1) &= \frac{1}{4} + \frac{1}{4}e^{-4u} - \frac{1}{2}e^{-4t} \\ P(\text{xyx}|t, u, \tau_1) &= P(\text{yxx}|t, u, \tau_1) = \frac{1}{4} - \frac{1}{4}e^{-4u}, \end{aligned} \quad (4)$$

where the equality for xyx and yxx follows by symmetry of A and B with respect to C in tree τ_1 . We have dropped the subscript σ in $P_\sigma(\cdot|\cdot)$ because the probability of a site pattern is independent of the species tree given the gene genealogy τ_i . Similarly,

$$\begin{aligned} P(\text{xxy}|t, u, \tau_3) &= P(\text{xxy}|t, u, \tau_2) = P(\text{xyx}|t, u, \tau_1) \\ P(\text{yxx}|t, u, \tau_3) &= P(\text{xyx}|t, u, \tau_2) = P(\text{xxy}|t, u, \tau_1) \\ P(\text{xyx}|t, u, \tau_3) &= P(\text{yxx}|t, u, \tau_2) = P(\text{yxx}|t, u, \tau_1). \end{aligned} \quad (5)$$

We next derive the full distribution of site patterns for a given species tree σ . Using the symmetries in equa-

tion (5),

$$\begin{aligned} P_\sigma(\text{xxy}) &= \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} P(\text{xxy}|t, u, \tau_1) g_\sigma(t, u, \tau_1) du dt \\ &\quad + \int_{\rho_0}^{\infty} \int_{\rho_0}^t \left\{ [P(\text{xxy}|t, u, \tau_1) + 2P(\text{xyx}|t, u, \tau_1)] \right. \\ &\quad \left. \times f_\sigma(t, u, \tau_i) \right\} du dt \\ &= \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} \frac{1 + e^{-4u} - 2e^{-4t}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(\rho_0 + \rho_1 - t - u)} du dt \\ &\quad + \int_{\rho_0}^{\infty} \int_{\rho_0}^t \frac{3 - e^{-4u} - 2e^{-4t}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(2\rho_0 + \rho_1 - 2u - t)} du dt \\ &= \frac{1 + 2\theta + e^{-4\rho_1} - 2e^{-4\rho_0}}{4 + 8\theta}. \end{aligned} \quad (6)$$

Analogously,

$$\begin{aligned} P_\sigma(\text{xyx}) &= \int_{\rho_0}^{\infty} \int_{\rho_1}^{\rho_0} \frac{1 - e^{-4u}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(\rho_0 + \rho_1 - t - u)} du dt \\ &\quad + \int_{\rho_0}^{\infty} \int_{\rho_0}^t \frac{3 - e^{-4u} - 2e^{-4t}}{4} \frac{4}{\theta^2} e^{\frac{2}{\theta}(2\rho_0 + \rho_1 - 2u - t)} du dt \\ &= \frac{1 + 2\theta - e^{-4\rho_1}}{4 + 8\theta}. \end{aligned} \quad (7)$$

By symmetry we have that $P_\sigma(\text{yxx}) = P_\sigma(\text{xyx})$ and by the law of total probability,

$$\begin{aligned} P_\sigma(\text{xxx}) &= 1 - P_\sigma(\text{xxy}) - P_\sigma(\text{xyx}) - P_\sigma(\text{yxx}) \\ &= \frac{1 + 2\theta + e^{-4\rho_1} + 2e^{-4\rho_0}}{4 + 8\theta}. \end{aligned} \quad (8)$$

The probability in equation (6) is greater than the probability in equation (7) if and only if $\rho_0 > \rho_1$, i.e., the root of the species tree is more ancient than the divergence of species A and B. Therefore, the probabilities of the segregating site patterns given the species tree σ are related by $P_\sigma(\text{xxy}) > P_\sigma(\text{yxx}) = P_\sigma(\text{xyx})$. Hence, the most probable segregating site pattern is the pattern that supports the species tree.

It is possible to extend the above derivation to other substitution models by modifying the expression in equation (4) and including a term for $P(\text{xyz})$ if there are more than two possible character states. Extending to site pattern probabilities for four or more taxa is also accomplished using the same approach but is considerably more tedious. For example, with four taxa, there are 15 rooted gene tree topologies rather than three, and the form of the joint density of coalescent times and gene tree topology depends on the *coalescent history*, a list of ancestral populations from the species tree where each coalescence occurs (Degnan and Salter 2005; Rosenberg 2007). For four-taxon trees, there are up to five coalescent histories for a given gene tree in a species tree, in contrast to the two expressions for three taxa (eqs. (1) and (2)). Thus, the probability of a site pattern \mathbf{x} is found by summing over

gene trees and computing triple integrals of $P(\mathbf{x})$ with respect to each of the algebraic expressions taken by the joint densities of coalescent times and gene tree topologies. Because SMRT-ML only uses alignments of three taxa, we have only derived three-taxon site pattern probabilities. We next provide two lemmas which aid in the proof of the theorem that SMRT-ML is consistent.

Lemma 1 essentially says that the alignment lengths do not matter asymptotically (under reasonable conditions), because the proportion of sites with any given pattern \mathbf{x} will approach the probability of the site pattern. In practice the length of the alignments could affect the rate at which a method using concatenated data (e.g., SM-ML or SMRT-ML) converges to a particular species tree. Lemma 2 says that because the most likely segregating pattern supports the species tree, SM-ML is consistent on concatenated three-taxon alignments under some assumptions (e.g., a clocklike species tree, constant ancestral θ s, and the CFN substitution model). Theorem 3 puts these ideas together and states that, because SMRT-ML constructs the species tree from several SM-ML estimates restricted to rooted triples, SMRT-ML is a statistically consistent estimator of clocklike species trees under the CFN model. The proofs for the theorem and the lemmas are provided in the Appendix.

We begin by stating assumptions used for proving the lemmas and theorem that follow:

1. Let the gene tree for the i th locus have topology $\tau^{(i)} \in \{\tau_1, \tau_2, \tau_3\}$ and coalescent times u_i and t_i (Figure 9), where the joint distribution of topology and coalescent times is given by equations (1) and (2). Assume that each site j in locus i is independent given the gene tree and coalescent times and has site pattern probability $P(\mathbf{x} | t_i, u_i, \tau^{(i)})$, given by equations (6)–(8), where the mutation parameter θ is constant for each ancestral population in the species tree. This derivation for site pattern probabilities depends on the following assumptions:

- Mutations occur under the CFN substitution model.
- The species tree is clocklike.
- Incomplete lineage sorting is the source of discordance between gene trees and species trees.
- There is no hybridization, horizontal gene transfer, or other gene flow between species.
- There is no population subdivision within species.

2. Consider a concatenated alignment of m non-recombining loci that are conditionally independent given the species tree, each with finite length $L_i \geq 1$ for $i = 1, 2, \dots, m$. Define $q_m = (\sum_{i=1}^m L_i^2) / (\sum_{i=1}^m L_i)^2$ and assume that, for any site pattern \mathbf{x} , $q_m \rightarrow 0$ as $m \rightarrow \infty$.

3. A supertree algorithm is used with the property that if the input trees are compatible, then the supertree is a rooted phylogenetic tree which displays all input trees.

The condition under assumption 2 that $q_m \rightarrow 0$ as $m \rightarrow \infty$ allows a version of the Law of Large Numbers (see Appendix) to be applied to site pattern probabilities for concatenated alignments with different lengths and ensures that the length of the concatenated alignment does not grow too rapidly. For example, if we concatenate loci of constant length L , then $q_m = mL^2 / (mL)^2 \rightarrow 0$ as $m \rightarrow \infty$. Similarly, if the gene length is bounded, so that $1 \leq L_i \leq B$, for some upper bound B , then $q_m \leq mB^2 / m^2 \rightarrow 0$. Since real genomes are finite, this assumption is reasonable for biological data. However, if every new locus were twice the length of the previous locus, say $L_i = 2^i$ for $i = 1, 2, \dots, m$, then $q_m \rightarrow 1/3$ as $m \rightarrow \infty$. Thus, if the concatenated alignment grows too quickly, Lemma 1 does not apply.

Assumption 3 states that the only characteristic of the supertree method that is necessary to prove Theorem 3 is that the method must return a tree which displays all input trees when they are compatible. Hence, if all rooted triples are inferred correctly, then the tree that displays those rooted triples is the species tree topology. A broad class of supertree algorithms can be used to prove this result including BUILD (Aho et al. 1981), matrix representation using parsimony (Baum 1992; Ragan 1992), mincut (Semple and Steel 2000), MMC (Page 2002), matrix representation using flipping (Chen et al. 2003), and normalized triplet supertrees (Willson 2009).

Lemma 1. Under assumptions 1 and 2, the proportion of sites with pattern \mathbf{x} converges in probability to $P(\mathbf{x})$.

Lemma 2. Under assumptions 1 and 2, SM-ML is a statistically consistent estimator of a three-taxon clocklike species tree.

Theorem 3. Under assumptions 1–3, SMRT-ML is a statistically consistent estimator of a clocklike species tree with three or more taxa.

Discussion

Overview of results and implications

In this study, we have shown that combining concatenation and supertree methods on rooted triples can overcome the problems caused by incomplete lineage sorting for concatenation-based ML inference of species trees. From theory, we find that SMRT-ML is a consistent estimator of species trees when sequences are generated under a CFN substitution model assuming a molecular clock and equal values of θ over the species tree.

Although neither SM-ML nor SMRT-ML performs uniformly better than the other, a scan of Table 1 shows that SMRT-ML often outperforms SM-ML when no single gene tree has high probability (typically $\leq 25\%$), and

when two gene trees have very similar probabilities. Because the yeast dataset has considerably less gene discordance than these cases, it is not surprising that SM-ML needs fewer loci than SMRT-ML to obtain the same species tree that was inferred from all 106 loci. The yeast data analysis also suggests that it may take a large number of genes for SMRT-ML to have a high probability of recovering the species tree, and therefore that SMRT-ML may have an advantage with sizeable genomic datasets. Simulations show that large amounts of data may also be necessary to resolve phylogenies when no single gene tree topology predominates.

Through simulations, we find that SMRT-ML is not misleading and often outperforms SM-ML given sufficiently severe gene tree discordance when sequences are generated under JC and GTR substitution models. This finding suggests that SMRT-ML is consistent when assuming models that are more complex than CFN. However, analytical results for three-taxon trees under more complex models are difficult to obtain. For example, Chor et al. (2006) found that the exact ML solution for a rooted three-taxon Jukes-Cantor problem required finding roots of an 11^{th} degree polynomial.

An attractive property of the SMRT method is computational efficiency. For each rooted triple, the tree space contains only three trees and the number of branch lengths needed to optimize is small. Therefore, the total number of trees examined is $3\binom{n}{3} = n(n-1)(n-2)/2$. In contrast, the total number of rooted tree topologies in an n -taxon tree space is $(2n-3)!!$. Although there are methods, such as Branch and Bound (Felsenstein 2004), that can ignore the irrelevant part of the tree space, finding globally optimal trees under criteria such as likelihood or parsimony is NP-hard (Day et al. 1986; Chor and Tuller 2005; Roch 2006). Because MMC is a polynomial-time algorithm (Page 2002), and only a polynomial number of trees is evaluated using SMRT, both steps of inferring triples and constructing the tree are polynomial in the number of taxa. Thus, at least under a simple substitution model, SMRT-ML is a polynomial-time algorithm for inferring the species tree and is statistically consistent when gene tree discordance is described by the multispecies coalescent model.

Taxon sampling for species tree inference

An issue that has received a lot of attention in phylogenetics is whether increased taxon sampling can improve the accuracy of species tree inference. Some researchers argue that increased taxon sampling generally improves phylogenetic inference (Zwickl and Hillis 2002; Hedtke et al. 2006), and others argue that it often does not (Poe and Swofford 1999; Rosenberg and Kumar 2001; Rokas and Carroll 2005). These studies have all focused on the effect of taxon sampling on the estimation of gene trees, prompting the need for investigating its effects on species tree estimation (Degnan and Rosenberg 2009).

Some of our results imply that the performance of SM-ML can either be improved or impaired when ex-

tra taxa are sampled, depending on the branch lengths and topology of the species tree. In general, SMRT-ML is less sensitive to taxon sampling than SM-ML for the range of species trees examined. As an example where SM-ML performs worse with more taxa, consider the species trees $((AB)(CD))$ with branch lengths $(x,y) = (0.1,1.0)$ (Figure 4C) and $((((AB)(CD))E))$ with branch lengths $(w,x,y) = (0.1,0.1,1.0)$ (Figure 7C). For the four-taxon species tree, SM-ML recovered the species tree topology $\sim 99\%$ of the time with 1000 loci. The addition of the E taxon with a short branch length separating the root of the tree from the most recent common ancestor of A, B, C, and D impaired the performance of SM-ML, making it incorrectly group E with (AB) 38% of the time with 1000 loci. In contrast, adding the E taxon to the same four-taxon tree had a much smaller influence on the performance of SMRT-ML (compare Figure 4K with Figure 7G).

To investigate this effect further, we added a sixth taxon separated from the root of the tree $((((AB)(CD))E))$ by 0.1 coalescent units to create the species tree $(((((AB)(CD))E)F))$ (Figure S12A). Adding the sixth taxon caused the probability that SM-ML inferred the AGT $((((AB)(CD))(EF))$ to approach 1 as more genes were added (Figure S12B). On the other hand, SMRT-ML had a similar performance with this six-taxon tree on taxa A–F as with the four- and five-taxon subtrees on taxa A–D and A–E, respectively.

For the five-taxon species tree $((((AB)C)D)E)$ with branch lengths $(w,x,y) = (1.0,0.1,0.1)$ (Figure 5D), SM-ML recovered the species tree 100% of the time given enough loci. However, when taxon E was removed from this species tree, SM-ML was misleading on the subtree $((((AB)C)D))$ with branch lengths $(x,y) = (0.1,0.1)$ (Figure S13), with a probability approaching 1 of returning the AGT $((AB)(CD))$. SMRT-ML was not as influenced by the presence of taxon E for this example, though the extra taxon slightly hindered the speed of convergence to the species tree. This example shows not only that increased taxon sampling had a less dramatic influence on SMRT-ML than SM-ML, but also that the same parameters can produce opposite effects that aid one method while hurting the other.

Rooted triple consensus

A recent study used rooted triples estimated at each locus as input to the quartet puzzling algorithm (Ewing et al. 2008) by treating a fourth taxon as a known outgroup. In quartet puzzling, maximum likelihood trees for all $\binom{n}{4}$ quartets of a set of n species are estimated and a heuristic algorithm is used to construct the tree from the inferred quartets (Strimmer and von Haeseler 1996). The R^* consensus method (Bryant 2003; Degnan et al. 2009) is similar in that it uses rooted triples at each locus, although these are generated by first inferring gene trees on the full set of taxa. R^* consensus then applies a different non-heuristic algorithm from that of the quartet puzzling based rooted triple consensus to construct the

tree from the estimated rooted triples. Like R^* consensus, rooted triple consensus construct the estimated species tree from $m \binom{n}{3}$ rooted triples, where m is the number of loci. Neither method requires the estimation of coalescent or population parameters, and each avoids the problem of AGTs due to incomplete lineage sorting through the use of rooted triples. R^* consensus given known gene trees at each locus is proven to be statistically consistent when gene tree discordance is due to incomplete lineage sorting (Degnan et al. 2009). A more general approach shows that supertree methods that have rooted triples as input can be statistically consistent in this setting under certain assumptions (Steel and Rodrigo 2008, see Proposition 5). SMRT is different from rooted triple consensus and supertree methods in that only the $\binom{n}{3}$ rooted triples from a supermatrix are inferred. SMRT also differs from rooted triple consensus in that rooted triples are input into a supertree algorithm to construct the estimated species tree.

One advantage of rooted triple consensus and R^* over SMRT is that they use the information of all available taxa at a given locus to infer a gene tree whereas SMRT only uses information on three taxa. Because there may be a lack of phylogenetic signal among the three taxa analyzed by SMRT, the extra information about the relationship between taxa used by rooted triple consensus and R^* can aid in more accurate estimates of species tree when the total amount of sequence is small. However, SMRT has the advantage that it is both fast and tractable on a large number of taxa. Because ML inference of phylogenetic trees is NP-hard (Chor and Tuller 2005; Roch 2006), if gene trees are inferred using ML at each locus, then both rooted triple consensus and R^* are NP-hard whereas SMRT is polynomial in the number of taxa.

Bayesian approaches

Recent methods, such as BEST (Liu and Pearl 2007; Liu 2008) and BUCKY (Ané et al. 2007), for inferring species trees from multilocus data take a Bayesian approach. The program BEST simultaneously estimates a joint posterior distribution of gene trees and species trees (Rannala and Yang 2008) assuming that gene trees are distributed according to the coalescent process and that gene tree discordance is due solely to incomplete lineage sorting. In contrast to BEST, which models discordance among of gene trees using the coalescent process, BUCKY uses a prior to model the correlation between gene trees without assuming the source (e.g., incomplete lineage sorting) of discordance. These methods are attractive in that they are designed to handle gene-tree discordance. However, both are computationally intensive, relying on MCMC runs for separate loci and for estimates of the species tree and are therefore tractable only for small numbers of taxa and loci (Edwards 2009). Because SMRT is polynomial in the number of taxa and not heavily affected by the number of loci, it is especially well-suited for genomic-level data and large numbers of taxa.

Other sources of discordance

SMRT gains its strength from the fact that when gene trees are distributed according to the multispecies coalescent, there are no anomalous three-taxon trees when the source of gene-tree discordance is due only to incomplete lineage sorting. However, in the presence of other sources of discordance, such as hybridization, horizontal gene transfer, gene duplication, recombination, and population structure the most probable three-taxon gene tree might not match the species tree (Slatkin and Pollack 2008). Hence, if there are forces acting strongly to create gene-tree discordance other than incomplete lineage sorting, then SMRT may not have enough information to obtain the correct tree. However, because SMRT has the ability to infer partially unresolved trees, then it may be the case that forces such as horizontal gene transfer will cause SMRT to infer a partially unresolved tree. Future studies are needed to assess how SMRT and other methods perform under various types and degrees of gene tree discordance.

Summary

When genetic data from multiple loci are concatenated, the distribution of site patterns is a mixture that depends on the distribution of gene trees over the loci. Such mixture distributions on site patterns make it difficult to obtain analytical results for concatenated data and therefore to understand theoretical properties of phylogenetic methods that use concatenated data. We have obtained the distribution of site patterns for three-taxon concatenated sequences under a mixture distribution due to the multispecies coalescent using the CFN substitution model. Thus, despite the poor performance of SM-ML for some species trees, there is enough information in the concatenated alignment, and therefore in the distribution of site patterns, to recover the species tree topology. SMRT-ML uses this information in the concatenated alignment to consistently recover the species tree.

The consistency of SMRT-ML shows that the species tree topology is identifiable from concatenated data in the sense that two distinct species trees (with either different topologies or the same topology but different branch lengths) cannot have the same distribution of site patterns. The analytic framework in this paper could be extended to either more complex substitution models or to larger numbers of taxa to yield further insights into some of the properties of concatenated data.

As a tool for inferring species trees, SMRT-ML could be extended to cases where there are multiple individuals sampled per species. Here, there could be multiple inferred triples for each choice of three species, where one individual from within each of the three species is chosen randomly, or all possible combinations with one individual per species are used. If there are n species and i individuals sampled per species, this procedure would result in $i^3 \binom{n}{3}$ inferred rooted triples from which the species tree could be constructed using a supertree method such

as MMC. With multiple rooted triples estimated on the same choice of three taxa, a supertree algorithm designed for high levels of conflict in the input triples might be useful, for example, Normalized Triplet Supertree (Willson 2009).

We have not investigated the performance of SMRT when combined with methods of inferring gene trees other than ML, such as parsimony and distance methods. Liu and Edwards (2009) show that for concatenated data, under similar assumptions as in this paper, distance methods and in many cases parsimony methods recover the species tree when SM-ML is misleading. Although, because of long branch attraction (Felsenstein 1978), maximum parsimony is not consistent for trees with five or more taxa, even when there is a molecular clock (Hendy and Penny 1989). However, for cases in which rooted three-taxon gene trees can be inferred consistently from concatenated data—including distance and parsimony methods under a molecular clock—SMRT is also consistent for larger trees because of the fact that rooted triples identify a tree, independently of how those rooted triples were inferred. Future studies using simulation and real data will be needed to further assess the performance of SMRT methods and its extensions.

Acknowledgments

We thank Noah A. Rosenberg, Raquel Assis, Elizabeth S. Allman, and Liang Liu for their valuable comments. This work was supported by NSF grant DEB-0716904 and NIH training grant T32 GM070449.

Literature Cited

- Aho AV, Sagiv Y, Szymanski TG, Ullman JD. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.* 10:405–421.
- Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance factors. *Mol. Biol. Evol.* 24:412–426.
- Baum BR. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon.* 41:3–10.
- Bininda-Emonds ORP. 2004. The evolution of supertrees. *Trends Ecol. Evol.* 19:315–322.
- Bryant D. 2003. A classification of consensus methods for phylogenies. In: Janowitz M, Lapointe FJ, McMorris FR, Mirkin B, Roberts FS, editors. *BioConsensus*, pp. 163–183. DIMACS. AMS.
- Buckley TR, Cordeiro M, Marshall DC, Simon C. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada dugdale*). *Syst. Biol.* 55:411–425.
- Chen D, Diao L, Eulenstein O, Fernández-Baca D, Sanderson M. 2003. Flipping: a supertree construction method. Pages 135–160 in *Bioconsensus* (Vol. 61) (M. F. Janowitz et al., eds.). American Mathematical Society.
- Chor B, Hendy M, Penny D. 2007. Analytic solutions for three taxon ML trees with variable rates across sites. *Discrete Appl. Math.* 155:750–758.
- Chor B, Hendy M, Snir S. 2006. Maximum likelihood Jukes-Cantor triplets: analytic solutions. *Mol. Biol. Evol.* 23:626–632.
- Chor B, Tuller T. 2005. Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics.* 21:i97–i106.
- Chung KL. 1974. *A course in probability theory* (2nd edn). San Diego, CA: Academic Press.
- Cormen TH, Leiserson CE, Rivest RL, Stein C. 2001. *Introduction to Algorithms* (2nd edn). Cambridge, MA: The MIT Press.
- Day W, Johnson D, Sankoff D. 1986. The computational complexity of inferring rooted phylogenies by parsimony. *Math. Biosci.* 81:33–42.
- de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol. Evol.* 22:34–31.
- Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. 2009. Properties of consensus methods for estimating species trees from gene trees. *Syst. Biol.* 58:35–54.
- Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2:e68.
- Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24:332–340.
- Degnan JH, Salter LA. 2005. Gene tree distributions under the coalescent process. *Evolution.* 59:24–37.
- Edwards SV. 2009. Is a new and general theory of systematics emerging? *Evolution.* 63:1–19.
- Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. *Proc. Natl. Acad. Sci. USA.* 104:5936–5941.
- Ewing GB, Ebersberger I, Schmidt HA, von Haeseler A. 2008. Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol.* 8:118.
- Felsenstein J. 1978. The number of evolutionary trees. *Syst. Zool.* 27:27–33.
- Felsenstein J. 2004. *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.
- Gadagkar SR, Rosenberg M, Kumar S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *J. Exp. Zool.* 304B:64–74.
- Gatesy J, Baker RH. 2005. Hidden likelihood support in genomic data: Can forty-five wrongs make a right? *Syst. Biol.* 54:483–492.
- Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst. Biol.* 55:522–529.
- Hendy MD, Penny D. 1989. A framework for the study of evolutionary trees. *Syst. Zool.* 38:297–309.
- Holland BR, Benthin S, Lockhart PJ, Moulton V, Huber KT. 2008. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evol. Biol.* 8:202.
- Jennings WB, Edwards SV. 2005. Speciation history of Australian grassfinches (*Poephila*) inferred from thirty gene trees. *Evolution.* 59:2033–2047.
- Kolaczowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature.* 431:980–984.
- Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56:17–24.
- Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics.* 24:2542–2543.
- Liu L, Edwards SV. 2009. Phylogenetic inference in the anomaly zone. *Syst. Biol.* in press:doi:10.1093/sysbio/syp034.
- Liu L, Pearl DK. 2007. Species trees from gene trees: recon-

- structing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56:504–514.
- Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–536.
- Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor. Pop. Biol.* 75:35–45.
- Mossel E, Vigoda E. 2005. Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science.* 309:2207–2209.
- Nei M. 1987. *Molecular Evolutionary Genetics*. NY: Columbia University Press.
- Neyman J. 1971. Molecular studies in evolution: a source of novel statistical problems. In: Gupta SS, Yackel J, editors. *Statistical Decision Theory and Related Topics*, pp. 1–27. NY: Academic Press.
- Page RDM. 2002. Modified mincut supertrees. pages 537–552 in lecture notes in computer science. In: Guigó R, Gusfield D, editors. *Algorithms in Bioinformatics, Second International Workshop, WABI, 2002, Rome, Italy, September 17–21, 2002, Proceedings (Vol. 2452)*, pp. 537–552. Berlin: Springer.
- Page RDM, Charleston MA. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.* 7:231–240.
- Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–583.
- Poe S, Swofford DL. 1999. Taxon sampling revisited. *Nature.* 398:299–300.
- R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.* 1:53–58.
- Rambaut A, Grassly NC. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosc.* 13:235–238.
- Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics.* 164:1645–1656.
- Rannala B, Yang Z. 2008. Phlogenetic inference using whole genomes. *Annu. Rev. Genom. Human Genet.* 9:217–231.
- Roch S. 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE ACM T. Comput. BI.* 3:92–94.
- Rokas A, Carroll SB. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol. Biol. Evol.* 22:1337–1344.
- Rokas A, Williams B, King N, Carroll S. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature.* 425:798–804.
- Rosenberg MS, Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. USA.* 98:10751–10756.
- Rosenberg NA. 2007. Counting coalescent histories. *J. Comp. Biol.* 14:360–377.
- Rosenberg NA, Tao R. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst. Biol.* 57:131–140.
- Semple C, Steel M. 2000. A supertree method for rooted trees. *Discrete Appl. Math.* 105:147–158.
- Slatkin M, Pollack JL. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. *Mol. Biol. Evol.* 25:2241–2246.
- Steel M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *J. Classification.* 9:91–116.
- Steel M, Rodrigo A. 2008. Maximum likelihood supertrees. *Syst. Biol.* 57:243–250.
- Strimmer K, von Haeseler A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13:964–969.
- Swofford DL. 2003. *PAUP*. Phylogenetic analysis using parsimony (* and other methods)*. Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Than C, Ruths D, Innan H, Nakhleh L. 2007. Confounding factors in HGT detection: Statistical error, coalescent effects, and multiple solutions. *J. Comput. Biol.* 14:517–535.
- Willson SJ. 2009. Robustness of topological supertree methods for reconciling dense incompatible data. *IEEE-ACM Trans. Comput. Biol. Bioinf.* 6:62–75.
- Yang Z. 2000. Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. Lond. B.* 267:109–116.
- Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51:588–598.

Appendix

Lemma 4 is a version of the Weak Law of Large Numbers that does not require identically distributed random variables. This lemma is used to prove Lemma 1.

Lemma 4 (Modified Theorem 5.2.3 of Chung (1974)). Consider the sequence X_1, X_2, \dots, X_n , where $X_i > 0$, of independent random variables each with their own distribution function. Define $S_n = \sum_{i=1}^n X_i$. Further, let $\{b_n\}$ be a sequence that approaches infinity and assume that $X_i \leq b_n$ for each $i = 1, 2, \dots, n$. If $\lim_{n \rightarrow \infty} \sum_{i=1}^n E[X_i^2]/b_n^2 = 0$, then $S_n/b_n \xrightarrow{P} E[S_n/b_n]$.

Proof of Lemma 1. First we show that the expected proportion of sites with a given site pattern \mathbf{x} is equal to the probability of that site pattern, $P(\mathbf{x})$. We consider the expected proportion of sites with each pattern and note that by Lemma 4, as the number of loci approaches infinity, the probability approaches 1 that the proportion of sites with a given pattern approaches the expected proportion. Let m denote the number of loci and let L_i denote the number of sites at locus i . The total number of sites is $\sum_{i=1}^m L_i$. For a site pattern \mathbf{x} , let $\delta_{\mathbf{x},i,j} = 1$ if site j in locus i has site pattern \mathbf{x} ; otherwise $\delta_{\mathbf{x},i,j} = 0$. Let $M_{\mathbf{x},i} = \sum_{j=1}^{L_i} \delta_{\mathbf{x},i,j}$ denote the number of sites in locus i that have site pattern \mathbf{x} . Let $S_m = \sum_{i=1}^m M_{\mathbf{x},i}$ and let $b_m = \sum_{i=1}^m L_i$. Because the length of the concatenated alignment is increasing with each additional locus, we have that $b_m \rightarrow \infty$ as $m \rightarrow \infty$. Note that $E[\delta_{\mathbf{x},i,j}^2] = P(\mathbf{x})$. Also note that $E[\delta_{\mathbf{x},i,j} \delta_{\mathbf{x},i,k}] = P(\mathbf{x}, \mathbf{x})$ for $j \neq k$ where $P(\mathbf{x}, \mathbf{x})$ is the probability of getting pattern \mathbf{x} at two different sites. Note that we do not need to know the actual value of $P(\mathbf{x}, \mathbf{x})$ —only that it is between 0 and 1. Then it follows that

$$\begin{aligned}
 & \frac{1}{b_m^2} \sum_{i=1}^m E[M_{\mathbf{x},i}^2] \\
 &= \frac{1}{b_m^2} \sum_{i=1}^m \left(\sum_{j=1}^{L_i} E[\delta_{\mathbf{x},i,j}^2] + 2 \sum_{j=1}^{L_i-1} \sum_{k=j+1}^{L_i} E[\delta_{\mathbf{x},i,j} \delta_{\mathbf{x},i,k}] \right) \\
 &= \frac{1}{b_m^2} \sum_{i=1}^m [P(\mathbf{x})L_i + P(\mathbf{x}, \mathbf{x})(L_i^2 - L_i)] \\
 &= \frac{P(\mathbf{x}) - P(\mathbf{x}, \mathbf{x})}{b_m} + P(\mathbf{x}, \mathbf{x}) \frac{\sum_{i=1}^m L_i^2}{b_m^2}. \tag{9}
 \end{aligned}$$

The quantity in equation (9) approaches 0 as $m \rightarrow \infty$ only if $q_m = \sum_{i=1}^m L_i^2 / b_m^2 \rightarrow 0$ as $m \rightarrow \infty$. We assumed that $q_m \rightarrow 0$ as $m \rightarrow \infty$. Thus, Lemma 4 applies and therefore,

$$\begin{aligned} \frac{\sum_{i=1}^m M_{\mathbf{x},i}}{\sum_{i=1}^m L_i} &= \frac{S_m}{b_m} \xrightarrow{P} E \left[\frac{\sum_{i=1}^m M_{\mathbf{x},i}}{b_m} \right] \\ &= \frac{\sum_{i=1}^m \sum_{j=1}^{L_i} P(\mathbf{x})}{\sum_{i=1}^m L_i} \\ &= \frac{\sum_{i=1}^m L_i}{\sum_{i=1}^m L_i} P(\mathbf{x}) = P(\mathbf{x}). \quad \square \end{aligned}$$

Proof of Lemma 2. Let m denote the number of loci. The total number of sites is $\sum_{i=1}^m L_i$. Let $M_{\mathbf{x},i}$ denote the number of sites in locus i that have site pattern \mathbf{x} . Further, let $M_{\mathbf{x}} = \sum_{i=1}^m M_{\mathbf{x},i}$ be the number of sites with pattern \mathbf{x} in the concatenated alignment. Suppose three species, A, B, and C, have the species tree $\sigma = ((A:\rho_1, B:\rho_1):\rho_0 - \rho_1, C:\rho_0)$, where ρ_0 and ρ_1 are measured in coalescent units, and the two ancestral populations each have the same θ . Further, suppose there are M_{xxx} , M_{xxy} , M_{xyx} , and M_{yxx} sites with site patterns xxx , xxy , xyx , and yxx , respectively. By Lemma 1, we know that the relative frequency of pattern xxy ($M_{xxy} / \sum_{i=1}^m L_i$) converges in probability to $P(xxy)$. Because xxy is the most likely segregating site pattern (eqs. (6)–(8)), it follows that the probability that xxy is the most frequently occurring segregating site pattern (i.e., $M_{xxy} > M_{xyx}, M_{yxx}$) approaches 1 as $m \rightarrow \infty$. Theorem 3 of Chor et al. (2007) states that if $M_{xxy} > M_{xyx}, M_{yxx}$, then ((AB)C) is the inferred ML under a molecular clock. Utilizing this theorem, the probability that the

ML tree topology is ((AB)C) approaches 1 as $m \rightarrow \infty$. \square

Proof of Theorem 3. Suppose we have an n -taxon species tree. There are $\binom{n}{3}$ subsets of three taxa. Let the rooted triples on the species tree be enumerated $\sigma_1, \sigma_2, \dots, \sigma_J$, where $J = \binom{n}{3}$. Let σ_j^* denote a rooted triple defined on the same taxa as σ_j but which is not a rooted triple on the species tree. From equations (6) and (7) and from equation (5), if \mathbf{x} is the most probable segregating site pattern for σ_j , then $P_{\sigma_j}(\mathbf{x}) > P_{\sigma_j^*}(\mathbf{x})$. Let the most frequently occurring segregating site pattern for supermatrix rooted triple j be \mathbf{x} . Applying Lemma 1, for any $\varepsilon > 0$, we can choose the number of loci m such that the probability of $P_{\sigma_j}(\mathbf{x}) > P_{\sigma_j^*}(\mathbf{x})$ is greater than $1 - \varepsilon / \binom{n}{3}$. By Lemma 2, the ML estimate for each of these J sets of three taxa is σ_j , $j = 1, 2, \dots, J$. Therefore the probability that all J rooted triples in the species tree are inferred by SMRT-ML is greater than $1 - \varepsilon$. In this case, the rooted triples will be compatible and the tree with the same topology as the species tree is uniquely identified by these J triples by Proposition 4 of Steel (1992). Applying a supertree algorithm to these J rooted triples with the property that if the input trees are compatible, then the supertree method returns a tree that displays all of its input trees, the supertree algorithm is guaranteed to return the matching species tree with probability greater than $1 - \varepsilon$. Thus, the supertree method applied to the J supermatrix rooted triples returns the species tree with probability greater than $1 - \varepsilon$. Therefore, SMRT-ML is statistically consistent under the CFN substitution model when the species tree is clocklike. \square

**Supplementary material: Fast species trees using supermatrix
and consistent estimation of rooted triples**

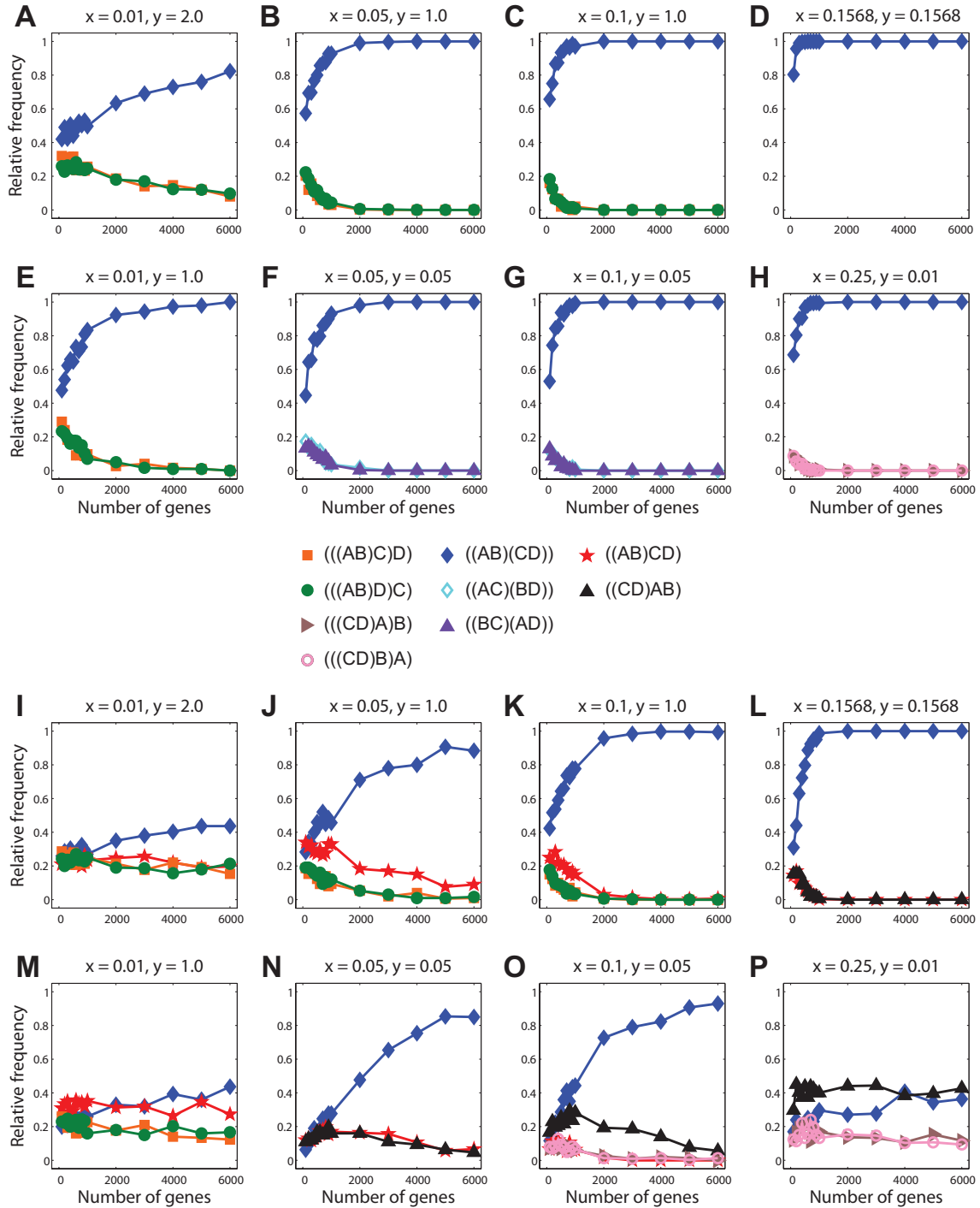


FIG. S1.—Results of simulations for the four-taxon tree $((AB)(CD))$ (Figure 1B) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

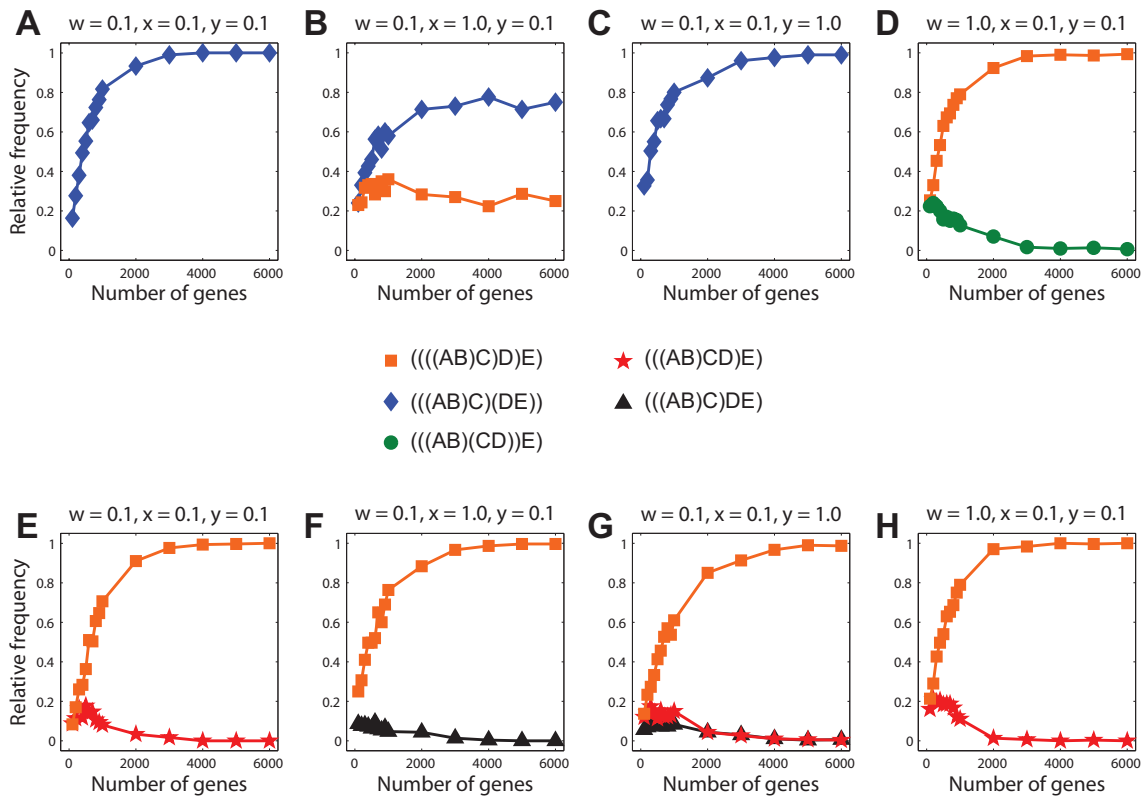


FIG. S2.—Results of simulations for the five-taxon tree $(((AB)C)D)E$ (Figure 1C) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

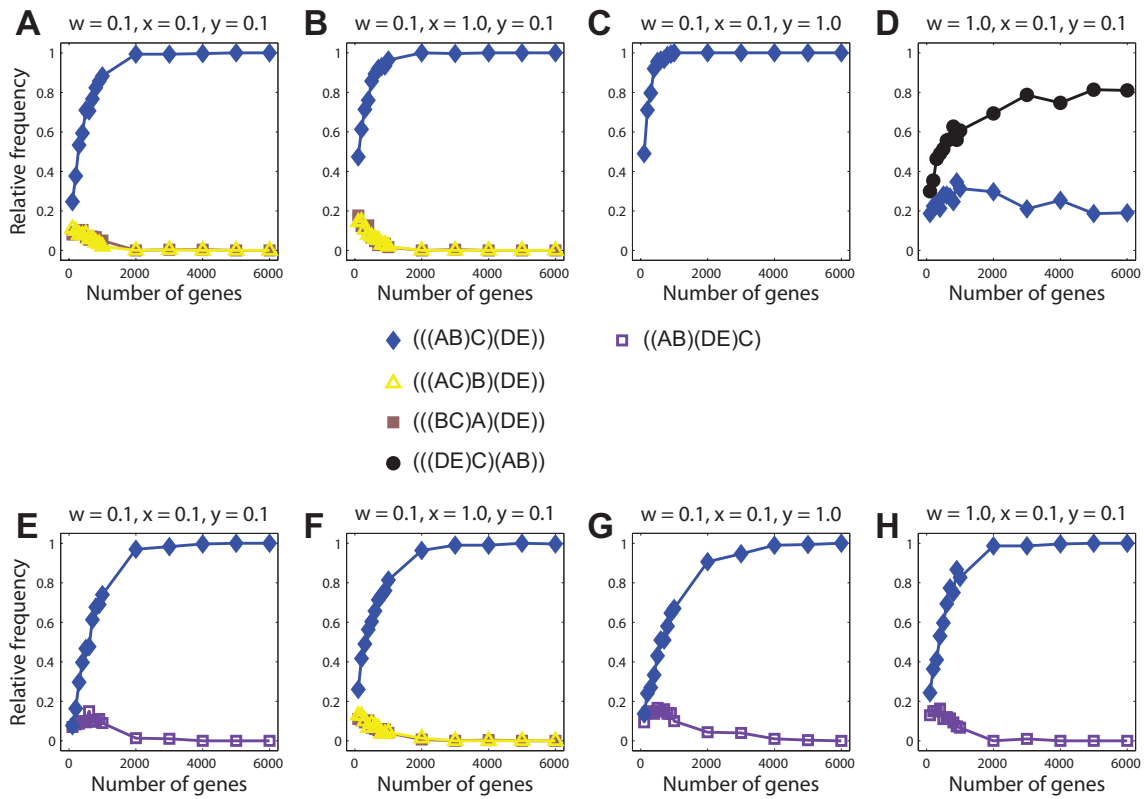


FIG. S3.—Results of simulations for the five-taxon tree $((AB)C)(DE)$ (Figure 1D) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

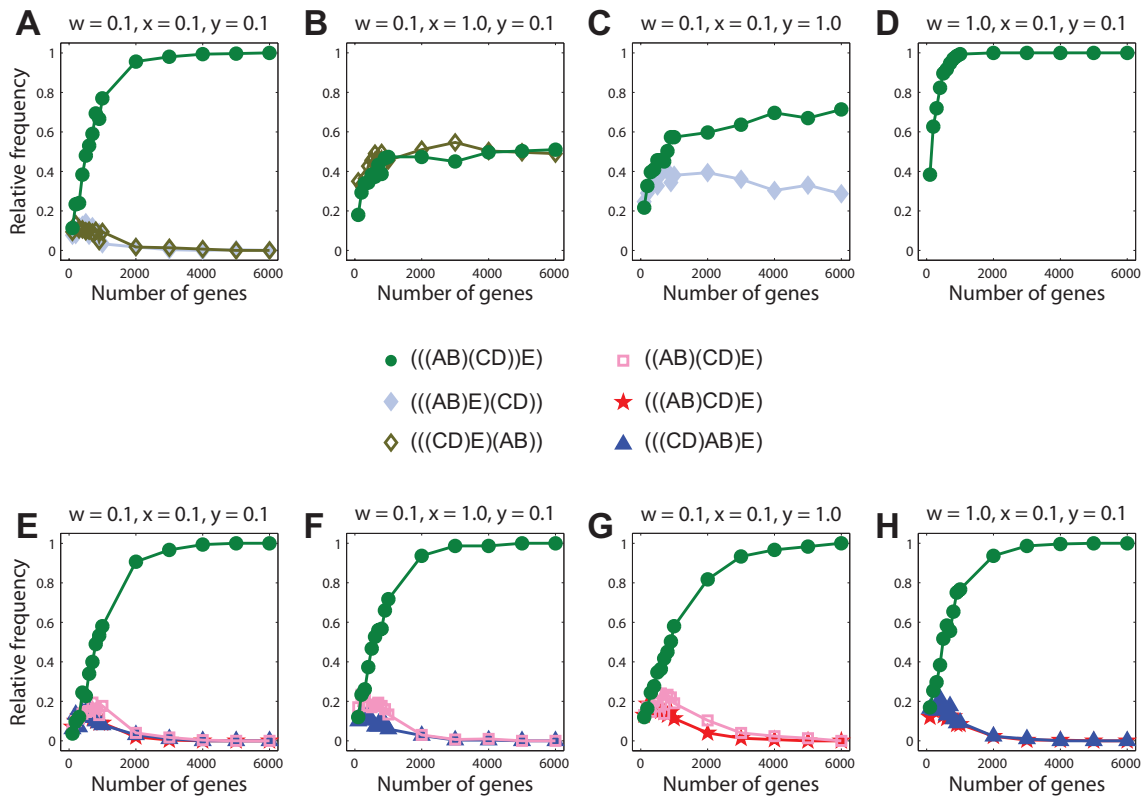


FIG. S4.—Results of simulations for the five-taxon $((AB)(CD))E$ (Figure 1E) generated under a Jukes-Cantor model with $\theta = 0.01$ and a violation of the molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

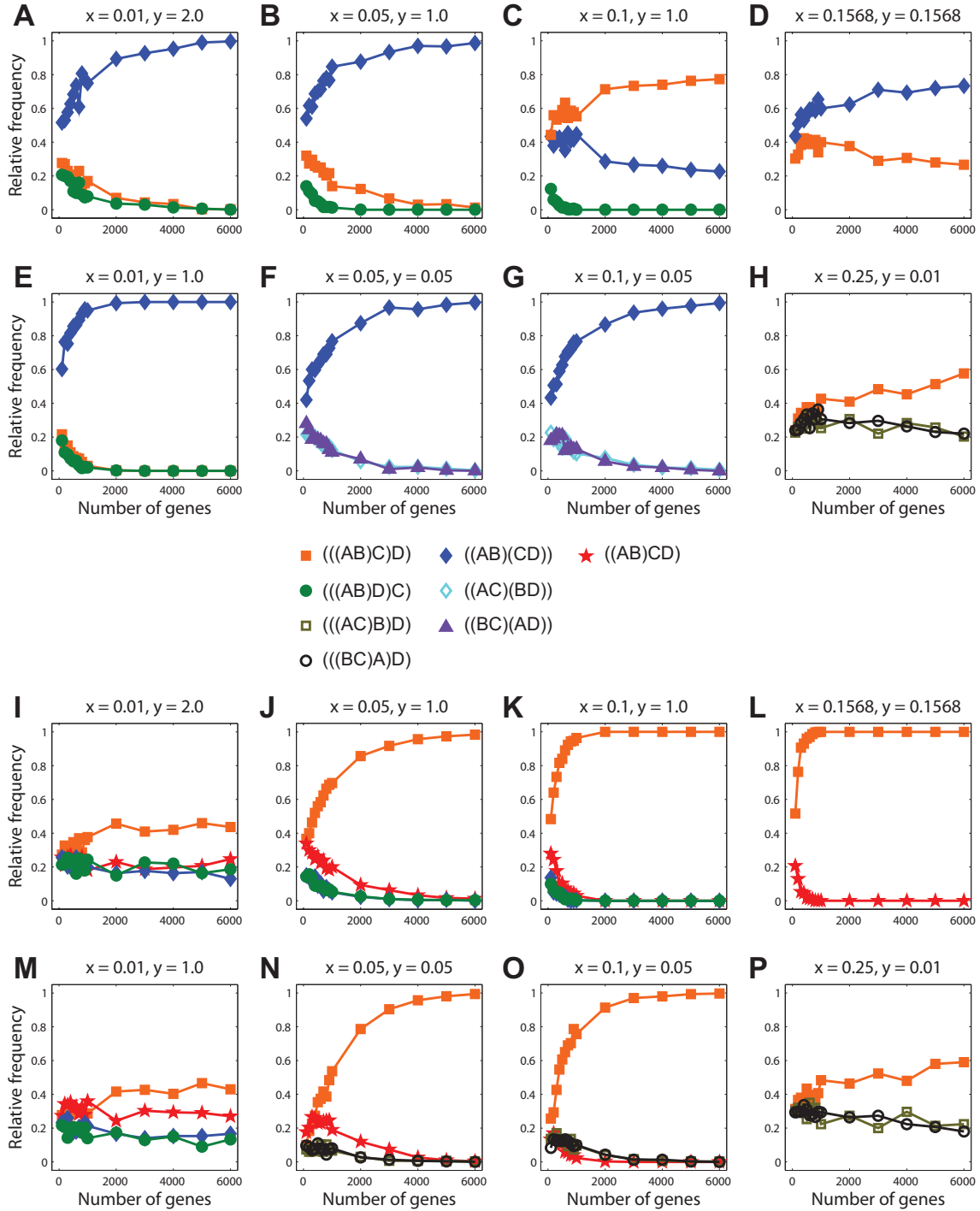


FIG. S5.—Results of simulations for the four-taxon tree $((AB)C)D$ (Figure 1A) generated under a General Time-Reversible model with shape parameter $\alpha = 1$, relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0), $\theta = 0.01$, and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

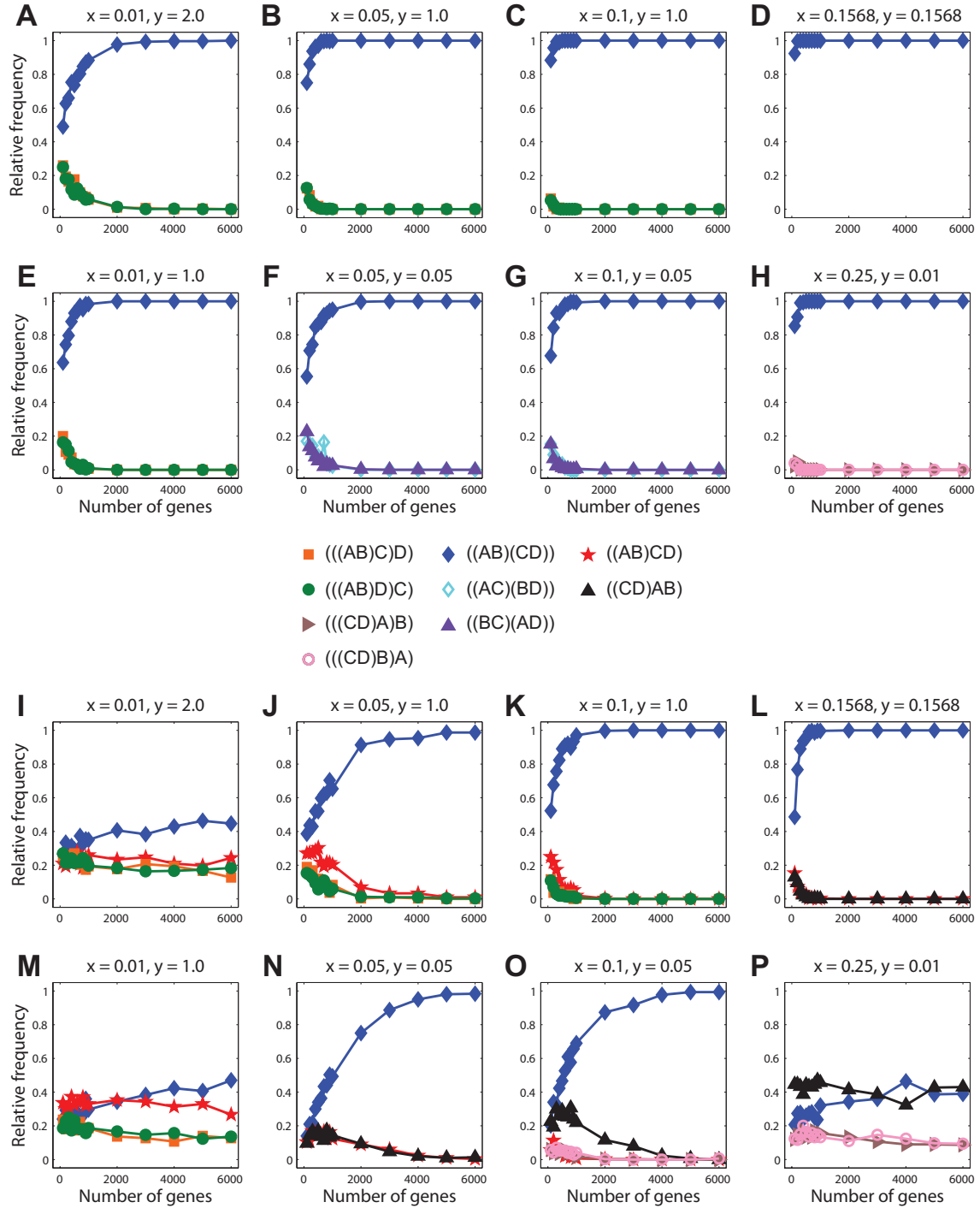


FIG. S6.—Results of simulations for the four-taxon tree $((AB)(CD))$ (Figure 1B) generated under a General Time-Reversible model with shape parameter $\alpha = 1$, relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0), $\theta = 0.01$, and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-H) SM-ML. (I-P) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

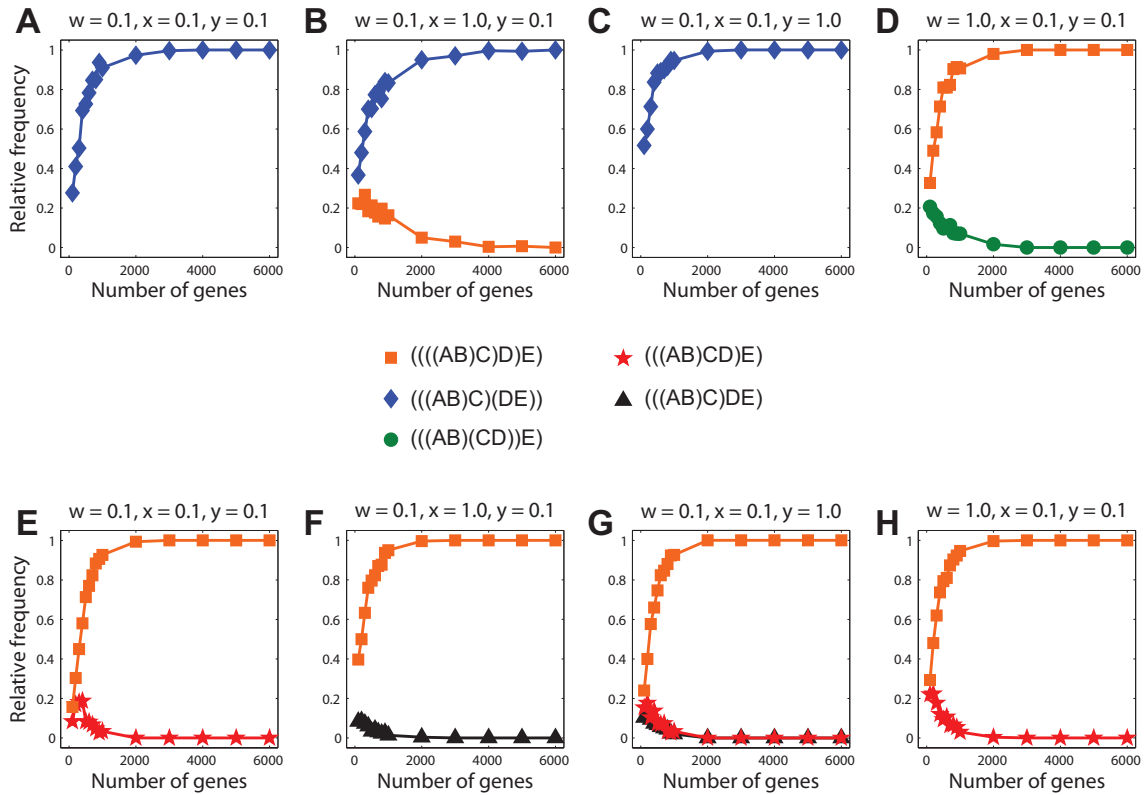


FIG. S7.—Results of simulations for the five-taxon tree $(((AB)C)D)E$ (Figure 1C) generated under a General Time-Reversible model with shape parameter $\alpha = 1$, relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0), $\theta = 0.01$, and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

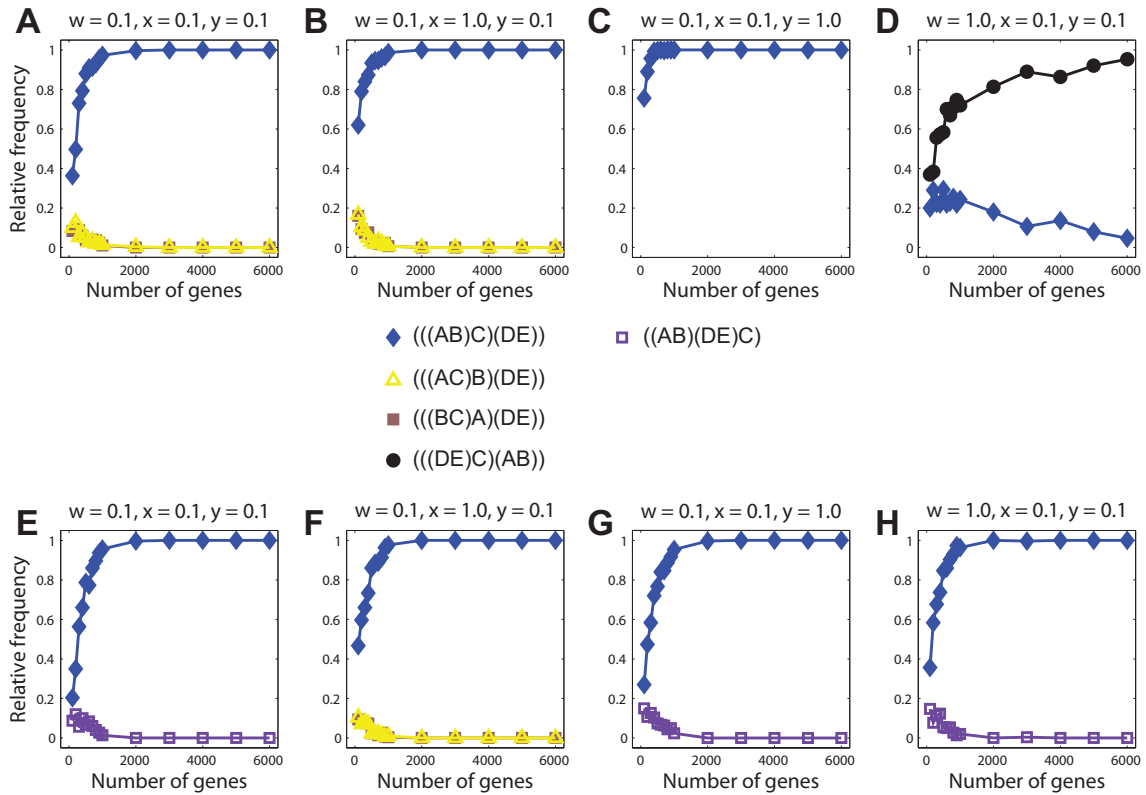


FIG. S8.—Results of simulations for the five-taxon tree $((AB)C)(DE)$ (Figure 1D) generated under a General Time-Reversible model with shape parameter $\alpha = 1$, relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0), $\theta = 0.01$, and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

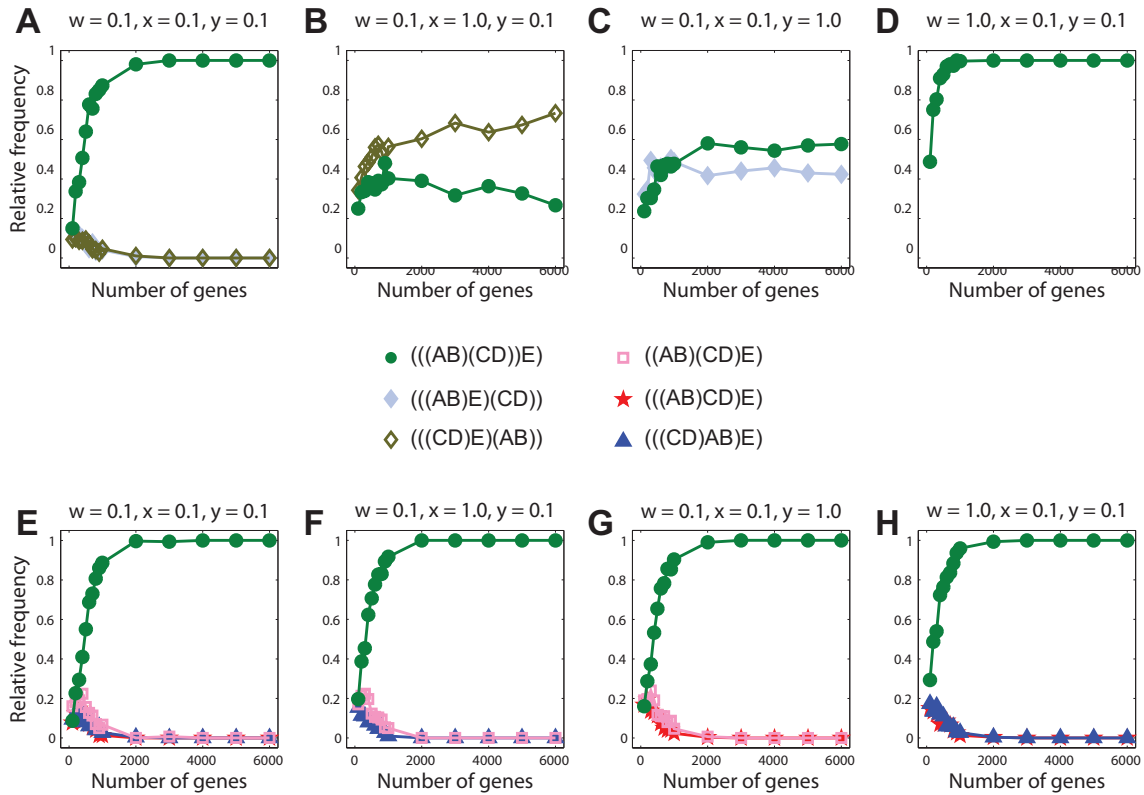


FIG. S9.—Results of simulations for the five-taxon tree $((AB)(CD))E$ (Figure 1E) generated under a General Time-Reversible model with shape parameter $\alpha = 1$, relative frequencies for nucleotides (A, C, G, T) = (0.1, 0.2, 0.3, 0.4), relative rates of substitutions (A \leftrightarrow C, A \leftrightarrow G, A \leftrightarrow T, C \leftrightarrow G, C \leftrightarrow T, G \leftrightarrow T) = (1.5, 1.5, 0.5, 10.5, 1.0, 6.0), $\theta = 0.01$, and a molecular clock, and analyzed under maximum likelihood assuming a molecular clock and a Jukes-Cantor model. (A-D) SM-ML. (E-H) SMRT-ML. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

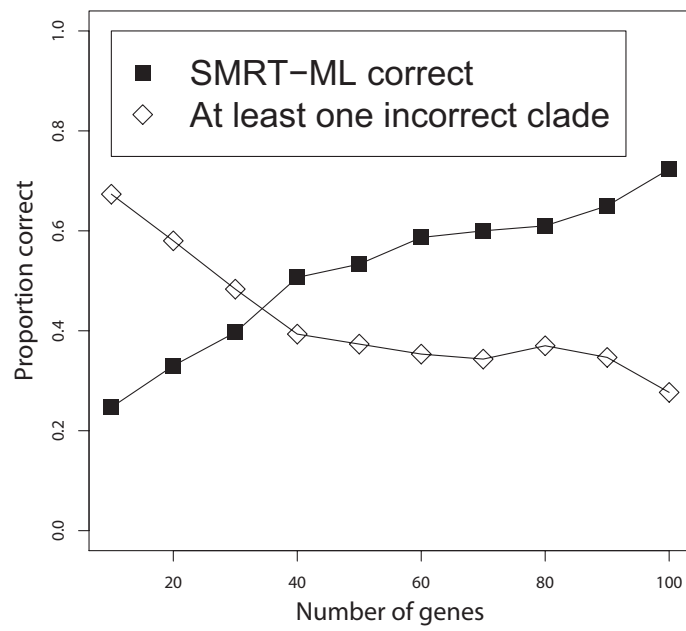


FIG. S10.—Proportion of times SMRT-ML recovers the estimated species tree or at least one false clade for random subsets of genes from the original data set. The two proportions do not add up to 100% because in some cases a partially unresolved tree, which does not have any false clades, is returned by SMRT-ML.

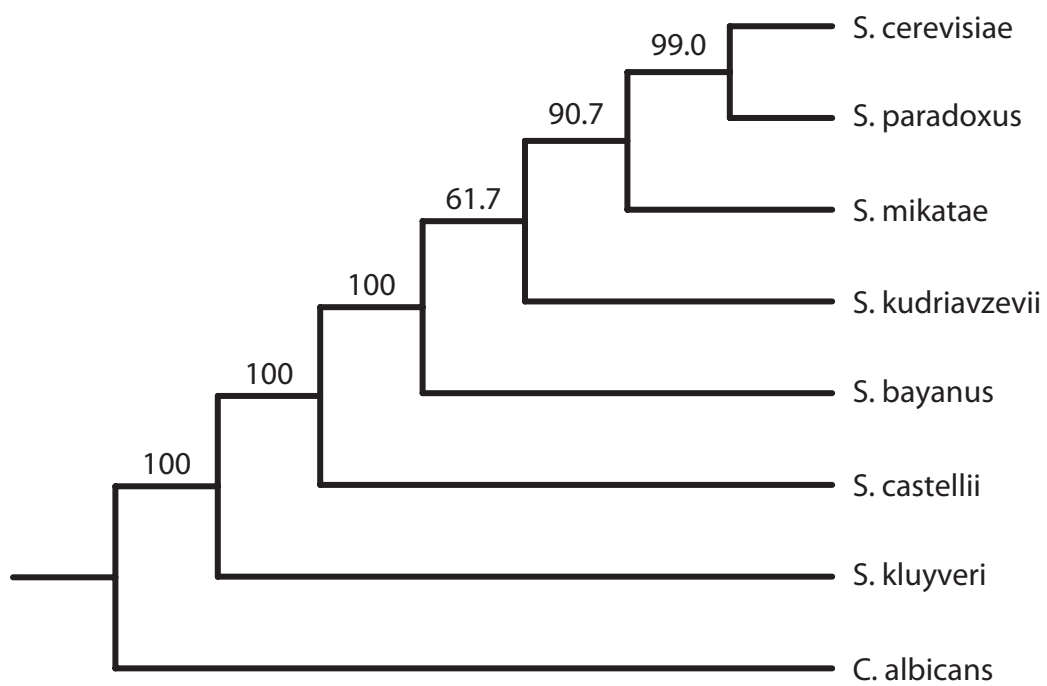


FIG. S11.—Bootstrap support percentages for nodes in the SMRT-ML yeast analysis using the 106-gene data set. Proportions are based on 300 bootstrap replicates.

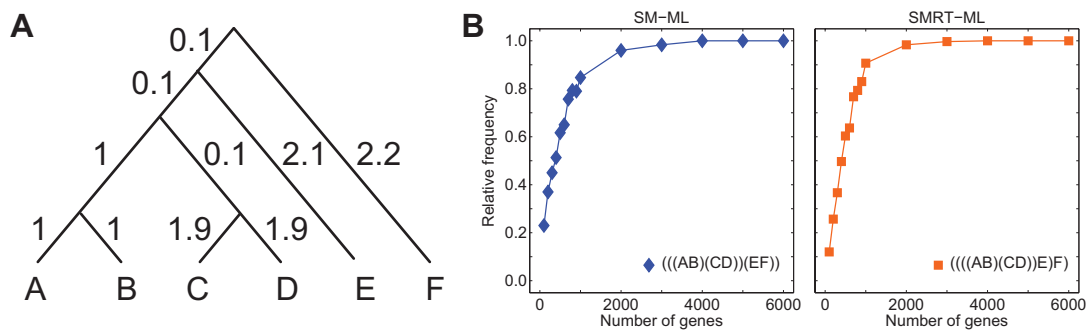


FIG. S12.—Results of simulations for a six-taxon tree with topology $((((AB)(CD))E)F)$. (A) Species tree. (B) SM-ML and SMRT-ML applied to simulated data under a Jukes-Cantor model with $\theta = 0.01$ satisfying a molecular clock analyzed assuming a molecular clock. Data for each combination of branch lengths and number of loci were generated from 300 independent simulations.

