Syst. Biol. 60(2):117–125, 2011

DOI:10.1093/sysbio/syq072
Advance Access publication on December 24, 2010

Genome-Scale Phylogenetics: Inferring the Plant Tree of Life from 18,896 Gene Trees

J. GORDON BURLEIGH[1,2,*], MUKUL S. BANSAL[3,4], OLIVER EULENSTEIN[3], STEFANIE HARTMANN[5,6], ANDRÉ WEHE[3], AND TODD J. VISION[2,5]

[1]*Department of Biology, University of Florida, Gainesville, FL 32609, USA;* [2]*NESCent, Durham, NC 27705, USA;* [3]*Department of Computer Science, Iowa State University, Ames, IA 50011, USA;* [4]*School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel;* [5]*Department of Biology, University of North Carolina, Chapel Hill, NC 27599, USA; and* [6]*Institute for Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany;* **Correspondence to be sent to: Department of Biology, University of Florida, PO Box 118526, Gainesville, FL 32611, USA; E-mail: gburleigh@ufl.edu.*

Received 25 February 2009; reviews returned 3 July 2009; accepted 17 August 2010
Associate Editor: Michael Charleston

Abstract.—Phylogenetic analyses using genome-scale data sets must confront incongruence among gene trees, which in plants is exacerbated by frequent gene duplications and losses. Gene tree parsimony (GTP) is a phylogenetic optimization criterion in which a species tree that minimizes the number of gene duplications induced among a set of gene trees is selected. The run time performance of previous implementations has limited its use on large-scale data sets. We used new software that incorporates recent algorithmic advances to examine the performance of GTP on a plant data set consisting of 18,896 gene trees containing 510,922 protein sequences from 136 plant taxa (giving a combined alignment length of >2.9 million characters). The relationships inferred from the GTP analysis were largely consistent with previous large-scale studies of backbone plant phylogeny and resolved some controversial nodes. The placement of taxa that were present in few gene trees generally varied the most among GTP bootstrap replicates. Excluding these taxa either before or after the GTP analysis revealed high levels of phylogenetic support across plants. The analyses supported magnoliids sister to a eudicot + monocot clade and did not support the eurosid I and II clades. This study presents a nuclear genomic perspective on the broad-scale phylogenic relationships among plants, and it demonstrates that nuclear genes with a history of duplication and loss can be phylogenetically informative for resolving the plant tree of life. [Gene tree–species tree reconciliation; gene tree parsimony; plant phylogeny; phylogenomics.]

The rapidly increasing amount of genomic and cDNA sequence data from non-model organisms provides an abundance of potential information for phylogenetic analyses, but only a small percentage of existing sequence data have been used to infer phylogenetic relationships. One of the challenges of using much of the genomic data is to reconcile the discord between gene trees and the species phylogeny that can result from evolutionary processes such as horizontal transfer, incomplete lineage sorting, or gene duplication and loss (e.g., Goodman et al. 1979; Maddison 1997).

In plants, nuclear genomes are characterized by a particularly high rate of gene duplication and loss (e.g., Lynch and Conery 2003; Adams and Wendel 2005; Sterck et al. 2007), which creates complex patterns of orthology and paralogy within gene families. Consequently, relatively few deep level plant phylogenetic analyses in plants have used low-copy (non-rDNA) nuclear genes (but see Frohlich and Parker 2000; Mathews and Donoghue 2000; Driskell et al. 2004; de la Torre et al. 2006; de la Torre-Bárcena et al. 2009).

One approach to inferring species trees from genes with complex evolutionary histories is to identify the species tree that implies the minimum number of events that cause conflict among the gene trees (e.g., Goodman et al. 1979; Maddison 1997). Gene tree parsimony (GTP) takes a collection of gene trees and seeks a species tree that contains all taxa represented in the gene trees and implies the fewest gene duplications or duplications and losses (e.g., Guigó et al. 1996; Page and Charleston 1997; Slowinski et al. 1997; Slowinski and Page 1999). With incomplete gene sampling, it is difficult to distinguish a gene loss from the absence of sequence data for a gene. Therefore, unless there is complete genomic sequence data for all taxa, it is appropriate to count only gene duplications for the reconciliation cost (e.g., Page and Charleston 1997). GTP analyses have performed well in analyses of snakes (Slowinski et al. 1997), vertebrates (Page 2000; Cotton and Page 2002), sharks (Martin and Burg 2002), *Drosophila* (Cotton and Page 2004), plants (Sanderson and McMahon 2007), and whales (McGowen et al. 2008). However, the run time performance of GTP implementations has limited the size of such studies.

Recent algorithmic advances and software have improved the speed of GTP heuristics and allow, for the first time, analyses using genome-scale data sets from large numbers of taxa (Bansal et al. 2007; Wehe et al. 2008). We use these new computational advances to examine the performance of GTP with a 136 taxon plant data set using over 500,000 protein sequences that comprise 18,896 nuclear gene trees. Our analyses not only demonstrate the phylogenetic utility of nuclear genes but also provide a strongly supported plant phylogeny from large-scale nuclear genomic data.

METHODS

Sequence Assembly

Sequence alignments for plant gene families were obtained from Phytome, an online comparative genomics

database of publicly available sequences (primarily from expressed sequence tag [ESTs]) for 136 plant species (Hartmann et al. 2006). Phytome version 2 contains 793,706 protein sequences that were clustered into 25,763 gene families of size 2 or greater using methods described in the online documentation (www.phytome.org). In brief, sequences were assigned to families using a combination of Tribe-MCL (Enright et al. 2002) and HMMer (Eddy 1998). Multiple sequence alignments of the families were generated using either MAFFT (Katoh et al. 2005) or HMMer (Eddy 1998). The full sequence alignments were masked using the program REducing Alignments prior to Phylogenetic reconstruction (REAP) to ensure positional homology of columns in the alignments (Hartmann et al. 2006; Hartmann and Vision 2008). To build the gene trees used in this study, we selected the masked amino acid alignments from all Phytome version 2 unipeptide families having at least four sequences from at least three taxa. All unipeptide family masked alignments used in the analysis are available at the Dryad data repository (http://dx.doi.org/10.5061/dryad.7881).

Gene Tree Construction

We first performed maximum-likelihood (ML) phylogenetic analyses on each of the masked gene family alignments using RAxML-VI-HPC version 2.2.3 (Stamatakis 2005). The ML analyses used the JTT amino acid substitution model (Jones et al. 1992) with the default settings for the optimization of individual per-site substitution rates and classification of these rates into rate categories ("JTTMIX model"; see Stamatakis 2005). There was no obvious outgroup for most of the gene families, and therefore, the gene trees were initially rooted at their midpoints using the PHYLIP program Retree (Felsenstein 2005). The initial midpoint rootings are subject to be updated in subsequent steps, as described below. We did not perform ML bootstrapping because it was prohibitively time-consuming for all gene families.

ML represents a computationally intensive method of phylogenetic inference, but we also wanted to explore the performance of GTP when the input gene trees were built using much faster methods, such as neighbor-joining (NJ; Saitou and Nei 1987). Therefore, for all gene alignments in which all sequences shared homologous amino acid characters, we calculated pairwise distance matrices for the sequences in the alignment based on the JTT amino acid substitution matrix using the PHYLIP program Protdist (Felsenstein 2005) and input the distance matrix into the PHYLIP program neighbor (Felsenstein 2005) to obtain NJ trees. If an alignment of a gene family contained any pairs of non-overlapping sequences, we estimated optimal gene trees using the generally slower protein parsimony (PP) method, implemented in the PHYLIP program ProtPars (Felsenstein 2005). We also performed 100 nonparametric bootstrap (Felsenstein 1985) replicates for each gene alignment using the same NJ or PP methods that were used to infer the optimal tree for that alignment. The bootstrap data sets were generated using the PHYLIP program Seqboot (Felsenstein 2005).

GTP Analysis

Species tree construction.—To estimate the optimal species tree, we used the fast local rooted subtree pruning and regrafting (rSPR) algorithm of Bansal et al. (2007) as implemented in the software program Duptree (Wehe et al. 2008). All GTP analyses consisted of three stepwise addition replicates to build starting trees and a local rSPR tree search from each starting tree.

GTP requires rooted gene trees, but it is often difficult to determine the true root of a gene tree, especially if there is a possible history of ancient gene duplications and losses. One strategy for rooting gene trees is to use a root that minimizes the duplication cost. Several studies have examined every possible rooting for each gene tree for every candidate species tree (Górecki and Tiuryn 2007; Sanderson and McMahon 2007); however, this is not computationally feasible for our large data set. Therefore, we developed a heuristic strategy to balance the benefits of not assuming a fixed root for the gene trees, with the computational burden of examining many different possible roots for each gene tree. Each rSPR tree search began using midpoint rooted gene trees. After a locally optimal tree was found, the program checked to see if any alternate rootings for the gene trees reduced the duplication score. If the score could be reduced, then all gene trees were optimally rerooted and the local rSPR heuristic search was repeated using the new rootings. In practice, the number of rerooting steps depends on the initial topologies of the gene and species trees, but analyses of our data set typically included 9–12 rerooting steps. The heuristic strategy for examining alternate gene tree rootings is now implemented in Duptree using the "-r opt" option (see Wehe et al. 2008).

We used supertree bootstrapping to examine the effect of uncertainty in the gene tree estimates on the GTP inference (e.g., Cotton and Page 2002). Each supertree bootstrap replicate consisted of a GTP analysis using a single bootstrap tree from each gene family. For example, in the first supertree bootstrap replicate, we used the trees resulting from the first bootstrap replicate for each gene family. We performed 100 supertree bootstrap replicates.

Examining phylogenetic support among species.— We quantified the relative phylogenetic support of each taxon in the bootstrap GTP analyses using quartet distances (e.g., Estabrook et al. 1985). A quartet is a set of four species, and the quartet similarity (or 1−quartet distance) is the proportion of all (unrooted) quartets with identical topologies in two trees. We measured the average quartet similarity between bootstrap trees for quartets

containing each taxon. In other words, for each taxon, we found the average quartet similarity between all pairs of bootstrap trees for only the quartets that contain the taxon. Taxa that are not well supported in the bootstrap trees will have lower average quartet similarity scores because the quartets containing that taxon are more likely to be different in the bootstrap trees. Quartet distance was computed with QDist (Mailund and Pedersen 2004) and normalized to range between 0 and 1 by dividing the total number of identical quartets between trees by the total number of quartets. Thus, a quartet similarity score of 0 corresponds to no shared quartets between trees, and a score of 1 corresponds to identical sets of quartets in the trees.

Results

Data Set

We identified 18,896 gene family alignments from the Phytome database (Hartmann et al. 2006) that had at least four sequences and sequences from at least three taxa. In total, these alignments included 510,922 amino acid sequences, with a combined alignment length of 2,901,617 characters. Sequences from 136 taxa were present in the alignments, with sequences from each taxon appearing in between 30 and 8984 alignments. Overall, 86 of the 136 taxa had sequences in over 1000 of the alignments, and 67 taxa had sequences in over 2000 of the alignments.

Phylogenetic Relationships among All Taxa

GTP analysis using the ML gene trees required 248,757 gene duplications, whereas analysis using a combination of NJ/PP gene trees required 281,052 gene duplications. The species trees built from ML gene trees are available as supplementary data (see http://www.sysbio.oxfordjournals.org/). Despite the difference in the reconciliation cost, there were few major differences in the species tree inferred from ML gene trees and the species tree inferred using NJ/PP gene trees (Fig. 1).

The placement of most taxa in both species trees was largely consistent with their classifications and other large-scale analyses of plant phylogeny (e.g., Soltis et al. 2000; Angiosperm Phylogeny Group 2003; Jansen et al. 2007). Still, there were a few anomalous results in the analyses from NJ/PP gene trees. For example, the two cycads (*Zamia fisheri* and *Cycas rumphii*) were not monophyletic, and *Acorus americanus* and *Zantedeschia aethiopica* were not placed with the other monocots in the majority rule consensus of the bootstrap trees (Fig. 1). Neither of these results occurred in the species tree built from ML gene trees.

We used a supertree bootstrapping approach to incorporate uncertainty in the gene tree topologies into the analysis (e.g., Page and Cotton 2002). The supertree bootstrap replicates consist of GTP analyses that include a single bootstrap gene tree from each gene family. Due to computational limitations, we could not perform ML bootstrap analyses for all 18,896 gene trees. Therefore, we performed supertree bootstrap replicates only using NJ/PP. The overall supertree bootstrap support was low, with only 56% of the possible clades having $\geqslant$70% support (Table 1; Fig. 1). Support for the seed plant clade was 62%, the angiosperm clade was 88%, and the eudicot clade was 72% (Fig. 1).

Relationship between Amount of Data and Phylogenetic Support

We examined the relationship between the amount of data for a taxon and the support for that taxon's placement in the supertree bootstrap analysis. In other words, were the taxa with little data (represented in few gene trees) relatively unsupported within the bootstrap trees? The average taxon quartet similarity (the average similarity of all quartets containing a specified taxon) among the bootstrap trees was lower than 0.89 for quartets with 18 of the 136 taxa. All but 2 of these 18 (*Vitis vinifera* and *V. shuttleworthii*) were represented in less than 1300 gene trees (Fig. 2). Thus, while some taxa with relatively little data were strongly supported, almost all the least supported taxa had relatively little data (Fig. 2). Generally, taxa present in >1300 gene trees were relatively strongly supported (Fig. 2).

In order to examine the underlying relationships among the taxa with the most data, we pruned all taxa for which there were less than 1300 gene trees from the bootstrap trees. The one exception was *Chlamydomonas reinhardtii*, which was present in less than 1300 gene trees but retained nonetheless because it represented the root of the species tree. We then built the majority rule consensus tree from the pruned bootstrap trees, each now containing only 82 taxa; this is a reduced consensus tree (e.g., Wilkinson 1996). In the reduced consensus tree, all but one bipartition had $\geqslant$50% bootstrap support and over half of the clades had 100% bootstrap support (Table 1; Fig. 3).

We compared the reduced consensus approach with an alternative approach in which we excluded the taxa with little data from the original gene tree construction. Specifically, we pruned the 54 low-data taxa (again, retaining *Chlamydomonas*) from the amino acid alignments, performed NJ/PP bootstrapping analyses on the pruned alignments, and repeated the supertree bootstrap analysis using the resulting gene trees. The overall levels of bootstrap support from this analysis were similar to those in the reduced consensus tree (Table 1). In general, the topologies from the two approaches were largely congruent with only a few weakly supported differences. The bootstrap consensus tree from the analysis in which the low-data taxa were excluded from the original alignments is included as supplementary data.

Phylogenetic Relationships among Taxa with Most Sequence Data

Well-supported major clades in the reduced consensus tree of the 136 taxon analysis included seed

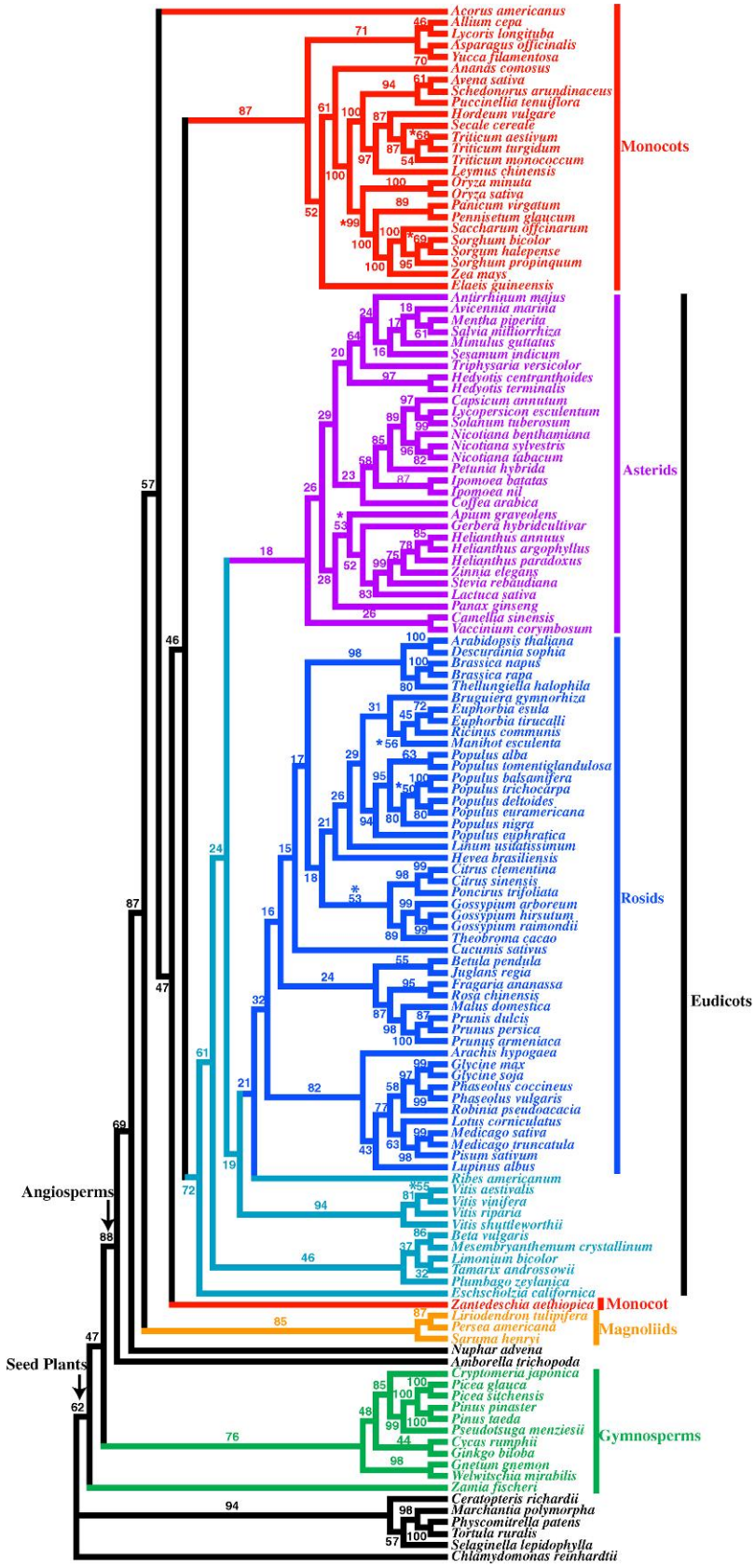

FIGURE 1. Consensus of the 136 taxon NJ/PP GTP supertree bootstrap trees. Numbers represent supertree bootstrap percentages. Asterisk represents branches with at least 50% supertree bootstrap support that are not present in the species tree inferred using ML gene trees.

TABLE 1. Summary of supertree bootstrap support from the GTP analysis

	Number of Taxa	Clades With Bootstrap Support (%)			
		100	⩾90	⩾70	⩾50
136-Taxon Cons.	136	9.8	30.8	56.4	74.4
Reduced Cons.	82	50.6	70.9	89.9	98.7
82-Taxon Cons.	82	53.1	72.2	84.8	96.2

Notes: This displays the percentage of total clades at or above a given level of bootstrap support for 1) the majority rule consensus of all bootstrap trees from the NJ/PP analysis of 136 taxa (136-Taxon Cons.), 2) the reduced consensus of all bootstrap trees for the 82 taxa present in at least 1300 of the gene trees (Reduced Cons.), and 3) the majority rule consensus of all bootstrap trees from the NJ/PP analysis of the same 82 taxa as above (82-Taxon Cons.).

plants (99% support), gymnosperms (100% support), angiosperms (99% support), eudicots (99% support), core eudicots (99% support), and asterids (100% support; Fig. 3). Within gymnosperms, Gnetales were sister to the conifers (100% support; Fig. 3). *Amborella* was sister to all other angiosperms, and *Nuphar* (Nympheales) was sister to all angiosperms except *Amborella* (Fig. 3). Magnoliids were sister to a monocot + eudicot clade (Fig. 3). Within monocots, the Poaceae (grass family) had 100% support, and within the grasses, the Panicoideae clade had 100% bootstrap support (Fig. 3). In the core eudicot clade, the Caryophyllales (100% support) were sister to the rosids (99% support) and the asterids (100% support) (Fig. 3).

There were several differences in the species tree obtained using ML gene trees versus NJ/PP gene trees. For example, the relationships among eurosid lineages differed slightly; however, in both analyses, Malpighiales (eurosid I) were nested in a clade with eurosid II taxa (Figs. 1 and 3). The BEP-clade (Bambusoideae, Ehrhartoideae, and Pooideae) was not supported in the analysis using NJ/PP gene trees, but it was when using ML gene trees (Fig. 3). *Acorus americanus* was not placed with other monocots in the NJ/PP analysis, but it was in a monocot clade when using ML gene trees (Fig. 3).

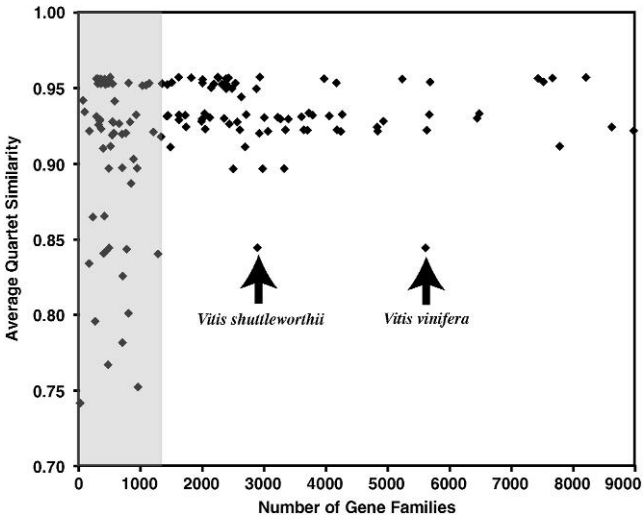

FIGURE 2. Average quartet similarity for each taxon among bootstrap trees. Each point in the graph represents a single taxon. The *x*-axis shows the number of gene families trees that have data from the taxon. The *y*-axis shows the average percentage frequency of quartets (four taxon statements) containing the taxon that are identical between two bootstrap trees. The shaded area in the graph contains all taxa that are present in less than 1300 gene trees.

DISCUSSION

Frequent gene and whole-genome duplications have, in the past, limited the use of nuclear genes for deep level phylogenetic analyses in plants and other clades with highly duplicated genomes. GTP provides a way to exploit the phylogenetic information inherent not only in the relationships among orthologous genes but also the rare gene duplications that produce paralogous gene family members. Rather than treating gene tree discordance as a nuisance, it seeks the species tree that provides the best reconciliation among the many discordant gene trees.

In this study, we used GTP to find species trees that minimize the total number of duplications across a collection of nearly 18,896 plant gene trees. The sequence sampling includes extensive collections of existing EST data that have rarely before been used for plant phylogenetics (but see de la Torre et al. 2006; Sanderson and McMahon 2007; de la Torre-Bárcena et al. 2009). Thus, this study provides a new nuclear genomic perspective on the plant tree of life.

Overall, the phylogenetic relationships inferred from gene duplications are largely consistent with previous large-scale molecular studies of plant phylogeny (e.g., Soltis et al. 2000; Hilu et al. 2003; Jansen et al. 2007). Yet the GTP analysis also provides support for some relationships that are unresolved or conflicting in previous analyses. For example, the results support the placement of magnoliids sister to monocots + eudicots, making eudicots (possibly with *Ceratophyllum*, which was not included in this study) sister to monocots (Figs 1 and 3). The relationships among these major clades are unclear from analyses using few genes (e.g., Soltis et al. 2000, 2007; Hilu et al. 2003), but our result is consistent with recent analyses using 81 plastid genes (Jansen et al. 2007). The placement of Malpighiales within a eurosid I clade (Figs. 1 and 3) generally conflicts with previous large-scale angiosperm analyses (e.g., Soltis et al. 2000; Hilu et al. 2003; Jansen et al. 2007). Given the novelty of the result, it should be interpreted with great caution.

Our results indicate that data from many gene trees may be required to produce a well-supported phylogeny using GTP (Table 1; Figs. 2 and 3), suggesting that GTP may not use data as efficiently as more traditional phylogenetic analyses of concatenated multigene data sets. For example, in plants, recent analyses of up to 83 plastid genes have apparently resolved enigmatic relationships in the backbone angiosperm phylogeny, whereas our analyses appear to require data from >1000 genes (Jansen et al. 2007; Moore et al. 2007, 2010). Like

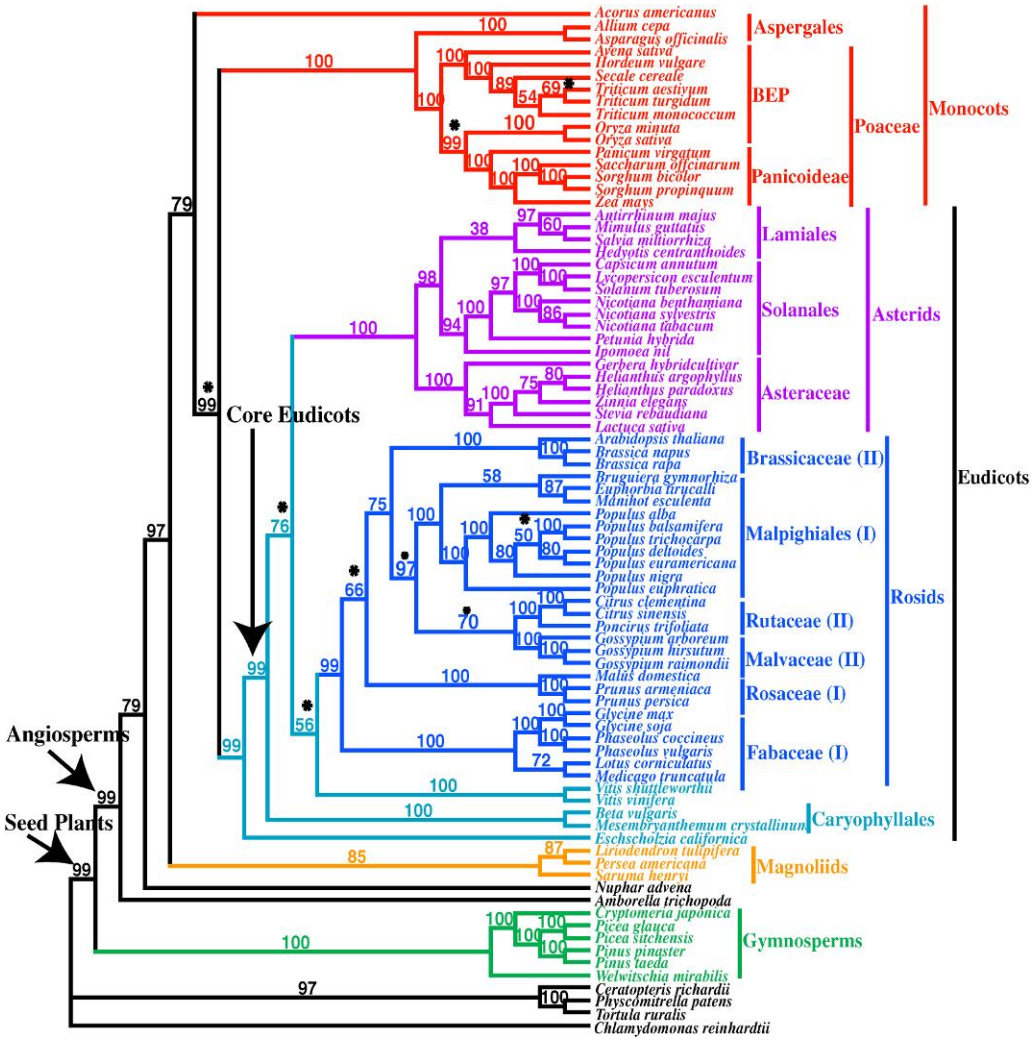

FIGURE 3. Reduced consensus of the GTP supertree bootstrap trees. To build the reduced consensus, 54 taxa that were present in less than 1300 of the gene trees were pruned from the bootstrap tree, and then a consensus was made from the pruned trees. Numbers above the branches are the supertree bootstrap percentages. Asterisk represents branches that are not present in the species tree inferred using ML gene trees.

other supertree methods, GTP discards the primary sequence data and relies on only a summary tree topology for each gene (e.g., Bininda-Emonds 2004). While GTP may not use gene sequence data as efficiently as conventional alignment-based phylogenetic methods, GTP makes use of data from large gene families that are not easy to use in conventional phylogenetic analyses at all. The large number of gene trees needed to produce a highly supported topology also suggests high levels of stochastic error associated with the gene tree topologies. Genome-scale data sets may include enough data to reduce the effects of stochastic error, resulting in strong support. This does not mean that there is not systematic error; the high levels of support could be positively misleading (e.g., Phillips et al. 2004).

Although the taxa that were the most difficult to place in the GTP analysis also were present in relatively few gene trees (Fig. 2), there is not necessarily a direct relationship between the amount of data and support in GTP. Not all taxa with data in few gene trees were difficult to place in the analysis, and conversely, the two *Vitis* taxa had a low average similarity for quartets containing these species despite appearing in over 2000 gene trees (Fig. 2). Other factors, such as the amount of overlap with other taxa in the gene trees, and the support within the gene trees likely affect the support for a taxon's placement in the species tree. For example, the placement of *Vitis* has received low support in previous large-scale angiosperm analyses (e.g., Hilu et al. 2003; Soltis et al. 2007), including recent analyses of a 36 gene data set (Wang et al. 2009) and an 83 gene plastid genome data set (Moore et al. 2010). If the position of *Vitis* is unsupported or varies among single-gene phylogenetic analyses, it is unlikely that a gene tree reconciliation of the gene trees will be able to place *Vitis*. Similarly, if the gene trees all strongly support an erroneous topology, we would expect the reconciled tree also to include this erroneous topology.

For example, previous analyses have demonstrated that long branch attraction can erroneously place *Acorus* outside the monocots (e.g., Stefanovic et al. 2004), and this may explain our similar result in the GTP analyses that used the NJ/PP, but not ML, gene trees (Figs. 1 and 3).

With the new GTP heuristics, the speed of phylogenetic analyses of the gene trees may be a greater limitation to GTP analyses than that of the GTP analysis. Erroneous gene trees can easily mislead a GTP analysis. For example, Sanderson and McMahon (2007) found that the optimal species tree inferred using maximum parsimony (MP) gene trees for a seven taxon test data set was incorrect, whereas the species tree inferred from ML gene trees was correct. Similarly, the position of *Acorus* outside the monocots and *Oryza* outside the remaining BEP clade in GTP analyses using NJ/PP gene trees also may be a case of gene tree error affecting the GTP inference (Figs. 2 and 3). While in some cases, likelihood-based phylogenetic methods may produce less systematic error than NJ or PP (e.g., Huelsenbeck 1995; Swofford et al. 2001), they often are more computationally intensive than NJ or MP analyses, and performing ML bootstrapping on large numbers of gene trees may be computationally prohibitive (but see Stamatakis et al. 2008). Methods that filter potentially problematic regions of the gene family alignments, such as REAP (Hartmann et al. 2006; Hartmann and Vision 2008) and GBlocks (Talavera and Castresana 2007), may reduce error in NJ or PP analyses gene family alignments (Hartmann and Vision 2008) and thus increase the accuracy of GTP.

Our GTP results demonstrate the utility of large-scale nuclear genomic data sets for resolving organismal phylogeny, but they also suggest some future directions for gene tree–species tree reconciliation methods. For example, our implementation of GTP reconciles trees only based on duplication, and it does not account for other evolutionary processes that can confound gene tree and species tree topologies, such as incomplete lineage sorting (e.g., Maddison and Knowles 2006), horizontal transfer (Ge et al. 2005), and error in gene tree construction (Sanderson and McMahon 2007). The performance of GTP might be improved by considering some of these additional factors that can confound gene tree and species tree topologies. Recently developed heuristics now allow GTP analyses of extremely large-scale data sets using the duplication loss and deep coalescence cost (Bansal et al. 2010), but these methods are largely uncharacterized.

While the parsimony approach is computationally tractable for very large data sets, a parsimony approach that minimizes the number of duplications, duplications plus losses, or deep coalescence events may not be appropriate when the rates of these events are high. In such cases, a model-based reconciliation approach may be needed. A likelihood-based method could also potentially jointly estimate the gene tree and species tree topologies. Recently, there have been several model-based methods for estimating the species tree from a set of gene trees based on a coalescent process (e.g., Liu and Pearl 2007; Liu et al. 2008; Kubatko et al. 2009). However, these methods do not explicitly address gene duplication and loss, which is pervasive in nuclear gene evolution. Probabilistic models of gene tree–species tree reconciliation that incorporate gene duplication and loss do exist (e.g., Arvestad et al. 2004; Åkerborg et al. 2009), but because they are computationally complex have not been incorporated into any species tree search heuristics.

Our study demonstrates that the growth of genomic sequence data from a large number of organisms, along with recent algorithmic and methodological advances in GTP, provides new opportunities for genome-scale phylogenetic analyses. In the present case, even with opportunistic sampling of existing and largely incomplete genome sequence data, GTP analyses revealed a strong phylogenetic signal. The results from this study should further motivate the refinement of gene tree reconciliation methods for inferring the tree of life from nuclear genomic data.

Supplementary Material

Supplementary material can be found at http://www.sysbio.oxfordjournals.org/.

Funding

This work was supported by the National Science Foundation (EF0334832 to O.E., DB0227314 to T.J.V., and EF0423641 to J.G.B. and T.J.V.) and the National Institutes of Health (R01-GM078991 to T.J.V.).

Acknowledgments

We thank Mike Sanderson for extensive discussion of GTP. Michael Charleston, James Cotton, and James McInerney provided very useful comments on this manuscript.

References

Adams K.L., Wendel J.F. 2005. Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. 8:135–141.

Åkerborg Ö., 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.

Angiosperm Phylogeny Group. 2003. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. Bot. J. Linn. Soc. 141:399–436.

Arvestad L., Berglund A.-C., Lagergren J., Sennblad B. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. RECOMB. 2004:326–335.

Bansal M.S., Burleigh J.G., Eulenstein O. 2010. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. BMC Bioinformatics. 11:S42.

Bansal M.S., Burleigh J.G., Eulenstein O., Wehe A. 2007. Heuristics for the gene-duplication problem: a $\Theta(n)$ speed-up for the local search. RECOMB 2007, LNCS. 4453:238–252.

Bininda-Emonds O.R.P. 2004. The evolution of supertrees. Trends Ecol. Evol. 19:315–322.

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.
Cotton J.A., Page R.D.M. 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. P. Roy. Soc. Lond. B Biol. 269:1555–1561.
Cotton J.A., Page R.D.M. 2004. Reconciled trees for supertree construction. Pages 107-125 in Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht (The Netherlands): Kluwer Academic Press.
Cotton J.A., Wilkinson M. 2009. Supertrees join the mainstream of phylogenetics. Trends Ecol. Evol. 24:1–3.
de la Torre J.E.B., Egan M.G., Katari M.S., Brener E.D., Stevenson D.W., Coruzzi G.M., DeSalle R. 2006. ESTimating plant phylogeny: lessons from partitioning. BMC Evol. Biol. 6:48.
de la Torre-Bárcena J.E., Kolokotronis S.-O., Lee E.K., Stevenson D.W., Brenner E.D., Katari M.S., Coruzzi G.M., DeSalle R. 2009. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. PLoS One. 4:e5764.
Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science. 306:1172–1174.
Eddy S.R. 1998. Profile hidden Markov models. Bioinformatics. 14: 755–763.
Enright A.J., Van Dongen S., Ouzounis C.A. 2002. An efficient algorithm for large-scale detection of protein families. Nuc. Acids Res. 30:1575–1584.
Estabrook G.F., McMorris F.R., Meacham C.A. 1985. Comparison of undirected phylogenetic trees based on subtrees of 4 evolutionary units. Syst. Zool. 34:193–200.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 39:783–791.
Felsenstein J. 2005. PHYLIP: phylogeny inference package version 3.6. Seattle (WA): University of Washington.
Frohlich M.W., Parker D.S. 2000. The mostly male theory of flower evolutionary origins: from genes to fossils. Syst. Bot. 25:155–170.
Ge F., Wang L.-S., Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3: e316.
Goodman M., Czelusniak J., Moore G.W., Romero-Herrera A.E., Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed by globin sequences. Syst. Zool. 28:132–163.
Górecki P., Tiuryn J. 2007. Urec: a system for unrooted reconciliation. Bioinformatics. 23: 511–512.
Guigó R., Muchnik I., Smith T.F. 1996. Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. 6:189–213.
Hartmann S., Lu D., Phillips J., Vision T.J. 2006. Phytome: a platform for plant comparative genomics. Nuc. Acids Res. 34:D724–D730.
Hartmann S., Vision T.J. 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol. Biol. 8:95.
Hilu K.W., Borsch T., Müller K., Soltis D.E., Soltis P.S., Savolainen V., Chase M.W., Powell M.P., Alice L.A., Evans R., Sauquet H., Neinhuis C., Slotta T.A.B., Rohwer J.G., Campbell C.S., Chatrou L.W. 2003. Angiosperm phylogeny based on *matK* sequence information. Am. J. Bot. 90:1758–1776.
Huelsenbeck J.P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.
Jansen R.K., Cai Z., Raubeson L.A., Daniell H., dePamphilis C.W., Leebens-Mack J., Müller K.F., Guisinger-Bellian M., Haberle R.C., Hansen A.K., Chumley T.W., Lee S.-B., Peery R., McNeal J.R., Kuehl J.V., Boore J.L. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. U. S. A. 104: 19369–19374.
Jones D.T., Taylor W.R., Thornton J.M. 1992. The rapid generation of mutation data matrices from protein sequences. Comp. Appl. Biosci. 8:25–282.
Katoh K., Kuma K., Toh H., Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nuc. Acids Res. 33:511–518.
Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 25:971–973.
Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56: 504–514.
Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution. 62:2080–2091.
Lynch M., Conery J.S. 2003. The evolutionary demography of duplicate genes. J. Struc. Func. Genomics. 3:35–44.
Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46: 523–536.
Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.
Mailund T., Pedersen C.N.S. 2004. QDist—quartet distance between evolutionary trees. Bioinformatics. 20:1636–1637.
Martin A.P., Burg T.M. 2002. Perils of paralogy: using HSP70 genes for inferring organismal phylogeny. Syst. Biol. 51:570–587.
Mathews S., Donoghue M.J. 2000. Basal angiosperm phylogeny inferred fro duplicate phytochromes A and C. Int. J. Plant Sci. 161:S41–S55.
McGowen M.R., Clark C., Gatesy J. 2008. The vestigial olfactory receptor subgenome of odontocete whales: phylogenetic congruence between gene-tree reconciliation and supermatrix methods. Syst. Biol. 57:574–590.
Moore M.J., Bell C.D., Soltis P.S., Soltis D.E. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. U. S. A. 104:19363–19368.
Moore M.J., Soltis P.S., Bell C.D., Burleigh J.G., Soltis D.E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc. Natl. Acad. Sci. U. S. A. 107: 4623–4628.
Page R.D.M. 2000. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol. Phylogenet. Evol. 14:89–106.
Page R.D.M., Charleston M.A. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol. Phylogenet. Evol. 7:231–240.
Page R.D.M., Cotton J.A. 2002. Vertebrate phylogenomics: reconciled trees and gene duplication. Pac. Symp. Biocomput. 7:536–547.
Phillips M.J., Delsuc F., Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21: 1455–1458.
Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–425.
Sanderson M.J., McMahon M.M. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 7:S3.
Slowinski J.B., Knight A., Rooney A.P. 1997. Inferring species trees from gene trees: a phylogenetic analysis of the elapidae (serpentes) based on the amino acid sequences of venom proteins. Mol. Phylogenet. Evol. 8:349–362.
Slowinski J.B., Page R.D.M. 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48:814–825.
Soltis D.E., Gitzendanner M.A., Soltis P.S. 2007. A 567-taxon data set for angiosperms: the challenges posed by Bayesian analyses of large data sets. Int. J. Plant Sci. 168:137–157.
Soltis D.E., Soltis P.S., Chase M.W., Mort M.E., Albach D.C., Zanis M., Savolainen V., Hahn W.H., Hoot S.B., Fay M.F., Axtell M., Swensen S.M., Prince L.M., Kress W.J., Nixon K.C., Farris J.S. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Bot. J. Linn. Soc. 133:381–341.
Stamatakis A. 2005. RAxML-VI-HPC: maximum likelihood phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.
Stamatakis A., Hoover P., Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web-servers. Syst. Biol. 75:758–771.
Stefanovic S., Rice D.W., Palmer J.D. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol. Biol. 4:35.

Sterck L., Rombauts S., Vandepoele K., Rouzé P., Van de Peer Y. 2007. How many genes are there in plants (... and why are they there)? Curr. Opin. Plant Biol. 10:199–203.

Swofford D.L., Waddell P.J., Huelsenbeck J.P., Foster P.G., Lewis, P.O. Rogers J.S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.

Wang H., Moore M.J., Soltis P.S., Bell C.D., Brockington S.F., Alexandre R., Davis C.C., Latvis M., Manchester S.R., Soltis D.E. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc. Natl. Acad. Sci U. S. A. 106:3853–3858.

Wehe A., Bansal M.S., Burleigh J.G., Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 24:1540–1541.

Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13:437–444.