How Many Genes are Needed to Resolve Phylogenetic Incongruence?

Bin Ai and Ming Kang

Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China.

ABSTRACT: The question how many genes are needed to resolve phylogenetic incongruence has been investigated at various taxonomic levels, yet few studies have investigated the minimum required numbers of selected genes based on single-gene tree performance at the genus level or lower. We conducted resampling analyses by compiling transcriptome-based single-copy nuclear gene sequences of 11 species of *Primulina* (*Gesneriaceae*) to investigate the minimum numbers of both random and selected genes needed to resolve the phylogeny. Only 8 of the 26 selected genes were sufficient for full resolution, while 175 genes were needed if all 830 random genes were used. Our results provided a baseline for future sampling strategies of gene numbers in molecular phylogenetic studies of speciose taxa. The gene selection strategies based on single-gene tree performance are strongly recommended in phylogenic analyses.

KEYWORDS: phylogenetic incongruence, consensus network, gene number, resampling analyses, *Primulina*

CITATION: Ai and Kang. How Many Genes are Needed to Resolve Phylogenetic Incongruence?. *Evolutionary Bioinformatics* 2015:11 185–188 doi: 10.4137/EBO.S26047.

TYPE: Rapid Communication

RECEIVED: March 18, 2015. **RESUBMITTED:** June 16, 2015. **ACCEPTED FOR PUBLICATION:** June 22, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,397 words, excluding any confidential comments to the academic editor.

FUNDING: This work was supported by the National Natural Science Foundation of China (31270427, 31300178).The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: mingkang@scbg.ac.cn

Introduction

Understanding the phylogenetic relationships among living organisms is fundamental to any comparative research in biology.[1] In the era of phylogenomics, phylogenetic incongruence has been widely documented in phylogeny construction with a large amount of available low-copy genes and is often ascribed to stochastic error, systematic error, and/or biological factors.[2–5] To overcome phylogenetic incongruence, a practical question is how many genes should be used to generate a robust phylogenetic hypothesis.[6] This question has been investigated at various taxonomic levels.[7–13] For example, one notable resampling analysis showed that a minimum of 20 concatenated genes were required to provide 95% bootstrap support for all nodes by compiling a data set from 106 random orthologous genes for seven *Saccharomyces* species and one out-group species.[7] In a recent study, among 59 carefully selected low-copy nuclear genes based on single-gene tree performance, fewer than 50 were enough to solve the deep angiosperm phylogeny.[13] Nevertheless, few studies have investigated the minimum required numbers of selected genes based on single-gene tree performance at the genus level or lower.

Primulina, one of the largest genera of the Old World *Gesneriaceae*, is a monophyletic group comprising more than 140 species widely distributed throughout southern China and adjacent countries in Southeast Asia, which is a biodiversity hot spot of the World.[14,15] Nevertheless, interbreeding through artificial experiments can still succeed among many *Primulina* species pairs,[16] suggesting that the genus has undergone recent rapid speciation or population differentiation. Although the entire *Primulina* phylogeny is far from resolved to date, a rough genus framework has been presented with sequences of four loci from 104 *Primulina* species,[17] and the phylogeny of 11 representative *Primulina* species has been fully resolved based on 834 putative single-copy nuclear genes identified from transcriptome data.[18] The high species richness and endemism, and low interspecific genetic isolation, together with comparative transcriptomic resources, make *Primulina* an ideal model system to study the gene numbers needed to obtain a robust phylogeny. In this study, we conducted resampling analyses by compiling the previously published transcriptome-based single-copy nuclear gene sequences of 11 representative *Primulina* species[18] to determine the minimum numbers of both random and selected genes needed to resolve the phylogeny and to provide a baseline for future sampling strategies of gene numbers in molecular phylogenetic studies of speciose taxa.

Materials and Methods

The aligned sequence data for 830 putative single-copy nuclear genes (Supplementary Table 1) from 11 *Primulina* species[18] were used for analysis, while the other four loci were discarded due to poor alignment performance. To test the utility of individual genes, all single-gene trees were constructed with

1,000 bootstrap replicates based on nucleotide sequences, using the maximum-likelihood (ML) method implemented in PhyML v3.0.[19] *Primulina swinglei* was chosen as out-group according to the *Primulina* phylogeny.[17] The nucleotide substitution model was specified as GTR+I+Γ and the branching-swapping method was set to subtree pruning and regrafting. To consider the limitations of the concatenation approach for phylogenetic inference,[20] the average rank of coalescences (STAR) method[21] implemented in the species tree analysis web server STRAW[22] was conducted based on the 830 single-gene trees to infer the species tree of the 11 *Primulina* species. Those genes based on which the topologies of the 11 *Primulina* species were congruent with the species tree were screened as good genes. To visualize the phylogenetic incongruence among different topologies, SplitsTree v4.13.1[23] was used to infer the consensus network from 830 source single-gene trees with a set of threshold values (0.01, 0.05, 0.1, 0.15, 0.2, and 0.25). Different threshold values were set to control the visual complexity of resulting networks by using only the splits that occurred in more than a given proportion of all trees.

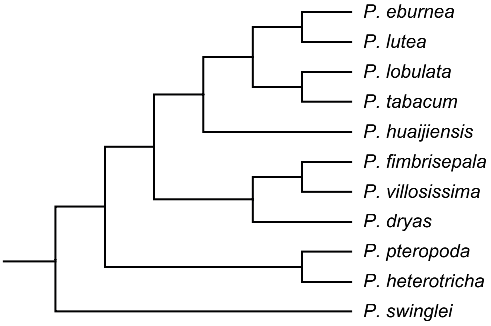

Figure 1. Species tree of the 11 *Primulina* species inferred from the 830 single-gene trees with the average rank of coalescences (STAR) method.

To investigate the effect of the number of sampled genes or nucleotides on the probability of the inferred species tree, random resampling without replacement from the original 830-gene (631,686-nucleotide) data set was performed with a custom java script. The number of sampled genes started from 25 with increments of 25, until the percentage of correct trees (see below) reached 100. Similarly, resampling by site started from 10 kb with increments of 10 kb. For each sample, 1,000 replicates were generated and the ML gene trees were inferred by RAxML v8[24] with the model GTR+Γ and 100 fast bootstrap replicates. Those trees with the same topology as the species tree and all ML bootstrap values larger than a threshold (95% or 75%) were counted as correct trees. Gene resampling among the selected good genes was also conducted and started from three with increments of one.

Results

The inferred STAR species tree (Fig. 1) showed the same topology as the previously published *Primulina* topology based on concatenated data[18] and all bootstrap values were 100. Only 26 of the 830 single-gene trees exhibited the same topology as Figure 1 (Supplementary Tables 1 and 2) and there were 25 topologies supported by more than three genes (Supplementary Table 3), indicative of extensive single-tree phylogeny incongruence, which was also strongly supported by the consensus network constructed from 830 source single-gene trees (Fig. 2). Network complexity decreased

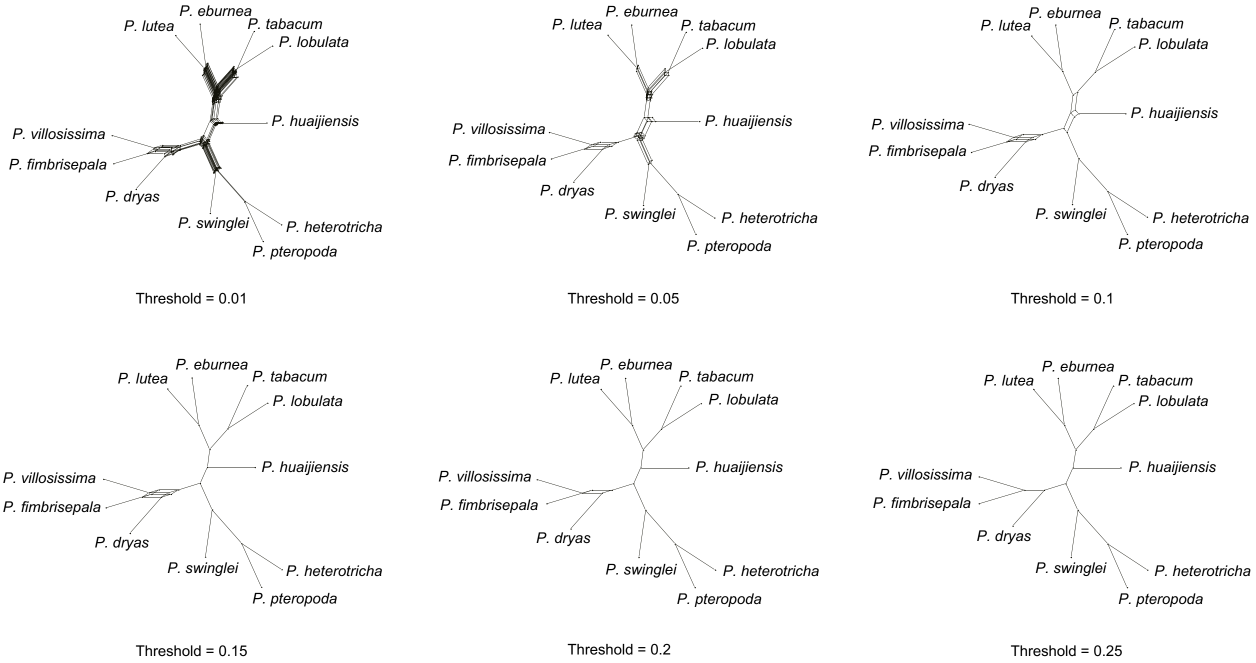

Figure 2. Consensus network of the 11 *Primulina* species constructed from 830 single-gene trees at six different threshold values.

as increasing threshold values and ended up with a topology identical to that shown in Figure 1 when the threshold reached 0.25, in which all boxes were collapsed. Among the 11 species, the triplet of *Primulina fimbrisepala*, *Primulina dryas*, and *Primulina villosissima* was the most difficult clade to resolve. Of the 830 single-gene trees, 404 (48.7%) supported the correct sister relationship between *P. fimbrisepala* and *P. villosissima*, while those supporting the alternative sister relationship between *P. fimbrisepala* and *P. dryas* or between *P. dryas* and *P. villosissima* numbered 206 (24.8%) and 220 (26.5%), respectively (Supplementary Table 1), suggesting extremely rapid triplet speciation. These results further support the previously published *Primulina* phylogeny[18] and highlight the phylogenetic incongruence of the triplet of *P. fimbrisepala*, *P. dryas* and *P. villosissima*.

The resolution performance through resampling by gene or site demonstrated that the probability of getting a solid topology steadily increased with the number of genes or sites sampled (Table 1). Using 95% correct trees with a 95% bootstrap threshold in 1,000 replicates as the criterion, about 175 genes were needed to resolve all 11 species, and about 75 genes were needed when ignoring the triplet complexity. When resampling by nucleotide site, about 70 kb of nucleotides yielded sufficient resolution, equivalent to 92 sampled genes given the average length of 761 bp per gene, and it took about 30 kb (about 39 genes) for the other clades except the triplet. Nevertheless, when resampling by gene among the 26 selected genes (Supplementary Table 2), only 8 genes were needed to resolve all the clades. As expected, all the counts decreased somewhat when using a bootstrap threshold of 75% rather than 95%. Under a bootstrap threshold of 75%, 4 selected genes were sufficient for full resolution, while 125 genes were needed if all 830 random genes were used.

Table 1. Summary of the percentages of correct trees at a bootstrap threshold of 95% or 75% (in parentheses) for all 11 *Primulina* species, the triplet, and other species, through random resampling by gene or site from the original 830-gene (631,686-nucleotide) data set.

N	ALL	TRIPLET	OTHER
Gene			
25	26.6 (51.2)	41.0 (63.4)	57.2 (77.1)
50	50.1 (75.0)	59.0 (78.9)	82.7 (94.4)
75	68.9 (85.9)	72.7 (87.2)	95.2 (98.7)
100	79.9 (92.6)	81.3 (92.8)	98.5 (99.8)
125	87.5 (95.6)	87.8 (95.6)	99.5 (100)
150	93.8 (98.7)	93.9 (98.7)	99.9 (100)
175	96.0 (98.9)	96.1 (98.9)	99.9 (100)
200	97.3 (99.6)	97.3 (99.6)	100 (100)
225	98.2 (99.7)	98.2 (99.7)	100 (100)
250	99.3 (99.8)	99.3 (99.8)	100 (100)
275	99.8 (100)	99.8 (100)	100 (100)
300	100 (100)	100 (100)	100 (100)
Site			
10 k	10.4 (54.7)	22.1 (61.3)	50.7 (88.8)
20 k	39.4 (81.4)	44.3 (82.6)	88.4 (98.6)
30 k	64.7 (92.2)	66.2 (92.3)	97.6 (99.9)
40 k	75.7 (96.7)	76.0 (96.7)	99.7 (100)
50 k	87.7 (98.7)	87.8 (98.7)	99.9 (100)
60 k	92.7 (99.5)	92.7 (99.5)	100 (100)
70 k	96.1 (99.9)	96.1 (99.9)	100 (100)
80 k	97.7 (99.9)	97.7 (99.9)	100 (100)
90 k	99.5 (99.9)	99.5 (99.9)	100 (100)
100 k	99.5 (100)	99.5 (100)	100 (100)
110 k	99.6 (100)	99.6 (100)	100 (100)
120 k	100 (100)	100 (100)	100 (100)

Discussion

As proposed previouly,[6–8,13] the inferred minimum gene number needed to resolve the phylogeny should be related to bootstrap threshold values to determine a correct tree, topological complexity of sampled species, and selection of source gene sets. The differences caused by the two bootstrap threshold values, 95% and 75%, were substantial in this study (Tables 1 and 2). Similar to our results, another study found that the gene number dropped from 20 to 8 when decreasing the bootstrap threshold value in a set of eight yeast species.[7] Nevertheless, the stricter criterion with a 95% threshold will provide a better estimate of the minimum gene number.

As expected, the required gene numbers for the *Primulina* phylogeny fell sharply from 175 to 75 when ignoring the most difficult triplet (Table 1). Similarly, the suggested gene number of 20 based on the eight yeast species set[7] was unnecessary for fewer taxa.[6] In fact, it is difficult to judge whether the test species set in this study is representative for other work, especially when focusing on topological complexity characterized by the number of sampled species (11) and the pairwise interspecies genetic distances (Ks: 0.027–0.064, Supplementary

Table 2. Summary of the percentages of correct trees at a bootstrap threshold of 95% or 75% (in parentheses) for all 11 *Primulina* species, the triplet and other species, through random resampling by gene from the selected 26-gene data set.

N	ALL	TRIPLET	OTHER
3	26.3 (87.1)	63.1 (96.9)	40.2 (90.1)
4	54.6 (96.3)	78.4 (99.2)	69.7 (97.1)
5	72.9 (99.5)	87.8 (99.8)	83.2 (99.7)
6	88.0 (99.9)	94.8 (100)	92.8 (99.9)
7	93.9 (100)	97.2 (100)	96.5 (100)
8	98.2 (100)	99.2 (100)	99.0 (100)
9	99.3 (100)	99.6 (100)	99.7 (100)
10	99.8 (100)	100 (100)	99.8 (100)
11	99.9 (100)	99.9 (100)	100 (100)
12	100 (100)	100 (100)	100 (100)

Table 4). Nevertheless, the required gene number of 75 when ignoring the triplet complexity in this study is still unrealistic for routine laboratory work with large-scale species sampling. Although unlinked nucleotide sites could provide more efficient phylogenetic performance than genes in terms of better data independence, as shown in this study (Table 1) and the previous studies,[7,11] most experimental data in molecular phylogenetics are still in the form of amplified gene fragment sequences.

To minimize the effect of hidden paralogs and to identify the most probable orthologs, one optimized sampling strategy is to select genes by checking single-gene trees, rather than to use all available random genes as in most previous studies.[13] In this study, we followed such a gene selection method based on single-gene tree performance[13] and found that only 8 of 26 selected genes were sufficient to resolve all 11 *Primulina* species, while 175 genes were needed if all 830 random genes were used (Tables 1 and 2). Compared with the two factors discussed above, selection of source gene sets has greater effect on the inferred minimum gene number. Based on our results, we strongly recommend the gene selection strategies based on single-gene tree performance of a few representative species with known or well-resolved phylogenies, and the subsequent use of the selected genes for phylogenetic reconstruction with larger scale species sampling.

Acknowledgments

We thank Xinhui Zou and Fumin Zhang for their help in the data analysis.

Author Contributions

Conceived and designed the experiments: BA, MK. Analyzed the data: BA. Wrote the first draft of the manuscript: BA. Contributed to the writing of the manuscript: BA, MK. Agree with manuscript results and conclusions: BA, MK. Jointly developed the structure and arguments for the paper: BA, MK. Made critical revisions and approved final version: BA, MK. Both authors reviewed and approved of the final manuscript.

Supplementary Materials

Supplementary Table 1. List of alignment lengths, proportions of variable sites, single-gene topologies, and function associations of the 830 genes.

Supplementary Table 2. List of alignment lengths, proportions of variable sites, and function associations of the selected 26 genes.

Supplementary Table 3. List of 25 topologies supported by more than three genes.

Supplementary Table 4. List of the pairwise interspecies divergence (Ks) values among the 11 *Primulina* species, estimated by Ai et al.[18]

REFERENCES

1. Felsenstein J. Phylogenies and the comparative method. *Am Nat*. 1985;125(1):1–15.
2. Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*. 2005;6(5):361–75.
3. Philippe H, Delsuc F, Brinkmann H, Lartillot N. Phylogenomics. *Annu Rev Ecol Evol Syst*. 2005;36:541–62.
4. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? *Trends Genet*. 2006;22(4):225–31.
5. Philippe H, Brinkmann H, Lavrov DV, et al. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 2011;9(3):e1000602.
6. Gatesy J, DeSalle R, Wahlberg N. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Syst Biol*. 2007;56(2):355–63.
7. Rokas A, Williams BL, King N, Carroll SB. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*. 2003;425(6960):798–804.
8. Collins TM, Fedrigo O, Naylor GJP. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenies. *Syst Biol*. 2005;54(3):493–500.
9. Rokas A, Carroll SB. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol*. 2005;22(5):1337–44.
10. Wortley AH, Rudall PJ, Harris DJ, Scotland RW. How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. *Syst Biol*. 2005;54(5):697–709.
11. Zou XH, Zhang FM, Zhang JG, et al. Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol*. 2008;9(3):R49.
12. Wen J, Xiong Z, Nie ZL, et al. Transcriptome sequences resolve deep relationships of the grape family. *PLoS One*. 2013;8(9):e74394.
13. Zeng LP, Zhang Q, Sun RR, Kong HZ, Zhang N, Ma H. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat Commun*. 2014;5:4956.
14. Wang YZ, Mao RB, Liu Y, et al. Phylogenetic reconstruction of *Chirita* and allies (*Gesneriaceae*) with taxonomic treatments. *J Syst Evol*. 2011;49(1):50–64.
15. Weber A, Middleton DJ, Forrest A, Kiew R, LIM CL, Rahman RA. Molecular systematics and remodelling of *Chirita* and associated genera (*Gesneriaceae*). *Taxon*. 2011;60(3):767–90.
16. Wen F. *Studies on Investigation and Introduction of Wild Ornamental Resources of Gesneriaceae in Guangxi (in Chinese)* [PhD dissertation]. Beijing Forestry University; 2008.
17. Kang M, Tao JJ, Wang J, et al. Adaptive and nonadaptive genome size evolution in Karst endemic flora of China. *New Phytol*. 2014;202(4):1371–81.
18. Ai B, Gao Y, Zhang XL, Tao JJ, Kang M, Huang HW. Comparative transcriptome resources of eleven *Primulina* species, a group of 'stone plants' from a biodiversity hot spot. *Mol Ecol Resour*. 2015;15(3):619–32.
19. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307–21.
20. Kubatko LS, Degnan JH. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*. 2007;56(1):17–24.
21. Liu L, Yu LL, Pearl DK, Edwards SV. Estimating species phylogenies using coalescence times among sequences. *Syst Biol*. 2009;58(5):468–77.
22. Shaw TI, Ruan Z, Glenn TC, Liu L. STRAW: species TRee analysis web server. *Nucleic Acids Res*. 2013;41:W238–41.
23. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2006;23(2):254–67.
24. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312–3.