Title: Identifying equally scoring trees in phylogenomics with incomplete data using Gentrius

Authors: O. Chernomor[1]*, C. Elgert[1], A. von Haeseler[1,2]

Affiliations:

[1]Center for Integrative Bioinformatics Vienna (CIBIV), Max Perutz Laboratories, University of Vienna and Medical University of Vienna, Vienna Bio Center (VBC); Vienna, Austria.

[2]Department of computer science, University of Vienna; Vienna Austria.

*Corresponding author. Email: o.chernomor@gmail.com

Abstract: Phylogenetic trees are routinely built from huge and yet incomplete multi-locus datasets often leading to multiple equally scoring trees under many common criteria. As typical tree inference software output only a single tree, identifying all trees with identical score challenges phylogenomics. Here, we introduce Gentrius – an efficient algorithm that tackles this problem. We showed on simulated and biological datasets that Gentrius generates millions of trees within seconds. Depending on the distribution of missing data across species and loci and the inferred phylogeny, the number of equally good trees varies tremendously. The strict consensus tree computed from them displays all the branches unaffected by the pattern of missing data. Thus, Gentrius provides an important systematic assessment of phylogenetic trees inferred from incomplete data.

One-Sentence Summary: Gentrius - the algorithm to generate a complete stand, i.e. all binary unrooted trees compatible with the same set of subtrees.

Main Text: In molecular phylogenomics one infers a tree for a group of species using genetic information from many different loci. Contemporary datasets often comprise hundreds of species and hundreds of loci, combining organisms from distant taxonomic groups, and well-studied species, whose genomes were sequenced completely, together with non-model organisms, for which only a handful of loci are available. As a result, such diverse datasets often exhibit missing data, i.e. for some species, sequences for some loci are not available, either as a consequence of species-specific gene losses/acquisitions or incompletely sequenced genomes. The availability of genetic sequences is summarised in a species per locus presence-absence matrix with 1's or 0's indicating presence or absence of a sequence, respectively (e.g., Fig. 1A). Typical percentage of zeros (i.e. missing data) in such matrices ranges from 30% to 80% (e.g. Table 1). The impact of missing data on the reliability of phylogenetic relationships is poorly understood and, importantly, not assessed routinely due to the lack of methods and efficient bioinformatic tools.

Independent of the phylogenomic approaches (e.g. supermatrix (*1*), supertrees (*2*)) to infer a binary fully-resolved phylogenetic tree from data with missing sequences, one can extract from the tree the set of induced binary (locus) subtrees (Fig.1B). Here, each induced subtree is obtained from the tree by removing species with zero entries for the corresponding locus in the presence-absence matrix. Notably, topologically distinct trees may lead to the same set of induced subtrees (*3*). The collection of all such trees, which induce (i.e. are compatible with) the same set of subtrees, is called a stand (*4*). Thus, if the objective function to evaluate the tree quality is computed from the induced subtrees independently of the complete tree, all trees from the same stand have identical score (*5*). For instance, this was shown for parsimony and likelihood (*4*, *6*) in the scope of supermatrix inference and for various criteria (*5*, *7*) used in supertree methods. Thus, missing data can directly lead to multiple equally scoring trees (*4–7*) in widely applied phylogenomic approaches. However, common phylogenetic software (e.g. IQ-TREE (*8*), RAxML (*9*), ASTRAL (*10*)) infer and output a single tree. Obviously, using only one out of many equally optimal trees may lead to overconfident if not false conclusions about evolutionary relationships. The main challenge is that generating trees from the same stand is computationally intractable (*11*) and existing methods (*6*, *12*) are limited to special cases. Other theoretical work focused on existence of multiple trees (*13*) with neither generating nor enumerating them all and, thus, not permitting any post-analysis of equally scoring trees.

To provide a universal and practically useful solution we developed Gentrius - a deterministic algorithm to generate binary unrooted trees from incomplete unrooted subtrees. For a tree inferred with any phylogenomic method and a species per locus presence-absence matrix, Gentrius generates all trees from the corresponding stand. Thus, Gentrius systematically assesses the influence of missing data on phylogenomic analysis and enhances the confidence of evolutionary conclusions one may draw from the data. When all trees from a stand are generated, one can subsequently study their topological differences employing routine phylogenetic approaches. Moreover, we also provide a handy summary and visualisation script to ease the post-analysis for non-technical experts. To foster a widespread application, Gentrius was implemented in IQ-TREE 2 (*8*). In the following we elucidate the impact of missing data on stand size, briefly describe the Gentrius algorithm, show its applicability on simulated and biological data and conclude by discussing approaches to assure the robustness of phylogenomic analysis in the presence of missing data.

Results

On missing data and stands

To illustrate the influence of missing data on phylogenomic inference, we first discuss three fictive examples (Fig. 1A) with seven species, five loci and presence-absence matrices A_1, A_2, A_3 with 29%, 29%, and 43% of missing data, respectively. For each matrix and each of the possible 945 binary unrooted seven species trees we computed induced subtrees (Fig. 1B) and partitioned the 945 trees with the same set of induced subtrees into stands (Fig. 1C). For matrix A_1 each stand contained one tree. For matrix A_2, with the same percentage of missing data as A_1, we identified 360 stands of size one and 165 stands comprising two or more trees. This indicates that stand sizes depend not only on the percentage, but also on the spread of zeros in the matrix. Finally, for matrix A_3, with the largest percentage of missing data, stand sizes varied from 17 to 59 trees. Thus, predicting the stand size from the presence-absence matrix alone is challenging and requires appropriate computational approaches, which motivated development of Gentrius.

The overview of Gentrius

For simplicity, we explain the modus operandi of Gentrius with an illustrative example (Fig. 2), while the formal description and corresponding theoretical results are provided in (*14*). To generate a stand Gentrius uses the information provided by the (induced) subtrees (Fig. 2A,B). It selects one of the subtrees as initial subtree (black subtree in Fig. 2B) and then inserts on it the missing species (3, 4, and 7) sequentially, starting with species 3. The initial subtree has nine branches and, thus, nine possibilities to insert species 3. The remaining input subtrees (red and blue) constrain the placement for insertion. The red constraint subtree and the initial subtree have species 1, 2, and 9 in common (Fig. 2C left). Further, according to the red subtree species 1 and 9 are more closely related in contrast to 2 and 3. Thus, to preserve this relationship, it is not allowed to insert species 3 on the path connecting 1 and 9 in the initial subtree, allowing only seven branches for insertion (marked by red dots in Fig. 2C left). Similarly, to agree with the blue constraint subtree, species 3 can be inserted only on five branches (marked by blue dots in Fig. 2C middle). Finally, their intersection, the three branches with red/blue dots are allowed by both constraint subtrees (Fig. 2C right). They constitute admissible branches for insertion of species 3. Inserting species 3 on them generates three new intermediate subtrees (Fig. 2D, second row). This procedure is continued with the next missing species and each intermediate subtree (Fig. 2D, third row) until all missing species are inserted generating all seven trees from the corresponding stand.

The efficient identification of admissible branches (for a formal description see (*14*) and Figs. S1-S2), as well as the choice of the initial subtree and the insertion order of missing species are crucial for the performance of Gentrius (*14*). Also note, that detecting all admissible branches is the core of Gentrius (*14*) and assures generating the stand completely. However, the number of trees on the stand can be exponential (*3*) and in such cases generating a complete stand in feasible time is not possible. Therefore, Gentrius employs a user-defined threshold, MaxStandTrees, on the maximal number of trees generated from a stand. If MaxStandTrees is reached, Gentrius stops with a partially generated stand. Moreover, the maximum number of intermediate subtrees is bounded by the threshold MaxIntermediate. This is important, since also the number of intermediate subtrees can be exponential (*11*). In the worst case MaxIntermediate might be reached even before any tree from a stand is generated. To tackle such computationally complex cases, we provide an alternative setting for the initial subtree (*14*), which comes with a limitation that Gentrius generates a stand only partially. The above cases define different difficulty levels for Gentrius: a stand is generated completely; partially with excessive number of trees (MaxStandTrees was triggered); partially,

triggering MaxIntermediate with either stand size > 0 or empty stand (i.e. complex dataset). We consider the application of Gentrius to a given data as successful, if either Gentrius generates a stand completely, or partially, thus, providing a lower bound on the stand size.

Performance on simulated data

To evaluate the feasibility of Gentrius we performed an extensive simulation study with varying data sizes, different percentages and spread of missing data. Namely, species number ranged between 20 and 700, locus number between 5 and 100, while the percentage of missing data took values 30%, 50% and 70%. To simulate different spread of missing data we developed a custom matrix simulator (*14*) and controlled the distribution of zeros across matrix via various parameters (Fig. S3, Table S1). We simulated in total 6,120 presence-absence matrices and for each matrix sampled five random trees (*15*), thereby generating 30,600 simulation-instances. For each instance we ran Gentrius with MaxStandTrees and MaxIntermediate set to 100 million (M) trees each.

Figure 3A and Figure S4A summarise simulation results. Importantly, our simulation-instances covered all possible computational cases discussed in the previous section (see also caption of Fig. 3). Notably, all runs finished in reasonable time (from milliseconds up to ~31 hours, Fig. 3A, Fig. S4A, Table S2). Moreover, the runtime depends strongly on the stand size and to a lesser extent on species and locus numbers (Figs. 3A, S4A). Instances of the highest complexity (i.e. runs stopped reaching MaxIntermediate without generating any tree from the stand) were further re-analysed with an alternative approach (i.e. alternative initial subtree (*14*)) confirming a lower bound of 100M trees for all corresponding stands. Thus, Gentrius successfully tackled all instances providing either a complete or a partial stand.

Finally, we compared the runtimes of Gentrius with Terraphast (*12*). Terraphast requires at least one species with no missing loci as input. This leaves us with 10,293 simulated-instances having stands with less than MaxStandTrees (100M) trees (*14*), to make it comparable. Overall, the runtimes of both software is very similar (Fig. S6). Notably, Gentrius tends to be faster, when the species number is larger than locus number. The biggest advantage of Gentrius, however, is that it does not impose any requirements on the input data.

Topological differences

For biological applications the most important question is how different or similar the trees of a stand are. The stand sizes vary tremendously for all parameters and percentages of missing data (Fig. 3B, Fig. S4B). To investigate topological differences of stand trees we selected all non-trivial stands (size > 1) for datasets with 20, 50 and 100 species, and stand size up to 100K trees (in total 8,440 stands). For each stand we computed the strict consensus tree (*14*) to identify conflicting branching patterns in its corresponding trees. The strict consensus tree displays branches occurring in all considered trees (here, all trees from a stand). A fully resolved binary tree with n species has $n-3$ internal branches, while the occurrence of multifurcating nodes (i.e. nodes with more than three outgoing branches) reduces the number of internal branches and, thus, tree resolution caused by the missing data. To reflect this loss of resolution for a strict consensus tree we computed the percentage of its internal branches compared to $n-3$ internal branches on fully resolved tree.

Figure 3C shows the resolutions of 2,067 strict consensus trees computed for stands with 100 species (see also Fig. S7). We observed the full range of resolutions (between 3%-99% and up to 10 multifurcating nodes), where 93% of the consensus trees showed a resolution ≥ 85%.

We also note that, importantly, the resolution cannot be predicted from the stand size. For instance, the three strict consensus trees in Figure 3D computed for stands with similar sizes (from ~20K to

~26K trees) show a resolution of 3%, 44% and 87% (Fig. 3D). Therefore, if a stand consists of more than one tree, it is crucial to consider all these trees to explore the amount of phylogenetic uncertainty due to missing data. Note, that this uncertainty should not be mixed with bootstrap support (*16*), which is, in fact, also affected by missing data (*4*).

Avoiding huge stands

Computing strict consensus tree is a straightforward approach to summarise non-trivial stands. However, it only makes sense, if a stand is generated completely. A partial stand even with millions of trees would not be representative for potentially an exponentially large stand. Even more so, since consecutive trees generated by Gentrius are more similar to each other (*14*).

Predicting stand size from a presence-absence matrix alone is only possible in special cases. One trivial case is when at least one locus has no missing data, then independent of the percentage of missing data in the matrix, each tree forms its own stand (i.e. all stands have size one). However, such complete loci are rare in large biological datasets with taxonomically diverse species. Here, we demonstrate, that in the absence of complete loci, the easiest strategy in avoiding huge stands is having as many as possible well-sampled loci and as little as possible poorly-sampled species.

We simulated a new set of presence-absence matrices by assigning the same number of missing loci per species and systematically controlling the spread of missing data across loci, and vice versa. Namely, for 100 species we generated in total 66 matrices with 10 or 30 loci and 30% to 70% of missing data (*14*). For example, Figures 4A and 4D show simulated matrices with 30 loci and with varied spread of missing data across loci (Fig. 4B, for 70% of missing data) and species (Fig. 4E, for 50% of missing data) respectively (see also Figs. S8, S9). Using the same set of 50 random trees we calculated their stand sizes for each matrix (in total 330 simulation-instances).

The more well-sampled loci (i.e. with small amount of missing data) there were, the smaller the stand sizes were (Fig. 4A-C). While the stand sizes and their variation for different trees increased with the number of poorly-sampled species (Fig. 4D-F, Figs. S8, S9), with the most drastic increase in the presence of minimally covered species (i.e. represented by a single locus, Fig. 4D-F bottom, matrices 9-11). Note, that minimally covered species frequently occur in practice (e.g. Table 1). Thus, given the huge variation among stand sizes for different trees (Fig. 4C,F, Fig. S8D, S9D), if the actual stand is huge or not, depends on the inferred phylogenetic tree.

Application to biological data

Finally, we applied Gentrius to 10 published alignments (*17–26*) from various taxonomic groups (Table 1) with 180 to 767 species and 3 to 79 loci. The percentage of missing data was between 34% and 80% (Table 1, Fig. S10). For each alignment we inferred a maximum likelihood (ML) tree with IQ-TREE 2 (*8*) assuming a partition model (*27–29*), for more details see (*14*). Next, for each ML tree and the corresponding presence-absence matrix we generated its stand with Gentrius. In accordance with our simulation results, the resulting stand sizes were highly variable (Table 1), ranging from one (for Frogs) to more than 100M trees (for Grasses, Salamander and Sedges).

To clarify whether these stand sizes are typical for the corresponding presence-absence matrices, we additionally sampled 100 random trees for each alignment and generated their stands. For nine datasets the majority of stands have more than 100M trees (Table S4). The only exception was Frogs dataset with stand sizes varying between one to seven trees (Table S4). This observation corroborates our simulations, that it is difficult to predict the stand size from a presence-absence matrix alone.

For six datasets with complete stands (for ML trees) we computed the strict consensus trees (*14*). Despite generally high resolution of the strict consensus trees (from 99% for Snakes to 78% for Carnivora, Fig. 5, and Fig. S11-16), each multifurcation should be investigated to understand, if it affects the conclusions of the studies. For instance, if a dataset contains species from diverse taxonomic groups and each multifurcating node only affects species from the same genus (e.g. see subtrees in Figs. 5B, D), then one can still make statements about the evolution of genera, even if the number of multifurcations is high. However, if the evolutionary question concerns species from the same genus/family (e.g. in Fig. 5C small subtree shows an example of poor resolution among species from Drosophila genus), then no conclusions can be drawn. The biggest issue is, of course, when multifurcating nodes involve species from different taxonomic groups (e.g. Fig. 5A,D,E,F).

In general, large number of multifurcating nodes and high node degree are indicators of analysis strongly affected by missing data. The above findings anew demonstrate that for a robust phylogenomic analysis taking into account missing data systematically is very important. Here, we advocate including Gentrius into a phylogenomic workflow to increase the confidence of findings.

Discussion

The main result of the paper is the development of Gentrius, an algorithm to assist phylogenomic analyses of large contemporary datasets with missing sequences. An exhaustive evaluation of Gentrius feasibility on simulated and biological datasets evidenced that Gentrius can deal with large datasets, generating millions of trees within reasonable time (Fig. 3A, Fig. S7A, Table 1, Table S2). Currently, Gentrius is the only algorithm available that generates complete stands for unrooted trees without any constraints on the structure and type of input data.

Importantly, Gentrius has direct practical application to the avalanche of phylogenomic data we are facing. When a phylogenetic tree is inferred by common phylogenomic methods (e.g. supermatrix, supertree), Gentrius can generate its corresponding stand. Since all trees from the stand have identical score under many commonly applied objective functions (*4*–*7*) generating stands should be routine in phylogenetic workflow.

Here, we have demonstrated, that stand sizes are strongly affected by missing data and can vary substantially for the same dataset. Moreover, trees from the same stand can be topologically very diverse. Therefore, if the stand size is larger than one, it is important to investigate topological differences of stand trees by constructing a strict consensus tree and subsequently investigating its unresolved parts. If the evolutionary relationships of interest are not affected by multifurcations, then the strict consensus tree can be used to substantiate evolutionary hypotheses. Otherwise, the results have to be considered with care. Ideally, the input data should be amended and reanalysed.

When dealing with missing data, the following considerations should be taking into account. If at least one locus contains data for all species, then stand size equals one for all possible trees. Thus, including a single locus with no missing data in the analysis avoids multiple equally optimal trees and, thus, no unresolved evolutionary relationship purely due to missing data. However, such data also require additional care and have to be investigated on potential artefacts, such as existence of many trees with non-equal, yet very similar scores, termed phylogenetic islands (*30*). Another alternative for likelihood methods is to use different partition schemes or more restrictive partition models (*4*), since then trees from the same stand have different likelihoods. However, this approach also does not eliminate the problem of trees with very similar likelihoods (*31*). If missing data is unavoidable, the best strategy is to combine a large number of nearly complete loci and avoiding species with a lot of missing data.

We would like to point out that Gentrius complements, but does not replace standard methods to assess clade confidence, such as bootstrap (*16*). Importantly, the information provided about a stand and its corresponding strict consensus tree is valuable, since bootstrap is also affected by missing data and can misleadingly report high bootstrap scores (*4*). One of the possibilities to combine bootstrap and Gentrius is generating stands for each bootstrap tree, constructing strict consensus trees for bootstrap stands and using them to compute clade support (see also (*4*)). However, further research is needed to find best strategies to assess statistical confidence in the presence of missing data.

Apart from direct practical application, Gentrius can be also employed to enhance our theoretical understanding of tree spaces (Fig. 1C). For instance, it may help studying how trees from the same stand and also between different stands are connected via common topological rearrangements (e.g. Nearest Neighbour Interchange, see also (*32*)), which has a great potential for development of better tree search strategies in the presence of missing data. Note, that for supertree methods distinct trees can have identical score even without missing data (*5*, *33*). Thus, also trees from multiple stands can have identical score. Hence, understanding the connectivity between stands is crucial in identifying all stands with identical score in supertree methods.

References and Notes

1. A. de Queiroz, J. Gatesy, The supermatrix approach to systematics. *Trends Ecol. Evol.* **22** (2007), , doi:10.1016/j.tree.2006.10.002.
2. S. Mirarab, L. Nakhleh, T. Warnow, Multispecies Coalescent: Theory and Applications in Phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* **52** (2021), doi:10.1146/annurev-ecolsys-012121-095340.
3. S. Böcker, Exponentially many supertrees. *Appl. Math. Lett.* **15** (2002), doi:10.1016/S0893-9659(02)00054-X.
4. M. J. Sanderson, M. M. Mcmahon, A. Stamatakis, D. J. Zwickl, M. Steel, Impacts of terraces on phylogenetic inference. *Syst. Biol.* **64** (2015), doi:10.1093/sysbio/syv024.
5. M. J. Sanderson, M. M. McMahon, M. Steel, *bioRxiv* [Preprint] (2020), doi:10.1101/2020.04.17.047092.
6. M. J. Sanderson, M. M. McMahon, M. Steel, Terraces in phylogenetic tree space. *Science*. **333** (2011), doi:10.1126/science.1206357.
7. M. Habib, A. H. Rahman, M. S. Bayzid, *bioRxiv* [Preprint] (2022), doi:10.1101/2022.11.21.517454.
8. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, R. Lanfear, E. Teeling, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37** (2020), doi:10.1093/molbev/msaa015.
9. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*. **35** (2019), doi:10.1093/bioinformatics/btz305.
10. S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, T. Warnow, "ASTRAL: Genome-scale coalescent-based species tree estimation" in *Bioinformatics*

(2014), vol. 30.

11. M. Bordewich, C. Semple, J. Talbot, Counting consistent phylogenetic trees is #P-complete. *Adv. Appl. Math.* **33** (2004), doi:10.1016/j.aam.2003.08.006.

12. R. Biczok, P. Bozsoky, P. Eisenmann, J. Ernst, T. Ribizel, F. Scholz, A. Trefzer, F. Weber, M. Hamann, A. Stamatakis, Two C++ libraries for counting trees on a phylogenetic terrace. *Bioinformatics*. **34** (2018), doi:10.1093/bioinformatics/bty384.

13. S. Böcker, D. Bryant, A. W. M. Dress, M. A. Steel, Algorithmic Aspects of Tree Amalgamation. *J. Algorithms*. **37** (2000), doi:10.1006/jagm.2000.1116.

14. Materials and methods are available as supplementary materials.

15. E. F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. Appl. Probab.* **3** (1971), doi:10.2307/1426329.

16. J. Felsenstein, Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution (N. Y).* **39** (1985), doi:10.2307/2408678.

17. K. Van Der Linde, D. Houle, G. S. Spicer, S. J. Steppan, A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet. Res. (Camb).* **92** (2010), doi:10.1017/S001667231000008X.

18. K. Nyakatura, O. R. P. Bininda-Emonds, Updating the evolutionary history of Carnivora (Mammalia): A new species-level supertree complete with divergence time estimates. *BMC Biol.* **10** (2012), doi:10.1186/1741-7007-10-12.

19. L. Y. Echevarría, I. De la Riva, P. J. Venegas, F. J. M. Rojas-Runjaic, I. R. Dias, S. Castroviejo-Fisher, Total evidence and sensitivity phylogenetic analyses of egg-brooding frogs (Anura: Hemiphractidae). *Cladistics*. **37** (2021), doi:10.1111/cla.12447.

20. P. H. Fabre, A. Rodrigues, E. J. P. Douzery, Patterns of macroevolution among primates inferred from a supermatrix of mitochondrial and nuclear DNA. *Mol. Phylogenet. Evol.* **53** (2009), doi:10.1016/j.ympev.2009.08.004.

21. Y. Bouchenak-Khelladi, N. Salamin, V. Savolainen, F. Forest, M. van der Bank, M. W. Chase, T. R. Hodkinson, Large multi-gene phylogenetic trees of the grasses (Poaceae): Progress towards complete tribal and generic level sampling. *Mol. Phylogenet. Evol.* **47** (2008), doi:10.1016/j.ympev.2008.01.035.

22. M. S. Springer, R. W. Meredith, J. Gatesy, C. A. Emerling, J. Park, D. L. Rabosky, T. Stadler, C. Steiner, O. A. Ryder, J. E. Janečka, C. A. Fisher, W. J. Murphy, Macroevolutionary Dynamics and Historical Biogeography of Primate Diversification Inferred from a Species Supermatrix. *PLoS One*. **7** (2012), doi:10.1371/journal.pone.0049521.

23. A. F. Jaramillo, I. De La Riva, J. M. Guayasamin, J. C. Chaparro, G. Gagliardi-Urrutia, R. C. Gutiérrez, I. Brcko, C. Vilà, S. Castroviejo-Fisher, Vastly underestimated species richness of Amazonian salamanders (Plethodontidae: Bolitoglossa) and implications about plethodontid diversification. *Mol. Phylogenet. Evol.* **149** (2020), doi:10.1016/j.ympev.2020.106841.

24. A. Stamatakis, N. Alachiotis, Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics*. **26** (2010),

doi:10.1093/bioinformatics/btq205.

25. C. E. Hinchliff, E. H. Roalson, Using supermatrices for phylogenetic inquiry: An example using the sedges. *Syst. Biol.* **62** (2013), doi:10.1093/sysbio/sys088.

26. R. A. Pyron, F. T. Burbrink, G. R. Colli, A. N. M. de Oca, L. J. Vitt, C. A. Kuczynski, J. J. Wiens, The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. *Mol. Phylogenet. Evol.* **58** (2011), doi:10.1016/j.ympev.2010.11.006.

27. Z. Yang, Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42** (1996), doi:10.1007/BF02352289.

28. O. Chernomor, A. Von Haeseler, B. Q. Minh, Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst. Biol.* **65** (2016), doi:10.1093/sysbio/syw037.

29. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods*. **14** (2017), doi:10.1038/nmeth.4285.

30. A. S. Silva, M. Wilkinson, On Defining and Finding Islands of Trees and Mitigating Large Island Bias. *Syst. Biol.* **70** (2021), doi:10.1093/sysbio/syab015.

31. P. Breitling, A. Stamatakis, O. Chernomor, B. Bettisworth, L. Reszczynski, Empirical Analysis of Phylogenetic Quasi-Terraces. *bioRxiv* [Preprint] (2019), doi:10.1101/810309.

32. S. Mark, J. C. McLeod, M. Steel, A navigation system for tree space. *J. Graph Algorithms Appl.* **20** (2016), doi:10.7155/jgaa.00392.

33. I. T. Farah, M. Islam, K. T. Zinat, A. H. Rahman, S. Bayzid, Species Tree Estimation from Gene Trees by Minimizing Deep Coalescence and Maximizing Quartet Consistency: A Comparative Study and the Presence of Pseudo Species Tree Terraces. *Syst. Biol.* **70** (2021), doi:10.1093/sysbio/syab026.

34. A. D. Gordon, Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labeled leaves. *J. Classif.* **3** (1986), doi:10.1007/BF01894195.

35. M. Constantinescu, D. Sankoff, An efficient algorithm for supertrees. *J. Classif.* **12**, 101–112 (1995).

Acknowledgments: We thank Heiko Schmidt and Clement Bader for helpful comments on strict consensus trees.

Funding:

Austrian Science Fund - FWF grant number I-4686 (OC, AvH)

Author contributions:

Conceptualization: OC, CE, AvH

Methodology: OC

Software: OC

Formal Analysis: OC

Investigation: OC

Visualization: OC

Funding acquisition: OC, AvH

Writing – original draft: OC

Writing – review & editing: OC, CE, AvH

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The implementation of Gentrius is available in IQ-TREE 2 (since version 2.2, GitHub: https://github.com/iqtree/iqtree2, Website: http://www.iqtree.org/). The custom matrix simulator is available at GitHub repository: https://github.com/OlgaChern/MatrixSimulator. All simulated and biological data used in this manuscript as well as auxiliary scripts are available at GitHub repository: https://github.com/OlgaChern/Gentrius_2023SM. Note, that all empirical alignments were published previously and are also available via original papers.

Supplementary Materials

Materials and Methods

Figs. S1 to S16

Tables S1 to S4

References (*34, 35*)

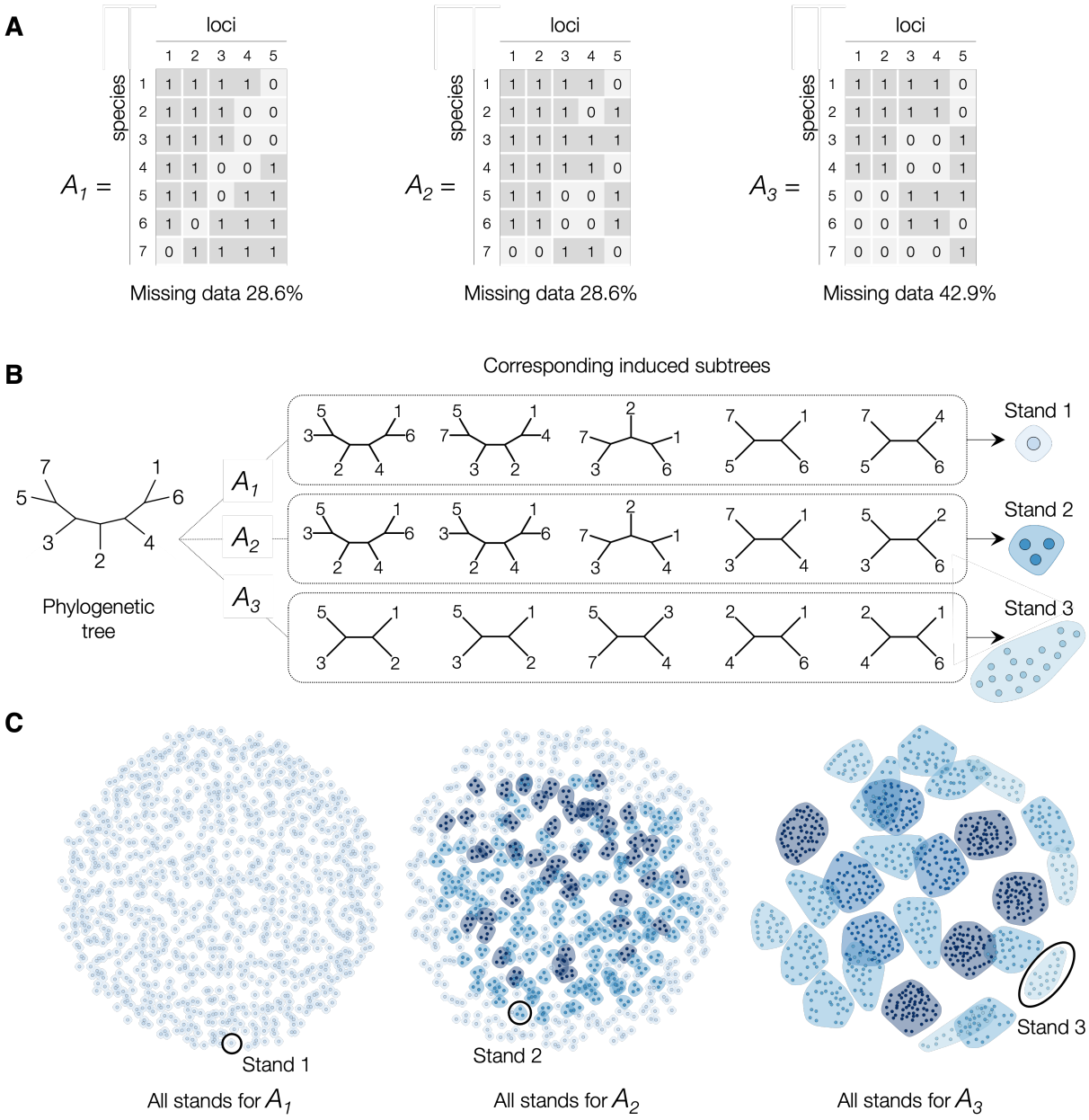

Fig. 1. Influence of missing data on phylogenomic inference. (**A**) Examples of species per locus presence-absence matrices. Here, "1" stands for presence and "0" for absence of sequence for corresponding species and locus. (**B**) A random tree with its induced subtrees and corresponding stands for each matrix. Each dot in a stand is a different tree, i.e. stands 1, 2, and 3 consist of 1, 3 and 17 trees, respectively. (**C**) For each considered matrix all 945 trees for seven species were grouped based on their induced subtrees into stands.

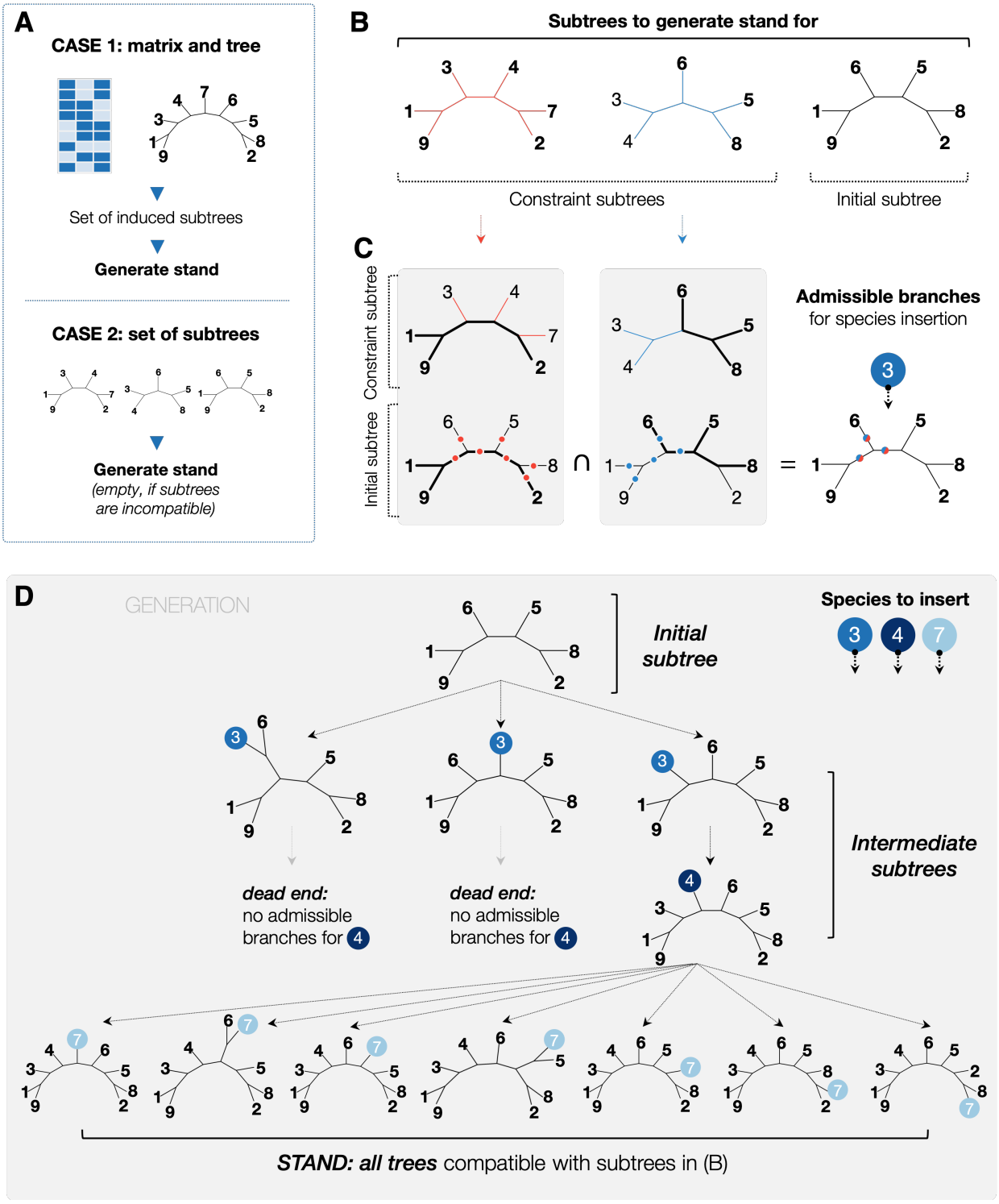

Fig. 2. The overview of Gentrius. (**A**) Gentrius generates stands from a set of binary unrooted subtrees. They can be obtained either as induced subtrees of a tree and a presence-absence matrix (CASE 1, anticipated practical application in a typical phylogenomic workflow), or as a set of subtrees inferred separately (CASE 2). (**B**) The set of subtrees to generate a stand for. The black subtree is selected as initial subtree. The remaining subtrees (red and blue) serve as constraints.

(**C**) Identification of admissible branches to insert species 3. First, we detect branches, which are allowed by each constraint subtree separately (left, middle) and then compute their intersection (right). Bold black branches connect species in common for corresponding pairs of initial and constraint subtree. Red, blue and red-blue dots mark branches on initial subtree allowed by red, blue and both constraint subtrees, respectively. (**D**) Generation of a stand. Species 3, 4 and 7 are inserted sequentially. Each insertion generates an intermediate subtree. After species 3 is inserted, Gentrius identifies admissible branches for species 4. If there are no admissible branches, a dead end is reached, i.e. the intermediate subtree cannot be extended without constraint violation, and Gentrius continues with the next intermediate subtree. After species 4 is inserted, we identify the admissible branches for species 7. By iterating over all admissible branches and inserting all missing species Gentrius generates a complete stand, i.e. all trees compatible with an input set of subtrees in (B).

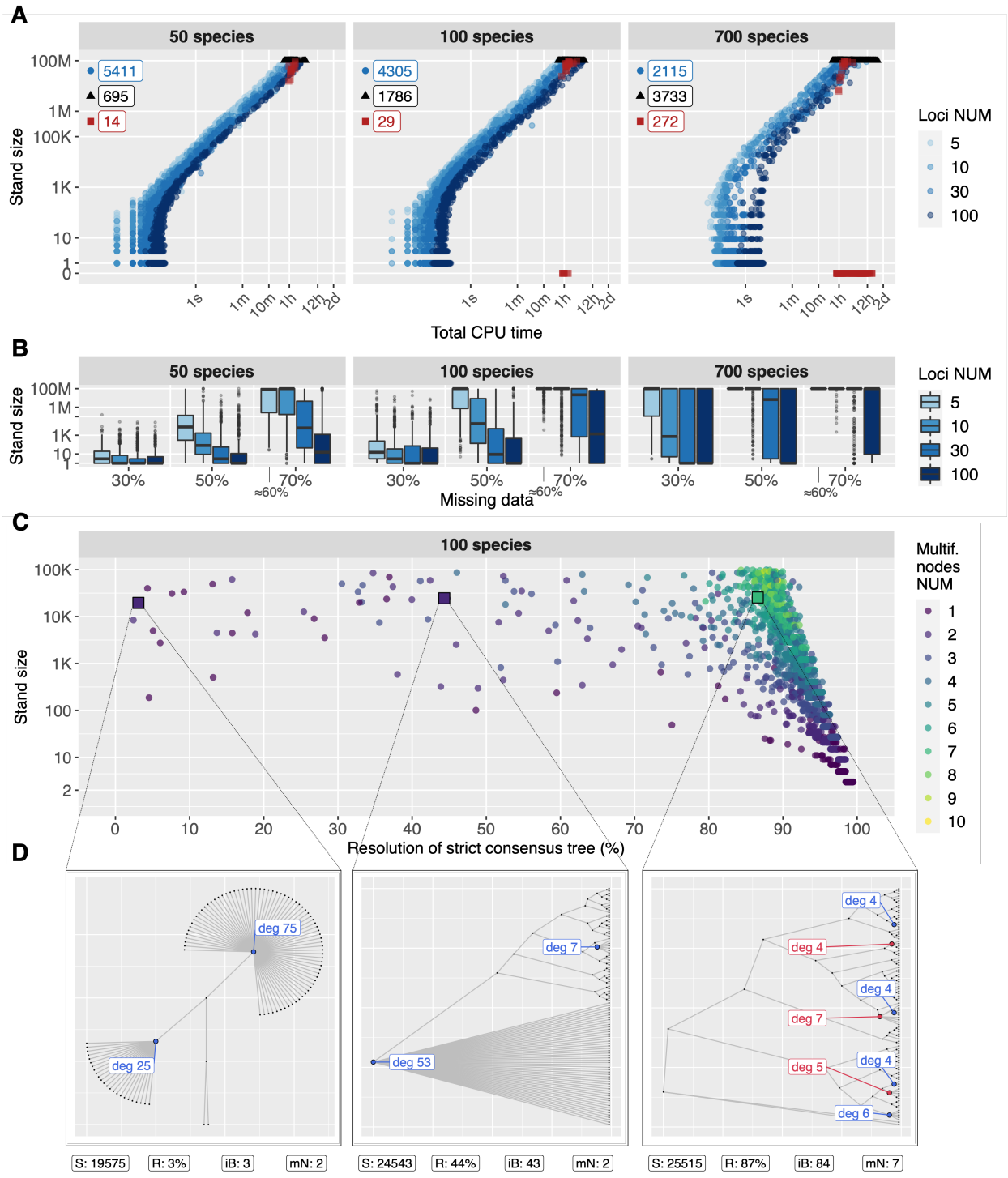

Fig. 3. Results of simulation study. (**A**) Stand size vs. total CPU runtime. Shapes correspond to different computational cases: circles denote complete stands, triangles and squares denote incomplete stands, i.e. triggering MaxStandTrees or MaxIntermediate, respectively. The numbers (top left) specify how often each case occurred. The squares at zero represent complex cases, where MaxIntermediate was triggered, but no tree from a stand was generated yet. For all datasets with red squares, we subsequently ran Gentrius with an alternative initial subtree (*14*), confirming a lower bound of 100M trees on the stand size. (**B**) Influence of the percentage of missing data on

stand sizes. Note, that for 5 loci with the constrains imposed (*14*), the resulting presence-absence matrices had around 60% of missing data instead of 70% as required (Fig. S5). (**C**) Resolution of strict consensus trees for 100 species. Each point corresponds to one stand. All stands have up to 100K trees. The resolution refers to the percentage of internal branches on a strict consensus tree compared to the $n-3$ internal branches on a fully resolved tree with n species. (**D**) Selected examples of strict consensus trees with different resolution. Multifurcating nodes are marked on the tree by coloured circles and degree of the node. Blue nodes are incident to exactly one internal branch and red nodes are incident to at least two internal branches. Large node degrees represent highly unresolved parts of the tree due to missing data. S – stand size, R – tree resolution, iB – number of internal branches, mN – number of multifurcating nodes.

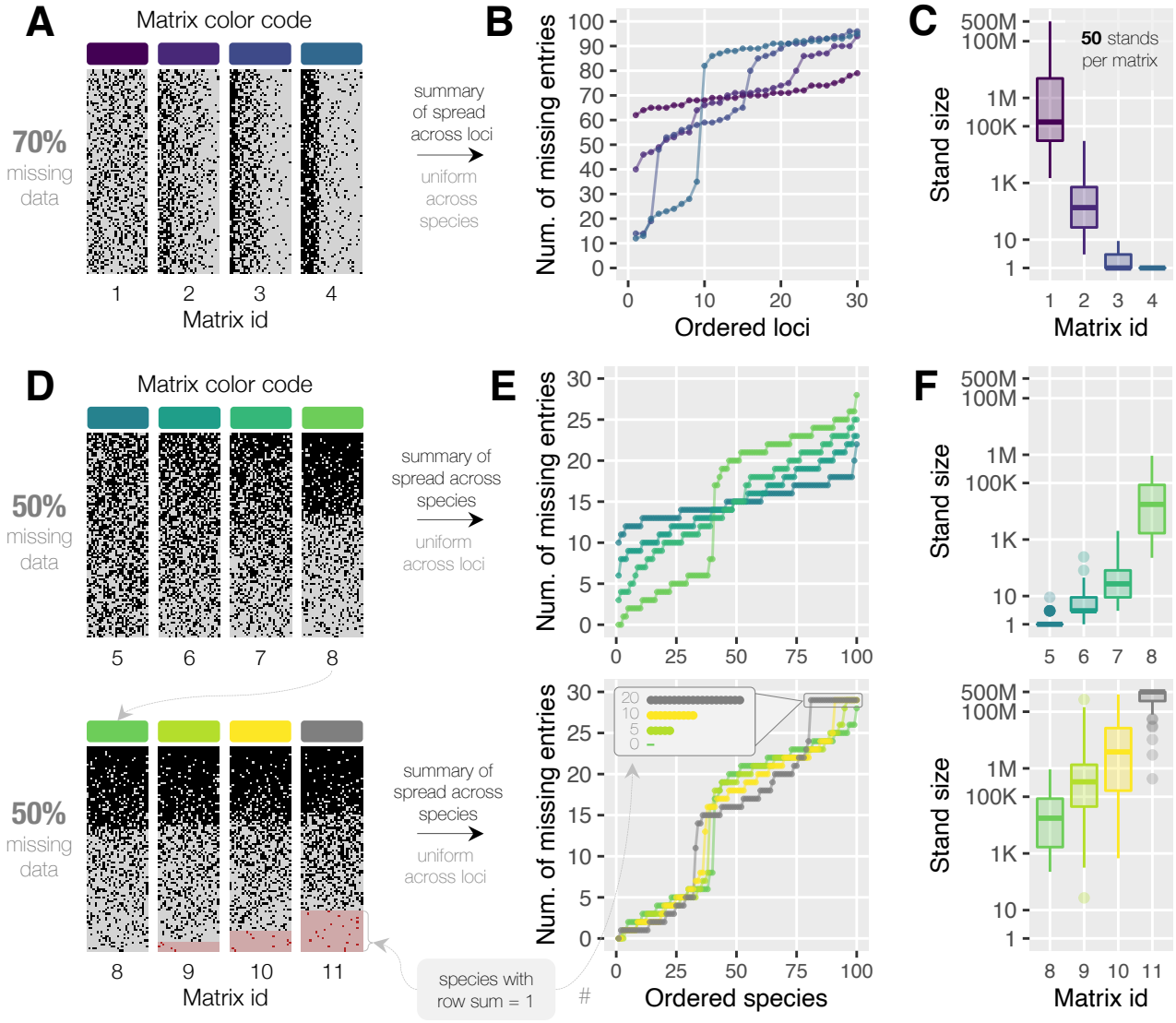

Fig. 4. Dependence of stand size on missing data across species and loci. The simulated presence-absence matrices have 100 species and 30 loci. Black and grey dots indicate presence or absence ("1" or "0") of sequence for corresponding species and locus. For stand size computation the 50 trees were fixed across all matrices. (**A**) Matrices with 70% of missing data. Each species has exactly 21 missing entries, while spread of missing data across loci varies among matrices. (**B**) For each matrix from (A) the loci are ordered by increasing missing data. (**C**) Stand sizes for matrices from (A). (**D**) Matrices with 50% of missing data. Each locus has exactly 50 missing entries, while spread of missing data across species varies among matrices. Species represented by a single locus are marked by pink row with a red dot indicating presence of sequence. Such minimally covered species are only present in matrices 9, 10, and 11 which have 5, 10 and 20 minimally covered species respectively. (**E**) For each matrix from (D) the species are ordered by increasing missing data. (**F**) Stand sizes for matrices from (D).

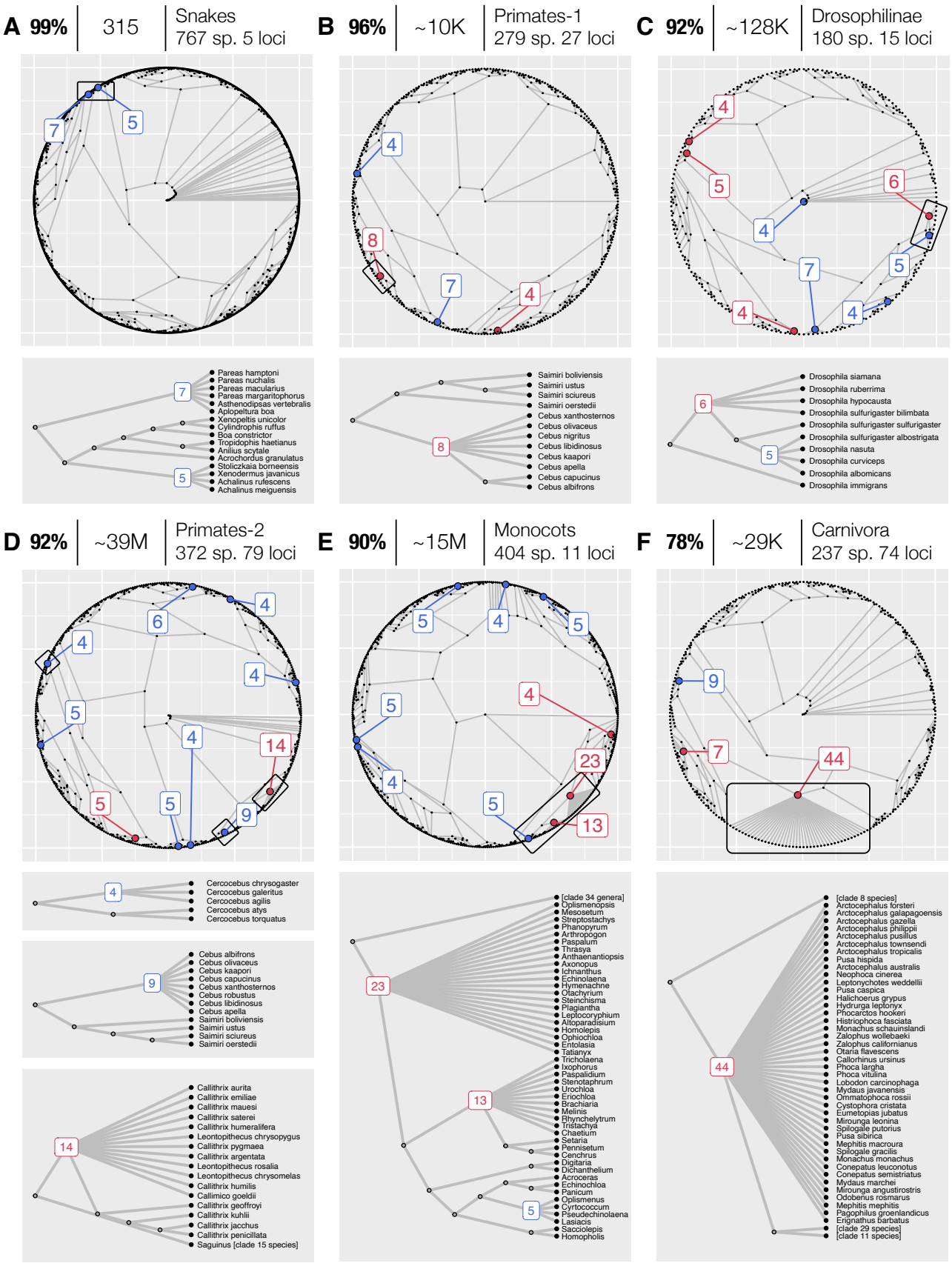

Fig. 5. Strict consensus trees for (complete) stands generated for ML trees for six biological datasets. The subplots are ordered by the decreasing resolution (in %) of strict consensus trees (left), followed by stand size (middle), species group, numbers of species and loci (right). The numbers in blue and red indicate the degree of multifurcation, where blue nodes are incident to exactly one internal branch and red nodes are incident to at least two internal branches. Black boxes on the circular trees denote the subtrees displayed below each subplot. These subtrees were chosen to illustrate different cases of uncertainties (different types of nodes (red/blue), node degree, species involved). (**A**) Strict consensus tree for Snakes with 48% of missing data. (**B**) Primates-1, 74%. (**C**) Drosophilinae, 60%. (**D**) Primates-2, 63%. (**E**) Monocots, 63%. (**F**) Carnivora, 75%.

Table 1. Summary for biological datasets. Column "Max locus coverage" for a given dataset indicates information about the largest number of species in one locus (i.e. the largest column sum in a presence-absence matrix). Column "Sp. min coverage" contains the number of species, represented by a single locus, thus have minimal coverage (i.e. row sum equals to one). The lower bound on stand size for partially generated stands is 100M trees.

ID	Species group	Sp.	Loci	Missing data	Max locus coverage	Sp. min coverage	Stand size	Total CPU time	Pub.
D1	Drosophilinae	180	15	60%	144 (80%)	6	127,575	5s	(*17*)
D2	Carnivora	237	74	75%	202 (85%)	22	29,133	6s	(*18*)
D3	Frogs	267	20	59%	266 (99.6%)	3	1	0.04s	(*19*)
D4	Primates-1	279	27	74%	202 (72%)	34	9,963	9s	(*20*)
D5	Grasses	298	3	34%	221 (74%)	113	>100 M	26m:24s	(*21*)
D6	Primates-2	372	79	63%	276 (74%)	57	39,355,875	1h:20m:3s	(*22*)
D7	Salamander	381	13	60%	314 (82%)	77	>100 M	1h:22m:40s	(*23*)
D8	Monocots	404	11	63%	290 (72%)	45	14,529,375	10m:53s	(*24*)
D9	Sedges	435	18	80%	329 (76%)	93	>100 M	7h:39m:22s	(*25*)
D10	Snakes	767	5	48%	716 (93%)	59	315	0.064s	(*26*)