Gentrius: Generating Trees Compatible With a Set of Unrooted Subtrees and its Application to Phylogenetic Terraces

Olga Chernomor [1,4,*] Christiane Elgert [1] Arndt von Haeseler [1,2,3]

[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

[3]Ludwig Boltzmann Institute for Network Medicine, University of Vienna, Vienna, Austria

[4]Present address: Global Computational Biology and Digital Sciences, Boehringer Ingelheim RCV GmbH, Vienna, Austria

***Corresponding author:** E-mail: o.chernomor@gmail.com.

Associate editor: Andrey Rzhetsky

Abstract

For a set of binary unrooted subtrees generating all binary unrooted trees compatible with them, i.e. generating their stand, is one of the classical problems in phylogenetics. Here, we introduce Gentrius—an efficient algorithm to tackle this task. The algorithm has a direct application in practice. Namely, Gentrius generates phylogenetic terraces—topologically distinct, equally scoring trees due to missing data. Despite stand generation being computationally intractable, we showed on simulated and biological datasets that Gentrius generates stands with millions of trees in feasible time. We exemplify that depending on the distribution of missing data across species and loci and the inferred phylogeny, the number of equally optimal terrace trees varies tremendously. The strict consensus tree computed from them displays all the branches unaffected by the pattern of missing data. Thus, by solving the problem of stand generation, in practice Gentrius provides an important systematic assessment of phylogenetic trees inferred from incomplete data. Furthermore, Gentrius can aid theoretical research by fostering understanding of tree space structure imposed by missing data.

Key words: **compatible subtrees, phylogenetic stands, phylogenetic terraces, phylogenomics, missing data.**

Introduction

For a set of (binary unrooted) subtrees with a total of n leaves, the collection of all (binary unrooted) trees with n leaves each and compatible with the subtrees is called a stand (also span) (Sanderson et al. 2015). For compatible subtrees (i.e. there exists at least one tree that displays all subtrees [Gordon 1986]), the stand may have exponentially many trees (Böcker 2002). While if the subtrees are incompatible (i.e. subtrees with contradicting branching patterns), their stand is empty. For unrooted binary subtrees, constructing a stand (or determining that it is empty) is computationally intractable (Bordewich et al. 2004). Current methods (Sanderson et al. 2011; Biczok et al. 2018), adaptations from the algorithm for rooted trees (Constantinescu and Sankoff 1995), are limited to a special case—when at least one species occurs in all subtrees. Other theoretical work focused on the existence of multiple trees (Böcker et al. 2000) on the stand with neither generating nor enumerating them all and thus not allowing any post analysis. Hence, while generating a stand is one of the classical problems in phylogenetics, to the best of our knowledge, there is no algorithm to generate a stand for a set of unrooted binary trees without any constraint on the subtrees.

To close this gap and to provide a practically useful solution, we developed Gentrius—a deterministic branch-and-bound type algorithm to generate complete stands from binary unrooted subtrees. Our motivation stems from an important practical application of stands to phylogenetic terraces. Generally speaking, terraces are collections of equally scoring trees, which arise in phylogenomics due to incomplete data (e.g. missing sequences). The concept has been introduced back in 2011 by Sanderson et al. (2011). It has been discussed in theoretical papers (Chernomor et al. 2015; Sanderson et al. 2015; Mark et al. 2016; Dobrin et al. 2018; Breitling et al. 2019) and used for improving and speeding up tree search routines (Stamatakis and Alachiotis 2010; Chernomor et al. 2016).

Methods

Received: July 08, 2024. **Revised:** September 30, 2024. **Accepted:** October 11, 2024

However, the lack of an algorithm to generate a complete terrace without any constraint on the data has prohibited routine analysis of terraces in empirical studies. By providing Gentrius, we bridge theory and practice for phylogenetic terraces. To foster a widespread application, Gentrius was implemented in IQ-TREE 2 (Minh et al. 2020).

We would like to point out the various potential applications of Gentrius (e.g. generating phylogenetic terraces to see the impact of missing data on the landscape of equally scoring trees; characterizing tree space structure in the presence of missing data; identifying if multiple incomplete subtrees are compatible or not, etc.) each requiring different approaches for simulations, different datasets, and post analyses. As we are primarily interested in generation of phylogenetic terraces, to fully elucidate and ease the application for practitioners, the remainder of the paper including simulations, illustration on biological examples and various insights are discussed in regards to phylogenetic terraces. All the theoretical details about tree compatibility and stands are provided in the Supplementary Material online.

Phylogenetic Terraces as a Practical Need and Motivation for Gentrius Development

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 summarized 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). The percentage of zeros (i.e. missing data) in such matrices sometimes reaches even 80% (e.g. Table 1). Missing data affects phylogenetic methods in various ways (e.g. Sanderson et al. 2011; Simmons 2012; Sanderson et al. 2015a; Hosner et al. 2016; Xi et al. 2016; Molloy and Warnow 2018; Nute et al. 2018).

One of the crucial impacts of missing data on tree inference is phylogenetic terraces—topologically distinct trees, which due to combinatorial structure and the considered scoring function, have equal scores (Sanderson et al. 2011). Terraces arise for parsimony and likelihood (Sanderson et al. 2011, 2015), as well as quartet consistency (Habib et al. 2022) (see also Sanderson et al. 2020 for reconciliation methods). Thus, they affect two popular tree inference approaches: supermatrix and supertree. Terraces influence the efficiency of tree searches (Stamatakis and Alachiotis 2010; Chernomor et al. 2015, 2016; Breitling et al. 2019) and introduce biases to bootstrap (Sanderson et al. 2015). Finally, if a best-found tree is on a large terrace (Sanderson et al. 2011; Dobrin et al. 2018), ignoring this fact may lead to overconfident conclusions about evolutionary relationships (Sanderson et al. 2011; Dobrin et al. 2018; Habib et al. 2022). Thus, the knowledge of all trees on a terrace will improve the reliability of a phylogenetic analysis.

Note, that even for complete datasets multiple equally scoring trees can occur (e.g. well-known for parsimony and supertree methods; Mickevich 1978; Vachaspati and Warnow 2018; Goloboff et al. 2022). In such cases, the researcher has little influence on the scoring landscape. In contrast, phylogenetic terraces are caused by missing data, which can be modified (e.g. removing or adding species/loci to the dataset) to minimize its negative effect on phylogenetic inference.

How do we determine the size of a terrace and trees on it (which can be exponentially many (Böcker 2002))? Trees from a terrace have a well-defined combinatorial structure: they all share (i.e. compatible with) the same set of induced per locus subtrees. Here, each induced subtree is obtained from a tree (any terrace tree) by removing species with zeros for the corresponding locus in the presence–absence matrix (Fig. 1b). Thus, generating terrace trees is equivalent to generating a stand (Sanderson et al. 2015) for the induced subtrees. Not to be confused with gene trees that have no influence on which trees belong to one terrace (for more details see Sanderson et al. (2011) and Habib et al. 2022).

Applied to phylogenetic terraces, for a tree inferred with any phylogenomic method (e.g. IQ-TREE 2 (Minh et al. 2020), RAxML (Kozlov et al. 2019), ASTRAL (Mirarab et al. 2014) etc.) and a species per locus presence–absence matrix, Gentrius generates all trees (if feasible) on the corresponding stand (thus, terraces for parsimony, likelihood, quartet consistency score). As terraces are equally scoring trees due to missing data, Gentrius systematically assesses the influence of missing data on their number and topologies and enhances the confidence of evolutionary conclusions one may draw from the data. When all trees from a stand/terrace are generated, one can subsequently study their (topological) differences employing routine phylogenetic approaches. Moreover, we provide a summary visualization script to ease the post analysis. In the following, we elucidate the impact of missing data on stand size, briefly describe the Gentrius algorithm, test its feasibility on simulated data, and show application to empirical data identifying phylogenetic terraces equipped with likelihood scoring function. We conclude with a discussion on approaches to ensure the robustness of phylogenomic analysis when faced with missing sequences.

Results

On Missing Data and Stands

To illustrate the impact of missing data on phylogenomic inference, we discuss three fictive examples (Fig. 1a) with seven species, five loci and presence–absence matrices

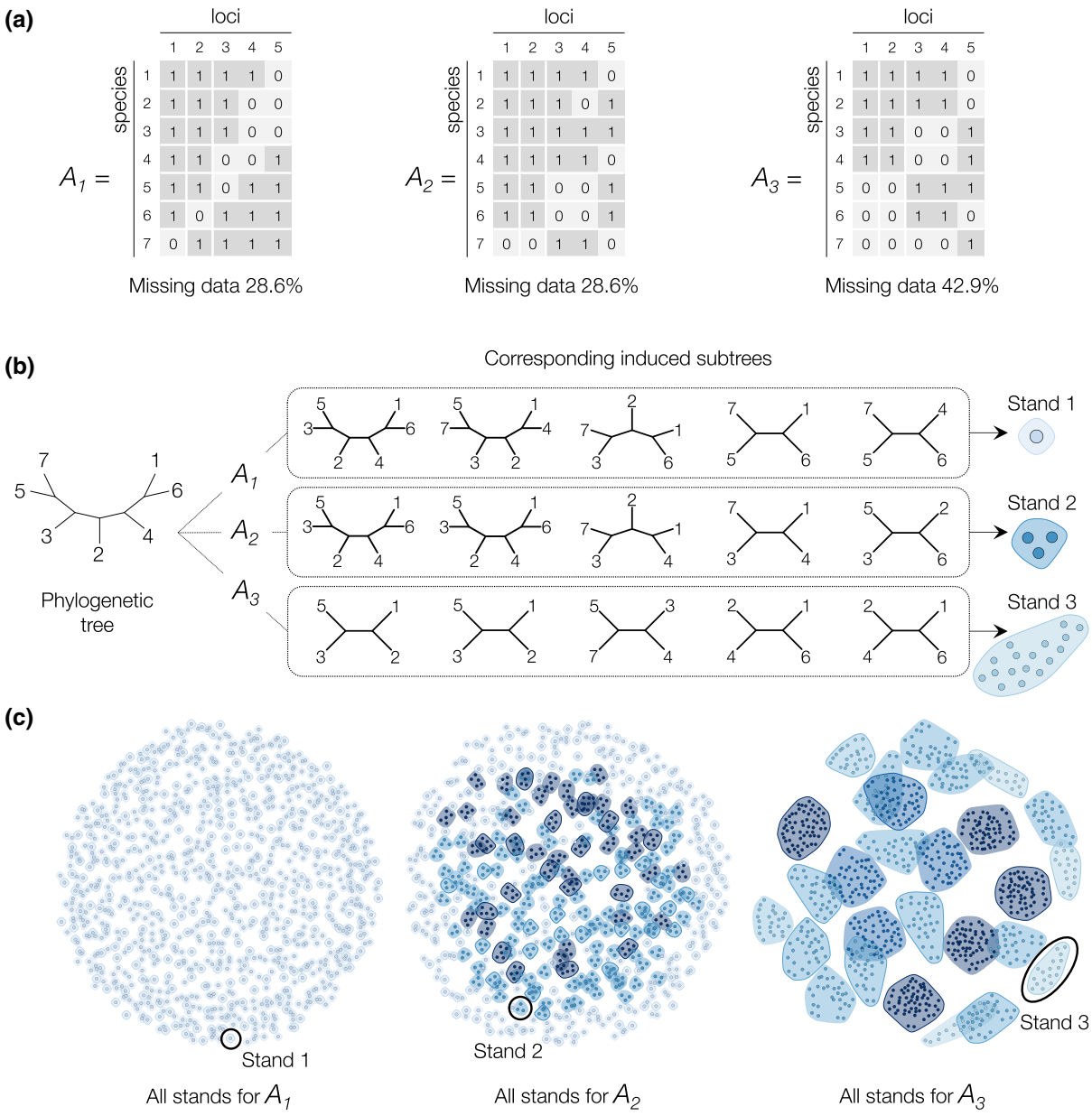

Fig. 1. Influence of missing data on phylogenomic inference. a) Examples of species per locus presence–absence matrices. Here, "1" stands for presence and "0" for absence of sequence for corresponding species and locus. b) A random tree with its induced subtrees and corresponding stands for each matrix. 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.

A_1, A_2, A_3 with 29%, 29%, and 43% missing data, respectively. For each presence–absence matrix and for each of the 945 binary unrooted seven-species trees, we computed the 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. Matrix A_2, with 29% missing data like A_1, leads to 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 presence–absence matrix. Finally, for matrix A_3, with the largest percentage of missing data (43%), 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.

Gentrius in a Nutshell

For simplicity, we explain the modus operandi of Gentrius with an illustrative example (Fig. 2) (for a detailed description see supplementary note S1, Supplementary Material online). To generate a stand Gentrius uses the information provided by the (induced) subtrees (Fig. 2a and b). It selects one of the subtrees as initial subtree (black subtree

Table 1 Summary for biological datasets

ID	Species group	Sp.	Loci	Missing data (%)	Max locus coverage	Sp. min coverage	Stand size	Total CPU time
D1	Snails	62	1027	37	58 (94%)	0	1	2 s
D2	Land plants	103	852	34	100 (97%)	0	1	2 s
D3	Drosophilinae	180	15	60	144 (80%)	6	127,575	5 s
D4	Carnivora	237	74	75	202 (85%)	22	29,133	6 s
D5	Frogs	267	20	59	266 (99.6%)	3	1	0.04 s
D6	Primates-1	279	27	74	202 (72%)	34	9,963	9 s
D7	Grasses	298	3	34	221 (74%)	113	>100 M	26 m:24 s
D8	Primates-2	372	79	63	276 (74%)	57	39,355,875	1 h:20 m:3 s
D9	Salamander	381	13	60	314 (82%)	77	>100 M	1 h:22 m:40 s
D10	Monocots	404	11	63	290 (72%)	45	14,529,375	10 m:53 s
D11	Sedges	435	18	80	329 (76%)	93	>100 M	7 h:39 m:22 s
D12	Snakes	767	5	48	716 (93%)	59	315	0.064 s

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 100 M (Million) trees.

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 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 (red dots in Fig. 2c left). Similarly, to agree with the blue subtree, species 3 can be inserted only on five branches (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.

In the above example, the subtrees are induced subtrees, i.e. there is at least one tree displaying them all, thus they are compatible and their stand is nonempty. For noninduced subtrees, first Gentrius tests if each pair of initial and constraint subtrees agrees on the relationship of their common species (see supplementary note S2, Supplementary Material online). If this is not satisfied, the subtrees are incompatible and Gentrius terminates with an empty stand. Otherwise, Gentrius proceeds with the species insertion. Note, that pairwise compatibility does not assure compatibility of a set of subtrees (Semple and Steel 2003). If during the insertion step one of the species cannot be inserted in any of the intermediate trees (i.e. there are no admissible branches), this indicates that the subtrees are incompatible and Gentrius stops.

To detect admissible branches, we developed edge maps (supplementary note S2, Supplementary Material online, supplementary figs. S1 to S3, Supplementary Material online) and a data structure (supplementary note S3, Supplementary Material online) that considers all subtrees simultaneously. This ensures the detection of all admissible branches and, hence, guarantees that all generated trees belong to the corresponding stand and if Gentrius stops before hitting a user stopping threshold (see next paragraph) then it is guaranteed to be a complete stand. The efficient identification of admissible branches and their updating during species insertion (and deletion), as well as the choice of the initial subtree and the insertion order of missing species are crucial for Gentrius runtime. Current settings were selected for the best performance among various alternatives. For instance, instead of defining and fixing the insertion order at the initial step, Gentrius employs dynamic selection of species to be inserted next. Namely, each time a new species has to be chosen, Gentrius decides on the species with the least number of admissible branches at the current step (supplementary note S1, Supplementary Material online).

While various techniques were used to make Gentrius efficient (supplementary note S1.3, Supplementary Material online), the number of trees on the stand can be exponential (Böcker 2002). In such cases, generating a complete stand in feasible time is not possible. Therefore, Gentrius employs a user-defined threshold, MaxStandTrees (supplementary note S3, Supplementary Material online), on the maximal number of trees generated from a stand. If MaxStandTrees is reached, Gentrius outputs a partial stand. Moreover, the maximum number of intermediate subtrees is bounded by the threshold MaxIntermediate, since also the number of intermediate subtrees can be exponential (Bordewich et al. 2004). In the worst case, MaxIntermediate might be reached before any tree from a stand is generated. To tackle such computationally complex cases, we provide an alternative setting for the initial subtree (supplementary note S1.2, Supplementary Material online), which comes with the limitation that Gentrius generates a stand only partially. The above cases define different difficulty levels for Gentrius: a stand is generated completely; partially with

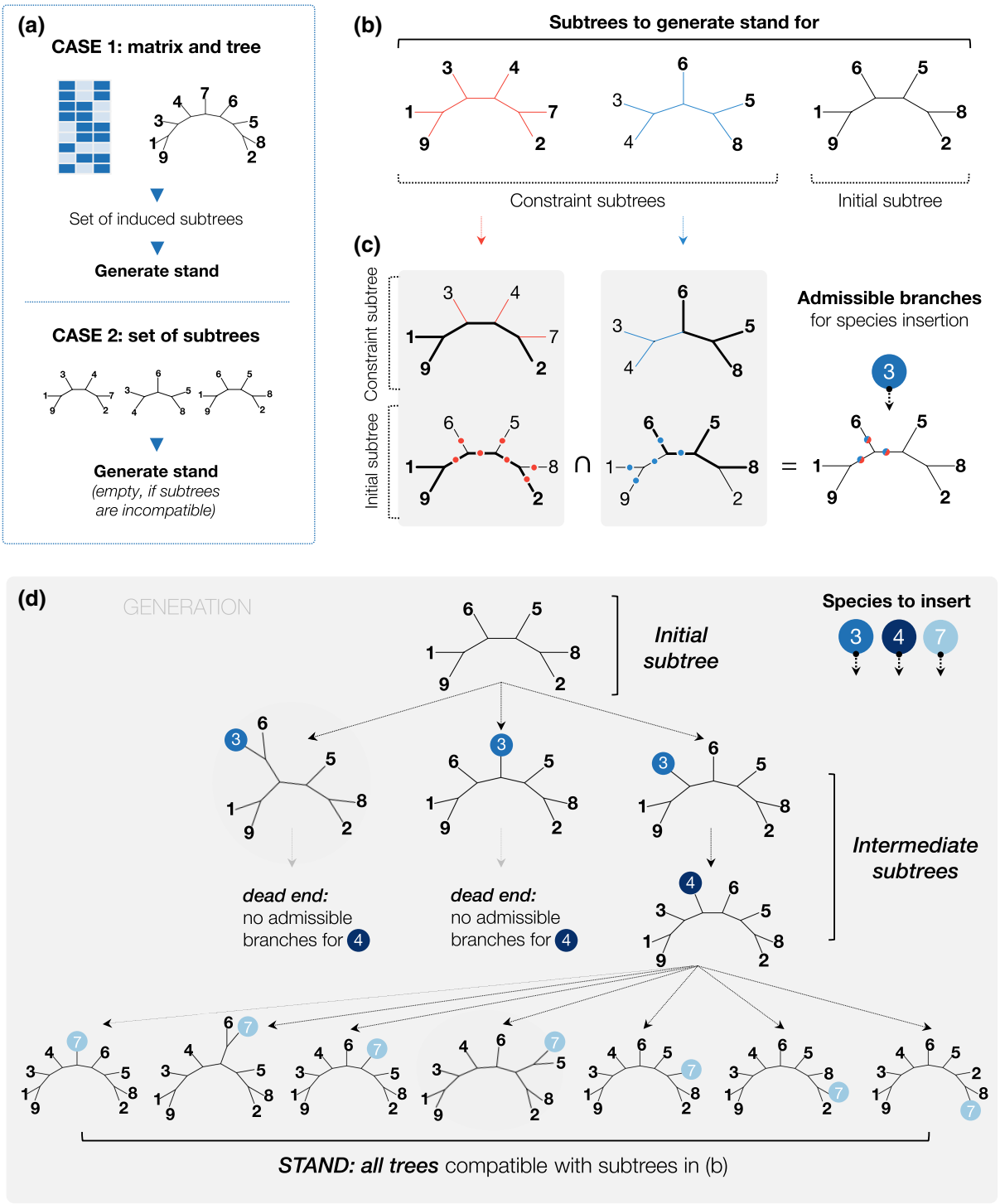

Fig. 2. The overview of Gentrius exemplified on a small dataset. a) Gentrius generates stands from a set of binary unrooted subtrees. They can be obtained either as induced subtrees of a tree and a presence–absence matrix (CASE 1, anticipated practical application in a typical phylogenomic workflow), or as a set of subtrees inferred separately (CASE 2). b) The set of subtrees to generate a stand for. One of the subtrees is selected as initial subtree. The remaining subtrees 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 branches connect species in common for corresponding pairs of initial and constraint subtree. The dots mark branches on initial subtree allowed by corresponding and both constraint subtrees. 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).

tree number equal to MaxStandTrees; 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 (in the context of phylogenetic terraces), if either Gentrius generates a stand completely, or partially, thus, providing a lower bound on the stand size. In the current implementation, when both MaxStandTrees and MaxIntermediate are "turned off," the maximum stand size, which Gentrius can generate is bound by the maximal value of unsigned integer (e.g. 4,294,967,295 trees for 32-bits, see supplementary note S3, Supplementary Material online).

Performance on Simulated Data

To evaluate the feasibility of Gentrius we performed an extensive simulation study with varying species number (20 to 700), locus number (5 to 100), different percentages (30%, 50%, 70%), and coverage patterns of missing data. To simulate different coverage patterns, we developed a custom matrix simulator (supplementary note S4, Supplementary Material online) and controlled the distribution of zeros across the matrix via various parameters (Materials and Methods). We simulated in total 6,120 presence–absence matrices and for each matrix sampled five random trees (Harding 1971), thereby generating 30,600 pairs of presence–absence matrix and tree. For each instance, we ran Gentrius with MaxStandTrees and MaxIntermediate set to 100 million (M) trees each.

Figure 3 and supplementary fig. S6, Supplementary Material online summarize key aspects of the simulation. Importantly, our simulation-instances covered all complexity levels of the input (i.e. feasible stand size, excessively large stand size, excessive number of intermediate trees with partial stand or empty stand). Despite varying complexity all runs finished in reasonable time (from milliseconds up to ~31 h, Fig. 3a, supplementary fig. S6a, Supplementary Material online; supplementary table S2, Supplementary Material online). Contrary to classical *tree inference* methods (i.e. supertree and supermatrix), where the runtime depends strongly on the species and locus numbers, Gentrius' runtime predominantly depends on the stand size and, thus, on the complexity level of the input data (Fig. 3a, supplementary fig. S6a, Supplementary Material online). This point will be further illustrated with empirical datasets.

If for a simulated dataset a run stopped hitting MaxIntermediate without generating any tree from the stand, we re-run Gentrius employing the alternative initial subtree approach (supplementary note S1 and S2, Supplementary Material online). In all such cases, the alternative approach generated a partial stand with 100 M trees (hitting MaxStandTrees). Thus, Gentrius successfully tackled all instances providing either a complete or a partial stand.

Finally, on 10,293 simulated instances we compared the runtimes of Gentrius with Terraphast (Biczok et al. 2018), which generates stands only if the data contain at least one reference species (i.e. a species with no missing data). Overall, the runtimes of both software are very similar (supplementary fig. S7, Supplementary Material online). 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 (e.g. a reference species).

Topological Differences

In a phylogenomic analysis, all trees belonging to the phylogenetic terrace of the highest scoring tree are plausible hypotheses about the relationship of organisms (i.e. equally optimal trees). Thus, generating terraces is important to understand the influence of missing data on phylogenetic relationships and to elucidate reliable branching patterns in the multiple best-found trees.

To investigate the similarity of trees from the same terrace, we analyzed trees from 8,440 stands (computed for simulated datasets above) with stand sizes between 2 and 100 K and with 20 to 100 species. For each stand, we computed the strict consensus tree to identify common branching patterns. Here, the strict consensus tree displays branches occurring in 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. To measure the loss of resolution (here, caused by missing data) for each strict consensus tree, we computed the percentage of its internal branches compared to $n-3$ internal branches on a fully resolved tree.

Figure 3c shows the relation between resolutions of the strict consensus trees and stand size for 2,067 stands for datasets with 100 species each (see also supplementary fig. S8, Supplementary Material online). We observed the full range of resolutions (between 3% and 99% and up to 10 multifurcating nodes), where 93% of the consensus trees showed a resolution ≥ 85%. Importantly, we observed that the resolution cannot be predicted from the stand size. For instance, the strict consensus trees for stands with similar sizes (from ~20 K to ~26 K trees) show a resolution of 3%, 44% and 87% (Fig. 3d, see also supplementary fig. S8, Supplementary Material online).

Therefore, for nontrivial phylogenetic terraces (i.e. stand size > 1), it is crucial to know all its trees to precisely compute the amount and location in a tree of phylogenetic uncertainty due to missing data. Note, that this uncertainty should not be confused with bootstrap support (Felsenstein 1985), which is, in fact, also affected by missing data (Sanderson et al. 2015).

Note that, while a strict consensus tree is a good measure for the congruence/compatibility of trees from a terrace/stand, it should only be used if a stand is generated completely. A partial stand even with millions of trees would not be representative for potentially an exponentially large stand. In particular, since consecutive trees generated by Gentrius are more similar to each other. Thus, if

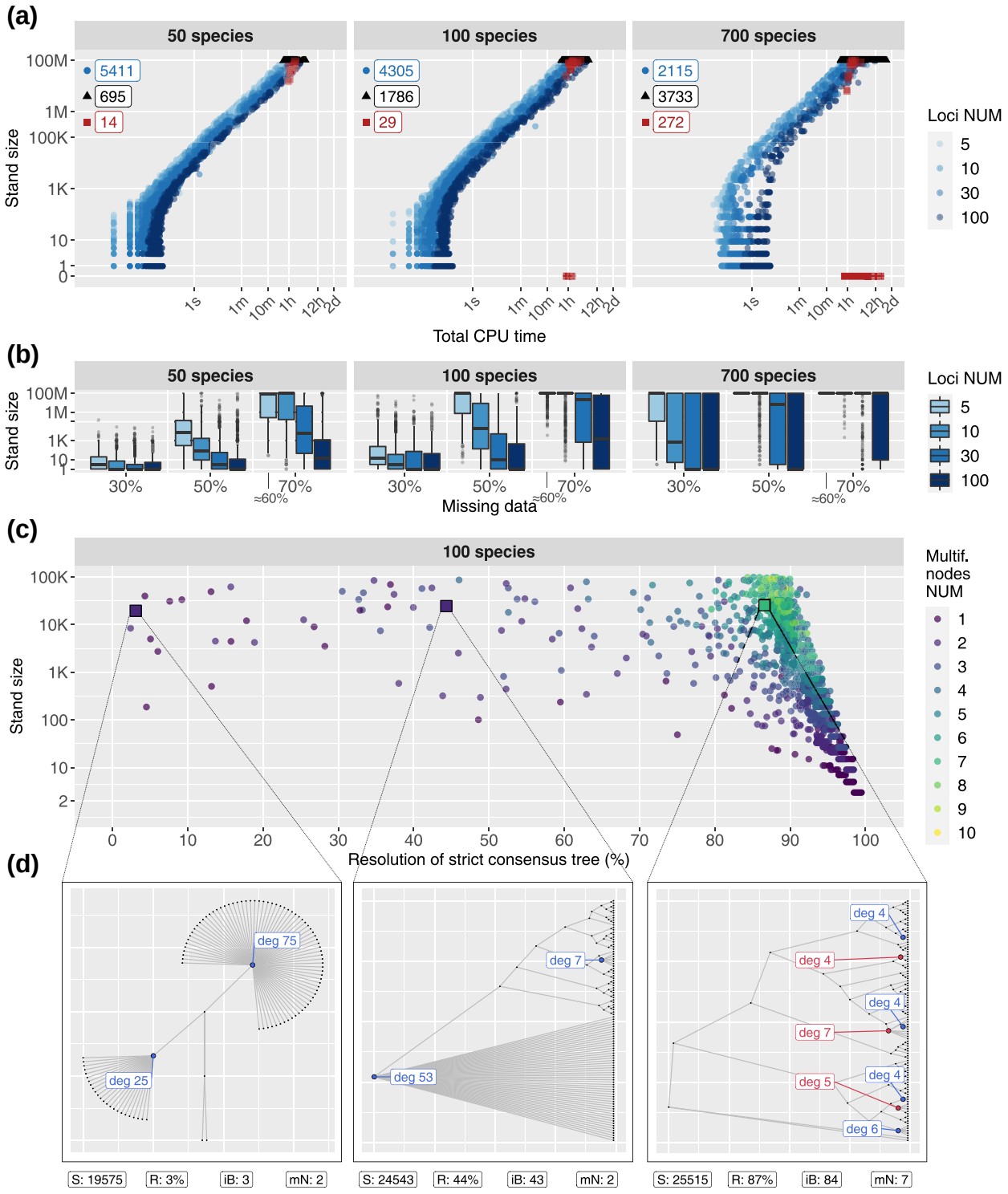

Fig. 3. Results of simulation study. a) Stand size versus total CPU runtime. Shapes correspond to different computational cases: circles denote complete stands, triangles, and squares denote incomplete stands, i.e. triggering MaxStandTrees or MaxIntermediate, respectively. The numbers (top left) specify 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 squares, we subsequently ran Gentrius with an alternative initial subtree (Materials and Methods), confirming a lower bound of 100 M trees on the stand size. b) Influence of the percentage of missing data on stand sizes. Note, that for five loci with the constrains imposed (Materials and Methods), the resulting presence–absence matrices had around 60% of missing data instead of 70% as required (supplementary fig. S5, Supplementary Material online). c) Resolution of strict consensus trees for 100 species. Each point corresponds to one stand. All stands have up to 100 K 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 colored 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.

one is interested in the consensus tree and Gentrius outputs a partial stand, it is advised to increase the stopping thresholds to increase the chances of generating a complete stand.

Practical Advice to Reduce Stand Size

Knowing which presence–absence matrices lead to huge stands would help designing appropriate datasets before performing computationally intensive phylogeny inference and then evaluating the impact of missing data and terraces. However, predicting stand size even for a specific tree is generally not feasible. One trivial case is when at least one locus has no missing data, then stands for all trees have size one (i.e. only trivial terraces). However, complete loci are rare in empirical datasets with large number of taxonomically diverse species.

Figure 3b shows that stand size depends on the % of missing data, species, and locus numbers (see also supplementary fig. S6b, Supplementary Material online). However, simply increasing the number of loci does not necessarily reduce the stand size (e.g. results for 700 species in Fig. 3b). As exemplified in Fig. 1, stand sizes also depend on the spread of missing data across the presence–absence matrix.

To get a better understanding, we simulated presence–absence matrices with different coverage patterns, but fixed the amount of missing data across either species or loci to uniform.

Figure 4a–c shows the effect of presence–absence matrices if their locus coverage pattern changes, while the amount of missing data across species is uniform. Figure 4b shows the changing locus coverage pattern from a "uniform" distribution of missing data (horizontal line) to different bi-modal distributions (step-type functions). The amount of missing data across species is fixed (i.e. uniform coverage). Not surprisingly having a few densely sampled loci reduces the stand size.

Figure 4d–f shows the complementary analysis. Now, the species coverage pattern changes and the amount of per locus missing data is fixed. Note, that some species here are only represented by a single locus. Now, the number of trees on a stand increases as the deviation of the missing data from the uniform distribution increases. Thus, species with sparse data increase the stand size. For other sizes of matrices, see also supplementary figs. S9 and S10, Supplementary Material online.

Thus, stand/terrace sizes are likely to be small, when the missing data per species are uniformly distributed and if at least some loci are very well sampled. For real data, it can be helpful to exclude sparsely sampled species and if possible, sequence/collect additional loci to increase the coverage for some loci. For the latter case Gentrius helps to systematically evaluate, which changes in the presence–absence matrix will lead to a size reduction of the stand/terrace without too much extra work or without excluding too many data.

Application to Biological Data and Phylogenetic Terraces With Likelihood

The direct practical application of Gentrius is identifying phylogenetic terraces—equally scoring trees due to missing data. For likelihood, terraces exist for partition model with independent branch length estimation per locus (Sanderson et al. 2011, 2015), i.e. edge unlinked partition model (Yang 1996). For other partition models (e.g. with equal or proportional branches across loci; Yang 1996), stand trees have different, yet very similar likelihoods (Breitling et al. 2019). Thus, Gentrius can be helpful in quickly generating alternative candidate trees with similar scores, instead of using multiple runs with expensive likelihood computations.

To illustrate the application of Gentrius to terraces, we analyzed phylogenetic terraces for likelihood inference with the edge unlinked partition model (Sanderson et al. 2011, 2015). Namely, we generated a terrace for maximum likelihood (ML) tree, thus, identifying all equally optimal ML trees for each dataset. The main questions are how many such optimal trees per dataset exist (i.e. terrace size) and how similar are their topologies. This provides a quantitative analysis, missing data hampers phylogenetic analysis via multiple equally scoring trees and, thus, if the quality of the input dataset is sufficient to make reliable phylogenetic statements.

To this end, we analyzed 12 published alignments (Table 1, supplementary fig. S11, Supplementary Material online) from various taxonomic groups. For each alignment, its corresponding presence–absence matrix and inferred ML tree (Materials and Methods) we generated the stand/terrace with Gentrius.

In accordance with our simulations, the resulting stand sizes were highly variable (Table 1), ranging from one (for snails, land plants, and frogs) to more than 100 M trees (for grasses, salamander, and sedges). We would like to stress that the results reinforce the conclusions drawn from simulations (Fig. 3a), that the computing times (Table 1) are by and large determined by the stand size, rather than locus number. To elucidate the influence of species number on the runtime, in preliminary tests, we computed stands for presence–absence matrix obtained from the Angiosperms dataset (Zanne et al. 2014) with 32,223 species and 7 loci together with random trees (failing to infer the ML tree in feasible time; therefore, this dataset was excluded from empirical examples in the present manuscript). In these tests, Gentrius generated millions of stand trees within 2 min. Hence, large species and locus numbers do not necessarily lead to long runtimes. Gentrius' runtime depends on the dataset at question and the stand complexity (i.e. feasible stand size, excessively large stand size, etc.).

In the simulations, we have observed that stand sizes for the same presence–absence matrix vary a lot for different trees (Fig. 4c and f, supplementary figs. S9d and S10d, Supplementary Material online). To confirm this on empirical presence–absence matrices, we additionally sampled 100 random trees for each alignment and generated their

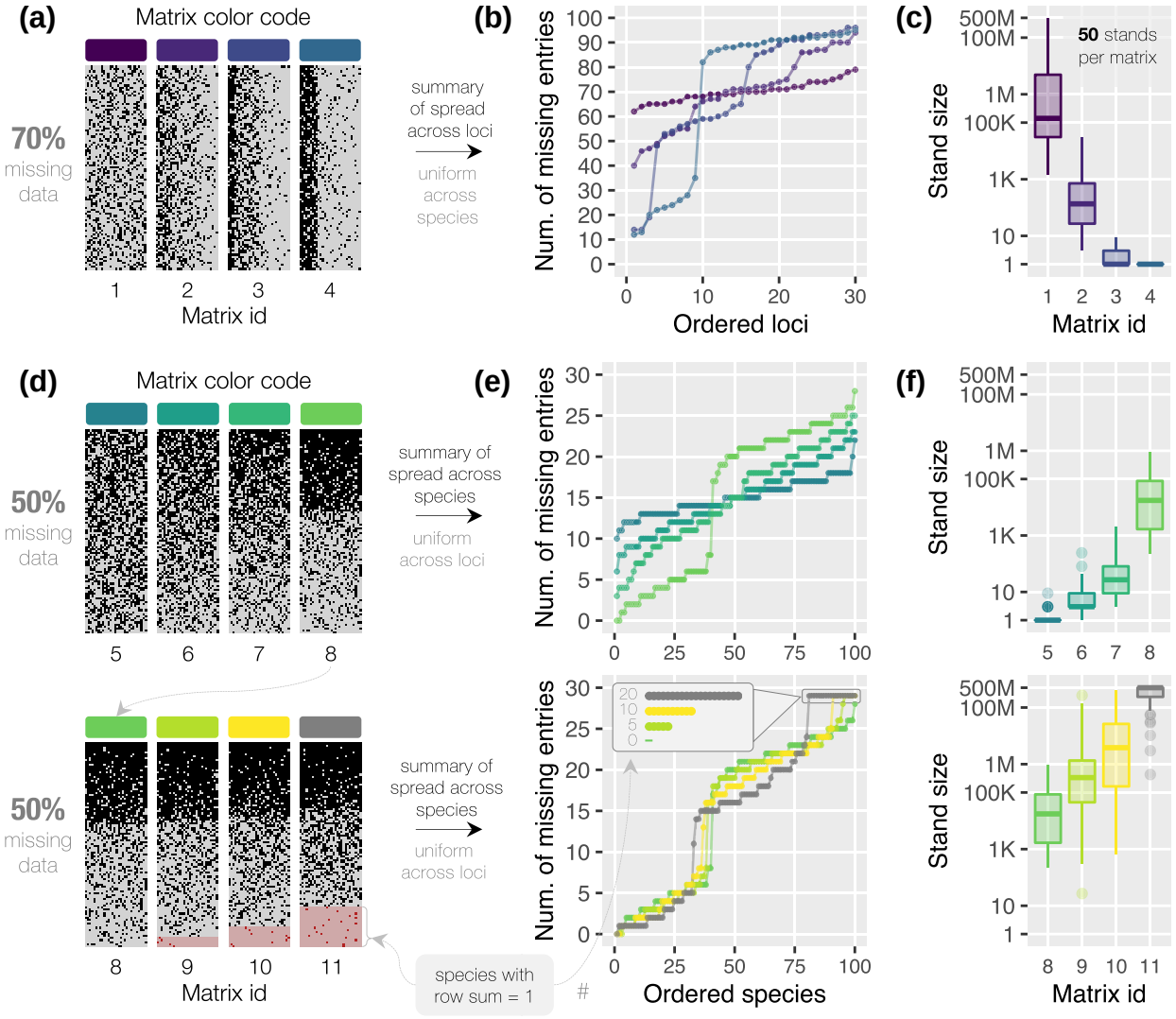

Fig. 4. Dependence of stand size on missing data across species and loci. The simulated presence–absence matrices have 100 species and 30 loci. Black and gray dots indicate presence or absence ("1" or "0") of sequence for corresponding species and locus. For stand size computation, the 50 trees were 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).

terraces. For snails and land plants, all terraces have size one (supplementary table S4, Supplementary Material online). This is expected, since both datasets contain many loci with few missing sequences (supplementary fig. S11, Supplementary Material online). For the frogs dataset, terrace sizes varied between one and seven trees. Finally, for nine datasets the majority of terraces have more than 100 M trees. This observation corroborates our simulations, that it is difficult to predict the stand/terrace size from a presence–absence matrix alone, but that the computation of complete and partial stands is possible and computationally not very expensive.

For six datasets with complete terraces (for ML trees) and with sizes from 315 to ~40 M trees, we computed their strict consensus trees. Despite generally high resolution of the strict consensus trees (from 99% for snakes to 78% for carnivora, Fig. 5, and supplementary figs. S12 to S17, Supplementary Material online), for a careful interpretation of the results, the characteristics of each multifurcation, as indicators of uncertain branching patterns due to missing data, are important.

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 Fig. 5b and d), then one can still make reliable 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, like in

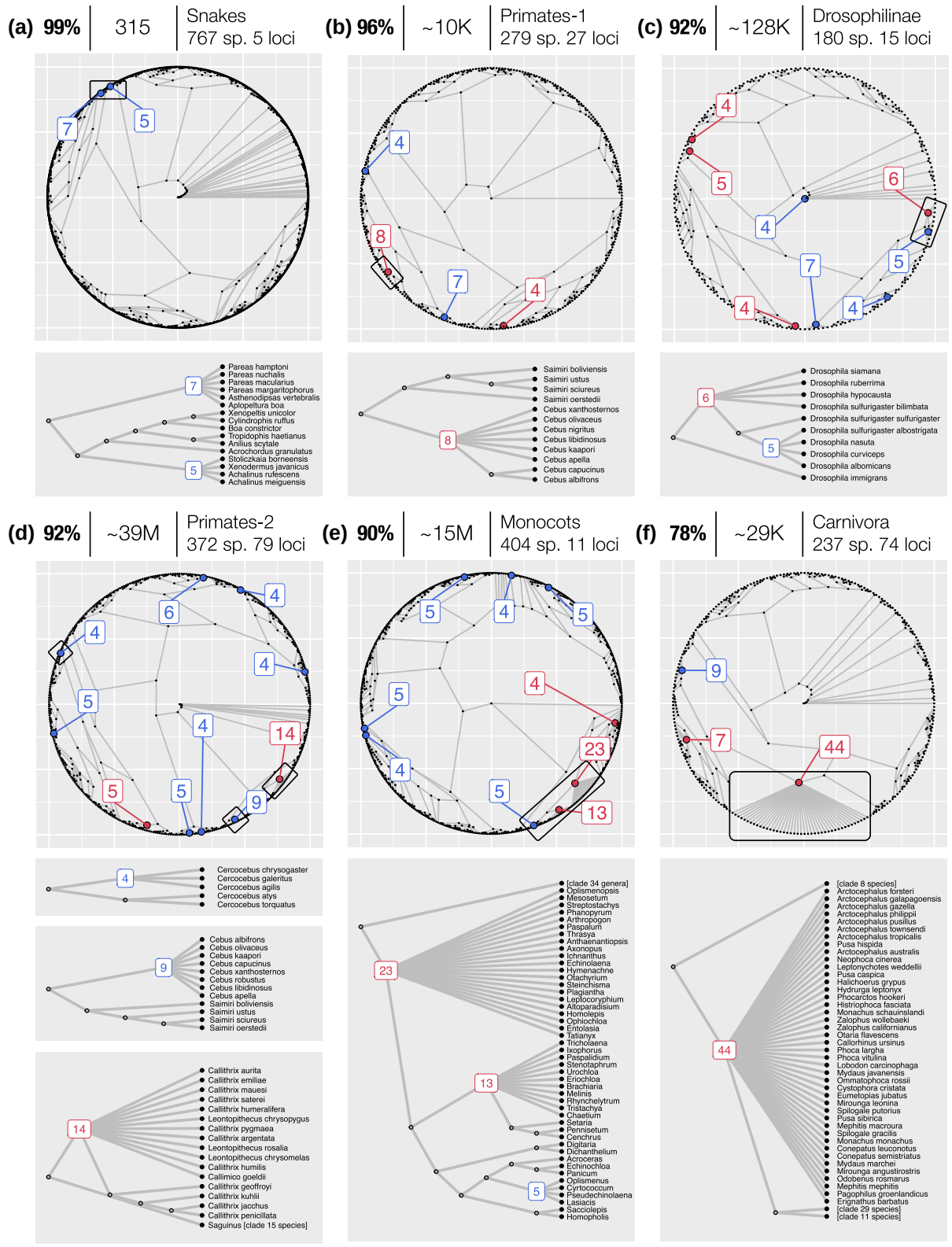

Fig. 5. Strict consensus trees for (complete) stands generated for ML trees for six biological datasets. The subplots are ordered by the decreasing resolution (in %) of strict consensus trees (left), followed by stand size (middle), species group, numbers of species and loci (right). The numbers in blue 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. The 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%.

Drosophilinae (Fig. 5c), then multifurcations prohibit a detailed analysis of the phylogenetic relationship of the species. The biggest issue is, of course, when multifurcating nodes involve species from different taxonomic groups and no conclusions can be drawn about their evolution as a result of missing data and terraces (e.g. Fig. 5a, d, e, and f).

In general, large number of multifurcating nodes and high node degree (i.e. the number of incident branches for a node; e.g. Fig. 5e and f) are indicators of analysis strongly affected by missing data. The findings above further demonstrate the critical importance of systematically accounting for missing data to achieve a robust phylogenomic analysis. Here, we advocate including Gentrius into a phylogenomic workflow to increase the confidence of findings.

Of note, the empirical datasets included in the manuscript aim at illustrating the terrace concept and the impact of missing sequences on the number of terrace trees as well as tree resolution. Importantly, since the stand size is the main driver for the runtime, the datasets provided challenging test cases for Gentrius. To further elucidate the occurrence of phylogenetic terraces in modern day studies, an exhaustive analysis of presence–absence matrices from contemporary datasets collected in recently published (and constantly growing) database RAxML Groove (Höhler et al. 2022) will be published elsewhere.

Discussion

The main result of the paper is the development of Gentrius, to the best of our knowledge, the first algorithm to generate complete stands for binary unrooted subtrees without any constraints on the structure and type of input data. Gentrius is a deterministic branch-and-bound like algorithm. While being the first in class for the stand generation problem for unrooted subtrees (i.e. no constraints on the input data), Gentrius is also very efficient. The evaluation of its feasibility on simulated and biological datasets evidenced that Gentrius deals with various datasets, generating millions of trees within reasonable time. It is worth mentioning that the runtime of Gentrius depends primarily on the stand size and not, for instance, on the species and locus numbers, like in tree inference methods (e.g. supertree and supermatrix). Our simulations and application to empirical datasets elucidated these points.

Importantly, Gentrius has direct practical application generating phylogenetic terraces—equally scoring trees due to missing data. As the number of species with diverse genome composition (e.g. species-specific gene losses and acquisitions) grows, collecting complete datasets with thousands of species is challenging. Therefore, missing data are an intrinsic property of datasets with highly diverse species. Thus, generating terraces and investigating impact of missing data should be routine in phylogenetic workflow to assure reliable phylogenetic inference.

Here, we have demonstrated, that sizes of stands (and, thus, terraces) are strongly affected by missing data. 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, for instance, 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 reanalyzed. It is expected that terraces are small or even trivial for datasets with locus number much greater than species number (e.g. D1 and D2). While more and more contemporary studies exhibit such characteristics, according to RAxML Groove database (Höhler et al. 2022), there is still a substantial number of datasets with missing data and with locus number (much) smaller than species number. For such datasets, Gentrius is crucial in identifying the impact of missing data on phylogenetic inference.

Huge nontrivial terraces are a sign that the input data is not sufficient for reliable phylogenetic conclusions. However, a terrace size of one does not exclude the existence of other equally or almost equally good trees (Maddison 1991; Silva and Wilkinson 2021). For instance, note that for supertree methods and parsimony, trees from multiple terraces can have identical scores. Thus, generating one stand/terrace for these methods gives only a lower bound on the total number of equally scoring trees.

In likelihood methods, all stand trees have identical score (i.e. stand is terrace) under edge unlinked partition model (Sanderson et al. 2015). To avoid terraces, one may employ different partition schemes or more restrictive partition models (Sanderson et al. 2015). Then, trees from one stand have similar but not identical likelihood scores (Breitling et al. 2019). Thus, to generate trees with similar quality for alternative partition models or schemes, instead of expensive multiple runs of likelihood inference, one can use stands, and subsequently analyze stand trees by computing their scores and using tree tests (e.g. Shimodaira 2002). If missing data are unavoidable, the best strategy is the inclusion of a large number of nearly complete loci and reducing the number of species with a lot of missing data (e.g. D1, D2, D5 datasets).

Importantly, phylogenetic terraces and missing data affect not only tree inference but also bootstrap (Felsenstein 1985)—a classical method for assessing clade confidence. Namely, a tree for each bootstrap replicate (i.e. resampled alignment) might lie on a large terrace of its own (Sanderson et al. 2015), which may cause misleading bootstrap scores (Sanderson et al. 2015). Even summarizing bootstrap trees becomes nontrivial (Sanderson et al. 2015). Importantly, bootstrap cannot explicitly say anything about the effect of missing data on equally optimal trees. In contrast, by generating all terrace trees for the best-found tree Gentrius provides an unbiased deterministic summary of the effect of missing data on the number and topology of equally optimal trees.

Apart from direct practical application to phylogenetic terraces, Gentrius fosters our theoretical understanding of

tree spaces (set of all trees for a given species number, e.g. Fig. 1c). For instance, it allows investigating scoring landscapes imposed by missing data as well as studying how trees from the same stand and between different stands are connected via topological rearrangements (e.g. Nearest Neighbor Interchange, see also Mark et al. 2016). This knowledge has a great potential for development of better tree search strategies in the presence of missing data (Chernomor et al. 2015, 2016).

Materials and Methods

Extended Description of Gentrius Algorithm

See supplementary note S1, Supplementary Material online.

Identifying Admissible Branches

See supplementary note S2, Supplementary Material online.

Implementation, Complexity and Validation

See supplementary note S3, Supplementary Material online.

Custom Matrix Simulator

See supplementary note S4, Supplementary Material online.

Comparison With Terraphast

See supplementary note S5, Supplementary Material online.

Simulated Presence–Absence Matrices and Trees

SIM1: To test feasibility of Gentrius we simulated 6,120 binary matrices using a custom matrix simulator (supplementary note S4, Supplementary Material online). For each triplet of species number (rows), locus number (columns) and the percentage of missing data (zeros in a matrix), we varied spread of zeros across rows and columns. Starting with a matrix of one's, the zeros were distributed using row and column sampling probabilities, subject to constraints on row/column sums. The sampling probabilities were chosen to include various combinations of rows and columns with high, medium and low amounts of missing data. Basic constraints on row/column sums included: (i) column sum ≥ 4 (i.e. locus covers at least four species); (ii) each column has missing data; (iii) row sum ≥ 1 (i.e. each species has to be covered by at least one locus); (iv) matrix has either at least one complete row or no complete rows. We generated 3,060 matrices that could be analyzed with both Terraphast and Gentrius (i.e. with at least one complete row) and 3,060 only with Gentrius (i.e. no complete rows). Additionally, we varied minimal amount of missing data per locus (column) and the number of minimally covered species (rows with row sum = 1). For more details on parameters and tested values, see supplementary fig. S4, Supplementary Material online and supplementary table S1, Supplementary Material online. For each of the presence–absence matrix, we generated five distinct random trees IQ-TREE 2 (Minh et al. 2020). The resulting 30,600 simulated instances were analyzed with Gentrius to obtain corresponding stands.

SIM2: To elucidate the dependence of stand size on the coverage pattern in species and loci, we simulated matrices fixing the amount of missing data in loci or species. Namely, we generated 66 matrices with 100 species, 10 and 30 loci, and 30%, 50%, and 70% of missing data. First, we fixed the number of missing entries per species to uniform (e.g. each species has exactly 3, 5, or 7 missing entries for 10 loci and 30%, 50%, and 70% of missing data respectively). The spread across loci varied from uniform (supplementary figs. S9a and S10a, Supplementary Material online, matrices with id 1) to datasets with high and low locus coverage (matrices 2 to 4). Then missing data per loci were fixed and varied for species (supplementary figs. S9 and S10, matrices 5 to 11). For all parameter combinations, matrices 1 to 8 did not have species represented by a single locus (i.e. row sum = 1), while matrices 9, 10, 11 had 5, 10, 20 such minimally covered species, respectively. All parameters used to generate each matrix are listed in supplementary table S3, Supplementary Material online. We next generated 50 distinct random trees with IQ-TREE 2 and computed stands for these trees for each matrix using Gentrius. The same 50 trees were used for each simulated matrix. Note, that we initially simulated multiple matrices per parameter setting. The stand sizes were very similar for similar type of matrices. Hence, the matrices shown in the article are representative and the conclusions about the dependence of stand sizes on missing data across species and loci are consistent.

Biological Datasets and Their Phylogenetic Trees

We analyzed 12 published alignments (Table 1; D1—Snails (Cunha et al. 2022); D2—Land plants (Wickett et al. 2014); D3—Drosophilinae (Van Der Linde et al. 2010); D4—Carnivora (Nyakatura and Bininda-Emonds 2012); D5—Frogs (Echevarría et al. 2021); D6—Primates-1 (Fabre et al. 2009); D7—Grasses (Bouchenak-Khelladi et al. 2008); D8—Primates-2 (Springer et al. 2012); D9—Salamander (Jaramillo et al. 2020); D10—Monocots (Stamatakis and Alachiotis 2010); D11—Sedges (Hinchliff and Roalson 2013); D12—Snakes (Pyron et al. 2011)). For each alignment, we inferred a ML tree using IQ-TREE 2 (version 2.2.2.4 for D1 and D2, version 2.1.2 for D3–D12) with branch unlinked partition model (Yang 1996; Chernomor et al. 2016). Best fit evolutionary models for each partition were identified using ModelFinder (Kalyaanamoorthy et al. 2017). For each alignment, we computed presence–absence matrix. Next, using Gentrius we generated stands for all 12 inferred ML trees and corresponding matrices. To explore other stands for these 12 empirical presence–absence matrices, for each dataset we also generated 100 distinct random trees using IQ-TREE 2 and computed their stands with Gentrius.

Topological Differences of Trees from the Same Stand

For simulated and empirical datasets D12, D6, D3, and D4, we constructed strict consensus trees with IQ-TREE 2. For

datasets D8 and D10 with ∼39 M and ∼15 M trees on corresponding stands, we used TreeShredder (Bader 2023), which allows efficient analysis for such large tree sets. Importantly, strict consensus trees were only constructed for complete stands. The number of internal branches in strict consensus trees, the tree resolution and the number of multifurcating nodes, as well as their types were computed using a custom R script. The trees were visualized with the same script.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

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

Author Contributions

Conceptualization: O.C., C.E., and A.v.H.; methodology: O.C.; software: O.C.; formal analysis: O.C.; investigation: O.C.; visualization: O.C.; funding acquisition: O.C. and A.v.H.; writing—original draft: O.C.; writing—review and editing: O.C., C.E., and A.v.H.

Funding

O.C. and A.v.H. acknowledge funding from the Austrian Science Fund—FWF grant number I-4686.

Conflict of Interest

The authors declare that they have no competing interests.

Data 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/, Gentrius Manual at GitHub: https://github.com/iqtree/iqtree2/tree/master/terrace). 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.

References

Bader C. 2023. TreeShredder: a program for phylogenetic analysis of large sets of trees based on splits [master's thesis] University of Vienna. https://doi.org/10.25365/thesis.74014.

Biczok R, Bozsoky P, Eisenmann P, Ernst J, Ribizel T, Scholz F, Trefzer A, Weber F, Hamann M, Stamatakis A. Two C++ libraries for counting trees on a phylogenetic terrace. Bioinformatics. 2018:**34**(19):3399–3401. https://doi.org/10.1093/bioinformatics/bty384.

Böcker S. Exponentially many supertrees. Appl Math Lett. 2002:**15**(7):861–865. https://doi.org/10.1016/S0893-9659(02)00054-X.

Böcker S, Bryant D, Dress AWM, Steel MA. Algorithmic aspects of tree amalgamation. Algorithms. 2000:**37**(2):522–537. https://doi.org/10.1006/jagm.2000.1116.

Bordewich M, Semple C, Talbot J. Counting consistent phylogenetic trees is #P-complete. Adv Appl Math. 2004:**33**(2):416–430. https://doi.org/10.1016/j.aam.2003.08.006.

Bouchenak-Khelladi Y, Salamin N, Savolainen V, Forest F, van der Bank M, Chase MW, Hodkinson TR. Large multi-gene phylogenetic trees of the grasses (Poaceae): progress towards complete tribal and generic level sampling. Mol Phylogenet Evol. 2008:**47**(2):488–505. https://doi.org/10.1016/j.ympev.2008.01.035.

Breitling P, Stamatakis A, Chernomor O, Bettisworth B, Reszczynski L. Empirical analysis of phylogenetic quasi-terraces. bioRxiv. 2019. https://www.biorxiv.org/content/10.1101/810309v1.full, preprint: not peer reviewed.

Chernomor O, Minh BQ, von Haeseler A. Consequences of common topological rearrangements for partition trees in phylogenomic inference. J Comput Biol. 2015:**22**(12):1129–1142. https://doi.org/10.1089/cmb.2015.0146.

Chernomor O, von Haeseler A, Minh BQ. Terrace aware data structure for phylogenomic inference from supermatrices. Syst Biol. 2016:**65**(6):997–1008. https://doi.org/10.1093/sysbio/syw037.

Constantinescu M, Sankoff D. An efficient algorithm for supertrees. J Classif. 1995:**12**(1):101–112. https://doi.org/10.1007/BF01202270.

Cunha TJ, Reimer JD, Giribet G. Investigating sources of conflict in deep phylogenomics of vetigastropod snails. Syst Biol. 2022:**71**(4):1009–1022. https://doi.org/10.1093/sysbio/syab071.

Dobrin BH, Zwickl DJ, Sanderson MJ. The prevalence of terraced treescapes in analyses of phylogenetic data sets. BMC Evol Biol. 2018:**18**(1):46. https://doi.org/10.1186/s12862-018-1162-9.

Echevarría LY, De la Riva I, Venegas PJ, Rojas-Runjaic FJM, Dias IR, Castroviejo-Fisher S. Total evidence and sensitivity phylogenetic analyses of egg-brooding frogs (Anura: Hemiphractidae). Cladistics. 2021:**37**(4):375–401. https://doi.org/10.1111/cla.12447.

Fabre PH, Rodrigues A, Douzery EJP. Patterns of macroevolution among primates inferred from a supermatrix of mitochondrial and nuclear DNA. Mol Phylogenet Evol. 2009:**53**(3):808–825. https://doi.org/10.1016/j.ympev.2009.08.004.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. 1985:**39**(4):783–791. https://doi.org/10.2307/2408678.

Goloboff PA, Catalano SA, Torres A. Parsimony analysis of phylogenomic datasets (II): evaluation of PAUP*, MEGA and MPBoot. Cladistics. 2022:**38**(1):126–146. https://doi.org/10.1111/cla.12476.

Gordon AD. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. J Classif. 1986:**3**(2):335–348. https://doi.org/10.1007/BF01894195.

Habib M, Rahman AH, Bayzid MdS. Terraces in species tree inference from gene trees. bioRxiv. 2022. https://www.biorxiv.org/content/10.1101/2022.11.21.517454v2, preprint: not peer reviewed.

Harding EF. The probabilities of rooted tree-shapes generated by random bifurcation. Adv Appl Probab. 1971:**3**(1):44–77. https://doi.org/10.2307/1426329.

Hinchliff CE, Roalson EH. Using supermatrices for phylogenetic inquiry: an example using the sedges. Syst Biol. 2013:**62**(2):205–219. https://doi.org/10.1093/sysbio/sys088.

Höhler D, Pfeiffer W, Ioannidis V, Stockinger H, Stamatakis A. RAxML grove: an empirical phylogenetic tree database. Bioinformatics. 2022:**38**(6):1741–1742. https://doi.org/10.1093/bioinformatics/btab863.

Hosner PA, Faircloth BC, Glenn TC, Braun EL, Kimball RT. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol Biol Evol. 2016:**33**(4):1110–1125. https://doi.org/10.1093/molbev/msv347.

Jaramillo AF, De La Riva I, Guayasamin JM, Chaparro JC, Gagliardi-Urrutia G, Gutiérrez RC, Brcko I, Vilà C, Castroviejo-Fisher S. Vastly underestimated species richness of Amazonian salamanders (Plethodontidae: Bolitoglossa) and implications about plethodontid diversification. Mol Phylogenet Evol. 2020:**149**:106841. https://doi.org/10.1016/j.ympev.2020.106841.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017:**14**(6):587–589. https://doi.org/10.1038/nmeth.4285.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019:**35**(21):4453–4455. https://doi.org/10.1093/bioinformatics/btz305.

Maddison DR. The discovery and importance of multiple islands of most-parsimonious trees. Syst Biol. 1991:**40**(3):315–328. https://doi.org/10.1093/sysbio/40.3.315.

Mark S, McLeod JC, Steel M. A navigation system for tree space. J Graph Algorithms Appl. 2016:**20**(2):247–268. https://doi.org/10.7155/jgaa.00392.

Mickevich MF. Taxonomic congruence. Syst Zool. 1978:**27**(2):143–158. https://doi.org/10.2307/2412969.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R, Teeling E. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020:**37**(5):1530–1534. https://doi.org/10.1093/molbev/msaa015.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson M S, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 2014:**30**(17):i541–i548. https://doi.org/10.1093/bioinformatics/btu462.

Molloy EK, Warnow T. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst Biol. 2018:**67**(2):285–303. https://doi.org/10.1093/sysbio/syx077.

Nute M, Chou J, Molloy EK, Warnow T. The performance of coalescent-based species tree estimation methods under models of missing data. BMC Genomics. 2018:**19**(S5):286. https://doi.org/10.1186/s12864-018-4619-8.

Nyakatura K, Bininda-Emonds ORP. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol. 2012:**10**(1):12. https://doi.org/10.1186/1741-7007-10-12.

Pyron RA, Burbrink FT, Colli GR, de Oca ANM, Vitt LJ, Kuczynski CA, Wiens JJ. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. Mol Phylogenet Evol. 2011:**58**(2):329–342. https://doi.org/10.1016/j.ympev.2010.11.006.

Sanderson MJ, Mcmahon MM, Stamatakis A, Zwickl DJ, Steel M. Impacts of terraces on phylogenetic inference. Syst Biol. 2015:**64**(5):709–726. https://doi.org/10.1093/sysbio/syv024.

Sanderson MJ, McMahon MM, Steel M. Terraces in phylogenetic tree space. Science. 2011:**333**(6041):448–450. https://doi.org/10.1126/science.1206357.

Sanderson MJ, McMahon MM, Steel M. Terraces in gene tree reconciliation-based species tree inference. Biorxiv. 2020. https://www.biorxiv.org/content/10.1101/2020.04.17.047092v1, preprint: not peer reviewed.

Semple C, Steel M. Phylogenetics. New York: Oxford University Press; 2003.

Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002:**51**(3):492–508. https://doi.org/10.1080/10635150290069913.

Silva AS, Wilkinson M. On defining and finding islands of trees and mitigating large island bias. Syst Biol. 2021:**70**(6):1282–1294. https://doi.org/10.1093/sysbio/syab015.

Simmons MP. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics. 2012:**28**(2):208–222. https://doi.org/10.1111/j.1096-0031.2011.00375.x.

Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, Stadler T, Steiner C, Ryder OA, Janečka JE, et al. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS One. 2012:**7**(11):e49521. https://doi.org/10.1371/journal.pone.0049521.

Stamatakis A, Alachiotis N. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. Bioinformatics. 2010:**26**(12):i132–i139. https://doi.org/10.1093/bioinformatics/btq205.

Vachaspati P, Warnow T. SIESTA: enhancing searches for optimal supertrees and species trees. BMC Genomics. 2018:**19**(S5):252. https://doi.org/10.1186/s12864-018-4621-1.

Van Der Linde K, Houle D, Spicer GS, Steppan SJ. A supermatrix-based molecular phylogeny of the family Drosophilidae. Genet Res (Camb). 2010:**92**(1):25–38. https://doi.org/10.1017/S001667231000008X.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 2014:**111**(45):E4859–E4868. https://doi.org/10.1073/pnas.1323926111.

Xi Z, Liu L, Davis CC. The impact of missing data on species tree estimation. Mol Biol Evol. 2016:**33**(3):838–860. https://doi.org/10.1093/molbev/msv266.

Yang Z. Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol. 1996:**42**(5):587–596. https://doi.org/10.1007/BF02352289.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, Fitzjohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, et al. Three keys to the radiation of angiosperms into freezing environments. Nature. 2014:**506**(7486):89–92. https://doi.org/10.1038/nature12872.