Terraces in Gene Tree Reconciliation-Based Species Tree Inference

Michael J. Sanderson[1,*], Michelle M. McMahon[2], and Mike Steel[3]

[1] *Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA*
[2] *School of Plant Sciences, University of Arizona, Tucson, AZ 85721*
[3] *Biomathematics Research Centre, University of Canterbury, Christchurch, NZ*

*Michael J. Sanderson, Dept. of Ecology and Evolutionary Biology, University of Arizona; Tucson, AZ

85721, USA; 520-626-6848; sanderm@email.arizona.edu

Abstract

1 Terraces in phylogenetic tree space are sets of trees with identical optimality scores for a

2 given data set, arising from missing data. These were first described for multilocus

3 phylogenetic data sets in the context of maximum parsimony inference and maximum

4 likelihood inference under certain model assumptions. Here we show how the mathematical

5 properties that lead to terraces extend to gene tree - species tree problems in which the

6 gene trees are incomplete. Inference of species trees from either sets of gene family trees

7 subject to duplication and loss, or allele trees subject to incomplete lineage sorting, can

8 exhibit terraces in their solution space. First, we show conditions that lead to a new kind

9 of terrace, which stems from subtree operations that appear in reconciliation problems for

10 incomplete trees. Then we characterize when terraces of both types can occur when the

11 optimality criterion for tree search is based on duplication, loss or deep coalescence scores.

12 Finally, we examine the impact of assumptions about the causes of losses: whether they are

13 due to imperfect sampling or true evolutionary deletion.

14 *Key words*: gene tree reconciliation, terraces, phylogenomics, span

15

A long standing and still largely dominant paradigm in phylogenetic tree inference is based on optimization of some score derived from data over candidate tree solutions. In addition to familiar maximum parsimony, maximum likelihood, and (certain) distance-based methods like minimum evolution and Fitch-Margoliash (Felsenstein 2004), all commonly used to infer a tree from a sequence alignment, optimization methods also are employed in a diverse set of methods aimed at solving other tree inference problems, such as supertree construction, gene tree reconciliation, species tree inference using likelihood or pseudo-likelihood, and network reconstruction. Computational obstacles in optimization include the problem of multiple optima and regions where the solution space is flat, which can both impede algorithms to find optima and make circumscription of solutions more complex. One contributor to this problem in phylogenetics is missing data, and a particularly direct example of this is the phenomenon of "terraces"—regions of tree space having identical optimality scores purely due to certain patterns of missing data among the taxa sampled (Sanderson et al. 2011, 2015).

The properties of terraces have been elucidated mostly in the context of large multilocus data sets, where the pattern of missing data can be described by the "taxon coverage" of data—which loci are sampled for which taxa. If a tree is inferred for a concatenated multilocus alignment by maximum parsimony, or by maximum likelihood with certain model assumptions, the pattern of taxon coverage alone can be used to infer the number and sizes of terraces having identical optimality scores. Surveys of empirical studies indicate that terraces can be astronomically large in large trees (Dobrin et al. 2018), and they can degrade other aspects of phylogenetic inference aside from tree search, such as estimation of confidence intervals (Sanderson et al. 2015). They are likely to have an impact on comparative methods and other studies that employ statistical inference methods that depend on accurate evaluation of the confidence set of trees.

The conditions required for terraces to be observed in phylogenetic inference problems are quite general (Steel 2016). In this paper we describe their impact on a large

43 additional set of phylogenetic problems involving building species trees from gene trees

44 using gene tree reconciliation methods. This includes methods that have been used to infer

45 species trees from gene families that have undergone a history of gene duplication and

46 deletion, as well as methods that have been used to infer species tree from allele trees in

47 which population genetic processes lead to deep coalescence events. This area is

48 undergoing active development in the phylogenomics community and our results pertain

49 most directly to discrete parsimony-like methods arising from the reconciliation

50 framework, but potential connections to model based approaches may exisit by analogy to

51 previous findings with maximum likelihood in concatenation methods (Sanderson et al.

52 2011). We find that terraces are not only expected in this new context, but that an

53 additional *type* of terraces can be seen in certain versions of reconciliation problems.

54 ## *Trees, Subtrees, and Display*

55 We assume all trees, T, are rooted, binary, and have edges, $E(T)$ and nodes, $V(T)$,

56 and root node, $r(T) \in V(T)$. Nodes are partitioned into the set of internal nodes, $\dot{V}(T)$

57 (having outdegree two or more), and the set of leaves, $L(T)$ (having indegree one), each of

58 which is labeled by a distinct element of the leaf set, X. If node v' is an ancestor of v, we

59 write $v' > v$. The set of leaf labels descended from v is $C_T(v)$. The node of the most recent

60 common ancestor of a set of leaves, A, is $\mathrm{MRCA}_T(A)$.

61 Given a rooted binary tree, T with leaf label set, X, and with $X_g \subseteq X$, we define

62 two kinds of subtrees of T having leaf label set, X_g (Fig. 1):

63 1. The "homomorphic subtree", $T|_{X_g}$, is the smallest subtree of T having leaf label

64 set X_g and suppressing any interior nodes with outdegree one ("unary" nodes). When the

65 context is clear, we abbreviate this to $T|_g$.

66 2. The "restriction subtree", $T||_{X_g}$, is the smallest subtree having leaf label set X_g

67 but keeping any unary nodes. When the context is clear, we abbreviate this to $T||_g$.

68 A tree T *displays* a binary subtree g having leaf label set $X_g \subseteq X$ if $T|_{X_g} = g$ (Steel

69 2016). [If g is not binary then this definition can be extended to allow $T|_g$ to be a

70 refinement (resolution) of g.] This is the conventional definition of "display", but it can be

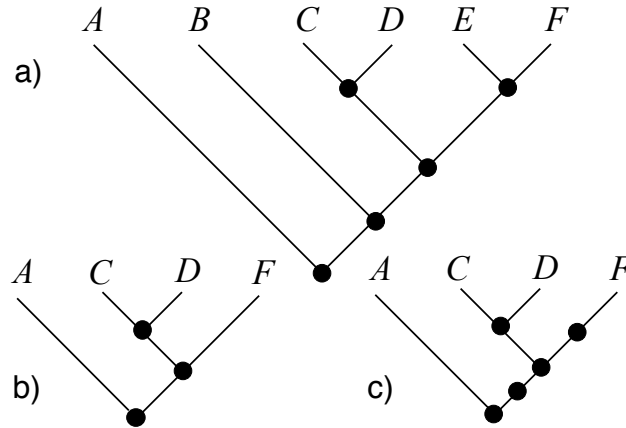71 generalized to use the restriction subtree when appropriate (see below).

Fig. 1. Illustration of two subtree operations, retaining leaf label set $X_g = \{A, C, D, F\}$: a) Original species tree, T; b) homomorphic subtree, $T|_{X_g}$; c) restriction subtree, $T||_{X_g}$.

Spans and Terraces

73   In general, given a sequence of rooted binary trees, $\mathcal{T} = (T_1, ..., T_k)$, with leaf label

74 sets X_i ($X = \cup_i X_i$), define the *span*, $\langle \mathcal{T} \rangle$, as the set of all rooted binary trees having label

75 set X that display every tree in $\mathcal{T}$ (Fig. 2). The span may be empty.

76   Consider the special case of a rooted binary *parent tree*, T, with leaf label set, X,

77 along with a sequence of leaf label subsets, $\chi = (X_1, ..., X_k)$, $X_i \subseteq X$. Label set X_i may be

78 thought of as the set of leaves present in a subtree of T or as the set of leaves of T that

79 have some kind of data present, say the sequence of the ith gene in a multigene multiple

80 sequence alignment. Let $\mathcal{T}|_\chi = (T|_{X_1}, ..., T|_{X_k})$, then $\langle \mathcal{T}|_\chi \rangle$ is the span of this sequence of

81 subtrees derived from the parent tree (Fig. 2). Clearly the size of this span is at least one

82 since it contains T.

83   A *terrace* is a span, $\langle \mathcal{T}|_\chi \rangle$, in which all the trees have the same *score*. The

84 properties of this score determine whether this is a possibility. Given some cost function,
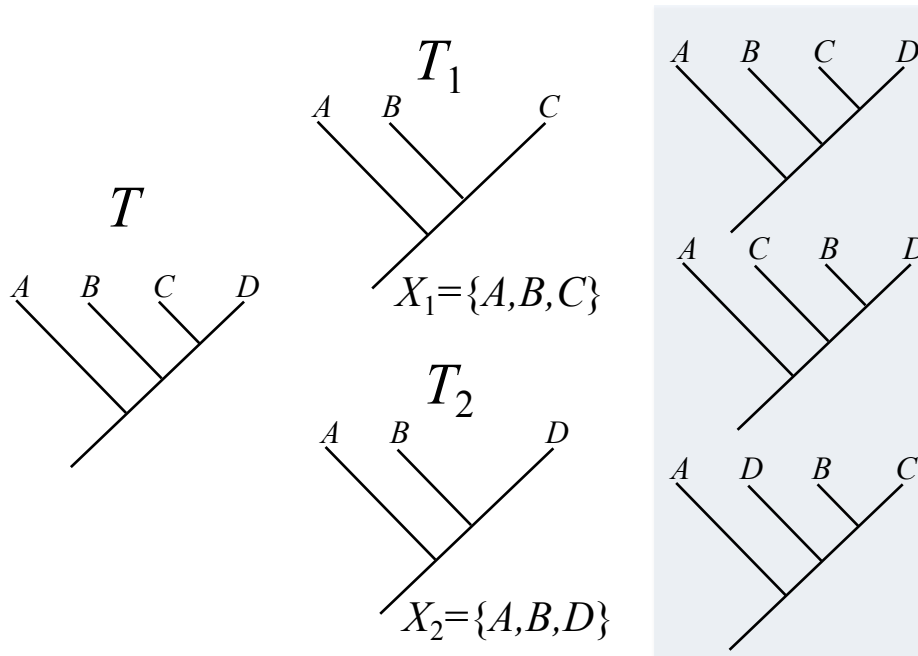
Fig. 2. Span of a set of subtrees. Original rooted parent tree is T having leaf label set, $X = \{A, B, C, D\}$. Subsets of label sets, X_1 and X_2, induce a sequence of subtrees, $\mathcal{T}|_\chi = (T|_{X_1}, T|_{X_2}) = (T_1, T_2)$. Each of the trees on X in the shaded box displays both T_1 and T_2 and this set of three trees is the *span* of $\mathcal{T}|_\chi$, $\langle \mathcal{T}|_\chi \rangle$.

$c(S, g)$ for a tree, T, with leaf label set, X, based on some data associated with "gene", g, valid even if only a subset of leaves, $X_g \subseteq X$, have data present; and given some score function, $s(c(T, g_1), ..., c(T, g_m))$, that combines these costs across a set of genes, $\mathcal{G} = \{g_1, ..., g_m\}$, the following two properties of c and s are sufficient to cause all trees in the span to have the same score (Sanderson et al. 2011; Steel 2016):

Condition 1. The cost function is the same on the full tree as it is on the subtree pruned to those taxa that have data.

$$c(T, g) = c(T|_{X_g}, g). \tag{1}$$

Condition 2. The score function, s, is a linear sum of costs for each gene:

$$s(T, \mathcal{G}) = \sum_i c(T, g_i). \tag{2}$$

Theorem 1 (Terraces—Type I). *If the score function, s, satisfies the two conditions described above, then every tree in the span, $\langle \mathcal{T}|_\chi \rangle$, has the score, $s(T, \mathcal{G})$.*

Proof. Following Steel (2016), if $T' \in \langle \mathcal{T}|_\chi \rangle$, Then $T'|_{X_i} = T|_{X_i}$ for $i = 1, ..., k$, so

$$s(T', \mathcal{G}) = \sum_i c(T', g_i) = \sum_i c(T'|_{X_i}, g_i) = \sum_i c(T|_{X_i}, g_i) = \sum_i c(T, g_i) = s(T, \mathcal{G}) \quad (3)$$

$\square$

This is most interesting if $|\langle \mathcal{T} \rangle| >> 1$. In fact, the size of terraces can be astronomically large for some data sets for large parent trees (Sanderson et al. 2011; Dobrin et al. 2018) when there is a sizable amount of missing data.

In Sanderson et al. (2011, 2015) and Steel (2016) the data set associated with gene g was taken to be a multiple sequence alignment over some block of sites, and the scores considered included maximum parsimony and maximum likelihood scores. The exact circumscription of "blocks" is relevant, but for simplicity we conceptualize it as data associated with a single gene in the genome (see Sanderson et al. 2015).

Terraces—Type II

We now show that terraces arise when using the restriction subtree operation, which will be relevant for discussion of gene tree reconciliation below.

Define a "restriction span" by (i) using the restriction subtree in the above definition of "display", and then (ii) letting $\mathcal{T}||_\chi = (T||_{X_1}, ..., T||_{X_k})$. The resulting span can be labeled $\langle\langle \mathcal{T}||_\chi \rangle\rangle$ (see example in Fig. 3). [The double angle brackets reinforce the idea that the definition of "display" has changed]

Theorem 2 (Terraces—Type II). *If the score function, s, satisfies the second condition required for Theorem 1 (linear sum of costs), and the first condition is modified so that* $c(T, g) = c(T||_{X_g}, g)$, *then every tree in the restriction span,* $\langle\langle \mathcal{T}|_\chi \rangle\rangle$, *has the score,* $s(T, \mathcal{G})$.

Proof. The proof follows easily by substituting the restriction span and restriction subtree operations in the proof for Theorem 1 $\square$

The example in Fig. 3 shows that there can be more than one tree on this kind of terrace, as with the first type of terrace. The elements of $\langle\langle \mathcal{T}||_\chi \rangle\rangle$ and $\langle \mathcal{T}|_\chi \rangle$ are not
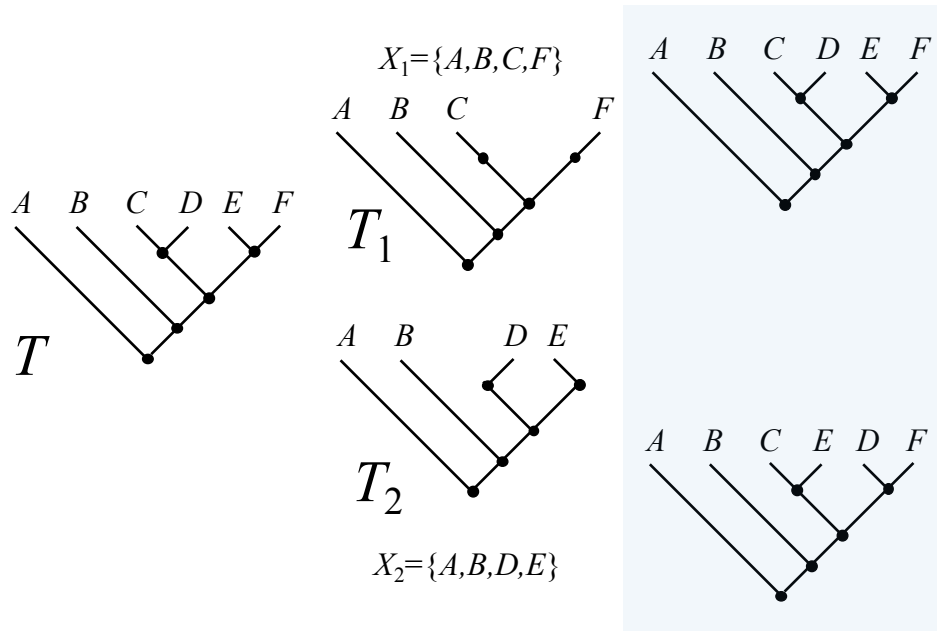
Fig. 3. Restriction span of a set of subtrees. Original rooted parent tree is T having leaf label set, $X = \{A, B, C, D, E, F\}$. Subsets of label sets, $X_1, X_2 \subseteq X$, induce "restriction" subtrees (Fig. 1), $\mathcal{T}||_\chi = (T||_{X_1}, T||_{X_2}) = (T_1, T_2)$. Each of the trees in the shaded box displays both T_1 and T_2, and this set of two trees is the *restriction span*, $\langle\langle \mathcal{T}||_\chi \rangle\rangle$.

necessarily the same. For example, in Fig. 3 there are 15 trees in $\langle \mathcal{T}|_\chi \rangle$ but only two in

$\langle\langle \mathcal{T}||_\chi \rangle\rangle$. In general, the former set of trees contains the latter, as we now state:

Theorem 3. *For any rooted binary parent tree, T, having leaf label set X, and given label*

sets $\chi = (X_1, ..., X_k)$, $X_i \subseteq X$,

$$\langle\langle \mathcal{T}||_\chi \rangle\rangle \subseteq \langle \mathcal{T}|_\chi \rangle. \tag{4}$$

Proof. Suppose that $T' \in \langle\langle \mathcal{T}||_\chi \rangle\rangle = \langle\langle T||_{X_1}, \ldots, T||_{X_k} \rangle\rangle$. Then $T'||_{X_i} = T||_{X_i}$ for all

$i = 1, \ldots, k$ and so $T'|_{X_i} = T|_{X_i}$ for each i (since restriction display implies ordinary

display) and thus $T' \in \langle T|_{X_1}, \ldots, T|_{X_k} \rangle = \langle \mathcal{T}|_\chi \rangle$.

$\square$

In parallel with the terminology for terraces, we refer to these two spans as "Type

I" and "Type II" respectively.

Species Tree Inference by Reconciliation

¹³⁰

¹³¹ Different regions of the genome can have different phylogenetic histories, and these

¹³² "gene trees" can also differ from the species tree in which they are imbedded (Bravo et al.

¹³³ 2019; Liu et al. 2019). The framework of gene tree "reconciliation" provides an analytically

¹³⁴ rich and empirically powerful toolkit to understand this mosaic of histories and to infer

¹³⁵ species trees from discordant gene trees (Goodman et al. 1979; Page 1994). Here we extend

¹³⁶ our previous results on terraces to the reconciliation setting by discussing conditions under

¹³⁷ which reconciliation-based score functions lead to terraces when there are missing leaves in

¹³⁸ the gene trees.

¹³⁹ Reconciliation algorithms are based on costs that reflect (at least) four kinds of

¹⁴⁰ explicitly *evolutionary* processes in the phylogenetic history of a gene tree imbedded in a

¹⁴¹ species tree, "speciation", "duplication", "deletion", and "deep coalescence". In addition,

¹⁴² costs may reflect some aspect of *sampling* of leaves (Page 1994; Page and Charleston

¹⁴³ 1997). Additional evolutionary events, such as lateral transfer, have been studied to a

¹⁴⁴ lesser extent (Bansal et al. 2012) but are not considered here. Assume there is a rooted

¹⁴⁵ binary species tree, S, leaf labeled by $\mathcal{L}_S$, and an imbedded rooted binary gene tree, g,

¹⁴⁶ labeled by $\mathcal{L}_g \subseteq \mathcal{L}_S$. In general, all reconciliation cost functions can be written as $c(S, g)$,

¹⁴⁷ and these are generally combined additively across gene trees, so the score function $s(T, \mathcal{G})$

¹⁴⁸ discussed above extends easily in principle, even though the underlying sequence data are

¹⁴⁹ no longer relevant.

¹⁵⁰ Define a *reconciliation* of S and g, loosely, as an annotation of g that indicates the

¹⁵¹ locations of speciation events, duplications and losses. [More formal definitions include

¹⁵² defining a *reconciled tree* as an extension of g made by inserting "lost" subtrees so that the

¹⁵³ resulting tree is consistent with speciation and duplication alone (Chauve and El-Mabrouk

¹⁵⁴ 2009).]

¹⁵⁵ Given a set of gene trees, $\mathcal{G}$, the *species tree inference problem* is then to find a

156 species tree, $\hat{S}$, that minimizes

$$s(S, \mathcal{G}) = \sum_{g_i \in \mathcal{G}} c(S, g_i) \tag{5}$$

157 over all S, for various choices of the cost function, $c(S, g_i)$.

Incompleteness, missing data, absence, deletion and loss

159 Let $\mathcal{L}_{S \backslash g}$ be the set of leaf labels of S missing from g. If $\mathcal{L}_{S \backslash g} \neq \varnothing$, then g is

160 *incomplete.* The concept of "missing data" in the original notion of terraces equates to

161 incompleteness in one or more gene trees, and this can lead to terraces in the species tree

162 inference problem.

163 However, distinguishing between evolutionary events and sampling processes can be

164 confounded in incomplete gene trees. In a gene tree having no leaves from species x, were

165 these genes evolutionarily deleted, or was x never sampled? For $x \in \mathcal{L}_{S \backslash g}$, by

166 "incompleteness due to deletion", we mean that the absence of x from g is considered

167 positive evidence that no such leaves exist in g (x has been well studied and its absence is

168 meaningful)—absence implies deletion. By "incompleteness due to sampling", we mean x

169 has not been studied, and we make no claim a priori about whether further study of x

170 would reveal it has leaves in g—absence implies no more than failure to sample or bad lab

171 technique (Bayzid and Warnow 2018).

172 Page and Charleston (1997) pointed out that the loss cost could influence species

173 tree inference under reconciliation and thereby raised the issue of how its interpretation

174 could matter. Bayzid and Warnow (2018) showed, in fact, that the correct loss score of a

175 gene tree reconciled with a species tree depends directly on this assumption about the

176 meaning of loss. We therefore qualify the term "loss" by whichever mechanism is assumed

177 to generate it.

Reconciliation Costs, Incompleteness and Spans

When g is complete, it is well known that an optimal reconciliation exists that minimizes the number of duplications, losses, and duplications plus losses (Chauve and El-Mabrouk 2009). The solution can be found using the *MRCA-mapping*, $\mathcal{M}$, from node u in g to a node in S: $\mathcal{M}_S(u) = \mathrm{MRCA}_S(C_g(u))$, which identifies the lowest node on the species tree that can have the subtree of g rooted at v imbedded within it (Fig. 4). Briefly, in this optimal reconciliation, a node u on g is annotated as a duplication if $\mathcal{M}_S(u)$ maps to the same node of S that one of the child nodes of u does. The duplication cost, c_{dup}, is the number of such duplications. Determining the location of, and the number of losses, c_{loss}, is just slightly more complicated, and depends on lengths of paths mapped onto S from g using $\mathcal{M}_S(v)$. The duplication plus loss cost is then $c_{D+L} = c_{dup} + c_{loss}$. All of these computations have been described in detail elsewhere (e.g. Zhang 2011; Bayzid and Warnow 2018) and we omit them here.
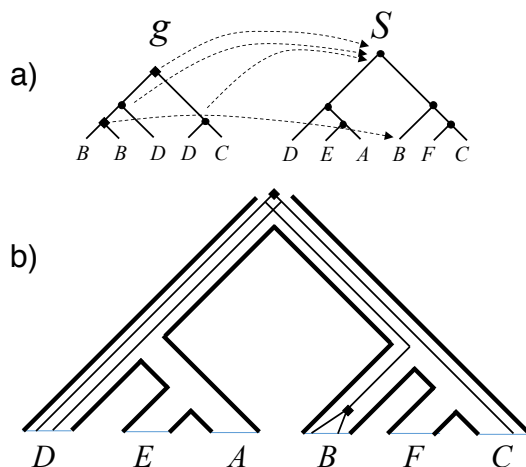

Fig. 4. Reconciliation of a gene tree, g, and species tree, S. Note the set of "missing leaves" on g, $\mathcal{L}_{S \setminus g} = \{A, E, F\}$. a) The MRCA mapping, $\mathcal{M}$, is shown for all internal nodes of g. Nodes annotated as duplications on g indicated by diamonds; nodes annotated as speciation indicated by dots. b) Imbedding of g in S. For clarity, only the duplication nodes on g are shown as annotated.

There are multiple formulations of the deep coalescence cost (Zhang 2011; Steel 2016), c_{DC}, which is the number of "extra edges" of the gene tree imbedded within the species tree for this reconciliation. Any edge $e = (u, v) \in E(g)$ is imbedded in $k_S(e) \geqslant 0$

194 edges of S where $k_S(e)$ is the number of edges in the path from $\mathcal{M}_S(u)$ to $\mathcal{M}_S(v)$, so the

195 number of extra edges is,

$$\sum_{e \in E(g)} k_S(e) - |E(S)| \tag{6}$$

196 since there is minimally one edge in each edge of $E(S)$.

197   The computation of these reconciliation costs for incomplete gene trees depends in

198 part on the interpretation of incompleteness.

199   *Incompleteness due to sampling.—* The trees g and $S|_g$ have the same label sets

200 and standard reconciliation algorithms can be used to compute $c(S|_g, g)$, but how do we

201 compute $c(S, g)$? If the absence of a gene tree leaf for some species taxon is assumed to

202 arise because of failure to sample it, we compute the cost, $c(S, g)$, by "completing" the

203 gene tree so it has the same label set as the species tree (Bayzid and Warnow 2018). A

204 *completion*, g', is the tree obtained from g by adding subtree(s) having all the leaves,

205 $x \in \mathcal{L}_{S \setminus g}$. An *optimal completion*, g^*, with respect to some cost function, c, is the

206 completion that minimizes c over all g'. For example, the optimal completion with respect

207 to the duplication cost would be that gene tree, g^*, that has smallest number of

208 duplications among all possible completions. The optimal completion is a guess about

209 unsampled leaves that disturbs the cost the least. From here on, when discussing optimal

210 completions for incomplete gene trees, we use as shorthand, $c(S, g)$, instead of $c(S, g^*(g))$.

211   Bayzid and Warnow (2018) show that there are optimal completions for the

212 duplication and loss scores, which have the same duplication and loss scores as that for

213 $c(S|_g, g)$, so that:

$$c_{dup}(S, g) = c_{dup}(S|_g, g) = c_{dup}(S||_g, g), \tag{7a}$$

$$c_{loss}(S, g) = c_{loss}(S|_g, g), \tag{7b}$$

214 which implies that the duplication plus loss cost, c_{D+L} can be obtained simply by adding

215 these together. Notice the subtree operation for the loss cost is the homomorphic subtree

216   operation.

217   This can be shown by the following argument. Each unary node, u, of $S||_g$

218   corresponds to a binary node on S that has one pendant subtree, t, the leaves of which are

219   all in $\mathcal{L}_{S\backslash g}$ (i.e., missing). If there are l edges of g passing through u, then we attach l

220   replicate instances of t along their respective edges of g at new nodes $v_i, i = 1, ..., l$, of g.

221   For each v_i, the mapping, $\mathcal{M}(v_i)$ must be to a different node in S than either the parent of

222   v_i maps to, or any of its children map to, because the leaves of t are in $\mathcal{L}_{S\backslash g}$. Thus there

223   are no additional duplications inferred by this completion. This argument holds equally for

224   both subtree operations, so $c_{dup}(S|_g, g) = c_{dup}(S||_g, g)$.

225   This completion also avoids adding l losses, by "imputing" just the right leaves to

226   be present as subtrees of g (Fig. 5). The loss cost is computed on $S|_g$ but not $S||_g$. The

227   intuition for this is that adding these "ghost" subtrees, t, at edges of g passing through u

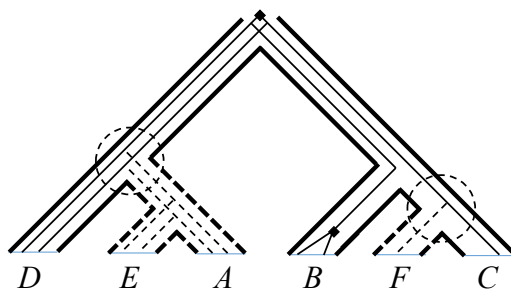228   allows us to pretend, with respect to losses, that node u was never there.

Fig. 5. Optimal completion under the duplication-loss costs. Trees are the same as in Fig. 4. Dashed border for some species tree edges indicates parts of S that are missing from g because of absence of leaves A, E, F. Dashed circles are unary nodes on the restriction subtree, $S||_g$. Dashed edges of g represent the optimal completion of g that allows leaves of g to be "present" at leaves of S, while adding as few duplications and losses as possible to the reconciliation. Here, no duplications or losses are added by this completion. Note that for the unary node at left, two edges of g traverse it and two subtrees are added, whereas for the unary node at right having one imbedded edge, only one subtree need be added.

229   For the deep coalescence cost when incompleteness is assumed due to sampling we

230   can also use the optimal completion approach (Bayzid and Warnow 2012, 2018)(Fig. 6).

231   First note that the argument above regarding duplications remains true and no new

232   duplications need be added by this completion. Now at each unary node, u, of $S||g$,

233   however, we attach only one instance of t to the $l \geqslant 1$ edges of g passing through u. The

²³⁴ remaining $l-1$ edges split at u but immediately are lost (and the interpretation of loss does

²³⁵ not matter). This guarantees that at most one edge of g will be present in t and therefore

²³⁶ the DC cost will not increase relative to what it was with respect to $S||g$, which means

$$c_{DC}(S, g) = c_{DC}(S||_g, g), \tag{8}$$

which, from Eq. 6, means

$$c_{DC}(S, g) = \sum_{e \in E(g)} k_{S||_g}(e) - |E(S||_g)|.$$

²³⁷   Note that this is not necessarily an optimal completion for *losses*, as it is fine to add

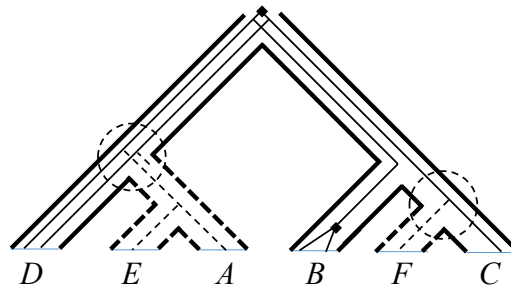²³⁸ losses in order to keep deep coalescences to a minimum.

Fig. 6. Optimal completion under the deep coalescence cost. Layout is the same as in Fig. 5. Here the optimal completion must add as few "extra edges" as possible (i.e., over and above one edge of g per edge of S), regardless of how many losses are then required. Note the loss of one edge of g immediately after its split within the unary node of S at lower left. Contrast with Fig. 5.

²³⁹   *Incompleteness due to deletion.—*   Here the absence of a gene tree leaf is assumed

²⁴⁰ to be caused by evolutionary deletion of that gene somewhere in the tree. With this

²⁴¹ assumption, we still have

$$c_{dup}^{-}(S, g) = c_{dup}^{-}(S|_g, g) = c_{dup}^{-}(S||_g, g), \tag{9}$$

²⁴² (Chauve et al. 2008; Górecki and Tiuryn 2006; Bayzid and Warnow 2018), where we add

²⁴³ the superscript $(^{-})$ to c to refer to the deletion interpretation of absence. Duplications are

²⁴⁴ unaffected by the choice of subtree operations, because $\mathcal{M}$ does not map nodes of g to

²⁴⁵ unary nodes of S, and these nodes are the only difference between $S|_g$ and $S||_g$.

246   But the same is not true for the loss cost (Bayzid and Warnow 2018). The correct

247 count of losses when incompleteness is due to deletion is computed by substituting S for

248 $S|_g$ in Theorem 3 of Bayzid and Warnow (2018). This computation uses lengths of paths

249 on S that run from some node $\mathcal{M}_S(u)$ to $\mathcal{M}_S(v)$, where u, v is a child and parent node on

250 g. Together, all paths in S of this type comprise the tree, $S||_g$, which means

$$c_{loss}^-(S, g) = c_{loss}^-(S||_g, g). \tag{10}$$

251 So, unlike the duplication cost, the loss cost is not necessarily the same under the two

252 interpretations of incompleteness. In fact, they differ exactly by which kind of subtree

253 operation is appropriate. Bayzid and Warnow (2018, Thm. 4) give an example in which

254 these different interpretations of loss lead to different optimal species trees when using the

255 loss cost.

256   Bayzid and Warnow (2018) also discussed the special case in which missing leaves

257 are distributed on the gene tree in a way that might imply a gene is absent from the root

258 of the species tree (Fig. 7). This is not a problem when incompleteness is assumed to arise

259 from sampling, but if it is assumed to arise from deletion, it is necessary to make an

260 assumption about whether genes are present at the root. Let $u^* = \mathcal{M}_S(r(g))$ be the

261 location on S of the root of g. If $r(S) > u^*$, should we assume the gene is present at $r(S)$?

262 In the above treatment, the definition of $S||_g$ solves the loss problem under the assumption

263 that a gene is unambiguously present only in the subtree of S having u^* as its root. To

264 make the stronger assumption that it is present at $r(S)$, we can define yet another subtree

265 operator, $S||_g^r$, which is the smallest subtree of S containing $r(S)$ and the leaf labels of g.

266 The original restriction operator, $S||_g$, can be interpreted in this problematic boundary

267 case to imply losses are due to deletion in the subtree of S having u^* as its root, and losses

268 are due to sampling elsewhere. We do not pursue this further here, but see Bayzid and

269 Warnow (2018).

270   The deep coalescence cost when incompleteness is assumed to arise from deletions

271 parallels the case for incompleteness due to sampling. On $S||g$, any unary node, u, has
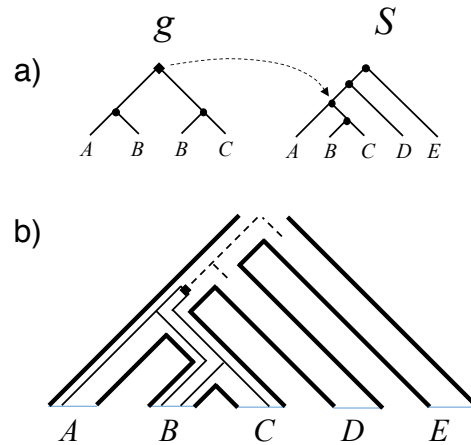
Fig. 7. Special case in which presence of gene at the root of the species tree is ambiguous for incompleteness due to deletion. a) Incomplete gene tree and species tree with MRCA mapping ($\mathcal{M}$) from root of g shown. b) Imbedding of g in S leaving uncertainty toward the root of S. Dashed lines represent presence of gene if it is forced present at root of S and location of two losses entailed as a consequence.

$l \geqslant 1$ edges of g passing through it. To explain the missing leaves present in the pendant subtree, t, of u that is present on S but not $S||g$, each of these l lineages must branch at u and end in a deletion somewhere in t. Moreover, at most one edge of g can persist in t, else there is a deep coalescence and the DC cost is increased. This lineage must end in a deletion before reaching a leaf node of S, and in fact it can be deleted immediately after splitting without changing the DC score, because if an edge of S has no imbedded edges of g, its DC score is still zero. In this respect the final form of g can differ from the optimal completion under sampling, but the net result is the same. All missing leaves can be accounted for without adding to the DC cost (though perhaps adding to the number of losses), and thus the DC cost is the same as for incompleteness due to sampling:

$$c_{DC}^-(S, g) = c_{DC}^-(S||_g, g) = c_{DC}(S||_g, g) \tag{11}$$

Reconciliation and Terraces

Based on the reconciliation costs for incomplete gene trees in Eqs. 7-11 above and Theorem 2, we can summarize which combinations of costs and interpretations of losses can lead to terraces (Table 1). First, recall that, from Theorem 3, for a given input, the

Type II span of trees is a subset of the Type I span. This implies that if a given reconciliation setting induces Type I terraces (so that scores are the same for all trees in its span), then all trees in the corresponding Type II span will also have the same score and be a Type II terrace. The reverse is not necessarily true.

In view of this, all four of the reconciliation costs can lead to Type II terraces and some, in addition, lead to Type I terraces. The duplication cost can lead to both kinds of terraces regardless of the interpretation of incompleteness. In contrast, the DC cost can only lead to Type II terraces (also regardless of the interpretation of incompleteness). For the loss cost, the nature of the assumption about incompleteness determines which type of terrace is possible.

DISCUSSION

Terraces in tree space can have several deleterious impacts on phylogenetic inference (reviewed in Sanderson et al. 2015), and the basic finding that terraces can arise in tree inference problems under very general conditions (Steel 2016) suggests some care is warranted. Loosely speaking, these conditions involve missing data and an optimality score that is decomposable into additive contributions from different subsets of the data. Originally, terraces were shown to arise in multilocus sequence alignments using maximum parsimony or maximum likelihood optimization scores (Sanderson et al. 2011). In the latter, certain subsets of models lead to terraces and others do not (Sanderson et al. 2011, 2015). Theory (Sanderson et al. 2010) predicted that terraces in these contexts are most likely for data sets with many taxa and few loci. A recent meta-analysis (Dobrin et al. 2018) confirms this expectation and indicates the number of trees on terraces can be very large.

It is no surprise that other classes of data sets and optimality criteria used in phylogenetics that also give rise to terraces. In this paper we explored how terraces can arise in gene tree reconciliation approaches to species tree inference. Here the gene trees

312 take the place of the multiple sequence alignments of individual loci studied in our earlier

313 work (Sanderson et al. 2011), and the optimality scores are additive functions of

314 reconciliation costs that depend on the processes assumed to connect gene trees and

315 species trees. These vary depending on whether processes of gene duplication, loss or

316 incomplete lineage sorting are assumed to occur.

317   In considering these divergent biological processes, we found a variant of our

318 original phylogenetic terrace that is relevant to some but not all of these processes. For

319 example, under the deep coalescence score, there can be terraces of the second type, but

320 not of the original type described in Sanderson et al. (2011). The composition of trees on a

321 terrace—the "span"—is determined by the pattern of missing data in the input. The span

322 of the trees on a terrace for the deep coalescence score is a subset of the span for the

323 original type of terrace (which would be a terrace for a different optimality score), and

324 therefore the size of this kind of terrace is always less than or equal to terraces of the first

325 type. This implies that the "problem" of terraces should be less for the deep coalescence

326 score than the duplication score, for example.

327   The phylogenomics literature contains relatively few studies that infer species trees

328 from gene trees subject to duplication and loss (Sanderson and McMahon 2007; Burleigh

329 et al. 2011), compared to those that infer species trees in the context of deep coalescence

330 (e.g. Copetti et al. 2017). For the latter, however, there are a variety of model based and

331 discrete algorithm approaches to inference (Liu et al. 2019), whereas our results are most

332 directly relevant to the simple method of minimizing the deep coalescence score (Maddison

333 1997; Ma et al. 2001; Zhang 2011; Nakhleh 2013). However, we expect that the general

334 idea of terraces will likely extend to certain model based methods of species tree inference

335 by analogy with how it extends to likelihood based inference from multi-locus sequence

336 alignments (Sanderson et al. 2011, 2015). It remains to be seen whether the ultimate

337 impact of this on species tree inference methods is as dramatic as it is for many large but

338 sparse multiple sequence alignments (Dobrin et al. 2018).

339   REFERENCES

340   Bansal, M., E. Alm, and M. Kellis. 2012. Efficient algorithms for the reconciliation problem

341   with gene duplication, horizontal transfer and loss. Bioinformatics 28:i283–i291.

342   Bayzid, M. S. and T. Warnow. 2012. Estimating optimal species trees from incomplete

343   gene trees under deep coalescence. Journal of Computational Biology 19:591–605.

344   Bayzid, M. S. and T. Warnow. 2018. Gene tree parsimony for incomplete gene trees:

345   addressing true biological loss. Algorithms for Molecular Biology 13:1–12.

346   Bravo, G. A., A. Antonelli, C. D. Bacon, K. Bartoszek, M. P. K. Blom, S. Huynh,

347   G. Jones, L. L. Knowles, S. Lamichhaney, T. Marcussen, H. Morlon, L. K. Nakhleh,

348   B. Oxelman, B. Pfeil, A. Schliep, N. Wahlberg, F. P. Werneck, J. Wiedenhoeft,

349   S. Willows-Munro, and S. V. Edwards. 2019. Embracing heterogeneity: coalescing the

350   Tree of Life and the future of phylogenomics. PeerJ 7.

351   Burleigh, J. G., M. S. Bansal, O. Eulenstein, S. Hartmann, A. Wehe, and T. J. Vision.

352   2011. Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 gene trees.

353   Systematic Biology 60:117–125.

354   Chauve, C., J. P. Doyon, and N. El-Mabrouk. 2008. Gene family evolution by duplication,

355   speciation, and loss. Journal of Computational Biology 15:1043–1062.

356   Chauve, C. and N. El-Mabrouk. 2009. New Perspectives on Gene Family Evolution: Losses

357   in Reconciliation and a Link with Supertrees vol. 5541 of *Lecture Notes in Computer*

358   *Science* Pages 46–+.

359   Copetti, D., A. Búrquez, E. Bustamante, J. L. M. Charboneau, K. L. Childs, L. E.

360   Eguiarte, S. Lee, T. L. Liu, M. M. McMahon, N. K. Whiteman, R. A. Wing, M. F.

361   Wojciechowski, and M. J. Sanderson. 2017. Extensive gene tree discordance and

362   hemiplasy shaped the genomes of North American columnar cacti. Proc. Natl. Acad.

363   Sci., USA 114:12003–12008.

364 Dobrin, B. H., D. J. Zwickl, and M. J. Sanderson. 2018. The prevalence of terraced

365   treescapes in analyses of phylogenetic data sets. BMC Evolutionary Biology 18.

366 Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Press, Sunderland, MA.

367 Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romeroherrera, and G. Matsuda. 1979.

368   Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by

369   cladograms constructed from globin sequences. Systematic Zoology 28:132–163.

370 Górecki, P. and J. Tiuryn. 2006. DLS-trees: A model of evolutionary scenarios. Theoretical

371   Computer Science 359:378–399.

372 Liu, L., C. Anderson, D. Pearl, and S. Edwards. 2019. Modern phylogenomics: Building

373   phylogenetic trees using the multispecies coalescent model. Methods in Molecular

374   Biology 1910:211–239.

375 Ma, B., M. Li, and L. Zhang. 2001. From gene trees to species trees. SIAM J. Comput.

376   30:729–752.

377 Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46:523–536.

378 Nakhleh, L. 2013. Computational approaches to species phylogeny inference and gene tree

379   reconciliation. Trends in Ecology and Evolution 28:719–728.

380 Page, R. D. M. 1994. Maps between trees and cladistic analysis of historical associations

381   among genes, organisms and areas. Systematic Biology 43:58–77.

382 Page, R. D. M. and M. A. Charleston. 1997. From gene to organismal phylogeny:

383   Reconciled trees and the gene tree/species tree problem. Molecular Phylogenetics and

384   Evolution 7:231–240.

385 Sanderson, M., M. M. McMahon, A. Stamatakis, D. J. Zwickl, and M. Steel. 2015. Impacts

386   of terraces on phylogenetic inference. Syst. Biol. 64:709–726.

387 Sanderson, M. J. and M. M. McMahon. 2007. Inferring angiosperm phylogeny from EST
388   data with widespread gene duplication. BMC Evolutionary Biology 7 (Suppl. 1):S3.

389 Sanderson, M. J., M. M. McMahon, and M. Steel. 2010. Phylogenomics with incomplete
390   taxon coverage: the limits to inference. BMC Evolutionary Biology 10.

391 Sanderson, M. J., M. M. McMahon, and M. Steel. 2011. Terraces in phylogenetic tree
392   space. Science 333:448–450.

393 Steel, M. 2016. Phylogeny: Discrete and Random Processes in Evolution. SIAM,
394   Philadelphia.

395 Zhang, L. X. 2011. From gene trees to species trees II: Species tree inference by minimizing
396   deep coalescence events. IEEE-ACM Transactions on Computational Biology and
397   Bioinformatics 8:1685–1691.

Table 1. *Reconciliation Costs and Possible Terrace Types*

Cost/Interpretation of loss	Sufficient Conditions for Terraces of Type
Duplication/Sampling	I,II
Duplication/Deletion	I,II
Loss/Sampling/	I,II
Loss/Deletion	II
D+L/Sampling	I,II
D+L/Deletion	II
DC/Sampling	II
DC/Deletion	II