Supplemental Material for Aggregated Dendrograms for Visual Comparison Between Many Phylogenetic Trees

Zipeng Liu, Shing Hei Zhan, and Tamara Munzner

December 3, 2018

Contents

S1 View coordination details

Table S1 documents the view coordination discussed in Section 6.6 of the main paper, showing which aspect of the data is visually encoded across all five levels of detail for each of the eight views. The table has six columns since we break out branches from leaves for clarity; both of these are the lowest level of detail. We duplicate Figure 7 and put it here for a quick reference showing all of these views, as Figure S1.

View	Level of detail (LoD)					
	Tree collection	Subset of trees	Individual tree	Subtree	Branch & its attributes	Leaf node
Reference Dendrogram			whole view	color background	line & tooltip	text label & black dot
Tree Distribution	row	segment in row				
Cluster AD	whole view	one cluster				
Individual AD			one AD	block	line or collapsed	
Pairwise comparison		consensus tree	butterfly layout	color background	line & tooltip	text label & black dot
Tree Similarity	t-SNE scatterplot		dot			
Ref. Br. Attr & Corr. Br. Attr.					attribute table; histogram	
Tree List			text label			

Table S1: View coordination of visual encoding across all levels of detail and all views.

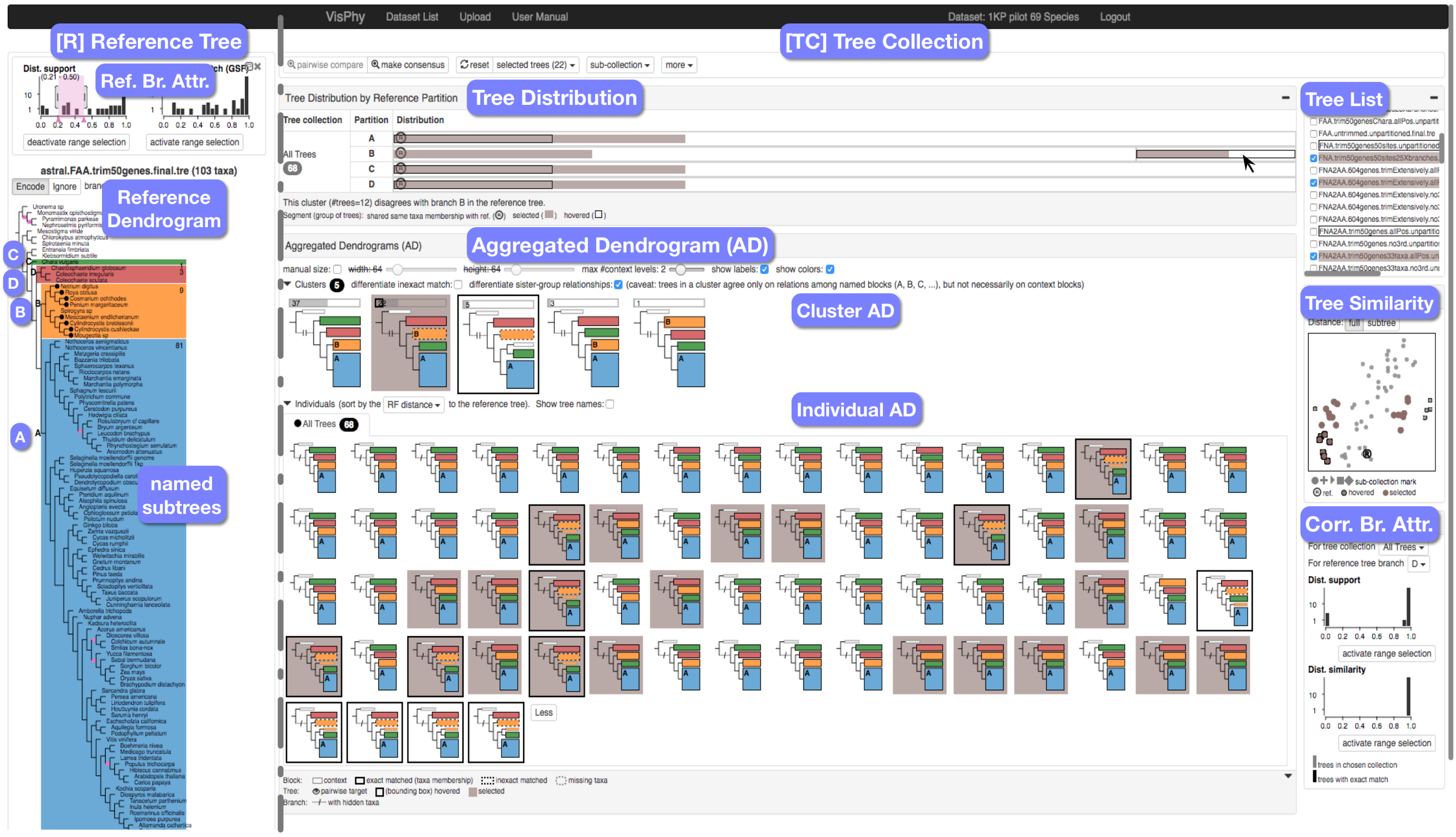


Figure S1: A screenshot of ADView as a reference to Table S1.

S2 Cluster AD gradient coloring

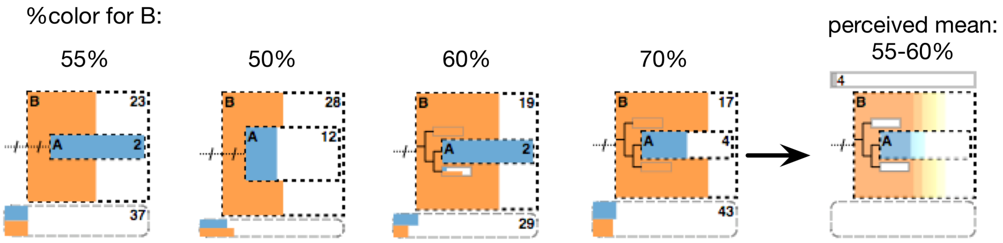


Figure S2: Illustration of gradient coloring in a cluster AD. The percentage represents the proportion of highlighted taxa in a block, and is encoded as a solid color in the individual AD, but gradient color in the resulting cluster AD.

When grouping topologically identical individual ADs into a cluster AD, we use one of the AD layouts as a proxy for the cluster AD, but use gradient color to convey the uncertainty of the proportion of taxa instead of the solid color in an individual AD. The percentage of the color fill represents the proportion of highlighted taxa in an AD block, which is fixed for individual AD but typically covers a variable range within a cluster AD. As shown in Figure S2, block B has different proportions of orange color in the four individual ADs on the left, which can be considered as a distribution [55%, 50%, 60%, 70%]. In the cluster AD on the right, the orange in block B has a fuzzy edge, which is perceived as the variance or uncertainty of the color proportion.

We encode the complementary cumulative distribution function (c-CDF) of the percentage of color fill with color saturation to achieve the fuzzy visual effect, which was first introduced as density strip by Jackson [1]. We do not choose the probability density function (PDF), which is usually rendered as a histogram, nor the cumulative distribution function (CDF), because both of them could be misleading when mapped to color saturation in our context, as illustrated in Figure S3.

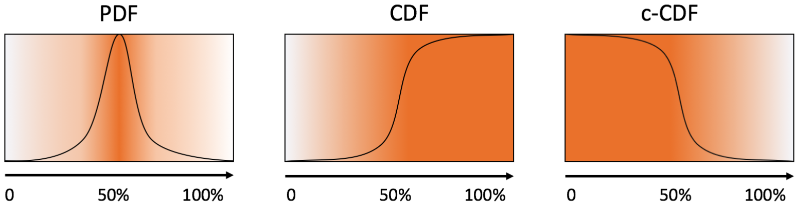


Figure S3: Illustration of different encoding representation for the same normal distribution (mean=50%, standard deviation is small). Color saturation is mapped to the value of the probability function shown inside the block. Probability functions from left to right are the density function, the cumulative distribution function, and the complementary cumulative distribution function.

S3 Algorithm details

Besides the algorithms we described in the paper, in this section we would like to supplement more detail about the aggregated dendrograms generation and rendering for the purpose of replication.

S3.1 AD layout function parameters

Figure S4 illustrates the parameters used in generating AD layouts. We use a best-effort mechanism to adapt the AD layout to user-specified resolution, described in the algorithm section of the main paper. If an AD layout does not pass our legibility test, we will shrink some of the flexible parameters and re-generate a layout.

There are two cascading sets of flexible parameters that can be changed: 1) the number of context levels, a metric to control how many context blocks to show; 2) inter-block gaps, branch lengths, block sizes.

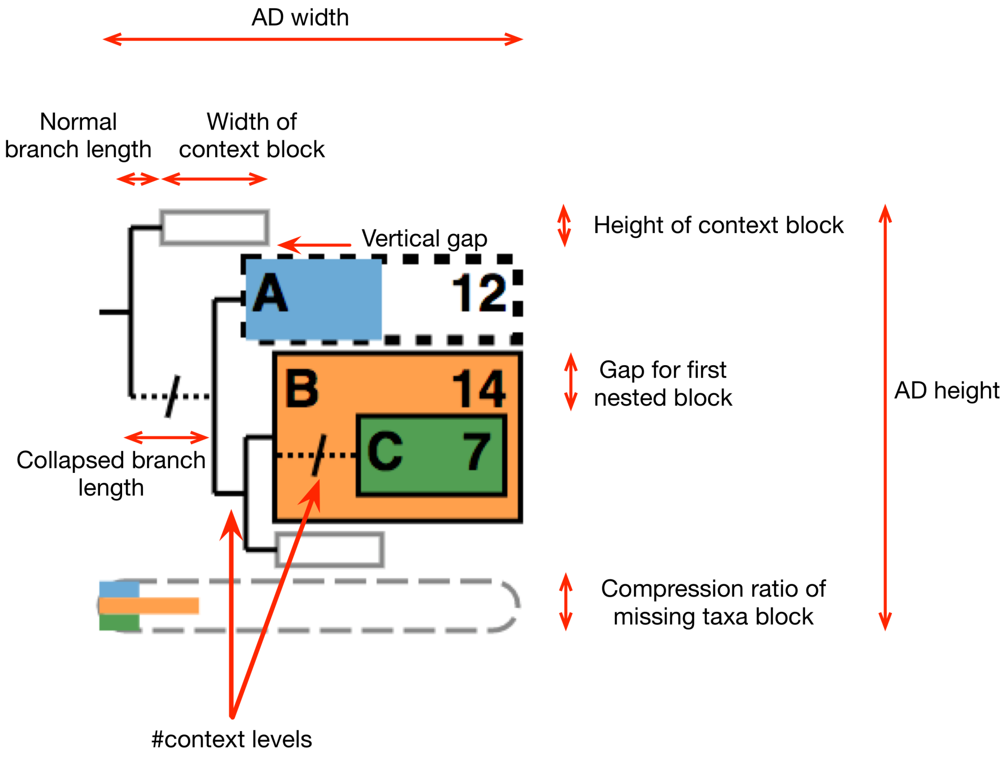


Figure S4: Illustration of parameters in an aggregated dendrogram layout.

S3.2 Front-end caching for rendering ADs

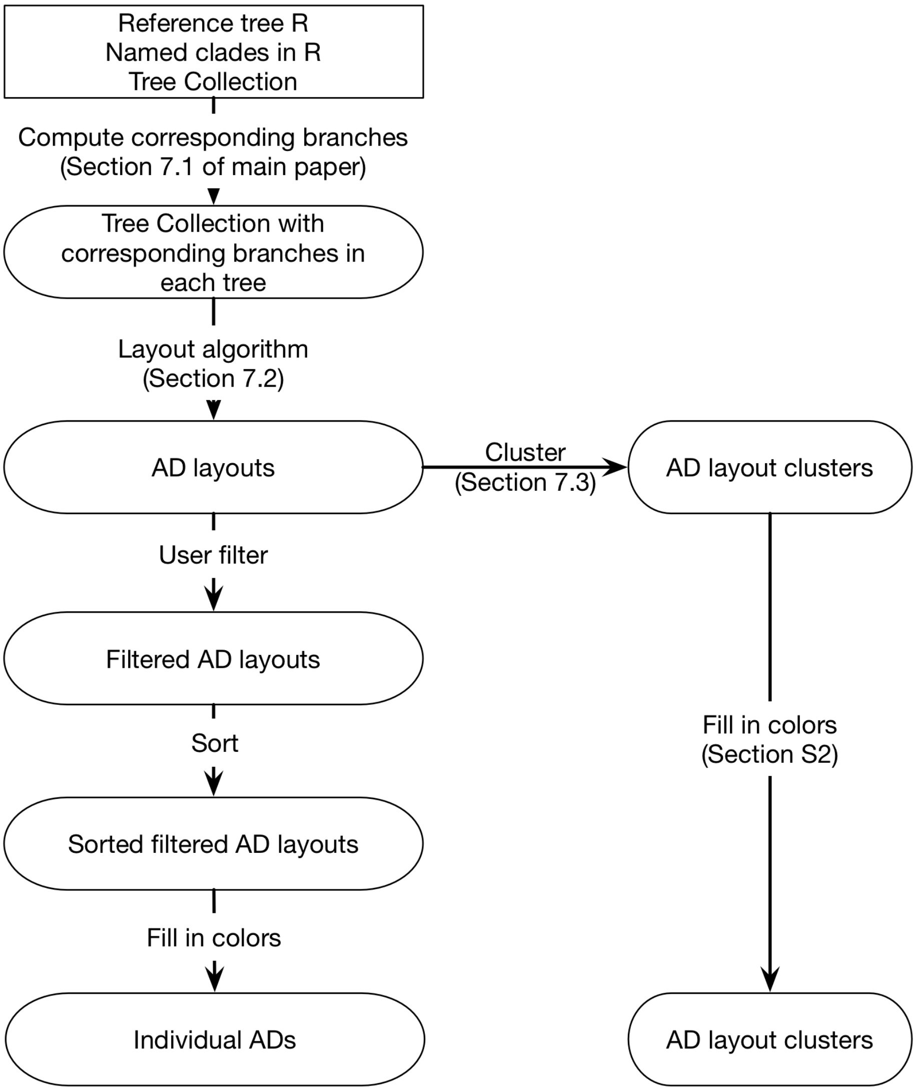


Figure S5: Pipeline on the frontend to render the aggregated dendrograms. The results in ovals are cached so that certain kinds of frequent user interaction can take place without triggering re-computation of intermediate results that remain useable.

The rendering of aggregated dendrograms is handled by the front end; that is, the browser. It is not a trivial process, as shown in Figure S5, and there might be hundreds of ADs to compute. Therefore, to achieve reasonable response time to user interactions, especially for frequent ones, such as hovering over an AD, we cannot afford to run the whole pipeline from start to end.

Notice that many frequent interactions do not affect the layout of ADs: for example, hovering only draws a black border around the AD, and resorting only affects the order of ADs. We can cache the intermediate results, namely, the ovals in Figure S5, so that a user interaction only triggers re-computation of the necessary steps reusing some of the cached results.

We use ReactJS [1] and Redux [2] as our rendering and interaction pipeline, and the memoization functionality provided by the Reselect library [3]. The memoization keeps an internal mapping between the state of data elements specified by the developer and the computation results such as AD layouts. It detects if the state changes, that is, if the state object is different from

[1] https://reactjs.org/

[2] https://redux.js.org/docs/introduction/

[3] https://github.com/reactjs/reselect

the previous one, and determines to whether reuse stored results or trigger a re-computation. In Figure S5, selecting a subtree in the reference tree leads to a re-computation of the whole pipeline because it changes the corresponding branches; changing the sorting order of AD will reuse the results of filtered AD layouts.

S4 Expert Interview Study

S4.1 Participants

Here, we summarize the relevant information of all participants and their datasets we recruited.

Participant	P1	P2	P3	P4a, P4b
Position	PhD student	Bioinformatician	Principal Investigator	2 PhD students
In-person interview	Yes	Yes	Skype	Yes
Interviewers	ZP,SZ,TM	ZP,SZ,TM	ZP,SZ	ZP,SZ
Previous experience with ADView	chauffeured demo 2 months ago	chauffeured demo 8 months ago	screenshots	None
Own dataset	Yes	Yes	Yes	No; use multiple public datasets
Type of comparison	Species tree vs gene trees	Consensus tree vs bootstrapping trees	Species tree vs gene trees	Species tree vs gene trees; Consensus tree vs bootstrapping trees
Number of Trees	260	100	71	dozens to hundreds
Type of organisms	Plants	Parasites	Algae	Human pathogens; parasites

Table S2: Relevant information of participants and datasets used in the study.

S4.2 Interview Questions

1. Before today, how long have you been analyzing this dataset and with what tools? Did you generate it yourself or get it from someone else?

2. Did you find any interesting biological insights in this dataset?

   (a) Could you confirm things that you already knew? And how long did it take to see this compared to other tools?

   (b) Did you notice anything new in this dataset?

3. What capabilities of the tool are useful for your research?

4. Is there any functionality that is missing from the tool that would be useful?

5. Are there aspects of the tool that are confusing, misleading or awkward?

6. In our corresponding branch matching algorithm, we currently compute the similarity metrics between every pair of branches and find the most similar branch in the tree to the one in the reference tree.

   (a) Does this computation make sense to you? How closely does it match your mental model of how you compare two trees? Is there some other way of thinking about support and conflicts to a specific clade in the reference tree that is an alternative to this kind of matching?

   (b) Our corresponding branch matching algorithm assumes that all trees have a root. We know this assumption breaks in many cases, including when the outgroup taxa are missing or you have a non-monophyletic outgroup. Here's an example of a tree where the root is wrong and you can see that the matching is very messy. Which clade do you think is the real match to A here? Or do you think that this question doesn't make sense and we shouldn't even try to find a match in this case?

   (c) Do you think we should differentiate between the case where two clades are very similar (although not exactly the same) from the case where two clades are very different? At some point is it no longer useful to show a "best match" past some cutoff value of "too different"?

7. Does the visual design of AD make sense to you and match your mental model of how you think of a phylogenetic tree? The same problem with incorrect rooting affects the visual layout of the ADs. Do you think an AD with incorrect rooting could mislead you?

8. Does the tree distribution view make sense to you and match your mental model of how you think about agreement and conflict between clades respectively?
   The tree distribution view sometimes has a very long tail with many sets that contain only a single tree. Is seeing all of these tiny sets helpful in terms of conveying any interesting information to you, or would it be better to leave that out? Do you have thoughts about whether it's conveying information about the true biology, or about systematic errors in tree reconstruction, or whether it could just be an artifact of the matching computation in our own software?

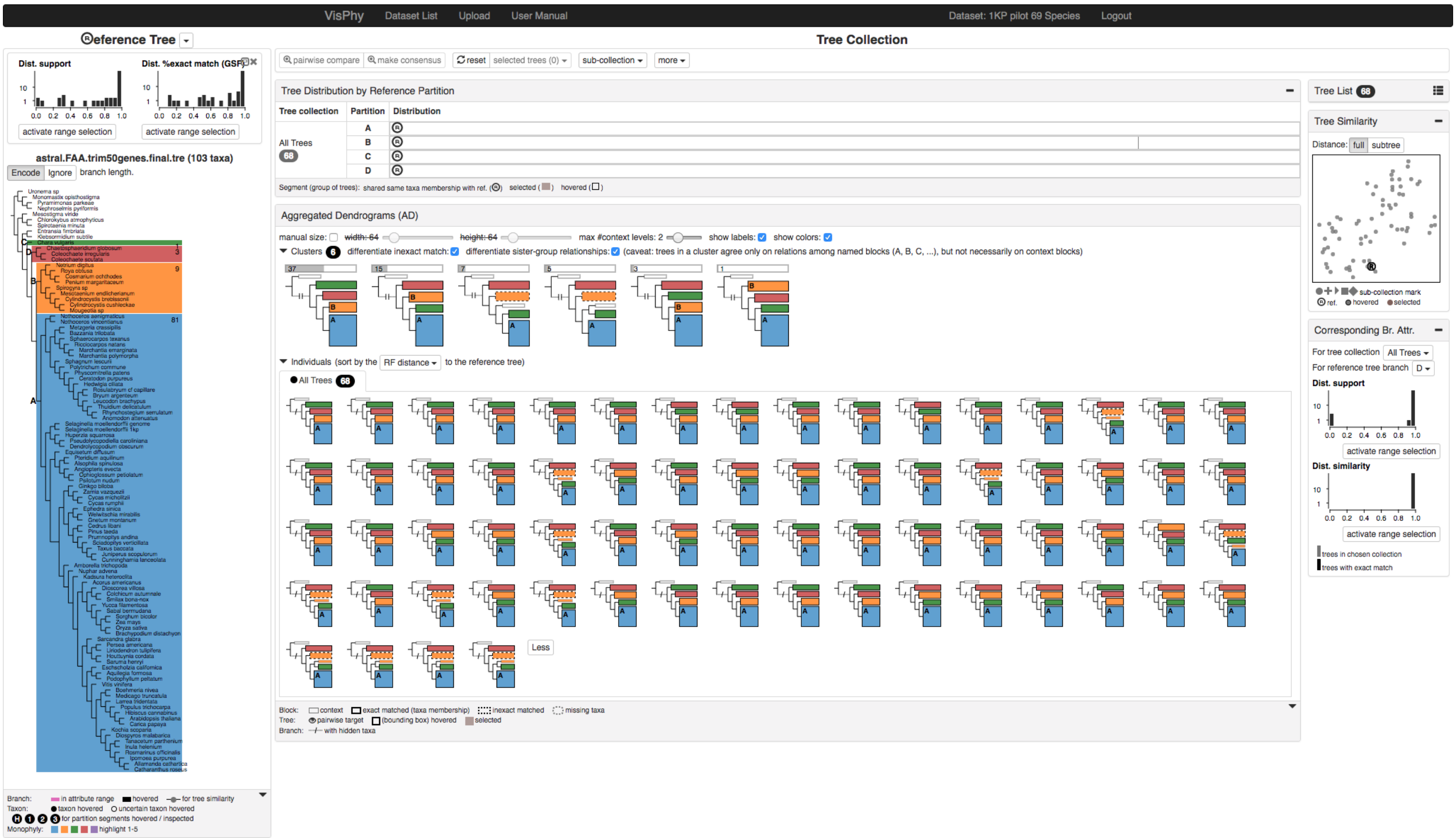
9. Do think this tool (or an improved version of it) could be helpful for other scenarios such as horizontal gene transfer? What kind of scenarios do you think would be suitable for this tool?
   How often would you find a tool like this useful in the work that you do? (For example, every day, once a month, once a year. . . ?) When in a project life cycle might you use something like this - at the beginning of a new project, after a specific phase of it, just before paper writing. . . ?)

10. Do you have other comments about the tool?

S5 Full Screenshots of Usage Scenario 1: 1KP pilot study

Figures S6, S8, S9 and S11 are the full screenshots of our first usage scenario in the main paper. We explored the two research questions in the 1KP pilot study [2] in ADView.

S5.1 Sister group of land plants



Tree collection	Partition	Distribution
All Trees 68	A	Ⓡ
	B	Ⓡ
	C	Ⓡ
	D	Ⓡ

Segment (group of trees): shared same taxa membership with ref. (Ⓡ) selected (■) hovered (□)

manual size: □ ~~width: 64~~ ~~height: 64~~ max #context levels: 2 show labels: ☑ show colors: ☑

Clusters 6 differentiate inexact match: ☑ differentiate sister-group relationships: ☑ (caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)

Individuals (sort by the RF distance to the reference tree)

Block: □ context □ exact matched (taxa membership) ⬚ inexact matched ◌ missing taxa
Tree: ● pairwise target □ (bounding box) hovered ■ selected
Branch: ┈/┈ with hidden taxa

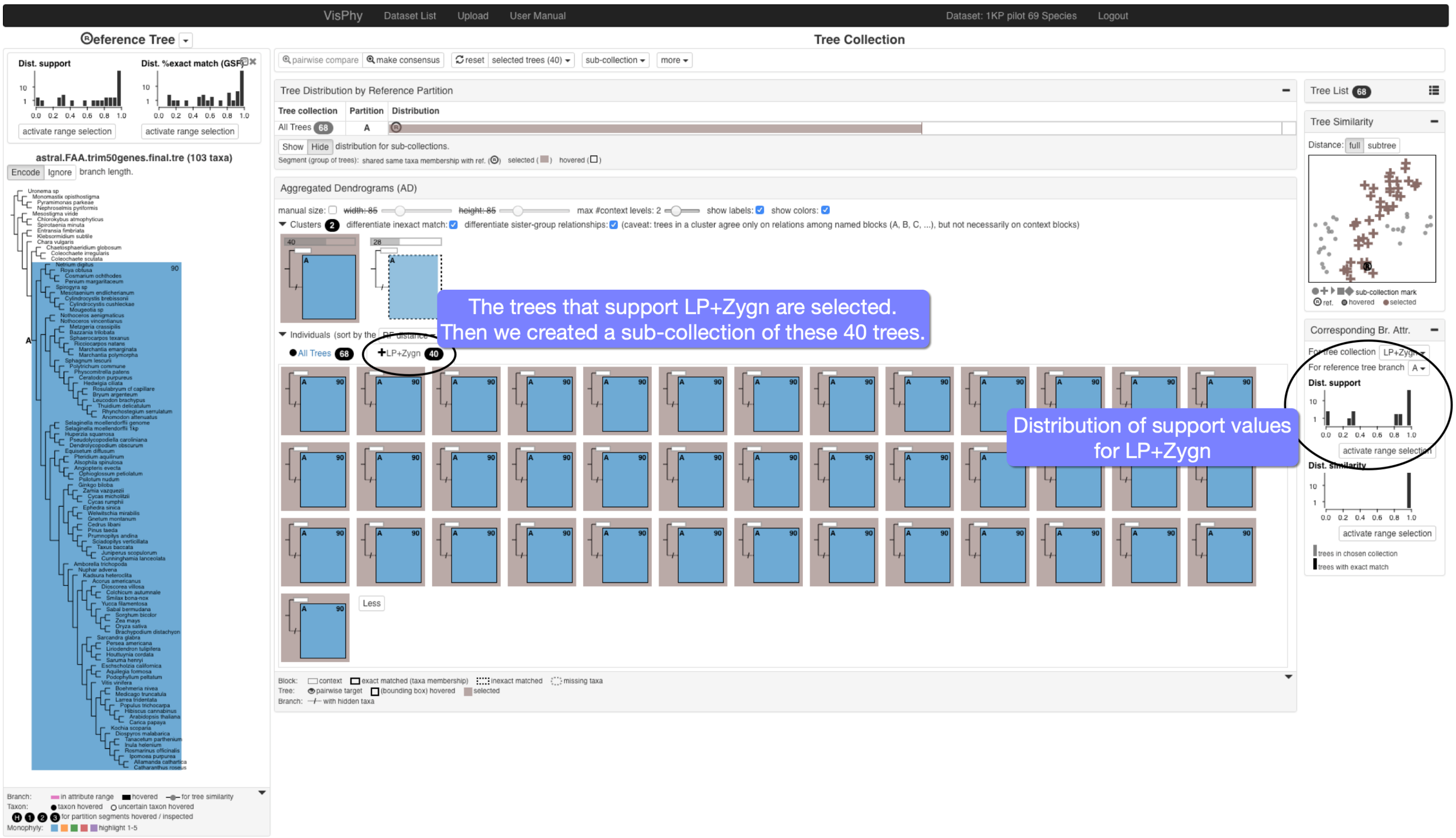
Branch: ▬ in attribute range ■ hovered —●— for tree similarity
Taxon: ● taxon hovered ○ uncertain taxon hovered
H 1 2 3 for partition segments hovered / inspected
Monophyly: 🟦 🟧 🟩 🟥 🟪 highlight 1-5

●+▸■◆ sub-collection mark
Ⓡ ref. ● hovered ● selected

For tree collection All Trees
For reference tree branch D

▮ trees in chosen collection
▮ trees with exact match

Figure S6: After selecting four focal clades: A (blue): LAND PLANTS (LP), B (orange): ZYGNEMATOPHYCEAE (ZYGN), C (green): CHARALES (CHAR), D (red): COLEOCHAETALES (COL). We compared the cluster ADs with the previous hypotheses to investigate which one (orange or green or red group) is the sister of LAND PLANTS.



The trees that support LP+Zygn are selected. Then we created a sub-collection of these 40 trees.

Distribution of support values for LP+Zygn

Figure S7: We checked the distribution of support values for the most popular hypothesis (LP + ZYGN) by selecting the trees that have the LP + ZYGN clade and created a sub-collection out of them. In the *Corresponding Branch* view, some trees have low support values for this hypothesis, which may render some doubts on whether LP + ZYGN is truly strongly supported by this dataset.

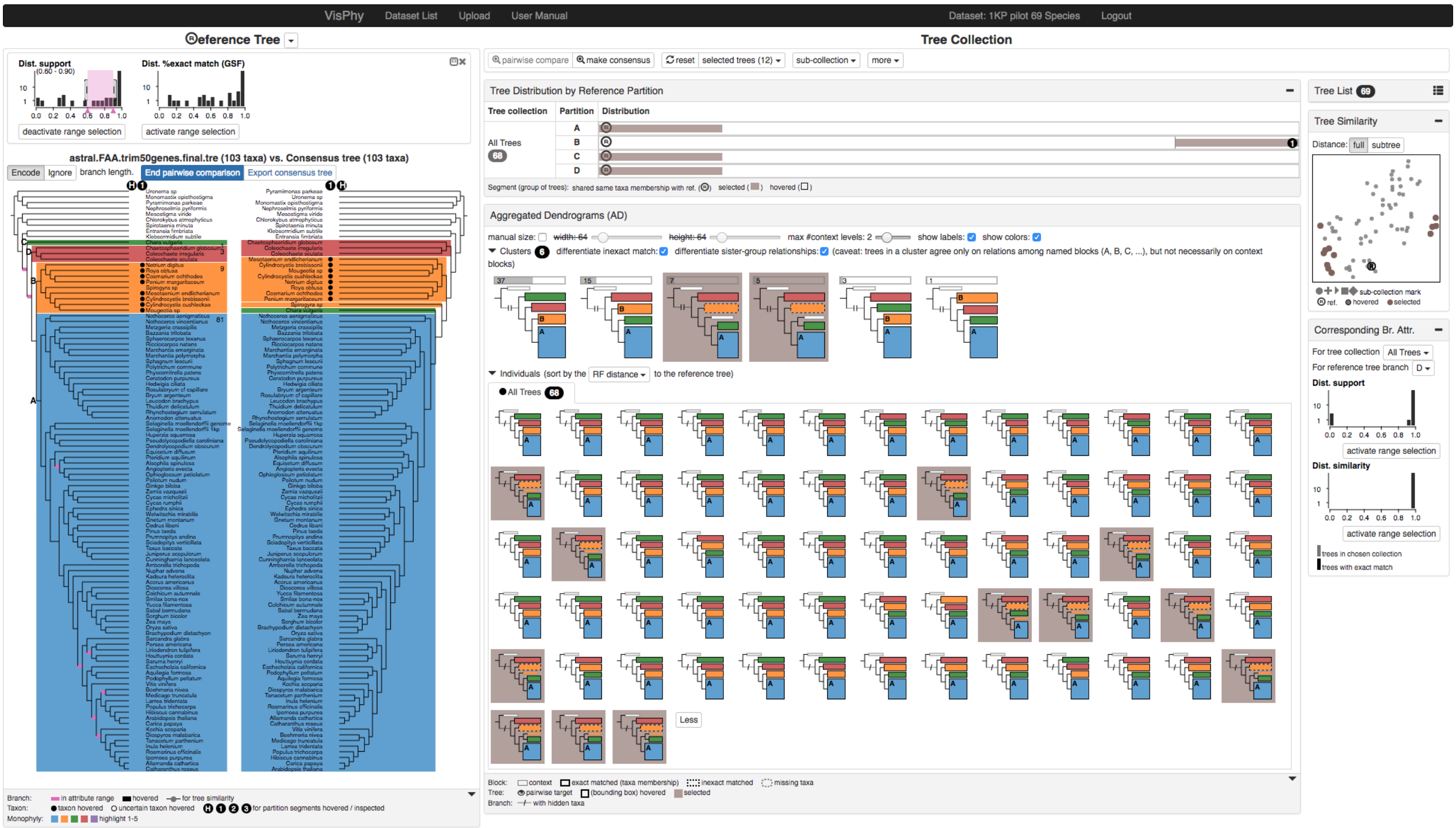
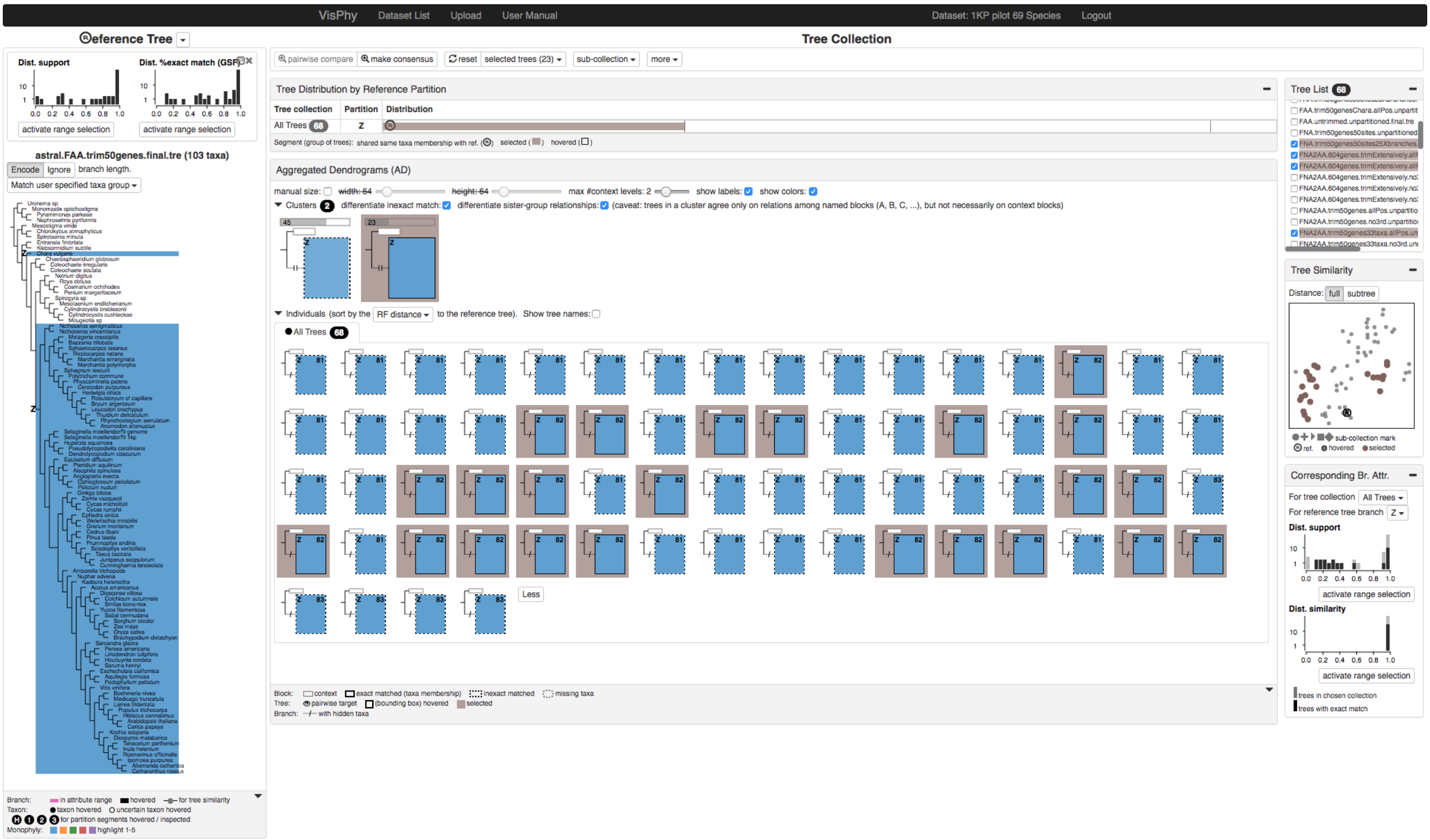

VisPhy
Dataset List
Upload
User Manual
Dataset: 1KP pilot 69 Species
Logout
Reference Tree
Tree Collection
Dist. support
Dist. %exact match (GSF)
deactivate range selection
activate range selection
astral.FAA.trim50genes.final.tre (103 taxa) vs. Consensus tree (103 taxa)
Encode
Ignore
branch length
End pairwise comparison
Export consensus tree
pairwise compare
make consensus
reset
selected trees (12)
sub-collection
more
Tree Distribution by Reference Partition
Tree collection
Partition
Distribution
All Trees
68
A
B
C
D
Segment (group of trees): shared same taxa membership with ref. selected hovered
Aggregated Dendrograms (AD)
manual size:
width: 64
height: 64
max #context levels: 2
show labels:
show colors:
Clusters
6
differentiate inexact match:
differentiate sister-group relationships:
(caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)
Individuals (sort by the RF distance to the reference tree)
All Trees 68
Less
Tree List 69
Tree Similarity
Distance: full subtree
sub-collection mark
ref.
hovered
selected
Corresponding Br. Attr.
For tree collection All Trees
For reference tree branch D
Dist. support
Dist. similarity
activate range selection
trees in chosen collection
trees with exact match
Block: context exact matched (taxa membership) inexact matched missing taxa
Tree: pairwise target (bounding box) hovered selected
Branch: with hidden taxa
Branch: in attribute range hovered for tree similarity
Taxon: taxon hovered uncertain taxon hovered for partition segments hovered / inspected
Monophyly: highlight 1-5

Figure S8: We selected the conflicting segment in the second row (B) in tree distribution (where trees shown in brown background) and made a consensus tree, which is being compared head-to-head against the reference tree in full details. We presented the markers for the selected trees in the butterfly dendrograms, which showed an outlier species SPIROGYRA is not included in the orange (ZYGN) group. Notice that selected trees are highlighted across multiple views in the interface.



astral.FAA.trim50genes.final.tre (103 taxa)

Encode | Ignore branch length.

Tree collection	Partition	Distribution
All Trees 68	Z	

Segment (group of trees): shared same taxa membership with ref. (Ⓡ) selected (▨) hovered (□)

manual size: □ ~~width: 64~~ ~~height: 64~~ max #context levels: 2 show labels: ☑ show colors: ☑

Clusters 2 differentiate inexact match: ☑ differentiate sister-group relationships: ☑ (caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)

Individuals (sort by the RF distance to the reference tree). Show tree names: □

Block: □ context ■ exact matched (taxa membership) ⬚ inexact matched ◌ missing taxa
Tree: pairwise target □ (bounding box) hovered ▨ selected
Branch: ┄/┄ with hidden taxa

Tree List 68
- □ FAA.trim50genesChara.allPos.unpartit
- □ FAA.untrimmed.unpartitioned.final.tre
- □ FNA.trim50genes50sites.unpartitioned
- ☑ FNA.trim50genes50sites25Xbranches.
- ☑ FNA2AA.604genes.trimExtensively.allP
- ☑ FNA2AA.604genes.trimExtensively.allP
- □ FNA2AA.604genes.trimExtensively.no3
- □ FNA2AA.604genes.trimExtensively.no3
- □ FNA2AA.604genes.trimExtensively.no3
- □ FNA2AA.trim50genes.allPos.unpartitio
- □ FNA2AA.trim50genes.no3rd.unpartitior
- ☑ FNA2AA.trim50genes33taxa.allPos.un
- □ FNA2AA.trim50genes33taxa.no3rd.un

● + ▶ ■ ◆ sub-collection mark
Ⓡ ref. ● hovered ● selected

For tree collection All Trees
For reference tree branch Z

▮ trees in chosen collection
▮ trees with exact match

Branch: ▬ in attribute range ■ hovered –●– for tree similarity
Taxon: ● taxon hovered ○ uncertain taxon hovered
H 1 2 3 for partition segments hovered / inspected
Monophyly: 🟦 🟧 🟩 🟥 🟪 highlight 1-5

Figure S9: We combined LP and CHAR into a user-specified taxa group. This feature is used for exploring hypotheses that are not presented in the reference tree, for example that there is no single subtree that consists of LP and CHAR in this reference tree. The second cluster we selected represent the subset of trees that group LP and CHAR together (solid border indicates exact matches). With this, we checked the support values for their corresponding branches (the overarching branch above LP + CHAR) in the corresponding branch attribute view on the bottom right. We found that there are a lot of low-support corresponding branches. In other words, some trees are not certain on grouping LP and CHAR together although they appear so.

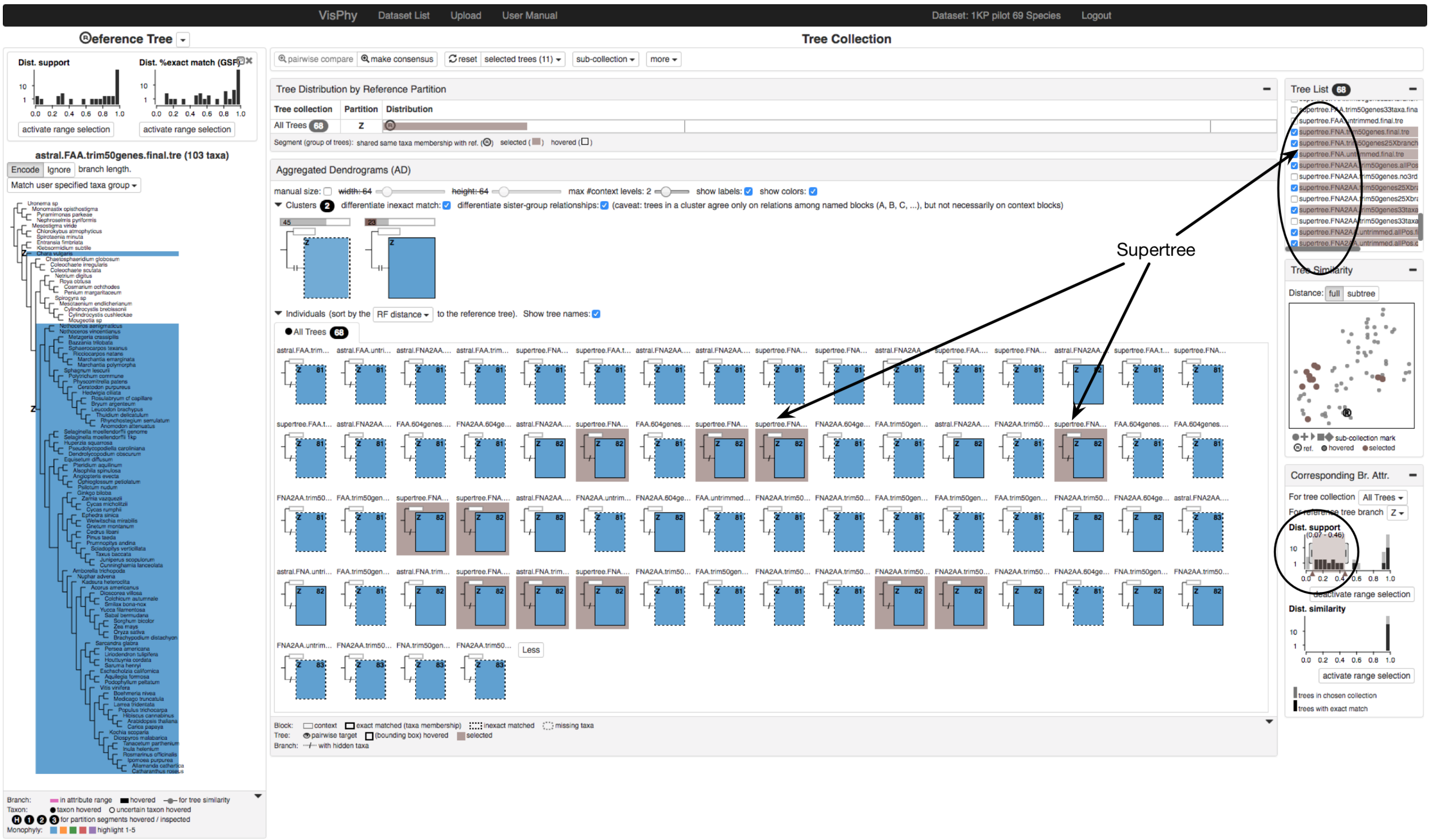


Figure S10: We noticed that some trees have low support values (below 0.5), so we selected these trees using the *Corresponding Branch Attribute* view, shown in the black circle at the bottom right. By displaying the tree names above the ADs, we found that most of them are generated with the "Supertree" method, which might be worth investigation later.

S5.2 Early diversification of land plants

In Figure S11, it is easy to find out that the monophylies of the four early lineages of LAND PLANTS (HORNWORTS, MOSSES, LIVERWORTS, and VASCULAR PLANTS) are strongly supported because almost all trees agree with the reference tree in the *Tree Distribution* view. The widely accepted hypothesis *Lv-basal* in Figure S12, that is, LIVERWORTS is the sister-group of all other LAND PLANTS, is rejected by most of the trees in this dataset. We notice that only two trees (the 4th and 6th cluster AD) support *Lv-basal*.

According to the 1KP paper, the widely accepted view that LIVERWORTS, MOSSES, and HORNWORTS are, respectively, successive sister groups to VASCULAR PLANTS, are not recovered in this dataset. In ADView, we can see that there are no such trees presenting this topology: (blue, (orange, (green, red))), which is a direct evidence to support the statement in the 1KP paper. The first cluster AD is exactly the *Hw-basal* hypothesis in Figure S12, and the second cluster AD is compatible with the *Bryo monophyletic group* hypothesis.

Biologists can then connect this evidence with their domain knowledge such as what substitution models are used to generate these trees and analyze their pros and cons. ADView presents the relevant information to them, in hopes that they interpret the visualization with biological judgment.

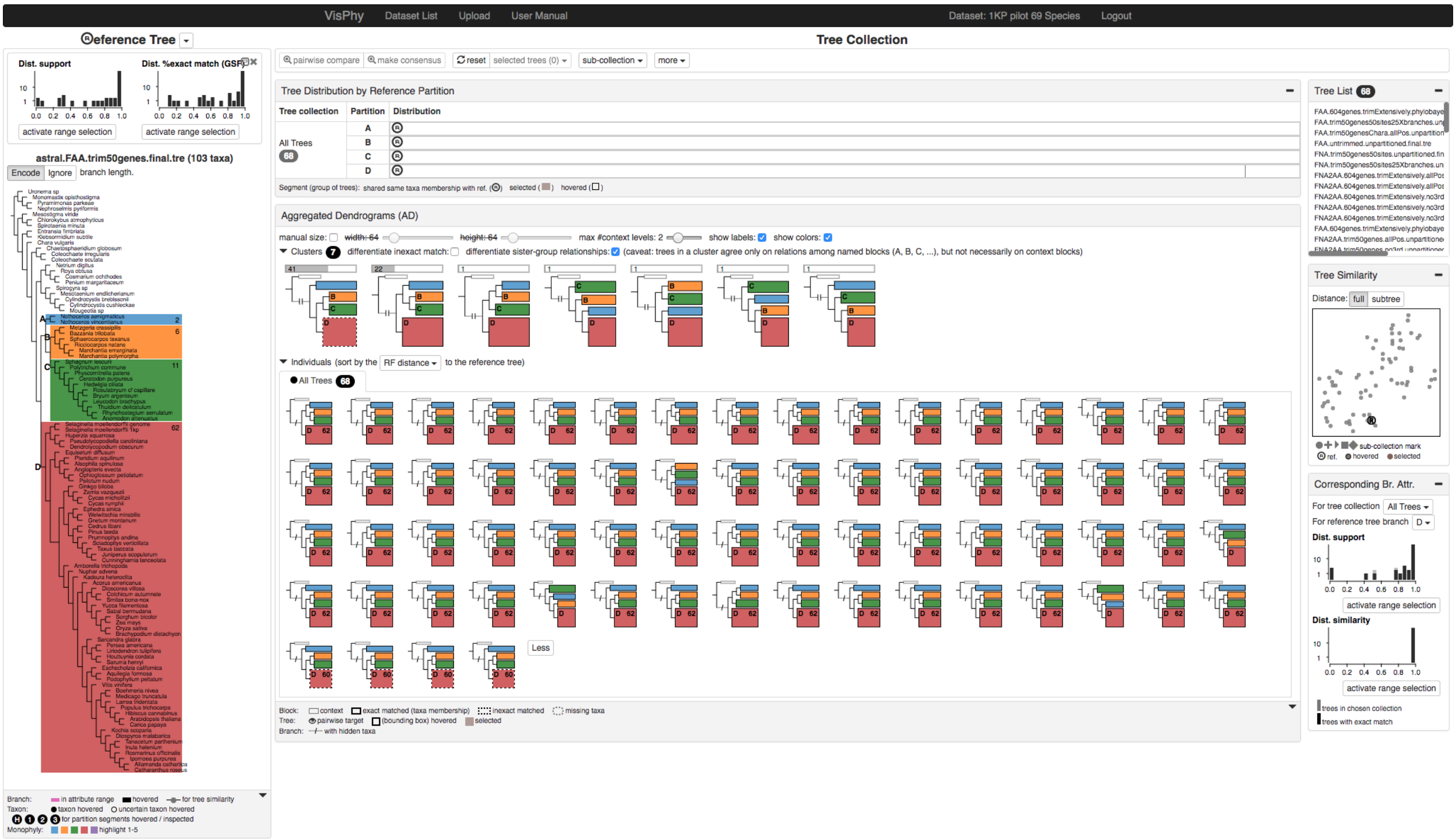


Figure S11: Screenshot of ADView exploring early lineage of LAND PLANTS. A (blue): HORNWORTS (Hw); B (orange): MOSSES (Mo); C (green): LIVERWORTS (Lv); D (red): VASCULAR PLANTS (VP).

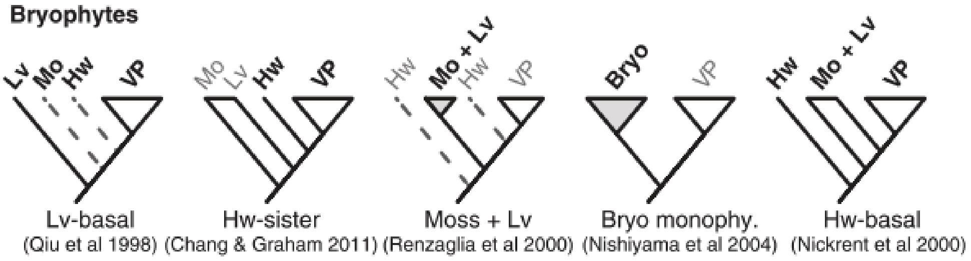


Figure S12: Previous hypotheses about early diversification of LAND PLANTS. Figure excerpted from the 1KP paper [2].

S6 Usage scenario 2: TreeFam

TreeFam [3] is a database of animal gene trees built from the genome sequences of representative animal species. The data set spans the entire evolutionary history of the phylum ANIMALIA. In this case, we utilized the TreeFam data to demonstrate how ADView may be used to discover gene trees that are concordant with well-established evolutionary relationships. More specifically, we examined and identified gene trees consistent with two widely accepted splits in animal taxonomy: PROTOSTOMIA and DEUTEROSTOMIA [4]; ECDYSOZOA and LOPHOTROCHOZOA [5]. Both of these views have received support from multi-gene phylogenetic and phylogenomic studies [6, 7].

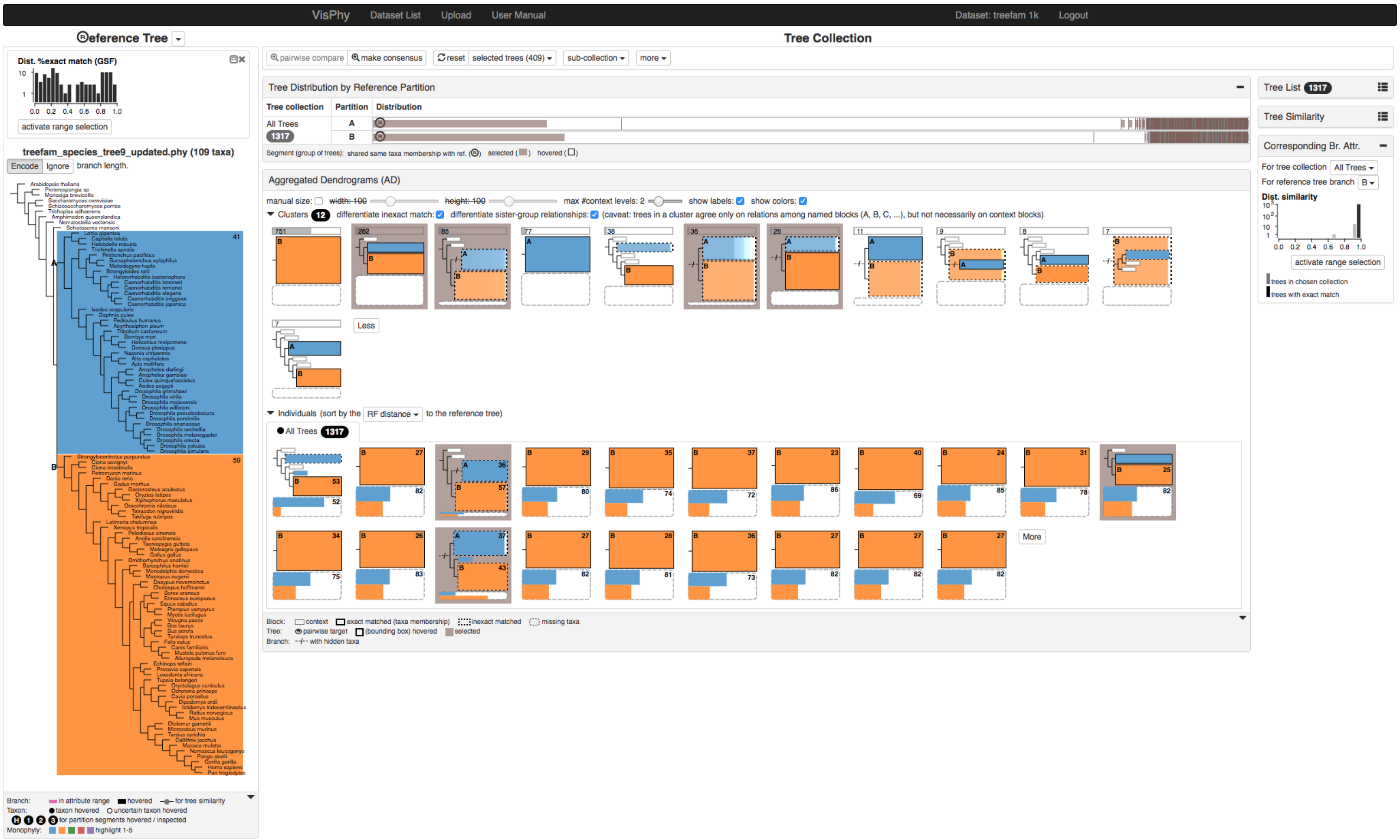
Since the initial publication in 2006, there were several updates to TreeFam. Here, we used the latest release (4.0), which includes genome sequence data from 108 animal species and one outgroup plant species (ARABIDOPSIS THALIANA). We downloaded the individual gene trees as well as a species tree that captures major evolutionary relationships in ANIMALIA from `http://treefam.genomics.org.cn/`. Because ADView cannot yet handle duplicate genes, we excluded gene trees with duplicate genes. Also, we only included gene trees with at least 20% of the taxa represented. A final set of 1,317 gene trees was taken as input to ADView.

Note that we collapsed the *Individual AD* view when we were only looking at the *Cluster AD* and the *Tree Distribution* view, because there are more than 500 individual ADs. Using only the *Cluster AD* view led to much better responsiveness to user interaction because there were so many fewer visual elements to handle. Also, this choice kept the full screenshots from being extremely long.

S6.1 PROTOSTOMIA and DEUTEROSTOMIA

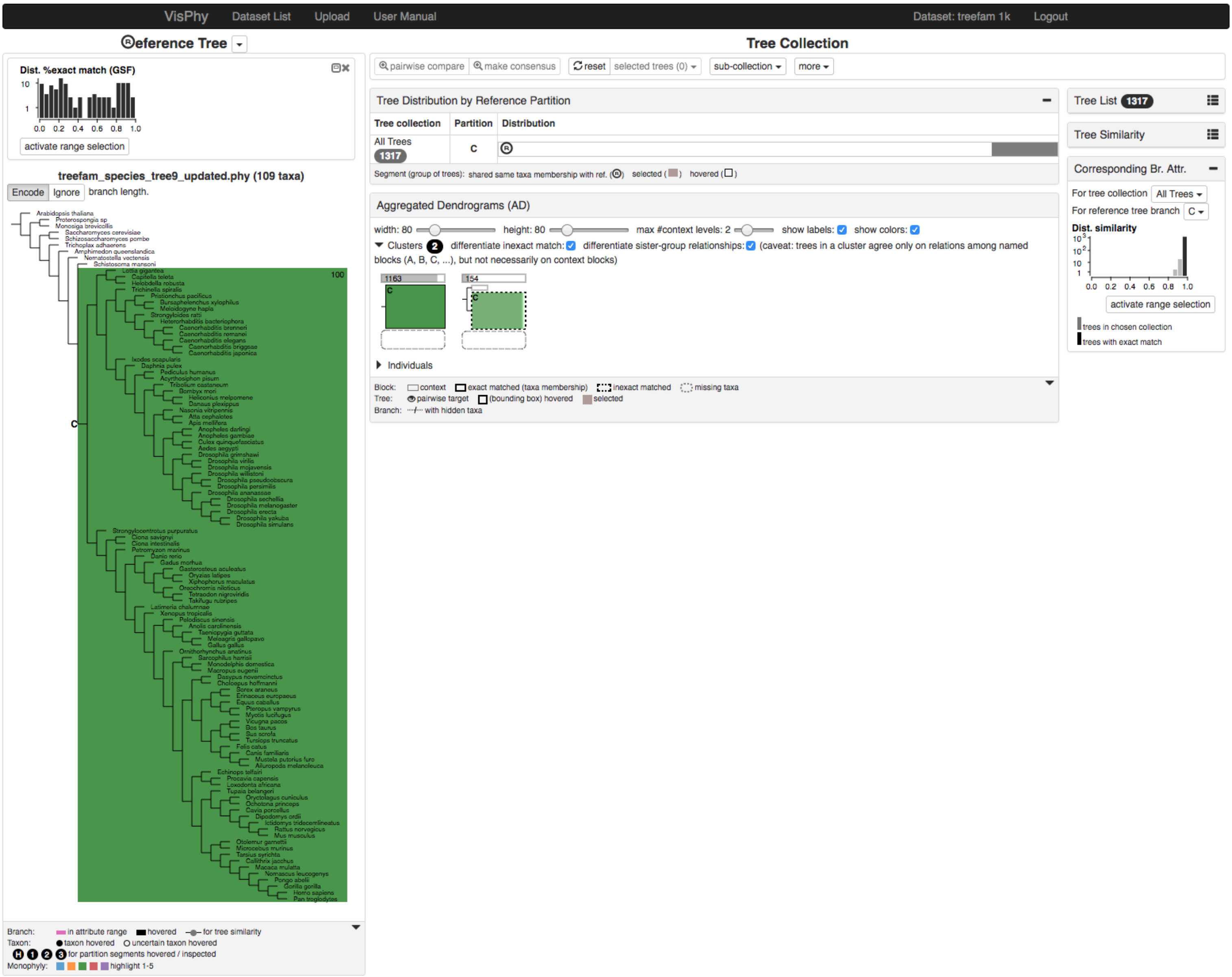
Figure S13 and Figure S14 illustrate how the user can use ADView to visually confirm that many of the gene trees are consistent with the classical taxonomic thought that PROTOSTOMIA and DEUTEROSTOMIA are the two major monophyletic branches of BILATERIAN animals. As illustrated in Figure S13, by selecting the two clades as A (PROTOSTOMIA) in blue and B (DEUTEROSTSOMIA) in orange, we found subsets of gene trees entirely or partially consistent (i.e., with missing taxa) with the species tree. The evolutionary relationship between A and B is reflected in the clustered ADs shaded with brown backgrounds. We also observed several other clustered ADs in which we did not find any support for RQ3. The first clustered AD (leftmost) contains 751 gene trees having only taxa from B; similarly, the fourth clustered AD contains 77 gene trees having only taxa from A. This situation immediately reveals genes unique to either A or B, due to a biological process (clade-specific gene gain or loss) or incomplete sampling (imperfect data collection). The other cluster ADs capture seemingly discordant gene trees; however, further exploration of the gene trees under the *Individual Aggregated Dendrogram* view and pairwise comparison of gene trees, as shown in Figure S15 and S16, revealed that the discordant signals likely resulted from inadequate post-processing of the gene trees. The TreeFam pipeline did not remove "rogue" taxa, which are outlier taxa spuriously inserted into the wrong clades. TreeFam was assembled a decade ago, therefore the results of the pipeline do not reflect the best practices in phylogenomics today. Removal of rogue taxa (e.g., using RogueNaRok [8]) should eliminate most of the discordant gene trees.

Additionally, we checked whether PROTOSTOMIA and DEUTEROSTOMIA are sister clades; that is, whether A and B form a monophyletic group. By selecting C (Figure S14), we found most of the gene trees (1,163) to be concordant with the monophyly of PROTOSTOMIA + DEUTERSTOMIA. We further explored the gene trees under the *Individual AD* view (not shown), and found that the discordant gene trees might be caused by a variety of reasons (e.g., rogue taxa, incomplete sampling, or stochastic and systematic errors in phylogenetic methods).



VisPhy
Dataset List
Upload
User Manual
Dataset: treefam 1k
Logout
Reference Tree
Tree Collection
Dist. %exact match (GSF)
activate range selection
treefam_species_tree9_updated.phy (109 taxa)
Encode
Ignore
branch length.
pairwise compare
make consensus
reset
selected trees (409)
sub-collection
more
Tree Distribution by Reference Partition
Tree collection
Partition
Distribution
All Trees
1317
A
B
Segment (group of trees): shared same taxa membership with ref. selected hovered
Aggregated Dendrograms (AD)
manual size: □
~~width: 100~~
~~height: 100~~
max #context levels: 2
show labels: ☑
show colors: ☑
Clusters 12
differentiate inexact match: ☑
differentiate sister-group relationships: ☑ (caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)
Less
Individuals (sort by the RF distance to the reference tree)
All Trees 1317
More
Block: context; exact matched (taxa membership); inexact matched; missing taxa
Tree: pairwise target; (bounding box) hovered; selected
Branch: with hidden taxa
Tree List 1317
Tree Similarity
Corresponding Br. Attr.
For tree collection All Trees
For reference tree branch B
Dist. similarity
activate range selection
trees in chosen collection
trees with exact match
Branch: in attribute range; hovered; for tree similarity
Taxon: taxon hovered; uncertain taxon hovered
for partition segments hovered / inspected
Monophyly: highlight 1-5

Figure S13: The major clades PROTOSTOMIA (A) and DEUTEROSTSOMIA (B) are selected. This view provides overall gene tree support for the monophyly of PROTOSTOMIA and that of DEUTEROSTSOMIA. The supporting clustered ADs are highlighted with brown backgrounds. The *Individaul AD* view is collapsed because we were not focusing on any individual trees at this point. Collapsing the *Individual AD* view also results in faster response because there were many fewer elements to render and keep track of for the browser.



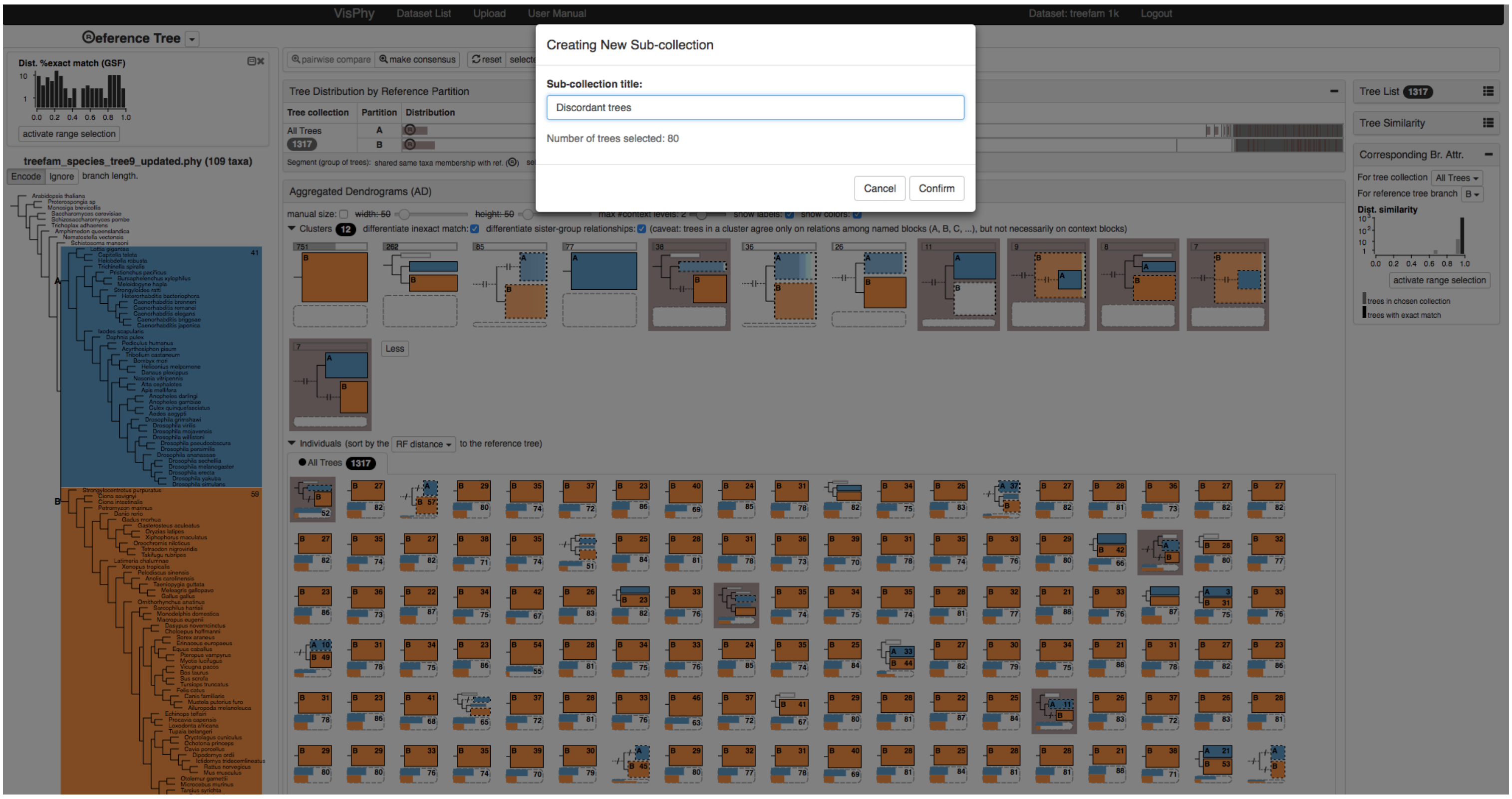
treefam_species_tree9_updated.phy (109 taxa)

Encode Ignore branch length.

Segment (group of trees): shared same taxa membership with ref. Ⓡ selected (▨) hovered (□)

☑ differentiate inexact match: ☑ differentiate sister-group relationships: ☑ (caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)

Figure S14: Both PROTOSTOMIA and DEUTEROSTSOMIA are selected as a single group (C). This combination reveals most of the gene trees are consistent with the monophyly of BILATERIAN animals (C).

 Creating New Sub-collection

Sub-collection title: Discordant trees

Number of trees selected: 80

Figure S15: We selected the discordant trees excluding trees that are missing PROTOSTOMIA or missing DEUTEROSTSOMIA, and created a sub-collection, that is, a subset of the tree collection. We wanted to find out why these trees do not behave as expected.

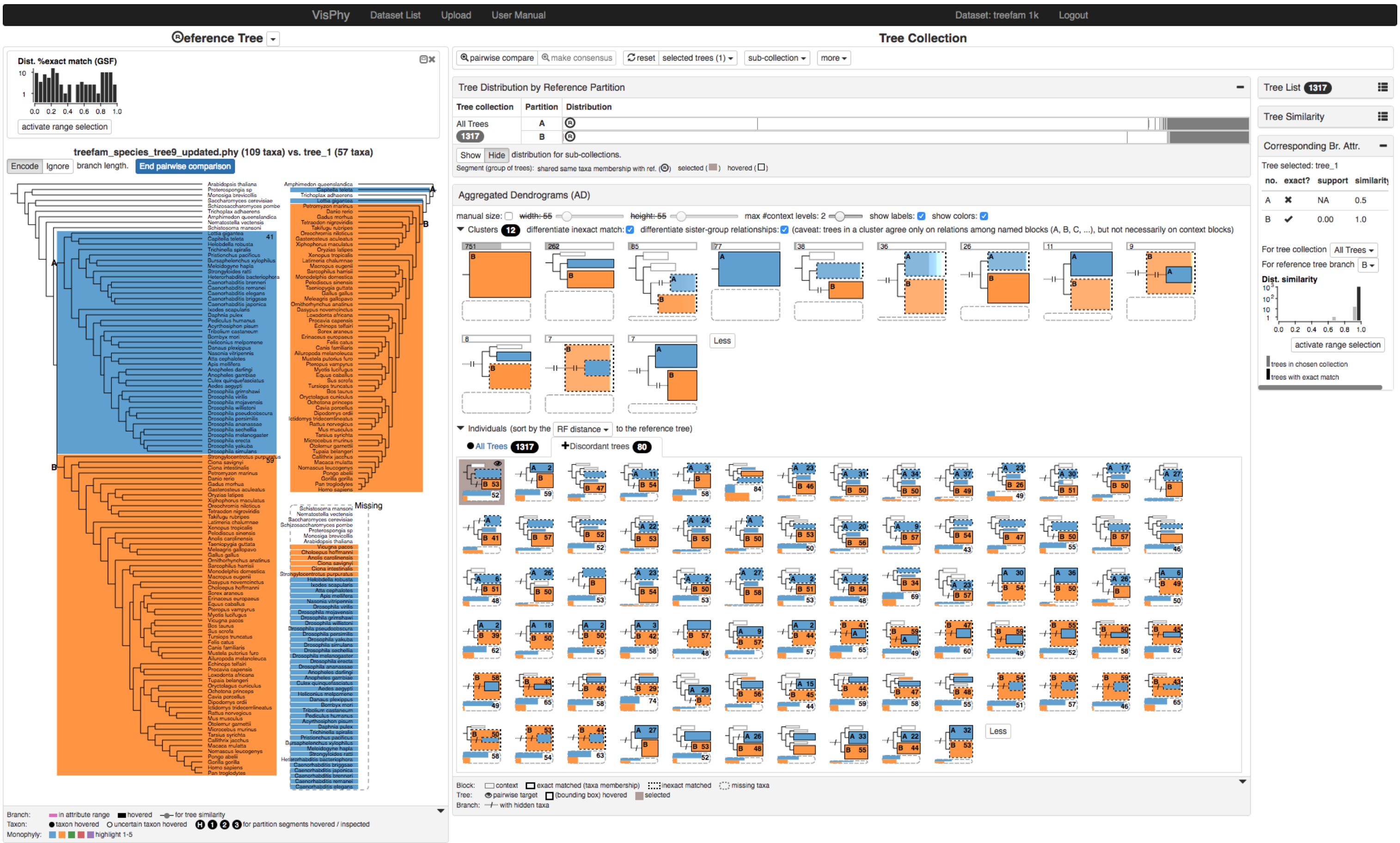
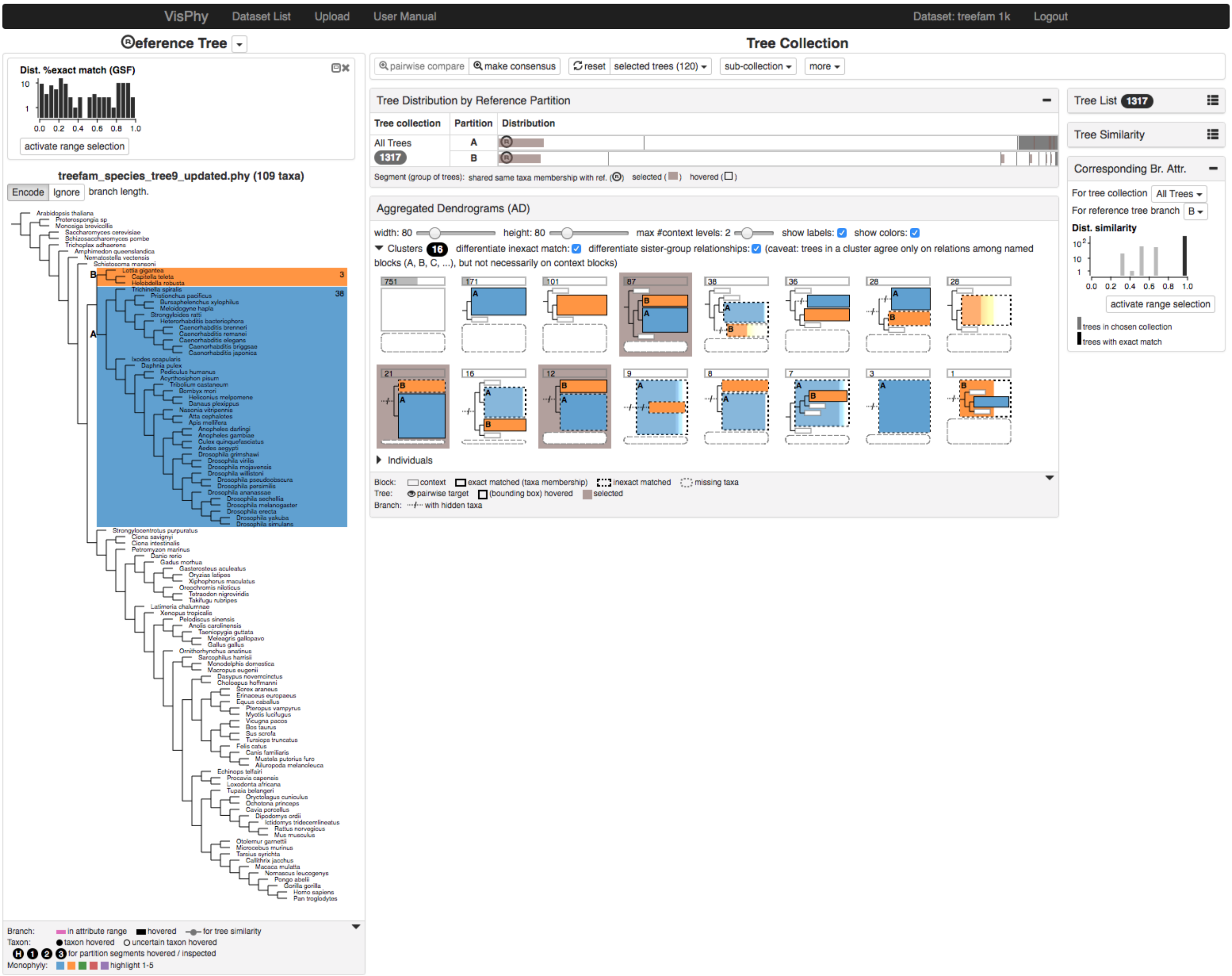


Figure S16: Pairwise comparison is helpful in locating the potential causes of the discordance. The orange outlier at the top shows that B has a rogue gene in the gene tree (right dendrogram) compared to the species tree (left dendrogram).

S6.2 Ecdysozoa and Lophotrozoa

Next, we performed the same tasks, but we revisited the well-established ideas of Ecdysozoa and Lophotrozoa being the two major monophyletic branches of Protostomia. We selected A (Ecdysozoa) and B (Lophotrozoa), but found only small clustered ADs (highlighted in brown background) consistent with their monophyly, as shown in Figure S17. We then selected only Ecdysozoa or Lophotrozoa separately. When we selected only Ecdysozoa (A), we found strong support for its monophyly, corroborated by most of the gene trees, as shown in Figure S18. When we selected only Lophotrozoa (not shown), however, we observed no strong support in TreeFam for its monophyly, probably because there are too few representative taxa included to yield robust phylogenetic signals. Indeed, later studies involving more Lophotrozoa taxa produced strong support for its monophyly [6, 7].



Dataset: treefam 1k

treefam_species_tree9_updated.phy (109 taxa)

Encode | Ignore branch length.

width: 80 height: 80 max #context levels: 2 show labels: ☑ show colors: ☑

Clusters 16 differentiate inexact match: ☑ differentiate sister-group relationships: ☑ (caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)

Segment (group of trees): shared same taxa membership with ref. (R) selected (▨) hovered (□)

Tree collection	Partition	Distribution
All Trees 1317	A	
	B	

For tree collection: All Trees

For reference tree branch: B

Block: □ context □ exact matched (taxa membership) ⬚ inexact matched ⬚ missing taxa

Tree: ● pairwise target □ (bounding box) hovered ▨ selected

Branch: —/— with hidden taxa

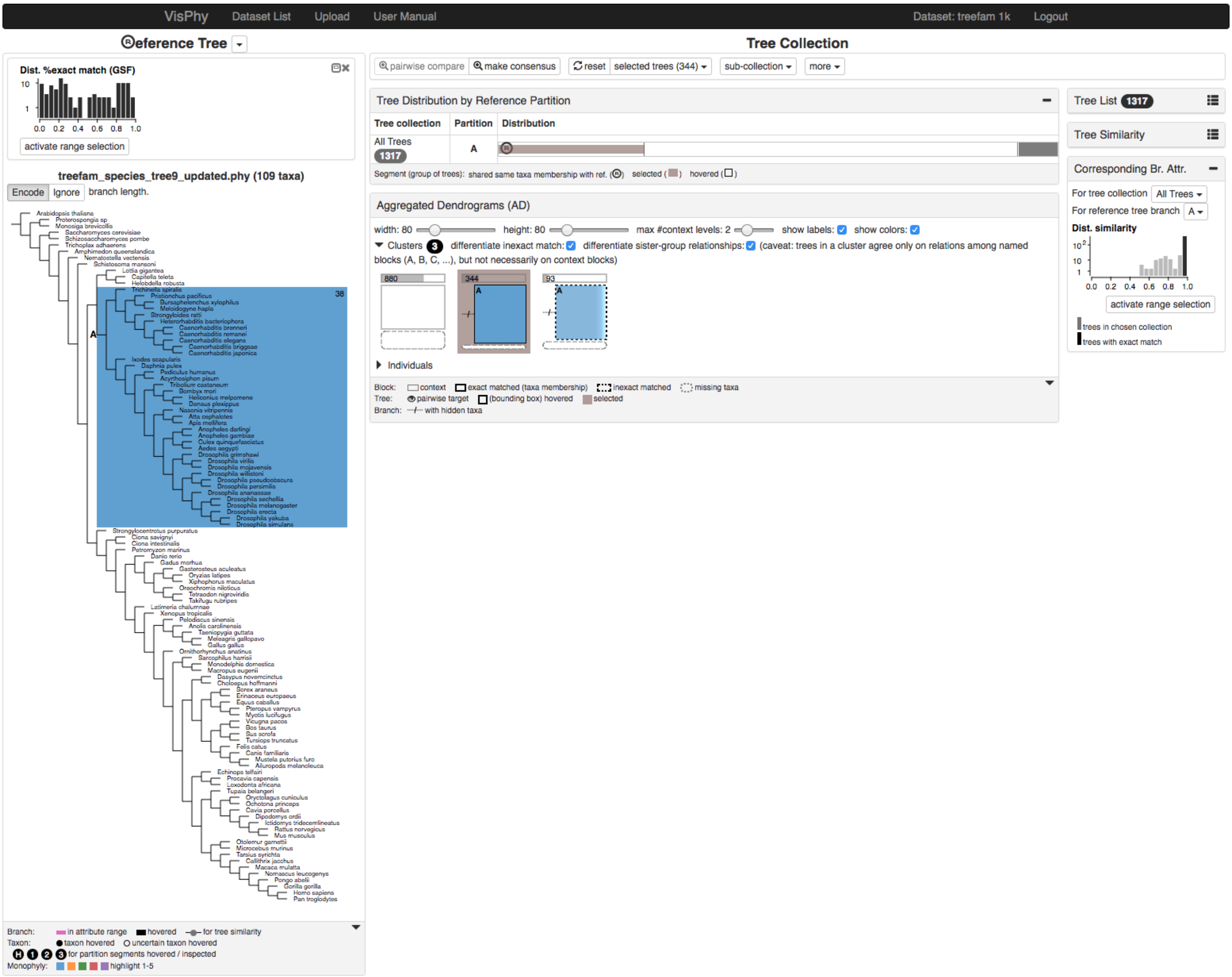
Branch: ▬ in attribute range ▬ hovered —●— for tree similarity

Taxon: ● taxon hovered ○ uncertain taxon hovered

H 1 2 3 for partition segments hovered / inspected

Monophyly: 🟦 🟧 🟩 🟥 🟪 highlight 1-5

Figure S17: The major clades Ecdysozoa (A) and Lophotrozoa (B) are selected. The clustered ADs in support of their monophylies are highlighted with brown backgrounds.



treefam_species_tree9_updated.phy (109 taxa)

width: 80 height: 80 max #context levels: 2 show labels: ☑ show colors: ☑

differentiate inexact match: ☑ differentiate sister-group relationships: ☑ (caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)

Figure S18: Only Ecdysozoa (A) is selected. The clustered AD, which contains most of the gene trees, corroborating its monophyly is highlighted with a brown background.

S7 Screenshots of Information Density Comparison

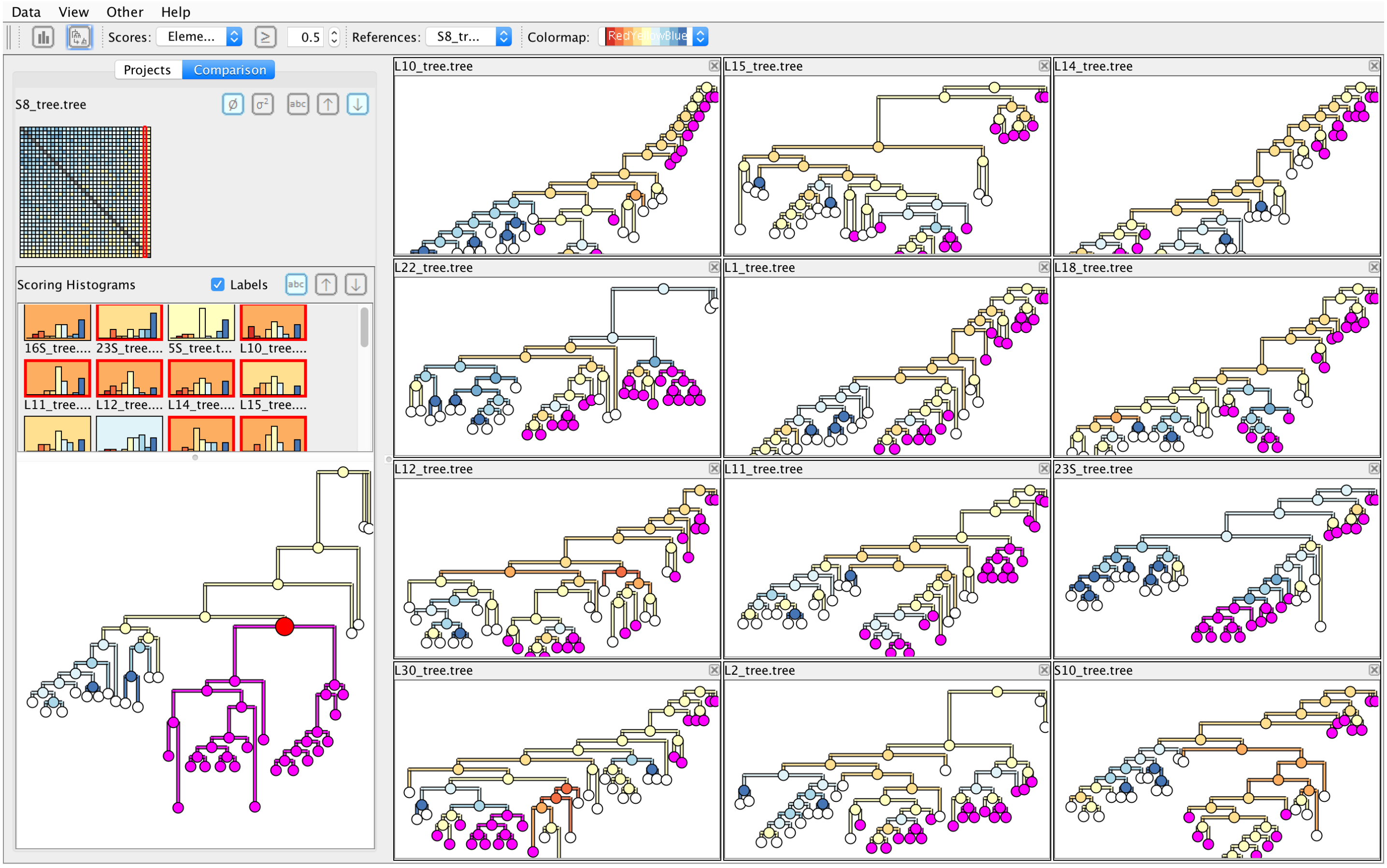


Figure S19: Screenshot of the tool developed by Bremm et. al [9], on the 35-tree dataset from their paper.

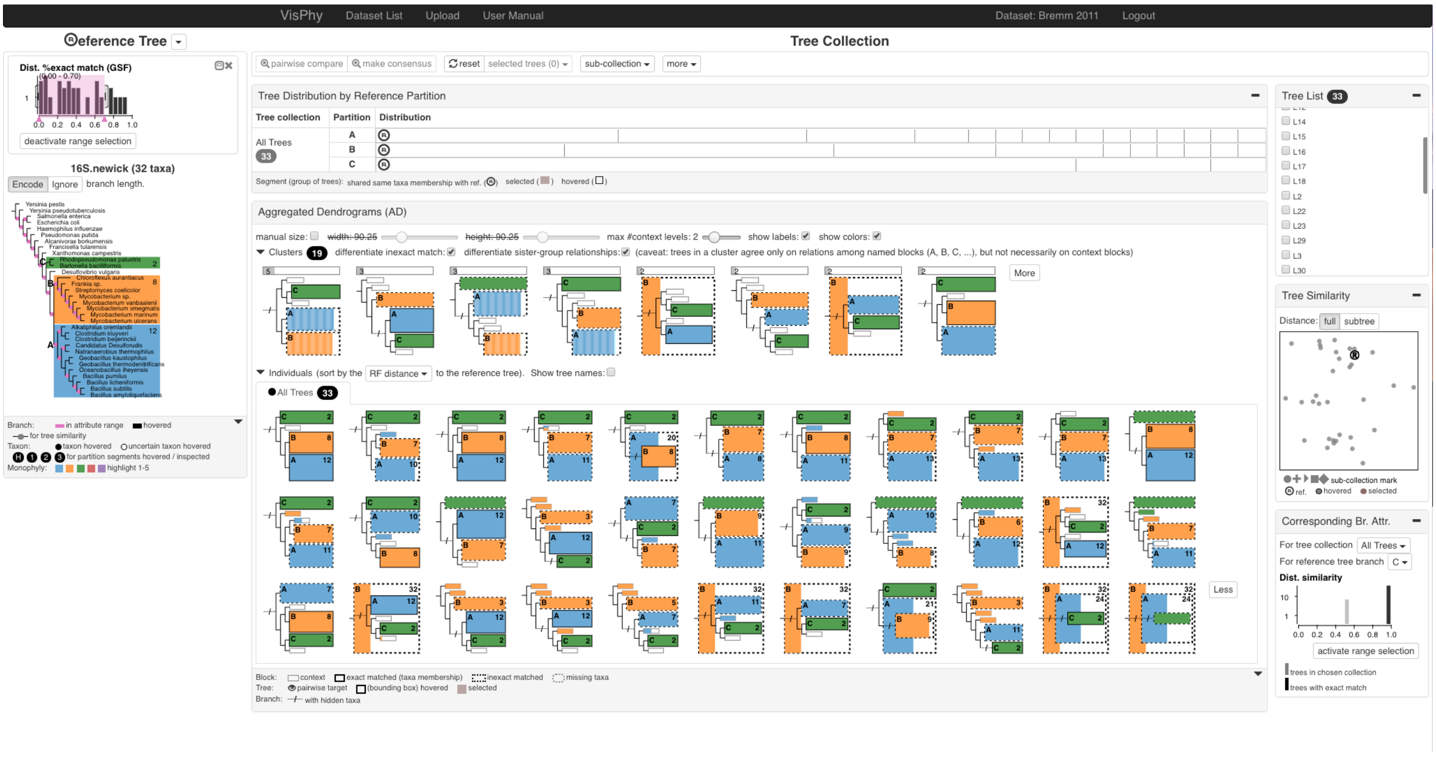


Figure S20: Screenshot of ADView with the same dataset at the same screen resolution, with three subtrees selected. Information about exactly what taxa their domain experts explored is not available from the paper, so we randomly picked three subtrees.

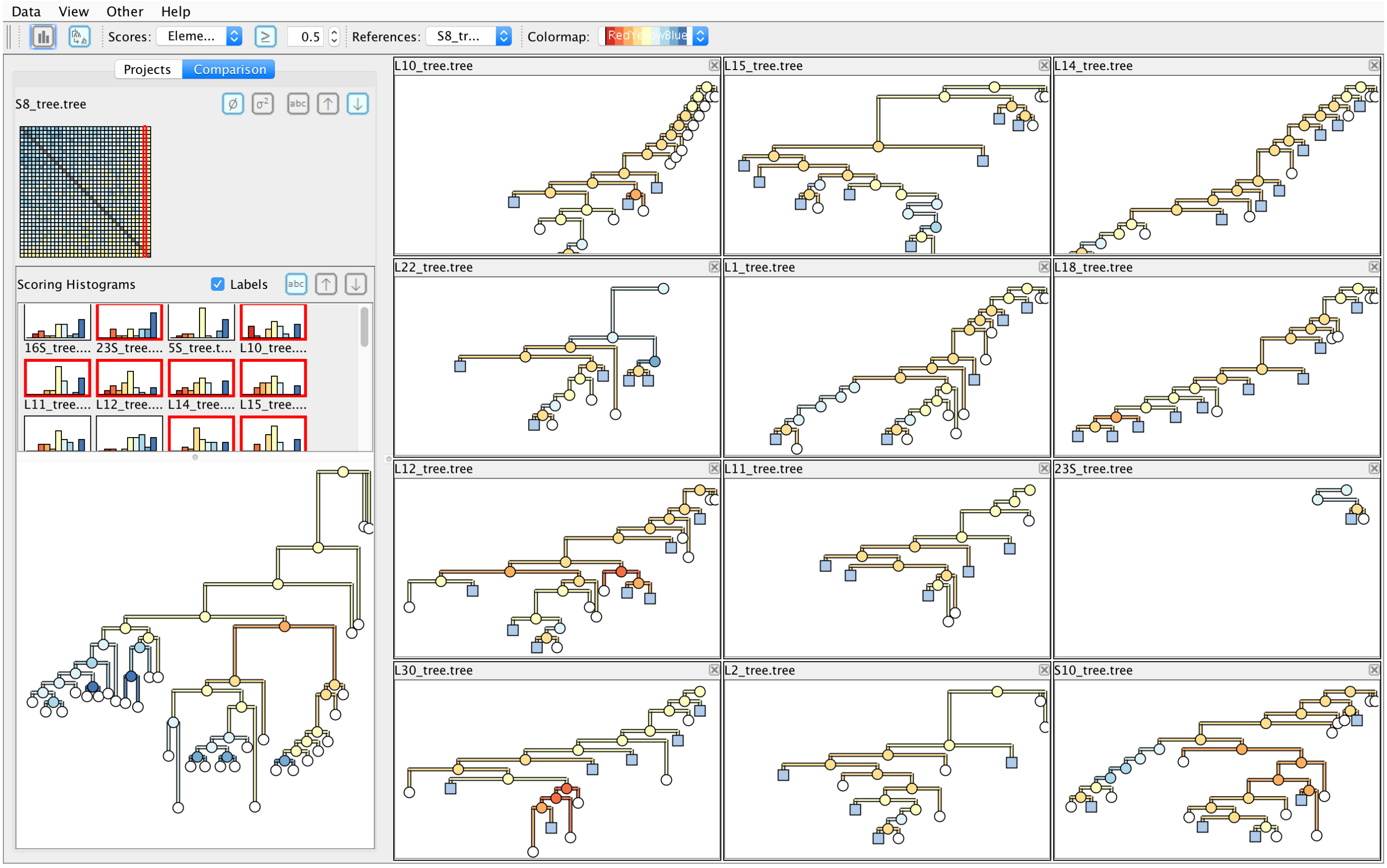
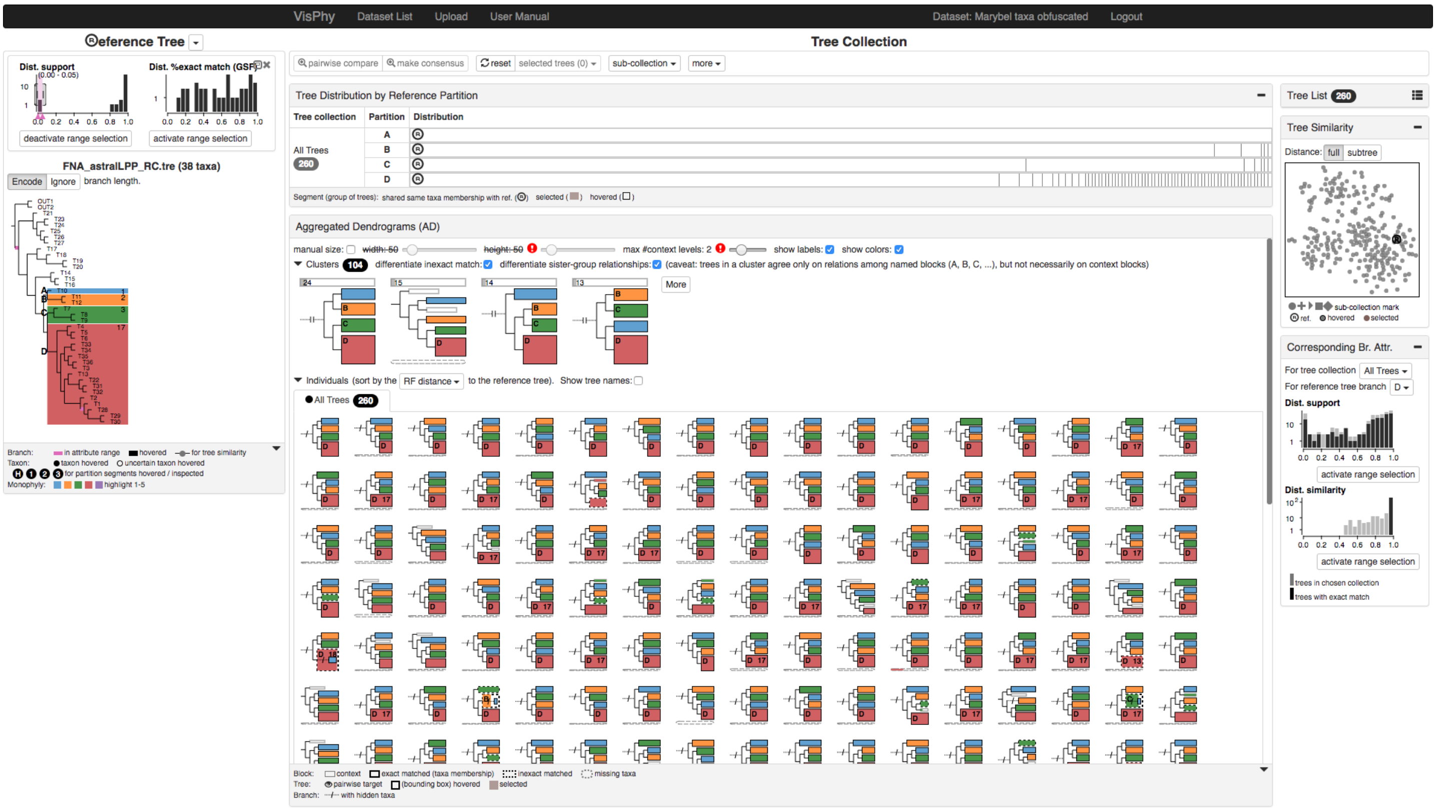


Figure S21: Screenshot of the tool developed by Bremm et. al [9]. Elements with similarity score below 0.5 are filtered out.

S8 Case Studies

S8.1 Full Screenshots of Case Study 1



VisPhy
Dataset List
Upload
User Manual
Dataset: Marybel taxa obfuscated
Logout
Reference Tree
Tree Collection
Dist. support
Dist. %exact match (GSF)
deactivate range selection
activate range selection
FNA_astralLPP_RC.tre (38 taxa)
Encode
Ignore
branch length.
pairwise compare
make consensus
reset
selected trees (0)
sub-collection
more
Tree Distribution by Reference Partition
Tree collection
Partition
Distribution
All Trees 260
A
B
C
D
Segment (group of trees): shared same taxa membership with ref. selected hovered
Aggregated Dendrograms (AD)
manual size:
width: 50
height: 50
max #context levels: 2
show labels:
show colors:
Clusters 104
differentiate inexact match:
differentiate sister-group relationships:
(caveat: trees in a cluster agree only on relations among named blocks (A, B, C, ...), but not necessarily on context blocks)
24
15
14
13
More
Individuals (sort by the RF distance to the reference tree). Show tree names:
All Trees 260
Block: context exact matched (taxa membership) inexact matched missing taxa
Tree: pairwise target (bounding box) hovered selected
Branch: with hidden taxa
Branch: in attribute range hovered for tree similarity
Taxon: taxon hovered uncertain taxon hovered
for partition segments hovered / inspected
Monophyly: highlight 1-5
Tree List 260
Tree Similarity
Distance: full subtree
sub-collection mark
ref. hovered selected
Corresponding Br. Attr.
For tree collection All Trees
For reference tree branch D
Dist. support
Dist. similarity
activate range selection
trees in chosen collection
trees with exact match

Figure S22: P1 was exploring the position of the T10 taxon with regard to a bigger monophyletic group: the direct parent of the four focal subtrees, namely T10 itself (blue) and three sibling groups (orange, green, red). The position of the blue T10 group varies; there are many cluster ADs with a relatively small number of trees in each. She also observed the same by skimming the individual ADs.

S8.2 Full Screenshots of Case Study 2

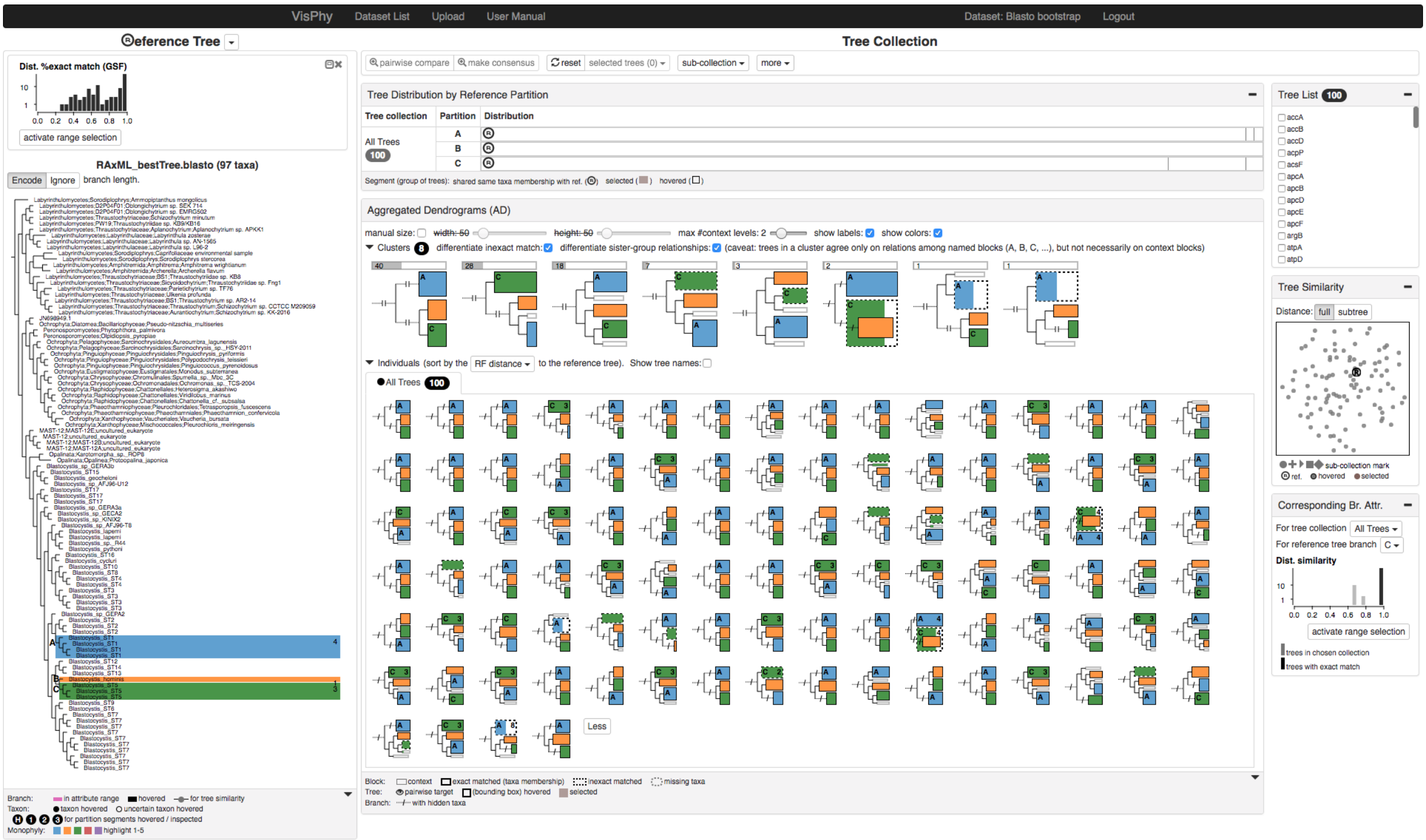


Figure S23: P2 identified an interesting taxon Blastocystis hominis, shown as the orange group, that usually lives in pigs but was placed close to human related strains.

S8.3 Case Study 3

P3 is a frequent collaborator of our domain expert co-author, and they were working together on a dataset comparing 1 species tree against 71 gene trees of 115 red algaes. During the study, they were able to locate some misbehaved missing taxa in the reference tree and outliers in the gene tree, but could not continue the analysis further due to a missing engineering feature of the interface (selecting a tree by its name). After we added that feature, they used ADView independently for several weeks and made some interesting biological discoveries.

They first compared four species trees created with different methods and kinds of sequences, mainly using the *Pairwise Comparison* view, to identify the one that is most consistent with the current literature (T3), as shown in Figure S24. Next, using that best tree among the four as a reference, they compared it against 155 gene trees built from amino acid sequences. They sought artifacts, such as unusually long branches, and anomalies, primarily "rogue taxa" where outlier leaves are spuriously inserted into the wrong subtree. They quickly scanned over the gene trees to get some insights about where discrepancies typically arise (T4 and T5). Unsurprisingly, discrepancies seem to be concentrated at deep nodes (e.g., Cyanidiales), because in many genes the information needed to infer deep relationships has been eroded or lost. They also observed that a three-taxa subclade of Polysiphonia, has undergone exceptionally high molecular evolution in a gene "rpl21", as shown in Figure S25. The conflict between rpl21 gene tree and the reference tree on the Polysiphonia subclade is an alert to be cautious when interpreting signals from Polysiphonia, well-known for uncertain phylogenetic placement. Their future work is to aggregate the results of the molecular evolution analysis across the 155 genes, and they plan to check the gene trees in ADView for details that may affect their interpretation of the results.

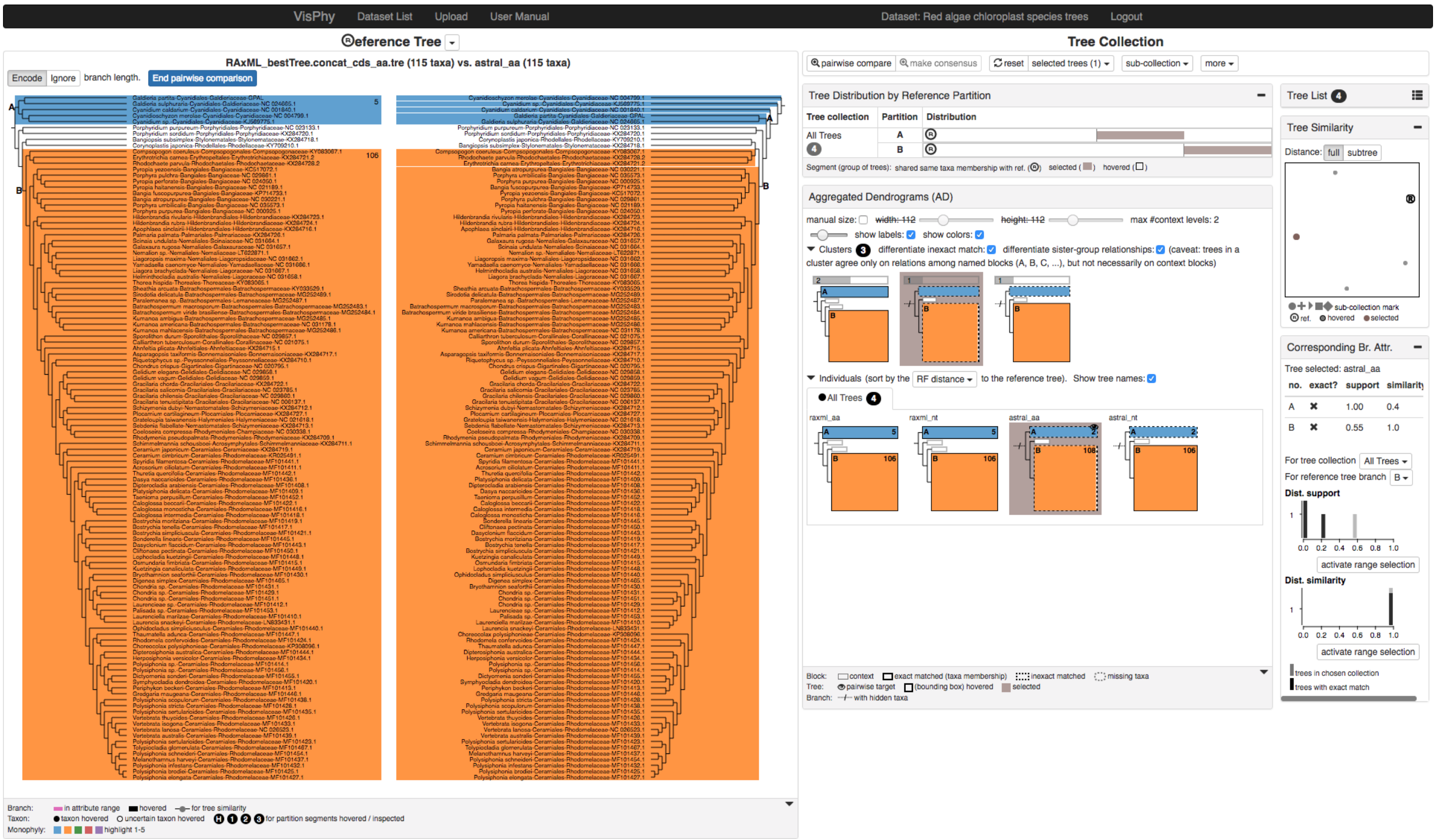


Figure S24: P3 and our biologist co-author first compared four species trees inferred from different methods and different kinds of molecular sequences in order to choose the best one for later analysis.

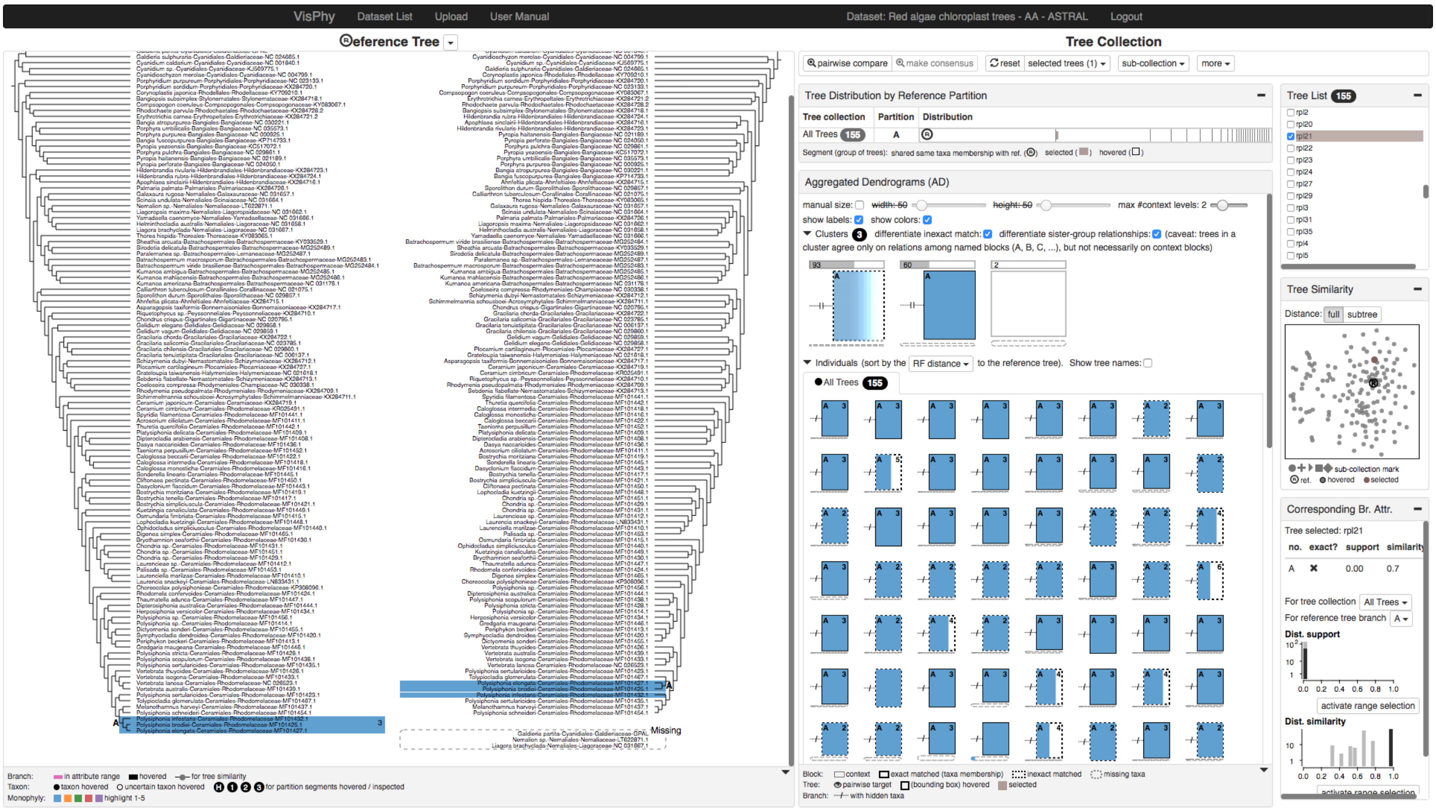
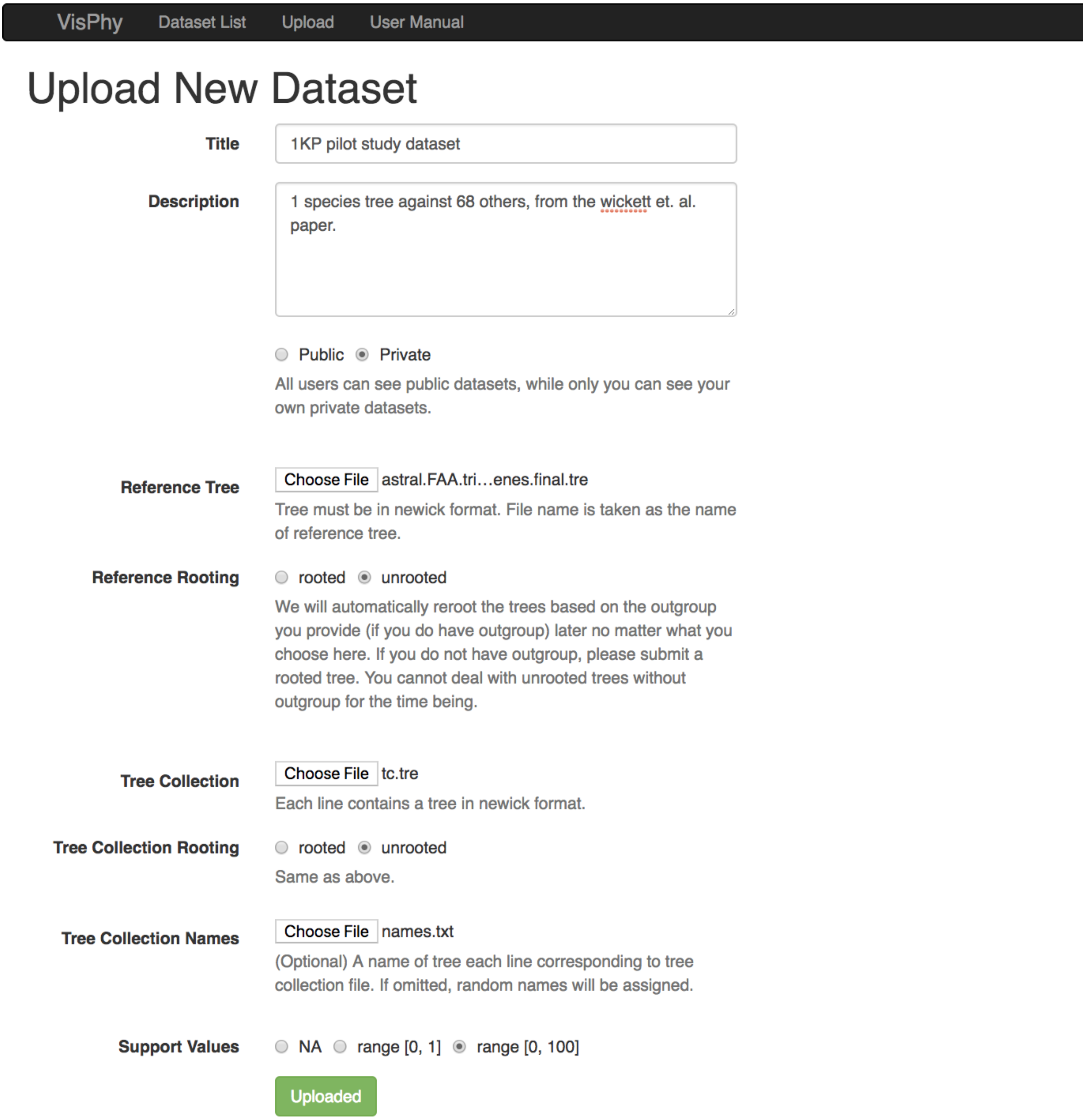

VisPhy
Dataset List
Upload
User Manual
Dataset: Red algae chloroplast trees - AA - ASTRAL
Logout
Reference Tree
Tree Collection
pairwise compare
make consensus
reset
selected trees (1)
sub-collection
more
Tree Distribution by Reference Partition
Tree collection
Partition
Distribution
All Trees 155
A
Aggregated Dendrograms (AD)
manual size:
width: 50
height: 50
max #context levels: 2
show labels:
show colors:
Clusters 3
differentiate inexact match:
differentiate sister-group relationships:
Individuals (sort by the RF distance to the reference tree). Show tree names:
All Trees 155
Tree List 155
rpl2
rpl20
rpl21
rpl22
rpl23
rpl24
rpl27
rpl29
rpl3
rpl31
rpl35
rpl4
rpl5
Tree Similarity
Distance: full subtree
sub-collection mark
ref.
hovered
selected
Corresponding Br. Attr.
Tree selected: rpl21
no. exact? support similarity
A ✗ 0.00 0.7
For tree collection All Trees
For reference tree branch A
Dist. support
Dist. similarity
activate range selection
Missing
Branch: in attribute range hovered for tree similarity
Taxon: taxon hovered uncertain taxon hovered for partition segments hovered / inspected
Monophyly: highlight 1-5
Block: context exact matched (taxa membership) inexact matched missing taxa
Tree: pairwise target (bounding box) hovered selected
Branch: with hidden taxa

Figure S25: P3 and our biologist co-author used the best tree from the previous step as the reference tree, and compared it against 155 gene trees. They found the blue group is inconsistent across gene trees. An example (gene “rpl21”) is shown in the figure.

S9 Screenshots of Dataset Upload

Users can upload their own datasets through a dedicated tab in the client browser interface. There are two steps to upload a dataset: upload the tree files, as shown in Figure S26, and specify an outgroup if the trees are unrooted, as shown in Figure S27. The server then performs indexing, tree distance calculation, and corresponding branch matching, as shown in Figure S28.

Finally, the newly uploaded dataset is added to the list of all uploaded datasets, available through another tab.



Upload New Dataset

Title: 1KP pilot study dataset

Description: 1 species tree against 68 others, from the wickett et. al. paper.

○ Public ● Private

All users can see public datasets, while only you can see your own private datasets.

Reference Tree: Choose File astral.FAA.tri…enes.final.tre

Tree must be in newick format. File name is taken as the name of reference tree.

Reference Rooting: ○ rooted ● unrooted

We will automatically reroot the trees based on the outgroup you provide (if you do have outgroup) later no matter what you choose here. If you do not have outgroup, please submit a rooted tree. You cannot deal with unrooted trees without outgroup for the time being.

Tree Collection: Choose File tc.tre

Each line contains a tree in newick format.

Tree Collection Rooting: ○ rooted ● unrooted

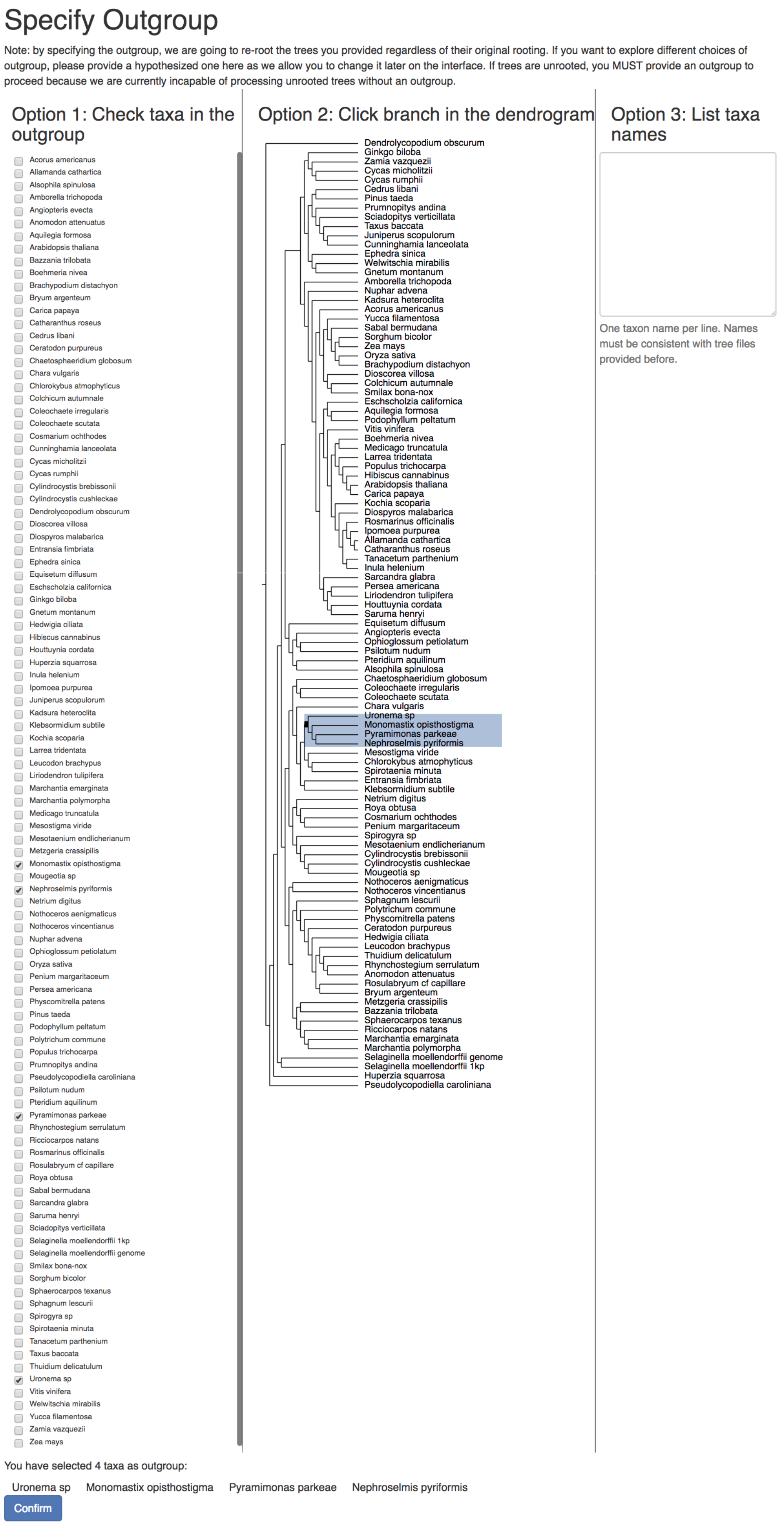
Same as above.

Tree Collection Names: Choose File names.txt

(Optional) A name of tree each line corresponding to tree collection file. If omitted, random names will be assigned.

Support Values: ○ NA ○ range [0, 1] ● range [0, 100]

Figure S26: The first step of uploading a dataset: specify the files for a reference tree and a collection of trees.



Specify Outgroup

Note: by specifying the outgroup, we are going to re-root the trees you provided regardless of their original rooting. If you want to explore different choices of outgroup, please provide a hypothesized one here as we allow you to change it later on the interface. If trees are unrooted, you MUST provide an outgroup to proceed because we are currently incapable of processing unrooted trees without an outgroup.

Option 1: Check taxa in the outgroup

- □ Acorus americanus
- □ Allamanda cathartica
- □ Alsophila spinulosa
- □ Amborella trichopoda
- □ Angiopteris evecta
- □ Anomodon attenuatus
- □ Aquilegia formosa
- □ Arabidopsis thaliana
- □ Bazzania trilobata
- □ Boehmeria nivea
- □ Brachypodium distachyon
- □ Bryum argenteum
- □ Carica papaya
- □ Catharanthus roseus
- □ Cedrus libani
- □ Ceratodon purpureus
- □ Chaetosphaeridium globosum
- □ Chara vulgaris
- □ Chlorokybus atmophyticus
- □ Colchicum autumnale
- □ Coleochaete irregularis
- □ Coleochaete scutata
- □ Cosmarium ochthodes
- □ Cunninghamia lanceolata
- □ Cycas micholitzii
- □ Cycas rumphii
- □ Cylindrocystis brebissonii
- □ Cylindrocystis cushleckae
- □ Dendrolycopodium obscurum
- □ Dioscorea villosa
- □ Diospyros malabarica
- □ Entransia fimbriata
- □ Ephedra sinica
- □ Equisetum diffusum
- □ Eschscholzia californica
- □ Ginkgo biloba
- □ Gnetum montanum
- □ Hedwigia ciliata
- □ Hibiscus cannabinus
- □ Houttuynia cordata
- □ Huperzia squarrosa
- □ Inula helenium
- □ Ipomoea purpurea
- □ Juniperus scopulorum
- □ Kadsura heteroclita
- □ Klebsormidium subtile
- □ Kochia scoparia
- □ Larrea tridentata
- □ Leucodon brachypus
- □ Liriodendron tulipifera
- □ Marchantia emarginata
- □ Marchantia polymorpha
- □ Medicago truncatula
- □ Mesostigma viride
- □ Mesotaenium endlicherianum
- □ Metzgeria crassipilis
- ☑ Monomastix opisthostigma
- □ Mougeotia sp
- ☑ Nephroselmis pyriformis
- □ Netrium digitus
- □ Nothoceros aenigmaticus
- □ Nothoceros vincentianus
- □ Nuphar advena
- □ Ophioglossum petiolatum
- □ Oryza sativa
- □ Penium margaritaceum
- □ Persea americana
- □ Physcomitrella patens
- □ Pinus taeda
- □ Podophyllum peltatum
- □ Polytrichum commune
- □ Populus trichocarpa
- □ Prumnopitys andina
- □ Pseudolycopodiella caroliniana
- □ Psilotum nudum
- □ Pteridium aquilinum
- ☑ Pyramimonas parkeae
- □ Rhynchostegium serrulatum
- □ Ricciocarpos natans
- □ Rosmarinus officinalis
- □ Rosulabryum cf capillare
- □ Roya obtusa
- □ Sabal bermudana
- □ Sarcandra glabra
- □ Saruma henryi
- □ Sciadopitys verticillata
- □ Selaginella moellendorffii 1kp
- □ Selaginella moellendorffii genome
- □ Smilax bona-nox
- □ Sorghum bicolor
- □ Sphaerocarpos texanus
- □ Sphagnum lescurii
- □ Spirogyra sp
- □ Spirotaenia minuta
- □ Tanacetum parthenium
- □ Taxus baccata
- □ Thuidium delicatulum
- ☑ Uronema sp
- □ Vitis vinifera
- □ Welwitschia mirabilis
- □ Yucca filamentosa
- □ Zamia vazquezii
- □ Zea mays

Option 2: Click branch in the dendrogram

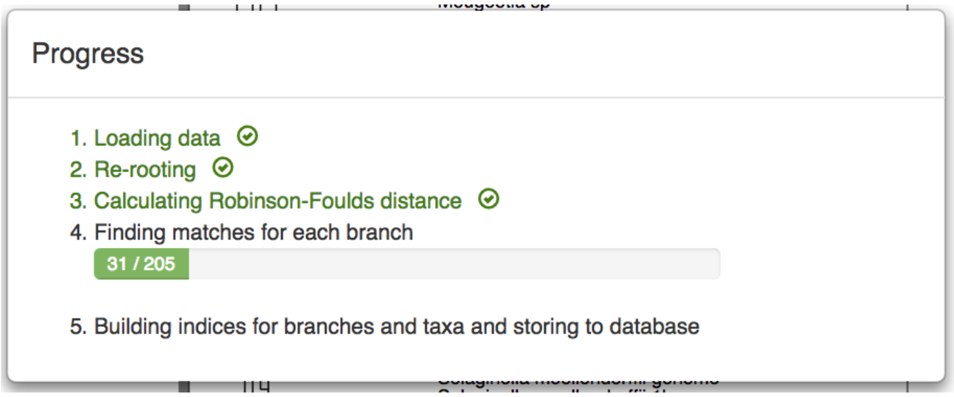
Option 3: List taxa names

One taxon name per line. Names must be consistent with tree files provided before.

You have selected 4 taxa as outgroup:

Uronema sp Monomastix opisthostigma Pyramimonas parkeae Nephroselmis pyriformis

Figure S27: The second step of uploading a dataset: select an outgroup by either checking the names, clicking on the branches (as shown here), or paste a text file.



Progress

1. Loading data ✓
2. Re-rooting ✓
3. Calculating Robinson-Foulds distance ✓
4. Finding matches for each branch

31 / 205

5. Building indices for branches and taxa and storing to database

Figure S28: Server is pre-processing data.

References

[1] Christopher H Jackson. Displaying uncertainty with shading. *The American Statistician*, 62(4):340–347, 2008.

[2] Norman J. Wickett et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. National Academy of Sciences*, 111(45):E4859–E4868, 2014.

[3] Jue Ruan, Heng Li, Zhongzhong Chen, Avril Coghlan, Lachlan James M Coin, Yiran Guo, Jean-Karim Heriche, Yafeng Hu, Karsten Kristiansen, Ruiqiang Li, et al. Treefam: 2008 update. *Nucleic acids research*, 36(suppl_1):D735–D740, 2007.

[4] Maximilian J Telford. The animal tree of life. *Science*, 339(6121):764–766, 2013.

[5] Anna Marie A Aguinaldo, James M Turbeville, Lawrence S Linford, Maria C Rivera, James R Garey, Rudolf A Raff, and James A Lake. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632):489–493, 1997.

[6] Hervé Philippe, Nicolas Lartillot, and Henner Brinkmann. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular biology and evolution*, 22(5):1246–1253, 2005.

[7] Martin Helmkampf, Iris Bruchhaus, and Bernhard Hausdorf. Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the lophotrochozoa concept. *Proc. Royal Society of London B: Biological Sciences*, 275(1645):1927–1933, 2008.

[8] Andre J Aberer, Denis Krompass, and Alexandros Stamatakis. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic biology*, 62(1):162–166, 2012.

[9] Sebastian Bremm, Tatiana Von Landesberger, Martin Heß, Tobias Schreck, Philipp Weil, and Kay Hamacher. Interactive visual comparison of multiple trees. In *IEEE Conf. Visual Analytics Science and Technology (VAST)*, pages 31–40, 2011.