Supplement Materials: Nonparametric Estimation of Phylogenetic Tree Distributions

Grady S. Weyenberg[1], Peter M. Huggins[2], Christopher L. Schardl[3], Daniel K. Howe[4], and Ruriko Yoshida[1]

[1]*Department of Statistics, University of Kentucky, Lexington, KY, USA;*
[2]*Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA;*
[3]*Plant Pathology Department, University of Kentucky, Lexington, KY, USA;*
[4]*Department of Veterinary Science, University of Kentucky, Lexington, KY, USA.*

Corresponding Author: Ruriko Yoshida, Department of Statistics, University of Kentucky, Lexington, KY, USA; E-mail: ruriko.yoshida@uky.edu.

Table S1. Analysis of Apicomplexa gene-sets identified as outliers

Gene ID	Functional Annotation	Analysis
PF08_0086	RNA-binding protein, putative	Significant sequence length disparity (164 a.a. for Ta vs 1075a.a. for Pf). Generally good sequence alignment in one region of 100 residues; otherwise, alignment is poor.
PF13_0228	40S ribosomal subunit protein S6, putative	Tt sequence much longer than all others; long N-terminal and C-terminal extensions. Very good alignment in blocks, but with lengthy insertions for outgroup Tt. Possible incorrect annotation of Tg sequence.
PFA0390w	DNA repair exonuclease, putative	Short sequences for Et and Cp. Several homopolymer stretches in Et. Modest to good alignment in multiple blocks, Et being an exception in several regions. Possible incorrect annotation of Et sequence.
PFF0285c	DNA repair protein RAD50, putative	Poor alignment in general. Three modest blocks (50-100 aa) of reasonable sequence alignment. Et sequence contains long homopolymeric stretches. Pf and Pv have long insertions that might be translated introns.
PFL1345c	Radical SAM protein, putative	Relatively short sequence for Et. Homopolymeric stretch at N-terminus of Tg. Modest to good alignment in blocks.
PFE0750c	hypothetical protein, conserved	Large difference in sequence lengths; 269 residues for Et vs. 848 for Pf. Central region with modest to good alignment; Et exhibited poor sequence identity suggestion it might not be a homologue.
PF10_0043	ribosomal protein L13, putative	80 residue N-terminal extension in Tg. Good sequence alignment, with Tt (outgroup) being an exception. Tt sequence might not be a homologue.
PF11_0463	coat protein, gamma subunit, putative	Multiple homopolymer stretches in Et sequence. Generally good alignment for all but Et; sequence might not be homologous.
MAL13P1.22	DNA ligase 1	Homopolymer stretches in Et sequence with poor alignment to other sequences. Et sequence might be incorrectly annotated and/or might not be homologous.
PFB0550w	Peptide chain release factor subunit 1, putative	Short sequence for Et (132 residues), with long homopolymer stretch. Other sequences are approximately 425 a.a. in length. Generally good alignment, even for Et over a short region (50 residues). Possible incorrect annotation of Et sequence.
PFF0120w	putative geranylgeranyltransferase	Two homopolymer stretches (serine) in Et sequence. Moderately good alignment. Possible incorrect annotation of Et sequence.
PFD0420c	flap exonuclease, putative	Very discrepant sequence lengths; 179 a.a. for Et vs. 2213 a.a. for Tt. All other sequences $500 - 600$ residues in length. Good alignment over several regions, although sequence for Et is absent in portions of these regions. Very long N-terminal extensions and insertions in Tt sequence. Possible incorrect annotations for Et and Tt.

Pf = *Plasmodium falciparum*, Pv = *Plasmodium vivax*, Bb = *Babesia bovis*, Ta = *Theileria annulata*, Et = *Eimeria tenella*, Tg = *Toxoplasma gondii*, Cp = *Cryptosporidium parvum*, and Tt = *Tetrahymena thermophila* (outgroup).

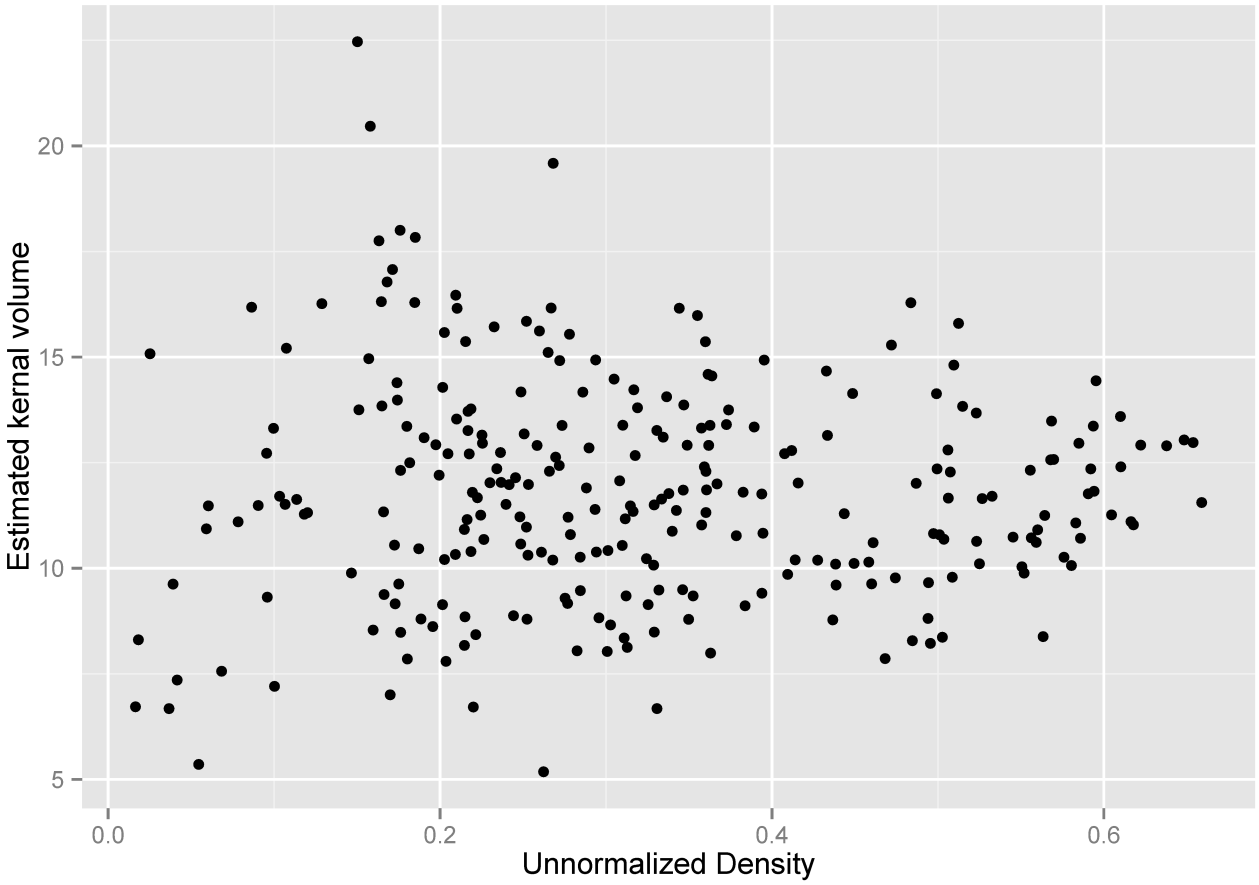

Figure S1. Monte Carlo estimates of $\sum_{T \in \mathcal{T}} k(T, T')$ are plotted against the unnormalized tree score for each tree T' in the Apicomplexa data. There is no significant evidence that the sum is related to the tree score ($p = 0.72$).

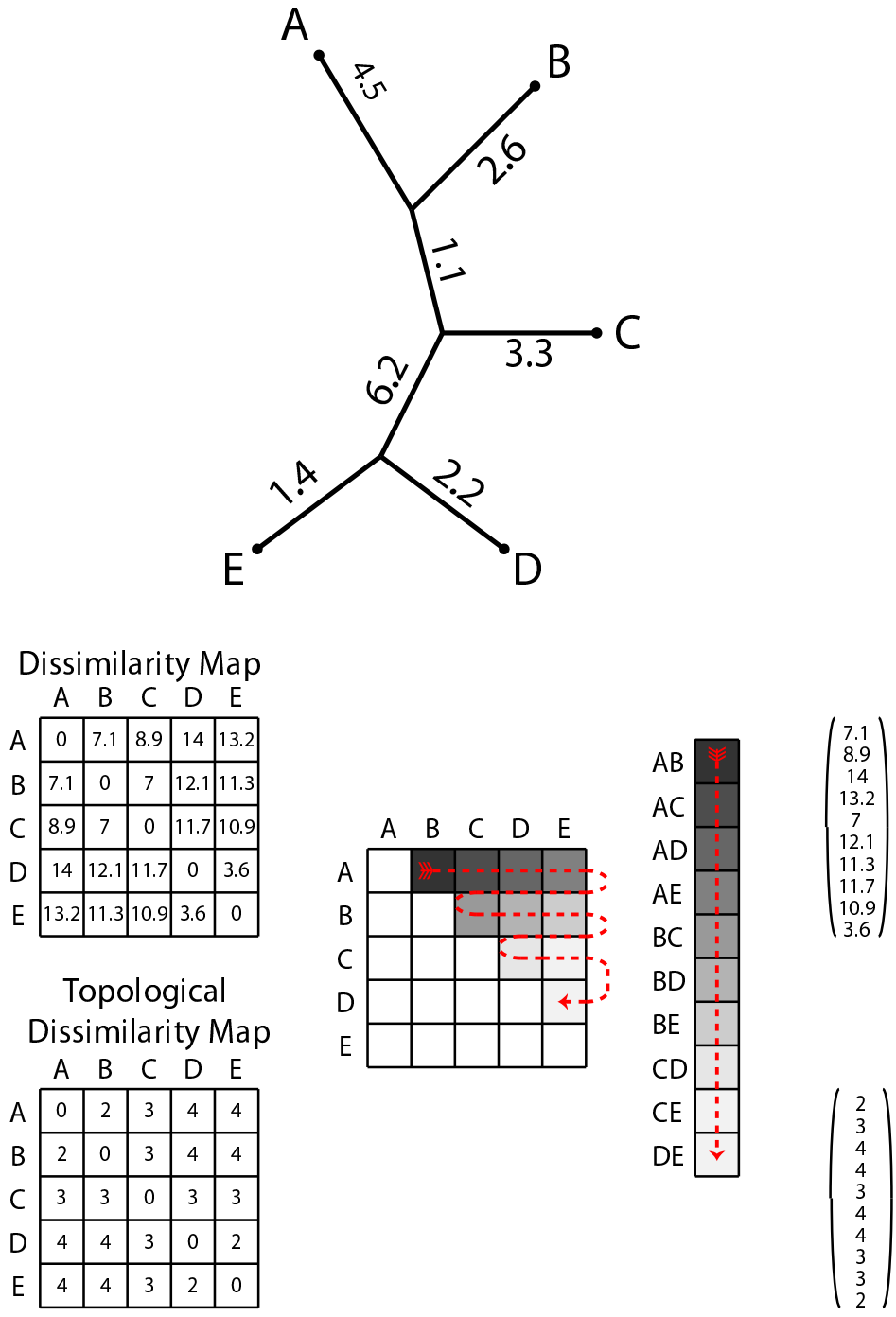

Figure S2. Schematic of how trees are converted to vectors. Numbers on branches in the unrooted tree are branch lengths. In this example, the tree is first converted to either a branch length-based dissimilarity map (matrix of distances between tips) or topological dissimilarity maps (matrix of number of edges between tips). Moving from left to right across rows in one half of a matrix, values are placed into a single column to yield a vector of distances between tips in the tree.

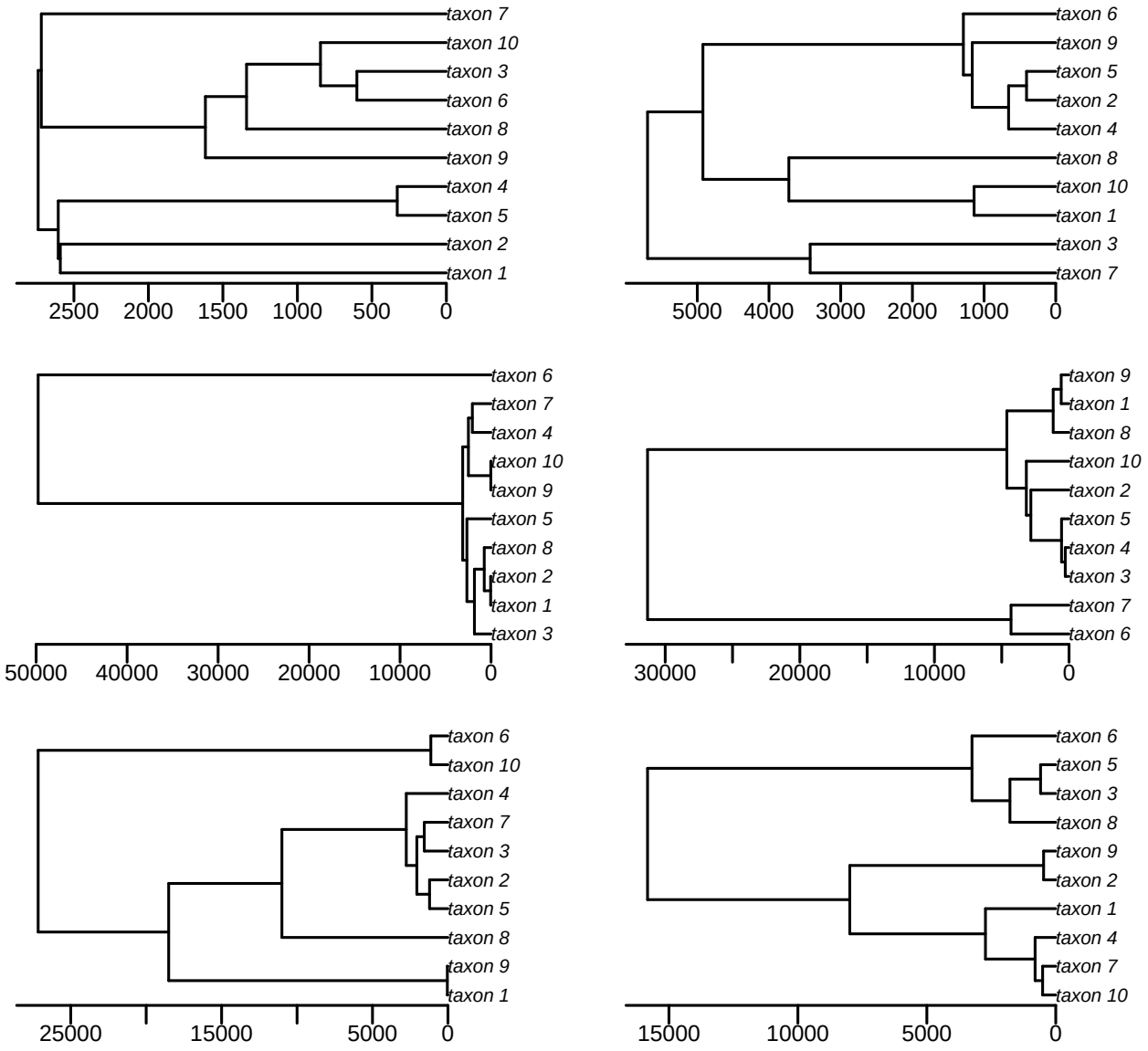

Figure S3. The species trees used to generate gene trees under the coalescent model for the simulation experiments. At top-left is the tree used for the "single" coalescent distribution simulations, while the other trees are used in the "mixed" simulations.

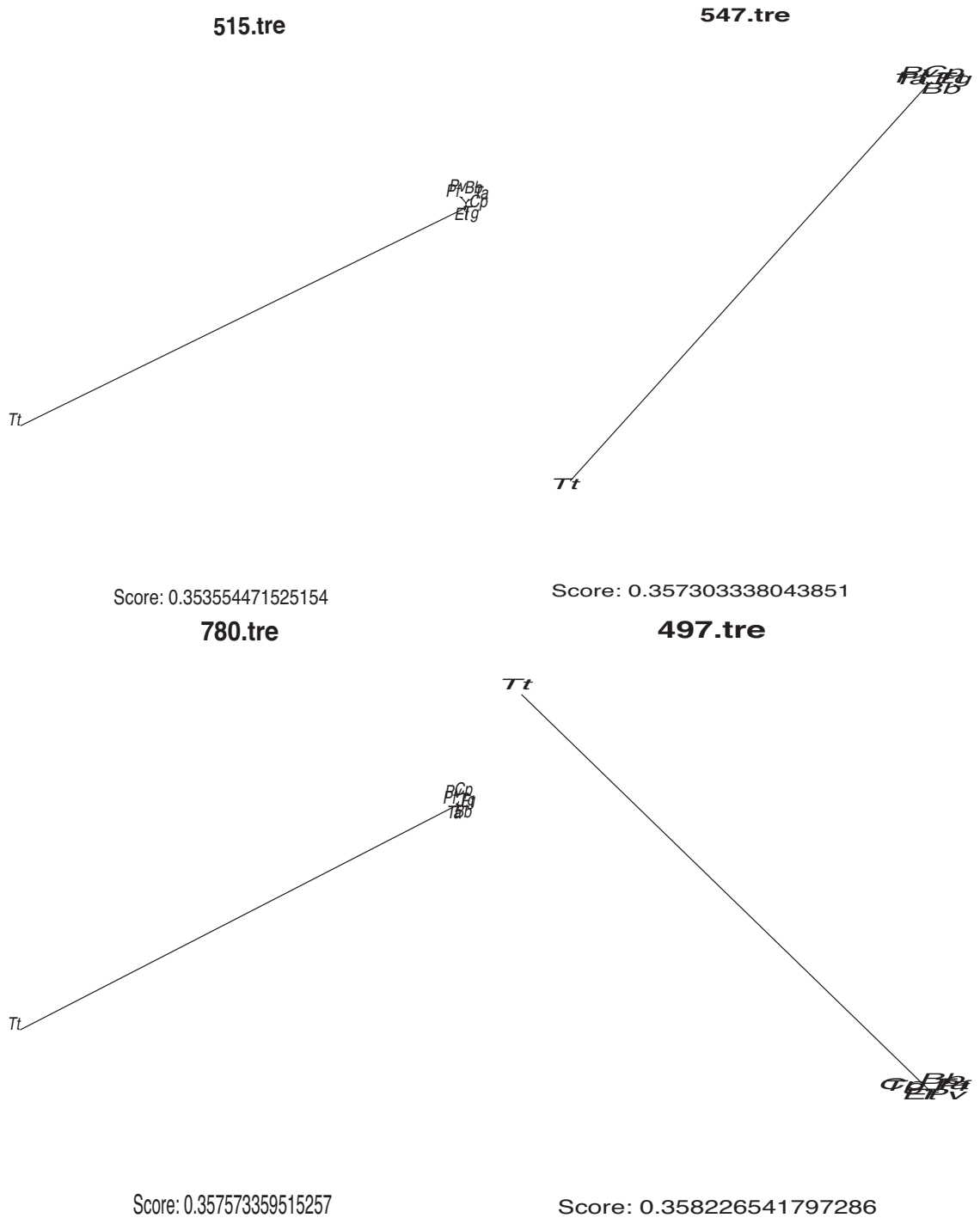

Figure S4. Plots of the first 4 Apicomplexa gene trees identified as outliers. The extremely long branches lead to the identification as outliers, and are likely the result of incorrect annotations of the original sequences.

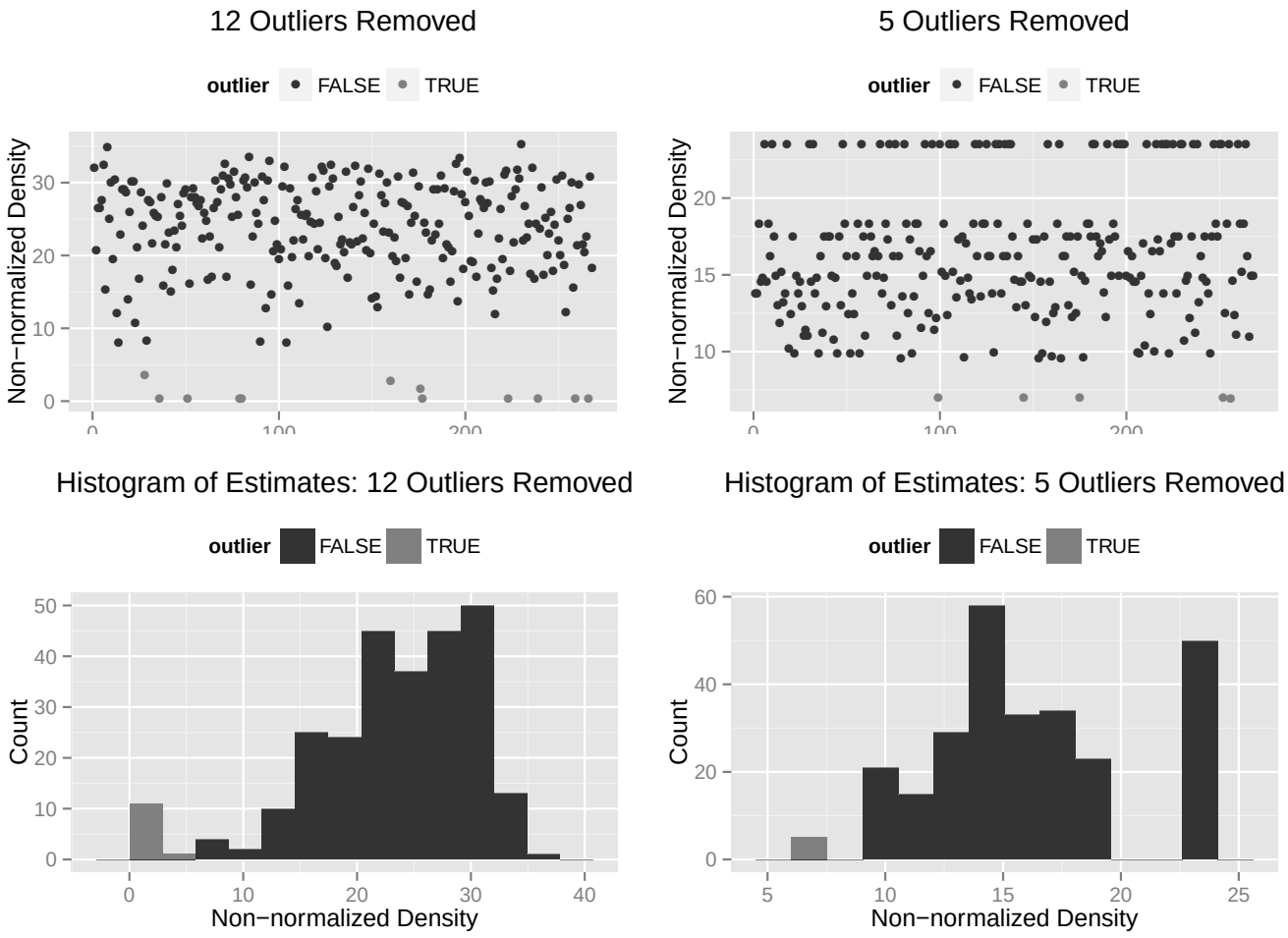

Figure S5. Summary of tree scores for the Apicomplexa data set. In the top row the scores of individual trees are shown. "Tree Index" refers to the ordering of the trees in the input files. In the bottom row, the scores are summarized as a histogram. In the left column are the results computed with branch-length information, while the topology-only results are shown at right.