Supplement for: Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny

Stephen A. Smith, James B. Pease

Supplementary Methods

Simulations

Phylogenies and molecular sequence data were simulated under a number of different scenarios (see Fig. 2 and Fig. S1A). Pure-birth trees of 100 and 1000 taxa were generated using the phyx package and the pxbd program (birth = 1, death = 0, https://github.com/FePhyFoFum/phyx). Only one parameter set was used, as examination of specific diversification scenarios is beyond the scope of this work. Dataset sizes of 100 and 1000 taxa were examined because, while many analyses include larger samples, this data set size allows for faster and more complete all-by-all comparisons. However, *E*- and *p*-values are impacted by the size and composition of the database, and so some additional analyses were conducted on much larger datasets including 100 runs with 100,000 amino acid sequences to demonstrate that the general results hold under scenarios with more sequences present in the database (see Fig. S2). Both nucleotide and amino acid sequences were simulated over 100 replicates over the pure birth trees (i.e., no extinctions) using the program indelible [version 1.03, 1]. Tree height, the distance from the longest tip to the root, is an important variable in these simulations as they determine the substitutions per site and potential saturation in the sequences. Trees with a longer distance from the root to tip will have more substitutions than a shorter one. For both nucleotide and amino acid sequences, simulations were conducted with trees heights of 0.5, 1, 2, 5, and 10 (measured in substitutions per site). For nucleotide sequences, the JC model of molecular evolution was used and sequences had length of 1000 bp (see Fig. S1). For amino acid sequences, simulations were conducted under the WAG model of sequence evolution with sequence length 400 (see Fig. S1A).

Indels are insertions or deletions within gene regions that can introduce noise and errors alignments between two sequences. To examine the effect of indels, datasets for each of the five previously described tree heights for nucleotides and amino acids were generated with indels under the Pascal insertion length distribution with parameters *q* = 0.2, *r* = 4, insertion rate = 0.5, and deletion rate = 0.5. While different and more complex models of nucleotide evolution could be conducted, this was not the primary question intended to be addressed here. There is also no reason to expect that different distributions of indels within the sequence would have different impacts on the relationship of these pairwise sequence alignments to phylogeny.

The trees used for the sequence simulations (described above) all assume a strict molecular clock. Rate heterogeneity of many types can impact phylogenetic reconstruction. Whether distance, parsimony, or likelihood based, phylogenetic methods vary significantly in how they may handle or perform in the face of heterogeneity in the rates of evolution. Here, two different lineage-specific rate heterogeneity scenarios were explored (Fig. 2). In both cases, the pure birth trees were transformed with uncorrelated log normal model [2] of lineage-specific rate heterogeneity using the R package NELSI (version 0.21, https://github.com/sebastianduchene/nelsi). The scenarios, abbreviated RH+ and RH++, had lognormal distributions with mean of -1 and standard deviations of 0.7 and 1.1, respectively. The second scenario was meant to reflect a strong deviation from a strict clock. These were performed for both the nucleotide and amino acid datasets. Datasets with 1000 taxa were only created with lineage-specific rate heterogeneity included. These datasets were not specifically intended to examine increases in dataset size, but rather to facilitate an increased amount of heterogeneity within the dataset.

In addition to lineage-specific rate heterogeneity, base composition bias may complicate phylogenetic analyses [3]. Here, one scenario of bias was examined for both nucleotides and amino acids. For a single random tree created from the same birth-death parameters described above, 100 biased and unbiased datasets of 100 taxa were generated for the tree heights mentioned above. The tree was kept the same to reduce one source of variation. In each of the biased datasets, two clades were allowed to have biased base composition. For amino acids, the bias was Ala = 0.26, Gln = 0.16, Lys = 0.16, Ser = 0.26, and all others with 0.01. For nucleotides, the bias both G and C were give composition of 0.4 with A and T given 0.1. Neither of these are meant to reflect a particular biological system, but instead are intended to demonstrate an example extreme base composition bias.

These simulations were conducted with p4 [3].

Supplementary Results

Sequence similarity correlates broadly with phylogenetic distance

In order to demonstrate potential problems with using similarity scores to infer phylogenetic relatedness, we first needed to characterize the relationship between sequence similarity and phylogenetic relatedness. Using Spearman's ρ, BLAST $-log_{10}(E)$ was found to be correlated with phylogenetic distance using both nucleotide and amino acid sequences (Fig. 3, Table S1). While SWIPE values also show this relationship, lower phylogenetic distances were associated with somewhat lower $-log_{10}(E)$ values from SWIPE compared to BLAST. Specifically, SWIPE *E*-values tend to be lower than BLAST *E*-values for nucleotides (Fig. S2). We also investigated the effect of indels, which may be the result of a number of biological processes. Inserting or deleting segments of a gene region causes pairwise alignment algorithms to have less information on which to obtain a significant hit (i.e., indels break up longer, higher scoring alignments). The relationship between the *E*-value and phylogenetic distance observed without indels remained with simulations including indels (Fig. S3, Table S1). In addition to indels, lineage-specific rate heterogeneity of many types causes significant error for many molecular phylogenetic analyses [4, 5]. Many simulations that include rate heterogeneity show a similar relationship to simulations without rate heterogeneity between *E*-value and phylogenetic distance (Fig. S3, Table S1). No significant difference was shown with rate heterogeneity runs including 1000 taxa. Simulations that included biases in base composition are better compared to the same datasets without biased composition ("CB+" vs "CB0") as these use a single tree on which to simulate data and both use the p4 simulation engine (instead of indelible; Fig. S3, Table S1). Collectively, these results demonstrate that, in general, sequence similarity and phylogenetic distance are grossly correlated though composition bias and rate heterogeneity may somewhat weaken this correlation.

Missing BLAST hits under variable rates and compositions

There were several scenarios in which BLAST missed hits that, as measured by SWIPE, were significant hits (Figure S1B–F, Table S2). SWIPE tends to give lower *E*-values than blastn but highly consistent *E*-values to blastp (as demonstrated in Figure S3). Therefore, judging missed hits based on SWIPE *E*-values was considered conservative, especially for nucleotides. We also found there were few runs that resulted in missed hits for nucleotides *versus* amino acids. In part, this could be due to either the lower *E*-values given to nucleotide hits from SWIPE or the more consistent performance of BLAST. Despite the stronger performance of amino acid searchers, we still found many scenarios with missed hits in amino acids. Particularly problematic were the biased base composition runs that recorded as many as 9217 missed hits (for tree height = 5) and the indel runs with as many as 16367 missed hits (for tree height = 1) (see Table S2). The parameters explored here were not meant to be comprehensive, and so other parameterizations likely also would have generated missed hits. The phylogenetic distance of the missed hits was lower for those simulations with rate heterogeneity than other simulations. While the focus of the analyses here are related to phylogenies and BLAST, the completeness of hits can be particularly important for analyses that presume complete sampling, such as phylostratigraphy. In these analyses, lack of a significant hit is sometimes interpreted as no homology and therefore the point of gene origination. However, as demonstrated here, in simulations where all sequences in a tree are generated from one source, a missed hit would be better interpreted as “no detectable homology”. In contrast, clustering procedures that use all-by-all BLAST results to generate graphs may not require complete hits because missed hits from one query may be compensated for by hits in other queries. Therefore, the incompleteness of hits may not adversely impact the final graph, though this topic in particular merits further investigation.

Ordering of BLAST hits

While completeness can be important for some analyses, the exact order can be important for many others (e.g., reciprocal best hit analyses). In order to examine error in the order of hits, the first significant BLAST hit was examined. This serves not only to address procedures that specifically use the first BLAST hit (e.g., reciprocal best hit analyses), but also gives a simple measure to describe errors in the order of hits. The lowest error rates, 0.3% for nucleotides and 2% for amino acids, was with tree height equal to 10 as generally only very closely related sequences would have successful hits (Figure 7B). As with the other measures, introducing indels, lineage-specific rate heterogeneity, and biased composition increased the error. The highest error rates were found with lower root heights, but this could simply be the result of sequences being very similar at lower root heights. Every dataset with extreme rate heterogeneity in amino acids produced error rates higher than 20% and as high as 24%.

The source of the error in the order of BLAST hits can be demonstrated on a small simulated dataset (Fig. 3). By examining the distribution of the *E*-values for of the incorrect hit, it is easier to discern if the incorrect result is obtained because of sequences being similar (i.e., little molecular evolution between parts of the tree). In many cases, the distribution of *E*-values for incorrect first hits overlapped significantly with those *E*-values for correct first hits (Fig. S4). This highlights the first hit often is not the most phylogenetically related sequence and should not be used for any analyses attempting to identify orthologs. Orthology is a fundamentally a phylogenetic question, and therefore a phylogeny should be constructed to infer orthology [e.g., 6].

References

[1] Fletcher W, Yang Z. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol* 2009; **26**:1879–1888.

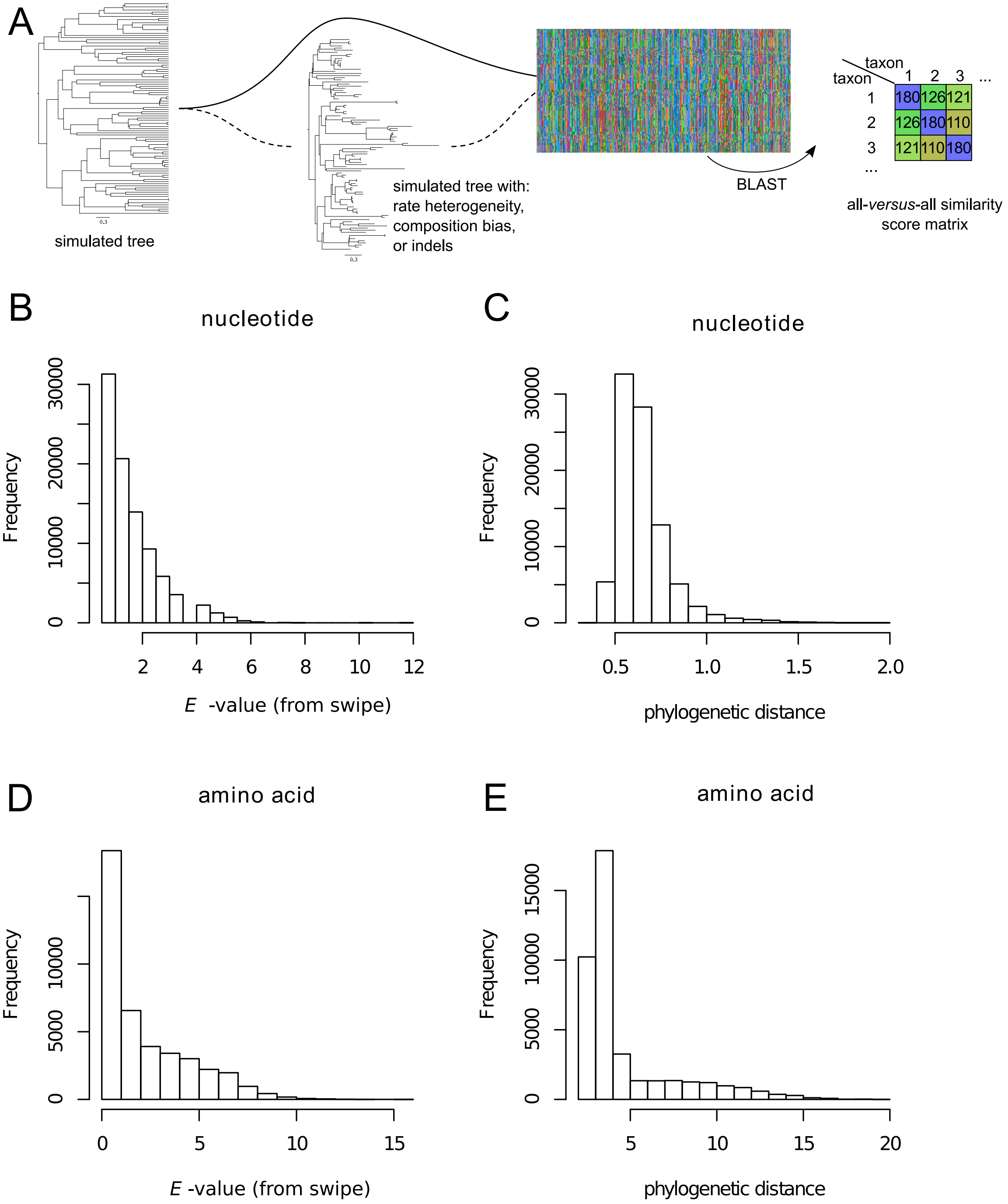
[2] Drummond AJ, Ho SY, Phillips MJ, *et al.* Relaxed phylogenetics and dating with confidence. *PLoS Biol* 2006;**4**:699.

[3] Foster PG. Modeling compositional heterogeneity. *Syst Biol* 2004;**53**:485–495.

[4] Kolaczkowski B, Thornton JW. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 2004;**431**:980–984.

[5] Holder MT, Zwickl DJ, Dessimoz C. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. *Philos Trans Roy Soc B* 2008;**363**:4013–4021.

[6] Yang Y, Smith SA. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol Biol Evol* 2014;**31**:3081–3092.



taxon \ taxon	1	2	3	...
1	180	126	121	
2	126	180	110	
3	121	110	180	
...				

Figure S1: (A) The simulation pipeline including the phylogeny simulated, potential to add heterogeneity, generation of the alignment, and all by all BLAST with $log_{10}(E)$-values shown in the partial matrix. (B–E) Frequencies of $log_{10}(E)$ values (calculated by SWIPE) and phylogenetic distances for hits missed by BLAST. Missed blastn hits for nucleotides with the tree height $= 1$ (B–C) and blastp hits for amino acids and tree height $= 10$ with rate heterogeneity (D–E). More information is provided in Table S1.

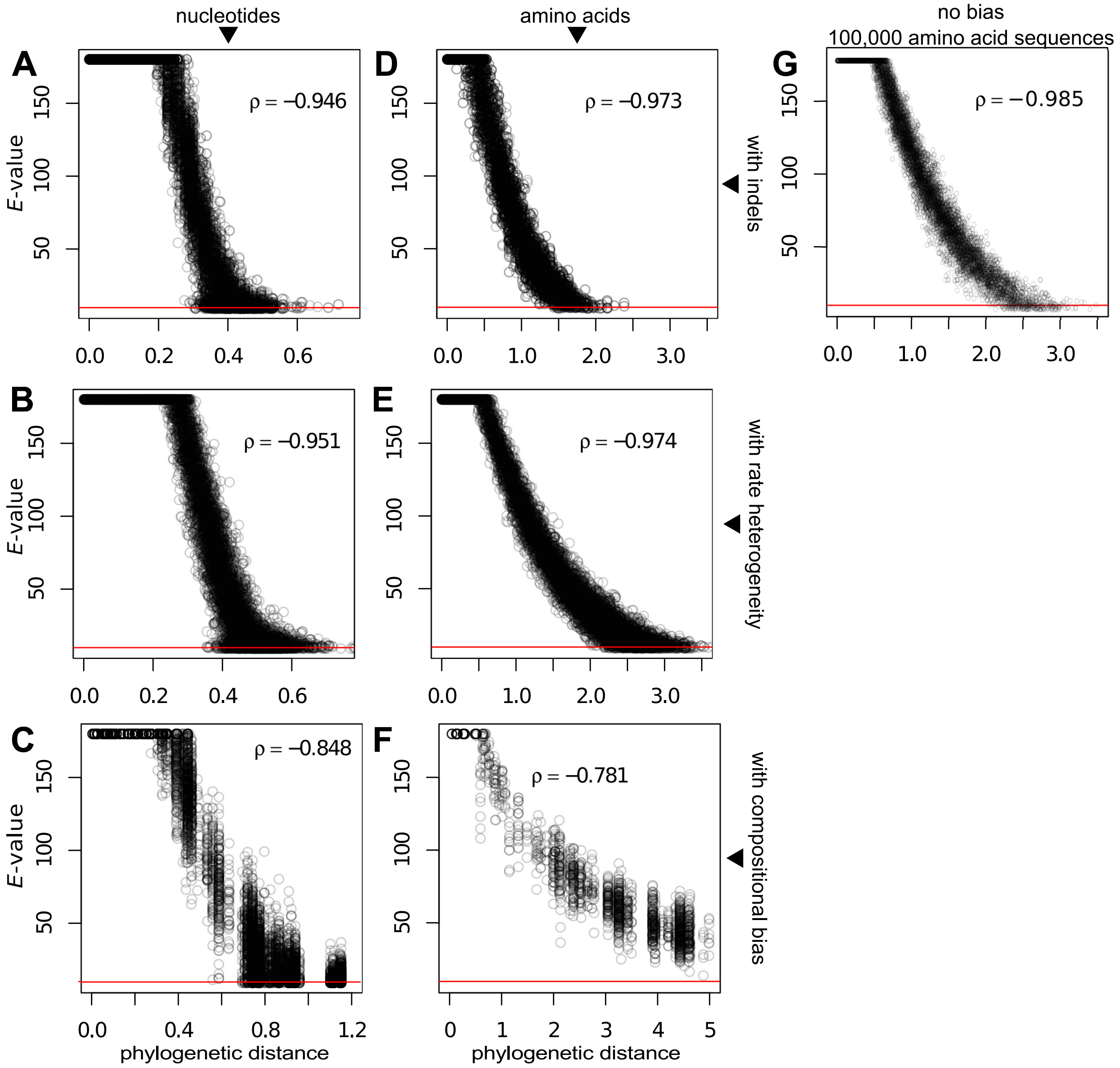


Figure S2: Correlation of phylogenetic distance and *E*-values for nucleotide (A–C) and amino acid (D–G) datasets, shown for tree height of 1 and 10, respectively. Both nucleotides and amino acids are tested under conditions of indels (A & D), rate heterogeneity (B & E), and compositional bias (C & F). The results of no sequence bias for an expanded 100,000 amino acid sequence set is also shown (G). The red line identifies an *E*-value cutoff of $E \leq 10^{-10}$, and BLAST has an implicit maximum of 10^{-180}. Because of the density of points, a random sample of 10,000 points for each plot is shown. Spearman's rank correlation (ρ) is shown on each plot

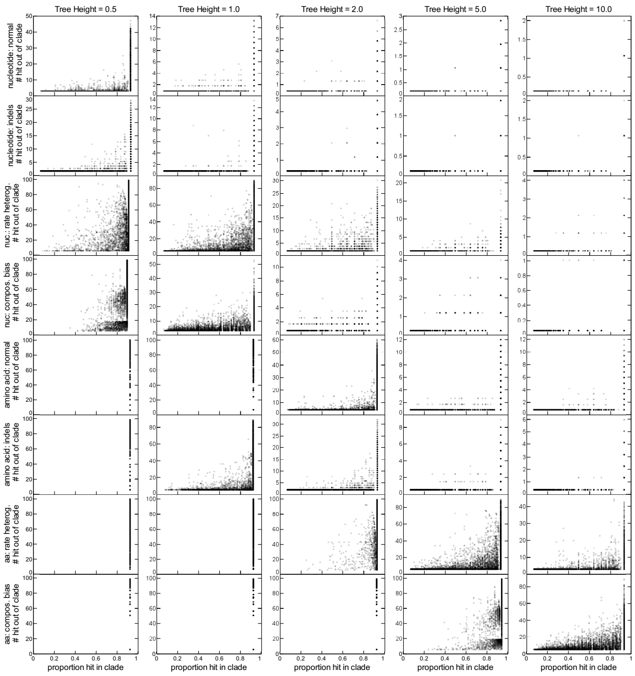


Figure S3: Proportion of those species hit within a clade plotted against the number of hits hit outside of the clade (with annotations in top left chart). (A) and (B) are nucleotides with root height = 1 and (C) and (D) are amino acids with root heights = 5. (B) and (D) include rate heterogeneity. (Points with proportion of hits equal to 1 and number of hits out of clade greater than 0 represent simulations where all sequences inside the clade hit and also additional sequences outside the clade. If points fall at 0 on the vertical axis, no hits were recorded outside the focal clade. If points fall below 1 on the horizontal axis and above 0 on the vertical axis, then hits were incomplete within the clade but hit sequences outside the clade. For visualization purposes, only 20,000 points are plotted for each dataset. See also Figure S4.

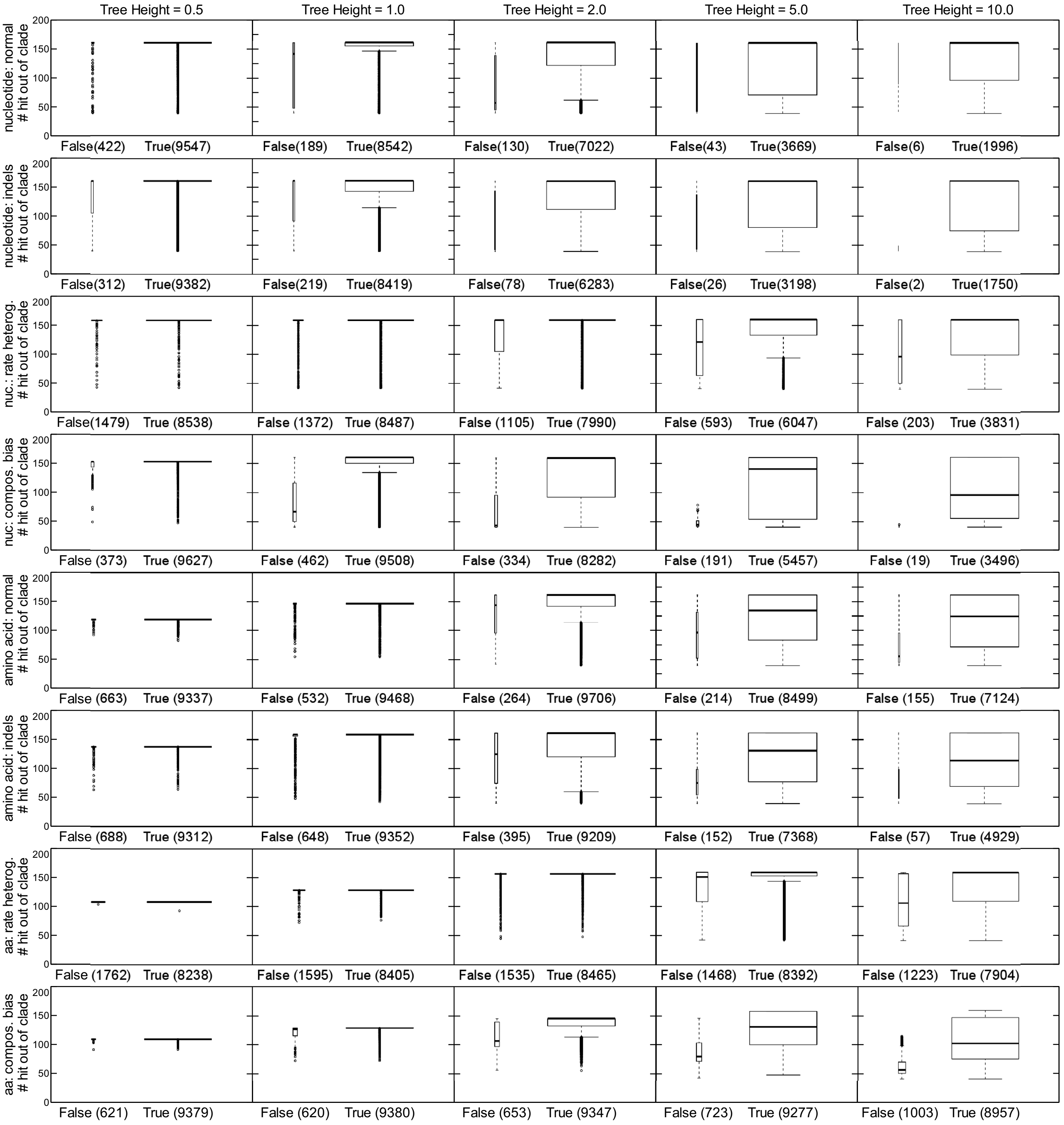


Figure S4: Proportion of those species hit within a clade plotted against the number of hits hit outside of the clade for nucleotide normal datasets. Points with proportion of hits equal to 1 and number of hits out of clade greater than 0 represent simulations where all sequences inside the clade hit and also additional sequences outside the clade. If points fall at 0 on the vertical axis, no hits were recorded outside the focal clade. If points fall below 1 on the horizontal axis and above 0 on the vertical axis, then hits were incomplete within the clade but hit sequences outside the clade. For visualization purposes, only 20,000 points are plotted for each dataset

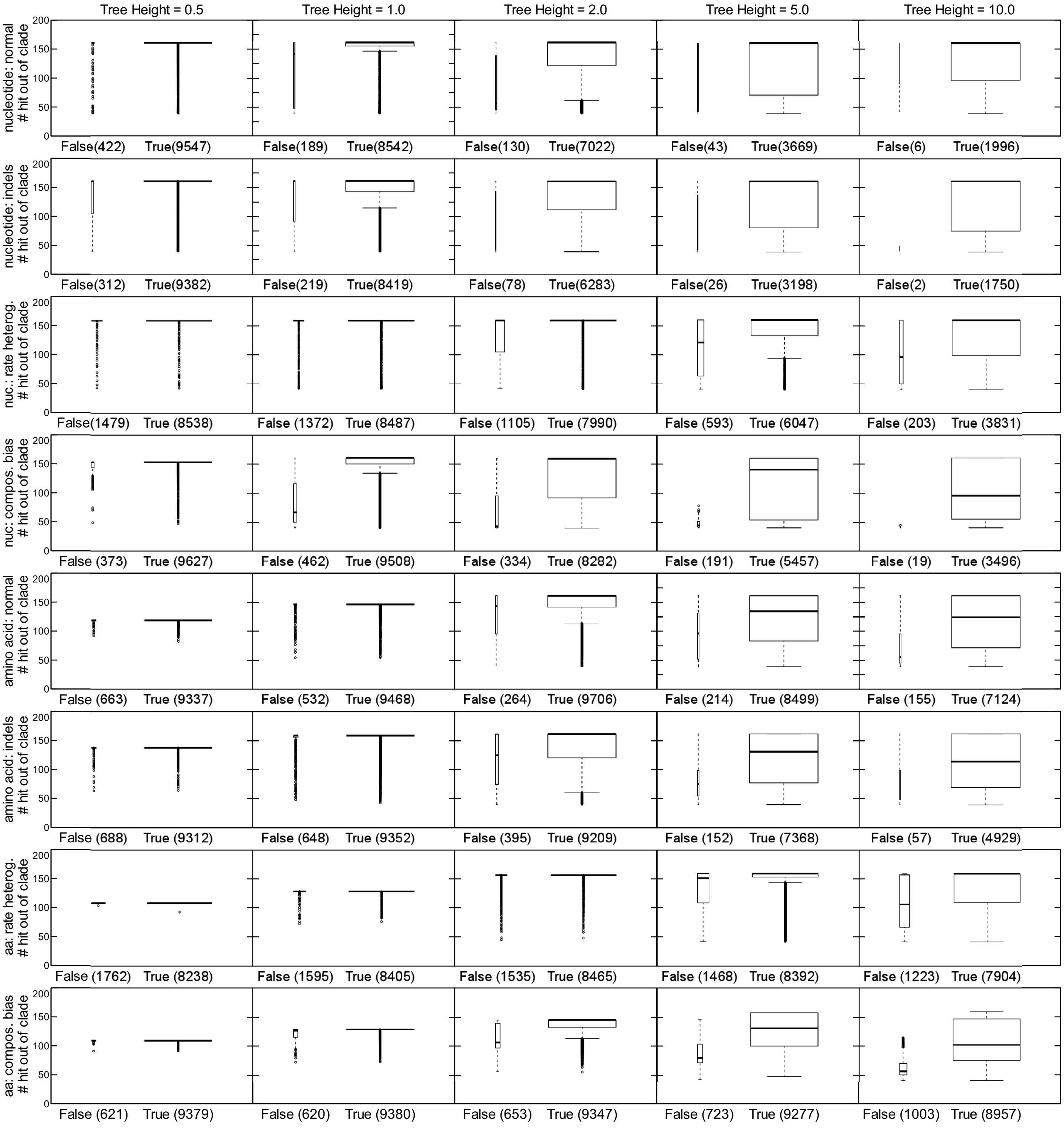


Figure S5: Boxplots for the distribution of *E*-values for phylogenetically correct (True) and incorrect (False) first hits for nucleotide normal datasets. The number of results is denoted in parentheses and by the width of the boxplot

Table 1: Spearman's rank correlation test of phylogenetic distance and $-log_{10}(E)$ using BLAST and SWIPE for amino acid runs and nucleotides identified by tree height, whether the run included indels, lineages-specific rate heterogeneity (RH+ and RH++), and compositional bias (CB+ = comp bias, CB0 = same datasets without comp bias). All had P-values < 0.05.

	height	0.5		1		2		5		10	
		BLAST	SWIPE	BLAST	SWIPE	BLAST	SWIPE	BLAST	SWIPE	BLAST	SWIPE
AA	Normal	-0.907	-0.908	-0.898	-0.899	-0.957	-0.91	-0.973	-0.965	-0.981	-0.953
	Indel	-0.866	-0.865	-0.959	-0.905	-0.97	-0.947	-0.975	-0.955	-0.973	-0.869
	RH+	-0.662	-0.972	-0.969	-0.982	-0.978	-0.981	-0.959	-0.954	-0.974	-0.967
	RH++	-0.406	-0.977	-0.91	-0.986	-0.984	-0.991	-0.981	-0.984	-0.974	-0.974
	CB+	-0.682	-0.683	-0.682	-0.694	-0.66	-0.679	-0.553	-0.585	-0.778	-0.568
	CB0	-0.683	-0.687	-0.682	-0.686	-0.662	-0.666	-0.581	-0.575	-0.813	-0.563
Nuc	Normal	-0.935	-0.871	-0.942	-0.914	-0.937	-0.925	-0.936	-0.852	-0.91	-0.665
	Indel	-0.943	-0.889	-0.946	-0.915	-0.92	-0.91	-0.92	-0.784	-0.923	-0.589
	RH+	-0.955	-0.926	-0.95	-0.892	-0.949	-0.906	-0.934	-0.914	-0.925	-0.862
	RH++	-0.853	-0.962	-0.947	-0.94	-0.95	-0.899	-0.95	-0.922	-0.937	-0.92
	CB+	-0.643	-0.545	-0.85	-0.585	-0.904	-0.768	-0.927	-0.749	-0.906	-0.687
	CB0	-0.64	-0.547	-0.937	-0.744	-0.922	-0.747	-0.936	-0.77	-0.854	-0.588

Table 2: Number of missed BLAST hits with SWIPE $log_{10}(E) > 10$ for amino acid runs and nucleotides identified by tree height, whether the run included indels, linage-specific rate heterogeneity (RH+ and RH++), and compositional bias (CB+ = comp bias, CB0 = same datasets without comp bias). Average phylogenetic distance reported in parentheses.

	Tree Height	0.5	1	2	5	10
AA	Normal	0	60 (2)	672 (2.68)	81 (2.54)	31 (2.48)
	Indel	20 (0.99)	16367 (1.55)	3469 (1.58)	777 (1.56)	197 (1.56)
	RH+	0	0	948 (2.5)	728 (2.6)	121 (2.62)
	RH++	0	0	399 (2.42)	947 (2.53)	545 (2.58)
	CB+	0	0	2 (4)	9217 (8.98)	8991 (13.54)
	CB0	0	0	6 (4)	6642 (9.06)	3934 (14.6)
Nuc	Normal	0	0	0	0	0
	Indel	0	0	0	0	0
	RH+	0	4 (0.46)	0	0	0
	RH++	0	0	0	0	0
	CB+	16 (0.89)	0	0	0	0
	CB0	8 (0.92)	2 (0.77)	0	4 (0.89)	0