Libertas Academica
FREEDOM TO RESEARCH

**Open Access:** Full open access to this and thousands of other papers at http://www.la-press.com.

Evolutionary Bioinformatics

# TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information

Torsten H. Struck

Zoological Research Museum Alexander Koenig, Bonn, Germany.

**ABSTRACT:** Phylogenies of species or genes are commonplace nowadays in many areas of comparative biological studies. However, for phylogenetic reconstructions one must refer to artificial signals such as paralogy, long-branch attraction, saturation, or conflict between different datasets. These signals might eventually mislead the reconstruction even in phylogenomic studies employing hundreds of genes. Unfortunately, there has been no program allowing the detection of such effects in combination with an implementation into automatic process pipelines. TreSpEx (Tree Space Explorer) now combines different approaches (including statistical tests), which utilize tree-based information like nodal support or patristic distances (PDs) to identify misleading signals. The program enables the parallel analysis of hundreds of trees and/or predefined gene partitions, and being command-line driven, it can be integrated into automatic process pipelines. TreSpEx is implemented in Perl and supported on Linux, Mac OS X, and MS Windows. Source code, binaries, and additional material are freely available at http://www.annelida.de/research/bioinformatics/software.html.

**KEYWORDS:** phylogenomics, paralogy, biases, long-branch attraction, conflict detection, tree based

**CITATION:** Struck. TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. *Evolutionary Bioinformatics* 2014:10 51–67 doi: 10.4137/EBO.S14239.

**RECEIVED:** January 13, 2014. **RESUBMITTED:** February 25, 2014. **ACCEPTED FOR PUBLICATION:** February 27, 2014.

**ACADEMIC EDITOR:** Jike Cui, Associate Editor

**TYPE:** Original Research

**FUNDING:** This study was funded by DFG grants STR-683/7-1 and STR-683/8-1 as well as the National Science Foundation USA grant DEB-1036537.

**COMPETING INTERESTS:** Author discloses no potential conflicts of interest.



**CORRESPONDENCE:** torsten.struck.zfmk@uni-bonn.de

## Introduction

In the past years, the reconstruction of phylogenetic data sets has changed from using a single or few genes toward phylogenomic analyses exploiting hundreds of genes in a single study. To deal with the sequence data of so many genes, automatic process pipelines are required. There are two divergent views of the consequence of such vast data sets. First, it has been proposed that congruence between phylogenetic analyses is increasing as random noise in the data sets is strongly reduced.[1] Second, it has been concluded that because of increased artificial signal, which is not canceled out like random noise, true and strong cases of incongruence will now be detected more often.[2] In this case, two different kinds of artificial signal can be distinguished.

First, a crucial step in phylogenomic studies is the determination of orthologous genes across the different species present in the analysis. Usually, automated orthology prediction methods are used at this step.[3–7] However, these prediction methods can erroneously group paralogous sequences as sets of orthologous sequences. As a consequence, this can result in the reconstruction of gene trees rather than species trees.[8–14] For example, Philippe et al.[10] reanalyzed the data sets of Schierwater et al.[15] and Dunn et al.[16] with respect to artificial signal, including the use of manual means to detect paralogous genes in supposed sets of orthologous genes. They were able to detect several cases of paralogy in the first data set. Pruning these sequences from the first data set substantially reduced the very strong support (ie, bootstrap value of 100) for the monophyletic group of Porifera, Ctenophora, Cnidaria, and Placozoa present in the original data set. Owing to the pruning, Porifera was instead placed as sister to all other metazoans, and Cnidaria as sister to Bilateria.[10] An improvement of the Dunn et al.[16] data set also revealed a sister group relationship of Porifera to all other metazoan taxa, instead of Ctenophora being sister to all other metazoan taxa.[10] Similarly, for an annelid data set eight sets of orthologous genes could be

found containing paralogous sequences.[11] In two of these eight sets, the paralogous sequences included had a strong impact on the reconstruction of the concatenated data. Specifically, the taxa affected by the presence of a paralogous sequence—ie *Scoloplos armiger*, *Sthenelais boa*, and *Eurythoe complanata* (all three Annelida) as well as *Owenia fusiformis* (Annelida) and *Cerebratulus lacteus* (Nemertea)—were grouped together with strong nodal support.[11] Thus, the gene tree rather than the species tree drove the phylogenetic placement of these five taxa. Pruning the paralogous sequences of *S. armiger*, *S. boa*, and *E. complanata* or *O. fusiformis* and *C. lacteus* resulted in different placements of the taxa, more in line with their traditional placement. For example, the annelid *O. fusiformis* was now placed within Annelida and not with Nemertea as in the original data set. All other paralogous sequences present in the other six partitions had no or only minimal influence on the analysis of the concatenated data set. Thus, even in phylogenomic studies of several hundreds of genes the artificial signal present only in a few wrongly compiled sets of orthologous genes might mislead the analysis of the entire data set.[11] Unfortunately, the discovery of such cases has so far relied at best on semi-automated detection means, including manual curating of data sets.

Second, numerous studies in the past decades using both real and simulated data sets have shown that systematic biases like increased substitution rates or saturation can positively mislead phylogenetic reconstructions resulting, for instance, in long-branch attraction artifacts.[17–33] Recently, evidence is accumulating that this is also the case for phylogenomic studies. For example, the reconstruction of the eukaryotic tree of life is affected by the presence of both rapidly evolving species[8] and saturation at fast-evolving sites across all taxa.[34] It could also be shown that the placement of Ctenophora as sister to all other metazoan taxa is most likely a long-branch artifact because of increased substitution rates in some species.[9,10] Moreover, this position of Ctenophora is likewise affected by saturation at fast-evolving sites across all taxa.[10,33] Similar analyses revealed that the support for the monophyly of Tardigrada and Nematoda also stemmed from both long-branch attraction and fast-evolving sites across all taxa.[35] Finally, Salichos and Rokas[36] explored the effects of different parameters such as rapidly evolving species, slowly evolving genes, or phylogenetic signal on the reconstruction of the yeast phylogeny using phylogenomic data. They found that selecting genes based on strong phylogenetic signal would decrease incongruence within the final concatenated data set.

Finally, different methods have been proposed to detect conflict in the phylogenetic reconstruction between different partitions of a data set without any *a priori* assumption of the source of conflict.[36–57] Hence, these investigations are driven by the data and not a general hypothesis like saturation or long-branch attraction. This has the advantage that unexpected conflicts can be detected. Such methods are the ILD (incongruence length difference) test, reciprocal Shimodaira and Hasegawa tests, PBS (partitioned Bremer support), or PABA (partition addition bootstrap alteration).[41,44,47–49,51–53,58–61] Partition-by-partition and node-by-node approaches like PABA proved hereby to be the most powerful.[47,54–56] For example, in a study addressing the phylogeny of salamanders using morphological and molecular data the PABA approach could show that the morphological data of pedomorphic species introduced strong conflict regarding their placements as larval or juvenile characters in these species had been compared with adult characters in the other species.[56] The PABA approach could also aid in the decision about the best strategy to ameliorate this problem. Surprisingly, the PABA approach also revealed that over all nodes, the partition of mitochondrial data introduced much more conflict than the morphological data.[56] When the phylogeny of Urostylida (Ciliophora) was investigated using three partitions, the PABA approach revealed that the 18S partition was mainly driving the reconstruction of the concatenated dataset, while alpha-tubulin introduced conflicts at several nodes.[62] In a study addressing the evolutionary history of enterobacterial plant pathogens, the PABA approach substantiated the conclusion that the observed incongruences stemmed from horizontal gene transfer.[63]

Although all these methods have been shown to aid the vindication of artificial signal in phylogenetic and phylogenomic studies, they were usually conducted at best in a semi-automated way, which still required several manual analytical steps during the analyses. Manual exploration of hundreds of genes or trees in the course of phylogenomic studies is time-consuming, not very feasible, and likely to miss an instance. Tree-manipulation programs such as Phyutility[64] or tools calculating systematic biases from alignment data such as BaCoCa[65] exist and allow for the implementation in automatic analysis pipelines. But there has not previously been any program that implements the different methods used in the above studies. These methods comprise a screening procedure for paralogous sequences based on single-gene trees,[8–11] detection of conflict using partition-by-partition and node-by-node approaches utilizing nodal support values,[56] or measurements for saturation and long-branch attraction based on patristic distances (PDs) in the tree.[30,32,33,36] Because all these methods rely on tree-based information such as nodal support or PDs, the program TreSpEx (Tree Space Explorer) has been written in Perl and is presented herein. As it is command-line driven, it can be easily incorporated into automatic pipelines of, for example, phylogenomic studies.

## Implementation of Different Methods in TreSpEx

**Detection and pruning of paralogous sequences.** The procedure to detect paralogous sequences implemented in TreSpEx[8–11] is invoked after an initial determination of sets of supposedly orthologous genes using, for example, HaMStR, OrthoMCL, ReMark, MultiMSOAR 2.0, and PhyloTreePruner.[4–7,66] As discussed above, these automated

orthology predictions have some chance of grouping paralogous genes together. Such misclassifications might subsequently mislead the analysis of the combined data to inferring a gene tree rather than the desired species tree. Thus, before further analyses, additional screening of the sets of supposedly orthologous genes for paralogous sequences should follow the first orthology prediction.[8–11]

To detect such paralogous sequences, TreSpEx implements a screening procedure based on the phylogenetic reconstruction of single partitions (eg, genes) of a phylogenomic data set[8–11] (Fig. 1). For the best tree of each single-partition analysis, this screening identifies all clades possessing a bootstrap value equal to or larger than a certain threshold (eg, 95). These detected clades are regarded as potential indications of paralogy separating paralogs from each other within gene trees. However, strong bootstrap for a clade within a single-partition tree might also be because of true phylogenetic signal for a group of taxa (eg species from the same genus). To separate cases of most likely true signal from cases of paralogy, two different strategies have been proposed. First, clades congruent with clades present in the best tree obtained from the concatenated data set were regarded as exhibiting true phylogenetic signal and "masked" (ie eliminated from further analyses).[8–10] Second, only clades congruent with clades for which independent a priori evidence of monophyly from other sources of data can be shown were masked to avoid circularity[11] (Fig. 1). If required, both masking strategies can be invoked with TreSpEx. However, it should be noted that this masking strategy is not a prerequisite for the screening procedure, especially given the automatic BLAST search option in TreSpEx (see below). One reason for using the masking in the previous studies was to scale down the number of cases requiring further manual inspections such as BLAST searches.

The next step is to decide if the clades so far identified are truly the results of paralogy (Fig. 1). The first criterion is that, in addition to the strong nodal support that first suggested paralogy, a long-branch leads to the suspect clade[8–10] (Fig. 2A). The second is that taxa from a clade with independent a priori evidence of monophyly are found along with taxa outside this clade both within and outside the suspect clade[8–10] (Fig. 2B). Finally, TreSpEx marks very short branches leading to one of the terminal taxa in a suspect clade as indicative of potential cross-contamination[11] (Fig. 2C).

To gain further evidence for paralogy, BLAST searches can be applied using TreSpEx.[8–11] For each partition containing a suspect clade, TreSpEx conducts BLAST searches of all sequences of the partition against two reference databases (Fig. 1). Although different pre-compiled reference databases (eg of *Apis mellifera*, *Bos taurus*, *Branchiostoma floridae*, or *Helobdella robusta*) are provided along with the program, the users can provide their own databases as well. After the BLAST searches, TreSpEx examines whether the best hits of the sequences of the suspect clade are the same as for the other ones of the partition.

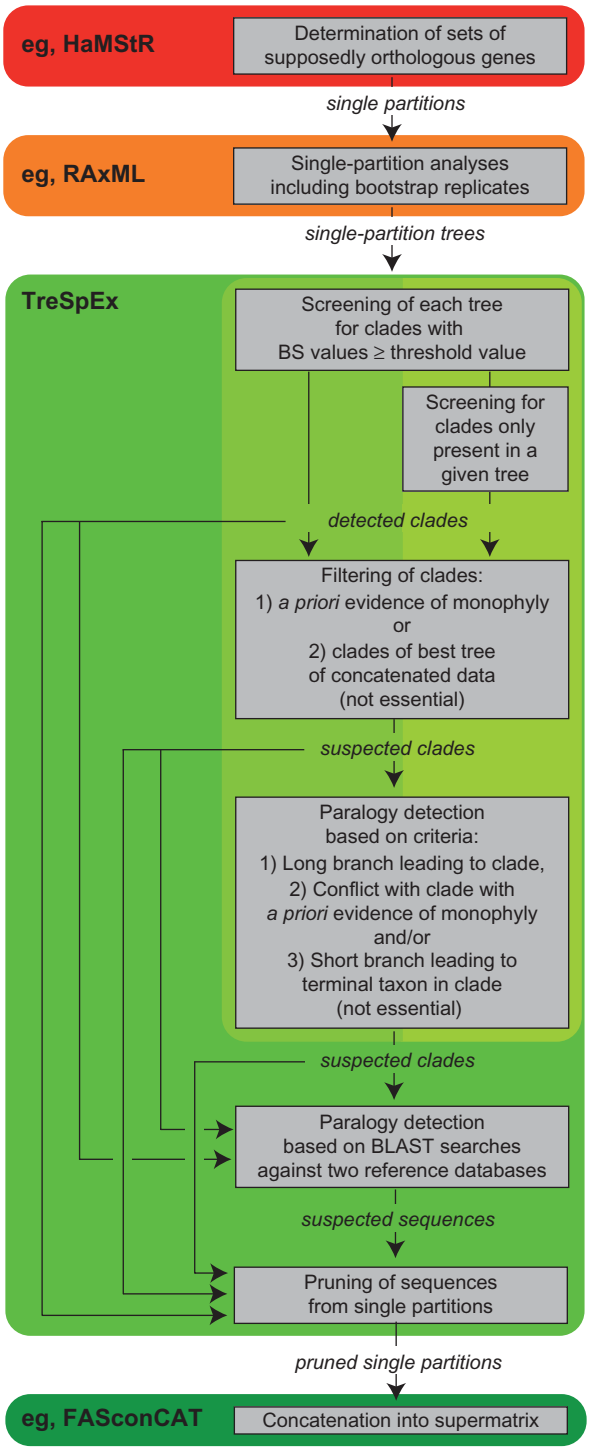


**Figure 1.** Flow-chart of the implementation of TreSpEx in an analytical procedure conducting a screening for paralogy. Programs other than TreSpEx are only examples, and any other program for orthology prediction, phylogenetic reconstruction, and data concatenation can be used.

Different best hits would indicate the presence of paralogy in this partition (ie, set of supposedly orthologous sequences).[11] As part of this comparison, TreSpEx automatically sorts the suspect clades into four different categories: no hits at all in both searches, certain paralogy, no paralogy, and uncertain cases. For the sorting into the latter three categories and for each sequence

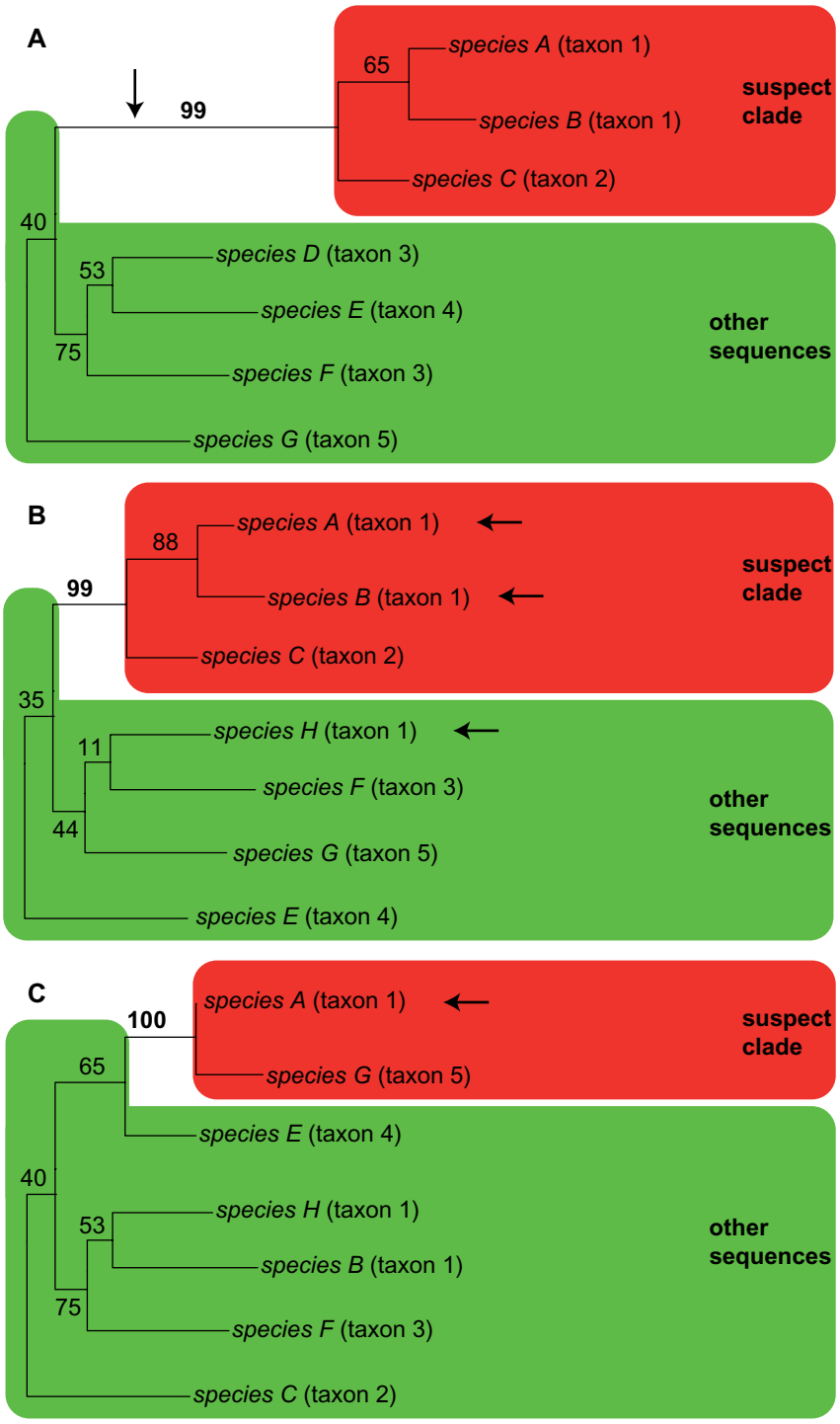


**Figure 2.** Theoretical examples of the sorting criteria in the paralogy screening of TreSpEx. Sorting based on (**A**) an additional long branch leading to the suspect clade (indicated by an arrow), (**B**) taxa from a clade with independent a priori evidence of monophyly are found along with taxa outside this clade both within and outside the suspect clade (indicated by arrows), and (**C**) very short branches leading to one of the terminal taxa in a suspect clade (indicated by an arrow).

of the suspect clade, the proportion of the non-suspect sequences with the same best hit ($p_{ident}$) is determined:

$$p_{\text{ident}} = \frac{n_{\text{ident}}}{n_{\text{other}}} \quad (1)$$

with $n_{\text{ident}}$ the number of non-suspect sequences with the same best hit as the suspect sequence and $n_{\text{other}}$ the number of all non-suspect sequences (ie, not part of the suspect clade). If $p_{\text{ident}}$ of each sequence of the suspect clade is smaller than or equal to a particular threshold value (eg, 0.1) in at least one of the two BLAST searches, this is regarded as a certain case of paralogy. For example, if the suspect clade contains three sequences and all three have a proportion of identical best hits $p_{\text{ident}}$ of 0.1 or lower in one of the two BLAST searches, this

is definitely a case of paralogy. On the other hand, if $p_{ident}$ of each sequence of the suspect clade is higher than or equal to a particular threshold value (eg, 0.85) in both BLAST searches, this suspect clade is regarded as not being paralogous. For example, if for all three sequences of a suspect clade $p_{ident}$ is 0.85 or higher in both BLAST searches, the paralogy has been ruled out. All remaining clades with values between these two thresholds are regarded as uncertain.

Instead of considering, in principle, all possible clades for the paralogy screening, TreSpEx also provides the option to test if the support for a clade or clades in a given tree (eg, the best tree of the concatenated data set) stems from paralogous sequences rather than true phylogenetic signal.[11] This is called a posteriori screening, as it is conducted after a phylogenetic reconstruction of some kind. In contrast, the other option described above considering all possible clades is named a priori screening, as it can be conducted before any phylogenetic reconstruction. Finally, TreSpEx allows for the pruning of affected sequences from the partitions of the data set (Fig. 1).

**Case study I.** To exemplify the potential of TreSpEx to detect paralogous sequences, I used the analysis and single-partition trees of Struck,[11] which are publically available. Struck[11] found that 24 out of 229 partitions contained clades with bootstrap support of 95 or higher, which could not be attributed to clades with a priori evidence of monophyly (Table 1). The clades with a priori evidence of monophyly comprised only members of Clitellata, Sipuncula, Myzostomidae, Terebelliformia, Capitellidae/Echiura, or Serpulidae/Spionidae. Struck[11] used the sequences of each suspect clade as well as those of *Lottia gigantea*, *Capitella teleta*, and *Helobdella robusta* for tblastn 2.2.26+ searches in NCBI against databases of *B. taurus*, *B. floridae*, or *Homo sapiens* to detect cases of paralogy. The sequences of *L. gigantea*, *C. teleta*, and *H. robusta* were part of both the core set for the orthology prediction using HaMStR and the final data set. These BLAST searches and other means showed that 8 out of the 24 partitions constituted sets of orthologous sequences containing paralogous sequences (Table 1). However, only the paralogous sequences present in two of the eight sets had a strong impact on the reconstruction using the concatenated data (see above).[11]

**Table 1.** Comparison of the paralogy screening based on the BLAST search of TreSpEx using the data set of Struck[11] to the original results of the study of Struck.[11] The numbers in the brackets provide the number of partitions found by TreSpEx regarded as cases of paralogy (first position) or non-paralogy (second position) by Struck.[11]

| | STRUCK[11] | TRESPEX |
|---|---|---|
| Suspect partitions | 24 | 25 |
| Paralogy | 8 | 5 (5/0) |
| Uncertain cases | na | 6 (3/3) |
| No paralogy | 16 | 14 (0/13) |
| No hits | na | 0 |

Using TreSpEx for the paralogy screening, only clades with bootstrap values of 95 or higher (Fig. 1) were detected and masked for the same clades with a priori evidence of monophyly as in Struck.[11] This first screening returned all 24 partitions found by the more or less manual screening of Struck[11] and one additional partition (Table 1). The next step was to blast all sequences of the suspicious 25 partitions against the reference databases of *B. taurus* and *B. floridae* and to automatically sort the results. The parameters for the BLAST search and sorting in TreSpEx were an *e* value of 10, and a lower threshold value of $p_{ident}$ of 0.1 and a higher one of 0.85. Thus, if for each sequence of a suspect clade the proportion of the non-suspect sequences with the same best hit (ie, $p_{ident}$) was 10% or less in the search against the database of either *B. taurus* or *B. floridae*, it was assumed that this partition was affected by paralogy. On the other hand, if for each sequence of a suspect clade the proportion of the non-suspect sequences with the same best hit was 85% or higher in the searches against both databases of *B. taurus* and *B. floridae*, it was assumed that this partition was not affected by paralogy. For each partition, BLAST searches returned hits for both the suspect clade sequences and the remaining sequences (no hits = 0 in Table 1). TreSpEx indicated 5 of the 25 partitions as cases of paralogy and 14 as not affected by paralogy; only 6 as could not be placed with certainty (Table 1). More importantly, all five cases of paralogy were also regarded as being affected by paralogy in Struck[11] including the two partitions with strong impact on the analysis of the concatenated data. Similarly, the 14 cases of not being affected by paralogy contained 13 cases already indicated as not being affected by paralogy by Struck[11] as well as the one additional partition found by TreSpEx (Table 1). Thus, no case of paralogy has been erroneously indicated as unaffected by paralogy and vice versa.

However, six partitions could not be assigned with certainty. Three of these partitions were indicated as paralogous by Struck[11]. Struck[11] differentiated two classes of paralogy. In one class, taxa of the core set of the orthology prediction (ie, *L. gigantea*, *C. teleta*, and *H. robusta*) were present within and outside the suspect clade instead of being only outside. Thus, for this partition the core set of the prediction was already a mixture of paralogous sequences, and hence, the hidden Markov model used for the orthology determination was a mixture as well. This resulted already in the analyses of Struck[11] in less clear-cut BLAST results than the results obtained for the other class of paralogy. In the other class, all core-set taxa were placed outside the suspect clade. For the cases with core-set taxa present within and outside the suspect clade, additional evidence of paralogy such as signature amino acids or differences in *e* values had to be used to determine paralogy.[11] Moreover, Struck[11] usually pruned only the paralogous sequences of the suspect clade from the partition, retaining all other sequences for further analyses. But when core-set taxa were present within and outside the suspect clade, the entire partition was excluded from further analyses

as already the core set of the orthology prediction was affected by paralogy.[11] Interestingly, all three partitions indicated as uncertain by TreSpEx and as cases of paralogy by Struck[11] were the ones in which the taxa of the core set were present within and outside the suspect clade, and where the entire partition had been excluded.[11] In contrast, the five partitions indicated as paralogous by both studies were those in which the taxa of the core set were placed outside the suspect clade and where only affected sequences were pruned.[11] TreSpEx had thereby indirectly separated the cases of paralogy requiring different kinds of data exclusion.

However, this separation was not perfect as also three partitions indicated as non-paralogous by Struck[11] were marked as uncertain by TreSpEx. In these three cases, in one of the two searches $p_{ident}$ was below 0.85, whereas it was above 0.85 in the other one. In the searches with the low $p_{ident}$, two highly similar hits were returned by the BLAST searches against the database of either *B. taurus* or *B. floridae*, most likely indicating inparalogs within this reference database. For example, for the suspect sequences of partition 23680, the BLAST searches against *B. floridae* returned the gene ID 43208 as the best hit and 25071 as the second best with *e* values of, e.g., $1 \times e^{-62}$ and $3 \times e^{-62}$, respectively. Both are probably NM23/nucleoside diphosphate kinase subunits. The blast results of the non-suspect sequences also returned these two gene IDs as the two best hits, and in some cases, 25071 was slightly better than 43208. The maximal difference between the two was an *e* value of $6 \times e^{-54}$ for 25071 and $2 \times e^{-53}$ for 43208. Using different reference databases might circumvent the problem of inparalogs to a certain degree. However, even using very closely related reference taxa the chance of, for example, species-specific inparalogs will still be present. As for the uncertain cases of paralogy above, in these cases additional evidence is required. Therefore, TreSpEx not only sorts the partition into different categories but also provides additional results related to the BLAST searches such as the actual result of the BLAST searches and compilations in the results for each partition comprising gene IDs, max scores, and *e* values for both the best hit and a confidence set of best hits. With this information, it is easier to assess whether uncertain cases are cases of paralogy or not. Moreover, instead of screening 229 partitions manually for paralogy, only 6 partitions would have to be analyzed with more scrutiny based on the results of TreSpEx.

**Detection of conflict.** TreSpEx in combination with a program for phylogenetic reconstruction such as RAxML[67] or PhyloBayes[68] can also be used to detect conflicts in data sets based on the PABA principle.[55,56] Using this principle, conflict is detected on a node-by-node and partition-by-partition basis utilizing nodal support values. The PABA principle was first proposed using bootstrap values,[55] and can also be employed with any other nodal support values such as Bremer support (PABSA, partition addition Bremer support alteration) or posterior probabilities (PAPPA, partition addition posterior probability alteration).[56] For reasons of simplicity, it will be referred to as PABA herein. For each node and partition, the alteration of nodal support is determined as partitions are added to the data set. During this process, as each partition is added the order of addition is also taken into account, that is if a partition is added as first, second, or third partition and so on. To condense the results, the mean values of alteration are calculated for each partition and position of addition. These results then allow the alteration of support values to be examined for indications of conflicts. For example, if a partition always decreases the support for a node regardless of its position of addition, this would indicate a conflict between this partition and the other partitions in this data set concerning this particular node (for more details, refer to Struck[56]).

Except for the phylogenetic reconstructions, TreSpEx provides the first implementation of the other three steps of the PABA approach. First, TreSpEx generates all possible combinations of partitions of a data set as Phylip files for phylogenetic reconstructions in the second step (Fig. 3). For example, if a data set comprises six partitions, TreSpEx will generate all possible data sets comprising only one, two, three, four, five, or six of the six partitions. An option at this step is to generate only a range of possible combinations. For example, instead of generating all possible data sets containing one to six of the six partitions, only all possible data sets with four or five of the six partitions can be generated. Second, after the phylogenetic reconstructions of the data sets with the different combinations of partitions, TreSpEx summarizes bootstrap support values or posterior probabilities across all data sets for each of the nodes that can be found in at least one of the trees. Third, for each node, partition, and position of addition, TreSpEx calculates the alteration in nodal support and averages the results in accordance with the position of addition (eg, added as fourth or fifth partition).

Two different statistical tests proposed by Struck[56] can also be conducted by TreSpEx (Fig. 3). To assess whether the positive contribution of a partition outweighs, if present, its negative impact on a given set of nodes, a Wilcoxon-Signed-Rank test[69–73] is conducted. A given set of nodes can, for example, be all nodes of the best or an alternative tree. The results of this test can be used to guide the decision if an entire partition should be excluded from the analysis instead of just a few sequences. To test the significance of the results of each partition at each node and position of addition, a permutation test similar to ILD or LILD (localized ILD) tests[41,49] is implemented in TreSpEx. For this permutation test, TreSpEx randomly assigns positions to partitions of the same sizes as the predefined partitions used for the calculation of the original values. Then the same analyses are conducted as for the original partitions. Thus, the test can reveal if the value found for a partition at a node and position of addition can be obtained just by chance because of randomly partitioning the data set. Such tests were lacking in the first proposal of this approach.[55]

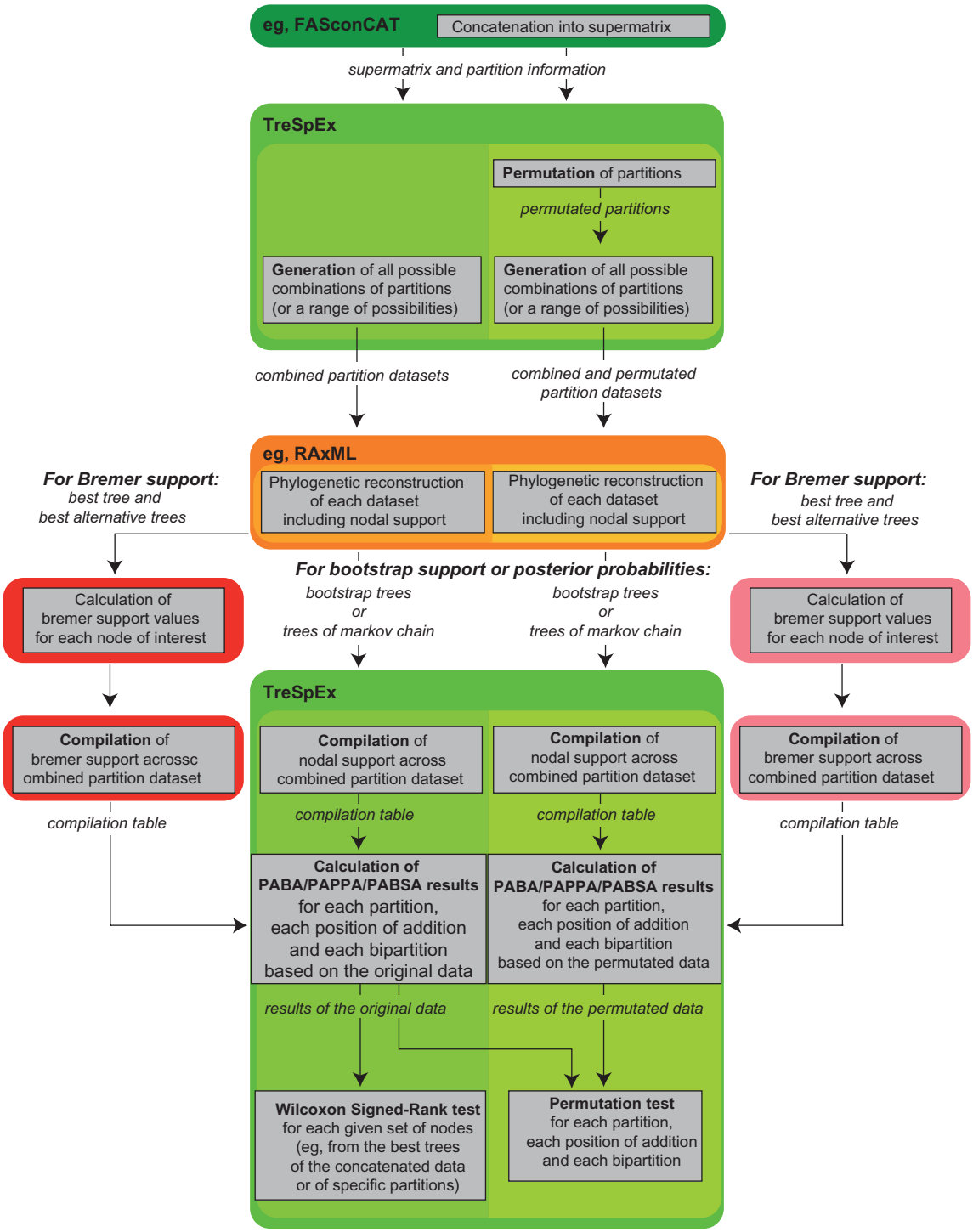


**Figure 3.** Flow-chart of the implementation of TreSpEx in an analytical procedure conducting a detection of conflict using the PABA, PAPPA, or PABSA approach. Programs other than TreSpEx are only examples, and any other program for data concatenation and phylogenetic reconstruction can be used.

**Case study II.** Herein I exemplified the potential of TreSpEx to detect conflict based on the PABA principle using the data set of Struck et al.[55]. By manual inspection of trends in alteration of nodal support, Struck et al.[55] highlighted three cases in the COI (cytochrome oxidase I) partition and three in the 28S partition as interesting (Table 10 in Struck et al.[55], Figure 4). The COI partition introduced a strong conflict at node 4 and a slight conflict at node 13. Hidden support was revealed at node 12[55] (Fig. 4). The 28S partition introduced the strongest conflict at node 9 and a slight conflict at node 13. Again, hidden support was revealed at node 8[55] (Fig. 4). For the present demonstration, I generated all 15 possible combinations of the four partitions 16S, 18S, 28S, and COI using TreSpEx. In addition, the 15 possible combinations for each of the 100 permutated data sets were also created with TreSpEx, resulting in an additional 1,500 datasets. In the second step,

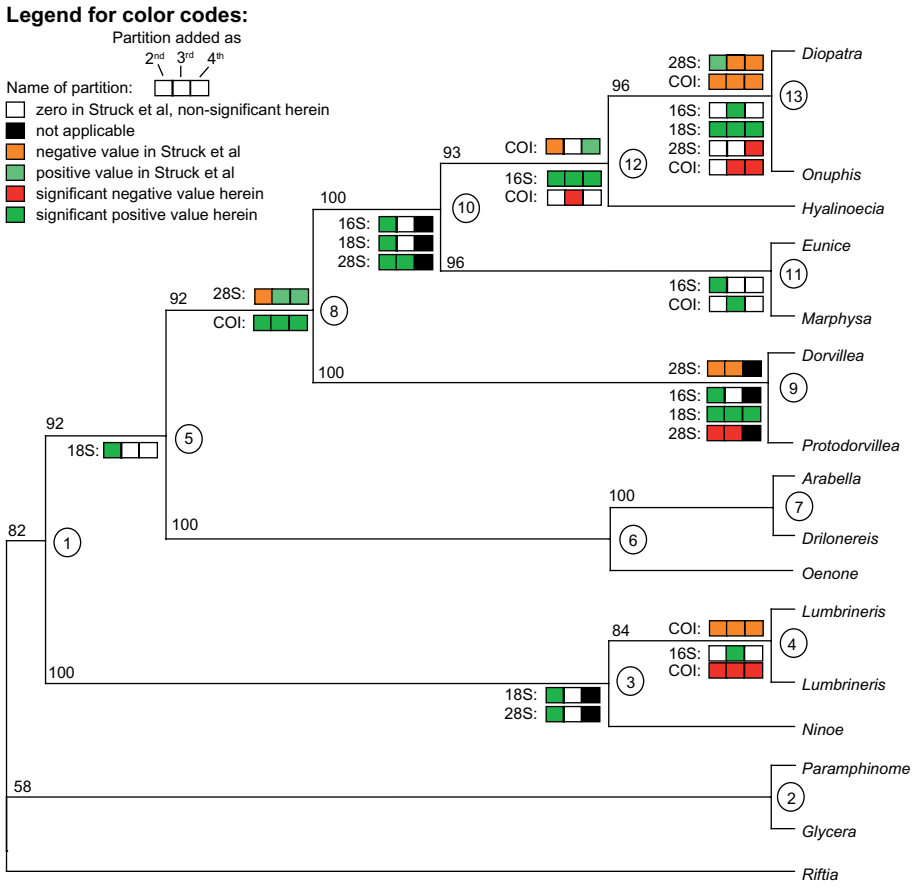


**Figure 4.** Cladogram of the best maximum likelihood tree based on the analysis of all four partitions herein (same topology as in Struck *et al.*[55]). The nodes are labeled in circles to the right of the node with the same number as in Struck *et al.*[55] Bootstrap support values for the nodes from the ML analysis herein are given above the branch at the beginning. The partitions and nodes, which were highlighted as interesting alterations in Table 10 in Struck *et al.*,[55] are indicated above the branches. The order of the cells after the partition name is in the order of addition from second to fourth. Orange indicates a negative value in Struck *et al.*,[55] light green a positive value, and white a value of 0. A black cell means that the PABA approach was not applicable as the node was maximally supported before and after the addition of the partition. The partitions with significant PABA results determined by TreSpEx are shown below the branches. Green indicates a positive value and red a negative value. If a partition is not shown at a node at all below the branch, the values obtained for this node were not significant and/or the approach was not applicable.

phylogenetic analyses were conducted for each of the 1,515 datasets using RAxML 7.3.1[67] with the GTR + Γ + I substitution model and 100 bootstrap replicates.[74] For the original as well as the 100 permutated data, the bootstrap values were individually summarized in the third step. In the fourth and final step, the PABA results were calculated, and both a permutation test and a Wilcoxon-Signed-Rank test were conducted to test the significance of the individual PABA results and the overall contribution of a partition to given sets of nodes, respectively. Therefore, two sets of nodes were tested. The first set comprised all nodes of the ML (maximum likelihood) tree of the concatenated data set of all four partitions and the second all nodes of the ML tree of the data set with only the 28S data.

Although a different ML algorithm was used herein than by Struck et al.,[55] of the six instances discussed by Struck et al.[55] five were indicated here as showing significant conflicts. Especially, the nodes 4, 9, and 13 were affected (Fig. 4). Only node 8, which was regarded as revealing hidden support[55], was not indicated. Interestingly, as in Struck et al.[55] the 28S partition was not able to overwhelm the support for node 9 present in the other three partitions, when added as fourth (see black box in Figure 4). Furthermore, the permutation test of TreSpEx revealed partitions that contributed significantly more to a node than would have been expected given their size. The value obtained by the original data was significantly higher than the values obtained by just randomly assigning positions to a partition of the same size, for example, 18S and 28S contributed strongly to node 3 when added as second and similarly, 16S, 18S, and 28S to node 10. Thus, using the permutation test of TreSpEx allowed also the detection of strong support for a particular node by a partition. For example, given its size COI was significantly contributing to the bootstrap support of node 8 independent of the position of addition. The 18S partition was also positively contributing to this node as bootstrap support increased by an amount of 33–47%; but given the size of the 18S partition, this contribution was not significant. For the other two partitions, the contribution was

also generally positive, but close to zero. Therefore, support for this node did stem from COI and 18S, but considering its size COI contributed more to the support.

The Wilcoxon-Signed-Rank test showed that over all nodes all partitions contributed positively to the ML tree of the concatenated data set. For each partition, its contribution significantly outweighed its negative impact at least at one position of addition (Table 2). Interestingly, although the 28S partition introduced a strong conflict at node 9[55] (Fig. 4) over all nodes, its contribution significantly outweighed its negative impact when added, for example, as second. Its contribution was even stronger than that of the 18S partition when added as second, despite the fact that the 18S partition did not introduce any conflict. Only when COI was added as fourth partition, its negative impact at two nodes outweighed its positive contribution, but this was still not significant. The stronger negative impact of COI when added as fourth was because of two reasons. First, because of the other three partitions most nodes were already maximally or nearly maximally supported and the COI partition could not add any more measurable bootstrap support to these nodes when added as fourth. On the other hand, the conflicts at nodes 4 and 13 persisted. However, this was also the case for the 28S partition and node 9. This led to the second reason. While other partitions (namely 16S and 18S) significantly contributed to node 9 and to a certain degree 13, this was less prominent at node 4. More specifically, maximal bootstrap support was already achieved at node 9 by the concatenation of 16S, 18S, and COI, and the conflict introduced by the 28S partition when added as fourth was not strong enough to decrease the bootstrap value below maximal support (see black box in Figure 4 at node 9). Hence, the Wilcoxon-Signed-Rank test can help to reveal very strong conflicts in a partition when nodal support values with a maximal support value like bootstrap values or posterior probabilities are used.[56] This is different when, for example, Bremer support values are used, which do not have a maximum value.[56]

**Table 2.** Results of the Wilcoxon-Signed-Rank test for the analyses of the data of Struck *et al.*[55] as well as the nodes of the best ML tree of all four partitions and only the 28S partition (Figs. 3 and 4 in Struck et al.[55]). Significant results of the test at $\alpha = 0.05$ are indicated by a star (*). – = over all considered nodes, the impact of the partition was negative; + = over all considered nodes, the impact of the partition was positive.

| SET OF NODES | PARTITION | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 18S | | | 16S | | | COI | | | 28S | | |
| | Added as | | | | | | | | | | | |
| | 2nd | 3rd | 4th | 2nd | 3rd | 4th | 2nd | 3rd | 4th | 2nd | 3rd | 4th |
| 28S only | – | – | –* | + | – | + | – | – | –* | + | +* | + |
| All four genes | + | + | +* | + | +* | + | +* | + | – | +* | +* | + |

In contrast to the nodes of the ML tree of the concatenated data set, the 28S partition was not surprisingly the only partition that over all nodes contributed positively to the nodes of the ML tree of the 28S data set (Table 2). When added as third partition, this contribution was significant. The 16S partition was also contributing to this set of nodes to a certain degree, but the 18S and COI partitions clearly had an overall negative impact, which was significant when they were added as fourth.

**Detection of long-branched taxa and partitions.** To assess long-branch attraction based on tree-specific properties, two means have been mainly used. Average evolutionary rates of complete data sets or their partitions have been calculated as a proxy for long-branch attraction, and faster evolving partitions were excluded in favor of slower evolving ones.[75] This is also called the slow-fast method. However, the problem of long-branch attraction stems from heterogeneous branch length and, hence, evolutionary rates between taxa within a data set or partition.[30,32] Therefore, distances from the root of the tree to each taxon (ie, tip-to-root distances) are used as a taxon-specific measurement for long-branch attraction.[8] However, the recognition of long-branched taxa by tip-to-root distances heavily depends on the root of the tree by definition. For automatic process pipelines, this can pose severe problems in the recognition of long-branched taxa or data sets severely affected by long-branch attraction. When the root of the tree cannot be objectively placed as different outgroup taxa root the tree differently, it cannot assess which root-based distance would be trustworthy. Therefore, TreSpEx also calculates a new long-branch score, the LB (long branch) score[76] (Fig. 5). The score utilizes PDs, ie, the distance between two taxa based on the connecting branches, and is based on the mean pairwise PD of a taxon $i$ to all other taxa in the tree relative to the average pairwise PD across all taxa ($a$):

$$LB_i = \left( \frac{\overline{PD_i}}{\overline{PD_a}} - 1 \right) \times 100 \qquad (2)$$

In particular, the score measures for each taxon the percentage deviation from the average PD. Moreover, it is independent of the root of the tree. As both tip-to-root distances and LB scores are taxon specific, direct comparisons between different data sets are not possible. To facilitate comparisons between data sets or partitions, TreSpEx provides two values for each data set or partition. First, the standard deviation of either the tip-to-root distances or the LB scores is calculated as a measure of heterogeneity. Second, the average of the upper quartile of either the tip-to-root distances or the LB scores is determined as a representative value for the taxa with the longest branches.

**Case study III.** The annelid taxon Myzostomidae is well known for its long-branch problem.[77–80] In many molecular phylogenetic studies, it is attracted to the longest outgroup taxon. In the analyses of Struck,[11] Myzostomidae was also placed

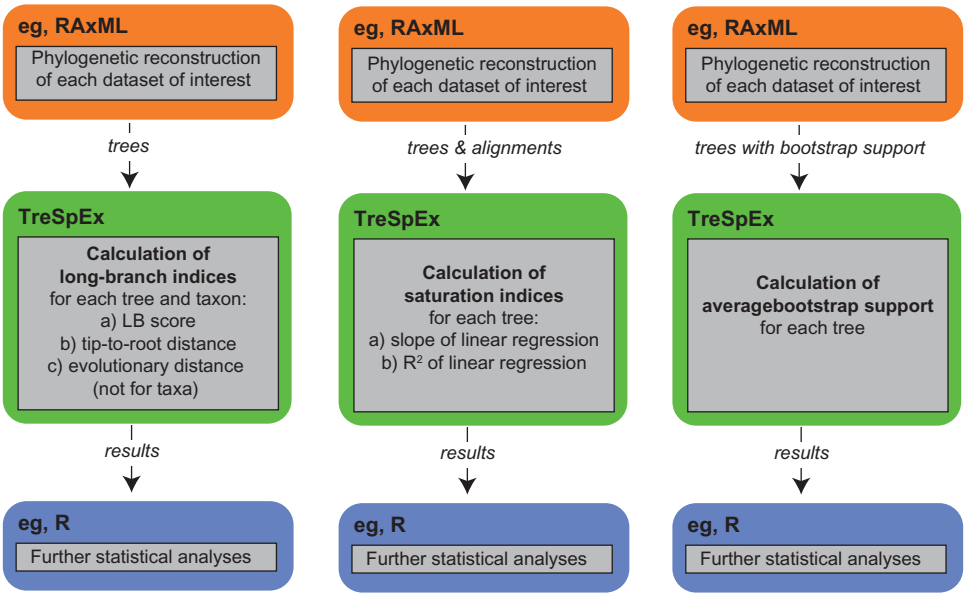


**Figure 5.** Flow-chart of the implementation of TreSpEx in an analytical procedure analyzing long-branch attraction (**A**), saturation (**B**), or phylogenetic signal (**C**). Programs other than TreSpEx are only examples, and any other program for statistical analyses and phylogenetic reconstruction can be used.

with the longest outgroup taxon, the ectoproct *Bugula*. Therefore, the capability of TreSpEx to detect long-branched taxa is shown using the tree in Figure 9 of Struck.[11] Taxon-specific LB scores and tip-to-root distances were calculated with TreSpEx and density plots were generated with *R*. The values of the LB score or tip-to-root distances generally followed a normal distribution, but the curve was slightly skewed toward higher values and additional smaller optima could be observed (Fig. 6A, B, and D). Irrespective of the index, the two myzostomid species were the taxa with the highest values (Fig. 6A, B, and D). However, these analyses also showed the power of the LB score; the ectoproct *Bugula* exhibited the highest LB score of all other taxa (Fig. 6A), so that the long-branch attraction of Myzostomidae toward the ectoproct *Bugula* is clearly indicated. Using tip-to-root distances with the original root (the brachiopod *Terebratalia* and the nemertean *Cerebratulus*), this attraction is not quite as obvious. The ectoproct *Bugula* is part of the skewed right part of the distribution, but not clearly set apart (Fig. 6B). Second, rerooting the tree with the ectoproct *Bugula* did not alter the results of the LB score, but those of the tip-to-root distances. Now the ectoproct *Bugula* is not part of the skewed right part of the distribution, but is placed in close vicinity to the global optimum of the distribution (Fig. 6D), so that in this case the long-branch attraction between Myzostomidae and the ectoproct *Bugula* would be concealed. Recently, Ryan et al.[81] proposed that Ctenophora is the sister group to all other Metazoa, instead of the traditional view that Porifera is the sister group to all other Metazoa including Ctenophora. Such a position of Ctenophora had been suggested before,[16] but on the other hand, it had been shown that Ctenophora was affected by long-branch attraction[9,10,33] (see above). Thus, to assess if Ctenophora was affected again by long-branch attraction, the LB scores were calculated in TreSpEx using the tree shown in Figure 3 of Ryan et al.[81] Again the distribution generally appeared normal (Fig. 6C), but also had a shoulder starting at about 7 and two smaller optima at higher values. Whereas the values of most Porifera species were part of the normal distribution, the values of all outgroup and ctenophore species were placed in the skewed part of the distribution starting at the shoulder. Although this is not as clear-cut as in the myzostomid example, this result might indicate that the position of Ctenophora in the study of Ryan et al.[81] could again be affected by long-branch attraction. This is further substantiated by the fact that support for the position of Ctenophora as sister group to the other metazoan taxa substantially decreased from posterior probabilities of 0.71 to 0.02 as distantly related outgroup taxa were excluded from the data set (Table 1 in Ryan et al.[81]). The tendency of long-branch attraction to increase support with the addition of distant outgroups is well known and has been suggested as an indication of possible long-branch attraction.[9,10,30] On the other hand, support for the traditional position of Porifera as sister group to all other metazoan taxa strongly increased from 0.29 to 0.98 with the removal of outgroup taxa (again Table 1 in Ryan et al.[81]).

In addition to taxa genes, data sets or partitions can be analyzed with respect to long-branch attraction as well. For example, the 229 genes of the data set of Struck[11] were analyzed using TreSpEx, and density plots were generated with the aid of *R*. All five indices calculated by TreSpEx showed a normal distribution with a skewed and ragged right tail (Fig. 7). For example, the distribution of the values of the standard deviation of LB scores showed a small shoulder close to the global optimum and a clear shoulder at a value of 58.1

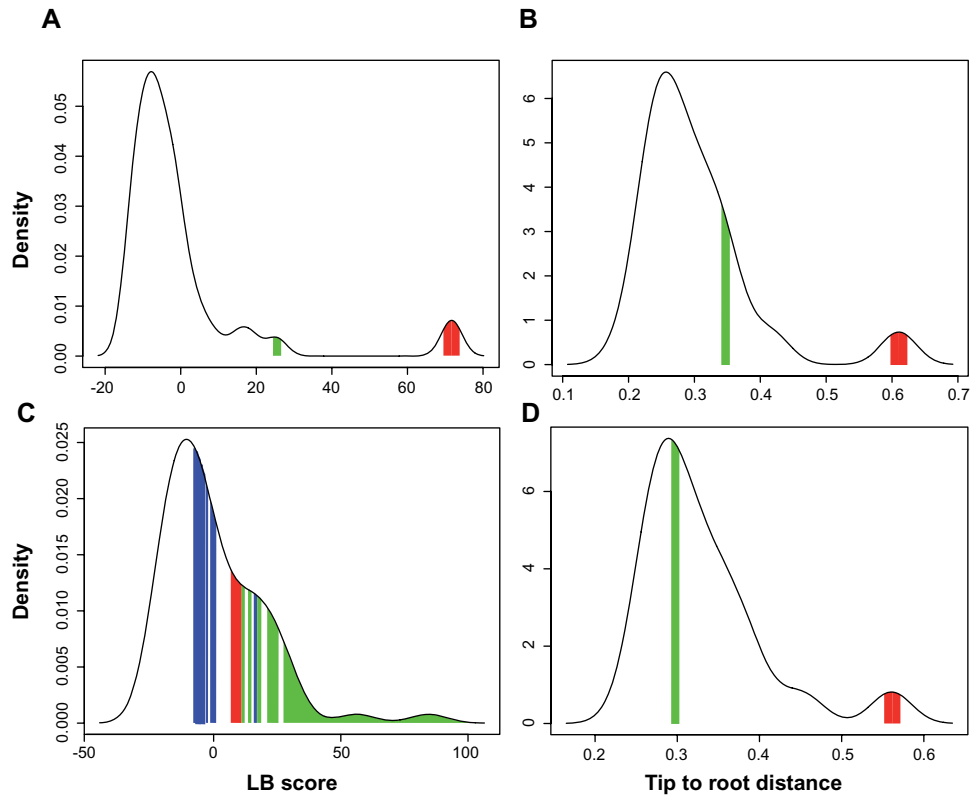


**Figure 6.** Density plots generated with *R* of taxon-specific LB scores (**A** and **C**) or tip-to-root distances (**B** and **D**) of the trees shown in Figure 9 of Struck[11] (**A** and **B**) or Figure 3 of Ryan et al.[81] (**C**). Tip-to-root distances in (**D**) are based on a tree rerooted with the ectoproct *Bugula* instead of the brachiopod *Terebratalia* and the nemertean *Cerebratulus* as in Struck[11] and (**B**).

**Notes:** Red = values of either Myzostomidae (**A**, **B**, and **D**) or Ctenophora (**C**). Green = values of either the outgroup taxon *Bugula* (Ectoprocta) (**A**, **B**, and **D**) or all outgroup taxa (**C**). Blue = values of Porifera (**C**).

(Fig. 7A). This was followed by five small local optima. Starting at the 58.1 shoulder, the distribution comprised 24 of the 229 genes. Similar results were obtained for the other four indices, with a maximum of 41 genes being part of the skewed and ragged right tail (Fig. 7). Even more detailed insights can be gained by analyzing taxa versus gene matrices, for example, of LB scores with the aid of heat-map analyses in combination with hierarchical clustering (Fig. 8). For example, the heat map showed that Myzostomidae was long branched more or less across all genes. On the other hand, the mollusks *Crassostrea* and *Lottia* were long branched only in a few genes. Thus, the results of TreSpEx allow very thorough investigations of the long-branch problem even using phylogenomic data sets. By this latter approach, affected sequences can be specifically pruned from the data set instead of either entire partitions or taxa.

**Detection of saturation and phylogenetic signal.** Saturation is known to influence phylogenetic reconstructions even using phylogenomic data sets.[10,33] Assessment of the degree of saturation can be determined either by the visual inspection of saturation plots or based on specific values measuring the degree of saturation.[10,23,31,33] These values are either the slopes or the $R^2$ values of linear regressions of PDs against uncorrected distances $p$ for each gene, data set, or partition.[10,33] These specific values have the advantage that their calculation can be automated and, thus, implemented into processing pipelines. This is not possible for the visual inspection of plots.[23] Therefore, TreSpEx calculates the slope of and the $R^2$ fit of the linear regression of PDs against uncorrected distances $p$ for each gene, data set, or partition (Fig. 5).

Finally, it has been proposed to assess the resolution power of partitions or genes within a larger phylogenomic data set using the average bootstrap support of each partition.[36] Moreover, average bootstrap support has also been used to determine whether alterations to the data sets like exclusion of data or taxa were beneficial or detrimental to the phylogenetic reconstruction.[80] Hence, TreSpEx also calculates average bootstrap values across all nodes of a given tree (Fig. 5).

**Case study IV.** To exemplify these two features of TreSpEx, the 229 genes of the data set of Struck[11] have been used again in combination with density plots generated with *R*. The rationale for both saturation indices is that the lower the value, the higher is the degree of saturation.[33] The slope of the linear regression was generally in a range of 0.1–0.4 (Fig. 9A). Only very few genes showed higher slope values, and 32 genes possessed a slope value of less than 0.089. At 0.089, a slight shoulder could be detected at the left-hand side of the distribution. The distribution of the $R^2$ values showed more

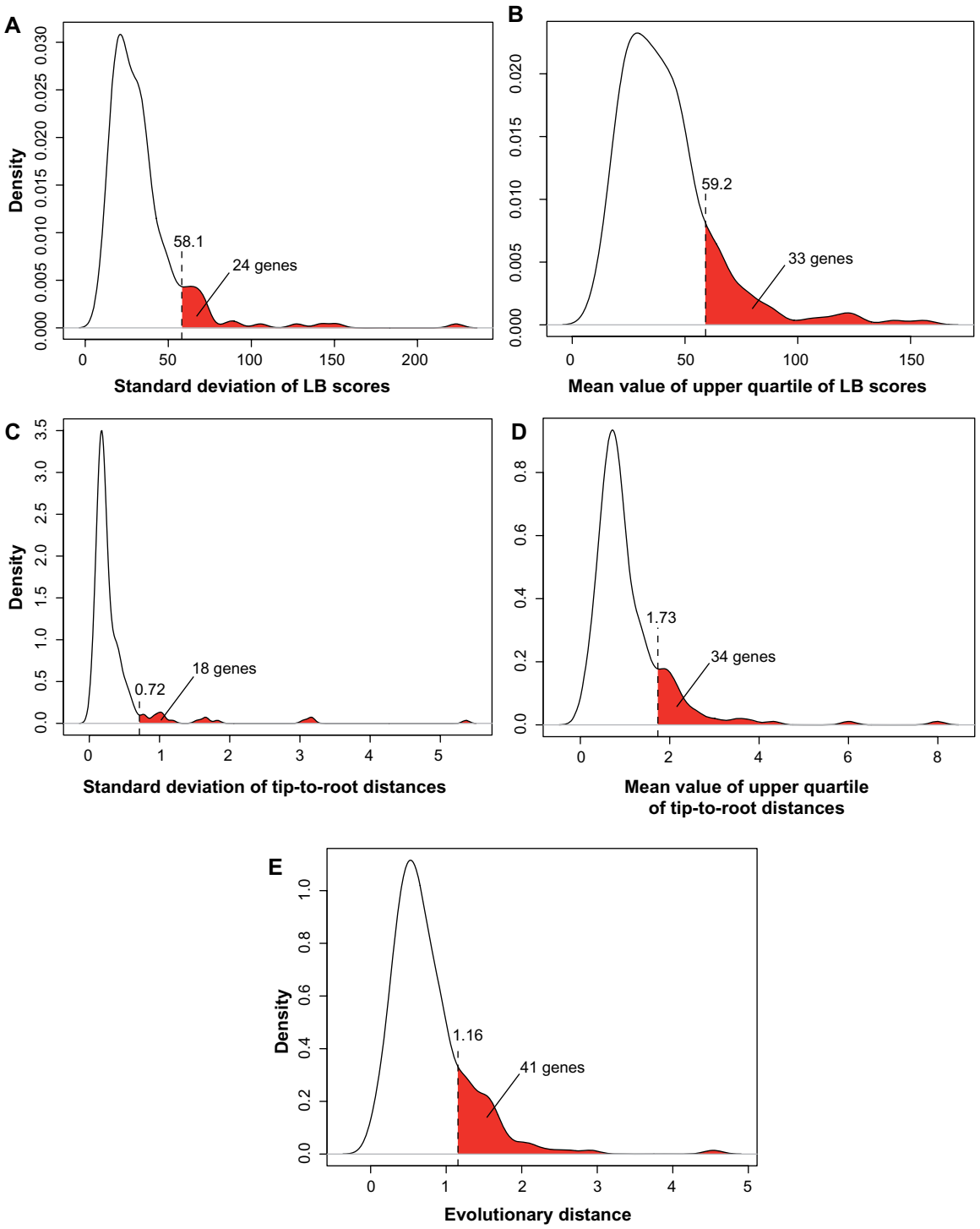


**Figure 7.** Density plots generated with *R* of different gene-specific long-branch indices for the 229 genes present in the data set of Struck.[11] (**A**) Standard deviation of LB scores measuring heterogeneity; (**B**) average of the upper quartile of LB scores representing the taxa with the longest branches; (**C**) standard deviation of tip-to-root distances measuring heterogeneity; (**D**) average of the upper quartile of tip-to-root distances representing the taxa with the longest branches; and (**E**) average PD as a proxy for genes affected by long-branch attraction. Red areas indicate deviations from the normal distribution.

pronounced shoulders than that of slope values (Fig. 9B). A total of 91 genes had an $R^2$ value of less than 0.469. On the other hand, this also meant that more than 100 genes had an $R^2$ value above 0.5 and, hence, a relatively good fit to a linear regression. Without saturation the expectation is that PDs and uncorrected distances $p$ show a perfect linear correlation, as no adjustment for multiple substitutions along the branches is necessary. On the other hand, in case of saturation with multiple substitutions convergence of the curve is expected and thus, a deviation from the linear regression.[33] Therefore, the better the fit to a linear regression, the less saturated the data. Analysis of the average bootstrap support

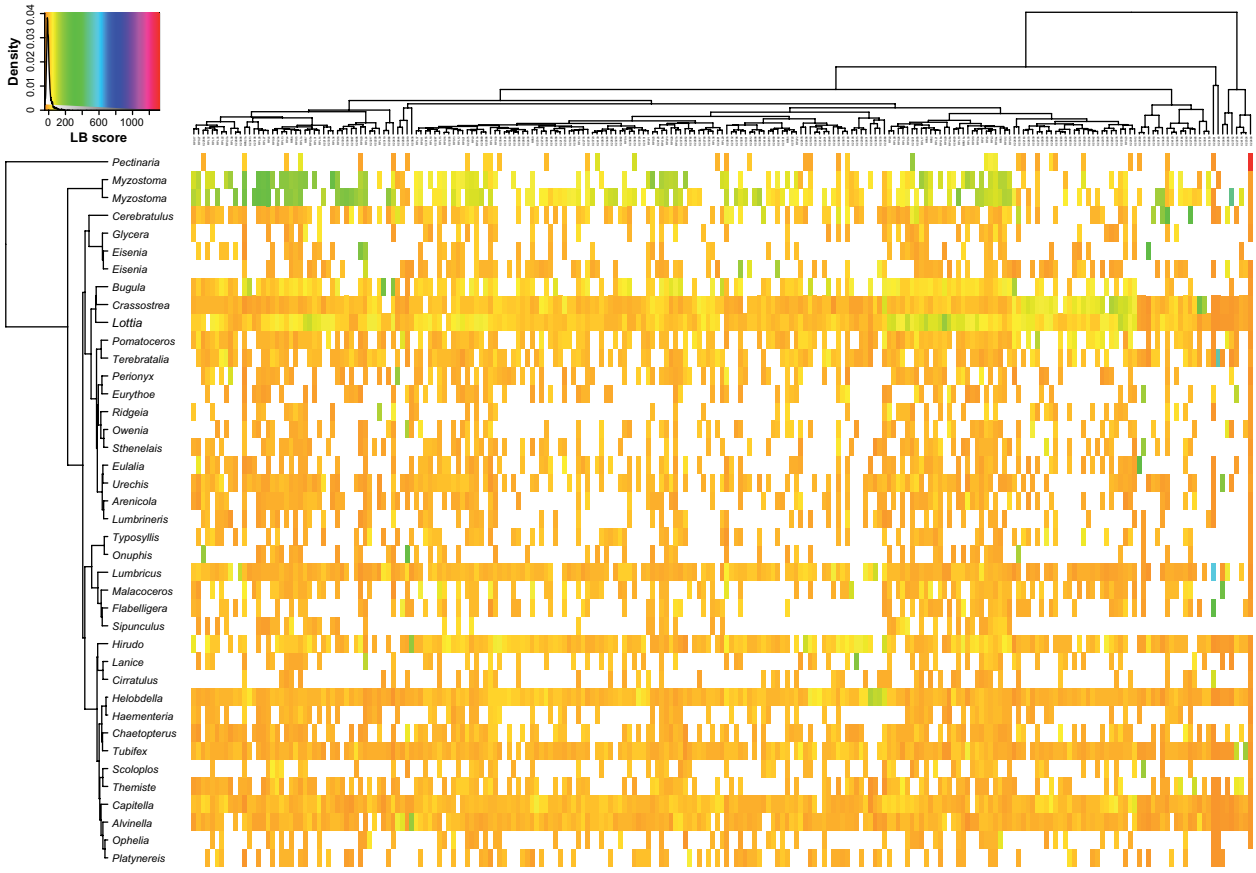


**Figure 8.** Heat map in combination with hierarchical clustering generated with *R* of the taxon-specific LB scores for each of the 229 genes of Struck.[11] Rows show taxa and columns genes. Color key and density plot of LB scores are provided in the upper left corner of the heat map. White cells in the matrix indicate that the taxon was lacking in that gene.

shows that most genes had very low average bootstrap support values in the range from 20 to 60% with a global optimum at 32% (Fig. 9C). Only 20 genes had an average bootstrap support of 60% or higher. This number is relatively low, but not surprising as it had been shown before that single genes will not be able to resolve the annelid phylogeny, but that substantially increased numbers of genes are necessary.[31,82,83]

## Run-time Statistics of TreSpEx

For each data point, the calculation of the run-time statistics was repeated 10 times to assess the variability of the run time. For all steps of the paralogy screening, the run time shows a linear increase with an increasing number of trees or data sets (Fig. 10A). The pruning and a priori screening options are the fastest, requiring less than 0.5 seconds even for 200 data sets or trees, respectively. The a posteriori screening takes about three times longer than the a priori screening, but still needs less than 1.5 seconds. Interestingly, the masking option has no substantial influence on the run time of the screening (Fig. 10A). By far and not surprisingly, the longest time is taken by the BLAST searches of the sequences of the partitions with suspect clades against the two reference databases. When the screening procedure started with 200 trees, this step took about 140 seconds. Thus, a complete paralogy screening procedure starting with 200 trees and, thus, 200 data sets requires a total time of less than three minutes to finish, including the BLAST searches and cleaning of the data sets.

Similarly, the run times of the determination of the average bootstrap support, long-branch, or saturation indices also increase more or less linearly with the number of trees and data sets (Fig. 10C). Calculation of the long-branch indices of 200 trees requires less than a quarter second, and the average bootstrap supports only about 0.3 seconds. The calculation of the saturation indices takes substantially longer, but is still achieved in about 1.5 minutes for 200 trees and data sets. This is because of the fact that for this index, the pairwise PDs have to be calculated from the trees as well as in parallel the uncorrected pairwise distances $p$ from the alignments.

Finally, the run times of the three PABA options follow an exponential growth as the number of partitions increases (Fig. 10B). This is because of the exponential growth of the number of possible combinations of data sets with an increasing number of partitions[56] (Fig. 10B), so that the correlation of the run time and the number of data sets is linear for the generation of all possible combinations as well as the compilation of bootstrap summaries (data not shown). However, for the calculation of the PABA results itself, the correlation between run time and number of data sets is still exponential. This

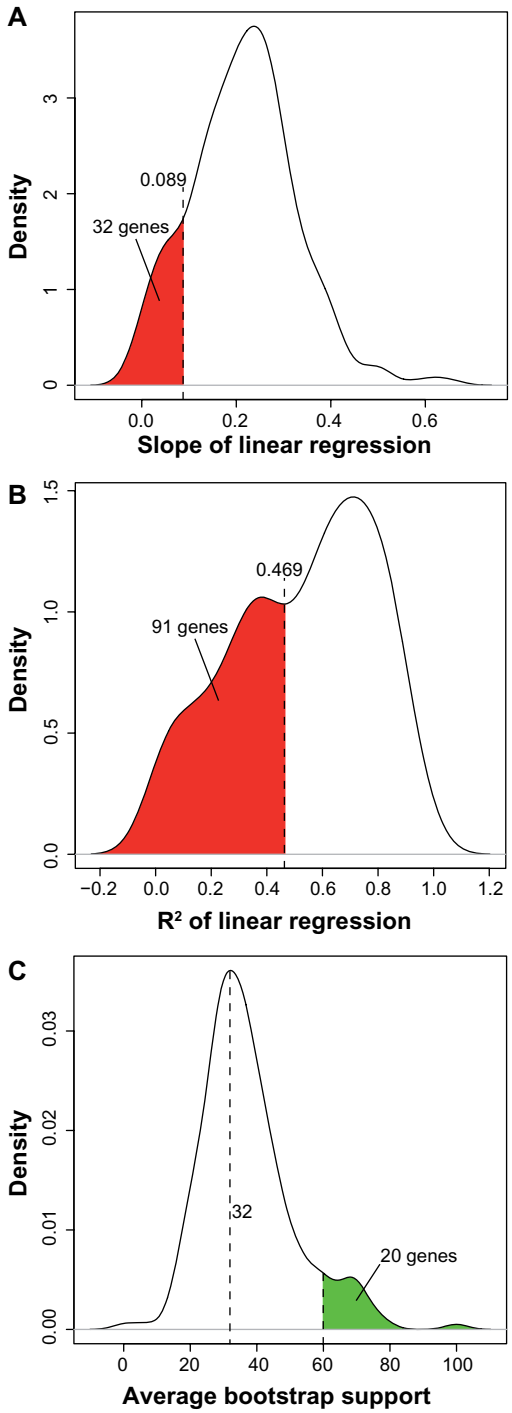


**Figure 9.** Density plots generated with *R* of different gene-specific saturation indices (**A** and **B**) as well as phylogenetic signal (**C**) for the 229 genes present in the data set of Struck.[11]

**Notes:** (**A**) Slope of the linear regression between patristic and uncorrected pairwise distances; (**B**) $R^2$ of the linear regression between patristic and uncorrected pairwise distances; and (**C**) average bootstrap support across all nodes of the best ML tree of each gene. Red and green areas indicate obvious deviations from the normal distribution.

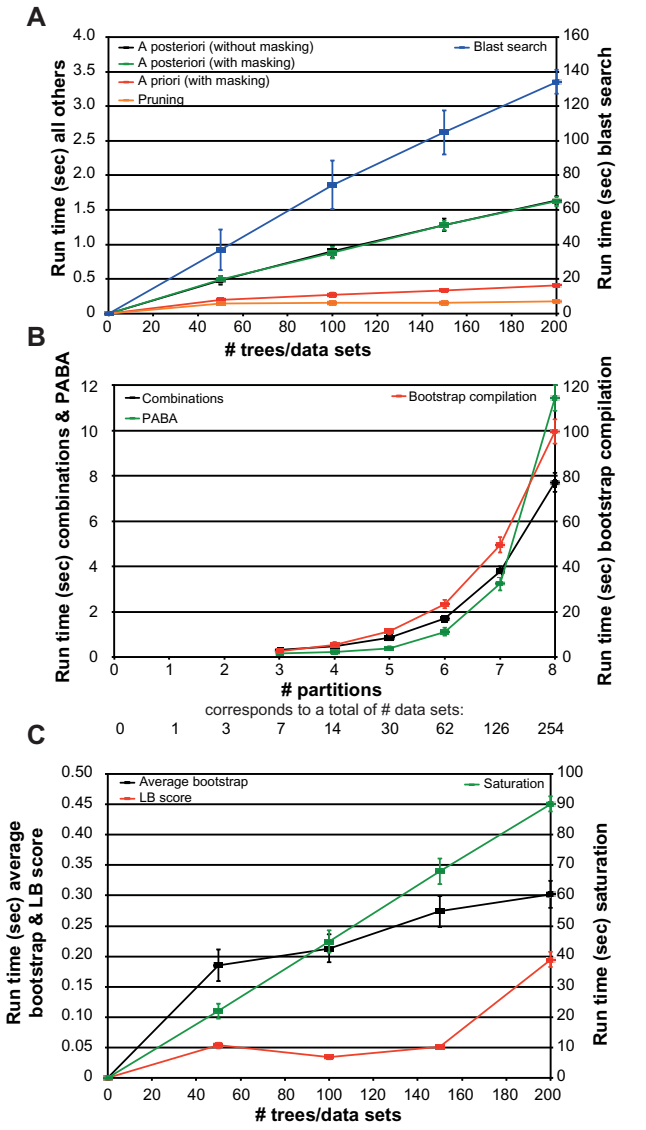


**Figure 10.** Run-time statistics of different options of TreSpEx as part of (**A**) the paralogy screening, (**B**) PABA analyses, and (**C**) determination of long-branch indices, saturation index, or phylogenetic signal as assessed by average bootstrap support.

**Notes:** Standard deviation in run time is indicated by error bars. The legends in each subfigure indicate the options and were placed in close vicinity to the corresponding *y*-axis.

difference can also be observed in the plot against the number of partitions. While the generation of all possible combinations as well as the compilation of bootstrap summaries show similar curves, the curve for the calculation of the PABA results is much steeper (Fig. 10B). The difference is because of the fact that with an increasing number of data sets, the number of possible additions of a partition to data sets without the partition also increases exponentially. However, even given this exponential growth the calculation of the PABA results for eight partitions takes less than 12 seconds, only about 50% longer than the generation of all possible combinations of eight partitions in the first step (Fig. 10B). For eight partitions, the longest time in performing the steps of the PABA analysis in TreSpEx is used for the summary and compilation of the bootstrap results of all data sets. This step requires a little less than two minutes. However, given the double exponential correlation of the calculations of the PABA results to

the number of partitions at a certain number of partitions, this calculation will take the longest. For example, with only up to seven partitions the generation of the possible data sets takes longer than the calculation of the PABA results, but after that it is vice versa (Fig. 10B). Regarding run-time requirements, though, the bottleneck in the PABA analysis will be none of the three steps, but the actual phylogenetic reconstruction including, for example, a bootstrap analysis.[56] For instance, even performing a parallel RAxML analysis on 15 threads with 100 bootstrap replicates, the shortest phylogenetic reconstruction took about 0.5 minutes, so that with 8 partitions and 254 data sets this would be at best about 2 hours. The three steps of the PABA analysis performed by TreSpEx together took less than two minutes for eight partitions and, thus, only about one-sixtieth of the time of the phylogenetic reconstructions. In case of a permutation test, the time for the generation of all possible combinations, the phylogenetic reconstructions, as well as the compilation of bootstrap summaries multiplies by the number of the permutated data sets plus one as the three steps have to be conducted for each permutated data set and the original data set. However, the calculation of the PABA results increases only slightly.

## Conclusion

TreSpEx allows the detection of artificial signal because of paralogy, long-branch attraction, or saturation, as well as conflict between different data sets, by utilizing tree-based information like nodal support or PDs. TreSpEx enables the parallel analysis of hundreds of trees and/or predefined gene partitions in very short to reasonable amounts of time. For example, the analysis of the sister group relationship of Ctenophora to all other Metazoa[81] using TreSpEx indicated that the support for this relationship might stem from long-branch attraction of Ctenophora toward the outgroup taxa in the analysis. Hence, more thorough analyses in how far this affects the position of Ctenophora are still necessary and to this end, taxon sampling of Ctenophora should also be substantially increased in future phylogenomic studies. Moreover, after increasing the number of taxa the analyses should be complemented by thorough investigations of the individual genes of the data set with respect to biases such as saturation and heterogeneous substitution rates. TreSpEx could be a useful tool in such analyses.

Generally, the results of TreSpEx provide the foundation and raw data for further analyses of different properties of the data set and the influence of these properties on the phylogenetic reconstructions. The partitions of a data set or different data sets can be ranked or grouped together. Additionally, taxa can be excluded based on the results of TreSpEx. The influence of individual properties like long-branch or saturation indices on the phylogenetic reconstruction can be assessed in combination with additional phylogenetic analyses. Hence, regardless of whether the studies are based on single, a few, or hundreds of genes the reliability of phylogenetic reconstructions can be increased using TreSpEx. This will improve the robustness of phylogenies and therefore also the conclusions drawn in many areas of comparative biological studies that rely on robust phylogenies.

TreSpEx will be kept up to date in the next years if changes in the Perl environment occur, and new tree-based methods will be incorporated. Moreover, on request different input and output formats can be added. The program is open source and released under the terms of GNU General Public License (GPL) 3.0.

## Acknowledgements

I would like to thank Andreas Müller (University of Osnabrück) and Patrick Kück (Zoological Research Museum Alexander Koenig) for advice in Perl programming. I also acknowledge the input of the editor and two anonymous reviewers to the manuscript.

## Author Contributions

Conceived and designed the experiments: THS. Analyzed the data: THS. Wrote the first draft of the manuscript: THS. Made critical revisions: THS. The author reviewed and approved of the final manuscript.

**DISCLOSURES AND ETHICS**

As a requirement of publication the author has provided signed confirmation of compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

## REFERENCES

1. Gee H. Ending incongruence. *Nature*. 2003;425:782.
2. Jeffroy O, Brinkmann H, Delsuc F, Philippe H. Phylogenomics: the beginning of incongruence? *Trends Genet*. 2006;22:225–31.
3. Enright AJ, Dongen SV, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30:1575–84.
4. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13:2178–89.
5. Ebersberger I, Strauss S, von Haeseler A. HaMStR: profile hidden Markov model based search for orthologs in ESTs. *BMC Evol Biol*. 2009;9:157.
6. Kim K, Kim W, Kim S. ReMark: an automatic program for clustering orthologs flexibly combining a recursive and a Markov clustering algorithms. *Bioinformatics*. 2011;27:1731–3.
7. Shi G, Peng M-C, Jiang T. MultiMSOAR 2.0: an accurate tool to identify ortholog groups among multiple genomes. *PLoS One*. 2011;6:e20892.
8. Rodríguez-Ezpeleta N, Brinkmann H, Burger G, *et al*. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Curr Biol*. 2007;17:1420–5.
9. Philippe H, Derelle R, Lopez P, et al. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol*. 2009;19:706–12.
10. Philippe H, Brinkmann H, Lavrov DV, et al. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 2011;9:e1000602.
11. Struck TH. The impact of paralogy on phylogenomic studies—a case study on annelid relationships. *PLoS One*. 2013;8:e62892.
12. Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*. 2006;440:341–5.
13. Scannell DR, Frank AC, Conant GC, et al. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A*. 2007;104:8397–402.
14. Sémon M, Wolfe KH. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet*. 2007;23:108–12.

15. Schierwater B, Eitel M, Jakob W, et al. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol*. 2009;7:e1000020.
16. Dunn CW, Hejnol A, Matus DQ, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 2008;452:745–50.
17. Kuhner MK, Felsenstein J. A simulation comparison phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*. 1994;11:459–68.
18. Lake JA. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proc Natl Acad Sci U S A*. 1994;91:1455–9.
19. Lockhart PJ, Steel MA, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*. 1994;11:605–12.
20. Simon C, Frati F, Beckenbach AT, et al. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Ann Entomol Soc Am*. 1994;87:651–701.
21. Milinkovitch MC, LeDuc RG, Adachi J, et al. Effects of character weighting and species sampling on phylogeny reconstruction: a case study based on DNA sequence data in cetaceans. *Genetics*. 1996;144:1817–33.
22. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, eds. *Molecular Systematics*. 2nd ed. Sunderland, MA: Sinauer Associates, Inc.; 1996:407–514.
23. Halanych KM, Robinson TJ. Multiple substitutions affect the phylogenetic utility of cytochrome b and 12S rDNA data: examining a rapid radiation in leporid (Lagomorpha) evolution. *J Mol Evol*. 1999;48:369–79.
24. Lopez P, Forterre P, Philippe H. The root of the tree of life in the light of the covarion model. *J Mol Evol*. 1999;49:496–508.
25. Philippe H, Forterre P. The rooting of the universal tree of life is not reliable. *J Mol Evol*. 1999;49:509–23.
26. Nickrent DL, Parkinson CL, Palmer JD, Duff RJ. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol Biol Evol*. 2000;17:1885–95.
27. Struck TH, Hessling R, Purschke G. The phylogenetic position of the Aeolosomatidae and Parergodrilidae, two enigmatic oligochaete-like taxa of the 'Polychaeta', based on molecular data from 18SrDNA sequences. *J Zool Syst Evol Res*. 2002;40:155–63.
28. Xia X, Xie Z, Salemi M, Chen L, Wang Y. An index of substitution saturation and its application. *Mol Phylogenet Evol*. 2003;26:1–7.
29. Jördens J, Struck TH, Purschke G. Phylogenetic inference regarding parergodrilidae and *Hrabeiella periglandulata* ("Polychaeta", Annelida) based on 18S rDNA, 28S rDNA and COI sequences. *J Zool Syst Evol Res*. 2004;42:270–80.
30. Bergsten J. A review of long-branch attraction. *Cladistics*. 2005;21(2):163–93.
31. Struck TH, Nesnidal MP, Purschke G, Halanych KM. Detecting possibly saturated positions in 18S and 28S sequences and their influence on phylogenetic reconstruction of Annelida (Lophotrochozoa). *Mol Phylogenet Evol*. 2008;48:628–45.
32. Kück P, Mayer C, Wägele J-W, Misof B. Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS One*. 2012;7:e36593.
33. Nosenko T, Schreiber F, Adamska M, et al. Deep metazoan phylogeny: when different genes tell different stories. *Mol Phylogenet Evol*. 2013;67:223–33.
34. Rodriguez-Ezpeleta N, Brinkmann H, Roure B, *et al*. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol*. 2007;56:389–99.
35. Campbell LI, Rota-Stabelli O, Edgecombe GD, *et al*. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc Natl Acad Sci U S A*. 2011;108:15920–4.
36. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*. 2013;497:327–31.
37. Barrett M, Donoghue MJ, Sober E. Against consensus. *Syst Zool*. 1991;40:486–93.
38. Davis J. Character removal as a means for assessing stability of clades. *Cladistics*. 1993;9:201–10.
39. Chippindale PT, Wiens JJ. Weighting, partitioning, and combining characters in phylogenetic analysis. *Syst Biol*. 1994;43:278–87.
40. Olmstead RG, Sweere JA. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. *Syst Biol*. 1994;43:467–81.
41. Farris JS, Kallersjo M, Kluge AG, Bult C. Constructing a significance test for incongruence. *Syst Biol*. 1995;44:570–2.
42. Huelsenbeck JP, Bull JJ, Cunningham CW. Combining data in phylogenetic analysis. *Trends Ecol Evol*. 1996;11:152–8.
43. Mason-Gamer RJ, Kellogg E. Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). *Syst Biol*. 1996;45:524–45.
44. Baker RH, DeSalle R. Multiple sources of character information and the phylogeny of Hawaiian *Drosophila*. *Syst Biol*. 1997;46:654–73.
45. Cunningham CW. Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. *Syst Biol*. 1997;46:464–78.
46. Halanych KM. Considerations for reconstructing metazoan history: signal, resolution, and hypothesis testing. *Am Zool*. 1998;38:929–41.
47. Gatesy J, O'Grady P, Baker RH. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics*. 1999;15:271–313.
48. Reed R, Sperling F. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus *Papilio*. *Mol Biol Evol*. 1999;16:286–97.
49. Thornton JW, DeSalle R. A new method to localize and test the significance of incongruence: detecting domain shuffling in the nuclear receptor superfamily. *Syst Biol*. 2000;49:183–201.
50. O'Grady PM, Remsen J, Gatesy JE. Partitioning of multiple data sets in phylogenetic analysis. In: DeSalle R, Giribet G, Wheeler WC, eds. *Methods and Tools in Biosciences and Medicine*: *Techniques in Molecular Evolution and Systematics*. Berlin: Birkhauser Verlag; 2002:176–248.
51. Nygren A, Sundberg P. Phylogeny and evolution of reproductive modes in Autolytinae (Syllidae, Annelida). *Mol Phylogenet Evol*. 2003;29:235–49.
52. Hipp A, Hall J, Sytsma K. Congruence versus phylogenetic accuracy: revisiting the Incongruence Length Difference test. *Syst Biol*. 2004;53:81–9.
53. Passamaneck YJ, Schander C, Halanych KM. Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. *Mol Phylogenet Evol*. 2004;32:25–38.
54. Bond JE, Hedin M. A total evidence assessment of the phylogeny of the diverse North American trapdoor spider subfamily Euctenizinae (Araneae, Mygalomorphae, Cyrtaucheniidae). *Mol Phylogenet Evol*. 2006;41:70–85.
55. Struck TH, Purschke G, Halanych KM. Phylogeny of Eunicida (Annelida) and exploring data congruence using a partition addition bootstrap alteration (PABA) approach. *Syst Biol*. 2006;55:1–20.
56. Struck TH. Data congruence, paedomorphosis and salamanders. *Front Zool*. 2007;4:22.
57. Struck TH. Direction of evolution within Annelida and the definition of Pleistoannelida. *J Zool Syst Evol Res*. 2011;49:340–5.
58. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*. 1999;16:1114–6.
59. Lyons-Weiler J, Hoelzer GA. Null model selection, composition bias, character state bias, and the limits of phylogenetic information. *Mol Biol Evol*. 1999;16:1400–5.
60. Hall JS, Adams B, Parsons TJ, et al. Molecular cloning, sequencing, and phylogenetic relationships of a new potyvirus: sugarcane streak mosaic virus, and a reevaluation of the classification of the Potyviridae. *Mol Phylogenet Evol*. 1998;10:323–32.
61. Holmdahl OJ, Morrison DA, Ellis JT, Huong LT. Evolution of ruminant *Sarcocystis* (Sporozoa) parasites based on small subunit rDNA sequences. *Mol Phylogenet Evol*. 1999;11:27–37.
62. Yi Z, Song W. Evolution of the order Urostylida (Protozoa, Ciliophora): new hypotheses based on multi-gene information and identification of localized incongruence. *PLoS One*. 2011;6:e17471.
63. Naum M, Brown EW, Mason-Gamer RJ. Is a robust phylogeny of the enterobacterial plant pathogens attainable? *Cladistics*. 2011;27:80–93.
64. Smith SA, Dunn CW. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics*. 2008;24:715–6.
65. Kück P, Struck TH. BaCoCa—a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Mol Phylogenet Evol*. 2014;70:94–8.
66. Kocot KM, Citarella MR, Moroz LL, Halanych KM. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol Bioinform Online*. 2013;9:429–35.
67. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22:2688–90.
68. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*. 2004;21:1095–109.
69. Templeton AR. Phylogenetic inference from restriction site endonuclease cleavage site maps with particular reference to the human and apes. *Evolution*. 1983;37:221–44.
70. Macey JR, Schulte JA, Larson A, et al. Molecular phylogenetics, tRNA evolution, and historical biogeography in anguid lizards and related taxonomic families. *Mol Phylogenet Evol*. 1999;12:250–72.
71. Whitlock BA, Baum DA. Phylogeny of cacao (*Theobroma cacao* L., Sterculiaceae) and its wild relatives based on sequences of the nuclear-encoded gene *VICILIN*. *Syst Bot*. 1999;24:128–38.
72. Lee MS. Tree robustness and clade significance. *Syst Biol*. 2000;49:829–36.
73. Lee MS, Hugall AF. Partitioned likelihood support and the evaluation of data set conflict. *Syst Biol*. 2003;52:15–22.
74. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. 1985;39:783–91.

75. Brinkman H, Philippe H. Animal phylogeny and large-scale sequencing: progress and pitfalls. *J Syst Evol*. 2008;46:274–86.
76. Weigert A, Helm C, Meyer M, *et al*. Illuminating the base of the annelid tree using transcriptomics. *Mol Biol Evol*. 2014. Doi: 10.1093/molbev/msu080.
77. Bleidorn C, Eeckhaut I, Podsiadlowski L, *et al*. Mitochondrial genome and nuclear sequence data support Myzostomida as part of the annelid radiation. *Mol Biol Evol*. 2007;24:1690–701.
78. Bleidorn C, Podsiadlowski L, Zhong M, *et al*. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol Biol*. 2009;9:150.
79. Hartmann S, Helm C, Nickel B, *et al*. Exploiting gene families for phylogenomic analysis of myzostomid transcriptome data. *PLoS One*. 2012;7:e29843.
80. Struck TH, Paul C, Hill N, *et al*. Phylogenomic analyses unravel annelid evolution. *Nature*. 2011;471:95–8.
81. Ryan JF, Pang K, Schnitzler CE, *et al*. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science*. 2013;342(6164):1242592.
82. Dordel J, Fisse F, Purschke G, Struck TH. Phylogenetic position of Sipuncula derived from multi-gene and phylogenomic data and its implication for the evolution of segmentation. *J Zool Syst Evol Res*. 2010;48:197–207.
83. Struck TH. Phylogeny of Annelida. *Zool Online*. 2012:23. Available at http://www.degruyter.com/view/Zoology/bp_029147–6_1