Revising transcriptome assemblies with phylogenetic information - Supplementary Information

Contents

Supplementary methods

The code for the analyses presented here (including an executable version of this document) are available in a git repository at https://github.com/caseywdunn/ms_treeinform. The phylogenetic analyses considered here are based on a 7 taxon siphonophore (Dunn 2009) dataset. This dataset originated as the regression test dataset for the Agalma automated phylogenetics workflow (Dunn, Howison, and Zapata 2013) and was selected for its well-resolved species tree as well as the fact that treeinform is implemented in Agalma. The gene trees were built with Agalma1.1, and bash scripts for the run can be found at https://github.com/caseywdunn/ms_treeinform/code/revisions. Most analyses are loaded from saved Rdata objects to speed up reproduction of the notebook, with the code to generate the Rdata objects commented out. To reproduce the notebook in its entirety from the raw Agalma output will take around 5 hours. Including the posterior distribution from Gibbs sampling (see Selecting a threshold for transcript reassignment) rather than the parameter estimates will take around 30 hours.

The phylogenetic analyses in Agalma followed standard approaches with default settings. Speciation and duplication nodes were identified in the gene trees with phyldog (Boussau et al. 2013). Bash scripts and associated code for the runs can be found at in the same directory as above. Agalma uses the transcriptome assembler Trinity (Grabherr et al. 2011), version 2.5.1.

Assessing the extent of transcript assignment errors

We first examined the prevalence of transcript misassignment. For each node in each of the 5187 gene phylogenies, we calculated the length of the corresponding subtree. This is the sum of the length of all branches in the subtree defined by the node. An excess of very short subtrees would be a strong indication of assigning different transcripts of the same gene, which have very similar sequences and therefore short branches connecting them in phylogenetic trees, to different genes. This is the pattern we found (Supplementary Figure 1).

Two issues could create a misleading impression in the histogram of subtree lengths for internal nodes (Supplementary Figure 1). First, it considers all subtrees, including those defined by both speciation and duplication nodes. Misassigning transcripts from the same gene to multiple genes will artificially inflate only the number of duplication nodes, since variation across transcripts within a gene are essentially misassigned to gene duplication events. Examining just the duplication events in the gene trees therefore provides a more direct perspective on the problem we investigate here. Second, subtree lengths are in units of expected numbers of substitution, which depend on both rates of molecular evolution and time. Because the rates of evolution can vary within and between gene phylogenies, variation in rates could confound the interpretation of gene tree sublength.

We therefore performed a calibrated analysis and focused only on duplication nodes. We first created a time calibrated species tree, with all tips with age 0 and the root node with age 1. We then transformed the branch lengths of the gene trees so that each speciation node in each gene tree had the same age as the corresponding node in the species tree (see source code for this document). A histogram of the calibrated duplication times (Supplementary Figure 2) indicates there is a large excess of recent duplications. This provides additional evidence for the frequent misassignment of transcripts from the same gene to artefactual recent gene duplicates.

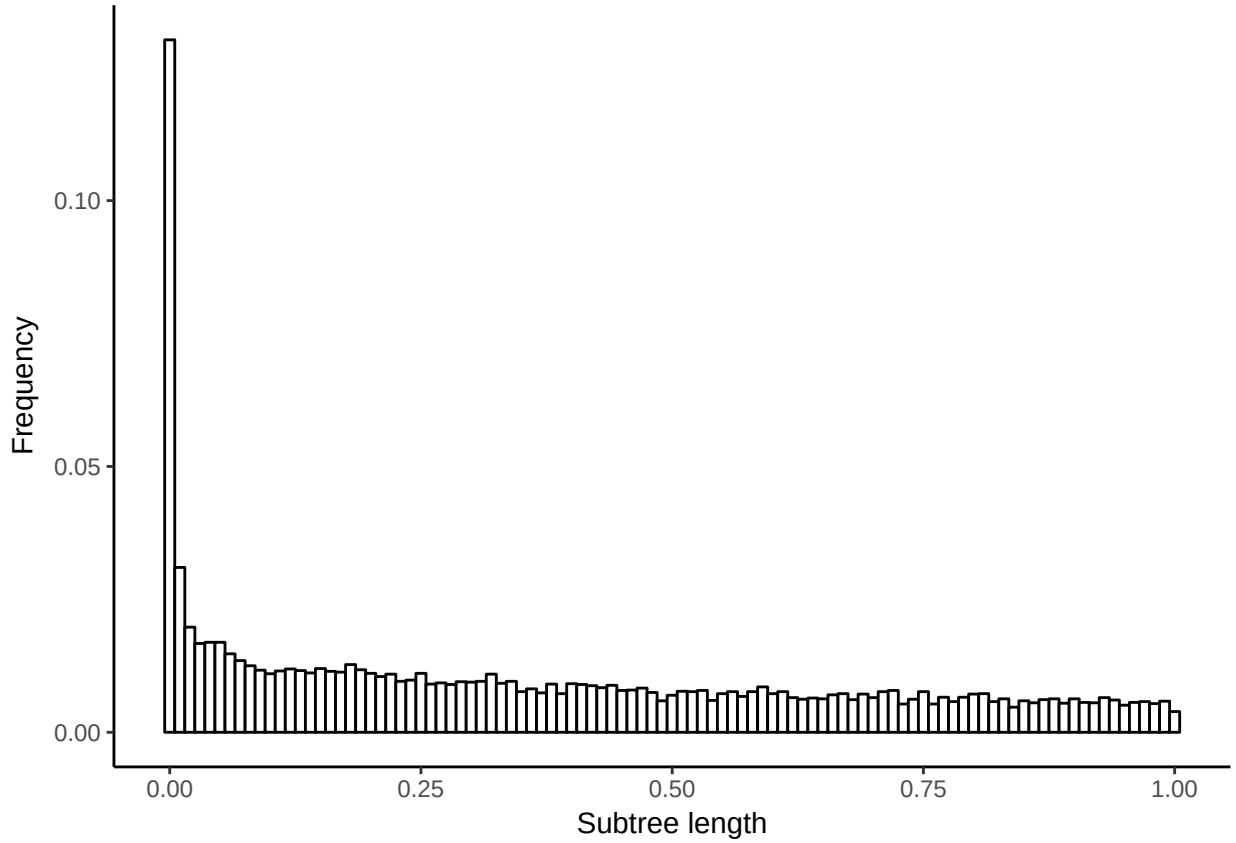

Figure 1: Histogram of subtree lengths for internal nodes in each Siphonophora subset gene tree from Agalma containing tip descendants from the same species. Subtree lengths greater than 1 were filtered out for clarity.

Comparison to other transcript clustering methods

In order to get a sense of whether the transcript misassignment errors were localized to Trinity or are a more general problem to transcriptome assembly, we compared Trinity transcript clustering results with another transcript clustering tool, Corset (Davidson and Oshlack 2014) for the 5 species that had to be assembled. For 3 of the samples (SRX288285, SRX288430, SRX288431) we also ran cd-hit (Fu et al. 2012) to remove transcripts with 100% identity in order to address some speed issues in Corset. The distribution of cluster sizes (Supplementary Figure 2) suggests that Corset tends to overcluster compared to Trinity, which would lead to similar misassignment errors.

Additionally, we computed the Adjusted Rand Index (Rand 1971) to get a sense of the similarity between the Trinity and Corset clusterings. (Supplementary Table 1) The Adjusted Rand Index computes the proportion of pairs that either both belong to the same cluster, or both belong to different clusters, corrected for chance. The ARI ranges from 0 to 1, with 0 meaning the clusterings are maximally dissimilar and 1 meaning that the clusterings are exactly the same. Though we have no ground truth here, the ARI suggests that Trinity

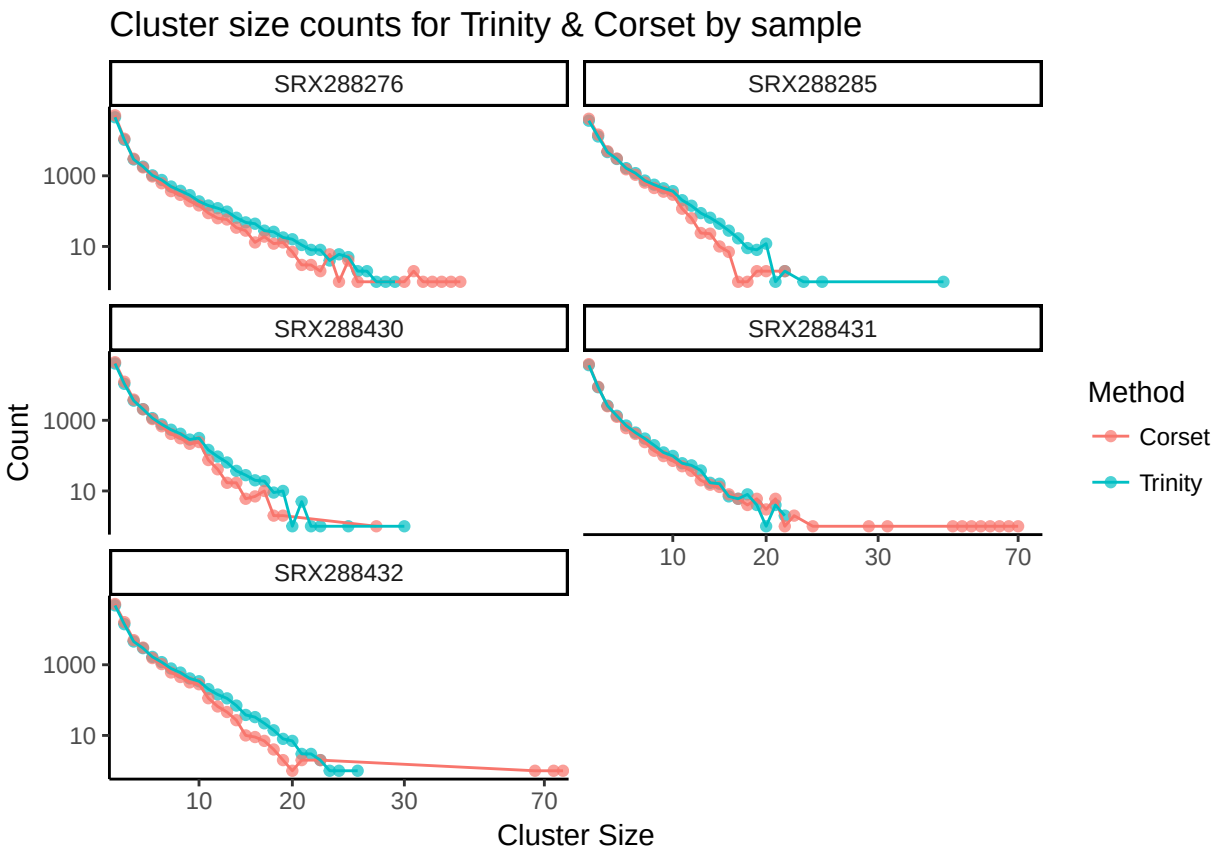

Figure 2: Cluster size counts for Trinity assembly and Corset clustering algorithm on Trinity contigs. There are 3 Trinity clusters with size greater than 30, while there are 20 Corset clusters with size greater than 30. This suggests that the same misassignment errors are generated by other transcriptome assemblers and clustering algorithms as well.

and Corset clusterings are more similar than dissimilar, and in the case of SRX288285, SRX288430, and SRX288431 are very similar.

Table 1: Adjusted Rand Index between Trinity and Corset clusterings by sample.

Sample	Adjusted.Rand.Index
SRX288276	0.5805402
SRX288285	0.8353263
SRX288430	0.8094703
SRX288431	0.8089121
SRX288432	0.7835530

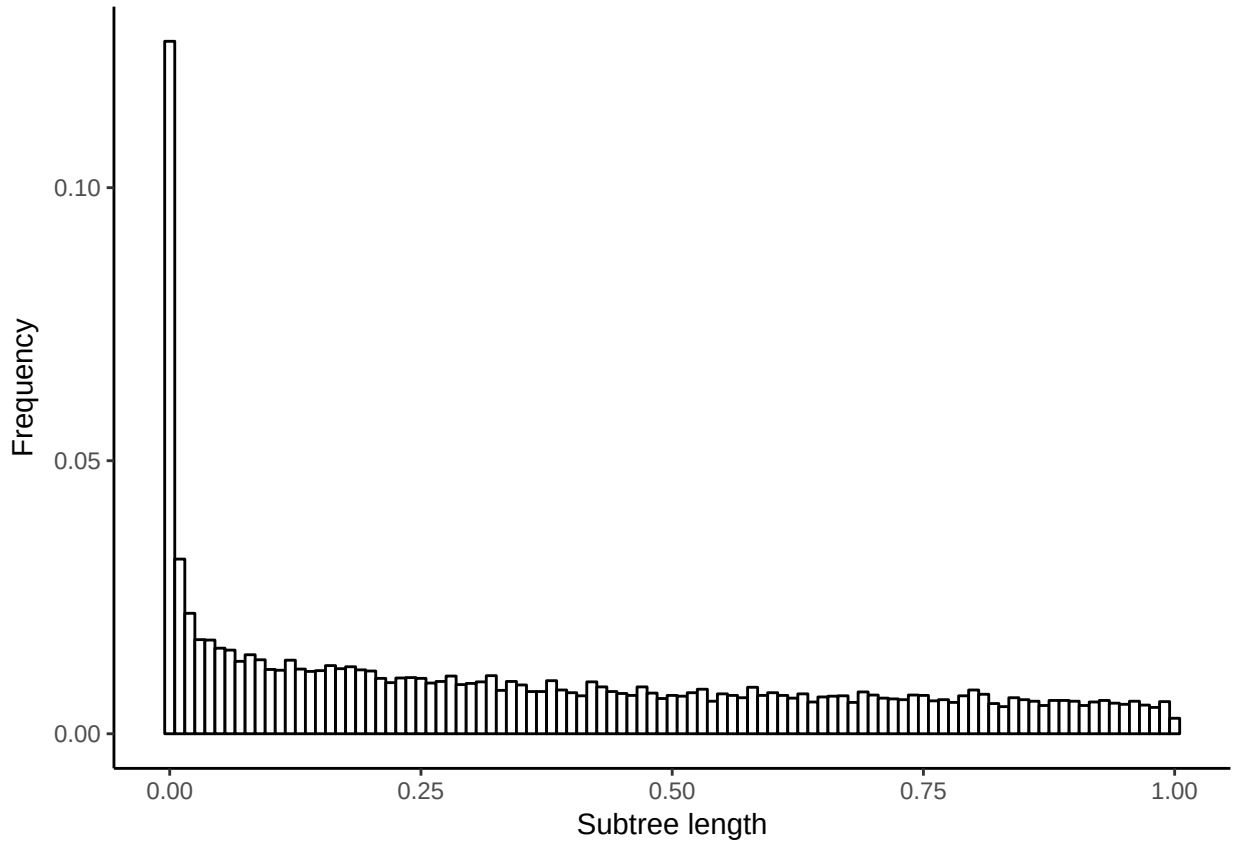

Figure 3: Histogram of subtree lengths for internal nodes in each Siphonophora subset gene tree from Agalma with Corset clusterings containing tip descendants from the same species. Subtree lengths greater than 1 were filtered out for clarity.

Supplementary Figure 3 shows the histogram of subtree lengths for internal nodes in each Siphonophora subset gene tree with Corset clusterings. It shares the same characteristics as in Supplementary Figure 1, and the two-sample Kolmogorov-Smirnov test returns a test statistic of D=0.0104 which is less than the rejection threshold at level 0.05 for the samples, 0.0125. This indicates that regardless of which transcript clustering method is chosen, the same transcript misassignment errors persist. Given the intrinsic challenges of assigning assembled transcripts to genes it is likely that the same misassignment errors are generated by other transcriptome assemblers as well.

Selecting a threshold for transcript reassignment

A visual inspection of the histogram of subtree lengths (Supplementary Figure 1) suggested that the frequency of subtree length for internal nodes was high below 0.01 but leveled out above it. Numerically, 8.413% of internal nodes containing tip descendants from the same species had a subtree length less than 0.01, with an additional 1.389% having a subtree length between 0.01 and 0.02. It is unlikely that all of these clades are gene duplication events.

These observations suggest that two different processes are operating simultaneously to generate the observed subtree lengths, one for the misassigned transcripts and one for the correctly assigned transcripts. To model this pattern, we applied a mixture model to the inferred duplication times (equivalent to branch lengths; see main text) from the gene trees. One component modelled duplication events and associated times arising from transcripts assigned to different genes that belong to the same gene (i.e., misassigned transcripts), and the other component modelled duplication events and associated times arising from transcripts assigned to different genes that in fact do belong to different genes (i.e., corrrectly assigned transcripts) (Supplementary Figure 4).

We expected the implied duplication events of transcripts of the same gene that were misassigned to different genes to have very short implied duplication times approaching 0, and thus chose to model that component (Component 1) as a gamma distribution with parameters shape= α and rate= β. To model duplication events and associated times arising from the correctly assigned transcripts (Component 2), we used a birth-death process (Gernhard 2008), which is well studied and often applied to gene analyses of duplication and loss. The probability distribution function in the model we used has parameters birth rate λ, death rate μ, and tree time of origin t_{or}. Because we fitted a chronogram with time of origin 1 onto the gene trees $G = \{G_1, G_2, \ldots, G_K\}$, we made the assumption that all gene tree times of origin are $t_{or} = 1$. Some gene trees have duplication events predating the first speciation event, thus when we fitted chronograms onto those gene trees they had times of origin greater than 1. We chose to filter these gene trees out of the mixture model and subsequent analyses.

Let $x_{i,k}$ represent duplication time i from gene tree G_k, with $z_i \in \{1, 2\}$ representing whether $x_{i,k}$ is drawn from the 1st component ($z_i = 1$) or the 2nd component ($z_i = 2$). Then if π_1 and π_2 denote the overall probability that a duplication time belongs to the 1st and 2nd component respectively, $\Gamma(x_{i,k}|\alpha, \beta)$ is the probability density function for the gamma distribution, and $f(x_{i,k}|t_{or,k} = 1, \lambda, \mu)$, we get the expression

$$P(x_{i,k}) = \pi_1 \Gamma(x_{i,k}|\alpha, \beta) + \pi_2 f(x_{i,k}|t_{or,k} = t, \lambda, \mu)$$

We used Just Another Gibbs Sampler (JAGS) (Plummer 2003) to perform Bayesian Gibbs Sampling in order to infer the parameters α, β, λ, and μ as well as the mixing proportions π_1 and π_2. This gave us the parameter estimates in Table 2.

Table 2: Summary of parameter estimates from JAGS.

	Lower95	Mean	Upper95	MCerr
α	0.2418870	0.2548908	0.2628810	0.0010395
β	1.7442500	1.9488374	2.1686600	0.0147482
μ	0.0000009	0.0119074	0.0356818	0.0001564
λ	2.7344800	2.8621597	2.9990300	0.0008479
π_1	0.3167680	0.3386287	0.3604700	0.0001405
π_2	0.6395300	0.6613713	0.6832320	0.0001405

The posterior probability that a duplication time x is drawn from component 1 or component 2, i.e. $P(z|x)$ gives us a way to determine the probability of error. It can be inferred from Gibbs sampling as well, although it can also be estimated from the parameters of the mixture model. If we decide x is drawn from the 2nd component, then $P(z = 1|x)$ will be the error probability, and if we decide x is drawn from the 1st

component, then $P(z = 2|x)$ is the error probability. If we care more about having fewer correctly assigned transcripts being erroneously flagged as misassigned, then we can use the posterior probability to select an appropriate threshold for treeinform by selecting T such that $P(z = 2|x) < \alpha$ for all $x < T$, where α is the error rate. Here we decided to use $\alpha = 0.05$. In Bayesian decision theory this is equivalent to a loss matrix of $\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 19 & 0 \end{bmatrix}$, where each entry λ_{mn} is the penalty for selecting component n when $x_{i,k}$ is actually drawn from component m. This gives us the threshold 0.0003255.

From that threshold, we can back-calibrate to determine a subtree branch length threshold for use in treeinform. Specifically, we took all duplication events with times below the intersection point on all chronogram-fitted gene trees, mapped them to the equivalent events on the phyldog-outputted gene trees, computed the subtree length of all events, and then took the maximum of those subtree lengths. From the intersection point 0.0003255, this gave us a threshold of 0.000562, which we approximate with 0.0005.

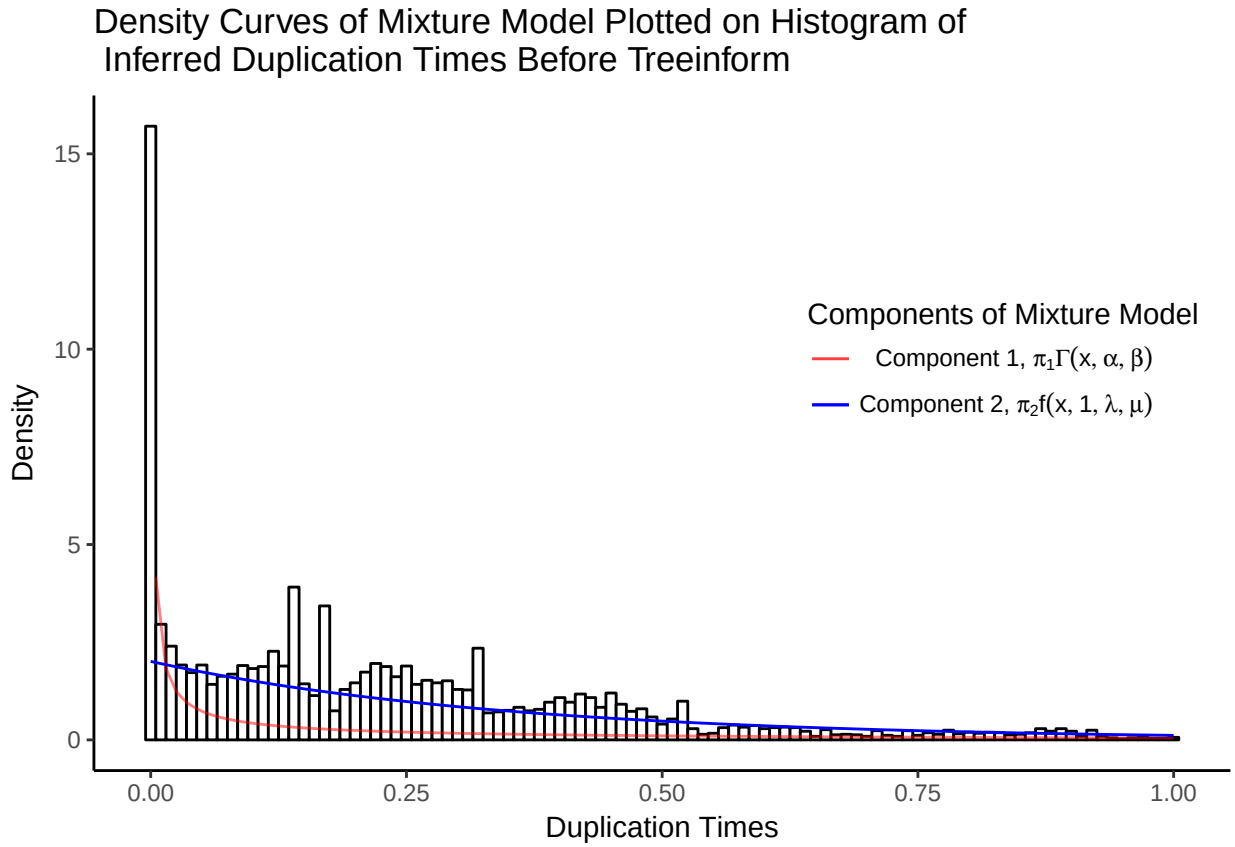

Figure 4: Histogram of the inferred duplication times with an overlaid mixture model. Component 1 of the mixture model captures the technical issues we address where transcripts from the same gene are assigned to different genes, and component 2 captures the biological pattern, i.e. transcripts from different genes correctly assigned so. We first ran phyldog (Boussau et al. 2013) on the Siphonophora subset multiple sequence alignments from Agalma and a user-inputted species tree. This provided gene trees with internal nodes annotated as duplication or speciation events. We then fitted chronograms onto these gene trees with our user-inputted species tree.

Validating the effectiveness of treeinform

In order to validate that treeinform improves the accuracy of assigning transcripts to genes under the specified threshold, we performed two analyses. First, we plotted the percentage of reassigned genes at different

thresholds to assess the performance of the default threshold value of 0.0005 (Supplementary Figure 5). Below the default value, the percentage of reassigned genes begins to plateau, while above the default value the percentage of reassigned genes increases very quickly, increasing the likelihood of treeinform to reassign transcripts from different genes to the same gene in addition to reassign transcripts from the same gene together.

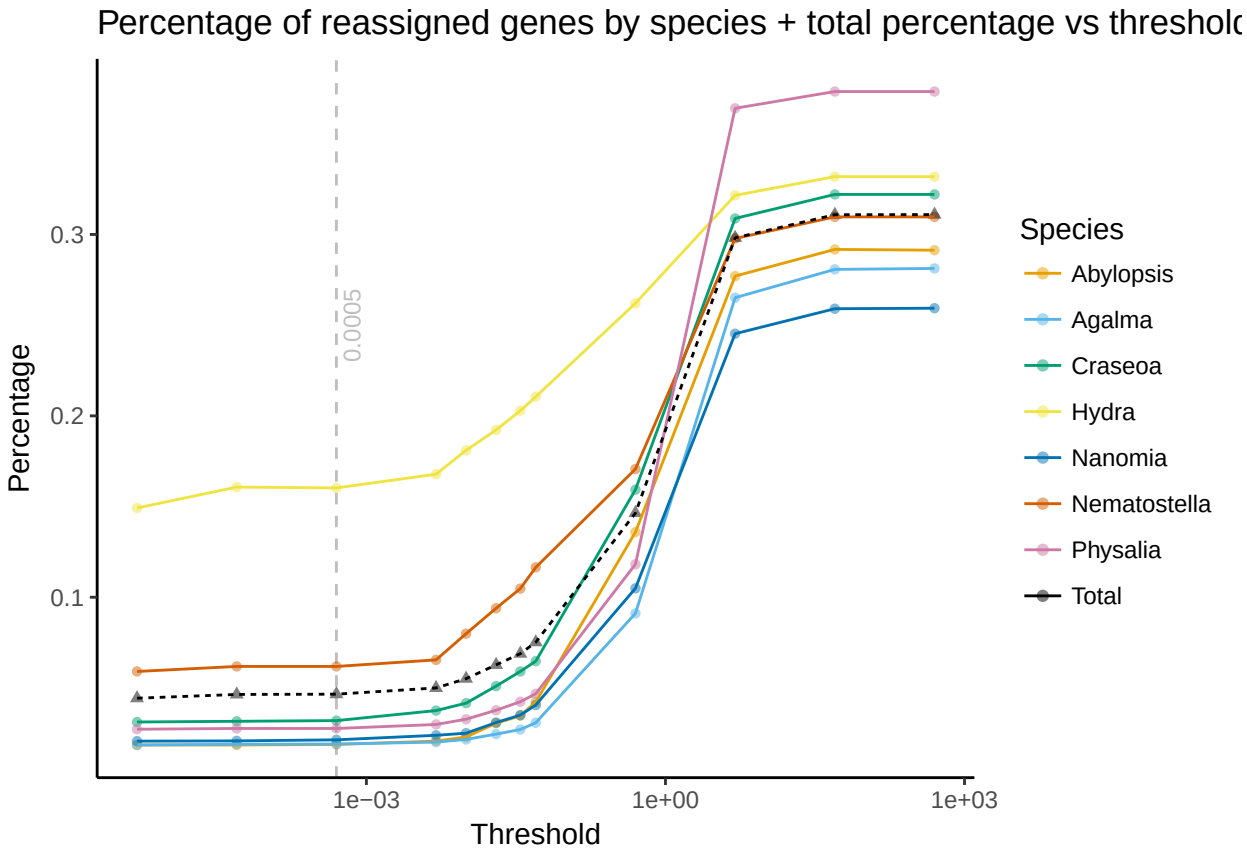

Figure 5: The percentage of reassigned tips is plotted above on a log scale. The original assembly had 433,071 genes, with 47,688 included in the gene trees after Agalma filtering criteria, and at most 23,396 possible candidates (49.06% of genes) for reassignment. The default threshold for treeinform is marked by the grey vertical dashed line.

We also looked at the percentage of reassigned genes for each species in order to get a sense of how variable by species transcript misassignment was. The percentage of reassigned genes for each species was rather variable, with Hydra magnipapillata having a much higher proportion of reassigned genes (16.03%) at the threshold. This affected the total proportion of reassigned genes as well, with the majority (46-47%) of reassigned genes at and around the treeinform threshold coming from Hydra magnapapillata. However, even for the other species, 1.88-6.18% of genes were reassigned at the threshold.

Second, we compared the density of duplication times under the model provided for Component 2 of the mixture model to the distribution of estimated duplication times for gene trees from Agalma before treeinform, and gene trees from Agalma after treeinform under 3 different thresholds: 0.05, 0.0005, and 0.07 (Supplementary Figure 6). We again fitted chronograms with the same Siphonophora species tree onto all gene trees from Agalma and filtered out those gene trees with time of origin greater than 1, so that duplication times were comparable between trees. Visually, the analyses with the 0.0005 threshold comes closest to the theoretical.

Additionally, we computed the Kullback-Leibler distance (Kullback and Leibler 1951) between the distributions of duplication times under different thresholds and the theoretical distribution of duplication times (Table

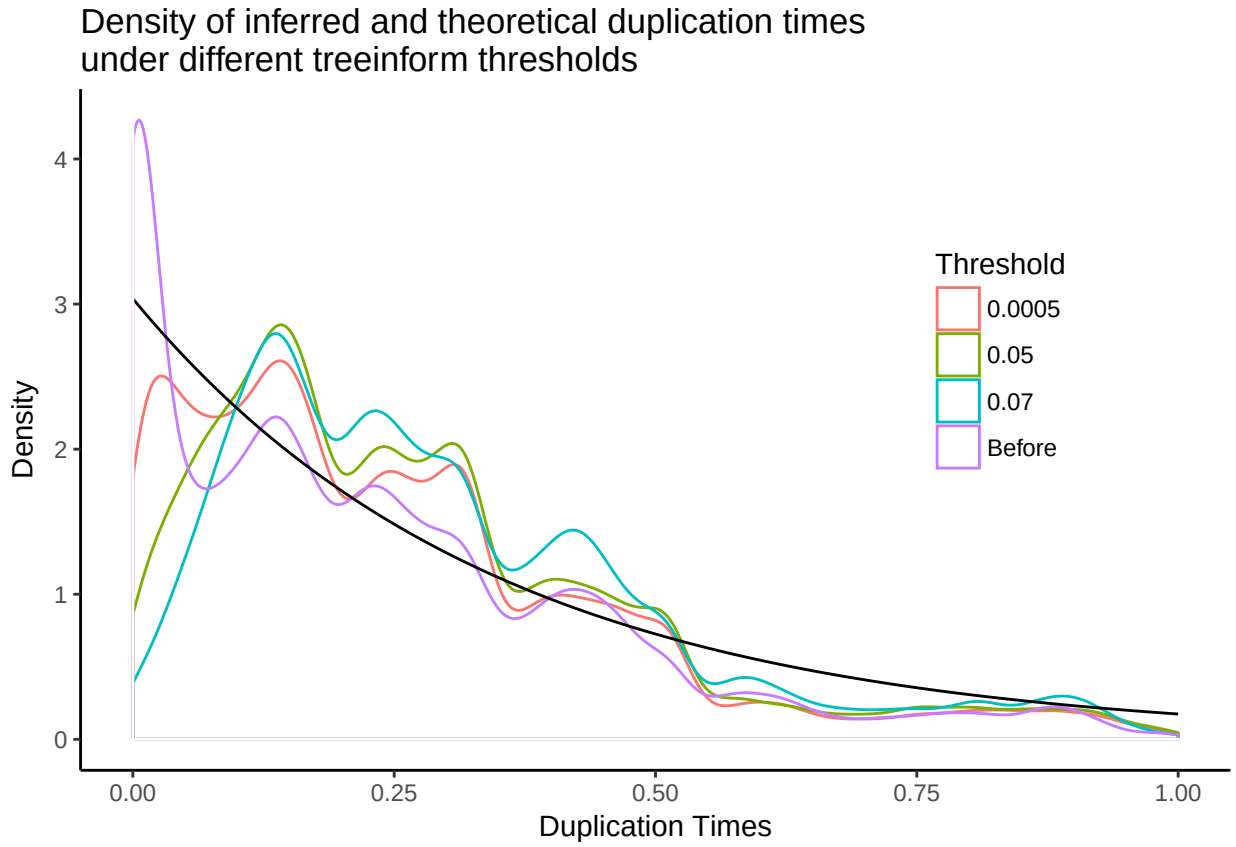

Figure 6: Density from theoretical and the empirical density under the 3 different thresholds as well as before treeinform was run. The distribution before treeinform has a large peak on the left that is removed by treeinform with all examined thresholds.

3). Kullback-Leibler distance, otherwise known as relative entropy, measures the distance between two distributions. The KL distance between the distribution of duplication times after running treeinform with the default threshold value of 0.0005 is not minimal, but is robust as compared to both threshold levels below and above the default value. This indicates that treeinform produces more accurate gene trees with appropriate threshold selection.

Table 3: Kullback-Leibler distances between duplication times after running treeinform with different thresholds and theoretical duplication times.

	KL.Distance
Before	0.2543703
0.07	0.1395099
0.05	0.1101704
0.005	0.0946684
0.0005	0.0997587
5e-05	0.1013232

Software versions

This manuscript was computed on Mon Apr 30 19:55:18 2018 with the following R package versions.

```
R version 3.4.0 beta (2017-04-08 r72499)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS  10.13.4

Matrix products: default
BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] bindrcpp_0.2     entropy_1.2.1    runjags_2.0.4-2  knitr_1.17
 [5] treeio_1.0.2     hutan_0.0.0.9000 dplyr_0.7.4      reshape2_1.4.2
 [9] ggplot2_2.2.1    scales_0.5.0     ape_4.1

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.13     highr_0.6        compiler_3.4.0   plyr_1.8.4
 [5] bindr_0.1        tools_3.4.0      digest_0.6.12    jsonlite_1.5
 [9] evaluate_0.10.1  tibble_1.3.4     gtable_0.2.0     nlme_3.1-131
[13] lattice_0.20-35  pkgconfig_2.0.1  rlang_0.1.2      rvcheck_0.0.9
[17] yaml_2.1.14      coda_0.19-1      stringr_1.2.0    rprojroot_1.2
[21] grid_3.4.0       glue_1.1.1       R6_2.2.2         rmarkdown_1.6
[25] magrittr_1.5     backports_1.1.1  htmltools_0.3.6  assertthat_0.2.0
[29] colorspace_1.3-2 labeling_0.3     stringi_1.1.5    lazyeval_0.2.0
[33] munsell_0.4.3
```

References

Boussau, Bastien, Gergely J Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. 2013. "Genome-Scale Coestimation of Species and Gene Trees." *Genome Research* 23 (2). Cold Spring Harbor Lab: 323–30.

Davidson, N M, and A Oshlack. 2014. "Corset: enabling differential gene expression analysis for." *Genome Biol* 15 (7): 410. doi:10.1186/PREACCEPT-2088857056122054.

Dunn, Casey W. 2009. "Siphonophores." *Current Biology* 19 (6). Cell Press: R233–R234. doi:10.1016/J.CUB.2009.02.009.

Dunn, Casey W, Mark Howison, and Felipe Zapata. 2013. "Agalma: an automated phylogenomics workflow." *BMC Bioinformatics* 14: 330. https://doi.org/10.1186/1471-2105-14-330.

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. "CD-HIT: Accelerated for clustering the next-generation sequencing data." *Bioinformatics* 28 (23): 3150–2. doi:10.1093/bioinformatics/bts565.

Gernhard, Tanja. 2008. "The Conditioned Reconstructed Process." *Journal of Theoretical Biology* 253 (4). Elsevier: 769–78.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-length transcriptome assembly from RNA-Seq data without a reference genome." *Nat Biotech* 29 (7). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 644–52. http://dx.doi.org/10.1038/nbt.1883.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *Ann. Math. Statist.* 22 (1). The Institute of Mathematical Statistics: 79–86. doi:10.1214/aoms/1177729694.

Plummer, Martyn. 2003. "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling."

Rand, William M. 1971. "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association* 66 (336): 846–50. doi:10.1080/01621459.1971.10482356.