Revising transcriptome assemblies with phylogenetic information in Agalma1.0 - 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 Agalma1.0, and bash scripts for the run can be found at https://github.com/caseywdunn/ms_treeinform/code/agalma.

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 https://github.com/caseywdunn/ms_treeinform/code/phyldog.

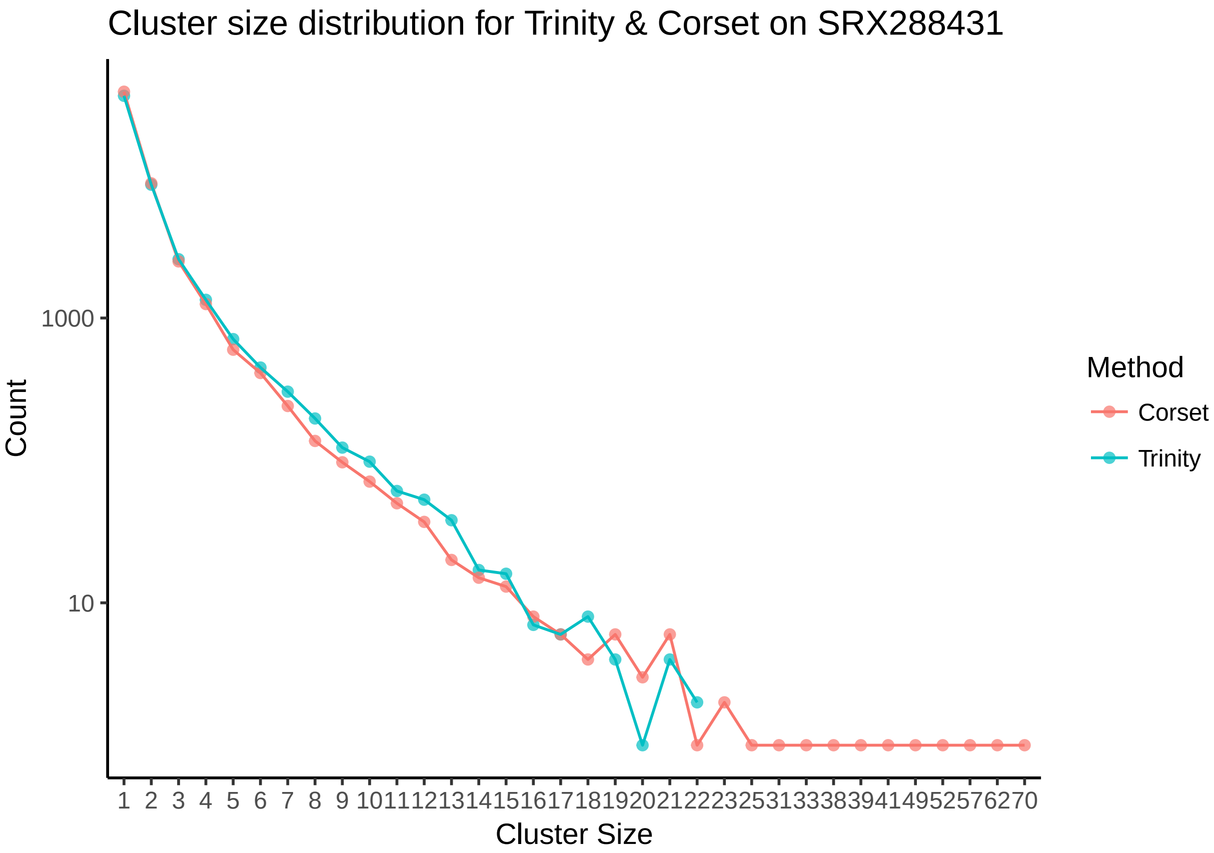
Agalma uses the transcriptome assembler Trinity (Grabherr et al. 2011). Given the intrinsic challenges of assigning assembled transcripts to genes it is likely that the same misassignment errors are generated by other transcriptome assemblers as well.

Comparison to other transcript clustering methods

We ran some corset files…

```
corset_file <- read.csv("data/revisions/corset/SRX288431-clusters.mod.txt", sep="\t", col.names = c("Tr
```

```
trinity_cluster <- cluster_size_distribution(corset_file$Trinity.gene)
corset_cluster <- cluster_size_distribution(corset_file$Corset.gene)
size_dist_df<-rbind(trinity_cluster, corset_cluster)
method <- rep("Trinity",nrow(trinity_cluster)+nrow(corset_cluster))
method[nrow(trinity_cluster)+1:nrow(corset_cluster)]<-"Corset"
size_dist_df$'Method' <- method
ggplot(data=size_dist_df, aes(x=size,y=freq,color=Method,group=Method)) + geom_point(alpha=0.7) + scale
```



Cluster size distribution for Trinity & Corset on SRX288431

Assessing the extent of transcript assignment errors

We first examined the prevalence of transcript misassignment. For each node in each of the 5202 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.

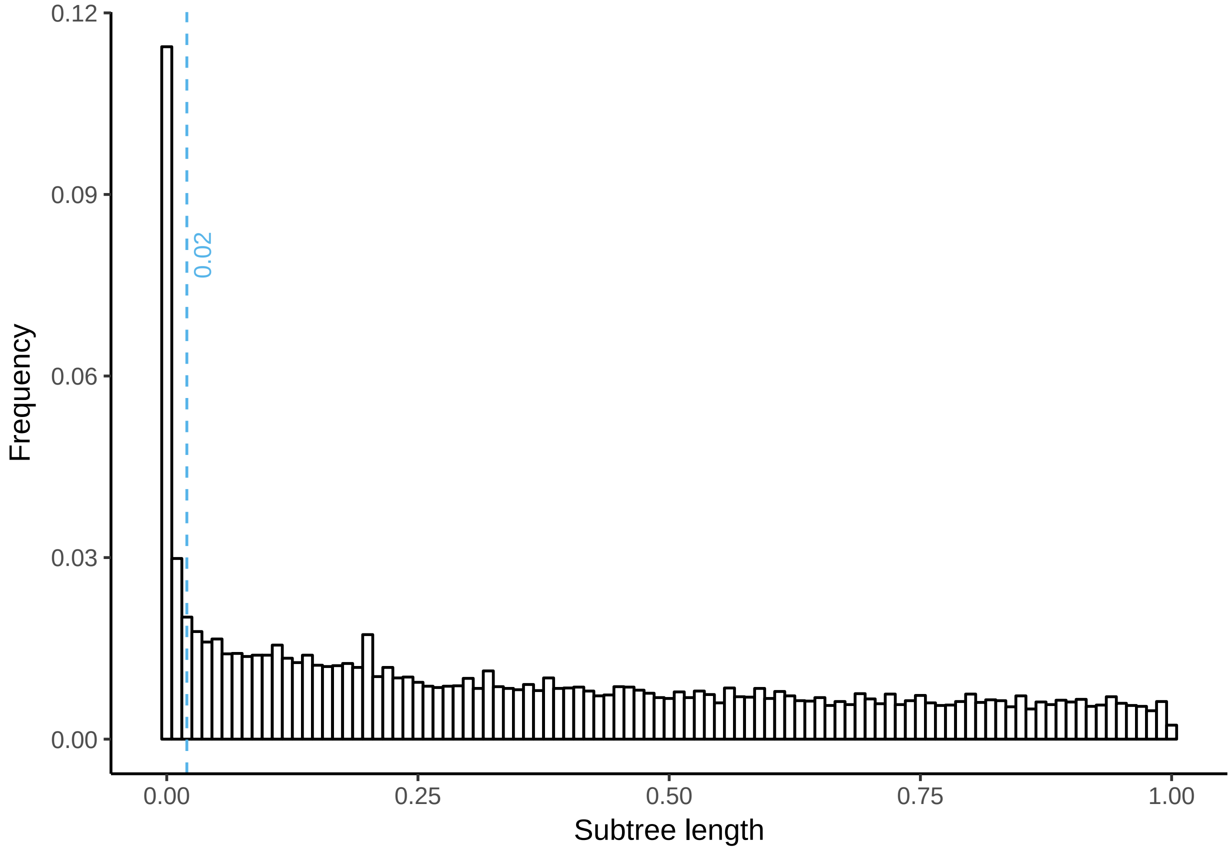


Figure 1: Histogram of subtree lengths for internal nodes in each Siphonophora subset gene tree from Agalma1.0 containing tip descendants from the same species. Subtree lengths greater than 1 were filtered out for clarity.

Selecting a threshold for transcript reassignment

A visual inspection of the histogram of subtree lengths (Supplementary Figure 1) suggested that 0.02 *(TODO: new thresh is 1.65)* was an appropriate threshold for this particular dataset, as the frequency of subtree length for internal nodes was high below such threshold but leveled out above it. In addition, 7.257% of internal nodes containing tip descendants from the same species had a subtree length less than 0.01, with an additional 1.395% 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 2).

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.

If $x_{i,k}$ represents duplication times i from gene tree G_k, π_1 and π_2 denote the 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 $f(x_{i,k}|t_{or,k} = 1, \lambda, \mu)$, then 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 1.

Table 1: Summary of parameter estimates from JAGS.

	Lower95	Mean	Upper95	MCerr
α	0.3633600	0.3864139	0.4027500	0.0028330
β	2.2433200	2.5022162	2.7950000	0.0189497
μ	0.0000006	0.0141640	0.0426011	0.0001946
λ	2.8090300	2.9706293	3.1302000	0.0014825
π_1	0.3828620	0.4197476	0.4547310	0.0004575
π_2	0.5452690	0.5802524	0.6171380	0.0004575

As the intersection point of the two components of the mixture model signals the duplication time point at which more duplication events are likely to arise from transcripts from different genes assigned to different genes, back-calibrating that intersection point provided a threshold for use in treeinform. Specifically, we took all duplication events with times below the intersection point on all chronogram-fitted gene trees, mapped

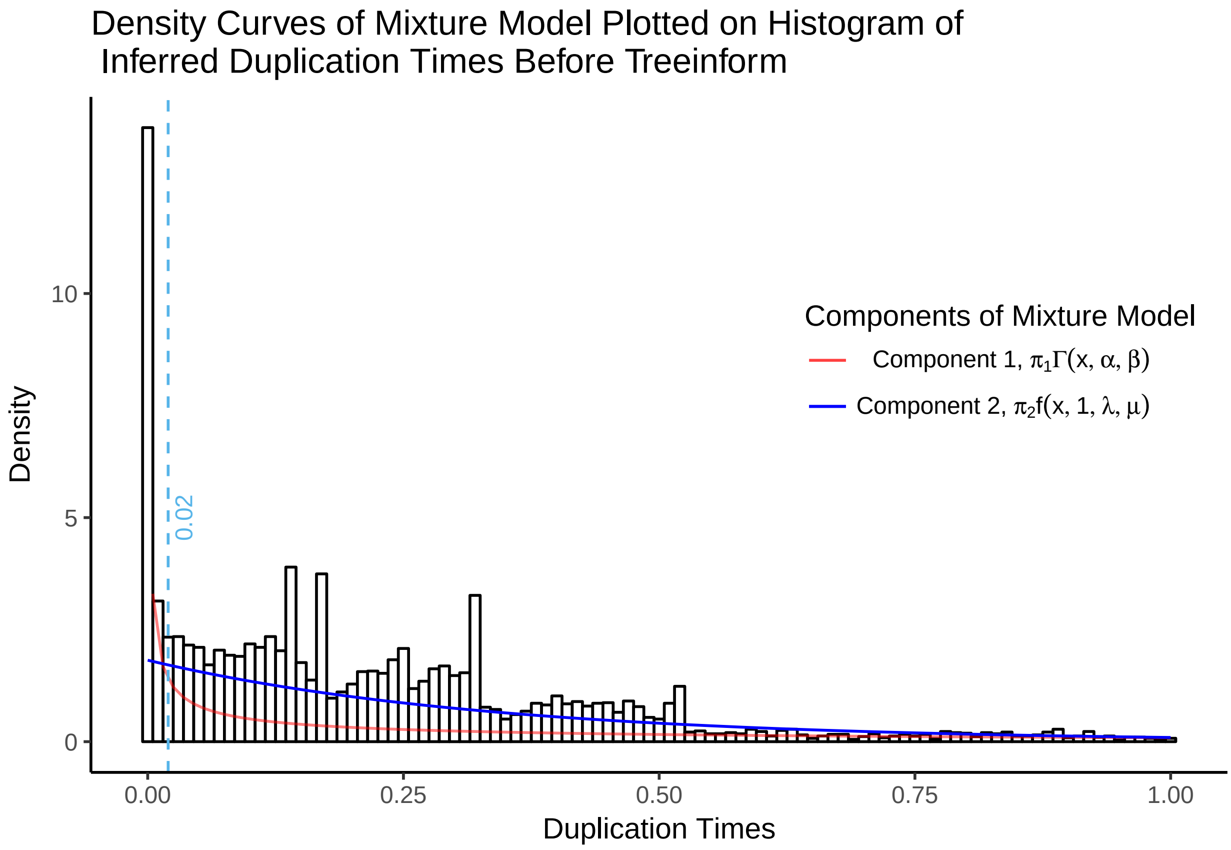


Figure 2: Histogram of the inferred duplication times. We first ran phyldog (Boussau et al. 2013) on the Siphonophora subset multiple sequence alignments from Agalma1.0 and a user-inputted species tree. This provided gene trees with internal nodes annotated as duplication or speciation events. We then fitted chronograms onto these gene trees with our user-inputted species tree. In the overlaid mixture model, the intersection point between the two distribution curves was 0.009.

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 0.01407, this gave us a threshold of 1.65. This suggested that a threshold choice of 0.02 was appropriate, consistent with the threshold sugged by Supplementary Figure 1.

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 0.02 *(TODO:threshold is 1.65 in revision)* (Supplementary Figure 3). 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.

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 Agalma1.0 before treeinform, and gene trees from Agalma1.0 after treeinform under 3 different thresholds: 50, 0.05, and the default value 0.02 (Supplementary Figure 4). We again fitted chronograms with the same Siphonophora species tree onto all gene trees from Agalma1.0 and filtered out those gene trees with time of origin greater

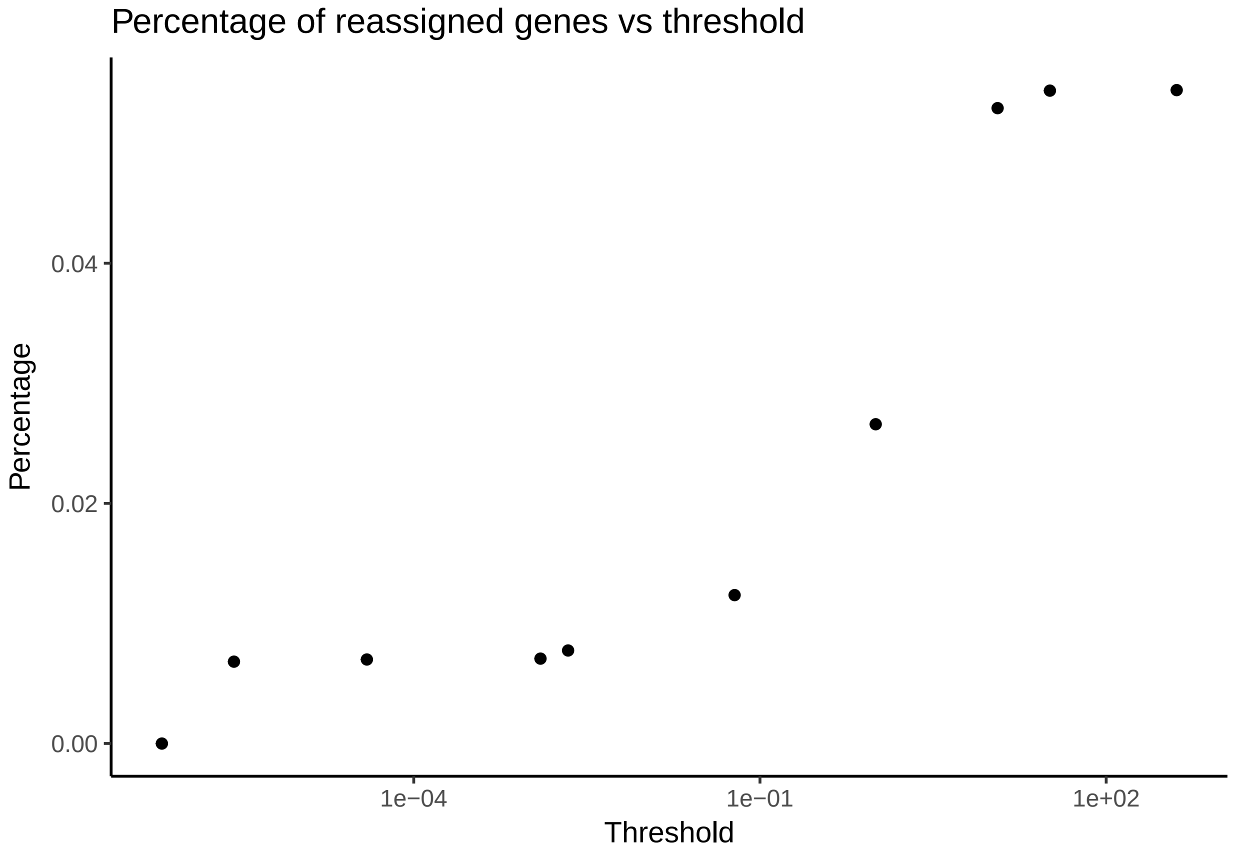


Figure 3: The percentage of reassigned tips is plotted above on a log scale. The original assembly had 315,041 genes, with at most 23,396 possible candidates (7.43% of genes) for reassignment. The default threshold for treeinform is highlighted in light blue. This threshold value is robust over a wide range from 0.02 down several orders of magnitude.

than 1, so that duplication times were comparable between trees. Visually, the analyses with the 0.02 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 2). 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 0.02 come closest to the theoretical distribution 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.

Software versions

This manuscript was computed on Fri Mar 30 16:29:21 2018 with the following R package versions.

```
R version 3.4.3 (2017-11-30)
Platform: x86_64-apple-darwin15.6.0 (64-bit)
Running under: macOS High Sierra 10.13.3

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_1.20     
 [5] treeio_1.2.2    hutan_0.5.1     dplyr_0.7.4     ggplot2_2.2.1  
 [9] scales_0.5.0    ape_5.0        

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.16     highr_0.6        pillar_1.2.1     compiler_3.4.3  
 [5] plyr_1.8.4       bindr_0.1.1      tools_3.4.3      digest_0.6.15   
 [9] jsonlite_1.5     evaluate_0.10.1  tibble_1.4.2     gtable_0.2.0    
[13] nlme_3.1-131.1   lattice_0.20-35  pkgconfig_2.0.1  rlang_0.2.0     
[17] rvcheck_0.0.9    yaml_2.1.18      coda_0.19-1      stringr_1.3.0   
[21] rprojroot_1.3-2  grid_3.4.3       glue_1.2.0       R6_2.2.2        
[25] rmarkdown_1.9    magrittr_1.5     backports_1.1.2  htmltools_0.3.6 
[29] assertthat_0.2.0 colorspace_1.3-2 labeling_0.3     stringi_1.1.7   
[33] lazyeval_0.2.1   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.

Dunn, Casey W. 2009. "Siphonophores." *Current Biology* 19 (6). Cell Press:R233–R234. https://doi.org/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. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3840672.

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. https://doi.org/10.1214/aoms/1177729694.

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