Revising transcriptome assemblies with phylogenetic information in Agalma1.0 - Supplementary Information

Contents

Assessing the extent of transcript assignment errors **1**

Selecting a threshold for transcript reassignment through treeinform **1**

Validating the effectiveness of treeinform **3**

References **7**

Assessing the extent of transcript assignment errors

To assess how often transcripts from the same gene were assigned to different genes, we built histograms of both subtree lengths (Figure 1) and gene duplication times (Figure 2) of 5304 gene trees from a phylogenetically well-resolved 7 species subset of Siphonophora. This provided a broad look at how prevalent transcript misassignment was, as large peaks close to zero length or time would suggest that transcript misassignment was very common across the gene trees.

The histogram of the subtree length provided perspective on the occurrence of the problem in the gene trees from Agalma1.0, while the histogram of the duplication times provided perspective on the impact of the problem. On both histograms, a very large peak close to zero length or time exists, suggesting that transcript misassignment is very common across this set of gene trees.

Selecting a threshold for transcript reassignment through treeinform

A preliminary visual inspection of the histogram of subtree lengths (Figure 1) suggested that 0.02 was an appropriate threshold for this particular dataset, as bin counts were very high under that threshold, and leveled out in frequency after. However, subtree lengths are in units of expected numbers of substitution, which combine rates of molecular evolution and time. Because we expect gene families to evolve at different rates, we cannot directly compare features such as subtree length between gene trees without calibrating the trees to account for the different rates.

As the Siphonophore subset we analyzed is well-resolved, we fitted chronograms onto the gene trees with a user-inputted species tree with height 1, transforming the branch lengths into units of time relative to 1. This made it possible to directly compare between gene trees and perform a more rigorous analysis to select a threshold. Given that birth-death models are well studied and often applied to gene duplication and loss, we decided to fit a mixture model onto the inferred duplication times 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, and one component modelled duplication events and associated times arising from transcripts assigned to different genes that in fact do belong to different genes (Figure 2).

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. To be more precise, we took all duplication events with times below the intersection point on all chronogram-fitted gene

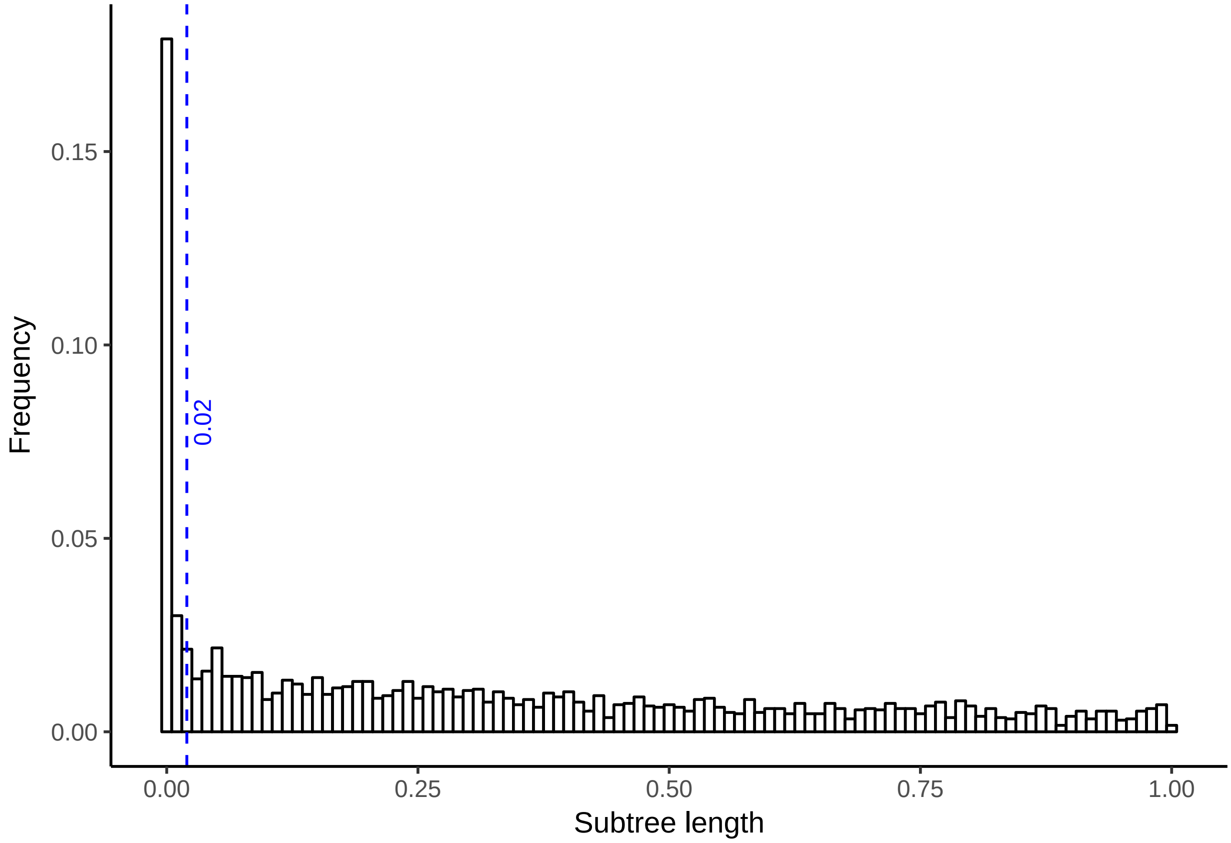

Figure 1: Histogram of subtree lengths for internal nodes in each Siphonophora subset gene tree from Agalma1.0 containing tip descendants from the same species. Subtree lengths greater than 1 were filtered out for clarity. 19.12% of internal nodes containing tip descendants from the same species from the same species have a subtree length of less than 0.01, with an additional 2.17% having a subtree length between 0.01 and 0.02. It is unlikely that all of these clades are gene duplication events. After 0.02, the distribution of subtree lengths levels out.

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 0.00979974606909549, this gave us a threshold of 0.1848778. This suggested that a threshold choice of around 0.02 was appropriate.

We expected duplication events from transcripts assigned to different genes that belong to the same gene to have very short duplication times approaching 0, and thus chose to model that component (Component 1) as a gamma distribution with parameters shape= α and rate= β. For duplication events from transcripts assigned to different genes that in fact do belong to different genes, we used the probability distribution function given by Gernhard (Gernhard 2008) (Component 2) which 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 $t_{or} = 1$, although technically it is an incorrect estimate of the age of both the gene tree and the species tree. 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.

In mathematical terms, 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 pdf 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	0.2271265	0.2424239	0.2580250	0.0007320
β	1.1420843	1.7861715	2.5024476	0.0184331
μ	0.0000006	0.0789256	0.2355944	0.0010044
λ	2.5932256	2.9634860	3.3280288	0.0018998
π_1	0.2373131	0.2840289	0.3337098	0.0003343
π_2	0.6662902	0.7159711	0.7626869	0.0003343

Validating the effectiveness of treeinform

In order to validate that treeinform produces more accurate gene trees, 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, 0.02, and 0.05 (Figure 4). We again fitted chronograms with the same user-inputted Siphonophora subset species tree onto all gene trees from Agalma1.0 and filtered out those gene trees with time of origin greater than 1, so that duplication times were comparable between trees.

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 come closest to the theoretical distribution. This confirms that treeinform produces more accurate gene trees.

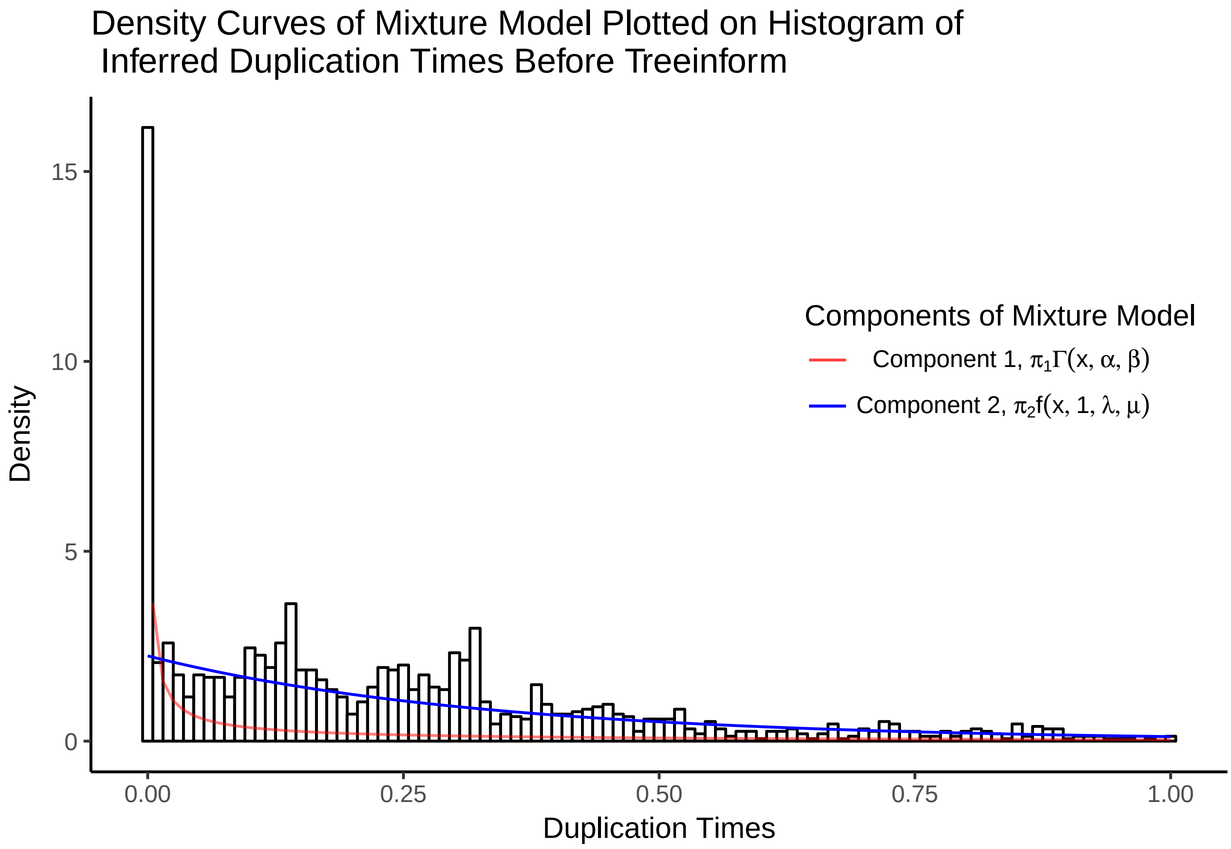

Figure 2: Histogram of the inferred duplication times. We first ran phyldog (Boussau et al. 2013) on the Siphonophora subset multiple sequence alignments from Agalma1.0 and a user-inputted species tree. This provided gene trees with internal nodes annotated as duplication or speciation events. We then 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.

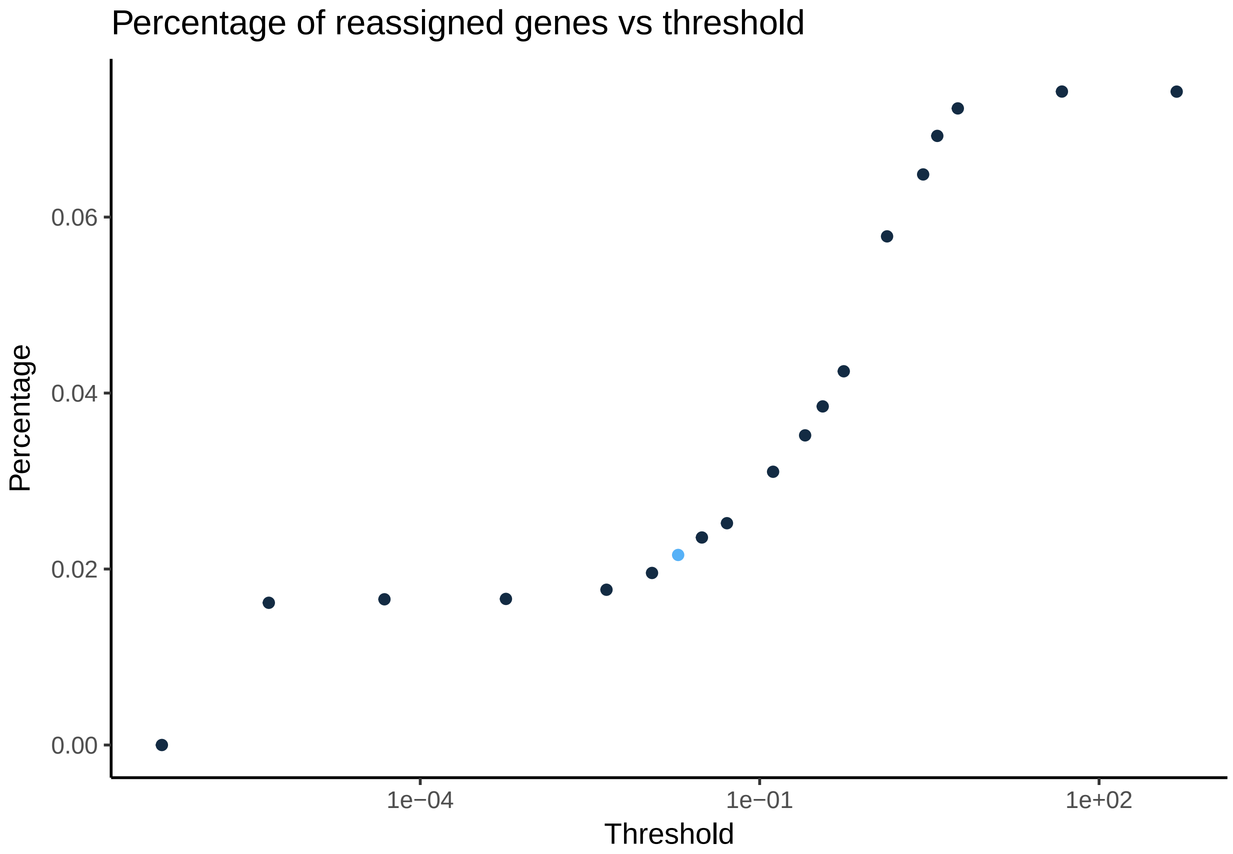

Figure 3: The percentage of reassigned tips is plotted above. The original assembly had 315,041 genes, with at most 23,396 possible candidates (7.43% of genes) for reassignment. The default threshold for treeinform is highlighted in blue.

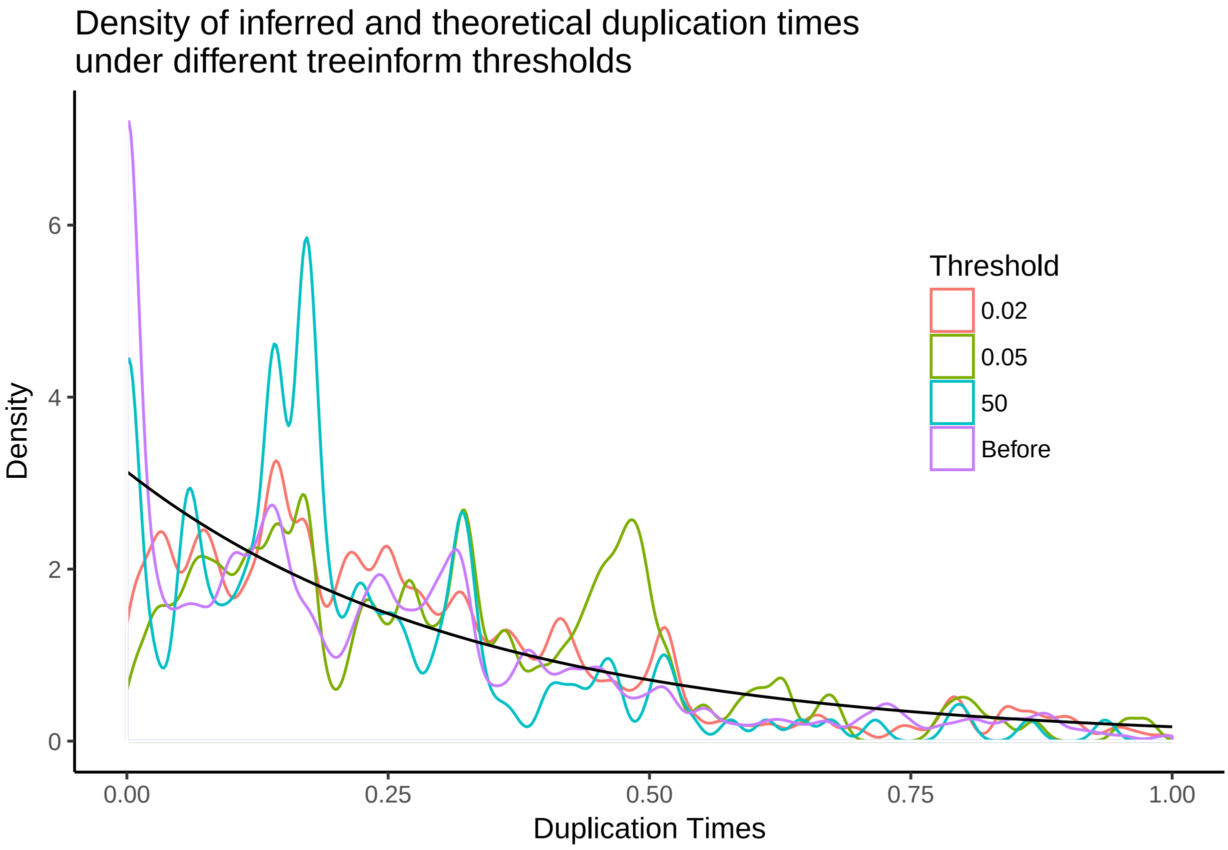

Figure 4: Density from theoretical and the empirical density under the 3 different thresholds. It appears that 0.02 looks the closest to the theoretical density.

Table 2: Kullback-Leibler distances between duplication times after running treeinform with different thresholds and theoretical duplication times.

	KL.Distance
Before	0.2871064
50	0.6468972
0.02	0.1606074
0.05	0.3216297

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.

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

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."