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

Contents

Treeinform motivation and threshold selection analysis

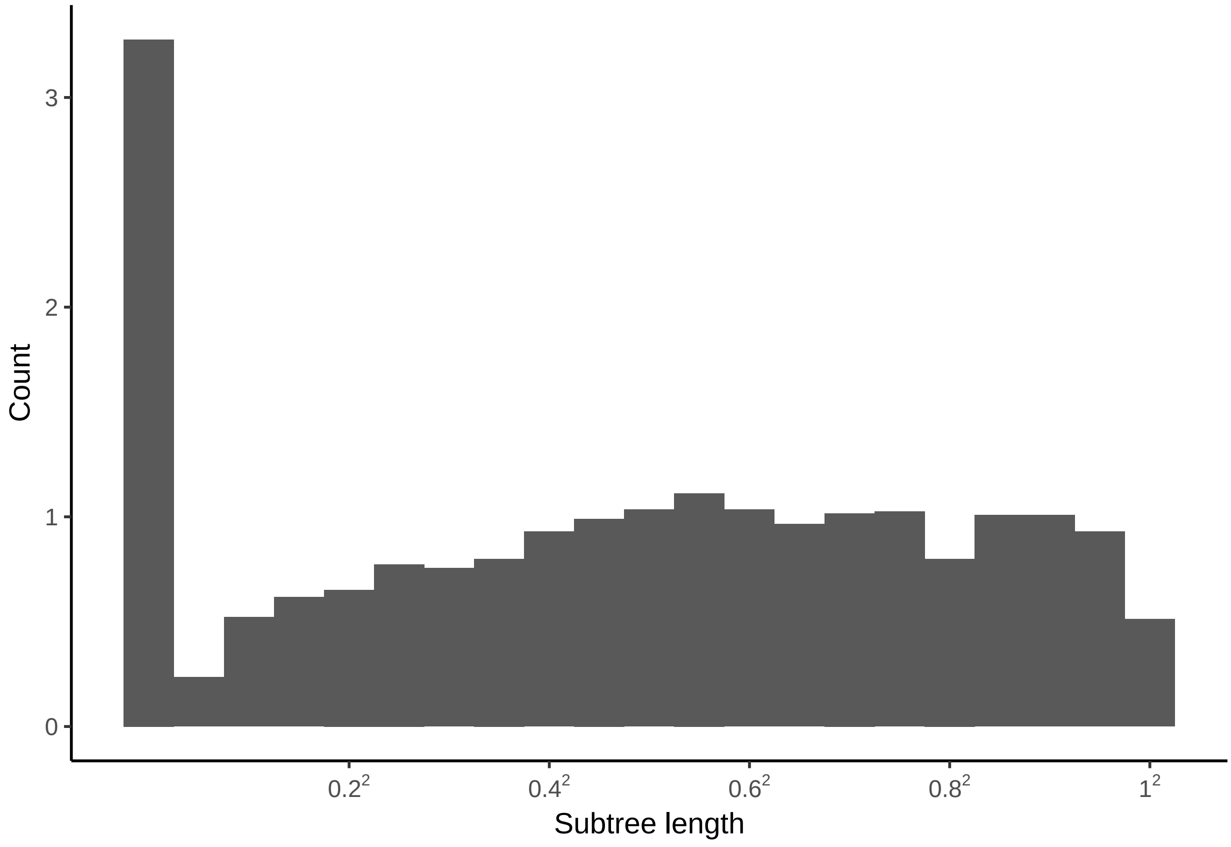

Figure 1: We plotted a histogram of subtree lengths for internal nodes in each Siphonophora subset gene tree from Agalma1.0 containing tip descendants from the same species. We filtered subtree lengths greater than 1 out for clarity. A high proportion of genes from the same species have a subtree length of close to 0. It is unlikely that all of these clades are gene duplication events. Points are plotted on a sqrt scale. From this histogram, we visually selected 0.05 based on the fact that bin counts were very high under that threshold, with a noticeable gap after that threshold.

To more closely examine threshold selection, we ran phyldog (Boussau et al. 2013) on the Siphonophora subset multiple sequence alignments from Agalma1.0 and a user-inputted species tree in order to infer 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. This allowed us to fit a mixture model onto the duplication times, where 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 provides a threshold for use in treeinform. To be more precise, we take all duplication events with times below the intersection point on all chronogram-fitted gene trees, map them to the equivalent events on the phyldog-outputted gene trees, compute the subtree length of all events, and then take the maximum of those subtree lengths. This gives us a threshold of 0.0208. This intersection point suggests that a threshold choice of around 0.05 is 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 can use Bayesian Gibbs Sampling with the Bernoulli Ones trick to infer the parameters α, β, λ, and μ as well as the mixing proportions π_1 and π_2. The package we used to do this analysis was Just Another Gibbs Sampler (JAGS).

From JAGS we get the parameter estimates as:

	Lower95	Median	Upper95	Mean	MCerr
α	0.2271265	0.2424161	0.2580250	0.2424239	0.0007320
β	1.1420843	1.7535964	2.5024476	1.7861715	0.0184331
μ	0.0000006	0.0552104	0.2355944	0.0789256	0.0010044
λ	2.5932256	2.9634120	3.3280288	2.9634860	0.0018998
π_1	0.2373131	0.2832335	0.3337098	0.2840289	0.0003343
π_2	0.6662902	0.7167665	0.7626869	0.7159711	0.0003343

Threshold selection validation

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, 0.05, and 5e-05. We again fitted chronograms with the same

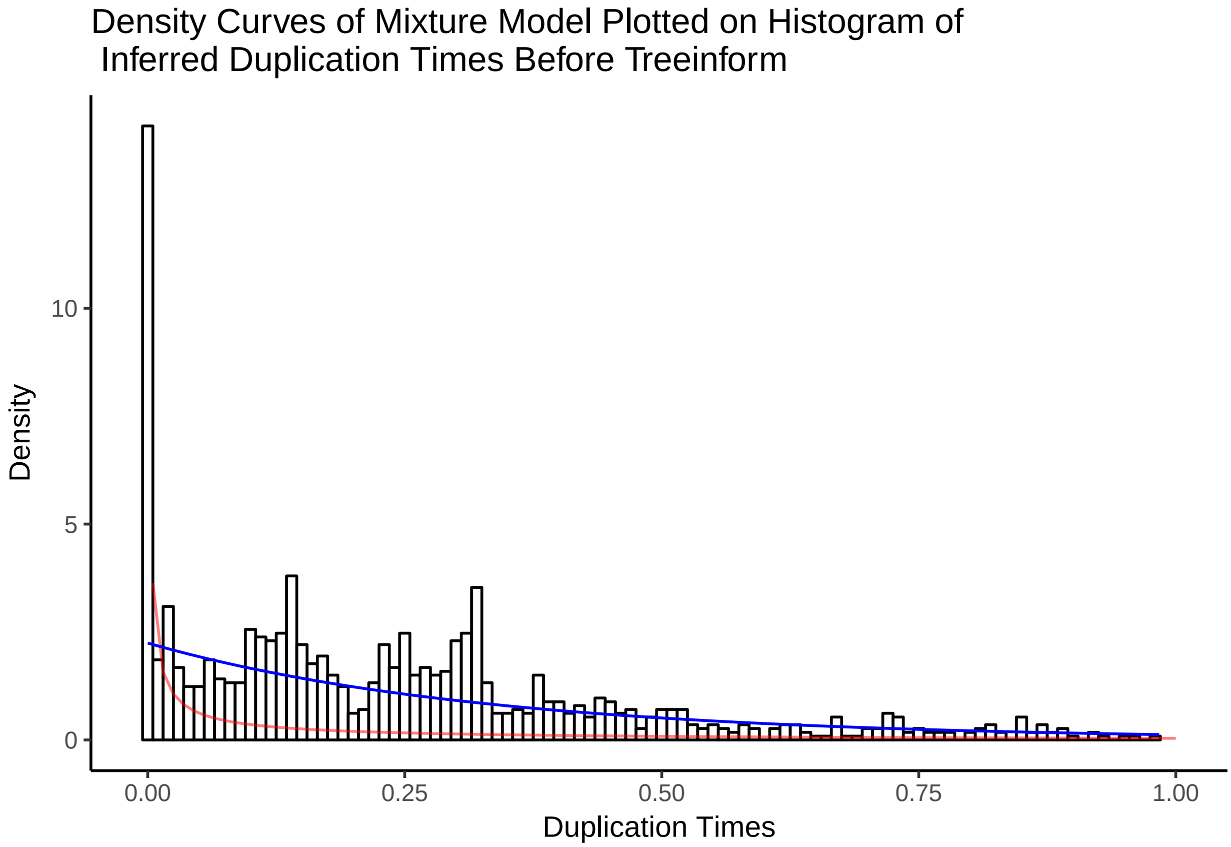

Figure 2: Plotted onto a histogram of the inferred duplication times, we see that most of the duplication times attributed to splice variants are < 0.05 with the actual intersection point between the two distribution curves being 0.009799746.

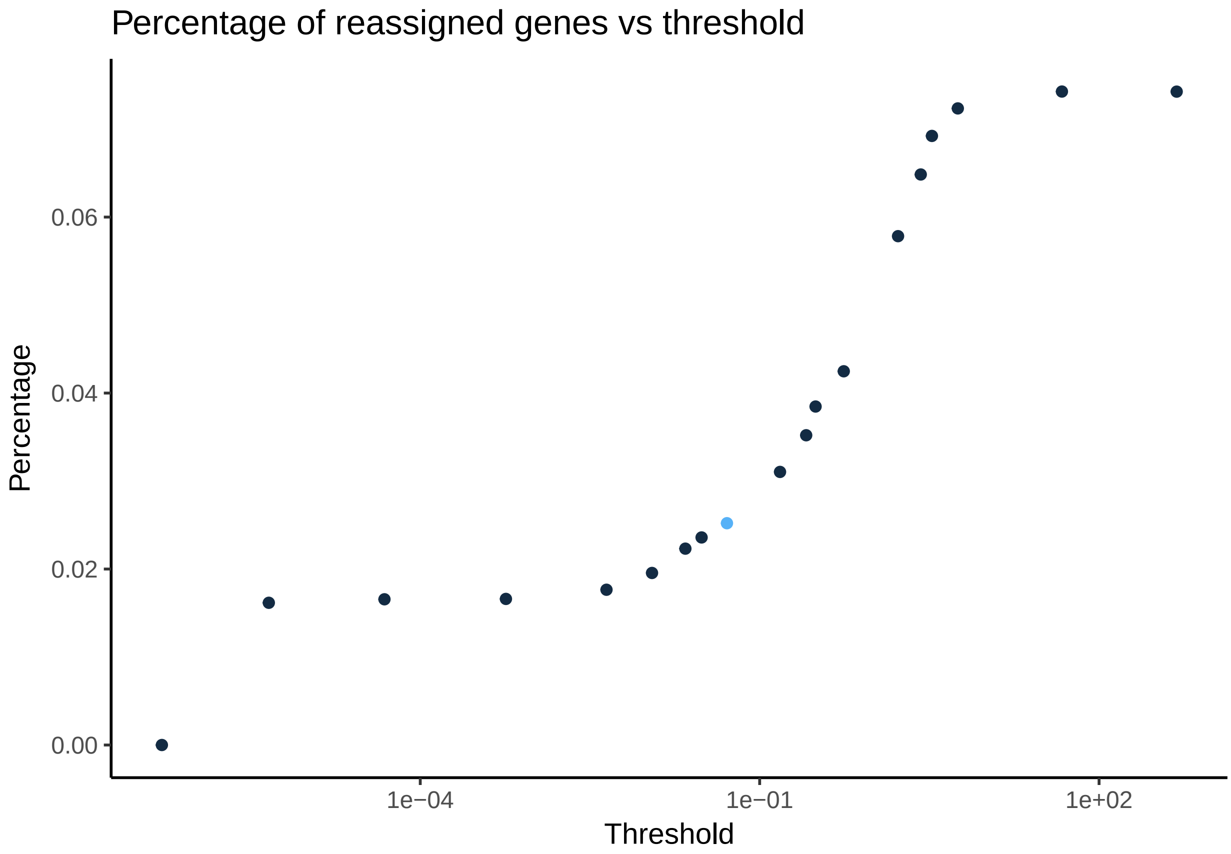

Figure 3: The percentage of reassigned tips is plotted above. The original assembly had 315,041 genes, with at most 23,396 possible candidates (7.43% of genes) for reassignment. The default threshold for treeinform is highlighted in blue.

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.

For the duplication times under the model provided for Component 2 of the mixture model, we first estimated λ and μ using CAFE3 (Han et al. 2013) with the same fixed $t_{or} = 1$ and $G = \{G_1, \ldots, G_K\}$=gene family sizes for gene trees $1, \ldots, K$. We then plugged the λ and μ estimates from CAFE3 in along with the same calibrated $t_{or,k} = t_{or} = 1$s.

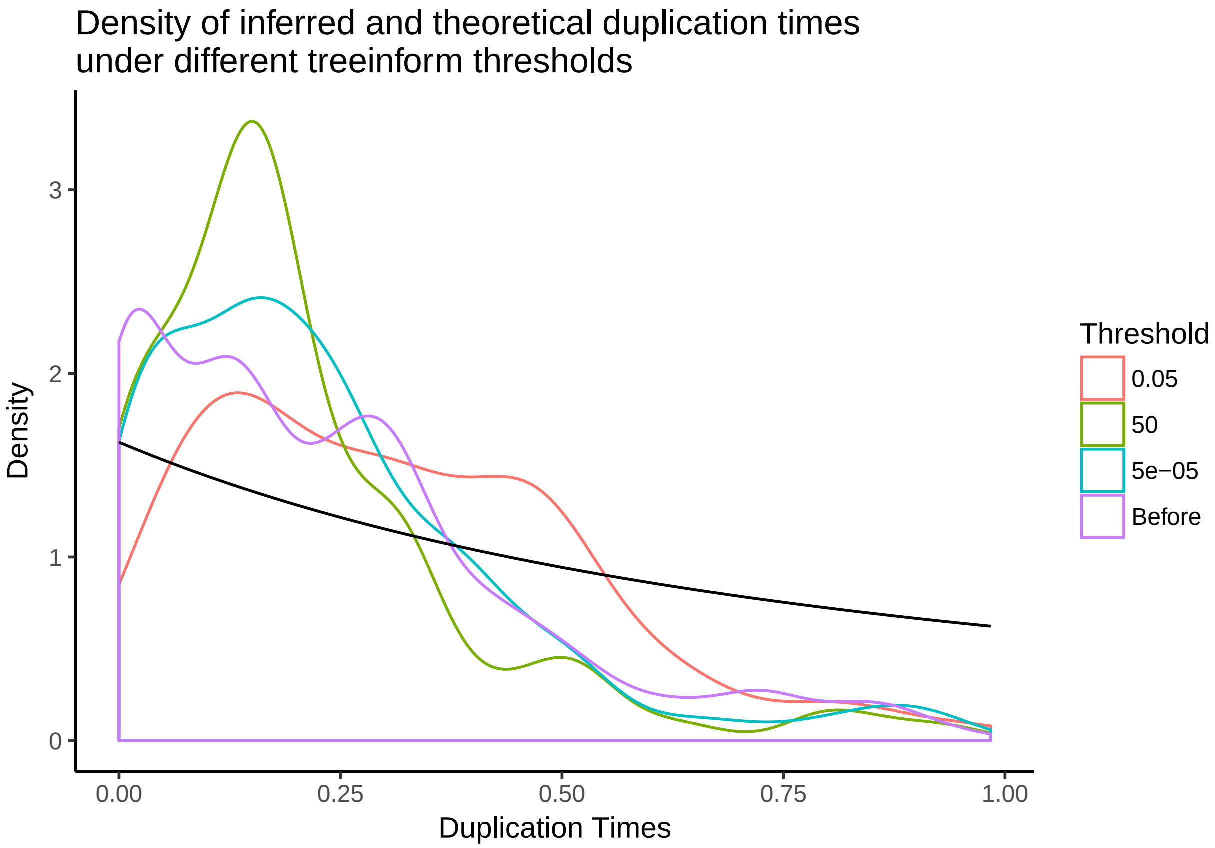

Figure 4: We plotted the density from (Gernhard 2008) and the empirical density under the 3 different thresholds together to get a visual of how accurate estimated duplication times are after running treeinform with different thresholds. It appears that 0.05 looks the closest to the density from (Gernhard 2008).

Additionally, we ran the Kolmogorov-Smirnov test on the cumulative distribution functions of the duplication times under different thresholds against the cumulative distribution function of the duplication times under (Gernhard 2008) to compare. Although the K-S tests fail for all thresholds, 0.05 comes the closest.

	Statistic	P-Value
Before	0.3134862	0
50	0.4228759	0
0.05	0.2440439	0
5e-05	0.3308351	0
0.01	0.3093898	0

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.

Han, Mira V, Gregg WC Thomas, Jose Lugo-Martinez, and Matthew W Hahn. 2013. "Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using Cafe 3." *Molecular Biology and Evolution* 30 (8). SMBE: 1987–97.