Supplementary Figure proposals

This is a sketch of an alternative default threshold which is in accordance with the parameters for λ and μ, birth and death rates for duplication events, inferred from the mixture model. In that model, $\lambda = 2.9634860$, and $\mu = 0.0789256$. The intersection point from that model was 0.009 in units of time, giving a threshold of approximately 0.02 in units of expected numbers of substitution.

How this would differ from how we originally laid out treeinform and the supplementary material:

- λ and μ rates would come from the mixture model inference. In the original supplementary information, λ and μ rates came from a program called CAFE3, which was run on the set of gene trees pre-treeinform.
- The default threshold would change from 0.05 to 0.02 based on the back-calibration of the threshold from the mixture model. The mixture model would serve as the justification for the default threshold. In the original supplementary information, 0.05 was chosen based on an analysis of subtree lengths for internal nodes with multiple descendants from the same species.
- Figure 3b and Table 3c would change. We would plot the theoretical pdf of duplication times using λ and μ from the mixture model. With this change, Table 3c shows that using 0.02 as a threshold brings us closest to the theoretical cdf of duplication times. In the original supplementary information, Table 3c showed that using 0.05 as a threshold brought us closest to the theoretical cdf of duplication times. Figure 3b and Table 3c were also used as validation for the selection of 0.05 as a threshold.

One drawback of this is that the birth (λ) and death (μ) rates inferred from the mixture model don't seem realistic. The birth rate is very high and the death rate is very low. This suggests that the mixture model was still skewed by the presence of what are likely transcript assignment errors, i.e. transcripts belonging to the same gene being assigned to different genes.

One thing we could do is rerun the mixture model with different initial gamma parameters (we used a gamma distribution to model transcripts belonging to the same gene being assigned to different genes) that have a larger area under the curve, or rerun the mixture model with bounds on the birth and death rates.

Below is a view of how Figure 3b and Table 3c would change. These analyses were done with phyldog run on gene trees from the 7taxa Siphonophora dataset after treeinform with threshold=0.025, as I am currently running Agalma1.0 with treeinform threshold=0.02 and waiting on the results.

Figure 3b

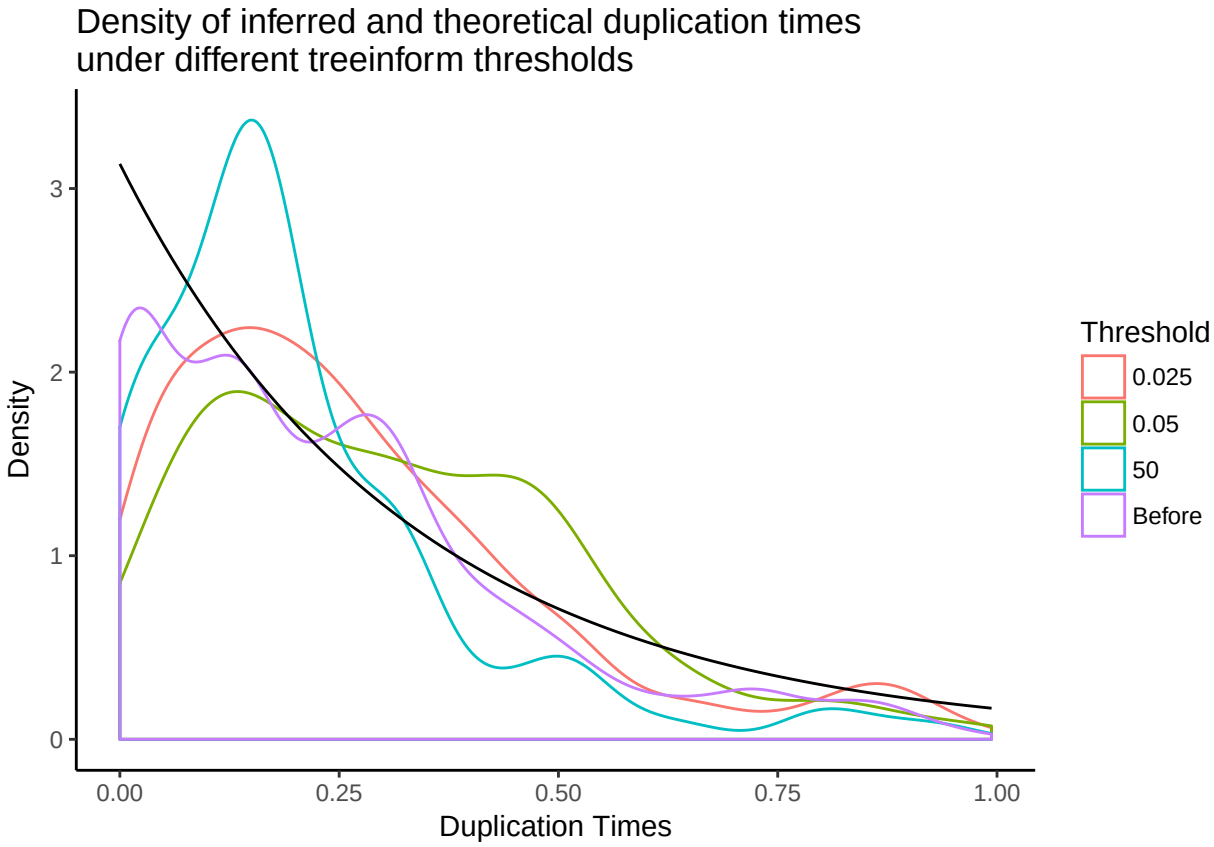

Table 3c

	Statistic	P-Value
Before	0.1293732	0.0000000
50	0.2118081	0.0000206
0.05	0.1138487	0.0231532
0.025	0.0740171	0.0105115