treeinform: Revising assemblies with phylogenetic information - Supplementary Materials

Contents

Figure 1: Gene tree before and gene tree after

Still have to do this...

```
tree = read.tree("data/1006.newick")
plot(tree)
```

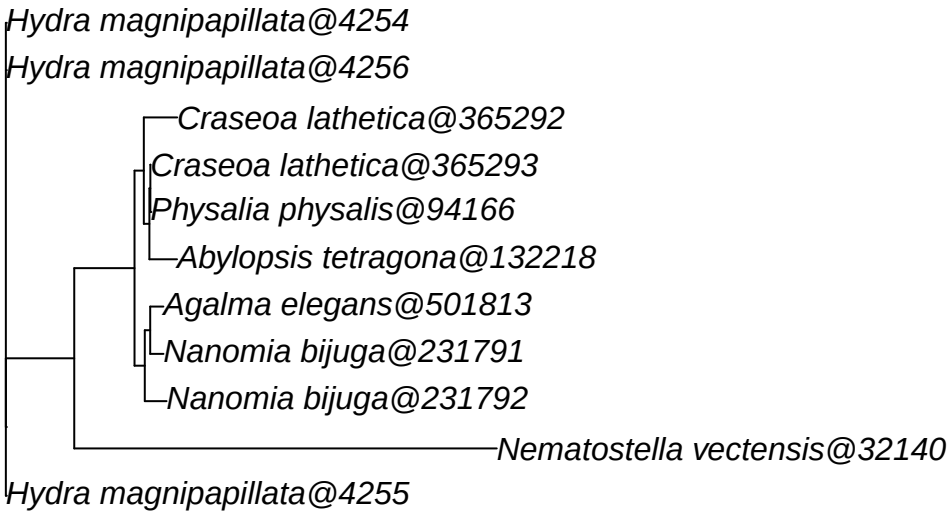


Figure 2: Threshold selection analysis

Figure 2a: Percentage of reassigned tips

The percentage of reassigned tips is plotted below. The original assembly had 315,041 genes, with at most 23,396 possible candidates (7.43% of genes) for reassignment.

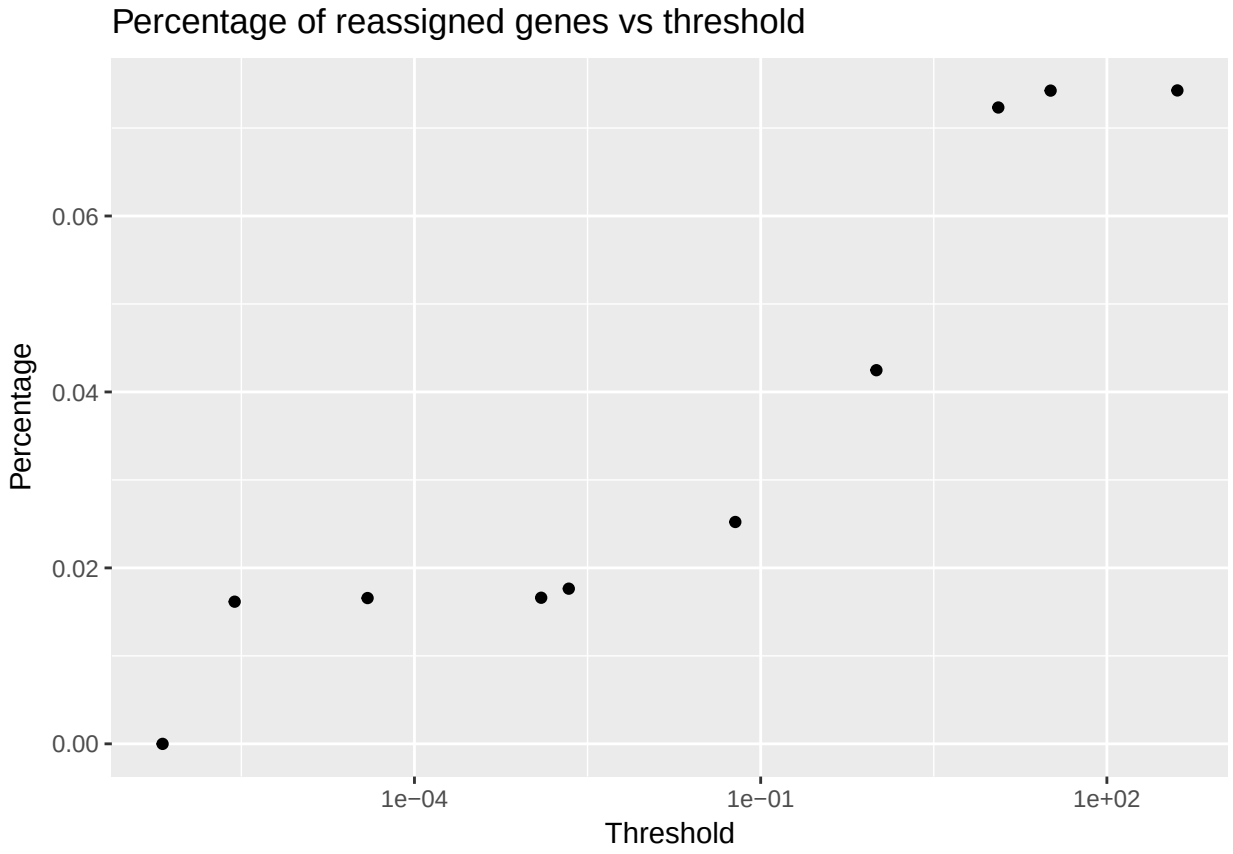


Figure 2b: Theoretical duplication times compared to empirical duplication times from before treeinform and after treeinform

In order to validate that treeinform produces more accurate gene trees, we compare the density of duplication times under a birth-death model with parameter λ=birth rate, μ=death rate, and $t_{or} = \{t_{or,1}, t_{or,2}, \ldots, t_{or,K}\}$=time of tree origin for gene families $1, \ldots, K$ to the distribution of estimated duplication times for gene trees from before treeinform, and gene trees from after treeinform under 3 different thresholds: 50, 0.05, and 5e-05.

For the distribution of estimated duplication times of gene trees before and after treeinform, we used phylDOG (Boussau et al. 2013) to estimate duplication times in units of expected number of substitutions, then calibrated branch lengths to units of time based on the Siphonophora species tree that was calibrated to 1 using the chronos function from ape. (TODO: expand on this)

The pdf of duplication times under a birth-death model and given a time of origin is:

$$f(x_{i,k}|t_{or,k}, \lambda, \mu) = (\lambda - \mu)^2 \frac{e^{-(\lambda-\mu)x_i}}{(\lambda - \mu e^{-(\lambda-\mu)x_i})^2} \frac{\lambda - \mu e^{-(\lambda-\mu)t}}{1 - e^{-(\lambda-\mu)t}}$$

where $x_{i,k}$ represents duplication times i from gene tree G_k, as derived by Gernhard 2008. To compute this, we first estimate λ and μ using CAFE3 (Han et al. 2013) with the same fixed $t_{or} = 1,000,000$ and $G = \{G_1, \ldots, G_K\}$=gene family sizes for gene trees $1, \ldots, K$, then plug the λ and μ estimates from CAFE3 in along with the same calibrated $t_{or,k}$s. A key assumption here is that the fixed $t_{or,k}$s allows for accurate comparisons between the theoretical density of duplication times and the estimated duplication times even if it is technically an incorrect estimate of the age of the gene trees (and species tree).

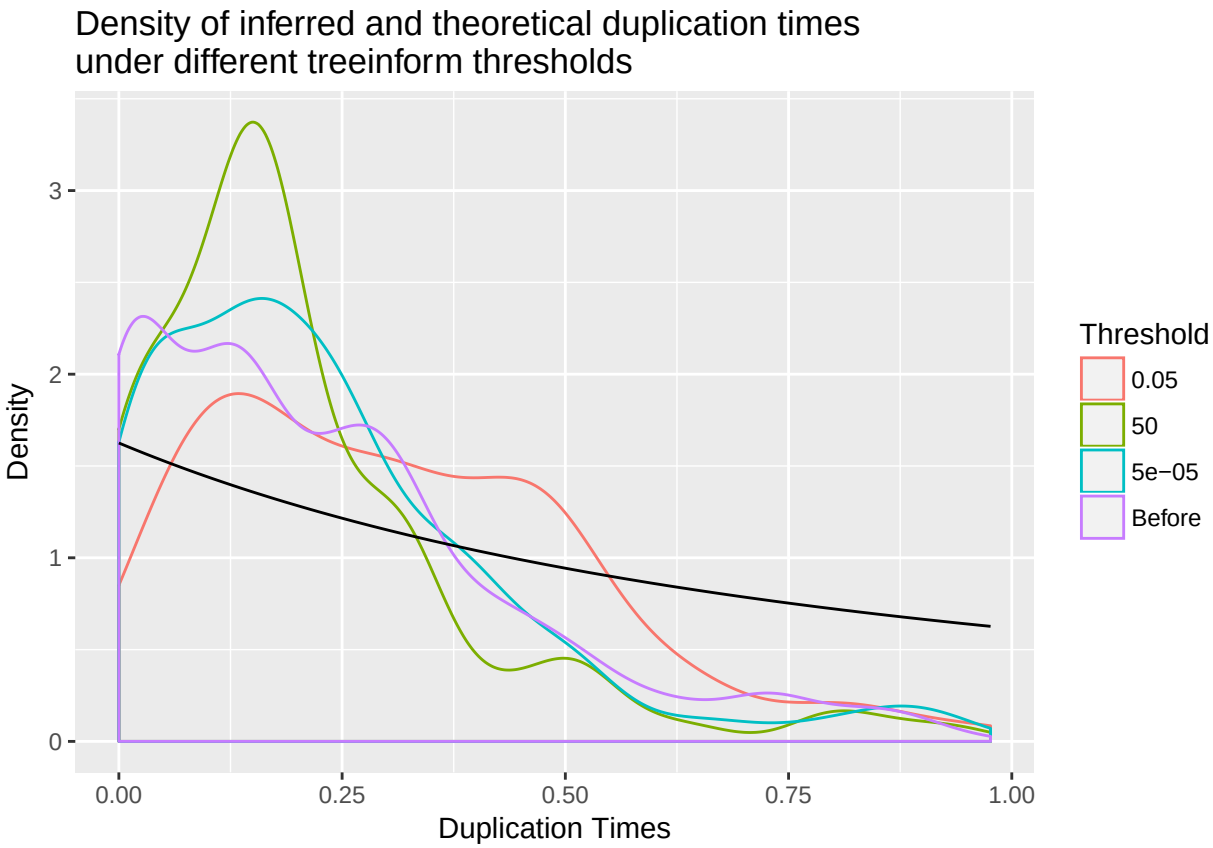


Table 2c: Kolmogorov-Smirnov statistics for before treeinform and after treeinform

Additionally, we can run the Kolmogorov-Smirnov test on the CDFs of the thresholds to compare. The CDF of the theoretical distribution is:

$$F(x_{i,k}|t = t_{or,k}, \lambda, \mu) = \frac{1 - e^{-(\lambda-\mu x_{i,k})}}{\lambda - \mu e^{-(\lambda-\mu)x_{i,k}}} \frac{\lambda - \mu e^{-(\lambda-\mu)t}}{1 - e^{-(\lambda-\mu)t}}$$

	Statistic	P-Value
Before	0.3196327	0
50	0.4228759	0
0.05	0.2440439	0
5e-05	0.3308351	0

The default threshold (0.05) comes closest to the theoretical distribution.

Figure 3: Mixture model fit

Under the assumption that transcript annotation errors bias duplication times towards 0 and can be modeled as a gamma distribution, we can view the empirical distribution of duplication times before treeinform as a 2-component mixture of the gamma distribution and the theoretical provided by Gernhard:

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

where $\Gamma(x_{i,k}|\alpha, \beta)$ is the pdf for the gamma distribution with shape α and rate β, and $f(x_{i,k}|t_{or,k} = t, \lambda, \mu)$ is the pdf for the duplication times under birth rate λ and death rate μ from above. π_1 and π_2 denote the mixing proportions, thus $\pi_1 + \pi_2 = 1$.

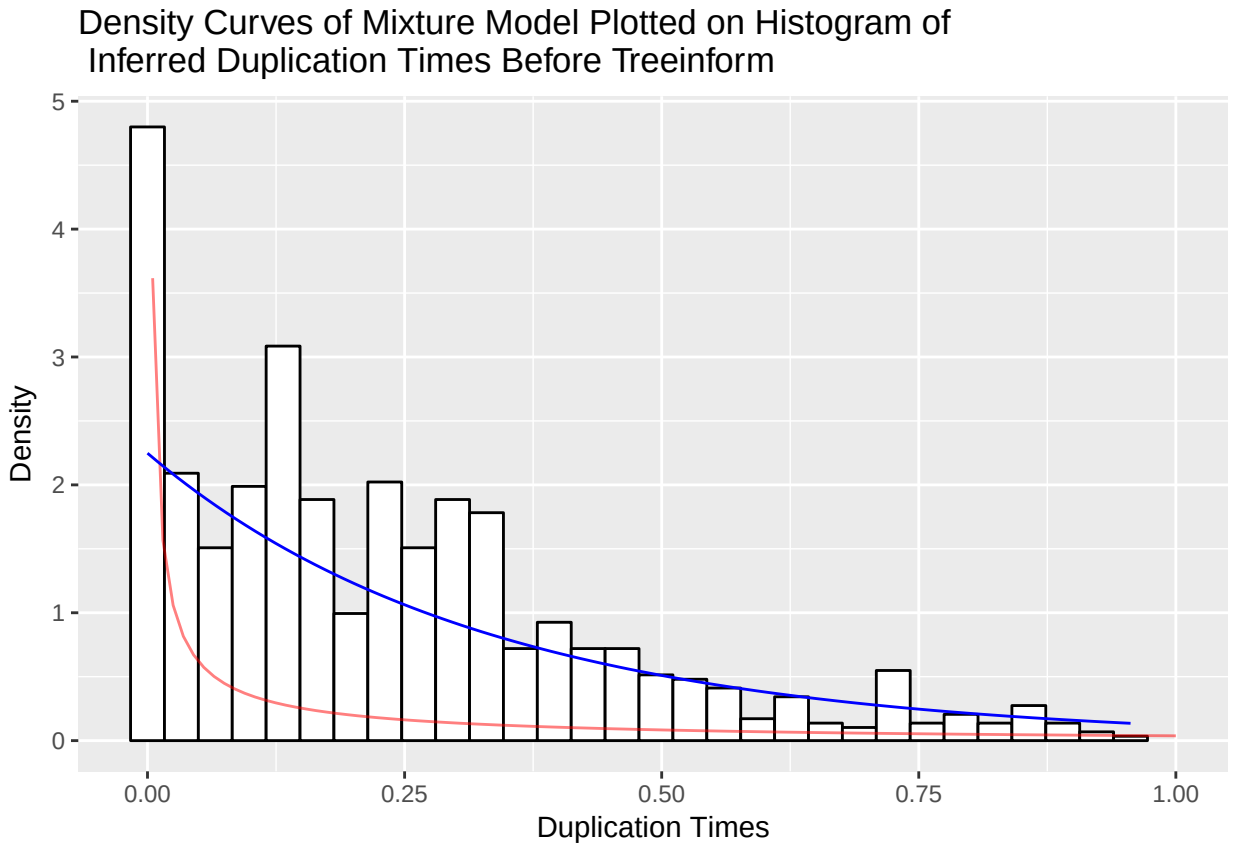
We use Bayesian Gibbs Sampling with the Bernoulli Ones trick to infer the parameters α, β, λ, and μ as well as the mixing proportions p_1 and p_2. The package we use is Just Another Gibbs Sampler (JAGS).

From JAGS we get the parameter estimates as:

	Lower95	Median	Upper95	Mean	MCerr
α	0.2270312	0.2412745	0.2575032	0.2416420	0.0005433
β	1.1484215	1.7723067	2.5840863	1.8109248	0.0204710
μ	0.0000026	0.0565047	0.2360077	0.0799409	0.0009941
λ	2.5952231	2.9618346	3.3411465	2.9639478	0.0019906
π_1	0.2345487	0.2823350	0.3314127	0.2829988	0.0003145
π_2	0.6685873	0.7176650	0.7654513	0.7170012	0.0003145

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. This intersection point suggests that a threshold choice of around 0.05 is appropriate. (NOTE: duplication times is not the same as subtree height since it is calibrated vs uncalibrated, so the times are not directly comparable with threshold choice)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

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.