

Supplemental Materials: Ecological Insights from the Evolutionary History of Microbial Innovations

Mario E. Muscarella¹ & James P. O'Dwyer^{1,2}

¹Department of Plant Biology, University of Illinois

²Carl R. Woese Institute for Genomic Biology, University of Illinois

Supplemental I: Simulation Model

Our empirical results suggest that that observed discrepancy between the conservation and gain-loss frameworks may be related to the loss rate inferred under the gain-loss framework. To test this prediction and estimate the upper bound of accuracy for our inferences, we used an idealized simulation model (Fig. S1).

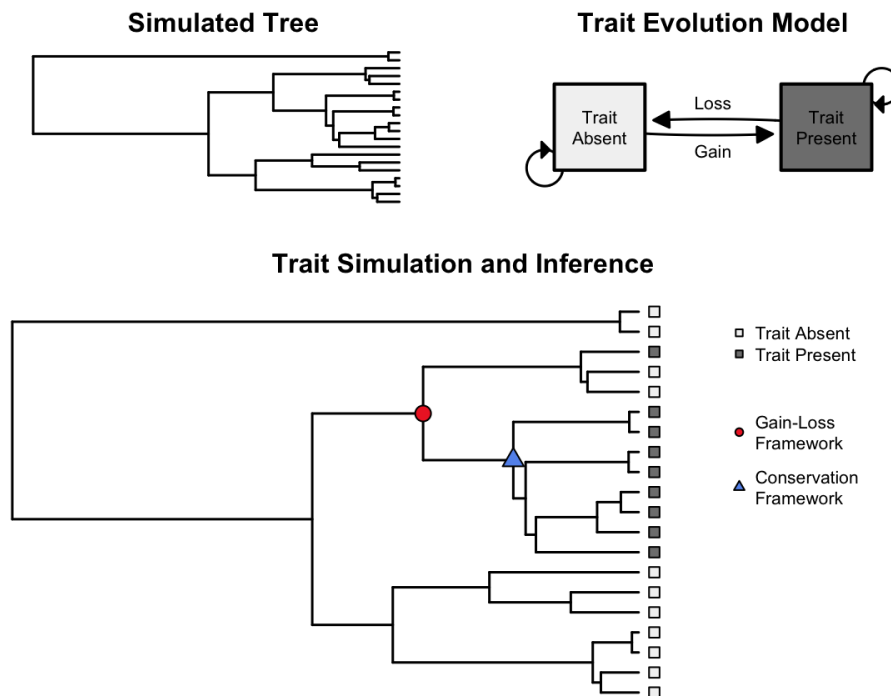


Figure S1: **Conceptual model used in our trait simulation.** Our trait simulations use phylogenetic trees which are the same size and age as the empirical data. We simulated discrete traits which evolved (*i.e.*, switched between states: present and absent) according to a Markov process. Based on simulated trait states at the tips, we used our frameworks to predict trait innovations across a range of transition rates which reflect the observed variation in the empirical data.

5 Methods

First, we simulated a phylogenetic tree equivalent to our empirical data. We simulated a phylogenetic tree using a Yule process (*i.e.* species death rate = 0) implemented using the `sim.bdtree()` function in the `geiger` R package¹. Our tree consisted of 2000 tips and was normalized to 4000 time steps. This was done to resemble the empirical data.

- 10 Next, we evolved traits based on an idealized system where traits evolve according to a discrete two-state Markov process. This Markov process is identical to the mathematics used in the gain-loss framework. For each simulation, the trait was initially absent and evolved as the tree was traversed. Briefly, the trait state at a parent node evolved along each daughter branch and the probability of each trait state at the daughter nodes is determined according to the following equation:

$$P(t) = u \cdot \exp(Qt), \quad (1)$$

$$Q = \begin{bmatrix} -x & x \\ y & -y \end{bmatrix}, \quad (2)$$

- 15 where u is the trait state at the parent node, t is the branch length, x is the transition rate from absent to present (*i.e.*, gain), and y is the transition rate from present to absent (*i.e.*, loss). The trait state at the daughter node is then determined given these probabilities, $P(t)$. This process was then repeated until the trait states at all nodes and tips were determined. We ran this simulation across a range of transition rates that spanned those estimated from the empirical data.
- 20 Last, we used the trait states at the tips to infer the evolutionary history of our simulation using the conservation and gain-loss frameworks. For the conservation framework, we identified the deepest nodes at which 90% of the tips had the trait. For the gain-loss framework, we used our ancestral state reconstruction method (*i.e.*, the Mk2 model) to identify the most likely node where the trait first appeared in a lineage. This was done by sampling the internal posterior probabilities and identifying nodes where the trait was
- 25 absent in the parent but present in the daughter according to a threshold of 0.5. Origins were determined by finding the deepest nodes at which this transition was found.

Results

Using these predictions, we compared the frameworks. First, we compared the predictions for origins. We found that, similar to the empirical data, the gain-loss framework predicts origins which are more ancestral
30 (Fig. S2).

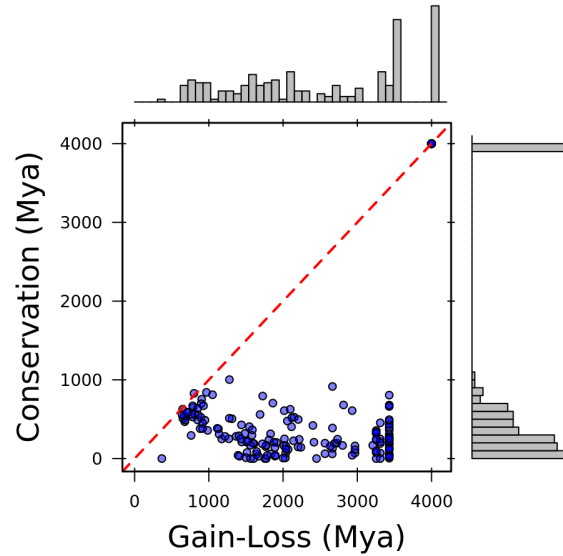


Figure S2: **Innovation predictions under the gain-loss and conservation frameworks.** Across simulation runs, we find that trait innovation predictions using the gain-loss framework are more ancestral than those using the conservation framework. These findings support the findings in the empirical data.

Next, we determined if the discrepancy between the two frameworks was related to the inferred loss rate. We found that as the gene loss rate increased, the agreement between the two frameworks decreased (Fig. S3). This finding supports the observations in the empirical data and suggests that there is a fundamental relationship between the gene loss rate and the difference between where a gene originated and the node at which it is still conserved.

35

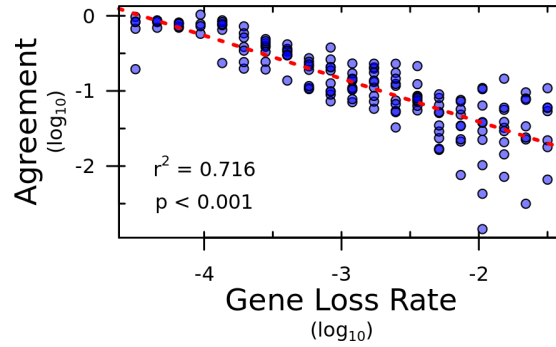


Figure S3: Relationship between gene loss rate and framework agreement. The negative relationship between framework agreement and loss rate suggests that predictions based on the conservation and gain-loss frameworks match when loss rates are low ($< 10^{-4}$), but as rates increase the frameworks support different conclusions. Under these scenarios, the conservation framework predicts more recent innovations.

Last, to make sure that our inferences were not due to issues associated with our ancestral state reconstruction method or simulation model design, we tested our ability to estimate the correct transition rates (*i.e.*, gain and loss rates). In previous studies, it has been shown that ancestral state reconstructions may produce biased estimates under certain conditions². However, we found that across the range of transition rates used in this study we were able to accurately estimate the rates used in our simulation (Fig. S4). For the majority of parameter combinations, estimates matched the true parameters. In some cases, however, we greatly underestimated the parameter. These cases appear to be situations where at least one of the two transition rates was very low ($< 10^{-5}$). This result is likely due to a limited opportunity to observe a switching event given the size of tree. However, these rates were very uncommon in the empirical data (See Fig. 2).

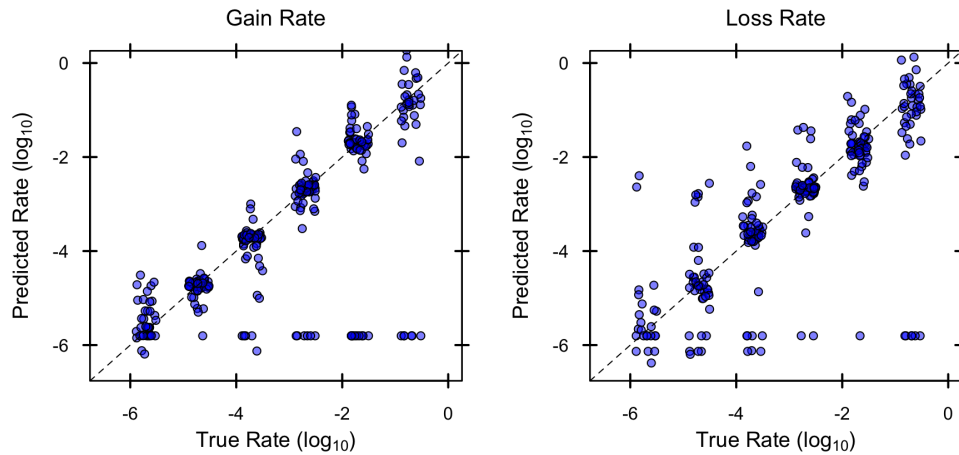


Figure S4: Ability to predict the true rate parameters given our simulation model and ancestral state reconstruction method. Across the transition rates in the empirical data, we could accurately predicted the true rates used in our simulations. In some cases (*i.e.*, when one rate was $< 10^{-5}$), we underestimated the true rates, but these rates were very uncommon in the empirical data (See Fig. 2).

References

- [1] Harmon, L., Weir, J., Brock, C., R.E., G. & W., C. Geiger: investigating evolutionary radiations. *Bioinformatics* **24**, 129–131 (2008).
- [2] Goldberg, E. E. & Igić, B. On phylogenetic tests of irreversible evolution. *Evolution* **62**, 2727–2741 (2008).

50