Ecological Insights from the Evolutionary History of Microbial Innovations: Supplemental

Mario E. Muscarella[a,1] **and James P. O'Dwyer**[a,b]

[a]Department of Plant Biology, University of Illinois; [b]Carl R. Woese Institute for Genomic Biology, University of Illinois

Simulation Model

Our emperical results suggest that that observed discrepency between the conservation and gain-loss frameworks may be related to the loss rate infered under the gain-loss framework. To test this prediction and estimate the upper bound of accurancy for our inferences, we used an idealized simulation model (Fig. 1).

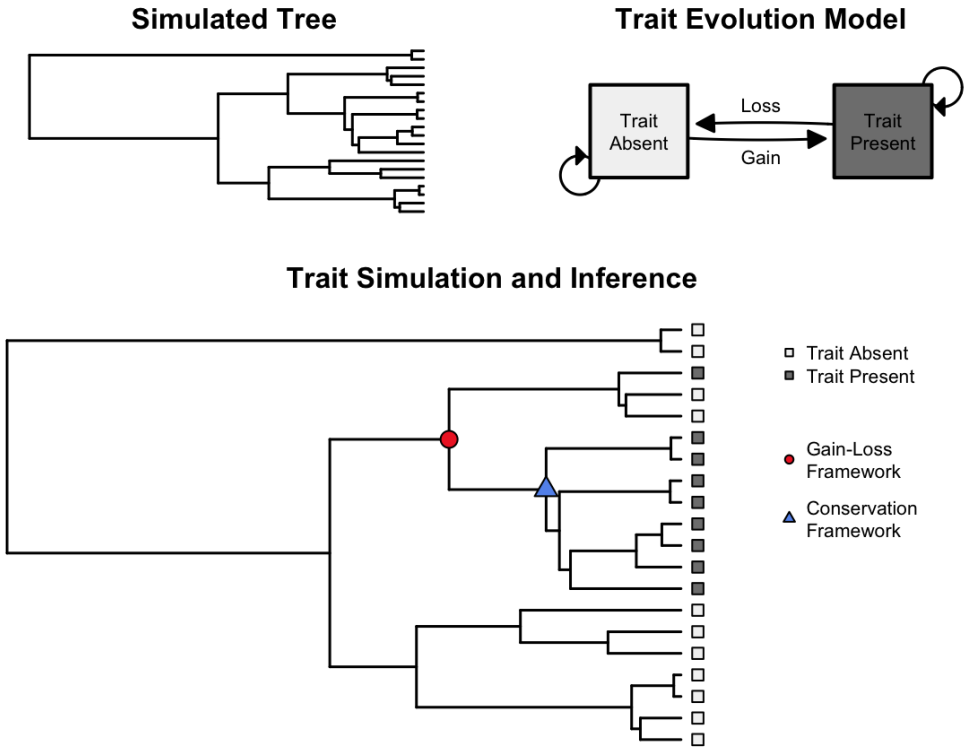

Fig. 1. Conceptual model used in our trait simulation. Our trait simulations use phylogenetic trees which are the same size and age as the empirical data. We simulated discrete traits which evolved (*i.e.*, switched between states: present and absent) according to a two-state Markov process. Based on 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.

Methods

First, we simulated a phylogenetic tree equivelent to our emperical 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 package (1). Our tree consisted of 2000 tips and was normalized to 4000 time steps. This was done to resemble the emperical data.

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. Therefore, we assume that diversification rates and trait-state are independent. For each simulation, the trait was initally 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:

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

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

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 emperical data.

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 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 emperical data, the gain-loss framework predicts origins which are more ancestral (Fig. 2).

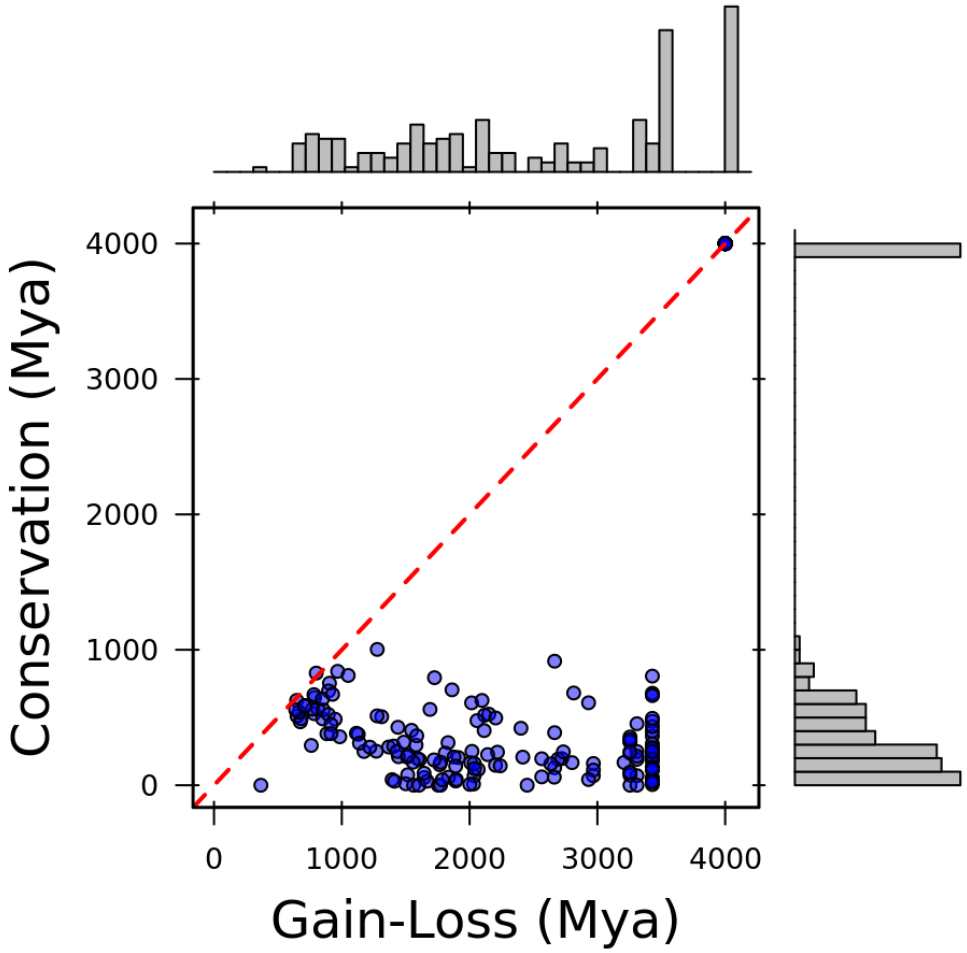

Fig. 2. Innovation predictions under the gain-loss and conservation frameworks. Across simulation runs, we find that trait innovation predictions using the gain-loss framework are more ancestral than those using the conservation framework. These findings support the findings in the emperical data.

Next, we determined if the discrepency between the two frameworks was related to the infered loss rate. We found that as the gene loss rate increased, the agreement between the two frameworks decreased (Fig. 3). This finding supports the observations in the emperical 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.

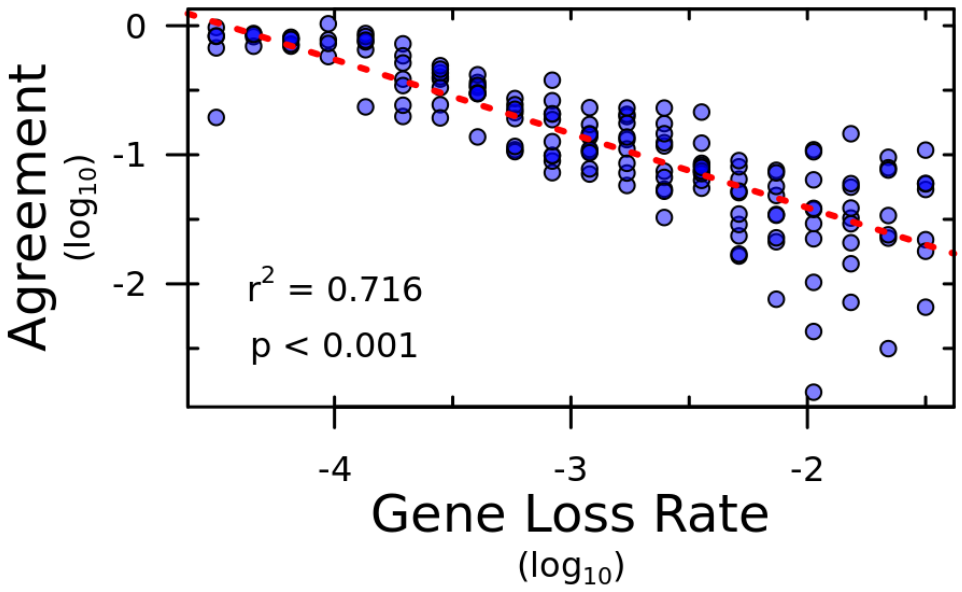

Fig. 3. Relationship between gene loss rate and framework agreement. The negative relationshp between framework agreement and loss rate suggests that predictions based on the conservation and gain-loss frameworks match when loss rates are low ($< 10^{-4}$), but as rates increase the frameworks support 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 conditions (2). 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. 4). For the majority of parameter combinations, estimates matched the true parameters. In some cases, however, we greatly underestimted the parameter. These cased 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 emperical data (See Fig. 2).

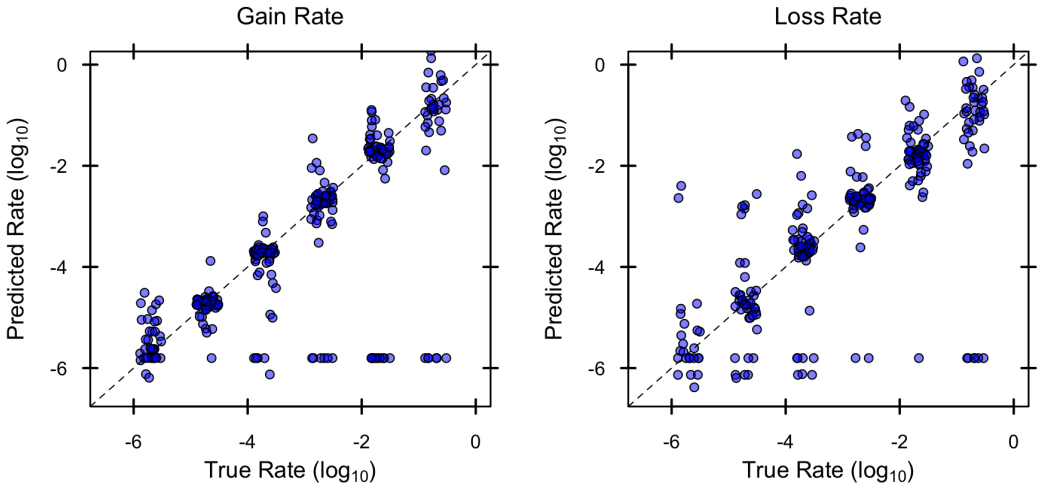

Fig. 4. Ability to predict the true rate parameters given our simulation model and ancestral state reconstruction method. Across the transition rates in the empirical data, we could accurately predicted the true rates used in our simulations. In some cases (*i.e.*, when one rate was $< 10^{-5}$), we underestimated the true rates, but these rates were very uncommon in the emperical data (See Fig. 2).

Gene Loss Rates

To further explore the various in loss rate within gene pathway groups, we seperated genes using a mixture of KEGG BRITE modules and gene operon. We first explored genes related the energy metabolism because the distribution appears to be bimodal. We split the genes based on BRITE module level C which split the genes into: ATP synthesis, methane metabolism, nitrogen metabolism, and sulfur metabolism. We plotted each as an independent distribution. We found that the rates associated with ATP synthesis and methane metabolism are lower than those for nitrogen and sulfur metabolism (Fig. 5).

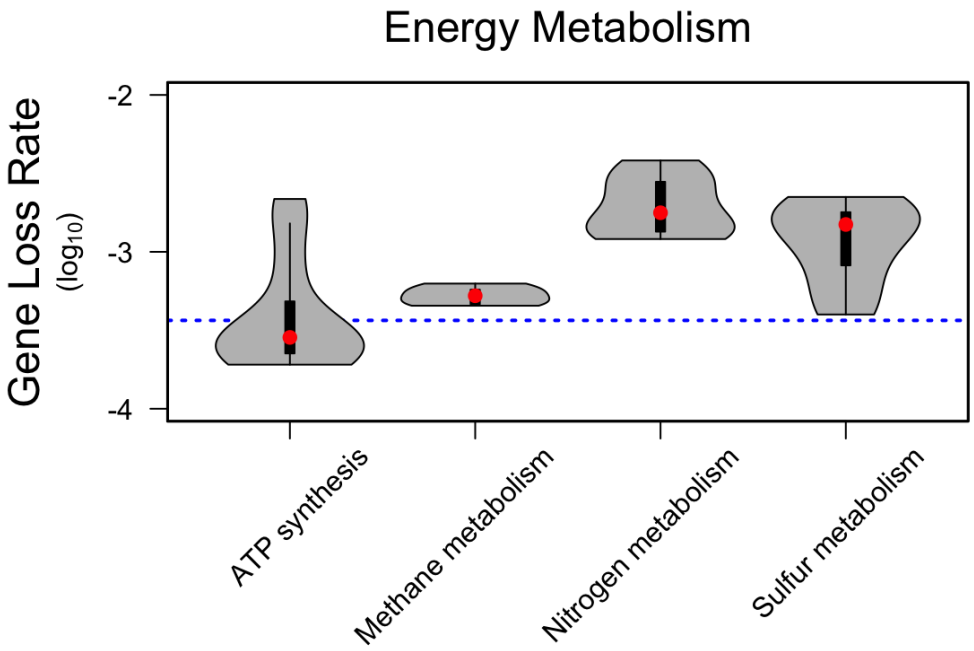

Fig. 5. Variation in gene loss rates for energy matabolism genes. When energy metabolism genes are broken into BRITE modules we find that the bimodal distribution can be explained by the different types of energy metabolism genes. Genes related to ATP synthesis and methane metabolism have lower loss rates than genes related to nitrogen and sulfur metabolism. The blue line represents the median loss rate across all genes. The red dots represent the median loss rate for the different modules of energy metabolism genes genes.

These results suggest that that ecologiaclly distinct traits have different gene loss rates. As such, those traits with low loss rates will show strong concordence between trait innovation and contemporary trait conservation. For these traits, conservation is a good predictor of current trait presence – though quantitative trait variation may still be substantial since we are only using trait presence–absence. On the other hand, traits belonging to ecological functions with elevated loss rates will show strong disagreement between innovation and contemporary trait conservation. For these traits, conservation is not an accurate predictor of trait history and traits are likely highly dispersed throughout the tree of life. This pattern can be seen for the nitrogen metabolism gene *nirK* as shown in Fig 1. While current data show conserved nodes for these genes, due to the high loss rates it will be uncertain if these nodes represent true evoutionary conservation or just a lack of data. Furthermore, the expectations across conserved nodes will be uncertain. As such, there are not well established expectations for phylogenetic trait signal across these nodes.

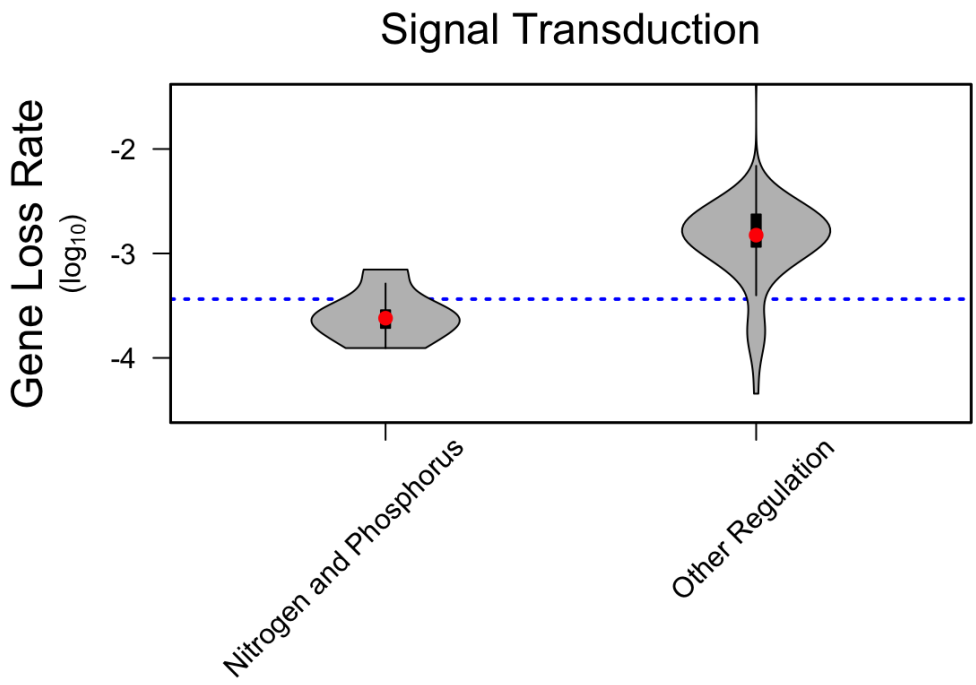

Fig. 6. Variation in gene loss rates for signal transduction genes. When signal transduction genes are broken into finer resolution groups based on ecological function we find that the tail of the distribution can be explained by the different types of genes. The genes are the tail of this distribution are primarily relate to maintaining cellular levels of nitrogen and phosphorus which have a lower loss rate than other signal transduction genes. The blue line represents the median loss rate across all genes. The red dots represent the median loss rate for the two groups of genes.

Bibliography

1. Harmon L, Weir J, Brock C, R.E. G, W. C (2008) Geiger: investigating evolutionary radiations. *Bioinformatics* 24:129–131.
2. Goldberg EE, Igić B (2008) On phylogenetic tests of irreversible evolution. *Evolution* 62(11):2727–2741.