Ecological Insights from the Evolutionary History of Microbial Innovations

Mario E. Muscarella[1,*] & James P. O'Dwyer[1,2]

[1]Department of Plant Biology, University of Illinois

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

* corresponding author: mmuscar@illinois.edu

Bacteria and Archaea represent the base of the evolutionary tree of life and contain the vast majority of phylogenetic and functional diversity[1,2]. These microorganisms and their traits directly impact ecosystems and human health[3,4]. As such, a focus on functional traits has become increasingly common in microbial ecology and these trait-based approaches have the potential to link microbial communities and their ecological function[5,6]. But what is missing is how, why, and in what order microorganisms acquired the traits we observe in the present day. These are important questions because they relate to the evolution, selective advantage, and trait similarity of extant organisms. Here, we reconstruct the evolutionary history of microbial traits using genomic data. We use the geological timeline and physiological expectations to provide independent evidence in support of this evolutionary history. Finally, we show that gene transition rates can be used to make predictions about the size and type of genes in a genome: generalist genomes comprise many labile genes while specialist genomes comprise more highly conserved functional genes. Our results provide a framework for understanding the evolutionary history of extant microorganisms, and provide insights into the evolution, selective advantage, and phylogenetic patterns of microbial traits. We anticipate that our work will improve our understanding of microbial trait variation and help identify microbial functional groups. In doing so, the evolutionary history of microbial traits will shed new light on our understanding of microbial communities in environmental and human ecosystems.

Global estimates predict upwards of 10^{12} bacterial and archaeal species[7,8]; however, the origin and maintenance of traits that define their phenotypes remains largely unknown. Trait evolutionary history, which captures these processes, is the foundation describing differences between species and links evolutionary relatedness and ecological function[9]. The true evolutionary history is unknown, but observed traits provide a window into evolutionary processes through their distributions across phylogenic trees. The simplest model would be that traits are conserved and therefore maintained by descendants. As such, taxonomy and/or phylogeny could be used as a trait proxy because specific traits would be conserved within distinct groups[10,11]. Here, we refer to this as the *conservation framework*. This framework is common in microbial ecology[5]. It has been used to predict traits[10], identify functional groups[12], and is often used to describing differences between communities based on indicator taxa. A more complex model would be that traits can be gained through innovation but also lost by descendants due to changes in selection and environmental context[13]. In this model, the dynamics of trait gain and loss can be represented as a stochastic process[14,15]. Here, we refer to this as the *gain-loss framework*. This framework is a common way to infer ancestral states[16] and to predict traits for unknown organisms[17].

These frameworks differ in the trait history they predict, but the differences provide an opportunity to understand trait innovation and selection. For example, the conservation framework assumes that closely related organisms have conserved traits and functions[18,19]. This assumption leverages the hierarchical evolutionary ancestry of organisms and the idea that related organisms should resemble each other[20,21]. In contrast, the gain-loss framework assumes that both trait gain and loss are common evolutionary mechanisms that can confer fitness advantages[22,23,24]. While there has not been a wide-scale quantitative comparison of these frameworks, there is no reason to expect one framework across all traits. For example, while we know there is a basic set of essential genes required for cellular function[25], high levels of genome reduction have been documented[26]. As such, it may be more informative to compare predictions across traits. For example, when the frameworks agree, it would suggest that traits are essential and thus maintained by descendants once they originate. In contrast, if the frameworks disagree then it would suggest that traits are less essential and can be lost by descendants. These traits would provide mechanisms for evolutionary diversification between related species.

In this study, we use 3179 bacterial and archaeal genomes to explore trait innovation based on individual genes. In total, 2950 orthologous genes were associated with the genome collection. Genes represent the raw genetic information underlying traits; therefore, while not phenotypic traits themselves, genes can be used to understand trait distributions. Using these genomes, we describe microbial innovations using the frameworks outlined above and illustrated in Fig. 1. Specifically, the conservation framework identifies phylogenetic nodes where 90% of the descendent genomes contain a gene of interest[18]. In contrast, the gain-loss framework estimates the nodes where traits first arose using a two-state Markov process, allowing for traits to be both acquired and lost over time[14]. Using these predictions, we compare the inferred innovation dates across genes to understand the evolutionary history and to make predictions regarding trait selection and genome composition.

Across genes, we find that predictions under the gain-loss framework are more ancestral than those for the conservation framework (Fig. 2A). However, there is a large range in the discrepancy (Fig. 2B). This suggests that the frameworks are part of a spectrum describing gene evolutionary history. At one end, genes are essential and highly conserved after they originate. On the other end, genes are non-essential and can be easily lost by descendants. Our results suggest that the essential to non-essential gradient is related to inferred gene loss rates (Fig. 2C). To confirm this possible mechanism, we simulated an idealized system where the true dynamics are controlled by a Markov process (See Supplemental). In this simulation, we used a tree with similar size and age, and a range of transition parameters which were similar to the inferred rates. In addition, our simulation tested the prediction accuracy across the range of transition rates we observed. Our simulation confirms that the observed discrepancy can be linked to high loss rates, and provides further evidence to suggest that the conservation framework is a special case within the gain-loss framework where loss rates are extremely low. Our findings also suggest that loss rates which lead to discrepancy are common across most genes (85%), which is consistent with studies showing high levels of genome reduction throughout evolutionary time[26]. Therefore, our findings suggest that the inferred gene loss rate is a quantitative measure of gene evolutionary lability, and can be used as a proxy for how essential a gene is for cellular function.

To explore the evolutionary lability predictions, we compared the predictions for specific genes. For example, there are about 290 genes with a strong agreement between the frameworks (Fig. 2B). Genes predicted to originate at the base of the tree are involved in cellular processes such as growth, information processing, and central metabolism which are essential for all organisms. Likewise, genes involved with oxygenic photosynthesis are known to be essential for specific groups of microorganisms (*i.e.*, cyanobacteria). We expect these genes to be essential, and indeed they have low loss rates and strong agreement between the frameworks. In contrast, there are >2500 genes with disagreement between the frameworks. The majority of these genes (>70%) are involved with cellular metabolism and 17% are involved with environmental information processing. While these genes are not needed for core cellular functions (*e.g.*, DNA replication, cell division), they are used to acquire resources and contend with environmental variation. We expect these genes to be more evolutionarily labile, and we find that they have high loss rates and low agreement between the frameworks. Therefore, these genes may represent ways organisms experiment with new traits and diversify into new ecological niches.

To explore the predictions regarding the origin of innovations, we compared the predictions for genes within broad pathways. Pathway characteristics (*e.g.*, required for cell division) and Earth's geological history can be used to provide expectations for our inferences. For example, while some cellular processes are required by all organisms and thus show strong agreement at the base of the tree (*e.g.*, *fitZ* – encodes the cell division ring) other processes range from group specific (e.g., *mutH* – encodes a sequence specific endonuclease) to evolutionarily labile (*e.g.*, *solR* – a quorum-sensing system regulator) (Fig. 3A). Likewise, we can leverage Earth's geological history as independent evidence. For example, current estimates suggest that life originated during the

Hadean (> 4000 Mya)[27], and that chemoautotrophic organisms dominated during the Archaean (4000 – 2500 Mya)[28]. During the Archaean anoxygenic photosynthesis originated[29] and towards the end of the eon oxygenic photosynthesis originated thus ensuing the Great Oxidation Event[28]. This series of geological events qualitatively supports the ordering of our predictions (Fig. 3B). Furthermore, the discrepancy between the frameworks suggests that while oxygenic photosynthesis is essential for a specific group (*i.e.*, cyanobacteria), anoxygenic photosynthesis is highly labile and often lost by descendants. Together, our results qualitatively recapitulate the evolutionary history of microbial traits and further suggest that the differences between frameworks is a signal of evolutionary lability.

Documenting the evolutionary history of traits also allows us to gain insights into the ecology and evolution of organisms. For example, genome size and the number of metabolic pathways may distinguish generalists and specialist organisms[30]. Furthermore, gene transition rates may provide evidence into evolutionary strategies and high rates may be a signature of evolutionary diversification. Across genomes, we find a strong positive linear relationship between the estimated gene loss rates and genome size (Fig. 4). These findings suggest that larger genomes contain genes which, on average, have higher loss rates and are thus more evolutionarily labile (Fig. 2C). Therefore, if genome size is a signature of the generalist–specialist gradient, then generalists contain more evolutionarily labile genes and specialists contain more physiological essential genes. Furthermore, while generalists maintain more genes, they are not hoarding traits but rather experimenting with new ecological functions. These findings suggest that generalists are undergoing ecological diversification through a process of trait experimentation while specialists are relying on conserved functional genes.

In this study, we cataloged the evolutionary history of microbial traits. We explored two frameworks which differ in their assumptions. We demonstrated that the framework which included loss met expectations for microbial evolutionary history. Last, we found that organisms with large genomes contain genes with higher loss rates. We interpret the estimated loss rate as a measure of trait necessity and therefore needed in a generalizable evolutionary framework. As such, the conservation framework is best when used as a heuristic to group extant organisms without direct evolutionary implications or assumptions about trait loss. While the gain-loss framework lacks mechanisms such as horizontal gene transfer, varying rates, and the non-independence of trait gain and loss, it qualitatively recapitulates the evolutionary history of major metabolic processes. In addition, while some traits may confer changes in diversification rates[31], it is implausible that these would be sustained indefinitely. Together, our results provide a tractable framework for understanding microbial evolutionary history and describing the evolutionary underpinnings for ecological differences.

Methods

Genomes were downloaded from the Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) data warehouse. Briefly, we downloaded the full list of publicly available bacterial and archaeal genomes (accessed April 2017, Supplemental Table 1). Using the JGI project ID, we searched the JGI project status page for the database name and the names of any larger project databases. Using the JGI Genome Portal API, we then searched for project archives and downloaded the most recent complete archive. We extracted each archive and only included those representing genomes with KEGG annotations and high quality 16S rRNA gene sequences (> 1200 bp). The 16S rRNA gene sequences were identified by searching the annotation file (*.gff) for entries based on the following search criteria: ‘rRNA.*product=16S‘. We parsed the entries for annotated gene IDs and searched the associated sequence file (*.fna) for the sequence entry. The KEGG KO annotations were found in the KO table file associated with each genome (*.ko.tab.txt). We used KEGG annotations because they are hierarchically organized into pathways and modules[32]. In total, 3179 genomes met all the criteria to be included in our study and together these genomes contained 2950 orthologous genes based on KEGG annotations. Database parsing and downloading was done using a custom automated script which retrieved the required information and downloaded the genomes using the JGI API. Genome parsing was done using a mixture of custom bash scripts and code implemented in R[33].

Using the 16S rRNA genes, we created a representative phylogenetic tree. We aligned the 16S rRNA genes based on the GreenGenes reference phylogeny (v. 13.8.99) using mothur[34]. We only included sequences which aligned to the reference. We used FastTree to generate a phylogenetic tree assuming the general time reversible model of nucleotide evolution[35]. We applied midpoint rooting to our tree and used treePL to estimate divergence times[36]. We standardized the tree by setting the root at 4000 ± 200 Mya[27], estimates at this date should be regarded as evolving prior to the bacterial–archaeal divergence. To prevent bias when comparing predictions to geological events, we did not internally calibrate our tree. Therefore, the dates inferred in this study should only be used as qualitative estimates. To check the accuracy of our tree reconstruction, we compared taxonomic assignments with tree topology.

We used KEGG annotations to infer the evolutionary history of traits. While genes do not represent phenotypic traits, they represent the genomic underpinning for traits and provide a standardized method to compare organisms. We treated genes as discrete traits based on presence-absence. We then inferred the evolutionary history using our proposed frameworks (Fig. 1). Under the conservation framework, we identified the nodes where 90% of the downstream genomes contained the gene of interest[18]. Under the gain-loss framework, we fit a continuous time discrete two-state Markov model to the observed trait states using maximum likelihood estimation[16]. This is the commonly used Mk2 model and we used the joint likelihood for maximum likelihood estimation. Other models of ancestral state reconstruction exist, including those which allow for state dependent diversification rates[31], but we assumed that changes in diversification rate would not be sustained at the evolutionary timescale of our tree. Using the inferred rate parameters, we estimated the probability of trait states at each node using the posterior probabilities at each internal node. We identified the trait *innovation* as the first node at which the trait most likely went from absent to present in a lineage using a posterior threshold of 0.5. Both methods were implemented in R using code adapted from the custom *ConsenTrait* function[18] and the *fitMK* function from the *phytools* R package[37] in addition to custom scripts.

Acknowledgements. We thank K. Webster, S. Kembel, and J. Martiny for critical feedback on an earlier version of this manuscript. JOD acknowledges the Simons Foundation Grant #376199, McDonnell Foundation Grant #220020439, and NSF Grant #DEB1557192.

Competing interests statement. The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to JOD (email: jodwyer@illinois.edu) or MEM (email: mmuscar@illinois.edu).

References

[1] Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).

[2] Hug, L. A. *et al.* A new view of the tree and life's diversity. *Nature Microbiology* **1**, 16048 (2016).

[3] Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459**, 193–9 (2009).

[4] Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012). NIHMS150003.

[5] Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: a phylogenetic perspective. *Science* **350**, aac9323–aac9323 (2015).

[6] Krause, S. *et al.* Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Frontiers in microbiology* **5**, 251 (2014).

[7] Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* **113**, 1–6 (2016).

[8] Schloss, P. D., Girard, R. A., Martin, T., Edwards, J. & Thrash, J. C. Status of the archaeal and bacterial census: An update. *mBio* **7**, 1–10 (2016).

[9] Cavender-Bares, J., Kozak, K. H., Fine, P. V. A. & Kembel, S. W. The merging of community ecology and phylogenetic biology. *Ecology Letters* **12**, 693–715 (2009).

[10] Louca, S. *et al.* High taxonomic variability despite stable functional structure across microbial communities. *Nature Ecology & Evolution* **1**, 0015 (2016).

[11] Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology* **8**, 523–9 (2010).

[12] Amend, A. S. *et al.* Microbial response to simulated global change is phylogenetically conserved and linked with functional potential. *The ISME Journal* **10**, 109–118 (2016).

[13] Nilsson, A. I. *et al.* Bacterial genome size reduction by experimental evolution. *Proceedings of the National Academy of Sciences USA* **102**, 12112–12116 (2005).

[14] Pagel, M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology* **48**, 612–622 (1999).

[15] O'Meara, B. C. Evolutionary inferences from phylogenies: A review of methods. *Annual Review of Ecology, Evolution, and Systematics* **43**, 267–285 (2012).

[16] Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).

[17] Langille, M. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**, 814–21 (2013).

[18] Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal* **7**, 830–838 (2013).

[19] Goberna, M. & Verdú, M. Predicting microbial traits with phylogenies. *The ISME Journal* 959–967 (2015).

[20] Blomberg, S. P., Garland, T. & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).

[21] Wiens, J. J. & Graham, C. H. Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics* **36**, 519–539 (2005).

[22] Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: Evolution of dependencies through adaptive gene loss. *mBio* **3**, e00036–12 (2012).

[23] Wolf, Y. I. & Koonin, E. V. Genome reduction as the dominant mode of evolution. *BioEssays* **35**, 829–837 (2013).

[24] Albalat, R. & Cañestro, C. Evolution by gene loss. *Nature Reviews Genetics* **17**, 379–391 (2016).

[25] Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences* **103**, 425–430 (2006).

[26] David, L. A. & Alm, E. J. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93–96 (2011).

[27] Tashiro, T. *et al.* Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. *Nature* **549**, 516–518 (2017).

[28] Judson, O. P. The energy expansions of evolution. *Nature Ecology & Evolution* **1**, 0138 (2017).

[29] Canfield, D. The early history of atmospheric oxygen: Homage to Robert M. Garrels. *Annual Review of Earth and Planetary Sciences* **33**, 1–36 (2005).

[30] Livermore, J. A., Emrich, S. J., Tan, J. & Jones, S. E. Freshwater bacterial lifestyles inferred from comparative genomics. *Environmental Microbiology* (2013).

[31] Maddison, W. P., Midford, P. E., Otto, S. P. & Oakley, T. Estimating a binary character's effect on speciation and extinction. *Systematic Biology* **56**, 701–710 (2007).

[32] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic acids research* **32**, D277–80 (2004).

[33] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017). URL `https://www.R-project.org/`.

[34] Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537–7541 (2009).

[35] Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5** (2010).

[36] Smith, S. A. & O'Meara, B. C. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).

[37] Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).

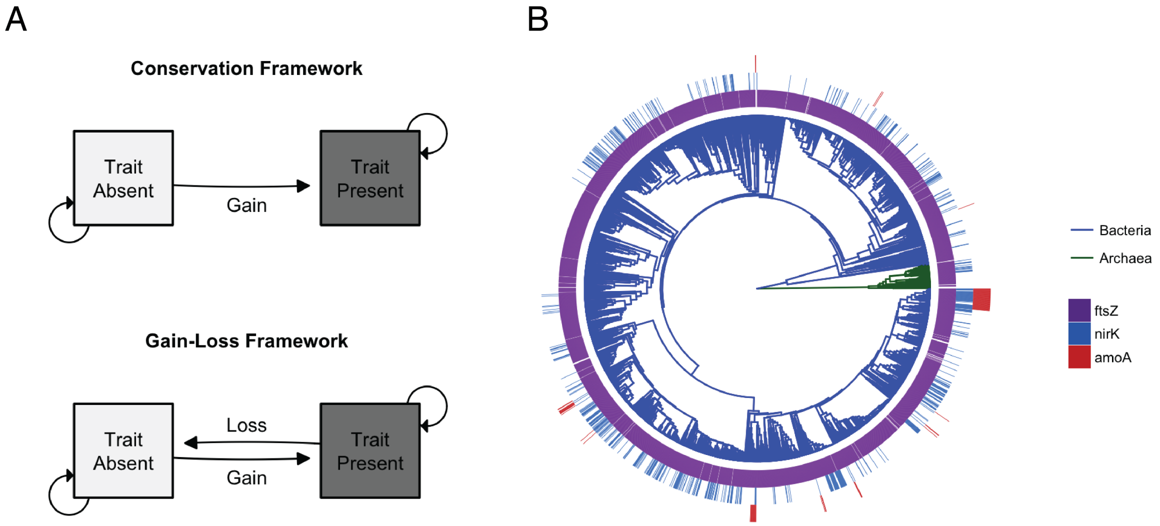

Figure 1: **Trait innovations in the microbial tree of life**. **A** The two frameworks for trait innovation. Under the conservation framework traits are gained but not lost. Absent traits are maintained with probability of *A* or gained at probability 1 - *A*. Once a trait is gained, it is maintained by descendants. Under the gain-loss framework traits are gained and lost. Absent traits are maintained with probability *A* or gained at probability 1 - *A*. Once a trait is gained, it is maintained at probability *B* or lost at probability 1 - *B*. **B** Phylogenetic tree of bacteria (n = 3108) and archaea (n = 71). The tree is based on the 16S rRNA sequences. Example genes are plotted at the tips. Some genes (e.g., *ftsZ*: cell division ring) are found in almost all genomes. Other genes (e.g., *nirK*) are abundant on the tree, but highly dispersed across taxa. Finally, some genes (e.g., *amoA*) are found in only a few groups.

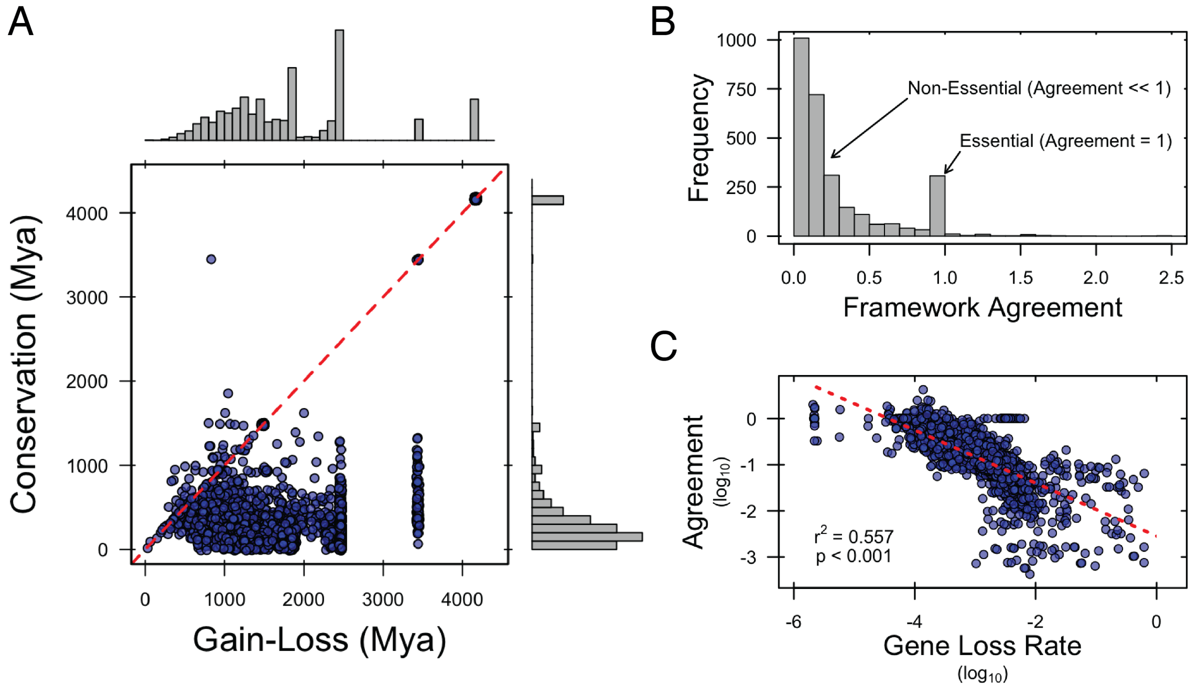

Figure 2: **Innovation predictions under the gain-loss and conservation frameworks**. **A: Comparison between frameworks.** Across genes, we find that gain-loss (GLF) predictions are more ancestral than those for conservation (CF). The CF identifies two major groups of genes. The first group is at the tree root contains 192 genes, including *ftsZ* (cell division), *dnaA* (DNA replication), *GPI* (glycolosis), as well as other genes for cell growth and division, translation, and oxidative phosphorylation. The second group includes 2666 genes at more recent nodes ($\sim$ 222 Mya). Most (95%) are associated with metabolism. In contrast, the GLF predicts more variation. First, a group of 229 genes is predicted at root, with a 99% overlap with the CF. We find peaks of innovation around 3400, 2400, and 1800 Mya. Finally, we find 1577 recent innovations ($\sim$ 1200 Mya). Similar to the CF, most are associated with metabolism. **B: Agreement between the frameworks.** The agreement between the frameworks was calculated as the CF estimate divided by GLF estimate. If the frameworks agree, then the agreement would be equal to 1. Across genes, we find 290 genes with agreement between the frameworks. This suggests that once these genes evolve, they are maintained by descendants. However, about 2500 genes have an agreement below 1. This suggests that these genes are not maintained after they originate. **C: Lower agreement is associated with higher gene loss rate.** Loss rates and agreement were $\log_{10}$ transformed and a linear regression model was used to determine the relationship between loss rate and agreement. A significant negative relationship was found ($F_{1,2796}$ = 3388, p < 0.001), suggesting that the difference in agreement between the frameworks is related to the loss rate inferred by ancestral state reconstruction.

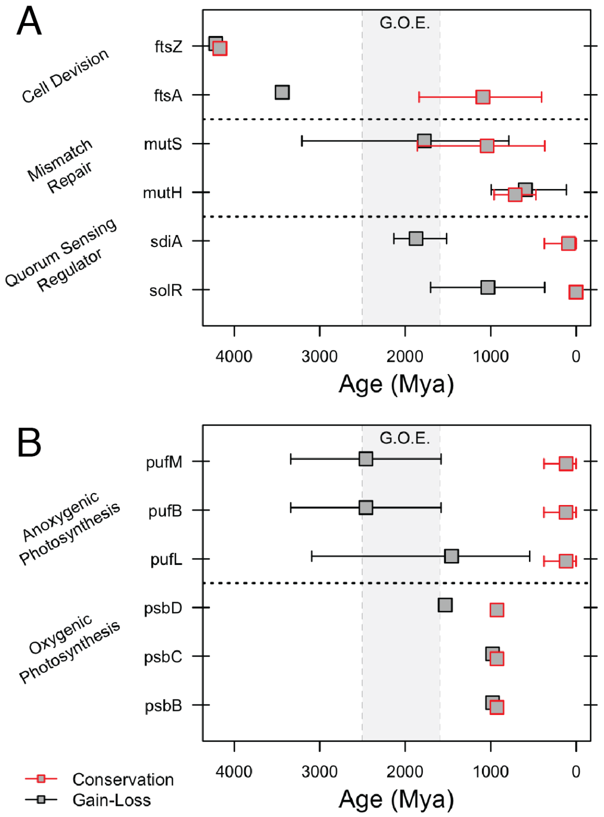

Figure 3: **Predicted innovation dates for genes in various pathways**. Example gene innovations predicted under the conservation (*red boxes*) and gain-loss (*gray boxes*) frameworks are shown for cellular processes (**A**) and photosynthesis (**B**). For some genes, there is a strong agreement between the two frameworks and the predictions overlap (*e.g.*, *ftsZ*). For other genes, there is little agreement between the frameworks. For example, genes related to anoxygenic photosynthesis show a much earlier origin under the gain-loss framework. In addition, predictions under the gain-loss framework are in better agreement with geological evidence for the appearance of specific metabolic processes. For example, the Great Oxidation Event is predicted to have taken place between 1800 and 2500 Mya[29]. Before this event, oxygen was a trace element in the atmosphere and anaerobic processes dominated. In general, the predictions from the gain-loss framework qualitatively recapitulate the ordering of these predictions.

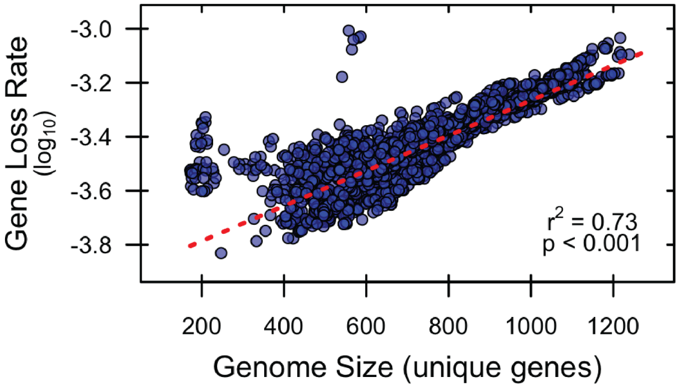

Figure 4: **Larger genomes contain genes with a higher gene loss rate.** There is a linear relationship between the number of unique annotated genes in a genome and the median gene loss rate of the genes contained in the genome. The loss rate of each gene is the maximum likelihood estimate for the gene switching rate based on the ancestral state reconstruction. Loss rates were $\log_{10}$ transformed and a linear regression model was used to determine the relationship between genome size and loss rate. A significant positive relationship was found ($F_{1,3123} = 8657$, $p < 0.001$).