UNIVERSITY OF CALIFORNIA

Los Angeles

# Reconstructing Phylogenetic Rings

# and Their Application

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Biomathematics

by

Joseph Robert Larsen

2016

ABSTRACT OF THE THESIS

Reconstructing Phylogenetic Rings

and Their Application

by

Joseph Robert Larsen

Master of Science in Biomathematics

University of California, Los Angeles, 2016

Professor Janet S. Sinsheimer, Chair

Phylogenetic rings represent evolution on a taxanomic scale through convergent ( genome fusion events) and divergent (phylogenetic tree-like events) gene flows. Rings have the potential to reconcile the inconsistent phylogenetic tree models that have been constructed from phenotypic evidence versus genotypic evidence. In order to exemplify this potential phylogenetic rings were applied here to investigate the origins of photosynthesis in the Proteobacteria. Another opportunity with-in phylogenetic ring research is developing a quantitative method to reconstruct the rings. The two methods explored here are *Occam's Ring*, the simplest ring reconstruction, and the Ring Identification for Non-Generalized Structures (R.I.N.G.S.) method, a more in-depth ring reconstruction based on quantitative methods. Phylogenetic rings have the potential to help resolve many of the conundrums in modeling evolution, which phylogenetic trees have been unable to address. This thesis is another step in solving these issues by further developing phylogenetic rings.

The thesis of Joseph Robert Larsen is approved.

James A. Lake

Kenneth L. Lange

Janet S. Sinsheimer, Committee Chair

University of California, Los Angeles

2016

This thesis is dedicated to my

father, Charles E. Larsen Jr.,

and mother, Harriet Larsen.

TABLE OF CONTENTS

# ACKNOWLEDGEMENTS

Chapter One is a version of James A. Lake, Joseph Larsen, Brooke Sarna, Rafael R. de la Haba, Yiyi Pu, HyunMin Koo, Jun Zhao, and Janet Sinsheimer. Rings reconcile genotypic and phenotypic evolution within the Proteobacteria. Genome Biol Evol. 2015. 7:3434-3442. doi: 10.1093/gbe/evv221. In-press.

(Early conceptualization, such as theory and initial draft, for chapter one's publication was contributed by Brooke Sarna with support from Jun Zhao, HyunMin Koo, Yiyi Pu, and Rafael R. de la Haba. Final development, such as figures and final draft, was handled by Joseph Larsen, Lead Researcher, and James A. Lake, P.I., with consultation from Janet Sinsheimer.)

## Introduction

Phylogenetic rings were first presented in 2004 by Maria Rivera and James Lake in the *Nature* article *The ring of life: evidence for a genome fusion origin of eukaryotes* (Rivera MC. and Lake JA. 2004). The motivation for the phylogenetic rings was to resolve how genome fusions and horizontal gene fusions were leading to inconsistent signals in gene sequence data thus confounding the reconstruction of the “Tree of Life” (Rivera MC. and Lake JA. 2004). This led to what is now considered the cornerstone of the Rings of Life Hypothesis, which is a contender for replacing the Three Domain Hypothesis (McInerny JO, O'Connell MJ., and Pisani D. 2014 and McInery JO, Pisani D., O'Connell MJ. 2015), known as the Rings of Life. This is presented in *Rings reconcile genotypic and phenotypic evolution within the Proteobacteria* by Lake *et al.* Although rings have contributed to the field of evolutionary biology in the last twelve years, ring analyses are still relatively new with immense potential to grow. Here I outline my contribution to the phylogenetic rings through assisting the study of their application towards the proteobacteria and developing and outlining novel reconstruction methods.

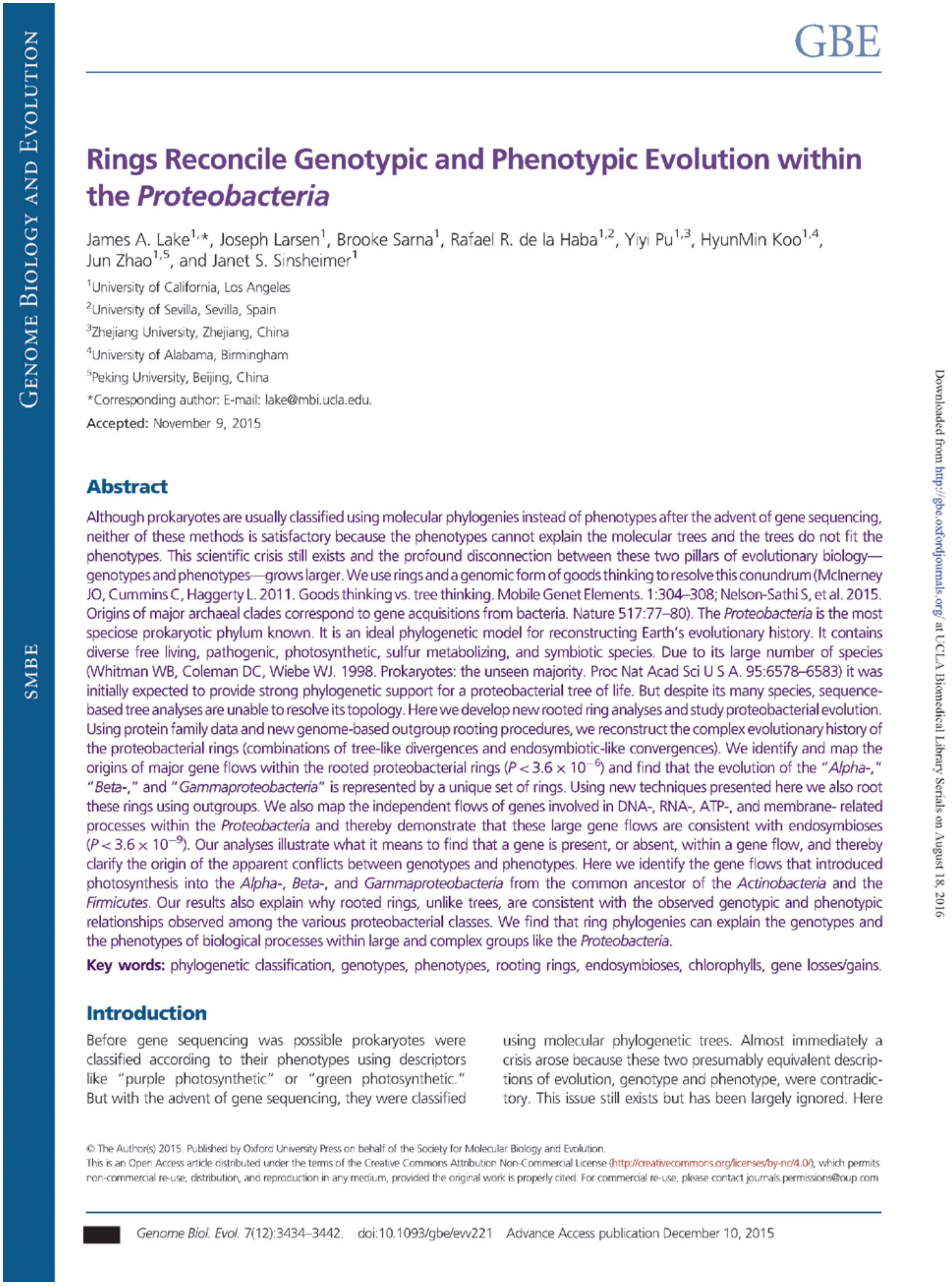
This thesis is a collection of my work towards furthering the ideals and depth of phylogenetic rings. The first chapter is the entire article *Rings reconcile genotypic and phenotypic evolution within the Proteobacteria* which was published in GBE, for which I am second author. The second chapter is the current manuscript for my first author publication titled *Reconstructing Phylogenetic Rings through Occam's Ring Structures and the R.I.N.G.S. Method* which is in its final stages before being submitted for publication. Finally, I take the time to speak of future directions for my work and what I believe my contribution lent to the field of phylogenetic rings. All in all, I have dedicated the last year of my life to better understanding and developing phylogenetic rings, which I have cherished and deeply appreciated every step of the way.

I first became involved in the study of phylogenetic rings through my work on the 2015 GBE article *Rings reconcile genotypic and phenotypic evolution within the Proteobacteria*. My contribution ranged from being involved in reshaping the paper, after initial reviews, and helping in the refocus of the article from Alpha-,Beta-, Delta-, and Gamma- proteobacteria to Alpha-,Beta-, and Gamma- proteobacteria. We found a new focus for the work in revealing possible origins for biological processes, such as photosynthesis. In order to identify how these biological processes flowed through the rings, I helped write a program to see what protein families were present in what flows of the ring and therefore where they were generated. Another contribution to this article was my assistance in determining the key assumption to phylogenetic rings, which was how rings are sensitive to gene gain but unaffected by gene loss. Finally, I drew all the figures found in the paper.. My efforts garnered me a second author position on this article.

My second major contribution to the phylogenetic rings is in outlining two reconstruction methods for the rings. Both of these methods utilize a newly defined phylogenetic tree called the Duplicate Taxa Tree (DTT) which is a phylogenetic tree that has at least one taxon that appears at least twice at the end of a terminal branch. This is possible due to the fact that every DTT transforms into a unique phylogenetic ring by combining branches with duplicate taxa, creating converging paths. The first of the two methods introduced in this work is called the *Occam's Ring* structure, which has the minimum number of paths and taxa on its respective DTT that still depicts the significant-flow paths for the set of taxa being investigated. The other method, named Ring Identification for Non-Generalized Structures (R.I.N.G.S.) method, is the first step toward a quantitative technique for reconstructing phylogenetic rings. This method assumes that a ring structure holds, contains a major assumption for a constant rate of evolution between taxa and makes a large approximation of applying a z-test to counts that have questionable independence. This work is an initial step and it has shown promise in determining a realistic phylogenetic ring structure by finding the significant-flow counts and number of each taxa on the ring's DTT. This work is the first ever quantitative approach for reconstructing phylogenetic rings. My contribution to this manuscript, that is in its final stages of drafting, earned me a first-authorship position.

# CHAPTER ONE

## Rings Reconcile Genotypic and Phenotypic Evolution within the *Proteobacteria*



GBE

# Rings Reconcile Genotypic and Phenotypic Evolution within the *Proteobacteria*

James A. Lake[1,*], Joseph Larsen[1], Brooke Sarna[1], Rafael R. de la Haba[1,2], Yiyi Pu[1,3], HyunMin Koo[1,4], Jun Zhao[1,5], and Janet S. Sinsheimer[1]

[1]University of California, Los Angeles

[2]University of Sevilla, Sevilla, Spain

[3]Zhejiang University, Zhejiang, China

[4]University of Alabama, Birmingham

[5]Peking University, Beijing, China

*Corresponding author: E-mail: lake@mbi.ucla.edu.

**Accepted:** November 9, 2015

## Abstract

Although prokaryotes are usually classified using molecular phylogenies instead of phenotypes after the advent of gene sequencing, neither of these methods is satisfactory because the phenotypes cannot explain the molecular trees and the trees do not fit the phenotypes. This scientific crisis still exists and the profound disconnection between these two pillars of evolutionary biology—genotypes and phenotypes—grows larger. We use rings and a genomic form of goods thinking to resolve this conundrum (McInerney JO, Cummins C, Haggerty L. 2011. Goods thinking vs. tree thinking. Mobile Genet Elements. 1:304–308; Nelson-Sathi S, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80). The *Proteobacteria* is the most speciose prokaryotic phylum known. It is an ideal phylogenetic model for reconstructing Earth's evolutionary history. It contains diverse free living, pathogenic, photosynthetic, sulfur metabolizing, and symbiotic species. Due to its large number of species (Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. Proc Nat Acad Sci U S A. 95:6578–6583) it was initially expected to provide strong phylogenetic support for a proteobacterial tree of life. But despite its many species, sequence-based tree analyses are unable to resolve its topology. Here we develop new rooted ring analyses and study proteobacterial evolution. Using protein family data and new genome-based outgroup rooting procedures, we reconstruct the complex evolutionary history of the proteobacterial rings (combinations of tree-like divergences and endosymbiotic-like convergences). We identify and map the origins of major gene flows within the rooted proteobacterial rings ($P < 3.6 \times 10^{-6}$) and find that the evolution of the "*Alpha*-," "*Beta*-," and "*Gammaproteobacteria*" is represented by a unique set of rings. Using new techniques presented here we also root these rings using outgroups. We also map the independent flows of genes involved in DNA-, RNA-, ATP-, and membrane- related processes within the *Proteobacteria* and thereby demonstrate that these large gene flows are consistent with endosymbioses ($P < 3.6 \times 10^{-9}$). Our analyses illustrate what it means to find that a gene is present, or absent, within a gene flow, and thereby clarify the origin of the apparent conflicts between genotypes and phenotypes. Here we identify the gene flows that introduced photosynthesis into the *Alpha*-, *Beta*-, and *Gammaproteobacteria* from the common ancestor of the *Actinobacteria* and the *Firmicutes*. Our results also explain why rooted rings, unlike trees, are consistent with the observed genotypic and phenotypic relationships observed among the various proteobacterial classes. We find that ring phylogenies can explain the genotypes and the phenotypes of biological processes within large and complex groups like the *Proteobacteria*.

**Key words:** phylogenetic classification, genotypes, phenotypes, rooting rings, endosymbioses, chlorophylls, gene losses/gains.

## Introduction

Before gene sequencing was possible prokaryotes were classified according to their phenotypes using descriptors like "purple photosynthetic" or "green photosynthetic." But with the advent of gene sequencing, they were classified using molecular phylogenetic trees. Almost immediately a crisis arose because these two presumably equivalent descriptions of evolution, genotype and phenotype, were contradictory. This issue still exists but has been largely ignored. Here

*Genome Biol. Evol.* 7(12):3434–3442. doi:10.1093/gbe/evv221 Advance Access publication December 10, 2015

we show that rings, unlike trees, allow one to see the connections between genotypes and phenotypes as alternative views of one evolutionary roadmap.

The importance of ring-like evolution has increasingly been recognized because, unlike molecular trees, rings can simultaneously accommodate two major modes of evolution: Tree-like bifurcations and endosymbiotic-like fusions. Thus rings can provide extremely general representations of evolutionary history. To illustrate their ability to provide a framework for understanding the evolution of life, consider the major gene flows present in the rings of life summarized in figure 1.

In the upper ring, the green path represents genes (Rivera and Lake 2004) flowing from the double membrane prokaryotes into the eukaryotes, shown in purple at the top of the rings. This flow includes the photosynthetic gene flow (Nelson-Sathi et al. 2012) that subsequently produced the chloroplasts, mitochondria, and possibly a host organism for the eukaryotic nucleus (Rivera and Lake 2004). The flow shown in magenta at the top right of the upper ring represents the informational gene flow into the Eocytes and the Eukaryotes (Lake et al. 1984; Lake 1988; Cox et al. 2008; Williams et al. 2013; McInerney et al. 2014), and the phototrophic gene flow shown in yellow represents the beginnings of light-driven ATP (AdenosineTriPhosphate) biosynthesis (Lake et al. 1985). At the bottom of figure 1 the root of the rings of life is represented by the three rooted rings shown in blue, yellow, orange, green, and white (Lake and Sinsheimer 2013).

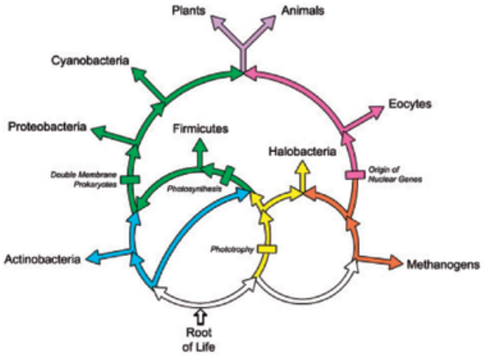


Fig. 1.—The rings of life are summarized in this figure. The eukaryotes, shown in purple at the top of the rings, are the result of the convergence of multiple gene flows. The *Proteobacteria* are present in the upper left green ring representing the flow from the double membrane prokaryotes into the eukaryotes (Lake 2009a, 2009b) that introduced mitochondria and chloroplasts into the eukaryotes (shown in purple). A second flow of genes into the eukaryotes is shown in cyan at the top right. It corresponds to the gene flow that transported informational genes into the eukaryotic nucleus from the eocytes. This gene flow includes many proteins and RNAs that are involved in fundamental cell/molecular processes that are unique to eukaryotes and eocytes. Examples include the eocyte/eukaryotic ribosomal apparatus for protein synthesis, the mechanisms for RNA transcription, and the unique chromatins that are used for the bundling of chromosomes into nucleosomes. The root of the rings of life is shown at the lower left of the figure. This set of rings leads to the *Actinobacteria*, to the *Firmicutes*, to the *Halobacteria*, and to the double membrane prokaryotes, including the *Proteobacteria*.

Note that the *Proteobacteria* emerge from a gene flow that is formed by the merger of two ancestral gene flows, the *Actinobacterial* (blue) and the *Firmicute* (green) gene flows. The rooted rings of life predict that the *Actinobacteria* and the *Firmicutes* fused to form the bouble membrane prokaryotes (Lake 2009a, 2009b). Because it was a fusion it defines "two" independent taxa, the *Actinobacteria* and the *Firmicutes*. Either of these two can be used to root the *Proteobacteria*. Thus gene presence–absence analyses of proteobacterial evolution that use either the *Actinobacteria* or the *Firmicutes* as immediate outgroups are predicted to support identical graphs. In contrast, the *Halobacteria* is a partial outgroup that is derived from two gene flows, only one of which flows into the *Proteobacteria*.

Reconstructing the evolution of the *Proteobacteria* is an important scientific goal in itself. Few other prokaryotic phyla, aside from the *Cyanobacteria*, have influenced Earth's evolution so dramatically. For example, the *Proteobacteria* impacted eukaryotic evolution by producing the ancestral mitochondrion, thought to have been an *Alphaproteobacterium*. Furthermore, the *Proteobacteria* is the most speciose prokaryotic phylum on Earth and 44% of all known prokaryotic species are contained within it (Whitman et al. 1998). It consists of diverse free living, pathogenic, photosynthetic, sulfur metabolizing, and symbiotic species. Its history can tell us much about the diversification of life on Earth.

## Proteobacterial History

Early classifications of photosynthetic prokaryotic diversity (Stanier et al. 1976) were based on prokaryotic phenotypes represented by processes such as photosynthesis and sulfur metabolisms. The two photosynthetic groups identified in these early studies were called the purple sulfur bacteria and the purple nonsulfur bacteria. The purple sulfur bacteria use sulfide or elemental sulfur as reducing agents and bacterial chlorophyll a for photosynthesis, whereas the purple nonsulfur bacteria use hydrogen and bacterial chlorophyll b for photosynthesis.

When the polymerase chain reaction made 16S ribosomal RNAs easy to sequence, new Proteobacterial classes were proposed on the basis of tree reconstructions and the purple bacteria were renamed the *Proteobacteria*. But the Proteobacterial classes did not fit the phenotypic classifications because some, but not all, *Alpha-*, *Beta-*, and

*Gammaproteobacteria* are photosynthetic. Even today the analyses of entire genomes can neither resolve the phylogenetic relationships among proteobacterial classes, nor can they explain the phylogenetic distributions of well-known proteobacterial phenotypes such as photosynthesis. For example, the group originally known as the purple sulfur bacteria is present in two distinct classes (the *Beta*- and *Gammaproteobacteria*), and the group originally known as the purple nonsulfur bacteria is present in a different set of classes, the *Alpha*- and *Betaproteobacteria*. These two phenotypic classifications clearly conflict with all possible trees, because the *Betaproteobacteria* contain both purple sulfur and purple nonsulfur bacteria.

But how and why this happened remained unknown. The initial optimism that genomics could pinpoint major events in the evolution of the *Proteobacteria* vanished when neither ribosomal RNA- nor whole genome- based trees could explain the mutually contradictory distributions of photosynthesis and bacterial chlorophylls within the *Proteobacteria*. Even with large numbers of proteobacterial species available for analysis, no statistically significant tree-like phylogenetic signals could relate the proteobacterial classes to each other (Lerat et al. 2004), and sophisticated tree reconstructions (Creevey et al. 2004) could only resolve the relationships "within" the proteobacterial classes located at the tips of trees. Some suggested that this might be due to lateral gene transfers (LGTs), "…there is too little phylogenetic signal to permit firm conclusions about the mode of inheritance. Although there is clearly a central tendency in this data set…lateral gene transfers cannot be ruled out" (Susko et al. 2006). Recently, a comprehensive study showed that highly asymmetric "…transfers from bacteria to archaea are more than fivefold more frequent than *vice versa*" (Nelson-Sathi et al. 2015). Others recognized this problem and referred to it as the "Tree of One Percent" (Dagan and Martin 2006). In another comprehensive analysis of 329 proteobacteria genomes, the *Gammaproteobacteria* were categorized as showing "…the most chameleon-like evolutionary characteristics" (Kloesges et al. 2011). New evidence for a large photosynthetic flow of more than a thousand genes (Nelson-Sathi et al. 2012) and for the related phototrophic flow (Lake et al. 1985; Lake and Sinsheimer 2013), however, suggested that it might be possible to reconstruct the flow of photosynthesis within the *Proteobacteria* (Archibald 2008).

Motivated to understand the evolutionary origin of these major conflicts in terms of known evolutionary processes, we asked whether rings could explain the differences between proteobacterial genotypes and phenotypes. Using genome and protein family presence/absence analyses (Lake 2009a, 2009b; Lake and Sinsheimer 2013) and by devising new methods to root rings we reconstruct the evolution of the *Alpha*-, *Beta*-, and *Gammaproteobacteria*.

## Results

### An Overview of the Proteobacterial Rings

Ring analyses (Lake 2009a, 2009b; Rivera and Lake 2004) have been used to reconstruct major evolutionary gene flows within the rings of life. Using new, but related, methods we reconstruct the rings describing the evolution of the *Alpha* (A)-, *Beta* (B)-, and *Gamma* (Γ)-*proteobacteria*.

In the overview of the rings shown in figure 2 (Lake and Sinsheimer 2013), the gene flow originating from the local root (shown by the yellow arrow at the bottom of the rings) first divides into a yellow gene flow (on the left) and an orange gene flow (on the right). The yellow gene flow then bifurcates to form the cyan and the magenta gene flows that lead to the *Alphaproteobacteria* and the *Betaproteobacteria*, respectively. Subsequently, these two gene flows converge and form the purple gene flow which then merges with the orange flow and they ultimately form the *Gammaproteobacteria*.

The presence–absence counts that accompany these flows are shown in table 1. The three largest gene flows, marked in red in table 1, correspond to the flows of 619 Pfams into the *Beta*- and *Gammaproteobacteria* (−,+,+); 389 Pfams into the *Alpha*- and *Gammaproteobacteria* (+,−,+), and 3511 Pfams into the *Alpha*-, *Beta*-, and *Gammaproteobacteria* (+,+,+). It should be noted that, similar to three taxon tree reconstructions, the counts for the +++, + − −, − + −, and − − + ring terms are phylogenetically uninformative. This is because all rooted trees and rings have roots, represented by the term +++, and because all rooted trees and rings have external branches (represented by the terms + − −, − + −, and − − +).

Thus when analyzing significant and nonsignificant patterns, only the patterns with two +'s are topologically informative. By using chi-squared probability ratios to evaluate whether 71 and 368, or 368 and 619 are drawn from the same normally distributed populations, we find that 368 and 619 are 1.33225 × $10^{30}$ times more likely to have been drawn

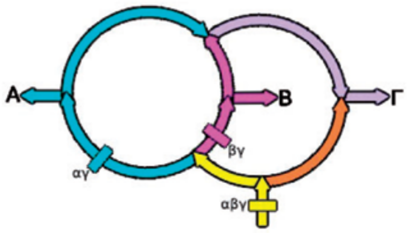


**Fig. 2.**—The gene flows representing the evolution of the A-, B-, and Γ- *Proteobacteria* are labeled and color coded. The start of the AΓ gene flow is marked by the cyan rectangle labeled αγ, the start of the ABΓ gene flow is marked by the yellow rectangle labeled αβγ, and the start of the BΓ gene flow is marked by the magenta rectangle labeled βγ.

from the same population than are 71 and 368. Hence the alpha–gamma (368) and beta–gamma (619) gene flows are inferred to be present in figure 2. The start sites for these two gene flows are labeled in figure 2.

## Rooting the Rings

When roots are known, ring reconstructions are simplified. Recently, indels (inserts/deletions in genes) were used to root the rings of life shown in figure 1 (Lake and Sinsheimer 2013). Because the most reliable genomic-based rooting information is often provided by indels and because those indels used for the rooting in figure 1 had extremely strong statistical support, this provided an opportunity to test whether our ring analyses can provide additional support for the rooted rings.

Our analyses utilize ring outgroup rooting, a new algorithm developed here, to further test the rooted proteobacterial rings. Ring outgroup rooting allows one to test whether potential roots are valid or not. Outgroups to the *Proteobacteria* were discovered when the root of life was localized "to a segment of the deepest ring ($P < 10^{-21}$ and $P < 10^{-194}$)" using indel rooting (Lake and Sinsheimer 2013). Based on this rooting we obtained evidence that two lineages (one from the *Actinobacteria* and the other from the *Firmicutes*) merged to form the gene flow leading to the *Proteobacteria* (fig. 1).

Because gene flows from the *Actinobacteria* and from the *Firmicutes* merge to form the stem lineage leading to the *Proteobacteria*, either can be used to root the proteobacterial rings. The merger of these two gene flows makes it possible to test whether the indel-based root of the proteobacterial rings will also be recovered from ring analyses. The rings shown in figure 1 predict that the *Halobacteria* cannot be used to root the *Proteobacteria* because only one of the gene flows leading to the *Halobacteria* (the yellow flow) directly connects with the *Proteobacteria*. Although the orange gene flow also enters the *Halobacteria* (shown in fig. 1), it does not flow into the *Proteobacteria* and hence cannot be used to root the proteobacterial rings. Although the *Halobacteria* is not a valid outgroup, it nevertheless serves as a negative control for our analyses.

## Rooting the Proteobacteria

To test whether the *Actinobacteria*, the *Firmicutes*, and the *Halobacteria* are outgroups to the *Proteobacteria*, we analyzed the relevant four-taxon Pfam presence/absence tables shown in table 2. Subtable 1, on the left, relates the *Proteobacteria* to the *Actinobacteria*; subtable 2, in the middle, relates the *Proteobacteria* to the *Firmicutes*; and subtable 3, on the right, relates the *Proteobacteria* to the *Halobacteria*. As in table 1, the statistically significant gene flows in table 2 are marked in red. Background gene levels, thought to be due to horizontal gene transfer (HGT)/lateral gene transfer (LGT), are identified by the largest gap between large and small gene flows using chi-squared analyses. Note that the same six significant gene flows (marked in red) are present when either the *Firmicutes* or the *Actinobacteria* are included in the analyses (subtables 1 and 2). This demonstrates that the *Firmicutes* and the *Actinobacteria* are immediate outgroups to the *Proteobacteria* because they have the same topological relationship to the proteobacterial rings. The probability that the same six signal patterns were chosen by chance from the set of 10 informative patterns for both the Firmicute- and the Actinobacterial outgroups is low ($P < .0048$ by the hypergeometric test), thus providing strong evidence that both outgroups have the same phylogenetic relationship to the *Proteobacteria*. In contrast, when the *Halobacteria* are included in the analyses only three of the six signal gene flows are present and there is no support for the *Halobacteria* having the same relationship to the *Proteobacteria* that was found for the *Actinobacteria* and the *Firmicutes* ($P < 0.923$, by the hypergeometric test). We conclude that the *Firmicutes* and the *Actinobacteria* are immediate outgroups to the *Proteobacteria* but that the *Halobacteria* is not an outgroup.

**Table 1**
Gene Presence/Absences

| A | B | Γ | Pfams |
|---|---|---|---|
| + | + | + | 3511 |
| + | + | − | 71 |
| + | − | + | 368 |
| − | + | + | 619 |

Note.—Significant Pfam flows are in red.

In contrast, because all three of the signals present within the *Halobacteria* in table 1 [(+,+,+), (+,−,+), (−,+,+)] are also the only signals present in table 2 [(+,+,+,−), (+,−,+,−), (−,+,+,−)], we conclude that this signal is generated solely by the proteobacterial rings and not from connections between the *Halobacteria* and the *Proteobacteria*. The observation that the findings presented in subtables 1 and 2 independently support the Firmicute/Actinobacterial fusion previously reported in the rooted rings of life (Lake and Sinsheimer 2013) is consistent with the *Firmicutes* and the *Actinobacteria* (but not the *Halobacteria*) being immediate outgroups to the *Proteobacteria*.

### *The Rooted Proteobacterial Rings*

Because the *Actinobacteria* and the *Firmicutes* are outgroups to the *Proteobacteria* in the rings of life in figure 1 (Lake and Sinsheimer 2013), this suggests that genes flow from the root defined by the *Actionbacteria* and the *Firmicutes* into the *Proteobacteria*. With this rooting information we can now formally test the evolutionary origins of the *Proteobacteria* within the rings of life.

The rooted proteobacterial rings reconstructed from the Firmicute and the Actinobacterial subtables are shown in

Genome Biology and Evolution SMBE

**Table 2**
Rooting the Proteobacterial Rings with Outgroups

| A | B | Γ | Ac | Pfams | A | B | Γ | Fi | Pfams | A | B | Γ | H | Pfams |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | + | + | – | 816 | + | + | + | – | 816 | + | + | + | – | 2388 |
| + | + | – | + | 30 | + | + | – | + | 35 | + | + | – | + | 9 |
| + | + | – | – | 41 | + | + | – | – | 36 | + | + | – | – | 62 |
| + | – | + | + | 211 | + | – | + | + | 238 | + | – | + | + | 60 |
| + | – | + | – | 157 | + | – | + | – | 130 | + | – | + | – | 308 |
| + | – | – | + | 52 | + | – | – | + | 63 | + | – | – | + | 21 |
| – | + | + | + | 241 | – | + | + | + | 328 | – | + | + | + | 47 |
| – | + | + | – | 378 | – | + | + | – | 291 | – | + | + | – | 572 |
| – | + | – | + | 41 | – | + | – | + | 65 | – | + | – | + | 8 |
| – | – | + | + | 233 | – | – | + | + | 361 | – | – | + | + | 66 |

Note.—The outgroups are as follows: *Actinobacteria*, $A_c$; *Firmicutes*, $F_i$; *Halobacteria*, *H*. Significant Pfam flows are in red.

figure 3. In table 2, these gene flows are highlighted in red and nonsignificant gene flows, consistent with the background of HGT/LGT, are black. Note that the three red gene flows present in table 1 [(+,+,+), (+,–,+), and (–,+,+)] are also present in all three subtables in table 2 [(+,+,+,–), (+,–,+,–), and (–,+,+,–)]. Thus the αγ, βγ, and αβγ gene flow patterns that are explained by the graph in figure 2 are also present in the double rings structure at the top of figure 3. When either the *Actinobacteria* or the *Firmicutes* are included in the analyses, the top rings are connected to their Actinobacterial/Firmicute roots by the additional complex gene flows shown in gray. Accordingly, the data in table 2 are consistent with the proteobacterial rings and with the *Actinobacteria* and the *Firmicutes* being outgroups.

In contrast, there are no large gene flows in subtable 3 that directly connect the *Proteobacteria* to the *Halobacteria*, because all three informative patterns (those with at least two +s) which connect the *Halobacteria* with the *Proteobacteria* lack statistical support. Thus the *Actinobacteria* and the *Firmicutes* are outgroups in the ring sense, but the *Halobacteria* is not.

Although the details of the proteobacterial part of the ring shown in figure 3 are identical to those in figure 2, the deeper connections of the *Proteobacteria* to the *Firmicutes* and to the *Actinobacteria* involve additional gene flows. Those flows, shown in gray in figure 3, connect the *Alpha-*, *Beta-*, and *Gammaproteobacteria* to their *Firmicute* and *Actinobacterial* outgroups. Because the same six large gene flows, that is, the same connections, are present when either the *Firmicutes* or the *Actionbacteria* are used as outgroups, this further confirms by the hypergeometric test (population size = 10, successes in: A population = 6, sample size = 6, and successes in sample = 6, $P < 0.00477$) that they are sister outgroups as previously shown by indel rooting. Specifically, because *Firmicutes* and the *Actionobacteria* are supported by the same set of gene flows the graph representing the Firmicute data set shown in figure 3 must be the same as that representing the Actinobacterial data set with the *Firmicutes* replaced by the *Actinobacteria*.

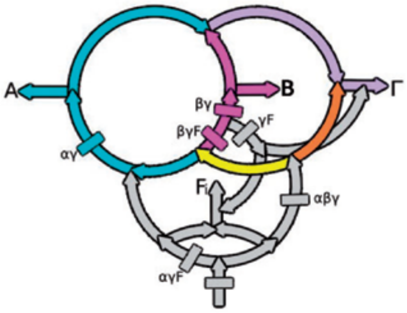


**Fig. 3.**—The deeper gene flows that connect the ABΓ *Proteobacterial* to their Firmicute/Actinobacterial outgroups are shown in grey. The start sites of gene flows introduced by outgroup rooting are marked in color. They are the αγF (grey), the βγF (magenta), and the γF (grey) gene flows. Identical rings and similar gene flow counts are produced when these rings are rooted using the *Actinobacteria* as the outgroup, and the corresponding gene flows are labeled the αγA, βγA, and γA gene flows, respectively.

The outgroups define the directions of gene flows as follows. Genes flow from the root at the bottom of figure 3 and subsequently bifurcate. The flows on the left and the right then divide a second time so that one path leads to the outgroup and the other to the *Proteobacteria*. Note that the directions of the arrows indicate the flow of genes and of time. We interpret the gene flows shown in presence/absence tables 1 and 2 as representing gene gains, as discussed below.

## Detecting Gene Gains

We find that net gene gains can be reliably measured in presence–absence studies of large populations. We illustrate how this differential sensitivity to gene loss and gene gain

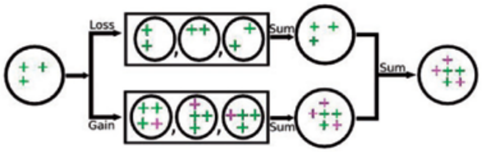


FIG. 4.—The differential effects of gene losses and gains on the measurement of presence/absence counts. The large circle at the left represents genes initially present within the founding gene flow. Over time genes will be lost from some members of the population as shown by genomes (circles) within the box at the top. Similarly, over time genes will be gained by other members of the population, as shown by genomes (circles) within the box at the bottom. The large circle on the top line (to the right of the box) represents the number of different genes present in all organisms that have "lost" genes. For large populations like the ones studied here, few, if any, genes will be lost from every single member of the population. In contrast, the large circle on the bottom line represents the number of different genes present in organisms that have "gained" genes (shown by red +'s). This sum will increase over time as new genes emerge, even if only a very small percent of individuals within the population carry new genes. When the gene inventory from cells with losses is added to that from cells with gains, the net change will be an "increase" in the number of novel genes within the population. Thus the totals calculated in presence/absence analyses represent new genes. This is also consistent with the results of our rooting analyses using Firmicutes and Actinobacteria. In addition, a background of HGT/LGT will introduce new genes over time; however, the numbers of genes introduced are small as estimated by the difference between the large statistical differences between the gene flow and the background counts.

arises when databases containing large numbers of individual organisms are studied. In figure 4, gene losses occurring within individual organisms (the circles in the upper box) are shown as missing +'s. Even though the loss of a particular gene from individual cells may be frequent, the elimination of that gene from an entire taxon is rare. It is because the gene must be lost from all individuals, which for even small populations rapidly becomes highly unlikely. Thus when genes are summed over large populations, as illustrated in the top box, it becomes highly unlikely for all of the organisms to have lost the same gene (upper row). In contrast, gene gain in even one organism, shown in red in the lower box, would be obvious when genes are tallied over all the members of the population (as in the lower right circle). When gene gains and losses are added together in presence/absence tables, the net result is that gene gains are detected whereas gene losses are hidden, as shown in the circle on the far right.

In summary, even though genes may be readily lost over time from individual organisms, it is extremely improbable for genes to be completely lost from large populations. In contrast, the gain of a single new gene by even one species can be detected when many taxa are sampled within a large population. For these reasons, we have used protein families (Pfams) for our analyses because, unlike genomes, Pfams can represent tens of thousands of species. This makes this database ideal for detecting gene gains, and for being insensitive to gene losses. Protein family PF00009 (GTP_EFTU), for example, contains 69,868 sequences from 24,054 discrete species. By using large databases the probability that ring analyses will detect new genes is exponentially increased, and the probability that genes which are easily lost will be scored as missing is greatly decreased. Consistent with these ring findings and with previous indel rooting studies (Lake and Sinsheimer 2013), our results provide strong support for the Rooted Proteobacterial Rings shown in figure 3.

### Identifying Endosymbiotic Flows within the Rings

Rings can simultaneously describe divergent and convergent evolution. Divergences are responsible for tree-like evolution and the underlying tree-like evolutionary mechanisms responsible for them are well known, but convergences are only beginning to be understood.

In rings convergences may be caused by endosymbioses or by HGT/LGT. In the past it has been difficult to distinguish between these two alternative modes of evolution (Doolittle 2007). Traditionally, evidence for endosymbioses has come from membrane systems and from phylogenetic trees reconstructed from organellar DNA. For example, the endosymbiotic origins of mitochondria and chloroplasts were initially based on the observations that those organelles 1) were surrounded by inner and outer membranes and 2) had gene sequences that differed from the nuclear genes of their host cells. Subsequently, analyses of mitochondrial and chloroplast DNA sequences showed that they were related to the *Alphaproteobacteria* and to the *Cyanobacteria*, respectively. Even the nucleus has been proposed to have endosymbiotic origins (Lake and Rivera 1994) in the *Eocyta* (Lake 1988; Cox et al. 2008; Williams et al. 2013; McInerney et al. 2014) and viable mechanisms have been suggested for its acquisition (Martin and Muller 1998).

Within prokaryotes, endosymbioses are much harder to identify because separate compartments for host and guest DNA are not normally present. However, precedents exist for prokaryotic endosymbioses. For example, some eukaryotes contain endosymbiotic prokaryotes (*Gammaproteobacteria*), which contain their own endosymbionts (*Betaproteobacteria*), much like a set of nesting Russian dolls (von Dohlen et al. 2001). Additionally, even the inner and outer membranes of the double membrane, that is, gram negative, prokaryotes may have been derived as the result of an endosymbiosis between two ancient prokaryotes, a Firmicute and an Actinobacteria (Lake 2009a, 2009b). But prokaryotic examples of endosymbiosis are relatively rare, so that new computational methods are needed to distinguish endosymbiotic gene flows from LGT/HGT.

Given the subjective aspects of interpreting membrane organization within prokaryotes, we present a genomic-based method for discriminating between endosymbioses and gene transfers. In these analyses, the functions of the genes being transferred provide a basis for distinguishing endosymbioses from gene transfers. The essence of the test lies in determining the functions of the genes being transferred. Horizontally/laterally transferred genes tend to have specialized functions. For example, organisms living in aquatic environments are more likely to exchange genes with other organisms living in that environment, and so on. In contrast, endosymbioses transfer entire cells complete with all the genes necessary to survive as free living entities. They pass on genes that are essential for fundamental life processes such as translation, replication, energy production, and cellular compartmentalization (Jain et al. 2003).

Here we use these fundamental properties to test whether gene flows within the proteobacterial rings are consistent with endosymbiotic transfers, or whether they are consistent with HGT/LGT. This is accomplished by operationally defining endosymbiosis as a process that can be recognized by the simultaneous transfer of statistically significant numbers of genes responsible for fundamental cellular processes. Our tests explicitly follow the paths of inheritance of genes involving DNA, RNA, ATP, and membranes. We reason that if a gene flow involves just one or two (or even three?) of these cellular processes, then it might be the result of multiple LGT/HGT. If significant numbers of genes are transferred into gene flows for each of these four fundamental processes, then it is statistically highly unlikely that they were transferred by multiple independent LGT events. In contrast, endosymbioses are predicted to share similar patterns of gene flows for DNA, RNA, ATP, and membrane-related processes.

To test for endosymbioses, we explicitly search all Pfam descriptors for the appearance of these four terms representing fundamental life processes. From these we count the number of Pfams in which only one, two, or three of these four descriptors ares used. (For example, if a Pfam was to refer to three, or fewer, of the four descriptors shown in table 3, such as "DNA" and "RNA," then that Pfam flow would not be counted as being consistent with an endosymbiotic flow. This procedure enables us to compute statistically independent counts of gene gains within each of these four categories.) Then we ask whether all four independent categories have the same evolutionary history, as measured by gene presence tables. If all four processes have the same evolutionary histories then we infer that they were transported as a single cellular unit, that is, that they represent endosymbiotic transfers. Alternatively, if any of these processes have different histories, then we infer that mechanisms other than endosymbioses, such as HGT, were responsible. By including only Pfams that refer to just one of these four descriptors, we independently measure support for each of the processes, that is, a Pfam referring to DNA and RNA, or to "ATP and membrane," and so on would not be counted.

**Table 3**
Distributions of Pfams and Cell Processes

| A | B | G | $A_c$ | All Pfams | DNA | RNA | ATP | Membrane |
|---|---|---|---|---|---|---|---|---|
| + | + | + | − | 816 | 86 | 21 | 13 | 248 |
| + | + | − | + | 30 | 2 | 1 | 0 | 8 |
| + | + | − | − | 41 | 4 | 1 | 0 | 4 |
| + | − | + | + | 211 | 7 | 11 | 9 | 56 |
| + | − | + | − | 157 | 9 | 10 | 8 | 37 |
| + | − | − | + | 52 | 3 | 3 | 2 | 7 |
| − | + | + | + | 241 | 19 | 12 | 4 | 54 |
| − | + | + | − | 378 | 28 | 17 | 8 | 97 |
| − | + | − | + | 41 | 2 | 3 | 0 | 8 |
| − | − | + | + | 233 | 21 | 14 | 5 | 59 |

Note.—Significant Pfam flows are in red.

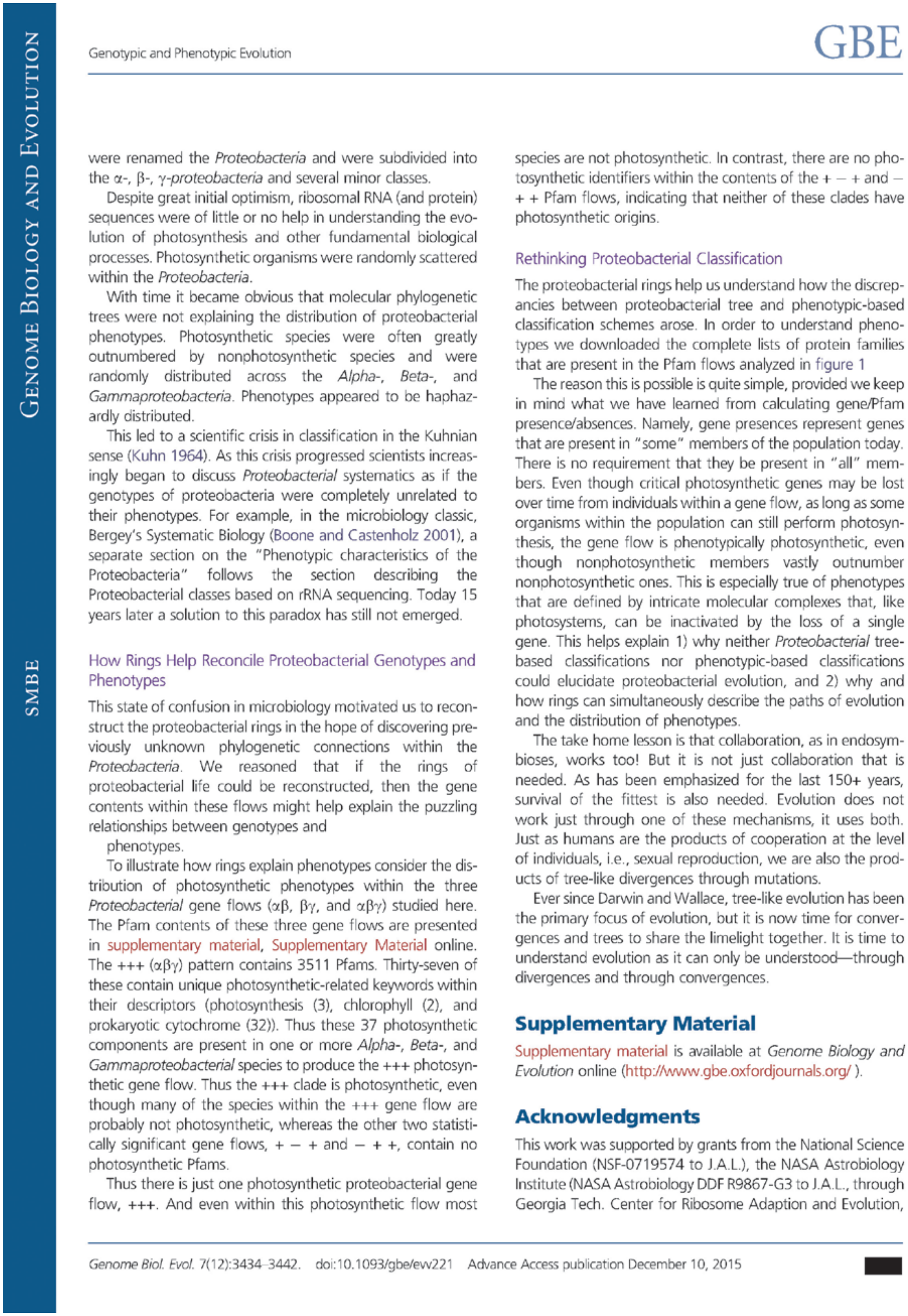
From the four independent sets of gene counts analyzed in table 3, we calculate lists of the numbers of informative patterns found in the proteobacterial rings. The six largest Pfam flows (shown in red) are present in the same rows for all four categories: DNA, RNA, ATP, and membranes. Because the six largest informative patterns in the DNA, RNA, ATP, and membrane columns are statistically independent and because they correspond to the same six largest informative patterns in the "All Pfams" column, we conclude that endosymbioses are responsible for the identical patterns observed for all four significant gene flows. The small probability that all four categories support the same rings happened by chance, $P < 5.15 \times 10^{-10}$, operationally identifies endosymbioses as the process responsible for the proteobacterial rings, and excludes LGT/HGT-related mechanism.

## Discussion

### Proteobacterial Genotypes and Phenotypes

Before ribosomal RNA and DNA sequencing was possible the phylogenetic relationships of the *Proteobacteria*, then known as the "purple bacteria," were based on phenotypes. The purple bacteria consisted of two photosynthetic groups: The "purple sulfur bacteria" and the "purple nonsulfur bacteria." One type contained "bacterial chlorophyll a" and the other contained "bacterial chlorophyll b" (Stanier et al. 1976). Thus photosynthesis initially seemed to provide a reasonable functional basis for classification within the purple bacteria.

However, when Margaret Dayhoff and collaborators published the first ribosomal RNA trees (Dayhoff 1972), the study of proteobacterial evolution was transformed. Two of the three 5S ribosomal RNA sequences analyzed in that work were from purple bacteria, and the third was from a human cell line. Her pioneering work, although not highly publicized or promoted, had a major effect on molecular phylogenomics. As more 5S and subsequently 16S rRNA (Ribosomal Ribonucleic Acid) sequences appeared the purple bacteria



were renamed the *Proteobacteria* and were subdivided into the α-, β-, γ-*proteobacteria* and several minor classes.

Despite great initial optimism, ribosomal RNA (and protein) sequences were of little or no help in understanding the evolution of photosynthesis and other fundamental biological processes. Photosynthetic organisms were randomly scattered within the *Proteobacteria*.

With time it became obvious that molecular phylogenetic trees were not explaining the distribution of proteobacterial phenotypes. Photosynthetic species were often greatly outnumbered by nonphotosynthetic species and were randomly distributed across the *Alpha-*, *Beta-*, and *Gammaproteobacteria*. Phenotypes appeared to be haphazardly distributed.

This led to a scientific crisis in classification in the Kuhnian sense (Kuhn 1964). As this crisis progressed scientists increasingly began to discuss *Proteobacterial* systematics as if the genotypes of proteobacteria were completely unrelated to their phenotypes. For example, in the microbiology classic, Bergey's Systematic Biology (Boone and Castenholz 2001), a separate section on the "Phenotypic characteristics of the Proteobacteria" follows the section describing the Proteobacterial classes based on rRNA sequencing. Today 15 years later a solution to this paradox has still not emerged.

## How Rings Help Reconcile Proteobacterial Genotypes and Phenotypes

This state of confusion in microbiology motivated us to reconstruct the proteobacterial rings in the hope of discovering previously unknown phylogenetic connections within the *Proteobacteria*. We reasoned that if the rings of proteobacterial life could be reconstructed, then the gene contents within these flows might help explain the puzzling relationships between genotypes and

phenotypes.

To illustrate how rings explain phenotypes consider the distribution of photosynthetic phenotypes within the three *Proteobacterial* gene flows (αβ, βγ, and αβγ) studied here. The Pfam contents of these three gene flows are presented in supplementary material, Supplementary Material online. The +++ (αβγ) pattern contains 3511 Pfams. Thirty-seven of these contain unique photosynthetic-related keywords within their descriptors (photosynthesis (3), chlorophyll (2), and prokaryotic cytochrome (32)). Thus these 37 photosynthetic components are present in one or more *Alpha-*, *Beta-*, and *Gammaproteobacterial* species to produce the +++ photosynthetic gene flow. Thus the +++ clade is photosynthetic, even though many of the species within the +++ gene flow are probably not photosynthetic, whereas the other two statistically significant gene flows, + − + and − + +, contain no photosynthetic Pfams.

Thus there is just one photosynthetic proteobacterial gene flow, +++. And even within this photosynthetic flow most species are not photosynthetic. In contrast, there are no photosynthetic identifiers within the contents of the + − + and − + + Pfam flows, indicating that neither of these clades have photosynthetic origins.

## Rethinking Proteobacterial Classification

The proteobacterial rings help us understand how the discrepancies between proteobacterial tree and phenotypic-based classification schemes arose. In order to understand phenotypes we downloaded the complete lists of protein families that are present in the Pfam flows analyzed in figure 1

The reason this is possible is quite simple, provided we keep in mind what we have learned from calculating gene/Pfam presence/absences. Namely, gene presences represent genes that are present in "some" members of the population today. There is no requirement that they be present in "all" members. Even though critical photosynthetic genes may be lost over time from individuals within a gene flow, as long as some organisms within the population can still perform photosynthesis, the gene flow is phenotypically photosynthetic, even though nonphotosynthetic members vastly outnumber nonphotosynthetic ones. This is especially true of phenotypes that are defined by intricate molecular complexes that, like photosystems, can be inactivated by the loss of a single gene. This helps explain 1) why neither *Proteobacterial* tree-based classifications nor phenotypic-based classifications could elucidate proteobacterial evolution, and 2) why and how rings can simultaneously describe the paths of evolution and the distribution of phenotypes.

The take home lesson is that collaboration, as in endosymbioses, works too! But it is not just collaboration that is needed. As has been emphasized for the last 150+ years, survival of the fittest is also needed. Evolution does not work just through one of these mechanisms, it uses both. Just as humans are the products of cooperation at the level of individuals, i.e., sexual reproduction, we are also the products of tree-like divergences through mutations.

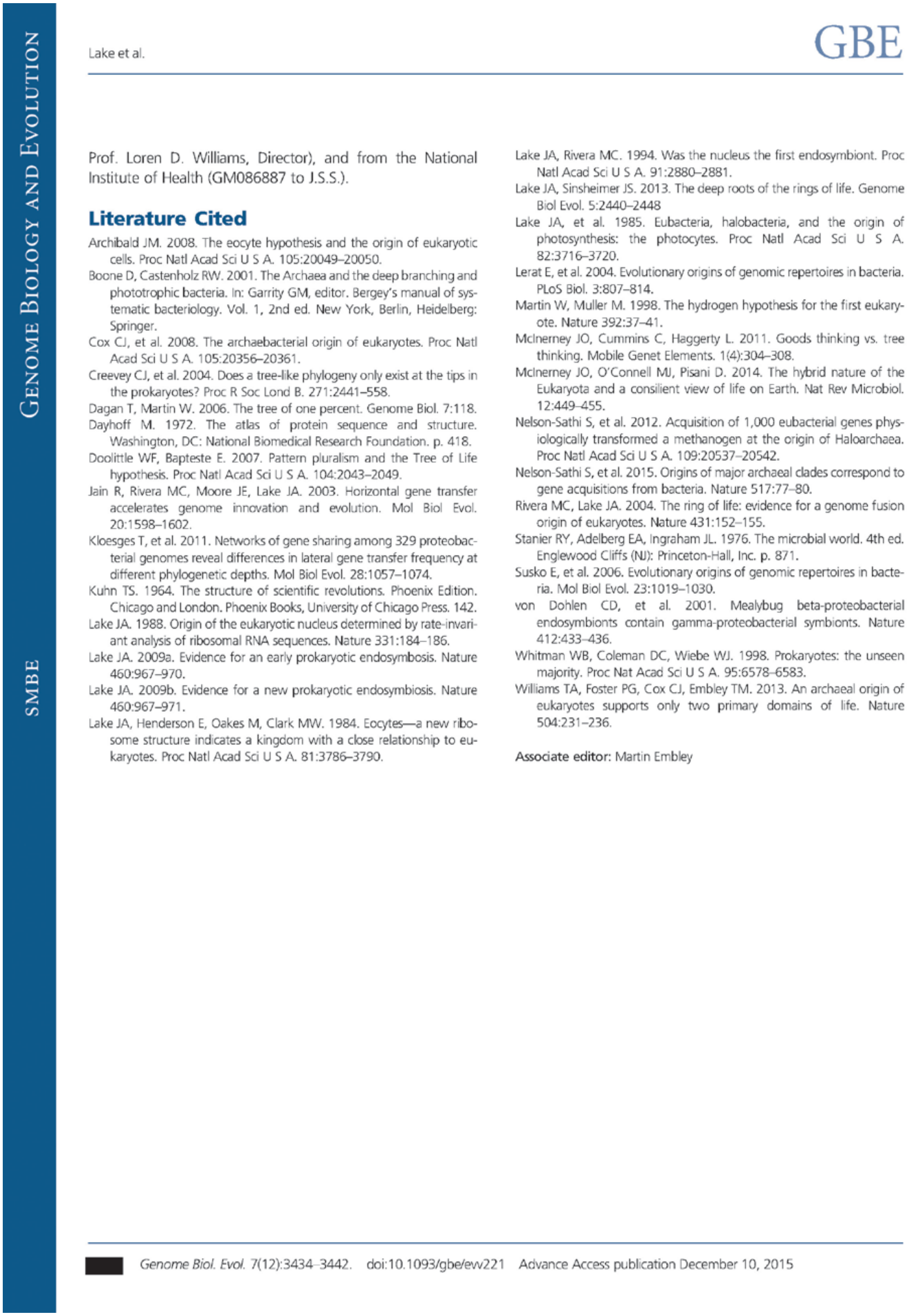
Ever since Darwin and Wallace, tree-like evolution has been the primary focus of evolution, but it is now time for convergences and trees to share the limelight together. It is time to understand evolution as it can only be understood—through divergences and through convergences.

## Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/ ).

## Acknowledgments

This work was supported by grants from the National Science Foundation (NSF-0719574 to J.A.L.), the NASA Astrobiology Institute (NASA Astrobiology DDF R9867-G3 to J.A.L., through Georgia Tech. Center for Ribosome Adaption and Evolution,



Prof. Loren D. Williams, Director), and from the National Institute of Health (GM086887 to J.S.S.).

## Literature Cited

Archibald JM. 2008. The eocyte hypothesis and the origin of eukaryotic cells. Proc Natl Acad Sci U S A. 105:20049–20050.

Boone D, Castenholz RW. 2001. The Archaea and the deep branching and phototrophic bacteria. In: Garrity GM, editor. Bergey's manual of systematic bacteriology. Vol. 1, 2nd ed. New York, Berlin, Heidelberg: Springer.

Cox CJ, et al. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A. 105:20356–20361.

Creevey CJ, et al. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc R Soc Lond B. 271:2441–558.

Dagan T, Martin W. 2006. The tree of one percent. Genome Biol. 7:118.

Dayhoff M. 1972. The atlas of protein sequence and structure. Washington, DC: National Biomedical Research Foundation. p. 418.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A. 104:2043–2049.

Jain R, Rivera MC, Moore JE, Lake JA. 2003. Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol. 20:1598–1602.

Kloesges T, et al. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol. 28:1057–1074.

Kuhn TS. 1964. The structure of scientific revolutions. Phoenix Edition. Chicago and London. Phoenix Books, University of Chicago Press. 142.

Lake JA. 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of ribosomal RNA sequences. Nature 331:184–186.

Lake JA. 2009a. Evidence for an early prokaryotic endosymbosis. Nature 460:967–970.

Lake JA. 2009b. Evidence for a new prokaryotic endosymbiosis. Nature 460:967–971.

Lake JA, Henderson E, Oakes M, Clark MW. 1984. Eocytes—a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci U S A. 81:3786–3790.

Lake JA, Rivera MC. 1994. Was the nucleus the first endosymbiont. Proc Natl Acad Sci U S A. 91:2880–2881.

Lake JA, Sinsheimer JS. 2013. The deep roots of the rings of life. Genome Biol Evol. 5:2440–2448

Lake JA, et al. 1985. Eubacteria, halobacteria, and the origin of photosynthesis: the photocytes. Proc Natl Acad Sci U S A. 82:3716–3720.

Lerat E, et al. 2004. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:807–814.

Martin W, Muller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37–41.

McInerney JO, Cummins C, Haggerty L. 2011. Goods thinking vs. tree thinking. Mobile Genet Elements. 1(4):304–308.

McInerney JO, O'Connell MJ, Pisani D. 2014. The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nat Rev Microbiol. 12:449–455.

Nelson-Sathi S, et al. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc Natl Acad Sci U S A. 109:20537–20542.

Nelson-Sathi S, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80.

Rivera MC, Lake JA. 2004. The ring of life: evidence for a genome fusion origin of eukaryotes. Nature 431:152–155.

Stanier RY, Adelberg EA, Ingraham JL. 1976. The microbial world. 4th ed. Englewood Cliffs (NJ): Princeton-Hall, Inc. p. 871.

Susko E, et al. 2006. Evolutionary origins of genomic repertoires in bacteria. Mol Biol Evol. 23:1019–1030.

von Dohlen CD, et al. 2001. Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. Nature 412:433–436.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. Proc Nat Acad Sci U S A. 95:6578–6583.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–236.

Associate editor: Martin Embley

# CHAPTER TWO

## Reconstructing Phylogenetic Rings: *Occam's Ring* Structures and the R.I.N.G.S. Method

# Reconstructing Phylogenetic Rings: *Occam's Ring* Structures and the R.I.N.G.S. Method

**Authors and Affiliations:**

Joseph R. Larsen, Janet S. Sinsheimer, Dan Thy Tran, and James A. Lake

University of California, Los Angeles

## Abstract

Phylogenetic rings represent evolution on a taxanomic scale by showing both convergent endosymbiotic-events and divergent processes representing tree-like events producing different species. Although rings present a novel way to reconstruct models that may expand the knowledge of how life evolved, there has yet to be an quantitative method on how to construct phylogenetic rings. The goal of this article is to introduce two methods for reconstructing phylogenetic rings. The first method is defined as *Occam's Ring*, which is the simplest ring structure that contains all taxa being investigated as well as all statistically significant informative patterns. The second method is proclaimed here as Ring Identification for Non-Generalized Structures (R.I.N.G.S.) method, which is an expansion on *Occam's Ring* which produces a model of evolution that is more specific and dynamic. Introducing these two methods is a first step in encouraging the exploration of new and informative reconstructions of evolution.

## Introduction

Phylogenetic rings model evolution on a taxanomic scale by depicting divergent and convergent gene flows. The diverging pathways represent when two species have evolved separately to the point they are considered to be unique from one another, which is the speciation so commonly depicted in phylogenetic trees. When a converging pathway is observed it is believed that a genome fusion has occurred, such as when endosymbiosis there has been a genome transfer event similar to endosymbiosis. These convergences cause the unidirectional flows to come back in on themselves and produce a ring like structure. Through this reconstruction of evolutionary history, a more realistic depiction of how certain taxa have evolved can be modeled.

The phylogenetic rings were first introduced in 2004 by James Lake and Maria Rivera in their Nature article, *The ring of life provides evidence for a genome fusion origin of eukaryotes*. The motivation for the phylogenetic rings were to resolve how genome fusions and horizontal gene fusions were obscuring gene sequences in the reconstruction of the "Tree of Life" (Rivera MC. and Lake JA. 2004). This led to the first depiction of an unrooted Ring of Life, and eventually would mature into its current rooted form, known as the "Rings of Life" (**figure 2-1)** (Lake JA. and Sinsheimer JS. 2013 and Lake *et al.* 2015). This representation of how early life evolved on Earth is the cornerstone of the Ring of Life Hypothesis, which is a contender to replace the Three Domain Hypothesis, and has been gaining support through evidence from current research (McInerny JO, O'Connell MJ., and Pisani D. 2014 and McInery JO, Pisani D., O'Connell MJ. 2015).

Although it is evident that the reconstruction of phylogenetic rings is capable of contributing to the modeling and understanding of evolutionary history, no formal ring reconstruction protocol has been outlined. This article is to present a procedure for producing basic phylogenetic ring structures and then introduce a new quantitative technique to explicitly develop ring models. The first procedure, called Occam's Ring, constructs a ring with the minimum number of gene flows consistent with the statistically informative patterns. A new exploratory technique is introduced, called the Ring Identification for Non-Generalized Structures (R.I.N.G.S.) method. This process is being developed to help reconstruct phylogenetic rings that have specific gene flows in order to produce a realistic depiction of how life evolved. By presenting these two methods we hope to introduce phylogenetic ring reconstruction to the field of phylogenetic modeling in order to better summarize evolutionary histories.

## Methods

### Occam's Ring

The objective when reconstructing phylogenetic rings using the *Occam's Ring* fomalization is to model the simplest representation of how we believe genes flowed through diverging and converging paths. In order to make the appropriate ring structure, a new type of phylogenetic tree is introduced here known as the Duplicate Taxa Tree (DTT), which is a phylogenetic tree that has at least one taxon that appears at least twice at the end of a terminal branch. Every DTT can be transformed into a ring by combining all taxa of the same type, which appear more than once, into a converging path producing a phylogenetic ring (**figure 2-2)**. Branches of a DTT are analogous to the flows in the corresponding phylogenetic ring. We will show that a DTT can be a useful tool in phylogenetic ring reconstruction.

To reconstruct *Occam's Rings*, we generate a presence-absence table of protein family counts of all taxa currently being modeled. Protein family (pfam) counts are currently used due to their ability to represent tens of thousands of species, so a protein family database allows a vast sample size to be generated quickly and reliably. The protein family database is an ideal one for reconstructing phylogenetic rings because the assumption of gene gain detection and gene loss insensitivity applies (Lake *et al.* 2015). This assumption states that for a gene gain to be detected only a single cell in the population has to produce a new gene while a gene loss is detected only when every member of the taxon being studied experiences the same loss. When using large sample sizes gene gains are easily detected, where gene loss detection is highly improbable and so this assumption holds for ring reconstruction.

Using the counts generated in our presence absence table we must identify the counts associated to the root pattern, the statistically significant informative patterns, and singleton patterns. The root pattern is the pattern containing all pluses and the singleton patterns are the patterns that have only one plus and n-1 minuses when there are n taxa. The statistically significant informative patterns were introduced in a previous paper (James *et al.* 2015*).* They are patterns that have at least two pluses and at least one minus, which can be identified using the chi-square analysis described below. The counts attributed to statistically significant informative patterns are believed to be protein families involved in genome fusion events.

The chi-square analysis used here was derived by defining two positive integers x and y where

$$x < y$$

These values are the observed counts in a goodness-of-fit chi-square test. The expected values for both these observed values is the midpoint,

$$\frac{(x+y)}{2}$$

From here we can calculate the test statistic for the goodness-of-fit and find

$$\chi^2 = \frac{\left(x - \left(\frac{(x+y)}{2}\right)\right)^2}{\left(\frac{(x+y)}{2}\right)} + \frac{\left(y - \left(\frac{(x+y)}{2}\right)\right)^2}{\left(\frac{(x+y)}{2}\right)} = \frac{(x-y)^2}{(x+y)}$$

The test statistic we have derived for our chi-square is identical to that of the McNemar Test. This approach is appropriate for our analysis because we may assume the evolution of a protein family is a random and rare event, and therefore follows a Poisson distribution. Since the patterns are unique from one another, our counts may be considered independent. Using this chi-square test for proximity the degrees of freedom are always one, the null hypothesis is the two values are the same, and the alternative hypothesis is the terms are different.

Thus, the chi-square test is applied to the protein family counts associated to all patterns with at least two pluses and at least one minus, in their rank order. Where the strongest test statistic appears is where the divide between statistically significant informative patterns, the larger of the pair of counts tested and all counts greater than that, and the supposed noise, the smaller of the pair tested and all counts less that that, is believed to exist. This divide is the border between those patterns that represent flows involved in divergences and in a genome fusions, with the higher counts being defined as the statistically significant informative patterns, and are present in our model.

The Occam's Rings procedure we have outlined here is designed to reconstruct rings when there is at least one informative pattern that is significantly less observed than the other patterns. When all informative patterns are observed in roughly equal numbers then the Occam's Rings procedure alone can not determine whether a ring is appropriate or whether there just isn't enough data to determine which of the standard phylogenetic trees is the best representation of the evolutionary history. Additional analyses are needed to distinguish between these possibilities

In order to represent the *Occam's Ring* structure, a DTT is drawn that has the minimum number of branches and taxa of each type that represents the statistically significant informative patterns. Most taxa will have a single gene path, from root to terminal end, leading to it unless it is absolutely necessary to add taxa, and therefore branches, to incorporate the statistically significant informative patterns. Once the DTT with the necessary number of minimal branches and taxa is depicted, then it is transformed into its ring representation, which is the *Occam's Ring* structure for this set of taxa. An example of what is and is not an *Occam's Ring* for 4 taxa may be observed in **figure 2-3.**

## Ring Identification for Non-Generalized Structures (R.I.N.G.S.) Method

The goal of the Ring Identification for Non-Generalized Structures (R.I.N.G.S.) method is to

develop a detailed reconstruction of how taxonomic groups evolved throughout history. This method picks up where Occam's Ring left off in that we only use R.I.N.G.S. after finding evidence for some gene flows using Occam's Rings.

A major underlying assumption of this process is that all taxa in the set undergo a constant rate of evolution. This means that the point from where the first protein family present in this set of taxa starts the root, known as the generating point, then the rate of accumulation of genes along any given path to a single terminal end is assumed to have a constant rate *C*. A path is defined as a set of branches on the DTT that always begin at the generating point and can be traveled to a single terminal end.

To begin this process the shortest path must be identified, by finding the lowest count value. A path starts in the root, where the first pfam in this set of taxon is generated, travels a set of branches, never turning back on itself, ending at a terminal branch for a single taxon. Of all the ways to travel from the root to a terminal end, at a taxon. We are curious to identify the shortest path assoicated to the least number of protein counts over the traveled branches. In order to identify this path, we observe each taxon individually and identify which patterns have the current taxon present. Once these protein family counts are identified they are summed to produce this particular taxon's sum of counts. This is repeated for each taxon being investigated. The lowest sum of counts is the smallest rate that *C* must fit and therefore is defined as *C* for this set. The taxon defined as *C* is believed to have the least amount of gene flows passed into it due to its low rate of protein families produced over time, and therefore has one taxon on the DTT.

The other sum of counts will be compared to *C*, if they are found to be approximately the same then they are believed to also have a single taxon on the DTT. If the sum of counts and *C* are not the same then following analyses is required.

Again because we are comparing a large sample size, due to using protein families, our statistic  z has a normal distribution with mean zero and variance one.

$$z=\left(\frac{(x-y)}{\sqrt{(x+y)}}\right)$$

We recognize this to be only an approximation, because we understand these paths are independent paths we also recognize that the sum of counts used in this analysis are generated from common branches. Therefore the independence of this analysis is questionable.

For the initial comparison, we set the null to be *C* equal to the sum of counts of the path(s) leading to a  taxon currently being compared, with an alternative of not equal. If the result of the analysis is that there is not enough evidence to reject the null, then we assume that the path(s) leading to this taxon is similar enough to *C* and there is also a single path leading to this taxon on the DTT, though if the null is rejected then we will test the possibility that the sum of counts represents two paths. This is done by multiplying *C* by two and then adding another root count, otherwise known as the count associated to the root pattern, to the sum of counts being tested. The additional root count is added because every individual path starts at the generating point, which is the initial value that is the root count, and travels some, yet to be deduced, path. Although it is unknown whether every branch traveled for this taxon, it is assumed that the sum of counts for this taxon contains the counts for most of these paths, which is why it failed the assumption of one path initially.

The same comparison is run again, but now the null is that two times *C* is equal to the sum of counts of the current taxon with the additional root count added on, and an alternative hypothesis of not equal. If this null is not rejected then it is assumed this taxon appears twice on the DTT. Though if the sum of counts plus the root count is less than two times *C*, then we may also assume two taxon on the DTT because the sum of counts and the additional root counts may be missing from the other duplicate paths that will be identified in a later step. Therefore if we believe the sum of counts, with the appropriate number of roots added on, to be the same or less than *C* times the number of paths being considered then this process is complete for this particular taxon.

If the sum of counts with the added root count is still greater than two times *C*, and the null is rejected then *C* is multiplied by 3 and another root count is added to the sum of counts. The same inference process is run, with a null hypothesis of three multiplied by *C* being equal to the sum of counts summed with two root counts, and once again an alternative hypothesis of not equal. This process is continued in this fashion until the null is unable to be rejected or the sum of counts with the additional root counts is found to be less than *C* times the number of supposed paths.

This inference is done for all taxa being studied and the final number of paths is approximated for each individual taxon. Next, the necessary counts and the approximate number of taxa present on the DTT are obtained. From here a subset of all possible DTT are constructed that contain these counts and approximate number of taxa. Due to the assumption of a constant rate of evolution, each terminal branch from the same node has equal counts, and similarly all branches from the same node to the terminal node should sum to the same count. By applying this logic all the branch counts are calculated for all possible trees. By using the counts like branch lengths and assuming they follow a Poisson distribution we may derive the maximum log-likelihood for each DTT by setting the derivative of the log-likelihood to zero as follows

$$\ln(L(\lambda_i)) = \sum_{i=1}^{b} -\lambda_i + c_i \ln(\lambda_i)$$

$$0 = \frac{d}{d\lambda_i}(\ln(L(\lambda_i))) = -1 + \frac{c}{\lambda_i}$$

For some tree, $T_{k,}$ in the subset of DTT, *b* is the number of branches on the particular DTT and $c_i$ is the count associated to the pfam count, a constant, associated to the current branch. Although when we evaluate this maximum log-likelihood for any branch *i* it is found

$$\lambda_i = c_i$$

Meaning that when evaluating the maximum log-likelihood for any given tree, we may assume that each branch's mean is equal to the branch's count. Therefore, the maximum log-likelihood for any DTT may be defined by

$$\ln(L(T_k)) = \sum_{i=1}^{b} -c_i + c_i \ln(c_i) = -\left(\sum_{i=1}^{b} c_i(1 - \ln(c_i))\right)$$

Taking this into consideration, in order to pick the most likely tree in the subset of DTT's we use AIC. Because all the trees in our subset have the same number of branches, after solving for them earlier using our assumption of a constant rate of evolution, we have equal numbers of λ for each likelihood. Therefore the penalty term of each of the AICs is the same. Thus we drop the penalty term and let the AIC be for all $T_k$

$$AIC = -2\ln(L(T_k)) = 2\left(\sum_{i=1}^{b} c_i(1 - \ln(c_i))\right)$$

The tree, $T_k$, with the lowest AIC is determined to be the most likely tree DTT. Finally, the selected DTT is transformed into its phylogenetic ring form, producing a more realistic model of evolution.

**Results**

To illustrate both of these methods they will be applied to the Alpha-,Beta-, and Gammaproteobacteria classes, previously studied in *Rings Reconcile Genotypic and Phenotypic Evolution within Proteobacteria* by Lake *et al.* (Lake *et al.* 2015).

***Occam's Ring***

To begin, the presence absence table for the Alpha-,Beta-, and Gammaproteobacteria classes are generated, and may be found in **table 2-1**. The count associated with the root pattern appears in the table as 4396 and the counts associated with the singleton patterns are 303 for Alphaproteobacteria, 132 for Betaproteobacteria, and 881 for the Gammaproteobacteria. In order to identify the statistically significant informative patterns, all the counts associated to patterns with at least two pluses and at least one minus are placed in ascending order. Then the chi-square analysis is applied to all the adjacent counts. The most substantial test statistic is 235.81 and is found between the flows for Alpha/Gamma, 419, and Alpha/Beta, 77. This means we believe any count 419 or above was involved in a gene flow that underwent a genome fusion. Therefore the statistically significant informative patterns for this set of classes is Alpha/Gamma and Beta/Gamma.

Now that we have all necessary counts, we can draw a DTT with the minimum number of taxa and branches that incorporate the Alpha/Gamma and Beta/Gamma flows. It is necessary to have a flow that leads to only Alpha and Gamma but not Beta as well as a flow that leads to a Beta and Gamma but not Alpha. So the DTT for this construct will have a branch leading to terminal branches containing Alpha and Gamma and another branch leading to terminal branches containing Beta and Gamma. The simplest DTT we can construct,  in Newick format, would be ((Α,Γ),( Γ,Β)).

This DTT requires an extra Gammaproteobacteria flow to be present to represent the two statistically significant informative patterns and represents the minimum number of flows that are consistent with the observed results for these taxa. Since we have at least two statistically significant informative patterns for three taxa, then a ring is believed to be an appropriate model By this conclusion we believe we found the DTT analogous to *Occam's Ring* for Alpha-,Beta-, and Gammaproteobacteria. The last step is to transform the DTT reconstructed here into its ring form, which may be found in **figure 2-4**.

**R.I.N.G.S. Method**

Once we have established that a phylogenetic tree is insufficient to explain the observed counts, then we can apply the R.I.N.G.S. method.  When applying the R.I.N.G.S. method, in order to identify a less generalized model of evolution for the Alpha-,Beta-, and Gammaproteobacteria, it is necessary to expand on the *Occam's Ring* analysis. That means the only information obtained up to this step is the  count associated to the root pattern (4396), the counts associated to the singleton patterns (303, 132, and 881 for Alpha-,Beta-,and Gammaproteobacteria, respectively), and counts associated to the statistically significant informative patterns ( 419 and 547 for the Alpha/Gamma and Beta/Gamma flows, respectively). The next step is to approximate how many paths for each taxon likely appear on the DTT.

In order to approximate the number of each taxon, the sum of counts for each taxon is found by summing all the counts for each taxon for which that taxon is present. The sum of counts are found to be 5118  for the Alphaproteobacteria, 5075 for the Betaproteobacteria, and 6243 for the

Gammaproteobacteria. Due to our assumption of a constant rate of evolution between the Alpha-,Beta-, and Gammaproteobacteria, then the lowest sum of counts is the smallest count that the constant rate, *C*, must fit. This means we define *C* for this set of classes to be the Betaproteobacteria sum of counts, 5075.

Now that *C*, the count attributed to a single path on the DTT, has been identified the remaining members of the set of classes must be compared to *C* to see if they have one or more paths. First, the Alphaproteobacteria sum of counts, 5118, is compared to *C* using the z-test. A p-value of .3336 is found, meaning the null hypothesis, stating that the two values are the same, does not have enough evidence to be rejected for a critical value of 0.05. Therefore it is approximated that one Alphaproteobacteria appears on the likely DTT. Next, the Gammaproteobacteria sum of counts of 6243 is compared to *C*, and a p-value of less than .0001 is found and the null of the two paths being the same is rejected. So *C* is multiplied by two, checking for two paths for Gammaproteobacteria, producing a count of 10150 and another root count is added onto the sum of counts of the Gammaproteobacteria, producing a count of 10639. This is tested by the z-test and a p-value of 0.0003 and once again the null hypothesis of the counts being the same is rejected. So, *C* is multiplied by three, checking for three paths, producing a count of 15225 and two root counts are added to the sum of counts of the Gammaproteobacteria, producing a count of 15035. Once again these values are tested by the z-test, but a p-value of .138 is found and the null hypothesis does not have enough evidence to be rejected. Therefore, it is approximated that the likely DTT has one Alphaproteobacteria, one Betaproteobacteria, and three Gammaproteobacteria.

Next, a subset of all possible DTT's is generated with the statistically significant informative patterns and number of each taxon as a constraint. Since we are assuming a constant rate of evolution among taxa, we recognize that any pair of branches diverging from a node have equal sum of counts from the starting node to any terminal end it leads to. Due to this assumption, the unknown branch counts may be deduced from the necessary counts identified earlier. The subset of possible trees with their appropriate branch counts may be found in **figure 2-5**.

In order to select the most likely tree in this subset of DTT's, the AIC is calculated for each tree. Although prior to that, tree (III) may be ruled out, because for it to have equal branch sums from the first node to any terminal end it would require a negative protein family count over one of its branches. Since this is not possible, tree (III) is not possible. Once the AIC is calculated for each tree, which are shown in **figure 2-5**, it is discovered that tree (I) is the most likely DTT for the Alpha-,Beta-, and Gammaproteobacteria. The final step in this process is to transform the selected DTT into its phylogenetic ring form, as depicted in **figure 2-6**.

## Discussion

The Occam's Ring shown in **figure 2-4** is the simplest representation of the evolutionary history of the Proteobacteria based on the pfam data. This model shows that the two major gene flows that led to the Gammaproteobacteria were due to an endosymbiotic like event that occurred between ancestors of the Alphaproteobacteria and Betaproteobacteria. The genome fusion depicted in this general model is only a single additional piece of information, yet it is important. Information about what protein families were passed into the Gammaproteobacteria could help better inform what protein families evolved earlier in history or how they contributed to the Gammaproteobacteria class. However due to the generality of the *Occam's Ring* method of reconstructing evolutionary histories, there is little detail about the evolutionary history one can deduce from this construct.

In order to better understand the history of the Alpha-,Beta-, and Gammaproteobacterial, we utilized a newly formulated way to detect additional flows. This construction was derived from our new  method, R.I.N.G.S.. Although the rings presented here are slightly different from Lake *et al.* (Lake *et al.* 2015), the result does not contradict any assertions made in that work and only improves the model with additional evidence. The postulated genome fusion events carry precedent due to the prevalence of endosymbiosis in the history of proteobacteria classes (Sagan L. 1967 and von DohlenI*et al.* 2001). We summarize the evidence that supports the phylogenetic ring found using the R.I.N.G.S. method.

The first phylogenetic ring for the Alpha-,Beta-, and Gamma- proteobacteria, presented by Lake *et al.*, recognized the presence of a Alpha-Gamma and Beta-Gamma flows. This information is presented here but further explored, using the R.I.N.G.S. method, derived by detecting the number of taxa on the ring's DTT. Although there is a different double ring structure present in Lake's 2015 study, it still represents multiple flows into the Gammaproteobacteria, which the Alpha- and Beta-proteobacteria produced. The only adjustment we made here is the placement of a Gammaproteobacteria on the likely DTT based on the probability of our result and the impossibility of the DTT for the previous ring structure in Lake's 2015 depiction, which corresponds to (III) in **figure 2-5**. This slight adjustment alters the root placement but retains all the critical conclusion and assertions made by the earlier ring, such as the flow of protein families, what the structure elicits about the proteobacteria class, and allowing for additional statistically significant informative patterns of either the Firmicute and Actinobacteria taxon. In other words, the new representation only adds more quantifiable evidence to the earlier ring studies of Proteobacteria while ultimately creating a more likely ring reconstruction that supports the prior assertions.

Traditional phylogenetic trees are limited in their ability to portray endosymbiotic events; summarizing evolution using rings is a preferred approach to include these events. Along with the evidence provided by Lake *et al.*(Lake *et al.* 2015), it is evident that Proteobacteria have undergone a multitude of endosymbiotic events throughout their history. Such events include those described by the Endosymbiotic Theory (Sagan L. 1967) and findings of endosymbiotic events of Eukaryotes with Gammaproteobacteria, who themselves went through endosymbiosis with Betaproteobacteria (von Dohlen *et al.* 2001). This gives the reconstructions portrayed here lends precedence that genome fusions, possibly endosymbiosis, led to the genesis of the Gammaproteobacteria.

Therefore, taking the empirical evidence presented here, along with the DTT constructed earlier we believe we have found the appropriate multiple flow ring structures for the Alpha-,Beta-, and Gamma- proteobacteria. This information may be used to help us understand which of these three evolved earlier in evolutionary history and how common traits among these Proteobacteria were inherited due to these gene flows. Further research into these classes will be needed, but we hope this model will help provide a road map for future studies.

**Conclusion**

Both *Occam's Ring* and the R.I.N.G.S. method are important tools in understanding evolutionary history. *Occam's Ring* provides a general but reliable ring reconstruction, while R.I.N.G.S. identifies additional flows and brings more specificity to help researchers better understand how genes flowed into a particular taxon and from where once phylogenetic trees can be rejected in favor of rings. The *Occam's Ring* approach necessarily cannot determine just how many gene flows have occurred.  At this early stage, the R.I.N.G.S. method leaves room for further development. Problems with the R.I.N.G.S. method include the assumption of equal rates of evolution along all branches. Further work is necessary to determine how reasonable this assumption is, how much the method depends on this assumption, what biases can occur with the conclusions from R.I.N.G.S. if the assumption is violated and how the assumption might be relaxed. Another fundamental problem is that the estimation of the number of influential flows depends on a sequential testing scheme that ends with a failure to reject a null hypothesis. Statistically, failure to reject a null hypothesis does not necessarily mean that the null hypothesis is true but that is what we are essentially assuming here. The Occam's Ring procedure also has its own problems. The most notable problem is that the method breaks down if all influential patterns are large. Despite these problems, we believe that using these two methods of phylogenetic ring reconstruction provide insights into understanding the complicated evolutionary histories of the Proteobacteria.

There is no question that depicting evolutionary relationships as phylogenetic rings is still in an early stage. When used alongside traditional phylogenetic trees and networks, rings can help provide a more full and encompassing image of what paths life took to be what it is today. Researchers have found it difficult to depict the history of life in a simple tree structure, phylogenetic rings may be the answer. Life certainly is not only about the strongest surviving, which are what phylogenetic trees primarily depict, but also about the most cooperative surviving, which rings incorporate beautifully.

## Acknowledgements

Partial funding for this work was provided by the National Science Foundation grant DMS 1264153 . A major thanks goes out to the support and guidance of the UCLA Biomathematics Department, especially Dr. Kenneth Lange.

## Figures

**Figure 2-1**

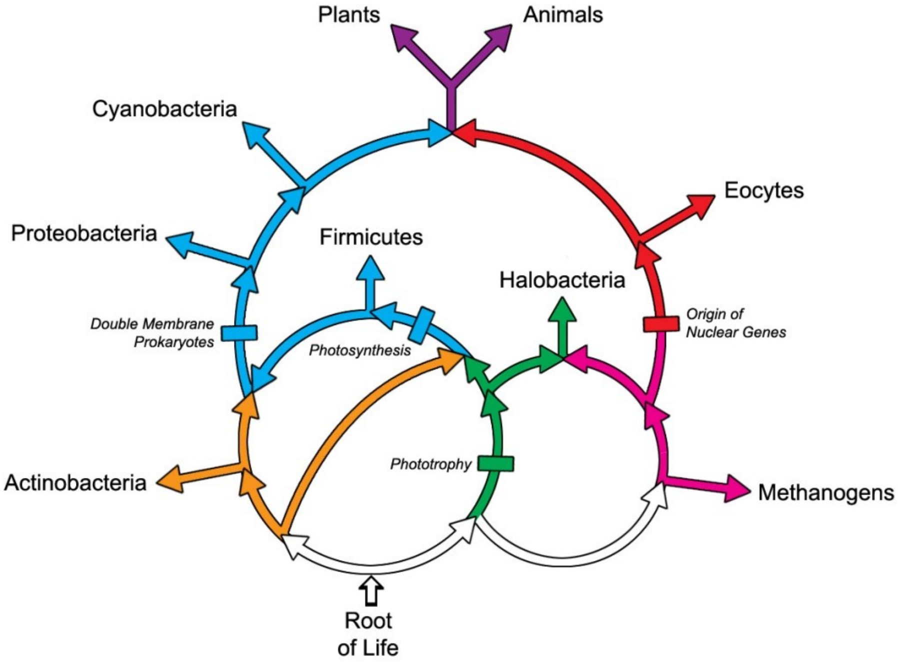


**Figure 2-2**

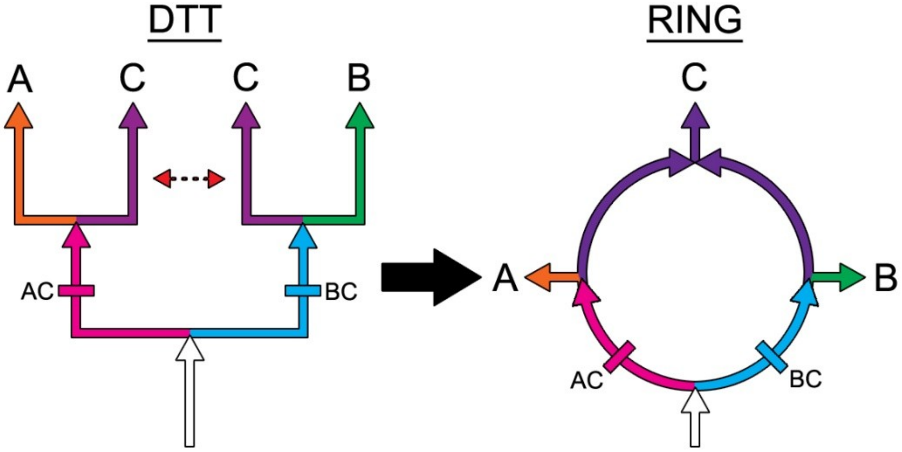


**Figure 2-3**

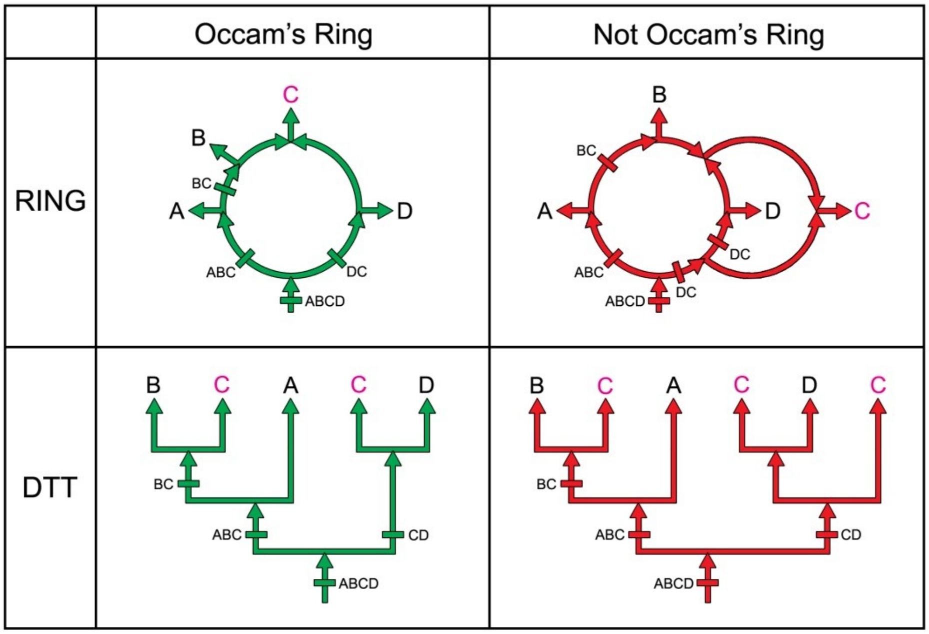


**Figure 2-4**

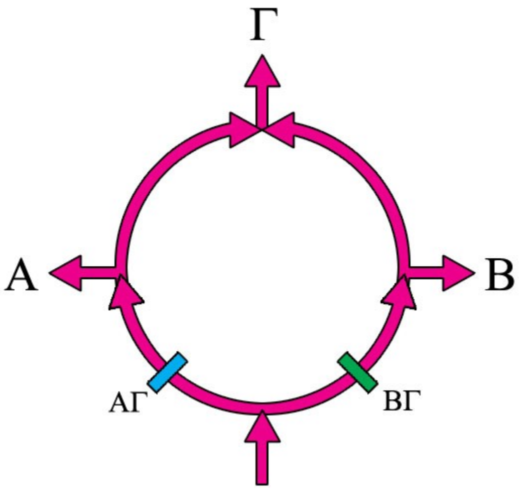
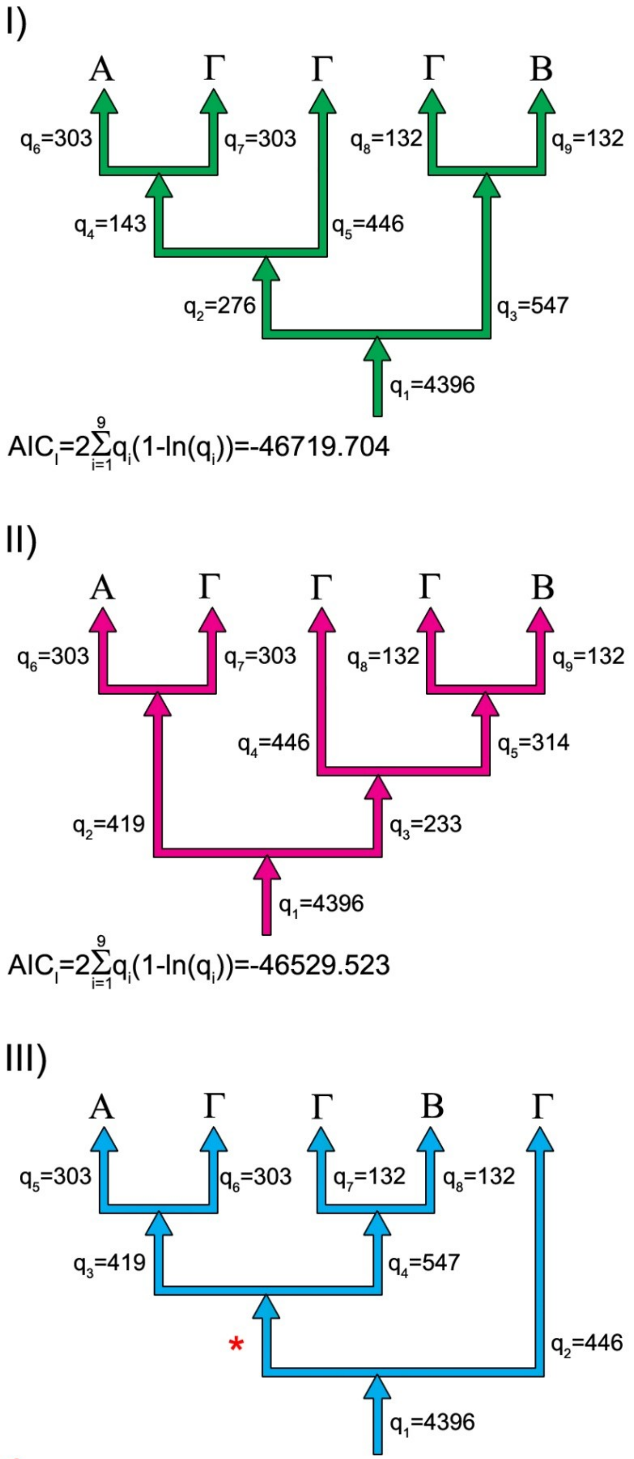
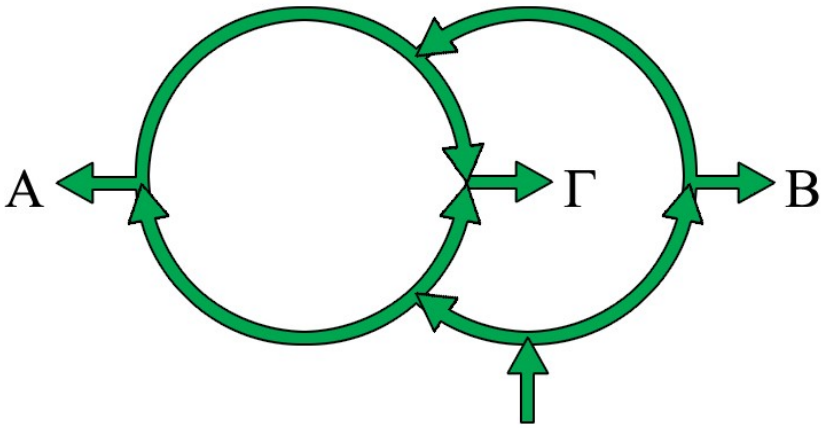


**Figure 2-5**



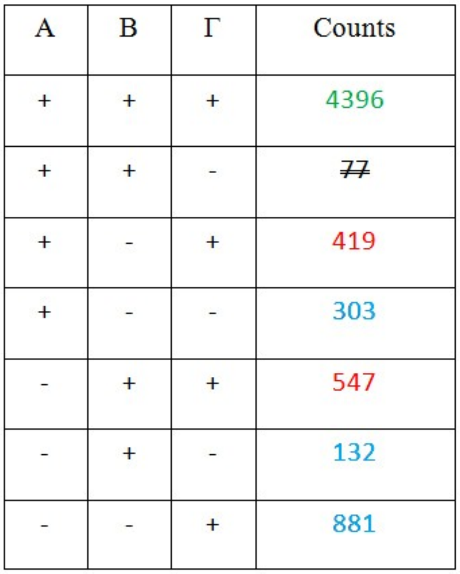
*Would require a negative count, which is not possible

**Figure 2-6**



## Table

### Table 2-1



| A | B | Γ | Counts |
|---|---|---|---|
| + | + | + | 4396 |
| + | + | - | ~~77~~ |
| + | - | + | 419 |
| + | - | - | 303 |
| - | + | + | 547 |
| - | + | - | 132 |
| - | - | + | 881 |

## Figure Legends

**Figure 2-1. Rings of Life.** The current reconstruction of the rooted Rings of Life. Introduced in the article *The Deep Roots of the Rings of Life* by Lake JA. and Sinsheimer JS. and presented in its current form in *Rings reconcile genotypic and phenotypic evolution within the Proteobacteria* by Lake *et al.*

**Figure 2-2. An example of a DTT and its analogous phylogenetic ring.** A Duplicate Taxa Tree (DTT) is a phylogenetic tree with at least one taxa that appears more than once on the phylogenetic tree. Though this is not possible on a traditional phylogenetic tree, a DTT allows this due to repetitive taxa being combined into a converging path, creating a phylogenetic ring. The DTT on the left has a duplicate C present and the red arrows represent how they are combined to produce its respective phylogenetic ring, found on the right.

**Figure 2-3. An example of *Occam's Ring.*** The left column presents an *Occam's Ring* and its respective DTT while the right column shows a ring structure that does not adhere to the *Occam's Ring* criteria. Though both rings have the same statistically significant informative patterns, notice that the DTT in the right column has one more C than the left column. Therefore the right DTT does not have the minimum number of taxa or paths to represent the same flows while the left does. Therefore the left column ring, and its respective DTT, is the *Occam's Ring* for these taxa and statistically significant informative patterns.

**Figure 2-4. The Occam's Ring for Alpha-,Beta-,Gammaproteobacteria.** The DTT with the minimum number of taxa and paths with an Alpha/Gamma and Beta/Gamma flow, in Newick format, is ((Α,Γ),( Γ,Β)). The phylogenetic ring presented here is the structure associated with this DTT.

**Figure 2-5. Subset of all possible Duplicate Taxa Tree with statistically significant**

**informative patterns and likely number of taxon.** Presented here are the three possible reconstructions for a Alpha/Gamma and Beta/Gamma flow as well as one Alpha-proteobacteria, one Beta-proteobacteria, and three Gamma-proteobacteria taxa on the DTT. The counts were found by using the numbers in the presence-absence table and recognizing that the sum of branches for the two branches leaving any node must be equal. DTT (III) is eliminated immediately due to the impossibility of having a negative branch length. The AIC formulated here is performed on the remaining two trees and it is found that DTT (I) is the most likely reconstruction.

**Figure 2-6. The ring structure for Alpha-,Beta-,Gammaproteobacteria from R.I.N.G.S. method.**

Transforming DTT (I) from figure 2-5, which was found to be the most likely DTT in the subset of possible DTT, into its ring form produces the phylogenetic ring structure presented here.

**Table Legends**

**Table 2-1. The Pattern-Absence Table for Alpha-,Beta-, and Gammaproteobacteria.**

In green is the count associated to the root pattern, in blue are the counts associated to the singleton counts, and in red are the counts associated to the statistically significant informative patterns. After applying the chi-square analysis, it was found that 77 is believed to be due to gene transfers and is uninformative for our reconstruction.

## Conclusion

My contribution to phylogenetic rings has been studying their application and developing reconstruction techniques. The applications I helped investigate involved studying the proteobacteria class and the possible origins of photosynthesis. The reconstruction techniques I helped develop were Occam's Ring and the R.I.N.G.S. method. These methods are the first ever to be developed, and therefore created with the intention of sharing their methods in order to be used by fellow scientist.

The work I have accomplished up to this point has help further the legacy of phylogenetic rings, but future research is still necessary. Although major work in the application of phylogenetic rings and their reconstruction has been undertaken in the last year, there is still immense work left to be done. The future work for the application of phylogenetic rings would be to take the techniques used in the manuscript Rings reconcile genotypic and phenotypic evolution within the Proteobacteria and apply them to other classes and other biological pathways. A possible class that may be better understood by these techniques is the cyanobacteria. Also, further investigation needs to be done to validate the origin of photosynthesis. Additional work regarding the first quantitative method for reconstructing rings, R.I.N.G.S. method also needs to be done. The R.I.N.G.S. method has immense potential, but has assumptions and makes some approximations. The major assumptions in the R.I.N.G.S. method is that there is a constant rate of evolution between taxa and that genes can be added but are not eliminated. Depending on the data sets being considered these assumptions are not always true and require future work in finding ways to compare taxon paths while relaxing these assumptions. The approximation that needs to be worked out is the z-test that compares the taxon paths, although the paths being compared share overlapping components. Due to the shared paths, there is a question whether the compared values are independent and therefore may not satisfy the independence condition for the z-test. A possible solution for this may be found in modeling the dependency of these paths.

## References

Lake JA. e*t al.* 2015. Rings reconcile genotypic and phenotypic evolution within the Proteobacteria. Genome Biol Evol.7:3434-3442.

Lake JA., Sinsheimer JS. 2013. The deep roots of the rings of life. Genome Biol Evol. 5:2440-8

McInery JO., O'Connell MJ., Pisani D. 2014. The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nature Review Microbiology. 12:449-455.

McInery JO., Pisani D., O'Connell MJ. 2015. The ring of life hypothesis for Eukaryotes origins is supported by multiple kinds of data. Phil Trans R Soc B.370:20140323

Rivera MC., Lake JA. 2004. The ring of life: evidence for a genome fusion origin of eukaryotes. Nature. 431:152-155.

Sagan L. 1967. On the origin of mitosing cells. J Theor Biol. 14:255-74.

von Dohlen CD *et al.* 2001. Mealybug beta-proteobacterial endosymbionts contain gamma-` proteobacterial symbionts. Nature 412:433–436.