# Uncovering the Genomic Origins of Life

James A. Lake[1,2,3,*], Joseph Larsen[1,2,4], Dan Thy Tran[1,2], and Janet S. Sinsheimer[3,4]

[1]Molecular, Cell and Developmental Biology, University of California, Los Angeles

[2]Molecular Biology Institute, University of California, Los Angeles

[3]Human Genetics, University of California, Los Angeles

[4]Biomathematics, University of California, Los Angeles

*Corresponding author: E-mail: jimucla@gmail.com.

Accepted: June 25, 2018

Data deposition: All data obtained in this study are presented in table 1, "Pattern Counts for Origin of Life Ring Analyses." The python script used for downloading the Pfam data will be uploaded to the Dryad, or to a similar website upon publication of this article.

## Abstract

The Origin of Life Domain (OLD) is the period during which life on Earth began. Here, we derive and use a new phylogenetic algorithm to analyze Protein Families in order to reconstruct the chronological steps by which the OLD evolved. During this period, life began with the appearance of the fundamental components of life such as RNAs, DNAs, amino acids, and membranes. Chronologically, the Origin of Life preceded the Last Universal Common Ancestor, which then subsequently engendered modern life on Earth. Our phylogenetic algorithm allows us to explicitly answer previously unknown origin of life questions. Specifically, we explain and illustrate our computational methods by reconstructing the rings describing the evolution of the RNA and DNA worlds. We phylogenetically reconstruct how the RNA and DNA worlds evolved, infer the origins and chronological order of appearance of the first genetic codes, test whether the Ribosomal RNA world preceded the Membrane world, and interpret these new findings with respect to the experimental and theoretical origin of life studies by others.

**Key words:** genetic codes, LUCA, membranes, Rings of Life, evolution, RNAs.

## Introduction

As the evolution of life is being investigated in increasing detail, our understanding of the beginnings of life is being pushed back to ever earlier times through phylogenetic reconstructions (Cox et al. 2008; McInerney et al. 2014). Previously, phylogenetic studies have reconstructed important aspects of life within Last Universal Common Ancestor (LUCA) (Koonin and Martin 2005; Martin et al. 2008), the last universal common ancestor that gave rise to all extant life on Earth. However, it is thought to be impossible to reconstruct evolution explicitly within the Origin of Life Domain (OLD) which preceded LUCA.

Despite this inability to reconstruct the OLD, origin of life studies nevertheless prospered as they followed an indirect but highly productive path. Scientists experimentally searched for biochemical reactions that could produce primal molecules of life present at life's beginnings. These studies, described below, successfully discovered various in vitro biosynthetic pathways for producing molecules that were likely to have been present during the early evolution of life on Earth, including nucleic acids, membranes, amino acids, and ribosomes.

Experimental origin of life studies began with the Miller–Urey experiments (Miller 1953) that synthesized amino acids by using electrical spark discharges under reducing atmospheric conditions that were thought to exist on early Earth. Those experiments were subsequently followed by the "RNA World" proposal (Gilbert 1986). Together both approaches ultimately stimulated researchers to discover primitive chemistries on Earth that could have produced the first amino acids, nucleic acids, and membranes.

In the last two decades, biochemical searches for origin of life reactions have further extended those earlier results and have produced a broad understanding of early biological reactions (Mansy and Szostak 2008; Powner et al. 2009) and potential environments (Weiss et al. 2016) that likely affected life's early evolution during LUCA. Despite these successes, most origin of life scientists thought that it would be

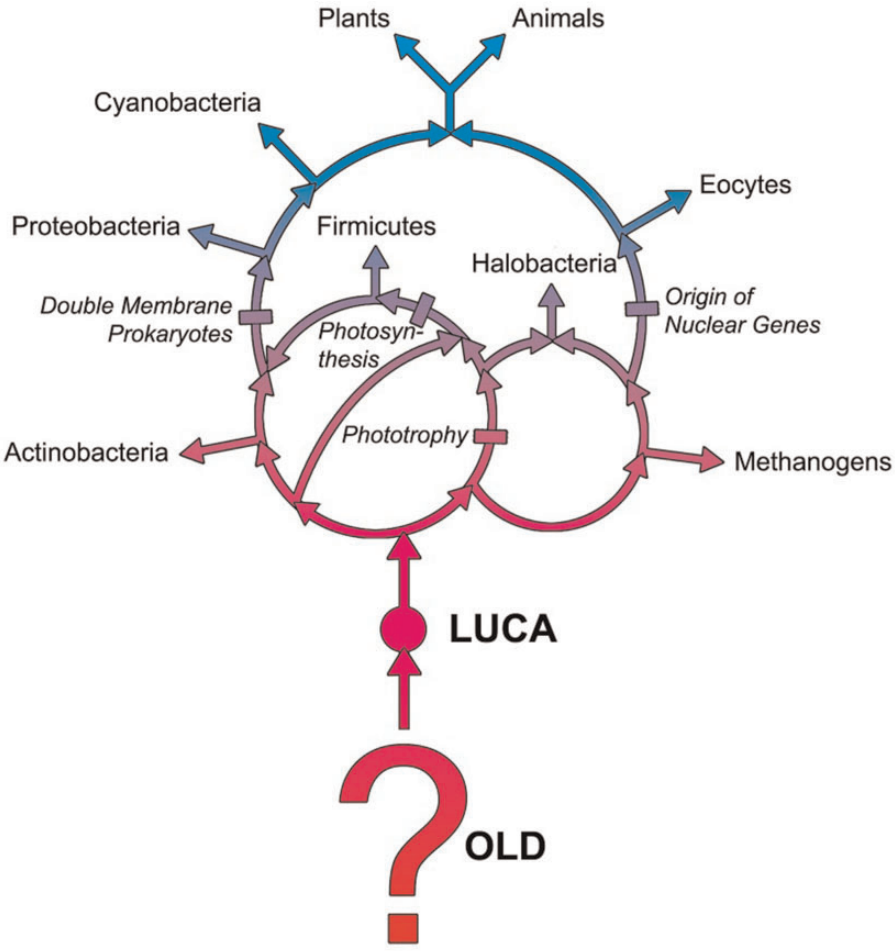


**Fig. 1.**—Relationship of the rings of life to the last common ancestor and to the origin of life domain. During the early stages of life, genes/Pfams first evolved and flowed as "organisms" from the Origin of Life Domain (OLD), shown as a magenta question mark, into the Last Universal Common Ancestor (LUCA). Subsequently, the organisms present in LUCA then flowed into the Rings of Life representing extant life on Earth.

impossible to reconstruct the evolution of life before LUCA by using genomic and sequence data. If it were possible to phylogenetically reconstruct the origin of life, then this would greatly extend our understanding of this earliest stage of biological evolution.

Here, we search the Sanger Pfam database, consisting of ~12 million sequences, 16,245 Protein Families, and 35 million Uniprot Descriptors, and demonstrate how to explicitly reconstruct the evolutionary steps by which membranes, the RNA world, and the DNA world evolved. Using our genomic phylogenetic reconstruction methods, we reconstruct steps in the evolution of the genetic code ($P < 0.05$), and establish that the ribosomal RNA world preceded the RNA/membrane world ($P < 0.05$).

## Reconstructing Milestones in the Origin of Life

Over the last decade, tree-based phylogenetic methods have been employed to reconstruct those genes that were present in the post-LUCA period. During this time, we have learned that genome evolution proceeds by both divergences (tree-like branchings) and by convergences (mergers of gene flows to form rings). Here, we demonstrate how ring based techniques can use Pfam data to phylogenetically reconstruct the chronology of the origin of life gene flows that produced the first RNAs, DNAs, and membranes.

The complex gene flows shown in figure 1 describe three phases in the evolution of life on Earth. The rings shown at the top were reconstructed from Protein Families, or Pfams (protein domains that share a common evolutionary origin). Individual Pfams are recognized by their similar sequences, structures, and functions. Because protein families represent complex structures, they evolve much more slowly than either nucleotides or amino acids. This makes them ideal for performing deep evolutionary reconstructions, such as those used here to reconstruct the evolution of life from the first origin of life "organisms," to LUCA, and then to the Rings of Life (shown at the top of fig. 1). Because these gene flows can

converge to form rings and diverge to form trees, they provide a mathematically complete representation of the major evolutionary events that occurred during the evolution of life on Earth (Lake et al. 2015). LUCA preceded and gave rise to the Rings of Life and is represented by the magenta circle in figure 1. As a result of molecular studies of LUCA, we now understand in considerable detail which Pfams were present in LUCA (Weiss et al. 2016).

Over the last decade, phylogenetic reconstructions have provided evidence for the evolutionary rings that describe the recent evolution of life on Earth (the upper rings in fig. 1). Because many genes are shared by all known organisms it is generally accepted that extant life has evolved from LUCA (shown in red in the middle of the figure).

Life, however, did not begin originate at LUCA, and a diverse group of organisms existed before LUCA. This third component of this graph, the Origin of Life Domain or OLD, consists of those organisms that existed at the beginnings of life. The OLD is shown schematically by the magenta question mark at the bottom of figure 1. The OLD describes the evolution of life during the period when life was first emerging and beginning to utilize new biochemical inventions like DNAs, RNAs, membranes, and amino acids.

Although the rings of life (at the top of fig. 1) are relatively easy to reconstruct from Pfams (Lake et al. 2015) and while many of the protein families present in LUCA are known (Weiss et al. 2016), in contrast far less is known about the phylogeny of life within the OLD.

LUCA was commonly thought to impede us from to reconstructing the OLD because it stands between the Rings of Life and the OLD and thus its location would seem to prevent us from seeing the rings of the OLD. Fortunately, this is not the case. In the following sections, we describe how LUCA can be used to facilitate reconstructing life within the OLD, demonstrate how to reconstruct evolution within the OLD, and interpret these first reconstructions of the OLD in light of what is currently known about the origin of life on Earth.

## Material and Methods

In order to reconstruct the OLD it is important to understand the evolutionary processes that regulate the evolution of Protein Families. Specifically, understanding the population based evolutionary mechanisms that directly influence the evolution of gene flows is essential because these mechanisms allow populations of organisms to acquire new Protein families easily and simultaneously prevent the loss of Pfams from gene flows (Lake et al. 2015). Figure 2 illustrates how this happens.

The large black circle shown on the left side of figure 2 contains three green +'s that represent unique Pfams that were present within the initial gene flow. Over time, some Pfams within this gene flow will be lost by chance from individuals within this population, marked by the missing green pluses within the black circles in the upper rectangle. Simultaneously, other individuals will acquire new genes from the environment and from other organisms, as shown by the red pluses in the lower rectangle. Hence the large circle that is immediately to the right of the upper box represents the total number of different Pfams that are present in all organisms in the gene flow. Even though some, or possibly many, individual organisms will have lost genes, those genes nevertheless still remain within the population represented by the rectangle because it would have been mathematically impossible for all organisms to lose all copies of any one gene in a large population. This is particularly true in the rings of life, since some prokaryotic phyla are estimated to contain 1,029 individuals or more (Whitman et al. 1998).

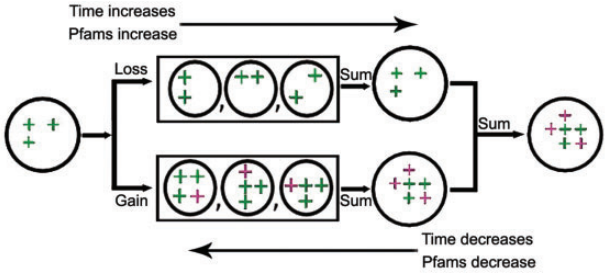


**Fig. 2.**—Gene population mechanisms monotonically increase the size of gene flows, over time. In this example, the population of organisms shown on the far left side of the figure initially contains three protein families, marked in green. As described in the text, individual Pfams can only rarely be completely lost from large gene flows, whereas new Pfams are continuously being invented and added to large gene flows as these flows move from the OLD, then into LUCA, and finally into the Rings of Life. Thus, the numbers of unique genes, that is, Pfams, must monotonically increase in gene flows over time as explained in the text.

In contrast, it is easy for phylogenetic flows to gain genes as shown by the large box on the bottom line. If a gene containing a new Pfam (shown as a red + in fig. 2) is gained by even one member of the population, for example, perhaps by an Alphaproteobacterium, then that gene will be identified by Pfam searches as being "present" within the Alphaproteobacteria. Thus, over time, the number of distinct Pfams that are present in the Alphaproteobacteria will continuously increase as new Pfams are gained in individuals. Thus, when the Pfam inventory from cells with losses is added to that from cells with gains, the net result is a continual increase in the number of Protein Families present within that gene flow over time, as shown by the increase in acquired (red) genes at the right of figure 2.

Our reconstructions of the OLD are based on this observation that the numbers of Pfams within gene flows continually increase over time. The corollary to this is that if one looks back to earlier times, then the numbers of Pfams within gene flows will decrease. As a result, there will be many Pfams in the ROL gene flows, fewer Pfams in LUCA, and still fewer Pfams in the OLD gene flows (the oldest).

Together gene flows within the OLD, LUCA, and ROL contain all of the chronological information that is needed to reconstruct the evolution of life from the Origin of Life, to LUCA, and then to the Rings of Life. Because the oldest Pfam gene flows are supported by the smallest numbers of Pfams they can be statistically identified as the group containing the smallest gene flows, just as we previously identified the largest gene flows as being members of the modern Rings of Life (Rivera and Lake 2004). Thus, the smallest flows contain the information that is needed to reconstruct evolution within the OLD.

## Categorizing and Labeling Organisms in the OLD

In order to reconstruct evolution within the OLD, we first label and categorize the organisms in the OLD into groups that are analogous to the orders etc. of extant organisms, and then analyze these labeled groups.

Because origin of life organisms preceded modern life by billions of years, we categorize them by molecules that are generally thought to have been present when life was first emerging. Some of the prime suspects for origin of life organisms are those that utilized amino acids, DNAs, RNAs, and membranes (Miller 1953; Gilbert 1986; Koonin and Martin 2005; Mansy and Szostak 2008; Martin et al. 2008). For example, if an origin of life organism contains RNA, Membranes, and DNA, we categorize it as an "RNA, DNA, Membrane" organism. Similarly, we label and categorize the OLD organisms that utilized RNA and Membranes by placing them in the "RNA, Membrane" category. We then use this system to label OLD organisms and to reconstruct their evolutionary paths from Pfams, similar to the way in which the Rings of Life were reconstructed from Pfams.

## Reconstructing the Origin of Life Rings

Here, we phylogenetically reconstruct the Origin of Life rings from the three Pfam presence/absence data sets that are shown in table 1. The three Pfam data sets listed below represent the number of times that Pfam descriptors contain the following presence/absence patterns. The first line of the first data set shows that all four terms, "rRNA," "mRNA," "tRNA," and "Membrane" appear in 49 different Pfams, the second line shows that 15 Pfams contain the three terms, "rRNA," "mRNA," and "tRNA." The three presence/absence tables shown below allow us to reconstruct three unique origin of life rings.

In order to interpret the data shown in the columns of table 1, we first identify the statistically largest gap between the smallest set of informative counts and the next smallest set of informative counts in that column. (Note: singleton presence/absence patterns, such as + − − −, or − + − −, are not topologically informative and thus were not used in our analyses.) In the reconstruction of the column on the left, the largest statistical gap between the small counts and the set of intermediate counts occurs between 23 and 40 and defines the boundary of the OLD. The probability that this gap happened by chance, *P*, is $<0.0323$. Thus, all nonsingleton patterns supported by counts of 23 or less in this column are informative, and these counts were used to reconstruct the rings describing the mutual evolution of rRNAs, mRNAs, tRNAs, and Membranes.

**Table 1**
Pattern Counts for Origin of Life Ring Analyses

| rRNA<br>mRNA<br>tRNA<br>Membrane<br>r m t M | Adenine<br>Guanine<br>Uracil<br>Cytosine<br>A G U C | Adenine<br>Guanine<br>Thymine<br>Cytosine<br>A G T C |
|---|---|---|
| + + + + 49 | + + + + 7 | + + + + 4 |
| + + + − 15 | + + + − 5 | + + + − 1 |
| + + − + 7 | + + − + 14 | + + − + 17 |
| + + − − 3 | + + − − 14 | + + − − 18 |
| + − + + 12 | + − + + 3 | + − + + 3 |
| + − + − 23 | + − + − 5 | + − + − 0 |
| + − − + 10 | + − − + 35 | + − − + 35 |
| + − − − 22 | + − − − 97 | + − − − 102 |
| − + + + 46 | − + + + 2 | − + + + 4 |
| − + + − 40 | − + + − 13 | − + + − 14 |
| − + − + 64 | − + − + 26 | − + − + 24 |
| − + − − 58 | − + − − 111 | − + − − 110 |
| − − + + 160 | − − + + 5 | − − + + 2 |
| − − + − 188 | − − + − 38 | − − + − 9 |
| − − − + 1,004 | − − − + 28 | − − − + 31 |
| $P<0.0323$ | $P<0.0593$ | $P<0.0184$ |

Note.—The informative patterns used in graph reconstructions are shown in red. Singleton patterns, that is, (− + − −) and (− − + −) do not affect these reconstructions. The probabilities that the red sets of counts are statistically significant are given below each of the columns.

Figure 3 depicts the major periods in the origin and evolution of life on Earth. These are shown as three Super Domains that are linked by horizontal lines. The red line, on the left, separates the OLD from LUCA, and the magenta line, on the right, separates LUCA from the Rings of Life. Because gene flows increase with time, large gene flows are present within the ROL, intermediate gene flows are present within LUCA, and small gene flows are present within the OLD. As figure 3 illustrates, statistically significant gaps exist between LUCA and the small gene flows that are present in the OLD. Thus, knowledge of these three domains allows us to reconstruct life in the OLD.

## Reconstructing the Origin of Life

LUCA is the population of organisms that connects the ROL to the OLD. It is represented by the circle that is present midway between the ROL and the OLD in figure 3. LUCA's intermediate location between the ROL and the OLD thus provides a reference point for relating the OLD to the Rings of Life.

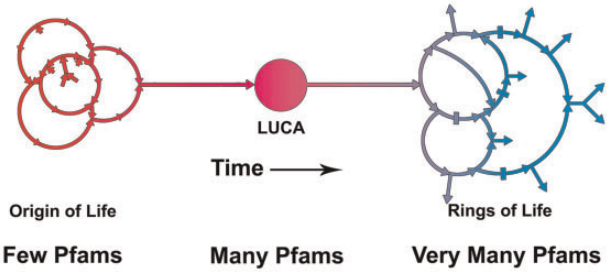


**Fig. 3.**—The evolution of life progressed through three stages. In the earliest stage of life's evolution, that is, during the OLD, shown at the left, DNAs, RNAs, ribosomes, and membranes were first being synthesized and assembled into cells. This was followed by the middle stage, LUCA, which represents the idealized population that gave rise to all subsequent life, and the subsequent Rings of Life that map out the great diversification of life and led to the formation of the kingdoms and phyla that subsequently colonized Earth.

Reconstructing evolution within the OLD is based on a modification of the techniques that were used to reconstruct the ROL (Lake et al. 2015). These new techniques are explicitly described in the next Section.

### Using Protein Family Identifiers to Reconstruct the OLD Rings

Because the numbers of Protein families monotonically increase with time as life evolved from the OLD to LUCA, as illustrated in figure 3, we first identify the statistically significant gap (shown in red in table 1) that separates the smallest set of Pfam patterns from the larger counts shown in black. These red, informative counts uniquely define the topology of the rings of the Origin of Life and are used to reconstruct life in the OLD.

### Steps in Reconstructing the OLD

Reconstructing the OLD involves four steps. In the first step, we identify the smallest set of gene flows. This set is separated from the larger gene flows by a statistically significant gap that can be identified by using $\chi^2$ analyses. Once the set of smallest counts has been identified, we analyze these data as if we were reconstructing the modern Rings of Life starting from LUCA (fig. 4*a*). However, since the smallest gene counts, that is, the OLD counts, are being reconstructed, we must also time-reverse the gene flows so that genes flow from the OLD into LUCA. This is shown by the transition between figure 4*a* and *b*. In the third, and final, step the OLD is rotated 180 degrees in order to be viewed in the conventional view that is shown in figure 4*c*. By using this time reversal reconstruction process, we can reconstruct evolution in the OLD by using only simple graphical modifications of those methods that were used to reconstruct the Rings of Life (Lake et al. 2015).

This procedure is successful because Ring reconstructions are founded on the observation that the population behavior of Pfam evolution dictates: a) that the number of Pfams present within evolutionary gene flows will continuously increase over time so that Pfams can almost never be lost from Pfam gene flows, and b) that gene flows will therefore monotonically increase in size from the smallest (oldest) to the largest (newest).

In traditional ring analyses, we identify the gap between LUCA and the ROL by using $\chi^2$ analyses to find the statistically most significant Pfam gap that exists between the larger counts from the Rings of Life and the smaller Pfam counts from LUCA. Similarly, in our Origin of Life studies, we use $\chi^2$ analyses to find the statistically significant gap that exists between the smaller counts from the OLD and the larger counts from LUCA. We then use the smaller OLD counts to reconstruct evolution within the OLD.

The method illustrated in figure 4 requires a minimum of phylogenetic manipulations. In the first step, all of the arrows shown in the ring reconstructed in figure 4*a* are reversed, as shown in figure 4*b*, so that our gene flows will continually increase in accord with to the population biology model developed and described in figure 2. In the second reconstruction step, figure 4*b* is rotated by 180 degrees about the center of LUCA so that the OLD will be rotated to the bottom of the figure, and the ROL will be rotated to the top of the figure as shown in the traditional view which is shown in figure 4*c*. Thus, by rotating the rings and reversing the direction of time, we can use traditional ring methods to reconstruct the evolution of the OLD.

## Results

### Reconstructing the RNA and DNA Worlds

Few hypotheses were more intriguing and controversial than the RNA world hypothesis (Gilbert 1986) and the ability to reconstruct evolution in the OLD provides an opportunity to test this theory. The RNA world theory posits that RNA may have appeared before DNA in the first nucleic acid organisms. Alternatively, DNA may have appeared first, or possibly both molecules may have coevolved. In order to test these theories, we reconstruct the rings describing the evolution of the four standard bases found in RNAs (Adenine, Guanine, Uracil, and Cytosine), and compare them with the rings describing the evolution of the four standard bases found in DNAs (Adenine, Guanine, Thymine, and Cytosine).

As shown in figure 5, our evolutionary reconstructions of the RNA-world bases and the DNA-world bases have the same topology and hence both reconstructions support the same set of rings. The only difference is that the RNA gene flow starts with a Uracil (U) whereas the DNA gene flow starts with a Thymine (T). In all other aspects, once both flows merge to form the combined Thymine/Uracil gene flow, they follow the same evolutionary paths.

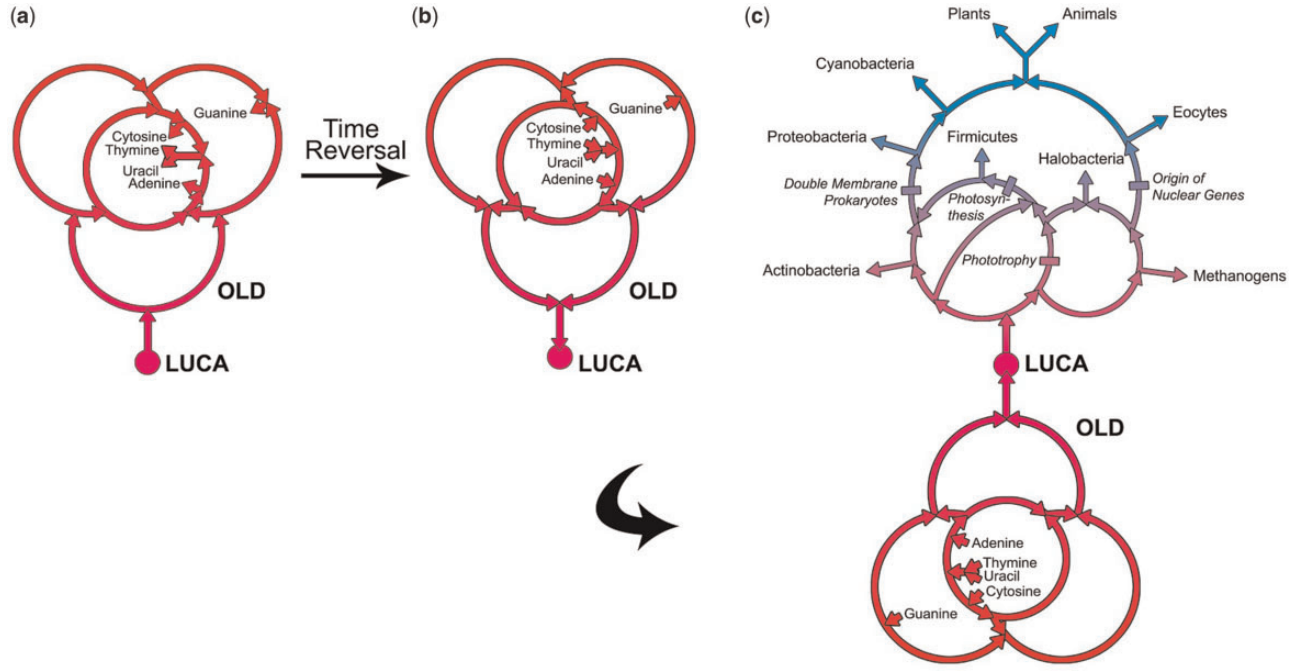


**Fig. 4.**—Reconstructing evolution within the OLD. Since Pfams from LUCA did not flow into the Origin of Life Domain, but rather flowed from the OLD into LUCA, in step 1 we reverse the flow of time, $t \rightarrow -t$, from that shown in (*a*) to that shown in (*b*). Thus, in (*b*), the genes flow from the Origin of Life Domain into LUCA (as they did in the distant past). In the third panel, (*c*), we rotate the graph in (*b*) by 180° and connect the Rings of Life to the OLD through LUCA. This new graph now represents the flow of genes from the OLD into LUCA, and from LUCA into the Rings of Life (*c*).

As the DNA and RNA gene flows represented by Thymine and Uracil merged, the resulting flows bifurcated and began to form the central ring that is present in figure 5. First Adenine joined the upward U/T flow to form the A/U/T gene flow, and Cytosine joined the downward U/T flow to form the C/U/T gene flow. Subsequently, the Guanine gene flow, shown at the lower left of figure 5, bifurcated so that one part flowed upward (toward the top of fig. 5) and joined the A/U/T gene flow, and produced the resulting G/A/U/T gene flow that ultimately joined LUCA.

The lower path of the Guanine gene flow joined the T/U/C gene flow and together these two flows merged to form the G/T/U/C gene flow. Subsequently, the G/T/U/C flow merged with the A/T/U gene flow to form the G/T/U/C/A gene flow (shown in the upper right quadrant of fig. 5). This flow ultimately joined the A/U/T flow from the upper left quadrant of the figure and subsequently the two gene flows at the top of the figure merged to form LUCA. Our interpretations of these results are described in the Discussion Section.

### Reconstructing the Evolution of Membranes, Messenger RNAs, Transfer RNAs, and Ribosomal RNAs

The origins of membranes, messenger-, transfer-, and ribosomal-RNAs are shown as reconstructed in figure 6. These rings are rooted by observing that "rRNA" can flow in both directions from the point where rRNA enters the rings, whereas mRNA-, tRNA-, and membrane-gene flows can travel in only one direction from their entry points. As the rRNA gene flow moves down toward the bottom of figure 6, it merges with the membrane gene flow and produces the first rRNA containing, membrane bound cells. These results provide strong evidence ($P < 0.0323$) that rRNA, a critical part of the functional scaffolding of ribosomes (Blattner et al. 1997; Simonson and Lake 2002; Tamura and Schimmel 2003; Ramakrishnan 2014) existed before mRNAs, tRNAs, and membranes.

## Discussion

### The First Genetic Codes

From the RNA and DNA gene flows reconstructed in figure 5, we can infer the earliest genetic codes based on the temporal order in which the bases emerged. These gene flows predict that the earliest RNA/DNA genetic codes were either the UA/TA or the UC/TC codes, but we cannot determine whether the UA/TA or the UC/TC codes evolved first. Because it would be nonparsimonious to assume that later codes were first, we describe the evolution of the most parsimonious codes.

The importance of Uracil and Thymine within the RNA- and DNA-World rings is apparent in figure 7. Uracil and Thymine are the first bases to appear in their respective rings, and hence they define the beginnings of the RNA and DNA worlds, respectively. Uracil (U) has a number of remarkable

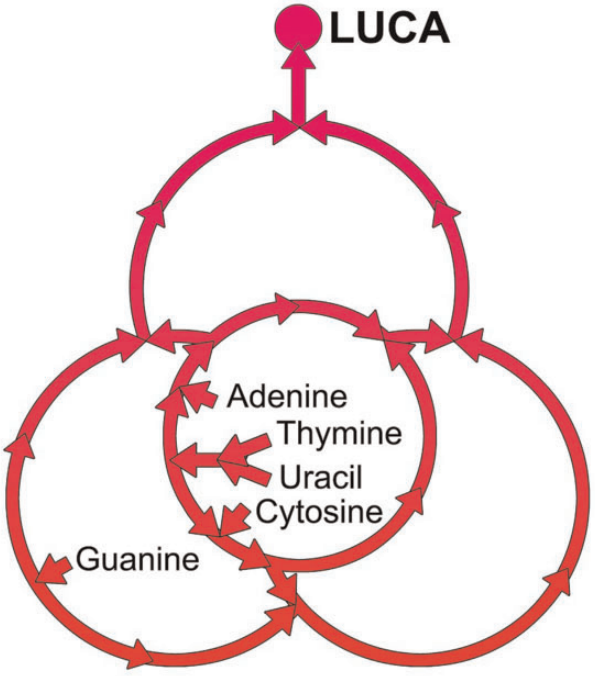


**Fig. 5.**—Evolution of the nucleic acid bases A, T, U, C, and G. Evolution of the nucleic acid bases started with the appearance of Thymine and Uracil, shown by the arrows in the middle of the central ring, and was followed by subsequent gene flow mergers, and by the ultimate emergence of the five bases: Adenine, Thymine, Uracil, Cytosine, and Guanine.

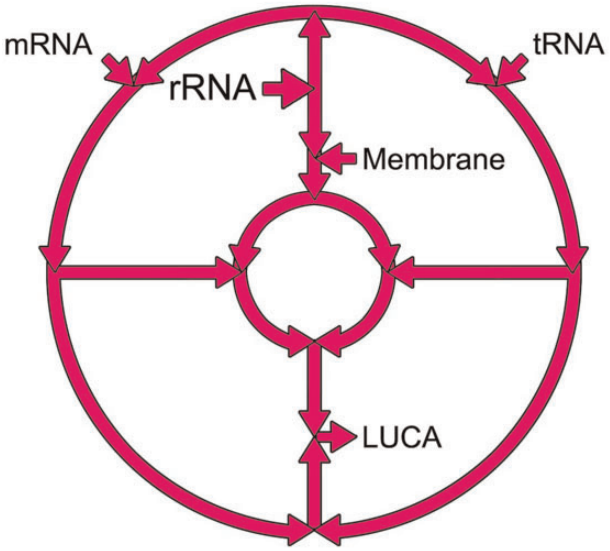


**Fig. 6.**—Ribosomal RNA preceded messenger RNA, transfer RNA, and Membranes. The ribosomal RNA (rRNA) gene flow, shown at the top center of the figure, was present at the beginning of these rings which represent the chronological origins of messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and membranes. Because genes flow in both directions from the entry point of ribosomal RNA, this reconstruction provides direct evidence that rRNA preceded messenger RNA, transfer RNAs, and even preceded the first membrane bound cells. In this sense, the ribosomal RNA world was very early. The relative chronological positions of "mRNA," "tRNA," and "membrane" are such that it is not possible to decide from this graph which of those three gene flows evolved first. But their flows are consistent with the approximately coeval emergence of mRNAs, tRNAs, and membranes.

properties that are not found in the other bases. Uracil (U) is capable of base pairing with Adenine (A), Guanine (G), and Cytosine (C) through traditional, wobble, and/or other non-canonical base pairings (Tanaka et al. 2000). Hence the UA and the TA genetic codes appear to be acceptable candidates for the first codes.

Knowing that these RNA bases can pair, we parsimoniously predict the first genetic codes utilized those nucleotides that were present within the earliest gene flows. Specifically, we envisage the following relative order of appearance of the first RNA genetic codes.

The chronological information contained within the reconstructed RNA rings parsimoniously predicts that the first two genetic codes were the UA- and the UC-codes. The UA code shown in figure 7, could have potentially coded for six amino acids: phenylalanine, leucine, tyrosine, asparagine, lysine, and isoleucine. These six amino acids have the following diverse chemical properties. They are nonpolar-hydrophobic, polar neutral-, hydrophobic neutral-, acid hydrophilic-, and basic hydrophilic-amino acids, respectively (Dickerson and Geis 1969) and thus these "organisms" could potentially have functioned in a variety of environments. The UC code is predicted to have initially lacked a stop codon, and thus termination of protein synthesis would have presumably been inefficient. However, the UC flow could have subsequently gained a stop codon once it merged with the UA gene flow.

In addition, the UA code contains the UAA stop codon (ochre). This stop codon could have made it possible for UA organisms to terminate protein synthesis at defined sites and to suppress frame shifting mutations. Subsequently, other stop codons like GUG and UUG also appeared, but our phylogenetic reconstructions parsimoniously indicate that the UAA stop codon was very early, and possibly the first. If so, this would mark a critical innovation within the UA lineage.

The UC code is parsimoniously predicted to code for four amino acids: phenylalanine, leucine, serine, and proline, since there is no evidence that any amino acid has ever been replaced by another amino acid in the genetic code. Chemically, these amino acids are nonpolar-hydrophobic, nonpolar-hydrophobic, polar neutral, and nonpolar-hydrophobic, respectively. Like the UA code, the UC code does not contain an AUG start codon. Thus, we predict that these very early organisms lacked the modern genetic machinery that is now utilized for the initiation of protein synthesis.

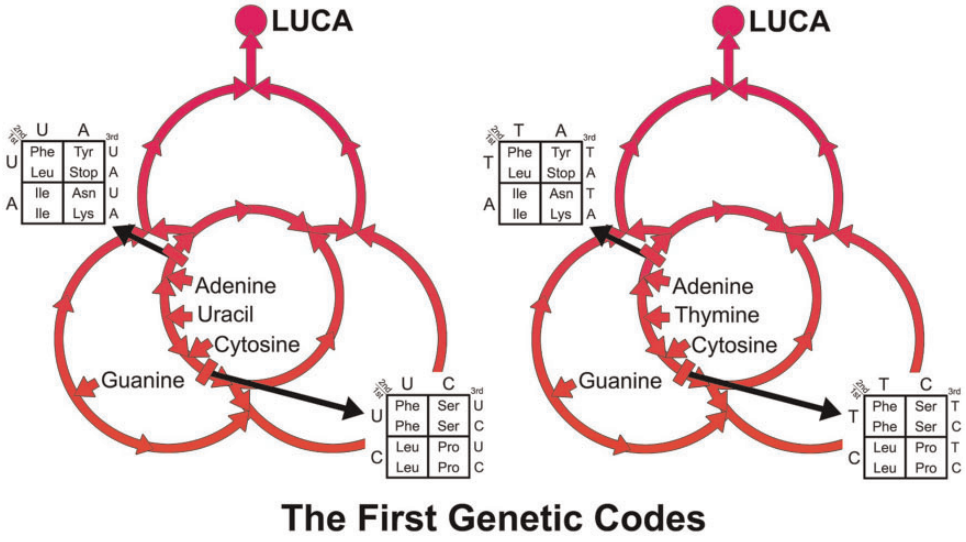


**Fig. 7.**—The First Genetic Codes. The first RNA and DNA genetic codes are parsimoniously inferred from the gene flows present in the RNA world. They are shown in the left and the right sides of figure 6, respectively. The UA code, shown at the left of the figure, is one of the two earliest codes, as is the UC code that is present in the lower right of the figure. Those codes shown in this figure map steps in the chronological evolution of the modern genetic code.

However, later codes, like the UAG gene flow that enters LUCA from the upper left side of the graph, or the UACG gene flow that enters LUCA from the upper right side of the graph, are potential sources for the AUG start codons. Hence this parsimoniously indicates that control of the initiation of protein synthesis appeared relatively late in the evolution of life during the OLD.

## The Origin of the First Cells: how the rRNA World Preceded the Membrane World.

The origin of the first membrane was a momentous step in the evolution of life. This evolutionary advance separated cellular from noncellular life. Membranes made it possible to protect the contents of cells from being diluted, altered, or otherwise adversely modified by the external environment. Biochemical origin of life experiments have shown that within simple membrane vesicles it would have been feasible to modify DNA via multiple cycles of external heating and cooling whereby "...double-stranded DNA could be separated by denaturation at high temperatures while being retained within the vesicles (Mansy and Szostak 2008)," and similarly the common origins of "RNA, protein, and lipid precursors" have been demonstrated (Patel et al. 2015). Once membranes had enveloped the first cell, emerging life could then begin to regulate its cytoplasmic environment.

Before now, however, there's been no direct phylogenetic evidence indicating whether membranes or RNAs came first. Given our new ability to generate genomic based phylogenies within the OLD, we can now ask whether RNAs came before or after membranes. If RNAs came before DNA, we would also like to know which of the RNAs appeared first (mRNA, tRNA, or rRNA).

## The First Membrane Bound Cells

Origin of life experiments have clearly demonstrated the importance of the first membranes and revealed the enormous advantages that they provided (Mansy and Szostak 2008). This suggests that the acquisition of membranes produced a strong selective advantage for the emergence of these first cells. But until now, the absence of direct phylogenetic evidence for the evolutionary steps leading to the first cells has made it impossible to ascertain whether rRNAs, mRNAs, tRNAs, or membranes came first. By using phylogenetic reconstructions of the OLD, we can now directly address these and similar questions. We now have statistically significant evidence that rRNA came before mRNA, tRNA, and membranes.

## A Future of Synergistic Origin of Life Studies

The future of reconstructing major events in the origin of life looks bright, because it provides a unified view of evolution on Earth as shown in figure 8, and especially because biochemical proof of process experiments can now be directly compared with phylogenetic ring reconstructions within the OLD. We predict that this new phylogenetic tool for reconstructing and understanding life's origin may open the door to new discoveries. Future advances will hopefully involve synergistic collaborations between laboratories that are skilled in experimental origin of life studies and laboratories that are proficient in reconstructing genome evolution in the OLD.

Such collaborations have the potential to integrate our understanding of the evolution of: nucleic acids, protein synthesis, membranes, and amino acids, and we suggest that they are potentially capable of producing a new evolutionary

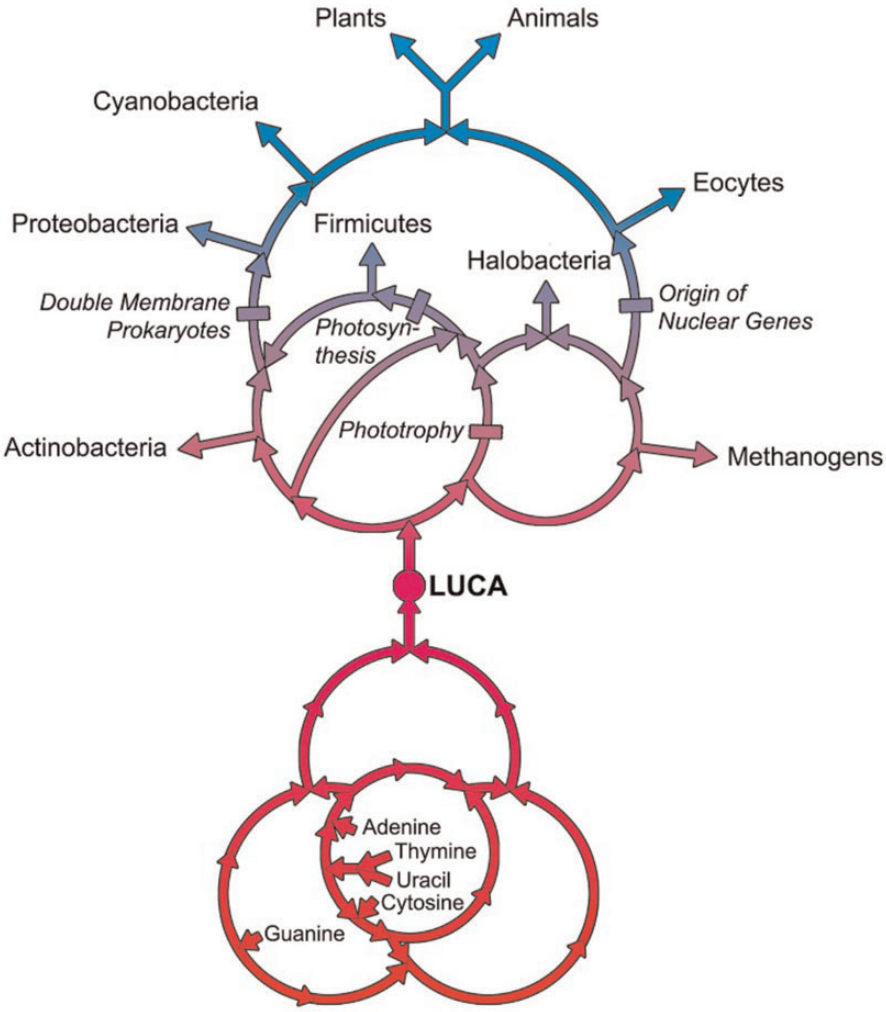


**Fig. 8.**—The phylogenetic relationships between the Rings of Life, LUCA, and the Origin of Life Domain. The Rings of Life, representing the evolution of extant life on Earth, are shown in shades of blue at the top of the figure. These rings are connected to LUCA which is represented by the magenta circle at the center of the figure, and the rings at the bottom of the figure represent the evolution of Adenine, Thymine, Uracil, Cytosine, and Guanine within the Origin of Life Domain.

synthesis of the origin of life. We anticipate that future reconstructions of the OLD will reveal new origin of life pathways and provide a more comprehensive, integrated knowledge of the biochemical-, energetic-, geological-, and genomic-events that occurred during the evolution of life on Earth.

In the past, reconstructing the origin of life appeared to be beyond the reach of phylogenetic analyses and it was universally assumed to be impossible to reconstruct evolution before LUCA. Initially, major advances like the abiotic synthesis of amino acids (Miller 1953) and the RNA world hypothesis (Gilbert 1986) provided hope that someday we could understand early events in the OLD. Over time feasible biosynthetic pathways have been discovered that help explain early events in the origin of life, including the formation of: amino acids (Miller 1953), membranes (Mansy and Szostak 2008), and RNA, protein, and lipid precursors (Patel et al. 2015). We anticipate that the new genomic methods described here will complement those studies, point to new research directions, and allow us to chronologically order additional major biological origin of life transitions. We envision that the synergistic combinations of genomics, biochemical-feasibility experiments, early Earth paleontological studies, and studies on other planets can, and will, greatly accelerate progress in discovering and understanding the evolution of life within the OLD.

## Acknowledgments

Initial support for this work was provided by grants from the National Science Foundation (NSF-0719574 to J.A.L. and NSF-DMS 126 4153 NIH-GM053275 to J.S.S) and from the NASA Astrobiology Institute (NASA Astrobiology DDF R9867-G3 to

J.A.L., through the Georgia Tech Center for Ribosome Adaption and Evolution, Prof. Loren D. Williams, Director).

## Author Contributions

J.L. modified and wrote the programs for downloading the Pfam data and helped calculate the rings; D.-T.T. helped design the figures, calculate rings, produce the final figures, and edit the text. J.S.S. provided statistical advice on performing the reconstructions and extensively edited the article. J.A.L. designed the method for reconstructing the origin of life before LUCA, calculated rings, tested the concept, and wrote the article.

## Literature Cited

Blattner FR, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277(5331):1453.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci USA. 105(51):20356–20361.

Dickerson RE, Geis I. 1969. The structure and action of proteins. New York: Harper & Row. p. 16–17.

Gilbert W. 1986. The RNA world. Nature 319(6055):618.

Koonin EV, Martin W. 2005. On the origin of genomes and cells within inorganic compartments. Trends Genet. 21(12):647.

Lake JA, et al. 2015. Rings reconcile genotypic and phenotypic evolution within the Proteobacteria. Genome Biol Evol. 7(12):3434.

Mansy SS, Szostak JW. 2008. Thermostability of model protocell membranes. Proc Natl Acad Sci USA. 105(36):13351.

Martin W, Baross J, Kelley D, Russell MJN. 2008. Hydrothermal vents and the origin of life. Nat Rev Microbiol. 6(11):805–814.

McInerney JA, O'Connell MJ, Pisani D. 2014. The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nat Rev Microbiol. 12(6):449–455.

Miller SL. 1953. Production of amino acids under possible primitive Earth Conditions. Science 117(3046):528–529.

Patel BH, Percivalle C, Ritson DJ, Duffy CD, Sutherland JD. 2015. Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. Nat Chem. 7(4):301–307.

Powner MW, Gerland B, Sutherland JD. 2009. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. Nature 459(7244):239.

Ramakrishnan V. 2014. The ribosome emerges from a black box. Cell 159(5):979–984.

Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431(7005):152–154.

Simonson AB, Lake JA. 2002. The transorientation hypothesis for codon recognition during protein synthesis. Nature 416(6878):281.

Tamura K, Schimmel P. 2003. Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. Proc Natl Acad Sci USA. 100(15):8666.

Tanaka Y, et al. 2000. Solution structure of an RNA duplex including a C-U base pair. Biochemistry 39(24):7074.

Weiss MC, et al. 2016. The physiology and habitat of the last universal common ancestor. Nat Microbiol. 1(9):16116.

Whitman MC, Coleman D, Wiebe W. 1998. Prokaryotes: the unseen majority. Proc Natl Acad Sci USA. 95(12):6578–6581.

**Associate editor:** Bill Martin