 doi:10.20944/preprints202009.0162.v2

Review

Chaos, order and systematics in evolution of the genetic code

Lei Lei [1] and Zachary F. Burton [2] *

[1] Department of Biology, University of New England, ME, USA; llei@une.edu
[2] Department of Biochemistry and Molecular Biology, Michigan State University; burton@msu.edu
* Correspondence: burton@msu.edu; Tel.: +01-517-881-2243

Abstract: The genetic code evolved by parallel tracks of chaotic and ordered processes. Liquid-liquid phase separation (hydrogels), a chaotic process, constructs diverse membraneless compartments within cells, resulting in regulated hydration and sequestration and concentration of reaction components. Hydrogels relate to chaotic amyloid fiber production. We propose that polyglycine and related hydrogels (i.e. GADV; G is glycine), phase separations, membraneless droplets and amyloid accretions organized protocell domains to drive the earliest evolution of the genetic code and the pre-life to cellular life transition. By contrast, evolution of tRNA, tRNAomes, aminoacyl-tRNA synthetases and translation systems followed highly ordered and systematic pathways, described by well-defined mechanisms and rules. The pathway of evolution of aminoacyl-tRNA synthetases, which tracked evolution of the genetic code, is clarified. Hydrogels and amyloids form a chaotic component, therefore, that complemented otherwise systematic processes. We describe with detail a pre-life world in which hydrogels and amyloids provided the selections of the first life.

Keywords: amyloids; frozen accident; genetic code; hydrogels; liquid-liquid phase separation; mRNA; polyglycine; rRNA; ribosomes; translational fidelity; tRNA.

1. Introduction

Eukaryotic cells divide into compartments. Some compartments are set aside by membranes but others are membraneless and divided instead by regulation of liquid phases [1-5]. Components of membraneless compartments concentrate through local interactions and selection and exclusion of defining components. Hydrogels and liquid-liquid phase separation (LLPS) form these functional units. In human neurological disease and cancer, hydrogel compartments can assemble improperly and, in some cases, lead to generation of amyloid accretions [1,3,6-11]. Although not yet as extensively studied, hydrogels and LLPS are also becoming recognized in prokaryotic systems [12-14]. In this paper, we explore hydrogel and LLPS compartments as drivers of the establishment of the genetic code.

Evolution of life on Earth required a small number of key transitions (Figure 1). In this paper, we concentrate on the pre-life to cellular life transition and the evolution of coding systems, but we use examples from later evolution to highlight very early events. We consider evolution of life on Earth to be a fairly simple outline with overwhelming relevant detail. The first cellular life on Earth is described as LUCA (the Last Universal Common cellular Ancestor), which we consider to be the first organisms with an intact DNA genome and an intact cell [15-18]. The second major transition is the great divergence of Archaea and Bacteria [19-21]. Based on our analyses and those of others, we consider Archaea to be most similar to LUCA and Bacteria to be more diverged. The third major transition is the genetic fusion of multiple Archaea and multiple Bacteria to generate Eukaryota [22-25]. Endosymbiosis of an *α-proteobacterium* taken up by an Asgard Archaea was a key step in eukaryogenesis [26]. The tortured evolutionary path of Eukarya describes the evolution of eukaryote complexity. Subsequently, another endosymbiotic event involving a *cyanobacterium* invading an alga gave rise to plants. Evolution of animal complexity required evolution of a promoter-proximal

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

pausing mechanism of RNA polymerase II and the RNA polymerase II CTD (carboxy-terminal domain) [1,5,27]. The best descriptions of the key stages relate to evolution of biological coding, translation and transcriptional mechanisms. We discuss the importance of hydrogels (liquid-liquid phase separation; LLPS) in transitions. In this paper, when we refer to hydrogels or LLPS, we consider these features in all their complexity, including membraneless organelles and associated amyloids [2,6,28]. It is our opinion that peptide disorder compartmentalized and regulated hydration and caused separation and concentration of reactants in protocells and was a major driving force in the early evolution of life and, specifically, in the evolution of the genetic code, which is the most central feature of evolution of complex life on Earth.

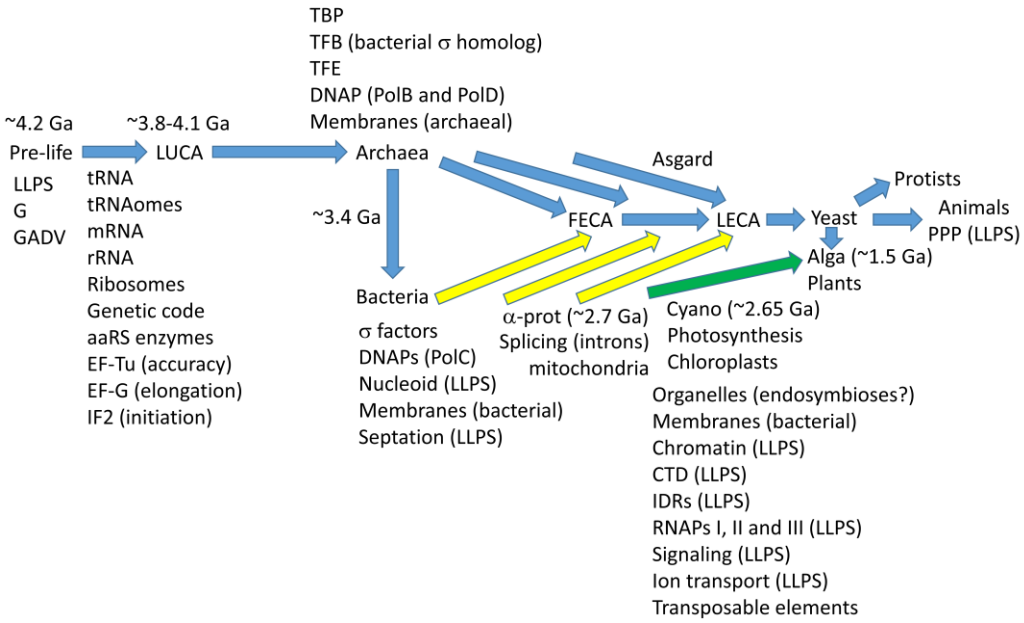


Figure 1. A working outline for the main transitions in evolution of life on Earth and the involvement of hydrogels (i.e. LLPS). PPP) promoter-proximal pausing; cyano) *cyanobacterium*; α-prot) *α-proteobacterium*; TBP) TATA box-binding protein; TFB) Transcription Factor B; TFE) Transcription Factor E; FECA) First Eukaryotic Common Ancestor; LECA) Last Eukaryotic Common Ancestor; RNAP) RNA polymerase; EF) Translation Elongation Factor; IF) Translation Initiation Factor [26,29].

To understand the pre-life to life transition requires bottom-up and top-down approaches [30-34]. In a bottom-up approach, a goal is to develop plausible prebiotic, coacervate and self-replicating polymerization systems. The top-down approach, by contrast, is intended to infer some major pathways in the pre-life world often from analyses of conserved sequences. The advantage of the bottom-up approach is that many prebiotic reactions are interesting and potentially on-pathway. The potential disadvantage of a bottom-up approach is that too many pathways are possible and too many plausible pathways may be dead ends or may not result in dominant pathways. The potential advantage of the top-down strategy is that inferences based on sequence are likely to reflect dominant and successful pathways. The limitation of a top-down strategy is that many important pathways may not be represented or may not be recognized in existing sequence data sets. Because authors of this manuscript are molecular biologists, our approach has been sequence-based and top-down. We find that top-down strategies describe the early evolution of translation systems and transcription systems and the divergence of Archaea and Bacteria. Top-down approaches also enrich bottom-up views. For instance, based on top-down methods, we posit models for pre-biotic chemistry (see below). When top-down approaches and bottom-up approaches meet, a richer analysis of the pre-life to life transition has been achieved.

We posit that the major event in the divergence of Archaea and Bacteria was the evolution of bacterial σ transcription factors (Figure 1) [21,35]. In Bacteria, σ factors bind to RNA polymerase to facilitate binding to the promoter. σ helix-turn-helix (HTH) factors are homologs of archaeal TFB,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

which itself includes two very regular HTH motifs, termed "cyclin-like repeats" [36]. Comparing bacterial σ factors to archaeal TFB, however, σ factors alter bacterial promoter recognition and transcriptional control in fundamental ways. This radical shift in core transcriptional mechanisms, promoters and control caused Bacteria to become significantly different from Archaea, while Archaea remained very similar to LUCA. Bacteria also adopted a new replicative DNA polymerase (PolC) relative to Archaea (PolB and PolD), so this is another fundamental difference comparing Bacteria and Archaea [37].

Much is not yet known about the evolution of Eukaryota. We view eukaryotes as genetic fusions of multiple Archaea and multiple Bacteria without a very clear model for how this transition occurred. We view the transition as a multi-stage process of endosymbiotic or other large horizontal gene transfer events (i.e. by feeding: referred to as "foodchain gene adoption") [25,26,38-40]. It is our opinion that horizontal transfers of small packets of genes were generally less successful than larger transfers. In Figure 1, we indicate a few possible events in the FECA (First Eukaryotic Common Ancestor) to LECA (Last Eukaryotic Common Ancestor) transition, focusing on evolution of cell architectures and transcriptional mechanisms. Splicing appears to have evolved near the time of LECA contributing to genome complexity [41]. It appears that Eukarya evolved a new use for hydrogels (liquid-liquid phase separation; LLPS) involving intrinsically disordered regions (IDRs) of proteins [5,6,42,43]. Histone tails and the carboxy-terminal domain (CTD) of RNA polymerase II are IDRs with regulated interactions and protein readers [1,5,27]. Other factors with IDRs cooperate in the transcription cycle, sequestering complexes in discrete hydrogels. Many LLPS compartments depend upon RNA and RNA-interacting proteins [1-3,8,11,44,45]. Because life appears to have originated from a RNA-dominated world, roles of RNA in ancient evolution are a central consideration. LLPS compartments are disordered with complex interactions and components, so disruption of LLPS compartments is associated with human diseases such as cancers and neurological impairments [6-11,45-47]. So far as we can ascertain, prokaryotes utilize LLPS probably utilizing short disordered protein regions, RNAs and DNA [12-14].

There is some redundancy in this paper compared to previously published work from our laboratory on evolution of tRNAs, tRNAomes, aminoacyl tRNA synthetases (aaRS enzymes) and the genetic code. Because the paper describes and combines multiple complex subjects, some redundancy was inevitable. The current paper provides refined models, insights and perspectives. A dominant theme of this review is that hydrogels, amyloids and LLPS drove the early evolution of the genetic code. Specifically, we posit that hydrogels and related assemblies provided the main driving force behind genetic code evolution. We provide highly detailed and connected models for key intermediates in the pre-life→life transition. Specifically, our tRNA evolution model, aaRS evolution model and genetic code evolution model are mutually reinforcing and highly predictive.

2. Artificial intelligence in evolution of life

At some level, evolution of life on Earth can be described according to principles of artificial intelligence [33,48]. A system is capable of "learning" (teaching itself) if it can build up intellectual property that enhances its subsequent capabilities. To evolve complex biology, a robust genetic code was necessary. Evolution of the genetic code, therefore, was the core bottleneck in evolution of complex life on Earth. As we show, tRNA was the central driver to evolve biological coding. Ligation of minihelices led to evolution of tRNA, and tRNA replaced minihelices in primitive translation systems, because, utilizing tRNA, a robust 3-nt genetic code could evolve. Minihelices, by contrast, supported a mixed set of anticodon-codon interactions (i.e. 1-nt and 3-nt codes; see below), consistent with peptide bond formation but inconsistent with a robust code. We consider tRNA, therefore, to be the core intellectual property driving evolution of translation systems, including: 1) tRNAomes (all of the tRNAs in an organism) [49]; 2) the genetic code [33,48]; 3) mRNA; 4) aminoacyl-tRNA synthetases (aaRS; i.e. GlyRS-IIA; IIA indicates the aaRS I or II structural class and A-E subclass) [32,50]; 5) rRNA; and 6) ribosomes [51]. The system taught itself to code, centered on tRNAs, and then vastly enriched the code and its expression by coevolving proteins. Biological coding expands the capacity of the system to create highly functional proteins and protein assemblies and then to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

evolve complex organisms. To reiterate the computer system comparison, translation systems appear to comprise an operating system for living systems on Earth.

Evolution of tRNA is also a story of artificial intelligence. tRNA evolved from ligation of 3-31-nt minihelices, as described below. Furthermore, the minihelices were comprised of repeating sequences and inverted repeats, so minihelices and tRNAs were constructed from highly ordered ancient sequences from about 4 Ga ago. Many have considered earliest evolution from random biopolymers, but tRNA did not evolve from random sequences, and evolution of tRNA is the central issue in evolution of translation systems and the genetic code. For computational studies, it remains an important question of why and how evolution from ordered sequences gave rise to some of the most central biomolecules.

3. The pre-life to life transition: evolution of translation

According to our vision for evolution of life on Earth, tRNA was the core intellectual property [33,48]. Evolution of tRNA directed evolution of mRNA, rRNA and the genetic code, which evolved around the tRNA anticodon. Evolution of aminoacyl-tRNA synthetases (aaRS), which are the enzymes that attach amino acids to tRNA, tracked the evolution of the genetic code. At least in part, some of the oldest rRNA segments appear to have evolved from amalgamations of tRNAs and tRNA-like sequences, indicating that tRNA predates rRNA, at least in rRNA's current form [52-56]. The primitive ribosome we consider to be a decoding center scaffold (pre-16S rRNA) and a mobile and separate peptidyl transferase center (PTC; pre-23S rRNA) [51]. The PTC appears to be a dehydration chamber for formation of peptide bonds utilizing diverse amino acid substrates [57,58]. We, in part, describe how ribosomes could have evolved to prokaryotic forms before LUCA.

4. Evolution of tRNA

A number of models have been advanced to describe tRNA evolution. We favor the three 31-nt minihelix model advanced by our laboratory (Figure 2), which we find to be best supported by sequence and statistical analyses and also most predictive [49,59-62]. The model fully accounts for the sequences of type I and type II tRNAs. Most tRNA are type I with a 5-nt V loop (V for variable). Type II tRNAs (i.e. $tRNA^{Leu}$ and $tRNA^{Ser}$ in Archaea) have an expanded V loop that initially was 14-nt, although type II V loops have expanded and contracted in evolution. The three 31-nt minihelix model is strongly predictive to describe evolution of the genetic code and evolution of ribosomes. We identify a tRNA primordial sequence ($tRNA^{Pri}$) from which tRNAs radiated. We showed that, in Archaea, $tRNA^{Pri}$ is very close in sequence to $tRNA^{Gly}$, indicating that glycine was the first encoded amino acid [49]. Remarkably, $tRNA^{Pri}$ is very close in sequence to a typical tRNA sequence (similar to a consensus sequence) from Archaea [63].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

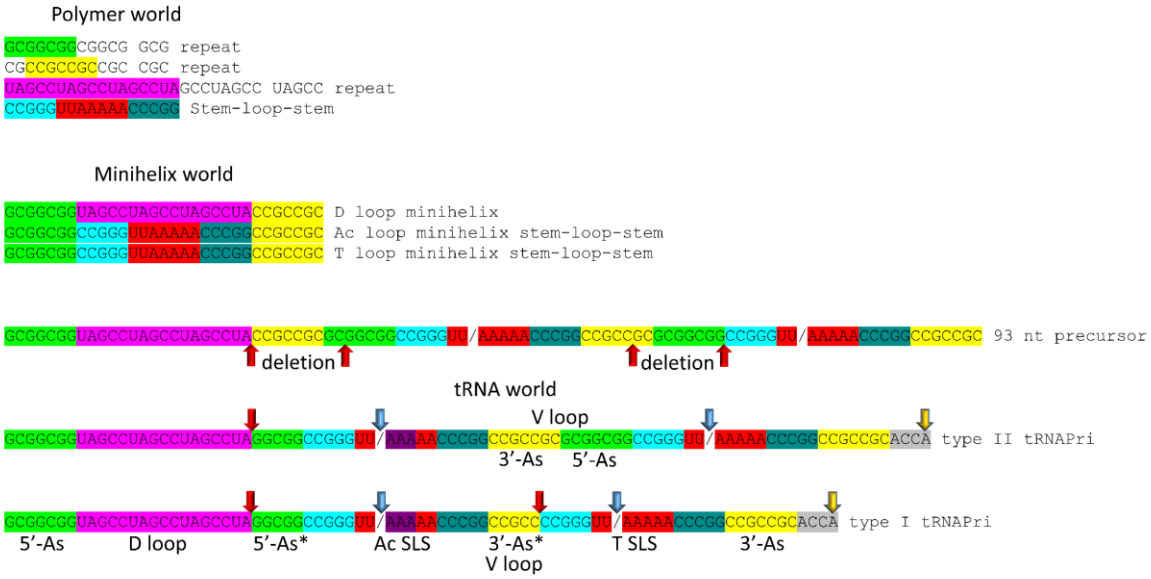


Figure 2. Evolution of tRNA. Polymer world evolved to minihelix world that evolved to tRNA world. tRNA evolved by ligation of three 31-nt minihelices of almost known sequence forming a 93-nt tRNA precursor. Type I tRNA was formed by two internal 9-nt deletions in the 93-nt precursor. Type II tRNA was formed by a single internal 9-nt deletion. Arrow colors: red) deletion endpoints; blue) U turns; gold) site of amino acid attachment. SLS indicates stem-loop-stem. / indicates a U turn. Colors are consistent with Figures 3-5.

4.1. Evolution of type I tRNA

A type I tRNA^{Pri} was formed by ligation of three 31-nt minihelices of almost completely known sequence (Figure 2). Figure 3A shows the structure of a type I tRNA colored according to the model. Figure 3B shows a typical type I tRNA (similar to a consensus sequence) from *Pyrococcus furiosis*, an ancient Archaea [63]. In the model (Figure 2), a 93-nt tRNA precursor was formed by ligation of three 31-nt minihelices. The 93-nt precursor was then processed into type I and type II tRNAs. The 31-nt minihelices that became the anticodon stem-loop-stem and the T stem-loop-stem were initially identical (~GCGGCGGCCGGGUU/AAAAACCCGGCCGCCGC; stem-loop-stem microhelix core: ~CCGGGUU/AAAAACCCGG; there is slight sequence ambiguity in the primordial ~UU/AAAAA (/ indicates a U turn) loop; there is no ambiguity in the 5-nt stems). The minihelix that became the D loop sequence is distinct because it has a 17-nt microhelix core based on a UAGCC repeat (initially GCGGCGGUAGCCUAGCCUAGCCUACCGCCGC; 17-nt UAGCC repeat microhelix core: UAGCCUAGCCUAGCCUA). Remarkably, the UAGCC repeat in the D loop is apparent in typical tRNAs from ancient Archaea (Figure 3B; UAGCNUAGCCUGGUNNA). To generate a type I tRNA^{Pri} requires two internal 9-nt deletions in the 93-nt precursor surrounding the anticodon stem-loop-stem within ligated acceptor stems (type I tRNAs missing 3′-ACCA were initially 75-nt) (Figure 2). In type I tRNA in Archaea, only a few small D loop deletions (i.e. 1-4 nt) and 1-nt deletions in the 5-nt V loop were tolerated [63]. To form a functional tRNA that could attach an amino acid, ligation (or genetic and/or enzymatic attachment) of 3′-ACCA was necessary (Figures 2 and 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

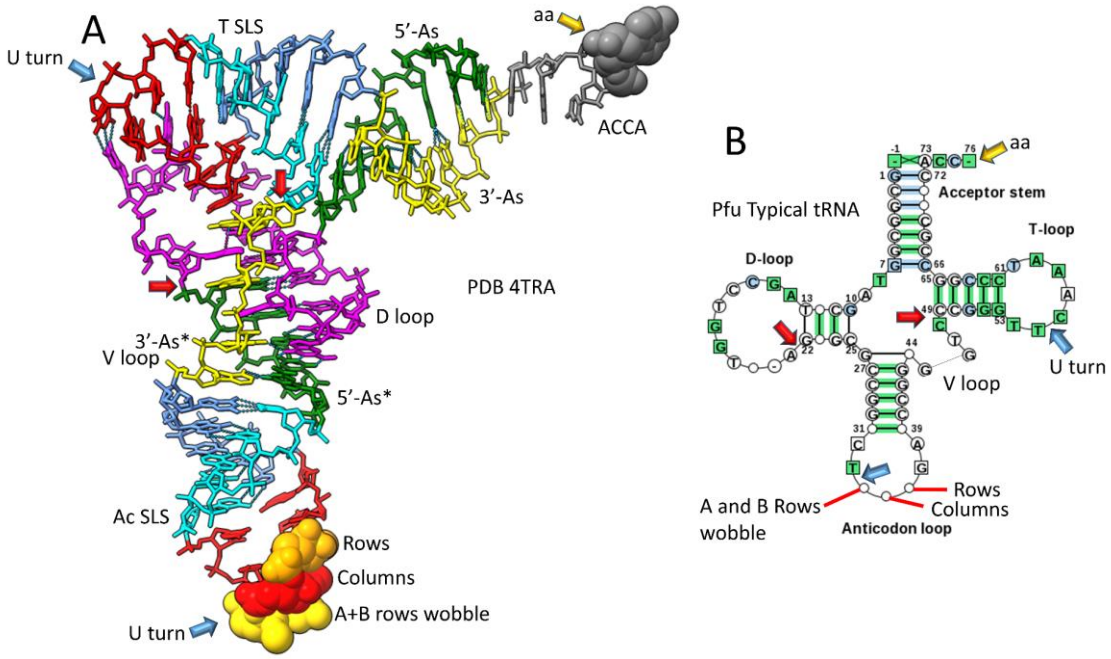


Figure 3. Type I tRNAs. A) $tRNA^{Phe}$ from *Saccharomyces cerevisiae*. B) a typical type I tRNA from *Pyrococcus furiosis* (Pfu) [63]. Molecular graphics was done using the program UCSF ChimeraX [64-66]. Arrows and labels are as in Figure 2. The anticodon is labeled according to the genetic code (see below).

4.2. Archaeal tRNAs radiated from $tRNA^{Gly}$

In support of the three 31-nt minihelix model for tRNA evolution, we show that $tRNA^{Pri}$, $tRNA^{Gly}$ and $tRNA^{Typical}$ are closely related sequences (Figure 4). Figure 4A shows a typical $tRNA^{Gly}$ from three *Pyrococcus* species. Figure 4B shows an annotated multiple sequence alignment of $tRNA^{Pri}$, $tRNA^{Gly}$ and $tRNA^{Typical}$. Despite ~4 Ga (1 Ga = 1 billion years) of evolution, the three sequences are nearly identical [49]. Sequence deviations from $tRNA^{Pri}$ can be explained based on tRNA folding [59].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

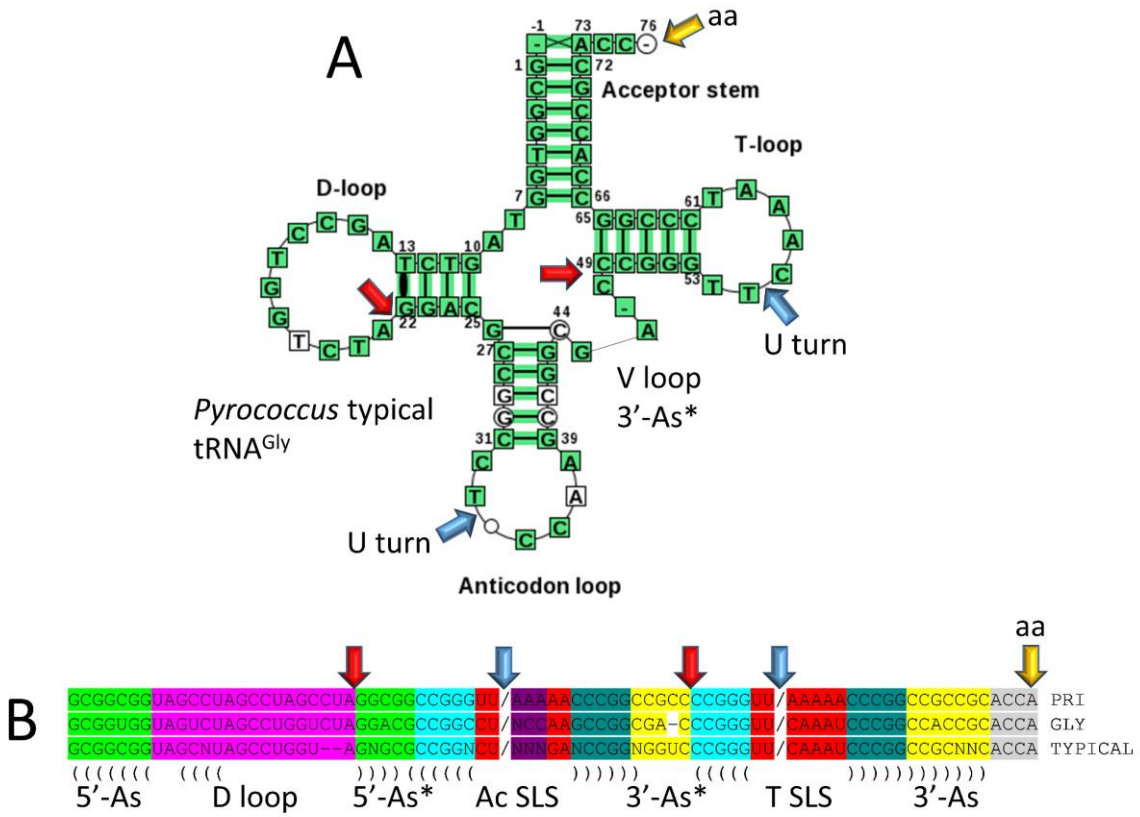


Figure 4. tRNA^{Gly} was the primordial tRNA (tRNA^{Pri}) from which other tRNAs radiated [49]. A) A typical tRNA^{Gly} from 3 *Pyrococcus* species [63]. B) An annotated sequence alignment. PRI) tRNA^{Pri}; GLY) tRNA^{Gly} (as in A); and TYPICAL) tRNA^{Typical} from *Pyrococcus furiosis* (Figure 3B). Arrows are as in Figures 2-5. Purple indicates the anticodon. / indicates a U turn. Other colors are as in Figure 2.

4.3. Evolution of type II tRNAs with an expanded V loop

The same model describes evolution of type I tRNA and type II tRNA with an expanded V loop (Figure 2), indicating that both models are correct (in Archaea, tRNA^{Leu} and tRNA^{Ser} are type II tRNAs) [60]. To generate a type II tRNA^{Pri} required a single internal 9-nt deletion corresponding precisely to the more 5'-deletion in generation of type I tRNA^{Pri} (type II tRNAs were initially 84-nt without 3'-ACCA) (Figure 2). Figure 5A shows a structure of tRNA^{Leu}. The expanded V loop was generated from a 3'-acceptor stem ligated to a 5'-acceptor stem, as indicated in the model (Figure 2). Figure 5B shows a typical tRNA^{Leu} from three ancient archaeal *Pyrococcus* species [63]. In ancient Archaea, tRNA^{Leu} is closer in sequence to a type II tRNA^{Pri} (Figure 2) than tRNA^{Ser} [49]. In Figure 5B an alignment of the primordial type II tRNA V loop and the typical tRNA^{Leu} V loop is shown. The length of the tRNA^{Leu} V loop is 14-nt, as predicted from the model (Figure 2). In Figure 5B, the V loop (numbered V1-V14) was selected to form a G26~UV1 wobble pair and a G15=CV14 reverse Watson-Crick base pair (referred to as the Leavitt base pair), as indicated [60]. The primordial V loop could pair along its entire length. In type II tRNAs, by contrast, the V loop is evolved to form a loop with a short stem. Also, the sequence of the tRNA^{Leu} V loop has diverged from the tRNA^{Ser} V loop, which is a direct determinant for SerRS-IIA serine addition, to avoid tRNA charging errors [67]. The tRNA^{Leu} V loop is evolved to be an anti-determinant for SerRS-IIA [60]. To attach amino acids to tRNAs, we posit that, initially, ACCA was ligated to tRNAs, minihelices, microhelices and other RNAs, utilizing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

a ribozyme ligase. In the ancient world, RNAs covalently linked to amino acids were frequently utilized as substrates in catalysis [32,68]. This mode of catalysis remains apparent today (see below).

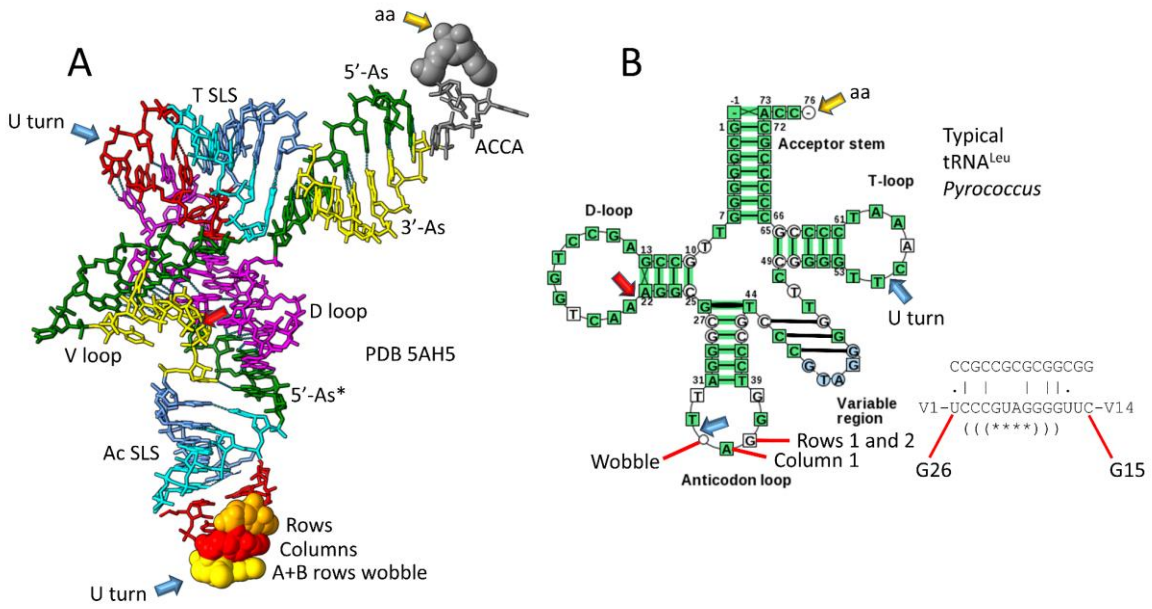


Figure 5. Type II tRNAs. A) A type II tRNA structure ($tRNA^{Leu}$) colored and labeled as in Figure 2. 4-nt in the anticodon loop were missing from the structure, so the anticodon loop shown is from PDB 4TRA ($tRNA^{Phe}$). B) A typical $tRNA^{Leu}$ from three *Pyrococcus* species. The sequence alignment shows a comparison of the typical $tRNA^{Leu}$ V loop to the primordial sequence.

4.4. Demonstration of the model

The evidence for the three 31-nt minihelix model is compelling. For instance, statistical analysis shows p-values of 0.001 (highest indication of homology) for: 1) homology of the anticodon and T stem-loop-stems (17-nt microhelix segments); 2) homology of the last 5-nt of the D loop (5′-As*; As for acceptor stem) and the last 5-nt of the 5′-acceptor stem; and 3) homology of the 5-nt V loop (3′-As*) and the first 5-nt of the 3′-acceptor stem [59,61]. Inspection of the typical tRNA (Figure 3B) is sufficient to confirm the homology of the anticodon stem-loop-stem (27-CCGGNCU/NNNGANCCGG-43) and the T stem-loop-stem (49-CCGGGUU/CAAAUCCCGG-65) (see also Figure 4). Standard tRNA numbering does not match $tRNA^{Pri}$ because standard numbering is based on tRNAs with a 3-nt deletion in the D loop (based on eukaryotic tRNAs). In some ancient Archaea, $tRNA^{Gly}$ (Figure 4A) and $tRNA^{Leu}$ (Figure 5B) have full-length D loops. Homology of the anticodon and T stem-loop-stems, which is obvious from inspection, is sufficient to confirm the three 31-nt minihelix model (Figures 2-5). We showed that the expanded V loop of type II tRNAs was initially a 3′-acceptor stem ligated to a 5′-acceptor stem, as predicted by the model for processing of the 93-nt tRNA precursor (Figures 2 and 5) [60].

We consider these analyses to prove our tRNA evolution model is correct and to falsify alternate models [59]. We showed that tRNAomes in ancient Archaea cluster tightly around $tRNA^{Pri}$ [49]. Ancient Bacteria (i.e. *Thermus thermophilus*) have fairly compact tRNAomes centered on $tRNA^{Pri}$. More derived Bacteria (i.e. *Escherichia coli*) have more diverged and less ordered tRNAomes centered on $tRNA^{Pri}$. Because of internal homologies in tRNA sequences, no accretion model (involving random insertions-deletions; indels) can be correct for tRNA evolution [59]. Other tRNA evolution models (i.e. 2-minihelix and Uroboros) are accretion models with random indels [69,70]. In both the 2-minihelix and Uroboros models, random indels would have to evolve to ordered and repeated sequences in tRNAs [59]. Given rules of genetics, we do not know how this is possible. We do not know of any reasonable predictions that can be made based on the 2-minihelix model or the Uroboros model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

By contrast, we consider the three 31-nt minihelix model to be highly robust and predictive. Given archaeal tRNA sequences, we are unsure how anyone can seriously question the three 31-nt minihelix model (Figures 2-5) [48,49,59-61]. So far, every prediction of the model has been justified (see above). The model describes type I and type II tRNA sequences (Figures 3 and 5) [59-61]. The model accurately describes internal tRNA homologies including sequence repeats and stem-loop-stems (Figures 2-5). The model supports U turn loop (anticodon and T loop) structural similarities [59-61]. The model has also been used to root tRNAome structures [49].

Because tRNA evolution indicates an ancient polymer world and minihelix world preceding the current tRNA world (Figure 2), about 200-300 million years of pre-life evolution are described by the top-down generated three 31-nt minihelix model. Surprisingly, polymer, microhelix, minihelix and tRNA sequences were derived from ordered sequences: repeats and inverted repeats (Figures 2-5). The ancient, pre-life world, therefore, included ordered polymers from which tRNAs evolved. Features of the model derive from tRNA sequences and structure.

5. Evolution of the genetic code (overview)

We posit that the genetic code evolved around the tRNA anticodon following a simple set of rules, which appear never to have been violated [33,48]. For the 2nd and 3rd positions of the anticodon, the rules are C>G>U>>A. Preferences are much stronger for the 3rd anticodon position than for the 2nd anticodon position, because the 2nd anticodon position is most central and, therefore, the easiest to read [51]. Consistent with the rule, however, C is strongly preferred in the 2nd position, just as it is in the 3rd position, as evidenced by the position of glycine in the code (see below). We posit that the genetic code initially sectored on the 2nd anticodon position, because the 2nd position was easiest to read on a primitive ribosome. Essentially, the system was teaching itself to encode proteins by accurately matching and reading codons and anticodons. Furthermore, we posit that, on a primitive ribosome, the 1st and 3rd anticodon positions were initially wobble positions. At a wobble position, only pyrimidine-purine discrimination was initially possible, so tRNA wobbling in translation limited the size of the code. Because of wobbling, tRNA, not mRNA, limited the final size of the genetic code. Considering a genetic code of 64 assignments in mRNA, therefore, is not reasonable. Because of wobbling in the 1st anticodon position, the genetic code has a maximum complexity in tRNA of 32 assignments (2x4x4). To encode additional amino acids, 6- and 4-codon sectors must be split into 2-codon sectors. Because fidelity mechanisms (i.e. aaRS editing and aaRS anticodon recognition) limit the splitting of large codon blocks, the standard genetic code evolved to 20 amino acids plus stops (21 assignments) rather than encoding additional amino acids (up to 32 assignments).

At the 3rd anticodon position, wobbling was abolished by evolution of the elongation factor (EF)-Tu "latch" (also referred to as conformational closing of the 30S ribosome subunit) [33,71-75]. tRNA enters the ribosome bound to the GTPase chaparonin EF-Tu. On the ribosome, EF-Tu holds the tRNA until GTP is hydrolyzed and the 30S ribosome subunit tightens its conformation and the EF-Tu GTPase latch is set. Then EF-Tu dissociates, allowing the verified tRNA with its tightened mRNA codon attachment to rotate its 3'-aa end into the ribosome PTC A site (addition or aminoacyl site). Setting the latch allows 4-base discrimination at the 3rd anticodon position. 4-base resolution was readily achieved at the 2nd anticodon position, because the 2nd anticodon position is most central and the easiest to read. 4-base resolution, therefore, was obtained at the 3rd anticodon position through evolution of the EF-Tu latch. The latch includes *Thermus thermophilus* (Tth) rRNA positions 16S rRNA G530, A1492 and A1493 and 23S rRNA A1913. The latch checks for Watson-Crick pairing to the mRNA codon at the anticodon 2nd and 3rd positions. The latch also checks the accuracy of pairing at the wobble position. Wobbling is necessary to evolve a genetic code based on RNA, and wobbling is a major story in the evolution of the code. The EF-Tu latch was a major determinant of translational accuracy and an essential evolutionary advance in building the code.

At the wobble 1st anticodon position, the sequence preference rule is G>(U~C)>>>>>A [33,48]. Only purine versus pyrimidine discrimination was initially possible at a wobble position. Wobble G appears to be favored over U~C, because Asp (wobble G) appears to have entered the code before Glu (wobble U/C) (see below). A is seldom or never used in the wobble anticodon position in Archaea.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

When wobble A is encoded in Bacteria and Eukarya, A is modified by deamination to inosine [50,76]. Essentially, A is not tolerated in the tRNA anticodon wobble position. Partly, A is not tolerated because A in the tRNA wobble position does not pair well with U in the mRNA wobble position. As noted above, before evolution of the EF-Tu latch, A was also poorly tolerated in the anticodon 3rd position, probably for the same reason. A is not necessary in the anticodon wobble position because G pairs with C (Watson-Crick pairing) and with U (wobble pairing) [77].

In the wobble anticodon position, U and C are read degenerately. Initially, one might expect anticodon wobble C to show reasonable specificity for codon G. Also, anticodon U might be expected to read codon A (Watson-Crick pairing) and codon G (wobble pairing), resulting in anticodon wobble ambiguity. Generally, Archaea use both anticodon wobble C and U tRNAs to encode the same amino acid, indicating that anticodon wobble ambiguity was too high a barrier in evolution to easily separate wobble C and U tRNAs to encode two different amino acids. In principle, such separation of functions might be achieved by tRNA wobble modifications [77,78]. To encode tryptophan, the anticodon CCA is used. The UCA anticodon, however, is not generally utilized because UCA corresponds to the UGA stop codon, which is recognized in mRNA by a protein release factor [79]. To encode methionine, anticodon CAU is utilized to read AUG codons. In Archaea, isoleucine also utilizes CAU with C modified to agmatidine to read only codon AUA (Ile) and not AUG (Met) [80-83]. To avoid ambiguity in coding, anticodon UAU is rarely utilized in Archaea and Bacteria [50]. With very few exceptions, tRNA wobble modifications cause U and C to be read with more ambiguity than expected for an unmodified base. Generally, tRNA wobble modifications support broader reading of synonymous codons rather than evolving higher tRNA specificity in coding [77,78].

We strongly support the concept that the genetic code evolved toward a maximum 32-assignment limit, primarily around the tRNA anticodon [33,48,51,69]. To make sense out of the genetic code, therefore, requires a view centered on tRNA and the tRNA anticodon. By contrast, 64-assignment codes, based on mRNA codons, are not reasonable nor descriptive of the evolutionary process. Below, we describe a detailed pathway for evolution of the genetic code based on these ideas.

6. Evolution of ribosomes

We support the model that rRNA arose from tangles of ligated RNAs that included amalgamations of tRNAs, as also has been proposed by others [52-56]. We imagine an ancient world in which RNAs were replicated by ligation, catalyzed by a ribozyme ligase, followed by complementary replication catalyzed by a template-dependent ribozyme replicase. Attaching a snap-back primer to RNAs would prime their complementary replication. 31-nt minihelices can function as snap-back primers. 17-nt microhelices (i.e. anticodon and T stem-loop-stems) can also function as snap-back primers (Figure 2). Minihelices and microhelices can be removed from larger RNAs via endonucleolytic cleavage of the RNA catalyzed by a ribozyme (i.e. cutting at the base of stems). In such a world, long RNAs with diverse sequences were generated, and some of these could function as a primitive decoding center scaffold and others as a mobile PTC [51,84]. Such tangled RNAs were also an incubator for evolution of novel ribozymes.

The patterns of rRNAs were established before LUCA. One indication of this conclusion is that 16S and 23S rRNAs in Archaea and Bacteria are very similar in sequence and have similar functional RNA motifs [57,58]. Archaeal and bacterial rRNA sequences align essentially over their entire lengths without frequent insertion-deletion. As some examples, in 16S rRNA, both Archaea and Bacteria have similar sequences for: 1) the decoding center; 2) the EF-Tu latch; and 3) the ribosome attachment site. In 23S rRNA, Archaea and Bacteria have similar sequences for: 1) the A-site (addition or aminoacyl site); 2) the P-site (peptidyl site); 3) the EF-Tu latch; and 4) the SRL (sarcin-ricin loop). We conclude, therefore, that 16S and 23S rRNAs were largely established before LUCA and persisted in Archaea and Bacteria with only minor changes and few large insertions-deletions.

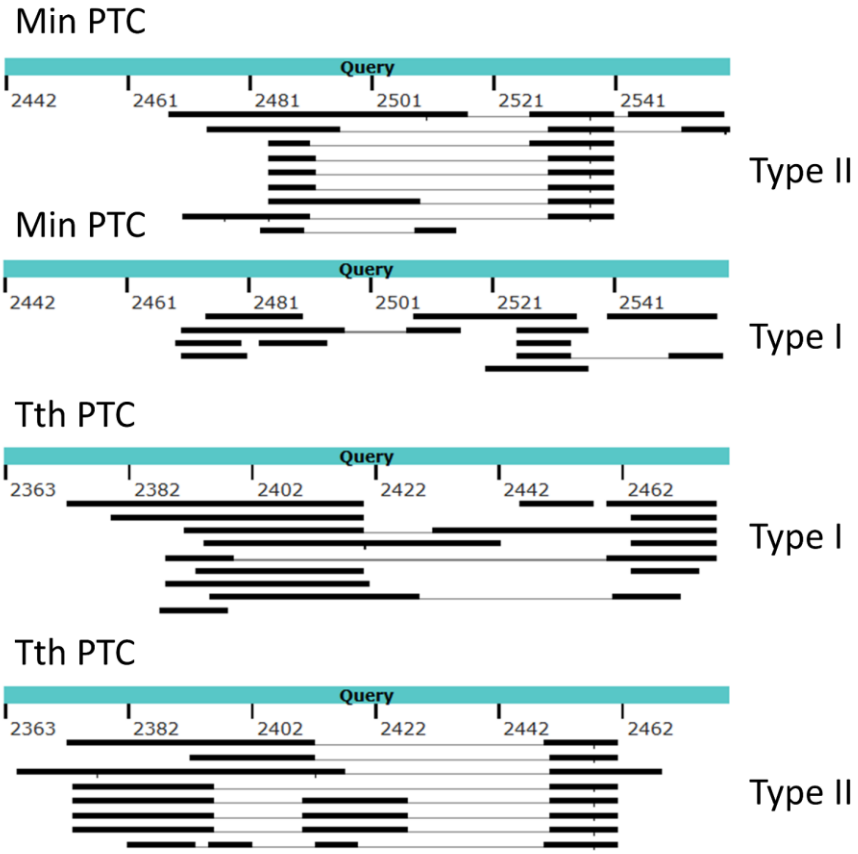
6.1. rRNA may be derived in part from amalgamated tRNAs

In support of the idea that segments of rRNAs may initially have been generated from ligated tRNAs, we show Figure 6. We searched an aligned region of the archaeal and bacterial PTC, located between the tRNA 3'-end CCA-binding segments named the P-loop and the A-loop, using the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

Pyrococcus furiosis (Pfu) tRNAome, which is very similar to a LUCA tRNAome [49]. We searched aligned segments of archaeal (*Methanocaldococcus infernus*; Min) and bacterial (*Thermus thermophilus*; Tth) PTCs (Figure 6A). We find tRNA-like sequences that were identified using multiple tRNA probes that align in both archaeal and bacterial sequences. For this search, the smallest (most likely homologous) e-value obtained was 7×10^{-4} (~1 chance in 1400 of being due to random chance) for an alignment of a Pfu tRNA (Arg (UCU)) to the Tth PTC (Figure 6B). The same region is detected as tRNA-like with aligned tRNA segments in the archaeal Min PTC using multiple Pfu tRNA probes. The alignment appears to extend in the plus/plus orientation from the tRNA D loop across the anticodon stem-loop-stem and a 5-nt (type I tRNA) V loop to the first base of the T loop of the tRNA, indicating that full-length tRNAs rather than minihelices or microhelices (Figure 2) were present for evolution of the PTC. Probably, the homology is to a type I tRNA because it appears to extend over a 5-nt V loop. This same alignment can be obtained using a search with a typical type I tRNA sequence from ancient Archaea. We conclude, therefore, that type I tRNAs probably evolved prior to the 23S rRNA PTC, and tRNA sequences probably contributed to PTC evolution. We have done similar analyses with 16S rRNA and other segments of the 23S rRNA with similar results. We detect both plus/plus and plus/minus alignments to tRNAs, indicating that complementary replication predates evolution of rRNAs. In Figure 6, plus/plus alignments are prominent. Others have reported similar findings using other bioinformatics approaches [52-56]. We note that in the tRNA-aligned segment of the PTC no tRNA-like stem-loop-stems were detected (not shown). Long RNAs tend to fold according to longer range RNA contacts rather than local sequences, so this result was not unexpected.

A



B

```
2373 AGCCGAGCAGGGGCGAAAGC-CGGGCCUAGUGAACCGGUGG 2412 Min 23S PTC
   9 AGCCUAGCCAGGG--AAGGCGCGGGCCUGGAGAGCCCGUGG 47   tRNASer (GGA) e=0.004

2373 AGCCGAGCAGGGGCGAAAGC-CGGGCCUAGUGAACCGGUGGUCCCGUGU 2420 Tth 23S PTC
   9 AGCCUAGCAGGAUAGGGCGC-CGG-CCUUCUAAGCCGGAGGUCCGGGU 55   tRNAArg (UCU) e=7x10-4
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

Figure 6. A tRNA-like segment of the PTC of 23S rRNA. A) Alignments of Pfu tRNA sequences (black bars) to aligned archaeal Min (top) and bacterial Tth (bottom) PTC fragments. Type I and type II tRNAs were searched separately. B) A top alignment in this search that was identified using multiple probes. tRNA Colors: Magenta) D loop; green) 5′-acceptor stem remnant; cyan) 5′-anticodon and T stem; red) anticodon and T loop; purple) anticodon; cornflower blue) 3′-anticodon stem; and yellow) 3′-acceptor stem remnant (V loop). The e-value is dependent on the size of the PTC fragment used in the search, which in this case is short (~117 nt), decreasing the e-value compared to longer PTC fragment searches.

In addition to the decoding center of the ribosome (16S rRNA; 30S subunit), which forms a scaffold on which to run the mRNA, and the PTC (23S rRNA; 50S subunit), at which amino acids are joined to a peptide chain, the ribosome has additional features, which we consider to be subsequent evolutionary add-ons [51]. Remarkably, the ribosome must be coevolved with the genetic code, and a model for evolution of the code parallels these advances, most of which occurred prior to LUCA. Ribosomes, of course, continue to evolve in Eukarya, but these enhancements are generally regulatory to support cell- and organism-specific functions [85,86].

6.2. The Prokaryotic Ribosome

A recent cryo-electron microscopy paper reveals the *Thermus thermophilus* ribosome and its dynamics and fidelity in amazing detail [75]. We highly recommend this paper to any with an interest in general translational mechanisms and fidelity. Here, we provide a general description of the translational mechanism with particular attention to evolution of the EF-Tu GTPase "latch", which we consider to be the fundamental advance in evolution of translation systems and the genetic code. In the paper referenced above, the "latch" is described as conformational closing of the 30S ribosome subunit. Also, of importance is the recognition that IF2, EF-Tu and EF-G are ancient homologous GTPases that function in translation initiation, fidelity, accommodation and translocation.

So far as we can discern, the prokaryotic ribosome was evolved before LUCA and the same basic functional design was maintained in Archaea and Bacteria. Initiation occurs on the small 30S subunit aided by initiation factors (IF1, IF2 and IF3). IF2, elongation factor (EF)-Tu and EF-G are homologous GTPases that function as chaparonins in the translation process [51]. Many Archaea and Bacteria have a UCCU sequence near the 3′-end of the 16S rRNA to orient the sequence ~AGGA on mRNA (the ribosome attachment site) relative to the AUG start codon sequence, which must be positioned in the ribosome P site for translation initiation. Incoming tRNAs first associate with EF-Tu before binding to mRNA. The 16S rRNA (30S subunit) has a "head", "neck" and "body". The head can adjust its rotation to orient the mRNA for initiation and to help by reversible swiveling and mRNA sliding with forward translocation during elongation. The mRNA runs along the neck where it is ratcheted forward via reversible swiveling of the head. The mRNA is held forward in part by bound tRNAs, maintaining the translation register.

For elongation, the 23S rRNA (50S subunit) associates with the 16S rRNA (30S subunit) with bound mRNA and the IFs then dissociate. The aa-tRNA-EF-Tu complex enters, GTP is hydrolyzed, the ribosome latch tightens (the 30S subunit closes) and the aa-tRNA rotates its 3′-XCCA-aa end (X is the discriminator base for aaRS discrimination and amino acid placement on tRNA) into the A-site (addition or aminoacyl site), also aligning peptide-tRNA in the P-site (peptidyl site). During tRNA rotation, EF-Tu dissociates from the A-site aa-tRNA and EF-G binds to the same ribosome site that had been occupied by its homolog EF-Tu during previous steps. EF-G hydrolyzes GTP and stimulates forward translocation. There is limited reversible rotation of the 23S rRNA (50S subunit) versus the 16S rRNA (30S subunit), facilitating forward translocation. tRNAs advance from the A-site to the P-site to the E-site (exit site). Having 2-3 tRNAs bound to the mRNA during elongation helps to maintain the translation frame.

EF-Tu tightens the ribosome "latch", which is a central feature of ribosome evolution [71-74,87]. The latch closes around the aa-tRNA-mRNA helix bound in the ribosome A-site. Enclosure includes interactions with 16S rRNA G530, A1492 and A1493 and 23S rRNA A1913 (Tth numbering). The closed conformation of the ribosome confirms 4-base recognition at the 2nd and 3rd tRNA anticodon

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

positions. Evolution of the latch, therefore, allowed evolution of the genetic code to advance beyond ~8 amino acids (i.e. 2x4 assignments; we posit that only a single wobble position could be read at one time on the primitive ribosome) [33,48]. Prior to evolution of the EF-Tu GTPase latch, both the 1st and 3rd anticodon positions were wobble positions, limited to pyrimidine versus purine resolution. Evolution of the EF-Tu GTPase latch, therefore, "teaches" the ribosome to potentially read a 32-assignment code, which froze at a 20-amino acid + stop codon standard code [16,17,88]. In order for the A-site tRNA to advance to the P-site, the latch must open. Because tRNAs bound in the A-site, P-site and E-site have anticodon interactions paired at the mRNA, associated with the 16S rRNA (30S subunit), and also 3'-end interactions with 23S rRNA (50S subunit), multiple intermediate structures (referred to as hybrid states) are possible [75]. Hybrid states appear to rotate around the EF-Tu latch. Setting of the latch results in dissociation of EF-Tu and is followed by a large rotation of the 3'-end of the verified aa-tRNA into the PTC A-site, a step referred to as "accommodation". Rotation of the deacylated P-site tRNA into the E site, by contrast, is associated with rotation of the 3'-end of the tRNA associated with opening of the latch and forward translocation. Depending on the step, therefore, tRNAs ratchet independently at their 3'-ends and anticodon ends, creating the hybrid states.

For chemistry, a P-site tRNA has XCCA-peptide at its 3'-end (X=the discriminator base). The A-site tRNA has XCCA-aa at its 3'-end. In the 23S rRNA (50S subunit) P-site, the sequence 2248-CUGGGGCGG-2256 presents 2251-GG-2252 to form Watson-Crick pairs with the 3'-CC of the P-site peptide-tRNA. In the 23S rRNA (50S subunit) A-site, the sequence 2548-GGGCUGUUCGCCC-2560 presents 2553-G to pair with 3'-CC (the 2nd C), to orient the A-site aa-tRNA. It appears that proximity of P-site and A-site tRNAs in the dehydrating environment of the PTC may be sufficient to form the next peptide bond [57].

The peptide chain elongates by its transfer to the A-site tRNA, resulting in deacylation of the P-site tRNA. After deacylation, the P-site tRNA can advance its 3'-end to the E site. Because the peptide chain is transferred to the A-site tRNA from the P-site tRNA, the peptide was lengthened by one amino acid. Once the P-site tRNA releases the peptide chain to the A-site tRNA, the deacylated P-site tRNA can then translocate to the E-site, displacing and releasing the E-site tRNA. So the march of tRNAs aided by EF-G through a compact tRNA-shaped tunnel in 23S rRNA helps to ensure forward translocation and maintenance of the translation frame. Because the 23S rRNA apparently evolved to match the shapes of advancing tRNAs, we posit that the final conformation of 23S rRNA and the 50S subunit evolved around tRNAs, and that tRNAs evolved prior to the final evolved shape of the prokaryotic ribosome.

The exiting peptide chain extends from the active site A-peptide (before translocation) or P-peptide site (after translocation), through a channel in the ribosome. When the peptide exits the ribosome, it can begin to fold or it can be targeted to a membrane for transport or excretion. Translation termination occurs when protein release factors bind to stop codons in the mRNA. Because there are no tRNAs corresponding to stop codons, anticodons that are complementary to stop codons are not represented in the tRNA-centric standard genetic code.

The genetic code expanded by two mechanisms we can identify: 1) tRNA charging errors; and 2) modification of amino acids bound to tRNAs (eg. Asp→Asn, Glu→Gln and pSer→Cys (pSer for phosphoserine)). We posit that the first amino acids to enter the code filled large sectors of the code, and these sectors were then invaded by other amino acids. Invasion follows a strict set of rules that we describe in more detail above and below. Significantly, because invasion by incoming amino acids required tRNA charging errors or amino acid modifications, translational fidelity is very important for the eventual freezing of the code. Ribosome fidelity, for instance, evolution of the EF-Tu GTPase latch, was fundamental to first expand and then to freeze the code.

7. Aminoacyl-tRNA Synthetases (aaRS)

We posit an updated model for aaRS evolution [33,48,50,51,89]. Remarkably, the pattern of aaRS evolution matches the pattern of genetic code evolution, providing a pathway for evolution of the genetic code. Also, apparent coevolution of the genetic code and aaRS enzymes indicates that the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

models we present for aaRS enzyme evolution and genetic code evolution are mutually-reinforcing, reliable and predictive. Remarkably, the evidence we cite of coevolution has been maintained during ~4 Ga of evolution with significant potential for divergence. aaRS evolution patterns show coevolution with genetic code columns, which represent the 2nd tRNA anticodon position, the most important position for translational accuracy. As we discuss below, amino acids appear to add into the genetic code by rows, which represent the 3rd anticodon position. Recently, our laboratory clarified aaRS evolution using the Phyre2 protein-fold recognition server, which utilizes sequence and structure to align sequences available in the Protein Data Base with a seed sequence [90]. Phyre2 provides evolutionary relationships of all (or most) aaRS enzymes in a structural class at once. The results provide a road map for evolution of the genetic code. Here, we provide an explanation for the radiation of the aaRS enzymes according to models for genetic code and tRNAome evolution.

7.1 aaRS structural classes

There are two structural classes of aaRS (class I and class II) with multiple structural subclasses (i.e. A-E) [67]. Class I and class II aaRS have incompatible folds but are homologs by sequence (see below). Class I aaRS enzymes have an active site arranged on a set of parallel β-sheets. As a result, class I aaRS have been referred to as a "Rossmann-like" fold. Class I aaRS, however, are not homologs of Rossmann fold proteins. Class II aaRS mount their active sites on a set of antiparallel β-sheets. Both class I and class II aaRS enzymes are among the very first proteins to evolve on Earth, so aaRS are ancient proteins that evolved before LUCA and coevolved with the genetic code. Both class I and class II aaRS enzymes can have an extra editing domain to remove an inappropriately attached amino acid from a tRNA. Remarkably, in Archaea, only amino acids found in the left half of the genetic code (columns 1 and 2) have editing active sites (see below).

Archaeal GlyRS-IIA was the first aaRS enzyme to evolve. Identifying GlyRS-IIA as the primordial aaRS indicates, once again, that glycine may have been the first encoded amino acid and that glycine maintained the dominant position in the evolving code. GlyRS-IIA is a product of protein encoding, so significant evolution of the genetic code must have been supported initially by ribozymes charging tRNAs (i.e. GlyRS-RBZ; RBZ for ribozyme) [16,91-94]. We posit that, because of coevolution, divergence of aaRS enzymes followed the pattern of evolution of the evolving tRNAome, and $tRNA^{Gly}$ held the most favored position in the code. GlyRS-IIA, therefore, evolved as the first protein aaRS, and all class I and class II aaRS diverged from GlyRS-IIA.

A schematic of a multiple sequence alignment comparing GlyRS-IIA, ValRS-IA and IleRS-IA is shown in Figure 7. Red blocks in the alignment indicate sequence similarity. Two Zn motifs are identified in the multiple alignment. One is specific for ValRS-IA and IleRS-IA, mostly within their N-terminal extensions. The other Zn motif is shared among GlyRS-IIA, ValRS-IA and IleRS-IA. Both of these Zn motifs tend to disappear with further divergence from LUCA. In these ancient Archaea, however, the Zn motifs help determine the alternate folding of class I and class II aaRS. Also, the N-terminal extension of ValRS-IA and IleRS-IA includes elements of the active site scaffold, so the N-terminal extension in class IA aaRS helps determine the class IA fold. Similarly, the shared Zn motif organizes a set of surrounding β-sheets that is different in the class IIA and class IA folds [50].

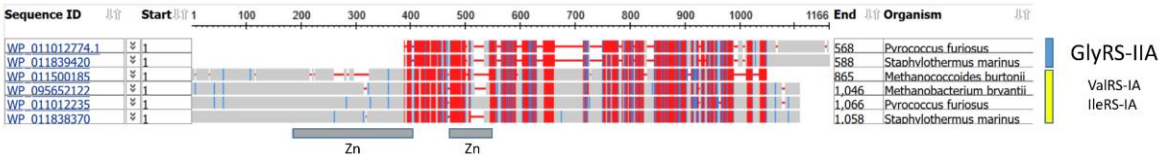


Figure 7. GlyRS-IIA, ValRS-IA and IleRS-IA are homologs by sequence. ValRS-IA and IleRS-IA have a N-terminal extension relative to GlyRS-IIA [50]. Red blocks indicate sequence homology in the multiple alignment. The figure was prepared with NCBI multiple alignment tools.

All class II aaRS enzymes derive in lineage from GlyRS-IIA. To form class IA aaRS enzymes (i.e. IleRS-IA and ValRS-IA), a primitive GlyRS-IIA was extended at its N-terminus by an upstream transcription and translation start and then refolded. Because of the N-terminal extensions, class IA enzymes are longer than GlyRS-IIA. Other class I aaRS enzymes were derived from a class IA aaRS (probably ValRS-IA) [33,48,50,51]. Comparing sequence alignments gave e-values of $3x10^{-13}$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

(*Candidatus Methanoperedens nitroreducens* GlyRS-IIA to *Methanococcoides burtonii* IleRS-IA) and $6x10^{-11}$ (*Methanobacterium congolense* GlyRS-IIA to *Methanobacterium bryantii* ValRS-IA). The chances of these independent alignments of homologous GlyRS-IIA regions being due to random events (i.e. convergent evolution) would be ~1 to 10^{23} against. We conclude that class IIA and class IA aaRS enzymes are homologs by sequence that have subsequently diverged in evolution. Other models (i.e. Carter-Ohno-Rodin) for class I and class II aaRS evolution have been published [95-98], but these models are not correct. In the Carter-Ohno-Rodin model, "urzymes" for class I and class II aaRS were posited to be generated from both strands of a primordial bidirectional gene. Bi-directional genes are very rare, they are found mostly in small phage genomes and they are typically very short in length. By contrast, aaRS genes are long and their overall class I and class II folds are preserved in evolution. The Carter-Ohno-Rodin model is unlikely and inconsistent with simple homology of class I and class II aaRS, as we demonstrate (Figure 7) [50].

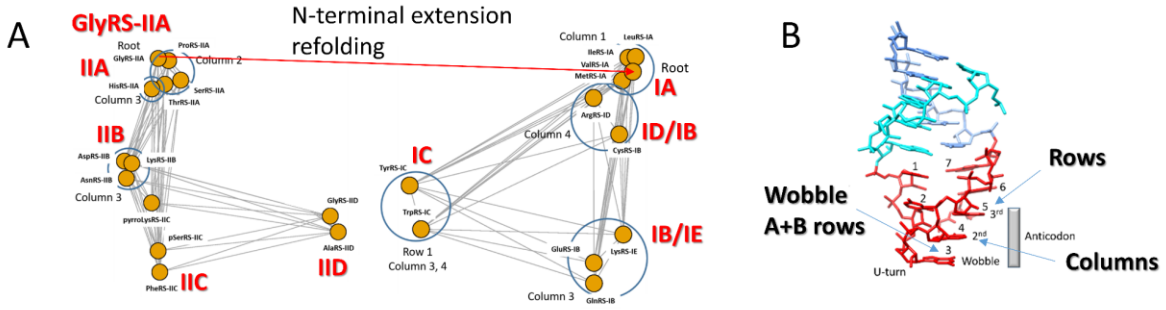
We posit that hydrogels sequestering tRNAs may have been involved in the earliest folding and class divergence of aaRS enzymes. Class IA and class IIA aaRS folding was initially directed by Zn-binding and the N-terminal extension of class IA enzymes, which comprises part of the class I aaRS active site (Figure 7). Because class I and class II aaRS bind opposite faces of their cognate tRNAs, tRNA binding might have also promoted appropriate aaRS folding. Hydrogels can sequester RNAs and cognate tRNAs could have promoted early aaRS class I and class II folds.

7.2. The pattern of aaRS evolution gives the pattern of amino acid placements in the standard genetic code

Figure 8 shows divergence of aaRS enzymes as they relate to the standard genetic code in Archaea. The graph represents the closest homologs in the Protein Data Base identified using the Phyre2 protein-fold recognition server (Figure 8A), so alignments and homology models represent sequence similarity and structural modeling [48,90]. Distances in the map represent evolutionary distances, so clustered aaRS are closely related. Remarkably, all class I aaRS enzymes were connected using the Phyre2 server. Many of these connections were not detected using sequence alignments. By contrast, some of the nodes in the class II aaRS map could not be connected using Phyre2. For instance, no relevant connection of class IIA and class IID enzymes could be obtained.

Other approaches to aaRS evolution have not provided as clear a picture or one so clearly correlated with genetic code evolution [89]. Figure 8B shows how the tRNA anticodon relates to the genetic code. The anticodon 2nd position relates to code columns. The anticodon 3rd position relates to code rows (1-4). The anticodon 1st wobble position relates to A and B rows. In Figure 8C, the standard genetic code (codon-anticodon table) is shown for Archaea with coloring for closely related aaRS enzymes, strongly indicating genetic code evolution within code columns (anticodon 2nd position). Because the genetic code evolved around the tRNA anticodon, and because genetic code evolution is tracked by aaRS evolution, we strongly advocate presenting the code as a codon-anticodon table including aaRS evolutionary data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2



C

row	1st codon position	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	3rd codon position
Column		1		2		3		4		
2nd codon position			U		C		A		G	
1A	U	PHE-IIC	A/GAA	SER-IIA	A/GGA	TYR-IC	A/GUA	CYS-IB	A/GCA	U/C
1B		LEU-IA	U/CAA	SER-IIA	U/CGA	STOP	U/CUA	TRP-IC	U/CCA	A/G
2A	C	LEU-IA	A/GAG	PRO-IIA	A/GGG	HIS-IIA	A/GUG	ARG-ID	A/GCG	U/C
2B		LEU-IA	U/CAG	PRO-IIA	U/CGG	GLN-IB	U/CUG	ARG-ID	U/CCG	A/G
3A	A	ILE-IA	A/GAU	THR-IIA	A/GGU	ASN-IIB	A/GUU	SER-IIA	A/GCU	U/C
3B		MET-IA	U/CAU	THR-IIA	U/CGU	LYS-IE	U/CUU	ARG-ID	U/CCU	A/G
4A	G	VAL-IA	A/GAC	ALA-IID	A/GGC	ASP-IIB	A/GUC	GLY-IIA	A/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IID	U/CGC	GLU-IB	U/CUC	GLY-IIA	U/CCC	A/G

Figure 8. aaRS and standard genetic code coevolution in Archaea. A) aaRS evolution. Distances between nodes indicate evolutionary distance. The red arrow indicates that GlyRS-IIA is homologous to ValRS-IA and IleRS-IA (Figure 7). Structural classes and subclasses of aaRS enzymes are indicated. B) Relationship between the tRNA anticodon and the genetic code. C) The standard genetic code as a codon-anticodon table. Coloring of aa-aaRS is meant to stress evolutionary relatedness mostly in genetic code columns. Ac for anticodon. Grey shading indicates aaRS enzymes with editing active sites in Archaea. Versions of this figure were previously published and the figure is reprinted here with permission [33,48].

7.3. A model for Code Sectoring based on aaRS coevolution

We posit that the genetic code coevolved with aaRS enzymes and tRNAomes and that a record of that coevolution is maintained in the pattern of aaRS divergence and the distributions of amino acids in the code. Here, we posit models for the sectoring of the genetic code correlated with aaRS evolution in Archaea and for modifications of the model in Bacteria [33,48]. In Figure 1, we indicate that Bacteria may have been derived from Archaea [26,29]. As a first consideration, most genetic code sectoring is within code columns, indicating powerful coevolution of the genetic code, aaRS enzymes and the 2nd anticodon position of tRNA. In column 1 (2nd anticodon position A), valine, leucine and isoleucine are hydrophobic amino acids, and ValRS-IA, MetRS-IA, IleRS-IA and LeuRS-IA are all closely related class IA aaRS enzymes (Figure 8). From metabolic pathways, valine can be converted to leucine in 5 enzymatic steps. We posit that this conversion may have initially occurred with valine bound to tRNA. So, Val-tRNA^{Leu}→Leu-tRNA^{Leu} (catalyzed by 5 enzymes) prior to evolution of LeuRS-IA, which we posit was derived from ValRS-IA after duplication. PheRS-IIC may have been derived from a similar enzyme to pSerRS-IIC (pSer for phosphoserine) [99-101]. pSerRS-IIC was probably the route by which cysteine was first introduced into the genetic code (see below). To suppress translation errors, the aaRS enzymes in genetic code column 1 have separate editing active sites to remove an inappropriately attached amino acid [67], further demonstrating their similarity and their evolution in genetic code columns. Significantly, in Archaea, only amino acids found on the left half of the genetic code (columns 1 and 2) utilize aaRS enzymes with editing active sites (Figure 8C) [33,48,50]. SerRS-IIA is a partial exception. SerRS-IIA has an editing active site and is found in columns 2 and 4 of the code [67].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

In column 2 (2nd anticodon position G), serine and threonine are similar amino acids, and SerRS-IIA, ProRS-IIA and ThrRS-IIA are closely related class IIA aaRS enzymes. We posit that AlaRS-IID may have been derived from a similar enzyme to pSerRS-IIC, i.e. by duplication, mutation and re-purposing. Probably, AlaRS-IIA was replaced by AlaRS-IID early in code evolution (i.e. before LUCA), in order to enhance the fidelity of tRNA charging. SerRS-IIA, ThrRS-IIA and AlaRS-IID have editing active sites. In Archaea, AlaX is a tRNA editing function homologous to AlaRS-IID but without a synthetic active site to add alanine to $tRNA^{Ala}$. We consider these observations to strongly support coevolution of amino acids, aaRS enzymes and the genetic code within column 2.

In column 3 (2nd anticodon position U), aspartate and asparagine are related amino acids, and AspRS-IIB, AsnRS-IIB and HisRS-IIA are reasonably closely related enzymes. We posit that AspRS-IIB was initially AspRS-IIA from which HisRS-IIA was derived (see below). AsnRS-IIB was derived from AspRS-IIB. In some ancient Archaea, Asp-$tRNA^{Asn}$, charged by AspRS-IIB, is converted enzymatically to Asn-$tRNA^{Asn}$ by Asp-$tRNA^{Asn}$ amidotransferase, indicating an important mechanism for evolution of the genetic code through modification of amino acids bound to tRNAs [102-106]. Of course, aspartate and glutamate are closely related amino acids, drawing a further linkage of most of the amino acids in column 3. In column 3, glutamate and glutamine are closely related amino acids and GluRS-IB, LysRS-IE and GlnRS-IB are closely related enzymes. The structural classification of LysRS-IE is deceptive (Figure 8A). Despite the different sub-classifications, LysRS-IE is very similar to GluRS-IB and GlnRS-IB. In some ancient Archaea, Glu-$tRNA^{Gln}$, charged by GluRS-IB, is converted to Gln-$tRNA^{Gln}$ by modification utilizing a Glu-$tRNA^{Gln}$ amidotransferase, indicating an evolutionary intermediate leading to replacement with GlnRS-IB [102-106]. Tyrosine and TyrRS-IC are late additions to the genetic code. No column 3 aaRS enzymes in Archaea have editing active sites. In Bacteria, LysRS-IIB replaced archaeal LysRS-IE. In Bacteria, LysRS-IIB edits [67]. Bacterial LysRS-IIB appears to be derived from AspRS-IIB, further indicating evolution within code columns, even when an aaRS appears to have been replaced in evolution. Surprisingly, there appears to be very little chaos in evolution of the genetic code.

In column 4 (2nd anticodon position C), a jumble of amino acids and aaRS enzymes is found. Archaeal GlyRS-IIA is the aaRS enzyme from which all class II and class I aaRS enzymes were derived (Figure 8A). Class I aaRS enzymes were generated by refolding an ancient GlyRS-IIA probably to ValRS-IA (Figure 7). In Bacteria, GlyRS-IID replaced archaeal GlyRS-IIA. GlyRS-IID, in Bacteria, is probably derived from AlaRS-IID, which arose prior to LUCA (Figure 8A). Using Phyre2, no direct homology was detected linking class IIA and class IID aaRS enzymes (without intermediates), indicating that GlyRS-IID and AlaRS-IID were reinvented. We posit that, in Bacteria, archaeal GlyRS-IIA and, before LUCA, AlaRS-IIA were replaced in evolution to increase translational fidelity. ArgRS-ID and subclass IA enzymes are closely related despite the sub-classification of ArgRS-ID. ArgRS-ID is also closely related to CysRS-IB, and cysteine and arginine are found in nearby sectors, in column 4. In some ancient Archaea, pSer-$tRNA^{Cys}$ is converted to Cys-$tRNA^{Cys}$ by Sep-tRNA:Cys-tRNA synthase (Sep for phosphoserine), indicating how cysteine first entered the genetic code and how CysRS-IB arose [99,101]. Cysteine was needed in proteins from an early time in evolution to ligate metals [18]. Subsequently, CysRS-IB could have evolved from ArgRS-ID to charge $tRNA^{Cys}$ directly. Tryptophan and TrpRS-IC are posited to be the final additions to the genetic code [107,108]. TrpRS-IC was probably derived from TyrRS-IC. We posit that serine invaded column 4 of the genetic code by jumping from column 2 (see below).

7.4. aaRS accuracy

The genetic code evolved around the tRNA anticodon. To an extent, this coevolution relates to aaRS enzymes recognizing the tRNA anticodon as a direct determinant to accurately place an amino acid on the cognate tRNA. Exceptions to this general rule, however, are also of interest for understanding evolution of the code (see below). The accuracy of amino acid placement by aaRS enzymes is a complicated issue that is only addressed briefly here. Because tRNAs are so similar in form and sequence, subtle determinants and anti-determinants are recognized by aaRS enzymes. As examples, aaRS enzymes may recognize some of the following determinants and anti-determinants in tRNAs: 1) the discriminator base; 2) the acceptor stems; 3) the anticodon; 4) the V loop; and 5) the D loop [67]. Generally, class I and class II aaRS enzymes recognize opposite faces of the tRNA, so

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

they may recognize different features as determinants and/or anti-determinants for discrimination and amino acid placement. The active site of the aaRS also has particular properties that accept the appropriate amino acid and reject incorrect substrates. For instance, the size of the active site pocket is appropriate for the amino acid substrate rejecting larger substrates. Amino acids with greater character, i.e. charge, hydrogen bonding and flexibility or rigidity, tend to more easily be discriminated in the aaRS active site. Hydrophobic and neutral amino acids, by contrast, are associated with aaRS enzymes with editing active sites. Remarkably, aaRS enzymes that edit were largely restricted to the left half of the genetic code (columns 1 and 2). Also, hydrophobic and neutral amino acids that, generally, require aaRS editing are found in columns 1 and 2.

The aaRS enzymes that lack tRNA anticodon recognition include AlaRS-IID, LeuRS-IA and SerRS-IIA [67]. Significantly, these aaRS enzymes that lack anticodon recognition have editing active sites to suppress charging errors. In ancient Archaea, the editing function of AlaRS-IID is supplemented by AlaX enzymes that edit inappropriately aminoacylated $tRNA^{Ala}$ but lack an active site to add alanine. Apparently, a different evolutionary route was taken to support the accuracy of alanine charging on $tRNA^{Ala}$. Probably, AlaRS-IID evolved to replace a now extinct AlaRS-IIA to reduce $tRNA^{Ala}$ charging errors. Remarkably, $tRNA^{Leu}$ and $tRNA^{Ser}$ are the only type II tRNAs in the archaeal standard code. Furthermore, leucine, serine and arginine are the only amino acids with 6-codon sectors. Below, we propose a model to explain the evolution of the 6-codon sectors. In 6-codon sectors, multiple columns (serine) and rows (leucine, serine and arginine) are crossed. Because this causes ambiguities reading the anticodon, other strategies for tRNA discrimination became necessary. Recognition of the expanded type II V loops, for instance, aids $tRNA^{Leu}$ and $tRNA^{Ser}$ discrimination. Arginine is a large, stiff amino acid with fairly unique hydrogen-bonding potential, so ArgRS-ID active site specificity for arginine and other $tRNA^{Arg}$ determinants largely describes the specificity of $tRNA^{Arg}$ charging. ArgRS-ID does not edit. Also, ArgRS-ID does recognize the $tRNA^{Arg}$ anticodon 2nd position. In Archaea, only SerRS-IIA on the right half of the code (column 4) edits, and SerRS-IIA is also found, and probably initially resided, in the left half of the code (column 2). Other than SerRS-IIA, only column 1 and 2 aaRS enzymes edit in Archaea.

8. Pre-life to LUCA

The pathway of the pre-life to life transition on Earth is largely unknown [18,109]. Our contention has been that the key advance in evolving to cellular life was evolution of tRNA, leading to evolution of tRNAomes, the genetic code and translation systems [33,49,50]. As a guesstimate, we consider the evolution of the code to be a "frozen accident" [48,88,110] that might have taken place over about 200-300 million years. To be more accurate, the code was established systematically rather than accidentally. The code was "frozen" by translational fidelity mechanisms. According to our view, evolution of the genetic code was the dominant pathway to enable life, making other metabolic, energy and motor pathways [111-113] of potentially secondary importance. We have published detailed models that we consider to be highly informative and reliable for evolution of the genetic code [33,48].

The genetic code evolved around the tRNA anticodon (Figure 8B), and tRNA evolved from a highly patterned primordial sequence that is known almost to the last nucleotide (Figures 2-5) [59,61]. The patterning includes sequence repeats and inverted repeats (stem-loop-stems), with specialized 7-nt U turn loops. The pre-life world, therefore, was capable of accurately producing repeating RNA sequences. At a minimum, GCG repeats (5'-acceptor stems), CGC repeats (3'-acceptor stems) and UAGCC repeats (D loop microhelix) were generated. Because the pre-life world generated 31-nt minihelices (Figure 2), a capacity to "measure" the truncations of repeat units must also have existed. The pre-life world must have been capable of complementary RNA replication, otherwise inverted repeats found in tRNA would not be notable. So, before evolution of the first tRNA, ribozymes must have existed to: 1) generate RNA repeats; 2) to accomplish complementary replication; and 3) to excise functional RNAs from longer RNAs.

According to our view, the code in mRNA evolved from the code in tRNA, as we have described [33,48]. This conclusion follows from the hypothesis that the genetic code evolved initially to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

synthesize polyglycine [33,48-51,60]. Adoption of this model, yielded the following insight. The genetic code appeared to have evolved by filling in large sectors of the code, which were then invaded by incoming amino acids. Because the code initially encoded only polyglycine, $tRNA^{Gly}$ was the first tRNA. Essentially, all anticodons must then have mutated from $tRNA^{Pri}$, which is a primitive $tRNA^{Gly}$, to all possible sequences. This is easy to imagine. The anticodon loop is exposed in tRNA, so the anticodon could mutate without affecting tRNA structure. Mutations in other positions of the anticodon loop (i.e. loop positions 1, 2, 6 and 7; 3-5 is the anticodon), by contrast, may disrupt the 7-nt anticodon loop conformation, which has a characteristic U-turn between loop positions 2 and 3 [114]. If all anticodons encoded glycine, all mRNA encoded polyglycine. Remarkably, in ancient Archaea, the $tRNA^{Pri}$ (the primordial tRNA) is most closely related to $tRNA^{Gly}$ (Figure 4). As newly added amino acids invaded, displaced amino acids retreated, retaining the most favored anticodons and surrendering less favorable anticodons to the invader. According to this model, the entire genetic code can be populated, as the standard genetic code was populated and subsequently maintained for ~4 Ga.

The rules for anticodon preference are as follows. In the 2nd and 3rd anticodon positions, the preferences are C>G>U>>A. These preferences are most apparent in the 3rd anticodon position, rather than the 2nd position, which was easier to read on the primitive ribosome. In the 1st anticodon position (the wobble position), the preference is G>(U~C)>>>>>A. In Archaea, A is strongly disfavored in the anticodon wobble position, and A is rarely or never encoded. In Bacteria and Eukarya, wobble A can be modified by deamination to inosine [50,76]. In the wobble position, only purine/pyrimidine resolution was achieved [33,48].

Evidence for this model includes the following. In Archaea, $tRNA^{Gly}$ is closest in sequence to $tRNA^{Pri}$ (a primordial tRNA) (Figure 4). Archaeal GlyRS-IIA is the primordial aaRS from which all aaRS enzymes radiated (Figure 8). Glycine, which is the first amino acid encoded, retains the best anticodons (2nd and 3rd anticodon position C). Glycine, alanine, aspartic acid and valine appear to be the first four encoded amino acids [16,17,31,33,48], and they occupy the most favored row 4 (3rd anticodon position C). We posit that aspartic acid entered the code before glutamic acid, and Asp retained the preferred anticodon (1st anticodon position G (Asp) appears to be preferred over 1st anticodon U/C (Glu)). Some of the last amino acids to enter the code occupy disfavored 3rd position A (Phe, Tyr, Cys, Trp). Stop codons, which are read in mRNA by protein release factors [79], occupy disfavored row 1 (3rd position A).

9. Polyglycine World

We posit that the genetic code initially evolved to encode polyglycine. First of all, $tRNA^{Pri}$ is almost a $tRNA^{Gly}$ sequence in ancient Archaea (Figure 4). An archaeal typical tRNA sequence (similar to a consensus sequence) is essentially a $tRNA^{Gly}$, indicating that other archaeal tRNAs radiated from $tRNA^{Gly}$ [49]. Glycine is the simplest amino acid, and glycine was present early on Earth [115]. GlyRS-IIA in Archaea is the root of both the class I and class II aaRS trees (Figure 8A), indicating that GlyRS-IIA was the first aminoacyl-tRNA synthetase. We posit that when aaRS ribozymes were replaced by encoded protein enzymes, GlyRS-IIA was first because glycine occupied the dominant position in the code.

We further posit that the prior minihelix world that existed before tRNA world (Figure 2) also evolved to synthesize polyglycine. In tRNA world, two 31-nt minihelix sequences were preserved. We posit that, before tRNA, numerous other 31-nt minihelices with ligated 3'-ACCA supported polyglycine synthesis. Polyglycine appears to have been of strongly selected value in the prebiotic world. Interestingly, hemolithin (a polyglycine/hydroxyglycine polymer with coordinated metals) has been identified in meteor samples. Hemolithin appears to be a pre-biotic modified polyglycine transported from outer space that was not genetically encoded [115]. We consider identification of hemolithin to be evidence of a polyglycine world before evolution of biotic systems.

We posit selections for pre-biotic polyglycine that subsequently drove evolution of the genetic code. First, polyglycine could function as a hydrogel in forming LLPS droplets [28,116-118]. Also, polyglycine can form amyloid accretions that could act in a protocell as modified hydrogels [119].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

Furthermore, polyglycine can be a cross-linking agent to stabilize protocell internal and external architectures [120-122]. For instance, polyglycine (i.e. Gly_5) is a component of bacterial peptidoglycan cell walls. Cell walls include long glycan chains (i.e. [N-acetyl-glucosamine-N-acetylmuramic acid]$_n$) with covalently attached short peptides (i.e. L-Ala-D-Glu-L-Lys-D-Ala). In peptidoglycan, Gly_5 can cross-link D-Ala and L-Lys in two nearby short glycan-linked peptides.

In the ancient world, polyglycine could have functioned as a hydrogel to enhance protocell chemistry. In this regard, we note that some human transcription factors include long polyglycine tracts. Human transcription is highly dependent on hydrogel (LLPS) compartments [1,2,5,27]. One example is the human androgen receptor, which includes a Gly_{23} tract. Forkhead Box Protein F1 has a Gly_{11} tract. zinc finger homeobox protein 3 has a long polyglycine tract. Alpha-fetoprotein enhancer binding protein, AT-rich interactive domain-containing protein and SWI/SNF chromatin remodeling complex subunit OSA2 are other examples. UNC-80 (ion transport) and phosphatidylinositol 4-kinase (signaling) also have polyglycine tracts and may rely on hydrogels for functional compartmentalization. We posit that these human factors could be models for the functions of polyglycine in ancient systems. In studying ancient evolution, we posit that polyglycine included in protocell systems will improve their coacervate properties and structural stability. Silk fibroin is glycine-rich and includes polyalanine tracts [123]. Fibroin forms β-sheet amyloid-like assemblies and can form hydrogels. We posit that protocell systems packed with tRNAs, polyglycine, short peptides, large RNA assemblies and other early metabolites will be shown to have enhanced activities. Advancing to a GADV (Gly, Ala, Asp, Val) world, of course, would enhance the potential for forming hydrogels and related compartments (see below).

10. Evolution of the Genetic Code (a working model)

Figure 9 shows a proposed order for the entry of amino acids into the genetic code. Figure 10 gives a highly detailed model for evolution of the code, considering anticodons, codons and aaRS enzymes [33,48,50]. We posit that glycine was the first encoded amino acid. Then, the code sectored on the 2nd anticodon position to encode Gly, Ala, Asp and Val [16,17,31,124-130]. The 8-aa code may have encoded Gly, Arg, Asp, Glu, Ala, Ser, Val and Leu. The 8-aa code represents a bottleneck in evolution because the EF-Tu GTPase latch was necessary to push the code beyond 2x4 complexity (1 wobble (1st or 3rd anticodon position) + 2nd anticodon position). At the ~16 amino acid stage, we posit that the code may have included Gly, Arg, Asp, Glu, Asn, Gln, His, Lys, Ala, Thr, Pro, Ser, Val, Ile and Leu. From this stage, the standard genetic code evolved, mostly by filling row 1 (disfavored 3rd anticodon position A), which was the most difficult row to fill. Our proposed order for amino acid entry is very similar to models proposed by others [16,17,31,126-130]. According to our model, the code transitions from the simplest amino acids to more complex amino acids, indicating that amino acid metabolism and the genetic code co-evolved. Our model incorporates negatively charged and positively charged amino acids relatively early, in part, to evolve more complex proteins. Our model incorporates aromatic amino acids last, consistent with their late evolution, as proposed by others [31,108].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

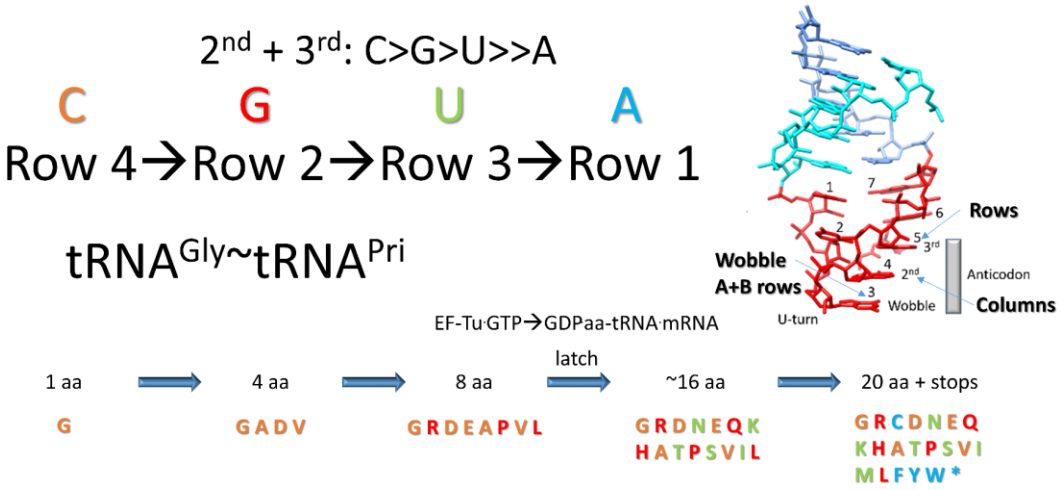


Figure 9. Proposed order of addition of amino acids to the genetic code. The code appears to fill by rows (anticodon 3rd position). Amino acids in 6-codon sectors that occupy more than one row (Leu, Ser and Arg) were scored on their most favored row. The asterisk indicates a stop codon.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

		1		2		3		4		column
		U		C		A		G		2nd codon position
1A	U	STOP	A/GAA	STOP	A/GGA	STOP	A/GUA	STOP	A/GCA	U/C
1B		STOP	U/CAA	STOP	U/CGA	STOP	U/CUA	STOP	U/CCA	A/G
2A	C	GLY-RBZ	A/GAG	GLY-RBZ	A/GGG	GLY-RBZ	A/GUG	GLY-RBZ	A/GCG	U/C
2B		GLY-RBZ	U/CAG	GLY-RBZ	U/CGG	GLY-RBZ	U/CUG	GLY-RBZ	U/CCG	A/G
3A	A	GLY-RBZ	A/GAU	GLY-RBZ	A/GGU	GLY-RBZ	A/GUU	GLY-RBZ	A/GCU	U/C
3B		GLY-RBZ	U/CAU	GLY-RBZ	U/CGU	GLY-RBZ	U/CUU	GLY-RBZ	U/CCU	A/G
4A	G	GLY-RBZ	A/GAC	GLY-RBZ	A/GGC	GLY-RBZ	A/GUC	GLY-RBZ	A/GCC	U/C
4B		GLY-RBZ	U/CAC	GLY-RBZ	U/CGC	GLY-RBZ	U/CUC	GLY-RBZ	U/CCC	A/G

A

1A	U	STOP	A/GAA	STOP	A/GGA	STOP	A/GUA	STOP	A/GCA	U/C
1B		STOP	U/CAA	STOP	U/CGA	STOP	U/CUA	STOP	U/CCA	A/G
2A	C	VAL-RBZ	A/GAG	ALA-RBZ	A/GGG	ASP-RBZ	A/GUG	GLY-RBZ	A/GCG	U/C
2B		VAL-RBZ	U/CAG	ALA-RBZ	U/CGG	ASP-RBZ	U/CUG	GLY-RBZ	U/CCG	A/G
3A	A	VAL-RBZ	A/GAU	ALA-RBZ	A/GGU	ASP-RBZ	A/GUU	GLY-RBZ	A/GCU	U/C
3B		VAL-RBZ	U/CAU	ALA-RBZ	U/CGU	ASP-RBZ	U/CUU	GLY-RBZ	U/CCU	A/G
4A	G	VAL-RBZ	A/GAC	ALA-RBZ	A/GGC	ASP-RBZ	A/GUC	GLY-RBZ	A/GCC	U/C
4B		VAL-RBZ	U/CAC	ALA-RBZ	U/CGC	ASP-RBZ	U/CUC	GLY-RBZ	U/CCC	A/G

B

1A	U	STOP	A/GAA	STOP	A/GGA	STOP	A/GUA	STOP	A/GCA	U/C
1B		STOP	U/CAA	STOP	U/CGA	STOP	U/CUA	STOP	U/CCA	A/G
2A	C	LEU-IA	A/GAG	SER-IIA	A/GGG	ASP-IIA	A/GUG	ARG-IA	A/GCG	U/C
2B		LEU-IA	U/CAG	SER-IIA	U/CGG	GLU-IA	U/CUG	ARG-IA	U/CCG	A/G
3A	A	VAL-IA	A/GAU	ALA-IIA	A/GGU	ASP-IIA	A/GUU	GLY-IIA	A/GCU	U/C
3B		VAL-IA	U/CAU	ALA-IIA	U/CGU	GLU-IA	U/CUU	GLY-IIA	U/CCU	A/G
4A	G	VAL-IA	A/GAC	ALA-IIA	A/GGC	ASP-IIA	A/GUC	GLY-IIA	A/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IIA	U/CGC	GLU-IA	U/CUC	GLY-IIA	U/CCC	A/G

C

1A	U	STOP	A/GAA	STOP	A/GGA	STOP	A/GUA	STOP	A/GCA	U/C
1B		STOP	U/CAA	STOP	U/CGA	STOP	U/CUA	STOP	U/CCA	A/G
2A	C	LEU-IA	A/GAG	SER-IIA	A/GGG	ASP-IIA	A/GUG	ARG-IA	A/GCG	U/C
2B		LEU-IA	U/CAG	SER-IIA	U/CGG	GLU-IA	U/CUG	ARG-IA	U/CCG	A/G
3A	A	LEU-IA	A/GAU	SER-IIA	A/GGU	ASP-IIA	A/GUU	ARG-IA	A/GCU	U/C
3B		LEU-IA	U/CAU	SER-IIA	U/CGU	GLU-IA	U/CUU	ARG-IA	U/CCU	A/G
4A	G	VAL-IA	A/GAC	ALA-IIA	A/GGC	ASP-IIA	A/GUC	GLY-IIA	A/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IIA	U/CGC	GLU-IA	U/CUC	GLY-IIA	U/CCC	A/G

D

1A	U	LEU-IA	A/GAA	SER-IIA	A/GGA	STOP	A/GUA	STOP	A/GCA	U/C
1B		LEU-IA	U/CAA	SER-IIA	U/CGA	STOP	U/CUA	STOP	U/CCA	A/G
2A	C	LEU-IA	A/GAG	PRO-IIA	A/GGG	HIS-IIA	A/GUG	ARG-ID	A/GCG	U/C
2B		LEU-IA	U/CAG	PRO-IIA	U/CGG	GLN-IB	U/CUG	ARG-ID	U/CCG	A/G
3A	A	ILE-IA	A/GAU	THR-IIA	A/GGU	ASN-IIB	A/GUU	SER-IIA	A/GCU	U/C
3B		ILE-IA	U/CAU	THR-IIA	U/CGU	LYS-IE	U/CUU	ARG-ID	U/CCU	A/G
4A	G	VAL-IA	A/GAC	ALA-IID	A/GGC	ASP-IIB	A/GUC	GLY-IIA	A/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IID	U/CGC	GLU-IB	U/CUC	GLY-IIA	U/CCC	A/G

E

1A	U	PHE-IIC	A/GAA	SER-IIA	A/GGA	TYR-IC	A/GUA	CYS-IB	A/GCA	U/C
1B		LEU-IA	U/CAA	SER-IIA	U/CGA	STOP	U/CUA	TRP-IC	U/CCA	A/G
2A	C	LEU-IA	A/GAG	PRO-IIA	A/GGG	HIS-IIA	A/GUG	ARG-ID	A/GCG	U/C
2B		LEU-IA	U/CAG	PRO-IIA	U/CGG	GLN-IB	U/CUG	ARG-ID	U/CCG	A/G
3A	A	ILE-IA	A/GAU	THR-IIA	A/GGU	ASN-IIB	A/GUU	SER-IIA	A/GCU	U/C
3B		MET-IA	U/CAU	THR-IIA	U/CGU	LYS-IE	U/CUU	ARG-ID	U/CCU	A/G
4A	G	VAL-IA	A/GAC	ALA-IID	A/GGC	ASP-IIB	A/GUC	GLY-IIA	A/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IID	U/CGC	GLU-IB	U/CUC	GLY-IIA	U/CCC	A/G

F

row | 1st codon position | aa-aaRS | Ac | aa-aaRS | Ac | aa-aaRS | Ac | aa-aaRS | Ac | 3rd codon position

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

Figure 10. A working model for evolution of the genetic code in Archaea. Colors help to track the orders of amino acid additions into the code. A) G-code; B) GADV-code; C) 8 aa-code; D) 8 aa-code with invasion of row 3; E) ~16 aa-code; F) standard code. Some of the last amino acids encoded on disfavored row 1 are indicated in charcoal (F). GlyRS-RBZ indicates a ribozyme GlyRS before evolution of GlyRS-IIA. Red letters in anticodons are (generally) not utilized in archaeal tRNAs.

10.1. Evolution of Stop Codons

The model shown in Figures 9 and 10 was designed in part to better model the evolution of stop codons and to better describe the evolution of the 1st row (disfavored 3rd anticodon position A). Also, the model provides potential insight into evolution of 6-codon sectors for leucine, serine and arginine. The genetic code appears to have filled from the 4th row (3rd anticodon position C) to the 2nd row (3rd anticodon position G) to the 3rd row (3rd anticodon position U) to the 1st row (3rd anticodon position A), so the 1st row was the most difficult to fill, and A was strongly disfavored in the 3rd anticodon position (Figure 9). We posit the following explanation. Before evolution of the EF-Tu GTPase latch, the 1st row may have been filled with tRNAs that were utilized inefficiently, often resulting in termination of translation and relatively short peptide release. In Archaea, A is rarely or never utilized in the 1st anticodon wobble position. We posit that A was very inefficiently utilized in the 3rd anticodon position, when the 3rd anticodon position was a wobble position, before evolution of the EF-Tu latch. We guess that row 1 tRNAs were charged with amino acids but were infrequently utilized. We posit, therefore, that stop codons recognized by protein release factors evolved after evolution of the EF-Tu latch. Because A was strongly disfavored in the 3rd anticodon position, the location of stop codons to row 1 is telling. Because row 1 (3rd anticodon position A) anticodons are disfavored and because stop codons are recognized as mRNAs, it makes sense that stop codons are located to row 1. Within the standard code, no tRNAs correspond to stop codons.

10.2. A Working Model

In Figure 10, we show a highly detailed model for evolution of the genetic code. This is a variation of models previously published by our laboratory [33,48]. Possible advantages of this model are: 1) an improved description of the evolution of stop codons; 2) an explanation for serine jumping from column 2 to column 4 of the code; 3) possible new insight into evolution of 6-codon sectors; and 4) disfavoring of A in the anticodon 3rd position. When we first tried to formulate these models, we thought that such detailed accounts of genetic code evolution were not reasonable. Now, we are convinced that these models are highly descriptive and likely highly accurate. We were surprised at how easily these models unfolded from a very small number of reasonable initial assumptions and, also, how readily hypotheses for tRNA, aaRS and genetic code evolution integrated.

10.2.1. Assumptions and background

The main initial assumption that we made was that glycine was the first encoded amino acid. Furthermore, we assumed that the entire genetic code (anticodons and codons) initially encoded polyglycine (Figures 9 and 10A). We considered that polyglycine could have multiple purposes in the ancient world. Polyglycine could function as a hydrogel, and polyglycine could act as a cross-linking agent, as it does in bacterial cell wall peptidoglycan layers [120-122]. So, polyglycine can be a structural and hydrogel (LLPS) component, enhancing protocell function (i.e. membrane function and ion transport) and rigidifying protocell membranes. One reason to believe this is a reasonable assumption is that $tRNA^{Pri}$ (the primordial type I tRNA) is essentially $tRNA^{Gly}$ in ancient Archaea (Figure 4). As noted above, GlyRS-IIA in Archaea is the primordial aaRS enzyme (Figure 8). If the entire genetic code initially encoded polyglycine, this provides a model for co-evolution of mRNA and tRNA. If all mRNA codons and tRNA anticodons initially encoded glycine, it was very simple to coevolve mRNA codons and tRNA anticodons with new invasions of amino acids from outside the code.

The concept of filling in the genetic code with glycine and then adding amino acids by invasion resulted in a simple model for additions of amino acids to the code. Glycine looked like the first encoded amino acid, and glycine utilizes anticodons GCC, UCC and CCC (2nd and 3rd anticodon

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

position C). As we discuss, C is favored in the 2nd and 3rd anticodon positions. A is (essentially) never found in the anticodon wobble 1st position in Archaea. Furthermore, the four simplest amino acids glycine, alanine, aspartate and valine appear to be the first four amino acids encoded, and these amino acids are all found in row 4 of the genetic code (3rd anticodon position C) [126-130]. It began to look as if C was a favored base in the tRNA anticodon. Because A is so strongly disfavored in the anticodon wobble position in Archaea, we began to wonder whether A was also disfavored in the 3rd anticodon position. Phenylalanine, tyrosine, tryptophan, cysteine and stop codons appear to be late additions to the genetic code that are found in row 1, which is 3rd anticodon position A. We began to think that A was disfavored in both the 1st (wobble) and 3rd anticodon positions. Stop codons are recognized by protein release factors as codons in mRNA, with no corresponding anticodon in tRNA, consistent with A being disfavored in the anticodon 3rd position. Protein release factors recognize stop codons to release the nascent peptide from the ribosome [79]. So, C is favored in the anticodon, in the 2nd and 3rd anticodon positions. A is disfavored in the anticodon and, probably, in all three anticodon positions, although the effect is not obvious in the 2nd anticodon position, which is the easiest to read. So, glycine occupies the most favored position in the genetic code (anticodons GCC, UCC, CCC), consistent with glycine being the first encoded amino acid.

10.2.2. Evolution in columns

Analysis of aaRS evolution indicates that much of the evolution of the genetic code occurred within code columns, which represent the 2nd anticodon position (Figure 8). On the ancient ribosome (i.e. before LUCA), the 2nd anticodon position was the easiest to read, in part because the 1st and 3rd anticodon positions were initially wobble positions. In column 1 of the genetic code, ValRS-IA, MetRS-IA, IleRS-IA and LeuRS-IA are all structural class IA enzymes (Figure 8A). Valine, isoleucine, and leucine are hydrophobic amino acids. Valine can be converted to leucine via a 5-step pathway. Methionine is a late invader that we posit attacked an isoleucine 4-codon sector to add methionine and to evolve translation start codons, but MetRS-IA is also structural class IA, indicating additional evolution in column 1 (i.e. IleRS-IA→MetRS-IA). In column 2 of the genetic code SerRS-IIA, ProRS-IIA and ThrRS-IIA are closely related enzymes of structural class IIA. Serine and threonine are closely related amino acids. There is substantial evidence, therefore, for coevolution of aminoacyl-tRNA synthetases and amino acids in column 2. In column 3, HisRS-IIA, AspRS-IIB and AsnRS-IIB enzymes are reasonably closely related, and aspartate and asparagine are related amino acids. Also, in column 3, GlnRS-IB, LysRS-1E and GluRS-IB enzymes are closely related, and glutamate and glutamine are closely related amino acids. In column 4, ArgRS-ID and CysRS-IB enzymes are closely related. In some cases, the assigned structural class for an aaRS does not represent its closest relatives. We take these data to overwhelmingly support evolution of the genetic code within columns (anticodon 2nd position) as indicated in the model in Figure 10.

In the model (Figures 9 and 10), the genetic code sectors from one that encodes only glycine (Figure 10A) to one that encodes glycine, alanine, aspartate and valine (Figure 10B). Others have supported these four simple amino acids as the first four encoded amino acids [16,31,126-130]. In the standard code (Figure 10F), glycine retreats from occupying the entire code (Figure 10A) to occupying only column 4 (Figure 10B) and, finally, to occupying only row 4 and anticodons GCC, UCC and CCC (2nd and 3rd anticodon position C) (Figure 10F). Aspartate retreats from initially occupying all of column 3 (Figure 10B) to occupying only anticodon GUC (Figure 10F). Alanine retreats from occupying all of column 2 (Figure 10B) to occupying only anticodons GGC, UGC and CGC (Figure 10F). Valine retreats from occupying all of column 1 (Figure 10B) to occupying only anticodons GAC, UAC and CAC (Figure 10F). In this way, the first four encoded amino acids land in the 4th row after occupying the entire code. Newly added amino acids, therefore, invaded previously occupied sectors of the code. Amino acids that entered the code first surrendered less-favored anticodons to invaders but retained the most favorable anticodons according to clear and inviolable rules.

10.2.3. An evolutionary bottleneck

On a primitive ribosome, the 1st and 3rd anticodon positions were initially wobble positions. The consequence of this bottleneck in genetic code evolution was that the complexity of the code froze at ~8 amino acids (Figures 10C and 10D). Wobble positions were read with only pyrimidine-purine discrimination, and only one wobble position could be read at one time, limiting the code complexity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

to 2x4=8 assignments. Columns 1, 2 and 4, initially sectored on the 2nd and 3rd (then wobble) anticodon positions. Column 3, by contrast, initially sectored on the 1st (wobble) and 2nd anticodon positions. In columns 1, 2 and 4, it appears that incoming amino acids leucine (column 1), serine (column 2) and arginine (column 4) may have first invaded the 2nd row (3rd anticodon position G) (Figure 10C). Interestingly, leucine, serine and arginine are the three amino acids that occupy 6-codon sectors in the fully evolved genetic code. Leucine, serine and arginine are posited to have occupied row 2 first, because row 1 was difficult to occupy (3rd anticodon position A is disfavored). It appears that the tRNA anticodon bases were selected for small size (pyrimidine>purine) and stronger hydrogen bonding potential (C>G>U>>A). These rules are most apparent in the 3rd anticodon position (Figures 9 and 10).

10.2.4. Evolution of 6-codon sectors

We posit that leucine (column 1), serine (column 2) and arginine (column 4), initially in row 2 (Figure 10C), may then have invaded row 3 (3rd position U; Figure 10D). Some reasons to think this invasion of row 3 might have occurred are as follows. First, serine jumps to row 3, column 4, from column 2 in the code. This jump is easiest to imagine if serine occupies column 2, row 3, before making the jump to column 4, row 3. According to the model, only a single base change in the tRNA anticodon (GGU→GCU) was necessary for serine to jump. Second, arginine occupies rows 2 and 3 in the code, as if arginine made the posited invasion of row 3. In this regard, it is notable that leucine, serine and arginine are the three amino acids to occupy 6-codon sectors in the final evolution of the standard code. Here, we suggest that 6-codon sectors may have arisen from the history of code sectoring.

10.2.5. Evolution of the EF-Tu latch

To proceed beyond the bottleneck of only 8 amino acids required the evolution of the EF-Tu GTPase latch. The latch sets a closed conformation of the codon-anticodon pair and the ribosome involving 16S rRNA (30S subunit) residues G530, A1492 and A1493 and 23S rRNA (50S subunit) residue A1913. Closing of the latch allows 4-base recognition at the 2nd and 3rd anticodon positions [71-74,131,132]. The latch also improves the accuracy of the wobble position, but the wobble 1st anticodon position only has purine versus pyrimidine resolution. Evolution of the EF-Tu latch allows a genetic code with up to 2x4x4=32 anticodon assignments.

10.2.6. Completion of column 1

To complete sectoring of the genetic code in column 1, we posit that isoleucine invaded row 3, displacing leucine, which retained row 2 (3rd anticodon position G (row 2) (Leu) was favored over U (row 3) (Ile)). Isoleucine formed a 4-codon sector with anticodons GAU, UAU and CAU. When methionine invaded CAU, isoleucine retained CAU, but isoleucine codon AUA is specified by anticodon wobble C→agmatidine modification, which does not read AUG methionine codons [82,83]. In Archaea, generally, isoleucine UAU is not utilized. So, methionine was a late invader of an isoleucine 4-codon sector. Methionine is utilized at start codons. In column 1, row 1A, phenylalanine was a late addition to the code.

10.2.7. Completion of column 2

To complete sectoring of column 2, serine first jumped to column 4 (Figures 10D and 10E), then, after evolution of the EF-Tu latch, the serine sector expanded to disfavored row 1 (disfavored 3rd anticodon position A). Then proline invaded row 2, displacing serine, and threonine invaded row 3, displacing serine. Because serine jumped to column 4, row 3, serine occupied a favorable anticodon (GUC), and serine, therefore, could give up otherwise favored anticodon positions in column 2 to proline and threonine. Serine is the only amino acid to have jumped in evolution of the genetic code, indicating the overall orderly evolution of the code. Other similar models are possible for evolution of column 2 [33,48].

10.2.8. Evolution of column 3

Because of sectoring on the 1st anticodon (wobble) rather than the 3rd position, genetic code column 3 is the most innovated column. We posit that sectoring on the wobble position caused this innovation and the pattern of sectoring. We posit that initially, aspartate filled column 3 (Figure 10B). Aspartate was displaced by the related amino acid glutamate in rows 4B, 3B and 2B. Aspartate retained rows 4A, 3A and 2A and surrendered rows 4B, 3B and 2B to glutamate, because, in the anticodon, wobble G is favored over wobble U and C, and aspartate entered the code first. Histidine

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

displaced aspartate in row 2A. We posit that HisRS-IIA evolved from a primitive AspRS-IIA. AspRS-IIA then evolved to AspRS-IIB to suppress translation errors. In row 3A, first an amidotransferase evolved to convert Asp-tRNA^{Asn}, charged by AspRS-IIB, to Asn-tRNA^{Asn}. Subsequently, AsnRS-IIB evolved from AspRS-IIB to replace the amidotransferase. Note that the order of invasions and modifications are indicated by the structural classes of the aaRS enzymes (Figures 8 and 10).

Glutamate occupied column 3, sectors 4B, 3B and 2B. GluRS-IA evolved to GluRS-IB. In sector 2B, glutamate bound to tRNA^{Gln} was modified to glutamine by an amidotransferase [102,103,133-135]. Subsequently, this system evolved a GlnRS-IB to substitute for the amidotransferase. We note that, metabolically, lysine can be derived from a pathway that utilizes glutamate, although the lysine carbon skeleton is derived from α-aminoadipic acid. It appears, therefore, that lysine invaded sector 3B from outside the code, displacing glutamate. Despite its different stated structural class, LysRS-IE in Archaea is very closely related to GluRS-IB and GlnRS-IB (Figure 8A). Probably, lysine invasion of the code occurred before full establishment of the GlnRS-IB system, which is incompletely evolved in some Archaea. In Bacteria, LysRS-IIB is found. We posit, therefore, that LysRS-IIB was a bacterial innovation that replaced the archaeal LysRS-IE. Bacterial LysRS-IIB is closely related to AspRS-IIB and AsnRS-IIB (Figure 8A). We posit that, in Bacteria, LysRS-IIB evolved within column 3 to better specify accurate tRNA^{Lys} charging. LysRS-IIB in Bacteria evolved an editing active site to remove improperly attached amino acids. It appears that LysRS-IIB mostly uses editing to discriminate against amino acids invading from outside the genetic code [67].

10.2.9. Column 4

Metabolically, arginine can be derived from ornithine, which may have been a more primitive positively charged amino acid utilized in pre-life [136]. It is possible, therefore, that arginine replaced encoded ornithine during genetic code evolution. We posit that arginine (or ornithine) occupied column 4, rows 2 and 3, displacing glycine, which, as the first encoded amino acid, retained the most favored anticodons, GCC, UCC and CCC (2nd and 3rd anticodon C) (Figure 10C and 10D). After evolution of the EF-Tu GTPase latch, row 1 could be occupied, and additional amino acids could be encoded. We posit that serine jumping from column 2 to column 4 occurred early in code evolution, for instance, before proline and threonine invasion of column 2. We note that serine jumping to column 4 could have been initiated by a single base change in the 2nd position of the tRNA anticodon (GGU→GCU). Also, SerRS-IIA is a very different enzyme than ArgRS-ID, facilitating the invasion of column 4 by limiting tRNA charging errors. Although ArgRS-ID is classified as a structural subclass ID enzyme, ArgRS-ID is closely related to subclass IA and Cys-IB enzymes (Figure 8A).

10.2.10. Late evolution of row 1

We posit that phenylalanine, tyrosine, cysteine, tryptophan and stop codons, all located on disfavored row 1, were late additions to the code. Here, we suggest that these amino acids and stop codons could not be added before evolution of the EF-Tu latch. Essentially, the disfavored first row (3rd anticodon position A) could not be efficiently occupied before evolution of the latch. Prior to evolution of the latch, the translation stop signal is posited to have been inefficiently functioning tRNAs with 3rd anticodon position A. After evolution of the latch, row 1 tRNAs could be efficiently utilized, and leucine and serine could effectively invade row 1. Phenylalanine then invaded column 1, row 1A, displacing leucine. Leucine retained favored row 2 anticodon positions with phenylalanine invasion. Uncharacteristically, within column 2, serine surrendered more favorable anticodons to proline and threonine, but serine retained a favorable anticodon in column 4, row 3A. Also, serine utilizes a type II tRNA^{Ser}. Type II tRNA^{Ser} has an expanded variable loop that functions as a positive determinant for accurate SerRS-IIA serine addition. Because tRNA^{Ser} is a type II tRNA, in which the expanded variable loop is a SerRS-IIA-contacted determinant for accurate charging, this facilitated serine jumping in the code, from column 2 to column 4, and allowed serine to maintain a favored anticodon in column 4 (GCU) and to surrender otherwise favored anticodons in column 2 to invading proline and threonine. At about the time of the evolution of the EF-Tu latch, we posit that protein release factors evolved to take over stop codon functions, allowing proteins to become longer and more complex with accurate starts and stops [79].

10.2.11. The genetic code model and perspectives

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

The model offered here is a variation of models published previously. The genetic code evolved around the tRNA anticodon. Taking a tRNA-centric view, therefore, simplifies the understanding of code evolution. The model presented here provides selections for the locations of all amino acids in the genetic code. Amino acids enter the code by two identifiable mechanisms: 1) invasion from outside the code; and 2) enzymatic modifications of amino acids bound to tRNAs (i.e. Asp→Asn, Glu→Gln and pSer→Cys) followed by subsequent evolution of aaRS enzymes (AsnRS-IIB, GlnRS-IB and CysRS-IB). Evolution occurred first in code columns because the 2nd anticodon position is most important and easiest to read on a primitive ribosome. Evolution also occurred by rows according to clear anticodon preference rules (Figure 9). Column 3 sectored differently than columns 1, 2 and 4. Column 3 sectored early on the 1st (wobble) and 2nd anticodon positions, between aspartate and glutamate. As a result of this sectoring strategy, column 3 became the most innovated column in the code, encoding the most amino acids. Columns 1, 2 and 4, which sectored on the 2nd and 3rd anticodon positions, are characterized by larger blocks of anticodons (i.e. 4- and 6-codon sectors). Columns 1 and 2 are characterized by aaRS enzymes with editing active sites. In Archaea, ProRS-IIA is an exception. Because editing is a fidelity mechanism and because amino acids invade the code through tRNA charging errors, editing probably protects larger blocks of anticodons (i.e. 4- and 6-codon sectors). Because arginine and glycine have unique characteristics, ArgRS-ID and GlyRS-IIA were under little selection pressure to evolve editing. Arginine is a stiff and bulky, positively charged amino acid with unique hydrogen-bonding capacity. By contrast, positively charged ornithine and lysine are very flexible. Arginine, therefore, forms more structured ion pairs, particularly with aspartate (ion pairs with glutamate are more flexible). Glycine is the smallest amino acid, so a compact active site in GlyRS-IIA limits mischarging of $tRNA^{Gly}$ with larger amino acids. We posit that aaRS editing protected 4- and 6-codon sectors in columns 1 and 2 by limiting mischarging of tRNAs. Editing was not necessary for the aaRS enzymes in column 4 because of glycine and arginine properties. Column 3 broke into 2-codon sectors because of early sectoring on the 1st anticodon (wobble) position.

Once the genetic code evolved and protein enzymes came to dominate, the potential to enrich metabolism and energy utilization exploded. Whatever systems predated current systems, therefore, were replaced by enzymatic and protein motor pathways. For these reasons, we do not favor models for genetic code evolution based primarily on metabolism. Of course, an amino acid must have been available in order to have been added to the code. On the other hand, the expanding code helped drive the evolution of metabolism to provide more amino acids because, with the advent of coding, amino acids were of enhanced selective value. With regard to energy transduction, it is clear that primitive energetic systems were sufficient to support the evolution of the standard genetic code. After code evolution, an explosion in energy transduction systems occurred, leading to modern systems. We note that multiple pathways are identified in which tRNA-bound amino acids are substrates for metabolic reactions. In the pre-life world, we posit that RNA-bound peptides and amino acids were substrates for many reactions [32,68]. One effect of covalent RNA-amino acid binding was to shield a potentially reactive group on the amino acid from unproductive side reactions.

The mutually reinforcing tRNA, tRNAome, aaRS and genetic code evolution models presented here make many testable predictions. The model for genetic code evolution follows strongly from the aaRS evolution model (Figure 8). Essentially, evolution of aaRS enzymes directs the genetic code model. Detailed hypotheses were generated for polyglycine world and GADV world, and these predictions can be tested experimentally (see below). The sequence analyses underlying the tRNA, tRNAome and aaRS evolution models described above can be further challenged using additional sequence data and more sophisticated bioinformatics and computation. We make suggestions about RNA-linked reactions in the ancient pre-life world that can be pursued. For instance, if Val-$tRNA^{Val}$ was converted to Leu-$tRNA^{Leu}$ through a series of tRNA-linked reactions and evolutionary steps, as we suggest, then the history of column 1 evolution becomes significantly richer and more interesting. Such a model for column 1 evolution reinforces the interaction between our views and those that support a metabolic coevolution theory. Also, such a view enriches the possibilities for RNA-linked and tRNA-linked reactions in the ancient world [32,68]. We posit the existence of diverse ribozymes

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

in the ancient world, some of which have not yet been generated by researchers. As an example, we imagine a telomerase-like ribozyme with a guide RNA template that accurately generated RNA repeats from an RNA 3'-end to synthesize tRNA precursor sequences (Figure 2). We also posit that diverse ribozyme aaRS enzymes with reasonable accuracy could be generated to initiate the genetic code before enough amino acids have joined the code to encode aaRS protein enzymes. We posit diverse RNA-linked reactions in the ancient world with many yet to be discovered.

11. The Great Divergence

How did Archaea and Bacteria diverge? Which domain is most similar to LUCA? Despite their many similarities, how are Archaea and Bacteria distinct? After LUCA, the great divergence occurred, which we posit resulted in the splitting of Archaea and Bacteria (Figure 1). Although this point has been argued, we identify LUCA as most similar to Archaea [26,29]. Interestingly, a recent paper placed LUCA in the midst of the archaeal domain. We noticed that tRNAomes (all of the tRNAs of an organism) are much more compact in ancient Archaea and much more diverged in Bacteria. Radiating tRNAomes from $tRNA^{Pri}$, archaeal $tRNA^{Gly}$ is most similar to $tRNA^{Pri}$ (Figure 4) [49]. Archaeal tRNAomes are much more similar to $tRNA^{Pri}$ than bacterial tRNAomes. Archaeal tRNAomes are also more highly structured than bacterial tRNAomes, as if convergent and divergent evolution have scrambled bacterial tRNAomes. We posit that ancient archaeal tRNAomes are structured similarly to LUCA tRNAomes. A similar argument could be made for archaeal aaRS trees and genetic code structures. These analyses support the hypothesis that Archaea are more closely rooted to LUCA than Bacteria. Archaea and Bacteria have distinct membrane lipids and replication systems that may in some way be linked [37]. Eukaryotes made an evolutionary choice of bacterial membrane systems over archaeal systems, indicating a potential selective advantage to the adopted bacterial membrane system at least in the evolving eukaryotic system.

12. Models to Describe Genetic Code Evolution

We suggest that the dominant models for description of genetic code evolution be re-evaluated. We found these models confusing and largely unhelpful. Formerly, views of genetic code evolution broke primarily into three main categories: 1) the stereochemical theory; 2) the coevolution (metabolism) hypothesis; and 3) the error-minimization theory [16,17,88]. Genetic code evolution has also been described as a "frozen accident" that occurred very rapidly and was fixed as the standard code because too many deviations from the code were lethal [88]. The stereochemical theory posits that originally nucleic acids and amino acids interacted chemically, resulting in the code between tRNA anticodons and amino acids. We find the stereochemical theory to have little predictive power for evolution of the code, although the stereochemical theory appears to reasonably describe evolution of riboswitches [137]. The coevolution hypothesis has some value. Of course, amino acid metabolism and the genetic code coevolved [138]. How could it be otherwise? If an amino acid could not be generated by primitive metabolism, it could not be added to the code. Amino acids, therefore, were added to the code from simple to complex, with glycine, alanine, aspartate and valine the first and simplest amino acids [126-130] and phenylalanine, tyrosine and tryptophan among the last and most complex additions [108]. We do not see metabolism of amino acids, however, as a strong driving force in selection of new amino acids added into the code. We also see the error minimization theory as a limiting idea. The error minimization theory indicates that the genetic code was structured to minimize translation errors. We do not think that idea is correct. Our opinion is that translational fidelity mechanisms drove the freezing of the code. Interestingly, the EF-Tu GTPase latch drove first the expansion of the code from an 8 amino acid bottleneck and later the freezing of the code at 20 amino acids plus stops (Figure 10).

Our view is that the standard genetic code evolved around tRNA and the tRNA anticodon [33,48-50]. The identification in existing tRNAs of 31-nt minihelices and 17-nt microhelices that can attach ACCA via a ribozyme ligase indicates a rich prebiotic chemistry involving covalent RNA-amino acid linkages and diverse ribozyme activities (Figure 2) [32,59-61,68]. The capacity for doing chemistry on tRNA-amino acid linkages persists, as we describe (i.e. pSer→Cys, Asp→Asn, Glu→Gln

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

and possibly Val→Leu) [99,101-103,105,106,134]. We posit that ACCA bound to amino acids at its 3'-end as a substrate for chemistry before the advent of 31-nt minihelices and tRNA. We strongly advocate for the tRNA-centric view, that evolution of tRNA from ligation of three 31-nt minihelices drove the evolution of the code, mRNA and rRNA. Therefore, tRNA was the central advance in biological intellectual property that enabled evolution of the code. We describe powerful selections driving evolution mostly within code columns but ultimately with amino acids distributing in an ordered manner in code rows, according to clear selection rules. We strongly argue that the history of amino acid additions to the code follows these interpretable patterns. For instance, column 3 of the genetic code becomes the most innovated column because of sectoring between Asp and Glu, initially utilizing the 1st and 2nd anticodon positions rather than the 2nd and 3rd anticodon positions, as for columns 1, 2 and 4 (Figure 10). Similarly, the history of evolution in columns 1, 2 and 4 appears to result in 6-codon sectors that encode leucine, serine and arginine. Of course, the failure to subdivide 6- and 4-codon sectors results in a code with fewer amino acids than could potentially be encoded. With regard to the freezing of the code, the genetic code was built by modifications of amino acids bound to tRNAs and by tRNA charging errors (invasions of amino acids from outside the code). tRNA charging errors and modifications, therefore, drove innovations of the code, and translational fidelity mechanisms froze the code. The EF-Tu latch is a major translational fidelity mechanism. The EF-Tu latch evolved to expand the code from an ~8 amino acid bottleneck to a richer code (Figure 10). Fidelity mechanisms such as the EF-Tu latch, aaRS editing, aaRS active site specificity, tRNA modifications and tRNA specialization drove the freezing of the code at 20 amino acids + stops. In this regard, we note that aaRS editing on the left half of the genetic code appears to protect 4- and 6-codon sectors from further divisions to encode additional amino acids. Anticodons specifying specific amino acids evolved as we describe through the coevolution of tRNAomes, aaRS enzymes and translational fidelity mechanisms. We show clearly that aaRS enzymes and the genetic code were powerfully coevolved (Figure 8).

13. Polyglycine and GADV Worlds (a working model)

13.1. The pre-life world

This paper reveals significant detail about the ancient pre-life and protocell worlds based mostly on the evolution of translation systems as inferred from conserved sequences. We describe fully the evolution of tRNAs and the genetic fragments and sequences from which tRNA was derived (Figures 2-5). We describe evolution and radiation of aaRS enzymes and the relationship between evolution of aaRS enzymes and sectoring of the genetic code (Figure 8). We describe how tRNA and hydrogels/LLPS could have contributed to the earliest aaRS folding. In this review, we attempt to link these ancient events to the activities of hydrogels, LLPS, membraneless compartments and amyloids. We believe the mechanisms and descriptions will lead to advances in analyses of hydrogels in pre-life reactions, protocell functions and prokaryotic systems. We consider hydrogels and related assemblies to be a formerly largely missing consideration in analyses of ancient evolution. For some future studies, we recommend increasing the system complexity and inclusion of hydrogels to help identify reactions of interest that may lead to an understanding of earlier events. Specifically, experimental probing of the richer chemistry of a GADV world (Figure 10B) would be expected to lead to insights into a prior polyglycine world (Figure 10A).

13.2. Coacervates

Significant work has been done with coacervate systems to enhance prebiotic chemistry. Examples of coacervates include clays, polymers and mica [30,31,139-141]. Such materials can concentrate reactants, control the access and activity of water, participate in wet-dry cycles and provide polar surfaces to help with enantiomer fractionations. Coacervates, therefore, are similar in their properties to biological hydrogels and LLPS. Here we propose that polyglycine and GADV polymers were potent prebiotic hydrogel components that drove the earliest stages in the evolution of the genetic code. Above, we describe a number of human proteins that may rely on hydrogels and that have polyglycine tracts. Shorter polyglycine tracts (i.e. length 6-8) can be found in archaeal, bacterial and phage proteins. We do not know the extent to which such short tracts can function to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

generate localized hydrogels. Based on our model for evolution of the genetic code, we further propose that polymers of Gly, Ala, Asp and Val may enhance hydrogel functions in prokaryotic systems, as indicated above. In pre-biotic systems, short peptide linkers can cross-link peptide chains to make more complex networks. Some human proteins (i.e. transcription factors) include polyalanine, polyhistidine and polyglutamine tracts. Because the ancient pre-life world was rich in RNA, including RNA polymers, minihelices and tRNAs, hydrogel assemblies evolved to include and concentrate or exclude RNAs [1,142-148].

13.3. Hydrogels

Hydrogels (LLPS) appear to be incompletely characterized and somewhat difficult to analyze in prokaryotic systems [12-14,44,136]. We guess that hydrogel compartments are important for many bacterial processes including cell division, coupling of transcription and translation, nucleoid body maintenance and rearrangements, ion transport and signaling. LLPS affects septation in Bacteria. In eukaryotic systems, LLPS has been more aggressively analyzed and is perhaps better understood. For instance, LLPS compartments tend to be larger and easier to visualize in eukaryotic systems. As a critical phase of their evolution, Eukaryotes appear to have powerfully enriched LLPS systems by increasing the use of proteins that include intrinsically disordered regions often with the potential for covalent modification and non-covalent bonding of diverse hydrogel components [1,2,5,6,27,42-44]. Examples of intrinsically disordered regions with these properties include histone tails and the carboxy-terminal domain of RNA polymerase II. Covalent modifications of these disordered regions alter activities and factor binding. As RNA polymerase II traverses the transcription cycle, polymerase can move between LLPS compartments that support different phases of the cycle. Transcriptional super-enhancers that direct cell-specific gene regulation programs organize hydrogel compartments, and, in some cancers, super-enhancers become disorganized and sometimes fuse together [1,6]. The nucleolis is a hydrogel compartment for RNA polymerase I transcription and ribosome assembly [3,144]. RNA polymerase III also separates into hydrogel compartments. Often RNA and RNA-binding factors are defining characteristics of hydrogels. In eukaryotic cells, hydrogel compartments appear to segregate and regulate many other activities including translational control, translational delay via microRNA, signaling and ion transport. Within cells, hydrogel compartmentalization is a mechanism to support diverse biological processes, the activity of water and the potency of acids and bases to support chemistry.

13.4. Polyglycine and GADV worlds

We try to imagine polyglycine world. In Figure 11, we show a working model for a protocell that utilized polyglycine as a hydrogel, cross-linking agent and component and stabilizer of protocell walls. As discussed in this paper, many of the predicted components of the protocell matrix are indicated. We posit that the interior of the protocell was packed with polyglycine hydrogels, membraneless compartments, polyglycine amyloid accretions, primitive metabolites, diverse RNAs, tRNAs, many ribozymes and pre-ribosomes. Our idea of a pre-ribosome is a pre-16S scaffold, on which to mount a mRNA (of any sequence), and a mobile and independent PTC [84]. We envision that tRNAs with essentially all anticodons represented are charged (essentially) only with glycine by a GlyRS-ribozyme. The idea is that the most rapid sequence to mutate successfully in tRNA is the anticodon, because most other changes cause structural defects in the L-shaped tRNA structure. We imagine that ACCA was ligated by a ribozyme ligase to early tRNAs, so the tRNA could be charged with glycine. ACCA is the most common 3'-end in archaeal tRNAs [63], indicating a pre-life function of ACCA ligated to RNAs. In such a system, essentially all encoded protein products were polyglycine of varying lengths. We posit that translation termination initially tended to occur at NNA anticodons, because the 3rd anticodon position was initially a wobble position, and 3rd position A had difficulty pairing to U in mRNA.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

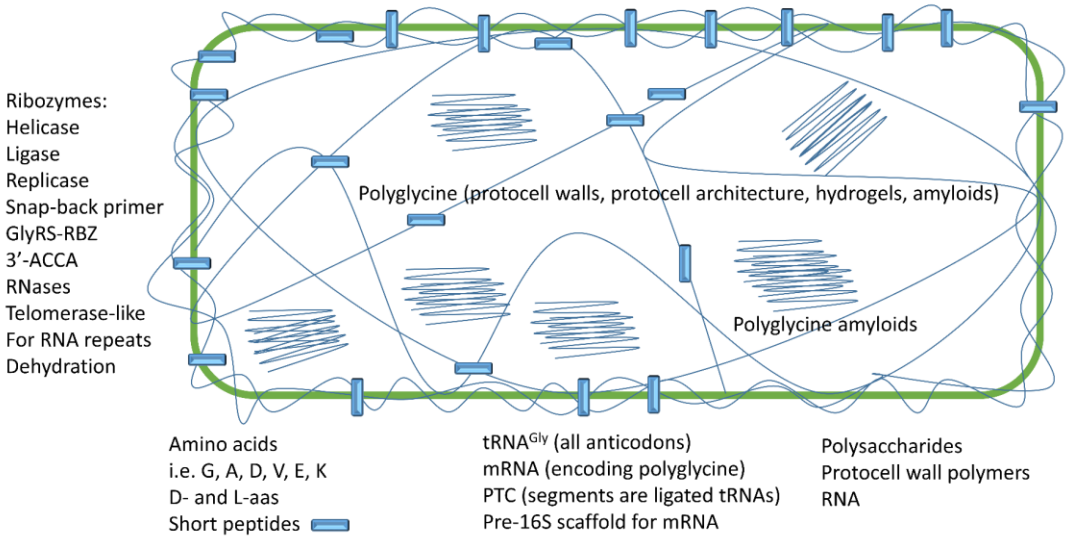


Figure 11. The polyglycine protocell (a working model).

13.5. Polyglycine world was first supported by 31-nt minihelices and then by tRNA

We imagine that a tRNA-based polyglycine world replaced a 31-nt minihelix polyglycine world as partly described in Figure 2. The advantages of tRNAs over minihelices to support a robust translation system are numerous. First, tRNA supports a 3-nt genetic code that evolved to the standard genetic code. Based on our understanding of genetic code evolution, it appears that type I and type II tRNA evolved from a 31-nt minihelix world rapidly, before the evolution of diverse 3-nt anticodons. One argument for our conclusion is that the anticodon loop and the T loop in tRNA are homologous, indicating that tRNA probably evolved from identical anticodon and T loop 31-nt minihelices (Figure 2). Very likely, tRNA could not evolve by ligation of three anticodon (or T loop) 31-nt minihelices because of the way the RNA would fold. Because of the strength of minihelix stem pairing, such an assembly would be processed accurately into three 31-nt minihelices rather than adopting a tRNA fold. Combination of a 5′ D loop minihelix with two anticodon minihelices, however, naturally folds as a tRNA L-shaped form, and refolding in the tRNA configuration blocks processing into 31-nt minihelices. Simply stated, to process to 31-nt minihelices requires full pairing of acceptor stems, and re-folding into a tRNA conformation blocks internal acceptor stem pairing. The D loop minihelix more easily folds into the alternate structures required for the tRNA fold. Very few sequence changes were required to stabilize the tRNA fold from the primordial tRNA sequence (Figures 2-5).

Two different 31-nt minihelix sequences are identified in existing tRNA sequences, the D loop minihelix and the initially identical anticodon loop and T loop minihelices (Figure 2). We imagine that many other 31-nt minihelix sequences existed that have not been preserved in evolution because they became dispensable after evolution of tRNA. If 3′-ACCA were ligated to 31-nt minihelices, these RNAs could then be aminoacylated with glycine by GlyRS-RBZ. We posit that 31-nt minihelices with attached ACCA (i.e. via ligation) evolved to synthesize polyglycine, and a tRNA-based polyglycine world would, therefore, have inherited many features of the prior minihelix world. The 31-nt minihelix model for a polyglycine protocell, therefore, is very similar to the tRNA-based model. The only difference is that the mRNA must be slightly different to interact with minihelices. A D loop 31-nt minihelix appears to present a single C to interact with a single G in mRNA. An anticodon loop 31-nt minihelix might present ~AAA to interact with ~UUU in mRNA. This anticodon-codon arrangement is slightly ambiguous because the primordial sequence of the anticodon loop and T loop minihelices (i.e. ~UU/AAAAA) is not known with certainty (Figure 2). So, we imagine a complex set

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 October 2020 doi:10.20944/preprints202009.0162.v2

of different minihelices with different codon recognition including 1-nt and 3-nt anticodons and possibly other interactions. The primitive ribosome could be approximated by a RNA decoding scaffold to hold a mRNA and a PTC ribozyme. Such a translation system could support peptide bond formation based on either minihelices or tRNAs.

13.6. Selections for polyglycine and GADV polymers

Polyglycine and GADV hydrogels and RNAs are expected to powerfully regulate the activity of water in protocells. Dehydration stimulates polymerization reactions including RNA synthesis, DNA synthesis, polypeptide synthesis and polysaccharide (i.e. cell wall) synthesis. RNAs bind water and, therefore, include dehydrating pockets to support chemistry of proximal reactants. For instance, the PTC of the ribosome has been considered a molecular crowding and dehydration chamber to support peptide bond formation utilizing chemically diverse amino acid substrates [57]. The "trigger loop" in RNA polymerase II closes over the active site expelling water to support RNA polymerization [149,150]. When the activity of water is decreased, acid-base reactions and polymerization reactions become more potent.

We imagine that polyglycine can be cross-linked to short polypeptide chains similar to those found in peptidoglycan bacterial cell walls and using similar chemistry. The protocell could have been encapsulated by peptidoglycan protocell walls. Capping polyglycine with other amino acids (i.e. lysine) and reduction of polyglycine C-termini to aldehyde groups would allow Schiff's base cross-linking, so a protocell matrix mostly of polyglycine with short peptide linkers could be constructed. Some reactions could be supported by the reducing environment and others by ribozymes. If protocells had cell walls, they also had long polysaccharides in addition to long RNA polymers. We hope that the model we outline can provide some utility in developing new approaches to enrich studies of pre-biotic chemistry. We see value in a top down experimental approach starting with more complex systems (i.e. GADV world), in which more diverse chemistry might be detected, and then moving to simper systems (i.e. polyglycine world).

14. Conclusions

Chaotic processes helped to initiate evolution of the genetic code by providing a powerful selection. Specifically, polyglycine and poly-GADV (i.e. polyalanine) hydrogels, LLPS and amyloids provided compartments and coacervates to stimulate polymerization reactions and novel chemistry within protocells. By interesting contrast, tRNAs and the genetic code evolved by highly ordered and systematic processes. Based on sequences in ancient Archaea, tRNA evolution appears to be a solved and highly systematic problem. tRNA-linked amino acid reactions in Archaea demonstrate the importance of diverse RNA-linked chemistries in early evolution of life and the genetic code. Unexpectedly, divergence of aaRS enzymes strongly indicates the pathway for evolution of the genetic code. This result was somewhat unexpected because how can protein enzymes evolve before they can successfully be encoded? We posit that the answer lies in coevolution of interacting systems, as we describe here. The EF-Tu GTPase latch (closing of the 30S ribosomal subunit), which suppressed wobbling at the 3rd anticodon position, appears to have been a major driving force required for code expansion beyond the first 8 amino acids. Wobbling, therefore, was of fundamental importance in code evolution. The standard code expanded to 20 amino acids, but the maximum theoretical complexity of the code in tRNA is 32 anticodon assignments. To generate more than 20 anticodon assignments would require division of 4- and 6-codon sectors. aaRS editing and other fidelity mechanisms, however, protected 4- and 6-codon sectors from further divisions.

Ancient evolution of the pre-life→life transition from about 4 Ga ago is becoming increasingly well understood. Evolution of the genetic code is a simpler problem in archaeal systems, which are closest to LUCA (Figure 1) [26,29]. Coevolution of tRNA, mRNA, rRNA, aaRS enzymes, the genetic code and ribosomes appears to be a largely outlined problem [33,48]. The pathway of evolution of translation systems was driven via the evolution of tRNA, which is a solved problem (Figure 2) [59-61]. Because of wobbling, the genetic code has a maximum complexity of 32 assignments, as in tRNA anticodons (2x4x4), not 64 assignments, as in mRNA codons (4x4x4). The standard genetic code froze at 20 amino acids + stops because of translational fidelity mechanisms [33,48,50]. Column 3 of the

 doi:10.20944/preprints202009.0162.v2

code is the most innovated column, encoding the most amino acids, because column 3 evolution was driven at an early stage by the tRNA anticodon 1st wobble position, rather than the 3rd anticodon position, as in columns 1, 2 and 4 (Figure 10). aaRS enzyme structural classes and some chemically similar amino acids align with genetic code columns, demonstrating the importance and centrality of the 2nd anticodon position. We posit that glycine, the first encoded amino acid, initially filled the genetic code. According to the model, after glycine, other amino acids enter the code by invading previously occupied sectors and, therefore, displacing previously encoded amino acids. Amino acids that first entered the code retained the most favored anticodons according to clear rules. Glycine lands in code column 4, row 4, indicating that C is preferred in the anticodon 2nd and 3rd positions. The genetic code, however, appears to fill via rows: row 4→row 2→row 3→row 1, indicating that the 3rd anticodon position has the following preference rules: C>G>U>>A. Filling row 1 (disfavored 3rd anticodon position A) appears to have required evolution of the EF-Tu latch, a major feature of translational fidelity that allowed expansion of the code to 20 amino acids + stops from an ~8 amino acid bottleneck. The preference rules for the anticodon wobble position are G>(U~C)>>>>>>A. Only purine-pyrimidine discrimination was initially achieved at the anticodon wobble position. Using these simple rules, the entire genetic code was populated, as observed in the standard code (Figure 10F). Other features of translation systems evolved around tRNAs. It appears that polyglycine and GADV may have constituted potent hydrogels, LLPS compartments and amyloid accretions driving strong selection of the standard code particularly during early stages before protein enzymes were sufficiently encoded. All features of the hypotheses arise from models for tRNA, tRNAome, aaRS and genetic code evolution and related literature.

Abbreviations:

aaRS	aminoacyl-tRNA synthetase (i.e. GlyRS)
Ac	Anticodon
As	Acceptor stem
CTD	RNA polymerase II carboxy terminal domain
DNAP	DNA polymerase
EF	Translation Elongation Factor
FECA	First Eukaryotic Common Ancestor
G	Glycine
Ga	billion years
GADV	Glycine, Alanine, Aspartic Acid, Valine
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
IF	Translation Initiation Factor
Indels	insertions, deletions
LECA	Last Eukaryotic Common Ancestor
LLPS	liquid-liquid phase separation
LUCA	Last Universal Common (cellular) Ancestor
Min	*Methanocaldococcus infernus*
PDB	Protein Data Bank
Pfu	*Pyrococcus furiosis*
PPP	Promoter Proximal Pausing
PTC	Peptidyl Transferase Center
RBZ	ribozyme
RNAP	RNA polymerase
Sel, pSer	phosphoserine
SLS	stem-loop-stem
TBP	TATA-box binding protein
TFB	Transcription Factor B (homolog of TFIIB in Eukaryotes)
TFE	Transcription Factor E (homolog of TFIIE in Eukaryotes)
$tRNA^{Pri}$	Primordial tRNA

Tth *Thermus thermophilus*
V loop Variable loop

Funding: This research received no external funding

Acknowledgements: The authors thank Helen Hansma (UC Santa Barbara, USA) and Bruce Kowiatek (Blue Ridge Community College, WV, USA) for helpful comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References:

1. Guo, C.; Che, Z.; Yue, J.; Xie, P.; Hao, S.; Xie, W.; Luo, Z.; Lin, C. ENL initiates multivalent phase separation of the super elongation complex (SEC) in controlling rapid transcriptional activation. *Sci Adv* **2020**, *6*, eaay4858, doi:10.1126/sciadv.aay4858.
2. Portz, B.; Shorter, J. Switching Condensates: The CTD Code Goes Liquid. *Trends Biochem Sci* **2020**, *45*, 1-3, doi:10.1016/j.tibs.2019.10.009.
3. Odeh, H.M.; Shorter, J. Arginine-rich dipeptide-repeat proteins as phase disruptors in C9-ALS/FTD. *Emerg Top Life Sci* **2020**, 10.1042/ETLS20190167, doi:10.1042/ETLS20190167.
4. Andre, A.A.M.; Spruijt, E. Liquid-Liquid Phase Separation in Crowded Environments. *Int J Mol Sci* **2020**, *21*, doi:10.3390/ijms21165908.
5. Boehning, M.; Dugast-Darzacq, C.; Rankovic, M.; Hansen, A.S.; Yu, T.; Marie-Nelly, H.; McSwiggen, D.T.; Kokic, G.; Dailey, G.M.; Cramer, P., et al. RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat Struct Mol Biol* **2018**, *25*, 833-840, doi:10.1038/s41594-018-0112-y.
6. Yoshizawa, T.; Nozawa, R.S.; Jia, T.Z.; Saio, T.; Mori, E. Biological phase separation: cell biology meets biophysics. *Biophys Rev* **2020**, *12*, 519-539, doi:10.1007/s12551-020-00680-x.
7. Li, W.; Hu, J.; Shi, B.; Palomba, F.; Digman, M.A.; Gratton, E.; Jiang, H. Biophysical properties of AKAP95 protein condensates regulate splicing and tumorigenesis. *Nat Cell Biol* **2020**, *22*, 960-972, doi:10.1038/s41556-020-0550-8.
8. Roden, C.; Gladfelter, A.S. RNA contributions to the form and function of biomolecular condensates. *Nat Rev Mol Cell Biol* **2020**, 10.1038/s41580-020-0264-6, doi:10.1038/s41580-020-0264-6.
9. Nozawa, R.S.; Yamamoto, T.; Takahashi, M.; Tachiwana, H.; Maruyama, R.; Hirota, T.; Saitoh, N. Nuclear microenvironment in cancer: Control through liquid-liquid phase separation. *Cancer Sci* **2020**, *111*, 3155-3163, doi:10.1111/cas.14551.
10. Zhang, H.; Ji, X.; Li, P.; Liu, C.; Lou, J.; Wang, Z.; Wen, W.; Xiao, Y.; Zhang, M.; Zhu, X. Liquid-liquid phase separation in biology: mechanisms, physiological functions and human diseases. *Sci China Life Sci* **2020**, *63*, 953-985, doi:10.1007/s11427-020-1702-x.
11. do Amaral, M.J.; Araujo, T.S.; Diaz, N.C.; Accornero, F.; Polycarpo, C.R.; Cordeiro, Y.; Cabral, K.M.S.; Almeida, M.S. Phase Separation and Disorder-to-Order Transition of Human Brain Expressed X-Linked 3 (hBEX3) in the Presence of Small

Fragments of tRNA. *J Mol Biol* **2020**, *432*, 2319-2348, doi:10.1016/j.jmb.2020.02.030.

12. Guilhas, B.; Walter, J.C.; Rech, J.; David, G.; Walliser, N.O.; Palmeri, J.; Mathieu-Demaziere, C.; Parmeggiani, A.; Bouet, J.Y.; Le Gall, A., et al. ATP-Driven Separation of Liquid Phase Condensates in Bacteria. *Mol Cell* **2020**, *79*, 293-303 e294, doi:10.1016/j.molcel.2020.06.034.
13. Shen, B.A.; Landick, R. Transcription of Bacterial Chromatin. *J Mol Biol* **2019**, *431*, 4040-4066, doi:10.1016/j.jmb.2019.05.041.
14. Monterroso, B.; Zorrilla, S.; Sobrinos-Sanguino, M.; Robles-Ramos, M.A.; Lopez-Alvarez, M.; Margolin, W.; Keating, C.D.; Rivas, G. Bacterial FtsZ protein forms phase-separated condensates with its nucleoid-associated inhibitor SlmA. *EMBO Rep* **2019**, *20*, doi:10.15252/embr.201845946.
15. Di Giulio, M. The last universal common ancestor (LUCA) and the ancestors of archaea and bacteria were progenotes. *J Mol Evol* **2011**, *72*, 119-126, doi:10.1007/s00239-010-9407-2.
16. Koonin, E.V.; Novozhilov, A.S. Origin and Evolution of the Universal Genetic Code. *Annu Rev Genet* **2017**, *51*, 45-62, doi:10.1146/annurev-genet-120116-024713.
17. Koonin, E.V.; Novozhilov, A.S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **2009**, *61*, 99-111, doi:10.1002/iub.146.
18. Weiss, M.C.; Preiner, M.; Xavier, J.C.; Zimorski, V.; Martin, W.F. The last universal common ancestor between ancient Earth chemistry and the onset of genetics. *PLoS Genet* **2018**, *14*, e1007518, doi:10.1371/journal.pgen.1007518.
19. Burton, Z.F.; Opron, K.; Wei, G.; Geiger, J.H. A model for genesis of transcription systems. *Transcription* **2016**, *7*, 1-13, doi:10.1080/21541264.2015.1128518.
20. Harish, A. What is an archaeon and are the Archaea really unique? *PeerJ* **2018**, *6*, e5770, doi:10.7717/peerj.5770.
21. Burton, S.P.; Burton, Z.F. The sigma enigma: bacterial sigma factors, archaeal TFB and eukaryotic TFIIB are homologs. *Transcription* **2014**, *5*, e967599, doi:10.4161/21541264.2014.967599.
22. Furukawa, R.; Nakagawa, M.; Kuroyanagi, T.; Yokobori, S.I.; Yamagishi, A. Quest for Ancestors of Eukaryal Cells Based on Phylogenetic Analyses of Aminoacyl-tRNA Synthetases. *J Mol Evol* **2017**, *84*, 51-66, doi:10.1007/s00239-016-9768-2.
23. Brueckner, J.; Martin, W.F. Bacterial Genes Outnumber Archaeal Genes in Eukaryotic Genomes. *Genome Biol Evol* **2020**, *12*, 282-292, doi:10.1093/gbe/evaa047.
24. Zachar, I.; Boza, G. Endosymbiosis before eukaryotes: mitochondrial establishment in protoeukaryotes. *Cell Mol Life Sci* **2020**, 10.1007/s00018-020-03462-6, doi:10.1007/s00018-020-03462-6.
25. Eme, L.; Spang, A.; Lombard, J.; Stairs, C.W.; Ettema, T.J.G. Archaea and the origin of eukaryotes. *Nat Rev Microbiol* **2017**, *15*, 711-723, doi:10.1038/nrmicro.2017.133.
26. Long, X.; Xue, H.; Wong, J.T.-F. Descent of Bacteria and Eukarya from an archaeal root of life. *biorixiv* **2019**, https://doi.org/10.1101/745372 doi:https://doi.org/10.1101/745372

27. Lu, H.; Yu, D.; Hansen, A.S.; Ganguly, S.; Liu, R.; Heckert, A.; Darzacq, X.; Zhou, Q. Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **2018**, *558*, 318-323, doi:10.1038/s41586-018-0174-3.
28. Harmon, T.S.; Holehouse, A.S.; Rosen, M.K.; Pappu, R.V. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *Elife* **2017**, *6*, doi:10.7554/eLife.30294.
29. Marin, J.; Battistuzzi, F.U.; Brown, A.C.; Hedges, S.B. The Timetree of Prokaryotes: New Insights into Their Evolution and Speciation. *Mol Biol Evol* **2017**, *34*, 437-446, doi:10.1093/molbev/msw245.
30. Mariscal, C.; Barahona, A.; Aubert-Kato, N.; Aydinoglu, A.U.; Bartlett, S.; Cardenas, M.L.; Chandru, K.; Cleland, C.; Cocanougher, B.T.; Comfort, N., et al. Hidden Concepts in the History and Philosophy of Origins-of-Life Studies: a Workshop Report. *Orig Life Evol Biosph* **2019**, 10.1007/s11084-019-09580-x, doi:10.1007/s11084-019-09580-x.
31. Chatterjee, S.; Yadav, S. The Origin of Prebiotic Information System in the Peptide/RNA World: A Simulation Model of the Evolution of Translation and the Genetic Code. *Life (Basel)* **2019**, *9*, doi:10.3390/life9010025.
32. Kunnev, D.; Gospodinov, A. Possible Emergence of Sequence Specific RNA Aminoacylation via Peptide Intermediary to Initiate Darwinian Evolution and Code Through Origin of Life. *Life (Basel)* **2018**, *8*, doi:10.3390/life8040044.
33. Lei, L.; Burton, Z.F. Evolution of Life on Earth: tRNA, Aminoacyl-tRNA Synthetases and the Genetic Code. *Life (Basel)* **2020**, *10*, doi:10.3390/life10030021.
34. Damer, B.; Deamer, D. Coupled phases and combinatorial selection in fluctuating hydrothermal pools: a scenario to guide experimental approaches to the origin of cellular life. *Life (Basel)* **2015**, *5*, 872-887, doi:10.3390/life5010872.
35. Burton, Z.F. The Old and New Testaments of gene regulation. Evolution of multi-subunit RNA polymerases and co-evolution of eukaryote complexity with the RNAP II CTD. *Transcription* **2014**, *5*, e28674, doi:10.4161/trns.28674.
36. Iyer, L.M.; Aravind, L. Insights from the architecture of the bacterial transcription apparatus. *J Struct Biol* **2012**, *179*, 299-319, doi:10.1016/j.jsb.2011.12.013.
37. Koonin, E.V.; Krupovic, M.; Ishino, S.; Ishino, Y. The replication machinery of LUCA: common origin of DNA replication and transcription. *BMC Biol* **2020**, *18*, 61, doi:10.1186/s12915-020-00800-9.
38. O'Malley, M.A.; Leger, M.M.; Wideman, J.G.; Ruiz-Trillo, I. Concepts of the last eukaryotic common ancestor. *Nat Ecol Evol* **2019**, *3*, 338-344, doi:10.1038/s41559-019-0796-3.
39. Fournier, G.P.; Poole, A.M. A Briefly Argued Case That Asgard Archaea Are Part of the Eukaryote Tree. *Front Microbiol* **2018**, *9*, 1896, doi:10.3389/fmicb.2018.01896.
40. Pittis, A.A.; Gabaldon, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **2016**, *531*, 101-104, doi:10.1038/nature16941.
41. Koonin, E.V. Intron-dominated genomes of early ancestors of eukaryotes. *J Hered* **2009**, *100*, 618-623, doi:10.1093/jhered/esp056.
42. McSwiggen, D.T.; Hansen, A.S.; Teves, S.S.; Marie-Nelly, H.; Hao, Y.; Heckert, A.B.; Umemoto, K.K.; Dugast-Darzacq, C.; Tjian, R.; Darzacq, X. Evidence for

DNA-mediated nuclear compartmentalization distinct from phase separation. *Elife* **2019**, *8*, doi:10.7554/eLife.47098.

43. Zhou, H.X.; Nguemaha, V.; Mazarakos, K.; Qin, S. Why Do Disordered and Structured Proteins Behave Differently in Phase Separation? *Trends Biochem Sci* **2018**, *43*, 499-516, doi:10.1016/j.tibs.2018.03.007.
44. Langdon, E.M.; Gladfelter, A.S. A New Lens for RNA Localization: Liquid-Liquid Phase Separation. *Annu Rev Microbiol* **2018**, *72*, 255-271, doi:10.1146/annurev-micro-090817-062814.
45. Frank, L.; Rippe, K. Repetitive RNAs as Regulators of Chromatin-Associated Subcompartment Formation by Phase Separation. *J Mol Biol* **2020**, *432*, 4270-4286, doi:10.1016/j.jmb.2020.04.015.
46. Ling, S.C. Synaptic Paths to Neurodegeneration: The Emerging Role of TDP-43 and FUS in Synaptic Functions. *Neural Plast* **2018**, *2018*, 8413496, doi:10.1155/2018/8413496.
47. Stochaj, U.; Weber, S.C. Nucleolar Organization and Functions in Health and Disease. *Cells* **2020**, *9*, doi:10.3390/cells9030526.
48. Kim, Y.; Opron, K.; Burton, Z.F. A tRNA- and Anticodon-Centric View of the Evolution of Aminoacyl-tRNA Synthetases, tRNAomes, and the Genetic Code. *Life (Basel)* **2019**, *9*, doi:10.3390/life9020037.
49. Pak, D.; Du, N.; Kim, Y.; Sun, Y.; Burton, Z.F. Rooted tRNAomes and evolution of the genetic code. *Transcription* **2018**, *9*, 137-151, doi:10.1080/21541264.2018.1429837.
50. Pak, D.; Kim, Y.; Burton, Z.F. Aminoacyl-tRNA synthetase evolution and sectoring of the genetic code. *Transcription* **2018**, *9*, 205-224, doi:10.1080/21541264.2018.1467718.
51. Opron, K.; Burton, Z.F. Ribosome Structure, Function, and Early Evolution. *Int J Mol Sci* **2018**, *20*, doi:10.3390/ijms20010040.
52. Root-Bernstein, R.; Root-Bernstein, M. The ribosome as a missing link in prebiotic evolution II: Ribosomes encode ribosomal proteins that bind to common regions of their own mRNAs and rRNAs. *J Theor Biol* **2016**, *397*, 115-127, doi:10.1016/j.jtbi.2016.02.030.
53. Root-Bernstein, M.; Root-Bernstein, R. The ribosome as a missing link in the evolution of life. *J Theor Biol* **2015**, *367*, 130-158, doi:10.1016/j.jtbi.2014.11.025.
54. Torres de Farias, S.; Gaudencio Rego, T.; Jose, M.V. Peptidyl Transferase Center and the Emergence of the Translation System. *Life (Basel)* **2017**, *7*, doi:10.3390/life7020021.
55. Farias, S.T.d.; Rêgo, T.G.; José, M.V. Origin of the 16S Ribosomal Molecule from Ancestor tRNAs. *MDPI* **2019**, https://doi.org/10.3390/sci1010008, doi:https://doi.org/10.3390/sci1010008.
56. Root-Bernstein, R.; Root-Bernstein, M. The Ribosome as a Missing Link in Prebiotic Evolution III: Over-Representation of tRNA- and rRNA-Like Sequences and Plieofunctionality of Ribosome-Related Molecules Argues for the Evolution of Primitive Genomes from Ribosomal RNA Modules. *Int J Mol Sci* **2019**, *20*, doi:10.3390/ijms20010140.

57. Bernier, C.R.; Petrov, A.S.; Kovacs, N.A.; Penev, P.I.; Williams, L.D. Translation: The Universal Structural Core of Life. *Mol Biol Evol* **2018**, *35*, 2065-2076, doi:10.1093/molbev/msy101.
58. Gulen, B.; Petrov, A.S.; Okafor, C.D.; Vander Wood, D.; O'Neill, E.B.; Hud, N.V.; Williams, L.D. Ribosomal small subunit domains radiate from a central core. *Sci Rep* **2016**, *6*, 20885, doi:10.1038/srep20885.
59. Burton, Z.F. The 3-Minihelix tRNA Evolution Theorem. *J Mol Evol* **2020**, 10.1007/s00239-020-09928-2, doi:10.1007/s00239-020-09928-2.
60. Kim, Y.; Kowiatek, B.; Opron, K.; Burton, Z.F. Type-II tRNAs and Evolution of Translation Systems and the Genetic Code. *Int J Mol Sci* **2018**, *19*, doi:10.3390/ijms19103275.
61. Pak, D.; Root-Bernstein, R.; Burton, Z.F. tRNA structure and evolution and standardization to the three nucleotide genetic code. *Transcription* **2017**, *8*, 205-219, doi:10.1080/21541264.2017.1318811.
62. Root-Bernstein, R.; Kim, Y.; Sanjay, A.; Burton, Z.F. tRNA evolution from the proto-tRNA minihelix world. *Transcription* **2016**, *7*, 153-163, doi:10.1080/21541264.2016.1235527.
63. Juhling, F.; Morl, M.; Hartmann, R.K.; Sprinzl, M.; Stadler, P.F.; Putz, J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* **2009**, *37*, D159-162, doi:10.1093/nar/gkn772.
64. Yang, Z.; Lasker, K.; Schneidman-Duhovny, D.; Webb, B.; Huang, C.C.; Pettersen, E.F.; Goddard, T.D.; Meng, E.C.; Sali, A.; Ferrin, T.E. UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *J Struct Biol* **2012**, *179*, 269-278, doi:10.1016/j.jsb.2011.09.006.
65. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **2004**, *25*, 1605-1612, doi:10.1002/jcc.20084.
66. Goddard, T.D.; Huang, C.C.; Meng, E.C.; Pettersen, E.F.; Couch, G.S.; Morris, J.H.; Ferrin, T.E. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **2018**, *27*, 14-25, doi:10.1002/pro.3235.
67. Perona, J.J.; Gruic-Sovulj, I. Synthetic and editing mechanisms of aminoacyl-tRNA synthetases. *Top Curr Chem* **2014**, *344*, 1-41, doi:10.1007/128_2013_456.
68. Gospodinov, A.; Kunnev, D. Universal Codons with Enrichment from GC to AU Nucleotide Composition Reveal a Chronological Assignment from Early to Late Along with LUCA Formation. *Life (Basel)* **2020**, *10*, doi:10.3390/life10060081.
69. Demongeot, J.; Seligmann, H. RNA Rings Strengthen Hairpin Accretion Hypotheses for tRNA Evolution: A Reply to Commentaries by Z.F. Burton and M. Di Giulio. *J Mol Evol* **2020**, 10.1007/s00239-020-09929-1, doi:10.1007/s00239-020-09929-1.
70. Di Giulio, M. A comparison between two models for understanding the origin of the tRNA molecule. *J Theor Biol* **2019**, *480*, 99-103, doi:10.1016/j.jtbi.2019.07.020.
71. Loveland, A.B.; Demo, G.; Grigorieff, N.; Korostelev, A.A. Ensemble cryo-EM elucidates the mechanism of translation fidelity. *Nature* **2017**, *546*, 113-117, doi:10.1038/nature22397.

72. Rozov, A.; Wolff, P.; Grosjean, H.; Yusupov, M.; Yusupova, G.; Westhof, E. Tautomeric G*U pairs within the molecular ribosomal grip and fidelity of decoding in bacteria. *Nucleic Acids Res* **2018**, *46*, 7425-7435, doi:10.1093/nar/gky547.
73. Rozov, A.; Demeshkina, N.; Westhof, E.; Yusupov, M.; Yusupova, G. New Structural Insights into Translational Miscoding. *Trends Biochem Sci* **2016**, *41*, 798-814, doi:10.1016/j.tibs.2016.06.001.
74. Rozov, A.; Westhof, E.; Yusupov, M.; Yusupova, G. The ribosome prohibits the G*U wobble geometry at the first position of the codon-anticodon helix. *Nucleic Acids Res* **2016**, *44*, 6434-6441, doi:10.1093/nar/gkw431.
75. Loveland, A.B.; Demo, G.; Korostelev, A.A. Cryo-EM of elongating ribosome with EF-Tu*GTP elucidates tRNA proofreading. *Nature* **2020**, 10.1038/s41586-020-2447-x, doi:10.1038/s41586-020-2447-x.
76. Saint-Leger, A.; Bello, C.; Dans, P.D.; Torres, A.G.; Novoa, E.M.; Camacho, N.; Orozco, M.; Kondrashov, F.A.; Ribas de Pouplana, L. Saturation of recognition elements blocks evolution of new tRNA identities. *Sci Adv* **2016**, *2*, e1501860, doi:10.1126/sciadv.1501860.
77. Agris, P.F.; Eruysal, E.R.; Narendran, A.; Vare, V.Y.P.; Vangaveti, S.; Ranganathan, S.V. Celebrating wobble decoding: Half a century and still much is new. *RNA Biol* **2018**, *15*, 537-553, doi:10.1080/15476286.2017.1356562.
78. Agris, P.F.; Narendran, A.; Sarachan, K.; Vare, V.Y.P.; Eruysal, E. The Importance of Being Modified: The Role of RNA Modifications in Translational Fidelity. *Enzymes* **2017**, *41*, 1-50, doi:10.1016/bs.enz.2017.03.005.
79. Burroughs, A.M.; Aravind, L. The Origin and Evolution of Release Factors: Implications for Translation Termination, Ribosome Rescue, and Quality Control Pathways. *Int J Mol Sci* **2019**, *20*, doi:10.3390/ijms20081981.
80. Satpati, P.; Bauer, P.; Aqvist, J. Energetic tuning by tRNA modifications ensures correct decoding of isoleucine and methionine on the ribosome. *Chemistry* **2014**, *20*, 10271-10275, doi:10.1002/chem.201404016.
81. Kohrer, C.; Mandal, D.; Gaston, K.W.; Grosjean, H.; Limbach, P.A.; Rajbhandary, U.L. Life without tRNAIle-lysidine synthetase: translation of the isoleucine codon AUA in Bacillus subtilis lacking the canonical tRNA2Ile. *Nucleic Acids Res* **2014**, *42*, 1904-1915, doi:10.1093/nar/gkt1009.
82. Voorhees, R.M.; Mandal, D.; Neubauer, C.; Kohrer, C.; RajBhandary, U.L.; Ramakrishnan, V. The structural basis for specific decoding of AUA by isoleucine tRNA on the ribosome. *Nat Struct Mol Biol* **2013**, *20*, 641-643, doi:10.1038/nsmb.2545.
83. Mandal, D.; Kohrer, C.; Su, D.; Russell, S.P.; Krivos, K.; Castleberry, C.M.; Blum, P.; Limbach, P.A.; Soll, D.; RajBhandary, U.L. Agmatidine, a modified cytidine in the anticodon of archaeal tRNA(Ile), base pairs with adenosine but not with guanosine. *Proc Natl Acad Sci U S A* **2010**, *107*, 2872-2877, doi:10.1073/pnas.0914869107.
84. Zhang, B.; Cech, T.R. Peptidyl-transferase ribozymes: trans reactions, structural characterization and ribosomal RNA-like features. *Chem Biol* **1998**, *5*, 539-553, doi:10.1016/s1074-5521(98)90113-2.

85. Petrov, A.S.; Gulen, B.; Norris, A.M.; Kovacs, N.A.; Bernier, C.R.; Lanier, K.A.; Fox, G.E.; Harvey, S.C.; Wartell, R.M.; Hud, N.V., et al. History of the ribosome and the origin of translation. *Proc Natl Acad Sci U S A* **2015**, *112*, 15396-15401, doi:10.1073/pnas.1509761112.
86. Petrov, A.S.; Bernier, C.R.; Hsiao, C.; Norris, A.M.; Kovacs, N.A.; Waterbury, C.C.; Stepanov, V.G.; Harvey, S.C.; Fox, G.E.; Wartell, R.M., et al. Evolution of the ribosome at atomic resolution. *Proc Natl Acad Sci U S A* **2014**, *111*, 10251-10256, doi:10.1073/pnas.1407205111.
87. Llacer, J.L.; Hussain, T.; Marler, L.; Aitken, C.E.; Thakur, A.; Lorsch, J.R.; Hinnebusch, A.G.; Ramakrishnan, V. Conformational Differences between Open and Closed States of the Eukaryotic Translation Initiation Complex. *Mol Cell* **2015**, *59*, 399-412, doi:10.1016/j.molcel.2015.06.033.
88. Koonin, E.V. Frozen Accident Pushing 50: Stereochemistry, Expansion, and Chance in the Evolution of the Genetic Code. *Life (Basel)* **2017**, *7*, doi:10.3390/life7020022.
89. O'Donoghue, P.; Luthey-Schulten, Z. On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol Mol Biol Rev* **2003**, *67*, 550-573.
90. Kelley, L.A.; Mezulis, S.; Yates, C.M.; Wass, M.N.; Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **2015**, *10*, 845-858, doi:10.1038/nprot.2015.053.
91. Illangasekare, M.; Yarus, M. Small aminoacyl transfer centers at GU within a larger RNA. *RNA Biol* **2012**, *9*, 59-66, doi:10.4161/rna.9.1.18039.
92. Yarus, M. The meaning of a minuscule ribozyme. *Philos Trans R Soc Lond B Biol Sci* **2011**, *366*, 2902-2909, doi:10.1098/rstb.2011.0139.
93. Turk, R.M.; Chumachenko, N.V.; Yarus, M. Multiple translational products from a five-nucleotide ribozyme. *Proc Natl Acad Sci U S A* **2010**, *107*, 4585-4589, doi:10.1073/pnas.0912895107.
94. Wolf, Y.I.; Aravind, L.; Grishin, N.V.; Koonin, E.V. Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* **1999**, *9*, 689-710.
95. Martinez-Rodriguez, L.; Erdogan, O.; Jimenez-Rodriguez, M.; Gonzalez-Rivera, K.; Williams, T.; Li, L.; Weinreb, V.; Collier, M.; Chandrasekaran, S.N.; Ambroggio, X., et al. Functional Class I and II Amino Acid-activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. *J Biol Chem* **2015**, *290*, 19710-19725, doi:10.1074/jbc.M115.642876.
96. Carter, C.W., Jr.; Li, L.; Weinreb, V.; Collier, M.; Gonzalez-Rivera, K.; Jimenez-Rodriguez, M.; Erdogan, O.; Kuhlman, B.; Ambroggio, X.; Williams, T., et al. The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. *Biol Direct* **2014**, *9*, 11, doi:10.1186/1745-6150-9-11.
97. Chandrasekaran, S.N.; Yardimci, G.G.; Erdogan, O.; Roach, J.; Carter, C.W., Jr. Statistical evaluation of the Rodin-Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases. *Mol Biol Evol* **2013**, *30*, 1588-1604, doi:10.1093/molbev/mst070.

98. Rodin, A.S.; Rodin, S.N.; Carter, C.W., Jr. On primordial sense-antisense coding. *J Mol Evol* **2009**, *69*, 555-567, doi:10.1007/s00239-009-9288-4.
99. Mukai, T.; Crnkovic, A.; Umehara, T.; Ivanova, N.N.; Kyrpides, N.C.; Soll, D. RNA-Dependent Cysteine Biosynthesis in Bacteria and Archaea. *MBio* **2017**, *8*, doi:10.1128/mBio.00561-17.
100. Turanov, A.A.; Xu, X.M.; Carlson, B.A.; Yoo, M.H.; Gladyshev, V.N.; Hatfield, D.L. Biosynthesis of selenocysteine, the 21st amino acid in the genetic code, and a novel pathway for cysteine biosynthesis. *Adv Nutr* **2011**, *2*, 122-128, doi:10.3945/an.110.000265.
101. Hauenstein, S.I.; Perona, J.J. Redundant synthesis of cysteinyl-tRNACys in Methanosarcina mazei. *J Biol Chem* **2008**, *283*, 22007-22017, doi:10.1074/jbc.M801839200.
102. Sheppard, K.; Akochy, P.M.; Salazar, J.C.; Soll, D. The Helicobacter pylori amidotransferase GatCAB is equally efficient in glutamine-dependent transamidation of Asp-tRNAAsn and Glu-tRNAGln. *J Biol Chem* **2007**, *282*, 11866-11873, doi:10.1074/jbc.M700398200.
103. Feng, L.; Sheppard, K.; Tumbula-Hansen, D.; Soll, D. Gln-tRNAGln formation from Glu-tRNAGln requires cooperation of an asparaginase and a Glu-tRNAGln kinase. *J Biol Chem* **2005**, *280*, 8150-8155, doi:10.1074/jbc.M411098200.
104. Kim, S.I.; Nalaskowska, M.; Germond, J.E.; Pridmore, D.; Soll, D. Asn-tRNA in Lactobacillus bulgaricus is formed by asparaginylation of tRNA and not by transamidation of Asp-tRNA. *Nucleic Acids Res* **1996**, *24*, 2648-2651, doi:10.1093/nar/24.14.2648.
105. Rogers, K.C.; Soll, D. Divergence of glutamate and glutamine aminoacylation pathways: providing the evolutionary rationale for mischarging. *J Mol Evol* **1995**, *40*, 476-481, doi:10.1007/bf00166615.
106. Schon, A.; Hottinger, H.; Soll, D. Misaminoacylation and transamidation are required for protein biosynthesis in Lactobacillus bulgaricus. *Biochimie* **1988**, *70*, 391-394, doi:10.1016/0300-9084(88)90212-x.
107. Mukai, T.; Reynolds, N.M.; Crnkovic, A.; Soll, D. Bioinformatic Analysis Reveals Archaeal tRNATyr and tRNATrp Identities in Bacteria. *Life (Basel)* **2017**, *7*, doi:10.3390/life7010008.
108. Fournier, G.P.; Alm, E.J. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. *J Mol Evol* **2015**, *80*, 171-185, doi:10.1007/s00239-015-9672-1.
109. Cantine, M.D.; Fournier, G.P. Environmental Adaptation from the Origin of Life to the Last Universal Common Ancestor. *Orig Life Evol Biosph* **2018**, *48*, 35-54, doi:10.1007/s11084-017-9542-5.
110. Ribas de Pouplana, L.; Torres, A.G.; Rafels-Ybern, A. What Froze the Genetic Code? *Life (Basel)* **2017**, *7*, doi:10.3390/life7020014.
111. Lane, N. Proton gradients at the origin of life. *Bioessays* **2017**, *39*, doi:10.1002/bies.201600217.

112. Sojo, V.; Pomiankowski, A.; Lane, N. A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol* **2014**, *12*, e1001926, doi:10.1371/journal.pbio.1001926.
113. Lane, N. Bioenergetic constraints on the evolution of complex life. *Cold Spring Harb Perspect Biol* **2014**, *6*, a015982, doi:10.1101/cshperspect.a015982.
114. Quigley, G.J.; Rich, A. Structural domains of transfer RNA molecules. *Science* **1976**, *194*, 796-806.
115. McGeoch, M.W.; Dikler, S.; McGeoch, J.E.M. Hemolithin: a Meteoritic Protein containing Iron and Lithium. *arXiv.org* **2020**.
116. Dunne, M.; Denyes, J.M.; Arndt, H.; Loessner, M.J.; Leiman, P.G.; Klumpp, J. Salmonella Phage S16 Tail Fiber Adhesin Features a Rare Polyglycine Rich Domain for Host Recognition. *Structure* **2018**, *26*, 1573-1582 e1574, doi:10.1016/j.str.2018.07.017.
117. Endow, J.K.; Rocha, A.G.; Baldwin, A.J.; Roston, R.L.; Yamaguchi, T.; Kamikubo, H.; Inoue, K. Polyglycine Acts as a Rejection Signal for Protein Transport at the Chloroplast Envelope. *PLoS One* **2016**, *11*, e0167802, doi:10.1371/journal.pone.0167802.
118. Inoue, K.; Keegstra, K. A polyglycine stretch is necessary for proper targeting of the protein translocation channel precursor to the outer envelope membrane of chloroplasts. *Plant J* **2003**, *34*, 661-669, doi:10.1046/j.1365-313x.2003.01755.x.
119. Lorusso, M.; Pepe, A.; Ibrisb, N.; Bochicchio, B. Molecular and supramolecular studies on polyglycine and poly-l-proline *Soft Matter* **2011**, *7*, 6327-6336
120. Pinho, M.G.; Kjos, M.; Veening, J.W. How to get (a)round: mechanisms controlling growth and division of coccoid bacteria. *Nat Rev Microbiol* **2013**, *11*, 601-614, doi:10.1038/nrmicro3088.
121. Zapun, A.; Vernet, T.; Pinho, M.G. The different shapes of cocci. *FEMS Microbiol Rev* **2008**, *32*, 345-360, doi:10.1111/j.1574-6976.2007.00098.x.
122. Scheffers, D.J.; Pinho, M.G. Bacterial cell wall synthesis: new insights from localization studies. *Microbiol Mol Biol Rev* **2005**, *69*, 585-607, doi:10.1128/MMBR.69.4.585-607.2005.
123. Yin, Z.; Wu, F.; Xing, T.; Yadavalli, V.K.; Kunduc, S.C.; Lu, S. A silk fibroin hydrogel with reversible sol–gel transition. *RSC Advances* **2017**.
124. Zamudio, G.S.; Jose, M.V. Phenotypic Graphs and Evolution Unfold the Standard Genetic Code as the Optimal. *Orig Life Evol Biosph* **2018**, *48*, 83-91, doi:10.1007/s11084-017-9552-3.
125. Jose, M.V.; Zamudio, G.S.; Morgado, E.R. A unified model of the standard genetic code. *R Soc Open Sci* **2017**, *4*, 160908, doi:10.1098/rsos.160908.
126. Ikehara, K. Evolutionary Steps in the Emergence of Life Deduced from the Bottom-Up Approach and GADV Hypothesis (Top-Down Approach). *Life (Basel)* **2016**, *6*, doi:10.3390/life6010006.
127. Ikehara, K. [GADV]-protein world hypothesis on the origin of life. *Orig Life Evol Biosph* **2014**, *44*, 299-302, doi:10.1007/s11084-014-9383-4.
128. Ikehara, K. Pseudo-replication of [GADV]-proteins and origin of life. *Int J Mol Sci* **2009**, *10*, 1525-1537, doi:10.3390/ijms10041525.

129. Oba, T.; Fukushima, J.; Maruyama, M.; Iwamoto, R.; Ikehara, K. Catalytic activities of [GADV]-peptides. Formation and establishment of [GADV]-protein world for the emergence of life. *Orig Life Evol Biosph* **2005**, *35*, 447-460, doi:10.1007/s11084-005-3519-5.
130. Ikehara, K. Possible steps to the emergence of life: the [GADV]-protein world hypothesis. *Chem Rec* **2005**, *5*, 107-118, doi:10.1002/tcr.20037.
131. Pellegrino, S.; Demeshkina, N.; Mancera-Martinez, E.; Melnikov, S.; Simonetti, A.; Myasnikov, A.; Yusupov, M.; Yusupova, G.; Hashem, Y. Structural Insights into the Role of Diphthamide on Elongation Factor 2 in mRNA Reading-Frame Maintenance. *J Mol Biol* **2018**, *430*, 2677-2687, doi:10.1016/j.jmb.2018.06.006.
132. Demeshkina, N.; Jenner, L.; Westhof, E.; Yusupov, M.; Yusupova, G. New structural insights into the decoding mechanism: translation infidelity via a G.U pair with Watson-Crick geometry. *FEBS Lett* **2013**, *587*, 1848-1857, doi:10.1016/j.febslet.2013.05.009.
133. Sheppard, K.; Sherrer, R.L.; Soll, D. Methanothermobacter thermautotrophicus tRNA Gln confines the amidotransferase GatCAB to asparaginyl-tRNA Asn formation. *J Mol Biol* **2008**, *377*, 845-853, doi:10.1016/j.jmb.2008.01.064.
134. Sheppard, K.; Yuan, J.; Hohn, M.J.; Jester, B.; Devine, K.M.; Soll, D. From one amino acid to another: tRNA-dependent amino acid biosynthesis. *Nucleic Acids Res* **2008**, *36*, 1813-1825, doi:10.1093/nar/gkn015.
135. Perona, J.J. Two-step pathway to aminoacylated tRNA. *Structure* **2005**, *13*, 1397-1398, doi:10.1016/j.str.2005.09.003.
136. Longo, L.M.; Despotovic, D.; Weil-Ktorza, O.; Walker, M.J.; Jablonska, J.; Fridmann-Sirkis, Y.; Varani, G.; Metanis, N.; Tawfik, D.S. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc Natl Acad Sci U S A* **2020**, 10.1073/pnas.2001989117, doi:10.1073/pnas.2001989117.
137. Bedard, A.V.; Hien, E.D.M.; Lafontaine, D.A. Riboswitch regulation mechanisms: RNA, metabolites and regulatory proteins. *Biochim Biophys Acta Gene Regul Mech* **2020**, *1863*, 194501, doi:10.1016/j.bbagrm.2020.194501.
138. Preiner, M.; Asche, S.; Becker, S.; Betts, H.C.; Boniface, A.; Camprubi, E.; Chandru, K.; Erastova, V.; Garg, S.G.; Khawaja, N., et al. The Future of Origin of Life Research: Bridging Decades-Old Divisions. *Life (Basel)* **2020**, *10*, doi:10.3390/life10030020.
139. Hansma, H.G. Better than Membranes at the Origin of Life? *Life (Basel)* **2017**, *7*, doi:10.3390/life7020028.
140. Hansma, H.G. The power of crowding for the origins of life. *Orig Life Evol Biosph* **2014**, *44*, 307-311, doi:10.1007/s11084-014-9382-5.
141. Hansma, H.G. Possible origin of life between mica sheets: does life imitate mica? *J Biomol Struct Dyn* **2013**, *31*, 888-895, doi:10.1080/07391102.2012.718528.
142. Chen, H.; Cui, Y.; Han, X.; Hu, W.; Sun, M.; Zhang, Y.; Wang, P.H.; Song, G.; Chen, W.; Lou, J. Liquid-liquid phase separation by SARS-CoV-2 nucleocapsid protein and RNA. *Cell Res* **2020**, 10.1038/s41422-020-00408-2, doi:10.1038/s41422-020-00408-2.

143. Cerase, A.; Tartaglia, G.G. Long non-coding RNA-polycomb intimate rendezvous. *Open Biol* **2020**, *10*, 200126, doi:10.1098/rsob.200126.
144. Lafontaine, D.L.J.; Riback, J.A.; Bascetin, R.; Brangwynne, C.P. The nucleolus as a multiphase liquid condensate. *Nat Rev Mol Cell Biol* **2020**, 10.1038/s41580-020-0272-6, doi:10.1038/s41580-020-0272-6.
145. Gotor, N.L.; Armaos, A.; Calloni, G.; Torrent Burgas, M.; Vabulas, R.M.; De Groot, N.S.; Tartaglia, G.G. RNA-binding and prion domains: the Yin and Yang of phase separation. *Nucleic Acids Res* **2020**, 10.1093/nar/gkaa681, doi:10.1093/nar/gkaa681.
146. Irastortza-Olaziregi, M.; Amster-Choder, O. RNA localization in prokaryotes: Where, when, how, and why. *Wiley Interdiscip Rev RNA* **2020**, 10.1002/wrna.1615, e1615, doi:10.1002/wrna.1615.
147. Ding, X.; Sun, F.; Chen, J.; Chen, L.; Tobin-Miyaji, Y.; Xue, S.; Qiang, W.; Luo, S.Z. Amyloid-Forming Segment Induces Aggregation of FUS-LC Domain from Phase Separation Modulated by Site-Specific Phosphorylation. *J Mol Biol* **2020**, *432*, 467-483, doi:10.1016/j.jmb.2019.11.017.
148. Chen, C.; Ding, X.; Akram, N.; Xue, S.; Luo, S.Z. Fused in Sarcoma: Properties, Self-Assembly and Correlation with Neurodegenerative Diseases. *Molecules* **2019**, *24*, doi:10.3390/molecules24081622.
149. Wang, B.; Predeus, A.V.; Burton, Z.F.; Feig, M. Energetic and structural details of the trigger-loop closing transition in RNA polymerase II. *Biophys J* **2013**, *105*, 767-775, doi:10.1016/j.bpj.2013.05.060.
150. Mazumder, A.; Lin, M.; Kapanidis, A.N.; Ebright, R.H. Closing and opening of the RNA polymerase trigger loop. *Proc Natl Acad Sci U S A* **2020**, *117*, 15642-15649, doi:10.1073/pnas.1920427117.