Order of amino acid recruitment into the genetic code resolved by Last Universal Common Ancestor's protein domains

Sawsan Wehbi[1], Andrew Wheeler[1], Benoit Morel[2], Bui Quang Minh[3], Dante S. Lauretta[4], Joanna Masel[5]

[1]Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, Arizona, 85721, USA
[2]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[3]School of Computing, Australian National University, Canberra, ACT, Australia
[4]Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ 85721, USA
[5]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA

Corresponding author: Joanna Masel

Significance

The order in which the amino acids were added to the genetic code was previously inferred from forty metrics. Many of these reflect abiotic abundance on ancient Earth. However, the abundances that matter are those within primitive cells that already had sophisticated RNA and perhaps peptide metabolism. Here we directly infer the order of recruitment from the relative amino acid frequencies of ancient protein sequences. Metal-dependent catalysis and a sulfur-rich environment shaped the early code, with implications for our understanding of early life and hence our search for life elsewhere in the universe.

Abstract

The current "consensus" order in which amino acids were added to the genetic code is based on potentially biased criteria such as absence of sulfur-containing amino acids from the Urey-Miller experiment which lacked sulfur. Even if inferred perfectly, abiotic abundance might not reflect abundance in the organisms in which the genetic code evolved. Here, we instead exploit the fact that proteins that emerged prior to the genetic code's completion are likely enriched in early amino acids and depleted in late amino acids. We identify the most ancient protein-coding sequences born prior to the archaeal-bacterial split. Amino acid usage in protein sequences whose ancestors date back to a single homolog in the Last Universal Common Ancestor (LUCA) largely matches the consensus order. However, our findings indicate that metal-binding (cysteine and histidine) and sulfur-containing (cysteine and methionine) amino acids were added to the genetic code much earlier than previously thought. Surprisingly, even more ancient protein sequences — those that had already diversified into multiple distinct copies in LUCA — show a different pattern to single copy LUCA sequences: significantly less depleted in the late amino acids tryptophan and tyrosine, and enriched rather than depleted in phenylalanine. This is compatible with at least some of these sequences predating the current genetic code. Their distinct enrichment patterns thus provide hints about earlier, alternative genetic codes.

Keywords:
Origins of life, early life, astrobiology, phylostratigraphy, antioxidants, metalloproteins

Introduction

The modern genetic code was likely assembled in stages, hypothesized to begin with "early" amino acids present on Earth before the emergence of life (possibly delivered by extraterrestrial sources such as asteroids or comets), and ending with "late" amino acids requiring biotic synthesis (Fried, Fujishima et al. 2022). For example, the Urey-Miller experiment (Miller 1953) has been used to identify which amino acids were available abiotically and are thus likely to have come earlier than those requiring biotic synthesis. The order of amino acid recruitment, from early to late, was inferred by taking statistical consensus among 40 different rankings (Trifonov 2000), none of which constitute strong evidence on their own. On the basis of this ordering, Moosmann (2021) hypothesized that the first amino acids recruited into the genetic code were those that were useful for membrane anchoring, then those useful for halophilic folding, then for mesophilic folding, then for metal binding, and finally for their antioxidant properties. However, a late role for metal-binding amino acids is puzzling; many metalloproteins date back to the Last Universal Common Ancestor's (LUCA)'s proteome, where they are presumed to be key to the emergence of biological catalysis (Nitschke, McGlynn et al. 2013).

Indeed, the rankings on which this inference is based are disputed (Fried, Fujishima et al. 2022). For example, the Urey-Miller experiment (Miller 1953) did not include sulfur, and so should not have been used to infer that the sulfur-containing amino acids cysteine and methionine were late additions. Homocysteine (a product of cysteine degradation) was detected in H_2S-rich spark discharge experiments, suggesting that cysteine could be abiotically produced (Parker, Cleaves et al. 2011). A nitrile-activated dehydroalanine pathway can produce cysteine from abiotic serine that is produced from a Strecker reaction (Foden, Islam et al. 2020), further demonstrating the possibility of its early chemical availability.

Histidine's classification as abiotically unavailable also contributed to its annotation as late (Trifonov 2000). While histidine can be abiotically synthesized from erythrose reacting with formamidine followed by a Strecker synthesis reaction (Shen, Yang et al. 1990), the reactant concentrations might have been insufficient in a primitive earth environment (Vázquez-Salazar, Becerra et al. 2018). More importantly, because histidine resembles a purine, even if histidine were abiotically unavailable, it would have had cellular availability at the time of genetic code construction (Goldman and Kacar 2021), in an organism that biotically synthesized ribosomes, and that might also have already utilized amino acids and peptides. Indeed, histidine is the most commonly conserved residue in the active site of enzymes (Ribeiro, Tyzack et al. 2020).

To directly infer the order of recruitment from protein sequence data, without reference to abiotic availability arguments, we consider that some of LUCA's proteins were born prior to the completion of the genetic code. Given sufficiently slow subsequent evolution, we predict that their modern descendants remain enriched in early amino acids and depleted in late amino acids (James, Willis et al. 2021). Previous analyses relied on conserved residues within a small number of LUCA proteins (Brooks, Fresco et al. 2002, Fournier and Gogarten 2007). Here, we classify a larger set of protein-coding domains that date back to LUCA, rather than being more recently born, e.g., *de novo* from non-coding sequences or alternative reading frames (Keese and Gibbs 1992, Van Oss and Carvunis 2019), and look at their overall amino acid frequencies. We use variations in contemporary amino acid frequencies among protein cohorts born at different times to deduce the sequence in which amino acids were incorporated into the genetic code.

A previous annotation of LUCA's proteins used only 1,847 bacterial and 134 archaeal genomes (Weiss, Sousa et al. 2016), creating an undersampling problem (Berkemer and McGlynn 2020) that yielded implausible results (Gogarten and Deamer 2016). For example, Weiss et al. (Weiss,

Sousa et al. 2016) claimed that LUCA had the genes needed for nitrogen fixation. However, more extensive phylogenetic reconstruction showed that nitrogen fixation emerged in the Last Bacterial Common Ancestor (LBCA) and was later horizontally transferred to archaea (Pi, Lin et al. 2022). Weiss et al. (2016) only recovered 10 of the aminoacyl-tRNA synthetases (aaRSs) that each specifically catalyze the attachment of one amino acid to its corresponding tRNAs, despite consensus that LUCA had a relatively mature genetic code (Woese, Olsen et al. 2000).

Here, we take advantage of improved gene tree inference methods (Morel, Kozlov et al. 2019), and broader species sampling, to infer LUCA's protein sequences. Different protein domains within the same gene might have different ages, so we assign ages not to whole proteins but to protein domains (see Methods), as classified by the Pfam database (Mistry, Chuguransky et al. 2021). We recognize Pfams present in LUCA by trimming horizontal gene transfer (HGT) events, and by exploiting long archaeal-bacterial branches (Figure 1). By analyzing ancient amino acid usage, we infer the order in which the code was constructed.

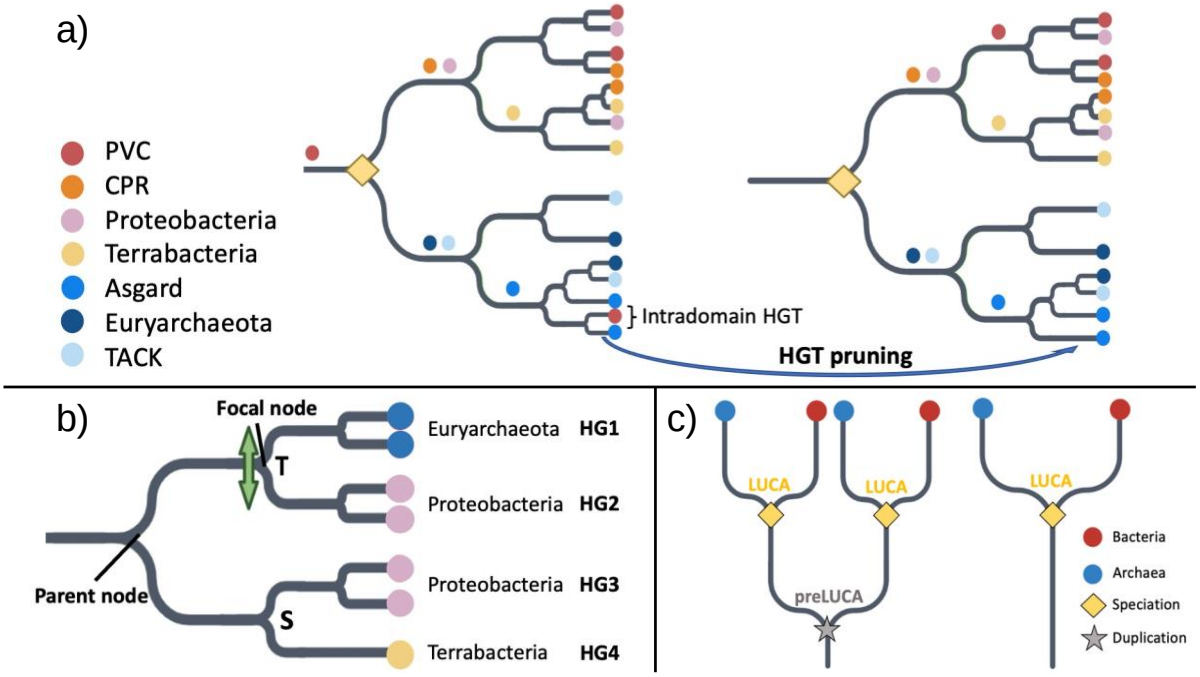


Figure 1. Criteria for (a) LUCA Pfam annotation, (b) identifying HGT to be filtered, and (c) pre-LUCA Pfam annotation. Details are in Methods, with a brief summary here. a) Pruning HGT between archaea and bacteria reveals a LUCA node as dividing bacteria and archaea at the root. Colored circles are indicated just upstream of the most recent common ancestor (MRCA) of all copies of that Pfam found within the same taxonomic supergroup. We recognize five bacterial supergroups (FCB, PVC, CPR, Terrabacteria and Proteobacteria (Rinke, Schwientek et al. 2013, Brown, Hug et al. 2015)) and three archaeal supergroups (TACK, DPANN, Asgard and Euryarchaeota (Baker, De Anda et al. 2020, Shu and Huang 2021)). The yellow diamond indicates LUCA as a speciation event between archaea and bacteria. Note that we do not assume that the LUCA coalescence timing was the same for every Pfam (Doolittle 2000). Prior to HGT pruning, PVC sequences can be found on either side of the two lineages divided by the root. After pruning intradomain HGT, four MRCAs are found one node away from the root, and three more MRCAs are found two nodes away from the root, fulfilling our other LUCA criterion described in the Methods, namely presence of at least three bacterial and at least two archaeal supergroup MRCAs

one to two nodes away from the root. b) Criteria for pruning likely HGT between archaea and bacteria (see Methods for details). We partition into monophyletic groups of sequences in the same supergroup; in this example, there are four such groups, representing two bacterial supergroups and one archaeal supergroup. There is one 'mixed' node, separating an archaeal group (HG1) from a bacterial group (HG2). It is also annotated by GeneRax as a transfer 'T'. The bacterial nature of groups 3 and 4 indicates a putative HGT direction from group 2 to group 1. Group 2 does not contain any Euryarchaeota sequences, meeting the third and final requirement for pruning of group 1. If neither Proteobacteria or Euryarchaeota sequences were present among the other descendants of the parent node, both groups 1 and 2 would be considered acceptors of a transferred Pfam and would both be pruned from the tree. c) Pre-LUCA Pfams have at least two nodes annotated as LUCA.

Results

Ancient proteins remain depleted in late amino acids

Our high-throughput set of LUCA sequences classify 12% of our Pfam dataset and 10% of our clan (sets of Pfams that are evolutionary related) dataset as present in LUCA (Figure 2, including both LUCA and pre-LUCA), compared to 3% of protein family clusters considered by Weiss et al. (Weiss, Sousa et al. 2016). Our age assignment approach provides 2.7 times as many Pfams and 1.8 times as many clans present in LUCA, with only modest overlap, albeit more than for earlier works (Crapitto, Campbell et al. 2022). Our annotations are also of higher quality than the Weiss et al. LUCA annotations (2016), as evidenced by agreement with previous detailed case studies (Supplementary Results).

Hydrophobic amino acids are more interspersed along the primary sequence in LUCA clans than in the still-ancient 'post-LUCA' clans that include Last Archaeal Common Ancestor (LACA) candidates and LBCA candidates (Supplementary Figure 1; Wilcoxon rank sum test; $p = 0.02$), and than in 'modern' clans that are specific to particular prokaryotic supergroups ($p = 0.0002$). Post-LUCA and modern clans also differ from each other (Wilcoxon rank sum test; $p = 0.02$). This confirms the antiquity of this long-term trend proposed by (James, Willis et al. 2021) on the basis of animal Pfam domains.

Clans present in LUCA were born before the divergence of Archaea and Bacteria, some potentially prior to the completion of the genetic code. If newly recruited amino acids were added slowly, the contemporary descendants of LUCA clans will have lower frequencies of the amino acids that were added late to the genetic code. In contrast, post-LUCA clans are predicted to reflect amino acid usage from the standard genetic code of all 20 amino acids. We first focus on clans present in one copy in LUCA (denoted "LUCA clans"), excluding those that had already duplicated and diverged into multiple sequences (denoted "pre-LUCA clans"). We score amino acid enrichment and depletion as relative to still-ancient post-LUCA clans. Amino acids enriched in LUCA tended to be classified by Trifonov (2000) as earlier additions to the genetic code (Figure 3a; weighted $R^2 = 0.37$, $p = 0.002$). To avoid potential circularity (James, Willis et al. 2021), we excluded those criteria of Trifonov (2000) that are associated with contemporary amino acid composition and found similar results (weighted $R^2 = 0.35$, $p = 0.003$). We interpret this statistical significance as evidence that amino acid usage in LUCA domains reflects the early genetic code in ways that 4 billion years of subsequent evolution have not yet been erased.

Weiss et al., (2016) \ This work	Pre-LUCA	LUCA	Post-LUCA	Modern	Unclassifiable	Total Pfams
LUCA	38	179	124	24	34	399
Not LUCA	63	689	2931	3624	576	7883
Total Pfams	101	868	3055	3648	610	8282

Weiss et al., (2016) \ This work	Pre-LUCA	LUCA	Post-LUCA	Modern	Unclassifiable	Total Clans
LUCA	79	85	53	11	14	242
Not LUCA	57	224	1172	2225	249	3931
Total Clans	137	309	1229	2235	263	4173

Figure 2. Our Pfam and clan age classifications overlap only modestly with those of Weiss et al., 2016. For Weiss et al. (2016) age assignments, we treat each Pfam or clan as having the age of the oldest protein of which it was part. Ancient post-LUCA Pfam classifications include 285 LACA candidates and 2770 LBCA candidates (more analysis would be required to rule out extensive HGT within archaea or bacteria). Modern Pfams are distributed among the prokaryotic supergroups as follows: 9 CPR, 210 FCB, 942 Proteobacteria, 51 PVC, 1111 Terrabacteria, 2 Asgard, 49 TACK, and 177 Euryarchaeota. In addition to supergroup-specific modern Pfams, we classified another 1097 Pfams, present in exactly two bacterial supergroups, as modern post-LBCA. We deemed 15 Pfams unclassifiable due to high inferred HGT rates, 397 due to uncertainty in rooting, and 198 due to ancient rooting combined with absence from too many supergroups (see Methods). Pre-LUCA clans contain at least two LUCA-classified Pfams or one pre-LUCA Pfam, whereas LUCA clans contain exactly one LUCA Pfam. Ancient post-LUCA clans contain no LUCA, pre-LUCA, or unclassified Pfams; they include an ancient post-LUCA Pfam or at least two modern Pfams covering at least two supergroups from only one of either bacteria or archaea. Modern clans include Pfams whose root is assigned at the origin of one supergroup. Finally, unclassifiable clans did not meet any of our clan classification criteria, e.g., because they included both post-LUCA and unclassifiable Pfams.

Sulfur-containing and metal-binding amino acids were added early to the genetic code

Cysteine (C), methionine (M), and histidine (H) are all enriched in LUCA, despite having been previously annotated as late additions to the genetic code (Figure 3a). Enrichment for sulfur-containing C and M suggests that LUCA's environment was rich in sulfur (Neubeck and Freund 2020). C can easily be converted to M via the transsulfuration pathway regulated by the cystathionine gamma-synthase protein (Clausen, Huber et al. 1998), whose one Pfam domain (PF01053) we annotate as present in LUCA.

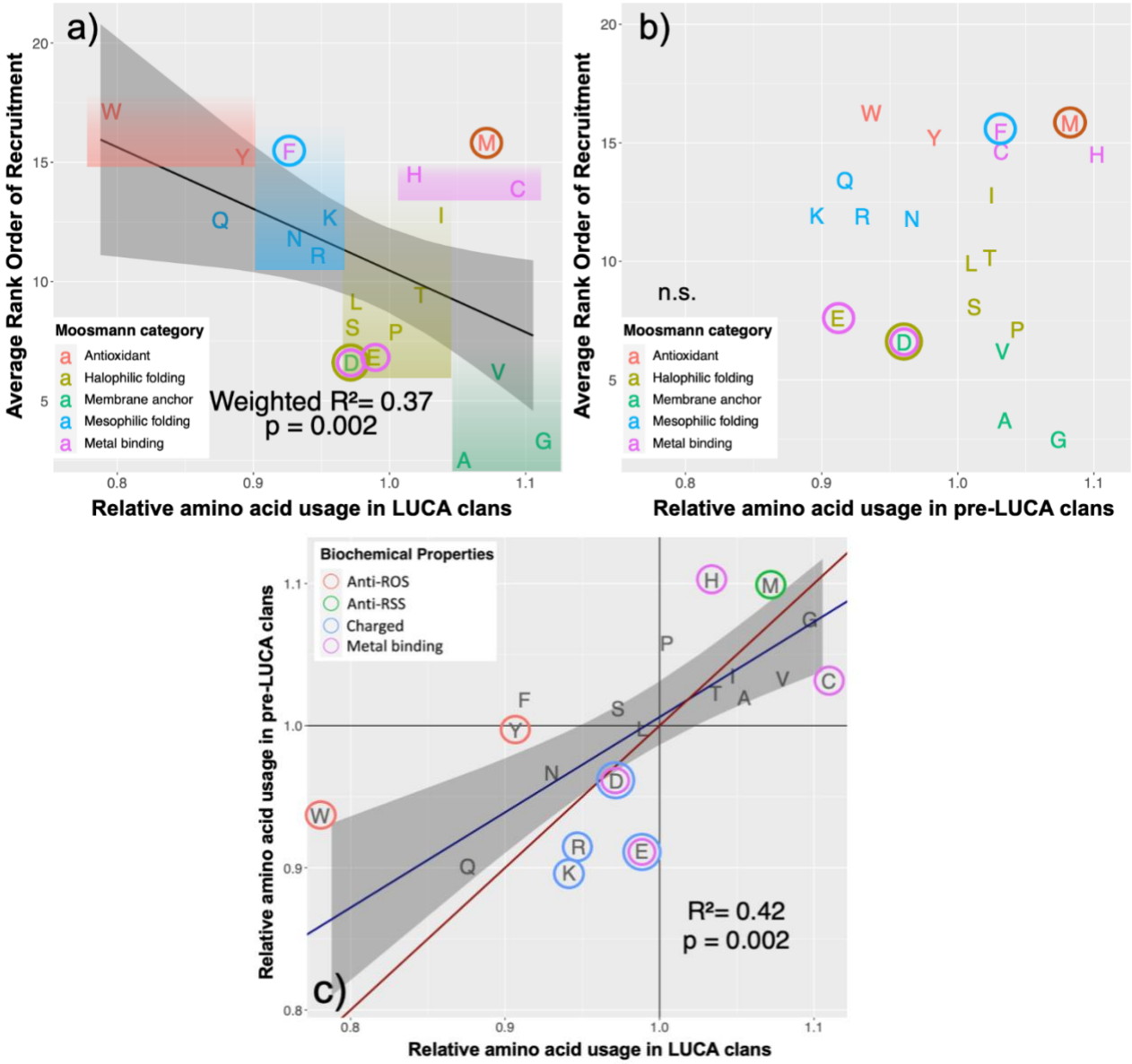


Figure 3. LUCA's single-copy sequences reflect the origins of the current genetic code, whereas multi-copy pre-LUCA sequences provide hints of extinct code(s). Amino acid usage is shown relative to that of ancient post-LUCA clans. The y-axis of a) and b) was calculated by Trifonov (2000) from 40 ranked metrics presumed informative. Associated weights were based on the standard errors in the average rank; because these reflect but considerably underestimate uncertainty, we treat Trifonov's (2000) rankings as the dependent variable in weighted model 1 regressions, shown by a black line and grey 95% shading in a) and omitted as non-significant in b). Character colors in a) and b) show the assignments of Moosmann (2021); colored circles indicate our re-assignments. We categorize the ability to bind transition metals on the basis of Figure 2D of (Li, He et al. 2023). We reclassify phenylalanine (F) because it is enriched in proteins in mesophiles compared to their orthologs in thermophiles and hyperthermophiles (Závodszky 2000). We reclassify aspartic acid (D) because the surfaces of proteins within halophilic bacteria are highly enriched in aspartic acid compared to in the surfaces of non-halophilic mesophilic and

thermophilic bacteria, in a manner that cannot be accounted for by the dinucleotide composition of the halophilic genomes (Fukuchi, Yoshimune et al. 2003). Although pink circles around D and glutamic acid (E) acknowledge their metal-binding capacity, we also retain classification of their halophilic folding property. The brown circle around methionine highlights its ancient antioxidant activity against RSS. c) Model 2 Deming regression in (dark blue line with grey 95% CI) shows that pre-LUCA enrichments are not more extreme versions of LUCA enrichments, lying on the wrong side of the y=x red line, with W, Y, and F as particular outliers from expectations. We used the deming() function in the deming R package (Therneau 2022) to account for standard errors of both variables (Supplementary Table 1). Colored circles show only our own assignments.

Moosmann (2021) classified M, tryptophan (W), and tyrosine (Y) as antioxidants, because in laboratory experiments, they protect the overall protein structure from reactive oxygen species (ROS) via sacrificial oxidization (Levine, Mosoni et al. 1996, Lim, Kim et al. 2019). However, proteins in aerobes are enriched in W and Y but not in M (Vieira-Silva and Rocha 2008). Surprisingly, our results separate early M from late Y and W (Figure 3a). Reactive sulfur species (RSS) rather than ROS would have been the primary oxidizing agents in early, sulfur-rich environments (Neubeck and Freund 2020). We therefore re-classify methionine as "anti-RSS" to distinguish it from the two antioxidants, Y and W, that are adaptations against ROS. We further speculate that methionine's anti-RSS role drove its early addition to the genetic code. Our results are compatible with (Granold, Hajieva et al. 2018)'s view that Y and W were added to complete the modern genetic code after ROS became the main oxidizing agents.

Our findings also point to an earlier role for binding the transition metals that are key to catalysis. Statistical analysis of metalloproteins shows that C and H are key to binding iron, zinc, copper, and molybdenum, that H, aspartic acid (D) and glutamic acid (E or Glu) are key to binding manganese and cobalt, and that all four amino acids bind nickel (Figure 2D of (Li, He et al. 2023)). We find ancient enrichment for C and H, albeit not for the charged metal-binding amino acids D and E (Figure 3a).

Ancient C enrichment could be directly connected to the late addition of W as the 20th amino acid, possibly even post-LUCA (Dong, Zhou et al. 2010). The sole W codon (UGG) differs from C codons (UGU & UGC) only at the third position. This suggests that prior to W's incorporation, UGG might have encoded C, which would have made C more common via mutation bias. We note that the enrichment of W and Y in aerobes is matched by a depletion in C (Vieira-Silva and Rocha 2008). Similarly, H codons could have been recoded to Q.

Although strong W and Y depletion in LUCA was expected, the strength of glutamine (Q or Gln) depletion was not. Consistent with the late addition of Q, the Gln-tRNA synthetase (GlnRS) is absent in many prokaryotes, and prokaryotes that do have GlnRS (e.g., *E. coli*) acquired it via horizontal gene transfer from eukaryotes (Lamour, Quevillon et al. 1994). Prokaryotes that lack GlnRS perform tRNA-dependent amidation of Glu mischarged to Gln-tRNA by GluRS, forming Gln-acylated Gln-tRNA via amidotransferase. The core catalytic domain (PF00587), shared between the GlnRS and GluRS paralogs, is present in LUCA and can indiscriminately acylate both Gln-tRNA and Glu-tRNAs with Glu (Lapointe, Duplain et al. 1986).

Pre-LUCA clans hint at more ancient genetic codes

Although pre-LUCA enrichments do not correlate with Trifonov's consensus order (Trifonov 2000) (Figure 3b), they do correlate with the amino acid usage of LUCA sequences (Figure 3c). Pre-LUCA, like LUCA, is strongly depleted in Q, supporting the inference that Q, not

Y, was the 19th amino acid recruited into the standard genetic code. We expected pre-LUCA enrichments and depletions to be more extreme than for LUCA; this is true for depletion of charged amino acids D, E, R, and K, and for enrichment of H, P and M, but of the 7, only H is statistically significant (Welch 2-sample t-test; $p = 0.02$ prior to correction for multiple comparisons).

In contrast, the anti-ROS amino acids W and Y are significantly less depleted in pre-LUCA than in LUCA (Welch 2-sample t-test; $p = 0.001$ and 0.0009, respectively; 1.4% vs 1.2% W frequencies, 3.2% vs. 2.9% Y frequencies). That the contemporary descendants of pre-LUCA clans have 17% more W than LUCA is particularly surprising, because there is scientific consensus that W was the last of the 20 canonical amino acids to be added to the genetic code.

We manually inspected the pre-LUCA Pfam with the highest tryptophan frequency (3.1%); PF00133 is the core catalytic domain of the tRNA synthetases of leucine (L), isoleucine (I), and valine (V). Each of these three synthetases has well-separated archaeal and bacterial branches, confirming its pre-LUCA dating (Supplementary Figure 2). Highly conserved tryptophan sites Trp449, Trp456, and Trp529 (Chen, Luo et al. 2021) regulate the size of the amino acid binding pocket, allowing the synthetases to discriminate among I, L, and V (Fukai, Nureki et al. 2000). There are also conserved I and V sites in the common ancestor of the I and V tRNA synthetases, indicating that discrimination between the two happened prior to the evolution of the synthetases currently responsible for the discrimination (Fournier, Andam et al. 2011). This suggests that an alternative, more ancient system predated the modern genetic code, and in particular predated the evolution of super-specific, cognate aaRSs (Fournier, Andam et al. 2011). Figure 4 shows how the prior existence of such a system could explain our data.

The strongest lack of concordance is for phenylalanine (F), which is enriched in pre-LUCA but depleted in LUCA (Welch 2-sample test; $p = 0.02$); this is the only amino acid to significantly switch sign. F is highly exchangeable with Y; they are both aromatic and hydrophobic. Y is also biochemically derived from F. Our results could be explained if F sites were replaced by Y after the latter was incorporated to the genetic code.

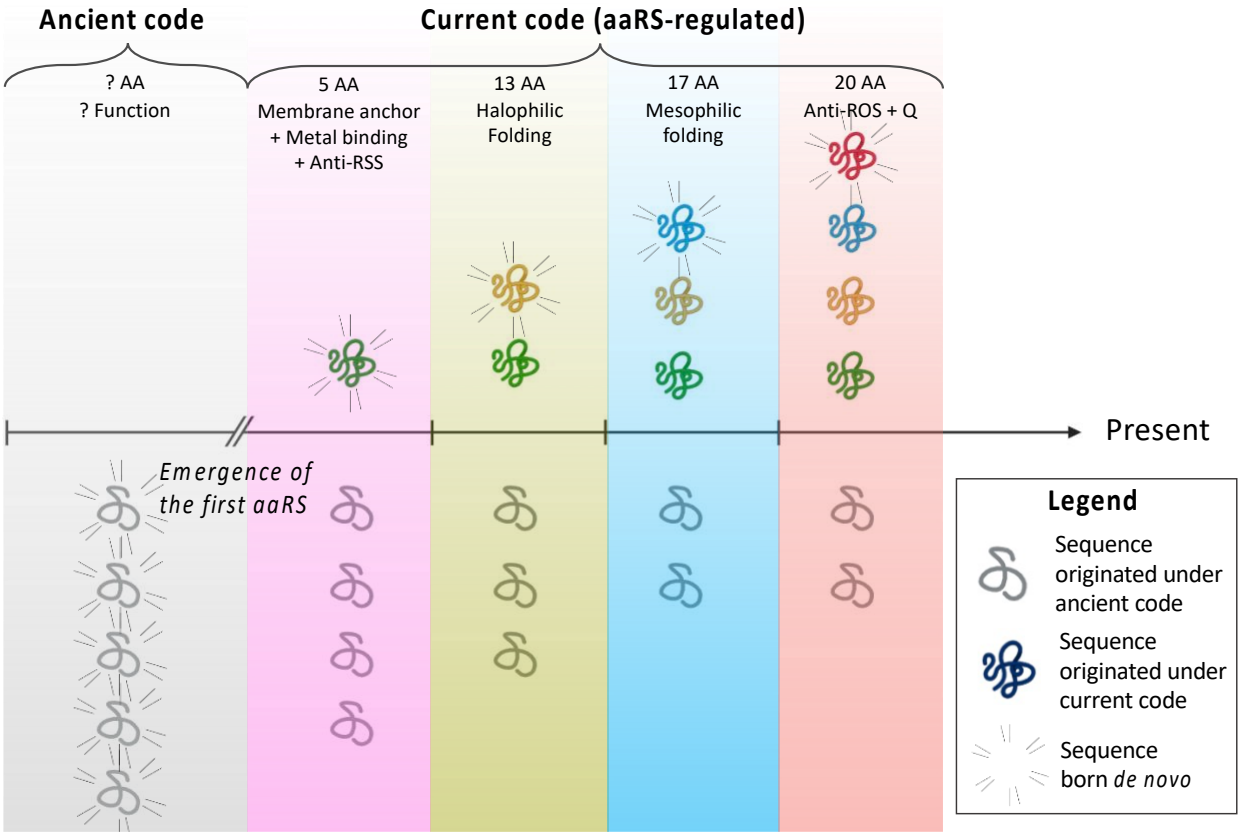


Figure 4. LUCA clans inform the origins of the current genetic code, whereas pre-LUCA clans reflect a still older system. Colored sequences were born at different stages during the construction of the current genetic code ("current" includes the canonical code and its direct ancestors). Grey sequences were born under some more ancient, unknown, now-extinct coding mechanism(s).

Discussion

The evolution of the current genetic code proceeded via stepwise incorporation and replacement of amino acids, driven in part by changes in early life's environment and requirements. Contemporary proteins remain shaped by which amino acids were part of the code at the moment of their birth, allowing us to infer the order of recruitment on the basis of enrichment or depletion in LUCA's protein domains. Our results suggest early incorporation of sulfur-containing amino acids (cysteine and methionine), with methionine protecting against RSS within an ancient sulfur-rich environment (Ranjan, Todd et al. 2018, Fairchild, Islam et al. 2024). Results also support early incorporation of metal-binding amino acids (cysteine and histidine). An early role for histidine also resolves the previous puzzle that the ancestral, conserved region of all Class I aaRSs contains a histidine-rich HIGH motif (Eriani, Delarue et al. 1990, Douglas, Bouckaert et al. 2024). Enrichment patterns place glutamine rather than tyrosine as the 19th amino acid, in agreement with data on glutamyl-tRNA synthetases. Amino acid usage of protein sequences even more ancient that had already duplicated and diverged pre-LUCA tell a different story, enriched rather than depleted in phenylalanine, and significantly less depleted in tyrosine and tryptophan.

Our results should be interpreted as the order of recruitment of amino acids into the code(s) under which contemporary protein-coding sequences were born – this need not match the importance or abundance with which amino acids were used by still earlier life forms, or during the prebiotic to biotic transition. Instead of using Trifonov's assignments (2000) to capture the order in which amino acids were recruited into our genetic code, we recommend using the LUCA amino acid enrichment values plotted on the x-axis of Figure 3a, which can be found together with their standard errors in Supplementary Table 1. Our work does not address abiotic availability, although we note that ongoing research on plausible prebiotic syntheses in cyanosulfidic environments (Fairchild, Islam et al. 2024) is reshaping our understanding of which amino acids were accessible to early life. Amino acid abundances obtained from asteroid sample returns will soon contribute (Lauretta, Adam et al. 2022, Lauretta 2023).

The surprisingly high frequency of tryptophan in the oldest, pre-LUCA sequences could be explained if tryptophan had an early anti-RSS role, fell subsequently out of favor as life migrated to a less sulfur-rich environment, and was later revived — rather than used for the first time — as an anti-ROS antioxidant (Granold, Hajieva et al. 2018). It is also possible that a conserved tryptophan appeared in a concerted manner when a codon previously coding for a different amino acid was recoded to translate tryptophan, with this scenario arising most often for the most ancient sequences.

More broadly, coding for different amino acids might have emerged at similar times but in different biogeochemical environments. The temporal order of recruitment that we infer for LUCA sequences is not the temporal order for coding as a whole, but for the ancestor of the modern genetic code, to which amino acids were likely added as much or more by horizontal gene transfer as by de novo evolution (Froese, Campos et al. 2018).

To explain the different enrichments of pre-LUCA versus LUCA sequences, as well as the surprising conservation of some sites prior to the emergence of the aaRSs that distinguish the relevant amino acids, we propose that some pre-LUCA sequences are older than the current genetic code, perhaps even tracing back to a peptide world at the dawn of precellular life (Fried, Fujishima et al. 2022). Stepwise construction of the current code and competition among ancient codes could have occurred simultaneously (Koonin and Novozhilov 2009, Morgens and Cavalcanti 2013). Ancient codes might also have used non-canonical amino acids, such as norvaline and norleucine

(Alvarez-Carreño, Becerra et al. 2013) which can be recognized by LeuRS (Tang and Tirrell 2002, Mascarenhas, An et al. 2008).

Perhaps the biggest mystery is how sequences such as the common ancestor of L/I/V-tRNA synthetase, which were translated via alternative or incomplete genetic codes, ended up being re-coded for translation by the direct ancestor of the canonical genetic code. Harmonization of genetic codes facilitated innovation sharing via HGT, making it advantageous to use the most common code, driving code convergence (Sella and Ardell 2006, Vetsigian, Woese et al. 2006). Only once a common code was established did HGT drop to levels such that a species tree became apparent, i.e. the LUCA coalescence point corresponds to convergence on a code (Vetsigian, Woese et al. 2006). Our identification of pre-LUCA sequences provides a rare source of data about early, alternative codes.

Materials and Methods

Pfam sequences

We downloaded genomes of 3562 prokaryotic species from NCBI that were present in the Web of Life (WoL): Reference phylogeny of microbes (Zhu, Mai et al. 2019) in August 2022. We classified them into five bacterial supergroups (FCB, PVC, CPR, Terrabacteria and Proteobacteria (Rinke, Schwientek et al. 2013, Brown, Hug et al. 2015)) and four archaeal supergroups (TACK, DPANN, Asgard and Euryarchaeota (Baker, De Anda et al. 2020, Shu and Huang 2021)). We included incomplete genomes, to enhance coverage of underrepresented supergroups.

Our analysis relies on protein domains instead of whole-gene orthologs. Proteins are often made of multiple protein domains, each of which might have originated at a different point in time. For the purpose of amino acid usage, what matters is the age of the protein domain, not that of the whole protein that it is part of. We used InterProScan (Jones, Binns et al. 2014) to identify instances of each Pfam domain (Mistry, Chuguransky et al. 2021) in our prokaryotic genomes. We excluded Pfams with fewer than 50 instances across all downloaded genomes. We also excluded 9 Pfams because they were marked "obsolete" starting July 2023. Among the remaining 8282 Pfams, 2496 Pfams had more than 1000 instances. We downsampled these to balance representation across the two taxonomic domains (archaea and bacteria). For instance, a Pfam with 2000 bacterial and 500 archaeal instances was downsampled by retaining all 500 archaeal sequences plus a subset (randomly sampled without replacement) of 500 bacterial sequences.

Not all Pfams are phylogenetically independent. The Pfam database includes annotations of "clans" of Pfams that share a common ancestor; for many analyses, we used clans rather than Pfams as our set of independent datapoints. We considered Pfams that were not annotated as part of a clan to be a single-entry clan, and assigned them a clan ID equal to their Pfam ID.

Calculating Pfam sequence properties

For each Pfam, we took the average amino acid frequencies across all instances (before downsampling species), weighted by their sequence length. For each clan, we took the average amino acid frequencies across Pfams within a clan, weighted by the median length of the Pfam. For each phylostratum (i.e., cohort of sequences born along the same phylogenetic branch), we took the average amino acid frequencies of the clans, weighted by the median length of each Pfam or clan. Median clan lengths were calculated across all Pfam instances within a clan. Ancient amino acid usage was calculated as a ratio of LUCA amino acid frequencies to ancient post-LUCA amino acid frequencies. Weighted linear model 1 regressions were estimated using the lm() function with the 'weights' argument in the 'stats' package in base R (R Core Team, 2021).

Hydrophobic clustering was calculated as a normalized index of dispersion for each Pfam instance (Irbäck et al., 1996). This involved assessing the ratio of the variance to the mean in the number of the most hydrophobic amino acids (leucine, isoleucine, valine, phenylalanine, methionine, and tryptophan) within consecutive blocks of six amino acids. The values of this index of dispersion were then normalized, to make them comparable across Pfams with different lengths and hydrophobicities. In cases where the Pfam length was not a multiple of 6, the average across all possible 6-amino acid frames was computed, trimming the ends as needed. For additional details, refer to Foy et al. (2019) or James et al. (2021). A clustering value of 1 corresponds to the hydrophobic nature of each site being independently sampled. Higher values indicate a tendency for hydrophobic residues to form clusters along the sequence, while lower values indicate a more interspersed distribution of hydrophobic residues than would be expected given independence. For

each Pfam, we took the average hydrophobic clustering across all its instances (prior to downsampling species).

Pfam trees

After aligning the downsampled Pfam sequences with MAFFT v.7 (Katoh and Standley 2013), we inferred preliminary trees for each Pfam using a time non-reversible amino acid substitution matrix trained on the Pfam database (NQ.PFAM) (Dang, Minh et al. 2022) implemented in the IQ-Tree software (Minh, Schmidt et al. 2020) with no rate heterogeneity among sites. Because most individual Pfam amino acid sequences are too short for reliable tree inference, we next reconciled these preliminary Pfam trees with a species tree using GeneRax (Morel, Kozlov et al. 2019). While there is no perfect species tree for prokaryotes, reconciliation even with a roughly approximate tree can still provide benefits. We used the WoL prokaryotic species tree (Zhu, Mai et al. 2019). We ran GeneRax twice. The first run used an LG amino acid substitution model, a gamma distribution with four discrete rate categories, and a Subtree Prune and Regraft (SPR) radius of 3, as recommended by the GeneRax developer. The second run used the output of reconciled trees from the first run as input, and used the Q.PFAM amino acid substitution model (Minh, Dang et al. 2021), which was trained on the Pfam dataset, instead of LG. We did not use NQ.PFAM, because time non-reversible models are only implemented in IQ-Tree (Dang, Minh et al. 2022) and not GeneRax. An SPR radius of 5 was set during the second run, allowing for a more exhaustive tree search. In both runs, the UndatedDTL probabilistic model was used to compute the reconciliation likelihood. The second run of GeneRax reduced the estimated transfer rates by 7% compared to the first run (Welch two sample t-test, $p = 10^{-12}$), indicating continued improvements to the phylogenies.

We re-estimated the branch lengths of the reconciled Pfam trees in IQ-Tree using the NQ.PFAM model with no rate heterogeneity and performed midpoint rooting using the phytools R package (Revell 2012) on these re-estimated branch lengths. As alternative rooting methods, we also explored and rejected minimum variance (Mai, Sayyari et al. 2017), minimal ancestral deviation (Tria, Landan et al. 2017), and rootstraps based on time non-reversible substitution models (Naser-Khdour, Quang Minh et al. 2022). The first two methods work best when deviations from the molecular clock average out on longer time-scales, which is not true for phylogenies, for instance, in which evolution at different temperatures causes sustained differences in evolutionary rate. Indeed, minimum variance failed to resolve the prokaryotic supergroups as separate clades, in visual inspection of PF00001, due to presumed genuine rate variation among taxa. The latter produced very low confidence roots. In contrast, midpoint rooting performed well on aaRSs once we implemented the procedure for outlier removal described under “Classifying Pfam domains into ancient phylostrata” below.

We then implemented a new --enforce-gene-tree-root option in GeneRax, and ran GeneRax in evaluation mode, with Q.PFAM+G as the substitution and rate heterogeneity models, respectively. Evaluation mode re-estimates the reconciliation likelihood and the duplication, transfer and loss (DTL) rates on a fixed tree, without initiating a tree search. Fifteen reconciled Pfam trees had inferred transfer rates higher than 0.6, three times the seed transfer rate implemented by GeneRax. We took this as a sign of poor tree quality, and annotated these 15 Pfams as of unclassifiable age.

Filtering out HGT between archaea and bacteria

Exclusion of likely products of horizontal gene transfer (HGT) between bacteria and archaea facilitates the classification of a Pfam into LUCA (Figure 1a). To achieve this, we divided sequences into "homogeneous groups", meaning the largest monophyletic group in the Pfam tree for which the corresponding species all belong to the same prokaryotic supergroup. Each homogeneous group was considered as a candidate for exclusion. We consider the "focal node" separating a homogeneous group from its sister lineage. To avoid over-pruning, we do not consider deep focal nodes that are 2 or fewer nodes away from the root.

We first require the focal node to be 'mixed', meaning its descendants are found within both Bacteria and Archaea, rather than 'unmixed', meaning its descendants, while spread across at least two supergroups, are found either within Bacteria or Archaea but not both. We next require the focal node to be labelled by GeneRax as most likely a transfer (T), rather than a duplication (D) or speciation (S). Finally, to identify homogeneous groups that are likely to be receivers rather than the donors of transferred Pfam sequences, we require the sister lineage to contain no sequences present in the same supergroup as that defining the homogeneous group in question. An example of filtering is shown in Figure 1b.

We ran the filtering process twice to address rare occasions of an intradomain HGT nested within another intradomain HGT group. In the second filter, we apply the third criterion after pruning the homogenous groups identified as HGT during the first filter run.

Classifying Pfam domains into ancient phylostrata

We re-rooted the HGT-pruned Pfam trees using the midpoint.root function in the 'phytools' R package (Revell 2012), before classifying them into phylostrata. Classification was based on the locations of the most recent common ancestors (MRCAs) of each supergroup. In the case of a LUCA Pfam, we require the root to separate the MRCAs of all bacterial supergroups from the MRCAs of all archaeal supergroups (Figure 1a).

If there were no horizontal transfer, and the tree of a Pfam present in one copy in LUCA were error-free, then the MRCAs for the nine supergroups would be two to four branches away from the root. This is true even if our Pfam tree and/or species tree do not correctly capture the true phylogenetic relationships among supergroups. However, we cannot ignore HGT; we have not filtered out the products of HGT between supergroups within Archaea or within Bacteria, only that of HGT between Archaea and Bacteria. HGT from a more derived supergroup to a more basal supergroup will move the inferred MRCA of the former further back in time. Given rampant HGT, whether real or erroneously implied by Pfam tree error, we required Pfams to have their supergroups' MRCA two branches away from the root (Figure 1a).

Phylogenies with three or more basal bacterial supergroups and two or more basal archaeal supergroups were classified as LUCA. In other words, we allow the absence of up to two supergroups per taxonomic domain, as compatible with ancestral presence followed by subsequent loss. Trees with three or more basal bacterial supergroups but fewer than two basal archaeal supergroups, as well as trees with two or more basal archaeal supergroups but fewer than three basal bacterial supergroups, were classified as ancient but post-LUCA. These are candidate Pfams for the Last Bacterial Common Ancestor (LBCA) and the Last Archaeal Common Ancestor (LACA) phylostrata, respectively, but the necessary HGT filtering for sufficient confidence in this classification is beyond the scope of the current work. If only one basal supergroup is present, then the Pfam is classified into the corresponding supergroup-specific phylostratum, meaning it emerged relatively recently (modern post-LUCA). If two basal bacterial supergroups (and no

archaeal supergroups) were present, the Pfam was classified as post-LBCA which was also considered modern post-LUCA (younger than LBCA but older than the supergroup-specific phylostrata). The remaining Pfams were considered unclassifiable.

We also classify into a pre-LUCA phylostratum the subset of LUCA-classified Pfams for which there is evidence that LUCA contained at least two copies that left distinct descendants. This is motivated by the assumption that LUCA domains that were born earlier are more likely to have had the chance to duplicate and diverge prior to the archaeal-bacterial split (Fournier and Alm 2015). We require that both the nodes that are only one branch from the root be classified as LUCA nodes. This means that each of these nodes should, after HGT filtering: i) split a pure-bacterial lineage from a pure-archaeal lineage, and ii) include as descendants at least three bacterial and two archaeal basal MRCAs no more than two nodes downstream of the potential LUCA nodes (Figure 1c).

Assignment of a Pfam to a phylostratum is sensitive to the root's position. Midpoint rooting is based on the longest distance between two extant sequences. A single sequence that is inaccurately placed within the Pfam tree can yield an abnormally long terminal branch, upon which the root is then based. Indeed, this phenomenon was readily apparent upon manual inspection of rooted Pfam trees. To ensure the robustness of our phylostratum classifications to the occasional misplaced sequence, we removed the Pfam instance with the longest root-to-tip branch length in each HGT-filtered tree as potentially faulty, re-calculated the midpoint root, and then re-classified each Pfam. We repeated this for ten iterations, then retained only those Pfams that were classified into the same phylostratum at least 7 out of 10 times. Our HGT filtering algorithm does not act on nodes near the root, making it robust to small differences in root position; we therefore did not repeat the HGT-filtering during these iterations.

We classified clans that contained at least two LUCA Pfams as pre-LUCA clans. Clans that contained both ancient archaeal and ancient bacterial post-LUCA Pfams (i.e. candidate LACA and LBCA Pfams) were classified as LUCA. Clans that contained at least two different archaeal but no bacterial supergroup-specific Pfams, or three different bacterial supergroup-specific Pfams but no archaeal supergroup-specific Pfams, were classified as ancient post-LUCA clans. Clans that meet neither of these criteria, and that contain at least one unclassified Pfam, were considered unclassifiable due to the possibility that the unclassified Pfam might be older than the classified Pfams present in the clan. All other clans were assigned the age of their oldest Pfam.

Data and Code Availability

Data files and R scripts (R Core Team, 2021) used to generate the results and figures are available at sawsanwehbi/Pfam-age-classification GitHub repository.

Author Contributions

S.W. and J.M. conceived the study, with input from D.S.L. S.W. conducted the analyses, with advice from J.M., A.W., B.Q.M., and B.M. B.M. implemented a new feature in GeneRax to assist the analyses. S.W. wrote the first draft of the manuscript, with subsequent editing from S.W., J.M., D.S.L., A.W., and B.Q.M.

Acknowledgements

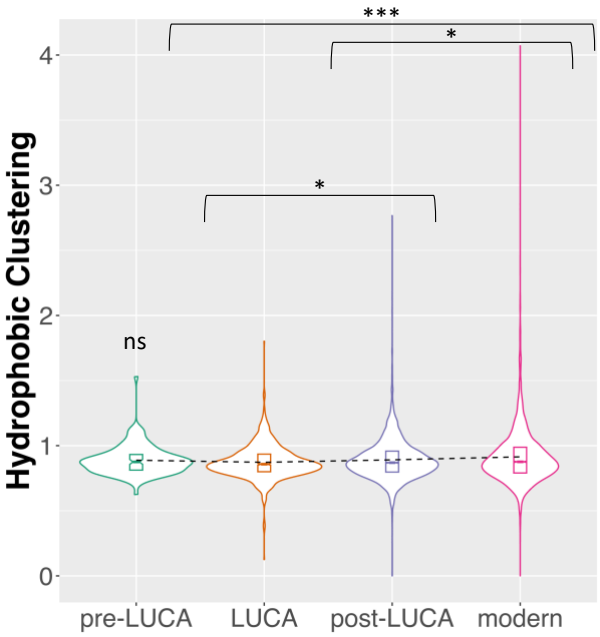
We thank NASA [80NSSC24K0384] and the John Templeton Foundation [62220] for funding SW and JM, Chan-Zuckerberg Initiative [EOSS4-0000000312] for funding BQM, the DFG [STA 860/6-2] for funding BM, and the NIH [T32GM132008] for funding AW. We thank Mike Barker, Alan Moses, and Elisa Tomat for helpful discussions, and Cat Wolner for comments on the manuscript.

Supplementary results: agreement with case studies

In agreement with (Pi, Lin et al. 2022)'s findings that nitrogen fixation evolved in bacteria, we annotate the nitrogenase component 1 type Oxidoreductase (PF00148), the main Pfam domain in NifN, NifD and NifK as a post-LUCA, LBCA candidate. We also classify the 4Fe-4S iron sulfur cluster binding Pfam domain within nitrogenase (PF00142) as unclassifiable but low-confidence candidate LBCA (the first five of our outlier removal iterations during midpoint rooting had it as unclassifiable, the next five iterations as LBCA).

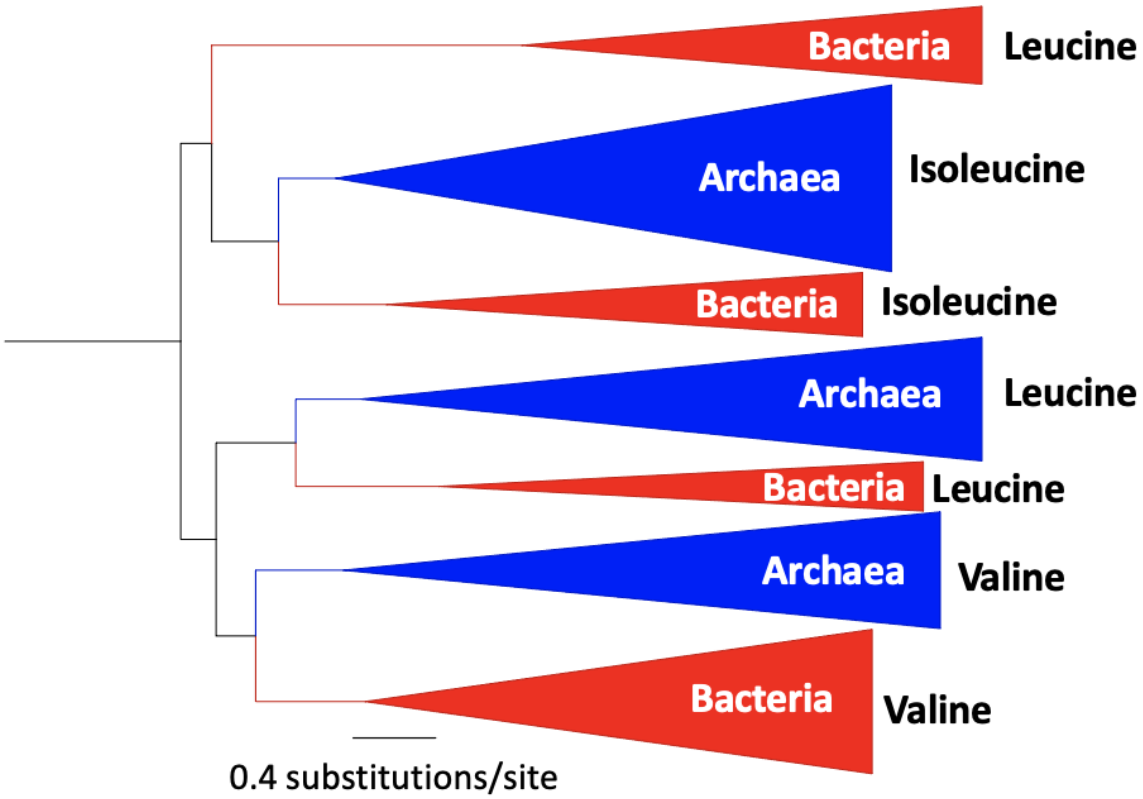
We also succeed in assigning to LUCA the complete set of amino acid-tRNA synthetase-associated anti-codon binding domains found in modern prokaryotes. Focusing on complete genes would have been problematic, because accessory domains (e.g. PF04073 and PF13603, which deacylate misacylated tRNA) were sometimes added later.

Supplementary files and figures



Supplementary Figure 1. Older protein clans have more interspersed hydrophobic amino acids. Hydrophobic clustering scores (see Methods; independent amino acid placement corresponds to a score of 1, values less than 1 indicate interspersed hydrophobic amino acids, values greater than 1 indicate clustered hydrophobic amino acids) are lower in LUCA Pfams than in either post-LUCA Pfams ($p = 0.02$) or modern Pfams ($p = 0.0002$). The latter two also differ from each other ($p = 0.02$). Pairwise Wilcoxon rank sum tests were used to test for statistical significance. Boxplots show the median, upper, and lower quartiles of the data. Violin plots were generated using the geom_violin function in the ggplot R package (Wickham 2016). Small notches around the medians define the 95% confidence intervals for the medians, and dashed lines join means. The hydrophobic clustering of pre-LUCA clans is not significantly different from any of the other phylostrata, which may be due to being fewer in number than other phylostrata. We note that unlike animal proteins analyzed in James et al. (2021), even the youngest, modern clans have

clustering significantly below the expected value of 1 for independent sites (Binomial test; $p < 10^{-16}$).



Supplementary Figure 2. Topology of the leucine, valine, and isoleucine tRNA synthetase core catalytic domain confirms pre-LUCA origin. Three archaeal-bacterial splits denoting LUCA are shown, representing leucine, valine and isoleucine specific tRNA synthetase homologs. Collapsed archaeal clades names are in blue and collapsed bacterial clades are in red. Branch lengths were estimated using the Pfam-trained time non-reversible amino acid substitution model NQ.pfam (Dang, Minh et al. 2022), with an R10 heterogeneity model. Tree depicted is HGT-filtered, and midpoint rooted following pruning of the first ten branch length outliers.

Amino Acid	LUCA clan usage	LUCA clan usage standard error	Pre-LUCA clan usage	Pre-LUCA clan usage standard error
A	1.062	0.018	1.027	0.02
C	1.102	0.06	1.024	0.056
D	0.98	0.021	0.969	0.033
E	0.981	0.025	0.905	0.032
F	0.92	0.024	1.025	0.032
G	1.106	0.022	1.082	0.019
H	1.026	0.028	1.095	0.036
I	1.042	0.019	1.029	0.023
K	0.949	0.027	0.904	0.029
L	0.982	0.017	1.004	0.025
M	1.064	0.025	1.091	0.031
N	0.938	0.018	0.974	0.025
P	0.997	0.021	1.051	0.018
Q	0.884	0.02	0.909	0.022
R	0.955	0.021	0.922	0.024
S	0.98	0.011	1.005	0.011
T	1.03	0.012	1.017	0.013
V	1.087	0.015	1.04	0.017
W	0.788	0.037	0.945	0.045
Y	0.899	0.02	0.99	0.021

Supplementary Table 1. LUCA and pre-LUCA clan amino acid frequencies are divided by post-LUCA clan amino acid frequencies to produce measures of relative usage. Standard errors of each of these ratios L/P use an approximation derived from a Taylor expansion of the ratio: $\sqrt{\frac{\sigma_L{}^2}{P^2} + \frac{L^2\sigma_P{}^2}{P^4}}$ (Seltman 2012). For each of the 20 amino acid frequencies, the standard errors σ_L and σ_P of the weighted means across all the clans within the LUCA and pre-LUCA phylostrata (weighted by the length of the median clan within each phylostratum) were calculated using the weighted_se() function in the diagis R package (Helske 2023).

Supplementary Table 2. Pfams with ages, amino acid frequencies, median length, clustering score are found within github.

Supplementary Table 3. Clans with ages, amino acid frequencies, median length, clustering score are found within github.

References

Alvarez-Carreño, C., A. Becerra and A. Lazcano (2013). "Norvaline and Norleucine May Have Been More Abundant Protein Components during Early Stages of Cell Evolution." Origins of Life and Evolution of Biospheres **43**(4-5): 363-375.

Baker, B. J., V. De Anda, K. W. Seitz, N. Dombrowski, A. E. Santoro and K. G. Lloyd (2020). "Diversity, ecology and evolution of Archaea." Nat Microbiol **5**(7): 887-900.

Berkemer, S. J. and S. E. McGlynn (2020). "A New Analysis of Archaea–Bacteria Domain Separation: Variable Phylogenetic Distance and the Tempo of Early Evolution." Molecular Biology and Evolution **37**(8): 2332-2340.

Brooks, D. J., J. R. Fresco, A. M. Lesk and M. Singh (2002). "Evolution of Amino Acid Frequencies in Proteins Over Deep Time: Inferred Order of Introduction of Amino Acids into the Genetic Code." Molecular Biology and Evolution **19**(10): 1645-1655.

Brown, C. T., L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams and J. F. Banfield (2015). "Unusual biology across a group comprising more than 15% of domain Bacteria." Nature **523**(7559): 208-211.

Chen, B., S. Luo, S. Zhang, Y. Ju, Q. Gu, J. Xu, X.-L. Yang and H. Zhou (2021) "Inhibitory mechanism of reveromycin A at the tRNA binding site of a class I synthetase." Nature communications **12**, 1616 DOI: 10.1038/s41467-021-21902-0.

Clausen, T., R. Huber, L. Prade, M. C. Wahl and A. Messerschmidt (1998). "Crystal structure of Escherichia coli cystathionine γ-synthase at 1.5 Å resolution." The EMBO Journal **17**(23): 6827-6838.

Crapitto, A. J., A. Campbell, A. Harris and A. D. Goldman (2022). "A consensus view of the proteome of the last universal common ancestor." Ecology and Evolution **12**(6).

Dang, C. C., B. Q. Minh, H. McShea, J. Masel, J. E. James, L. S. Vinh and R. Lanfear (2022). "nQMaker: Estimating Time Nonreversible Amino Acid Substitution Models." Syst Biol **71**(5): 1110-1123.

Dong, X., M. Zhou, C. Zhong, B. Yang, N. Shen and J. Ding (2010). "Crystal structure of Pyrococcus horikoshii tryptophanyl-tRNA synthetase and structure-based phylogenetic analysis suggest an archaeal origin of tryptophanyl-tRNA synthetase." Nucleic Acids Res **38**(4): 1401-1412.

Doolittle, W. F. (2000). "The nature of the universal ancestor and the evolution of the proteome." Current Opinion in Structural Biology **10**(3): 355-358.

Douglas, J., R. Bouckaert, C. W. Carter and R. Wills, Peter (2024). "Enzymic recognition of amino acids drove the evolution of primordial genetic codes." Nucleic Acids Research **52**(2): 558-571.

Eriani, G., M. Delarue, O. Poch, J. Gangloff and D. Moras (1990). "Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs." Nature **347**(6289): 203-206.

Fairchild, J., S. Islam, J. Singh, D.-K. Bučar and M. W. Powner (2024). "Prebiotically plausible chemoselective pantetheine synthesis in water." Science **383**(6685): 911-918.

Foden, C. S., S. Islam, C. Fernández-García, L. Maugeri, T. D. Sheppard and M. W. Powner (2020). "Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water." Science **370**(6518): 865-869.

Fournier, G. P. and E. J. Alm (2015). "Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code." J Mol Evol **80**(3-4): 171-185.
Fournier, G. P., C. P. Andam, E. J. Alm and J. P. Gogarten (2011). "Molecular Evolution of Aminoacyl tRNA Synthetase Proteins in the Early History of Life." Origins of Life and Evolution of Biospheres **41**(6): 621-632.
Fournier, G. P. and J. P. Gogarten (2007). "Signature of a Primitive Genetic Code in Ancient Protein Lineages." Journal of Molecular Evolution **65**(4): 425-436.
Fried, S. D., K. Fujishima, M. Makarov, I. Cherepashuk and K. Hlouchova (2022). "Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids." Journal of The Royal Society Interface **19**(187).
Froese, T., J. I. Campos, K. Fujishima, D. Kiga and N. Virgo (2018). "Horizontal transfer of code fragments between protocells can explain the origins of the genetic code without vertical descent." Scientific Reports **8**(1).
Fukai, S., O. Nureki, S.-I. Sekine, A. Shimada, J. Tao, D. G. Vassylyev and S. Yokoyama (2000). "Structural Basis for Double-Sieve Discrimination of L-Valine from L-Isoleucine and L-Threonine by the Complex of tRNAVal and Valyl-tRNA Synthetase." Cell **103**(5): 793-803.
Fukuchi, S., K. Yoshimune, M. Wakayama, M. Moriguchi and K. Nishikawa (2003). "Unique amino acid composition of proteins in halophilic bacteria." J Mol Biol **327**(2): 347-357.
Gogarten, J. P. and D. Deamer (2016). "Is LUCA a thermophilic progenote?" Nat Microbiol **1**: 16229.
Goldman, A. D. and B. Kacar (2021). "Cofactors are Remnants of Life's Origin and Early Evolution." Journal of Molecular Evolution **89**(3): 127-133.
Granold, M., P. Hajieva, M. I. Toşa, F.-D. Irimie and B. Moosmann (2018). "Modern diversification of the amino acid repertoire driven by oxygen." Proceedings of the National Academy of Sciences **115**(1): 41-46.
Helske, J. (2023). "diagis: Diagnostic Plot and Multivariate Summary Statistics of Weighted Samples from Importance Sampling."
James, J. E., S. M. Willis, P. G. Nelson, C. Weibel, L. J. Kosinski and J. Masel (2021). "Universal and taxon-specific trends in protein sequences as a function of age." Elife **10**.
Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez and S. Hunter (2014). "InterProScan 5: genome-scale protein function classification." Bioinformatics **30**(9): 1236-1240.
Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Mol Biol Evol **30**(4): 772-780.
Keese, P. K. and A. Gibbs (1992). "Origins of genes: "big bang" or continuous creation?" Proceedings of the National Academy of Sciences **89**(20): 9489-9493.
Koonin, E. V. and A. S. Novozhilov (2009). "Origin and evolution of the genetic code: The universal enigma." IUBMB Life **61**(2): 99-111.
Lamour, V., S. Quevillon, S. Diriong, V. C. N'Guyen, M. Lipinski and M. Mirande (1994). "Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer." Proceedings of the National Academy of Sciences **91**(18): 8670-8674.

Lapointe, J., L. Duplain and M. Proulx (1986). "A single glutamyl-tRNA synthetase aminoacylates tRNAGlu and tRNAGln in Bacillus subtilis and efficiently misacylates Escherichia coli tRNAGln1 in vitro." Journal of Bacteriology **165**(1): 88-93.

Lauretta, D. S., C. D. Adam, A. J. Allen, R.-L. Ballouz, O. S. Barnouin, K. J. Becker, T. Becker, C. A. Bennett, E. B. Bierhaus, B. J. Bos, R. D. Burns, H. Campins, Y. Cho, P. R. Christensen, E. C. A. Church, B. E. Clark, H. C. Connolly, M. G. Daly, D. N. Dellagiustina, C. Y. Drouet D'Aubigny, J. P. Emery, H. L. Enos, S. F. Kasper, J. B. Garvin, K. Getzandanner, D. R. Golish, V. E. Hamilton, C. W. Hergenrother, H. H. Kaplan, L. P. Keller, E. J. Lessac-Chenen, A. J. Liounis, H. Ma, L. K. Mccarthy, B. D. Miller, M. C. Moreau, T. Morota, D. S. Nelson, J. O. Nolau, R. Olds, M. Pajola, J. Y. Pelgrift, A. T. Polit, M. A. Ravine, D. C. Reuter, B. Rizk, B. Rozitis, A. J. Ryan, E. M. Sahr, N. Sakatani, J. A. Seabrook, S. H. Selznick, M. A. Skeen, A. A. Simon, S. Sugita, K. J. Walsh, M. M. Westermann, C. W. V. Wolner and K. Yumoto (2022). "Spacecraft sample collection and subsurface excavation of asteroid (101955) Bennu." Science **377**(6603): 285-291.

Lauretta, D. S. C., Harold C. Jr; Grossman, Jeffrey N. ; Polit, Anjani T. ; the OSIRIS-REx Sample Analysis Team (2023). "OSIRIS-REx Sample Analysis Plan -- Revision 3.0."

Levine, R. L., L. Mosoni, B. S. Berlett and E. R. Stadtman (1996). "Methionine residues as endogenous antioxidants in proteins." Proceedings of the National Academy of Sciences **93**(26): 15036-15040.

Li, J., X. He, S. Gao, Y. Liang, Z. Qi, Q. Xi, Y. Zuo and Y. Xing (2023). "The Metal-binding Protein Atlas (MbPA): An Integrated Database for Curating Metalloproteins in All Aspects." J Mol Biol **435**(14): 168117.

Lim, J. M., G. Kim and R. L. Levine (2019). "Methionine in Proteins: It's Not Just for Protein Initiation Anymore." Neurochem Res **44**(1): 247-257.

Mai, U., E. Sayyari and S. Mirarab (2017). "Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction." PLOS ONE **12**(8): e0182238.

Mascarenhas, A. P., S. An, A. E. Rosen, S. A. Martinis and K. Musier-Forsyth (2008). Fidelity Mechanisms of the Aminoacyl-tRNA Synthetases. Protein Engineering, Springer Berlin Heidelberg**:** 155-203.

Miller, S. L. (1953). "A Production of Amino Acids under Possible Primitive Earth Conditions."

Minh, B. Q., C. C. Dang, L. S. Vinh and R. Lanfear (2021). "QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution." Syst Biol **70**(5): 1046-1060.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler and R. Lanfear (2020). "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." Mol Biol Evol **37**(5): 1530-1534.

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, A. Salazar, Gustavo, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn and A. Bateman (2021). "Pfam: The protein families database in 2021." Nucleic Acids Research **49**(D1): D412-D419.

Moosmann, B. (2021). "Redox Biochemistry of the Genetic Code." Trends in Biochemical Sciences **46**(2): 83-86.

Morel, B., A. M. Kozlov, A. Stamatakis and G. J. Szöllősi (2019). GeneRax: A tool for species tree-aware maximum likelihood based gene family tree inference under gene duplication, transfer, and loss, Cold Spring Harbor Laboratory.

Morgens, D. W. and A. R. O. Cavalcanti (2013). "An Alternative Look at Code Evolution: Using Non-canonical Codes to Evaluate Adaptive and Historic Models for the Origin of the Genetic Code." Journal of Molecular Evolution **76**(1-2): 71-80.

Naser-Khdour, S., B. Quang Minh and R. Lanfear (2022). "Assessing Confidence in Root Placement on Phylogenies: An Empirical Study Using Nonreversible Models for Mammals." Systematic Biology **71**(4): 959-972.

Neubeck, A. and F. Freund (2020). "Sulfur Chemistry May Have Paved the Way for Evolution of Antioxidants." Astrobiology **20**(5): 670-675.

Nitschke, W., S. E. McGlynn, E. J. Milner-White and M. J. Russell (2013). "On the antiquity of metalloenzymes and their substrates in bioenergetics." Biochimica et Biophysica Acta (BBA) - Bioenergetics **1827**(8): 871-881.

Parker, E. T., H. J. Cleaves, J. P. Dworkin, D. P. Glavin, M. Callahan, A. Aubrey, A. Lazcano and J. L. Bada (2011). "Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment." Proceedings of the National Academy of Sciences **108**(14): 5526-5531.

Pi, H. W., J. J. Lin, C. A. Chen, P. H. Wang, Y. R. Chiang, C. C. Huang, C. C. Young and W. H. Li (2022). "Origin and Evolution of Nitrogen Fixation in Prokaryotes." Mol Biol Evol **39**(9).

Ranjan, S., Z. R. Todd, J. D. Sutherland and D. D. Sasselov (2018). "Sulfidic Anion Concentrations on Early Earth for Surficial Origins-of-Life Chemistry." Astrobiology **18**(8): 1023-1040.

Revell, L. J. (2012). "phytools: an R package for phylogenetic comparative biology (and other things)." Methods in Ecology and Evolution **3**(2): 217-223.

Ribeiro, A. J. M., J. D. Tyzack, N. Borkakoti, G. L. Holliday and J. M. Thornton (2020). "A global analysis of function and conservation of catalytic residues in enzymes." Journal of Biological Chemistry **295**(2): 314-324.

Rinke, C., P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz and T. Woyke (2013). "Insights into the phylogeny and coding potential of microbial dark matter." Nature **499**(7459): 431-437.

Sella, G. and D. H. Ardell (2006). "The Coevolution of Genes and Genetic Codes: Crick's Frozen Accident Revisited." Journal of Molecular Evolution **63**(3): 297-313.

Seltman, H. (2012). "Approximations for mean and variance of a ratio." unpublished note.

Shen, C., L. Yang, S. L. Miller and J. Oró (1990). "Prebiotic synthesis of histidine." Journal of Molecular Evolution **31**(3): 167-174.

Shu, W.-S. and L.-N. Huang (2021). "Microbial diversity in extreme environments." Nature Reviews Microbiology **20**(4): 219-235.

Tang, Y. and D. A. Tirrell (2002). "Attenuation of the editing activity of the Escherichia coli leucyl-tRNA synthetase allows incorporation of novel amino acids into proteins in vivo." Biochemistry **41**(34): 10635-10645.

Therneau, T. (2022). "Deming,Theil-Sen,Passing-BablockandTotalLeastSquares Regression."

Tria, F. D. K., G. Landan and T. Dagan (2017). "Phylogenetic rooting using minimal ancestor deviation." Nature Ecology & Evolution **1**(7).

Trifonov, E. N. (2000). "Consensus temporal order of amino acids and evolution of the triplet code." Gene **261**(1): 139-151.

Van Oss, S. B. and A.-R. Carvunis (2019). "De novo gene birth." PLOS Genetics **15**(5): e1008160.

Vetsigian, K., C. Woese and N. Goldenfeld (2006). "Collective evolution and the genetic code." Proceedings of the National Academy of Sciences **103**(28): 10696-10701.

Vieira-Silva, S. and E. P. C. Rocha (2008). "An Assessment of the Impacts of Molecular Oxygen on the Evolution of Proteomes." Molecular Biology and Evolution **25**(9): 1931-1942.

Vázquez-Salazar, A., A. Becerra and A. Lazcano (2018). "Evolutionary convergence in the biosyntheses of the imidazole moieties of histidine and purines." PLOS ONE **13**(4): e0196349.

Weiss, M. C., F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi and W. F. Martin (2016). "The physiology and habitat of the last universal common ancestor." Nat Microbiol **1**(9): 16116.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York.

Woese, C. R., G. J. Olsen, M. Ibba and D. SöLl (2000). "Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process." Microbiology and Molecular Biology Reviews **64**(1): 202-236.

Zhu, Q., U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J. G. Sanders, P. Belda-Ferre, G. A. Al-Ghalith, E. Kopylova, D. Mcdonald, T. Kosciolek, J. B. Yin, S. Huang, N. Salam, J.-Y. Jiao, Z. Wu, Z. Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J. T. Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S. Mirarab and R. Knight (2019). "Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea." Nature Communications **10**(1).

Závodszky, A. S. a. P. (2000). "Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits- results of a comprehensive survey." Structure **8**(5): 493-504.