Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code

Citation: Fournier, G. P., and E. J. Alm. "Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code." Journal of Molecular Evolution 80.3–4 (2015): 171–185.

As Published: http://dx.doi.org/10.1007/s00239-015-9672-1

Publisher: Springer US

Persistent URL: http://hdl.handle.net/1721.1/104053

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Title:

Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code.

Fournier GP[1]*, Alm EJ[2]

[1]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.
[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

**corresponding author: Gregory P. Fournier*
email: g4nier@mit.edu
telephone: (617) 324-6164

Abstract

The genetic code was likely complete in its current form by the time of the Last Universal Common Ancestor (LUCA). Several scenarios have been proposed for explaining the code's pre-LUCA emergence and expansion, and the relative order of the appearance of amino acids used in translation. One co-evolutionary model of genetic code expansion proposes that at least some amino acids were added to the code by the ancient divergence of aminoacyl-tRNA synthetase (aaRS) families. Of all the amino acids used within the genetic code, Trp is most frequently claimed as a relatively recent addition. We observe that, since TrpRS and TyrRS are paralogous protein families retaining significant sequence similarity, the inferred sequence composition of their ancestor can be used to evaluate this co-evolutionary model of genetic code expansion.

We show that ancestral sequence reconstructions of the pre-LUCA paralog ancestor of TyrRS and TrpRS have several sites containing Tyr, yet a complete absence of sites containing Trp. This is consistent with the paralog ancestor being specific for the utilization of Tyr, with Trp being a subsequent addition to the genetic code facilitated by a process of aaRS divergence and neofunctionalization. Only after this divergence could Trp be specifically encoded and incorporated into proteins, including the TyrRS and TrpRS descendant lineages themselves. This early absence of Trp is observed under both homogeneous and nonhomogeneous models of ancestral sequence reconstruction. Simulations support that this observed absence of Trp is unlikely to be due to chance or model bias. These results support that the final stages of genetic code evolution occurred well within the "protein world", and that the presence-absence of Trp within conserved sites of ancient protein domains is a likely measure of their relative antiquity, permitting the relative timing of extremely early events within protein evolution before LUCA.

Keywords:

ancestral sequence reconstruction
genetic code
tryptophanyl-tRNA synthetase
tyrosyl-tRNA sythetase
tryptophan

Introduction

Evolution of the genetic code and amino acid alphabet

The mechanism and history of the evolution of the genetic code is one of the most important questions surrounding the origin of life on Earth. It has been the subject of numerous investigations in the last several decades, with a wide variety of approaches using different lines of biochemical, evolutionary, schematic, and mathematical evidence (Hartman and Smith 2014; Higgs and Pudritz 2009; Vetsigian et al. 2006; Wong 1988{Cavalcanti, 2004 #63)reviews in (Koonin and Novozhilov 2009; Trifonov 2000). Along these lines, previous work has also shown that ancestral reconstructions of universally conserved ribosomal proteins contain a compositional bias that likely provides evidence of earlier stages in the evolution of the genetic code (Brooks and Fresco 2002; Brooks et al. 2002; Fournier and Gogarten 2007; Fournier and Gogarten 2010). Specifically, compositional analyses suggest that Gly, Ala, Asp, Asn, and Thr are among the most ancient amino acids in the code, while Glu, Gln, Phe, Tyr, Trp, Cys, and Ser are later additions. The results of this empirical sequence-based method are similar to those of consensus meta-studies (Trifonov 2000), classic prebiotic synthesis experiments (Miller 1953), and more recent selection-based models of the organization and evolution of the codon table schema (Higgs 2009).

While the ribosome is the RNA/protein complex responsible for mediating translation and elongation of the nascent peptide chain via codon/anticodon pairing and peptidyltransferase activity, many additional proteins are required for defining the genetic code and performing protein synthesis in living cells. Three major groups of these proteins could potentially have played a role in the development of the genetic code via their early evolution and diversification: (1) aminoacyl-tRNA synthetase (aaRS) proteins, which load specific tRNA with their cognate amino acids; (2) amino acid biosynthesis enzymes; and (3) tRNA modification proteins, essential for the partitioning of the genetic code via codon-anticodon interactions that allow for its current complex organization. Within each of these groups, many protein families with similar functions are paralogs, suggesting a common ancestor with a similar function. These relationships are especially apparent between several sets of aaRS proteins (TyrRS/TrpRS, IleRS/ValRS, GluRS/GlnRS, AspRS/AsnRS, CysRS/MetRS) (Brown and Doolittle 1995; Landes et al. 1995; Nagel and Doolittle 1995). A similar scenario is found in many amino acid biosynthetic pathways, such as the distinct, paralogous sets of enzymes involved in Lys, Arg, and Leu biosynthesis (Fondi et al. 2007). The situation is more complex for tRNA modification enzymes, which are often non-homologous across major domains of life (Grosjean et al. 2010).

To at least some extent, the evolution of the genetic code was likely dependent on the duplication and divergence of these groups of proteins, and the expansion of their functional roles. Evolution of novel aaRS and amino acid biosynthesis proteins could potentially permit new amino acids to be added to the code and used in polypeptide synthesis (Cavalcanti et al. 2004; Di Giulio 1992; Klipcan and Safro 2004; Wetzel 1978). Similarly, the partitioning of the genetic code into blocks of synonymous codons via tRNA codon-anticodon recognition is made possible by the evolution of tRNA modification proteins. However, even if the genetic code was shaped by protein evolution, its earliest origins must predate polypeptide synthesis as we know it, possibly arising within an RNA-based system consisting of ribozymes with diverse catalytic activities (Hartlein and Cusack 1995; Wetzel 1995).

Ancestral reconstruction of aaRS protein family paralog ancestors

Among these protein families, aaRS are unique in that their specific functions directly presuppose the specific encoding of certain amino acids, of which they, themselves, are composed. The ancestors of aaRS families could also only have been composed of amino acids specified within

the code at the time. Therefore, the identities of the amino acids found within the reconstructed sequences of aaRS family paralog ancestors can be used to constrain their ancestral functions, extending the tools of molecular evolution to a time before LUCA. Three possible functional histories are associated with the duplication and divergence of paralogous proteins from a common ancestor: *neofunctionalization*, with the addition of a novel function in one or the other paralog; *subfunctionalization*, with specialization of each paralog from a nondiscriminating ancestor; and *parafunctionalization*, with one or both descendants taking over pre-existing functions following their divergence. If function in aaRS families is defined as amino acid specificity, the genetic code itself will also have been changed in the cases of neofunctionalization or subfunctionalization, with amino acid sequence space co-evolving as enzymatic diversity and/or specificity increases.

Previous work has already applied this principle to the paralog ancestor of IleRS and ValRS (Fournier et al., 2011). IleRS and ValRS show a high level of sequence and structural similarity, recognize very similar amino acids that are frequently substituted for one another, and also use a similar set of codons. Furthermore, ValRS and IleRS are relatively abundant amino acids, providing many sites for analysis within the reconstruction. Interestingly, the probabilistic reconstruction of the IleRS/ValRS ancestor shows many sites with a high probability of being specific for Ile and Val, respectively. This result supports parafunctionalization, with specific coding for Val and Ile predating their cognate synthetases, possibly indicating an alternative aminoacylation or proofreading system in operation at this early time. It has been previously proposed that, as appears to be the case for Ile and Val, the genetic code was fixed by the time of the evolution of cognate aaRS families, and was mediated by an RNA-based system (Hartlein and Cusack 1995).

In this work, we extend and further develop this compositional reconstruction analysis to investigate the paralog ancestor of TyrRS and TrpRS. These aromatic amino acids are generally considered to be more recent additions to the code, by virtue of their complex biosynthetic pathways, and being encoded within the "stop" codon block, among other criteria (Trifonov 2000). In particular, the temporal ordering of Trp as one of the most recent additions to the code is one of the most longstanding and robust observations across independent lines of evidence (Jukes 1973; Jukes 1981; Osawa et al. 1992). As such, reconstructing the TyrRS/TrpRS paralog ancestor permits the investigation of a time before LUCA, recent enough that the genetic code was largely established, but still early enough that Trp and/or Tyr may not have yet existed as part of the canonical genetic code.

The presence or absence of Trp and/or Tyr residues at sites within the reconstructed paralog ancestor permits the discrimination of the previously described three possible functional histories (Figure 1). In the case of subfunctionalization, an earlier genetic code is probabilistic for some amino acids, incorporating similar amino acids at certain positions without discrimination. As selection would be unable to act to "fix" any individual AA within this set, proteins would simply evolve with this constraint, with functions being tolerant to the presence of Tyr or Trp at specific positions. One likely cause of such promiscuity could be nonspecific aminoacylation of tRNA. However, following aaRS duplication and divergence, "symmetry breaking" could occur, with selection for subfunctionalization driven by increases in fitness gained by fixing either Trp or Tyr at previously nondiscriminated positions. Similarly, codon space could also become subdivided via the partitioning of cognate tRNAs between each paralog. The fitness impact of such a transitional "partitioning phase" has been found to be favorable in specific instances (Higgs 2009). In such a case, it is expected that no sites with a high probability for being specifically Tyr or Trp would be observed within the reconstructed paralog ancestor, although sites with high nonspecific probabilities of being either Tyr or Trp would be expected to be observed. For example, sites

may be observed with a 45% probability of Trp, a 45% probability of Tyr, and only a 10% probability of being another amino acid. This suggests a tolerance to interchangeability between the two amino acids across protein sequences. If a large number of such ambiguous sites are observed, significantly more than are expected given amino acid substitution models within protein sequences, this likely indicates an inherited tolerance arising from ancestral ambiguity, and evidence of a non-specific genetic code at the ancestral node of the reconstruction.

In the case of neofunctionalization, an ancestral duplication and subsequent divergence of TyrRS and TrpRS lineage ancestors would permit addition of a new amino acid to the coding schema. This would be inferred by the absence of any sites reconstructing for either Tyr or Trp, respectively, within the paralog ancestor. An absence of Tyr and a presence of Trp would imply that TrpRS was the ancestral function, and the code lacked Tyr before the divergence occurred. Conversely, an absence of Trp and a presence of Tyr would suggest the code lacked Trp at this time.

As was the case with IleRS and ValRS, parafunctionaliztion would be inferred if specific sites individually reconstructing for Tyr and Trp are observed within the paralog ancestor, suggesting that specific encoding of Tyr and Trp predates their cognate aaRS, and must have been mediated by another, more ancient system. As Trp and Tyr are generally considered to be some of the most recent additions to the code, younger than than Ile or Val, this would suggest that a different and currently unknown aminoacylation regime persisted through the entire development of the canonical genetic code, independent of the evolution of the aaRS families.

Reliability of ancient protein ancestral reconstructions

To our knowledge, the most ancient protein attempted to be resurrected via ancestral sequence reconstruction is the bacterial ancestor of elongation factor Tu (EF-Tu), which is estimated to have existed >3 Gya (Gaucher et al. 2003). This resurrected protein ancestor was shown to have GDP binding activity comparable to its contemporary homologs. Furthermore, this activity was shown to be optimal at a temperature of ~65 °C, consistent with the range of ocean temperatures predicted on the early Earth based on interpretations of $\delta^{18}O$ isotopic ratios in Archaean rocks (Knauth 2005). These experimental results demonstrate the capacity for ancestral reconstruction methods to successfully and accurately recover the biological functionality of very ancient proteins, presumably via their fidelity in reconstructing correct ancestral sequences. While the pre-LUCA paralog ancestors of aaRS proteins may predate the bacterial ancestor by hundreds of millions of years, the similarly high sequence, structure, and functional conservation observed for EF and aaRS proteins suggests that the latter are similarly amenable to accurate reconstruction over these timescales.

Results

Phylogenetic Reconstruction

Maximum-likelihood phylogenetic reconstruction of TyrRS and TrpRS paralogs shows the monophyly of each group, and a clear division between bacterial and archaeal/eukaryal homologs in each case (Figure 2). While some previous investigations have suggested a paraphyletic relationship between TyrRS and TrpRS (Dong et al. 2010; Ribas de Pouplana et al. 1996), our result is consistent with other analyses that include sequence data from all three Domains, supporting the monophyly and pre-LUCA origin of each family (Brown et al. 1997;

Chandrasekaran et al. 2013). In further agreement with these analyses, we also find clear evidence of HGT within each protein family, especially within archaeal groups. Within TyrRS, there is a complex pattern of transfer between Crenarchaeota and other members of the TACK superphylum, and euryarchaeal clades including Thermoplasmatales, Thermococcales, and *Nanoarchaeum equitans*, which apparently acquired TyrRS from its symbiotic host, *Ignicoccus hospitalis* (or vice versa) (Podar et al. 2008). The ancestor of TyrRS within *I. hospitalis* and *N. equitans* itself appears to have also been transferred from Thermoproteales, or possibly vertically inherited within *N. equitans* as a sister to Thermococcales (Brochier et al. 2005), with independent transfers to Thermoproteales and *I. hospitalis*. A subset of Halobacteriales also appear to have received TyrRS from a deep TACK-associated lineage. Most notably, Eukarya are also polyphyletic within the TyrRS tree, with Animalia and Fungi (Opisthokonta) grouping with the TACK-associated halobacterial subset, while other eukaryotes form a monophyletic group within the TACK clade. This is consistent with previous work showing a horizontal gene transfer of the gene encoding TyrRS from Halobacteria to the opisthokont ancestor (Huang et al. 2005). A deep division is also apparent within bacterial TyrRS forms, as has been attributed to complex patterns of biased HGT (Andam et al. 2010). Several HGT events are also apparent within the TrpRS tree, including a transfer from within group II methanogens to Desulfurococcales, crenarchaeal transfers to *N. equitans* and Thaumarchaeota, and a deep TACK transfer to a subset of Halobacteriales, similar to that observed within TyrRS. For TrpRS, Eukarya form a monophyletic group, rooting deeply within the archaeal Domain closer to the TACK group. Additional HGT events within more recently diverging groups may also have occurred, but are less readily apparent and were not investigated.

The deep division between bacterial and archaeal variants of each paralog support, in both cases, the node ancestor of these domain lineages being congruent with LUCA. This is further supported by the reciprocal rooting of paralog sub-trees on congruent branches, showing a rooting of the ToL on the bacterial branch. As such, these aaRS paralog lineages diverged pre-LUCA, and a reconstruction of the likely sequence of the paralog ancestor provides information about a very early time in the history of life, possibly even during the later stages of the evolution of the genetic code itself.

Partial HGT of TyrRS within Eukarya and Halobacteriales

Phylogenetic trees of the TyrRS protein family show eukaryal and halobacterial TyrRS proteins as polyphyletic, with opisthokont orthologs grouping deeply on the crenarchaeal branch together with a subset of Halobacteriales, and all other eukaryal orthologs grouping within Crenarchaeota. The remaining Halobacteriales group together with group II methanogens. However, this HGT to Opisthokonta does not seem to be consistently evident across the full TyrRS protein sequence. Three regions were identified containing conserved amino acids supporting different bipartions for the placement of Eukarya and Halobacteriales. The first region (R1) consists of 20 amino acid sites. This region generates a phylogenetic tree that supports the monophyly and vertical placement of Halobacteriales within group II methanogens, as well as the monophyly of Eukarya, including Opisthokonta. The second region, located 100 AA sites downstream of the first, contains two sub-regions. The sub-region (R2) contains 11 sites, again supporting the monophyly of Halobacteriales, but with Opisthokonta grouping deeply within Crenarchaeota, congruent with its position in the gene tree, and not monophyletic with other Eukarya. Immediately downstream is a second 14 AA sub-region (R3) with sites supporting another distinct topology, where Halobacteriales is once again monophyletic within group II methanogens. However, in this region, Opisthokonta groups together with all Halobacteriales within the methanogens.

While these regions are small and do not contain many sites for phylogenetic inference, it is conspicuous that, in both cases, R1 and R2/R3 are at sites of $tRNA^{Tyr}$ recognition. This suggests a complex narrative of partial HGT and selection, wherein recombinations between donor and recipient genes in both eukaryal and halobacterial lineages preserved the regions of the ancestral, vertically inherited gene that had co-evolved with its cognate tRNA, to maintain the specificity of that interaction. This is not without precedent, as a partial HGT within halobacterial LeuRS and GluRS have also been reported in previous analyses. In both cases the recombined regions were also shown to be involved in tRNA recognition (Dasgupta and Basu 2014; Fang et al. 2014). These recombinations appear to have occurred following the HGT of a crenarchaeal TyrRS to a subset of Halobacteriales and, consequently, the HGT to the ancestor of Opisthokonta, so that both clades preserved their respective vertically inherited regions (Figure 3). In the case of R1, the recombinations within Halobacteria and Opisthokonta could have occurred independently (if the secondary HGT to Opisthokonts occurred before the R1 recombination within Halobacteria) or in a stepwise fashion, with an R1 recombined copy transferred to Opisthokonta, then replaced via recombination with the vertically inherited eukaryal copy. Within these halobacteria, the entire R2/R3 region appears to have been recombined following HGT from the crenarchaeal donor, preserving the vertically inherited sequence. In the secondary HGT to Opisthokonta, however, the crenarchaeal R2 region from the initial HGT appears to be retained, while the recombined halobacterial R3 region was inherited. Therefore, for the R3 region, but not the R2 region, Opisthokonta groups with Halobacteriales within group II methanogens. This could result from HGT to Opisthokonta after R3 recombination, but before R2 recombination within halobacerial populations.

These events require step-wise recombination of transferred and vertically inherited gene regions within halobacteria, which suggest that both copies of the gene persisted within halobacterial populations for some time. Interestingly, this seems to be the case for halobacterial TrpRS, in which several species contain two gene copies, with one version likely acquired via HGT (Figure 2). This scenario also makes the prediction that, while R1 and R2/R3 regions may both be important in halobacterial tRNA recognition, R1 is most important within opisthokont tRNA recognition, since R1 underwent recombination to preserve this interaction, but did not undergo additional recombination to preserve the eukaryal R2/R3 regions.

While HGT generally is not problematic for ancestral sequence reconstruction of individual proteins, since inferences are based on gene trees rather than organismal trees, partial HGT can confound ancestral reconstruction studies. Site probabilities of ancestral amino acids may be incorrectly inferred if the tree topology for one part of the gene does not match another part. Furthermore, if regions of partial HGT are large enough, the overall gene tree may be impacted, resulting in an artifactual topology different from any of the component "true" evolutionary histories. For this reason, the ancestral sequences for these regions were calculated using gene trees edited to reflect the unique topologies of R1, R2, and R3. Conversely, the gene tree used for ancestral reconstruction of the remaining majority of sites (Figure 2) was generated from an alignment omitting these regions.

Absence of Trp within paralog ancestor sequence reconstructions

Homogeneous and non-homogeneous reconstruction models return similar sequences and site probability distributions for the paralog ancestor node (Figure 4). Of 251 reconstructed aligned sites, 212 show the same amino acid maximum-likelihood identity. Both show a complete absence of sites with a maximum-likelihood identity of Trp (Trp_{ML}) within the paralog ancestor of TyrRS and TrpRS. Additionally, in both models the total expectation count for Trp sites (Trp_{Exp}) within paralog ancestor reconstructions is well below 1 (Figure 5). Importantly, this

absence does not seem to arise from an asymmetrical usage of Trp across TyrRS and TrpRS ancestors; each aaRS family ancestor shows a similar absence of Trp_{ML} sites, and similarly low Trp_{Exp} counts. In both families, Trp residues only begin to be incorporated along the branches leading to the major Domains. Four Trp_{ML} sites are acquired on the branch leading to the TyrRS bacterial ancestor, two on the branch leading to the TyrRS crenarchaeal/eukaryal ancestor, two on the branch leading to the TrpRS bacterial ancestor, and one each on branches leading to the TyrRS euryarchaeal ancestor, and TrpRS euryarchaeal, crenarchaeal, and eukaryal ancestors.

Avoidance in using cognate amino acids within amino acid biosynthetic enzymes has been observed in many pathways, including Trp synthesis (Alves and Savageau 2005; Tivorsak 2001; Xie and Reeve 2005). It has been proposed that this avoids translation attenuation due to a lack of charged tRNA under amino acid starvation conditions, under which these genes must be more highly expressed. Importantly, the inferred absence of Trp within the paralog ancestor does not seem to be a consequence of broad selection against including Trp within TrpRS descendant lineages. Rather, Trp acquisition appears to be near-universal across major groups (Figure 5).

Only groups of transferred halobacterial TrpRS and TyrRS, and associated opisthokont TyrRS appear to lack Trp in extant sequences, in each case likely due to loss after earlier archaeal acquisition, at sites 166 and 278, respectively. It may not be coincidental that two of these affected groups are halobacterial, if these sites are somehow functionally impacted by high salt concentrations. Along the stem branches to these groups, both homogeneous and non-homogeneous reconstructions show a substitution of Trp to Arg at site 166, and a substitution of Trp to Thr at site 278. These particular amino acid substitutions are not generally expected in response to halophily (Fukuchi et al. 2003). Nevertheless, if both are due to selection under halophilic conditions that favor the replacement of Trp, the substitution at site 278 informs the directionality of HGT of TyrRS between Opisthokonta and Halobacteriales. Since Opisthokonta also show a Thr at this site, a halophily-induced substitution would polarize the direction of transfer, suggesting that TyrRS was secondarily transferred from Halobacteria to the Opisthokont ancestor. This would be consistent with the schema inferred from analysis of partial HGT regions (Figure 3).

Per-site probability densities

The estimation of Trp_{Exp} within the paralog ancestor sequence is cumulative across sites, so that as permitted by the substitution model, given a sufficiently long sequence, there will be a substantial expectation of at least one site containing Trp. However, it is clear that the probability density of Trp is non-uniform across sites within the reconstructed paralog ancestor, with the vast majority of sites (>96%) having a very low probability of containing a Trp residue. Most probability density resides within a handful of sites that are consistent across each reconstruction model (Figure 6). Under each reconstruction model, over half of the total expectation of Trp within the paralog ancestor arises from only four sites. This suggests that the expected frequency of Trp within the paralog ancestor sequence does not arise from a lack of reconstruction information across most sites, or a consistent model bias excluding Trp from ancestral sites. Rather, as shown in Figure 4, these few sites contain Trp within descendant lineages, which directly contribute to the probability of Trp at ancestral nodes.

Within both homogeneous and non-homogeneous models, much of the probability density for Trp within the TyrRS/TrpRS paralog ancestor was found within two sites, 190 and 269, together accounting for 38% and 35% of the expectation of Trp within the inferred paralog ancestor sequence, respectively (Figure 6). For both cases, this arises from the derived substitution of Trp along specific branches at the Domain level, rather than a distributed low probability of Trp

across the tree. For site 269, the substitution occurs on the branch leading to the TyrRS bacterial ancestor. For site 190, this substitution occurs on two separate lineages within the tree, leading to the bacterial TrpRS ancestor and crenarchaeal/eukaryal TyrRS ancestor.

Site 269 is the first site of the R1 partial HGT region. However, the proposed differing tree topology for this region does not affect the inference of Trp for this residue within the reconstruction, as neither the halobacterial or eukaryal groups involved in the proposed HGT contain a Trp at this site.
The ancestral reconstruction for site 190 at both LUCA ancestors, as well as the paralog ancestor, is most likely a Leu residue, while the probability of Trp being present within the paralog ancestor at site 190 is relatively low in both homogeneous and non-homogeneous reconstructions (homogeneous: p_{Trp}=0.099, p_{Leu}=0.771; non-homogeneous: p_{Trp}=0.136, p_{Leu}=0.672). This suggests that incorporation of Trp was most likely convergent, and does not represent an ancestral Trp residue within the paralog ancestor. As this site is part of the dimerization interface for both TyrRS and TrpRS, these substitutions may have been independently advantageous consistent with changes in amino acid hydrophobicity and aromatic character being under selection within these protein regions.

Presence of other rare amino acids within the TyrRS/TrpRS ancestor

The absence of Trp within reconstructed sequence ancestors appears to be unique among the twenty amino acids, with other rare amino acids (Cys, His) being present in both TyrRS and TrpRS sequences, and also predicted within the reconstructed paralog ancestor (Table 1). Similar to Trp, Cys and His are also typically conserved in a small number of positions, resisting frequent substitution. This is further evidence against the absence of Trp in the paralog ancestor being the result of model bias due to its low equilibrium frequency.

While the equilibrium frequency of Cys within the TyrRS and TrpRS proteins sampled for this analysis is nearly identical to that of Trp (0.0096 vs. 0.0080, respectively), the paralog ancestor contains at least two and possibly three distinct Cys_{ML} sites, and Cys_{Exp} is least 2.75 times the observed value of Trp_{Exp}. These differences suggest that the absence of Trp cannot be explained purely as model bias arising from the low frequency of Trp usage. While His is more abundant within these proteins (2.5% of sites), it is still the third rarest amino acid within the dataset, and, like Cys, but unlike Trp, has a predicted usage in the paralog ancestor similar to its overall abundance within the dataset.

Table 1. Reconstructed usages of rare amino acids within TyrRS/TrpRS ancestors.

	EqFr	Exp_H	Exp_{NH}	ML_H	ML_{NH}
Trp	0.007	0.54	0.69	0	0
Cys	0.009	1.49	2.27	2	3
His	0.025	7.62	6.68	6	5

EqFr=alignment equilibrium frequency; H=homogeneous reconstruction model; NH=non-homogeneous reconstruction model; Exp=expectation value; ML=maximum-likelihood sites.

Simulated reconstructions correctly predict the absence of Trp in paralog ancestors

There is inherent uncertainty in ancestral sequence reconstruction, propagating from many sources, including phylogenetic uncertainty, nonpolarized characters, and the fitting of incorrect and/or inadequate evolutionary models. Each branch within a phylogeny also loses information about its ancestral state as substitutions continue to occur. For these reasons, the absence of rare

amino acids at sites within an inferred ancestor sequence may not be reflective of the true ancestral state. This is especially true of a non-homogeneous model, which may over-fit by reducing the equilibrium frequencies of very rare amino acids to near zero along some branches, effectively eliminating the possibility of observing them at the inferred ancestor. In order to test for the possibility of a false negative observation of Trp at the inferred paralog ancestor, two sets of 100 simulations were performed for each set of homogeneous and non-homogeneous model parameters inferred from the original sequence alignment and phylogeny. One set of simulations evolved sequences under an evolutionary scenario where root frequencies for Trp were set to 0, so that all simulated sequences evolved from an ancestral state of a total absence of Trp. Another set of simulations evolved sequences with a root frequency of Trp equal to that observed within the actual leaf sequences (0.8%), and a presence of at least one Trp residue in each root sequence. Ancestral reconstructions were then performed with both sets of sequences, using the exact model parameters inferred from the original data. In this way, the ability of the ancestral reconstruction model and method to accurately predict the presence or absence of Trp within the paralog ancestor could be tested.

Simulations under both homogeneous and non-homogeneous models support the hypothesis that the observed absence of Trp within the TyrRS/TrpRS paralog ancestor is conspicuous, and unlikely to be observed if there was a true presence of Trp at the root. Under the homogeneous model (Figure 7A), 64% of reconstructions from zero-Trp simulations correctly inferred zero Trp_{ML} positions within the ancestor. Of the remaining simulations, 27% had a false positive of one Trp_{ML}, 8% had a false positive of two Trp_{ML}, and a single simulation had a false positive of three Trp_{ML}. Only 15% of nonzero-Trp simulations showed a false negative, that is, an absence of Trp_{ML} sites. Under the non-homogeneous model (Figure 7B), 82% of reconstructions from zero-Trp simulations correctly inferred zero Trp_{ML} positions within the ancestor, while the remaining 18% inferred a single Trp residue within the ancestor. Conversely, only 7% of nonzero-Trp simulations incorrectly inferred zero Trp_{ML} sites in the ancestor. Therefore, given these models, it is substantially more likely to observe an absence of Trp_{ML} sites as a true negative, rather than a false negative (Bayes' factors $K_H = 4.27$, $K_{NH} = 11.71$). These simulations further suggest that these models of ancestral reconstruction are more likely to over-estimate the probability and presence of Trp within deep ancestors, rather than under-estimate it.

Since ancestral reconstructions are probabilistic, it is possible that an inferred ancestor with no Trp_{ML} sites could still be likely to contain at least one Trp residue, given the per-site likelihoods across the entire sequence, the sum of which equal the expectation value (Trp_{Exp}) for the count of Trp within the inferred ancestral sequence. Furthermore, since the probability of an ancestral Trp residue at each site is always nonzero, Trp_{Exp} continues to increase with sequence length; in a sufficiently long sequence, even if Trp_{ML} remains zero, Trp_{Exp} will increase to infinity. Comparison to simulated Trp_{Exp} values is therefore a useful additional metric. The homogeneous and non-homogeneous sequence reconstructions of the TyrRS/TrpRS ancestor show Trp_{Exp} values of 0.54 and 0.69, respectively. These values were compared to each set of simulations, in order to determine the relative likelihood of observing similarly low Trp_{Exp} under each hypothesis. Under the homogeneous model, 14% of nonzero-Trp simulations showed expectations lower than 0.54, compared to 63% of zero-Trp simulations. Under the nonhomogeneous model, only 8% of nonzero-Trp simulations showed expectations lower than 0.69, compared with 93% of zero-Trp simulations. (Bayes' factors $K_H = 4.50$, $K_{NH} = 11.63$) (Figure 7).

These simulations show that the inferred low probability of a Trp residue being present within the TyrRS/TrpRS paralog ancestor is substantially more likely to occur under a model where Trp is truly absent at the root. For each evolutionary model, K values for Trp_{ML} and Trp_{Exp} were similar, supporting that each metric is reflecting the same signal within the data.

Methods

Sequence collection, alignments and partial HGT

Amino acid sequences of 182 TyrRS and TrpRS proteins were collected from GenBank (Benson et al. 2014), with a representative sampling across all 3 Domains. Protein BLAST searches within each Domain were performed using the NCBI Non-redundant protein sequences database (Altschul et al. 1990), subsequently using the neighbor-joining tree visualization tool to confirm major clusters of homologs were represented in the sampling. All sequences were aligned in MUSCLE using default parameters (Edgar 2004). Proposed regions of partial HGT involving Opisthokonta and Halobacteriales were identified by visual inspection of amino acid site identities and confirmed with subsequent phylogenetic analysis (see Tree Reconstructions). Proposed partial HGT regions were removed from the alignment before phylogenetic analysis of the remaining sites. Abbreviated sequence name key and aligned sequence files, including partial HGT region files, are available as online resources (see Online Resource files 3-7).

Tree Reconstructions

Phylogenetic trees were generated using PhyML v3.0 (Guindon et al. 2010) with the WAG amino acid substitution model, estimated portions of invariable sites, estimated rate gamma distribution parameter alpha, 8 rate categories, estimated amino acid frequencies, and an NJ starting tree. For the tree generated from sites not involved in partial HGT (shown in Figure 2), 100 bootstrap replicates were generated. In the case of partial HGT regions, trees were initially calculated from each proposed recombined region. Differences in topology for halobacterial and opisthokont groups were then identified and used to edit the tree derived from the remainder of non-recombined sites (the majority of sites) to generate phylogenies that reflect the recombined topology, as well as preserve the phylogenetic relationships between unaffected groups, and within the descendant lineages of affected groups. Branch lengths on the edited trees were then re-estimated using TREEPUZZLE (Schmidt et al. 2002) under the same model parameters. The branch length of the branch leading to the transferred group from each recombined region tree was then used to replace the corresponding value from the edited trees. In this way, each resulting partial HGT tree used for ancestral reconstruction preserves both the topology and branch lengths across all groups using information from the full sequence, as well as reticulated branches associated with each recombined region (Online Resources 8-10). All trees were rooted between paralogs, as depicted in Figure 2.

Model parameter generation and ancestral reconstruction

All homogeneous and non-homogeneous model parameters were estimated using the bppML program belonging to the bppSuite of software (Gueguen et al. 2013). In the homogeneous case, the JTT substitution model was used, with equilibrium frequencies fixed at observed usage. A discrete Gamma distribution with an estimated alpha value of 0.890 and 4 categories was

employed to model variation of rates among sites, plus an additional category for invariant sites (estimated invariant site rate of 0.003). In the non-homogeneous case, the COaLA model (Groussin et al. 2013) was used. COaLA permits the variation of global amino acid composition between lineages. To do so, COaLA assigns branch-specific parameters to explore the space of equilibrium frequencies, in a sub-space defined by a correspondence analysis computed with the matrix of observed frequencies. Two parameters corresponding to the positions along the first two axes of the model were assigned to each branch of the tree, with branch-specific parameters being independent between branches. Two parameters were also assigned to the root.

Only positions from the alignment not predicted to be involved in partial HGT were used for estimating model parameters under the majority site phylogeny (the same sites used to calculate the phylogeny), so as to avoid biases resulting from the inclusion of sites likely evolving under differing tree topologies. Model parameters for partial HGT regions were estimated separately under the adjusted phylogenies for each region, as described in the previous section. Analogous to branch length estimation for the affected bipartitions, only the sets of parameters for the reticulate branches were used from these parameter estimates, with other unaffected branches in the partial HGT model remaining identical to those from the majority model. Ancestral reconstructions were subsequently performed on sequence alignments using bppAncestor (Gueguen et al. 2013). Per-site amino acid probabilities for reconstructed nodes are provided (Online Resources 1-2). A phylogenetic tree with internal node numbers mapping to the ancestral reconstructions is also provided (Online Resources 11). Alignment sites with an excessive number of gaps were automatically excluded from the reconstruction results.

Sequence simulations

Protein sequence evolution was simulated over the majority site TyrRS/TrpRS paralog tree using bppSeqGen (Gueguen et al. 2013) for 100 iterations under both homogeneous and non-homogeneous model parameters, with and without Trp sites in the root sequence. For iterations with Trp, root frequencies were set equal to the observed alignment frequency of 0.008, with at least one Trp site in the root sequence. For iterations without Trp, root frequencies of Trp were set to 0%. Alignment length was set equal to the number of reconstructed sites within the actual TyrRS/TrpRS alignment (n=251). Ancestral reconstruction was then performed for each resulting simulated alignment, using the same tree and model parameters as the TyrRS/TrpRS reconstructions using the majority sequence. Due to their short lengths and likely small impact on estimated root frequencies of Trp, the different inferred topologies arising from partial HGT regions were not used in simulations. The site counts of these regions are included in the simulated sequence alignment length.

Discussion

A limited co-evolution of aaRS and the genetic code

The time between the origin of life and LUCA is one of dramatic evolutionary change, encompassing the invention of all the core cellular machinery, including the translation system and the genetic code itself. Despite the importance in this interval in establishing fundamental biological processes, the tools of comparative genomics are limited in its investigation, save for the relationships between highly conserved paralogous gene families that can be inferred to have diverged pre-LUCA. Due to their ancient origins, high levels of structural, functional, and sequence conservation, and critical role within protein synthesis and the syntax of the genetic code, aaRS proteins are an especially informative in this regard.

Combined with our previous results showing the parafunctionaliztion of the ValRS/IleRS protein families (Fournier et al. 2011), this work points to the evolution of the genetic code being an extended process that continued throughout an increasingly complex and protein-based system. The many stages of code evolution likely included early stages that occurred within and were mediated by an RNA-based (or other) physiology. At later stages, possibly once proteins had increased in specificity and functionality due to this very same genetic code expansion, protein evolution could directly shape the genetic code itself. This stage may have been very late indeed, if the aaRS-mediated addition of Trp to the genetic code was the final stage in the evolution to its current form. The observation that all other amino acids are represented within the TyrRS/TrpRS paralog ancestor support this hypothesis. However, this does not preclude the possibility of subsequent parafunctionalization events within other aaRS lineages. In fact, the inferred ValRS/IleRS paralog ancestor does contain sites specific for both Tyr and Trp , suggesting that any takeover of aminoacylation activities for Val and/or Ile by these proteins would have occurred after the addition of Trp to the code. Given the relatively short branch lengths separating these pairs of aaRS paralogs in both cases (compared to other sets of aaRS families), both divergences were likely among the most recent among aaRS families, making this order of events a plausible scenario.

The results of this work also predict that other groups of protein families arising before the addition of Trp to the code should also lack Trp in their ancestral sequences. As such, there should be a clear delineation between pre-Trp and post-Trp proteins, as evidenced by the presence or absence of conserved Trp sites within their ancestral sequences. It may be possible to make such distinctions among other ancient paralogous gene families diverging pre-LUCA.

Partial HGT and ancestral reconstruction

Ancestral sequence reconstruction of a protein generally assumes that each site shares the same evolutionary history, as is generally assumed in phylogenetic reconstruction of single genes, even for those undergoing HGT. However, if different parts of a protein alignment evolved under different histories, as is the case following partial HGT, this assumption is invalid, and the ancestral reconstruction may be incorrect for the affected regions. If the recombination(s) contain sites that experienced substitutions along the impacted lineages, specious inferences will propagate to the site probabilities of ancestral nodes. Depending on the size of the recombined region and the phylogenetic depth of the reticulation, these events may be highly disruptive to both model parameter estimation and the accuracy of reconstructed ancestor sequences. In this analysis, we have attempted a maximum-information approach to deal with three predicted partial HGT regions, by reconstructing them individually under their own respective topologies, using as much phylogenetic and model information as possible from the remainder of the alignment. One alternative approach would involve removing the affected clades entirely, sacrificing phylogenetic information in order to prevent false inferences at the affected sites. To our knowledge, previously published reconstructions of ancient protein families have not been impacted by known partial HGT events. While it has been shown that partial HGT events have occurred within the EF-1a sequences of some archaeal lineages (Inagaki et al. 2006), these groups were not included in the published ancestral reconstruction of homologous bacterial EF-Tu proteins (Gaucher et al. 2003).

Future work

Detailed mechanistic analyses of amino acid recognition and aminoacylation within TyrRS and TrpRS have been performed (Doublie et al. 1995; Hogue et al. 1996; Praetorius-Ibba et al. 2000);

applying the results of these models to the inferred sequence of TyrRS/TrpRS paralog ancestor may further elucidate ancestral function. By the same principle, reconstructed sites involved with tRNA discrimination may also be useful in determining ancestral tRNA specificity, and, by extension, amino acid specificity (Bedouelle et al. 1993; Tsunoda et al. 2007). Direct biochemical specificity assays on resurrected ancestor protein sequences is another possible avenue of investigation (e.g., Gaucher et al. 2003). However, this approach would be challenged by very large evolutionary distances and correspondingly higher levels of uncertainty in pre-LUCA paralog ancestor amino acid site identities. Due to the combinatorics arising from even a subset of sites having an uncertain ancestral identity, it is likely that a very large number of potentially "true" ancestor sequences would need to be synthesized and tested in order to accurately and fully explore the likely phenotypic space of the ancestral function. However, in the case of an informative reconstruction, one would predict that the likelihood of each predicted resurrected ancestor (calculated from per site amino acid probabilities) would positively correlate with the true ancestral amino acid binding specificity/aminoacylation activity. Thus, a more sparse sampling within the space of possible ancestors may provide a tractable solution.

Acknowledgements

This work was supported by National Science Foundation Grant 0936234, NASA Astrobiology Institute Grant NNA08CN84A, and an appointment from the NASA Postdoctoral Program to GPF at the Massachusetts Institute of Technology. We thank Mathieu Groussin and Bastien Boussau for helpful discussions and their assistance with implementing non-homogeneous ancestral reconstruction models.

Conflict of Interest

The authors declare that they have no conflict of interest.

Figure Legends

Figure 1. Hypotheses for the ancestral amino acid specificity of TyrRS and TrpRS paralogs. Proteins (grey) and cognate amino acids (white) are shown in this schema. Tyr (Y) and Trp (W) represent the presence of these amino acids within the paralog ancestor and descendants under each hypothesis. W/Y represents an ambiguous specificity, in which Tyr and Trp are not discriminated by the genetic code during protein synthesis, or in aaRS binding. In the case of parafunctionalization, the cognate amino acid(s) cannot be inferred from composition analysis, as under this hypothesis specific Tyr and Trp usage in the genetic code both predate their cognate aaRS.

Figure 2. Maximum likelihood tree of TyrRS and TrpRS paralogs. Colors indicate broad taxonomic groups, as indicated. TACK refers to the TACK superphylum, in this tree including Crenarchaeota, Thaumarchaeora, and Korarchaeota. Black lines indicate lineages existing either earlier than Domain ancestors, or for which no taxonomic group could be inferred due to patterns of HGT. Tree reconstruction is described in *Methods*. Bootstrap support values are provided for major nodes. Branch length scale bar indicates the number of substitutions/site.

Figure 3. Proposed schema for HGT and recombination of gene encoding TyrRS within Halobacteriales and Opisthokonta. Given that many halobacterial genomes retain the vertically inherited TyrRS homolog, the initial transfer was either likely to within the halobacterial clade, or lineage sorting of the vertical and transferred copy occurred following HGT into the halobacterial ancestor. Subsequent HGT

to the ancestor lineage of Opisthokonta supports a stepwise series of recombination events within a halobacterial populations. Recombined regions R1-R3 are not shown to scale.

Figure 4. Site amino acid probabilities in reconstructed TyrRS/TrpRS paralog ancestor. Gaps indicate sites with protein family-specific indels that were not included in the final analysis. Partial HGT regions are labeled on the left. Sites corresponding to labeled Trp_{Exp} contributions in Figure 6 are numbered. Site probabilities for Trp are in bold red. (*) indicates sites with ML amino acid identities that differ between homogeneous and non-homogeneous reconstructions. Site amino acid probabilities for all internal nodes of each ancestral reconstruction are provided (Online Resources 1-2).

Figure 5. Expectation counts of Trp within reconstructed ancestors of the TyrRS/TrpRS protein families. Results from homogeneous (A) and non-homogeneous (B) reconstruction models consistently support an absence of Trp within the pre-LUCA paralog ancestor of TyrRS and TrpRS. Trp_{Exp} values for each branch are labeled and proportional to branch thickness. Additionally, bracketed values in bold indicate site specific acquisition (+) and loss (-) of Trp_{ML} along their respective branches. [0] values indicate no Trp_{ML} sites are observed along a branch. (*) an ambiguous gain of Trp_{ML} is observed at site 278, which may have preceded multiple gains and losses in descendent lineages. Terminal branches correspond to the stem lineages ancestral to each labeled clade.

Figure 6: Contribution of individual sites to paralog ancestor expectation of Trp. The majority of Trp_{Exp} within reconstructed TyrRS/TrpRS paralog ancestor sequences is within only a few sites, consistently recovered across both reconstruction models. In each model, the vast majority of sites (>96% of sites) contribute less than 33% of the total expectation value for Trp at the root. Sites contributing >5% Trp_{Exp} are labeled. Additional sites contributing remaining Trp_{Exp} are also consistent across homogeneous and non-homogeneous models. In order of decreasing contribution per model: *a* = 166, 265, 633, 277, 282; *b* =164, 166, 277, 265, 633, 282.

Figure 7: Ancestral reconstruction simulations support a conspicuous absence of Trp within the TyrRS/TrpRS paralog ancestor. Homogeneous (A) and non-homogeneous (B) models were tested. Simulations under each model (n=100) were performed with a 0% root frequency of Trp (black) and a 0.8% root frequency of Trp with at least 1 Trp residue within the simulation root ancestor node (grey), respectively. Actual Trp_{ML} and Trp_{Exp} values (Observed) were compared to distributions of corresponding values for each set of simulations. The Bayes' factor (K) for each test shows substantial support for the observed absence of Trp being more likely due to a true absence of Trp within the paralog ancestor sequence.

Online Resource Captions

Online Resource 1: Homogeneous ancestral reconstruction of TyrRS/TrpRS paralog ancestor, per-site amino acid probabilities. Sites within majority topology regions are labeled V(vertical), sites within proposed partial HGT regions are labeled R1-R3(recombined). Combined alignment sites refers to the native sequence alignment, and matches the numbering used throughout the manuscript. Topology-specific alignment sites refer to the provided FASTA format sequence alignments for each topology region (V, R1, R2, R3). As R regions were removed from the V alignment, these numberings differ. Numbering in the "prob" column headings refers to each internal node reconstruction. The mapping of these nodes to the phylogeny is provided in Online Resource 11. For each node, "max" refers to the maximum likelihood AA for each site.

Online Resource 2: Non-homogeneous ancestral reconstruction of TyrRS/TrpRS paralog ancestor, per-site amino acid probabilities. Sites within majority topology regions are labeled V(vertical), sites within proposed partial HGT regions are labeled R1-R3(recombined). Combined alignment sites refers to the native sequence alignment, and matches the numbering

used throughout the manuscript. Topology-specific alignment sites refer to the provided FASTA format sequence alignments for each topology region (V, R1, R2, R3). As R regions were removed from the V alignment, these numberings differ. Numbering in the "prob" column headings refers to each internal node reconstruction. The mapping of these nodes to the phylogeny is provided in Online Resource X. For each node, "max" refers to the maximum likelihood AA for each site.

Online Resource 3: Table of species names and abbreviations used in online trees and alignments.

Online Resource 4: FASTA format alignment of TyrRS/TrpRS protein sequences, with partial HGT regions removed.

Online Resource 5: FASTA format alignment of TyrRS/TrpRS protein sequences, proposed partial HGT region R1.

Online Resource 6: FASTA format alignment of TyrRS/TrpRS protein sequences, proposed partial HGT region R2.

Online Resource 7: FASTA format alignment of TyrRS/TrpRS protein sequences, proposed partial HGT region R3.

Online Resource 8: Phylogeny for proposed partial HGT region R1.

Online Resource 9: Phylogeny for proposed partial HGT region R2.

Online Resource 10: Phylogeny for proposed partial HGT region R3.

Online Resource 11: Phylogeny with mapping for ancestral node reconstruction numberings.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410

Alves R, Savageau MA (2005) Evidence of selection for low cognate amino acid bias in amino acid biosynthetic enzymes. Mol Microbiol 56:1017-1034

Andam CP, Williams D, Gogarten JP (2010) Biased gene transfer mimics patterns created through shared ancestry. Proc Natl Acad Sci U S A 107:10679-10684

Bedouelle H, Guez-Ivanier V, Nageotte R (1993) Discrimination between transfer-RNAs by tyrosyl-tRNA synthetase. Biochimie 75:1099-1108

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2014) GenBank. Nucleic Acids Res 42:D32-37

Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P (2005) Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol 6:R42

Brooks DJ, Fresco JR (2002) Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol Cell Proteomics 1:125-131
Brooks DJ, Fresco JR, Lesk AM, Singh M (2002) Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. Mol Biol Evol 19:1645-1655
Brown JR, Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci U S A 92:2441-2445
Brown JR, Robb FT, Weiss R, Doolittle WF (1997) Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. J Mol Evol 45:9-16
Cavalcanti AR, Leite ES, Neto BB, Ferreira R (2004) On the classes of aminoacyl-tRNA synthetases, amino acids and the genetic code. Orig Life Evol Biosph 34:407-420
Chandrasekaran SN, Yardimci GG, Erdogan O, Roach J, Carter CW, Jr. (2013) Statistical evaluation of the Rodin-Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases. Mol Biol Evol 30:1588-1604
Dasgupta S, Basu G (2014) Evolutionary insights about bacterial GlxRS from whole genome analyses: is GluRS2 a chimera? BMC Evol Biol 14:26
Di Giulio M (1992) The evolution of aminoacyl-tRNA synthetases, the biosynthetic pathways of amino acids and the genetic code. Orig Life Evol Biosph 22:309-319
Dong X, Zhou M, Zhong C, Yang B, Shen N, Ding J (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 38:1401-1412
Doublie S, Bricogne G, Gilmore C, Carter CW, Jr. (1995) Tryptophanyl-tRNA synthetase crystal structure reveals an unexpected homology to tyrosyl-tRNA synthetase. Structure 3:17-31
Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797
Fang ZP et al. (2014) Coexistence of bacterial leucyl-tRNA synthetases with archaeal tRNA binding domains that distinguish tRNA(Leu) in the archaeal mode. Nucleic Acids Res 42:5109-5124
Fondi M, Brilli M, Emiliani G, Paffetti D, Fani R (2007) The primordial metabolism: an ancestral interconnection between leucine, arginine, and lysine biosynthesis. BMC Evol Biol 7 Suppl 2:S3
Fournier GP, Andam CP, Alm EJ, Gogarten JP (2011) Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. Orig Life Evol Biosph 41:621-632
Fournier GP, Gogarten JP (2007) Signature of a primitive genetic code in ancient protein lineages. J Mol Evol 65:425-436
Fournier GP, Gogarten JP (2010) Rooting the ribosomal tree of life. Mol Biol Evol 27:1792-1801

Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K (2003) Unique amino acid composition of proteins in halophilic bacteria. J Mol Biol 327:347-357

Gaucher EA, Thomson JM, Burgan MF, Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285-288

Grosjean H, de Crecy-Lagard V, Marck C (2010) Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. FEBS Lett 584:252-264

Groussin M, Boussau B, Gouy M (2013) A branch-heterogeneous model of protein evolution for efficient inference of ancestral sequences. Syst Biol 62:523-538

Gueguen L et al. (2013) Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol Biol Evol 30:1745-1750

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-321

Hartlein M, Cusack S (1995) Structure, function and evolution of seryl-tRNA synthetases: implications for the evolution of aminoacyl-tRNA synthetases and the genetic code. J Mol Evol 40:519-530

Hartman H, Smith TF (2014) The Evolution of the Ribosome and the Genetic Code. Life 4:227-249

Higgs PG (2009) A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biol Direct 4:16

Higgs PG, Pudritz RE (2009) A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology 9:483-490

Hogue CW, Doublie S, Xue H, Wong JT, Carter CW, Jr., Szabo AG (1996) A concerted tryptophanyl-adenylate-dependent conformational change in Bacillus subtilis tryptophanyl-tRNA synthetase revealed by the fluorescence of Trp92. J Mol Biol 260:446-466

Huang J, Xu Y, Gogarten JP (2005) The presence of a haloarchaeal type tyrosyl-tRNA synthetase marks the opisthokonts as monophyletic. Mol Biol Evol 22:2142-2146

Inagaki Y, Susko E, Roger AJ (2006) Recombination between elongation factor 1alpha genes from distantly related archaeal lineages. Proc Natl Acad Sci U S A 103:4528-4533

Jukes TH (1973) Possibilities for the evolution of the genetic code from a preceding form. Nature 246:22-6

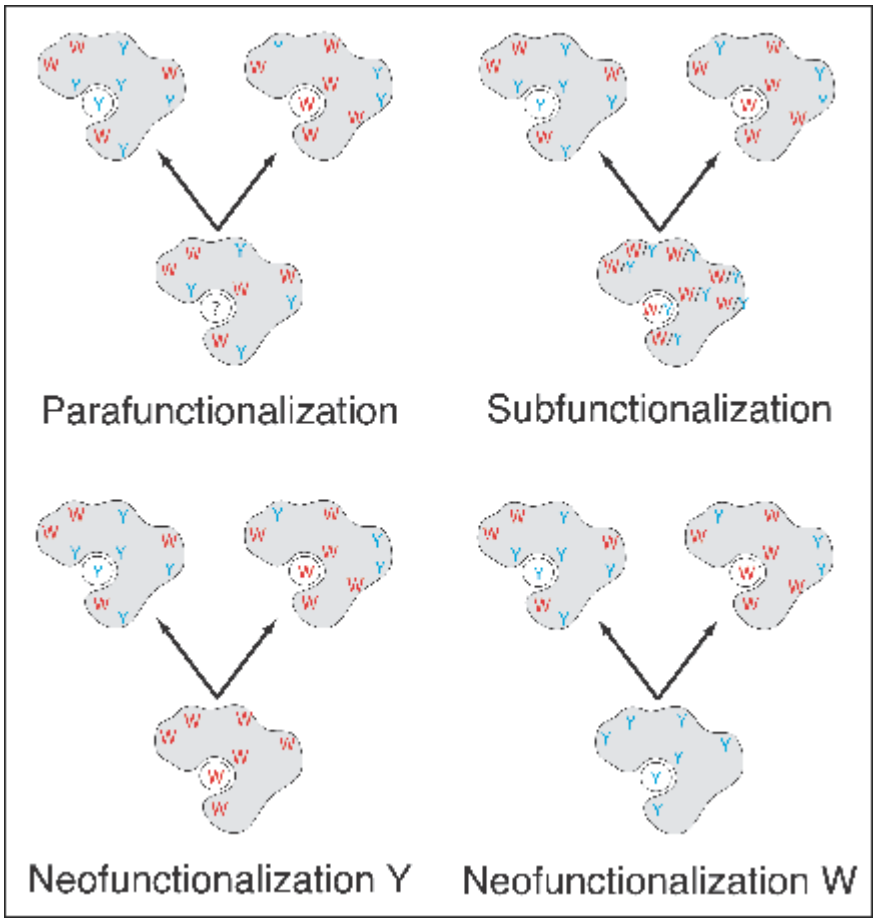
Jukes TH (1981) Amino acid codes in mitochondria as possible clues to primitive codes. J Mol Evol 18:15-17

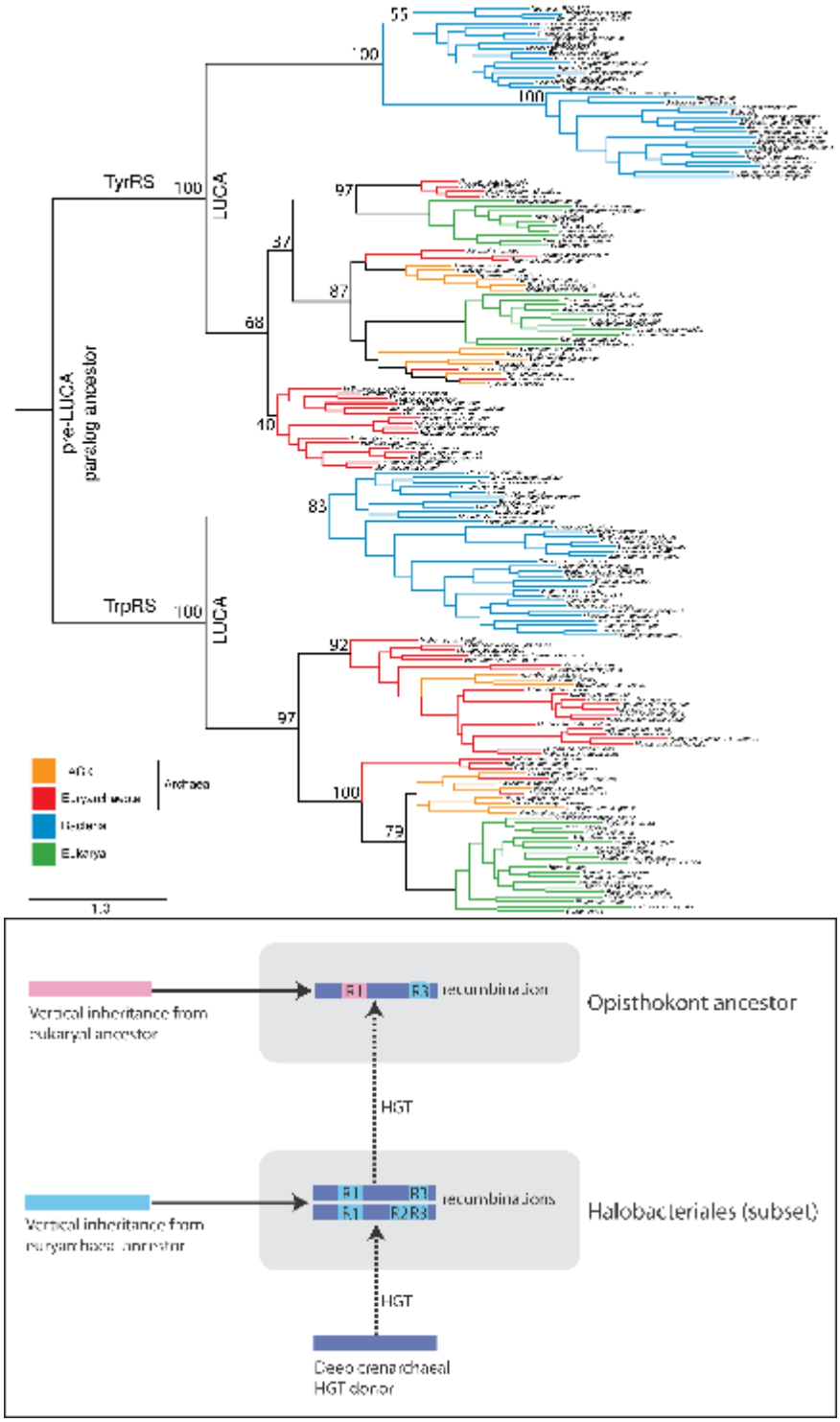
Klipcan L, Safro M (2004) Amino acid biogenesis, evolution of the genetic code and aminoacyl-tRNA synthetases. J Theor Biol 228:389-396

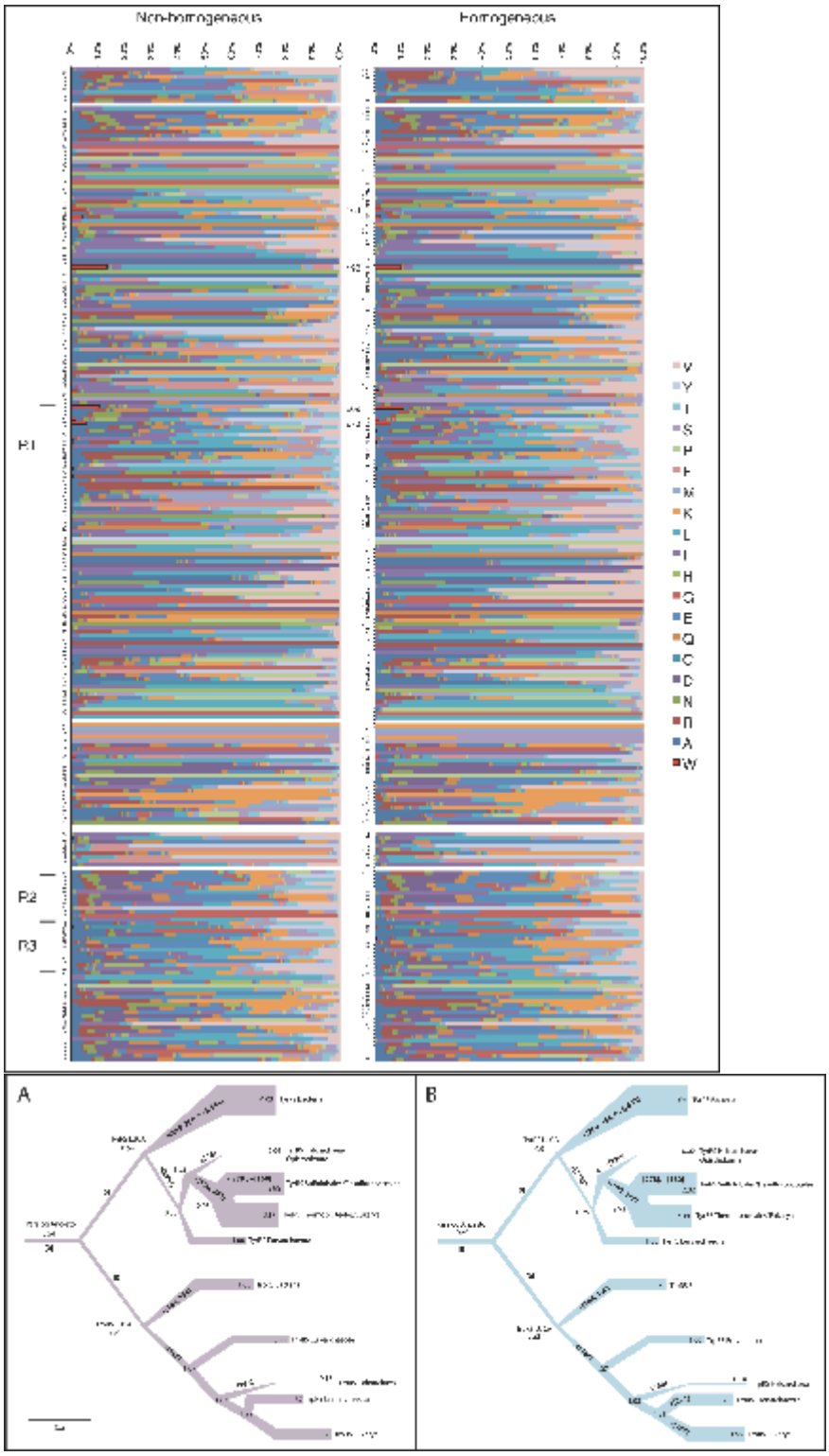
Knauth LP (2005) Temperature and salinity history of the Precambrian ocean: implications for the course of microbial evolution. Palaeogeogr Palaeoclimatol Palaeoecol 219:53-69

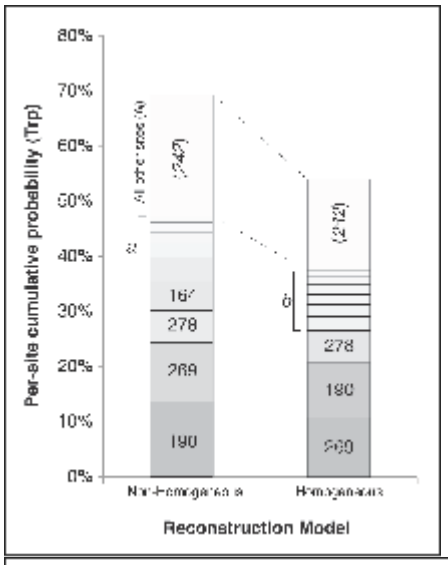
Koonin EV, Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. IUBMB Life 61:99-111

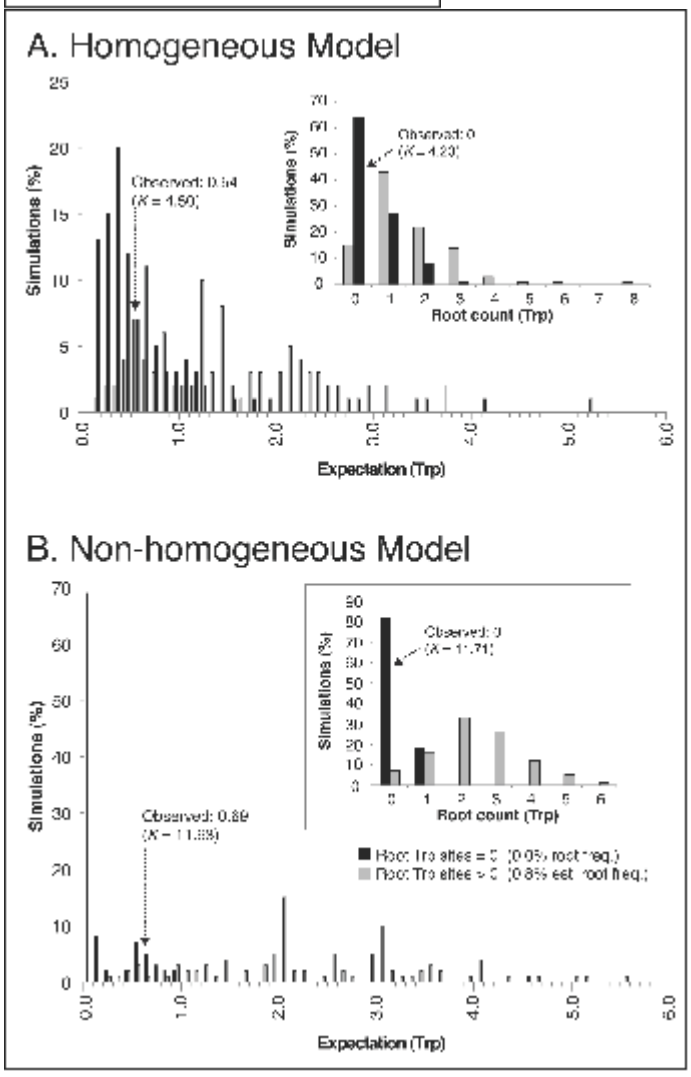
Landes C, Perona JJ, Brunie S, Rould MA, Zelwer C, Steitz TA, Risler JL (1995) A structure-based multiple sequence alignment of all class I aminoacyl-tRNA synthetases. Biochimie 77:194-203
Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117:528-529
Nagel GM, Doolittle RF (1995) Phylogenetic analysis of the aminoacyl-tRNA synthetases. J Mol Evol 40:487-498
Osawa S, Jukes TH, Watanabe K, Muto A (1992) Recent evidence for evolution of the genetic code. Microbiol Rev 56:229-264
Podar M et al. (2008) A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. Genome Biol 9:R158
Praetorius-Ibba M et al. (2000) Ancient adaptation of the active site of tryptophanyl-tRNA synthetase for tryptophan binding. Biochemistry 39:13136-13143
Ribas de Pouplana L, Frugier M, Quinn CL, Schimmel P (1996) Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria. Proc Natl Acad Sci U S A 93:166-170
Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-504
Tivorsak TL (2001) Reconstructing Ancestral Biosynthetic Enzymes: An Approach to Explore the Evolution of the Genetic Code - Tryptophan Synthase as a Model. Senior Thesis, Princeton University
Trifonov EN (2000) Consensus temporal order of amino acids and evolution of the triplet code. Gene 261:139-151
Tsunoda M et al. (2007) Structural basis for recognition of cognate tRNA by tyrosyl-tRNA synthetase from three kingdoms. Nucleic Acids Res 35:4289-4300
Vetsigian K, Woese C, Goldenfeld N (2006) Collective evolution and the genetic code Proc Natl Acad Sci U S A 103:10696-10701
Wetzel R (1978) Aminoacyl-tRNA synthetase families and their significance to the origin of the genetic code. Orig Life 9:39-50
Wetzel R (1995) Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code. J Mol Evol 40:545-550
Wong JT (1988) Evolution of the genetic code. Microbiol Sci 5:174-181
Xie Y, Reeve JN (2005) Regulation of tryptophan operon expression in the archaeon Methanothermobacter thermautotrophicus. J Bacteriol 187:6419-6429


Parafunctionalization
Subfunctionalization
Neofunctionalization Y
Neofunctionalization W


TyrRS
TrpRS
LUCA
LUCA
pre-LUCA
paralog ancestor
100
100
55
100
97
37
87
88
10
83
92
97
100
79
Archaea
1.0
Vertical inheritance from
eukaryal ancestor
R1
R3
recombinations
Opisthokont ancestor
HGT
Vertical inheritance from
R1
R1
R1
R2R3
recombinations
Halobacteriales (subset)
HGT
Deep crenarchaeal
HGT donor


Non-homogeneous
Homogeneous
P1
P2
P3
V
Y
I
S
P
F
M
K
L
I
H
G
E
Q
C
D
N
T
A
W
A
B


80%
70%
60%
50%
40%
30%
20%
10%
0%
Per-site cumulative probability (Trp)
(2342)
(2132)
167
278
269
190
278
180
Reconstruction Model


A. Homogeneous Model
Simulations (%)
Observed: 0.64
(K = 1.50)
Observed: 0
(K = 4.23)
Root count (Trp)
Expectation (Trp)
B. Non-homogeneous Model
Simulations (%)
Observed: 0.89
(K = 11.89)
Observed: 0
(K = 11.71)
Root count (Trp)
Expectation (Trp)