Escherichia coli transforms the leucyl-tRNA synthetase gene *in vivo* into primordial genes

Guo Qing Tang and, Charles W. Carter, Jr

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260

ORCID ID:

C.W.C: 0000-0002-2653-4452

Summary

The Last Universal Common Ancestor (LUCA) lived about 4.2 billion years ago and had a nearly modern metabolism[1]. Genes and proteins must therefore have achieved modern lengths during a time comparable to the error in that date[2]. How did that happen? We show here that *E coli* transformed with a double mutant full-length Leucyl-tRNA synthetase (LeuRS) produces discrete sets of shorter genes. These lack the anticodon-binding domain and large parts of the catalytic domain. They connect remote active site parts in two different ways. Large pre-steady-state bursts confirm that they are the active enzymes. These *in vivo* results validate earlier designs for ancient aminoacyl-tRNA synthetase enzymes, greatly expanding the sequence space of active synthetases. One construct joins the 56-residue protozyme directly to a 25-residue segment containing the second catalytic signature. It catalyzes both activation and minihelix acylation with amino acids. AlphaFold3[3] predicts that the mRNA sequence encoding the latter fragment is a long hairpin. Thus, 3D structures in the gene itself may promote the deletions[4]. The deletions appear to reverse the modular evolution of full-length synthetases from simpler catalysts. The reverse evolution we describe could open broad access to primordial gene discovery.

Moody, et. al. proposed that the Last Universal Common Ancestor of all living things (LUCA) had a nearly modern metabolism[1]. The same study dated LUCA to about 4.2 billion years ago. Similar estimates place the origins of organic chemistry on primitive earth to about the same time. Thus, both nucleic acids (i.e., genes) and proteins must have arisen within the errors of the various dates[2]. Aminoacyl-tRNA synthetases (AARS)[5] and their cognate tRNAs needed for translation appear to have had modern sequences before LUCA appeared[5,6].

We created models of primordial AARS called "urzymes"[7-9] and "protozymes"[10,11] to help define that transition[12,13]. Those models for very early stages of AARS evolution[14-18] are based on structural biology. "Urzymes" have 120-130 residues and nearly intact catalytic sites. "Protozymes" are ~50-residue peptides containing the ATP binding sites. Their chief advantage is that their catalytic activities allow us still to perform experiments. Modern AARS, however, are nearly an order of magnitude longer than their primordial ancestors. That gap seems to have been filled by adding new modular blocks of genetic information to the ancestral forms[14].

A central question is how to connect the much shorter coding sequences of ancestral AARS quasispecies to conventional ancestral sequence reconstructions. The latter are currently limited to essentially full-length species that had already been assembled by the time of the last universal common ancestor, LUCA. It seems likely that some process or processes added new genetic material as modules of significant size. How did AARS genes undergo that modular growth?

We recently stumbled onto a novel *in vivo* process in *Escherichia coli* that seems to open a new window on how that happened. To better define the functional properties of AARS urzymes[15], we constructed a plasmid containing a mutant full-length *Pyrococcus horikoshii* leucyl-tRNA synthetase (LeuRS) gene in pMal c2x (https://addgene.org/214193). The mutant gene had alanine mutations to both HVGH(AVGA) and KMSKS(AMSAS) signatures[15]. *E. coli* generated many discrete deletions from that exogenous and invasive plasmid. Curation of those deletions revealed a nested hierarchy of sequences similar to those of the protozymes and urzymes we had made on purpose.

It is routine to isolate and sequence plasmids from multiple transformants to identify one with the desired sequence[19]. The polymerase chain reaction (PCR)[20] can allow single point mutations to accrue. Mutant plasmids can also rearrange in cells grown after transformation. Heretofore, such variation has generally been thought to be a nuisance, as for example were the earliest observations of microRNAs in gels and northern blots[21]. This is especially true in cases such as ours, that involve the orthogonality of the respective AARS[22]. Various precautions can increase the frequency of the desired plasmid[19] (see also[23]). To our knowledge, nobody has considered the spectrum of plasmids produced after transformation as a coherent source of insight. In our case, however, the surviving plasmids contained uniquely useful data.

These data and preliminary evidence from AlphaFold3[3] predictions of the intact LeuRS mRNA suggest that intrinsic structures in the LeuRS gene itself promoted the observed deletions. This *in vivo* system may thus offer an important new window on the modular assembly of protein genes.

E. coli creates a discrete population of LeuRS urzyme-like sequences *in vivo*.

The double mutant full-length sequence served as a megaprimer to preclude adventitious loss of sequences during the PCR[24]. We sequenced 97 plasmids from transformed single colonies to assess the degree of homogeneity in the ensemble. About 65% (63) of sequenced plasmids isolated had the desired, full-length double mutant. The raw sequence data for the insertions in the remaining 34 plasmids are in Supplemental Section A. A fasta file with sequences related to LeuAC urzymes are in Supplemental Section B. Careful curation of these sequences revealed three distinct clades (Fig. 1).

The shortened plasmids deleted major segments and re-sorted the remaining modular sequences from the full-length LeuRS gene in idiosyncratic ways, so that rather than ~2900 nt, they have between 290 – 450nt,

far shorter than the full length LeuRS gene. Fig. 1A reveals an unexpectedly coherent mosaic. A simple neighbor joining tree reveals three main clades (Fig. 1B).

Each clade differs in the same way from the analytical LeuAC urzyme[25,26] construct we studied earlier[15,27]. The longest ORF in Fig. 1A (300 – 450nt; 41% (14 sequences)), includes the second crossover[28] of the Rossmann fold[29] (green backgrounds in Fig. 1A). These sequences also have a truncated protozyme and an extra segment that is not part of LeuAC (red band above). About 24% (8) of the 34 sequences encode truncated protozymes, (blue backgrounds). Another ~20% (7) are intermediate in size, (tan backgrounds). These are ~81 residues long. Curiously, the lengths of the truncated protozyme ORFs match exactly the portion of the protozyme in the green, LeuAC-like ORFs. The protozyme segment is twice as long in the intermediate-sized ORF which lacks the entire green segment. Each clade in the ensemble has remarkable discreteness and reproducibility. Occasional point mutations are neither reproducible nor highlighted.

We will refer to the 81-residue ORFs as "Goldilocks" urzymes because the deletion is "just right" for antisense alignment with the shortest Class II urzymes[30].

Supplementary Section C contains figures showing the variety of urzyme, Goldilocks, and protozyme sequences that appeared in sequenced plasmid samples. The curious genomic features include single open-reading frame deletions, (Supplemental figure 1, 5); tandem-configurations, (Supplemental Figure 2, 3, 7); and multi-frame embedding within the same plasmid (Supplemental Figure 3, 4, 6, 7).

The bottom 4 sequences in the second half of the green LeuAC-like deletions have an inserted segment that derives from two different regions of the full-length sequence. Blue shading represents a duplication of the first half of the protozyme. It is unchanged in both protozyme-like and LeuAC-like sequences. A short N-terminal segment derives from the C-terminus of both the LeuAC-like and Goldilocks urzymes, C-terminal to the AMSAS signature is magenta. The bottom 5 sequences are aberrant deletions difficult to classify.

Interestingly, the temperature at which transformed colonies were incubated had a substantial effect on the distribution of these genomic properties. Notably, those incubated exclusively at 37° C had no Goldilocks deletions and far fewer tandemly repeated and multiframe encoding than did those at 4° C (Fig. 1C).

We selected excerpts from the latter two clades, Clone22 and Clone21 (Fig. 1B), respectively, for expression and functional characterization.

The deletions must occur *in vivo*.

Controls tell us a bit about what happened. Full-length mutants at either site do not produce the ensemble (data not shown). Nor does the reciprocal mutagenesis of double mutant plasmid with WT primers. The process is therefore one-way. The deletions derive not from the PCR mutagenesis itself. They were likely selected because they eliminate the toxic intracellular mis-acylation by the full-length double mutant protein. What is remarkable is that so many of the remaining sequences preserve AARS catalytic activity. Although we cannot suggest a detailed mechanism, the deletions must have been created and selected *in vivo*. The *E. coli* strain carries a RecA^{-} gene. The most likely creation thus entailed other recombination pathways[31]. There are likely clues as well in the temperature dependent appearance of recombinant deletions (Fig. 1C). We cannot, however, interpret their significance.

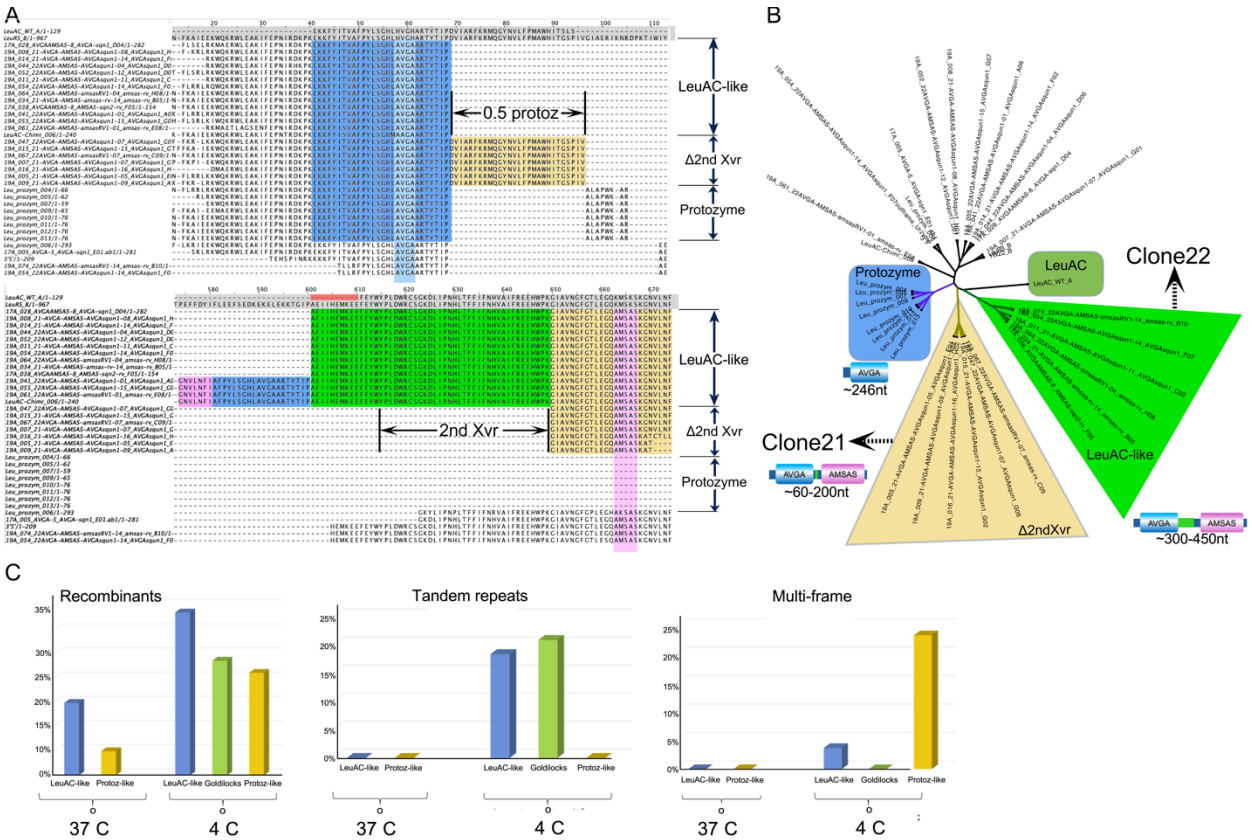


Figure 1. Plasmids isolated from 34 transformed colonies, excluding those that contain full-length double mutant LeuRS, form a discrete, coherent ensemble. **A.** Modular deletions and shuffling in excerpts from the multiple sequence alignment in Supplementary Data B. Both panels together show the LeuAC sequence and segments from full-length *P. horikoshii* leucyl-tRNA synthetase in rows 1 and 2 (light grey background). Vertical light blue and magenta columns indicate the two mutant signatures. Green horizontal backgrounds highlight sequences of the LeuAC-like urzymes, which lack the second half of the protozyme. Tan horizontal backgrounds highlight urzyme sequences that lack the second crossover connection of the Rossman dinucleotide-binding fold[29,32]. **B.** Neighbor-joining tree from the alignment in A highlights the three principal clades of related sequences. Coloring is that in A. For subsequent reference Clone21, the Goldilocks deletion variant, was excerpted from one of the second crossover deletions (denoted Δ2nd Xvr). Clone 22 contains the second crossover lacking in Clone21 and has an extended segment preceding it (red superbar). **C.** Temperature-dependence of genomic properties show that incubating transformant colonies at lower temperatures leads to the appearance of Goldilocks deletions and to increased tandem and multiframe repetition.

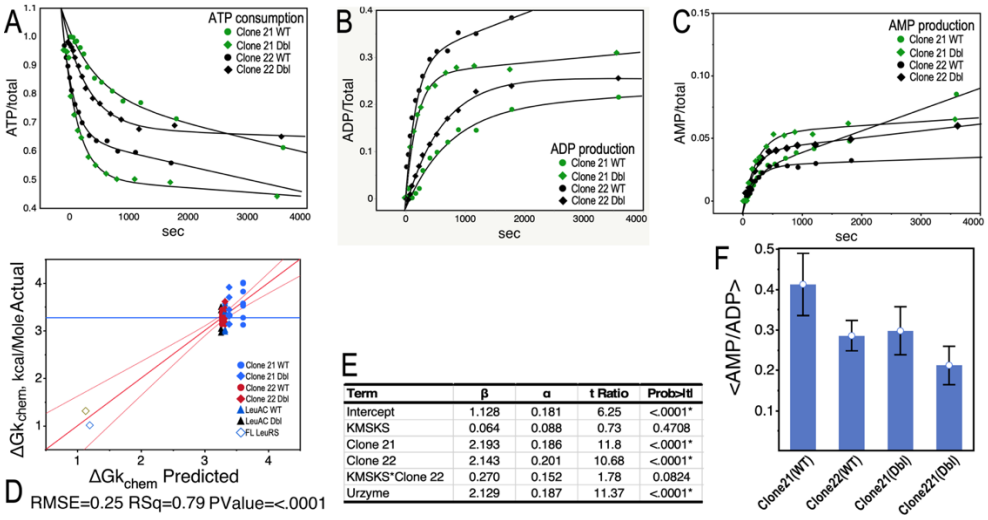
The deletions encode active tRNA synthetases.

AARS urzymes retained three catalytic properties seen in full length AARS. (i) They accelerated amino acid activation. (ii) They exhibited significant pre-steady state bursts in single-turnover kinetic assays[33,34]. And (iii) they accelerated aminoacylation of tRNA and minihelix substrates[27]. Clone21 and Clone22 (Fig. 1B), retain all three catalytic activities at levels comparable to those observed for the urzymes generated intentionally on the basis of structural biology[7].

Fig. 2 summarizes single-turnover experiments on wild type (WT), double (AVGA/AMSAS; Dbl) mutations of Clones 21, 22, and LeuAC urzyme[15,26,27] variants. Single-turnover experiments[34,35] use large enzyme concentrations to build the first-round product formation into a measurable signal. They use only a small excess of substrate; Most kinetic events are unimolecular, hence first-order. If product release is rate-limiting, the product builds up in the enzyme. Such experiments show a pre-steady-state burst. The burst size represents the fraction of active enzyme molecules contributing to the experimental signal.

Those data confirm both (i) and (ii). Time courses for the three adenine nucleotides of all four variants (Fig. 2A-C) are very similar to those observed previously for LeuAC[27]. The first-order rate constants for amino

acid activation (Supplementary Table S1) are very nearly the same as those for LeuAC (Fig. 2D, E) and orders of magnitude slower than those of full-length LeuRS. Moreover, as we found for LeuAC[15] the AVGA and AMSAS mutations actually enhance $k_{chem.}$ for Clone 21. But for Clone 22, which is slightly more active, that effect was reversed. P-values for the β-coefficients of the regression model in Fig. 2D (Fig. 2E) suggest that although the difference between the two variants is highly significant, the difference between mutant and WT active site residues is less so. Isoleucine is also a better amino acid substrate for Clones 21 and 22 as it is for LeuAC.



Term	β	α	t Ratio	Prob>|t|
Intercept	1.128	0.181	6.25	<.0001*
KMSKS	0.064	0.088	0.73	0.4708
Clone 21	2.193	0.186	11.8	<.0001*
Clone 22	2.143	0.201	10.68	<.0001*
KMSKS*Clone 22	0.270	0.152	1.78	0.0824
Urzyme	2.129	0.187	11.37	<.0001*

Figure 2. Single turnover analysis of the recombinant deletion urzymes. **A.** Time course for ATP consumption by all four variants. **B.** Time course for ADP production. **C.** Time course for AMP production. **D.** Regression model for the dependence of the first-order rate constant, k_{chem}, on variant and mutant properties. **E.** Table of regression coefficients for the model in **D**, their estimated errors and P values. **F.** Histogram of the ratio of AMP to ADP produced by the four urzymes. WT Clone 21 has the highest ratio, hence uses ATP most efficiently. This ratio is consistent with the fact that the burst size of this variant in ADP production is lower than the others and the fact that the steady-state rate of AMP production is higher.

Single-turnover experiments yield two rate constants. The first-order rate constant is k_{chem}, referring to the depletion of ATP in the chemical step. The turnover rate is called k_3. The mean ratio k_{chem}/k_3 for Clones 21 and 22 (230) is similar to that for LeuAC (223). Thus, the pre-steady state bursts are comparable. All constructs retain the activated Ile-5'AMP with an increased affinity of ~ –3.2 kcal/mole.

The ratio of AMP to ADP produced in such assays measures the efficiency ATP usage to produce activated amino acid[30]. Interestingly, the WT Clone 21 uniquely makes produced less ADP (green dots in Fig 2B) and produced AMP at a higher steady-state rate (Fig. 2F). It is more efficient than the others.

The deletion variant catalytic proficiencies validate our studies of AARS urzymes.

Steady-state kinetics confirm these observations. The essence of translation in biology is forming a covalent bond between an amino acid and an RNA substrate. Our previous work with assignment catalysts derived from LeuRS showed three crucial details. (i) The preferred RNA substrate was the 35-nucleotide minihelix containing the acceptor stem and the TΨC-minihelix. (ii) The preferred amino acid substrate was isoleucine, not leucine. (iii) Catalysis of acyl transfer to RNA was minimally impacted by alanine substitution of the polar side chains in HVGA and KMSKS signatures. For these reasons, for the new double mutant (Dbl) constructs generated *in vivo,* we focus on isoleucine and minihelix substrates. Triplicate assays of acyl transfer are shown in Fig. 3 and summarized in Supplementary Table S2.

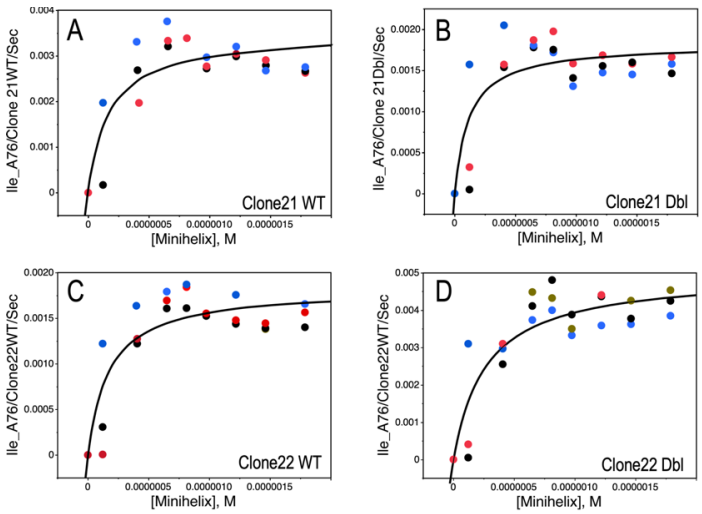


Figure 3. Michaelis-Menten plots of replicated aminoacylation assays. Aminoacylation of the leucine TΨC minihelix by Clone21, the Goldilocks urzyme with WT HVGH and KMSKS active site signatures (**A**) and the double mutant (**B**) sequences. **C, D.** Aminoacylation of the minihelix by Clone22 WT and DBL, respectively.

The curves in Fig. 3A, B, C suggest a decrease in observed rates at higher [minihelix]. The [substrate] necessary to detect dissociation between enzyme and minihelix RNA, thus the apparent dissociation constants for the enzyme•RNA complex, are an order of magnitude lower than the enzyme concentration. The RNA substrate is rapidly consumed, and the acylated RNA product may remain bound, further inhibiting the reaction by reducing [E_{free}].

The distinction between single-turnover and steady state assays becomes blurred because we study catalysts that represent increasingly ancestral forms. As turnover decreases, assays require longer times and higher enzyme concentrations to build an adequate signal-to-noise ratio. Unpublished time-course data also suggest that the observed rates are, at least partially, first-order processes. $\Delta G(k_{cat})$ values from aminoacylation correlate with a linear combination of $\Delta G(k_{chem})$ and $\Delta G(k_3)$ values from single turnover experiments ($R^2 \sim 0.97$; $P \sim 0.03$). Regression coefficients (see Supplemental Section D and Fig. S8) suggest that $\Delta G(k_{chem})$ contributes about 65% of the correlation and $\Delta G(k_3)$ 35%. We have not found any way around this. It is a cost of working close to the boundary where experiments are no longer possible.

The *in vivo* deletions (Fig.1) bear striking resemblance to the nested hierarchy of protozymes and urzymes we designed based on shared Class I AARS 3D structures. Those similarities and the catalytic proficiencies of Clone21 and Clone22 (Figs. 2,3) validate our previous work[7-10,13,18,36-38] in unexpected ways (Fig. S8). We had thought that the full protozyme and 2nd crossover segments were essential for acylation. Neither is actually required. Catalytic proficiencies of Clones 21, 22 are the same order of magnitude as those of the LeuAC double mutant (see K_{sp} values in Table S2).

mRNA structure predictions identify matching modularity in LeuRS mRNA.

We considered that 3D structures of the gene itself might promote recombination at close tertiary contacts. The new AlphaFold3[3] (AF3) has a rudimentary capacity to predict 3D RNA structures. Unfortunately, there are few, if any, related RNA structures in the training set, so results are fragmentary. Nonetheless, there are striking resonances with our hypothesis (Fig. 5).

AF3 predicted five very different folds for the mRNA. However, predicted structures for both coding regions of the Goldilocks urzyme, Clone21, are consistent in all predictions (Fig. 5B,C). One structure shared by all predictions is the elongated segment (bases 118-189; Fig. 5B) that codes for a 24 amino acid segment in the protozyme containing the HVGA/AVGA signature. That segment is part of all recombinant deletions. The protozyme-like and LeuAC-like deletions have exclusively that N-terminal fragment of the protozyme.

The second reproducible tertiary structure in all predictions is a stem-loop (bases 1905-1965; Fig. 5C). That stem-loop encodes residues 637-677 in the translated LeuRS. It coincides with the magenta β-strand that includes the KMSKS signature. This long stem-loop connects to the rest of the mRNA structure in a manner similar to the expansion signatures in ribosomal 23S RNA[39,40] (see Fig. 4 in[39]). We suggest on this basis that the deletions have reversed the evolutionary process by which the much shorter ancestral forms gained new domains by transposition-like processes and selected them because they enhanced their functions.

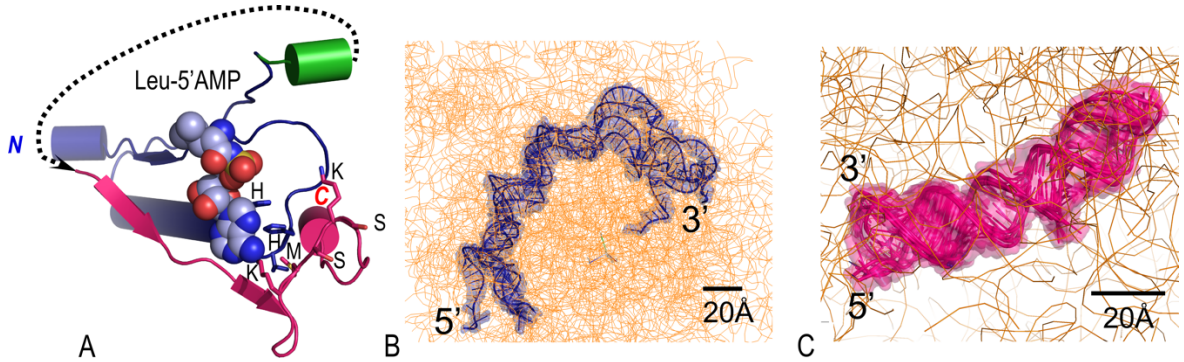


Figure 4. The recombinant deletions reflect the gene's tertiary structure. **A.** Segments of the "Goldilocks" urzyme from 1WZ2. The blue module, containing the HVGH/AVGA signature, is the 51-residue LeuRS protozyme. It is extended by four residues from the N-terminus of the CP1 insert (green). The magenta module is a 25-residue segment containing the KMSKS/AMSAS signature. The N and C- termini are labeled in blue and red. Side chains from the active-site signatures mutated here are shown as sticks and labeled with single-letter codes. **B.** Superposition of the coding region (bases 118-189) from five 3D structures predicted by AlphaFold3 [3] for the mRNA of full-length *P. horikoshii* LeuRS. The segment codes for amino acids 39-63 of the *P. horikoshii LeuRS,* the first half of the protozyme with the HVGH/AVGA signature. **C** `Superposition of bases 1905-1965 of the mRNA. That region codes for the entire magenta segment containing the KMSKS/AMSAS signature.

Class 1 and 2 AARS use similar secondary structures for aminoacylation.

Clone 21 is the smallest fully functional Class I AARS gene ever produced. Clones 21 and 22 share the AMSAS signature. They differ in both the protozyme and second crossover. The C-terminal fragment shared by both Clone22 and Clone21 is key to binding RNA substrates. It is strong evidence that acquisition of that signature converted Class I protozymes into aminoacylating enzymes. Curiously, the secondary structure of the segment is a β-strand followed by a long loop. The Motif 2 loop in Class 2 AARS also forms much of the RNA binding site. A striking resemblance of the RNA-binding secondary structures of Class 1 and 2 AARS is shown in Fig. 6. There are thus extraordinary similarities between the stereochemistry in both sets of polypeptide•RNA interactions leading to aminoacylation. We have suggested elsewhere how the contrasting substrate-binding interactions of Class I and II AARS can arise from the inversion symmetry of base-pairing in bidirectional ancestral genes[12,13,41-43].

The modest size of the Class I KMSKS-containing module is reminiscent of the N-terminal fragment of the protozyme, which also is ~25 residues long and has a similar β–>turn secondary structure and is key to the binding of ATP. In fact, the corresponding fragment in Class II AARS is actually the same fragment, residues 54-78 in Fig. 5. As we noted previously[15], neither the histidine side chains in the HVGH nor the lysine residues in the KMSKS signatures are necessary for amino acid activation or aminoacylation. The functionality of all urzyme Lys-to-Ala mutations implies that function arises solely from polypeptide secondary structure.

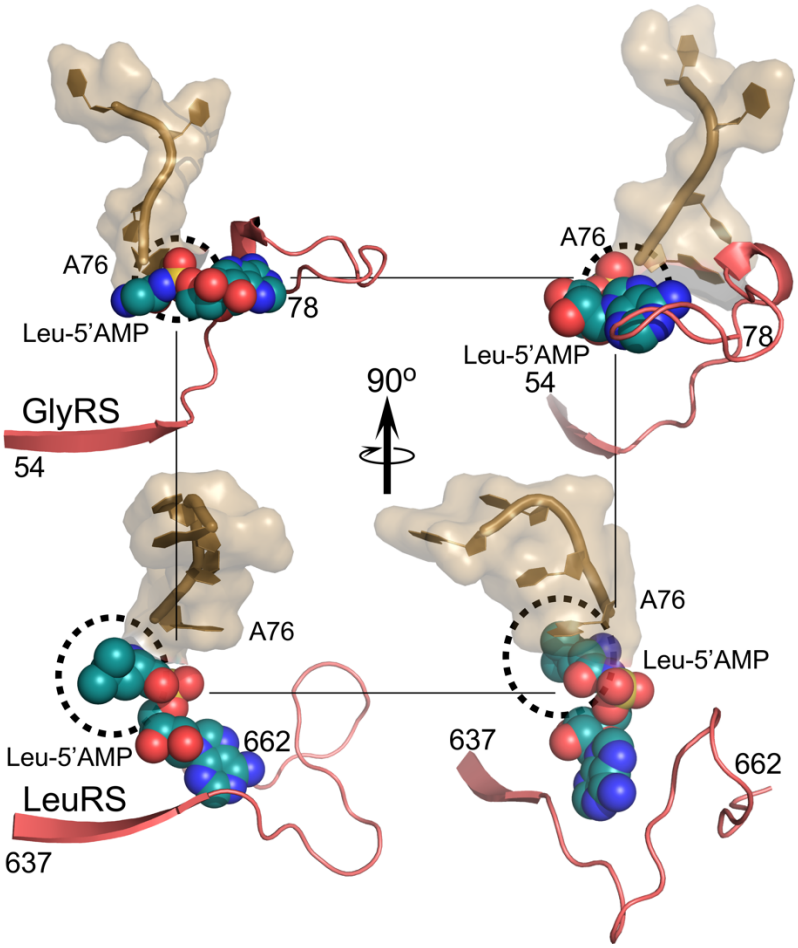


Figure 5. Corresponding segments of the active site that bind RNA substrates in Class 1 LeuRS (PDB ID 1WZ2) and Class II GlyRS (PDB ID 7YSE) have very similar secondary structures. Approximately orthogonal views of the complexes between RNA binding motifs in the two Classes with their respective activated amino acyl-5'AMP and RNA ligands. Dashed circles highlight the aminoacyl groups. The vertices of the square in the background indicate the 5'-phoshporyl group. They are located remarkably similarly with respect to the β-strand and loop. A chlorine atom that is part of the adenine ring in 7YSE has been omitted.

The Goldilocks urzyme, Clone 21, shows how one gene can encode both Class I, II AARS.

The Rodin-Ohno hypothesis that bidirectional ancestral AARS genes coded for Class I and II urzymes on opposite strands has led to a surprising number of retrodictions about the Class distinction[12]. The bidirectional protozyme gene was easy to construct[10] and others have verified its activation activity[11]. We have failed, however, to design a bidirectional gene encoding Class I and II AARS urzymes on opposite strands. The Goldilocks urzyme, Clone21, may show how to design such a gene. Clone 21 is the same length (81 residues) as the Class I GlyCA urzyme[30]. Thus, 81-residue excerpts of both Class I and II AARS are both catalytically active AARS. Moreover, their Class-defining signature sequences both align opposite one another in antiparallel alignments[12].

A preliminary study revealed likely pairing of homologous segments of three sets of Class I AARS Goldilocks-length with Class II AARS urzyme genes. Two of the three pairings match Class IA LeuRS and IIB AspRS; Class IIA ProRS with Class IB GlnRS and Class I TrpRS with Class II HisRS (Fig. 8). This pairing conflicts with the conventional sub-classification. Each such pairing brings along two additional AARS from the same respective subclasses. Thus, 16 of the 20 bacterial AARS sequence alignments show analogous pairings between Class I and II AARS genes[12].

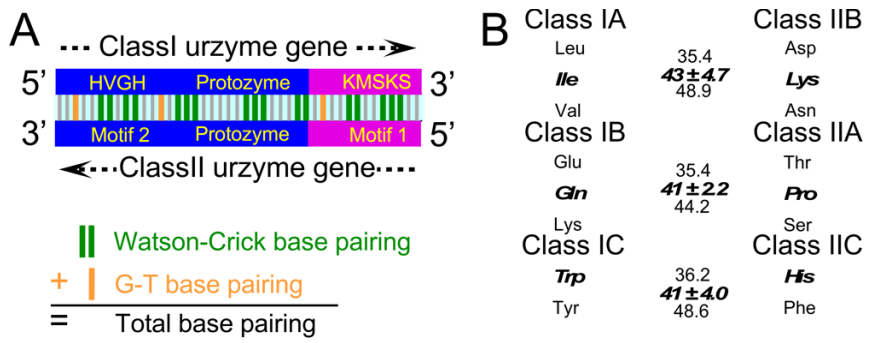


Figure 6. The Goldilocks urzyme, Clone21, suggests a paradigm for designing bidirectional urzyme genes for Class I and II ancestors on opposing DNA strands. **A.** Schematic of sense/antisense encoding of Class I and II AARS urzyme genes, consistent with the architecture of the two AARS Classes. Vertical lines between the two genes suggest bases that would be paired in a bidirectional gene but are only partially paired in sequences derived from contemporary genes. **B.** Frequencies of codon middle-base pairing in antiparallel alignments, calculated as indicated in **A** are summarized for three groups derived from the three consensus synthetase subclasses[12]. Values greater than 40% are highly unlikely under the null hypothesis.

Modular growth of AARS superfamilies tolerated unexpected structural variation.

The high conservation of the two Class I catalytic signatures was strong evidence they were essential for catalysis. That conclusion now seems to be false. Class I AARS urzyme catalysis does not depend either on histidine or lysine side chains. It seems to result instead from as yet unidentified secondary structural elements[44]. Supplementary Section E and Fig. S8 underscore that these elements also can be connected three quite different ways. Yet, each is about equally functional. We also provided evidence that Class I AARS urzymes are catalytically active molten globules[49]. The surprising tolerance, not only of sequence, but also of tertiary structural variation suggests that genetic coding and decoding arose from a very broad manifold of possibilities.

Clones 21 and 22 enhance studies of the evolution of protein modularity at unrivalled resolution.

How the ancestral AARS grew substantially to their contemporary length is one of the chief puzzles about the evolution of genetic coding. Regression modeling described more fully in Supplementary Section F and Fig. S9 shows that the second half of the protozyme and the 2nd crossover of the Rossmann fold have significant effects on the free energy for k_{cat}, the first-order process that takes place in the active site to transfer the acyl group to the cognate minihelix. Such effects measure selectable properties necessary for understanding the modular growth of ancient enzymes. They offer an experimental platform to assess why such modules were selected.

Modular growth of AARS superfamilies parallels that of the genetic code

Translation—readout of genes into proteins—is one of the earliest hallmarks of life. The genetic coding table is vital to biology because it makes sense of genes. The table stores information, the physical chemistry of amino acids, for a purpose, namely making functional proteins. Translation expands the capacities of genes by perhaps a billion-fold (See supplement for [45]). Proteins, in turn, enable genetics, metabolism, and free-energy transfer — all requisites for inheritance and to sustain life far from equilibrium.

The origins of coding have been obscure, but our experiments are changing that. A plausible and supported narrative is now within reach[12,13,46]. The growth of the two synthetase superfamilies parallels the expansion and refinement of the genetic code[14,47,48]. We add a new element to that narrative by suggesting that the protein modules also reflect a similar modularity in the genes themselves. That new window arises *in vivo* and so lies beyond the reach of current tools. The time-course of *in vivo* studies is thus now a key new goal.

Double mutations may generate similar nested deletions in a wider range of protein genes.

Analytical structure-based deconstruction took us ten years. Here, we found that uncommonly similar specific deletions can be created *in vivo.* Can similar deletions be observed for other Class I AARS genes? With Class II AARS genes? With other protein superfamilies? If the answers to any of these questions is yes, these results may provide a method to generate active miniature versions of full-length proteins from other classes. Deleting long sequences *in vivo* with other systems would extend the range of possible constructs. Protein engineering[51-54] can further enhance function[55].

Engineered enzyme and regulatory protein cores that retain function would also have widespread scientific and practical benefit. Miniaturized enzymes and regulatory proteins have practical use in research investigations into the evolution of the proteome, including the eventual search for a minimal gene set. Retention of the functionality and reducing the size of large and complex proteins may also enable packaging in viral and other delivery vectors and so have potential applications in gene therapy.

Methods and Materials

Materials

The plasmid vector was inserted into pMal-c2x (Plasmid #75286. New England Biolabs, Ipswich, MA.). DNA oligos were ordered from IDT (Integrated DNA Technologies, Coralville, IA, United States). The detailed sequences of oligo DNAs used for mutagenesis (*a1/b1*) and for sequencing (*seqn1/seqn2*) are provided in the Table 2, together with the sequence of full-length LeuRS gene. Phusion™ Plus PCR Master Mix (Cat# F631S) was purchased from Thermo Fisher Scientific (Waltham, MA, United States). *E coli* competent cells were partially from Agilent (XL10-Gold ultracompetent cells, Cat# 200315, Santa Clara, CA, United States), and partially using DH5a, home-made, prepared according to the method described by Sharma in 2017. Restrict enzyme DpnI (Cat# 500402, United States) was purchased from Agilent. Purifications and handling of DNA fragment and plasmid used the QIAquick Gel Extraction Kit (Cat# 28704, Qiagen, Hilden, Germany) and QIAprep Spin Miniprep Kit (Cat# 27104, Qiagen, Hilden, Germany) according to instruction, unless stated otherwise.

Table 1. Summary of DNA template sequences of *P. horikoshii* Ph-FL-LeuRS and Oligos for the study. Sequences encoding WT LeuAC urzyme are highlighted in yellow.

Oligo Identity	Sequence
Full-length Leucyl-tRNA synthetase	catatgcatcaccatcaccatcacgcggaactgaactttaaagcgattgaagaaaaa tggcagaaacgctggctggaagcgaaaatttttgaaccgaacattcgcgataaaccg aaagaaaaaaaatttatattaccgtggcgtttccgtatctgagcggccatctgcat gtgggccatgcgcgcacctataccattccggatgtgattgcgcgctttaaacgcatg cagggctataacgtgctgtttccgatggcgtggcatattaccggcagcccgattgtg ggcattgcggaacgcattaaaaaccgcgatccgaaaaccatttggatttatcgcgat gtgtataaagtgccggaagaaattctgtggacctttgaagatccgattaacattgtg aaatattttatgaaagcggcgaaagaaacctttattcgcgcgggctttagcgtggat tggagccgcgaattttataccaccagcctgtttccgccgtttagcaaatttattgaa tggcagttttggaaactgaaagaaaaaggctatattgtgaaaggcgcgcatcgcgtg cgctgggatccggtggtgggcaccccgctgggcgatcatgatctgatggaaggcgaa gatgtgccgattctggattatattattattaaatttgaactgcgcgaaaacggcgaa

	gtgatttatctgccggcggcgaccctgcgcccggaaaccgtgtatggcgtgaccaac atgtgggtgaacccgaacgcgacctatgtgaaagcgaaagtgcgccgcaaagataaa gaagaaacctggattgtgagcaaagaagcggcgtataaactgagctttcaggatcgc gaaattgaagtgattgaagaatttaaaggcgaaaaactgattggcaaatatgtgcgc aacccggtgagcggcgatgaagtgattattctgccggcggaatttgtggatccggat aacgcgaccggcgtggtgatgagcgtgccggcgcatgcgccgtttgatcatgtggcg ctggaagatctgaaacgcgaaaccgaaattctggaaaaatatgatattgatccgcgc attgtggaaaacattacctatattagcctgattaaactggaaggctatggcgatttt ccggcggtggaagaagtgaacaaactgggcattaaaagccagaaagataaagaaaaa ctggaacaggcgaccaaaaccatttataaagcggaatatcataaaggcatttttaaa gtgccgccgtatgaaggcaaaccggtgcaggaagtgaaagaagcgattgcgaaagaa atgctggaaaaaggcattgcggaaattatgtatgaatttgcggaaaaaaacgtgatt agccgctttggcaaccgcgcggtgattaaaattattcatgatcagtggtttattgat tatggcaacccggaatggaaagaaaaagcgcgcaaagcgctggaacgcatgaaaatt ctgccggaaacccgccgcgcgcagtttgaagcgattattgattggctggataaaaaa gcgtgcgcgcgcaaaattggcctgggcaccccgctgccgtgggatccggaatgggtg attgaaagcctgagcgatagcaccatttatatggcgtattataccattagccgccat attaacaaactgcgccaggaaggcaaactggatccggaaaaactgaccccggaattt tttgattatatttttctggaagaatttagcgaagataaagaaaaagaactggaaaaa aaaaccggcattccggcggaaattattcatgaaatgaaagaagaatttgaatattgg tatccgctggattggcgctgcagcggcaaagatctgattccgaaccatctgacctttt tttatttttaaccatgtggcgatttttcgcgaagaacattggccgaaaggcattgcg gtgaacggctttggcaccctggaaggccagaaaatgagcaaaagcaaaggcaacgtg ctgaactttattgatgcgattgaagaaaacggcgcggatgtggtgcgcctgtatatt atgagcctggcggaacatgatagcgattttgattggcgccgcaaagaagtgggcaaa ctgcgcaaacagattgaacgcttttatgaactgattagccagtttgcggaatatgaa gtgaaaggcaacgtggaactgaaagatattgatcgctggatgctgcatcgcctgaac aaagcgattaaagaaaccaccaacgcgctggaagaatttcgcacccgcaccgcggtg cagtgggcgttttatagcattatgaacgatctgcgctggtatctgcgccgcaccgaa ggccgcgatgatgaagcgaaacgctatgtgctgcgcaccctggcggatgtgtgggtg cgcctgatggcgccgtttaccccgcatatttgcgaagaactgtgggaaaaactgggc ggcgaaggctttgtgagcctggcgaaatggccggaaccggtggaagaatggtggaac gaaaccattgaagcggaagaagaatttattcgcagcgtgatggaagatattaaagaa attattgaagtggcgaaaattgaaaacgcgaaacgcgcgtatatttataccgcggaa gattggaaatggaaagtggcggaagtggtgagcgaaaaacgcgattttaaaagcagc atggaagaactgatgaaagatagcgaaattcgcaaacatggcaaagaagtggcgaaa attgtgcagaaactgattaaagaacgcacctttgatgtgaaacgcattaacgaagaa aaagcgctgcgcgaagcgaaagaatttatggaaaaagaactgggcattgaaattatt attaacccgaccgaagataaaggcggcaaaaaaaaacaggcgatgccgctgaaaccg gcgatttttattgaatgataagctgagc
LeuRS_AVGAa1_F (Forward mutagenesis primer)	cgtttccgtatctgagcggccatctgGCGgtgggcGCGgcgcgcacctataccattc cgg
LeuRS_AMSASb1_R (Reverse mutagenesis primer)	caataaagttcagcacgttgcctttgctTGCgctcatCGCctggccttccagggtgc caa

seqn1_AVGA (Forward Sanger sequencing primer)	atgcatcaccatcaccatcacgcggaactg
AMSAS_seqn2rv (Reverse Sanger sequencing primer)	tcatgttccgccaggctcataatatacag

314 **Methods**

315 *Method for preparation of double mutant plasmid.* DNA oligo sequences were designed according to the
316 manual of QuikChange II-E Site-Directed Mutagenesis Kit (Santa Clara, CA, United States) as published
317 online at https://www.agilent.com/cs/library/usermanuals/public/200555.pdf. In line with the oligo DNA
318 design, and for creating the double mutant plasmid DNA, we have implemented a megaprimer PCR method
319 based on that described by Picard[24] (Fig. 9).

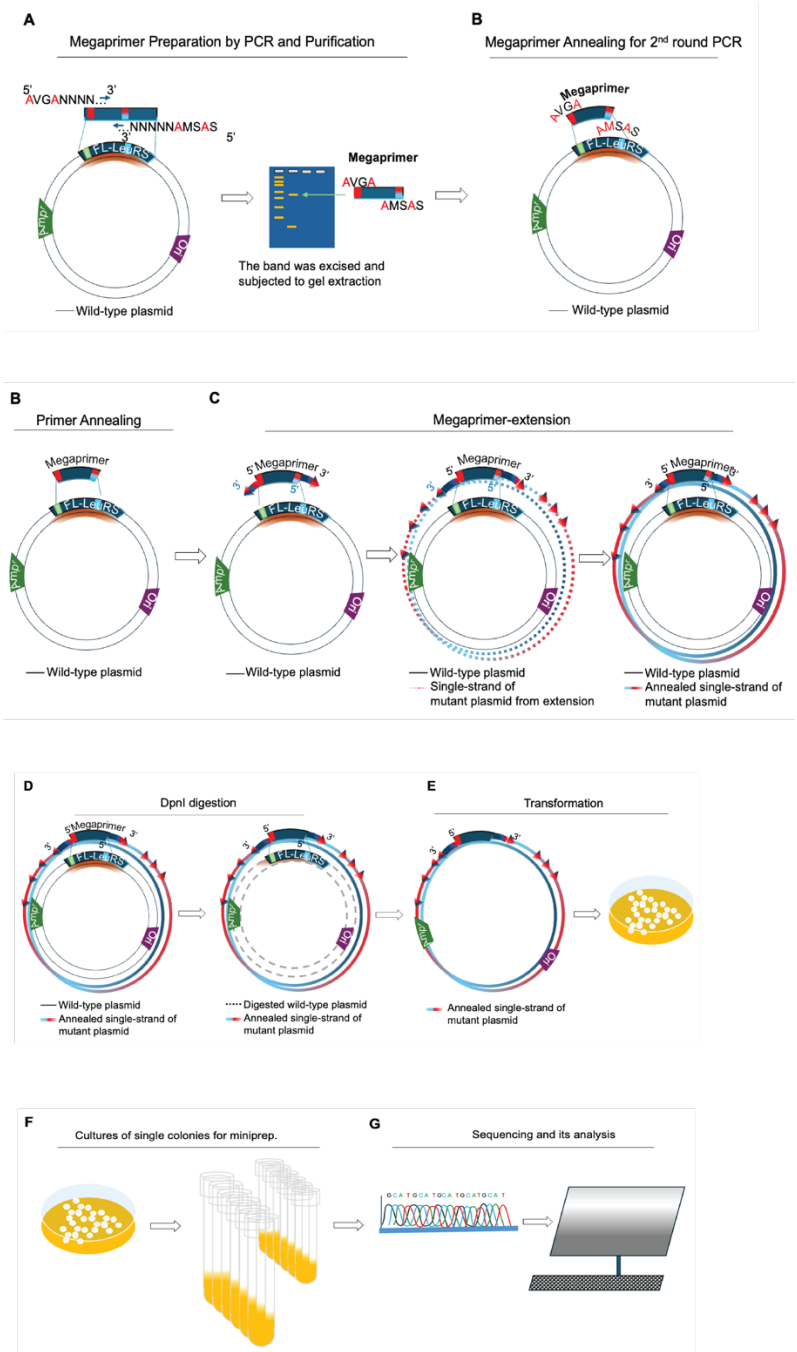


Figure 8. A flowchart illustration of megaprimer-PCR. The method employs a combination of the "Megaprimer-PCR" technique [24] with conventional mutagenesis products (e.g., Agilent, Santa Clara, CA[56]). **A**. Following a first round of PCR (left) with releasing a megaprimer from a target wild-type plasmid DNA, an agarose gel and the PCR product (namely, megaprimer) purification was carried out, in which the propriate mutations AVGA and AMSAS were installed as indicated in bended color squares and amino acid sequences. **B.** For the second round of PCR, annealing the newly purified and resultant megaprimer onto a target plasmid DNA. **C.** PCR extension of the second round of PCR is indicated. The desired PCR extension is initiated and proceeded using the sense-single strand and antisense-single strand of the megaprimer as indicated as colored dot circular lines. **D.** Restriction enzyme DpnI digestion of the second round PCR product, the destroy of original intact wild-type plasmid DNA is indicated as dash lines in grey. **E.** After DpnI digestion of the second round PCR mixture, single colonies of transformants on agar plates can be obtained via conventional transformation with DpnI-digested PCR mixture into competent *E. coli* cells. **F.** Following transformation, a variety of mutant plasmids can be identified through conventional plasmid miniprep. **E.** sequencing, sequence analysis and its identification.

333 First, megaprimer was prepared in a 1st round of PCR. Phusion™ Plus PCR Master Mix (Cat# F631S) was
334 employed according to instruction from manufacture (ThermoFisher Scientific, Waltham, MA, United
335 States). The primer pairs for the 1st round PCR was LeuRS_AVGAa1_F and LeuRS=AMSASb1_R as listed
336 in Table 1 above. In the PCR reaction mix, the plasmid template concentration was 1-4 ng/μL. The primers
337 were used at a concentration of 0.66 μM. The PCR was started at 98 C, 30 sec for initial template
338 denaturation and was following by 44 cycles (98 ºC, 12 sec; 61 ºC, 30 sec; 72 ºC, 3 min 20 sec). After the
339 final extension at 72 ºC, 5 min, the PCR reactions were kept at 4 ºC overnight. After the PCR run, following
340 by horizontal electrophoresis in a 1xTAE, 1.2% Agarose-gel, and visualization by UV transmitter, the
341 megaprimer fragment from the 1st round of PCR was collected from gel cutting out the desired band using
342 a clean razor blade. After DNA fragment purification procedures using QIAquick Gel Extraction Kit (Cat#
343 28704, Hilden, Germany), to yield the megaprimer for the downstream operations.

344 In the second phase, the entire plasmid DNA was amplified using the megaprimer to generate the resultant
345 plasmid harboring the mutant. A 2nd round of PCR was performed using Phusion™ Plus PCR Master Mix
346 (Cat#F631S) and the parameters were setup according to the instructions from the manufacture
347 (ThermoFisher Scientific, Waltham, MA). The PCR procedures were the same as the 1st PCR run. Very
348 importantly, instead of oligo primers, the megaprimer was added at specific desired concentration.
349 Basically, in a 25-40 μL PCR reaction volume, the amount of megaprimer fragment was within a range of
350 0.16μM-1.25 μM, and the template plasmid DNA was 0.5-1.5 ng.

351 After DpnI digestion of the PCR mixture at 37ºC from hours to overnight according to instruction provided
352 by Agilent (Santa Clara, CA, United States), the DpnI-digested PCR mixture is ready for *E. coli.*
353 transformation. The transformation was done using 0.3 μL DpnI-digested PCR mixture per transformation.

354 *Method for E. coli. transformation and the sample submission for sequencing. E coli* transformation was
355 proceeded according to conventional procedure [57] and spread onto LB (Lysogeny broth) agar plates with
356 antibiotics (Ampicillin 50 μg/mL, or Carbenicillin 100 μg/mL). After incubation overnight at 37 ºC,
357 individual colonies on the LB agar plate were readied for miniprep by following the standard protocol
358 described by QIAprep Spin Miniprep Kit (Cat# 27104, Qiagen, Hilden, Germany)

359 The resultant plasmid DNA samples were subjected to sequencing by following standard sequencing sample
360 preparation and submission described by Eton Bioscience (San Diego, CA) online
361 https://www.genewiz.com/en/Public/Resources/Sample-Submission-Guidelines/Sanger-Sequencing-
362 Sample-Submission-Guidelines/Sample-Preparation#sanger-sequence with minor modifications. Basically,
363 each sequencing sample was in 15 mL, containing 100-200 ng plasmid DNA with 5-10 picomole
364 sequencing primer, i.e. seqn1_AVGA or AMSAS_seqn2rv as detailed sequences in Table 1 above.

365 *Method for plasmid sequencing and sequencing analysis.* Sequence analysis was performed using
366 conventional methods as follows. First, all the DNA sequences were translated into 6 reading frames in
367 amino acid sequence using Expasy translate online at https://web.expasy.org/translate/. For screening the
368 tentative urzyme candidates, each amino acid sequence from 6 reading frames from a single sequenced
369 plasmid DNA sample were aligned with the sequence of urzyme template using conventional sequencing
370 alignment software such as MAFFT online at https://www.ebi.ac.uk/jdispatcher/msa/mafft?stype=protein
371 [58] or MUSCLE online at https://www.ebi.ac.uk/jdispatcher/msa/muscle?stype=protein[59]. The list of criteria
372 for selection of tentative candidate was according to the sequence length, similarity and identify in addition
373 to the highly conserved catalytic motifs such as motif I HVGH and motif II KMSKS or their-like.

374 The resulting new candidates were designated protozyme, urzyme-like or Δ2ndXvr (for deleted second
375 crossover) and goldilocks, according to their sequence content and were collected and subjected for further
376 downstream identification at the enzymology level after protein biochemistry procedures.

377 *2.2 $tRNA^{Leu}$ and TΨC-$minihelix^{Leu}$ preparation*

A plasmid encoding the *P. horikoshii* $tRNA^{Leu}$ (UAG anticodon) was synthesized by Integrated DNA Technologies and used as template for PCR amplification of the tRNA and upstream T7 promoter and downstream Hepatitis Delta Virus (HDV) ribozyme. The PCR product was used directly as template for T7 transcription. Following a 4-hour transcription at 37°C the RNA was cycled five times (90°C for 1 min, 60°C for 2 min, 25°C for 2 min) to increase the cleavage by HDV. The tRNA was purified by urea PAGE and crush and soak extraction. The tRNA 2'-3' cyclic phosphate was removed by treatment with T4 PNK (New England Biolabs) following the manufacturer's protocol. The tRNA was then phenol chloroform isoamyl alcohol extracted, filter concentrated, aliquoted, and stored at -20°C.

The original sequence information for composing and designing *Pyrococcus horikoshii* TΨC-minihelix-Leu was based on the mature *P. horikoshii* tRNA-Leu, the tRNA sequence was collected from the complement strand of the genomic sequence of *P. horikoshii* (OT3, GB# BA000001.2, between nucleotides 1448081 and 1448168). The designated minihelix sequence was obtained by combining the acceptor stem sequence including DCCA with the TΨC -stem-loop according to Schimmel[60]. A plasmid harboring the minihelix, an upstream T7 promoter, and downstream Hepatitis Delta Virus (HDV) ribozyme was acquired from previous co-worker Jessica Elder. To reduce challenges posed by the high GC content in the stem portion of minihelix during the preparation, we implemented the Phi29 DNA polymerase-mediated isothermal amplification approach according to the manufacture's protocol (New England Biolabs, Ipswich, MA). After the purification of the product from isothermal amplification, T7 transcription was carried out. Following a 4–6-hour transcription at 37 ˚C, the reaction mixture was subjected to urea PAGE fractionation, crush, and soak extraction. After 2'-3' cyclic phosphate removal by T4 PNK, the minihelix RNA was then phenol chloroform isoamyl alcohol extracted, filter-concentrated, quantitated, and aliquoted, and stored at −80 ˚C.

Expression and purification of variant recombinant deletions

We expressed all constructs as MBP fusions from pMAL-c2x in BL21Star (DE3) (Invitrogen) as described in[27]. The lysis buffer contained 20 mM Tris, pH 7.4, 1 mM EDTA, 5 mM β-mercatoethanol (βME), 17.5% Glycerol, 0.1% NP40, 33 mM $(NH_4)_2SO_4$, 1.25% Glycine, 300 mM Guanidine Hydrochloride) plus cOmplete protease inhibitor (Roche). After elution from Amylose FF resin (Cytiva) with 10 mM maltose in lysis buffer (200 mM HEPES, pH 7.4, 450 mM NaCl, 100 mM KCl, 10 mM β-ME. Fractions containing protein were concentrated and mixed to 50% glycerol and stored at -20°C. All protein concentrations were determined using the Pierce™ Detergent-Compatible Bradford Assay Kit (Thermo Scientific). We determined purity by running samples on PROTEAN® TGX (Bio-RAD) gels and active fractions for all variants were measured as described in the next section.

Single turnover active-site titration assays

Active-site titration assays were performed as described [33,34] with the exception that ^{32}P-ATP was labeled in the α position in order to follow the time-dependence of all three adenine nucleotides. of protein was added to 3 μM in 1x reaction mix (50 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 5 μM ATP, 50 mM amino acid, 1 mM DTT, inorganic pyrophosphatase, and α-labeled [^{32}P] ATP) to start the reaction. Timepoints were quenched in 0.4 M sodium acetate 0.1% SDS and kept on ice until all points had been collected. Quenched samples were spotted on TLC plates, developed in 850 mM Tris, pH 8.0, dried and then exposed for varying amounts of time to a phosphor image screen and visualized with a Typhoon Scanner (Cytiva). The time-dependent of loss (ATP) or de novo appearance (ADP, AMP) of the three adenine nucleotide phosphates were estimated using the ImageJ measure function. Time-dependence was fitted using the nonlinear regression module of JMP16PRO™ to equation (1):

$$\text{Product(calc)} = A*\exp(-k_{chem}*\text{seconds}) - k_3*\text{seconds} + C \quad (1)$$

where k_{chem} is the first-order rate constant, k_3 is the rate of turnover, A is the amplitude of the first-order process, and C is an offset.

For [ATP] decay curves, the fitted A value estimates the burst size, n, directly as n = A*[ATP]/[Enzyme]. C gives n = (1–C) * [ATP]/[Enzyme]. For exponentially increasing concentrations of product, the situation reverses, and n = (1–A) * [ATP]/[Enzyme]and n = C*[ATP]/[Enzyme. These approximations are justified by the very small variance of multiple estimates.

Aminoacylation assays

We determined the active fraction of minihelix RNA by following extended acylation assays using the AVGA LeuAC mutant until they reached a plateau. That plateau value was used to compute minihelixconcentrations in assays with all variants.

Aminoacylations withs minihelix were performed as described [27] in 50 mM HEPES, pH 7.5, 10 mM $MgCl_2$, 20 mM KCl, 5 mM DTT with indicated amounts of ATP and isoleucine. The high affinity of Clones 21 and 22 for minihelix meant that we had to mix increased amounts of [α^{32}P] A76-labeled tRNA for assays. The RNA substrates were heated in 30 mM HEPES, pH 7.5, 30 mM KCl to 90°C for 2 minutes and then cooled linearly (drop 1°C/30 seconds) until it reached 80°C when $MgCl_2$ was added to a final concentration of 10 mM. The tRNA continued to cool linearly until it reached 20°C. Michaelis-Menten experiments were performed by repeating these assays at the indicated RNA concentrations. Concentration-dependence fitted to $K_{sp} = k_{cat}/K_M$ and kcat according to the modified formula (Eqn. 2) introduced by Johnson[61].

$$Rate_{calc} = \frac{\frac{k_{cat}}{K_M} \times [RNA,M]}{\left(1 + \frac{k_{cat}}{K_M} * \frac{[RNA,M]}{k_{cat}}\right)} \qquad (2)$$

Data processing and Statistical analysis

Phosphor imaging screens of TLC plates were densitometered using ImageJ. Data were transferred to JMP16PRO™ Pro 16 via Microsoft Excel (version 16.49), after intermediate calculations. We fitted both active-site titration curves and Michaelis-Menten assays using the JMP16PRO™ nonlinear fitting module.

Factorial design matrices (e.g. Table 1) were processed using the Fit Model multiple regression analysis module of JMP16PRO™ Pro, using an appropriate form of equation (3) [62]ΔG‡

$$Y_{obs} = \beta_0 + \Sigma\beta_i * P_i + \Sigma\beta_{ij} * P_i * P_j + \varepsilon \qquad (3)$$

where Y_{obs} is a dependent variable, usually an experimental observation, β_0 is a constant derived from the average value of Y_{obs}, β_i and β_{ij} are coefficients to be fitted, $P_{i,j}$ are independent predictor variables from the design matrix, and ε is a residual to be minimized. All rates and apparent affinities were converted to free energies of activation or binding, ΔG = – RTln(k), before regression analysis. Free energies are additive, whereas rates and apparent affinities are multiplicative. For example, the activation free energy for the first-order decay rate in single-turnover experiments is ΔG‡(kc_{hem}).

Multiple regression analyses of factorial designs exploit the replication inherent in the full collection of experiments to estimate experimental variances. In most cases we estimated errors on the basis of t-test P-values. In some cases, we added error bars to histograms to show the variance of individual datapoints. Multiple regression analyses reported here also entail triplet experimental replicates, which enhance the associated analysis of variance.

Funding

This work was supported by the Matter-to-Life program of the Alfred P Sloan Foundation, Grant G-2021-16944.

References

1 Moody, E. R. *et al.* The nature of the last universal common ancestor and its impact on the early Earth system. *Nature Evolution and Ecology* **8**, 1654–1666 (2024). https://doi.org/10.1038/s41559-024-02461-1

2 Ball, P. in *The Guardian* (Guardian News & Media Limited, London, UK, 2025).

3 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **https://doi.org/10.1038/s41586-024-07487-w** (2024).

4 Petrov, A. S. & Williams, L. D. The Ancient Heart of the Ribosomal Large Subunit: A Response to Caetano-Anolles. *J Mol Evol* **80**, 166-170 (2015). https://doi.org/10.1007/s00239-015-9678-8

5 Fournier, G. P., Andam, C. P., Alm, E. J. & Gogarten, J. P. Molecular Evolution of Aminoacyl tRNA Synthetase Proteins in the Early History of Life. *Orig Life Evol Biosph* **41** 621–632 (2011).

6 Fournier, G. P. & Alm, E. J. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. *J. Mol. Evol.* **80**, 171-185 (2015). https://doi.org/10.1007/s00239-015-9672-1

7 Carter, C. W., Jr. Urzymology: Experimental Access to a Key Transition in the Appearance of Enzymes. *J. Biol. Chem.* **289**, 30213–30220 (2014). https://doi.org/10.1047/jbcR114.576495

8 Li, L., Weinreb, V., Francklyn, C. & Carter, C. W., Jr. Histidyl-tRNA Synthetase Urzymes: Class I and II Aminoacyl-tRNA Synthetase Urzymes have Comparable Catalytic Activities for Cognate Amino Acid Activation. *J. Biol. Chem.* **286**, 10387-10395 (2011). https://doi.org/10.1074/jbc.M110.198929

9 Pham, Y. *et al.* Tryptophanyl-tRNA synthetase Urzyme: a model to recapitulate molecular evolution and investigate intramolecular complementation. *J. Biol. Chem.* **285**, 38590-38601 (2010). https://doi.org/10.1074/jbc.M110.136911

10 Martinez-Rodriguez, L. *et al.* Functional Class I and II Amino Acid Activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. *J. Biol. Chem.* **290**, 19710–19725 (2015). https://doi.org/10.1074/jbc.M115.642876

11 Onodera, K. *et al.* Amino acid activation analysis of primitive aminoacyl-tRNA synthetases encoded by both strands of a single gene using the malachite green assay. *BioSystems* **208**, 104481 (2021).

12 Carter, C. W., Jr *et al.* Structural Enzymology, Phylogenetics, Differentiation, and Symbolic Reflexivity at the Dawn of Biology. *Genome Biology and Evolution* **In review** (2025).

13 Carter, C. W., Jr. & Wills, P. R. The Roots of Genetic Coding in Aminoacyl-tRNA Synthetase Duality *Annual Review of Biochemistry* **90**, 349-373 (2021). https://doi.org/10.1146/annurev-biochem-071620-021218

14 Douglas, J., Bouckaert, R., Carter, C. W. J. & Wills, P. Enzymic recognition of amino acids drove the evolution of primordial genetic codes. *Nucleic Acids Research* **52**, 558–571 (2024).

15 Tang, G. Q., Elder, J. J. H., Douglas, J. & Carter, C. W., Jr, . Domain Acquisition by Class I Aminoacyl-tRNA Synthetase Urzymes Coordinated the Catalytic Functions of HVGH and KMSKS Motifs. *Nucleic Acids Research* **51**, 8070–8084 (2023). https://doi.org/10.1093/nar/gkad590

16 Carter, C. W., Jr., Popinga, A., Bouckaert, R. & Wills, P. R. Multidimensional Phylogenetic Metrics Identify Class I Aminoacyl-tRNA Synthetase Evolutionary Mosaicity and Inter-modular

Coupling. *International Journal of Molecular Sciences* **23**, 1520 (2022). https://doi.org/10.3390/ijms23031520

17 Weinreb, V. *et al.* Enhanced Amino Acid Selection in Fully-Evolved Tryptophanyl-tRNA Synthetase, Relative to its Urzyme, Requires Domain Movement Sensed by the D1 Switch, a Remote, Dynamic Packing Motif *J. Biol. Chem.* **289**, 4367-4376 (2014). https://doi.org/10.1074/jbc.M113.538660

18 Li, L. & Carter, C. W., Jr. Full Implementation of the Genetic Code by Tryptophanyl-tRNA Synthetase Requires Intermodular Coupling. *J. Biol. Chem.* **288**, 34736–34745 (2013). https://doi.org/10.1074/jbc.M113.510958

19 Green, M. & Sambrook, J. *Molecular Cloning: A Laboratory Manual*. Vol. 1 256 (Cold Spring Harbor Laboratory Press, 2012).

20 Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia On Quantitative Biology* **51**, 263-273 (1986).

21 Varani, G. Twenty years of RNA: the discovery of microRNAs. *RNA* **21**, 751–752 (2015). https://doi.org/10.1261/rna.050237.115

22 Anderson, J. C. & Schultz, P. G. Adaptation of an Orthogonal Archaeal Leucyl-tRNA and Synthetase Pair for Four-base, Amber, and Opal Suppression. *Biochem.* **42**, 9598-9608 (2003).

23 Tomoiaga, D., Bubnell, J., Herndon, L. & Feinstein, P. High rates of plasmid cotransformation in E. coli overturn the clonality myth and reveal colony development. (2921).

24 Picard, V., Ersdal-Badju, E., Lu, A. & Bock, S. C. A rapid and efficient one-tube PCR-based mutagenesis technique using Pfu DNA polymerase. *Nucleic Acids Res.* **22**, 2587-2591 (1994).

25 Carter, C. W., Jr. *et al.* The Rodin-Ohno Hypothesis That Two Enzyme Superfamilies Descended from One Ancestral Gene: An Unlikely Scenario for the Origins of Translation That Will Not Be Dismissed. *Biology Direct* **9**, 11 (2014).

26 Hobson, J. J., Li, Z. & Carter, C. W., Jr. A leucyl-tRNA synthetase urzyme: authenticity of tRNA Synthetase urzyme catalytic activities and production of a non-canonical product. *International Journal of Molecular Sciences* **23**, 4229 (2022).

27 Tang, G. Q., Hu, H., Douglas, J. & Carter, C. W., Jr. Primordial aminoacyl-tRNA synthetases preferred tRNA minihelix substrates over full-length tRNA. *Nucleic acids Research* **52**, 1-24 (2024). https://doi.org/10.1093/nar/gkae417

28 Richardson, J. S. Handedness of Crossover Connections in B Sheets. *Proc. Nat. Acad. Sci. USA* **73**, 2619-1623 (1973).

29 Buehner, M., Ford, G. C., Moras, D., Olsen, K. W. & Rossmann, M. G. D-Glyceraldehyde 3-Phosphate Dehydrogenase: Three Dimensional Structure and Evolutionary Significance. *Proc. Nat. Acad. Sci. USA* **70**, 3052-3054 (1973).

30 Patra, S. K. *et al.* A genomic database furnishes minimal functional glycyl-tRNA synthetases homologous to other, designed class II urzymes. *Nucleic Acids Research* **52**, 13305–13324 (2024). https://doi.org/https://doi.org/10.1093/nar/gkae992

31 Ivancic-Bace, I., Salaj-Smic, E. & Brcic-Kostic, K. Effects of recJ, recQ, and recFOR Mutations on Recombination in Nuclease-Deficient recB recD Double Mutants of Escherichia coli. *J. Bact.* **187**, 1350–1356 (2005). https://doi.org/10.1128/JB.187.4.1350–1356.2005

32 Burbaum, J. & Schimmel, P. Structural Relationships and the Classification of Aminoacyl-tRNA Synthetases. *The Journal of Biological Chemistry* **266**, 16965-16968 (1991).

33 Fersht, A. R. *et al.* Active Site Titration and Aminoacyl Adenylate Binding Stoichiometry of Amionacyl-tRNA Synthetases. *Biochem.* **14**, 1-4 (1975).

34 Francklyn, C. S., First, E. A., Perona, J. J. & Hou, Y.-M. Methods for kinetic and thermodynamic analysis of aminoacyl-tRNA synthetases. *Methods* **44**, 100–118 (2008).

35 Fersht, A. R., Ashford, J.S., Bruton, C.J., Jakes, R., Koch, G.L.E. and Hartley, B.S. Active Site Titration and Amino Acyladenylate Binding Stoichiometry of Amino Acyl Synthetases. *Biochemistry* **14**, 1-4 (1975).

36 Carter, C. W., Jr. What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life* **5**, 294-320 (2015). https://doi.org/10.3390/life5010294

37 Li, L., Francklyn, C. & Carter, C. W., Jr. Aminoacylating Urzymes Challenge the RNA World Hypothesis. *J. Biol. Chem.* **288**, 26856-26863 (2013). https://doi.org/10.1074/jbc.M113.496125

38 Pham, Y. *et al.* A Minimal TrpRS Catalytic Domain Supports Sense/Antisense Ancestry of Class I and II Aminoacyl-tRNA Synthetases. *Mol. Cell* **25**, 851-862 (2007).

39 Petrov, A. S. *et al.* Evolution of the Ribosome at Atomic Resolution. *PNAS* **111** 10251–10256 (2014).

40 Bose, T. *et al.* Origin of life: protoribosome forms peptide bonds and links RNA and protein dominated worlds. *Nucleic Acids Res.* **50**, 1815–1828 (2022). https://doi.org/10.1093/nar/gkac052

41 Carter, C. W., Jr. Base Pairing Promoted the Self-Organization of Genetic Coding, Catalysis, and Free-Energy Transduction. *MDPI Life* **14**, 199 (2024).

42 Carter, C. W., Jr & Wills, P. R. Class I and II aminoacyl-tRNA synthetase tRNA groove discrimination created the first synthetase•tRNA cognate pairs and was therefore essential to the origin of genetic coding. *IUBMB Life* **71**, 1088–1098 (2019). https://doi.org/10.1002/iub.2094

43 Carter, C. W., Jr. & Wills, P. R. Experimental Solutions to Problems Defining the Origin of Codon-Directed Protein Synthesis. *BioSystems* **183**, 103979 (2019). https://doi.org/10.1016/j.biosystems.2019.103979

44 Hol, W. G. J. The Role of the α-Helix Dipole in Protein Function and Structure. *Prog. Biophys. molec. Biol.* **45**, 149-195 (1985).

45 Carter, C. W., Jr & Wills, P. R. Interdependence, Reflexivity, Fidelity, and Impedance Matching, and the Evolution of Genetic Coding. *Molecular Biology and Evolution* **35**, 269-286 (2018). https://doi.org/10.10-1093/molbev/msx265

46 Carter, C. W., Jr. in *Pre-Biotic Chemistry and Life's Origin* (ed Michele Fiore) Ch. 11, 317-346 (The Royal Society of Chemistry, 2022).

47 Douglas, J., Bouckaert, R., Carter, C. W., Jr. & Wills, P. R. Reduced amino acid substitution matrices find traces of ancient coding alphabets in modern day proteins. *BioRxiv* (2025). https://doi.org/10.1101/2025.02.19.63915

48 Douglas, J., Bouckaert, R., Harris, S., Carter, C. W., Jr & Wills, P. R. Evolution is coupled with branching across many granularities of life. *Proceedings of the Royal Society B* **In Press** (2025).

49 Sapienza, P. J., Li, L., Williams, T., Lee, A. L. & Carter, C. W., Jr. An Ancestral Tryptophanyl-tRNA Synthetase Precursor Achieves High Catalytic Rate Enhancement without Ordered Ground-State Tertiary Structures. *ACS Chemical Biology* **11**, 1661−1668 (2016). https://doi.org/10.1021/acschembio.5b01011

50 Leman, J. K. *et al.* Sequence-structure-function relationships in the microbial protein universe. *Nature Communications* **14**, 2-11 (2023).

51 Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *BioRxiv* (2023). https://doi.org/10.1101/2023.12.22.573103

52 Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378,** , 49–56 (2022) (2022).

53 Romeroa, M. L. R. *et al.* Simple yet functional phosphate-loop proteins. *Proc. Nat. Acad. Sci. USA* **115**, E11943–E11950 (2018). https://doi.org/10.1073/pnas.1812400115

54 Kuhlman, B. & Baker, D. Exploring folding free energy landscapes using computational protein design. *Curr. Op. Struct. Biol.* **14**, 89-95 (2004).

55 Patra, S. K. *et al.* Deep Learning Can Enhance the Amino Acid Sequences of Aminoacyl-tRNA Synthetase Urzymes. *Structural Dynamics* **In preparation** (2025).

56 Technologies, A. (https://www.agilent.com).

57 Fritsch, E. F., Sambrook, J. & Maniatis, T. *Molecular Cloning: A Laboratory Manual*. 545 (Cold Spring Harbor Laboratory, 1982).

58 Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160-1166 (2019). https://doi.org/10.1093/bib/bbx108.

59 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004). https://doi.org/10.1093/nar/gkh340

60 Schimmel, P. & Alexander, R. Diverse RNA substrates for aminoacylation: Clues to origins? *Proc. Nat. Acad. Sci. USA* **95**, 10351-10353 (1998). https://doi.org/10.1073/pnas.95.18.10351

61 Johnson, K. A. New standards for collecting and fitting steady state kinetic data. *Beilstein J. Org. Chem.* **15**, 16–29 (2019). https://doi.org/10.3762/bjoc.15.2

62 Box, G. E. P., Hunter, W. G. & Hunter, J. S. *Statistics for Experimenters*. (Wiley Interscience, 1978).

***Escherichia coli* transforms the leucyl-tRNA synthetase gene *in vivo* into primordial genes**

Guo Qing Tang and, Charles W. Carter, Jr

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7260

Supplemental Data

A. Raw sequencing data for 79 plasmids containing recombinant deletions from the Full-length double mutant LeuRS plasmid. These data can be found in the files:

Gene_Assembly_1.xltx
Gene_Assembly_2.xltx
Gene_Assembly_3.xltx
Gene_Assembly_4.xltx
Gene_Assembly_6.xltx

Note that in all of the excel files and subsequent supplemental figures, filenames beginning with 17A were incubated exclusively at 37o C and those beginning with 19A were incubated overnight at 37° C and then transferred to 4° C.

B. Fasta file of selected alignments excerpted in Fig. 1A of the paper. Sequences have been curated and aligned manually to facilitate comparison between one another.

```
>LeuAC_WT_A/1-129
-------------------------------------EKKFYITVAFPYLSGHLHVGHARTYTIPDVIA
RFKRMQGYNVLFPMAWHITSLS---------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------DSTIYMAYYTFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQKMSKSKGNVLNF----------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
-----------------------I
>LeuRS_B/1-967
M-------AELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLHVGHARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIVGIAERIKNRDPKTIWIYRDVYKVPEEILWTFEDPINIVKYFMKAAKET
FIRAGFSVDWSREFYTTSLFPPFSKFIEWQFWKLKEKGYIVKGAHRVRWDPVVGTPLGDHDLMEGEDVPILD
YIIIKFELRENGEVIYLPAATLRPETVYGVTNMWVNPNATYVKAKVRRKDKEETWIVSKEAAYKLSFQDREI
EVIEEFKGEKLIGKYVRNPVSGDEVIILPAEFVDPDNATGVVMSVPAHAPFDHVALEDLKRETEILEKYDID
PRIVENITYISLIKLEGYGDFPAVEEVNKLGIKSQKDKEKLEQATKTIYKAEYHKGIFKVPPYEGKPVQEVK
EAIAKEMLEKGIAEIMYEFAEKNVISRFGNRAVIKIIHDQWFIDYGNPEWKEKARKALERMKILPETRRAQF
EAIIDWLDKKACARKIGLGTPLPWDPEWVIESLSDSTIYMAYYTISRHINKLRQEGKLDPEKLTPEFFDYIF
LEEFSEDKEKELEKKTGIPAEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQKMSKSKGNVLNFIDAIEENGAD-------------------------------------
```

----------VV----------
----------RLYI--------MS----------LAEHD
SDFDWRRKEVGKLRKQIERFYELISQFAEYEVKGNVE--------LKDIDRWMLHRLNKAIKE-TTNALEEF
RTRTAVQWAFYSIMNDLRWYLR----------
----------RTEGRDDEAKRYVLRTLADVWVRLMAPFTPHICEELWEKLGGEGFVSLAKWP
EPVEEWWNETI----------EAEEEFIRSVMEDIKEIIEVAKIENA
KRAYIYTAEDWKWKVAEVVSEKRDFKSSMEEL----------MKDS------EIRKHGKEVAKIVQKLIKER
TFDVKRIN--EEKAL----------RE----------AKEFMEKELGIEIII
NPTEDKGGKKKQAMPLKPAIFIE
>17A_028_AVGAAMSAS-8_AVGA-sqn1_D04/1-282
----------FLSELRKMAERWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
---

----------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEENGADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQFA
EYEVKGNVELKDIDRWMLHRLNKAIKETTNALEEFRTRTAVQW--AFYSIMNDLRWYLRRTEGRDDEAKR--
----------YVLRTLADVWV----------RLMAPFTPHICEELWEKL--------
GG----------

----------E----------
----------AL
>19A_008_21-AVGA-AMSAS-AVGAsqun1-08_AVGAsqun1_H01/1-289
----------F-KRLRKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
---

----------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEENGADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQFA
EYEVKGNVEPKEKKFYITV----------AFPYLSGHLAVGA--A----------
----------RTYT--------IPAEIIHEMK-EEFEYWYPLDWRCSGKDLIPN--------
HLTFFIFTHVAIFREE----------
----------HLAE---RHCGERL---------W--HP--GGQAMIAT----------
------K----------

----------A----------
----------MC
>19A_014_21-AVGA-AMSAS-AVGAsqun1-14_AVGAsqun1_F02/1-238
----------RKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
---


```
------------------------------------------------------------------
------------------------------------------------------------------
------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEENGADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQFA
EYEVKGNVELKDIDRWMLHRLNKAIKETTNALEEFRTRTAVQW--AFYSIMNDLRWYLRRTE----------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
--------------R-----------------------------------------------------
--------------------PR
>19A_044_22AVGA-AMSAS-AVGAsqun1-04_AVGAsqun1_D06/1-348
-----------------RKWQKRWLEAKIFEPNIRHKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEENGADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQFA
EYEVKGNVELKDIDRWMLHRLNKAIKETTNALEEFRTRTAVQW--AFYSIMNDLRWYLRRTEGRDDEAKR--
---------------------YVLRTLADVWV-------------RLMAPFTPHICEELWEKL--------
GGEGFV-------SLA-------------------------------------------------
----------------KWPE---PVE--EW---------WNETI--EA-----EEEFIRSVMEDIKEIIEV
AKIETRNARIFNTAE--------------------------------------------------------
----------IEMESGGVV----S-----------------------------------------------
-----------------------------------------------ENA----------------------
------------ILK-----------------------------------------AA-----------
------------------------
>19A_052_22AVGA-AMSAS-AVGAsqun1-12_AVGAsqun1_D07/1-366
-----------T-FLSRLRKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------------------------------------------------------
------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEENGADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQFA
EYEVKGNVELKDIDRWMLHRLNKAIKETTNALEEFRTRTAVQW--AYYTISRHIN-KLRQ-EGKLDPEKLTP
EFFDYIFLEEFSEDKEKELEKKTG--------IPAEIIHEMK-EEFEYWYPLDWRCSGKDLIPN--------
HLTFFIFNHVAIFREE-------------------------------------------------------
----------------HWAE---RHCGERL---------W--HP--GSQAMSAS----------------
------KGNVLNFID-------------------------------------------------------
----------AIEENGADV----V-----------------------------------------------
------------------------------------------------------------------
------------RLL-----------------------------------------------YL--
------------------------
>19A_011_21-AVGA-AMSAS-AVGAsqun1-11_AVGAsqun1_C02/1-153
-----------------RKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------
```

```
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
-------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEEKW-QKR--------------------WL------EAKIF------
---EPNIR-------------------------------------------------------------
-----------------------------------D-KPK---------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
-----------------------------E-----------------------------K----------
-------------------------
>19A_054_22AVGA-AMSAS-AVGAsqun1-14_AVGAsqun1_F07outframe_Urzyme/1-227
-------------FLRRLRQWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
-------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIAFPY-------------------------------------------
-----------------------------------GHLAVGA---------------------------
-------------------ARTYT--------IPDVIARFKRMQGYNVLFPMAWHITGSPIV---------
------------------------------------------------------------------------
-----------------------GIAVNGF---------G--TL--EGQAMSAS----------------
-------KGNVLNFIA-----------------------------------------------------
-----------FP-VSERPSGGGRG-------------------------------------------
-------------------------------------------AHLYHS----------------
-------------GGN-------------------------------------------YS-------
-------------------------
>19A_064_22AVGA-AMSAS-amsasRV1-04_amsas-rv_H08/1-163
-EHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
-------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAMKKTA-TDG-------------------------------------
------------------------------------------------------------------------
-------------------------------WPXX----------------XX----------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
-------------------------------------------X--------------------
-------------XXX-------------F--------------------X------------------
```

```
------------------------
>19A_034_21-AVGA-AMSAS-amsas-rv-14_amsas-rv_B05/1-152
-MHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAMKKTA-RM--------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
----------------------------W----------------------------A----------
------------------------
>17A_038_AVGAAMSAS-8_AMSAS-sqn2-rv_F05/1-154
-MHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------AEIIHEMKEEFEYWYPLDWRCSGKDVIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAMKKMA-RRG--------------------G------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
----------------------------P-------------------------------------
---------------------G--
>19A_041_22AVGA-AMSAS-AVGAsqun1-01_AVGAsqun1_A06/1-377
-----------X-FLRRLRKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
----------------------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIP--------
NHLTFFIFNHV------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
--------------------AIFREEHWPKGIAVNGFGTLEGQAMS----ASK----------GNVLNFIA
FPYLSGHLAVGAARTYTIPAEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEENGADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQFA
EYEVKGNVELKDIDRWMLHRLNKAIKETTNALEEFRTRTAVQW--AFYSIMNDLRWYLRRTEAAMMKRNA--
---------------------MCCATLRCVGA-------------PD-GAVYPHICEEAVEKL--------
GR---------------------------------------------------------------------
------------------------------------------------------------------------
```

```
------------------------------------------------------------------------
--------------------------------RAL---------------------
>19A_055_22AVGA-AMSAS-AVGAsqun1-15_AVGAsqun1_G07/1-389
-----------H-FLSRLKIWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
----------------------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIP--------
NHLTFFIFNHV-----------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
--------------------AIFREEHWPKGIAVNGFGTLEGQAMS----ASK----------GNVLNFIA
FPYLSGHLAVGAARTYTIPAEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAIEENGADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQFA
EYEVKGNVELKDIDRWMLHRLNKAIKETTNALEEFRTRTAVQW--AFYSIMNDLRWYLRRTEAAMMKRNA--
--------------------MCCAPCGMLLG------------CADGAVTPHICEEAVGKL--------
GRRSFG-------ELX------------------------------------------------------
----------------NXXG-----------------------------------------------
------------------------------------------------------------------------
-------------------------XXW---------------------------
>19A_061_22AVGA-AMSAS-amsasRV1-01_amsas-rv_E08/1-232
-----------------RKMAETLAGSENFEPNIRDKPKEKKFYISVAFPYLSGHLAVGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
----------------------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIP--------
NHLTFFIFNHV-----------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
--------------------AIFREEHWPKGIAVNGFGTLEGQAMS----ASK----------GNVLNFIA
FPYLSGHLAVGAARTYTIPAEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAMKKT-TRMF-------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------PP-----------------
>LeuAC-Chimr_006/1-240
-L---------K-FKAIEEKWQKRWLEAKIFEPNTRDKPKEKKFYISVAFPYLSGHMAAGAARTYTIP----
------------------------------------------------------------------------
------------------------------------------------------------------------
----------------------------------AEIIHEMKEEFEYWYPLDWRCSGKDLIP--------
NHLTFFIFNHV-----------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
--------------------AIFREEHWPKGIAVNGFGTLEGQAMS----ASK----------GNVLNFIA
FPYLSGHLAVGAARTYTIPAEIIHEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFIDAMKKIGARLV-------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
------------------------------------------------------------------------
--------------------------------G"----------------------
>19A_047_22AVGA-AMSAS-AVGAsqun1-07_AVGAsqun1_G06/1-135
-----------F-FKR-LKKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIV----------------------------------------------
```


---GIAV
NGFGTLEGQAMSASKGNVLNFY-CVSVSE--RP----------
---SGGGRG----------
----------AHLYH---------SG-G----------NY----------

---S----------

----------A
>19A_015_21-AVGA-AMSAS-AVGAsqun1-15_AVGAsqun1_G02/1-140
----------TFFKA-IEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIV----------

---GIAV
NGFGTLEGQAMSASKGNVLNFI-AFPYLSG-HL----------
---AVGA----------
----------ARTYT--------IPVCLA----------PWKA----------

---R----------

----------R
>19A_067_22AVGA-AMSAS-amsasRV1-07_amsas-rv_C09/1-145
-HHHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIV----------

---GIAV
NGFGTLEGQAMSASKGNVLNFY-CVSVFE--RP----------
---SGGGRG----------
----------AHLYH---------SG-G----------NY----------

----------S
>19A_007_21-AVGA-AMSAS-AVGAsqun1-07_AVGAsqun1_G01/1-373
----------P-FKPI-EKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIV----------

----------GIAV
NGFGTLEGQAMSASKGNVLNFI-AFPYLSG-HL----------
----------AVGA----------
----------ARTYT--------IPDVIARFKRMQGYNVLFPMAWHITGSPIVGIAERIKN
RDPKTIWIYRDVYKVPEEILWTFEDPINIVKYFMKAAKETFIRAGFSV----------DWSREFYTTSLFP
PFS--KFIEWQF---------WKLKEKGYIVKGAHRVRWDPVVG--TPLGDHDLMEGE----------
--------DVPILDYIIIKFELRENG----------EVIYL-PAAT-------

-LRPETVYGV----------
TNMWVNPNATYVKAKVR----------RKDKEETWIVSKEAAYKLSFQIAKFEVIEEFKGEKLIG---
----------KYVRNP-XAAMK
>19A_016_21-AVGA-AMSAS-AVGAsqun1-16_AVGAsqun1_H02/1-116
----------DMAERWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIV----------

----------GIAV
NGFGTLEGQAMSASKATCTLLTLEGQAMSA-S----------

----------KA----------

----------T----------

----------C
>19A_005_21-AVGA-AMSAS-AVGAsqun1-05_AVGAsqun1_E01/1-106
----------N-FLRRLKKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIV----------

----------GIAV
NGFGTLEGQAMSASKAT----------

----------C
>19A_009_21-AVGA-AMSAS-AVGAsqun1-09_AVGAsqun1_A02/1-105
----------X-FKR-LRKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIPDVIA
RFKRMQGYNVLFPMAWHITGSPIV----------

----------GIAV
NGFGTLEGQAMSASKAT----------

----------C
>Leu_prozym_004/1-66
----------N-FLRRLRKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

----------R----------
----------"-
>Leu_prozym_005/1-62
----------RLRKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

---R----------
---"-
>Leu_prozym_007/1-59
----------EKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

---R-
>Leu_prozym_009/1-65
----------F-FKAI-EEMAERWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

---R----------
---"-
>Leu_prozym_010/1-76
-EHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

----------R----------
----------"-
>Leu_prozym_011/1-76
-EHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

----------R----------
----------"-
>Leu_prozym_012/1-76
-EHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

----------R----------
----------"-
>Leu_prozym_013/1-76
-EHHHHHHAELN-FKAIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------ALAPWK-AR----------

----------R----------
----------"-
>Leu_prozym_006/1-293
----------F-LKRIEEKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------EEIIHEMKEQFEYWYPLDWRCR----------

----------GKYLIPNPLTFFIFNRVAIFREKHWPKGIAV
NGFGPLEGHAKSASKGNVLNFI----------
----------AFPYLSGPLAGGA----------
----------ARTYT--------IPAEIIREMK-KKFEYWYPLEWRCRGKYLIPN------
--PLTFFIFNRVAIFREK----------
----------HWPK---GIAENGF---------G--TL--ERHAKSAS----------
--------KGNVLFHMA----------
----------WHITGSPS----VGIAERIK-HRYPNTIWIYRDVYKVREDILCTFEY----------
----------PINIVIYYMSAARDTYTRA----------
----------GFR----------V----------
----------"-
>17A_005_AVGA-5_AVGA-sqn1_E01.ab1/1-281 5'3'frame3
----------SRLRKWQKRWLEAKIFEPNIRDKPKEKKFYITVAFPYLSGHLAVGAARTYTIP----
----------AEIIHEMKEEFEYWYPLDWRCS----------

----------GKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFI-DAIEEKW-QKR----------WL------EAKIF-----
----EPNIRDKPKEKKFYITV----------AFPYLSGHLAVGA----------
----------ARTYT--------IPDVIARFKRMQGYNVLFPMAWHITGSPIV--------
----
----------GIAE----------

----------RIK-NRDPKTIWIYRDVYKVPEEILWTFED----------
----------PINIVKYFMKAAKETFIRA----------
----------GFSVDWSREFYTTRPVSA----------

----------V-
>3'5'/1-209 Frame 2_1+2
----------TEHSPINRKKKKFYITVAFPYLSGHLAVGAARTYTIP----
----------AEII----------

----------HEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFI-AFPY----------
----------GHLAVGA----------
----------ARTYT--------IPDVIARFKRMQGYNVLFPMAWHITGSPIV--------
---
----------GIAVNGF---------G--TL--EGQAMSAS----------
--------KGNVLNFIA----------
----------FP-VFERPSGGGRG----------
----------AHLYHS----------
----------GGN----------
----------YS-
>19A_074_22AVGA-AMSAS-amsasRV1-14_amsas-rv_B10/1-110
----------TLLRFPYLSGHLAVGAARTYTIP----
----------AEII----------

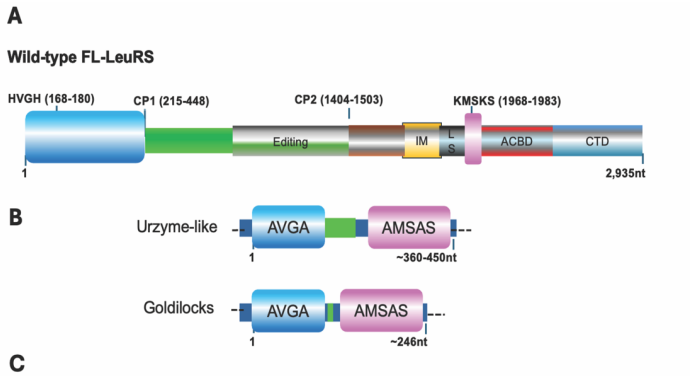
----------HEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLEGQAMSASKGNVLNFI-DAMKKTA-RMW----------

----------A----------
----------P
>19A_054_22AVGA-AMSAS-AVGAsqun1-14_AVGAsqun1_F07/1-151
----------TLLRFPYLSGHLAVGAARTYTIP----
----------AEII----------

----------HEMKEEFEYWYPLDWRCSGKDLIPNHLTFFIFNHVAIFREEHWPKGIAV
NGFGTLESQAMSASKGNVLNFI-DAIEENG-ADVVRLYIMSLAEHDSDFDWRRKEVGKLRKQIERFYELISQ
FADM----------

----------K

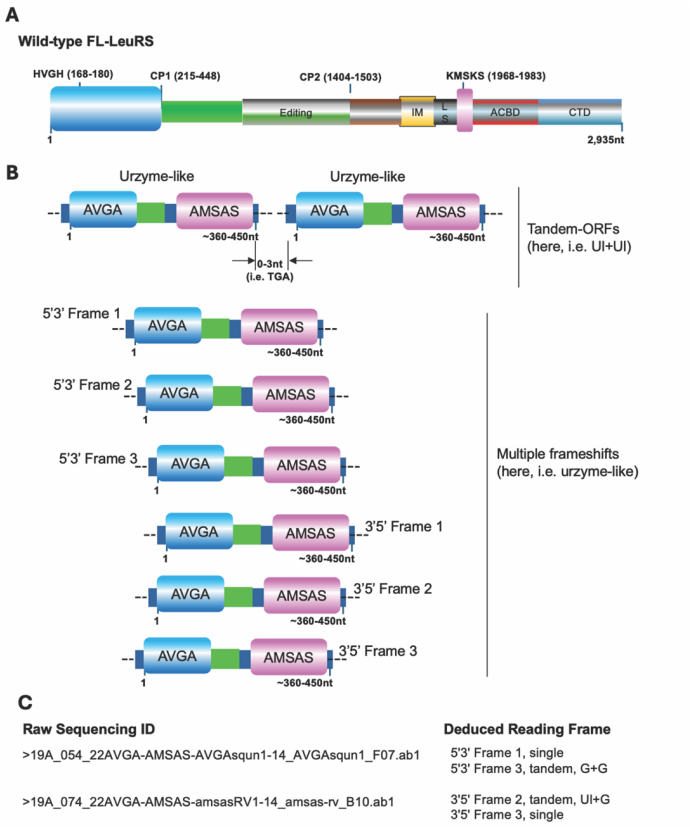
C. Varieties of genomic structure. The following figures illustrate the variety of urzyme, Goldilocks and protozyme sequences appearing in sequenced plasmid samples. These include single open-reading frame, (Supplemental figure 1, 5); tandem-configurations, (Supplemental Figure 2, 3, 7); and multiple embedding in multiple frames in the same plasmid, (Supplemental Figure 3, 4, 6, 7). Interestingly, Goldilocks and protozyme sequences appeared only at lower temperature 4º C.



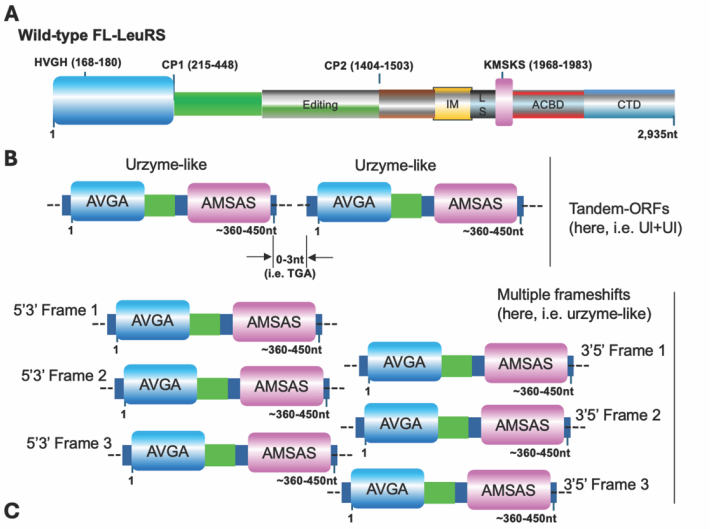
C

Raw Sequencing ID	Deduced Reading Frame
>17A_005_AVGA-5_AVGA-sqn1_E01.ab1	5'3' Frame 3, single
>17A_028_AVGAAMSAS-8_AVGA-sqn1_D04.ab1	5'3' Frame 2, single
>17A_038_AVGAAMSAS-8_AMSAS-sqn2-rv_F05.ab1	3'5' Frame 3, single
>17A_030_AVGAAMSAS-10_AVGA-sqn1_F04.ab1	5'3' Frame 3, single
>19A_005_21-AVGA-AMSAS-AVGAsqun1-05_AVGAsqun1_E01.ab1	5'3' Frame 1, single
>19A_007_21-AVGA-AMSAS-AVGAsqun1-07_AVGAsqun1_G01.ab1	5'3' Frame 3, single
>19A_008_21-AVGA-AMSAS-AVGAsqun1-08_AVGAsqun1_H01.ab1	5'3' Frame 1, single
>19A_009_21-AVGA-AMSAS-AVGAsqun1-09_AVGAsqun1_A02.ab1	5'3' Frame 3, single
>19A_014_21-AVGA-AMSAS-AVGAsqun1-14_AVGAsqun1_F02.ab1	5'3' Frame 3, single
>19A_015_21-AVGA-AMSAS-AVGAsqun1-15_AVGAsqun1_G02.ab1	5'3' Frame 2, single
>19A_019_21-AVGA-AMSAS-AVGAsqun1-19_AVGAsqun1_C03.ab1	5'3' Frame 1, single
>19A_026_21-AVGA-AMSAS-amsas-rv-6_amsas-rv_B04.ab1	3'5' Frame 1, single
>19A_034_21-AVGA-AMSAS-amsas-rv-14_amsas-rv_B05.ab1	3'5' Frame 1, single
>19A_047_22AVGA-AMSAS-AVGAsqun1-07_AVGAsqun1_G06.ab1	5'3' Frame 3, single
>19A_052_22AVGA-AMSAS-AVGAsqun1-12_AVGAsqun1_D07.ab1	5'3' Frame 2, single
>19A_055_22AVGA-AMSAS-AVGAsqun1-15_AVGAsqun1_G07.ab1	5'3' Frame 1, single
>19A_056_22AVGA-AMSAS-AVGAsqun1-16_AVGAsqun1_H07.ab1	5'3' Frame 2, single
>19A_061_22AVGA-AMSAS-amsasRV1-01_amsas-rv_E08.ab1	3'5' Frame 3, single
>19A_064_22AVGA-AMSAS-amsasRV1-04_amsas-rv_H08.ab1	3'5' Frame 3, single
>19A_075_22AVGA-AMSAS-amsasRV1-15_amsas-rv_C10.ab1	3'5' Frame 3, single

Supplemental Figure S1. Schematic diagram depicting genomic annotation of the sequenced insert from plasmids collected from two sets of small experiments. **A.** A cartoon describing Full-length Leucyl-tRNA synthetase (FL-LeuRS) labelled with domains and motifs. **B.** Symbolic scheme representation of urzyme-like sequence and goldilocks with AVGA and AMSAS mutations of the conventional catalytic signatures HVGH and KMSKS. **C.** The IDs of individual raw sequencing samples with single urzyme-like and goldilocks ORFs, corresponding to 6 possible ORFs. Raw sequencing identification (ID) on left; deduced open reading frame on right.

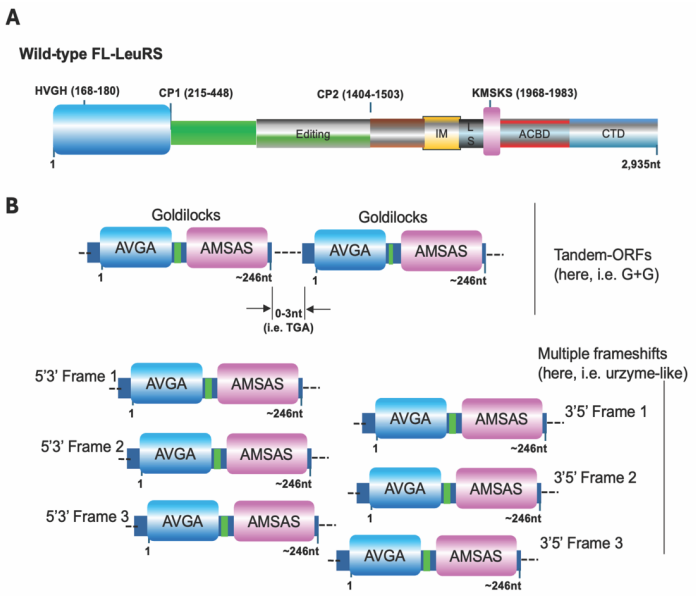


Supplemental Figure S2. Schematic representation of tandemly arranged urzyme-like elements in sequenced inserts at plasmids from two sets of experiments in small scales. **A.** Diagram indicates Full-length Leucyl-tRNA synthetase (FL-LeuRS) labelled with domains and motifs. **B.** Symbolic representation of the organization of urzyme-like ORFs in tandem arrays with a gap as indicated. **C.** A list of the sequenced plasmids with IDs of individual raw sequencing sample that include tandem-ORFs embedded, corresponding to 6 possible ORFs as indicated. Abbreviations: Ul, urzyme-like; G, goldilocks.



Raw Sequencing ID	Deduced Reading Frame
>19A_017_21-AVGA-AMSAS-AVGAsqun1-17_AVGAsqun1_A03.ab1	5'3' Frame 1, single
	5'3' Frame 3, single
>19A_065_22AVGA-AMSAS-amsasRV1-05_amsas-rv_A09.ab1	3'5' Frame 3, single
>19A_066_22AVGA-AMSAS-amsasRV1-06_amsas-rv_B09.ab1	3'5' Frame 3, single
>19A_067_22AVGA-AMSAS-amsasRV1-07_amsas-rv_C09.ab1	3'5' Frame 1, single
	3'5' Frame 2, single
>19A_054_22AVGA-AMSAS-AVGAsqun1-14_AVGAsqun1_F07.ab1	5'3' Frame 1, single
>19A_074_22AVGA-AMSAS-amsasRV1-14_amsas-rv_B10.ab1	3'5' Frame 3, single
>17A_005_AVGA-5_AVGA-sqn1_E01.ab1	5'3' Frame 3, single
>17A_028_AVGAAMSAS-8_AVGA-sqn1_D04.ab1	5'3' Frame 2, single
>17A_038_AVGAAMSAS-8_AMSAS-sqn2-rv_F05.ab1	3'5' Frame 3, single
>19A_008_21-AVGA-AMSAS-AVGAsqun1-08_AVGAsqun1_H01.ab1.	5'3' Frame 1, single
>19A_011_21-AVGA-AMSAS-AVGAsqun1-11_AVGAsqun1_C02.ab1	5'3' Frame 1, tandem, UI+UI
>19A_014_21-AVGA-AMSAS-AVGAsqun1-14_AVGAsqun1_F02.ab1	5'3' Frame 3, single
>19A_026_21-AVGA-AMSAS-amsas-rv-6_amsas-rv_B04.ab1	3'5' Frame 1, single
>19A_031_21-AVGA-AMSAS-amsas-rv-11_amsas-rv_G04.ab1	3'5' Frame 2, tandem, UI+UI
>19A_034_21-AVGA-AMSAS-amsas-rv-14_amsas-rv_B05.ab1	3'5' Frame 1, single
>19A_041_22AVGA-AMSAS-AVGAsqun1-01_AVGAsqun1_A06.ab1	5'3' Frame 2, tandem, UI+UI
>19A_052_22AVGA-AMSAS-AVGAsqun1-12_AVGAsqun1_D07.ab1	5'3' Frame 2, single
>19A_055_22AVGA-AMSAS-AVGAsqun1-15_AVGAsqun1_G07.ab1	5'3' Frame 1, single
>19A_061_22AVGA-AMSAS-amsasRV1-01_amsas-rv_E08.ab1	3'5' Frame 3, single
>19A_064_22AVGA-AMSAS-amsasRV1-04_amsas-rv_H08.ab1	3'5' Frame 3, single
>19A_075_22AVGA-AMSAS-amsasRV1-15_amsas-rv_C10.ab1	3'5' Frame 3, single

Supplemental Figure S3. A schematic display of genomic features of urzyme-like sequences found in either multiple-frame shift ORFs or tandemy-repeated elements in sequenced inserts. **A.** Full-length Leucyl-tRNA synthetase (FL-LeuRS) labelled with domains and motifs. **B.** Symbolic representation of organization of urzyme-like and goldilocks sequences in both of multiple-frameshifts and tandem arrays with a gap as indicated. **C.** List of individual raw sequencing samples with IDs on left, genomic features (both single and tandem-ORFs) were listed on right, corresponding to 6 possible ORFs as indicated in **B** and **C**.

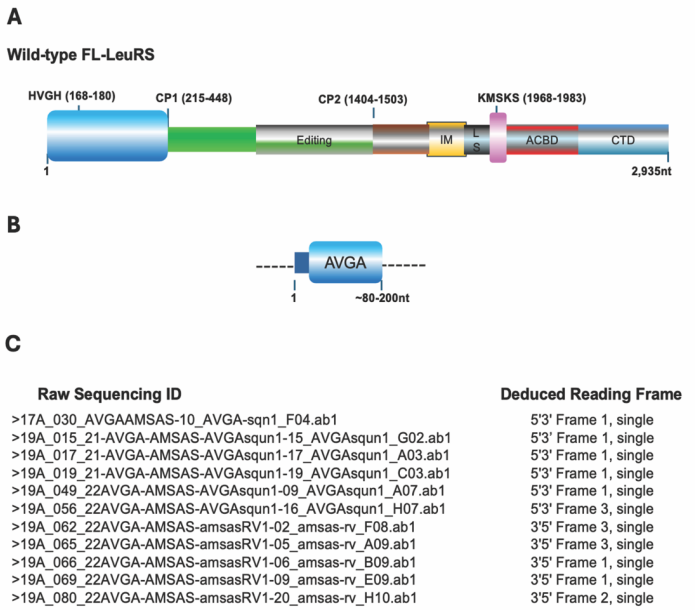


C

Raw Sequencing ID	Deduced Reading Frame
>19A_065_22AVGA-AMSAS-amsasRV1-05_amsas-rv_A09.ab1	3'5' Frame 2, single
>19A_066_22AVGA-AMSAS-amsasRV1-06_amsas-rv_B09.ab1	3'5' Frame 1, single
>19A_067_22AVGA-AMSAS-amsasRV1-07_amsas-rv_C09.ab1	3'5' Frame 1, single
>19A_054_22AVGA-AMSAS-AVGAsqun1-14_AVGAsqun1_F07.ab1	5'3' Frame 3, tandem, UI+G
>19A_074_22AVGA-AMSAS-amsasRV1-14_amsas-rv_B10.ab1	3'5' Frame 2, tendem, UI+G
>17A_030_AVGAAMSAS-10_AVGA-sqn1_F04.ab1	5'3' Frame 3, single
>19A_005_21-AVGA-AMSAS-AVGAsqun1-05_AVGAsqun1_E01.ab1	5'-3' Frame 1, single
>19A_007_21-AVGA-AMSAS-AVGAsqun1-07_AVGAsqun1_G01.ab1	5'3' Frame 3, single
>19A_009_21-AVGA-AMSAS-AVGAsqun1-09_AVGAsqun1_A02.ab1	5'3' Frame 3, single
>19A_015_21-AVGA-AMSAS-AVGAsqun1-15_AVGAsqun1_G02.ab1	5'3' Frame 2, single
>19A_019_21-AVGA-AMSAS-AVGAsqun1-19_AVGAsqun1_C03.ab1	5'3' Frame 1, single
>19A_047_22AVGA-AMSAS-AVGAsqun1-07_AVGAsqun1_G06.ab1	5'3' Frame 3, single
>19A_048_22AVGA-AMSAS-AVGAsqun1-08_AVGAsqun1_H06.ab1	5'3' Frame 3, tandem, G+G
>19A_056_22AVGA-AMSAS-AVGAsqun1-16_AVGAsqun1_H07.ab1	5'3' Frame 2, tandem, G+G
>17A_030_AVGAAMSAS-10_AVGA-sqn1_F04.ab1	5'3' Frame 3, single

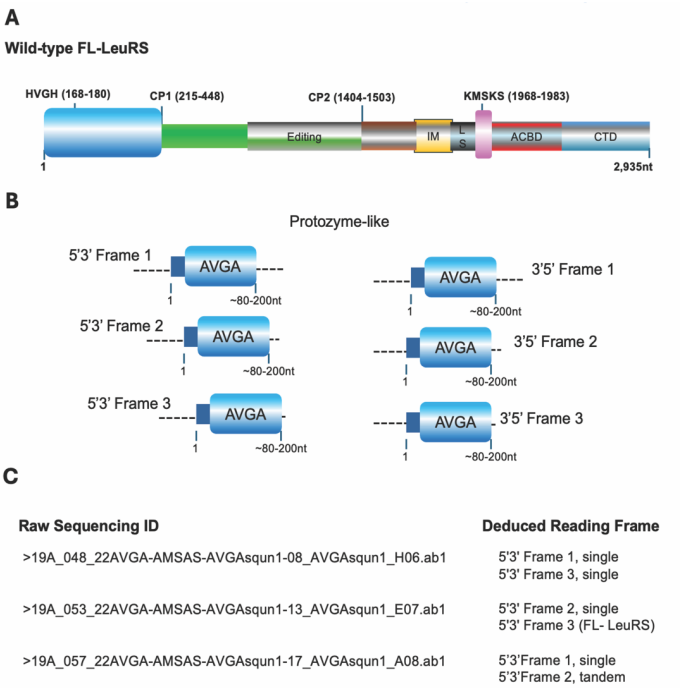
Supplemental Figure S4. Schematic diagram depicting Goldilocks sequences from sequenced insert regions of plasmid population with multiple Goldilocks ORFs in the same sequenced insert. **A.** Full-length Leucyl-tRNA synthetase (FL-LeuRS) labelled with domains and motifs. **B.** Symbolic representation of genomic features of Goldilocks sequences as indicated, showing combinations of multi-frames and ***6 reading frames***. **C.** The sequenced plasmids with IDs of individual raw sequencing sample, and genomic features of each individual Goldilocks element are summarized.

Commented [cwcjr1]: I fail to understand how these could occur in all six reading frames.



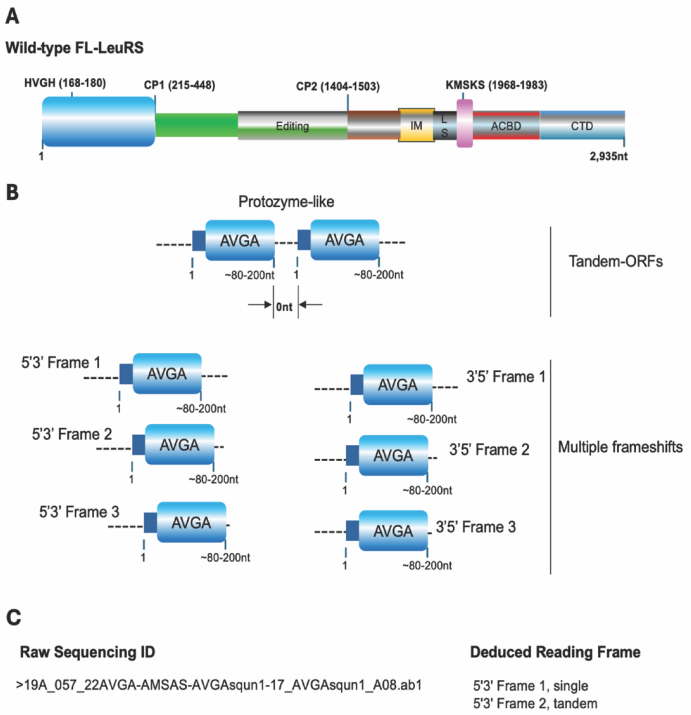
Raw Sequencing ID	Deduced Reading Frame
>17A_030_AVGAAMSAS-10_AVGA-sqn1_F04.ab1	5'3' Frame 1, single
>19A_015_21-AVGA-AMSAS-AVGAsqun1-15_AVGAsqun1_G02.ab1	5'3' Frame 1, single
>19A_017_21-AVGA-AMSAS-AVGAsqun1-17_AVGAsqun1_A03.ab1	5'3' Frame 1, single
>19A_019_21-AVGA-AMSAS-AVGAsqun1-19_AVGAsqun1_C03.ab1	5'3' Frame 1, single
>19A_049_22AVGA-AMSAS-AVGAsqun1-09_AVGAsqun1_A07.ab1	5'3' Frame 1, single
>19A_056_22AVGA-AMSAS-AVGAsqun1-16_AVGAsqun1_H07.ab1	5'3' Frame 3, single
>19A_062_22AVGA-AMSAS-amsasRV1-02_amsas-rv_F08.ab1	3'5' Frame 3, single
>19A_065_22AVGA-AMSAS-amsasRV1-05_amsas-rv_A09.ab1	3'5' Frame 3, single
>19A_066_22AVGA-AMSAS-amsasRV1-06_amsas-rv_B09.ab1	3'5' Frame 1, single
>19A_069_22AVGA-AMSAS-amsasRV1-09_amsas-rv_E09.ab1	3'5' Frame 1, single
>19A_080_22AVGA-AMSAS-amsasRV1-20_amsas-rv_H10.ab1	3'5' Frame 2, single

Supplemental Figure S5. Schematic diagram depicting genomic features of truncated protozyme sequences at the same sequenced insert region of same plasmid in a sequenced population. **A.** Full-length Leucyl-tRNA synthetase (FL-LeuRS) labelled with domains and motifs. **B.** Symbolic representation of genomic features of the truncated protozyme sequences as indicated in combinations of multi-frames and tandem-array in each sequenced insert. **C.** The sequenced plasmids with IDs of individual raw sequencing sample, and genomic features of each individual truncated protozyme are summarized.



Raw Sequencing ID	Deduced Reading Frame
>19A_048_22AVGA-AMSAS-AVGAsqun1-08_AVGAsqun1_H06.ab1	5'3' Frame 1, single 5'3' Frame 3, single
>19A_053_22AVGA-AMSAS-AVGAsqun1-13_AVGAsqun1_E07.ab1	5'3' Frame 2, single 5'3' Frame 3 (FL- LeuRS)
>19A_057_22AVGA-AMSAS-AVGAsqun1-17_AVGAsqun1_A08.ab1	5'3'Frame 1, single 5'3'Frame 2, tandem

Supplemental Figure S6. A schematic display of single ORF genomic features of truncated protozyme sequences in the list of sequenced inserts at tested pool of plasmids. **A.** Full-length Leucyl-tRNA synthetase (FL-LeuRS) labelled with domains and motifs. **B.** Symbolic representation of the organization of protozyme-like sequences in single ORFs as indicated. **C.** IDs of individual raw sequencing data containing the genomic features indicated in **B**.



Supplemental Figure S7. Schematic diagram depicting of genomic features with multiple-frames of truncated protozyme sequences at sequenced insert regions of plasmid population. **A.** Full-length Leucyl-tRNA synthetase (FL-LeuRS) labelled with domains and motifs. **B.** Symbolic representation of genomic features of protozyme sequences as indicated with combinations of 6 reading frames in each sequenced insert. **C.** The sequenced plasmids with IDs of individual raw sequencing sample, and genomic features of each individual truncated protozyme.

D. Design matrices for regression analyses.

Table S1. Design matrix for regression analysis of active-site titration parameters. Independent parameters in the last three columns are: WT, the presence of full LeuAC; 2ndXvr, the presence of the second crossover of the Rossmann dinucleotide binding fold; Protoz, the presence of the full length of the protozyme.

Variant	ΔGk_{chem}	ΔGk_3	$\Delta Gk_{chem}/k_3$	WT	2ND XVR	Protoz
WT 21–All	3.3	6.03	-2.73	1	0	1
WT 21–All	3.44	6.66	-3.22	1	0	1
WT 21–All	2.85	6.72	-3.87	1	0	1
WT 21–All	2.65	6.21	-3.56	1	0	1
WT 21–1	3.7	5.87	-2.17	1	0	1
WT 21–1	3.91	7.11	-3.2	1	0	1
WT 21–1	3.11	6.5	-3.39	1	0	1
WT 21–1	3.22	6.1	-2.88	1	0	1
WT 21-2	3.32	5.89	-2.57	1	0	1
WT 21-2	3.13	6.29	-3.16	1	0	1
WT 21-2	3.63	6.75	-3.12	1	0	1
WT 21-2	3.78	6.39	-2.61	1	0	1
WT 21-3	3.14	5.65	-2.51	1	0	1
WT 21-3	3.34	5.39	-2.05	1	0	1
WT 21-3	3.02	5.7	-2.68	1	0	1
WT 21-3	2.66	5.84	-3.18	1	0	1
DBL 21-All	3.26	5.82	-2.56	0	1	1
DBL 21-All	3.45	6.58	-3.13	0	1	1
DBL 21-All	2.82	6.51	-3.69	0	1	1
DBL 21-All	2.93	6.11	-3.18	0	1	1
DBL 21-1	3.46	5.88	-2.42	0	1	1
DBL 21-1	3.61	7.44	-3.83	0	1	1
DBL 21-1	3.2	6.34	-3.14	0	1	1
DBL 21-1	3.3	5.96	-2.66	0	1	1
DBL 21-2	3.13	5.96	-2.83	0	1	1
DBL 21-2	3.17	6.34	-3.17	0	1	1
DBL 21-2	2.9	6.82	-3.92	0	1	1
DBL 21-2	3.27	6.45	-3.18	0	1	1
DBL 21-3	3.16	5.74	-2.58	0	1	1
DBL 21-3	3.13	5.25	-2.12	0	1	1
WT 22–All	3.52	6.56	-3.06	1	0	0
WT 22–All	3.54	6.58	-2.58	1	0	0
WT 22–1	3.99	7.35	-3.04	1	0	0
WT 22–1	4.01	7.05	-3.04	1	0	0
WT 22-2	3.27	6.66	-4.66	1	0	0
WT 22-2	3.12	6.73	-4.36	1	0	0
WT 22-3	3.57	7.03	-3.36	1	0	0
WT 22-3	3.82	7.84	-3.04	1	0	0
DBL 22–All	3.26	6.05	-4.36	0	1	0
DBL 22–All	3.44	6.34	-3.39	0	1	0
DBL 22-1	3.37	6.04	-3.39	0	1	0
DBL 22-1	3.29	6.26	-3.61	0	1	0
DBL 22-2	3.13	6.3	-4.26	0	1	0
DBL 22-2	3.2	6.73	-3.9	0	1	0
DBL 22-3	3.14	5.65	-3.46	0	1	0
DBL 22-3	3.34	5.39	-4.02	0	1	0
LeuAC_WT_Ile	2.98	6.24	-4.5	1	1	1
LeuAC_DOUBLE_Ile	3.43	6.27	-3.99	0	1	1
LeuAC_WT_LEU	3.26	5.96	-2.79	1	1	1
LeuAC_DOUBLE_LEU	2.49	5.95	-2.9	0	1	1
LeuAC_WT_Ile_tRNA	3.67	6.58	-4.43	1	1	1
LeuAC_DOUBLE_Ile_tRNA	3.63	9.2	-4.01	0	1	1
LeuAC_WT_Ile_MINI	3.57	6.37	-2.67	1	1	1
LeuAC_DOUBLE_Ile_MINI	3.04	6.17	-2.97	0	1	1

Table S2. Design matrix for the analysis of steady-state kinetics of aminoacylation by Clones 21 and 22 as isolated (AVGA, AMSAS; Dbl) and corresponding constructs with wild-type catalytic signatures (HVGH, KMSKS; WT). Dependent variables are the free energies and were computed as –0.592*ln(Ksp, kcat, KM). Independent variables are coded as (0,1) according to the presence or absence of the WT catalytic signatures, the second crossover connection of the Rossmann dinucleotide binding fold, and the full ~50-residue protozyme.

Dataset	K_{sp}	k_{cat}	K_M	ΔGK_{sp}	ΔGk_{cat}	ΔGK_M	WT	2nd Xvr	Protoz
Clone21_WT	2.00E+04	3.51E-03	1.75E-07	-5.86E+00	3.35E+00	9.21E+00	1	0	1
Clone21_WT	1.46E+04	3.66E-03	2.51E-07	-5.67E+00	3.32E+00	9.00E+00	1	0	1
Clone21_WT	1.24E+04	3.74E-03	3.00E-07	-5.58E+00	3.31E+00	8.89E+00	1	0	1
Clone21_WT	6.00E+04	3.33E-03	5.55E-08	-6.51E+00	3.38E+00	9.89E+00	1	0	1
Clone21_DBL	1.54E+04	1.82E-03	1.18E-07	-5.71E+00	3.73E+00	9.44E+00	0	1	1
Clone21_DBL	7.60E+03	1.97E-03	2.60E-07	-5.29E+00	3.69E+00	8.98E+00	0	1	1
Clone21_DBL	9.66E+03	2.06E-03	2.13E-07	-5.43E+00	3.66E+00	9.09E+00	0	1	1
Clone21_DBL	4.44E+04	1.71E-03	3.86E-08	-6.34E+00	3.77E+00	1.01E+01	0	1	1
Clone22_WT	1.05E+04	1.83E-03	1.74E-07	-5.48E+00	3.73E+00	9.21E+00	1	0	0
Clone22_WT	6.24E+03	1.89E-03	3.03E-07	-5.17E+00	3.71E+00	8.89E+00	1	0	0
Clone22_WT	7.65E+03	2.00E-03	2.62E-07	-5.29E+00	3.68E+00	8.97E+00	1	0	0
Clone22_WT	4.07E+04	1.69E-03	4.16E-08	-6.28E+00	3.78E+00	1.01E+01	1	0	0
Clone22_DBL_Mt	1.87E+04	4.99E-03	2.66E-07	-5.82E+00	3.14E+00	8.96E+00	0	1	0
Clone22_DBL_Mt	1.18E+04	6.16E-03	5.24E-07	-5.55E+00	3.01E+00	8.56E+00	0	1	0
Clone22_DBL_Mt	1.60E+04	5.46E-03	3.42E-07	-5.73E+00	3.08E+00	8.81E+00	0	1	0
Clone22_DBL_Mt	1.13E+05	3.74E-03	3.32E-08	-6.89E+00	3.31E+00	1.02E+01	0	1	0
LeuAC_Ile_Mini_WT	7.18E+01	1.48E-04	2.06E-06	-2.53E+00	5.22E+00	7.75E+00	1	1	1
LeuAC_Ile_Mini_WT	5.93E+01	5.92E-04	9.99E-06	-2.42E+00	4.40E+00	6.82E+00	1	1	1
LeuAC_Ile_Mini_WT	1.73E+02	6.99E-05	4.04E-07	-3.05E+00	5.66E+00	8.72E+00	1	1	1
LeuAC_Ile_Mini_WT	7.73E+01	9.36E-05	1.21E-06	-2.57E+00	5.49E+00	8.07E+00	1	1	1
LeuAC_Ile_Mini_DOUBLE	3.79E+03	1.14E-03	3.00E-07	-4.88E+00	4.01E+00	8.89E+00	0	1	1
LeuAC_Ile_Mini_DOUBLE	3.04E+03	1.37E-03	4.51E-07	-4.75E+00	3.90E+00	8.65E+00	0	1	1
LeuAC_Ile_Mini_DOUBLE	5.11E+03	1.04E-03	2.03E-07	-5.05E+00	4.07E+00	9.12E+00	0	1	1
LeuAC_Ile_Mini_DOUBLE	3.11E+03	1.10E-03	3.54E-07	-4.76E+00	4.03E+00	8.79E+00	0	1	1

Table S3 Design matrix for regression analysis of the ratio of

Variant	<AMP/ADP>	Clone 21	WT
Clone21(WT)	0.412	1	1
Clone21(Dbl)	0.297	1	0
Clone22(WT)	0.285	0	1
Clone221(Dbl)	0.212	0	0

Table S4. Design matrix for estimating steady-state Δgkcat from single turnover data.

Variant	Δgkchem	ΔGk3	Δgkcat
21WT	3.26	6.19	3.34
21DBL	3.20	6.23	3.67
22WT	3.60	6.98	3.73
22DBL	3.27	6.10	3.14
LeuAC	3.26	6.59	4.60

E. Ancestral protein genes may have tolerated unexpected structural variation.

The *in vivo* deletions (Fig.1) bear striking resemblance to the nested hierarchy of protozymes and urzymes we designed based on the shared Class I AARS 3D structures. Those similarities and the catalytic proficiencies of Clone21 and Clone22 (Figs. 2,3) validate our previous work[1-9] in unexpected ways (Fig. S8). We had thought that the full protozyme and 2nd crossover segments were essential for acylation. Neither is actually required. There is little difference between the catalytic proficiencies of Clones 21, 22 and the LeuAC double mutant (see K_{sp} values in Table S2).

The long, dashed lines in Fig. S8D, E, F underscore the conceptual difficulty posed by their activity. Three discontinuous segments are connected by single peptide bonds in LeuAC. Neither of the two distinct deletions can be replaced by a single peptide bond or even a short oligopeptide. Making a continuous chain therefore seems to require different, drastic refolding, relative to full-length LeuRS (PDB ID 1WZ2). Yet, despite the long separations both constructs are single polypeptide chains that retain about the same activities as the double mutant LeuAC urzyme.

Some unknown aspect of the primary structure of the various modules enables at least three different linkages between the AVGA and AMSAS signatures to function catalytically in quite different contexts. An important clue may be that such short polypeptides are functional molten globules that hover around a catalytically competent configuration. Presented with appropriate substrates, they can adopt highly active arrangements complementary to the reaction transition state. The TrpAC urzyme has temperature-dependent melting curves and an NMR Heteronuclear Single Quantum Coherence (HSQC) spectrum characteristic of a molten globule[10]. More detailed studies on a monomeric variant of chorismate mutase[11-14] provide even more convincing evidence for such behavior. Its HSQC spectrum, like that of TrpAC has a narrow dispersion of proton resonances. However, in the presence of a transition-state analog inhibitor, the dispersion increases dramatically to that of a fully folded protein.

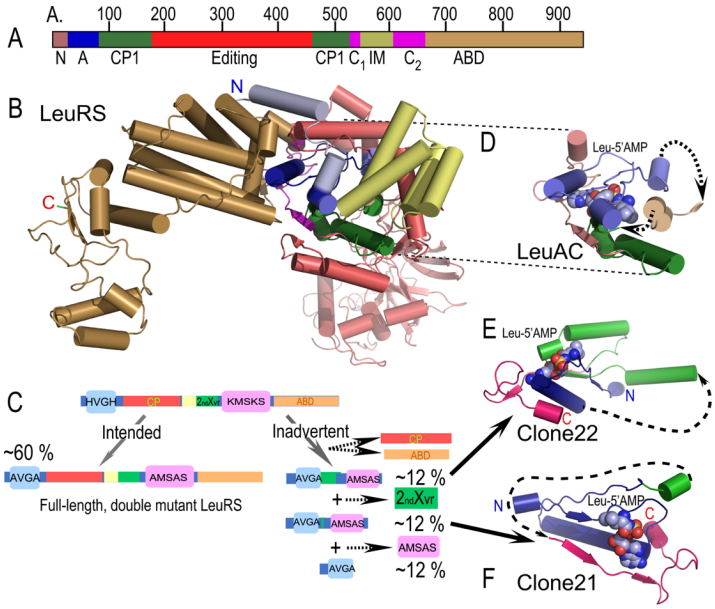


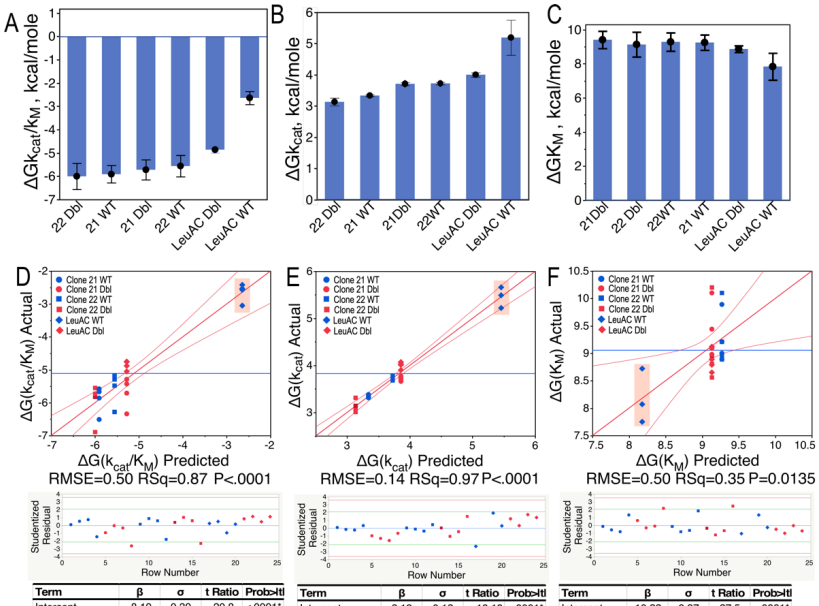
Figure S8. Deletion architectures entail different connections between active-site signatures. **A**. Linear schematic of the 3D structure of the mosaic structure of *P. horikoshii* LeuRS and the LeuAC urzyme. **B.** 3D cartoon of LeuRS and LeuAC which is excerpted in the inset (**D**). **C.** Comparison of the intended double mutant and the three subsets of short, fragmentary plasmids formed *in vivo* from the starting WT gene. **E.** The LeuAC-like urzyme. **F**. The intermediate Goldilocks urzyme. The vertical alignment of **D, E,** and **F** show that the signature sequences in the dark blue protozyme and the magenta C-terminal segment are linked together very differently. Yet, all have about the same catalytic proficiency.

F. Regression modeling shows how modules effect catalysis.

The Michaelis-Menten constants in Supplementary Table S2 vary in significant ways with the independent variables. The missing second crossover connection in Clone 22 and last β-strand of the protozyme in Clone 21, together with the WT and double mutant active site residues form a factorial design[6,15-17]. Clones 21 and 22 are significantly better at generating aminoacyl-3' minihelix (Fig. 7). The enhanced values of kcat/K_M result both from increased k_{cat} (Fig. 7B) and increased affinity for minihelix RNA (Fig. 7C). Clones 21 and 22 exhibit catalytic proficiencies similar to the AVGA/AMSAS double mutant of LeuAC.

The effects of active-site mutations on the two *in vivo* constructs are inverted. The WT Clone 21 is the best catalyst of aminoacylation, and it is nearly matched by the double mutant Clone 22. The Clone 21 double mutant is the least active variant. It has nearly the same $\Delta G(k_{cat}/K_M)$ as the double mutant LeuAC. The wild-type active-site variants bind more tightly to substrate minihelix in both Clone 21 and Clone 22, (Fig. 7C), but not in LeuAC. All four of the *in vivo* variants are better catalysts in all respects than the double mutant LeuAC. All of the effects influencing ΔGk_{cat} have highly significant P-values (Fig. 7B). The WT active side residues reduce the effect of all three parameters. The second crossover connection reverses the effect of active-site mutations on Clone 22. This is apparent from the placement of the various symbols on the graph in both Fig. 7D and 7E.

The rose shading in all three plots shows that the kinetic properties of WT LeuAC are all significantly worse than those of the other constructs. Estimation of higher order effects is limited by the poor reproducibility of K_M. The R^2 values are disappointing for ΔGK_M and consequently also for $\Delta Gk_{cat}/K_M$. Notwithstanding, significant evidence remains for the strength of interaction between the second crossover connection and the full length of the protozyme (Fig. 7D, E). That interaction reduces both ΔGk_{cat} and $\Delta Gk_{cat}/K_M$ by about 1 kcal/mole.



Term	β	σ	t Ratio	Prob>|t|
Intercept	-8.19	0.39	-20.8	<.0001*
WT	2.63	0.30	8.6	<.0001*
2ND XVR	2.19	0.47	4.7	0.0002*
Protoz	-0.35	0.35	-1.0	0.33
2ND XVR*Protoz	1.07	0.47	2.3	0.0329*

Term	β	σ	t Ratio	Prob>|t|
Intercept	2.13	0.12	18.19	<.0001*
WT	1.60	0.09	17.01	<.0001*
2ND XVR	1.01	0.14	7.42	<.0001*
Protoz	-0.39	0.10	-3.92	0.0010*
2ND XVR*Protoz	1.11	0.13	8.53	<.0001*

Term	β	σ	t Ratio	Prob>|t|
Intercept	10.22	0.37	27.5	<.0001*
WT	-0.95	0.33	-2.9	0.0084*
2ND XVR	-1.09	0.34	-3.2	0.0046*

Figure S9. Michaelis-Menten parameters of AARS urzymes depend on the structural differences between constructs. Each plot is ordered in descending favorability from left to right. **A**. Bar plot of $\Delta G(k_{cat}/K_M)$, the free energy for the second order rate constant for aminoacylation by reaction of free TΨC-minihelix with free substrate at saturating concentration of isoleucine, the amino acid substrate. **B**. $\Delta G(k_{cat})$, the free energy for the first-order rate constant. **C.** $\Delta G(K_M)$, the free energy for the apparent dissociation constant for the minihelix substrate. **D. E. F.** Regression models that relate steady-state parameters in **A-C** to the structural differences. Independent variables denote the following: WT refers to the presence of WT active-site histidine and lysine residues; 2ndXvr denotes the presence of the second crossover; Protoz denotes the presence of the intact protozyme at the N-terminus.

G. Clones 21 and 22 enhance studies of the evolution of protein modularity at unrivaled resolution.

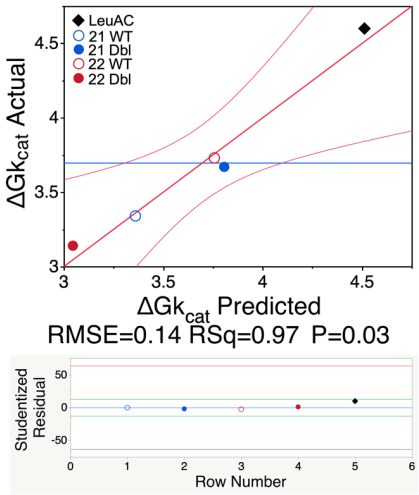
In what form(s) did that substantial mass arise? What functional improvements favored the addition of new modules? For Class I AARS, two distinct and substantial modules complement the urzyme. One is a long, variable insertion between the first and second halves of the catalytic apparatus. Much of that insertion is highly conserved in all 10 Class I AARS (see also [18]). It is often called connecting peptide 1 or CP1[19,20]. The other is an idiosyncratic set of modules—anticodon-binding domains, ABD—that bind to the tRNA anticodon stem-loop. It was highly unlikely that the urzyme gained both modules at the same time. We noted[21] that one of the two domains probably added first to the urzyme.

We made both possible evolutionary intermediates, adding either CP1 or the ABD to the TrpRS urzyme. Neither hybrid worked as well as the urzyme itself[6,22]. Thus, no selective advantage favored either pathway. That result falsified our hypothesis. We now measure the efficiency of ATP utilization using single turnover kinetics[23]. Efficiency joins catalysis and fidelity as a new metric of selective advantage.

An *in vivo* system that reflects the modularity of the nucleic acid structure of the gene can also help better define the actual extent of the ancestral modules. The differences between the deletions described here and the corresponding hierarchy we produced analytically may allow a more precise specification of the actual modularity of ancestral AARS. It will be of interest to follow the distribution of plasmid sizes as a function of time following transformation. The order in which modules are deleted *in vivo* may also suggest the order in which they were acquired. The results may thus open access to the chronology of events.

H. Estimating steady-state rates from single turnover rate constants.

To investigate further the evident difficulty of performing steady-state assays of aminoacylation by AARS urzymes, we developed a regression model expressing ΔGk_{cat} in terms of the free energy for the first order rate constant, ΔGk_{chem} and that for turnover, ΔGk_3. The data themselves are very noisy with respect to this question. However, as we had sufficient multiple estimates for all rate constants to reduce the standard error of the mean value for each variable to ~1%. The design matrix in Supplementary Table S4 shows this design matrix. The regression model is summarized in Supplementary Fig. 9. The table of regression coefficients implies that the proportional contributions to the ΔGk_{cat} value of the first-order and turnover are in the ratio of 2:1.



Term	β	σ	t Ratio	Prob>|t|
Intercept	3.54	1.40	2.53	0.13
ΔGkchem	-5.53	0.80	-6.87	0.02
ΔGk3	2.88	0.35	8.17	0.01

Supplementary Figure S10. Regression model for steady state $\Delta\Delta GkcagGk_{cat}$ in terms of the free energies for the two single-turnover rate constants. β and σ are regression coefficients and their standard deviations. The errors are thus < 15%. Their P values are more significant than the 95% level.

References

1 Carter, C. W., Jr. & Wills, P. R. The Roots of Genetic Coding in Aminoacyl-tRNA Synthetase Duality *Annual Review of Biochemistry* **90**, 349-373 (2021). https://doi.org/10.1146/annurev-biochem-071620-021218

2 Martinez-Rodriguez, L. *et al.* Functional Class I and II Amino Acid Activating Enzymes Can Be Coded by Opposite Strands of the Same Gene. *J. Biol. Chem.* **290**, 19710–19725 (2015). https://doi.org/10.1074/jbc.M115.642876

3 Carter, C. W., Jr. What RNA World? Why a Peptide/RNA Partnership Merits Renewed Experimental Attention. *Life* **5**, 294-320 (2015). https://doi.org/10.3390/life5010294

4 Carter, C. W., Jr. Urzymology: Experimental Access to a Key Transition in the Appearance of Enzymes. *J. Biol. Chem.* **289**, 30213–30220 (2014). https://doi.org/10.1047/jbcR114.576495

5 Li, L., Francklyn, C. & Carter, C. W., Jr. Aminoacylating Urzymes Challenge the RNA World Hypothesis. *J. Biol. Chem.* **288**, 26856-26863 (2013). https://doi.org/10.1074/jbc.M113.496125

6 Li, L. & Carter, C. W., Jr. Full Implementation of the Genetic Code by Tryptophanyl-tRNA Synthetase Requires Intermodular Coupling. *J. Biol. Chem.* **288**, 34736–34745 (2013). https://doi.org/10.1074/jbc.M113.510958

7 Li, L., Weinreb, V., Francklyn, C. & Carter, C. W., Jr. Histidyl-tRNA Synthetase Urzymes: Class I and II Aminoacyl-tRNA Synthetase Urzymes have Comparable Catalytic Activities for Cognate Amino Acid Activation. *J. Biol. Chem.* **286**, 10387-10395 (2011). https://doi.org/10.1074/jbc.M110.198929

8 Pham, Y. *et al.* Tryptophanyl-tRNA synthetase Urzyme: a model to recapitulate molecular evolution and investigate intramolecular complementation. *J. Biol. Chem.* **285**, 38590-38601 (2010). https://doi.org/10.1074/jbc.M110.136911

9 Pham, Y. *et al.* A Minimal TrpRS Catalytic Domain Supports Sense/Antisense Ancestry of Class I and II Aminoacyl-tRNA Synthetases. *Mol. Cell* **25**, 851-862 (2007).

10 Sapienza, P. J., Li, L., Williams, T., Lee, A. L. & Carter, C. W., Jr. An Ancestral Tryptophanyl-tRNA Synthetase Precursor Achieves High Catalytic Rate Enhancement without Ordered Ground-State Tertiary Structures. *ACS Chemical Biology* **11**, 1661−1668 (2016). https://doi.org/10.1021/acschembio.5b01011

11 Roca, M., Messer, B., Hilvert, D. & Warshel, A. On the relationship between folding and chemical landscapes in enzyme catalysis. *Proc. Nat. Acad. Sci., USA* **105** 13877–13882 (2008).

12 Schulenburg, C., Stark, Y., Künzle, M. & Hilvert, D. Comparative Laboratory Evolution of Ordered and Disordered Enzymes. *J. Biol. Chem.* **290,** , 9310–9320 (2015).

13 Vamvaca, K., Jelesarov, I. & Hilvert, D. Kinetics and Thermodynamics of Ligand Binding to a Molten Globular Enzyme and Its Native Counterpart. *J. Mol. Biol.* **382**, 971–977 (2008).

14 Pervushin, K., Vamvaca, K., Vogeli, B. & Hilvert, D. Structure and dynamics of a molten globular enzyme. *Nat. Struct. Mol. Biol.* **14**, 1202-1206 (2007).

15 Tang, G. Q., Hu, H., Douglas, J. & Carter, C. W., Jr. Primordial aminoacyl-tRNA synthetases preferred tRNA minihelix substrates over full-length tRNA. *Nucleic acids Research* **52**, 1-24 (2024). https://doi.org/10.1093/nar/gkae417

16 Tang, G. Q., Elder, J. J. H., Douglas, J. & Carter, C. W., Jr, . Domain Acquisition by Class I Aminoacyl-tRNA Synthetase Urzymes Coordinated the Catalytic Functions of HVGH and KMSKS Motifs. *Nucleic Acids Research* **51**, 8070–8084 (2023). https://doi.org/10.1093/nar/gkad590

17 Weinreb, V., Li, L. & Carter, C. W., Jr. A Master Switch Couples Mg^{2+}-Assisted Catalysis to Domain Motion in B. stearothermophilus Tryptophanyl-tRNA Synthetase. *Structure* **20**, 128-138 (2012). https://doi.org/10.1016/j.str.2011.10.020.

18 Douglas, J., Bouckaert, R., Carter, C. W. J. & Wills, P. Enzymic recognition of amino acids drove the evolution of primordial genetic codes. *Nucleic Acids Research* **52**, 558–571 (2024).

19 Burbaum, J. & Schimmel, P. Structural Relationships and the Classification of Aminoacyl-tRNA Synthetases. *The Journal of Biological Chemistry* **266**, 16965-16968 (1991).

20 Burbaum, J. J. & Schimmel, P. Assembly of a Class I tRNA Synthetase from Products of an Artificially Split Gene. *Biochemistry* **30**, 319-324 (1991).

21 Pham, Y. *et al.* A Minimal TrpRS Catalytic Domain Supports Sense/Antisense Ancestry of Class I and II Aminoacyl-tRNA Synthetases. *Mol. Cell* **25**, 851-862 (2007).

22 Weinreb, V. *et al.* Enhanced Amino Acid Selection in Fully-Evolved Tryptophanyl-tRNA Synthetase, Relative to its Urzyme, Requires Domain Movement Sensed by the D1 Switch, a Remote, Dynamic Packing Motif *J. Biol. Chem.* **289**, 4367-4376 (2014). https://doi.org/10.1074/jbc.M113.538660

23 Patra, S. K. *et al.* A genomic database furnishes minimal functional glycyl-tRNA synthetases homologous to other, designed class II urzymes. *Nucleic Acids Research* **52**, 13305–13324 (2024). https://doi.org/https://doi.org/10.1093/nar/gkae992