# Common evolutionary origins of the bacterial glycyl tRNA synthetase and alanyl tRNA synthetase

Claudia Alvarez-Carreño[1,2]; Marcelino Arciniega[3]; Lluis Ribas de Pouplana[4,5]; Anton S. Petrov[1,2]; Adriana Hernández-González[3]; Marco Igor Valencia-Sánchez[3,*]; Loren Dean Williams[1,2,*]; and Alfredo Torres-Larios[3]

[1]NASA Center for the Origin of Life, Georgia Institute of Technology, Atlanta, GA 30332-0400, USA

[2]School of Chemistry and Biochemistry, Georgia Institute of Technology, 901 Atlantic Drive, Atlanta, GA 30332, USA

[3]Departamento de Bioquímica y Biología Estructural, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Circuito Exterior s/n, Ciudad Universitaria, Apartado postal 70-243, Mexico City, 04510, México.

[4]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Parc Científic de Barcelona, C/ Baldiri Reixac 10, 08028 Barcelona, Catalonia, Spain.

[5]Catalan Institution for Research and Advanced Studies, Barcelona, Catalonia 08010, Spain and 6 University Pompeu Fabra, Barcelona, Catalonia 08003, Spain.

* Corresponding authors:

Marco Igor Valencia-Sánchez

**Email:** igorval@comunidad.unam.mx

Loren Dean Williams

**Email:** loren.williams@chemistry.gatech.edu

## Abstract

Aminoacyl-tRNA synthetases (aaRSs) establish the genetic code. Each aaRS covalently links a given canonical amino acid to a cognate set of tRNA isoacceptors. Glycyl tRNA aminoacylation is unusual in that it is catalyzed by different aaRSs in different lineages of the Tree of Life. We have investigated the phylogenetic distribution and evolutionary history of bacterial glycyl tRNA synthetase (bacGlyRS). This enzyme is found in early diverging bacterial phyla such as Firmicutes, Acidobacteria, and Proteobacteria, but not in archaea or eukarya. We observe relationships between each of six domains of bacGlyRS and six domains of four different RNA-modifying proteins. Component domains of bacGlyRS show common ancestry with i) the catalytic domain of class II tRNA synthetases; ii) the HD domain of the bacterial RNase Y; iii) the body and tail domains of the archaeal CCA-adding enzyme; iv) the anti-codon binding domain of the arginyl tRNA synthetase; and v) a previously unrecognized domain that we call ATL (Ancient tRNA latch). The ATL domain is found only in bacGlyRS and in the universal alanyl tRNA synthetase (uniAlaRS). Further, the catalytic domain of bacGlyRS is more closely related to the catalytic domain of uniAlaRS than to any other aminoacyl tRNA synthetase. The combined data suggest that the ATL and catalytic domains of these two enzymes are ancestral to bacGlyRS and uniAlaRS, which emerged from common protein ancestors by bricolage, stepwise accumulation of protein domains, before the last universal common ancestor of life.

## Keywords

Translation, GlyRS, tRNA, genetic code

## Abbreviations and symbols

aaRS: aminoacyl tRNA synthetase
ABD: anti-codon binding domain
arcGlyRS: glycyl aminoacyl tRNA synthetase of archaea and eukarya, and some bacteria
ATL: Ancient tRNA latch domain
bacGlyRS: bacterial glycyl aminoacyl tRNA synthetase

55 CPR: Candidate Phyla Radiation
56 GlyRS: glycyl aminoacyl tRNA synthetase
57 LysRS: Lysyl aminoacyl tRNA synthetase
58 mRNA: messenger RNA
59 MSA: multiple-sequence alignment
60 RNase Y: ribonuclease Y
61 tRNA: transfer RNA
62 uniAlaRS: universal alanyl tRNA synthetase
63 uniArgRS: universal arginyl-tRNA synthetase
64

## Introduction

Translation of mRNAs into coded polypeptides is universal to life (1). mRNA decoding during translation depends on the ribosome and a group of enzymes called aminoacyl tRNA synthetases (aaRSs) (2-5). aaRSs determine the genetic code by recognizing specific tRNAs and linking them to their cognate amino acids. aaRSs fall into two classes with independent origins in the deep evolutionary past (6; 7; 4). The catalytic domain of Class I aaRSs is a Rossmann fold while the catalytic domain of Class II aaRSs is an α/β three-layered sandwich. The early evolution of aaRSs has been linked to the origins of the genetic code and to the emergence of the first proteins (8-12).

Throughout the phylogenetic tree, a given tRNA is typically aminoacylated by a single type of aaRS. However, there are exceptions to this rule. There are two distinct lysyl-tRNA synthetases (LysRSs) (13). A class II LysRS is more frequent in bacteria, while a class I LysRS is found predominantly in archaea (14).

There are also two distinct GlyRSs (15-17). The GlyRS found in archaea and eukarya, and in some bacteria, is called here arcGlyRS. The GlyRS found in most in bacteria is called here bacGlyRS. Except for their catalytic domains, bacGlyRS and arcGlyRS are globally different (18) and use different strategies to recognize $tRNA^{Gly}$. arcGlyRS is relatively small, with a catalytic domain and a tRNA recognition domain. bacGlyRS is larger and more complex. Here we ask why there are two GlyRSs. How did they originate? Which GlyRS appeared first? Why is the distribution of the two GlyRSs over the phylogenetic tree patchy?

We and others (19; 17) have hypothesized that the component domains of bacGlyRS were serially exapted through bricolage. The bricolage hypothesis suggests that complex proteins can arise and evolve through stepwise exaptation and adaption of pre-existing domains or modules, which are combined to produce new functions (20). In this hypothesis, domains can be thought of as evolutionary building blocks that provide information on ancestry.

If the bricolage hypothesis is correct, homologies of individual domains can be used to help reconstruct the complex histories of multidomain proteins such as bacGlyRS and can help

us understand ancestral relationships. Accurate histories require accurate domain delineations. A domain is an quasi-independent and stable three-dimensional structure (21-23) and is a common unit of evolution (24; 25). Domains are encoded by genes. Duplicated genes, called paralogs (26), can combine to form a variety of distinct multi-domain proteins.

bacGlyRS contains $\alpha$ and $\beta$ subunits. The combined data are interpreted here to indicate that the $\beta$ subunit of bacGlyRS holds six component domains rather than five domains as proposed previously (27). The domain of bacGlyRS that escaped detection in previous work is informative about bacGlyRS ancestry and shows homology to a domain found in the universal alanyl tRNA synthetase (uniAlaRS), a class II aaRS. Sequence and structural comparisons suggest that bacGlyRS is more closely related to uniAlaRS than to arcGlyRS (28; 29; 19).

Our results are consistent with a model in which bacGlyRS evolved by bricolage. We show that six component domains of bacGlyRS are homologous to other proteins. These other proteins are universal, bacteria-specific or archaea-specific. bacGlyRS shows common ancestry with six domains of four proteins. Each of these proteins interacts with an RNA although none interact with $tRNA^{Gly}$. Component domains of bacGlyRS show sequence and structural similarities with 1) the catalytic and ATL domains (see below) of uniAlaRS; 2) the anti-codon binding domain (ABD) of the universal arginyl-tRNA synthetase (uniArgRS); 3) the body and tail domains of the archaeal CCA-adding enzyme; and 4) the HD domain of the bacterial RNase Y (in agreement with previous observations of structural similarity (27)). This conservation of RNA as substrate throughout a series of exaptation steps suggests an accessible mechanism by which RNA-associated proteins are repurposed for new RNA-associated functions.

# Results

## bacGlyRS has a complex evolutionary history.

The phylogenetic distribution of bacGlyRS appears irregular and discontinuous ('patchy') over the extant twigs of the bacterial Tree of Life. We examined a broad range of archaeal and bacterial proteomes, some of which are only partial (Figure 1). Bacterial species have either bacGlyRS or arcGlyRS. Bacteria containing both bacGlyRS and arcGlyRS are extremely rare. Of the 542 sampled proteomes, arcGlyRS (GenBank: PSL06957.1) and bacGlyRS (GenBank: PSL02860.1) are found together only once, in the bacterium *Haloactinopolyspora alba* (NCBI RefSeq assembly: GCA_003014555.1) isolated from sediment of the Dead Sea. Many bacterial phyla have an irregular distribution of the GlyRSs with multiple branches containing either bacGlyRS or arcGlyRS. Firmicutes, Actinobacteria, Chloroflexi, Proteobacteria and DST, PVC, and FCB groups each contain species with bacGlyRS and species with arcGlyRS. Candidate Phyla Radiation (CPR) uniformly contains arcGlyRS; Cyanobacteria uniformly contains bacGlyRS.

## bacGlyRS sequences are present in early diverging bacteria

The phylogenetic distribution of bacGlyRS suggests that it was present at the Last Bacterial Common Ancestor (LBCA). bacGlyRS is found in early diverging bacterial phyla (30) such as Firmicutes, Acidobacteria, and Proteobacteria (Figure 1). bacGlyRS is not found in the Candidate Phyla Radiation (CPR) phylum. The presence of arcGlyRS in CPR proteomes may be due to horizontal gene transfer rather than vertical inheritance. The CPR phylum contains putative symbiotic and parasitic bacteria with reduced genomes, large gene losses and extensive lateral gene transfer (31).

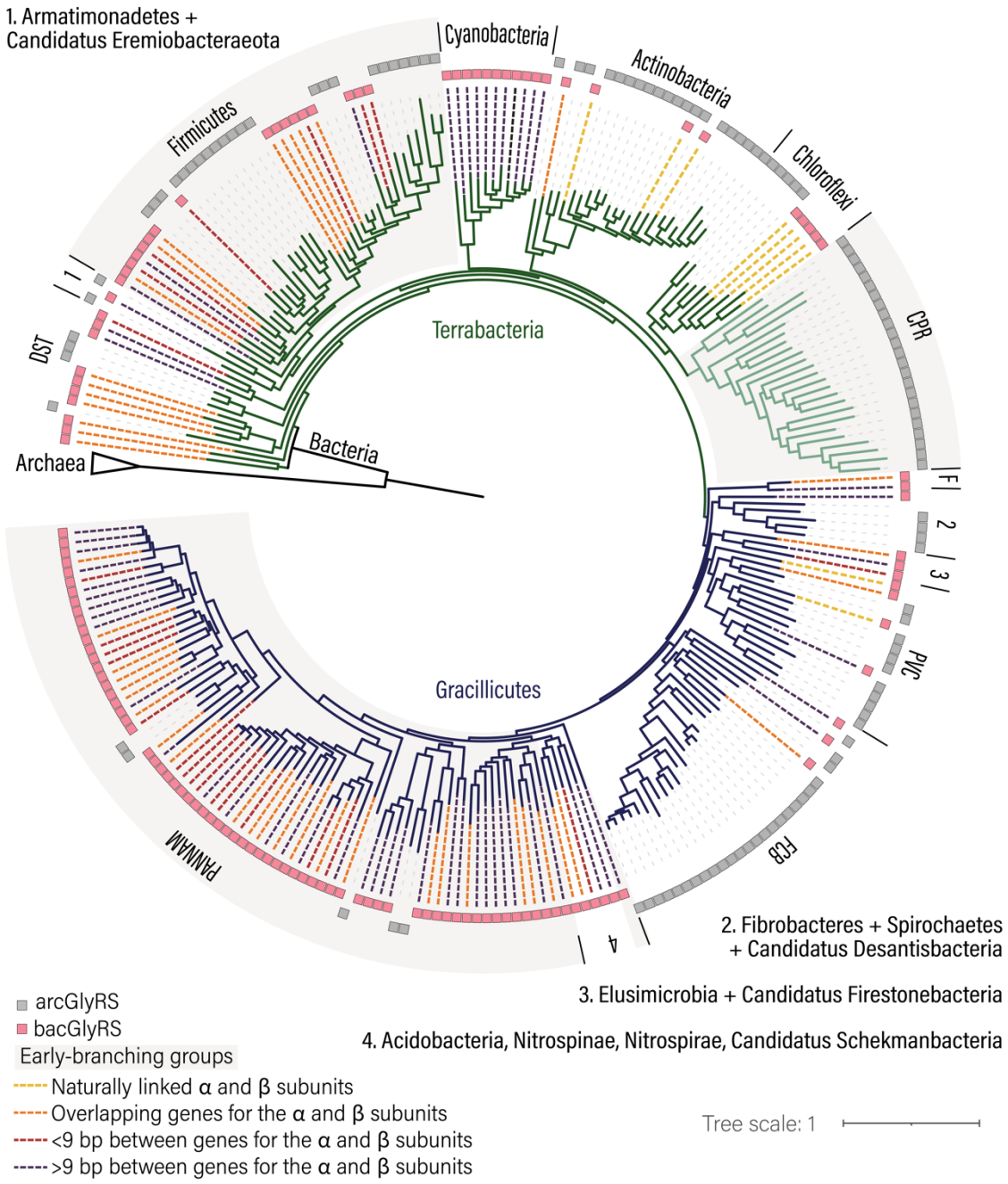


**Figure 1. Phylogenetic distribution of the two versions of Glycyl tRNA ligase in Bacteria.** The presence of bacGlyRS and arcGlyRS is mapped into a phylogeny of bacterial species. Yellow dotted lines: bacGlyRS with naturally linked α and β subunits form homodimers instead of heterotetramers. Pink dotted lines: species in which the genes for the α and β subunits overlap. Red dotted lines: species in which the genes for the α and β subunits are separated by nine base pairs or less. Purple dotted lines: species in which the genes for the α and β subunits are separated by more than nine base pairs. The topology of the tree was obtained from (32). DST: Deinococcus-Thermus, Synergistetes, Thermotogae, Caldiserica, Coprothermobacterota; CPR: Candidatus Phyla Radiation; PVC: Planctomycetes, Verrucomicrobia, Chlamydiae, Kiritimatiellaeota, Lentisphaerae, Candidatus Desantisbacteria, Candidatus Omnitrophica; FCB: Fibrobacteres, Chlorobi, Bacteroidetes, Gemmatimonadetes, Candidatus Cloacimonetes, Candidatus Fermentibacteria, Candidatus Glassbacteria; PANNAM: Proteobacteria, Aquificae, Nitrospinae, Nitrospirae, Acidobacteria, Chrysiogenetes, Deferribacteres, Schekmanbacteria and Thermodesulfobacteria. F. Fusobacteria. bp: base pairs.

## Domains in bacGlyRS are related to RNA-modifying enzymes

Each domain of bacGlyRS has an independent evolutionary history. Six domains in bacGlyRS display strong sequence similarity to domains of other proteins (Figure 2). We determined proteins that are the closest sequence matches to each of the six domains of bacGlyRS (Figure 2E). The catalytic domain shows sequence similarity to the catalytic domain of uniAlaRS (e-value $1.1\times10^{-07}$). The body and tail domains show sequence similarity to domains of the archaeal CCA-adding enzyme (e-value $1.4\times10^{-08}$). The HD domain shows sequence similarity to a domain of ribonuclease Y (RNase Y) (e-value $5.7\times10^{-08}$). The ABD shows sequence similarity to a domain of uniArgRS (e-value $2.9\times10^{-17}$).

We identified a previously unrecognized domain in bacGlyRS that we call Ancient tRNA latch (ATL) domain. Our determination that ATL is a distinct domain follows previous identification of this domain in uniAlaRS (33). The ATL domains of uniAlaRS and bacGlyRS are similar to each other in sequence and structure (Figure 3A). If ATL is a distinct domain in uniAlaRS, then it is also a domain in bacGlyRS. Previously, the ATL domain was described incorrectly as a dependent element of the HD domain (27; 17; 34); we infer here from sequence and structural considerations that the ATL is independent of other elements of bacGlyRS. The ATL domain appears to be unique to bacGlyRS and uniAlaRS, no other homologs of the ATL domain are identified in a comprehensive protein sequence database (UniRef90). The ATL domain of uniAlaRS contains a C-terminal $\alpha$-helical decoration that is absent from the ATL domain of bacGlyRS.

In addition, we observe that the function of the ATL domain has been mischaracterized. The ATL domain in uniAlaRS is called the "putative anticodon-binding domain" in some classification databases (SCOPe ID a.203.1.1, (35), and ECOD ID 613.1.1.1 (36)). However, neither the ATL domain of uniAlaRS nor the ATL domain of bacGlyRS contacts the anti-codon stem-loop of tRNA (37; 38). Both interact extensively with the amino acid acceptor stem. This mischaracterization of the ATL domain structure and function has obscured its evolutionary origins and molecular function.

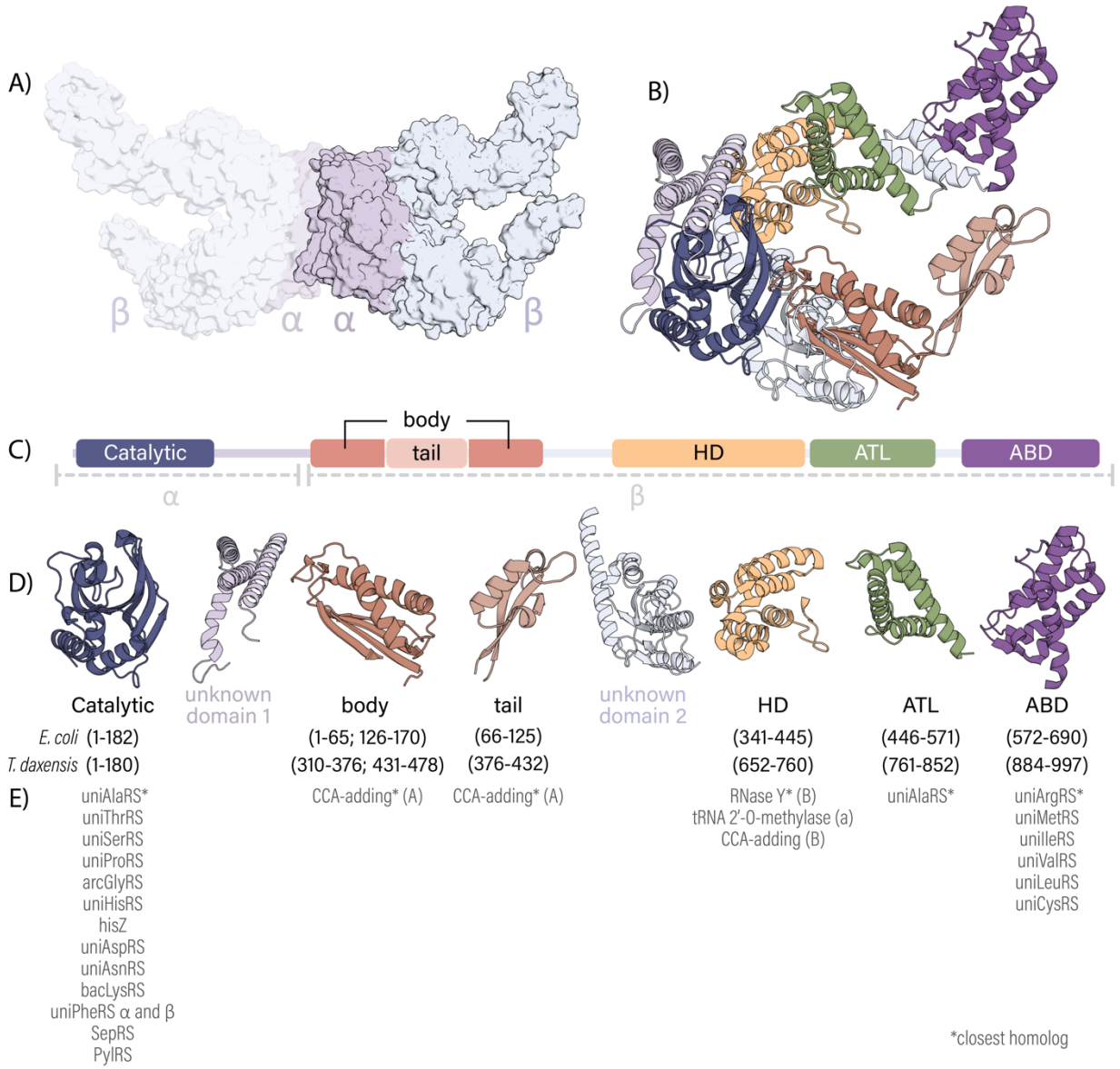


**Figure 2. Domain organization of bacGlyRS.** Structure of bacGlyRS of *Thermoanaerotrix daxensis* with naturally linked α and β-subunits (PDB: 7LU4). (A) Surface representation of bacGlyRS indicating the region that corresponds to the α- (light purple) and β- (light blue) subunits in other bacteria. (B) A ribbon representation of bacGlyRS shows structural units in different colors: dark blue, catalytic domain; light purple, unknown domain 1; dark orange, body domain; light orange, tail domain; light blue, unknown domain 2; yellow, HD domain; green, ATL domain; and dark purple, ABD domain. (C) Domains and structural units of bacGlyRS. (D) Proteins containing domains with high-scoring sequence similarity by bidirectional BLAST.

## bacGlyRS and uniAlaRS homology is observed in key functional domains

The primary functions of aaRSs are to recognize tRNAs and amino acids and to covalently join them together. The catalytic and ATL domains, which carry out these essential functions are homologous between bacGlyRS and uniAlaRS.

167 The ATL domains of uniAlaRS and bacGlyRS recognize tRNAs by variations of a common
168 mechanism. tRNAs are recognized by uniAlaRS and bacGlyRS by interactions of ATL with the
169 tRNA acceptor stem (27; 17; 34). The tRNA acceptor stem is the first seven base pairs of tRNA,
170 the unpaired nucleotide at position 73, and the 3'-CCA tail. In species containing bacGlyRS, the
171 acceptor stems of $tRNA^{Gly}$ and $tRNA^{Ala}$ differ in three pairs and in position 73. These three base
172 pairs and position 73 are signatures (39), meaning that they are highly conserved within each
173 group, $tRNA^{Gly}$ or $tRNA^{Ala}$, but differ between the two groups. The specific signatures of $tRNA^{Gly}$
174 are base pairs C2:G71, G3:C70, G4:C69, and U73. The signatures of $tRNA^{Ala}$ are G2:C71, G3:U70,
175 C4:G69, and A73. In uniAlaRS, the wobble base pair G3:U70 is the main identity determinant for
176 $tRNA^{Ala}$ aminoacylation by uniAlaRS (40; 33; 37; 41; 42; 34). The ATL domain interacts with the
177 base of the discriminator position (U73 in $tRNA^{Gly}$, and A73 in $tRNA^{Ala}$).

178 The ATL domain has adapted to distinguish features of the amino acid acceptor stems of
179 $tRNA^{Gly}$ or $tRNA^{Ala}$ (Figure 3). We have identified four distinct ATL adaptations, allowing a
180 common fold to recognize different tRNAs. To discriminate between the two different tRNAs,
181 the ATL domain (i) slightly shifts global position relative to the tRNA, (ii) shifts local position of
182 the secondary structural elements by conformational change, (iii) changes sequence and (iv)
183 acquires $\alpha$-helical decorations.

184 The ATL domains of bacGlyRS and uniAlaRS exhibit slightly different positions relative to
185 the tRNA. A distance of 4.1 Å between the centers of mass of $^{bacGly}$ATL and $^{Ala}$ATL after direct
186 superimpositions of the $tRNA^{Gly}$ in complex with bacGlyRS (PDB 7YSE) and the $tRNA^{Ala}$ in
187 complex with uniAlaRS (PDB 3WQY), indicates the changes in the global pose of the ATL
188 domains (Figure 3A).

189 The interactions between ATL and tRNA are locally different in $^{bacGly}$ATL and $^{Ala}$ATL. After
190 direct superimposition of the $^{bacGly}$ATL and $^{Ala}$ATL protein domains, a 3.1 Å root mean square
191 deviation (RMSD) between the superimposed backbone atoms quantitates local changes in
192 conformation (Figure 3A). To describe local positional shifts, we have reindexed the ATL
193 residues so that a common index (subscripted here) indicates the same column in the sequence
194 alignment (Figure 3D). $^{bacGly}ATL_{10}$ indicates the amino acid at the $10^{th}$ alignment position of ATL

195 domain of bacGlyRS while $^{Ala}ATL_{10}$ indicates the amino acid of the ATL domain of uniAlaRS at
196 the same alignment position. Differences in local positioning of amino acids within ATL are seen
197 by differences in ATL indexes of a common interaction. For example, local positional shifts are
198 seen in the interaction between $^{bacGly}ATL_{29}$ and $^{Ala}ATL_{28}$ with the G1 base and the discriminator
199 positions of the tRNAs (U73 in $tRNA^{Gly}$ and A73 in $tRNA^{Ala}$) and in the interaction between
200 $^{bacGly}ATL_{35}$ and $^{Ala}ATL_{34}$ with C72 and the discriminator. Base pair G1:C72 is conserved in both
201 $tRNA^{Gly}$ and $tRNA^{Ala}$. By contrast, there is no shift at index 26. Both $^{bacGly}ATL_{26}$ and $^{Ala}ATL_{26}$
202 interact with position 70, which is a C in $tRNA^{Gly}$ and a U in $tRNA^{Ala}$.

203 The $^{bacGly}$ATL interaction with tRNA appears intrinsically less stringent than that of $^{Ala}$ATL.
204 The $^{bacGly}$ATL domain contacts nucleotides of only one strand in the region of C68, C69 and C70.
205 It does not contact the opposing strand (G3, G4 or G5, Figure 3E). By contrast, $^{Ala}$ATL targets
206 both strands of $tRNA^{Ala}$, interacting with both bases of base-pairs G3:U70 and C4:G69 in part by
207 using the $^{Ala}$ATL-specific C-terminal decoration (Figure 3F).

208 The catalytic domain undergoes a radical change in global position between bacGlyRS
209 and uniAlaRS. In uniAlaRS, the catalytic and ATL domains are directly adjacent in sequence and
210 structure. By contrast, in bacGlyRS, the HD and body-tail domains are inserted between the
211 catalytic and ATL domains. The HD domain shifts and rotates the catalytic domain of bacGlyRS
212 by 130° compared to the catalytic domain of uniAlaRS (Figure 3B). To accommodate the change
213 in the position of the catalytic domain of bacGlyRS, the tRNA distorts. The conformation of the
214 four unpaired nucleotides of the acceptor arm of $tRNA^{Gly}$ bends (Figure 3A), such that the
215 position of the terminal adenosine is shifted by ~ 26 Å in the bacGlyRS complex compared to in
216 the uniAlaRS complex.

217 The conformation of the CCA tail and the position of the terminal adenosine appear to
218 be unique in bacGlyRS compared to other aaRSs. The CCA tail of $tRNA^{Gly}$ bends away from the
219 anti-codon in the complex with bacGlyRS (34), shifting the terminal A toward the minor groove
220 of the tRNA. All other class II aaRSs bend the CCA tail in the opposite direction, towards the
221 anticodon (Figure S1) or, in the case of the uniAlaRS-tRNA complex (Figures 3A and 3C),

maintain the relaxed linear structure (43). The position of the terminal adenosine is shifted by ~ 30 Å in the tRNA complex of bacGlyRS compared to in the PheRS complex (Figure S1).

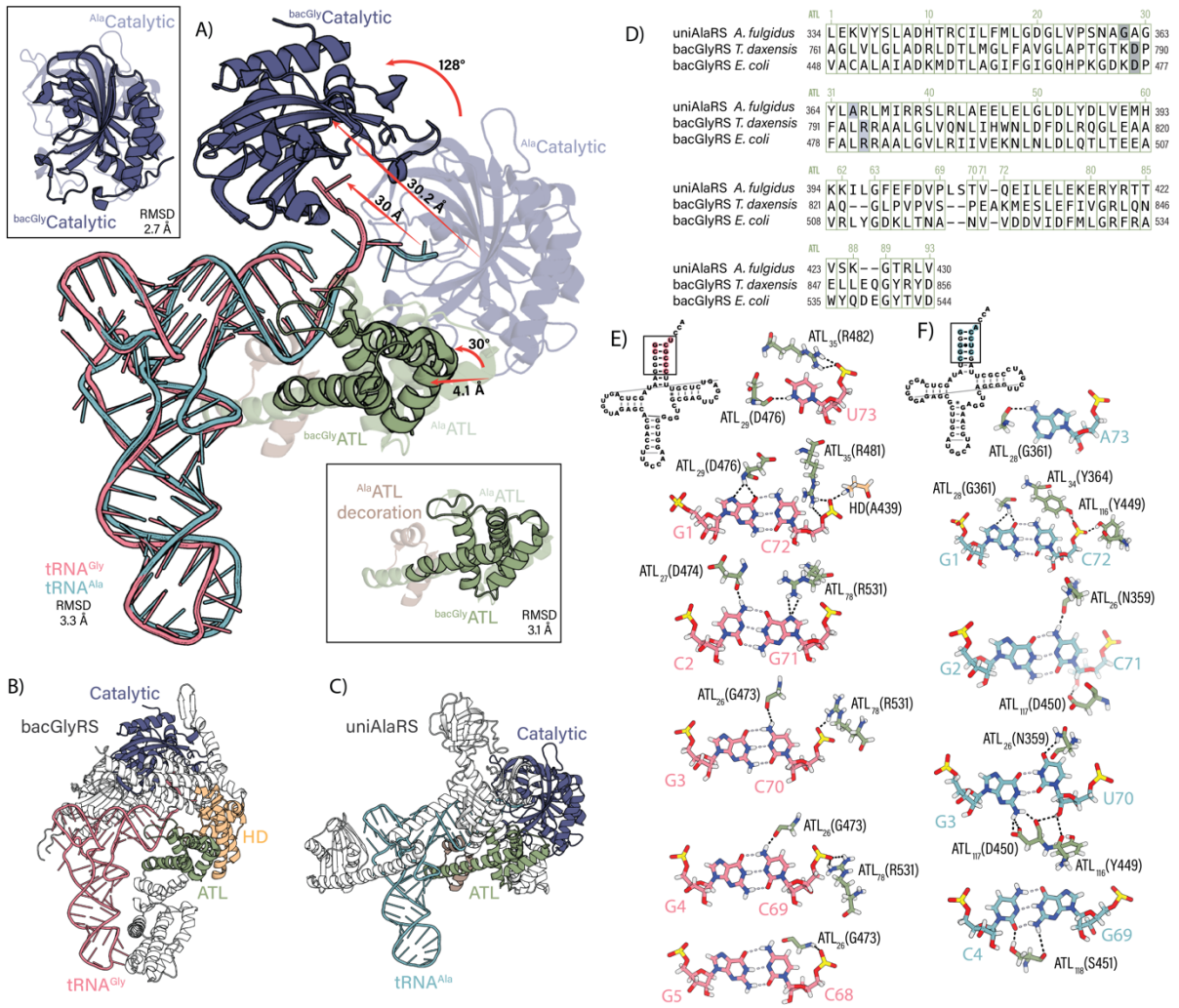


**Figure 3. Structures of bacGlyRS and uniAlaRS showing regions of homology.** (A) Comparison of the relative position of the catalytic and ATL domains of bacGlyRS (PDB: 7YSE) and uniAlaRS (PDB: 3WQY) with respect to their cognate tRNA molecules. Insets: Structure superimposition of the catalytic and ATL domains of bacGlyRS and uniAlaRS. (B) bacGlyRS of *Escherichia coli* in complex with tRNA. (C) uniAlaRS of *Archaeoglobus fulgidus* in complex with tRNA. (D) Sequence alignment and indexing of the ATL domains of uniAlaRS and bacGlyRS. (E) Hydrogen bonds between the ATL domain of bacGlyRS and its cognate tRNA. (F) Hydrogen bonds between the ATL domain of uniAlaRS and its cognate tRNA.

## The catalytic and ATL domains of bacGlyRS resemble archaeal uniAlaRS homologs

We formulated a tentative model in which bacGlyRS originated in bacteria via duplication of a two-domain uniAlaRS module. One domain is ATL and the other is the catalytic domain. This tentative model predicts that bacGlyRS component domains would be most similar to bacterial uniAlaRS homologs. Therefore, we investigated the similarity of bacGlyRS

230 domains to archaeal and bacterial uniAlaRS sequences. As shown here this model is not
231 supported by the data.

232 The sequences of the catalytic and ATL domains in bacterial uniAlaRS and archaeal
233 uniAlaRS are distinctive. A composite multiple-sequence alignment (MSA) containing both
234 bacterial and archaeal uniAlaRS sequences reveals lineage-specific signature positions and
235 insertions (Figure S2). Signature positions correspond to columns in a composite MSA that are
236 conserved within groups but are different between groups (39). Thus, differences between
237 bacterial and archaeal uniAlaRS sequences reflect deep divergence between bacteria and
238 archaea and are not due to a simple bias in amino acid usage. This type of pattern in a
239 composite MSA is called a block structure and is observed in many universal aaRS and other
240 proteins of the translation system (3; 44).

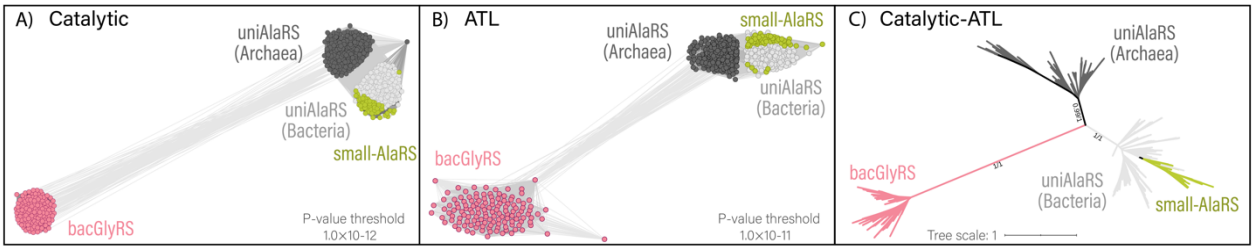


**Figure 4. Sequence analysis of bacGlyRS protein domain homologs.** Cluster map based on pairwise sequence similarities. Nodes represent sequences; connections between nodes indicate matches between pairs of sequences with a *P*-value lower than the threshold. (A) Pairwise sequence similarity between the catalytic domains of bacGlyRS and uniAlaRS. (B) Pairwise sequence similarity between the ATL domains of bacGlyRS and uniAlaRS. (C) Maximum-likelihood (LG+R4) unrooted tree of catalytic and ATL domains including bacGlyRS and uniAlaRS sequences. small-AlaRS are sequences from an anomalous subgroup of uniAaRS found in some archaea and bacteria.

241 The catalytic and ATL domains of bacGlyRS resemble archaeal uniAlaRS homologs more
242 closely than bacterial uniAlaRS homologs (Figures 4A and B). We compared the sequences of
243 the catalytic and ATL domains of bacGlyRS with those of uniAlaRS using the program CLANS (45;
244 46) and their profiles using the program HH-align (47). Profiles are calculated from multiple-
245 sequence alignments and reflect position-specific sequence variability. A cluster map of
246 sequences from the catalytic domains of bacGlyRS and uniAlaRS at a P-value lower than $10^{-12}$
247 shows connections between bacGlyRSs and archaeal uniAlaRSs. At this P-value threshold, the
248 cluster map lacks connections between bacGlyRSs and bacterial uniAlaRSs (Figure 4A). Similarly,

249 a cluster map of sequences from the ATL domains of bacGlyRS and uniAlaRS at a P-value
250 threshold of $10^{-11}$ shows connections between bacGlyRSs and archaeal uniAlaRSs. This cluster
251 map lacks connections between bacGlyRSs and bacterial uniAlaRSs (Figure 4B). Thus, the
252 catalytic and ATL domains of bacGlyRS are more similar to archaeal homologs than to bacterial
253 homologs. Therefore, our tentative model in which bacGlyRS originated by duplication of a
254 bacterial uniAlaRS module composed of catalytic and ATL domains (Figure 5D) is not supported
255 by the data (Figure 4).

## 256 The catalytic and ATL domains represent an evolutionary and functional core

257 Canonical uniAlaRS has five component domains; (i) the catalytic domain, which
258 aminoacylates the tRNA, (ii) the ATL domain, which recognizes $tRNA^{Ala}$, (iii) a cradle loop barrel
259 domain, (iv) an editing domain and (v) a C-terminal domain (C-Ala) that binds to the tRNA elbow
260 formed by the D- and T- loops (48; 33) (Figures 3C and S1). We identified a subgroup of
261 bacterial and archaeal uniAlaRS sequences that have only three of these domains: catalytic,
262 ATL, and editing (Figures 4 and S1). uniAlaRSs in this subgroup lack the cradle loop barrel
263 domain and the C-Ala domain of canonical uniAlaRSs (Figure S2). We call this anomalous
264 subgroup small-AlaRS. We identified small-AlaRS in the proteomes of the CPR bacterium
265 *Candidatus Thermoplasmatota*, and in DPANN, and Asgard groups. small-AlaRS proteins have
266 bacteria-specific insertions, suggesting that sequences were transferred to some archaea via
267 horizontal gene transfer. The maximum likelihood tree of catalytic and ATL domain sequences
268 supports a horizontal transfer hypothesis of small-AlaRS to some archaea (Figure 4C and S3).
269 Small-AlaRS sequences and recently discovered viral mini-AlaRS sequences (49), which lack the
270 cradle loop barrel, the editing and the C-Ala domains of canonical cellular uniAlaRSs, support
271 the hypothesis that a core composed of the catalytic and ATL domains can be functional.

272 The combined data support a model in which bacGlyRS and uniAlaRS emerged before
273 LUCA from common catalytic and ATL ancestors (Figure 5A and B). These catalytic and ATL
274 ancestors were single-domains and were independent of each other. Other possibilities put the
275 emergence of the bacGlyRS core after the divergence of LUCA. This late emergence of bacGlyRS
276 appears less parsimonious because it suggests that some bacteria lost the ancestral GlyRS and
277 replaced it with an RS protein that has no homologs in other bacteria. Possible scenarios are 1)

278 formation of the proto-bacGlyRS in the bacterial lineage independently from the uniAlaRS
279 catalytic-ATL core (Figure 5C); 2) formation of the proto-bacGlyRS by partial or complete
280 duplication of the uniAlaRS gene in the archaeal lineage, followed by a lateral gene transfer of
281 the proto-bacGlyRS from archaea to the bacterial lineage; and 3) formation of the proto-
282 bacGlyRS by partial or complete duplication of the uniAlaRS gene in the bacterial lineage (Figure
283 5D). This last scenario is not supported by the data (Figure 4).

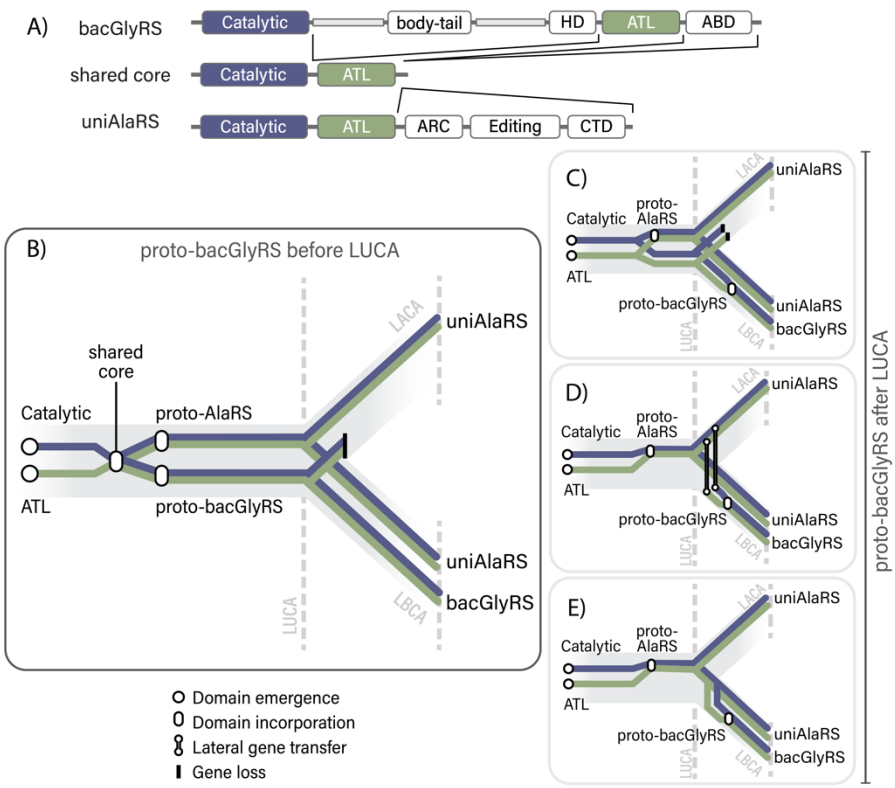


**Figure 5. Emergence of bacGlyRS through molecular bricolage.** (A) Comparative diagram of the multidomain arrangements of bacGlyRS and uniAlaRS. (B) Emergence of the common ancestor of bacGlyRS and uniAlaRS before the divergence of LUCA. Duplication of the common ancestor gives rise to the proto-uniAlaRS and proto-bacGlyRS. The ancestor of bacGlyRS is lost in archaea. (C) Emergence and duplication of the catalytic and ATL domains at LUCA. Before the divergence between archaea and bacteria, one copy of the catalytic domain and one copy of the ATL domain form the uniAlaRS core, the paralogs of the catalytic and ATL domains continue as single-domain proteins. Secondary loss of the copy of the catalytic and ATL domains in archaea. Formation of the bacGlyRS core in bacteria. (D and E) Evolution of the core of uniAlaRS before the divergence of LUCA. (D) The catalytic and ATL domains are horizontally transferred from the archaeal to the bacterial lineage and formation of the bacGlyRS core. (E) Duplication of the catalytic and ATL domains from uniAlaRS in bacteria and formation of the bacGlyRS core.

284 # Discussion

285 bacGlyRS has a deep and convoluted evolutionary history. We infer exaptation
286 relationships (50) between six bacGlyRS domains and six domains of four different RNA-

287 modifying proteins (Figure 2). Exaptation is the co-opting of an existing molecule, domain, trait
288 or system for new function. In our model, upon or after fusion of the ATL and catalytic domains
289 to form ancestors of both bacGlyRS and uniAlaRS, the catalytic domain was exapted to link
290 glycine and $tRNA^{Gly}$ to form glycyl-$tRNA^{Gly}$. The ATL domain was exapted to recognize the
291 acceptor-stem of the $tRNA^{Gly}$ (17). The prior function of the catalytic domain may have been
292 non-specific linkage of amino acids to RNAs. The prior function of ALT is unknown. After these
293 processes, insertion of the HD, body and tail domains impacted the relative positions of the ALT
294 and catalytic domains of bacGlyRS and the interaction of the catalytic domain with the CCA tail
295 of $tRNA^{Gly}$ (see below). The HD domain was exapted to interact with the discriminator position
296 of $tRNA^{Gly}$ (17). The body and tail domains of bacGlyRS were exapted to interact with the arm
297 and elbow of $tRNA^{Gly}$ (34). In the archaeal CCA-adding enzyme, the body and tail domains also
298 interact with the acceptor stem of the tRNA (51; 52). The ABD domain, which is homologous to
299 a domain in the class I uniArgRS, and shares ancestry with it, was exapted to interact with the
300 anti-codon loop of $tRNA^{Gly}$ (34).

301 Here we provide evidence that the ATL domain is a distinct, independent domain of
302 bacGlyRS rather than an integral component of the HD domain as previously proposed. This
303 distinction is important for understanding the deep evolutionary history of aaRSs. Homology
304 between the catalytic and ATL domains of bacGlyRS and the corresponding domains of uniAlaRS
305 suggests common ancestry of the core domains and functions of these enzymes. It is likely that
306 an ancestral class II catalytic domain and an ATL domain joined, and the resulting two-domain
307 aaRS gave rise to two separate enzymes, one for glycine and another for alanine. This simple
308 model is consistent with sequence and structural similarities between the core domains of
309 bacGlyRS and uniAlaRS.

310 The conservation of the pose of the ATL domain in uniAlaRS and bacGlyRS, but not the
311 pose of the catalytic domain, is consistent with a model in which the ATL domain is ancestral to
312 both uniAlaRS and bacGlyRS (Fig. 5B). It appears that the ATL domain played a role in primitive
313 RNA biology before it was incorporated into these two synthetases. We have identified four
314 adaptations of ATL that allowed a common fold in a common pose to recognize different tRNAs.
315 The ATL domain (i) slightly shifts global position relative to the acceptor stem of the tRNA, (ii)

316 shifts local positions of amino acid backbone and sidechains by conformation change, (iii)
317 changes sequence, and (iv) acquires a C-terminal decoration in the uniAlaRS that is absent from
318 the bacGlyRS.

319 bacGlyRS and uniAlaRS use different strategies to achieve final recognition of their
320 cognate tRNAs (Figure 3). ATL domains of both enzymes have similar interactions with the
321 discriminator base and the G1:C72 base pair which is conserved in $tRNA^{Gly}$ and $tRNA^{Ala}$.
322 uniAlaRS uses the ATL domain to interact intensively with the acceptor stem, employing the C-
323 terminal decoration to assist in targeting both strands of the acceptor step. The ATL domain in
324 bacGlyRS lacks C-terminal decoration and interacts less intensively with the acceptor stem. In
325 the bacGlyRS complex, the ATL domain only contacts nucleotides of one strand of the acceptor
326 stem. To compensate, bacGlyRS interacts with other regions of the tRNA. bacGlyRS but not
327 uniAlaRS exapted additional recognition domains through a process of bricolage.

328 The conformation of the CCA tail and the position of the terminal adenosine appear to
329 be unique in bacGlyRS compared to other aaRSs. In uniAlaRS, the catalytic and ATL domains are
330 directly adjacent to each other in sequence and structure. $tRNA^{Ala}$ interacts with the CCA tail of
331 uniAlaRS in a relaxed state, which is reasonably similar to the structure in the free tRNA (53). In
332 bacGlyRS, insertion of the HD domain caused a translation and rotation of the catalytic domain
333 relative to the ATL domain (Figure 3B). To accommodate this translation/rotation, the acceptor-
334 arm of $tRNA^{Gly}$ in the bacGlyRS complex bends acutely in the direction of the minor groove
335 (Figure 3C). This conformation of the tRNA appears to be unique to bacGlyRS.

336 The sequences of bacGlyRS and archaeal uniAlaRS share distinctive features. The
337 catalytic and ATL domains of bacGlyRS more closely resemble uniAlaRS archaeal homologs than
338 bacterial homologs (Figures 4A and B). Thus, a model in which bacGlyRS originated within the
339 bacterial lineage by duplication of a bacterial catalytic-ATL uniAlaRS module is not supported by
340 the data (Figure 5D). The combined results suggest that bacGlyRS and uniAlaRS share a
341 common ancestor composed of the catalytic and ATL domains (Figure 5A). The catalytic and ATL
342 domains represent an evolutionary and functional core. The domains in the shared common
343 core of bacGlyRS and uniAlaRS interact with the most ancient region of the tRNA (54; 55).

344 Here we propose a simple model, in which insertion of the body-tail and HD domains (or
345 possibly just the HD domain) into a gene closely related to the proto-AlaRS formed a novel
346 enzyme, a proto-bacGlyRS (Figure 5A). In this model, the interaction between ATL domain and
347 the tRNA remained relatively unchanged as the catalytic domain was displaced and the CCA tail
348 bends towards the minor groove. The proto-bacGlyRS enzyme underwent adaptations that
349 conferred a new amino acid specificity (Gly) and a new tRNA identity ($tRNA^{Gly}$). It is possible
350 that the new tRNA was better suited to the CCA tail orientation imposed by the HD insertion.

351 bacGlyRS is one of the largest aaRSs and is usually composed of two subunits, with a
352 catalytic domain in the α subunit and several recognition domains in the β subunit. In most
353 bacteria (Figure 1) bacGlyRS is a heterotetramer (4). In some bacteria, including
354 *Thermoanaerotrix daxensis*, the α and β subunits are linked in a single polypeptide chain (Figure
355 1); bacGlyRS with linked α and β subunits form homodimers (Figure 2A). The form of
356 oligomerization, the presence of α-β linked bacGlyRSs and the encoding of α and β bacGlyRS
357 subunits as a single (56; 57), and sometimes overlapping (Figure 1, Table S1), open reading
358 frame, could indicate that the ancestor of all bacGlyRSs was a single α-β protein that formed
359 homodimers. Fragmentation of bacGlyRS may have occurred later (57) and persists in most
360 species due to evolutionary pressures (27). Secondary fragmentation of uniAlaRS has also been
361 reported (58).

362 The distribution of bacGlyRS in a wide range of deeply rooted bacterial species suggests
363 origins before LBCA. The presence of bacGlyRS in Terrabacteria and Gracilicutes, the two major
364 bacterial groups (Figure 1) also supports this early origin. However, in the bacterial tree the
365 distribution of bacGlyRS and arcGlyRS appears patchy and irregular. The dominance of arcGlyRS
366 in CPR proteomes suggests that symbiotic bacteria with reduced genomes prefer a smaller
367 enzyme. Currently, we do not have a model that fully explains the discontinuous distribution
368 over evolutionary history.

369 Our observations here reflect deep evolutionary history. Glycine and alanine appear to
370 be ancient amino acids, both prebiotically available and biologically important. They are among
371 the most abundant amino acids produced in spark discharge experiments (59) and are found in

chondrite asteroids (60). We propose that bacGlyRS and uniAlaRS emerged from a shared core before LUCA (Figure 5A and B), from catalytic and ATL domains that had separate and distinct origins. These prebiotic amino acids share the first and third codon positions, and the proteins that aminoacylate $tRNA^{Gly}$ and $tRNA^{Ala}$ share sequence and structure relationships. It appears we are seeing evidence of an ancestral protein with non-specific catalytic activity that may have been able to link either glycine or alanine to an RNA, presumably an ancestor of tRNA. The deep relationship between alanine and glycine is seen in their shared chemistry and biology.

# Methods

## Sequence similarity search

A multiple sequence alignment (MSAs) of bacGlyRS sequences was calculated for a representative set of archaea and bacteria. The MSA was trimmed to the domain boundaries as defined by the structure analysis. The MSAs of the domains were converted to profile Hidden Markov Models using the HHsuite version 3.3.0 (31). A profile-profile similarity search on the Evolutionary Classification of Domains (ECOD) database was performed with HHsearch using the profiles of bacGlyRS domains as queries. Target domains in ECOD yielding HHsearch probabilities greater than 60% over more than 50 aligned columns were retrieved to serve as queries of a second search on a sequence database containing the proteomes of the representative set of bacterial and archaeal species. The second search was performed using jackhmmer from the HMMER suite version 3.3.1 (31), with default parameters (five iterations). The ECOD database only contains domains that are directly annotated on proteins with experimentally determined structures. The second search aims to identify sequence similarity in the UniProt Reference Cluster, UniRef90 (61); this database includes proteins without experimentally determined structures. Sequences passing the threshold values (E-value $< 1\times10^{-3}$, query coverage 80%) were retrieved and trimmed to the region of sequence similarity.

We broadly describe the phylogenetic distribution of the best hits according to the number of sampled proteomes containing homologous sequences as follows: universal (presence in > 50% of archaea and >50% of bacteria); archaea-specific (presence in >50% of

400 archaea and <10% of bacteria); bacteria-specific (presence in < 10% of archaea and <50% of
401 bacteria); lineage-specific within archaea (presence in <50% of archaea and <50% of bacteria);
402 lineage-specific within bacteria (presence in <10% of archaea and <50% of bacteria sampled); or
403 of complex distribution (presence in 10-50% of archaea and 10-50% of bacteria).

404 ## Cluster analysis

405 The query MSAs and trimmed target sequences retrieved from the second sequence
406 similarity search were clustered with CLANS (46) based on pairwise sequence similarities. To
407 find the best sequence match of each query domain, the cut-offs of the clustering were
408 adjusted to find the P-value for which the group of query sequences shows sequence
409 relationships with only one other group of orthologs. The subset of best matches was further
410 clustered to reveal possible differences between archaeal and bacterial sequences. Bacterial
411 and archaeal uniAlaRS sequences displayed differential similarity to bacGlyRS sequences on the
412 cluster analysis. To corroborate the clustering analysis results, additional analyses were
413 performed at the level of profiles.

414 ## Pairwise profile analysis: HHalign

415 An MSA containing orthologous sequences of uniAlaRS was further divided to include
416 only archaeal sequences or only bacterial sequences. The MSAs of archaeal and bacterial
417 sequences of uniAlaRS and the MSA of bacGlyRS were trimmed to the catalytic and ATL
418 domains. The MSA of the catalytic domain of bacGlyRS was aligned to the MSAs of the catalytic
419 domains of archaeal and bacterial uniAlaRS sequences using HHalign. The HHalign scores of
420 these MSA-MSA comparisons agree with the differential pairwise similarities observed in CLANS
421 between archaeal and bacterial sequences.

422 ## ML tree

423 The composite MSA of the catalytic domain containing sequences from bacGlyRS and
424 archaeal sequences from uniAlaRS was concatenated to the composite MSA of the ATL domain.
425 The concatenated MSA was trimmed with trimAl (62). A Maximum Likelihood tree was
426 calculated using the trimmed MSA using PhyML (63) and SMS model selection tool (64) on the

Montpellier Bioinformatics Platform. The tree was inferred using an LG+R4 model of evolution. Branch support corresponds to a Bayesian-like transformation of aLRT (aBayes) (65).

## Data Availability

Sequence alignments and CLANS input files associated with this manuscript have been deposited in the FigShare repository DOI: 10.6084/m9.figshare.23269007.

## Acknowledgements

The authors dedicate this work to the memory of Professor ATL who passed away on September 9th, 2021. Professor ATL worked on the structural determination of the bacterial glycyl tRNA synthetase and performed the initial structural analysis. The authors thank Jessica Bowman, Eric Smith and Arturo Becerra for discussions. This work was funded by the National Aeronautics and Space Administration grant 80NSSC18K1139. CAC was sponsored by the National Aeronautics and Space Administration (NASA) through a contract with ORAU.

*Author contributions*: Conceptualization, ATL, MIVS and CAC. Methodology, CAC. Formal analysis CAC, MA, AHG, ASP, MIVS and LDW. Writing – Original Draft, CAC and LDW. Writing – Review & Editing, ASP, LRP and MIVS.

## References

1. Bowman JC, Petrov AS, Frenkel-Pinter M, Penev PI, Williams LD (2020) Root of the Tree: The Significance, Evolution, and Origins of the Ribosome. Chem Rev 120:4848-4878.
2. Burbaum JJ, Schimmel P (1991) Structural relationships and the classification of aminoacyl-tRNA synthetases. J Biol Chem 266:16965-16968.
3. Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64:202-236.
4. Gomez MAR, Ibba M (2020) Aminoacyl-tRNA synthetases. RNA 26:910-936.
5. Giegé R, Eriani G (2023) The tRNA identity landscape for aminoacylation and beyond. Nucleic Acids Res 51:1528-1570.
6. Eriani G, Delarue M, Poch O, Gangloff J, Moras D (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature 347:203-206.
7. Moras D (1992) Structural and functional relationships between aminoacyl-tRNA synthetases. Trends Biochem Sci 17:159-164.
8. Hartman H (1995) Speculations on the evolution of the genetic code IV the evolution of the aminoacyl-tRNA synthetases. Orig Life Evol Biosph 25:265-269.

457 9. Ribas de Pouplana L, Schimmel P (2001) Aminoacyl-tRNA synthetases: potential markers of
458 genetic code development. Trends Biochem Sci 26:591-596.
459 10. Fournier GP, Andam CP, Alm EJ, Gogarten JP (2011) Molecular evolution of aminoacyl tRNA
460 synthetase proteins in the early history of life. Orig Life Evol Biosph 41:621-632.
461 11. Koonin EV, Novozhilov AS (2017) Origin and evolution of the universal genetic code. Annu
462 Rev Genet 51:45-62.
463 12. Ribas de Pouplana L. The evolution of aminoacyl-tRNA synthetases: From dawn to LUCA.
464 (2020) The Enzymes. Elsevier, pp. 11-37.
465 13. Terada T, Nureki O, Ishitani R, Ambrogelly A, Ibba M, Söll D, Yokoyama S (2002) Functional
466 convergence of two lysyl-tRNA synthetases with unrelated topologies. Nat Struct Biol
467 9:257-262.
468 14. Tumbula D, Vothknecht UC, Kim H-s, Ibba M, Min B, Li T, Pelaschier J, Stathopoulos C,
469 Becker H, Söll D (1999) Archaeal aminoacyl-tRNA synthesis: diversity replaces dogma.
470 Genetics 152:1269-1276.
471 15. Ostrem DL, Berg P (1974) Glycyl transfer ribonucleic acid synthetase from Escherichia coli:
472 purification, properties, and substrate binding. Biochemistry 13:1338-1348.
473 16. Tang SN, Huang JF (2005) Evolution of different oligomeric glycyl-tRNA synthetases. FEBS
474 Lett 579:1441-1445.
475 17. Han L, Luo Z, Ju Y, Chen B, Zou T, Wang J, Xu J, Gu Q, Yang X-L, Schimmel P (2023) The
476 binding mode of orphan glycyl-tRNA synthetase with tRNA supports the synthetase
477 classification and reveals large domain movements. Science Advances 9:eadf1027.
478 18. Logan D, Mazauric M, Kern D, Moras D (1995) Crystal structure of glycyl-tRNA synthetase
479 from Thermus thermophilus. EMBO J 14:4156-4167.
480 19. Dimas-Torres J, Rodríguez-Hernández A, Valencia-Sánchez M, Campos-Chávez E, Godínez-
481 López V, Rodríguez-Chamorro D, Grøtli M, Fleming C, Hernández-González A, Arciniega
482 M (2021) Bacterial glycyl tRNA synthetase offers glimpses of ancestral protein
483 topologies.
484 20. Jacob F (1977) Evolution and tinkering. Science 196:1161-1166.
485 21. Rossman MG, Liljas A (1974) Letter: Recognition of structural domains in globular proteins. J
486 Mol Biol 85:177-181.
487 22. Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261:552-558.
488 23. Porter LL, Rose GD (2012) A thermodynamic definition of protein domains. Proc Natl Acad
489 Sci USA 109:9420-9425.
490 24. Ponting CP, Schultz J, Copley RR, Andrade MA, Bork P (2000) Evolution of domain families.
491 Advances in protein chemistry 54:185-244.
492 25. Ohno S. 2013. Evolution by gene duplication, Springer Science & Business Media.
493 26. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99-113.
494 27. Ju Y, Han L, Chen B, Luo Z, Gu Q, Xu J, Yang XL, Schimmel P, Zhou H (2021) X-shaped
495 structure of bacterial heterotetrameric tRNA synthetase suggests cryptic prokaryote
496 functions and a rationale for synthetase classifications. Nucleic Acids Res 49:10106-
497 10119.
498 28. Smith TF, Hartman H (2015) The evolution of Class II Aminoacyl-tRNA synthetases and the
499 first code. FEBS Lett 589:3499-3507.

500 29. Valencia-Sánchez MI, Rodríguez-Hernández A, Ferreira R, Santamaría-Suárez HA, Arciniega
501 M, Dock-Bregeon A-C, Moras D, Beinsteiner B, Mertens H, Svergun D, Brieba LG, Grøtli
502 M, Torres-Larios A (2016) Structural Insights into the Polyphyletic Origins of Glycyl tRNA
503 Synthetases. J Biol Chem 291:14430-14446.
504 30. Moody ERR, Mahendrarajah TA, Dombrowski N, Clark JW, Petitjean C, Offre P, Szöllősi GJ,
505 Spang A, Williams TA (2022) An estimate of the deepest branches of the tree of life from
506 ancient vertically evolving genes. eLife 11.
507 31. Achsel T, Stark H, Lührmann R (2001) The Sm domain is an ancient RNA-binding motif with
508 oligo(U) specificity. Proc Natl Acad Sci USA 98:3685-3689.
509 32. Moody ER, Mahendrarajah TA, Dombrowski N, Clark JW, Petitjean C, Offre P, Szöllősi GJ,
510 Spang A, Williams TA (2022) An estimate of the deepest branches of the tree of life from
511 ancient vertically evolving genes. eLife 11:e66695.
512 33. Naganuma M, Sekine S, Fukunaga R, Yokoyama S (2009) Unique protein architecture of
513 alanyl-tRNA synthetase for aminoacylation, editing, and dimerization. Proc Natl Acad Sci
514 USA 106:8489-8494.
515 34. Yu Z, Wu Z, Li Y, Hao Q, Cao X, Blaha GM, Lin J, Lu G (2023) Structural basis of a two-step
516 tRNA recognition mechanism for plastid glycyl-tRNA synthetase. Nucleic Acids Res
517 51:4000-4011.
518 35. Chandonia JM, Fox NK, Brenner SE (2017) SCOPe: Manual Curation and Artifact Removal in
519 the Structural Classification of Proteins - extended Database. J Mol Biol 429:348-355.
520 36. Schaeffer RD, Liao Y, Cheng H, Grishin NV (2017) ECOD: new developments in the
521 evolutionary classification of domains. Nucleic Acids Res 45:D296-d302.
522 37. Naganuma M, Sekine S, Chong YE, Guo M, Yang XL, Gamper H, Hou YM, Schimmel P,
523 Yokoyama S (2014) The selective tRNA aminoacylation mechanism based on a single
524 G•U pair. Nature 510:507-511.
525 38. Kumar A, Åqvist J, Satpati P (2019) Principles of tRNA(Ala) Selection by Alanyl-tRNA
526 Synthetase Based on the Critical G3·U70 Base Pair. ACS Omega 4:15539-15548.
527 39. Penev PI, Alvarez-Carreño C, Smith E, Petrov AS, Williams LD (2021) TwinCons: Conservation
528 score for uncovering deep sequence similarity and divergence. PLoS Comput Biol
529 17:e1009541.
530 40. McClain WH, Foss K, Jenkins R, Schneider J (1991) Four sites in the acceptor helix and one
531 site in the variable pocket of tRNA (Ala) determine the molecule's acceptor identity.
532 Proc Natl Acad Sci USA 88:9272-9276.
533 41. Chong YE, Guo M, Yang X-L, Kuhle B, Naganuma M, Sekine S-i, Yokoyama S, Schimmel P
534 (2018) Distinct ways of G: U recognition by conserved tRNA binding motifs. Proc Natl
535 Acad Sci USA 115:7527-7532.
536 42. Nagato Y, Yamashita S, Ohashi A, Furukawa H, Takai K, Tomita K, Tomikawa C (2023)
537 Mechanism of tRNA recognition by heterotetrameric glycyl-tRNA synthetase from lactic
538 acid bacteria. The Journal of Biochemistry:mvad043.
539 43. Mohan S, Hsiao C, VanDeusen H, Gallagher R, Krohn E, Kalahar B, Wartell RM, Williams LD
540 (2009) Mechanism of RNA double helix-propagation at atomic resolution. J Phys Chem B
541 113:2614-2623.

542 44. Vishwanath P, Favaretto P, Hartman H, Mohr SC, Smith TF (2004) Ribosomal protein-
543 sequence block structure suggests complex prokaryotic evolution with implications for
544 the origin of eukaryotes. Mol Phylogenet Evol 33:615-625.
545 45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search
546 tool. J Mol Biol 215:403-410.
547 46. Frickey T, Lupas A (2004) CLANS: a Java application for visualizing protein families based on
548 pairwise similarity. Bioinformatics 20:3702-3704.
549 47. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J (2019) HH-suite3
550 for fast remote homology detection and deep protein annotation. BMC Bioinformatics
551 20:473.
552 48. Swairjo MA, Otero FJ, Yang X-L, Lovato MA, Skene RJ, McRee DE, de Pouplana LR, Schimmel
553 P (2004) Alanyl-tRNA synthetase crystal structure and design for acceptor-stem
554 recognition. Mol Cell 13:829-841.
555 49. Antika TR, Chrestella DJ, Tseng YK, Yeh YH, Hsiao CD, Wang CC (2023) A naturally occurring
556 mini-alanyl-tRNA synthetase. Commun Biol 6:314.
557 50. Frenkel-Pinter M, Petrov AS, Matange K, Travisano M, Glass JB, Williams LD (2022)
558 Adaptation and Exaptation: From Small Molecules to Feathers. J Mol Evol 90:166-175.
559 51. Schimmel P, Yang XL (2004) Two classes give lessons about CCA. Nat Struct Mol Biol 11:807-
560 808.
561 52. Tomita K, Fukai S, Ishitani R, Ueda T, Takeuchi N, Vassylyev DG, Nureki O (2004) Structural
562 basis for template-independent RNA polymerization. Nature 430:700-704.
563 53. Westhof E, Dumas P, Moras D (1988) Restrained refinement of two crystalline forms of
564 yeast aspartic acid and phenylalanine transfer RNA crystals. Acta Crystallogr A 44 ( Pt
565 2):112-123.
566 54. Schimmel P, Giegé R, Moras D, Yokoyama S (1993) An operational RNA code for amino acids
567 and possible relationship to genetic code. Proc Natl Acad Sci USA 90:8763-8768.
568 55. Schimmel P, de Pouplana LR (1995) Transfer RNA: from minihelix to genetic code. Cell
569 81:983-986.
570 56. Keng T, Webster TA, Sauer RT, Schimmel P (1982) Gene for Escherichia coli glycyl-tRNA
571 synthetase has tandem subunit coding regions in the same reading frame. J Biol Chem
572 257:12503-12508.
573 57. Wagar EA, Giese MJ, Yasin B, Pang M (1995) The glycyl-tRNA synthetase of Chlamydia
574 trachomatis. J Bacteriol 177:5179-5185.
575 58. Arutaki M, Kurihara R, Matsuoka T, Inami A, Tokunaga K, Ohno T, Takahashi H, Takano H,
576 Ando T, Mutsuro-Aoki H (2020) G: U-independent RNA minihelix aminoacylation by
577 Nanoarchaeum equitans alanyl-tRNA synthetase: an insight into the evolution of
578 aminoacyl-tRNA synthetases. J Mol Evol 88:501-509.
579 59. Parker ET, Cleaves HJ, Dworkin JP, Glavin DP, Callahan M, Aubrey A, Lazcano A, Bada JL
580 (2011) Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark
581 discharge experiment. Proc Natl Acad Sci USA 108:5526-5531.
582 60. Botta O, Glavin DP, Kminek G, Bada JL (2002) Relative amino acid concentrations as a
583 signature for parent body processes of carbonaceous chondrites. Orig Life Evol Biosph
584 32:143-163.

585 61. Lussi YC, Magrane M, Martin MJ, Orchard S, Consortium U (2023) Searching and Navigating
586 UniProt Databases. Current Protocols 3:e700.
587 62. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated
588 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973.
589 63. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms
590 and methods to estimate maximum-likelihood phylogenies: assessing the performance
591 of PhyML 3.0. Syst Biol 59:307-321.
592 64. Lefort V, Longueville JE, Gascuel O (2017) SMS: Smart Model Selection in PhyML. Mol Biol
593 Evol 34:2422-2424.
594 65. Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O (2011) Survey of branch support
595 methods demonstrates accuracy, power, and robustness of fast likelihood-based
596 approximation schemes. Syst Biol 60:685-699.

597