Supporting Information For:

Ancestral Archaea Expanded the Genetic Code with Pyrrolysine

Li-Tao Guo[1, †], Kazuaki Amikura[1,3 †], Han-Kai Jiang[4, 5, 6 †], Takahito Mukai[7], Xian Fu[8, 9], Yane-Shih Wang[4, 5, 10], Patrick O'Donoghue[11, 12], Dieter Söll[1, 2], Jeffery M. Tharp[1,13*]

[1]Department of Molecular Biophysics & Biochemistry, [2]Department of Chemistry, Yale University, New Haven, CT, USA
[3]Department of Interdisciplinary Space Science, Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency, Kanagawa, Japan
[4]Institute of Biological Chemistry, [5]Chemical Biology and Molecular Biophysics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan
[6]Department of Chemistry, National Tsing Hua University, Hsinchu, Taiwan
[7]Department of Life Science, College of Science, Rikkyo University, Tokyo, Japan
[8]BGI-Shenzhen, Shenzhen, China
[9]Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen, China
[10]Institute of Biochemical Sciences, National Taiwan University, Taipei, Taiwan
[11]Department of Biochemistry, [12]Department of Chemistry, The University of Western Ontario, London, Canada
[13]Current Address: Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA

[†]These authors contributed equally to this work.

* Correspondence should be addressed to Jeffery M. Tharp (jemtharp@iu.edu).

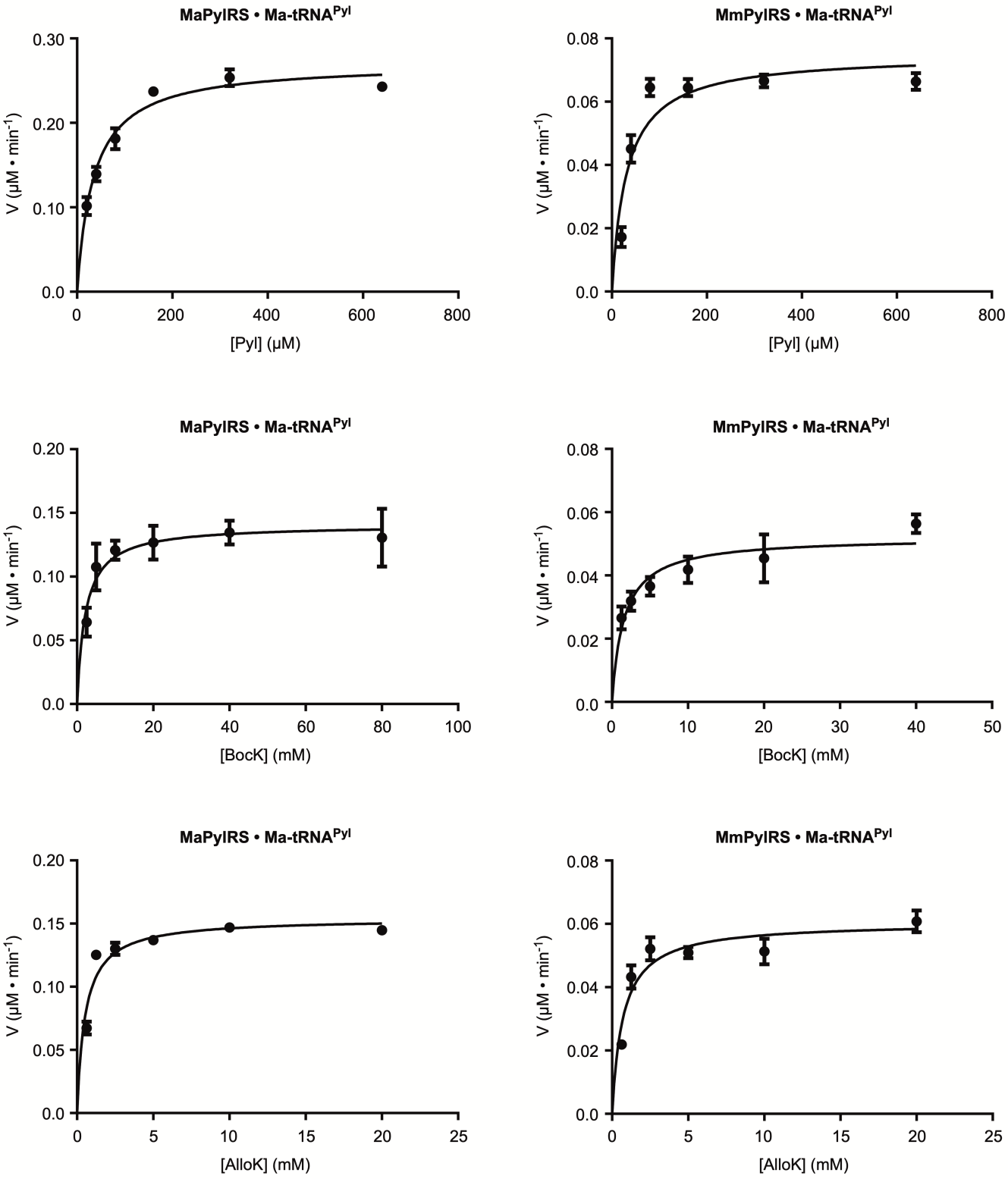


Figure S1. Kinetic plots for the aminoacylation of tRNA^{Pyl} by PylRS. Aminoacylation data are shown for L-pyrrolysine and two ncAA substrates (BocK and AlloK). Data were collected using 1 μM of purified recombinant enzymes and tRNA transcripts as described in the Experimental Procedures. Data points indicate the mean of three replicates. Error bars represent one standard deviation of the mean.

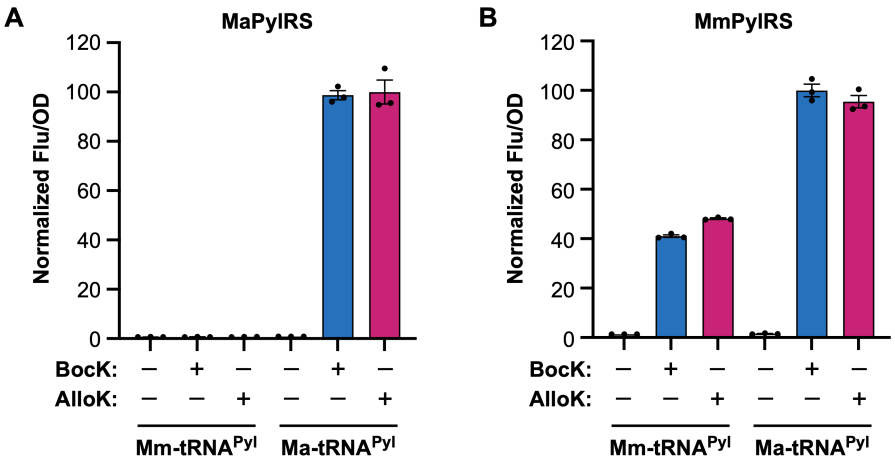


Figure S2. Activity of MaPylRS and MmPylRS with homologous and nonhomologous pyrrolysine tRNAs. PylRS activity was quantified by measuring *in vivo* UAG suppression using a superfolder green fluorescent protein reporter harboring a UAG codon at position 2. Relative fluorescence intensities were measured in *E. coli* DH10B cells co-expressing MaPylRS (*A*) or MmPylRS (*B*) and the pyrrolysine tRNA from *M. mazei* (Mm-tRNA^{Pyl}), or *Ca.* M. alvus (Ma-tRNA^{Pyl}). The data show that MaPylRS is only active with its homologous tRNA, whereas, MmPylRS is active with both Ma-tRNA^{Pyl} and Mm-tRNA^{Pyl}. Amber suppression increased ~2-fold when MmPylRS was paired with Ma-tRNA^{Pyl}, despite this pair showing poor *in vitro* aminoacylation activity. Data are presented as the mean ± SEM of three biological replicates.

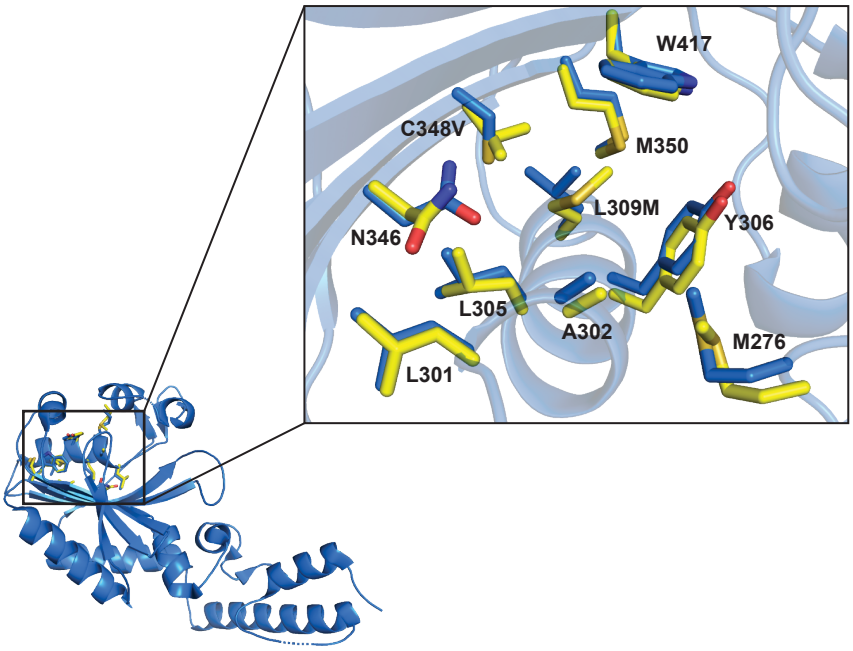


Figure S3. Alignment of the MmPylRS and MaPylRS amino acid binding pockets. The crystal structures of MmPylRS (2q7h) and MaPylRS (6jp2) were aligned using PyMOL. The MmPylRS structure is shown in blue. The inlay shows residues in the amino acid binding pocket, which are numbered according to the sequence of MmPylRS. Residues in the MaPylRS amino acid binding pocket are shown in yellow. The binding pockets of the two enzymes differ only at positions L309 and C348.

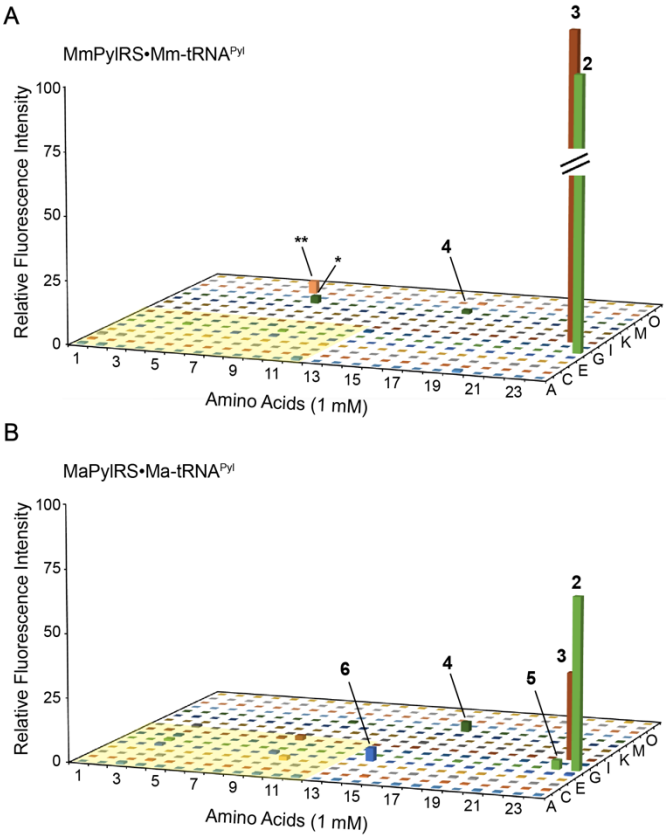


Figure S4. The substrate range of wild type PylRS from *M. mazei* and *Ca.* M. alvus measured with the sfGFP-2am reporter. Substrate recognition was determined by measuring *in vivo* amber suppression using a superfolder green fluorescent protein reporter harboring a UAG codon at position 2. Relative fluorescence intensities were measured in *E. coli* cells, co-expressing the MmPylRS•Mm-tRNA^{Pyl} pair (*A*) or the MaPylRS•Ma-tRNA^{Pyl} pair (*B*), in GMML medium supplemented with 1 mM of a noncanonical amino acid. The highlighted wells indicate phenylalanine analogs. Numbers on the bars correspond to the ncAAs shown in **Fig. 1**. * and ** indicate lysine or glutamine incorporation in the presence of (3*S*)-1, 2, 3, 4-Tetrahydroisoquinoline-7hydroxy-3-carboxylic acid or (3*R*)-1, 2, 3, 4-Tetrahydroisoquinoline-7-hydroxy-3-carboxylic acid, as determined by intact protein mass spectrometry.

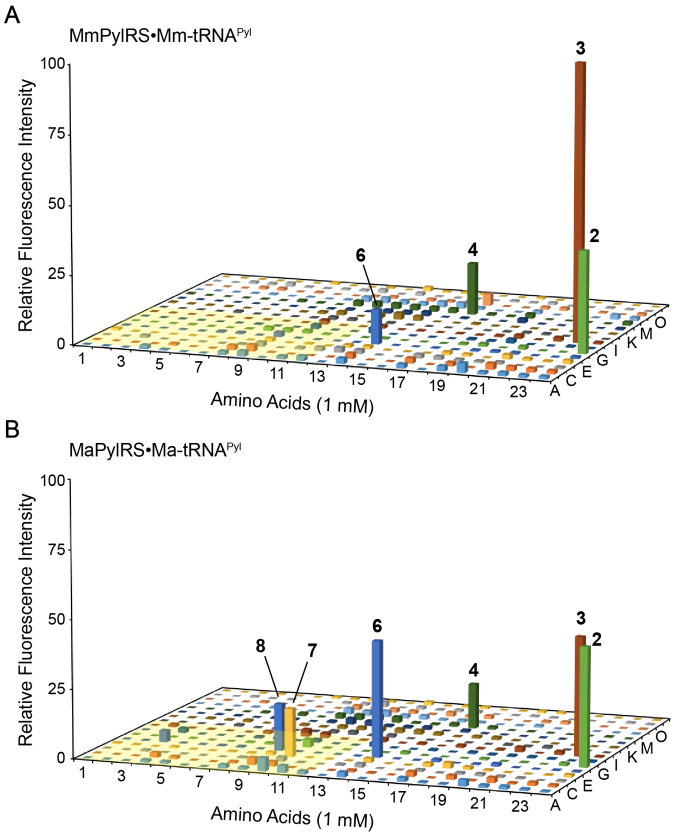
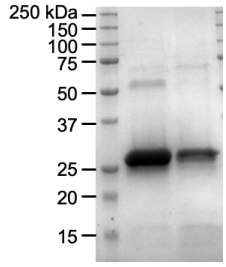


Figure S5. The substrate range of wild type PylRS from *M. mazei* and *Ca.* M. alvus measured with the sfGFP-27am reporter. Substrate recognition was determined by measuring *in vivo* amber suppression using a superfolder green fluorescent protein reporter harboring a UAG codon at position 27. Relative fluorescence intensities were measured in *E. coli* cells, co-expressing the MmPylRS•Mm-tRNA^{Pyl} pair (*A*) or the MaPylRS•Ma-tRNA^{Pyl} pair (*B*), in GMML medium supplemented with 1 mM of a noncanonical amino acid. The numbers on the bars correspond to the ncAAs shown in **Fig. 1**. The highlighted wells indicate phenylalanine analogs.

A



B

Replicate	Yield (mg/L)	
–	MaPylRS	MmPylRS
1	9.8	3.8
2	13.7	3.6
3	13.2	5.6
Average	12.2 ± 2.1	4.3 ± 1.1

Figure S6. The yield of sfGFP-27am expressed with MmPylRS•Mm-tRNA^{Pyl} and MaPylRS•Ma-tRNA^{Pyl}. (*A*) SDS-PAGE analysis of sfGFP-27am expressed with 1 mM BocK using the MaPylRS•Ma-tRNA^{Pyl} pair (Lane 1) and the MmPylRS•Mm-tRNA^{Pyl} pair (Lane 2). (*B*) The yield (mean ± SD) of sfGFP-27am obtained from three independent experiments using biological replicates. The protein was expressed and purified from *E. coli* BL21 (DE3) as described in the Experimental Procedures.

Figure S7. Phylogenetic tree based on 16S rRNA of 122 putative PylRS-encoding archaea and bacteria. The maximum-likelihood analysis was performed with 100 replicates using MEGAX. Branches with bootstrap support values greater than 75% are indicated with a black circle. The tree is rooted between bacterial and archaeal sequences. Colored circles indicate the domain architecture of the encoded PylRS enzymes.

Figure S8. Maximum-likelihood tree constructed using 156 PylSc protein sequences from bacteria and archaea. The tree was constructed with 100 replicates using MEGAX. Branches with bootstrap support values greater than 75% are indicated with a black circle. Colored circles indicate the domain architecture of the PylRS enzyme. PheRS sequences were used as an outgroup to root the tree. PheRS and PylRS sequences are given in **Supplementary File 2**.

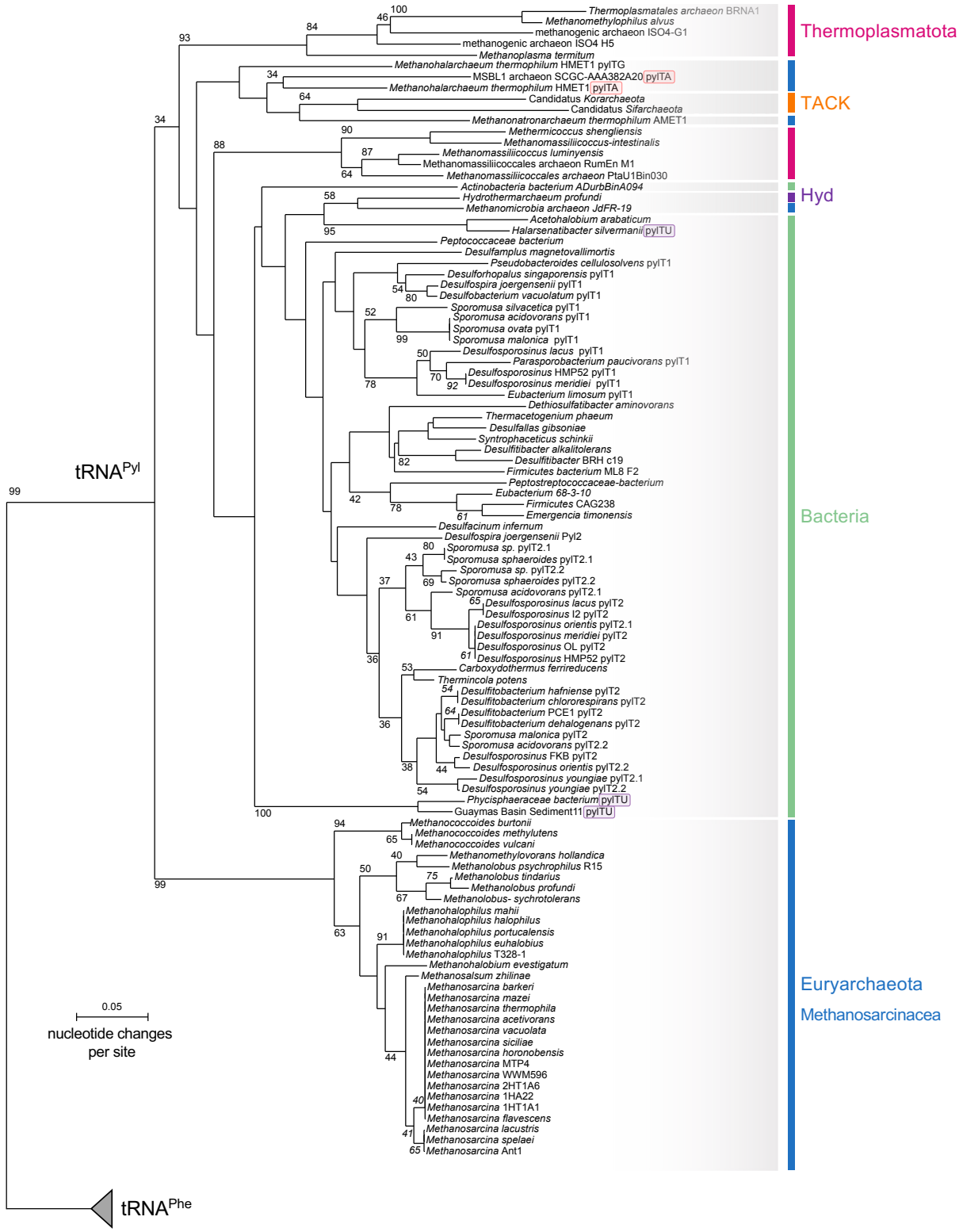


Figure S9. Distance-based phylogeny of tRNA^{Pyl}. The distance-based tree was computed from an alignment of 260 tRNA^{Phe} sequences and 105 tRNA^{Pyl} sequences. The phylogeny was calculated using the PhyML BioNJ algorithm PhyML in SeaView , and bootstrap supports were based on BioNJ trees calculated using 1000 pseudo-replicate data sets. Bootstrap values > 30/100 are shown. The position of the outgroup tRNA^{Phe} sequences is indicated (gray triangle), and sub-clades among the tRNA^{Pyl} sequences are highlighted, including the *Mathanosarcinaceae* (blue), TACK group (orange), Thermoplasmatota (red), and bacteria (green). tRNA^{Pyl} and tRNA^{Phe} sequences are given in **Supplementary File 3**.

Figure S10. Maximum-likelihood phylogeny of tRNA^{Pyl}. The maximum likelihood tree was computed from an alignment of 260 tRNA^{Phe} sequences and 105 tRNA^{Pyl} sequences. The phylogeny was calculated based on 100 randomly generated starting trees using the PhyML in SeaView . Bootstrap supports were calculated based on the Shimodaira-Hasegawa (1) approximate likelihood-ratio test (SH-aLRT) in PhyML . Bootstrap values > 30/100 are shown. The position of the outgroup tRNA^{Phe} sequences is indicated (gray triangle), and sub-clades among the tRNA^{Pyl} sequences are highlighted, including the *Methanosarcinaceae* (blue), TACK group (orange), Thermoplasmatota (red), and bacteria (green).

Figure S11. Comparison of the amino acid binding pocket of PylRS enzymes. The amino acid binding pocket (including residues M276, L301, A302, L305, Y306, L309, N346, C348, M350, Y384, V401, and W417, numbered according to the *M. mazei* PylRS sequence) of 156 putative PylRS enzymes were compared. Residues lining the amino acid binding pocket were highly conserved with the exception of L309, C348, and M350. The identity of the amino acids at these three positions is given as colored letters next to the species name in the above maximum-likelihood tree, which is derived from Fig. S8.

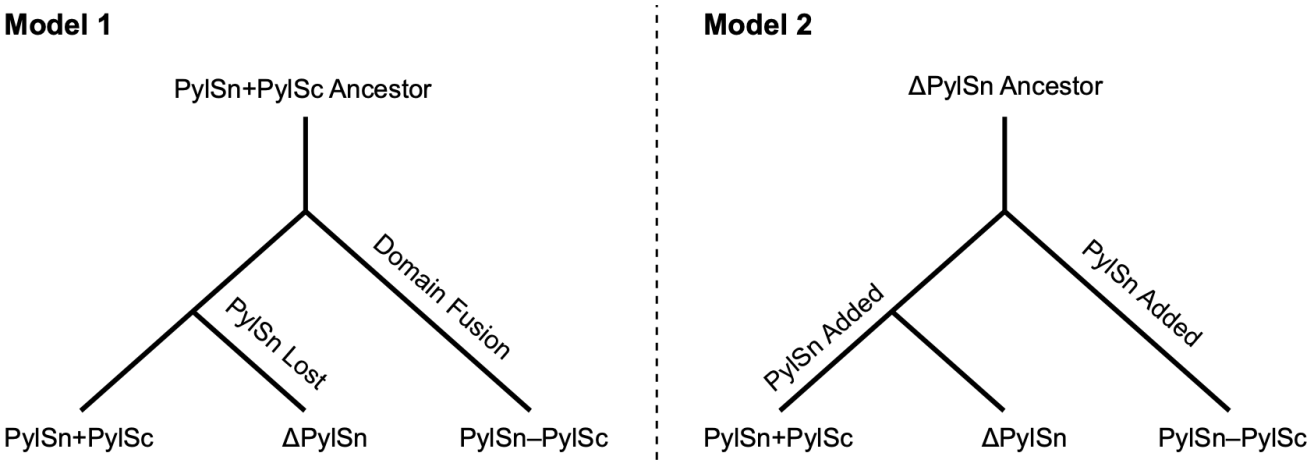


Figure S12. Two models of PylRS evolution. In Model 1, extant PylRS enzymes are derived from a PylSn+PylSc ancestor. In Model 2, extant enzymes are derived from a ΔPylSn ancestor. Model 2 requires that PylSn emerged twice during PylRS evolution. Therefore, the more parsimonious model, Model 1, is favored based on currently available data.

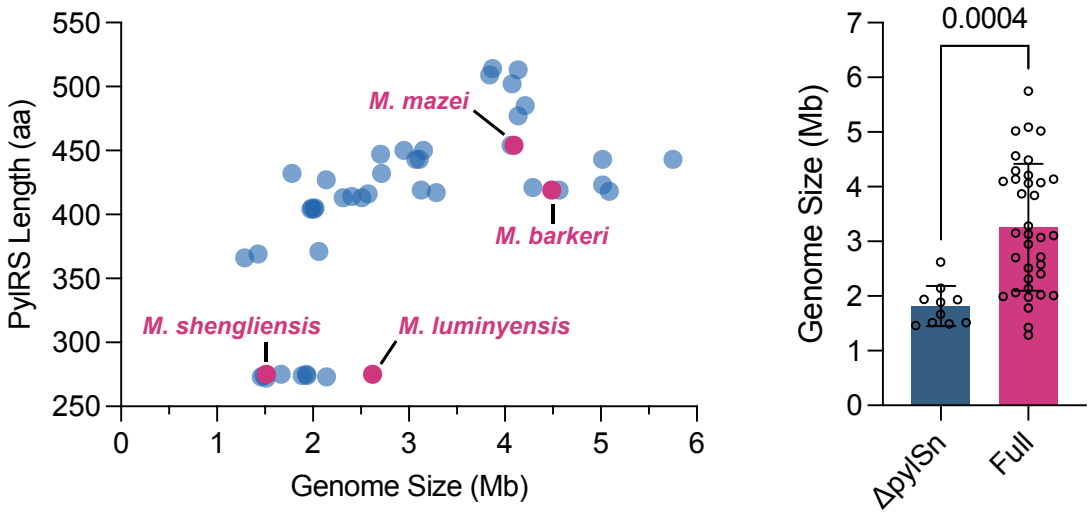


Figure S13. A comparison of genome size and PylRS length. On average, genomes of ΔPylSn PylRS-encoding organisms are 1.8-fold smaller than other PylRS encoding archaea (p = 0.0004, unpaired t-test). Only genomes with a CheckM completeness score of greater than 90% were included in the analysis. Error bars represent one standard deviation of the mean.

References

1. Shimodaira, H. (2001) Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. *Commun. Stat. - Theor. Met.* **30**, 1751–1772.