



HAL
open science

Pyrrolysine in archaea: a 22nd amino acid encoded through a genetic code expansion

Jean-François Brugere, John F. Atkins, Paul W. O'Toole, Guillaume Borrel

► **To cite this version:**

Jean-François Brugere, John F. Atkins, Paul W. O'Toole, Guillaume Borrel. Pyrrolysine in archaea: a 22nd amino acid encoded through a genetic code expansion. *Emerging Topics in Life Sciences*, 2018, 2 (4), pp.607 - 618. 10.1042/etls20180094 . hal-02996724

HAL Id: hal-02996724

<https://hal.science/hal-02996724>

Submitted on 9 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 **Pyrrolysine in Archaea, a 22nd amino acid encoded through a genetic code expansion**

2

3 Jean-François Brugère^{1,2}, John F. Atkins^{3,4}, Paul W. O'Toole² & Guillaume Borrel⁵

4

5 ¹ Université Clermont Auvergne, Clermont-Ferrand, France

6 ² School of Microbiology and APC Microbiome Institute, University College Cork, Cork,

7 Ireland

8 ³ School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

9 ⁴ Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

10 ⁵ Evolutionary Biology of the Microbial Cell, Department of Microbiology, Institut Pasteur,

11 Paris, France

12

13

14 Correspondence to J.-F. B.: 5e CBRV, Facultés de Médecine et de Pharmacie, Université

15 Clermont Auvergne, Place Henri Dunant, F-63000 Clermont-Ferrand, France

16 e-mail address: J-Francois.Brugere@uca.fr

17 **ABSTRACT:**

18 The 22nd amino acid discovered to be directly encoded, pyrrolysine, is specified by UAG. Until
19 recently, pyrrolysine was only known to be present in archaea from a methanogenic lineage
20 (*Methanosarcinales*) where it is important in enzymes catalysing anoxic methylamines
21 metabolism, and a few anaerobic bacteria. Relatively new discoveries have revealed wider
22 presence in archaea, deepened functional understanding, shown remarkable carbon source-
23 dependent expression of expanded decoding and extended exploitation of the pyrrolysine
24 machinery for synthetic code expansion. At the same time other studies have shown the
25 presence of pyrrolysine-containing archaea in the human gut and this has prompted health
26 considerations. The article reviews our knowledge of this fascinating exception to the
27 “standard” genetic code.

28 **1. Introduction**

29 It is now more than half a century ago since the genetic code was deciphered (1,2). The
30 understanding gained of how just four different nucleobases could specify the 20 universal
31 amino acids to synthesize effective proteins reveals a fundamental feature of extant life. It is
32 a remarkable feat of information transfer. Correct aminoacylation, supported by aminoacyl-
33 tRNA synthetases (aaRS), and highly discriminatory selection of tRNA whose anticodon is
34 complementary to the codon, are essential for faithful decoding.

35 Deviations from the genetic code known in most organisms have been gathered under two
36 different types termed “reassignment” and “recoding” (3). In specialized niches, such as
37 certain mitochondria (4,5), the meaning of particular codons is reassigned such that
38 wherever that codon occurs it has the new meaning, and only the new meaning. Many but
39 not all of these reassignments involve UAG or UGA encoding a specific one of the 20
40 universally encoded amino acids rather than signifying translation termination. This is
41 distinct from context dependent codon redefinition where the new meaning only applies at
42 particular occurrences of the codon. Such redefinition is dynamic with specification of the
43 new meaning being in competition with the standard meaning so that only a proportion of
44 the product reflects the new meaning. It is a type of recoding. When dynamic redefinition
45 occurs at UAG or UGA, which are by far the most frequent codons at which it occurs, one, or
46 several different, of the 20 universal amino acids can be specified. Such occurrences are
47 often termed stop codon readthrough as the identity of the amino acid specified is generally
48 unimportant and functional significance derives from a proportion of the product having a C-
49 terminal extension.

50 In addition to the 20 amino acids specified in all organisms, 2 additional amino acids,
51 pyrrolysine and selenocysteine, are directly encoded by some organisms but not by others.

52 In the organisms that encode selenocysteine, which includes several archaea (6), such
53 specification is, with few though interesting exceptions (7), by UGA (8,9) and is mRNA
54 context dependent. With just one exception (the 3' region of selenoprotein P mRNA), even
55 when such specification does occur, it is in competition with the release factor mediated
56 termination i.e. the redefinition is dynamic. In contrast, when the 22nd directly encoded
57 amino acid, pyrrolysine, was discovered by Krzycki and colleagues (10,11) to be encoded by
58 UAG in an archaeon, an early question was whether or not all ribosomes decoded all UAG
59 codons in that organism as pyrrolysine, i.e. whether UAG was reassigned to specify
60 pyrrolysine. One alternative considered was that all UAG codons present were sometimes
61 decoded as either pyrrolysine or termination, and another was that a proportion of UAG
62 codons were more prone than others to specify pyrrolysine because of a context feature of
63 their encoding UAGs (a particular mRNA stem loop 3' of such UAG codons). At issue is
64 whether, unlike the selenocysteine specification, pyrrolysine is sometimes specified
65 incidentally and the only selective advantage is a continuous synthesis (UAG reading)
66 without importance to the identity of the amino acid incorporated. Pertinent to this issue is
67 the frequency of utilized UAGs in mRNAs in organisms that encode pyrrolysine. While it was
68 infrequent in the initially studied pyrrolysine-containing archaea, it was frequent in the
69 anaerobic bacteria that also encode pyrrolysine. As discussed below some of the dilemma
70 posed led to an exciting regulatory discovery (12). Related to the natural expansion of
71 genetic decoding exemplified by pyrrolysine encoding, was the excitement that arose with
72 the realization that the nature of pyrrolysine tRNA and of its aminoacylation could be
73 exploited for synthetic code expansion. Another reason for high initial interest in pyrrolysine
74 was the functional role it plays in methylamine-dependent methyltransferases. Recent
75 discoveries of additional organisms, especially archaea that utilize pyrrolysine with

76 distinctive traits have substantially extended appreciation of its importance (e.g. climate
77 change and biogenic methane production or potential biofuel production; human health and
78 realization of potential importance of pyrrolysine in gut microbes metabolism), making a
79 review timely.

80

81 **2. Pyrrolysine, a non-dispensable AA linked to anaerobic methylamines metabolism**

82 From a biochemical perspective Pyl ($C_{12}H_{21}N_3O_3$; N^6 -{[(2*R*,3*R*)-3-methyl-3,4-dihydro-2*H*-
83 pyrrol-2-yl]carbonyl}-L-lysine) is a typical L-lysine amino acid to which a pyrrole ring is
84 branched on the lateral chain through an amide bond (biosynthesis in section 3.1). However,
85 while L-lysine derivatives are already known in proteins, encompassing those from archaea
86 like hypusine or methyllysine (13), Pyl does not originate from post-translational
87 modifications of L-lysine, but is translationally incorporated, representing the 22nd
88 proteinogenic amino acid. This presence has been mostly documented in three different
89 kinds of proteins (MtmB, MtbB and MttB) which are methyltransferases involved in the
90 metabolism of respectively mono-, di- and tri- methylamine (abbreviated in MMA, DMA and
91 TMA). Until recently, the only other known Pyl-containing proteins were certain
92 transposases (14) and a tRNA^{His}-guanylyltransferase Thg1 (15) both present in a subset of
93 *Methanosarcinales*. In methyltransferases, Pyl is found in the active site, thus arguing for a
94 non-dispensable role for biological activity. It captures methylamines (through its imine
95 electrophilic group) before transferring one methyl-group to a Co(I)-corrinoid cofactor of an
96 associated protein (MtmC/MtbC/MttC) (11). Recently, this non-dispensable role of Pyl
97 towards methylamines has been strengthened by the discovery of natural MttB analogs
98 without Pyl (16). These homologues of MttB have been identified in multiple bacteria as well
99 as methanogenic and non-methanogenic archaea. One of these enzymes (named MtgB) has

100 been characterized in *Desulfitobacterium hafniense* YF51 and catalyses the transfer of a
101 methyl group from glycine-betaine, but not trimethylamine, to a corrinoid protein. In
102 consequence, Pyl in archaea seems tightly linked to methanogenesis from methylamines
103 (MMA, DMA, TMA) and is non-dispensable for such a metabolism (17).

104

105 **3. Components of the Pyl system**

106 The first indication of unusual decoding of the *mtmB* gene of the archaeon *Methanosarcina*
107 *barkeri* was that its mRNA contained an in-frame UAG codon that was not interpreted as a
108 stop signal during translation (18). This was also the case in other *Methanosarcinales*, certain
109 anaerobic bacteria and even for genes encoding the different protein families, MtbB and
110 MttB (19). The crystal structure of *Methanosarcina barkeri* MtmB1 revealed the presence of
111 an unknown amino acid, which was named L-pyrrolysine (11). It was determined that it relies
112 on an UAG decoding tRNA (Pyl^T) charged by Pyl (10). With specification of a new non-
113 universally encoded amino acid, this dramatically showed a natural expansion of the genetic
114 code (20).

115 Pyl synthesis and insertion only relies on 5 genes, *pylTSBCD* (or 6 when tRNA^{Pyl} synthetase is
116 encoded by two genes, see below). In most cases, these genes are gathered in an operon-
117 like structure as shown in Figure 1 (light grey box, upper left) (21,22). However, some
118 exceptions to such genomic organisation have emerged (see some examples upper part of
119 Figure 1) (21,23). Evidence of this minimum set of genes was obtained from generating in *E.*
120 *coli* an MtmB protein containing Pyl at the correct location by co-expressing *pylTSBCD* from
121 *Methanosarcina barkeri* (24). Figure 1 gives a schematic overview of the Pyl system based on
122 the decoding of UAG as Pyl in the MttB protein of “*Candidatus Methanomethylophilus*
123 *alvus*”, a *Methanomassiliicoccales* representative.

124

125 3.1. Synthesis of Pyl from a standard AA

126 The synthesis of pyrrolysine relies on three genes, *pyBCD* (22). It was initially hypothesized
127 that L-lysine was one of the precursors, with the pyrrole ring added on its acyl chain derived
128 from either L-isoleucine, L-proline, L-methionine, L-glutamate or D-ornithine. Indeed,
129 supplying D-ornithine stimulates the readthrough of UAG codon of a reporter gene in *E. coli*
130 expressing *pyITSBCD* (25). D-ornithine supply leads however to a desmethylpyrrolysine
131 insertion instead of Pyl due to the respective enzymatic role of *pyICD* products that were
132 identified (26). This suggested that PylB activity was mandatory to get Pyl and led Krzycki
133 and colleagues to postulate that the reaction product catalysed by PylB was (3*R*)-3-methyl-
134 D-ornithine. To obtain such a product, they demonstrated that PylB was an L-lysine mutase
135 inverting the chiral centres in the pyrrole ring (27). Therefore, Pyl has only one kind of
136 precursor, which is surprisingly the common proteinogenic amino acid, L-lysine. It is used
137 twice in a three-reaction pathway (Figure 2). This pathway encompasses first the L-lysine
138 mutase PylB, a radical S-adenosylmethionine (SAM) protein with a TIM barrel (28). The
139 amide bond between a second L-lysine (at ϵ location) is then catalysed by PylC to form (3*R*)-
140 3-methyl-D-ornithine. Finally, PylD initiates an (NAD dependent) oxidation of this molecule
141 to form a semi-aldehyde, intermediary compound that self-reacts to form the pyrrole ring
142 by dehydration (Figure 2).

143

144 3.2. Linking Pyl to a specifically dedicated tRNA

145 The two other genes of the Pyl system, *pyIT* and *pyIS*, encode respectively a tRNA whose
146 anticodon is complementary to the UAG codon, and the subunit (PylS) of the tRNA^{Pyl}
147 synthetase. This tRNA synthetase / tRNA pair possesses a strict orthogonality towards the

148 other 20 proteinogenic amino acids, leading specifically to link Pyl to this dedicated tRNA
149 (10). The structural analysis of near complete PylS (PylSc from the bacterium
150 *Desulfitobacterium hafniense*, see below) was deduced either from the crystallisation of the
151 apoprotein alone or from its co-crystallisation with tRNA^{Pyl}. It revealed a pair highly distinct
152 from the other ones, even the most closely related ones (29), with both tRNA^{Pyl} synthetase
153 and tRNA^{Pyl} having original features compared to other known equivalents. Moreover,
154 differences are also observable among Pyl-decoding bacteria and archaea, for their
155 (unusually small) tRNA^{Pyl} and for PylS (see below).

156 The active tRNA^{Pyl} synthetase is a homodimer belonging to the class II family of aaRS. While
157 being very different from other class II aaRS, its catalytic core shares common traits with
158 tRNA^{Phe} synthetase, suggesting that these two aaRS are evolutionarily related (section 4.3).
159 This tRNA synthetase is encoded by a unique gene in *Methanosarcinales* (*pylS*) and two in
160 bacteria, one encoding the Cter portion of PylS (*pylSc*) and one encoding ~110 amino acids of
161 PylS corresponding to the Nter portion of *Methanosarcinales* PylS (*pylSn*). The N-ter part of
162 *Methanosarcinales* PylS (homolog to PylSn) is dispensable for correct aminoacylation *in*
163 *vitro*, but is necessary for correct activity *in vivo* (30). Due to its highly hydrophobic nature,
164 likely preventing its *in-vitro* crystallisation, there is no formal structural data on tRNA^{Pyl} with
165 complete PylS or PylSc+PylSn: initial work performed 10 years ago gave however interesting
166 features with several structural data encompassing the catalytic domain structure (C-ter
167 region 185–454) of the archaeal PylS from *Methanosarcina mazei* with Pyl (31) and the
168 bacterial PylSc structure (from *Desulfitobacterium hafniense* (29,32)). PylSn from bacteria
169 has been shown to interact and recognize tRNA^{Pyl} in a specific way, enhancing the affinity /
170 specificity with tRNA^{Pyl} (33). Recently, the crystal structure of the Nter domain of PylS (from
171 the archaeon *Methanosarcina mazei*) with tRNA^{Pyl} gave precious new insights, and notably

172 the determinants of its high specificity against canonical AAs not based on anticodon
173 recognition (34). These interactions are susceptible to nucleotides mutations of tRNA^{Pyl} that
174 are not involved in PylSc interactions with tRNA^{Pyl} (for example in the T-arm and variable
175 loop). PylS Cter part / PylSc are responsible at least for the catalytic amino-acylation and
176 partly for the recognition of tRNA^{Pyl}. Nonetheless, anticodon UCA is not directly involved in
177 its recognition by PylS (31), while only few specific bases on anticodon stem and loop
178 nucleotides of tRNA^{Pyl} seem necessary, like for example the bases immediately adjacent to
179 the anticodon in the archaeon *Methanosarcina barkeri* (35). However, recent genomic
180 identification of the Pyl system in archaeal lineages other than *Methanosarcinales*, especially
181 in the *Methanomassiliicoccales* gives indications that these properties are not shared by all
182 Pyl systems (see section 4.2).

183

184 **4. Recent Pyl discoveries**

185 As mentioned, work performed more than a decade ago showed that at least some Pyl gene
186 cassettes from archaea or bacteria are fully functional when expressed in *E. coli* (i.e. UAG
187 specification of Pyl) (24). Moreover, desmethylpyrrolysine was incorporated when D-
188 ornithine was provided to *E. coli* expressing a Pyl system, indicating a low susceptibility of
189 the side chain for aminoacylation of tRNA^{Pyl} by PylS. These properties were therefore used to
190 divert Pyl system from its natural function, allowing non-canonical amino acids (ncAAs) to be
191 included in proteins, and so in organisms, by expressing (wild type or mutated) PylS / tRNA^{Pyl}
192 in the presence of (free) ncAAs in the growth medium. Achievements and strategies for
193 using Pyl system components as a genetic code expansion tool have been recently reviewed
194 (e.g. (36-39)). To date, more than 100 ncAAs have been successfully engineered and
195 incorporated, in a broad phylogenetic spectrum of organisms ranging from bacteria and

196 unicellular eukaryotes (e.g. *Saccharomyces cerevisiae* (40)) to metazoans, like *Caenorhabditis*
197 *elegans* (41), *Drosophila melanogaster* (42), the Zebrafish (43) or the mammal *Mus musculus*
198 (44).

199

200 4.1. Pyl encoding, a widespread feature in the archaeal world?

201 For a long time, all known methanogens were part of six orders gathered into two large
202 classes (Class I: *Methanobacteriales*, *Methanococcales*, *Methanopyrales*; Class II:
203 *Methanosarcinales*, *Methanomicrobiales*, *Methanocellales*) (45). Recent data based on
204 culture and metagenomic approaches have largely modified our vision of the diversity of the
205 methanogens. Indeed, methanogenesis was predicted, and sometimes experimentally
206 confirmed, in several recently discovered lineages that are only distantly related to the Class
207 I and Class II methanogens: the *Methanomassiliicoccales* (46-47), the *Methanofastidiosa*
208 (48), possibly the *Bathyarchaeota* (49), the *Verstraetearchaeota* (50) and the
209 *Methanonatronarchaeia* (51) (Figure 3). A striking characteristic of the members of all these
210 lineages is their dependence on methyl-compounds for methanogenesis, a metabolism
211 present in a minority of previously known methanogens. Along with the methyl-dependent
212 hydrogenotrophic methanogenesis, the Pyl system was also found to be more widespread,
213 being present in some of these new methanogen lineages and for the first time in non-
214 methanogenic archaea (Figure 3). Among these lineages, the Pyl system was first described
215 in the *Methanomassiliicoccales* where it exhibits several unusual characteristics compared to
216 previously characterized Pyl systems (see section 4.2). In addition, one of the five currently
217 known genomes of the *Verstraetearchaeota* harbours the genes for the synthesis and
218 encoding of Pyl (Table 4 in supplementary data of (50)). In this genome, PylS is mentioned as
219 being encoded by two genes, *pylSn* and *pylSc*. This is different from *Methanosarcinaceae*

220 members, that have a single gene coding for both PylSn and PylSc domains, but similar to
221 what was observed in bacteria so far. Recently, the Pyl system was also reported from the
222 *Methanonatronarchaeia*, a class composed of extreme halophiles (51). In all these novel
223 lineages, the occurrence of the Pyl system was coupled with the presence of one or several
224 Pyl-containing methyltransferases (i.e. MtmB, MtbB, MttB) needed for methylamine
225 utilization. This strengthens previous observations that the Pyl system is dedicated to the
226 incorporation of Pyl in these methyltransferases and thus associated to methylamine
227 utilisation. Other methanogens missing the Pyl-containing methyltransferases (e.g.
228 *Methanosphaera* spp. (52), *Methanofastidiosa* spp. (48)), also lack the Pyl system, which
229 argues that this system is more linked to methylamine methyltransferases than to archaea
230 relying on a methanogenesis based on methyl-compounds. In addition, several members of
231 the candidate division MSBL-1 (or *Persephonarchaea*) harbour a complete set of genes for
232 Pyl synthesis and encoding, as well as *mtmB*, *mtbB* and *mttB* genes, as determined from
233 single-cell genomics (53). Interestingly, genomic analyses of these uncultured archaea
234 revealed that they are likely not methanogens, but rather sugar-fermenters. Moreover, the
235 components of the Pyl system in MSBL-1 are phylogenetically related to the bacterial ones,
236 and more specifically with the Pyl system found in the bacteria *Acetohalobium arabaticum*
237 (53). This possibly indicates a horizontal gene transfer (HGT) between these organisms living
238 in the same kinds of environment (hypersaline) and gives new missing pieces to the puzzle of
239 the unknown history, evolution and dissemination of Pyl. The main conclusions to be drawn
240 from these results are that i) archaea with a Pyl-encoding capacity do not only belong to the
241 *Methanosarcinales* but have a much wider distribution in the archaeal phylogeny, probably
242 not limited to the recently discovered lineages, and that ii) Pyl is not necessarily associated
243 to methanogenesis in archaea .

244

245 4.2. The case of Pyl in *Methanomassiliicoccales*

246 The Pyl system of *Methanomassiliicoccales* was the first archaeal one found outside of the
247 *Methanosarcinaceae* (54) and it revealed several originalities compared to
248 *Methanosarcinaceae* or to bacteria (23). It is not shared by all members of this methanogen
249 order, even among closely related species living in an identical environment, the human gut
250 (55). This patchy distribution of the Pyl system in the *Methanomassiliicoccales* is most likely
251 due to its complete loss in some representatives. In this order, all *pyl* genes are not always
252 grouped together in an operon-like structure. This is, for example, the case in “*Candidatus*
253 *Methanomethylophilus alvus*”, in which *pylB* is distant (more than 600 kbp) from *pylTSCD*
254 (Figure 1). It is notable that important genes for methanogenesis are present within a few
255 kilobases from either *pylB* or *pylTSCD*, those of the methyl-coenzyme M reductase catalysing
256 methane formation, and those of the methyltransferases dedicated to methylamines,
257 respectively. (23,56). Also, *Methanomassiliicoccus luminyensis* is the only known organism to
258 possess more than one copy of all *pyl* genes (23). The most amazing feature of the
259 *Methanomassiliicoccales* Pyl system concerns the nature of both tRNA^{Pyl} and PylS. Indeed,
260 the (UAG-recognizing) tRNA^{Pyl} shows a low number of universally conserved bases among
261 tRNA^{Pyl} (21,23). It has the usual cloverleaf secondary structure of tRNAs, although in a
262 condensed manner (Figure 4), and a longer anticodon stem (23). In *Methanomassiliicoccales*,
263 tRNA^{Pyl} shows the surprising feature of having one or two mismatches in the anticodon
264 “stem” (23) (Figure 4). This is a special feature among all natural tRNAs.
265 *Methanomassiliicoccales* PylS corresponds to the Cter part of *Methanosarcinaceae* PylS or to
266 the PylSc of Bacteria. No homolog of the Nter part of PylS / PylSn has been identified in all
267 complete genomes of *Methanomassiliicoccales* analysed so far. The structure of PylS in

268 *Methanomassiliicoccales* is thus likely different than those of *Methanosarcinaceae* members
269 and bacteria. This case seems also present in other archaea (e.g. *Persephonarchaea* (MSBL-
270 1)). Phylogenetic analyses revealed that the *Methanomassiliicoccales* Pyl system has an
271 intermediary position between the *Methanosarcinales* and the bacteria/MSBL-1 (57). In view
272 of these very atypical features, the functionality of the Pyl system in
273 *Methanomassiliicoccales* was therefore questionable. However, the utilization of
274 methylamines as a methanogenesis substrate by several members of the
275 *Methanomassiliicoccales* strongly suggests that they possess a functional Pyl system. TMA,
276 DMA and MMA utilization has been reported for a pure culture of *Methanomassiliicoccus*
277 *luminyensis* (58), as well as MMA utilisation by an enrichment culture of “*Ca.*
278 *Methanoplasma termitum*” (59) and TMA utilisation by “*Ca.* *Methanomethylophilus alvus*”
279 (55). The remaining question is therefore, how these organisms can accommodate such a
280 different tRNA^{Pyl} and the absence of Nter part of PylS / PylSn. One possibility is that in this
281 lineage, PylSc-like / shortened PylS efficiently binds and links Pyl to its dedicated tRNA^{Pyl} in
282 absence of the PylSn domain. Another one is that it requires an unknown factor that would
283 functionally replace PylSn and/or enhance the effectiveness of this restricted system. In any
284 case, this indicates that other mechanisms than those described so far are possible.

285 Finally, it seems that the presence of the Pyl system in *Methanomassiliicoccales* has led to
286 different genomic adaptations among clusters of *Methanomassiliicoccales* (56), some of
287 which are mentioned in section 4.4.

288

289 4.3. About the origin and evolution of Pyl

290 Recent data have revealed that both methanogenesis based on methylamines
291 (methylotrophic one and methyl-dependant hydrogenotrophic one) (60) and Pyl-encoding

292 behaviour are more widespread than stated just a few years ago (Figure 3). The hypothesis
293 that methanogenesis is a very ancient metabolism that arose before the last common
294 ancestor of *Archaea* has been strengthened by these data (61) and in a same way, a similar
295 hypothesis can be made for the origin of the 22nd amino acid.

296 It is generally accepted that LUCA (Last Unique Common Ancestor) did not have Pyl and that
297 the genetic code based on 20 proteinogenic amino acid catalogue was already established.

298 Among the hypotheses of the presence of Pyl in various and extremely distant lineages,
299 Gogarten and colleagues postulated that Pyl could have been encoded by cells living at the
300 time of LUCA, whose descendants are however either extinct or not yet discovered (62). In
301 this hypothesis, Pyl and associated coding behaviour persisted until present time because,
302 before their extinction, some of the descendants of this hypothetical lineage would have
303 spread the Pyl system by HGT to some LUCA descendants, an archaeon (leading to Pyl-
304 encoding in the lineage *Methanosarcinales*), and a bacterium. Molecular and structurally-
305 based phylogenetic studies seem to indicate that PylS arose from a duplication of PheRS-
306 coding gene anterior to LUCA (31). The extra copy of this gene could have been the matrix
307 for Pyl invention, once in one lineage, or possibly independently in several lineages, later
308 during the evolution (31). The evidence of Pyl in *Methanomassiliicoccales* and the nature of
309 its encoding may indicate a unique origin of the Pyl-system (in either archaea or bacteria)
310 and its later spread to the other domain of life by HGT(s). This would make its origin
311 subsequent to LUCA (23). The recent discovery and the nature of the Pyl system in
312 *Persephonarchaea* (MSBL-1) also supports this view. Irrespective of such considerations,
313 current data strongly suggests that Pyl-encoding is a very ancient trait, which arose likely just
314 once as a post LUCA invention possibly linked to methanogenesis. It could have then evolved
315 and been retained in some organisms for which methylamine metabolism was key to

316 survival and could have been further transferred across the bacterial and archaeal domains.
317 While our understanding is likely to deepen further with the probable discovery of new Pyl-
318 containing organisms, interesting questions can now be asked about how organisms have
319 adapted to this genetic code expansion. One type of informative analysis has come from
320 studying the consequences of deleting a functional Pyl system (63).

321

322 4.4. Implications and consequences of the natural presence of a UAG decoding tRNA

323 While genetic studies with mutants of tRNA (*amber* suppressors) that permit decoding of
324 UAG in competition with release factors have provided decoding insights, the organisms
325 studied have not adapted over long evolutionary time to the presence of the UAG-decoding
326 tRNA. In contrast, such adaptation has occurred in Pyl-containing organisms. Though the
327 situation in such organisms is not as extreme as the ciliate *Condolysoma magnum* where
328 UAG, UAA and UAG have been reassigned to be sense codons, with termination being
329 specified by a recoding event (64,65), since whole organisms are involved, the issues are
330 nevertheless more substantial than in the case of decoding small genomes such as
331 mitochondria.

332 The mechanism used for UGA specification of selenocysteine was first considered as a
333 potent paradigm from which Pyl insertion could be extrapolated. UGA specification of
334 selenocysteine is dependent on special features of the mRNA 3' of the UGA codon. UGA
335 codons that do not have the recoding signals involved function in the standard manner of
336 specifying termination (66,67). A counterpart Pyl insertion sequence (PYLIS) element was
337 first postulated (25). However, consistent with a detailed bioinformatics analysis (14), such a
338 *cis*-acting element was found to be not necessary for insertion of Pyl in engineered *E. coli*
339 (68, 25). The resulting hypothesis is that Pyl insertion relies “simply” on a competition

340 between tRNA^{Pyl} and aRF1 on UAG codon during mRNA translation at ribosome. This is likely
341 promoted by a peculiar aRF1, as some distinct patterns are observed in Pyl coding
342 *Methanosarcinaceae* compared to other archaeal aRF1 (14). In such a case, one can imagine
343 that UAG codon that does not need to be reassigned to a Pyl would be counter-selected in
344 favour of the other two non-sense codons. In agreement with this hypothesis, it is observed
345 that the UAG codon usage in Pyl-coding archaea is low (14,69) with *Methanosarcinaceae*
346 having only about 4 to 6 % CDSs with an in-frame UAG codon (12). Recent data indicate that
347 UAG usage is even lower in some *Methanomassiliicoccales* (in-frame UAG predicted in 2.8%
348 of CDSs in “*Ca. M. alvus*”). However, the UAG usage varies among *Methanomassiliicoccales*,
349 independently of GC% of considered genomes (56). More surprisingly, members of
350 *Methanomassiliicoccales* are predicted, through bioinformatics analysis, to use Pyl in a larger
351 repertoire of proteins than the usual methyltransferases described above (56). This
352 encompasses 16 non-MT proteins in “*Ca. M. alvus*” (56) and more than 40 non-MT proteins
353 in another representative (methanogenic archaeon ISO4-H5 (70)). This questions a possible
354 “genetic code domestication” of Pyl in *Methanomassiliicoccales* corroborating the
355 hypothesis of an on-going irreversible conversion of UAG sense from a stop sense to a Pyl-
356 encoding one proposed in *Methanosarcinales* (69). Indeed, the peculiar properties observed
357 in some *Methanomassiliicoccales* may reflect the on-going establishment of a genetic code
358 having 21 amino acids encoded by 62 codons and only 2 codons instead of 3 being non-
359 sense. This 21 proteinogenic AA repertoire would be currently invading the proteome of
360 some of these organisms by neutral selection on most proteins, while it cannot be excluded
361 that some cases are providing / will provide benefits for the cell.

362 The frequency of UAG codons is much higher in Pyl-coding bacteria than in archaea,
363 with about 20-28% noted in one study (14,12). In the absence of evidence for mRNA context

364 dependency for the natural Pyl specification that occurs in some bacteria, how, if it is, can
365 specification of Pyl by the numerous UAG codons be avoided? For *Acetohalobium*
366 *arabaticum*, a member of the phylum *Clostridia*, a fascinating resolution to the dilemma has
367 been discovered. While UAG specifies Pyl when the cells are grown in trimethylamine, when
368 instead the cells are grown on pyruvate as a carbon source, UAG only specifies termination
369 (12). This presumably means that when the cells are grown on trimethylamine, many
370 proteins for which pyrrolysine has no specific function have C-terminal extensions with
371 several amino acids encoded after pyrrolysine. Indeed, some such proteins have been
372 identified (12). Doubtless, selection has had ample time to adjust, where significant, further
373 downstream stop codon position. However, for most genes such selective pressure may be
374 weak since where analysed there is an abundance of proteases that act on incorrectly folded
375 peptides, and the efficiency of newly isolated genetically selected *E. coli* amber suppressors
376 with little or no discernible growth rate. Nevertheless, a decoding-wide strategy has evolved
377 for when trimethylamine is absent.

378

379 **5. Conclusion**

380 Recent data about Pyl in archaea have renewed our comprehension of this original and still
381 enigmatic exception in the “universal” genetic code. It was recently uncovered that this
382 system is not confined to the *Methanosarcinales* in the domain Archaea, but present in at
383 least 4 other orders, and not only methanogens. In members of those novel lineages, the
384 divergences from canonical Pyl systems offer new opportunities to understand its
385 functioning. These advances will likely facilitate improved utilization of the Pyl system as a
386 biotechnological tool for engineering of proteins / organisms by incorporating ncAAs in their
387 proteins. Also, Pyl is mandatory for methylamines utilisation under anoxic conditions which

388 may explain its presence and maintenance in diverse and phylogenetically distant organisms.
389 One of these methylamines, TMA, was recently shown to be a by-product of human
390 intestinal bacteria, and the unique precursor of plasma TMAO, an important
391 atherosclerogenic factor (71). Thus, autochthonous archaea able to metabolise TMA in the
392 gut (some *Methanomassiliicoccales* members), or engineered microbes with a Pyl-encoding
393 behaviour to “trap” TMA in the gut may be valuable auxiliaries for preventing cardiovascular
394 diseases in humans (58). Besides these biotechnological aspects, its study will likely facilitate
395 understanding of the genetic code evolution among lineages and the intimate mechanisms
396 of such important biological functions as translation. All of these bode well for increasing
397 interest in Pyl and its host organisms in the coming years.

398

399 **Acknowledgements**

400 J.-F. B. thanks the French national commission “64^e CNU” for agreeing with a sabbatical in
401 P.W.O.T.’s lab (School of Microbiology and APC Microbiome Institute, University College
402 Cork).

403

404 **Declaration of interest**

405 The authors declare no conflict of interest.

406

407 **Author contribution statement**

408 All authors contributed to manuscript writing.

409

410 **REFERENCE:**

- 411 1. Crick, F. H. C., *et al.* (1961). General nature of the genetic code for proteins. *Nature*. 192,
412 1227–1232.
- 413 2. “The Genetic Code” (1966). Collective work of proceedings of symposium held in Cold
414 Spring Harbor (USA) in 1966, *Cold Spring Harb Symp Quant Biol*. 31.
- 415 3. Atkins, J. F., & Baranov, P. V. (2010). The distinction between recoding and codon
416 reassignment. *Genetics*. 185(4), 1535-1536.
- 417 4. Barrell, B. G., Bankier, A. T., & Drouin, J. (1979). A different genetic code in human
418 mitochondria. *Nature*. 282(5735), 189.
- 419 5. Osawa, S., Jukes, T. H., Watanabe, K., & Muto, A. (1992). Recent evidence for evolution of
420 the genetic code. *Microbiol Rev*. 56(1), 229-264.
- 421 6. Rother, M. & Quitzke, V. (2018). Selenoprotein synthesis and regulation in Archaea.
422 *Biochim Biophys Acta*. S0304-4165(18), 30101-30106.
- 423 7. Mukai, T., Englert, M., Tripp, H.J., Miller, C., Ivanova, N. N., Rubin, E. M., Kyrpides, N. C.
424 & Söll, D. (2016). Facile recoding of selenocysteine in nature. *Angew Chem Int Ed Engl*. 55,
425 5337-5341.
- 426 8. Zinoni, F., Heider, J. & Böck A. (1990). Features of the formate dehydrogenase mRNA
427 necessary for decoding of the UGA codon as selenocysteine. *Proc Natl Acad Sci U S A*. 87,
428 4660–4664.
- 429 9. Chambers, I., Frampton, J., Goldfarb, P., Affara, N., McBain, W. & Harrison P. R. (1986).
430 The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active
431 site is encoded by the 'termination' codon, TGA. *EMBO J*. 5, 1221-1227.
- 432 10. Srinivasan, G., James, C. M., & Krzycki, J. A. (2002). Pyrrolysine encoded by UAG in
433 Archaea: charging of a UAG-decoding specialized tRNA. *Science*, 296(5572), 1459-1462.

- 434 11. Hao, B., Gong, W., Ferguson, T. K., James, C. M., Krzycki, J. A., & Chan, M. K. (2002). A
435 new UAG-encoded residue in the structure of a methanogen methyltransferase. *Science*,
436 296(5572), 1462-1466.
- 437 12. Prat, L., Heinemann, I. U., Aerni, H. R., Rinehart, J., O'Donoghue, P. & Söll, D. (2012).
438 Carbon source-dependent expansion of the genetic code in bacteria. *Proc Natl Acad Sci U S*
439 *A*. 109(51):21070-21075.
- 440 13. Eichler, J., & Adams, M. W. (2005). Posttranslational protein modification in Archaea.
441 *Microbiol Mol Biol Rev* , 69(3), 393-425.
- 442 14. Zhang, Y., Baranov, P. V., Atkins, J. F., & Gladyshev, V. N. (2005). Pyrrolysine and
443 selenocysteine use dissimilar decoding strategies. *J Biol Chem*. 280(21), 20740-20751.
- 444 15. Heinemann, I. U., O'Donoghue, P., Madinger, C., Benner, J., Randau, L., Noren, C. J., &
445 Söll, D. (2009). The appearance of pyrrolysine in tRNA^{His} guanylyltransferase by neutral
446 evolution. *Proc Natl Acad Sci U S A*. 106(50), 21103-21108.
- 447 16. Ticak, T., Kountz, D. J., Girosky, K. E., Krzycki, J. A., & Ferguson, D. J. (2014). A
448 nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is
449 a glycine betaine methyltransferase. *Proc Natl Acad Sci U S A*. 111(43), E4668-E4676.
- 450 17. Krzycki, J. A. (2004). Function of genetically encoded pyrrolysine in corrinoid-dependent
451 methylamine methyltransferases. *Curr Opin Chem Biol*. 8(5), 484-491.
- 452 18. Burke, S. A., Lo, S. L., & Krzycki, J. A. (1998). Clustered genes encoding the
453 methyltransferases of methanogenesis from monomethylamine. *J Bacteriol*. 180(13), 3432-
454 3440.
- 455 19. Paul, L., Ferguson, D. J., & Krzycki, J. A. (2000). The trimethylamine methyltransferase
456 gene and multiple dimethylamine methyltransferase genes of *Methanosarcina barkeri*
457 contain in-frame and readthrough amber codons. *J Bacteriol*. 182(9), 2520-2529.

- 458 20. Atkins, J. F., & Gesteland, R. (2002). The 22nd amino acid. *Science*. 296(5572), 1409-
459 1411.
- 460 21. Gaston, M. A., Jiang, R., & Krzycki, J. A. (2011). Functional context, biosynthesis, and
461 genetic encoding of pyrrolysine. *Curr Opin Microbiol*. 14(3), 342-349.
- 462 22. Krzycki, J. A. (2013). The path of lysine to pyrrolysine. *Curr Opin Chem Biol*. 17(4), 619-
463 625.
- 464 23. Borrel, G., Gaci, N., Peyret, P., O'Toole, P. W., Gribaldo, S., & Brugère, J. F. (2014). Unique
465 characteristics of the pyrrolysine system in the 7th order of methanogens: implications for
466 the evolution of a genetic code expansion cassette. *Archaea*, 2014-374146.
- 467 24. Longstaff, D. G., Larue, R. C., Faust, J. E., Mahapatra, A., Zhang, L., Green-Church, K. B., &
468 Krzycki, J. A. (2007). A natural genetic code expansion cassette enables transmissible
469 biosynthesis and genetic encoding of pyrrolysine. *Proc Natl Acad Sci U S A*. 104(3), 1021-
470 1026.
- 471 25. Namy, O., Zhou, Y., Gundllapalli, S., Polycarpo, C. R., Denise, A., Rousset, J. P., ... &
472 Ambrogelly, A. (2007). Adding pyrrolysine to the *Escherichia coli* genetic code. *FEBS lett*.
473 581(27), 5282-5288.
- 474 26. Cellitti, S. E., Ou, W., Chiu, H. P., Grünewald, J., Jones, D. H., Hao, X., ... & Lesley, S. A.
475 (2011). D-Ornithine coopts pyrrolysine biosynthesis to make and insert pyrroline-carboxy-
476 lysine. *Nat. Chem. Biol*. 7(8), 528.
- 477 27. Gaston, M. A., Zhang, L., Green-Church, K. B., & Krzycki, J. A. (2011). The complete
478 biosynthesis of the genetically encoded amino acid pyrrolysine from lysine. *Nature*.
479 471(7340), 647.

480 28. Quitterer, F., List, A., Eisenreich, W., Bacher, A., & Groll, M. (2012). Crystal structure of
481 methylornithine synthase (PylB): insights into the pyrrolysine biosynthesis. *Angew. Chem.*
482 *Int. Ed.* 51(6), 1339-1342.

483 29. Nozawa, K., O'Donoghue, P., Gundllapalli, S., Araiso, Y., Ishitani, R., Umehara, T., ... &
484 Nureki, O. (2009). Pyrrolysyl-tRNA synthetase-tRNA Pyl structure reveals the molecular basis
485 of orthogonality. *Nature*, 457(7233), 1163.

486 30. Herring, S., Ambrogelly, A., Gundllapalli, S., O'Donoghue, P., Polycarpo, C. R., & Söll, D.
487 (2007). The amino-terminal domain of pyrrolysyl-tRNA synthetase is dispensable in vitro but
488 required for in vivo activity. *FEBS Lett.* 581(17), 3197-3203.

489 31. Kavran, J. M., Gundllapalli, S., O'Donoghue, P., Englert, M., Söll, D., & Steitz, T. A. (2007).
490 Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation.
491 *Proc Natl Acad Sci U S A.* 104(27), 11268-11273.

492 32. Lee, M. M., Jiang, R., Jain, R., Larue, R. C., Krzycki, J., & Chan, M. K. (2008). Structure of
493 *Desulfitobacterium hafniense* PylSc, a pyrrolysyl-tRNA synthetase. *Biochem Biophys Res*
494 *Commun.* 374(3), 470-474.

495 33. Jiang, R., & Krzycki, J. A. (2012). PylSn and the homologous N-terminal domain of
496 pyrrolysyl-tRNA synthetase bind the tRNA that is essential for the genetic encoding of
497 pyrrolysine. *J Biol Chem.* jbc-M112.

498 34. Suzuki, T., Miller, C., Guo, L. T., Ho, J. M., Bryson, D. I., Wang, Y. S., Liu, D. R., & Söll, D.
499 (2017). Crystal structures reveal an elusive functional domain of pyrrolysyl-tRNA synthetase.
500 *Nat Chem Biol.* 13(12), 1261-1266.

501 35. Ambrogelly, A., Gundllapalli, S., Herring, S., Polycarpo, C., Frauer, C., & Söll, D. (2007).
502 Pyrrolysine is not hardwired for cotranslational insertion at UAG codons. *Proc Natl Acad Sci*
503 *U S A.* 104(9), 3141-3146.

- 504 36. Wan, W., Tharp, J. M., & Liu, W. R. (2014). Pyrrolysyl-tRNA synthetase: an ordinary
505 enzyme but an outstanding genetic code expansion tool. *Biochim Biophys Acta -Proteins and*
506 *Proteomics*. 1844(6), 1059-1070.
- 507 37. Brabham, R., & Fascione, M. A. (2017). Pyrrolysine Amber Stop-Codon Suppression:
508 Development and Applications. *Chembiochem*, 18(20), 1973-1983.
- 509 38. Mukai, T., Lajoie, M. J., Englert, M., & Söll, D. (2017). Rewriting the genetic code. *Annu*
510 *Rev Microbiol*. 71, 557-577.
- 511 39. Tharp, J. M., Ehnbohm, A., & Liu, W. R. (2017). tRNAPyl: Structure, function, and
512 applications. *RNA biol*. 1-12.
- 513 40. Chin, J. W., Cropp, T. A., Anderson, J. C., Mukherji, M., Zhang, Z., & Schultz, P. G. (2003).
514 An expanded eukaryotic genetic code. *Science*. 301(5635), 964-967.
- 515 41. Greiss, S., & Chin, J. W. (2011). Expanding the genetic code of an animal. *J Am Chem Soc*.
516 133(36), 14196-14199.
- 517 42. Bianco, A., Townsley, F. M., Greiss, S., Lang, K., & Chin, J. W. (2012). Expanding the
518 genetic code of *Drosophila melanogaster*. *Nat Chem Biol*. 8(9), 748.
- 519 43. Liu, J., Hemphill, J., Samanta, S., Tsang, M., & Deiters, A. (2017). Genetic Code Expansion
520 in Zebrafish Embryos and Its Application to Optical Control of Cell Signaling. *J Am Chem Soc*.
521 139(27), 9100-9103.
- 522 44. Han, S., Yang, A., Lee, S., Lee, H. W., Park, C. B., & Park, H. S. (2017). Expanding the
523 genetic code of *Mus musculus*. *Nat Commun*. 8, 14568.
- 524 45. Bapteste, É., Brochier, C., & Boucher, Y. (2005). Higher-level classification of the Archaea:
525 evolution of methanogenesis and methanogens. *Archaea*. 1(5), 353-363.

526 46. Borrel, G., O'Toole, P. W., Harris, H. M., Peyret, P., Brugere, J. F., & Gribaldo, S. (2013).
527 Phylogenomic data support a seventh order of methylophilic methanogens and provide
528 insights into the evolution of methanogenesis. *Genome Biol Evol.* 5(10), 1769-1780.

529 47. Petitjean, C., Deschamps, P., López-García, P., Moreira, D., & Brochier-Armanet, C.
530 (2015). Extending the conserved phylogenetic core of archaea disentangles the evolution of
531 the third domain of life. *Mol Biol Evol.* 32(5), 1242-1254.

532 48. Nobu, M. K., Narihiro, T., Kuroda, K., Mei, R., & Liu, W. T. (2016). Chasing the elusive
533 Euryarchaeota class WSA2: genomes reveal a uniquely fastidious methyl-reducing
534 methanogen. *ISME J.* 10, 2478-2487.

535 49. Evans, P. N., Parks, D. H., Chadwick, G. L., Robbins, S. J., Orphan, V. J., Golding, S. D., &
536 Tyson, G. W. (2015). Methane metabolism in the archaeal phylum Bathyarchaeota revealed
537 by genome-centric metagenomics. *Science*, 350(6259), 434-438.

538 50. Vanwonterghem, I., Evans, P. N., Parks, D. H., Jensen, P. D., Woodcroft, B. J., Hugenholtz,
539 P., & Tyson, G. W. (2016). Methylophilic methanogenesis discovered in the archaeal
540 phylum Verstraetearchaeota. *Nat Microbiol.* 1(12), 16170.

541 51. Sorokin, D. Y., Makarova, K. S., Abbas, B., Ferrer, M., Golyshin, P. N., Galinski, E. A., ... &
542 van Loosdrecht, M. C. (2017). Discovery of extremely halophilic, methyl-reducing
543 Euryarchaea provides insights into the evolutionary origin of methanogenesis. *Nat Microbiol.*
544 2(8), 17081.

545 52. Fricke, W. F., Seedorf, H., Henne, A., Krüer, M., Liesegang, H., Hedderich, R., ... & Thauer,
546 R. K. (2006). The genome sequence of *Methanosphaera stadtmanae* reveals why this human
547 intestinal archaeon is restricted to methanol and H₂ for methane formation and ATP
548 synthesis. *J Bacteriol.* 188(2), 642-658.

549 53. Guan, Y., Haroon, M. F., Alam, I., Ferry, J. G., & Stingl, U. (2017). Single-cell genomics
550 reveals pyrrolysine-encoding potential in members of uncultivated archaeal candidate
551 division MSBL-1. *Environ Microbiol Rep.* 9(4), 404-410.

552 54. Borrel, G., Harris, H. M., Tottey, W., Mihajlovski, A., Parisot, N., Peyretailade, E., ... &
553 Brugère, J. F. (2012). Genome sequence of "Candidatus Methanomethylophilus alvus"
554 Mx1201, a methanogenic archaeon from the human gut belonging to a seventh order of
555 methanogens. *J Bacteriol.* 194(24), 6944-6945.

556 55. Borrel, G., McCann, A., Deane, J., Neto, M. C., Lynch, D. B., Brugère, J. F., & O'Toole, P. W.
557 (2017). Genomics and metagenomics of trimethylamine-utilizing Archaea in the human gut
558 microbiome. *ISME J.* 11(9), 2059.

559 56. Borrel, G., Parisot, N., Harris, H. M., Peyretailade, E., Gaci, N., Tottey, W., ... & O'Toole, P.
560 W. (2014). Comparative genomics highlights the unique biology of Methanomassiliicoccales,
561 a Thermoplasmatales-related seventh order of methanogenic archaea that encodes
562 pyrrolysine. *BMC genomics.* 15(1), 679.

563 57. Mayumi, D., Mochimaru, H., Tamaki, H., Yamamoto, K., Yoshioka, H., Suzuki, Y., ... &
564 Sakata, S. (2016). Methane production from coal by a single methanogen. *Science.*
565 354(6309), 222-225.

566 58. Brugère, J. F., Borrel, G., Gaci, N., Tottey, W., O'toole, P. W., & Malpuech-Brugère, C.
567 (2014). Archaeobiotics: proposed therapeutic use of archaea to prevent trimethylaminuria
568 and cardiovascular disease. *Gut Microbes.* 5(1), 5-10.

569 59. Lang, K., Schuldes, J., Klingl, A., Poehlein, A., Daniel, R., & Brune, A. (2015). New mode of
570 energy metabolism in the seventh order of methanogens as revealed by comparative
571 genome analysis of "Candidatus Methanoplasma termitum". *Appl Environ Microbiol.* 81(4),
572 1338-1352.

- 573 60. Adam, P. S., Borrel, G., Brochier-Armanet, C., & Gribaldo, S. (2017). The growing tree of
574 Archaea: new perspectives on their diversity, evolution and ecology. *ISME J.* 11(11), 2407.
- 575 61. Borrel, G., Adam, P. S., & Gribaldo, S. (2016). Methanogenesis and the Wood–Ljungdahl
576 pathway: an ancient, versatile, and fragile association. *Genome Biol Evol.* 8(6), 1706-1711.
- 577 62. Fournier, G. P., Huang, J., & Gogarten, J. P. (2009). Horizontal gene transfer from extinct
578 and extant lineages: biological innovation and the coral of life. *Philos Trans R Soc Lond B Biol*
579 *Sci.* 364(1527), 2229-2239.
- 580 63. O’Donoghue, P., Prat, L., Kucklick, M., Schäfer, J. G., Riedel, K., Rinehart, J., ... &
581 Heinemann, I. U. (2014). Reducing the genetic code induces massive rearrangement of the
582 proteome. *Proc Natl Acad Sci U S A.* 111(48), 17206-17211.
- 583 64. Swart, E. C., Serra, V., Petroni, G., & Nowacki, M. (2016). Genetic codes with no
584 dedicated stop codon: context-dependent translation termination. *Cell*, 166(3), 691-702.
- 585 65. Heaphy, S. M., Mariotti, M., Gladyshev, V. N., Atkins, J. F., & Baranov, P. V. (2016). Novel
586 ciliate genetic code variants including the reassignment of all three stop codons to sense
587 codons in *Condylostoma magnum*. *Mol Biol Evol.* 33(11), 2885-2889.
- 588 66. Heider, J., Baron, C., & Böck, A. (1992). Coding from a distance: dissection of the mRNA
589 determinants required for the incorporation of selenocysteine into protein. *EMBO J.* 11(10),
590 3759-3766.
- 591 67. Berry, M. J., Banu, L., Harney J. W., & Larsen P. R. (1993). Functional characterization of
592 the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. *EMBO J.*
593 12, 3315–3322.
- 594 68. Longstaff, D. G., Blight, S. K., Zhang, L., Green-Church, K. B., & Krzycki, J. A. (2007). In vivo
595 contextual requirements for UAG translation as pyrrolysine. *Mol Microbiol.* 63(1), 229-241.

- 596 69. Alkalaeva, E., Eliseev, B., Ambrogelly, A., Vlasov, P., Kondrashov, F. A., Gundllapalli, S., ...
597 & Kisselev, L. (2009). Translation termination in pyrrolysine-utilizing archaea. FEBS lett.
598 583(21), 3455-3460.
- 599 70. Li, Y., Leahy, S. C., Jeyanathan, J., Henderson, G., Cox, F., Altermann, E., ... & Attwood, G.
600 T. (2016). The complete genome sequence of the methanogenic archaeon ISO4-H5 provides
601 insights into the methylotrophic lifestyle of a ruminal representative of the
602 Methanomassiliicoccales. Stand Genomic Sci. 11(1), 59.
- 603 71. Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., DuGar, B., ... & Wu, Y. (2011).
604 Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature.
605 472(7341), 57.
- 606 72. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., ... & Ong, Q.
607 (2015). The MetaCyc database of metabolic pathways and enzymes and the BioCyc
608 collection of pathway/genome databases Nucleic Acids Res. 44(D1), D471-D480.

609

610

611 **ABBREVIATIONS:**

612 AA: amino acid

613 aaRS: amino-acyl tRNA synthetase

614 aa-tRNA: amino-acyl charged tRNA

615 DMA: Dimethylamine

616 HGT: Horizontal gene transfer

617 IUBMB: International union for biochemistry and molecular biology

618 IUPAC: International Union of Pure and Applied Chemistry

619 LUCA: Last Unique Common Ancestor

- 620 MMA: monomethylamine
- 621 ncAA: non-canonical AA
- 622 Pyl: Pyrrolysine
- 623 PYLIS: Pyl insertion sequence
- 624 Sec: Selenocysteine
- 625 SECIS: Sec insertion sequence
- 626 TIM: Triose phosphate isomerase
- 627 TMA: trimethylamine

628 **FIGURE LEGENDS**

629 **FIGURE 1:** Overview of the components and functioning of the Pyl-system.

630 Some examples of genomic organisation of genes involved in Pyl synthesis and encoding are
631 provided in the upper part of the figure (adapted from (21), (23) and (53)). Examples from
632 Archaea and Bacteria are given respectively in a red delimited box at left and a green
633 delimited box at right. The “canonical” organisation in archaea is given in the upper light
634 grey box. The functioning of the Pyl-system is depicted here taken the example of Pyl
635 insertion by readthrough of UAG codon contained in *mttB* mRNA of the
636 *Methanomassiliicoccales* “*Candidatus* Methanomethylophilus alvus”. (A) It relies first on the
637 biosynthesis of Pyl (depicted with more details in Figure 2). Pyl is then engrafted on a specific
638 tRNA (encoded by *pylT*) whose anticodon recognizes UAG codons. This implies the specific
639 reaction catalysed by PylRS (in this example encoded by a truncated version of *pylS*
640 corresponding to *PylSc* (see section 4.2 for more details). (B) All the genes encoding the
641 (mono- di- or tri-) methylamine-corrinoid methyltransferases (respectively *mtmB*, *mtbB* and
642 *mttB*) show an in frame UAG codon, which should typically lead to a translational stop, i.e. a
643 truncated protein. (C) Due to tRNA^{Pyl} able to interact with UAG codon, translation is going on
644 by readthrough this codon which leads to insertion of Pyl in the growing polypeptide.
645 Functioning of such a system at the genomic and cellular level is discussed in section 4.4.

646

647 **FIGURE 2:** Biosynthesis of Pyl.

648 The complete biosynthesis pathway of L-pyrrolysine from 2 lysines catalysed by PylB, PylC
649 and PylD is depicted. This corresponds respectively to reactions #13074, #13080 and #13084
650 adapted from MetaCyc (72), as determined from original data in (26,27).

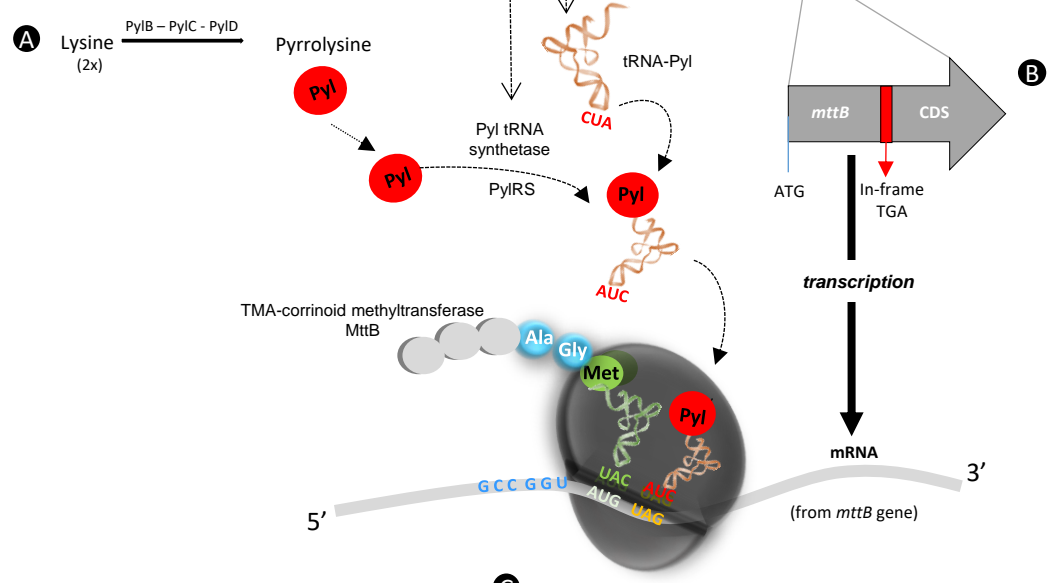
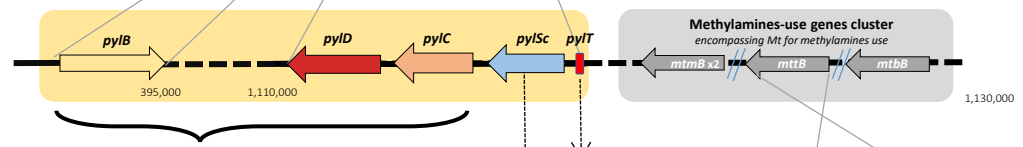
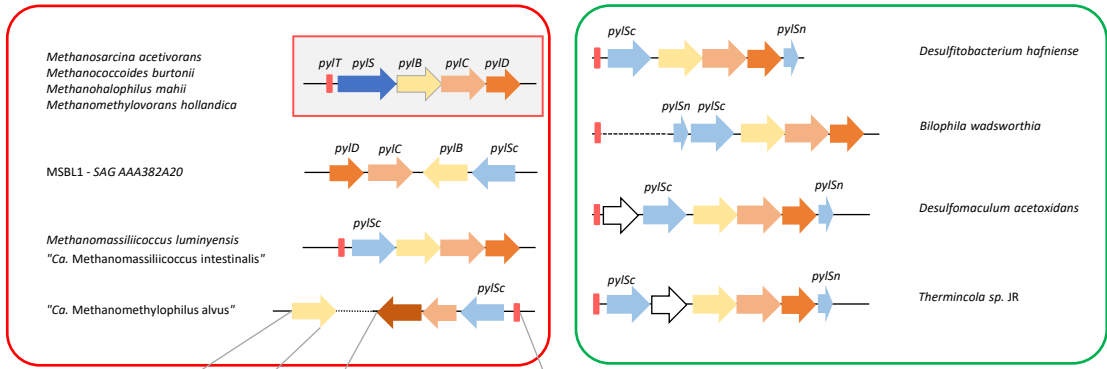
651

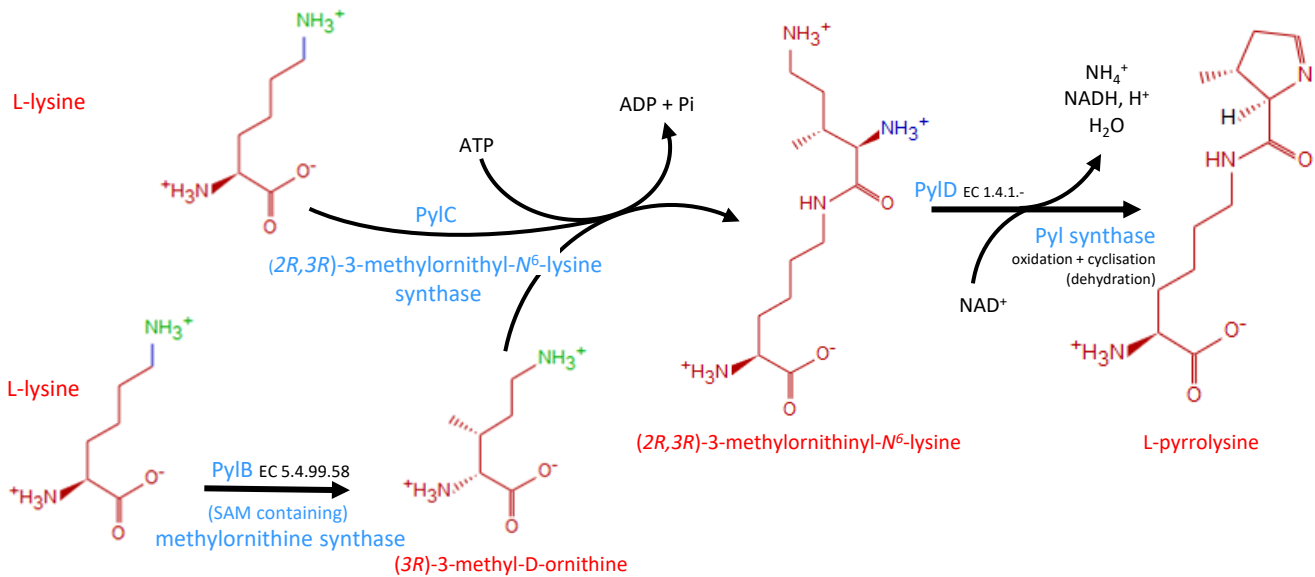
652 **FIGURE 3:** Distribution of the Pyl-encoding microorganisms in the domain Archaea, as
653 indicated by red dots. Lineages in bold contain predicted (based on genome analysis) or
654 experimentally characterized methanogens. “H₂ + R-CH₃” indicates that some members (e.g.
655 *Methanosphaera* in *Methanobacteriales*, *Methanosarcinaceae* in *Methanosarcinales*) or all
656 members of the lineage can gain energy through methyl-dependent hydrogenotrophic
657 methanogenesis. Depending on the methanogen species, one or several methylated
658 compounds (R-CH₃, e.g. methanol, TMA, DMA, TMA, methanethiol, glycine betaine) can be
659 used. Some of the names and the topology of the tree are derived from (56).

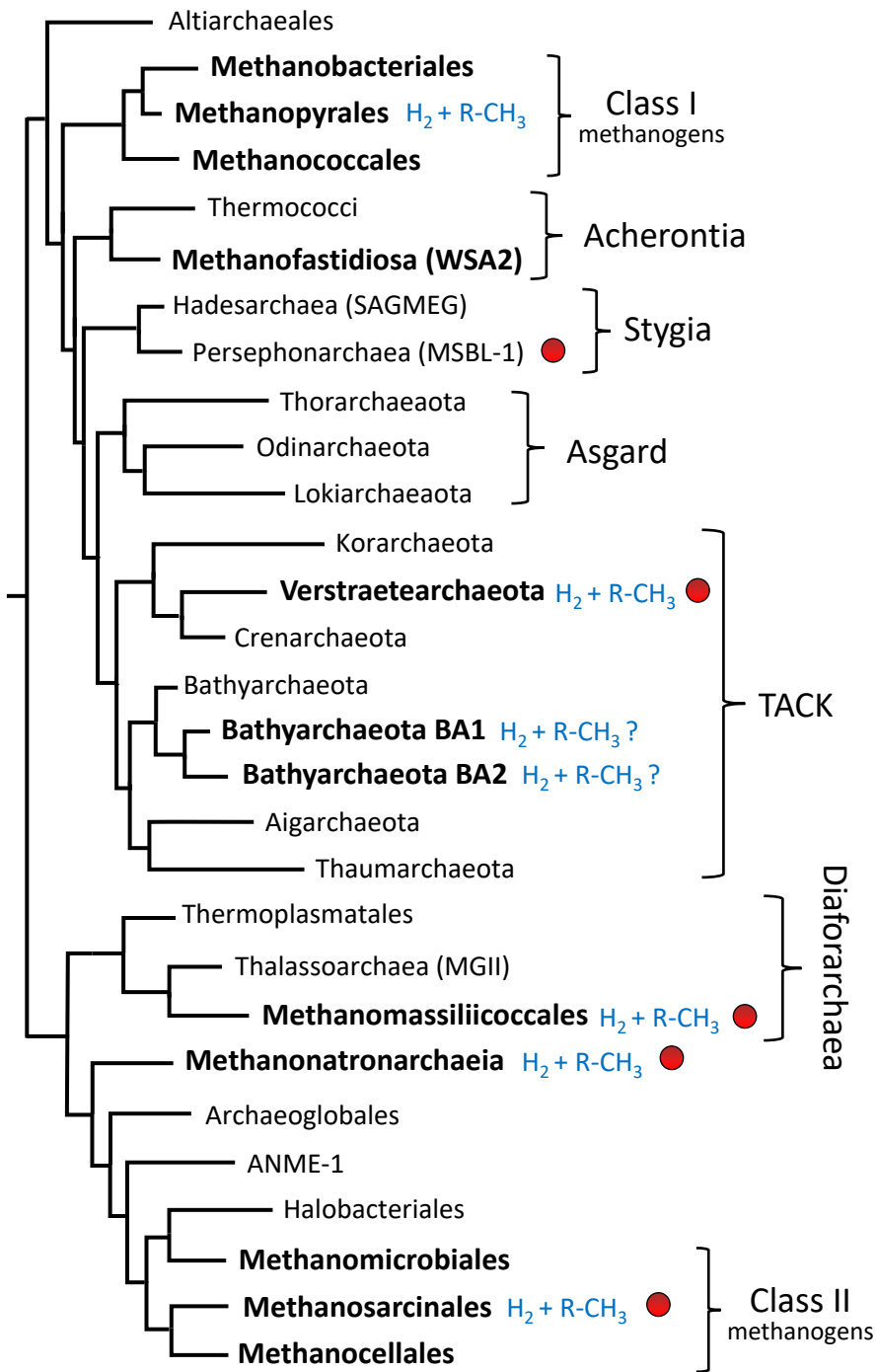
660

661 **FIGURE 4:** Sequence diversity among tRNA^{Pyl}.

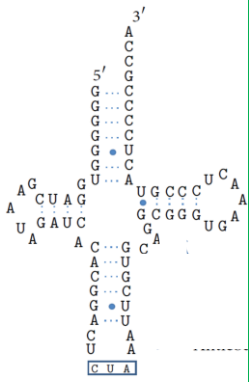
662 The sequence and structure of tRNA^{Pyl} are depicted here (with a cloverleaf representation)
663 using selected examples of known sequences, encompassing a bacterial one (green box) and
664 four archaeal ones (red box). Compared to other tRNAs, tRNA^{Pyl} is smaller and has marked
665 differences among Pyl lineages, especially between *Methanosarcinaceae* and other recently
666 discovered Pyl-encoding archaeal lineages. The MSBL-1 tRNA^{Pyl} shows a larger variable loop
667 (green arrow) and D-loop (blue arrow), than *Methanomassiliicoccales* tRNA^{Pyl} shown here
668 that has the unique characteristics of mismatches / a loop in their anticodon « arm » (red
669 arrow). This is especially evident in *M. luminyensis* and “*Ca. M. intestinalis*” (23). This figure
670 is an adaptation of the data available in (23,53). Only one of the 2 tRNA^{Pyl} recovered from *M.*
671 *luminyensis* genome is indicated (tRNA^{Pyl} #1 according to (23)).





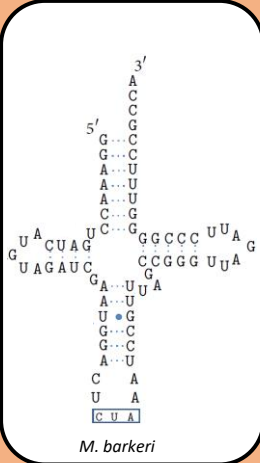


Bacteria

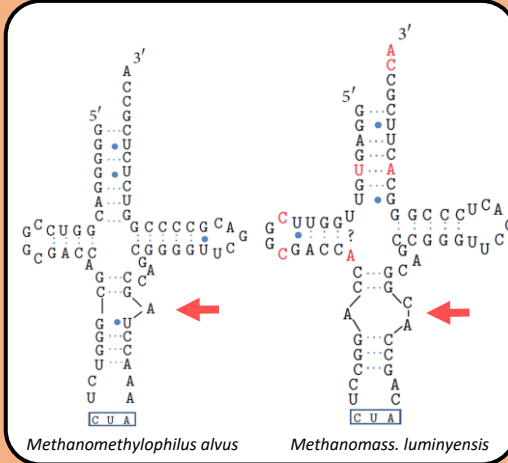


Desulfotobacterium hafniense

Archaea



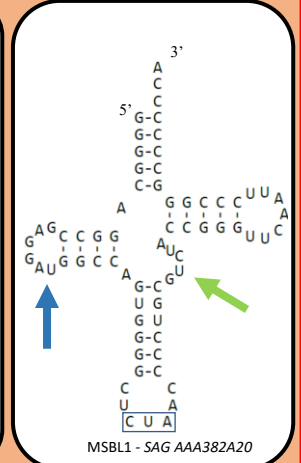
Methanosarcinaceae



Methanomethylophilus alvus

Methanomass. luminyensis

Methanomassiliicoccales



MSBL1 - SAG AAA382A20

Persephonarchaea