# Single-cell genomics reveals pyrrolysine-encoding potential in members of uncultivated archaeal candidate division MSBL1

**Yue Guan[1], Mohamed F. Haroon[1,2], Intikhab Alam[3], James G. Ferry[4], Ulrich Stingl[1,5*]**

[1]King Abdullah University of Science and Technology (KAUST), Red Sea Research Center, Thuwal, 23955-6900, Saudi Arabia

[2]Harvard University, Department of Organismic and Evolutionary Biology, Cambridge, MA 02138, USA

[3]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center, Thuwal, 23955-6900, Saudi Arabia

[4]Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, PA 16802, USA

[5]University of Florida/IFAS, Department for Microbiology & Cell Science, Fort Lauderdale Research and Education Center, Davie, FL 33314, USA

*For correspondence. E-mail: ustingl@ufl.edu

Originality-Significance Statement: This manuscript describes the discovery of a potential pyrrolysine system in a novel, uncultured group of extremophilic Archaea based on single cell genomics data.

Running title: MSBL1 potentially decodes pyrrolysine

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1111/1462-2920.12545

**Summary**

Pyrrolysine (Pyl), the 22nd canonical amino acid, is only decoded and synthesized by a limited number of organisms in the domains *Archaea* and *Bacteria*. Pyl is encoded by the amber codon UAG, typically a stop codon. To date, all known Pyl-decoding archaea are able to carry out methylotrophic methanogenesis. The functionality of methylamine methyltransferases, an important component of corrinoid-dependent methyltransfer reactions, depends on the presence of Pyl. Here, we present a putative *pyl* gene cluster obtained from single-cell genomes of the archaeal Mediterranean Sea Brine Lakes group 1 (MSBL1) from the Red Sea. Functional annotation of the MSBL1 single cell amplified genomes (SAGs) also revealed a complete corrinoid-dependent methyl-transfer pathway suggesting that members of MSBL1 may possibly be capable of synthesizing Pyl and metabolizing methylated amines.

## Introduction

Pyrrolysine (Pyl), the 22nd proteogenic amino acid, has been found in only a few organisms and a very small number of proteins (Prat *et al.*, 2012). Typically, substrate-specific methylamine methyltransferases (MtmB, MtbB, and MttB) incorporate Pyl into the in-frame and read-through amber codon (UAG), such that it translates as Pyl. In addition, Pyl was also found in the $tRNA^{His}$ guanylyltransferase from *Methanosarcina acetivorans* in a capacity unrelated to the methylamine metabolism. This observation is thought to be a product of neutral evolution, in which no selective advantage is conferred (Heinemann *et al.*, 2009).

Prior to this study, Pyl-decoding archaea were mainly phylogenetically affiliated to the family *Methanosarcinaceae* (Prat *et al.*, 2012) and the order *Methanomassiliicoccales* (Borrel *et al.*, 2014a; Lang *et al.*, 2015). In general, all Pyl-decoding archaea share the tendency to minimize the frequency of TAG codons in their genomes and possess the gene cassette *pylTSBCD* required for Pyl biosynthesis and for encoding of the UAG codon as Pyl (Prat *et al.*, 2012; Gaston *et al.*, 2011).

The candidate division Mediterranean Sea Brine Lakes group 1 (MSBL1) is one of the most prevalent phylotypes in the deep-sea anoxic brines in both the Mediterranean Sea and the Red Sea (van der Wielen *et al.*, 2005; Guan *et al.*, 2015). Although MSBL1 has yet to be cultivated, Mwirichia and colleagues (Mwirichia *et al.*, 2016) recently reported 32 MSBL1 single-amplified genomes (SAGs) with less than 57% estimated completeness from four deep-sea brines in the Red Sea (accession numbers LHXJ00000000-LHYO00000000). These SAGs have been decontaminated from foreign sequences based on GC content, size, phylogenetic affiliation, and tetranucleotide frequency. The study showed that the MSBL1 group clusters together (with SAG AAA382A20 branching off separately from the other SAGs) and

metabolic reconstruction analyses suggested that MSBL1 are likely sugar-fermenting organisms capable of autotrophic growth and not methanogens as hypothesized previously (van der Wielen *et al.*, 2005; Borin *et al.*, 2009; Yakimov *et al.*, 2013).

To expand our understanding on metabolic strategies of MSBL1, we analyzed these SAGs for the potential to encode and synthesize pyrrolysine.

**Results and discussion**

The presence of Pyl-containing methylamine methyltransferases (in SAGs: AAA382A03, AAA382A13, and AAA382A20) that catalyze the methyl-transfer reaction of methylated amines to coenzyme M suggests that MSBL1 is capable of metabolizing methylamines (Table 1 and Table S1). The majority of Pyl-containing methyltransferases from MSBL1 have higher sequence similarities to homologs in methanogenic species of the domain *Archaea* than to those in the domain *Bacteria* (Table S2). Proteins related to Pyl-containing trimethylamine:corrinoid methyltransferase (MttB; COG5598 Superfamily) in MSBL1 are related to archaeal and bacterial versions of MttBs and form a sister clade to MttBs reported previously (Figure S1). The non-Pyl methyltransferase KXA99101 from SAG AAA259M10 is the closest glycine betaine:corrinoid methyltransferase (DSY3156)-like protein (Ticak *et al.*, 2014) found in MSBL1, both by phylogenic relationship and by sequence similarity (32%).

In support of a potential Pyl metabolism in MSBL1, we found evidence of Pyl system genes in MSBL1 SAGs. We identified the full set of genes required for all the machinery necessary for Pyl biosynthesis and incorporation in SAG AAA382A20, including *pylBCD*, a predicted $tRNA^{Pyl}$ (Figure S2), and the canonical $tRNA^{Pyl}$ synthetase gene (*pylS*). In SAG AAA382A03, a transposase interrupts the *pyl* gene

cassette, leaving only genes for the Pyl biosynthesis radical SAM protein (PylB), PylS, and a Pyl-containing dimethylamine methyltransferase remaining in the adjacent region (Figure 1 and Table 2, MSBL1 SAGs are accessible via INDIGO http://www.cbrc.kaust.edu.sa/indigo/dataCategories.do).

Similar to the case of members belonging to the Pyl-decoding *Methanomassiliicoccales*, the domain organization of PylS in MSBL1 SAGs lacks an N-terminal extension (also a separate *pylSn* gene has not been found so far), unlike the complete version found in *Methanosarcinaceae* (Gaston *et al.*, 2011). The catalytic activity of PylS was not eliminated by the missing N terminal region (Jiang and Krzycki, 2012). In SAG AAA382A20, Pyl biosynthesis genes (*pylC* and *pylD*) are both transcribed in the opposite direction to *pylS* and *pylB* (Figure 1), different from all other Pyl-decoding microorganisms reported (Gaston *et al.*, 2011; Borrel *et al.*, 2014a).

To predict whether the *pyl* operons in SAG AAA382A20 can be transcribed with the archaeal machinery, six genes encoding Pyl-system and Pyl-utilizing enzymes on a single contig (LHYE01000044) were classified into three co-transcribed operons (*pylS* and *pylB*, *pylD* and *pylC*, and *mtmB* and *mtbC*) by using the FGENESB online server (Solovyev and Salamov, 2011). Figure 1 shows the promoter sequences and transcription factor B recognition elements (BRE) predicted from the genome sequence of SAG AAA382A20 (also see Table S3). There are at least two potential versions of PylD translation, but the default version was generated by INDIGO (Mwirichia *et al.*, 2016; Alam *et al.*, 2013) using GTG as a translation initiator. This PylD enzyme is comprised of 312 amino acids, of which the 5'-end has a fragment that is 32 to 40 amino acids longer than its homologs in the NCBI non-redundant GenBank CDS database. The transcript containing *pylD* and *pylC* does not contain an

obvious TATA box in its 5'-UTR. Alternatively, the PylD enzyme with ATG as the translation initiator comprises 272 amino acids and contains two predicted TATA-boxes in the upstream coding region, suggesting that these MSBL1 genes may be under the archaeal regulatory machinery. These results suggest an interesting avenue for further investigation of the regulatory properties of MSBL1.

The direct involvement of selenocysteine or pyrrolysine in enzymes that synthesize them has been reported previously, for example in selenophosphate synthetase (selenocysteine-containing) of *Methanococcus maripaludis* (Stock *et al.*, 2010) and in PylB protein (Pyl-containing) of *Methanosarcina acetivorans* (O'Donoghue *et al.*, 2014). Similarly, the annotation of MSBL1 SAGs suggests that some of the *pyl* genes in MSBL1 contain in-frame amber codons that could incorporate Pyl (Figure 1).

Pyl machinery proteins in MSBL1 SAGs AAA382A20 and AAA382A03 have low amino acid sequence similarities (<46%) to their homologs in other cultivated or *Candidatus* species. Phylogenetic analysis of each of the predicted MSBL1 PylB/C/D/S proteins in relation to their homologs revealed additional evidence on the evolutionary history of Pyl system genes involving lateral gene transfers among taxa (Figure 2) (Fournier, 2009; Borrel *et al.*, 2014a). All four Pyl system enzymes in MSBL1 are deep-branching with the PylC/D/S proteins being most closely related to those in *Acetohalobium arabaticum*. MSBL1 and *Acetohalobium* are both found in hypersaline environments and their genomes indicate that they are metabolically versatile. Previous studies revealed that *Acetohalobium* has chemolithoautotrophic, organotrophic, and methylotrophic metabolic potential (Sikorski *et al.*, 2010; Prat *et al.*, 2012), and that MSBL1 has sugar-fermenting and autotrophic capabilities (Mwirichia *et al.*, 2016). In this study, we found that (similar to *Acetohalobium*) the corrinoid-dependent methyl-transfer pathway is present in at least some of the SAGs

of MSBL1. Taken together, the similarities between MSBL1 and *Acetohalobium* in their Pyl enzymes may reflect their ability to use the Pyl system in response to methylamines as carbon sources and to thrive in hypersaline habitats. In addition, we found all four enzymes in the Pyl system of *Methermicoccus shengliensis*, a member of the order *Methanosarcinales*, are closely related to homologs in the order *Methanomassiliicoccales* rather than those in other *Methanosarcinales,* making Pyl systems in *Methanosarcinales* paraphyletic.

Pyl-decoding archaea (methanogens) constitutively encode UAG as Pyl regardless of their carbon source (Heinemann *et al.*, 2009) and generally minimize their amber codon usage frequency to only about 4% of the CDS (Prat *et al.*, 2012) (except for *Methanomassiliicoccus luminyensis* – 11.3% (Borrel *et al.*, 2014b) and *Methermicoccus shengliensis* – 25%. Unlike these methylotrophic methanogenic archaea, the amber codon comprises about 23% of the CDS of MSBL1 SAGs, in ratios that are more similar to Pyl-decoding species from the domain *Bacteria* (Prat *et al.*, 2012) than to those from the domain *Archaea* (Table 3). Thus, we speculate that the methylamine metabolism could be one of the factors contributing to the regulation of the Pyl machinery in MSBL1, similar to the strategy used by *Acetohalobium arabaticum* in response to trimethylamine as a growth substrate (Prat *et al.*, 2012).

Genomic analysis suggests that MSBL1 SAG AAA382A20 harbors all the necessary enzymes for methanogenesis from methylamines except for the elusive methyl coenzyme M reductase, which catalyzes the last step in the production of methane (Table S1). To add to work by Mwirichia *et al.* (Mwirichia *et al.*, 2016), we propose that MSBL1 SAG AAA382A20 is capable of oxidizing and using methyl groups to produce reducing equivalents.

A number of reasons may explain why only two MSBL1 SAGs have Pyl-decoding potential. It may be that genomic regions containing Pyl-systems could be missed due to the incompleteness of SAGs (estimated 0.5% – 54.5%, Mwirichia *et al.* (2016)). In addition, metabolic diversification and niche partitioning within the candidate division MSBL1 may also be responsible for the presence/absence of such a system. Both MSBL1 SAGs (AAA382A03 and AAA382A20) bearing *pyl* genes reported in this study were retrieved from Nereus Deep of the Red Sea. SAG AAA382A20 is phylogenetically divergent from any other MSBL1 SAG analyzed (Mwirichia *et al.*, 2016). Based on the current availability and quality of MSBL1 genome fragments, we predict that only certain MSBL1 sub-groups may potentially decode Pyl and be involved in methylamine metabolism, however this can only be confirmed by analyzing more complete MSBL1 genomes.

In summary, a greater abundance of sequenced genomes has enabled the discovery of additional organisms that genetically decode Pyl. Here, we report that members of MSBL1—one of the most abundant organisms from the domain *Archaea* in the deep-sea hypersaline brines of the Red Sea, and coincidentally, *Methermicoccus shengliensis*, a thermophilic and methylotrophic methanogenic archaeon previously isolated from oil-production water—putatively decode Pyl. We provide additional evidence for potential lateral gene transfer events in the evolution of Pyl system genes (Borrel *et al.*, 2014a). We also report that putatively Pyl-decoding members of MSBL1 as well as the methylotrophic methanogen *Methermicoccus shengliensis* do not minimize their TAG codon frequency in the genome, indicating that in some *Archaea* a potentially unusual mechanism may occur for the regulation of Pyl synthesis and its incorporation into proteins.

**List of figures:**

**Figure 1.** Schematic comparison of the putative Pyl synthesis and synthetase gene cluster region identified in the MSBL1 SAG AAA382A20 and SAG AAA382A03. *PylB/C/D*: Pyl biosynthesis genes *pylB*, *pylC* and *pylD*. *PylS*: Pyl synthetase gene. BRE: transcription factor B recognition elements. TAG with red bar underneath indicates the position of the stop codon TAG that is predicted to be translated as Pyl.

**Figure 2.** Phylogenetic inference of deduced peptide sequences of individual *pylS/B/C/D* genes. A) PylS. B) PylB. C) PylC. D) PylD. Sequences from MSBL1 SAGs are in red, sequences from *Methanosarcinales* are shown in blue, sequences of the 7$^{th}$ order methanogen *Methanomassiliicoccales* are shown in green, and bacterial sequences are in black. (Each set of PylB/C/D/S sequences were aligned by MergeAlign (Collingridge and Kelly, 2012) using MAFFT (Katoh *et al.*, 2005) and 91 other different amino acid substitution matrices. Only columns with a score ≥90% from the constituent alignments were kept for consensus alignment. Finally, maximum-likelihood phylogenetic inference was performed based on individual protein sequences reconstructed by PhyML (Guindon *et al.*, 2010) using Le and Gascuel's amino acid substitution matrix (Le and Gascuel, 2008). Branch support was calculated using 1000 bootstraps.

**List of tables:**

**Table 1.** Pyl-containing methylamine methyltransferases in MSBL1 SAGs. Pfam domains and E-values of enzymes were predicted using the online server (http://pfam.xfam.org) of the Pfam protein families database (Finn *et al.*, 2015).

**Table 2.** Pyl systems in MSBL1 SAGs. PylB/C/D are responsible for pyrrolysine biosynthesis. PylS/T are responsible for directing pyrrolysine incorporation into proteins.

**Table 3.** Percentage of open reading frames (ORFs) with predicted TAG codon relative to the total ORFs in the MSBL1 SAGs and various genomes of Pyl-decoding microorganisms.

**Supplementary documents**

**Figure S1.** PhyML phylogenetic tree (n = 500 bootstraps) showing the affiliations of Pyl-containing and nonpyrrolysine members of COG5598 from MSBL1. Sequences from MSBL1 SAGs are highlighted in red: The Pyl-containing MttBs are AAA382A13 (AKJ50_01720 AKJ50_01715), AAA382A03 (AKJ49_00205 AKJ49_00210) and AAA382A03 (AKJ49_00035 AKJ49_00040), while the nonpyrrolysine members are KXA91217.1, KXA97307.1, KXB05407.1, KXB04566.1 and KXA99101.1. The glycine betaine:corrinoid methyltransferase from *Desulfitobacterium hafniense* is shown in blue.

**Figure S2.** A) Predicted secondary structure of $tRNA^{Pyl}$ from MSBL1 SAG AAA382A20. The secondary structure of MSBL1 $tRNA^{Pyl}$ (GC-content 70%) was predicted using an Aragorn tRNA scan (Laslett and Canback, 2004). B) The nucleotide sequence of AAA382A20 contig 202. This contig contains a predicted $tRNA^{Pyl}$ (underlined), which had previously been removed from SAG AAA382A20 due to its short length that did not meet the original criteria of Mwirichia *et al*, 2016.

**Table S1.** Predicted enzymes related to the use of methylamines (mono-, di-, and tri-) and methanol in MSBL1 SAGs.

**Table S2.** Top 5 Blastp hits of Pyl-containing methylamine methyltransferases in MSBL1 SAGs AAA382A20, AAA382A03, and AAA382A13.

**Table S3.** Identified promoter sequences related to the Pyl systems in MSBL1 SAGs.

**Acknowledgements**

This study was supported by King Abdullah University of Science and Technology (KAUST) through baseline funding and the SEDCO Research Excellence award to US.

**References**

Alam I, Antunes A, Kamau AA, alawi WB, Kalkatawi M, Stingl U, *et al.* (2013). INDIGO – INtegrated Data Warehouse of MIcrobial GenOmes with examples from the Red Sea extremophiles. *PLoS ONE* **8**: e82210.

Borin S, Brusetti L, Mapelli F, D'Auria G, Brusa T, Marzorati M, *et al.* (2009). Sulfur cycling and methanogenesis primarily drive microbial colonization of the highly sulfidic Urania deep hypersaline basin. *Proc Natl Acad Sci USA* **106**: 9151–9156.

Borrel G, Gaci N, Peyret P, O'Toole PW, Gribaldo S, Brugere J-F. (2014a). Unique characteristics of the pyrrolysine system in the 7th order of methanogens: implications for the evolution of a genetic code expansion cassette. *Archaea* **2014**: 1–11.

Borrel G, Parisot N, Harris HMB, Peyretaillade E, Gaci N, Tottey W, *et al.* (2014b). Comparative genomics highlights the unique biology of *Methanomassiliicoccales*, a *Thermoplasmatales*-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics* **15**: 679.

Collingridge PW, Kelly S. (2012). MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinf* **13**: 117.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, *et al.* (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**: gkv1344–D285.

Fournier G. (2009). Horizontal gene transfer and the evolution of methanogenic pathways. *Methods Mol Biol* **532**: 163–179.

Gaston MA, Jiang R, Krzycki JA. (2011). Functional context, biosynthesis, and genetic encoding of pyrrolysine. *Current Opinion in Microbiology* **14**: 342–349.

Guan Y, Hikmawan T, Antunes A, Ngugi D, Stingl U. (2015). Diversity of methanogens and sulfate-reducing bacteria in the interfaces of five deep-sea anoxic brines of the Red Sea. *Research in Microbiology*. e-pub ahead of print, doi: 10.1016/j.resmic.2015.07.002.

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

Heinemann IU, O'Donoghue P, Madinger C, Benner J, Randau L, Noren CJ, *et al.* (2009). The appearance of pyrrolysine in $tRNA^{His}$ guanylyltransferase by neutral evolution. *Proceedings of the National Academy of Sciences* **106**: 21103–21108.

Jiang R, Krzycki JA. (2012). PylSn and the homologous N-terminal domain of pyrrolysyl-tRNA synthetase bind the tRNA that is essential for the genetic encoding of pyrrolysine. *J Biol Chem* **287**: 32738–32746.

Katoh K, Kuma K-I, Toh H, Miyata T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511–518.

Lang K, Schuldes J, Klingl A, Poehlein A, Daniel R, Brune A. (2015). New mode of energy metabolism in the seventh order of methanogens as revealed by comparative genome analysis of '*Candidatus* Methanoplasma termitum'. *Appl Environ Microbiol* **81**: 1338–1352.

Laslett D, Canback B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**: 11–16.

Le SQ, Gascuel O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol* **25**: 1307–1320.

Mwirichia R, Alam I, Rashid M, Vinu M, Ba-Alawi W, Anthony Kamau A, *et al.* (2016). Metabolic traits of an uncultured archaeal lineage -MSBL1- from brine pools of the Red Sea. *Sci Rep* **6**: 19181.

O'Donoghue P, Prat L, Kucklick M, Schäfer JG, Riedel K, Rinehart J, *et al.* (2014). Reducing the genetic code induces massive rearrangement of the proteome. *Proceedings of the National Academy of Sciences* **111**: 17206–17211.

Prat L, Heinemann IU, Aerni HR, Rinehart J, O'Donoghue P, Söll D. (2012). Carbon source-dependent expansion of the genetic code in bacteria. *Proceedings of the National Academy of Sciences* **109**: 21070–21075.

Sikorski J, Lapidus A, Chertkov O, Lucas S, Copeland A, Glavina Del Rio T, *et al.* (2010). Complete genome sequence of Acetohalobium arabaticum type strain (Z-7288). *Standards in Genomic Sciences* **3**: 57–65.

Solovyev V, Salamov A. (2011). Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In: *Metagenomics and its applications in agriculture, biomedicine and environmental studies | Clc*. pp 61–78.

Stock T, Selzer M, Rother M. (2010). In vivo requirement of selenophosphate for

selenoprotein synthesis in archaea. *Molecular Microbiology* **75**: 149–160.

Ticak T, Kountz DJ, Girosky KE, Krzycki JA, Ferguson DJ. (2014). A nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is a glycine betaine methyltransferase. *Proceedings of the National Academy of Sciences* 201409642.

van der Wielen PWJJ, Bolhuis H, Borin S, Daffonchio D, Corselli C, Giuliano L, *et al.* (2005). The enigma of prokaryotic life in deep hypersaline anoxic basins. *Science* **307**: 121–123.

Yakimov MM, La Cono V, Slepak VZ, La Spada G, Arcadi E, Messina E, *et al.* (2013). Microbial life in the Lake Medee, the largest deep-sea salt-saturated formation. **3**: 3554.

**Table 1.** Pyl-containing methylamine methyltransferases in MSBL1 SAGs. Pfam domains and E-values of enzymes were predicted using the online server (http://pfam.xfam.org) of the Pfam protein families database (Finn *et al.*, 2015).

| SAG ID | Contig | Predicted coding region | Locus tag | Functional annotation | Pfam domain | E-value |
|---|---|---|---|---|---|---|
| **AAA382A03** | LHYC01000006 | 8573 -> 7197 | AKJ49_00390, AKJ49_00385 and their intergenic region | Dimethylamine methyltransferase | Dimethylamine methyltransferase (MtbB) | 1.2E-172 |
| | LHYC01000001 | 5402 -> 6868 | AKJ49_00035, AKJ49_00040 and their intergenic region | Trimethylamine methyltransferase | Trimethylamine methyltransferase (MttB) | 1.4E-97 |
| | LHYC01000003 | 5148 -> 6587 | AKJ49_00205, AKJ49_00210 and their intergenic region | Trimethylamine methyltransferase | Trimethylamine methyltransferase (MttB) | 3.5E-99 |
| **AAA382A13** | LHYD01000032 | 2105 -> 678 | AKJ50_01720, AKJ50_01715 and their intergenic region | Trimethylamine methyltransferase | Trimethylamine methyltransferase (MttB) | 1.4E-100 |
| **AAA382A20** | LHYE01000044 | 3839 -> 2418 | AKJ51_03645, AKJ51_03640 and their intergenic region | Monomethylamine methyltransferase | Monomethylamine methyltransferase (MtmB) | 6.8E-187 |

**Table 2.** Pyl systems in MSBL1 SAGs. PylBCD are responsible for Pyl biosynthesis. PylS/T are responsible for directing Pyl incorporation into protein.

| SAG | Contig | Locus tag | Functional domain by NCBI Blastp | Pfam domain | E-value | tRNA by ARAGON scan |
|---|---|---|---|---|---|---|
| **AAA382A20** | LHYE01000044 | AKJ51_03665 | Pyrolysyl-tRNA synthetase, PylS | tRNA synthetases class II core domain | 3.7E-09 | |
| | LHYE01000044 | AKJ51_03660 | Pyrrolysine biosynthesis radical SAM protein, PylB | Radical SAM superfamily | 2.2E-17 | |
| | LHYE01000044 | AKJ51_03655 | Pyrrolysine biosynthesis protein PylC | ATP-grasp domain | 9.6E-16 | |
| | LHYE01000044 | AKJ51_03650 | Pyrrolysine biosynthesis protein PylD | - | - | |
| | 202 (INDIGO)[a] | AAA382A20_01503 | - | - | - | Pyl-tRNA |
| **AAA382A03** | LHYC01000006 | AKJ49_00405[b] | Pyrolysyl-tRNA synthetase, PylS | tRNA synthetases class II | 0.00013 | |
| | LHYC01000006 | AKJ49_00400 | Pyrrolysine biosynthesis radical SAM protein, PylB | Radical SAM superfamily | 1.4E-17 | |

[a] The nucleotide sequence of AAA382A20 contig 202 can be found in Figure S1 B.
[b] Partial gene

**Table 3.** Percentage of open reading frames (ORFs) with predicted TAG codon relative to the total ORFs in the MSBL1 SAGs and various genomes of Pyl-decoding microorganisms.

| **MSBL1 SAGs** | **TAG ORFs, %** |
|---|---|
| AAA259A05 | 23.7 |
| AAA259B11 | 24.1 |
| AAA259D14 | 22.4 |
| AAA259D18 | 23.5 |
| AAA259E17 | 27.5 |
| AAA259E19 | 26.6 |
| AAA259E22 | 24.7 |
| AAA259I07 | 23.9 |
| AAA259I09 | 24.3 |
| AAA259I14 | 23 |
| AAA259J03 | 24.2 |
| AAA259M10 | 22.4 |
| AAA259O05 | 25.8 |
| AAA261C02 | 22.4 |
| AAA261D19 | 27.2 |
| AAA261F17 | 23.4 |
| AAA261F19 | 26.8 |
| AAA261G05 | 21.5 |
| AAA261O19 | 22.5 |
| AAA382A03 | 17.0 |
| AAA382A13 | 20.9 |
| AAA382A20 | 21.1 |
| AAA382C18 | 26.1 |
| AAA382F02 | 26.4 |
| AAA382K21 | 29.0 |
| AAA382M17 | 25.2 |
| AAA382N08 | 14.9 |
| AAA385D11 | 22.7 |
| AAA385M02 | 14.9 |
| AAA385M11 | 18.1 |
| AAA833F18 | 22.0 |
| AAA833K04 | 20.2 |
| **Pyl-decoding *Methanosarcinales*** | **TAG ORFs, %** |
| *Methermicoccus shengliensis* | 25 |
| *Methanococcoides burtonii** | 5.8 |
| *Methanohalobium evestigatum** | 4.1 |
| *Methanohalophilus mahii** | 4.1 |
| *Methanosarcina acetivorans** | 5.7 |
| **Pyl-decoding methanogens within *Thermoplasmata*** | **TAG ORFs, %** |
| *Thermoplasmatales* archaeon BRNA1 | 2.0 |
| *Methanomassiliicoccales* archaeon Mpt1 | 1.8 |
| "*Ca.* Methanomethylophilus alvus", Mx1201* | 2.8 |
| "*Ca.* Methanomassiliicoccus intestinalis", Mx1-Issoire* | 5.3 |
| *Methanomassiliicoccus luminyensis** | 11.3 |
| **Pyl-decoding bacterial species*** | **TAG ORFs, %** |
| *Acetohalobium arabaticum* | 19.5 |
| *Desulfotomaculum acetoxidans* | 24.7 |
| *Desulfitobacterium dehalogenans* | 23.5 |
| *Desulfitobacterium hafniense DCB-2* | 22.4 |
| *Desulfitobacterium hafniense Y51* | 23.7 |
| *Desulfotomaculum gibsoniae* | 28.5 |
| *Desulfosporosinus meridiei* | 26.8 |
| *Desulfosporosinus orientis* | 22.7 |
| *Desulfosporosinus yougiae* | 23.6 |
| *Thermincola potens* | 24.2 |

* Data from Prat *et al.* (2012) and Borrel *et al.* (2014)

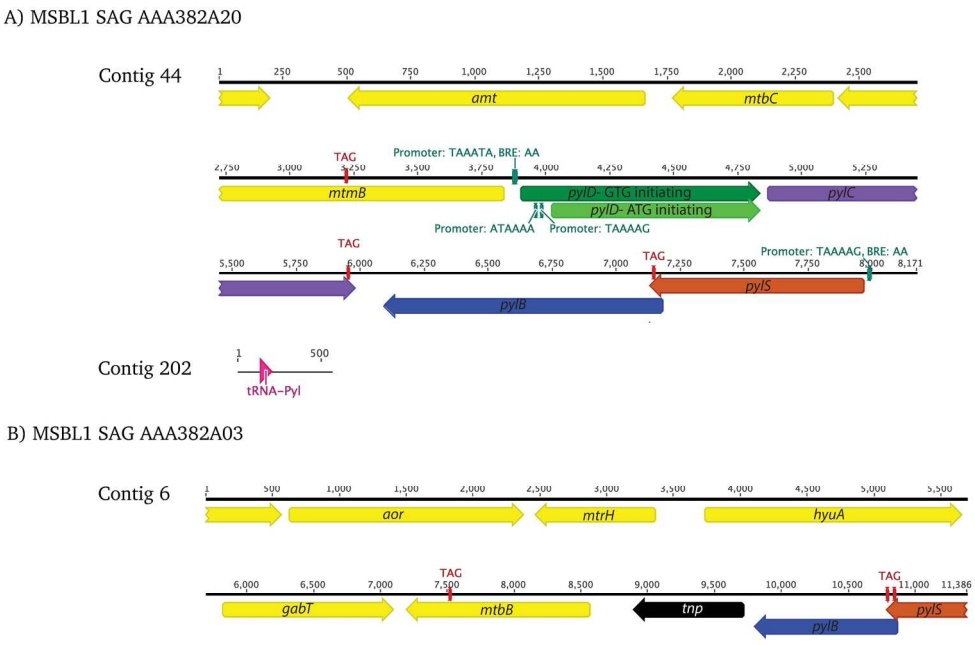


Figure 1. Schematic comparison of the putative Pyl synthesis and synthetase gene cluster region identified in the MSBL1 SAG AAA382A20 and SAG AAA382A03. PylB/C/D: Pyl biosynthesis genes pylB, pylC and pylD. PylS: Pyl synthetase gene. BRE: transcription factor B recognition elements. TAG with red bar underneath indicates the position of the stop codon TAG that is predicted to be translated as Pyl.

190x142mm (300 x 300 DPI)

Phylogenetic inference of deduced peptide sequences of individual pylS/B/C/D genes. A) PylS. B) PylB. C) PylC. D) PylD. Sequences from MSBL1 SAGs are in red, sequences from Methanosarcinales are shown in blue, sequences of the 7th order methanogen Methanomassiliicoccales are shown in green, and bacterial sequences are in black. (Each set of PylB/C/D/S sequences were aligned by MergeAlign (Collingridge and Kelly, 2012) using MAFFT (Katoh et al., 2005) and 91 other different amino acid substitution matrices. Only columns with a score ≥90% from the constituent alignments were kept for consensus alignment. Finally, maximum-likelihood phylogenetic inference was performed based on individual protein sequences reconstructed by PhyML (Guindon et al., 2010) using Le and Gascuel's amino acid substitution matrix (Le and Gascuel, 2008). Branch support was calculated using 1000 bootstraps.

612x584mm (600 x 600 DPI)