Gene orders in the upstream of 16S rRNA genes divide genera of the family *Halobacteriaceae* into two groups

Hiroaki Minegishi,¹ Masahiro Kamekura,² Tomomi Kitajima-Ihara,³ Kaoru Nakasone,⁴ Akinobu Echigo,^{1,5} Yasuhiro Shimane,⁶ Ron Usami,^{1,5,6} Takashi Itoh⁷ and Kunio Ihara³

Correspondence

Hiroaki Minegishi
minehiro@toyo.jp

¹Bio-Nano Electronics Research Center, Toyo University, Kawagoe, Saitama, Japan

²Halophiles Research Institute, Noda, Chiba, Japan

³Center for Gene Research, Nagoya University, Nagoya, Aichi, Japan

⁴Department of Chemistry and Environmental Technology, School of Engineering, Kinki University, Higashi-Hiroshima, Hiroshima, Japan

⁵Faculty of Science and Engineering, Toyo University, Kawagoe, Saitama, Japan

⁶Graduate School of Interdisciplinary New Science, Toyo University, Kawagoe, Saitama, Japan

⁷Japan Collection of Microorganisms, RIKEN BioResource Center, Wako, Saitama, Japan

In many prokaryotic species, 16S rRNA genes are present in multiple copies, and their sequences in general do not differ significantly owing to concerted evolution. At the time of writing, the genus *Haloarcula* of the family *Halobacteriaceae* comprises nine species with validly published names, all of which possess two to four highly heterogeneous 16S rRNA genes. Existence of multiple heterogeneous 16S rRNA genes makes it difficult to reconstruct a biological phylogenetic tree using their sequence data. If the orthologous gene is able to be discriminated from paralogous genes, a tree reconstructed from orthologous genes will reflect a simple biological phylogenetic relationship. At present, however, we have no means to distinguish the orthologous rRNA operon from paralogous ones in the members of the family *Halobacteriaceae*. In this study, we found that the dihydroorotate oxidase gene, *pyrD*, was present in the immediate upstream of one 16S rRNA gene in each of ten strains of the family *Halobacteriaceae* whose genome sequences have been determined, and the direction of the *pyrD* gene was opposite to that of the 16S rRNA genes. In two other strains whose genome sequences have been determined, the *pyrD* gene was present in far separated positions. We designed PCR primer sets to amplify DNA fragments encompassing a region from the conserved region of the *pyrD* gene to a conserved region of the tRNA-Ala gene or the 23S rRNA gene to determine the 16S rRNA gene sequences preceded by the *pyrD* gene, and to see if the *pyrD* gene is conserved in the immediate upstream of rRNA operon(s) in the type strains of the type species of 28 genera of the family *Halobacteriaceae*. Seventeen type strains, including the ten strains mentioned above, gave amplified DNA fragments of approximately 4000 bp, while eleven type strains, including the two strains mentioned above, did not give any PCR products. These eleven strains are members of the Clade I haloarchaea, originally defined by Walsh *et al.* (2004) and expanded by Minegishi *et al.* (2010). Analysis of contig sequences of three strains belonging to the Clade I haloarchaea also revealed the absence of the *pyrD* gene in the immediate upstream of any 16S rRNA genes. It may be scientifically sound to hypothesize that during the evolution of members of the family *Halobacteriaceae*, a *pyrD* gene transposition event happened in one group and this was followed by subsequent speciation processes in each group, yielding species/genera of the Clade I group and 'the rest' of the present family *Halobacteriaceae*.

The GenBank/EMBL/DDBJ accession numbers for the 16S rRNA gene sequences determined in this study are AB355982–AB355983, AB355985–AB355989 and AB605776–AB605778.

Three supplementary figures and five supplementary tables are available with the online version of this paper.

INTRODUCTION

The three rRNA genes, 16S, 23S and 5S rRNA, are typically linked together into an operon. In many prokaryotic species, rRNA operons are present in multiple copies and their sequences in general do not differ significantly owing to concerted evolution (Eickbush & Eickbush, 2007; Liao, 2000; Nei & Rooney, 2005). The extreme halophile *Haloarcula marismortui*, a species of the family *Halobacteriaceae*, however, was the first in the domain 'Archaea' to be shown to have highly heterogeneous (95 % similarity) 16S rRNA genes (Mylvaganam & Dennis, 1992; Dennis *et al.*, 1998). The genome sequence of *Har. marismortui* (Baliga *et al.*, 2004) confirmed the presence of three genes, *rrs-1*, *rrs-2* and *rrs-3*, and the similarity between *rrs-1* and *rrs-3* was 94.4 %. Members of the family *Halobacteriaceae*, the single family in the order *Halobacteriales* of phylum 'Euryarchaeota', have been classified within 119 species of 32 genera using polyphasic approaches. The family accommodates the largest number of genera in the domain 'Archaea'. Species of eleven other genera also possess two to four copies of the 16S rRNA gene, with high heterogeneity in some species: *Haladaptatus* (Savage *et al.*, 2007), *Halobaculum* (H. Minegishi, unpublished data), *Haloferax* (Hartman *et al.*, 2010), *Halomicrobium* (Cui *et al.*, 2009), *Haloquadratum* (Bolhuis *et al.*, 2006), *Halorubrum* (KEGG, see Methods), *Halosimplex* (Vreeland *et al.*, 2002), *Haloterrigena* (KEGG), *Natrialba* (KEGG), *Natrinema* (Boucher *et al.*, 2004) and *Natronoarchaeum* (Shimane *et al.*, 2010). The greatest difference is 9.4 % between the genes Hmuk_R0017 and Hmuk_R0036 of *Halomicrobium mukohataei* (retrieved from KEGG). At the time of writing, the genus *Haloarcula* comprises nine species with validly published names and a few species whose names have not been validly published, and most of them possess two to four highly heterogeneous 16S rRNA genes. Phylogenetic trees reconstructed from these gene sequences (Cui *et al.*, 2009), however, are not able to suggest which species are the closest relatives of each other. If the orthologous gene is discriminated from paralogous genes (Koonin, 2005), a tree reconstructed from orthologous genes only will give a simpler topology. At present, however, we have no means to distinguish the orthologous rRNA operon from paralogous ones in the members of the family *Halobacteriaceae*.

In 2006, we looked closely at the gene orders in the upstream and downstream of 16S rRNA gene(s) in genome sequences of *Halobacterium salinarum* strains R1 and NRC-1 (one copy), *Natronomonas pharaonis* DSM 2160 (one copy), *Haloquadratum walsbyi* HBSQ001 (two copies), and *Har. marismortui* ATCC 43049 (three copies) that were available at that time. It was noticed that the genes encoding two proteins, MC1 (non-histone chromosomal protein) (Chartier *et al.*, 1989; Laine *et al.*, 1991) and dihydroorotate oxidase (*pyrD*) (Jacquet *et al.*, 1985; Larsen & Jensen, 1985), were present in the immediate upstream (in the range 2400–3400 bp) of the 5' terminal of the single copy of *Hbt. salinarum* strains R1 (OE_r01) and NRC-1 (VNGr02), and *Nmn. pharaonis* (Npr03), as well as

one copy of *Hqr. walsbyi* (HQR01) and *Har. marismortui* (*rrs-1*) out of two and three copies, respectively. The MC1 gene of *Hqr. walsbyi* was located in the opposite direction compared to those of the three other species. In the upstream of the other copies of the 16S rRNA gene, HQR06 of *Hqr. walsbyi* and *rrs-2* and *rrs-3* of *Har. marismortui*, however, the MC1 and *pyrD* genes were not present. Subsequently, three more genome sequences became public, *Hmc. mukohataei* DSM 12286^T (three copies), *Halorhabdus utahensis* DSM 12940^T (one copy) and *Halorubrum lacusprofundi* ATCC 49239^T (three copies). The MC1 and *pyrD* genes were again present in the immediate upstream of one copy (out of three copies in *Hmc. mukohataei* and *Hrr. lacusprofundi*) of the 16S rRNA gene. These data seemed to suggest that the 16S rRNA gene preceded by the MC1 and *pyrD* genes was the orthologue we were looking for.

In early 2010, genome sequences of *Haloterrigena turkmenica* DSM 5511^T (three copies) and *Natrialba magadii* ATCC 43099^T (three copies) were made public. The MC1 and *pyrD* genes, however, were not detected in the immediate upstream of any 16S rRNA gene of either species. The two genes were located in clearly separated positions from each other and from the 16S rRNA genes. This fact seemed to suggest that the 'orthologous' 16S rRNA genes were missing in these two species. On the other hand, this fact reminded us that *Htg. turkmenica* and *Nab. magadii* were members of the Clade I haloarchaea, which has been disclosed in the phylogenetic tree reconstruction from 16S rRNA gene sequences and from RNA polymerase B' gene sequences (Walsh *et al.*, 2004; Enache *et al.*, 2007; Minegishi *et al.*, 2010; Fig. 1). Now, the question is whether the species belonging to the Clade I haloarchaea are distinguished from other species of the family *Halobacteriaceae* from the viewpoint of gene order phylogeny (Belda *et al.*, 2005; Luo *et al.*, 2009). As of January 2011, complete genome sequences of twelve species of the family *Halobacteriaceae* are publically available with annotation (KEGG), including the recently released *Halalkalicoccus jeotgali* B3^T, *Haloferax volcanii* DS2^T and *Halogeometricum borinquense* DSM 11551^T. Both the MC1 and *pyrD* genes were again present in the immediate upstream of a 16S rRNA gene of all three strains.

The *pyrD* gene encodes dihydroorotate oxidase (EC 1.3.98.1), required for *de novo* synthesis of uridine monophosphate. This is a single copy gene, at least in all archaeal species investigated so far, and the enzyme proteins consist of 349–357 amino acids in members of the family *Halobacteriaceae* and have highly conserved regions. The direction of the *pyrD* gene, when present, is opposite to that of the 16S rRNA genes, with no exceptions up to now (Table 1). In this paper, we designed PCR primer sets based on the conserved regions of the *pyrD* and tRNA-Ala genes or the 23S rRNA genes to determine the 16S rRNA gene sequences preceded by *pyrD*, and to see if the *pyrD* gene is conserved in the immediate upstream of rRNA operon(s) in strains of the type species of 28 genera

Fig. 1. Neighbour-joining tree inferred from sequences of 16S rRNA genes of type strains of type species of 28 genera used in this study, with sequence M60880 of *Methanospirillum hungatei* as an outgroup. Sequences of Clade I strains are surrounded by a broken line. Support for nodes corresponding to bootstrap values for 1000 replicates is shown; only values greater than 50 % are displayed. Bar, 0.02 substitutions per nucleotide position.

of the family *Halobacteriaceae*. We also extended the survey of genes further upstream of the 16S rRNA genes, as well as upstream and downstream of some other genes, to find any gene or gene array common to Clade I or members outside of Clade I, which is designated 'the rest' in this paper.

METHODS

Strains. The following type strains, except *Hqr. walsbyi*, of type species of 28 genera were cultivated in 3 ml media recommended by JCM (Japan Collection of Microorganisms) and genomic DNA was extracted and purified as described previously (Minegishi *et al.*, 2010): *Haladaptatus paucihalophilus* JCM 13897^T, *Halalkalicoccus tibetensis* JCM 11890^T, *Halarchaeum acidiphilum* MH1-52-1^T (=JCM 16109^T), *Haloarcula vallismortis* JCM 8877^T, *Hbt. salinarum* JCM 8978^T, *Halobaculum gomorrense* JCM 9908^T, *Halobiforma haloterrestris* JCM 11627^T, *Halococcus morrhuae* JCM 8876^T, *Hfx. volcanii* JCM 8879^T, *Hgm. borinquense* JCM 10706^T, *Hmc. mukohataei* JCM 9738^T, *Halopiger xanaduensis* JCM 14033^T, *Haloplanus natans* JCM 14081^T, *Hqr. walsbyi* DSM 16790, *Hrd. utahensis* JCM 11049^T, *Halorubrum saccharovororum* JCM 8865^T, *Halosarcina pallida* JCM 14848^T, *Halosimplex carlsbadense* JCM 11222^T, *Halostagnicola larsenii* JCM 13463^T, *Htg. turkmenica* JCM 9101^T, *Halovivax asiaticus* JCM 14624^T, *Natrialba asiatica* JCM 9576^T, *Natrinema pellirubrum* JCM 10476^T, *Natronobacterium gregoryi* JCM 8860^T, *Natronococcus occultus* JCM 8859^T, *Natronolimnobiobius baerhuensis* JCM 12253^T, *Nmn. pharaonis*

DSM 2160^T and *Natronorubrum bangense* JCM 10635^T. Members of the Clade I haloarchaea are surrounded by a broken line in Fig. 1. In addition, *Haloarcula amylolytica* JCM 13557^T, *Haloarcula japonica* JCM 7785^T, 'Haloarcula californiae' JCM 8912 and 'Haloarcula sinaiensis' JCM 8862 were also used in this study.

PCR primer sets and sequencing. The gene sequences of dihydroorotate oxidase, tRNA-Ala and 23S rRNA were downloaded from KEGG (Kyoto Encyclopedia of Genes and Genomes; <http://www.genome.jp/kegg/>) and two primer sets were designed.

pyrD primer set 1: forward primer (designed from tRNA-Ala gene sequences, corresponding to nt 45–72 of *Har. marismortui*) 5'-ATGGACCCACTGGGATTCTGAACCCAGGGC-3'; the reverse primer was a mixture of the following four primers, *pyrDrev1* 5'-GCGTCTCTGTCTGAAYCCYGCNGC-3', *pyrDrev2* 5'-GCGTCTCTGTCTGAAYC-CRGCNGC-3', *pyrDrev3* 5'-GCGTCTCTGTCTGAARCCYGCNGC-3' and *pyrDrev4* 5'-GCGTCTCTGTCTGAARCCRGCNGC-3', corresponding to nt 196–218 of *pyrD* of *Har. marismortui*. Amplification with KOD Fx DNA polymerase (Toyobo, Japan) was started by incubation at 96 °C for 1 min, followed by 30 cycles of 30 s at 94 °C, 30 s at 60 °C and 2 min at 72 °C, a final extension was performed for 2 min at 72 °C. Aliquots of 5 µl were analysed by agarose gel (1 %) electrophoresis to confirm the products, followed by sequencing of 16S rRNA gene regions with primers 1-f 5'-ATTCCGGTTGATCCTGCCGAGGCC-3', 1-r 5'-GG-CCTCCGGCAGGATCAACCGGAAT-3', 2-f 5'-TTATTGGGCCTAAA-GCGTCCGTAGC-3', 2-r 5'-GCTACGGACGCTTTAGGCCCAATAA-3', 3-f 5'-GCCGTCAGCTCGTACCGTGAGGCGT-3', 3-r 5'-ACGCCT-

Table 1. Gene array in the upstream of 16S rRNA genes of 18 species of the family *Halobacteriaceae*

MC1, non-histone chromosomal protein (102–109 aa); phe-tL, phe-tRNA ligase (EC 6.1.1.20) (1066–1149 aa); trp-tS, trp-tRNA Synthetase (EC 6.1.1.2) (513–614 aa); dld, dihydrolipoamide dehydrogenase (EC 1.8.1.4) (472–492 aa); E2, pyruvate dehydrogenase (EC 2.3.1.12) E2 component (478–563 aa); E1 α & β , pyruvate dehydrogenase E1 component subunit α & β (695–717 aa); LAsyn, lipoic acid synthase (EC 2.8.1.8) (308–321 aa); val-tS, val-tRNA synthetase (EC 6.1.1.9) (870–930 aa); pyrD, dihydroorotate oxidase (EC 1.3.98.1) (349–357 aa); mcO, multicopper oxidase type 3 [151(fragment?)–398 aa]; ND, gene not detected in the genome sequence. Arrows indicate the opposite direction of the genes.

16S rRNA gene	tRNA- 5S Cys	23S tRNA- 16S Ala														16S	tRNA- 23S Ala	5S	tRNA- Cys
<i>Halalkalicoccus jeotgali</i> HacjB3_r01			←dld	←E2	←E1β	←E1α	←LAsyn	4 genes	trp-tS	–	–	←pyrD	MC1	○	○	○	○	○	
‘ <i>Haloarcula californiae</i> ’ contig00131			←dld	←E2	←E1β	←E1α	←LAsyn		trp-tS	phe-tL	–	←pyrD	MC1	○					
<i>Haloarcula japonica</i> contig-rrnA									trp-tS	phe-tL	–	←pyrD	MC1	○	○	○	○	○	
<i>Haloarcula marismortui</i> rrs-1			←dld	←E2	←E1β	←E1α	←LAsyn	5 genes	trp-tS	phe-tL	–	←pyrD	MC1	○	○	○	○	○	
‘ <i>Haloarcula sinaiiensis</i> ’ contig00040										phe-tL	–	←pyrD	MC1	○	○	○			
<i>Haloarcula vallismortis</i> contig00026			←dld	←E2	←E1β	←E1α	←LAsyn	?	trp-tS	phe-tL	–	←pyrD	MC1	○					
<i>Halobacterium salinarum</i> VNGr02			–	–	–	–	–		–	phe-tL	–	←pyrD	MC1	○	○	○	○	○	
<i>Haloferax volcanii</i> HVO_3064			←dld	←E2	←E1β	←E1α	←LAsyn	5 genes	trp-tS	phe-tL	val-tS	←pyrD	MC1	○	○	○	○	–	
<i>Halogeometricum borinquense</i> Hbor_03170			←dld	←E2	←E1β	←E1α	←LAsyn	7 genes	trp-tS	phe-tL	val-tS	←pyrD	MC1	○	○	○	○	○	
<i>Halomicrobium mukohataei</i> Hmuk_R0017			–	←E2	←E1β	←E1α	←LAsyn	6 genes	trp-tS	phe-tL	–	←pyrD	MC1	○	○	○	○○	○	
<i>Haloquadratum walsbyi</i> HQr01			–	ND	ND	ND	ND		trp-tS	phe-tL	val-tS	←pyrD	←MC1	○	○	○	○	○	
<i>Halorhabdus utahensis</i> Huta_R0021			–	ND	ND	ND	ND		trp-tS	phe-tL	–	←pyrD	MC1	○	○	○	○	○	
<i>Halorubrum lacusprofundi</i> Hlac_R0001			–	–	–	–	–		–	–	val-tS	←pyrD	MC1	○	○	○	○	○	
<i>Natronomonas pharaonis</i> NPr03			–	–	–	–	–	←leu-tS	trp-tS	–	–	←pyrD	MC1	○	○	○	○	○	
<i>Haloterrigena turkmenica</i> Htur_R0063	←○	←○	←○	←○	←○	hyp x3	←mcO	←dld	←E2	←E1β	←E1α	←LAsyn	20 genes	trp-tS	phe-tL	–	–	MC1	
<i>Natrialba magadii</i> Nmag_R0026	←○	←○	←○	←○	←○	hyp x3	←mcO	←dld	←E2	←E1β	←E1α	←LAsyn	24 genes	trp-tS	phe-tL	–	–	MC1	
<i>Halopiger xanaduensis</i> contig60						←○	hyp x3	←mcO	←dld	←E2	←E1β	←E1α	←LAsyn	?	trp-tS	phe-tL	–	–	MC1
<i>Natrinema pellirubrum</i> contig246	←○	←○	←○	←○	←○	hyp x3	–	←dld	←E2	←E1β	←E1α	←LAsyn	?	trp-tS	phe-tL	–	–	MC1	

CACGGTACGAGCTGACGGC-3', 4-f 5'-TTAAGTCGTAACAAGGT-AGCCGTAG-3' and 4-r 5'-CTACGGCTACCTTGTTACGACTTAA-3'.

pyrD primer set 2: forward primer *pyrD* 5'-TCGTTGTTNARNCCCATNCGGTT-3' (corresponding to nt 346–368 of *pyrD* of *Har. marismortui*); reverse primer 23SRev 5'-GCTTWTGCG-AGCTTGG-3' (corresponding to nt 56–71 of the 23S rRNA gene of *Hbt. salinarum*). The reverse primer was designed from 23S rRNA gene sequences to avoid the possibility that the tRNA-Ala gene is missing in any 'orthologous' rRNA operons (see Results, Gene orders in downstream and further upstream). Amplification with Ex *Taq* DNA polymerase (TaKaRa) was started by incubation at 96 °C for 5 min, followed by 35 cycles of 20 s at 98 °C, 30 s at 60 °C and 1 min 30 s at 72 °C, a final extension was performed for 2 min at 72 °C. Aliquots of 5 µl were analysed by agarose gel (1%) electrophoresis to see if DNA fragments were amplified or not by PCR. In some cases, the products were purified for further sequencing as described above. Phylogenetic tree reconstruction was done as described by Minegishi *et al.* (2010).

Sources of downloaded sequences and annotations. Gene sequences and annotations were obtained from KEGG (Kanehisa & Goto, 2000; Kanehisa *et al.*, 2006). The contig sequences available at JGI - Genome Projects (Joint Genome Institute; <http://www.jgi.doe.gov/genome-projects/pages/projects.jsf?kingdom=Archaea>) and the Halophile genome site (<http://edwards.sdsu.edu/halophiles/>) were also downloaded for retrieval of genes. The rRNA operons and contigs analysed in this study are summarized in Supplementary Table S1 (available in IJSEM Online).

RESULTS

A single 16S rRNA gene is preceded by *pyrD* in strains of the genus *Haloarcula*

Contig sequences of three strains of the genus *Haloarcula*, *Har. vallismortis* ATCC 29715^T, '*Har. californiae*' ATCC 33799 and '*Har. sinaiiensis*' ATCC 33800, were obtained from the Halophile genome site and gene orders were analysed along with those from unpublished contigs of *Har. japonica* JCM 7785^T (see Supplementary Table S1). Amongst multiple 16S rRNA genes found in the contig sequences of each strain, a single 16S rRNA gene was preceded by the MC1 and *pyrD* genes. These data have demonstrated that one out of multiple 16S rRNA genes is preceded by the *pyrD* gene in all species of the genus *Haloarcula* without any exceptions. In order to confirm the presence of the *pyrD* gene, we designed *pyrD* primer set 1 to amplify the region from the *pyrD* gene to the tRNA-Ala gene and *pyrD* primer set 2 to amplify from the *pyrD* gene to the 23S rRNA gene. From seven species with validly published names and two additional species of the genus *Haloarcula*, approximately 4000 bp DNA fragments were amplified. Direct sequencing of the products gave unambiguous sequences of 16S rRNA genes, showing that regions containing the 16S rRNA gene preceded by the *pyrD* gene were selectively amplified. A *Haloarcula* 16S rRNA gene tree was reconstructed from sequences determined in this study, those retrieved from the Halophile genome site and high quality database-sequences (sequenced after cloning of PCR products)

(Supplementary Fig. S1). Sequences with many ambiguous nucleotides and sequences that were determined by direct sequencing of PCR products were not included in this analysis.

The tree clearly demonstrates that sequences determined in this study are identical to the sequences in genomes or contigs (in red). The multiple 16S rRNA gene sequences are divided into two groups: (a) containing all sequences preceded by *pyrD* and (b) containing all sequences not preceded by *pyrD*, plus *rrs-2* and *rrs-3* from *Har. Marismortui*, other sequences from *Har. Japonica* and *Har. Vallismortis*, and '*Har. sinaiiensis*' contig00146. A pruned tree reconstructed from selected sequences, nine from group (a) and five from group (b), is shown in Supplementary Fig. S2.

PCR amplification with *pyrD* primer set 2 from type species of 28 genera

The survey of the presence of the *pyrD* gene in the upstream of 16S rRNA genes was extended to other genera by PCR with *pyrD* primer set 2. If the *pyrD* gene is not present in the immediate upstream, in less than approximately 50 kbp, of any 16S rRNA gene, no PCR products will be obtained. While all type species of 11 genera belonging to the Clade I haloarchaea, *Hbf. haloterrestris*, *Hpg. xanaduensis*, *Hst. larsenii*, *Htg. turkmenica*, *Hvx. asiaticus*, *Nab. asiatica*, *Nnm. pellirubrum*, *Nbt. gregoryi*, *Ncc. occultus*, *Nln. baerhuensis* and *Nrr. bangense*, did not give any PCR products within our experimental conditions (Supplementary Fig. S3), the type species of 17 genera of 'the rest' gave approximately 4000 bp DNA fragments. These data strongly suggested that in genera of Clade I, 16S rRNA genes were not preceded by the *pyrD* gene, while in 'the rest', one 16S rRNA gene, out of multiple copies when present, is preceded by the *pyrD* gene.

Gene orders in the downstream of 16S rRNA genes

The genome maps of twelve species of the family *Halobacteriaceae* available at KEGG showed that the tRNA-Ala gene was present in the downstream of only one 16S rRNA gene, *rrs-1* or *Hmuk_R0017*, respectively, in *Har. marismortui* and *Hmc. mukohataei*, preceded by the *pyrD* gene (Supplementary Table S1). In the ten other species, including two in Clade I, the tRNA-Ala gene was present in all rRNA operons. The 23S rRNA genes were followed by one or two copies of the 5S rRNA genes without exception. The following tRNA-Cys gene, a single copy in all species of the genus *Halobacteriaceae*, is present only in the rRNA operon preceded by the *pyrD* gene, except in *Hfx. volcanii*. In the further downstream, no meaningful gene orders were detected, common to the twelve, or ten without Clade I, genome sequences.

Gene orders in the upstream of 16S rRNA genes of genera of 'the rest'

In the upstream of the ten 16S rRNA genes of the ten species where this gene is preceded by the *pyrD* gene, a conserved gene array was observed, in a region of 17 000 to 30 000 bp, in the order valyl-tRNA synthetase, phenylalanyl-tRNA synthetase, tryptophanyl-tRNA synthetase, lipoic acid synthase, pyruvate dehydrogenase E1 component subunit α and β , pyruvate dehydrogenase E2 component, dihydrolipoamide dehydrogenase. These are single copy genes in all strains of the family *Halobacteriaceae* whose genome sequences are reported. Some genes are not located within this region (Table 1) but placed in far separated positions in some genomes. These ten 16S rRNA genes preceded by the *pyrD* gene were all located on the main chromosome. More detailed gene orders between tryptophanyl-tRNA synthetase and lipoic acid synthase are summarized in Supplementary Table S2.

On the other hand, in the upstream regions of 16S rRNA genes not preceded by the *pyrD* gene (*rrs-2* and *rrs-3* of *Har. marismortui*, HVO_3038 of *Hfx. volcanii*, Hmuk_R0036 and Hmuk_R0060 of *Hmc. mukohataei*, HQr06 of *Hqr. walsbyi* and Hlac_R0031 and Hlac_R0058 of *Hrr. Lacusprofundi*), no meaningful common gene orders were detected. Furthermore, the following 16S rRNA genes were not located on the main chromosome: *rrs-2* on chromosome II, Hmuk_R0060 on plasmid pHmuk01 and Hlac_R0058 on chromosome 2. These observations overall seem to suggest that rRNA operons preceded by the *pyrD* and MC1 genes are orthologous in species of genera of 'the rest'.

Putative orthologous 16S rRNA gene in members of Clade I

In the two members of Clade I, the MC1 gene and the gene array of seven genes described above, from phenylalanyl-tRNA synthetase (phe-tS) to dihydrolipoamide dehydrogenase (*dld*), were preserved in the upstream of 16S rRNA genes (*Htg. turkmenica* Htur_R00063 and *Nab. magadii* Nmag_R0026) in a range of 39 to 49 kbp (Table 1). Furthermore, contig sequences of three species of the Clade I haloarchaea, *Hpg. xanaduensis* SH-6^T, *Halovivax ruber* XH-70^T and *Nnm. pellirubrum* DSM 15624^T, were made public from the JGI - Genome Projects. Analysis of these contigs showed that the gene array mentioned above was also detected in the upstream of the 16S rRNA gene in contig60 of *Hpg. xanaduensis* and contig246 of *Nnm. pellirubrum*, in exactly the same order as in *Htg. turkmenica* and *Nab. magadii* (Table 1). The *pyrD* gene was missing in this region in all strains of Clade I as far as we have investigated. This fact strongly suggests that a gene conversion event occurred in an ancestor of members of the Clade I haloarchaea, resulting in the change of direction of the rRNA operon, during the evolution of extremely halophilic archaea. It might be reasonable to conclude that the 16S rRNA genes, Htur_R0063 and Nmag_R0026, as well as those of contig60 and contig 246 are orthologous.

On the other hand, in the immediate upstream of two paralogous 16S rRNA genes, Htur_R0032 and Nmag_R0001, genes encoding a hypothetical protein and histidinol dehydrogenase were present, followed by many common genes (Supplementary Table S3) up to the PHP domain protein gene, Htur_1753 or Nmag_3538 (gene maps of KEGG) in *Htg. turkmenica* and *Nab. magadii*, respectively. Contig56 from *Hpg. xanaduensis* and contig232 from *Nnm. pellirubrum* also possessed this gene array (Supplementary Table S3). In the immediate downstream of the rRNA operon of *Htg. turkmenica* and *Nab. magadii*, there was an array of nine genes. These gene arrays were not detected in species of genera of 'the rest', suggesting the close relationship amongst species of the Clade I haloarchaea.

Positions of *pyrD* in members of Clade I and other Archaeal species

As described in the Introduction, the *pyrD* genes of *Htg. turkmenica* and *Nab. magadii* were located in definitely separated positions from the three 16S rRNA genes. Analysis of the gene maps of these two species and contig sequences of the three additional Clade I species revealed that the *pyrD* gene of all five Clade I species was 4–5 genes upstream of the Na⁺/H⁺ antiporter subunit gene cluster (Supplementary Table S4). The gene encoding site-specific recombinase XerD-like protein (XerD) in the upstream of the D1 subunit was also characteristic for the Clade I species. The similar Na⁺/H⁺ antiporter subunit gene clusters were also present in species of 'the rest', and the *dfp* gene was present in the upstream or downstream in all species except *Nmn. pharaonis*.

Ninety two archaeal genome sequences including members of the family *Halobacteriaceae* are now available at KEGG. In all sequences, the *pyrD* gene was a single copy gene, except '*Nanoarchaeum equitans*' in which the *pyrD* gene has not been detected. Only one methanogenic archaeon, *Methanospirillum hungatei* JF-1^T (Anderson *et al.*, 2009), was shown to possess the *pyrD* gene in the immediate upstream of a 16S rRNA gene, Mhun_R0027, out of four genes, although in the same direction as the 16S rRNA gene, contrarily to members of the family *Halobacteriaceae*. The gene array detected in the upstream of the *pyrD* gene in members of the family *Halobacteriaceae*, however, was not present in *M. hungatei* JF-1^T.

Other genes that distinguish the members of Clade I?

In the five species of the Clade I haloarchaea, a gene encoding response regulator receiver modulated GAF sensor protein was present in the immediate upstream of the tRNA-Asp gene, followed by genes encoding RNA polymerase subunits H, B'', B', A' and A''. The GAF genes were not present in this position in species of 'the rest' (Supplementary Table S5). In the upstream and downstream of a gene cluster coding for eight subunits of ATP

synthase, and four gene clusters encoding 25, 12, 4 and 3 ribosomal and related proteins, located in different places of genomes, no meaningful gene arrays were detected common to the Clade I species or species of 'the rest'.

Conclusion

Finally, we would like to propose the following conclusion. Genera of the family *Halobacteriaceae* are divided into two groups: (1) Clade I that does not possess the *pyrD* gene in the immediate upstream of any 16S rRNA genes, but probably possesses the dihydroliipoamide dehydrogenase gene, and (2) 'the rest' that possesses the *pyrD* gene in the immediate upstream of the single, or one out of multiple, 16S rRNA gene, which is the orthologous gene. The specific gene array partially found in the upstream of the *pyrD* gene of 'the rest' was also conserved in the Clade I haloarchaea.

DISCUSSION

Koonin (2005) defined orthologues and paralogues as follows: orthologues are genes derived from a single ancestral gene in the last common ancestor of the compared species, while paralogues are genes related via duplication. The genome and contig sequences of species of the family *Halobacteriaceae* recently made public have suggested that most of the database sequences of PCR amplified 16S rRNA genes are correct. They were identical in the case of *Hac. jeotgali* and *Hbt. salinarum*, and in most cases differ by one or two nucleotides. These differences should be regarded as within the range of experimental errors, owing to PCR mutation caused by low fidelity of DNA polymerase used, sequencing error, etc.

The sequence situation in the genera *Haloarcula*, *Halomicrobium* and *Halosimplex* is slightly different from other genera. It is now recognized that many sequences determined from strains of these genera are chimeras. Even the sequences determined after cloning of PCR products into plasmid vectors are different to some extent from those determined by genome sequencing. For example, EF645680 vs AB355983 from *Haloarcula argentinensis* JCM 9737^T, 9 nt differences; EF645684 and EF645685 vs Har.japonica_16SA and 16SB from *Har. japonica* JCM 7785^T, 22 and 17 nt differences, respectively; EF645687 vs the contig sequence from *Har. vallismortis* CGMCC 1.2048^T, 3 nt differences. The accuracy of the sequences (EF645680–EF645694) determined by Cui *et al.* (2009) is confirmed by the fact that some of the sequences are exactly the same as those from genome sequences (see Supplementary Fig. S1). Furthermore, fourteen sequences have been deposited for *Hsx. carlsbadense* ATCC BAA-75^T. They differ to some extent from each other and, if these sequences are correct, we have to assume the presence of at least six heterogeneous 16S rRNA genes.

These facts, altogether, strongly suggest that in the case of species possessing highly heterogeneous 16S rRNA genes,

PCR with non-specific primers that amplify all of the 16S rRNA genes fails to amplify them accurately, yielding chimeric DNA fragments. In this study we have demonstrated that only the orthologous 16S rRNA gene can be amplified accurately by PCR with the *pyrD* primer set from species of the genera *Haloarcula* and *Halomicrobium*. The reliability of the PCR was confirmed by the fact that the sequences of *Har. marismortui*, *Har. vallismortis*, *Har. japonica* and 'Har. sinaiensis' we determined in this study were exactly the same as the sequences rrs-1, contig0026, Har.japonica_16SA and contig00040, respectively.

Recently, we reported the presence of a coherent group, Clade I, consisting of species of eleven genera, in two phylogenetic trees inferred from 16S rRNA genes and from full-length sequences of the RNA polymerase subunit B' gene (Minegishi *et al.*, 2010). Besides the phylogeny based on multiple sequence alignments, another approach, gene order phylogeny, has been used recently, based on comparison of gene orders or gene contents (Belda *et al.*, 2005; Luo *et al.*, 2009). In this study we found that a gene array starting from the MC1 gene in the immediate upstream of a 16S rRNA gene is conserved throughout species of the family *Halobacteriaceae*. The presence or absence of the *pyrD* gene in this gene array was shown to divide the species into two groups.

The *pyrD* gene encodes dihydroorotate oxidase, with the systematic name (S)-dihydroorotate:oxygen oxidoreductase, catalysing the reaction, (S)-dihydroorotate + O₂ = orotate + H₂O₂, required for *de novo* synthesis of UMP. Why is the gene located in the immediate upstream of the 16S rRNA gene in species of 'the rest', while in the upstream of Na⁺/H⁺ antiporter subunit genes in species of Clade I? Genome sequences of prokaryotes reflect a long history of gene rearrangement. The fact that a methanogenic archaeon, *M. hungatei* JF-1^T, was the only other species in the domain 'Archaea' besides members of the family *Halobacteriaceae* shown to possess the *pyrD* gene in the immediate upstream of one 16S rRNA gene was interesting from the viewpoint of evolution of halophilic archaea. Methanogenic archaeal members are believed to be closely related to members of the family *Halobacteriaceae*, as revealed from concatenated gene sequences of transcription and translation machineries (Baptiste *et al.*, 2005; Brochier *et al.*, 2005). Four types of changes are known to affect the order of genes in prokaryote genomes, inversions and translocations, progressive disintegration, horizontal or lateral gene transfer, and partial duplication (Belda *et al.*, 2005). In-depth comparison of genomes of the family *Halobacteriaceae* with that of *M. hungatei* JF-1^T will shed light on the evolutionary history of methanogenic and halophilic archaea. Gene order information for another species of the genus *Methanospirillum*, *Methanospirillum lacunae* (Iino *et al.*, 2010) is highly expected.

It may be scientifically sound to hypothesize that during the evolution of members of the family *Halobacteriaceae*, a *pyrD* transposition event happened in one group, and this was followed by subsequent speciation processes in each group.

Definition of the Clade I haloarchaea will be made more inclusive by the genome sequencing in progress for the following species, *Hst. larsenii* XH-48^T, *Nbt. gregoryi* SP2, *Nln. baerhuensis* IHC-005^T and *Nrr. bangense* A33^T (information at JGI - Genome Projects). Thorough analysis of gene orders of species of Clade I and 'the rest' will elucidate the evolutionary history of the extremely halophilic Archaea. The time for proposal of a second family in the order *Halobacteriales* is not too far in the future.

REFERENCES

- Anderson, I., Ulrich, L. E., Lupa, B., Susanti, D., Porat, I., Hooper, S. D., Lykidis, A., Sieprawka-Lupa, M., Dharmarajan, L. & other authors (2009). Genomic characterization of *methanomicrobiales* reveals three classes of methanogens. *PLoS ONE* 4, e5797.
- Baliga, N. S., Bonneau, R., Facciotti, M. T., Pan, M., Glusman, G., Deutsch, E. W., Shannon, P., Chiu, Y., Weng, R. S. & other authors (2004). Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. *Genome Res* 14, 2221–2234. (Erratum: *Genome Res* 14, 2510).
- Baptiste, É., Brochier, C. & Boucher, Y. (2005). Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea* 1, 353–363.
- Belda, E., Moya, A. & Silva, F. J. (2005). Genome rearrangement distances and gene order phylogeny in γ -*Proteobacteria*. *Mol Biol Evol* 22, 1456–1467.
- Bolhuis, H., Palm, P., Wende, A., Falb, M., Rampp, M., Rodriguez-Valera, F., Pfeiffer, F. & Oesterhelt, D. (2006). The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. *BMC Genomics* 7, 169.
- Boucher, Y., Douady, C. J., Sharma, A. K., Kamekura, M. & Doolittle, W. F. (2004). Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. *J Bacteriol* 186, 3980–3990.
- Brochier, C., Forterre, P. & Gribaldo, S. (2005). An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol Biol* 5, 36.
- Chartier, F., Laine, B., Belaïche, D., Touzel, J. P. & Sautière, P. (1989). Primary structure of the chromosomal protein MC1 from the archaeobacterium *Methanosarcina* sp. CHTI 55. *Biochim Biophys Acta* 1008, 309–314.
- Cui, H. L., Zhou, P. J., Oren, A. & Liu, S. J. (2009). Intraspecific polymorphism of 16S rRNA genes in two halophilic archaeal genera, *Haloarcula* and *Halomicrobium*. *Extremophiles* 13, 31–37.
- Dennis, P. P., Ziesche, S. & Mylvaganam, S. (1998). Transcription analysis of two disparate rRNA operons in the halophilic archaeon *Haloarcula marismortui*. *J Bacteriol* 180, 4804–4813.
- Eickbush, T. H. & Eickbush, D. G. (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175, 477–485.
- Enache, M., Itoh, T., Fukushima, T., Usami, R., Dumitru, L. & Kamekura, M. (2007). Phylogenetic relationships within the family *Halobacteriaceae* inferred from *rpoB'* gene and protein sequences. *Int J Syst Evol Microbiol* 57, 2289–2295.
- Hartman, A. L., Norais, C., Badger, J. H., Delmas, S., Haldenby, S., Madupu, R., Robinson, J., Khouri, H., Ren, Q. & other authors (2010). The complete genome sequence of *Haloferax volcanii* DS2, a model archaeon. *PLoS ONE* 5, e9605.
- Iino, T., Mori, K. & Suzuki, K. (2010). *Methanospirillum lacunae* sp. nov., a methane-producing archaeon isolated from a puddly soil, and emended descriptions of the genus *Methanospirillum* and *Methanospirillum hungatei*. *Int J Syst Evol Microbiol* 60, 2563–2566.
- Jacquet, M., Kalekine, M. & Boy-Marcotte, E. (1985). Sequence analysis of a *Dictyostelium discoideum* gene coding for an active dihydroorotate dehydrogenase in yeast. *Biochimie* 67, 583–588.
- Kanehisa, M. & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28, 27–30.
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34 (Database issue), D354–D357.
- Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39, 309–338.
- Laine, B., Culard, F., Maurizot, J. C. & Sautière, P. (1991). The chromosomal protein MC1 from the archaeobacterium *Methanosarcina* sp. CHTI 55 induces DNA bending and supercoiling. *Nucleic Acids Res* 19, 3041–3045.
- Larsen, J. N. & Jensen, K. F. (1985). Nucleotide sequence of the *pyrD* gene of *Escherichia coli* and characterization of the flavoprotein dihydroorotate dehydrogenase. *Eur J Biochem* 151, 59–65.
- Liao, D. (2000). Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. *J Mol Evol* 51, 305–317.
- Luo, H., Sun, Z., Arndt, W., Shi, J., Friedman, R. & Tang, J. (2009). Gene order phylogeny and the evolution of methanogens. *PLoS ONE* 4, e6069.
- Minegishi, H., Kamekura, M., Itoh, T., Echigo, A., Usami, R. & Hashimoto, T. (2010). Further refinement of the phylogeny of the *Halobacteriaceae* based on the full-length RNA polymerase subunit B' (*rpoB'*) gene. *Int J Syst Evol Microbiol* 60, 2398–2408.
- Mylvaganam, S. & Dennis, P. P. (1992). Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaeobacterium *Haloarcula marismortui*. *Genetics* 130, 399–410.
- Nei, M. & Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet* 39, 121–152.
- Savage, K. N., Krumholz, L. R., Oren, A. & Elshahed, M. S. (2007). *Haladaptatus paucihalophilus* gen. nov., sp. nov., a halophilic archaeon isolated from a low-salt, sulfide-rich spring. *Int J Syst Evol Microbiol* 57, 19–24.
- Shimane, Y., Hatada, Y., Minegishi, H., Mizuki, T., Echigo, A., Miyazaki, M., Ohta, Y., Usami, R., Grant, W. D. & Horikoshi, K. (2010). *Natronoarchaeum mannanyticum* gen. nov., sp. nov., an aerobic, extremely halophilic archaeon isolated from commercial salt. *Int J Syst Evol Microbiol* 60, 2529–2534.
- Vreeland, R. H., Straight, S., Krammes, J., Dougherty, K., Rosenzweig, W. D. & Kamekura, M. (2002). *Halosimplex carlsbadense* gen. nov., sp. nov., a unique halophilic archaeon, with three 16S rRNA genes, that grows only in defined medium with glycerol and acetate or pyruvate. *Extremophiles* 6, 445–452.
- Walsh, D. A., Baptiste, E., Kamekura, M. & Doolittle, W. F. (2004). Evolution of the RNA polymerase B' subunit gene (*rpoB'*) in *Halobacteriales*: a complementary molecular marker to the SSU rRNA gene. *Mol Biol Evol* 21, 2340–2351.