The Methanosarcina barkeri genome: comparative analysis with Methanosarcina acetivorans and Methanosarcina mazei reveals extensive rearrangement within methanosarcinal genomes

Permalink
https://escholarship.org/uc/item/3g16p0m7

Authors
Maeder, Dennis L.
Anderson, Iain
Brettin, Thomas S.
et al.

Publication Date
2006-05-19

Peer reviewed
Title: The Methanosarcina barkeri genome: comparative analysis with Methanosarcina acetivorans and Methanosarcina mazei reveals extensive rearrangement within methanosarcinal genomes

Author(s): Dennis L. Maeder, Iain Anderson, et al
Division: Genomics

November 2006
Journal of Bacteriology
The *Methanosarcina barkeri* genome: comparative analysis with *Methanosarcina acetivorans* and *Methanosarcina mazei* reveals extensive rearrangement within methanosarcinal genomes

Dennis L. Maeder*, Iain Anderson†, Thomas S. Brettin†, David C. Bruce†, Paul Gilna†, Cliff S. Han†, Alla Lapidus†, William W. Metcalf‡, Elizabeth Saunders†, Roxanne Tapia†, and Kevin R. Sowers*.

*University of Maryland Biotechnology Institute, Center of Marine Biotechnology, Columbus Center, Suite 236, 701 E. Pratt St., Baltimore, Maryland 21202, USA
† Microbial Genomics, DOE Joint Genome Institute, 2800 Mitchell Drive, B400, Walnut Creek, CA 94598, USA
‡ University of Illinois, Department of Microbiology, B103 Chemical and Life Sciences Laboratory, 601 S. Goodwin Avenue, Urbana, Illinois 61801, USA

Running title: Comparative analysis of three methanosarcinal genomes

Keywords: *Methanosarcina barkeri*, archaeal genome, methanogenic Archaea
We report here a comparative analysis of the genome sequence of *Methanosarcina barkeri* with those of *Methanosarcina acetivorans* and *Methanosarcina mazei*. All three genomes share a conserved double origin of replication and many gene clusters. *M. barkeri* is distinguished by having an organization that is well conserved with respect to the other *Methanosarcinae* in the region proximal to the origin of replication with interspecies gene similarities as high as 95%. However it is disordered and marked by increased transposase frequency and decreased gene synteny and gene density in the proximal semi-genome. Of the 3680 open reading frames in *M. barkeri*, 678 had paralogs with better than 80% similarity to both *M. acetivorans* and *M. mazei* while 128 nonhypothetical orfs were unique (non-paralogous) amongst these species including a complete formate dehydrogenase operon, two genes required for N-acetylmuramic acid synthesis, a 14 gene gas vesicle cluster and a bacterial P450-specific ferredoxin reductase cluster not previously observed or characterized in this genus. A cryptic 36 kbp plasmid sequence was detected in *M. barkeri* that contains an *orc1* gene flanked by a presumptive origin of replication consisting of 38 tandem repeats of a 143 nt motif. Three-way comparison of these genomes reveals differing mechanisms for the accrual of changes. Elongation of the large *M. acetivorans* is the result of multiple gene-scale insertions and duplications uniformly distributed in that genome, while *M. barkeri* is characterized by localized inversions associated with the loss of gene content. In contrast, the relatively short *M. mazei* most closely approximates the ancestral organizational state.
INTRODUCTION

Biological methanogenesis by the methane-producing Archaea has a significant role in the global carbon cycle. This process is one of several anaerobic degradative processes that complement aerobic degradation by utilizing alternative electron acceptors in habitats where O₂ is not available (Sowers 2004). The efficiency of this microbial process is directly dependent upon the interaction of three metabolically distinct groups of microorganisms: the fermentative and acetogenic Bacteria and the methanogenic Archaea. The methanogenic Archaea have two pivotal roles in methanogenic consortia (Lovley and Klug 1982). By consuming hydrogen for methanogenesis and effectively lowering its partial pressure by the process of inter-species hydrogen exchange, the methanogens provide a thermodynamically favorable environment for the fermentative and acetogenic species to utilize protons as electron acceptors. This interaction enables fermentors to conserve more energy by producing a more oxidized product, acetate, which is a substrate for methanogenesis. The second role of the methanogens is the dismutation of acetate, which accounts for 70% of the global methane produced by biological methane production. The net effect of inter-species hydrogen exchange is the diversion of protons to hydrogen and carbon to acetate, which ultimately yields methane and carbon dioxide via methanogenesis.

The genus Methanosarcina includes the most metabolically diverse species of methanogens. Whereas most methanogenic species grow by obligate CO₂ reduction with H₂, methyl reduction with H₂, aceticlastic dismutation of acetate or methylotrophic
catabolism of methanol, methylated amines, and dimethylsulfide, most *Methanosarcina* spp. grow by all four catabolic pathways (Welander and Metcalf 2005).

*Methanosarcina acetivorans* was recently reported also to grow non-methanogenically with CO (Rother and Metcalf 2004). In addition to their appetency for all known methanogenic substrates most *Methanosarcina* spp. can grow in a minimal mineral medium and fix molecular nitrogen (Bomar and Knoll 1985; Lobo and Zinder 1988).

They also adapt to intracellular solute concentrations ranging from freshwater to three times that found in seawater (Sowers 1995) by osmoregulatory mechanisms that enable them to synthesize or accumulate osmoprotectants and modify their outer cell envelope (Sowers et al. 1993). This metabolic diversity is reflected in the relatively large genome sizes of *Methanosarcina acetivorans* (5.8 Mb) and *Methanosarcina mazei* (4.1 Mb) genomes and the relatively large number of number of putative coding sequences, 4,524 and 3,371 respectively, compared with other methanogenic Archaea (Deppenmeier et al. 2002; Galagan et al. 2002). The adaptive success of these species is further evident by the occurrence of multiple orthologs in the genomes including multiple catabolic methyltransferases and carbon monoxide dehydrogenases, all three known types of nitrogenases, and all four known chaperoning systems (Conway de Macario et al. 2003; Deppenmeier et al. 2002; Galagan et al. 2002).

*Methanosarcina barkeri* Fusaro was isolated from sediment from Lago del Fusaro, a freshwater coastal lagoon west of Naples, Italy (Kandler and Hippe 1977). This isolate utilizes all three catabolic pathways and exhibits a dichotomous morphology. When grown on freshwater medium this species grows as large multicellular aggregates imbedded in a heteropolysaccharide matrix (Figure 1).
composed primarily of D-galactosamine and D-glucuronic acid, termed methanochondroitin (Kreisl and Kandler 1986), whereas in marine medium these species grow as individual cells surrounded only by an S-layer (Sowers 1995). This isolate has been one of the most frequently studied methanosarcinal strains for the physiology, biochemistry and bioenergetics of methanogenesis (Sowers 2004). The development of a tractable methanosarcinal gene transfer system has led to a number of recent reports on the mechanisms of methanogenesis using genetic approaches (Rother and Metcalf 2005).

Herein we describe the genome of *M. barkeri*, which represents the third methanosarcinal genome sequenced. In addition to comparison of the genome annotation, this is the first three-way analysis of the complete genomes of closely related species in the methanogenic *Euryarchaeota*. Results reveal extensive gene rearrangements in *M. barkeri* relative to *M. acetivorans* and *M. mazei* and high degree of conservation within the fragments providing insight into the mechanisms of structural modification and the functional organization of the methanosarcinal genome.

**MATERIALS AND METHODS**

**Growth conditions**

The source for *Methanosarcina barkeri* Fusaro (=DSM 804) was described previously (Metcalf et al. 1996). *M. barkeri* was grown in F-medium (Sowers 1995) with 0.1 M trimethylamine, where described growth was tested with 0.1 M sodium formate or with a headspace of 200 kPa H$_2$-CO$_2$ (80:20) substituted for trimethylamine. Cultures
were incubated statically at 35 °C in the dark. Growth was monitored by measuring the optical density at 550 nm with a Spectronic 21 and by measuring methanogenesis by gas chromatography as described previously (Sowers et al. 1993; Sowers and Ferry 1983).

Genome sequencing, assembly and finishing

Genomic DNA was isolated from *M. barkeri* Fusaro as described previously (Boccazzi et al. 2000). The genome of *M. barkeri* was sequenced at the Joint Genome Institute (JGI) using a combination of 3 kb, 8 kb and 40 kb (fosmid) DNA libraries. All general aspects of library construction and sequencing performed at the JGI can be found at [http://www.jgi.doe.gov/](http://www.jgi.doe.gov/). Draft assemblies were based on 89216 total reads. All three libraries provided 13x coverage of the genome. The Phred/Phrap/Consed software package ([http://www.phrap.com](http://www.phrap.com)) was used for sequence assembly and quality assessment (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). After the shotgun stage, reads were assembled with parallel phrap (High Performance Software, LLC). Possible mis-assemblies were corrected with Dupfinisher (unpublished, C. Han) or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, custom primer walk or PCR amplification (Roche Applied Science, Indianapolis, IN). A total of 2389 additional reactions were necessary to close gaps and to raise the quality of the finished sequence. The completed genome sequences of *M. barkeri* contains 85812 reads, achieving an average of 12-fold sequence coverage per base with an error rate less than 1 in 100,000. The sequences of *M. barkeri*, including a chromosome and a
plasmid, can be accessed using the GenBank accession numbers CP000099 and CP000098 or from the JGI IMG site (http://img.jgi.doe.gov) as taxon ID 623520000.

Annotation and analysis

Genes were predicted with a combination of Glimmer and Critica (Badger and Olsen 1999; Delcher et al. 1999). These gene predictions were then run through a pipeline that identifies gene overlaps, missed genes, and incorrect start sites (Markowitz et al. 2006). The gene predictions were then manually curated. Functional predictions were generated automatically based on presence of hits to COGs (Tatusov et al. 2003), Pfam (Bateman et al. 2000), and Interpro (Mulder et al. 2005) families.

Whole genome alignment and analysis

Chromosome sequences (Table 1) in fasta format were used to build single sequence blast databases, which served as the subject sequences for comprehensive WuBlast (Gish, W., 1996-2004, http://blast.wustl.edu) blastn and tblastx paired comparisons both as whole sequences and as segmented comparisons using the following parameters: span2, noseqs, filter=none, hspmax =10000, gspmax=10000. Similarly all CDS sequence features were built into databases and blasted to generate a comprehensive set of pairwise comparisons. Blastn outputs were captured into a database of HSP features cross referenced to a sequence and sequence feature database. Outputs were also directly parsed by Cross (Maeder, D., 1998-2006, http://bigm.umbi.umd.edu/materials/software/Cross.pub/) for display and interactive examination of comparative features.
The paired comparison database, GRIT, runs under the database manager MySQL (MySQL AB) and consists of source, feature, fragment and link tables. The feature table was populated with predicted gene product features derived from GenBank or JGI (Table 1) with a foreign key pointing to a source table. The link table contains blastn HSP scores and identities with foreign keys pointing to entries in the fragment table, which contains positional information about HSPs with a foreign key pointing to the feature from which it was derived. This schema (Figure 2) allows the construction of a SQL query that directly and rapidly retrieves sets of features, which are either unique within a set of source chromosomes or describes a set of genes common at an arbitrary level of similarity between two or more sources. Blastn comparisons facilitate measurement of significant similarity in homologs of closely related organisms and manage non-coding sequences; tblastx or blastx comparisons are similarly applicable for comparison of less closely related sequences. Washu blastn was used with the parameters: span2, filter=none, hspmax =10000, gspmax=10000, and was wrapped in a perl script for automatic iteration through multiple pair-wise blasts. Output data were then parsed and HSPs stored in GRIT. A web interface to the queries and databases is available at (http://bigm.umbi.umd.edu/dat/genome/) and will be elaborated elsewhere.

Cumulative skew analysis was performed using skew (Maeder, 2001, http://bigm.umbi.umd.edu/materials/software/skew/) which implements the algorithm of Grigoriev (Grigoriev 1998). Repeat analysis emerged directly from unfiltered blast and was confirmed using Mummer (Delcher et al. 1999). Putative origins of replication were
explored by examining regions with locally separated inverted repeats in close upstream proximity to the \textit{orc1/cdc6} genes.

Chromosomal sequence similarity was calculated as a distance derived from blastn comparisons in the GRIT database using a perl script \texttt{cross match.pl} which generates distance matrices in mega2 format based on equation 1, where \( n \) is the length of the genome and HSP.ID is the maximal fractional identity at position \( n \) of sequence \( x \) where HSP.ID exceeds a threshold of e.g. 0.67. The mean distance \( D \) for both axes is calculated for both sequence axes by the equation:

\[
D_x = 1 - \frac{\sum_{i=1}^{n} (\text{MAX}(\text{HSP.ID}_i))}{n}
\]  

This measure of distance is comparable with hybridization techniques as it yields a fractional nucleotide similarity between organisms which considers stringency.

Synteny of any gene was measured by comparing the order of the gene’s left and right neighbors with those of their best matched paralogous genes in the comparable genome. Downstream synteny is expressed as the ratio of the ordinal distance between a gene, \( G \), and it's downstream neighbor, \( R \), (which is always 1) and the distance between a corresponding paralogous gene \( G' \) and the paralog of \( R \), \( R' \). This may be calculated as:

\[
SI = 1/\text{abs}(R' - G')
\]  

with \( 0 < SI \leq 1 \). Cumulative deviations from the mean of SI were calculated for intelligible display. Intergenic interval was calculated in the same manner.
For thin-section electron micrographs, cells were fixed with 2% glutaraldehyde and 2% osmium tetroxide, and dehydrated in a graded series of ethanol mixtures. Cells were embedded and sectioned in Epon resin, then post-stained with uranyl acetate and lead citrate as described previously (Sowers and Ferry 1983). A Joel JEM-1200 EX II transmission electron microscope at 80 kV was used to generate thin-section micrographs.

RESULTS & DISCUSSION

The genome of *Methanosarcina barkeri* was sequenced using a combination of whole genome shotgun and directed finishing as described in Methods. The genome consists of a circular chromosome of 4,837,408 base pairs (bp) and a 36,358 bp extrachromosomal element (Table 1). The *M. barkeri* genome, which is intermediate in size between *Methanosarcina acetivorans* (5.8 mb) and *Methanosarcina mazei* (4.1 mb), is the second largest genome among the Archaea. The extrachromosomal element is 6.7 times larger than the only other methanosarcinal extrachromosomal element, plasmid pC2A, from *M. acetivorans* (Metcalf et al. 1997).

*Methanosarcina barkeri* chromosome structure and content

A total of 3680 putative protein coding genes longer than 200 bp were identified (Table 1), which together cover 70% of the genome. The average protein coding region of *M. barkeri* at 921 bp is within 2% of *M. acetivorans* and *M. mazei* while its average intergenic region at 393 bp is considerably larger than those of *M. acetivorans* (328 bp).
and *M. mazei* (303 bp). A further 73 RNA features were identified including 3 sets of ribosomal RNAs (5S, 16S, 23S) and 62 tRNAs covering all amino acids and pyrrolysine that is encoded by the UAG codon in methylamine methyltransferase genes. 1780 hypothetical protein open reading frames accounted for nearly half of all protein features with 1837 putative functional proteins assignments based on similarity to identified protein sequences in public databases. Of hypothetical protein genes conserved at the 80% nucleotide level, 289 were shared with *M. acetivorans* and 249 with *M. mazei* of which 105 were common to both and should be considered highly conserved unidentified genes.

**Gene Annotation**

There were 128 unique orfs with sequence identities greater than 67\% to genes in the NCBI sequence database but without sequence identity to other methanosarcinal genomes ([http://bigm.umbi.umd.edu/materials/Methanosarcina/](http://bigm.umbi.umd.edu/materials/Methanosarcina/)). Some of these features are highlighted below.

The *M. barkeri* genome included the full complement of genes encoding enzymes in the hydrogenotrophic, methylotrophic and aceticlastic pathways (Deppenmeier et al. 2002; Galagan et al. 2002). In addition to these a complete formate dehydrogenase operon (MbarA 1561-1562), *fdhAB*, was detected with high sequence identity to catabolic formate dehydrogenase from several formate-utilizing methanogens. *Methanosarcina* spp. have never been reported to utilize formate for growth and *fdhAB* has not been detected previously in this genus (Boone et al. 2001). Attempts to grow *M. barkeri* on 50 mM formate in this study were unsuccessful and the
addition of sodium formate to cultures containing trimethylamine or hydrogen did not
enhance growth. *M. barkeri* lacks genes encoding a two subunit NDP-forming acetyl-
CoA synthetase (*acdAB*) that is found in *M. acetivorans* (MA3168 and MA3602) and *M.
mazei* (MM0358 and MM0493), but has a remnant of this enzyme, pseudogene
MbarA_3662. This enzyme catalyzes one of two pathways for generating acetyl-CoA;
the other is the CO dehydrogenase/acetyl-coenzyme A synthase that catalyzes
aceticlastic catabolism in *Methanosarcina* spp. The absence of the *acd* genes suggests
that the CO dehydrogenase/acetyl-coenzyme A synthase fulfills the function of both
enzymes in *M. barkeri*.

Among genes encoding biosynthetic functions a group of 14 sequential orfs
encode for predicted gas vesicles with highest identity to GvpANOFGJKLM (MbarA326-
339) in the haloarchaea, which includes the minimal gene set for expression of vesicle
in *Haloferax volcanii* (Offner and Pfeifer 1995). Although there are no reports of gas
vesicles in *M. barkeri* Fusaro, gas vesicles have been reported in another strain of
*M. barkeri*, FR-1, and in *Methanosarcina vacuolata*, which has 61% homology with the
type strain of *M. barkeri* (Archer and King 1984; Zhilina and Zavarzin 1979; Zhilina and
Zavarzin 1987). Interestingly, *M. barkeri* has three sequential copies of GvpA that
encodes the ribs of the vesicle wall and influences the strength and width of the vesicles
(Beard et al. 2002). The 33.5 kb region that includes the Gvp operon may have been
acquired from vesicle synthesizing species as it is flanked by transposons. Gas
vesicles are observed occasionally in *M. barkeri* cells grown with H₂-CO₂ on solidified
medium. *M. barkeri* also has orfs (MbarA_0022 and MbarA_0023) with high identity to
two enzymes required for N-acetylmuramic acid synthesis, which is unique among the
sequenced Archaea. However, prior analysis of the cell wall composition of *M. barkeri*
Fusaro failed to detect muramic acid (Kandler and Hippe 1977). *M. barkeri* also lacks a
low affinity phosphate transporter (MA2935) suggesting it originated in a phosphate-
poor environment. Two other transporters are missing in *M. barkeri*, a gluconate
transporter (MA0021) and a dicarboxylate transporter (MA2961). This suggests that *M.
barkeri* may have a lower ability to take up organic compounds than the other two.
Finally *M. acetivorans* and *M. mazei* have two copies of cheC (MA0012 and MA3065)
but *M. barkeri* does not have this protein. The role of this chemotaxis gene in
*Methanosarcina* spp. is currently unknown since motility has not been observed in these
species.

Another unique feature of the *M. barkeri* genome is the detection of a putative
operon encoding a bacterial P450-specific ferredoxin reductase (Mbar 1947-1945). The
family of heme protein monooxygenases known as cytochrome P450 plays a critical
role in the synthesis and degradation of many xenobiotics and physiologically important
compounds (Sono et al. 1966; Sono et al. 1996; Whitlock and Denison 1995). All
known P450s are multi-centre enzymes consisting of a heme, or P450, component with
associated reductase components. The gene encoding the putative cytochrome P450 in
*M. barkeri* is flanked immediately upstream by genes encoding a ferredoxin and
ferredoxin reductase, which is typical of bacterial class I three-component systems. For
catalytic activity, cytochrome P450 must be associated with the electron donor partner
proteins, ferredoxin/ferredoxin reductase complex (Takemori et al. 1993). Cytochrome
P450 has not been detected previously in the Archaea. Another putative operon
encoding oxygen dependent cytochrome *d* oxidase, *cydAB*, was also identified in the
genome of *M. barkeri* and the other two methanosarcinal genomes. The presence of these oxygen dependent genes along with one catalase and two superoxide dismutase suggests that these proteins protect methanosarcinal species from oxygen or they may support microaerophilic growth by a currently undescribed mechanism. As cytochrome P450 catalyzes an oxygen requiring reaction and has not been detected previously in an anaerobe, the detection of this gene in *M. barkeri* raises intriguing questions about the function of this gene product in this obligately anaerobic methanogen.

**Plasmid structure and content**

The 36.4 kb plasmid in *M. barkeri* has not been detected previously. In contrast to the smaller 5.4 kb plasmid pC2A in *M. acetivorans*, which appears to replicate by a rolling-circle mechanism (Metcalf et al. 1997), the *M. barkeri* extrachromosomal element lacks a putative *repA*. Instead it has a cdc-6 homolog in a region of highly repetitive sequence (discussed below), which suggests a novel mechanism of synchronous replication. Interestingly, one of the extrachromosomal orfs (MbarB 3749) has 44% sequence similarity to an ATPase associated with chromosomal partitioning. These combined characteristics suggest that the extrachromosomal element replicates with cell division. In addition to the putative cdc6 and partitioning protein, the orfs include 4 genes possibly associated with methanochondroitin synthesis, 7 hypothetical genes of unknown function and 5 putative transposases. None of the orfs had equivalent identities to orfs found in *M. acetivorans* and *M. mazei* genomes, but missing from the *M. barkeri* genome that might have suggested a critical function for the extrachromosomal element.
Features revealed by whole genome comparison

Whole genome distances (Table 2) based on maximal local alignments indicate that the genomes are quite similar in overall content with *M. acetivorans* and *M. mazei* marginally more closely related. This is in qualitative agreement with DNA-DNA hybridization experiments (Sowers and Johnson 1984), which showed 28% homology between *M. acetivorans* and *M. mazei* and 18% between these species and *M. barkeri*. This result underscores the comparability of these sequences with the exception of the plasmid sequence.

Location of origins of replication

In Archaea, origins of replication are invariably found in close proximity to the origin recognition complex gene (*orc1*) sometimes also referred to as cell division control protein 6 (*cdc6*) (Lopez et al. 1999). When genes are densely packed, searches for putative origins of replication are directed at proximal intergenic regions. In the three Methanosarcinae there are two highly conserved paralogous copies of these genes, in relatively close mutual proximity (about 100 kb or 300 kb in *M. barkeri*) situated on opposite strands and directed away from each other, a finding consistent with the observation of Kelman *et al.* (Kelmana and Kelman 2004). Flanking downstream ORFs are conserved (Table 3). The putative origins of replication are located in the upstream intergenic regions of approximately 1600 nt (ORI A) or 800 nt (ORI B) in extent and are somewhat conserved at the nucleotide level. In the chromosomal origin of replication region, gene products are approximately 95% identical across all species. Non-coding origin features (ORI A and ORI B) are not as well conserved (E <= 1e-44 in ORI A and
E <= 1e-8 in ORI B) and show only weak similarity between ORI A and ORI B. They are extremely AT rich (~70%) and may show unconserved inverted repeat structures.

A replication complex is initiated when orc1 (cdc6) protein binds to cognate DNA at the origin and allows the recruitment of MCM and the rest of the replication machinery. An approximate inverted repeat (Figure 3) could allow a pseudo-symmetrical double hairpin to form a crucifix motif similar to a Holliday junction thereby initiating bi-directional replication from a point, with complexes bypassing each other to replicate the origin at the beginning of replication. The concurrent presence of more than one active origin would cause contention for DNA, so there must be an implicit mechanism to control which origin and which origin recognition complex protein is dominant. It is notable that the downstream neighbor of the secondary orc1 B is a highly conserved hsp60 class heatshock protein as this suggests a possible stress-associated switching mechanism. The putative origins of replication are located centrally within the most highly conserved and syntenous regions of the respective genomes (Figure 4) consistent with the observation of Eisen et al. (Eisen et al. 2000) of symmetrical inversion about the origin of replication. GC skew analysis (results not shown) is not useful in this case as there is a high level of strand inversion and rearrangement.

The plasmid of *M. barkeri* presents a unique and distinct origin of replication characterized by an orc1 homolog (orc1 C) that is relatively weakly related to orc1 A and orc1 B, (37% / 66% with orc1 A, 21% / 44% with orc1 B) (Table 3). The immediately adjacent upstream region of the plasmid DNA contains a 5.6 kb non-coding
region (15.3% of the plasmid sequence) characterized by highly repetitive sequence
consisting of over 38 direct repeats of a 143 nt sequence with few variations between
them (Figure S-1, supplement). The consensus of the AT rich repeat sequence is:

\[
\text{ATCCCATTTCCCTAAGCAGAGAATTAGTTTCCTAAGCAGAAAAAAGATTTCTGgcttagA}
\]
\[
\text{GAGAATTAGTTTCCTAAGCAAAACGATATCAGAAGACATAACAAGTTAGAAGAaAAAtAaTT}
\]
\[
\text{AAAATTAGATATTAATCTGTATATAT, with internal repeats underscored and variable}
\]
\[
\text{regions in lowercase. This sort of arrangement (ORI C) is quite unlike that of the}
\]
\[
\text{chromosomal ORI A and ORI B, but has the capacity to present slideable bubbled-out}
\]
\[
\text{complete repeat motifs and retain a quasi-stable structure (Figure 5).}
\]

**Overall genomic organization**

A significant observation in the three-way comparison of the Methanosarcina

genomes (Figure 4) is the overall collinearity of *M. mazei* and *M. acetivorans* (lower left

panel). This attests to a history of conserved gene order and resistance to large scale

mosaicty. However closer examination (Figure 6) reveals that there is considerable

deviation from the expected 45° slope for a line of identity. This is maintained between

*M. barkeri* and *M. mazei*, indicating that *M. acetivorans* has been subject to uniformly

distributed local elongation, which may arise from gene duplication, elongation of

intergenic regions or insertion of sequence by transposition. This may explain the large

size of the *M. acetivorans* genome relative to the other Methanosarcina spp.

*M. barkeri* is distinguished by having an organization that is well conserved with

respect to the other Methanosarcinae in the region proximal to the origin of replication

where interspecies gene similarities are as high as 95% (Table 3). However there is
little apparent conservation of gene organization in the region most distal to the origin where large scale collinearity appears rare. The putative terminus of replication is observed to be a hotspot for reorganization (Myllykallio et al. 2000). Two properties of M. barkeri were measured: synteny, which measures the local paralog neighborhood with respect to comparable genomes and intergenic interval or the separation between successive genes which measures the relative content density. In the distal semi-genome the rate of change of synteny is negative in accord with the macroscopic observation of decreased collinearity, and the negatively correlated intergenic interval is greater than average indicating a loss of gene content in this region (Table 2). What might cause this wasteland effect? One possibility, given the symmetry with respect to the origin, is an accumulation of strand exchange failures in the replication process, and subsequent 'gene rot' of broken genes. The cross effect of random strand inversion noted by Eisen et al. (Eisen et al. 2000) gives way to a shotgun effect. Another possibility is infiltration by transposons with transposase mediated damage. Certainly there is an increased frequency of transposon genes in this area (Figure 4, trace d), but this may either be causative or opportunistic, with the organism tolerating infiltration of already dysfunctional sections of the chromosome.

CONCLUSIONS

Of the 3680 open reading frames in M. barkeri, 678 had paralogs with better than 80% similarity to both M. acetivorans and M. mazei while 256 were unique (non-paralogous) amongst these species. An etiology for genome rearrangement is revealed by whole genome comparison of three species of the genus Methanosarcina. The
inverse correlation of intergenic size and synteny demonstrates a mechanism for the
development of genome plasticity, which involves replication associated inversion with
concomitant gene damage and colonization by transposon elements. Gene
duplication is also observed as a mechanism for genome extension. The organization
of *M. barkeri* is well conserved with respect to the other *Methanosarcinae* in the region
proximal to the origin of replication with interspecies gene similarities as high as 95%. In
the half genome most distant from the origin, it is however disordered and marked by
increased transposase frequency and decreased gene synteny and gene density.
Furthermore we have observed a highly conserved double origin of replication which
suggests a mechanism for replication which allows a double start with pass through
which enables the origin itself to be replicated. The apparent genome plasticity likely
contributed to these species ability to adapt to a broad range of environments as a
result of genome elongation and enrichment for favorable phenotypes.
ACKNOWLEDGEMENTS

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF0098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36. KRS was supported in part by NSF MCB Division of Cellular and Bioscience grant #MCB0110762 and by DOE Energy Biosciences Program grant #DE-FG02-93-ER20106. WWM was supported in part by NSF MCB Division of Cellular and Biosciences grant #MCB12466 and by DOE Energy Biosciences Program grant #DEFG02-02ER15296.
LITERATURE CITED


Table 1. Comparison of Genome Features Among *Methanosarcina* spp.

<table>
<thead>
<tr>
<th>Organism</th>
<th><em>Methanosarcina acetivorans</em></th>
<th><em>Methanosarcina mazei</em></th>
<th><em>Methanosarcina barkeri fusaro</em></th>
<th><em>Methanosarcina barkeri fusaro</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>tax_id</td>
<td>188937</td>
<td>192952</td>
<td>269797</td>
<td>269797</td>
</tr>
<tr>
<td>Accession designation</td>
<td>NC_003552</td>
<td>NC_003901</td>
<td>NC_007355</td>
<td>NC_007349</td>
</tr>
<tr>
<td>length</td>
<td>5751494</td>
<td>4096345</td>
<td>4837408</td>
<td>36358</td>
</tr>
<tr>
<td>G+C%</td>
<td>42.7%</td>
<td>41.5%</td>
<td>39.2%</td>
<td>33.6%</td>
</tr>
<tr>
<td>feature count</td>
<td>4540</td>
<td>3371</td>
<td>3680</td>
<td>18</td>
</tr>
<tr>
<td>features length</td>
<td>4262934</td>
<td>3074712</td>
<td>3390164</td>
<td>21099</td>
</tr>
<tr>
<td>features coverage</td>
<td>74%</td>
<td>75%</td>
<td>70%</td>
<td>58%</td>
</tr>
<tr>
<td>featureless</td>
<td>26%</td>
<td>25%</td>
<td>30%</td>
<td>42%</td>
</tr>
</tbody>
</table>
| nt/feature mean     | 939                          | 912                    | 921                              | 1172                            | 922

Maeder et al. May 18, 2006
Table 2. Intergenomic distances calculated using eq 1 based on whole genome maximal local nucleotide sequence identity considering only HSPs with identity > 67% (lower left) or 55% (upper right).

<table>
<thead>
<tr>
<th>M. acetivorans</th>
<th>M. mazei</th>
<th>M. barkeri</th>
<th>M. barkeri plasmid</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. acetivorans</td>
<td>0</td>
<td>0.489</td>
<td>0.487</td>
</tr>
<tr>
<td>M. mazei</td>
<td>0.517</td>
<td>0</td>
<td>0.480</td>
</tr>
<tr>
<td>M. barkeri</td>
<td>0.552</td>
<td>0.569</td>
<td>0</td>
</tr>
<tr>
<td>M. barkeri plasmid</td>
<td>0.881</td>
<td>0.883</td>
<td>0.836</td>
</tr>
</tbody>
</table>
Table 3. Identification of conserved features for chromosomal origins of replication in *Methanosarcina* spp.

<table>
<thead>
<tr>
<th>Description</th>
<th>strand</th>
<th><em>M. acetivorans</em></th>
<th><em>M. barkeri</em></th>
<th><em>M. mazei</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>conserved</td>
<td>+</td>
<td>GI:20093437</td>
<td>GI:73668510</td>
<td>GI:21227413</td>
</tr>
<tr>
<td>hypothetical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conserved</td>
<td>+</td>
<td>GI:20093438</td>
<td>GI:73668511</td>
<td>GI:21227414</td>
</tr>
<tr>
<td>hypothetical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conserved</td>
<td>+</td>
<td>GI:20093439</td>
<td>GI:73668512</td>
<td>GI:21227415</td>
</tr>
<tr>
<td>hypothetical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>orc1 A</td>
<td>-</td>
<td>GI:20088900</td>
<td>GI:73668513</td>
<td>GI:21227416</td>
</tr>
<tr>
<td>ORI A</td>
<td></td>
<td>1529 - 3357</td>
<td>1189197 - 1191361</td>
<td>1564667 - 1566241</td>
</tr>
<tr>
<td>inter-origin</td>
<td>variable</td>
<td>~100Kbp</td>
<td>~300Kbp</td>
<td>~100Kbp</td>
</tr>
<tr>
<td>region</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conserved</td>
<td>-</td>
<td>GI:20088912</td>
<td>GI:73668731</td>
<td>GI:21227479</td>
</tr>
<tr>
<td>hypothetical</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORI B</td>
<td></td>
<td>104571 - 105272</td>
<td>1481179 - 1482170</td>
<td>1654267 - 1655075</td>
</tr>
<tr>
<td>orc1 B</td>
<td>+</td>
<td>GI:20088912</td>
<td>GI:73668732</td>
<td>GI:21227480</td>
</tr>
<tr>
<td>Hsp60</td>
<td>-</td>
<td>GI:20088912</td>
<td>GI:73668733</td>
<td>GI:21227481</td>
</tr>
</tbody>
</table>
FIGURE LEGENDS

Figure 1. Thin-section electron micrograph of *M. barkeri* Fusaro showing typical morphology consisting of multicellular aggregates embedded in a methanochondroitin matrix. Vacuole-like structures appear to be membrane bound. Bars: A, 1.0 μm, B, 0.2 μm.

Figure 2. GRIT database schema. Primary keys are capitalized, foreign keys are underlined. Arrows indicate foreign key primary key relationships.

Figure 3. The *M. barkeri* ORI A self-complementary region blastn alignment.

Figure 4. Asymmetric fragmentation in *M. barkeri*. The top panel shows cumulative deviations from the mean in *M. barkeri* genome for synteny (SI) with respect to *M. acetivorans* (a), *M. mazei* (b), or intergenic interval (c). The cumulative transposon count is superimposed (d). The bottom panel shows uniformly scaled blastn cross plots of *M. barkeri* chromosome with *M. mazei* and *M. acetivorans* with the origin regions circled.

Figure 5. Proposed mechanism for conserved repetitive sequence to provide bubbled-out repeat motifs for initiation of replication. Pairs of quasi-stable bubbles might occur in pairs at arbitrary locations on opposite strands. All motifs are essentially identical.

Figure 6. *M. acetivorans* is elongated due to distributed gene duplication events.
**Figure 1.** Thin-section electron micrograph of *M. barkeri* Fusaro showing typical morphology consisting of multicellular aggregates embedded in a methanochondroitin matrix. Vacuole-like structures appear to be membrane bound. Bars: A, 1.0 μm, B, 0.2 μm.
Figure 2. GRIT database schema. Primary keys are capitalized, foreign keys are underlined. Arrows indicate foreign key primary key relationships.
Figure 3. The *M. barkeri* ORI A self-complementary region blastn alignment

**Query:**
```
709 CAGAAAATGAAA-ATTTCTCAGAAC-A-TGTAATTTAGATTCT-CAATTT-TTT-GAAA 656
```

**Sbjct:**
```
1190360 CAAAAAAT-TATGAGTTTC-CAGAACCNAATTGTTAAATTTACCCCTTTCTTTCCGAT 1190471
```

**Query:**
```
655 ATCATACTTTTCTGA-T-AGATGAGTATCA-ATAAAAACTCAAAATAAAAATTTCA 599
```

**Sbjct:**
```
1190418 GGAATAAGTTATGAAAATTATAAT-ATTTTTACATAAAATTAAAAATAAAAATTT-AAA 1190475
```

**Query:**
```
598 TGAAATATTA-AATCA 583
```

**Sbjct:**
```
1190476 -GATAAAATAGGTAAT 1190491
```

--- Unaligned region of ~440 nt ---

**Query:**
```
141 TAATTCATATTTATC-TTAAA-ATTTTTATTTTTTTATTTATATTTAAAAAT-ATTATAA 85
```

**Sbjct:**
```
1190933 TGATT-TAATATTTTCATEGAATATTTTTTTATTTGATTGGTTTTAT-TGATACTCAATCTA 1190989
```

**Query:**
```
84 TTTCTATAACTTTCATCCATTCCCAAAGAAAAGGTTTTAAATTAAACCTACATTTGAATGTCG 26
```

**Sbjct:**
```
1190990 TCAGAAAAAGT-ATGATTTTTAC-AAA-AAATGGGAAATCTAAATTACAT-G-TTCTGG 1191043
```

**Query:**
```
25 AACTCATA-ATTTTTTG 10
```

**Sbjct:**
```
1191044 AAATT-TACATTTTCTG 1191059
```
Figure 4. Asymmetric fragmentation in *M. barkeri*. The top panel shows cumulative deviations from the mean in *M. barkeri* genome for synteny (SI) with respect to *M. acetivorans* (a), *M. mazei* (b), or intergenic interval (c). The cumulative transposon count is superimposed (d). The bottom panel shows uniformly scaled blastn cross plots of *M. barkeri* chromosome with *M. mazei* and *M. acetivorans* with the origin regions circled.
Figure 5. Proposed mechanism for conserved repetitive sequence to provide bubbled-out repeat motifs for initiation of replication. Pairs of quasi-stable bubbles might occur in pairs at arbitrary locations on opposite strands. All motifs are nearly identical.
Figure 6. *M. acetivorans* is elongated due to distributed gene duplication events.
SUPPLEMENT

Figure S-1. Plasmid DNA origin of replication containing a 5.6 kb non-coding region characterized by highly repetitive sequence consisting of over 38 direct repeats of a 143nt sequence with few variations between them.