Lawrence Berkeley National Laboratory
Lawrence Berkeley National Laboratory

Title

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

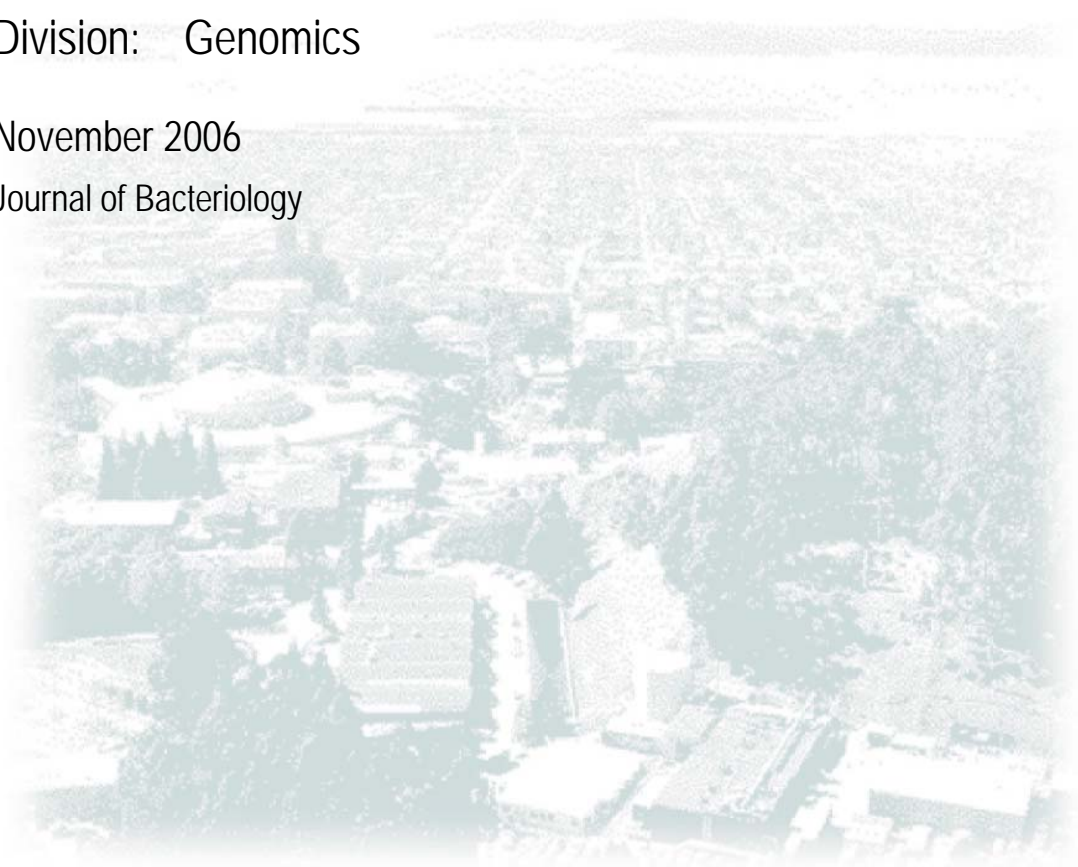ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

BERKELEY LAB

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

1

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

6

7

8

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*.

12

* 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

19

Running title: Comparative analysis of three methanosarcinal genomes

21

23 **ABSTRACT**

24

25 We report here a comparative analysis of the genome sequence of

26 *Methanosarcina barkeri* with those *of Methanosarcina acetivorans* and

27 *Methanosarcina mazei*. All three genomes share a conserved double origin of

28 replication and many gene clusters. *M. barkeri* is distinguished by having an

29 organization that is well conserved with respect to the other *Methanosarcinae* in the

30 region proximal to the origin of replication with interspecies gene similarities as high as

31 95%. However it is disordered and marked by increased transposase frequency and

32 decreased gene synteny and gene density in the proximal semi-genome. Of the 3680

33 open reading frames in *M. barkeri*, 678 had paralogs with better than 80% similarity to

34 both *M. acetivorans* and *M. mazei* while 128 nonhypothetical orfs were unique (non-

35 paralogous) amongst these species including a complete formate dehydrogenase

36 operon, two genes required for N-acetylmuramic acid synthesis, a 14 gene gas vesicle

37 cluster and a bacterial P450-specific ferredoxin reductase cluster not previously

38 observed or characterized in this genus. A cryptic 36 kbp plasmid sequence was

39 detected in *M. barkeri* that contains an *orc*1 gene flanked by a presumptive origin of

40 replication consisting of 38 tandem repeats of a 143 nt motif. Three-way comparison of

41 these genomes reveals differing mechanisms for the accrual of changes. Elongation of

42 the large *M. acetivorans* is the result of multiple gene-scale insertions and duplications

43 uniformly distributed in that genome, while *M. barkeri* is characterized by localized

44 inversions associated with the loss of gene content. In contrast, the relatively short *M.*

45 *mazei* most closely approximates the ancestral organizational state.

46 **INTRODUCTION**

47

48   Biological methanogenesis by the methane-producing Archaea has a significant

49   role in the global carbon cycle. This process is one of several anaerobic degradative

50   processes that complement aerobic degradation by utilizing alternative electron

51   acceptors in habitats where O_2 is not available (Sowers 2004). The efficiency of this

52   microbial process is directly dependent upon the interaction of three metabolically

53   distinct groups of microorganisms: the fermentative and acetogenic Bacteria and the

54   methanogenic Archaea. The methanogenic Archaea have two pivotal roles in

55   methanogenic consortia (Lovley and Klug 1982). By consuming hydrogen for

56   methanogenesis and effectively lowering its partial pressure by the process of inter-

57   species hydrogen exchange, the methanogens provide a thermodynamically favorable

58   environment for the fermentative and acetogenic species to utilize protons as electron

59   acceptors. This interaction enables fermentors to conserve more energy by producing a

60   more oxidized product, acetate, which is a substrate for methanogenesis. The second

61   role of the methanogens is the dismutation of acetate, which accounts for 70% of the

62   global methane produced by biological methane production. The net effect of inter-

63   species hydrogen exchange is the diversion of protons to hydrogen and carbon to

64   acetate, which ultimately yields methane and carbon dioxide via methanogenesis.

65   The genus *Methanosarcina* includes the most metabolically diverse species of

66   methanogens. Whereas most methanogenic species grow by obligate CO_2 reduction

67   with H_2, methyl reduction with H_2, aceticlastic dismutation of acetate or methylotrophic

68   catabolism of methanol, methylated amines, and dimethylsulfide, most *Methanosarcina*

69   spp. grow by all four catabolic pathways (Welander and Metcalf 2005).

70   *Methanosarcina acetivorans* was recently reported also to grow non-methanogenically

71   with CO (Rother and Metcalf 2004). In addition to their appetency for all known

72   methanogenic substrates most *Methanonsarcina* spp. can grow in a minimal mineral

73   medium and fix molecular nitrogen (Bomar and Knoll 1985; Lobo and Zinder 1988).

74   They also adapt to intracellular solute concentrations ranging from freshwater to three

75   times that found in seawater (Sowers 1995) by osmoregulatory mechanisms that enable

76   them to synthesize or accumulate osmoprotectants and modify their outer cell envelope

77   (Sowers et al. 1993). This metabolic diversity is reflected in the relatively large genome

78   sizes of *Methanosarcina acetivorans* (5.8 Mb) and *Methanosarcina mazei* (4.1 Mb)

79   genomes and the relatively large number of number of putative coding sequences,

80   4,524 and 3,371 respectively, compared with other methanogenic Archaea

81   (Deppenmeier et al. 2002; Galagan et al. 2002). The adaptive success of these species

82   is further evident by the occurrence of multiple orthologs in the genomes including

83   multiple catabolic methyltransferases and carbon monoxide dehydrogenases, all three

84   known types of nitrogenases, and all four known chaperoning systems (Conway de

85   Macario et al. 2003; Deppenmeier et al. 2002; Galagan et al. 2002)*.*

86   *Methanosarcina barkeri* Fusaro was isolated from sediment from Lago del

87   Fusaro, a freshwater coastal lagoon west of Naples, Italy (Kandler and Hippe 1977).

88   This isolate utilizes all three catabolic pathways and exhibits a dichotomous

89   morphology. When grown on freshwater medium this species grows as large

90   multicelluar aggregates imbedded in a heteropolysaccharide matrix (Figure 1)

91 composed primarily of D-galactosamine and D-glucuronic acid, termed

92 methanochondroitin (Kreisl and Kandler 1986), whereas in marine medium these

93 species grow as individual cells surrounded only by an S-layer (Sowers 1995). This

94 isolate has been one of the most frequently studied methanosarcinal strains for the

95 physiology, biochemistry and bioenergetics of methanogenesis (Sowers 2004). The

96 development of a tractable methanosarcinal gene transfer system has led to a number

97 of recent reports on the mechanisms of methanogenesis using genetic approaches

98 (Rother and Metcalf 2005).

99   Herein we describe the genome of *M. barkeri,* which represents the third

100 methanosarcinal genome sequenced. In addition to comparison of the genome

101 annotation, this is the first three-way analysis of the complete genomes of closely

102 related species in the methanogenic *Euryarchaeota*. Results reveal extensive gene

103 rearrangements in *M. barkeri* relative to *M. acetivorans* and *M. mazei* and high degree

104 of conservation within the fragments providing insight into the mechanisms of structural

105 modification and the functional organization of the methanosarcinal genome.

106 **MATERIALS AND METHODS**

107

108 Growth conditions

109   The source for *Methanosarcina barkeri* Fusaro (=DSM 804) was described

110 previously (Metcalf et al. 1996). *M. barkeri* was grown in F-medium (Sowers 1995) with

111 0.1 M trimethylamine, Where described growth was tested with 0.1 M sodium formate or

112 with a headspace of 200 kPa H_2-CO_2 (80:20) substituted for trimethylamine. Cultures

113   were incubated statically at 35 ^{o}C in the dark. Growth was monitored by measuring the

114   optical density at 550 nm with a Spectronic 21 and by measuring methanogenesis by

115   gas chromatography as described previously (Sowers et al. 1993; Sowers and Ferry

116   1983).

117   Genome sequencing, assembly and finishing

118   Genomic DNA was isolated from *M. barkeri* Fusaro as described previously

119   (Boccazzi et al. 2000). The genome of *M. barkeri* was sequenced at the Joint Genome

120   Institute (JGI) using a combination of 3 kb, 8 kb and 40 kb (fosmid) DNA libraries. All

121   general aspects of library construction and sequencing performed at the JGI can be

122   found at http://www.jgi.doe.gov/. Draft assemblies were based on 89216 total reads.

123   All three libraries provided 13x coverage of the genome. The Phred/Phrap/Consed

124   software package (http://www.phrap.com) was used for sequence assembly and quality

125   assessment (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). After the

126   shotgun stage, reads were assembled with parallel phrap (High Performance Software,

127   LLC). Possible mis-assemblies were corrected with Dupfinisher (unpublished, C. Han)

128   or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI).

129   Gaps between contigs were closed by editing in Consed, custom primer walk or PCR

130   amplification (Roche Applied Science, Indianapolis, IN). A total of 2389 additional

131   reactions were necessary to close gaps and to raise the quality of the finished

132   sequence. The completed genome sequences of *M. barkeri* contains 85812 reads,

133   achieving an average of 12-fold sequence coverage per base with an error rate less

134   than 1 in 100,000. The sequences of *M. barkeri*, including a chromosome and a

135   plasmid, can be accessed using the GenBank accession numbers CP000099 and

136   CP000098 or from the JGI IMG site (http://img.jgi.doe.gov) as taxon ID 623520000.

137   Annotation and analysis

138   Genes were predicted with a combination of Glimmer and Critica (Badger and

139   Olsen 1999; Delcher et al. 1999). These gene predictions were then run through a

140   pipeline that identifies gene overlaps, missed genes, and incorrect start sites (Markowitz

141   et al. 2006). The gene predictions were then manually curated. Functional predictions

142   were generated automatically based on presence of hits to COGs (Tatusov et al. 2003),

143   Pfam (Bateman et al. 2000), and Interpro (Mulder et al. 2005) families.

144   Whole genome alignment and analysis

145   Chromosome sequences (Table 1) in fasta format were used to build single

146   sequence blast databases, which served as the subject sequences for comprehensive

147   WuBlast (Gish, W., 1996-2004, http://blast.wustl.edu) blastn and tblastx paired

148   comparisons both as whole sequences and as segmented comparisons using the

149   following parameters: span2, noseqs, filter=none, hspmax =10000, gspmax=10000.

150   Similarly all CDS sequence features were built into databases and blasted to generate a

151   comprehensive set of pairwise comparisons. Blastn outputs were captured into a

152   database of HSP features cross referenced to a sequence and sequence feature

153   database. Outputs were also directly parsed by Cross (Maeder, D., 1998-2006,

154   http://bigm.umbi.umd.edu/materials/software/Cross.pub/) for display and interactive

155   examination of comparative features.

156   The paired comparison database, GRIT, runs under the database manager

157   MySQL (MySQL AB) and consists of source, feature, fragment and link tables. The

158   feature table was populated with predicted gene product features derived from

159   GenBank or JGI (Table 1) with a foreign key pointing to a source table. The link table

160   contains blastn HSP scores and identities with foreign keys pointing to entries in the

161   fragment table, which contains positional information about HSPs with a foreign key

162   pointing to the feature from which it was derived. This schema (Figure 2) allows the

163   construction of a SQL query that directly and rapidly retrieves sets of features, which

164   are either unique within a set of source chromosomes or describes a set of genes

165   common at an arbitrary level of similarity between two or more sources. Blastn

166   comparisons facilitate measurement of significant similarity in homologs of closely

167   related organisms and manage non-coding sequences; tblastx or blastx comparisons

168   are similarly applicable for comparison of less closely related sequences. Washu blastn

169   was used with the parameters: span2, filter=none, hspmax =10000, gspmax=10000,

170   and was wrapped in a perl script for automatic iteration through multiple pair-wise

171   blasts. Output data were then parsed and HSPs stored in GRIT. A web interface to the

172   queries and databases is available at (http://bigm.umbi.umd.edu/dat/genome/) and will

173   be elaborated elsewhere.

174   Cumulative skew analysis was performed using skew (Maeder, 2001,

175   http://bigm.umbi.umd.edu/materials/software/skew/) which implements the algorithm of

176   Grigoriev (Grigoriev 1998). Repeat analysis emerged directly from unfiltered blast and

177   was confirmed using Mummer (Delcher et al. 1999). Putative origins of replication were

178 explored by examining regions with locally separated inverted repeats in close upstream

179 proximity to the *orc1/cdc6* genes.

180 Chromosomal sequence similarity was calculated as a distance derived from

181 blastn comparisons in the GRIT database using a perl script cross match.pl which

182 generates distance matrices in mega2 format based on equation 1, where *n* is the

183 length of the genome and HSP.ID is the maximal fractional identity at position *n* of

184 sequence *x* where HSP.ID exceeds a threshold of e.g. 0.67. The mean distance *D* for

185 both axes is calculated for both sequence axes by the equation:

186 $$D_x = 1 - \sum_{1}^{n}(MAX(HSP.ID_n))/n \qquad\qquad (1)$$

187 This measure of distance is comparable with hybridization techniques as it yields a

188 fractional nucleotide similarity between organisms which considers stringency.

189 Synteny of any gene was measured by comparing the order of the gene's left

190 and right neighbors with those of their best matched paralogous genes in the

191 comparable genome. Downstream synteny is expressed as the ratio of the ordinal

192 distance between a gene, G, and it's downstream neighbor, R, (which is always 1) and

193 the distance between a corresponding paralogous gene G' and the paralog of R, R'.

194 This may be calculated as:

195 $$SI = 1/abs(R'-G') \qquad\qquad (2)$$

196 with $0 < SI <= 1$. Cumulative deviations from the mean of SI were calculated for

197 intelligible display. Intergenic interval was calculated in the same manner.

198   Microscopy

199   For thin-section electron micrographs, cells were fixed with 2% glutaraldehyde

200   and 2% osmium tetroxide, and dehydrated in a graded series of ethanol mixtures. Cells

201   were embedded and sectioned in Epon resin, then post-stained with uranyl acetate and

202   lead citrate as described previously (Sowers and Ferry 1983). A Joel JEM-1200 EX II

203   transmission electron microscope at 80 kV was used to generate thin-section

204   micrographs.

RESULTS & DISCUSSION
205
206
207
208   The genome of *Methanosarcina barkeri* was sequenced using a combination of

209   whole genome shotgun and directed finishing as described in Methods. The genome

210   consists of a circular chromosome of 4,837,408 base pairs (bp) and a 36,358 bp

211   extrachromosomal element (Table 1). The *M. barkeri* genome, which is intermediate in

212   size between *Methanosarcina acetivorans* (5.8 mb) and *Methanosarcina mazei* (4.1

213   mb), is the second largest genome among the Archaea. The extrachromosomal

214   element is 6.7 times larger than the only other methanosarcinal extrachromosomal

215   element, plasmid pC2A, from *M. acetivorans* (Metcalf et al. 1997).

Methanosarcina barkeri chromosome structure and content
216
217
218   A total of 3680 putative protein coding genes longer than 200 bp were identified

219   (Table 1), which together cover 70% of the genome. The average protein coding region

220   of *M. barkeri* at 921 bp is within 2% of *M. acetivorans* and *M. mazei* while its average

221   intergenic region at 393 bp is considerably larger than those of *M. acetivorans* (328 bp)

222   and *M. mazei* (303 bp). A further 73 RNA features were identified including 3 sets of

223   ribosomal RNAs (5S, 16S, 23S) and 62 tRNAs covering all amino acids and pyrrolysine

224   that is encoded by the UAG codon in methylamine methyltransferase genes. 1780

225   hypothetical protein open reading frames accounted for nearly half of all protein features

226   with 1837 putative functional proteins assignments based on similarity to identified

227   protein sequences in public databases. Of hypothetical protein genes conserved at the

228   80% nucleotide level, 289 were shared with *M. acetivorans* and 249 with *M. mazei* of

229   which 105 were common to both and should be considered highly conserved

230   unidentified genes.

231   **Gene Annotation**

232
233   There were 128 unique orfs with sequence identities greater than 67% to genes

234   in the NCBI sequence database but without sequence identity to other methanosarcinal

235   genomes (http://bigm.umbi.umd.edu/materials/Methanosarcina/). Some of these

236   features are highlighted below.

237   The *M. barkeri* genome included the full complement of genes encoding

238   enzymes in the hydrogenotrophic, methylotrophic and aceticlastic pathways

239   (Deppenmeier et al. 2002; Galagan et al. 2002). In addition to these a complete

240   formate dehydrogenase operon (MbarA 1561-1562), *fdhAB*, was detected with high

241   sequence identity to catabolic formate dehydrogenase from several formate-utilizing

242   methanogens. *Methanosarcina* spp. have never been reported to utilize formate for

243   growth and *fdhAB* has not been detected previously in this genus (Boone et al. 2001).

244   Attempts to grow *M. barkeri* on 50 mM formate in this study were unsuccessful and the

245   addition of sodium formate to cultures containing trimethylamine or hydrogen did not

246   enhance growth. *M. barkeri* lacks genes encoding a two subunit NDP-forming acetyl-

247   CoA synthetase (*acd*AB) that is found in *M. acetivorans* (MA3168 and MA3602) and *M.*

248   *mazei* (MM0358 and MM0493), but has a remnant of this enzyme, pseudogene

249   MbarA_3662. This enzyme catalyzes one of two pathways for generating acetyl-CoA;

250   the other is the CO dehydrogenase/acetyl-coenzyme A synthase that catalyzes

251   aceticlastic catabolism in *Methanosarcina* spp. The absence of the *acd* genes suggests

252   that the CO dehydrogenase/acetyl-coenzyme A synthase fulfills the function of both

253   enzymes in *M. barkeri.*

254   Among genes encoding biosynthetic functions a group of 14 sequential orfs

255   encode for predicted gas vesicles with highest identity to *Gvp*ANOFGJKLM (MbarA326-

256   339) in the haloarchaea, which includes the minimal gene set for expression of vesicle

257   in *Haloferax volcanii* (Offner and Pfeifer 1995). Although there are no reports of gas

258   vesicles in *M. barkeri* Fusaro, gas vesicles have been reported in another strain of

259   *M. barkeri*, FR-1, and in *Methanosarcina vacuolata*, which has 61% homology with the

260   type strain of *M. barkeri* (Archer and King 1984; Zhilina and Zavarzin 1979; Zhilina and

261   Zavarzin 1987). Interestingly, *M. barkeri* has three sequential copies of *Gvp*A that

262   encodes the ribs of the vesicle wall and influences the strength and width of the vesicles

263   (Beard et al. 2002). The 33.5 kb region that includes the *Gvp* operon may have been

264   acquired from vesicle synthesizing species as it is flanked by transposons. Gas

265   vesicles are observed occasionally in *M. barkeri* cells grown with H_2-CO_2 on solidified

266   medium. *M. barkeri* also has orfs (MbarA_0022 and MbarA_0023) with high identity to

267   two enzymes required for N-acetylmuramic acid synthesis, which is unique among the

268   sequenced Archaea. However, prior analysis of the cell wall compositon of *M. barkeri*

269   Fusaro failed to detect muramic acid (Kandler and Hippe 1977). *M. barkeri* also lacks a

270   low affinity phosphate transporter (MA2935) suggesting it originated in a phosphate-

271   poor environment. Two other transporters are missing in *M. barkeri*, a gluconate

272   transporter (MA0021) and a dicarboxylate transporter (MA2961). This suggests that *M.*

273   *barkeri* may have a lower ability to take up organic compounds than the other two.

274   Finally *M. acetivorans* and *M. mazei* have two copies of cheC (MA0012 and MA3065)

275   but *M. barkeri* does not have this protein. The role of this chemotaxis gene in

276   *Methanosarcina* spp. is currently unknown since motility has not been observed in these

277   species.

278   Another unique feature of the *M. barkeri* genome is the detection of a putative

279   operon encoding a bacterial P450-specific ferredoxin reductase (Mbar 1947-1945). The

280   family of heme protein monooxygenases known as cytochrome P450 plays a critical

281   role in the synthesis and degradation of many xenobiotics and physiologically important

282   compounds (Sono et al. 1966; Sono et al. 1996; Whitlock and Denison 1995). All

283   known P450s are multi-centre enzymes consisting of a heme, or P450, component with

284   associated reductase components. The gene encoding the putative cytochrome P450 in

285   *M. barkeri* is flanked immediately upstream by genes encoding a ferredoxin and

286   ferredoxin reductase, which is typical of bacterial class I three-component systems. For

287   catalytic activity, cytochrome P450 must be associated with the electron donor partner

288   proteins, ferredoxin/ferredoxin reductase complex (Takemori et al. 1993). Cytochrome

289   P450 has not been detected previously in the Archaea. Another putative operon

290   encoding oxygen dependent cytochome *d* oxidase, *cydAB*, was also identified in the

291   genome of *M. barkeri* and the other two methanosarcinal genomes. The presence of

292   these oxygen dependent genes along with one catalase and two superoxide dismutase

293   suggests that these proteins protect methanosarcinal species from oxygen or they may

294   support microaerophilic growth by a currently undescribed mechanism. As cytochrome

295   P450 catalyzes an oxygen requiring reaction and has not been detected previously in

296   an anaerobe, the detection of this gene in *M. barkeri* raises intriquing questions about

297   the function of this gene product in this obligately anaerobic methanogen.

298   **Plasmid structure and content**

299
300   The 36.4 kb plasmid in *M. barkeri* has not been detected previously. In contrast

301   to the smaller 5.4 kb plasmid pC2A in *M. acetivorans*, which appears to replicate by a

302   rolling-circle mechanism (Metcalf et al. 1997), the *M. barkeri* extrachromosomal element

303   lacks a putative *rep*A. Instead it has a cdc-6 homolog in a region of highly repetitive

304   sequence (discussed below), which suggests a novel mechanism of synchronous

305   replication. Interestingly, one of the extrachromosomal orfs (MbarB 3749) has 44%

306   sequence similarity to an ATPase associated with chromosomal partitioning. These

307   combined characteristics suggest that the extrachromosomal element replicates with

308   cell division. In addition to the putative cdc6 and partitioning protein, the orfs include 4

309   genes possibly associated with methanochondroitin synthesis, 7 hypothetical genes of

310   unknown function and 5 putative transposases. None of the orfs had equivalent

311   identities to orfs found in *M. acetivorans* and *M. mazei* genomes, but missing from the

312   *M. barkeri* genome that might have suggested a critical function for the

313   extrachromosomal element.

314 **Features revealed by whole genome comparison**

315
316   Whole genome distances (Table 2) based on maximal local alignments indicate

317 that the genomes are quite similar in overall content with *M. acetivorans* and *M. mazei*

318 marginally more closely related. This is in qualitative agreement with DNA-DNA

319 hybridization experiments (Sowers and Johnson 1984), which showed 28% homology

320 between *M. acetivorans* and *M. mazei* and 18% between these species and *M. barkeri.*

321 This result underscores the comparability of these sequences with the exception of the

322 plasmid sequence.

323 **Location of origins of replication**

324
325   In Archaea, origins of replication are invariably found in close proximity to the

326 origin recognition complex gene (*orc1*) sometimes also referred to as cell division

327 control protein 6 (*cdc6*) (Lopez et al. 1999). When genes are densely packed, searches

328 for putative origins of replication are directed at proximal intergenic regions. In the three

329 Methanosarcinae there are two highly conserved paralogous copies of these genes, in

330 relatively close mutual proximity (about 100 kb or 300 kb in *M. barkeri*) situated on

331 opposite strands and directed away from each other, a finding consistent with the

332 observation of Kelman *et al.* (Kelmana and Kelman 2004). Flanking downstream ORFs

333 are conserved (Table 3). The putative origins of replication are located in the upstream

334 intergenic regions of approximately 1600 nt (ORI A) or 800 nt (ORI B) in extent and are

335 somewhat conserved at the nucleotide level. In the chromosomal origin of replication

336 region, gene products are approximately 95% identical across all species. Non-coding

337 origin features (ORI A and ORI B) are not as well conserved (E <= 1e-44 in ORI A and

338   E <=1e-8 in ORI B) and show only weak similarity between ORI A and ORI B. They are

339   extremely AT rich (~70%) and may show unconserved inverted repeat structures.

340   A replication complex is initiated when orc1 (cdc6) protein binds to cognate DNA

341   at the origin and allows the recruitment of MCM and the rest of the replication

342   machinery. An approximate inverted repeat (Figure 3) could allow a pseudo-

343   symmetrical double hairpin to form a crucifix motif similar to a Holliday junction thereby

344   initiating bi-directional replication from a point, with complexes bypassing each other to

345   replicate the origin at the beginning of replication. The concurrent presence of more

346   than one active origin would cause contention for DNA, so there must be an implicit

347   mechanism to control which origin and which origin recognition complex protein is

348   dominant. It is notable that the downstream neighbor of the secondary orc1 B is a

349   highly conserved hsp60 class heatshock protein as this suggests a possible stress-

350   associated switching mechanism. The putative origins of replication are located

351   centrally within the most highly conserved and syntenous regions of the respective

352   genomes (Figure 4) consistent with the observation of Eisen et al. (Eisen et al. 2000) of

353   symmetrical inversion about the origin of replication. GC skew analysis (results not

354   shown) is not useful in this case as there is a high level of strand inversion and

355   rearrangement.

356   The plasmid of *M. barkeri* presents a unique and distinct origin of replication

357   characterized by an orc1 homolog (orc1 C) that is relatively weakly related to orc1 A

358   and orc1 B, (37% / 66% with orc1 A, 21% / 44% with orc1 B) (Table 3). The

359   immediately adjacent upstream region of the plasmid DNA contains a 5.6 kb non-coding

360   region (15.3% of the plasmid sequence) characterized by highly repetitive sequence

361   consisting of over 38 direct repeats of a 143 nt sequence with few variations between

362   them (Figure S-1, supplement). The consensus of the AT rich repeat sequence is:

363   ATCCCA<u>TTTCCTAAGCA</u>GAGAATTAG<u>TTTCCTAAG</u>CAAAAAAAAaGATTTCTGgcttagA

364   CCA<u>TTTCCTAAGCA</u>AAACGATATCAGAAGACATAACAAGTTAGAAGAaAAAtAAgTT

365   AAAATTAGATATTAATCTGTATATAT, with internal repeats underscored and variable

366   regions in lowercase. This sort of arrangement (ORI C) is quite unlike that of the

367   chromosomal ORI A and ORI B, but has the capacity to present slideable bubbled-out

368   complete repeat motifs and retain a quasi-stable structure (Figure 5).

369

370   **Overall genomic organization**

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

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

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

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

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

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

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

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

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

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

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

382   where interspecies gene similarities are as high as 95% (Table 3). However there is

383 little apparent conservation of gene organization in the region most distal to the origin

384 where large scale collinearity appears rare. The putative terminus of replication is

385 observed to be a hotspot for reorganization (Myllykallio et al. 2000). Two properties of

386 *M. barkeri* were measured: synteny, which measures the local paralog neighborhood

387 with respect to comparable genomes and intergenic interval or the separation between

388 successive genes which measures the relative content density. In the distal semi-

389 genome the rate of change of synteny is negative in accord with the macroscopic

390 observation of decreased collinearity, and the negatively correlated intergenic interval is

391 greater than average indicating a loss of gene content in this region (Table 2). What

392 might cause this wasteland effect? One possibility, given the symmetry with respect to

393 the origin, is an accumulation of strand exchange failures in the replication process, and

394 subsequent 'gene rot' of broken genes. The cross effect of random strand inversion

395 noted by Eisen et al. (Eisen et al. 2000) gives way to a shotgun effect. Another

396 possibility is infiltration by transposons with transposase mediated damage. Certainly

397 there is an increased frequency of transposon genes in this area (Figure 4, trace d), but

398 this may either be causative or opportunistic, with the organism tolerating infiltration of

399 already dysfunctional sections of the chromosome.

400

401 **CONCLUSIONS**

402

403   Of the 3680 open reading frames in *M. barkeri*, 678 had paralogs with better than

404 80% similarity to both *M. acetivorans* and *M. mazei* while 256 were unique (non-

405 paralogous) amongst these species. An etiology for genome rearrangement is revealed

406 by whole genome comparison of three species of the genus *Methanosarcina*. The

407   inverse correlation of intergenic size and synteny demonstrates a mechanism for the

408   development of genome plasticity, which involves replication associated inversion with

409   concommitant gene damage and colonization by transposon elements. Gene

410   duplication is also observed as a mechanism for genome extension. The organization

411   of *M. barkeri* is well conserved with respect to the other *Methanosarcinae* in the region

412   proximal to the origin of replication with interspecies gene similarities as high as 95%. In

413   the half genome most distant from the origin, it is however disordered and marked by

414   increased transposase frequency and decreased gene synteny and gene density.

415   Furthermore we have observed a highly conserved double origin of replication which

416   suggests a mechanism for replication which allows a double start with pass through

417   which enables the origin itself to be replicated. The apparent genome plasticity likely

418   contributed to these species ability to adapt to a broad range of environments as a

419   result of genome elongation and enrichment for favorable phenotypes.

420

421

422 **ACKNOWLEDGEMENTS**

423

434 **LITERATURE CITED**

435

436 Archer, D.B. and N.R. King. 1984. Isolation of gas vesicles from Methanosarcina

437   barkeri. *J. Gen. Microbiol.* **130:** 167-172.

438 Badger, J.H. and G.J. Olsen. 1999. CRITICA: Coding Region Identification Tool

439   Invoking Comparative Analysis. *Mol. Biol. Evol.* **16:** 512–524.

440 Bateman, A., E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, and E.L. Sonnhammer. 2000.

441   The Pfam protein families database. *Nucleic Acids Research* **28:** 263-266.

442 Beard, S.J., P.K. Hayes, F. Pfeifer, and A.E. Walsby. 2002. The sequence of the major

443   gas vesicle protein, GvpA, influences the width and strength of halobacterial gas

444   vesicles. *FEMS Microbiol. Lett.* **213:** 149-157.

445 Boccazzi, P., K.J. Zhang, and W.W. Metcalf. 2000. Generation of dominant selectable

446   markers for resistance to pseudomonic acid by cloning and mutagenesis of the

447   *ileS* gene from the archaeon *Methanosarcina barkeri* Fusaro. *J. Bacteriol.* **182:**

448   2611-2618.

449 Bomar, M. and K.W. Knoll, F. 1985. Fixation of molecular nitrogen by methanosarcina

450   barkeri. *FEMS Microbiol. Ecol.* **31:** 47-55.

451 Boone, D.R., W.B. Whitman, and Y. Koga. 2001. Genus *Methanosarcina*. In *Bergey's*

452   *Manual of Systematic Bacteriology* (eds. D.R. Boone and R.W. Castenholz), pp.

453   268-276. Springer, New York.

454 Conway de Macario, E., D.L. Maeder, and A.J.L. Macario. 2003. Breaking the mould:

455   archaea with all four chaperoning systems. *Biochemical and Biophysical*

456   *Research Communications* **301:** 811-812.

457   Delcher, A.L., D. Harmon, S. Kasif, O. White, and S.L. Salzberg. 1999. Improved

458   microbial gene identification with GLIMMER. *Nucleic Acids Research* **27:** 4636-

459   4641.

460   Deppenmeier, U., A. Johann, T. Hartsch, R. Merkl, R.A. Schmitz, R. Martinez-Arias, A.

461   Henne, A. Wiezer, S. Bäumer, C. Jacobi, H. Brüggemann, T. Lienard, A.

462   Christmann, M. Bömeke, S. Steckel, A. Bhattacharyya, A. Lykidis, R. Overbeek,

463   H.-P. Klenk, R.P. Gunsalus, H.J. Fritz, and G. Gottschalk. 2002. The genome of

464   *Methanosarcina mazei:* Evidence for lateral gene transfer between Bacteria and

465   Archaea. *J. Mol. Microbiol. Biotechnol.* **4:** 453-461.

466   Eisen, J., J. Heidelberg, O. White, and S. Salzberg. 2000. Evidence for symmetric

467   chromosomal inversions around the replication origin in bacteria. In *Genome*

468   *Biology*, pp. RESEARCH0011.

469   Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using

470   phred. II. Error probabilities. *Genome Research* **8:** 186-194.

471   Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated

472   sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8:**

473   175-185.

474   Galagan, J.E., C. Nusbaum, A. Roy, M.G. Endrizzi, P. Macdonald, W. FitzHugh, S.

475   Calvo, R. Engels, S. Smirnov, D. Atnoor, A. Brown, N. Allen, J. Naylor, N.

476   Stange-Thomann, K. DeArellano, R. Johnson, L. Linton, P. McEwan, K.

477   McKernan, J. Talamas, A. Tirrell, W.J. Ye, A. Zimmer, R.D. Barber, I. Cann, D.E.

478   Graham, D.A. Grahame, A.M. Guss, R. Hedderich, C. Ingram-Smith, H.C.

479   Kuettner, J.A. Krzycki, J.A. Leigh, W.X. Li, J.F. Liu, B. Mukhopadhyay, J.N.

480   Reeve, K. Smith, T.A. Springer, L.A. Umayam, O. White, R.H. White, E.C. de

481   Macario, J.G. Ferry, K.F. Jarrell, H. Jing, A.J.L. Macario, I. Paulsen, M. Pritchett,

482   K.R. Sowers, R.V. Swanson, S.H. Zinder, E. Lander, W.W. Metcalf, and B.

483   Birren. 2002. The genome of *Methanosarcina acetivorans* reveals extensive

484   metabolic and physiological diversity. *Genome Research* **12:** 532-542.

485   Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence

486   finishing. *Genome Research* **8:** 195-202.

487   Grigoriev, A. 1998. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids*

488   *Research* **26:** 2286–2290. .

489   Kandler, O. and H. Hippe. 1977. Lack of peptidoglycan in the cell walls of

490   Methanosarcina barkeri. *Arch. Microbiol.* **113:** 57-60.

491   Kelmana, L.M. and Z. Kelman. 2004. Multiple origins of replication in archaea *Trends in*

492   *Microbiology* **12** 399-430.

493   Kreisl, P. and O. Kandler. 1986. Chemical structure of the cell wall polymer

494   methanosarcina. *Syst. Appl. Microbiol.* **7:** 293-299.

495   Lobo, A.L. and S.H. Zinder. 1988. Diazotrophy and Nitrogenase Activity in the

496   Archaebacterium Methanosarcina barkeri 227. *Appl. Environ. Microbiol.* **54:**

497   1656-1661.

498   Lopez, P., H. Philippe, H. Myllykallio, and P. Forterre. 1999. Identification of putative

499   chromosomal origins of replication in Archaea. *Molecular Microbiology* **32:** 883-

500   886.

501   Lovley, D.R. and M.J. Klug. 1982. Intermediary metabolism of organic matter in the

502   sediments of a eutrophic lake. *Appl. Environ. Microbiol.* **43:** 552-560.

503   Markowitz, V., F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X.

504   Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N.

505   Ivanova, and N.C. Kyrpides. 2006. The Integrated Microbial Genomes (IMG)

506   System. *Nucleic Acids Research (special database issue)* **34:** D344-348.

507   Metcalf, W.W., J.K. Zhang, E. Apolinario, K.R. Sowers, and R.S. Wolfe. 1997. A genetic

508   system for Archaea of the genus *Methanosarcina*: Liposome-mediated

509   transformation and construction of shuttle vectors. *Proc Natl Acad Sci USA* **94:**

510   2626-2631.

511   Metcalf, W.W., J.K. Zhang, X. Shi, and R.S. Wolfe. 1996. Molecular, genetic, and

512   biochemical characterization of the serC gene of Methanosarcina barkeri fusaro.

513   *J Bacteriol* **178:** 5797-5802.

514   Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley,

515   P. Bork, P. Bucher, L. Cerutti, and e. al. 2005. InterPro, progress and status in

516   2005. *Nucleic Acids Research* **33:** D201–D205.

517   Myllykallio, H., P. Lopez, P. Lopez-Garcia, R. Heilig, W. Saurin, Y. Zivanovic, H.

518   Philippe, and P. Forterre. 2000. Bacterial Mode of Replication with Eukaryotic-

519   Like Machinery in a Hyperthermophilic Archaeon. *Science* **288:** 2212-2215.

520   Offner, S. and F. Pfeifer. 1995. Complementation studies with the gas vesicle-encoding

521   p-vac region of Halobacterium salinarium PHH1 reveal a regulatory role for the

522   p-gvpDE genes. *Mol. Microbiol.* **16:** 9-19.

523   Rother, M. and W.W. Metcalf. 2004. Anaerobic growth of Methanosarcina acetivorans

524   C2A on carbon monoxide: An unusual way of life for a methanogenic archaeon.

525   *Proceedings of the National Academy of Sciences of the United States of*

526   *America* **101:** 16929-16934.

527 Rother, M. and W.W. Metcalf. 2005. Genetic technologies for Archaea. *Curr Opin*

528   *Microbiol* **8:** 745-751.

529 Sono, M., M.P. Roach, E.D. Coulter, and J.H. Dawson. 1966. *Chem.Rev.* **96:** 2841-

530   2887.

531 Sono, M., M.P. Roach, E.D. Coulter, and J.H. Dawson. 1996. Heme-containing

532   oxygenases. *Chem. Rev.* **96:** 2841-2887.

533 Sowers, K.R. 1995. Growth of *Methanosarcina* spp. as single cells. In *Archaea: A*

534   *Laboratory Manual* (eds. F.T. Robb K.R. Sowers S. DasSharma A.R. Place H.J.

535   Schreier, and E.M. Fleischmann), pp. 61-62. Cold Spring Harbor Laboratory

536   Press, Cold Spring Harbor.

537 Sowers, K.R. 2004. Methanogenesis. In *The Desk Encylopedia of Microbiology* (ed. M.

538   Schaechter), pp. 659-679. Elsevier Academic Press, San Diego.

539 Sowers, K.R., J.E. Boone, and R.P. Gunsalus. 1993. Disaggregation of *Methanosarcina*

540   spp. and growth as single cells at elevated osmolarity. *Appl. Environ. Microbiol.*

541   **59:** 3832-3839.

542 Sowers, K.R. and J.G. Ferry. 1983. Isolation and characterization of a methylotrophic

543   marine methanogen, Methanococcoides methylutens gen. nov., sp. nov. *Appl.*

544   *Environ. Microbiol.* **45:** 684-690.

545 Sowers, K.R. and J.L.F. Johnson, J. G. 1984. Phylogenetic relationships among the

546   methylotrophic methane-producing bacteria and emendation of the

547   familyMethanosarcinaceae. *Int. J. Syst. Bacteriol.* **34:** 444-450.

548   Takemori, S., T. Yamazaki, and S. Ikushiro. 1993. Evolution and differentiation of P450

549   genes. In *Cytochrome P450* (ed. T. Omura, Ishimura, Y., and Fujii-Kuriyama, Y.).

550   VCH Publishers, Inc., New York.

551   Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M.

552   Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, and e. al. 2003. The

553   COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4:**

554   41.

555   Welander, P.V. and W.W. Metcalf. 2005. Loss of the mtr operon in Methanosarcina

556   blocks growth on methanol, but not methanogenesis, and reveals an unknown

557   methanogenic pathway. *PNAS* **102:** 10664-10669.

558   Whitlock, J.P.J. and M.S. Denison. 1995. In dusction of cytochrome P450 enzyme that

559   metabolize xenobiotics. In *Cytochrome P450: Structure, Mechanisms and*

560   *Biochemistry* (ed. P.R. Ortiz de Montellano), pp. 367-390. Plenum Press, New

561   York.

562   Zhilina, T.N. and G.A. Zavarzin. 1979. Comparative cytology of methanosarcinae and

563   description of Methanosarcina vacuolata sp. nova. *Microbiology* **48:** 279-285.

564   Zhilina, T.N. and G.A. Zavarzin. 1987. *Methanosarcina vacuolata* sp. nov., a vacuolated

565   Methanosarcina. *Int. J. Syst. Bacteriol.* **37:** 281-283.

566

567
568 **Table 1.** Comparison of Genome Features Among *Methanosarcina* spp.
569
570
571

Organism	*Methanosarcina acetivorans*	*Methanosarcina mazei*	*Methanosarcina barkeri fusaro*	*Methanosarcina barkeri fusaro*	*Methanosarcina barkeri fusaro*
tax_id	188937	192952	269797	269797	269797
Accession	NC_003552	NC_003901	NC_007355	NC_007349	
designation	chromosome	chromosome	chromosome	plasmid	genome
length	5751494	4096345	4837408	36358	4873766
G+C%	42.7%	41.5%	39.2%	33.6%	39.2%
feature count	4540	3371	3680	18	3698
features length	4262934	3074712	3390164	21099	3411263
features coverage	74%	75%	70%	58%	69%
featureless	26%	25%	30%	42%	31%
nt/feature mean	939	912	921	1172	922

572

573   **Table 2.** Intergenomic distances calculated using eq 1 based on whole genome
574   maximal local nucleotide sequence identity considering only HSPs with identity > 67%
575   (lower left) or 55% (upper right).

576

	M. acetivorans	*M. mazei*	*M. barkeri*	*M. barkeri* plasmid
M. acetivorans	0	0.489	0.487	0.570
M. mazei	0.517	0	0.480	0.591
M. barkeri	0.552	0.569	0	0.512
M. barkeri plasmid	0.881	0.883	0.836	0

577

578

579   **Table 3.** Identification of conserved features for chromosomal origins of replication in
580   *Methanosarcina* spp.

581
582
583

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

584

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.

609   **Figure 1.** Thin-section electron micrograph of *M. barkeri* Fusaro showing typical

610   morphology consisting of multicellular aggregates embedded in a methanochondroitin

611   matrix. Vacuole-like structures appear to be membrane bound. Bars: A, 1.0 μm, B, 0.2

612   μm.

613

614
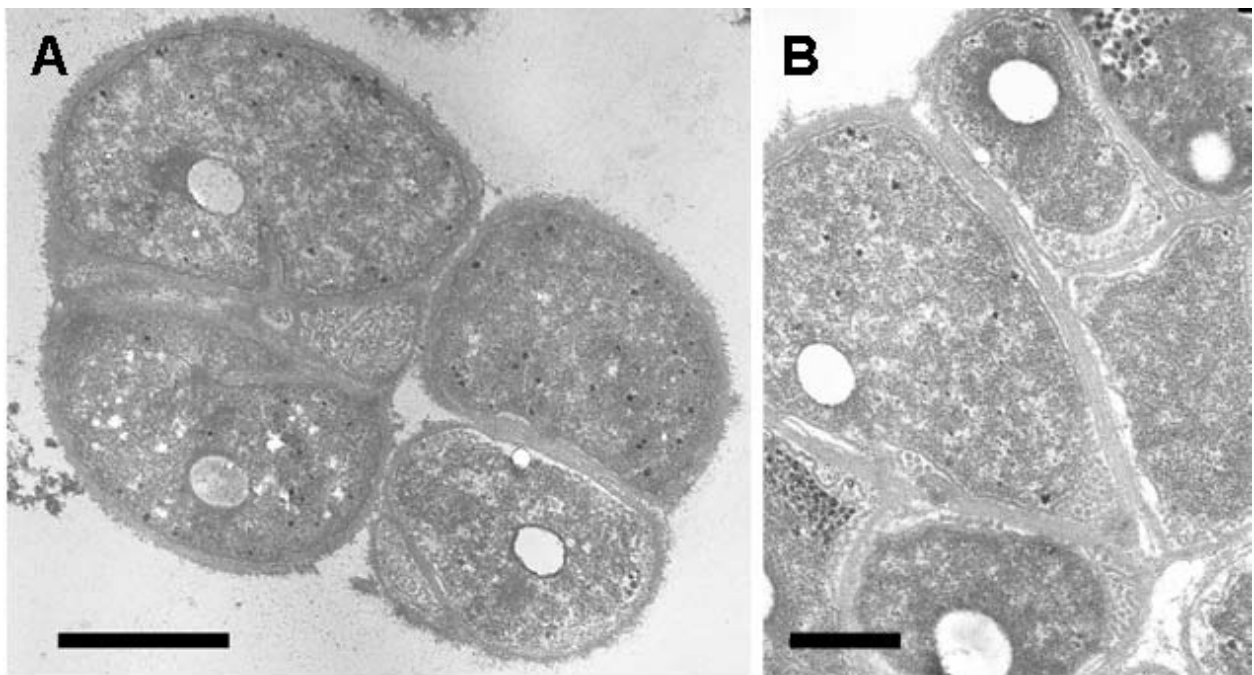
615

616

617
618
619

620 **Figure 2.** GRIT database schema. Primary keys are capitalized, foreign keys are
621 underlined. Arrows indicate foreign key primary key relationships.
622
623
624
625
626
627
628
629
630
631
632
633
634
635 .
636
637
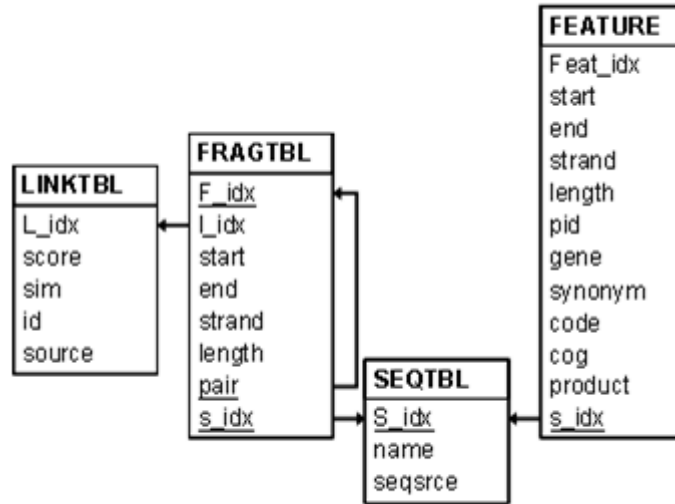
638   **Figure 3.** The *M. barkeri* ORI A self-complementary region blastn alignment
639
640
641
642   ```
643   Query: 709 CAGAAAATGTAA-ATTTCTCAGAAC-A-TGTAATTTAGATTTCT-CAATTT-TTT-GAAA 656
644 || |||||| || | ||| |||||| | |||| ||| |||| | ||| ||| |||
645   Sbjct: 1190360 CAAAAAAT-TATGAGTTC-CAGAACCAATGTAGGTTAAATTTAAACCCTTTCTTTCGAAT 1190417
646
647   Query: 655 ATCATACTTTTTCTGA-T-AGATTGAGTATCA-ATAAAAACTCAAAATAAAAATATTCAA 599
648 ||| || | | | | | | | | | | |||||| |||||||||||| || ||
649   Sbjct: 1190418 GGAATAAGTTATGAAAATTATAAT-ATTTTTACATAAAATTAAAAAAATAAAAATTTT-AA 1190475
650
651   Query: 598 TGAAAATATTA-AATCA 583
652 || || |||| ||| |
653   Sbjct: 1190476 -GATAAAATTAGAATTA 1190491
654
655   -- Unaligned region of ~440 nt --
656
657   Query: 141 TAATTCTAATTTTATC-TTAAA-ATTTTTATTTTTTAATTTTATGTAAAAAT-ATTATAA 85
658 | ||| |||| || || || || ||||||||||| |||||| | | | | | || ||
659   Sbjct: 1190933 TGATT-TAATATTTTCATTGAATATTTTTATTTTGAGTTTTTAT-TGATACTCAATCTA- 1190989
660
661   Query: 84 TTTTCATAACTTATTCCATTCGAAAGAAAGGGTTTAAATTTAACCTACATTGGTTCTG-G 26
662 | | || | || ||| ||| ||| | |||| ||| ||||| | |||||| |
663   Sbjct: 1190990 TCAGAAAAAGT-ATGATTTTC-AAA-AAATTGAG-AAATCTAAATTACAT-G-TTCTGAG 1191043
664
665   Query: 25 AACTCATA-ATTTTTTG 10
666 || | || |||||| ||
667   Sbjct: 1191044 AAATT-TACATTTTCTG 1191059
668
669
670   ```

671
672   **Figure 4.** Asymmetric fragmentation in *M. barkeri*. The top panel shows cumulative
673   deviations from the mean in *M. barkeri* genome for synteny (SI) with respect to
674   *M. acetivorans* (a), *M. mazei* (b), or intergenic interval (c). The cumulative transposon
675   count is superimposed (d). The bottom panel shows uniformly scaled blastn cross plots
676   of *M. barkeri* chromosome with *M. mazei* and *M. acetivorans* with the origin regions
677   circled.
678

679

680
681
682 **Figure 5.** Proposed mechanism for conserved repetitive sequence to provide bubbled-
683 out repeat motifs for initiation of replication. Pairs of quasi-stable bubbles might occur in
684 pairs at arbitrary locations on opposite strands. All motifs are nearly identical.
685
686
687

688 **Figure 6**. *M. acetivorans* is elongated due to distributed gene duplication events.
689
690
692
694

699
700
701

SUPPLEMENT

702
703
704 **Figure S-1.** Plasmid DNA origin of replication containing a 5.6 kb non-coding region
705 characterized by highly repetitive sequence consisting of over 38 direct repeats of a
706 143nt sequence with few variations between them.
707
708
709

```
atcccacttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagataaagaaattaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagataaagaaattaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaa-gatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaaagaagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagataaagaaattaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagataaagaaattaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagaagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcgagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagataaagaaattaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagataaagaaattaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagataaagaaattaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaa--gatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaa-gatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaa-gatttctggcttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaa[━━━━━━━━━━━━━━━━━━━━━━━━━━━]cgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattagtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacataacaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattggtttcctaagcaaaaaaaagatttctggtttagaccatttcctaagcaaacgatatcagaagacatagcaagttagaagaaaataagttaaaattagatattaatctgtatatat
atcccatttcctaagcagagaattggtttcctaagc[━━━━━━━━━━━━━━━━━━━━━━━]ctaagcaaa
```

749
750
751

752
753